

🚀⚡️ המאמר הימי של מיק 30.05.2024 ⚡️🚀 2BP: 2-Stage Backpropagation

אנו יודעים שהמודלים העמוקים גדולים היום כדי להיכנס לזיכרון ram של סעפ אחד. עקב לכך מחלקים את משקל המודל בין הדשנים השונים (sharding). זה פותר צוואר בקבוק אחד (זיכרון) אבל כתוצאה לכך נדרש בקבוק אחר בחישוב של backprop, המאמר הנסקר פיתח שיטה למצבי backprop את חישוב הגרדיאנטים במהלך backprop ובכך מקל על צוואר הבקבוק הזה.

מאמר: <https://arxiv.org/pdf/2405.18047.pdf>

🚀⚡️ המאמר הימי של מיק 31.05.24 ⚡️🚀 Transformers Can Do Arithmetic with the Right Embeddings

אנו יודעים שמודלי שפה גדולים לא מצטינים בלחשב ביטויים מתמטיים בטח כלו המכילים מספרים עם הרבה ספרות. גם אם מאמנים אותם על מילוני דוגמאות עדין מסתובבים להכליל אותם למספרים גדולים. המאמר מציע להוסיף positional encoding למספרים שמרתempם לספק למודל שפה מרחק של כל ספרה מתחילה המספר. וזה עובד לא רע.

רפו: <https://github.com/mcleish7/arithmetic>
מאמר: <https://arxiv.org/abs/2405.17399>

🚀⚡️ המאמר הימי של מיק 01.06.24 ⚡️🚀 The Evolution of Multimodal Model Architectures

אתם יודעים שאין אוחב לכתוב סקירות אבל בד"כ אני סוקר מאמר אחד. כאן יש לכם סקירה של תחום שלם שהוא מודלים מולטי-מודליים כלומר אליו שיודעים "לטפל" בסוגי DATA טוונות, אודיו ו視频ה. המאמר נותן סקירה היסטורית על ארכיטקטורות של מודלים מולטי-מודליים ומהלך אותם ל 4 קטגוריות רחבות שמתחולקות לתת-קטגוריות כמפורט. מאמר שיכל לעשות לכם קצת סדר בתחום המגןיב הזה.

טלגרם: <https://t.me/MathyAlwithMike/60>
טוויטר: https://x.com/MikeE_3_14/status/1796823310459666491
מאמר: <https://arxiv.org/abs/2405.17927>

🚀⚡️ המאמר הימי של מיק 02.06.24 ⚡️🚀 LLaMA-NAS: Efficient Neural Architecture Search for Large Language Models

פעם הנושא של Neural Architecture Search או NAS בקצרה שעסוק בחיפוש לאחר ארכיטקטורה אופטימלית של רשת נירונים עבור משימה/משימות/דומין היה די פופולרי אך בשנים האחרונות התחום נמצא בדיעיכה. אני שמח שנטקלהתי במאמר זהה שמנסה לפתח NAS עבור מודלי שפה. אני זוכר מאמרים ד' מגניבים שימושיים בשיטות RL ד' מגניבות לך. אולי בעתיד NAS תהופיע למתחרה רציניות של שיטות פרונינג וكونטיזיה.

מאמר: <https://arxiv.org/abs/2405.18377>
טלגרם: <https://t.me/MathyAlwithMike/69>



המאמר היום של מיק 03.06.24: Better & Faster Large Language Models via Multi-token Prediction

אתם בטח שידועים אנו רגילים לאמן מודל שפה גנרטיביים באמצעות חיזוי טוקן הבא בהינתם הטוקנים הקודמים (הקשר או קונטיקסט). המאמר הזה (שקיים די הרבה זו כשיצא) מציע לחזות כמה טוקנים עוקבים בו בזמןית בהינתן הקשר. המחברים הראו שזה יכול לשפר את ביצוע המודול - זה לא מפתיע (لتחשושתי) כי משימת חיזוי טוקנים מרובים דורשת מהמודול הבנה יותר עמוקה של השפה. השיטה גם עשויה לתרום להאצת זמן ריצה והרוחים גדלים עם גודל המודול.

מאמר: <https://arxiv.org/pdf/2404.19737.pdf>

טלגרם: <https://t.me/MathyAlwithMike/69>



המאמר היום של מיק 04.06.24: Are Emergent Abilities of Large Language Models a Mirage?

היום המאמר שנסקור הוא מלפני שנה בערך והוא משך את תשומת לבי בגל שהוא חוקר מה שנគרא emergent capabilities של מודלי שפה - לומר יכלתם ללמידה משימות חדשות. המאמר בוחן האם למודלי שפה אכן יש יכולת ללמידה משימות שהם אומנו עליהם בצורה מפורשת (פחות או יותר) או שזו אשלייה הנובעת מאייר שהוא מודד את יכולות האלו.

מאמר: <https://arxiv.org/abs/2304.15004>

טלגרם: <https://t.me/MathyAlwithMike/76>



GraphAny: A Foundation Model for Node Classification on Any Graph

כיצד לפתח מודלים foundational בתחום הגרפים?

מודלי שפה foundational שינו בצורה משמעותית את האופן שלנו בונים מודלים בתחום קוח: בהרבה מקרים הםאפשרים פיתוח מהיר יותר (פיניטון וכאל). מרחיב קלט משותף לכל המשימות (טוקנים) הוא מרכיב חיוני שדרכו LLMs foundational מגלמים יכולת הכללה שמאפשרת התאמתם היחסית לא מרכיבת למוגן משימות NLP.

לצערנו לגרפים אין תכמה מסווגת כמו טוקנים, כי כל גרף לרוב מאופין על ידי סמנטיקה מסוימת מחייבים לייבלים, דבר שמנע את פיתוח המודלים foundational של הגרפים. האם ניתן להתגבר על זה? יש לנו התשלה: המחברים מציעים GraphAny, ארכיטקטורה foundational לביצוע משימת סיווג קודקודים בgraf. המודול יכול להכליל לגרף חדש כלשהו עם מרחבי מאפיינים וליibiliים שריוןתיים, שונים בדרך כלל מآلה של הגרף שאימנו עליו.

מאמר: <https://arxiv.org/abs/2405.2044>

טלגרם: <https://t.me/MathyAlwithMike/78>



המאמר היום של מיק 06.06.24: Similarity is Not All You Need: Endowing Retrieval-Augmented Generation with Multi-layered Thoughts

בזמן האחרון גישות המשלבות מודלי שפה עם בסיסי נתונים חיצוניים הפכו למאוד פופולריים. גישות אלו לרוב שייכות למשפחה Retrieval Augmented Generation או RAG בקצרה. בגין בהינתן מודל שפה ומסמכים העשויים להכיל תשובה על שאלת משתמש, RAG קודם מחפש כמה מסמכים הרלוונטיים ביותר לשאלה ואז מזינה אותם יחד עם השאלה למודל שפה. המודל מרכיב את תשובתו על השאלה בהתאם שהזינו אליו.

אבל איך נבחר מסמכים הרלוונטיים יותר לשאלה? בדרך כלל בוחרים אותם לפי הקربה של האמבדינג (= "יצוג וקטורי") שלו לאמבדינג של השאלה. בדרך כלל הממציאות טיפה יותר מורכבת ממה שתיארתי: למשל אם המסמכים ארוכים צריך לחלק אותם לצ'אנקים איז הבחירה היא לפי דמיון האמבדינג של הצ'אנקים זהה לשאלה. כמובן שיש עוד גישות. הדמיון בין אמבדינגים בד"כ מחושב לפי דמיון קוסין (זווית בין הוקטורים). האם הבחירה הזאת היא אופטימלית - זו השאלה שהמאמר שנסקור היום מנסה לענות עליה.

כדי להבין האם הבחירה אופטימלית צריך להגיד מدد אופטימליות. הרעיון בסופו של דבר מטרתנו היא לחת תשובה נכונה לשאלת המשתמש. המאמר טוען שבבחירה מסמכים הרלוונטיים לפי דמיון אמבדינגים אינם אופטימלי בהתאם המدد הזה. אך המחברים מציעים גישה לשכלול הבחירה של המסמכים הרלוונטיים לשאלה. האמת הם מציעים שהוא די טבעי - בגין המטרה שלהם היא לאਪטם את הביצועים של RAG דרך "מקסום הסיכוי" לקבלת תשובה טובה אחרי בחירת מסמכים רלוונטיים על ידי RAG". המחברים מנוטים להשיג את המטרה בכמה שלבים:

שלב 1: אימון מודל utility. המטרה של מודל זה להעניק ציון ליכולת של מסמך נתון "لتת' תשובה טובה לשאלה כאשר הם (המסמך והשאלה) מוזנים למודל שפה ייחד. אבל איך נדע לשערר את איכות התשובה? בשביל זה המחברים לקחו מודל שפה חזק (גיגי4 gpt4) שמטרתו היא לתת ציון לתשובה עברו מסמך ושאלתה נתונים (כלכל שהתשובה טובה ציון גבוה יותר). המאמר לא מסביר איך זה נעשה אבל אני מניח שעבור דاطהסת המכיל תשיבות ניתן למדוד דמיון סמנטי בין תשובה אמיתי לתשובה מופקת על ידי Who (כלומר בין האמבדינגים), ניתן גם למדוד אותה על ידי הצעתם של המסמך, השאלה והתשובה -
- ועודידת נראות מירבית שלה (כלומר digits), בטח יש עוד שיטות. המחברים מאמנים model utility (שהוא מודל קל יחסית) להציג את אותה ההתפלגות של ציוני מסמכים (בהינתן שאלה) כמו המודל החזק. ככלומר ממדועים divergence KL בין התפלגות ציוניים של model utility לבין זו של מודל השפה (שהוא מוקפא - לא מאומן).

שלב 2: בחירת מסמכים עבור שאלה נתונה בוחרים רק מסמכים שיש להם ציון דמיון או ציון של model utility מוסף (בין k הגבוהים ביותר כל אחד).

שלב 3: אימון מודל תמצות מסמכים. המחברים טוענים שב"כ המסמכים שנבחרים מכילים לא מעט מידע לא רלוונטי לשאלה שמקשה על מודל שפה לחת תשובה טובה וגם מעלה עליות (צריכים להכניס הרובה טוקנים ל-LLM). בגין להתמודד עם הקושי הזה המחברים מציעים לאמן מודל שבהינתן שאלה מפיק מהמסמכים שנבחרו את המידע הרלוונטי לשאלה. זה נעשה ב-2 שלבים: בשלב הראשון עברו דאטהסת של שאלות והמסמכים הרלוונטיים מתשאלים מודל שפה חזק (gpt4) לתמצאת את המסמכים האלו (בעור שאלה נתונה). על הדאטהסת эта שאלה, מסמכים ותמצית) עושים פיניטון של מודל שפה לא כבד עם LoRa כמובן - כלומר עושים זאת בשלב השני עושים RLHF עם DPO כמו שמקובל היום. בשביל Fine-Tuning Supervised SFT או RLHF מודל שפה(האם לא מפרטים יותר מדי כאן) בונים דאטהסת של תשבות נכונות ולא נכונות בהינתן שאלה ותמצית מסמכים. בניית פונקציית תגמול (reward) מתבצעת בדיקן כמו ב- DPO הסטנדרטי.

אחריו שסימנו לאמן את מודל התמצות, ההיסק (אנפראנס) געשה בצורה מאוד טبيعית. לוקחים שאלה, מפיקים את המסתכים הרלוונטיים משלב 1, מתחמיצים אותם עם המודל משלב 3 ואז מזינים אותם לעוד מודל שפה (המחברים לא מפרטים עליו אבל מצינים שניתן לכיל אותו על דאטהסט כלשהו של שאלות ותשובות). והמודל מספק לנו את התשובה...

🚀⚡️ המאמר היום של מייק 24.06.07: Scaling and evaluating sparse autoencoders?

המאמר זהה של openai ממשיר את הקו המחקרי של antropic (<https://www.anthropic.com/news/mapping-mind-language-model>) המנסה לראות איך ניתן למצוא של קונספטיים (מסלולים ויזואליים) בתוך נירוניים של מודלי שפה מאומנים. המאמר של אנטרופיק בגודל טווען שיש נירוניים הנדלקים על קונספטיים (נגיד גשר הזהב) מסוימים ויש כאלו שמהווים ערבות של קונספטיים.

אבל איך ניתן להציג את קונספט באמצעות וקטור? מתרברר שניתן להציג כל קונספט באמצעות וקטור ארוך אך מאד דليل(sparse). אך נירוניים המהווים ערבות של קונספטיים ניתנת להציג בתור סכום משוקלל של וקטורים דילילים אלו אחורי שטטילים את הסכום על מרחב האמביד'ג המקורי של הטרנספורמה.

הוקטוריהם הדיליליים המתאימים לקונספטיים ניתנת להפיק באמצעות אימון של sparse autoencoder של שכבה אחת לכל כיוון כאשר הייצוג באמצעות (אחורי האנקודר) הוא וקטור דיליל: בmphלך האימון לוקחים ממנו את K הרכיבים הגדולים ביותר - אחורי ReLU

יש כמובן חוק Scaling מעניינים לגבי הייצוגים האלה. מאמר מעניין:
<https://cdn.openai.com/papers/sparse-autoencoders.pdf>

🚀⚡️ המאמר היום של מייק 24.06.08: Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality

למאמר זהה יש עוד שם והוא mamba-2. הוא מתמקד בשכלול הארכיטקטורה של מבנה המקורי שעשתה הרבה כותרות בחצי השנה האחרונות ואני הצלפתי לחגיה וסקרתי בערך 20 מאמרים בנושא המրתק זהה.

המאמר זהה של Albert Gu התוחח ממשיר להעшир את עולם המבנה והפעם הוא הגיע לכמה תובנות די מעניינות. הוא לראשונה מגדר SSM בעל תוכנה N-semi-separable של מעשה מגדר את צורתו של קרמל. קוונבולוציה המופעל על סדרת הקלט במודול הקונבולוציוני של SSM (כasher משתמשים ב-SSM לאימון ממוקבל). אלחש לכם בסוד שבסוףו של דבר זה מתנקז לצורתו של מטריצה A.

שניית מאמר חוקר מנגןוני attention בפרט השונים למשל הקלסטי הלינארי, כלומר ללא סופטמקס, ועם סדר שונה ביצוע פעולות בין מטריצות K, Q, V. המאמר מפרק את החישוב ל- 3 שלבים "אטומיים" (שכל אחת מהם הוא מכפלות מטריצות, אך לעיתים מאוד גדולות) השלב השני והחשוב ביותר הוא מיסוך (masking) שניתן לתאר אותו גם עלי ידי מכפלות מטריצות (Kernel trick). המיסוך הקוזזי (causal) הוא חלק ממנגןון ה-SSMs. הבדיקה זו אפשרה למחררים להוכיח סוג של שקילות בין מנגןוני חסויים ל-SSMs.

בנוסף הם מפתח שיטה לחישוב יעיל של קונבולוציה ארכוֹה (shape of the SSM) בחומרה עם מטריצות הנקרואט semi-separable-1 (מעבר למטריצה A מסוימת).

מה יצא לנו מכל הסיפור זהה? האצת אימון של מבנה (shape) למשה מובה (2) וגם פרימיוםרק תיאורטי למידול ארכיטקטורה העצמתית זו משותפת גם למנגוני-h-attention השונים.

קריאה מהנה!

<https://arxiv.org/abs/2405.21060>

⚡🚀 המאמר היום של מיק 09.06.24

What Do Language Models Learn in Context? The Structured Task Hypothesis.

המאמר הזה תפס את עיני כי הוא מנסה לפתרו את תעלומת context Learning in או ICL. היכולת של מודל שפה לבצע משימות שלא אומנו עליו באופן מפורש על לאחר הצגה של כמה דוגמאות (שאלה, תשובה) היא לא פחות ממדעית וудין אין תשובה חד משמעותה המסבירה מה אכן קורה שם.

המאמר בוחן 3 הסברים אפשריים ל-ICL:

1. מודל שפה אשכלה "מצאה" את המשימה מכמה דוגמאות ומבצע אותה לפורמט נתון
2. המודל לומד במהלך אימון מקדים (pre-training) לעשות meta-learning כלומר לומד למודד את המשימה מכמה דוגמאות שניתנו לו
3. המודל לומד לייצג משימה חדשה כ"שילוב" של כמה משימות שלמד במהלך אימון מקדים

המחברים מוכחים שההשערות 1 ו-2 לא מתקינות ולא משאיר הרבה אפשריות...

<https://arxiv.org/abs/2406.04216>

⚡🚀 המאמר היום של מיק 10.06.24

Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks

אחד התופעות המרתיקות בלמידה היא גראוקינג - שהוא מעבר "فتאומי" של רשותה עמוקות למצב של הכללה מהמצב של overfitting למשל אחרי אימון מאד ארוך. הרעיון שאמם עבר דאטאטס נתון ורשת עומקה בעלת יכולת ייצוג גבואה מספיק (representativeness) אחרי שלב מסוים באימון אנו נגיע ל-overfitting כלומר למצב שבו ביצועי המודול יילכו ישתפרו עבור סט האימון המקורי אולם הביצועים על סט הולידייז יספגו ירידת ביצועים.

מה שמניב ומפתיע בגרוקינג שעבור אימון ארוך מספיק מגע המצב שביצועי המודול על סט הולידייז מתחילה לעלות יחד עם אלה על סט האימון ככלומר המודול מגע לשלב של הכללה אמיינית. מעניין שתופעה דומה מתרחשת בתנאים מסוימים אם אנו מגדילים את קיבולת המודול (מספר הפרמטרים) כאשר גודל הדאטאטס ומשך האימון נותרים קבועים) וגם כאשר אנו מגדילים את גודל הדאטאטס תוך שמירה של מושך האימון קיבולת המודול קבועים.

למעשה תופעות אלו שייכות למשפחת double descent (יש גם descent multiple) שנחקרה רבות על ידי חוקר דגול מישא בלקון. התופעה עצמה נתגלתה לפני יותר מ-30 שנה (מי שרצה להתעמק בנושא תעקבו אחריו <https://www.linkedin.com/in/charlesmartin14>).

אוקי', אז מה עשה המאמר הנסקר? הוא חקר תופעת גראונינג כאשר מתרחשת אם מגדים את מספר המשימות (כל משימה היא סוג של רגסיה לינארית בשדה המודולו (שארית)) שבעורן אנו מאמנים את המודל (כਮובן לךו) מודל שפה). מתרבר כי יש כמה מושגים (מודים) של יכולת הכללה של המודל כאשר משחקים עם היחס של מספר הדוגמאות פר משימה ועם מספר המשימה. בגדול מאוד אם נתונים מספיק ממשימות גדול מספיק ומספר דוגמאות פר משימה גדול מספיק אך מוגעים להכללה אמיתית כאשר המודל אכן לומד את המשימה במלואה).

<https://arxiv.org/abs/2406.02550>
קריאה מהנה!

🚀⚡️: 11.06.24. המאמר היומי של מיק

The Geometry of Categorical and Hierarchical Concepts in Large Language Models

המאמר חוקר כיצד קונספטים ומושגים מקודדים במרחבוי הייצוג (embeddings) של מודלים של שפה גדולים. הכותבים חוקרים 2 שאלות מרכזיות: הייצוג של קונספטים קטגוריים והקידוד של יחסים היררכיים בין קונספטים.

הם מרחיבים את ההסתכלות הלינארית הרגילה על הקונספטים כדי להראות שהקונספטים קטגוריים מיוצגים כסימפלקסים, קונספטים היררכיים הם אורתוגונליים, וקונספטים מורכבים מיוצגים כפוליטופים שנבנו מסכומים ישירים של סימפלקסים.

המחקר בוחן 957 קונספטים היררכיים עם נתונים מ-WordNet באמצעות מודל ג'מה. הכותבים מראים שקונספטים סמנטיים high-level יכולים להיות מנוטרים ומנולים על ידי מדידה וערכה ישירה של הייצוגים הווקטוריים הפנימיים של ה-LLMs. התוצאות התיאורתיות מגלות מבנה פשוט שבו קונספטים קטגוריים מיוצגים גיאומטרית כסימפלקסים ומושגים היררכיים מקודדים אורתוגונליות.

<https://arxiv.org/pdf/2406.01506.pdf>

🚀⚡️: 12.06.24. המאמר היומי של מיק

Accelerating Feedforward Computation via Parallel Nonlinear Equation Solving

היום סוקרים קצירות מאמר עתיק (מלפני 3 שנים) אבל יש למאמר זהה אימפקט גדול (רק תמשיכו לעקב אחריו הסקירות היומיות). שימושתcls על שם המאמר הזה לא קל לקשר אותו למידה עמוקה. הרוי מה לפתרון מסוואות לא לינאריות ולמידה עמוקה? אולי מילה Parallel עשויה לرمוז לנו קלות על אייזהו קשר למידה עמוקה כי אנחנו מאד אוהבים לחשב דברים במקביל במהלך אימון או נספנס של המודלים העמוקים שלנו.

אוקי', זה כן קשור ותיק נבון למה. קודם כל נרענן טיפה את זכרונו על שיטות איטרטיביות לפתורן של מערכות מסוואות כמו שיטת Jacobi או Shitz(GS)-Seidel. שיטות אלו ניתן להפעיל גם במערכות מסוואות לינאריות ולא לינאריות כאחד. בכל שיטה מתחילה מניחוש אקראי לפתורן וمعدכנים אותו על ידי חישוב איטרטיבי עד התכנסות (שצריך כמובן להגדיר) על ידי עדכון וקטור הפתרון רכיב-רכיב. ד"א בשיטת יעקובי ניתן לעדכן את כל הרכיבים בצורה מקבילית ולעומת זאת GS פחות ניתן למקובל.

אבל איך כל זה קשור למודלים עמוקים? מתרבר שתהילר האינפרנס במודלי שפה (ונתמקד בהם למרות שהמאמר לא מගביל את עצמו אליהם אלא מדבר על מודלים אוטוגראטיביים כלילים) ניתן להציג על ידי מערכת מסוואות אשר כל מסוואה בעצם "בוחרת" את הטוון בעל נראות הגבואה ביותר בהינתן הטוקנים הקודמים. כלומר כל מסוואה מכילה פונקציית argmax על מרחב הטוקנים.

ב"כ האינפראנס מתבצע בצורה אוטורגסיבית כלומר טוקן אחריו טוקן זהה מבון מאט את מהירות האינפראנס. אנו מתחילה בסדרת טוקנים אקראי וממשיכים לעדכן אותה בצורה איטרטיבית עד ההתכנסות. מתרבר שבסוואות שלוב של שיטת יאקובי - GS ניתן לזרז את החיזוי.

<https://www.arxiv.org/pdf/2002.03629>

⚡️⚡️ המאמר היום של מיק 13.06.24

Break the Sequential Dependency of LLM Inference Using LOOKAHEAD DECODING

זכרים את המאמר שסקרנו קצרות אטמול שהציג גישה איטרטיבית לפתרון מקבילי של מערכות משוואות לא לינאריות. אחת הדוגמאות של פתרון מערכות משוואות كالו היא גנרט טקסט ממודלי שפה כאשר כל טוקן נבחר בתור argmax של התפלגות הטוקן בהינתן הטוקנים הקודמים (המופק באמצעות השכבה الأخيرة של מודל השפה).

יש בגודל שתי שיטות איטרטיביות שנitin לרמות אחרות דוגימה עיליה יותר ממודלי שפה: יאקובי וגאוס-סידל. שתי השיטות מתחילות מניחוש אקראי של כמה טוקנים בהינתן ההקשר ואז מאפטמים אותם על פתרון איטרטיבי של מערכת המשוואות עם argmax (שקלול לגנרט). אפשר די בקהלות לראות שבגלל שימוש המשוואות הן אוטורגסיביות שיטות אלו לא יכולות להתכנס ביותר מ ϵ איטרציות (מספר הטוקנים הנדגמים עם שיטה) ולפעמים אפשר להספיק פחות (נציין כי כל איטרציה דורשת קצת יותר משבבי החישוב).

הבעיה עם השימוש הנאבי בשיטה הוא שהרוווח המוצע על פני דוגימה אוטורגסיבית סטנדרטיבית ממודלי שפה הוא לא גדול ועומד על פחות מ 1.1 האצת קצב גנרט.

המאמר מציע שכלול לשיטה הנאבי למציע לשמר בזיכרון את הטוקנים של כמה איטרציות האחרונות. במקרה אם והיא מוצאת בזיכרון זה תת-סדרת טוקנים שבה הטוקן הראשון זהה לטוקן הראשון "הנקן" של האיטרציה(באיטרציה i טוקן i וקודמי נחזרים נכון) אנו לוקחים תת סדרה זו ומצביעים אותו במקום מה שנזהה באיטרציה الأخيرة.

זה מאפשר להקטין את כמות האיטרציות עוד טיפה

<https://arxiv.org/pdf/2402.02057>

⚡️⚡️ המאמר היום של מיק 14.06.24

CLLMs: Consistency Large Language Models

בשתי הסקירות הקודמות(כדי שתעבירו עליהם כי נתתי שם קצת הסברים) דיברנו על שיטות איטרטיביות מקבילות לדוגימה ממודלי שפה. השיטות האלו מבוססות על שיטות יאקובי או (GS) Gauss-Seidel. השיטות האלו מתחילות מכמה מסוימת ח של טוקנים שנדגמים באקראי (או בצורה קצת יותר מושכלת) ואז מעדכנים טוקנים אלו בביטחון איטרציות עד שתתנאי עצירה מתקיים(התכנסות). תנאי העצירה כאן הוא ב"כ שווין בין הפלטים של איטרציות עוקבות.

ובן לנו מעוניינים לסייע את התהילה' במשמעות פחות איטרציות ממספר הטוקנים שאנו חוזים בו בזמןית (ד"א ניתן להראות נדרשות לכל היותר ח איטרציות עד ההתכנסות).

שימוש לב במהלך האימון של מודלי שפה מותאם לשיטת הדגימה האוטו-רגרסיבית כאשר בוחרים טוקן בעל הסתsbירות הגבוהה ביותר ביחס לטוקנים הקודמים. אולם עכשו אנו דוגמים בצורה אחרת ואלו ניתן להתחשב בכך במהלך האימון. כמו כן במהלך האימוןuschair דוגמים עם השיטה זו (השילוב של יאקובי ו-GS). זה בדיק מה שנסקרו אותו היום עושים. המחברים מושפעים עוד יותר לLOS הריגל של מודלי שפה (הממקם את הנראות המירבית של הדטה). מטרת האיבור הזה היא לגרום למצער של מספר האיטרציות עד להתקנות של הדגימה האיטרטיבית.

המחברים בחנו שתי אופציונות לאיבר זהה:

1. מצער של מרחק KL הפוך לדעתך אך לא צלילי לעומק) בין התפלגיות הטוקנים בנקודת ההתקנות לבין התפלגיות טוקנים במהלך הדגימה האיטרטיבית (דוגמים האיטרציות באקראי).
2. מצער מרחק בין התפלגיות הטוקנים באיטרציות עוקבות.

ואם חשבתם שיש דמיון בין השיטה זו לבין המאמר של איליה סולזקייר ושותפיו "Consistency Models" - אכן הוא קיים ואני אצללו בו בקרוב.

<https://arxiv.org/abs/2403.00835>

🚀⚡️: 15.06.24 🚀⚡️: המאמר היומי של מיק

MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads

ב 3 הסקירות האחרונות ראיינו כמה שיטות איטרטיביות מקבילות, מבוססות על שיטות יאקובי ו-Gauss-Seidel המנסות להציג את מהירות גנרטט הטקסט (decoding) של מודלי שפה. היום נסקור קצרות מאמר המציג גישה אחרת לאותה הבעיה, שגם מבצעת גנרטט מקבילי של טקסט אבל בשיטה 'טיפה' אחרת.

בגדול המאמר מציע להוסיף ולאמן כמה "ראשים" (שכבה לינארית עם סופטמקס) למודל שפה מאומן. מטרתה של כל ראש זהה היא לחזות טקסט לא החל מהטוקן הבא אלא להתחילה לחזות מהטוקן ה-k אחריו הפרומפט (או הטוקן האחרון שנזהה). כמו כן בהינתן פרומפט באורך 10 טוקנים הראש מסדר 3 מגנרט טוקנים החל מהטוקן ה-14 בזמן שמודל שפה רגיל כוזה(מגנרט) החל מהטוקן ה-11. הראשים האלה מחוברים לשכבה האחורה (לפני שכבת החיזוי) של מודל שפה. ככלות המ פעלים טרנספורמציה לינארית על "יצוג(תליי קונטקט)" הטוקן המופק על ידי מודל שפה.

המחברים מציעים שתי דרכי לאמן מודל שפה עם הראשים אלו. הדרך הראשונה היא לאמן רק את הראשים כאשר מודל השפה עצמו נותר מוקפא. הדרך השנייה היא לעשות פין טון של מודל שפה מאומן (עם LoRa וכו'). במקרה השני הם משלבים את הLOS הסטנדרטי של מודלי שפה עם זה של הראשים האחרים.

באינפרנס המחברים לוקחים את החיזויים מהראשים השונים (כמה טוקנים החל מטוקן K לכל ראש) של הראשים השונים ומשלבים אותם בצורה דומה ל- beam search (כאן זה קצת יותר מורכב ונקרא tree-search) כדי לקבל את סדרות של טוקנים "התובות ביותר" ביותר. בעודם עושים משהו דומה למה שנעשה ב-decoding speculative קלאסי (טיפה נוספת מרכיב ממש sampling-and-rejection בעניין).

אז מה הרוח כאן אתם שואלים? שהראשים מופעלים באופן מקבילי ולפעמים בהפעלה אחת שלהם אינם חוזים כמו טוקנים ולא אחד כמו בגנרטט אוטורגרטיבי רגיל.

<https://arxiv.org/pdf/2401.10774.pdf>

的文章: 16.06.24 🔥🚀

STATISTICAL REJECTION SAMPLING IMPROVES PREFERENCE OPTIMIZATION

המאמר זהה וכמה הבאים שאס考ר בימים הקרובים מציעים שיטות שונות לשיטה Direct Preference או בקיצור DPO. למעשה PO בעצמה היא שדרוג של Optimization PPO שהפכה להיות מודול פופולרית אחרי שימושה בחברות השתמשו בה לישור מודלי שפה (alignment) או RLHF instruction tuning (foundational) בתור השלב האחרון של אימון מודל שפה foundational. השיטה שיצת למשפטת כי היא דורשת דאטא (שאלות ותשובות) המדורגות על ידי בני אדם - עבר כל שאלה הם (המתיגים) בוחרים מה התשובה איזה תשובה טוביה יותר.

למעשה PO בא ליתר את מודל התגמול (reward) גם חוסך גם משבבים לאימונו וגם מאפשר לא להחזיק מודל נוספת בשלב RLHF. למעשה PO מנצח את המבנה של פונקציית LOSS של PPO, שהוא מוקסם פונקציית התגמול עם איבר רגולרייזציה שבא לשמר את המודל המיושר קרוב למודל ההתחלתי, כדי להיפטר מפונקציית התגמול בפונקציית LOSS. זה מתאפשר עקב העובדה שקיים ביוטי מפורש לפוליס האופטימלי (מודל שפה "מושלים אחריו היישור") דרך הפוליסי אחריו-h-SFT (מודל שפה שאנו מתחילה ממנו את אימון היישור) ופונקציית התגמול.

אחרי שימושים במודל LOSS המשורה על ידי מודל BT (Bradley-Terry) המגדיר מהי הסתבותות העדפה של תשובה חיובית על תשובה שלילית (על אותה השאלה) מה- rewards שלהם, ואנו מגעים לביטוי עבר LOSS של RLHF שיכיל רק את הפוליסי ההתחלתי. זה למעשה PO והוא מזעער את פונקציית LOSS שלו על סט המכיל זוגות של תשבות טובות וגרועות.

המאמר שנסקור היום שואל את השאלה האם הדגימה האחדה מההסטה זהה היא אופטימלית (מבחינת איכות התוצאה שהיא הופolis ה壽피 או מודל שפה אחריו היישור). אולי אם היה לנו פונקציית תגמול הינו מעדיפים זוגות עם יחס מקסימלי בין reward של התשובה החיובית לשולית? אולי צריך לטעוף זוגות עם reward שלילי הנמור ביותר?

המאמר מציע את הגישה הבאה:

- מאמנים מודל text2text שבහינטן שאלת ושתי תשבות מוציא את התשובה המעודפת.
- באמצעות המודל הזה בונים את פונקציית התגמול דרך סמלוץ (על ידי דגימה של שאלת זוג תשבות) של הסתבותות העדפה של תשובה טוביה על תשובה גרועה.
- באמצעות פונקציית תגמול זו בונים פוליסי z_وك שולמעשה זה מודל שפה (האפשר לחשב הסתבותות של תשובה בהינתן שאלה)
- משתמשים בדגם rejection כדי לדגם z_וק באמצעות הפוליסי ההתחלתי (= מודל שפה) כדי למזעער את LOSS בדרך לפוליסי "המיושר".

הם גם משחקים עם כמה פונקציות LOSS כמו hinge loss (בטח כבר שכחתם אבל אוהבים להשתמש בו ב-SVM).
<https://arxiv.org/abs/2309.06657>

的文章: 17.06.24 🔥🚀

SSAMBA: SELF-SUPERVISED AUDIO REPRESENTATION LEARNING WITH MAMBA STATE SPACE MODEL

הסקירה נמצאת כאן:

https://docs.google.com/document/d/1zmMPssJsuvb_3zyXZf4uoehuR5GCuWXhuzrhDiCt3UE/edit
<https://arxiv.org/abs/2405.11831>

🚀⚡️ המאמר היומי של מיק 18.06.24 🚀⚡️
Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate REWARD HACKING

הסקירה זו ממשיכה את קו הסקירות האחרונות שכתבתי בנושא RLHF. כמו שאתם זוכרים פונקציית לוס ב-RLHF מכילה שני איברים: האיבר שמנסה למקסם את פונקציית התגמול (reward) והאיבר השני מנסה לשומר את מודל השפה אחרי טיב (פוליסי סופי) קרוב למודל שמתחלים את ה-RLHF ממנו. בעבר ייצאו מאמרים שהציגו לאמן כמה מודלי reward ואז למעט (או לקחת מקסימום) של כל ה-rewards של מודלים אלו עבור שאלה ותשובה נתנות של מודל שפה. זה לטענתם מקטין את הסיכוי שהמודל שפה ב-RLHF יבצע reward hacking כלומר יתכו נפוליים הממקסם תגמול אך בפועל מגנרט תשובות באיכות גורעה.

המאמר שנתקoor היום טוען שגישה זו אינה אופטימלית כי פונקציית לוס שאינה מאומנים מודלי reward (כזכור Bradley-Terry) גורמת לכך שכל שני מודלי reward שונים רק בקבוע שתלי רק בשאלתה א' קיבלו את אותו ערך של פונקציית לוס. בפועל זה אומר כי לכל לערכי ה-reward, המופקים על ידי המודלי, יכול להיות ממוצעים ובפועל הבחירה של המקיים או המוצע מכמה מודלי יכול עשויה להיות לא אופטימלית (כמו ממוצע של תפוזים ועגבניה). אז המאמר מציע לאמן פונקציית תגמול עם רגולרייזציה שבה "לרסס" את הקבוע זה שתלי רק בשאלתה ובכך "לסכרן" מודלי reward שונים.

<https://arxiv.org/abs/2312.09244>

🚀⚡️ המאמר היומי של מיק 19.06.24 🚀⚡️
INTRINSIC DIMENSIONALITY EXPLAINS THE EFFECTIVENESS OF LANGUAGE MODEL FINE-TUNING

כולכם מכירים את LoRa נכון? בטח גם שמעתם על שירות השכלולים השונים שלא כמו DoRa, MoRa וקדמיה. מתרברר כי היה מאמר שבצורה מסוימת הניח יסודות של משפחת הגישות זו.

למעשה מה זה LoRa? זה אופן שבו אנחנו עושים פיניטיון של מודלים מאומנים גדולים למשימה ספציפית בלבד' לעדכן את כל משקלים המודל. במקרה של LoRa אנו מאמנים מטריצת תוספות למשקלים של כל שכבה כאשר תוספת זו היא בעלת ראנק נמוך הרבה יותר מטריצת המשקלים המקורית. ככלומר ניתן ליאג אותה על ידי מכפלה שתי מטריצות בעלות ראנק נמוך (בגדלים מסוימים במקרה של LoRa).

מתרברר שגישה זו הייתה ידוע כבר ב 2020 ואףלו היו מאמרים שדיברו עליה ב 2018. אז המאים הציעו מספר דרכים לבניית מטריצת תוספת זו וביניהם הטלה ספארטית של וקטור במילדי נמוך למרחב בעל מספר מימדים גבוה דרך **Fastfood algorithm** (צורה של מטריצת ההטלה זו - תקראו עליו, זה חמוד).

בקיצור מאמר "היסטורי" מעוניין וקל לקרוא.

<https://arxiv.org/abs/2012.13255>

🚀⚡️ המאמר היומי של מיק 20.06.24:

WARM: On the Benefits of Weight Averaged Reward Models

הסקירה הזאת ממשיכה את קוו הסקירות בנושא שיפור ביצועי RLHF לטיב מודלי שפה. כבר דיברנו בסקירת הקודמות על כך שבמהלך RLHF המודל יכול לבצע reward hacking כלומר להתכנס לפוליס (משקל המודל) שמקסם את ה-reward ובאותו הזמן יוצר תשובות באיכות ירודה לפורומפטים.

המאמר שנסקור קוצרת היום מציע לאמן כמה מודלי reward שונים ולהשתמש בממוצע שלהם כ-reward יותר "יציב" שעשוי למנוע מהמודל לעשות reward hacking. הבעיה העיקרית בגישה הזאת נובעת מכך שהיא מצריכה להחזיק בזמן אימון RLHF כמה מודלי reward שונים יחדיו חישוב (ומיקיר את חשבון החשמל).

המחברים מציע לשלב את התוצאה של המודלים אלא הביצועים שלהם. בשפה פשוטה הם מאמנים כמה מודלי reward וממצאים את המשקלים שלהם. זה מסתמן על איזושהי תופעה שלא ידועה עלייה שנקראת "Linear LMC mode connectivity" או SFT. הטוענת שהביצועים של מודל עם סכום ממושקל של המשקלים של כמה מודלים אחרים הוא יותר טוב מסכום ממושקל (עם אותם משקלים) של ביצועים המודלים (אולי עמוק יותר בהמשך).

עכשו כדי לבצע את הפעולה הזאת הרשותות צרכות להיות בעלי אותה ארכיטקטורה ומה שונה בין מודלי reward כאן הם פרמטרי אימון כמו קצב למידה ודטופאות, סדר שונה של הכנסת נתונים לאימון (סיד שונה כנראה) וגם איתחולים שונים (לוקחים מודלים אחרים צ'קפונטים שונים ב-SFT).

התוצאה מתקבלים מודל reward אחד טוב יותר שמשמש אותו לאימון RLHF.

<https://arxiv.org/abs/2401.12187>

🚀⚡️ המאמר היומי של מיק 21.06.24:

Named Entity Recognition as Structured Span Prediction

היום נסקור מאמר בנושא שלא סקרתי הרבה מאוד זמן והוא *chon* Named Entity Recognition או NER. מטרת משימה זו היא לזהות בטקסט עצמים (מילים וקבוצות מילים רצופות) מסוימים כמו שמות פרטיים, כתובות מגוריים, מספרי ת"ז וכדומה. קיימים מודלי NER המתמחים בזיהוי שמות חברות, רשומות רפואיות וכדומה.

מהד גיסא משימת NER היא משימה דיסקרימינטיבית והتوزאה שלה היא סיווג של כל טוקן במשפט לקטגוריה שהוא שיר או לקטgorיה "O" אם הוא לא שיר לאף קטגוריה יעד (נציין כי היזהו מתבצע פר מילה ולא פר טוקן מכיוון שמילה עשויה להיות מורכבת מכמה טוקנים). מайдך גיסא ניתן בקלות להפוך אותה לבעיה גנרטטיבית כאשר המודל יגנרט את העצמים השיכים לקטגוריות יעד.

בעבר פותחו מגוון שיטות למשימה זו, חלקם rule-based, חלקם סטטיסטיים אך לאחרונה רשות השתלו לנו על NLP וגם המשימה זו לא הצלחה לבסוף מהן. הוצאו לא מעט רשותות שהגיעו לביצועים די יפים במשימה זו.

המאמר שנסקור היום מציע גישה מעניינת לבעית NER. כמו שאמרתי ניתן לפתור את הבעיה הזו באופן דיסקרימינטיבי וגןרטיבי אך המאמר הזה לוקח גישה בין אלו וקורא לה Structured Span Prediction.

בגדול הגישה עובדת באופן הבא. מעבירים את כל שמות הקטגוריות יחד (מופרדים עם טוקן מיוחד) דרך טוקנייזר שלהם. לאחר מכן מעבירים את הטקסט דרך טוקנייזר מסוילו ומכניסים את שניהם דרך מודל שפה דו-כיווני (bidirectional encoder או DeBerta) כמו BERT. המודל מפיק ייצוגי הטוקנים תלוי הקשר (גם עבר קטגוריות וגם בעבר הטקסט) בתור פלט.

החדש האמתי בא לאחר מכן. הרו המטרה של NER היא לזהות כמה מילים רצופות השייכים לאותה קטgorיה. נגיד אנו לוקחים את המילים מ 1 עד 4 ומנסים להזות מה הסיכוי שהם שייכים לקטגוריה C. המאמר מציע לחתת את הייצוגים תלוי הקשר של מילה 1, מילה 4 (המחברים מציעים להשתמש ביצוג של הטוקן הראשון של כל מילה לייצוג המילה) וגם קטגוריה C ובונים (מאמנים) מודל קטן לשערוך הסתברות זו. יש כמובן הרבה גישות לארכיטקטורה של מודל דليل זה. אפשר לעשות את עם רשות קונבולוציות פשוט, ניתן לנקוט את הייצוגים ולהוסיף שכבה לינארית ואני יכול לחשב על כמה אופציות נוספות.

עכשו השאלה האחרונה היא איך לבחור קטגוריות לכל המילים. השיטה הנאיבית היא לחשב את ההסתברויות האלו עבור כל תת סדרה של מילים רצופות החל מהמילה הראשונה ועבור כל תת סדרה לבחור את הקטגוריה בעלת הסתברות הגבוהה ביותר אם היא עולה על סף מסוים או קטגוריה אחרת אם זה לא. הבעיה עם הגישה הזאת שכך נוכל לפחות-spans ארוכים אחרים סימנו את ה-span קצר יותר שיש לו חיתוך עם ה-span האחרון.

عقب כך המאמר מציע כמה גישות שונות לבעה הלא פשוטה זו וביניהם Conditional Random Fields ו-*Maximum Weight Independent Set Method*. גם Exhaustive Search יכול לעבוד עבור טקסטים קצרים. יש כאן טעימה נחמדה של שיטות מעניות הקשורות לרשותה.

ואיך מאמנים את זה? האמת זה די פשוט - לוקחים את כל הקטגוריות המסווגנות בטקסט ומריצים עוקב cross-entropy loss על כלם.

מאמר מאד מעניין ומאיר עינים - מחר המשך...

<https://aclanthology.org/2022.umios-1.1/>

的文章: 22.06.24: GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer

המאמר הזה הוא שפוץ כל של המאמר שскаרנו אתמול 21.06.24. המאמר מציע גישה לאימון והיסק של מודל לזרוי NER המורכב משלבים הבאים:

1. מעבירים כל קטגוריה שברצוננו לזהות דרך טוקנייזר - הקטגוריות מופרדות על ידי טוקן מיוחד הנקרא "ENT"
2. מעבירים דרך הטוקנייזר את כל הטוקנים של הטקסט. ד"א הטוקנים של הקטגוריות מופרדות מהטוקנים של טקסט על ידי טוקן מיוחד "SEP"
3. מכנים את הטוקנים מהשלבים הקודמים לטרנספורמר דו-כיווני (encoder) כמו BERT או ROBERTA
4. מעבירים את הייצוגים תלוי הקשר של הקטגוריות דרך FFN דו שכבותי (יש צזה בטרנספורמר) כדי לקבל ייצוג של כל קטגוריה.
5. מפעילים את מה שנקרא במאמר הקודם: Structured Span Prediction כלומר כדי לזהות את הקטגוריה של הטוקנים ? עד ?+: לוקחים את הייצוג של טוקן ה-? ואת זה של טוקן ?+ ועוד מעבירים את השרשור שלהם דרך FFN דו שכבותי (מבנה דומה לסעיף הקודם) ורק מפיקים ייצוג של ה-span הזה.
6. כדי לשערק הסתברות ש-span (תת-סדרה של טוקנים רצופים) שייך לקטגוריה J מחשבים סיגמואיד של המכפלה פנימית של ייצוג הקטגוריה J מסעיף 4 עם ייצוג ה-span מהסעיף הקודם.

7. מפעילים אלגוריתמיים גרידים כדי לזרות spans השיכים לכל קטגוריה (המאמר לא מרחיב על כך, צריך להבitem בקוד)

<https://arxiv.org/abs/2311.08526>

🚀⚡️ המאמר היומי של מיק 23.06.24 TextGrad: Automatic “Differentiation” via Text

אני קצר שיכור אחריו כמה שוטרים ובירות באירוע המגניב של shot-shot-abl התמדה בסקרים יומיות גברת על כך. הסקירה של היום מדברת גישה ש”מיטילה”(project) את שיטת מורד גרדיאנט (gradient descent) או פשוט GD (gradient descent) למקרה שהמשתנה שאנחנו מפטים לפיו זה הפורטף ולא משקל המודל (שנותרונות קבועים). כמו שאתם זוכרים GD הסטנדרטי מזינים בכיוון הגרדיאנט החליל של פונקציית loss (מחסירים משקל המודל את הגרדיאנט מוכפל בקצב למדידה).

ב-GD הגרדיאנט מחושב בצורה ברורה (פחות מתמטית) כי פונקציית loss הינה גזירה ביחס למשקל המודל. ד”א בראייה מנווכח בתнатנו לגזר את פונקציית loss לפי הקלט (תמונה) מאותה הסיבה - לעיתים עושים זאת כדי לבנות תמונה המציגת את הלוס עבור קטגוריה מסוימת.

ABL איך לגזר את המודל ביחס לטקסט? הכוונה כאן לא לגזר את פונקציית loss לפי הייצוגים של טוקני הקלט (זה דואק אפשרי כמו במקרה של תמונה ונראה soft prompting). אך כאן מדובר ב”גזירה” אשכלה לפי הטקסט עצמו. כמובן שמדובר מתחם זה ד’ בעיתי כי טקסט הוא משתנה דיסקרטי.

המאמר הופך את מחליף גזירה מתמטית על ידי מה ”פידבק של שכבה ח לשכבה 1-ח” וכן לא מדובר בשכבות של מודלי שפה אלא בשכבות של כלים שונים המפעלים ומופעלים על ידי מודלי שפה (גיגי rag או כמו אגנטים). אך בכל שלב אנו שואלים מודל שפה (אם הוא מופעל) איך היה ניתן לשפר את הפורטף בשלב שלהם כדי לשפר את התוצאה וublisherים את הפידבק לשכבה הקודמת. כמובן שהאגרציה של פידבקים מתחילה ה-*W*|| בשכבה الأخيرة של המערכת ושיש לו סוג של פונקציית loss בתור ”שערך של איקות התשובה”. וכןן מתחילה האגרציה.

از textgrad זה בגודל פרופגציה של פידבק טקסטואלי ופחות טקסטABL עדין המאמר חמוד כי אפשר מערכות מורכבות מלא מעת כלים המערבים *also*.

<https://arxiv.org/abs/2406.07496>

🚀⚡️ המאמר היומי של מיק 24.06.24 Are you still on track!? Catching LLM Task Drift with Activations

הסקירה הזה הולכת להיות קצחה כי הרעיון העיקרי של המאמר הוא ד’ פשוט ואינטואיטיבי. אתם מדברים עם מודלי שפה שלכם באמצעות שאלות שבד”כ נקראות פרומפטים שהמודל עונה לכם. אבל מה קורה אם מודל השפה שלכם מחובר לעוד כל שmagנרט בשביilo פרומפטים למשל בהתקבוס על תוצאה של איזשהו חישוב על הפלט של מודל אחר או מתבסס על RAG או אולי אפילו תלוי בתוצאות חיפוש באינטרנט.

כמובן שגנרט אוטומטי של פרומפט יכול להתפרק (באגים, אולי פעילות זדונית) וזה ייחד עם שאלה לגיטימית המודל מקבל תופסת לא קשורה. בעיה ידועה, אה?

از המאמר שבנידון חקר את האקטיבציות של שכבות המודל (טרנספורמר כמובן) ומוצא הבדלים משמעותיים בין האקטיבציות הנוצרות על ידי שאלת לגיטימית לבין אלו שנוצרו עם שאלה "מקושחת". ואז הם בנו דאטהסט של שאלות טובות ושאלות מורעלות ואימנו מודל (קטן) שיודיע差异化 בין האקטיבציות של שאלות הטובות ולא טובות. המחברים לוקחים אקטיבציות של הטוקן האחרון של הפרומפט (השאלה) המלא

הם ניסו שתי שיטות: אחת היא אימון של שכבה לנארית המפרידה בין ייצוגים טובים ומורעלים. השיטה השנייה שהם מנסים נקראת metric learning שבמילים פשוטות מנסה ללמד ייצוג (המוחק על ידי המודל "המבדיל") המזכיר ייצוגים של העוגן (התחלתו השאלה ורחקיק אותו מהייצוג של השאלת המורעלת (התוספת המורעלת). אם מצליחים ב-*metric learning* אז בקהלות אפשר לתפור שכבה לנארית המבדילה בין הטובים ללא טובים.

<https://arxiv.org/pdf/2406.00799>

🚀⚡️:25.06.24⚡️🚀: המאמר היומי של מיק

Improving Reinforcement Learning from Human Feedback with Efficient Reward Model Ensemble

הסקירה זו ממשיכה את קו הסקירות על המאמרים שמנסים לשפר שיטות RLHF לטיב (instruction tuning) או פשטוט fine-tuning) של מודלי שפה. בחלק של שיטת RLHF (למשל PPO) אנו מאמנים מודל reward מבוסס על סט של שאלות ותשובות מדורגות על ידי המתאיגים האנושיים. מטרה של מודל זה לספק ציון לזוג (שאלה, תשובה) כאשר ציון גבוה מצביע על תשובה טובה ורצויה. לאחר כן אנו מאמנים (מטיבים) מודל שפה כאשר המטרה היא מקסום של פונקציה reward תוך שמירת של משקלים שהתחלנו מהם (نمداد על ידי KL divergence KL בין התפלגיות הטוקנים של שני המודלים). כל זה מתבצע *fly-the-kite* כאשר הדוגמאות נוצרות עלי ידי הגרסה העדכנית של המודל במהלך האימון.

הבעיה עם הגישה היא reward hacking כאשר למרות איבר הרגולריזציה (KL) המודל מתכוון למשקלים שמאימים לערכיהם גבוהים של פונקציית reward כאשר המודל עצמו "לא מספק את הסchorה". המאמר מציע להשתמש בכמה מודלי reward כי ensemble זה תמיד טוב. הבעיה שלהחזק יותר ממודל אחד בזמן האימון זה יקר מבחינת המשאבים. המאמר מציע שתי גישות להתגבר על זה:

- מתחילה מאותו המודל (שפה)
- לאמן מודלי reward זהים עם ראשים לנאריים (מאומנים) שונים. כך צריך לשמור רק מודל אחד והמשקלים עבור שכבה הליניארית עברו כל מודל.
- לאמן כמה מודלי reward בשיטה של LoRa - כך נדרש לשמור רק את תוספת המשקלים לכל שכבה שזה יכול להיות די זול מבחינת המשאבים

ואז אפשר לקחת ממוצע של rewards של כל המודלים או את המינימום ביניהם. יש לא מעט אופציות...

<https://arxiv.org/abs/2401.16635>

🚀⚡️:27.06.24⚡️🚀: המאמר היומי של מיק

Probing the Decision Boundaries of In-context Learning in Large Language Models

המאמר הזה מדגים בפעם מי יודע מה שיש możliwości שמודלי שפה מתקשים בהם מאוד אבל אם נסביר לו את המשימה ב"שפה" הוא די מצליח להסתדר אליה. הפעם המשימה היא סיווג מולטיקלאס - כלומר אנו מספקים למודל כמה זוגות של וקטורי x וויליבל שלו y . הווקטורים ניתנים להפרדה בצורה לנארית על ידי ישר מסוים

כלומר נמצאים בשני צידי (של הישר). זה ה-*context* שלנו. לאחר מכן המודל מקבל נקודה x ומנסה לחזות את הליבל שלו y .

המחברים ניסו להבין עד כמה המודל מצליח לזהות את הליבלים של הנקודות הנמצאות בין המינימום למקסימום של נקודות x שנדרשו לא-*context*. בשביל כך הם חקרו את את התפלגות הטוקן הבא אחרי השאלתה עבור כל הקטגוריות עבור ערכי x שונים. באופן לא מפתיע המודל לא למד להפריד את הנקודות על ידי הישר אלא התכנס לקו הפרדה די פרעו ורחוק מהישר המפריד. הגדלה מס' נקודות-*context* לא עזר ובנוסף קו ההפרדה היה רגיש לסדר הנקודות וגם לשימון של הליבלים השונים.

از המחברים גילו שפינטיון של המודל על מושימות דומות די עוזר למודל להתכנס לפתרון הנוכחי (גם לא מפתיע). יש עוד כמה דרכים Lagerom למודל לפתור את המשימה הפשוטה זו.

אם אתם שואלים אותי למשימות כאלו יש לכם גרסיה לוגיסטית....

<https://arxiv.org/abs/2406.11233>

🚀⚡️: 28.06.24 המאמר היומי של מיק

On-Policy Distillation OF LANGUAGE MODELS: LEARNING FROM SELF-GENERATED MISTAKES

זמן לא סקרתי מאמר על שיטות זיקוק של ידע (knowledge distillation) - לא נתקلت במאמרים מגניבים בנושא המעنى זהה. מה זה זיקוק ידע ממודל גדול למודל קטן יותר? למעשה זה ניסיון להעתיק למודל הקטן את הידע שיש למודל הגדל ככלותם לגרום להפגין ביצועים הדומים למודל הגדל.

יש כמה שיטות לעשות זאת - פשוטה ביותר זה לאמן אותו על הדטה שהמודל הגדל אומן עליה. יש שיטות המאמנת את המודל הקטן על הדטה המיוצר על ידי המודל הגדל. אם יש לנו גישה לתפלגות (של הטוקנים) אז מאמנים את המודל הקטן לחקות את התפלגות הטוקנים שהמודל הגדל מוציא. אם יש לנו אקטיבציות של השכבות של המודל הגדל ניתן לנסות לחקות גם אותם (אם המודל הקטן הוא בעל אותה ארכיטקטורה אבל עם פחות שכבות).

בכל גישות האלו אנו מאמנים (או פינטיון) את המודל הקטן בצורה supervised regression. כאמור יש לנו סט של דוגמאות ground-truth או שנוצרו על ידי המודל הגדל) אנו מאמנים את המודל הקטן עליהם. המאמר שנסקור היום מציעה להשתמש בגישה מעולמות למידה באמצעות חיזוקים (reinforcement learning) ממשפחota policy-policy. זה אומר שהאימון מתבצע על הדוגמאות שהרשת המאמנת עצמה יוצרת במהלך האימון (והיא משתנה כמובן).

המאמר הילך צעד אחד קדימה וחיליט לשלב את שיטת אימון policy-policy יחד עם האימון הסטנדרטי של זיקוק ידע. כאמור בהסתברותaphkaomega השיטה בוחרת דוגמאות אידאטיות האימון ובשאר המקרים היא מגירה דטה מהמודל הקטן. כל פעם המודל מנסה למצוור את המרחק בין התפלגות הטוקנים של הדוגמאות מהדאטס או מהמודל הקטן.

בד"כ ככל המרחק בין התפלגות של הטוקנים בשיטות זיקוק ידע נמדד על KL divergence (כלומר forward). המאמר מציע לשככל את הגישה זו עקב חולשה שיש לו forward. החולשה זו קשורה לעובדה ש-KL forward מנסה לקרב את התפלגות המודל המאמן לאזור המודל (mode) של התפלגות היעד (התפלגות

המודל הגדל במקורה שלנו. הכוונה כאן שההטפלגות המודל המאומן עלולה "להתרכז באזרע בעל מסה הסתברותיות גבוהה", נגיד ליד איזה מוד של ההטפלגות ומתעלמת מאייזרים אחרים שיש בהם מסה הסתברותית ליד מודים חלשים יותר של ההטפלגות.

למזהנו יש לנו KL reverse שהופך את המונח ואת המכנה בלוג של KL forward. ניתן להראות כי KL منסה "לכוסות" את כל האזרע בה הטפלגות היעד גדולה מאפס ובכך משלימה את KL forward. ניתן לשלב אותם לינארית (באופן קמור עם מקדם beta-1 ו-beta) ואז מקבל Jensen Shannon Convergence או JSJ שנותן מענה לבעה האינהרנטית של KL forward. ובה המאמר משתמש במקום KL forward הרגיל.

ניתן לשלב את פונקציית הלוס של המאמר עם עוד איבר האחראי על מיקסום פונקציית reward כלשהי עבור המודל הקטן (כמו ב-RLHF).

ושכחתי להגיד(לא קשור למאמר) ש- KL forward זה בדיק מה יש לנו בכל פונקציית loss המבוססת על cross entropy (נגיד במשימות סיווג).

<https://arxiv.org/abs/2306.13649>

🚀⚡️ המאמר היומי של מיק 29.06.24 ⚡️🚀

What Are the Odds? Language Models Are Capable of Probabilistic Reasoning

הסקירה הזאת הולכת להיות ממש קצרה. לפני ימיים (27.06) סקרתי מאמר שבדק האם מודלי שפה ענקאים מסוגלים לבצע רגסיה לוגיסטיבית והגיע למסקנה שבלי עזרה ורמזים מאוד שימושיים הם לא מצליחים לפתור אותה.

הפעם המחברים בדקו האם מודלי שפה מסווגים "לנתוח הטפלויות הסתברותיות". למשל אמורים למודל שפה שאיזשהו ערך מפוגג גאותית עם תוחלת 3 ושותות 2 ושאלים אותו מה האחזון ה-80 של הטפלות או לאיזה אחזון שייכת דוגמה בעלת ערך 4. באופן די מפתיע המודל מצליח לא רע בשאלות האלה למرات שקיבל הוראה לא להריץ קוד (זה יכול לעזור כמו שאתם מבינים).

אז מה לדעתכם קורה כאן? איך המודל מצליח לפתור את השאלות האלה?

<https://arxiv.org/abs/2406.12830>

🚀⚡️ המאמר היומי של מיק 01.07.24 ⚡️🚀

Grokfast: Accelerated Grokking by Amplifying Slow Gradients

המאמר הזה משך את ענייני משתי סיבות. הסיבה הראשונה היא הופעת מייל Grokking בקורסית. מה זה בעצם Grokking בהקשר של אימון רשותות. אתם בטח יודעים אם אנו מאמנים את הרשות שלנו ליותר מדי זמן (כלומר אפוקים) אז באיזושהי נקודה היא מגיעה למצב של אופורטיט. כלומר הלוס על טריין סט ממשיר לרדת בזמן שהLOS על סט ולידציה מתחילה לעלות כלומר יכולת הכללה של המודל נפגעת.

אבל אם אנו נשעיר לאמן את הרשות שלנו עוד ועוד באיזשהו שלב הלוס על סט ולידציה מתחילה לרדת לאט כלומר יכולת הכללה של המודל משתפרת. כלומר אנו יוצאים מ"משטר האoorpit" אחריו שלב מסוים של אימון זהה נקרא grokking. התופעה הזאת נחקרה רבות על ידי המדענים בתחום למידה عمוקה אבל אין הבנה מלאה למה זה קורה. השורשים של grokking והו נמצאים כנראה בתופעה שנקראה double descent.

הסיבה השנייה שבחרתי לסקור את המאמר כי נוכחתה של התמרת פורייה שם אלא אחריו התעמקות קלה התברר שניתן היה להסתדר גם בReLU ולהסביר את המאמר בצורה פשוטה יותר בהרבה (מה שאני עשו בסקירה זו).

גרוקינג זו תופעה מאוד נחמדה וכל אדם המאמין את המודלים שלו חפץ להגיע אליו אף הטעיה שצורך לאמן את הרשות למשך מאות אלפי ולפעמים יותר איפוקים זהה מאוד יקר. השאלה האם ניתן לזרז את התהילה זהה ולהגיע לגורוקינג מהר יותר.

זה בדיק מה זה המאמר רוצה לעשות. המאמר טוען שם נחlik טיפה את עדכון המשקלים של הרשות (כלומר את הגרדיאנטים) אז ניתן להגיע לגורוקינג מהר יותר. נשמע לא מופרך בגודל (למשל OPO בລמיה עם חיזוקים גם מרכיבת את עדכון הגרדיאנט וגם שיטות אימון כמו ADAM ומומנטום של נסטורוב) - אבל כמובן ההוכחה לא נמצא במאמר. וכך המחברים דוחפים התמרת פורייה מהטיבה הפושטה שהחלוקת זו היא למעשה העברת גרדיאנים דרך pass-wso אבל כאמור אפשר היה להסתדר בקלות בלבד.

בסיופה של דבר המאמר מציע למצע כמה גרדיאנטים, להחליק(להוציאף) באמצעות המוצע הזה את הגרדיאנט הנוכח והזען את משקל הרשות (עם adam למשל). כמובן שהה דרוש לשמר כמה גרדיאנטים וזה מזכיר הרבה זכרון והמחברים הציע החלקה מעריכית (exponential smoothing) במקום זה בלי כמעט לפגוע בתוצאות (התוצאה היא כמובן זירוז של הגעה לגורוקינג).

מאמר חמוץ אבל ציפיתי ממנו קצת יותר..

<https://arxiv.org/abs/2405.20233>

🚀⚡️:02.07.24 המאמר היומי של מיק

From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data

היום סוקרים מאמר קליל שלא דורש כל התעמקות מתמטית אבל עדיין יש בו רעיון נחמד. המאמר מציע גישה מאוד פשוטה לשיפור יכולת של מודל שפה להפיק מידע מテקסט בצורה מדויקת. למשל בהינתן טקסט ארוך המוזן למודל, המודל נדרש לנ强壮ן על שאלות עלי'ו (הטקסט) בלי קשר לאיפה נמצא פיסת הטקסט הרלוונטי לשאלת. מודלי שפה בד"כ מתקשים במשימה זה בהעדר אימון ייעוד.

שיטת פיניטיון מקובלת לתת למודל טקסטים ארוכים ולאמן אותו לענות על מגוון שאלות בטקסט זהה (למשל לוגחים פסקה לא קשורה, משתלים אותה לטקסט ושאלים אותה המודל לגביה. גישה זו מביאה לשיפור ביצועי המודל במשימה אבל כמה מחקרים הצביעו על כך שבמהלכה המודל למד "מידע ועובדות מיותרים" שהרע את יכולת reasoning שלו).

המחברים הציעו שיטה כדי להקל הטעיה זו. הם בנו דאטאסהט שהוא הרבה מיליון טקסטים שהפותחו והעריכו בהם הם מספרים. המודל מאומן להפיק נכון ערך של מפתח נתון. משימה יותר קשה להפיק ערך של מפתח מסוים המורכב מכמה מספרים כאשר אני מעבירים את המספרים מהפתח למודל בסדר שונה מאשר הם מופיעים באחד המילונים. היופי בכך שהดาטהסהט הזה לא מכיל מידע עובדתי בכלל והמודל לא יכול ללמוד אותו (המידע). ככה מונעים את "הרעלת המודל" במידע זר...

<https://arxiv.org/abs/2406.19292>

🚀⚡️:03.07.24 המאמר היומי של מיק

The Remarkable Robustness of LLMs: Stages of Inference?

מאמר מעניין החוקר איזה שכבות ניתן לזרוק ממודל השפה וудין לשומר על ביצועים נאותים. אתם אולי מכירים lottery ticket hypothesis כירשותות עתירות פרמטרים (overparameterized) בד:כ ניתן למצוא קטנה הרבה יותר עם ביצועים מאד קרוביים אך הבעה שאנו לא יודעים לאתרא אותה.

המאמר כאמור בוחן איזה שכבות הן סוג של מיותרות במודלי שפה והגיע לתופעות מעניינות לגבי תהליכי האינפראם שלהם. הם ציהו 4 שלבים עיקריים

1. דה-טוקניזציה או רכישה התחלתית של קשרים קונטקטואליים: טרנספורמציה ראשונית של "צוג-hraw" (מהamilion) של הטוקנים ל"צוג תליי הקשר" (חישובי attention כבדים לכל אורך הקונטקט).
2. הנדסת פיצ'רים התחלתיים מה"ציגים" תליי הקשר מהשלב הקודם ו"הכנת קרקע" לחיזוי של הטוקנים הבאים. עדין לא ניתן לחזות את הטוקנים הללו מהפיצ'רים בשלב זהה אבל המודל מתחילה "להבין" הקשרים מרחבים ועתים בטקסט (היא מחקר מעניין זהה)
3. בניית קבוצות נוירונים (אנSEMBל) לחיזוי הטוקן הבא. בשלב זהה הרשת מתחילה להתכנס ולבנות קבוצות "prediction neurons" שישולבו יחד למטרת חיזוי הטוקן הבא.
4. חידוד של prediction neurons: הרשת "בוחרת" את הנוירונים החשובים ביותר לחיזוי הטוקן הבא על ידי הדעכה של חלק מה-neurons prediction מהשלב הקודם.

והci חשוב שהשכבות מעורבות בשלב 1 ובשלב 4 הם הci חשובות לביצוע המודל כאשר חלק מהשכבות של שלב 2-3 ניתן להסיר ללא פגיעה משמעותית ביצועים.

הרבה טענות מעניינות במאמר זהה (חלקם הגדול זה סיכום של העבודות הקודמות בנושא זהה).

<https://arxiv.org/abs/2406.19384>

🚀⚡️ המאמר היומי של מילק 04.07.24 🚀⚡️

How Do Large Language Models Acquire Factual Knowledge During Pretraining?

המאמר חוקר נושא מתי מודלי שפה אשכרה רוכשים ידע עובדתי (למשל שעיר בירה של צרפת היא פריז) במהלך אימון מקדים. בנוסף המאמר גם בודק כמה זמן לוקח לשכוח ידע עובדתי. אוקי, אתם בטח זוכרים שאנו מאמנים מודלי שפה שלנו עם אחת הצורות של משפטת מורד הגרדיאנט (GD gradient descent) או (GD). בד"כ דוגמאות כמה דוגמאות הסט האימון שלנו (מינি-באץ') ומוציאים לינארית את משקלי המודל לכיוון הנגדי של הגרדיאנט המומוצע של מיני-באץ'.

המאמר בונה דוגמא של טקסט המכיל ידע עובדתי ומכויס אותו למיני-באץ' כל כמה איטרציות של GD. המחברים מצאו כמה דברים מעניינים. למשל כמה דatasheet שהמודול אומן עליו לפני התחלת הזרקת ידע עובדתי לא משפיע על מספר האיטרציות הנדרש ללמידה של ידע עובדתי. כלומר יותר "ידע" הנמצא כבר במודול תורם ללמידה+lמדידה.

שנית, המאמר מראה שמהירות הלמידה של ידע עובדתי לא מושפעת ממשי מתחילה להזrik למודול את הידע. ככלומר מודל מאמן לאו דואקא תלמיד יותר טוב. ויש עוד כמה תגליות מעניינות במאמר.

איך בודקים האם המודל אכן למד את הידע העובדתי שהזרקנו - המחברים לא מרחיבים על כך אבל נראה זה מחושב דרך likelihood של התשובה הנconaה על השאלה לגבי פיסת ידע עובדתי זה, למשל "מה עיר הבירה של צרפת".

<https://arxiv.org/abs/2406.11813>

🚀⚡️ המאמר היומי של מיק 05.07.24: A Survey of Large Language Models for Graphs

גרפים מודלי שפה גדולים: האם זה שידור מהחלומות? גרפים נמצאים בכל מקום, מרשומות חברותיות ועד למבנים מולקולריים ורשתות נוירונים על גרפים (GNNs) הם הפתרון הנפוץ לשימושים כמו יובי קישורים וסיג קודוקודים. אבל ל-GNNs יש מוגבלות: הם מתकשים עם דата דיליל ולעתים קרובות אינם מצליחים להכפיל היטב לגרפים בעל מבנה שלא נראה קודם.

מайдך גיסא LLMs מספקים פתרון משלים: הם מצטיינים בהבנה וסבירם טקסטים (זהה דата דיליל שהוא בעצם גרף - המתאר קשרים בין מילים או קבוצות של מילים) יותר מאשר גרפים. אז, מה אם נשלב את החזקות של GNNs ו-LLMs? מאמר סקר חדש חוקר לעומק את החיבור המבטיח הזה.

המחברים מציעים טקסטונומיה של ארבעה שילובים אפשריים בין LLM ל-GNNs: שימוש ב-GNNs בתור שלב מקדים ל-LLMs, שימוש ב-LLMs לפני GNNs, שילוב של LLMs וגרפים, ושימוש ב-LLMs בלבד למשימות גרפיות. לכל גישה יש יתרונות וחסרונות, אבל הפוטנציאל ברור. על ידי ניצול הכוח של LLMs, יוכל להתגבר על חלק מהמוגבלות של טכניקות למידה מסורתית על גרפים.

<https://arxiv.org/pdf/2405.08011>

🚀⚡️ המאמר היומי של מיק 07.07.24: The Road Less Scheduled

היום סוקרים מאמר שלא נראה כמו מאמר למידה عمוקה רגיל. בהתאם זה אולי יכול להיראות שהמאמר מציע עוד שככל מי ידוע מה ל-ADAM או שיטה אופטימיזציה של לוס אחרת. אבל זה לא בדיק. המאמר כן מציע שיטה אופטימיזציה (מציאת מינימום) לפונקציות קמורות אבל זה בא ממטרה לשפר את Adam או שהוא כזה אלא מציע שיטה לשיפור קצב ההתקנסות של אלגוריתם מורד הגראדיינט (GD) הידוע.

המאמר מתחילה מכך ש מבחינה תיאורטית האלגוריתם של PR (Polyak-Ruppert) הוא זה שאמור להביא התקנסות אופטימלי אבל בפרקטייה זה פחות קורה (לא ברור לאיזה פרקטיקה הם מתכוונים כי התוצאות שלהם נתנו מתייחסות לרשות עמוקות הלא קמורות). PR בעצם עשו אותו GD אבל העדכון האמייתי המוחלך מעריכת עם העדכון האחרון. ככלומר באיטרציה ℓ העדכון של GD נכנס עם המקדם $1/\ell$ (אפשר לשחק עם זה לפי המאמר אבל קשה להגיד לקצב החלקה אופטימלי).

המאמר מציע שיטה חדשה (3 שלבים במקום 2 ב-PR) שմפערת ההתקנסות של PR ללא צורך בבחירה של פרמטר ההחלה.

<https://arxiv.org/abs/2405.15682>

🚀⚡️ המאמר היומי של מיק 08.07.24: Mixture of A Million Experts

המאמר של היום מציע לקחת את שיטת MoE (Mixture of Experts) לבניית ארכיטקטורות של מודלים עמוקים פופולרית במיוחד במודלי שפה. מאד גדול ב- MoE הרשות מרכיבת מותת-רשנות (בד"כ מחלקים את שכבת FFN של הטרנספורטorm לכמה חלקים זרים). MoE מואמן להשתמש כל פעם בחלק מותת-רשנות או (הנקראות מומחים) כאשר רשות gating רדודה יחסית באיזה מומחים צריך להשתמש כל פעם. ככלומר יש לנו כן

סוג של מימוש הגישה שנקראת "osz lottery hypothesis" דינמי כאשר כל פעם בוחרים להריץ רק חלק מהרשת.

כנראה שככל יש ברשת יותר מומחים בעלי אותה הארכיטקטורה וכל פעם בוחרים אותו מספר של המומחים הביצועים אמורים להשתפר אולם המחיר הואemodel גדול יותר. המאמר מנסה לבדוק האם שווה להשתמש בהרבה מאוד במומחים רזים מאד. המחרים מציעים לעובוד עם מיליון של מומחים של כל אחד מהם היא דל במיוחד. כמובן שכל פעם צריך לבחון את המומחים כל פעם ומכיון שיש מיליון מומחים אז נדרש ממש חישובי לא קטן. המאמר מציע להשתמש בטכניקה הנקראת product key retrieval כדי להקטין את הסיבוכיות (בגודל זה חלוקה של וקטור המפתחות (keys) לשני חלקים, ביצוע חישוב לכל אחד בנפרד ושלובם).

וgility מהו מעניין במאמר זה - יש law scaling גם ל-*Mo*s. אולי אסקור אותו בקרוב.
<https://arxiv.org/abs/2407.04153>

🚀⚡️ המאמר היומי של מיק 09.07.24: Learning to (Learn at Test Time): RNNs with Expressive Hidden States

המאמר הזה המצahir שהוא לומד ב"זמן טוט" משך את עיני היום. המאמר מציע ארכיטקטורה חדשה ומעניינת לעיבוד נתונים סדרתי. בעיקרו הרשות ד' דומה ל-RNN מבחינת המהות אבל יש כמה הבדלים מהותיים.

אז מה יש לנו בארכיטקטורה זו? בדומה ל-RNN אנו מחשבים את היצוג עבור יחידת>Data בזמן t (נגיד טוקן t) אבל כאן עושים זאת בשיטה שונה. לפי המאמר במקום לחשב את היצוג עצמו אנו מחשבים את וקטור המשקלים שיאפשר לנו לחשב את ייצוגו של יחידת>Data t. ככלומר אנו מעדכנים את משקלות מודל בתנווה בהתאם לדאטה כלומר הרשות מתאਪטמת ומתחילה את עצמה לדאטה שעליה היא מופעלת. זה נעשה באמצעות הזזה של המקשיים בכיוון הנגדי של הגרדיאנט של פונקציית loss.

מה זה בעצם פונקציית loss ואיך מאמנים אותה? נניח שהייצוג של איבר>Data t מחושב על ידי פונקציית f. במקרה זה פונקציית loss יכולה להיות (למשל) נורמה של הפרש ריבוע של ייצוג>Data z (המחושב עם f) מהדאטה עצמו. ככלומר אם מאמנים את וקטור היצוג להיות מסוגל לשחרר (כלומר לזכור) את הדאטה עצמו t_x. כמובן שאין בזה הרבה משמעות אבל אם נאמן רשות עם קלט מושרע ונשווה את ייצוג עם הדאטה האמיתית נקבל סוג של רשות denoising שהרשת לומדת להפיך ייצוג המאפשר לזכור את הפיצ'רים המהותיים של דאטה הנחוצים לשחרר.

דרך אחרת המוצעת במאמר לאמן את רשות לשחרר הטלה למיד נמוך של דאטה להטלה אחרת כאשר שתי הטלות נלמדות גם כן. היצוג של דאטה במקרה הצעה במקורה המוחשב עם הטלה נלמדת שלישית (עם פונקציית f). ככלומר המטרה כאן ללמידה את ייצוג של דאטה אשר המשקלים מחושבים עם GD מהמשקלים הקודמים.

האררכיטקטורה קיבלה שם *ttt* וניתן לשלב אותו על שכבות אחרות (כמו טרנספורמרים או MSA). רעיון מגניב שבינתיים לא הפנתי אותו עד הסוף...
<https://arxiv.org/pdf/2407.04620.pdf>

🚀⚡️ המאמר היומי של מיק 11.07.24: DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS

המאמר שנסקרו היום הולך להיות די קליל. הוא מתמקד בהקענות היזמות (hallucinations) של מודלי שפה. מה זה היזה של מודל שפה? זו שאלה לא טריויאלית בכלל (יש כמה תרחישים). נתמקד בהזיה המתבטאת בכך

שהמודל נותן תשובה לא נכונה עובדתית. נגיד, כולם על השאלה מה עיר בירה של לטביה הוא עונה שזה ריגה בזמן שההתשובה הנכונה היא טאלין.

המחברים מציעו שיטה ה"מכילתי" את התפלגות הטוקנים בשכבה החיצונית (האחרונה) של מודל שפה. המאמר טוען כי בהרבה מקרים שבם הטוקנים הנכונים בתשובה מפגינים עליה משמעותית בהסתברות מהשכבות הראשונות ועד לאחרונות. זה בולט במיוחד בטוקנים הלא טריויאליים (לא מילות חיבור וכאלו) הדורשים ממודל שפה לגיאס את הידע העובדתי שלו. בהתאם לאובייקטיבציה זו המאמר מציע שיטה המורכבת משני שלבים. בשלב הראשון מזהים את השכבה הרחוקה ביותר מבחינת התפלגות הטוקנים (השכבה זו נקראת השכבה ה-*ci* פחותה בשלה) מהשכבה האחרונות. מרחק כאן מוגדר על ידי Jensen-Shannon divergence או JSDF בין התפלגות הטוקן.

בשלב השני מחסרים (ב-*scale* logo) את ההסתברויות של השכבה ה-*ci* פחותה בשלה מההסתברויות של השכבה האחרונות. בנוסף מAPSים את כל לוגיטים של הטוקנים בעלי הסתברות הקטנות ביותר (שמילא לא אמורים להיבחר). לאחר מכן עושים סופטמך ומשתמשים בשיטת decoding האהובה עליהם כדי לחזות את הטוקן הבא. <https://arxiv.org/abs/2309.03883>

🚀⚡️: המאמר היום של מיק To Believe or Not to Believe Your LLM

מאמר מאד מעניין מבית גוגל. המאמר מנסה להבין איך ניתן לזרזות עד כמה המודל בטוח בתשובתו לשאלת. כולם המאמר עוסק בכימיות של אי-ודאות של תשובות המודל. המאמר מנסה בין שני סוגים של אי-ודאות הידועים בתורת השערות: אלטורי (aleatoric) או אפיסטמי (epistemic). אי-הוודאות האפיסטמית מתרחשת כאשר המודל לא יודע מה התשובה לשאלת ומתייחס אליה (hallucinations או *hallucinations*). לעומת זאת אי-הוודאות אלט/orית מתרחשת כאשר יש כמה תשובות לשאלת נתונה והמודל בוחר אחת התשובות הנכונות.

המאמר מציע שיטת פרומפטינג המאפשרת להבדיל בין שני סוגים אי-ודאות. מאוד בגודל לשאלת נתונה מציגים למודל תשובות אחרות (לאו דווקא) נכונות לשאלת (*is other response*...). לאחר מכן בודקים האם ההסתברות של התשובה הנכונה מושפעת מכמות התשובות האחרות המזוננות למודל. אם הסתברות זו מתחילה לרדת זה הסימן שמודל שפה לא צזה "יודע מה התשובה" ואילו הוודאות האפיסטמית הינה גבוהה.

המאמר גם מציע פרימורק מתמטי המבוסס על כלים מתוך המידע לאנליה של אי-הוודאות האל. נשמע מ Amar שווה להתעמק בו.

<https://arxiv.org/pdf/2406.02543>

🚀⚡️: המאמר היום של מיק SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales

במשך לסקירה של אטמול, מאמר קليل יותר שמציע שיטה למד מודלי שפה לשערך אי-ודאות בתשובותם. המחברים מציעים שיטה מאוד אינטואיטיבית המורכבת משני שלבים עיקריים: ייצור דאטהסט למשימה זו (כימות אי-ודאות) וטיפוב (fine-tuning) של המודל על הדאטהסט זהה. בשלב השני ממשיכים לאמן את המודל עם שיטת PPO מועלם למידה באמצעות חיזוקים כדי לשיפור נוסף של ביצועיו.

בשלב הראשון לוקחים דאטהסט של שאלות ותשובות הנקרא HotpotQA ומצביעים את השאלות מהם למודל שפה ומקשים מהם לתת תשובה מלאה ב-*reasoning*. לאחר מכן מקליטים את תשובה המודל (יחד עם

ה-reasoning) לקלוסטרים לפי האמצעים שלהם ומחשבים את יחס של גודל הקלוסטר המכיל את התשובה הנכונה (מהדעתהסט) יחסית לכל התשובות. זה מدد איזה הוודאות שלנו שעלי נאמן את המודל בהמשך.

לאחר מכן מפלטים את השאלות ובסוף מבקשים מ-4gpt לחתה הסברים למה המודל היה עשוי לתת תשבות לא נכונות לשאלה (כלומר "הסבירה" לאו וודאות). בשלב האחרון מטיאבים (מאמנים) מודל שפה נתון קודם כל לתת תשובה נכון, לדיק במדוד של איזה הוודאות ובנוסף לתת reasoning נכון לנוכחות של איזה הוודאות. כל אלה נמצאים באופן פורש בפונקציית הלווי.

בשלב השני ממשיכים לאמן את המודל בשיטה PPO כדי למצער (או למקסם אותה עם מינוס) את ההפרש בין נכונות התשובה (0 או 1) ורמת-the-confidence של המודל לגבייה. כמו בכל שיטת PPO הדוגמאות נוצרות "חס the fly" אחרי כל עדכון של משקל המודל.

<https://arxiv.org/abs/2405.20974>

🚀⚡️ המאמר היומי של מייק : 15.07.24 ⚡️🚀

Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

אהבתי את המאמר הזה כי הרעיון מאחורי הוא מאד אינטואיטיבי ופשוט. המאמר מציע גישה להתרמודדות עם הזריות(hallucinations) של מודלי שפה. מאד בגודל הזריות של מודל שפה קורות כאשר מודל שפה עונה לא נכון לשאלת המשתמש. יש לכך סיבות: למשל המודל לא מסוגל לענות על התשובה כי היא פשוט לא נמצאת בזיכרון שלו" (למשל השאלה על אירוע עדכני שהמודול לא אומן על הדאטה לגביו). הסיבה השנייה היא העדר יכולות להבין את השאלה.

המחברים מנסים להתרמודד עם הזריות של מודל שפה על ידי ניתוח של משקל ה-h-attention של הפרומפט (השאלה) ושל תשובתו של המודל. נניח שהפרומפט מכיל N טוקנים וכרגע אנו חווים טוקן מס' ℓ של תשובתו של מודל שפה. קודם כל מחשבים את סכום מוקדי attention עבור N טוקנים של הפרומפט P ווסף מוקדי h-attention עבור כל ℓ הטוקנים של התשובה R . מחשבים את היחס בין $P - R + P$ ועבור כל שכבה של הטרנספורמר ועבור כל ראש (head) של בלוק הטרנספורמים.

לאחר מכן בונים וקטור מהיחסים הללו ומאמנים מודל המכיל שכבה אחת שמרתתו היא לחזות האם המודל הוזה או לא. כיוון האורך תשובתו של המודל יכול להיות כלשהו המוחברים מאמנים מודל עבור מספר קבוע של טוקני התשובה T . אם התשובה מכילה יותר מ- T טוקנים מפעילים את המודל עבור כמה פעמים בשבייל להזות הזריות בחלקים השונים של התשובה.

איך בונים דאטהסט לאימון של המסוג הזה? בגודל נתונים למודל שפה לענות על שאלה ומפעילים מודל שפה חזק כדי להזות תשבות נכונות ולא נכונות (הזריות).

<https://arxiv.org/abs/2407.07071>

🚀⚡️ המאמר היומי של מייק : 16.07.24 ⚡️🚀

How Does Quantization Affect Multilingual LLMs?

היום נסקור קצר מאמר שחוקר נושא חשוב לכלי מי שעוסק במודלי שפה. הנושא הזה הוא קוונטיזציה או קוונטוט של מודלי שפה שמאפשר לנו גם להקטין את כמות הזיכרון הנדרש לאחסון של המודל וגם מזרץ את האינפנס של המודל. אבל כמובן שהוא לא בא בלי המחיר והמחיר הוא ביצועי המודל. המאמר חוקר עד כמה

חמורה פגיעה בביטוי המודלי, כמו רמות ושיטות קוונטוט (ניתן לקוונטוט שכבות שונות בرمות שונות וגם ניתן לקוונטוט משקלי המודל והקטיביות בرمות שונות של קוונטוט).

המאמר נכתב על ידי מדען חברת *cohere* ובאופן טבעי מתמקד במודלי שלהם. המחברים לוקחים מודלים גדולים

שוניים ובודקם במספר ביצועים שונים וגם ביצעו אבלואציה של ביצועי המודלים על ידי בודקים אנושיים.

המחברים הגיעו למסגרת מסקנות מעניינות:

1. הפגיעה מהקוונטוט הנמדד על הביצ'מארקים משמעותית קטנה יותר מזו הנעשית על ידי בודקים אנושיים.
2. הפגיעה לרוב מחמירה כל שקוונטוט נהיה יותר קשה לומר לפחות לפחות ביתים
3. מודלים גדולים בד"כ עמידים יותר לקוונטוט מאשר מודלים קטנים יותר
4. מודלים מולטי-שפתיים (multilingual) סובלים יותר מקוונטוט מאשר מודלים חד שפתיים והביצועים על השפות הפחות נפוצות יותר מאשר על שפות נפוצות יותר
5. היכולת של המודלי ל-*reasoning* (למשל יכולת לפתור שאלות מתמטיות) נפגעת מאוד מהקוונטוט.

יש עוד כמה מציאות מעניינות...

<https://arxiv.org/pdf/2407.03211>

🚀⚡️: המאמר היומי של מאי 24 🚀⚡️ Learning Rate Curriculum

רוב המאמרים שסקרתי לאחרונה היו בנושא מודלי שפה והחלטתי לגאון טיפה ולסקור מאמרים בנושאים אחרים. מאמר שנתקור היום מדבר על שיטת אימון הנקרatta למידת curriculum שבא לנו מאננים את המודל כמו שאנו ממלדים חומר לתלמידים - מהקלקשה. יש כמה וריאציות של למידת curriculum: באחת מהם אנו מתחילהם לאמן מודל עם דוגמאות קלות ובהדרגה מעלים את קושי הדוגמאות. הוריאציה השנייה אנו מתחילהם ממשימה קלה יותר ומעלים את מרכיבותה בהדרגה. בשליית מאננים מודל יחסית פשוט ומעלים את "מורכבות" של המודל.

המאמר מציע גישת curriculum אבל רקציב למידה. המחברים מצינים שלמשל בראשות קונבולוציה עדיף להתחל להתמקד יותר בלמידה של השכבות הראשונות כי למעשה אם אלו לא נלמדו טוב ועדין קרוביים למצב האיתחול שלהם אז הם יוצרים דאטה "ירועש, מדי" שזרום גם לשכבות הבאות שמתפקידו להתמודד איתו (המאמר מצין כמה עבודות שעשו את הנושא והגיעו למסקנות האלו). תופעה דומה מתרחשת גם כאשר אנו עושים פין טיון למודל למשימה מסוימת כאשר המודל לפני זה אומן למשימה אחרת.

כדי להתמודד עם סוגיה זו המחברים מציעים להתחיל מקצב למידה גבוה עבור השכבות הראשונות (שיורד ככל שמתקדמים לשכבות עמוקות יותר). במהלך האיטרציות עלולות את קצב למידה בכל השכבות כך (קצב עלייה לא שווה בין השכבות) כך שעם הזמן (=AITRIZYOT) קצב הלמידה של כל השכבות משתווות. נציין שהמחברים מציעים במספר האיטרציות הנדרש להשוואת קצב הלמידה עבור כל השכבות צריך להיות ממשוערת קטן יותר מכמות האיטרציות הכול הנדרש לאימון המודל. כלומר כל השיטה הזאת מופעלת בשלב ה"חימום" של הרשת.

<https://arxiv.org/abs/2205.09180>

🚀⚡️: המאמר היומי של מאי 24 🚀⚡️ Trainable Highly-expressive Activation Functions

משיכים את קו הגיון וסוקרים מאמר לא קשור לשירות למודלי שפה. היום נסקור מאמר של כמה חוקרים ישראלים המציע דרך חדשה לבנות פונקציות אקטיביזציה בראשת נירונים. היום פונקציות אקטיביזציה הן לא נלמדות

לרוב (ReLU, GeLU, tanh ובודומה). לעיתים פונקציות אקטיבציה מכילות hyperparameter שלא נלמד במהלך האימון אלא נקבע מראש (ReLU, Swish ובודומה).

המאמר מציע פונקציות אקטיבציה שהן (הפרמטרים שלהן) אשכלה נלמדות במהלך האימון. ד"א לאחרונה ראיינו דוגמא נוספת לפונקציית אקטיבציה נלמדת ראיינו לא זמן במאמר המפורסם Kolmogorov-Arnold network או KAN - שם אלו היו ספליינים נלמדים. במאמר המשוקר אימצו שיטה אחרת לבניה של פונקציות אקטיבציה נלמדות. הבניה נעשה דרך שדות וקטורים שマגדירות את המסלול של נקודה למרחב.

במקרה זה אנו מתחילה מנוקודה x ובעזרת נגזרת של כיוון תנועת הנוקודה (=שדה וקטורי) ב- t ("זמן") (שמהחיל ב- $t=0$ ומסתיים ב- $t=1$) נבנה המסלול של נקודה x . המסלול מסתיים ב- $t=1$ לכל x שימושה מגדרי לנו פונקציית אקטיבציה $\phi(x)$. ניתן לתאר את התקדמות נקודה באמצעות משווה אינטגרלית (כמו שיטת אוילר לפתרון משוואות דיפרנציאליות).

המאמר מתבונן במקרה של שדה וקטורי נתון על ידי פונקציה רציפה המורכבת מפונקציות אפיניות (lienarit mozgat) באינטראול נתון. פונקציית זו מכיל פרמטרים נלדיים המגדירים את הפונקציות האפיניות. ניתן להראות כי פונקציות אקטיבציה היוצאות מהתהליך זהה הם diffeomorphism, כלומר פונקציה גזירה בעלת פונקציה הופכית גזירה גם כן. פונקציות אלו נקראות CPAB. דרך אגב פונקציות אלו שמשו בעבר לטרנספורמציות "локליות" של דטה בסדרות זמן או של תמונות (למשל ל-time warping דינמי של סדרות זמן).

המאמר מציע לשכלל את פונקציית אקטיבציה שתיארנו קודם ומגדירים אותה לכל x ולא באינטראול נתון. הם מגדירים באינטראול "הרגיל" פונקציית אקטיבציה שהרחיבנו עליהם לפני תוכפל ב-ReLU (זה התפלגות קומולטיבית של גausian המכפל ב- $\phi(x)$) ובשאר האינטראול תהיה שווה ל- x . יש גם גרסה שבה במקומם של פונקציית אקטיבציה שווה לReLU מעבר לאינטראול שלו.

בנוסף יש איבר רגוליזציה על הפרמטרים של CPAB של פונקציית האקטיבציה המוצעת. כדי לזרץ את החישובים (הרי כל פעם צריך לפתור משווה אינטגרלית לכל אקטיבציה) המחברים מציעים לבצע קווניטוט ולהשך את ערך הפונקציה רק ב- $t=0$ נקודות באינטראול CPAB שלו.

מאמר המקורי כתוב היטב - הנהניתי לקרוא.

<https://arxiv.org/abs/2407.07564>

🚀⚡️המאמר היום של מייק 19.07.24: DataDream: Few-shot Guided Dataset Generation

זמן לא סקרתי מאמר בנושא של מודלי דיפוזיה גנרטיביים - הנושא האהוב עליו לפני שנה - שנתיים. המאים בנושא זהה השתו מАЗ ובד"כ לוקח לי קצת זמן לצלול לעומק. המאמר הזה הוויה יוצא מן הכלל והוא די קל עקב האינטואיטיביות שלו ובנוסף שימוש בטכניקות דומות בתחום מודלי שפה.

המאמר מציע שיטה מעניינת לבניית מסוג לביעות למידת few-shot דרך יצירה של דטה סינטטי (מכאן בהרעיון העיקרי של המאמר). כלומר יש לנו מודל דיפוזיה מאומן, כמה תמונות בודדות מכמה קטגוריות והמטרה שלנו לבנות מסוג לתמונות קטגוריות אלו. כאשר יש לנו מעט תמונות פר קטgorיה וגם הקטגוריות עצמן הן לא טריויאליות ושכיחות איז המשימה ההזעולה להיות לא פשוטה בכלל.

כאמור המאמר מציע לגנרט דאטה סינטטי ולאמן עליו את המסוג. הרעיון הוא ליצור דאטה סינטטי באמצעות מודל דיפוזיה מאומן שעובר פין טיין על התמונות המעוטות מהקטגוריות שיש לנו ביד. ואז אנו מאמנים את המסוג על התמונות האלו. הבעה עם הגישה זו היא שהתפלגות התמונות המוגנרטות לא תמיד קרובה להתפלגות האמיתית של הקטגוריות עצמן וזה המאומן עליון לא מגין ביצועים גבוהים.

המאמר מציע גישה נחמדה להתגבר (או לפחות להקל) על הסוגיה זו. המאמר מציע לבצע שני סוגי של פין טיין של מודל דיפוזיה מאומן (שodium ליצור תמונה מטוקסט) על התמונות שיש לנו ביד. הפין טיין הראשון הוא פר קטgorיה כלומר המודל לומד ליצור תמונה פר קטgorיה (שיצור N מודלים כאשר N זה מספר הקטגוריות) והשני D_all לומד ליצור תמונה מהדאטהסת (לא מקטgorיה ספציפית).

הפיינטוניים מתבצעים בצורה של LoRA כך לומדים מטריצת תוספות בעלות רנק נמוך למטריצות key, value, query ומטריצות S_W (המשלבת את הפלט של כל ראשי הטרנספוררים שיש לנו במודל דיפוזיה גנרטיבי). לאחר מכן יוצרים דאטהסת סינטטי גדול באמצעות $1+N$ המודלים שאומנו (המאמר לא מפרט איך מסוגים קטגוריות של התמונות המיוצרות על ידי D_all האומן על כל הקטגוריות).

בשלב האחרון לוקחים את מודל CLIP (מודל פופולרי של openai לפני openai) וועושים פין טיין באמצעותו אותה LoRA לאנקודר של תמונות ולאנקודר של טקסט שלו על הדאטהסת המכיל את התמונות האמיתיות והtamונות המוגנרטות. המטרה היא לקרב את הייצוגים של התמונות ושל הקטגוריות שלהן בהתאם לדאטה המתיוג.

מאמר נחמד וקל לקרוא.

<https://arxiv.org/pdf/2407.10910>

🚀⚡️המאמר היומי של מיק 20.07.24: Consistency Models

המאמר הזה חיכה את תורו די הרבה זמן, קצת פחות משנה וחצי המאמר הזה נכתב על ידי Song Yang האגד' (זה שכתב מאמרם חזקים מאוד בתחום הדיפוזיה) ואחד המחברים הוא איליה סלוצקי שארני מניח שאתם מכירים היטב. המאמר נכתב עוד בתקופה שנייה המדענים הדגולים אלו עבדו ב-openai. דרך אגב שני המחברים אחרים גם תרמו לא מעט בתחום המודלים הגנרטיביים ושניהם עבדו ב-openai במשך שנים למשך .2023

המאמר מציג גישה חדשה לאימון מודלי דיפוזיה גנרטיבי סטנדרטי מורכב מטהיליך קדמי ומטהיליך האחורי (forward & backward). בטהיליך הקדמי אנו מוסיפים רעש (בב"כ גauso) לדאטה באופן הדרגתית עד שפיסת דאטה הופכת להיות רעש. בטהיליך האחורי אנו מאמנים את המודל להסיר רעש בצורה הדרגתית גם כן. ככלומר המודל לומד מה הרושץ צריך להחסיר מהדאטה המורעש באיטרציה t כדי לקבל את הדאטה באיטרציה t-1. אחרי שהמודל מאומן לעשות זאת אנו יכולים להשתמש בו ולבנות פיסת דאטה מרעש טהור על ידי הסרה של רעש בצורה הדרגתית.

מה הבעה בטהיליך זהה? הוא עלול להיות די ארכט (צריך להריץ מודל כמספר האיטרציות) ויצאו לא מעט מחקרים שונים להקטין את מספר האיטרציות בלי לפגוע באיכות הדאטה המוגנרט. מודלים קונסיסטנטיים (consistency models) זה עוד ניסיון לתקן את הבעיה המעניינת הזו. בגודל הרעיון כאן הוא שעבור פיסת דאטה נתונה x_0 שלא משנה מייזו איטרציה t (=דאטה מורעש t_x) נתחיל את הסרת הרושץ בסופו של דבר אנו חיבים לחזור לדאטה המקורי x_0 .

המאמר מציע שתי שיטות לאמן מודל דיפוזיה קונסיסטנטי. השיטה הראשונה מניחה שיש לנו ביד מודל דיפוזיה מאומן (consistency distillation) והשני מאמן את המודל מאפס (consistency training). כדי להסביר את השיטה הראשונה צריך טיפה לצלול למתמטיקה אבל געשה את זה לאט ובהירות.

נתחיל מזה התהילה הקדמי המודל דיפוזיה מתואר על ידי משווה דיפרנציאלית סטוכסטית המתארת את היצירה הדרגתית של הדאטה המורעש. ניתן להראות שהמשווה דיפרנציאלית רגילה (ODE) ל t - x . מעניין ש-ODE זהה מכיל לוגריתם של פונקציית ההסתברות של הדאטה המורעש $t-x$ (נקרא score function או SF). המשווה מתארת את בתהילך האחורי (הסירה הדרגתית של רעש). אז אם יש בידינו שערור של SF אנו יכולים לשזרז (הדרגתית) את הדאטה שלנו על ידי הפתרון הנומרי (באייטרציות) של ה-ODE זהה (נגיד Euler-Maruyama).

הדבר וכי מגניב שאם יש לנו מודל דיפוזיה מאומן (שערור של הריש באיטרציה t) אז ניתן בקלות לקבל שערור של SF (בתנאי של רעש גauso).

אבל איך כל זה קשור למודלים קונסיסטנטיים שצרכים להיכנס לאותה הנקודה לא משנה מייזו איטרציה של הרעשה מתחילה. אנו מאמנים את המודל, באופן הבא: לוקחים נקודה מושעתה \hat{t} , עושים איטרציה אחת של הפתרון הנומרי של ODE (עם SF) כדי לקבל את הדאטה באיטרציה $1-\hat{t}$. תזכירו שהמטרה שלנו היא לאמן את המודל לשזרז את הדאטה הנוכחי מכל איטרציה של הרעשה. אז מאמנים מודל למצער את ההפרש בין הדאטה המשוחזר מאייטרציה \hat{t} לזו של האיטרציה $1-\hat{t}$. בגדול יש כאן שני מודלים (בדומה לשיטה של למידת הייזוג הנקראת BUOL). המודל הראשון המוחלק שהפרמטרים שלו הם ממוצע מעריצי של המשקלים של המודלים מהאייטרציות אימן הקודמות (לא מאומן - יש stop gradient) והוא נקרא target והמודל השני שהוא למעשה מעשה מאומן עם מورد הגרדיאנט.

ניתן גם לאמן מודל ללא מודל דיפוזיה מאומן ובמקרה זה יוצרים את $1-t-x$ על ידי הורדת הרעש. ברגע שאימנו מודלי קונסיסטנטי ניתן ליצור דאטה נקי מרעש טהור באיטרציה אחת אך זה לא תמיד אופטימלי. ניתן לבצע מה שנקרא במאמר Multistep Consistency Sampling (בזמןה לשיטת טהור, ליצור דאטה נקי, להוציא רעש, שוב ליצור דאטה נקי ולזרז עד שאיכות הדאטה הוא לשביעת רצוננו. המאמר טוען שנדרש משמעותית פחות אייטרציות בתהילך זה מאשר במודלי דיפוזיה סטנדרטיים.

סימנו, מקווה שלא איבדתי אותכם כאן...

<https://arxiv.org/abs/2303.01469>

🚀⚡️המאמר היומי של מייק 22.07.24: TRAINING DIFFUSION MODELS WITH REINFORCEMENT LEARNING

אוקי, בסירה הקודמת סקרה אחד בנושא מודלי דיפוזיה גנרטיביים וקיבלתי תיאבן בסקור עד כמה-Cal. אז בחרתי במאמר המגניב הזה שאחד מחברי הוא סרגיי לוין האידי (בנוסף למאמרים רבים יש לו קורס DI מטורף מבחינת העומק בנושא reinforcement learning deep). באופן לא מפתיע המאמר שנסקרו קשור ללמידה עם חיזוקים (או RL בקצרה) אבל יחד עם זאת מופיע בשם גם מודלי דיפוזיה.

לדעתי בעבר כבר סקרה אחד המאים שלו המשלב גישות מעולם RL לאימון מודלי דיפוזיה. מתרבר שניתן לפחות אין אימון במודל דיפוזיה עם כלים מעולם RL כלומר ניתן לבנות תהליך החלטה מרקובי (MDP) מאד אינטואיטיבי עבור מודל דיפוזיה.

כמו שאתם זוכרים אימון של מודל דיפוזיה מסתכם בנית מודל שמשערך את הרוש שהתווסף לדאטה באיטרציה t של התהיליך הקדמי (של ההרעשה הדרגתית של דאטה). אם יש לנו את האומדן של הרוש שהתווסף לפיסט דאטה באיטרציה t אנו יכולים לאמוד את הדאטה המורוש באיטרציה הקודמת $t-1$. כלומר אנו מאמנים מודל denoising לבנייה של דאטה מרוש טהור.

המאמר למעשה מצא פרימורק מעולם RL (לומר P) למידול של אימון מודל דיפוזיה גנרטיבי. בשביל כך נגידר את כל הפרמטרים של ה- MDP באיטרציה t באופן פורמלי:

- המצב (state): השליישיה {פרומפר, מספר איטרציה t , הדאטה המורוש t_x }
- הפעולה (action) היא t_{-1-x}
- הפליסי היא הסתברות לקבל t_{-1-x} ומהפרומפט C
- המצב ההתחלתי מוגדר על ידי השליישיה: {רוש גואו סטנדרטי (ממנו מתחילה denoising), הסתברות על מרחב הפרומפטים, האיטרציה الأخيرة T }
- פונקציית תגמול (reward) שהמאמר מגידר ככמה צורות. היא מחושבת באיטרציה الأخيرة (על התמונה המשוחזרת).

עכשו אחריו שיש לנו הגדרת RL של אימון מודלי דיפוזיה אנו יכולים להשתמש בשיטות RL קלאסית כמו REINFORCE או PPO למקסום של פונקציית התגמול.

לגביה פונקציית התגמול המאמר מציע כמה אופציות. האופציה הראשונה היא לחשב את מה שנקרא BERT Score שבודק כמה התמונה מתאימה לפרומפט שלה (שווים את האմבדינגו שלהם). האופציה השנייה היא להשתמש במה שנקרא LAION aesthetics predictor שאוןן לשערך עד כמה התמונה היא אסתטית (זו למעשה שכבה לינארית על האמבדינג של CLIP המאמן עד דאטהסט של תוצאות המתואגות על ידי בני אדם).

מאמר מעניין ויחסית לאקשה לקרוא.

<https://arxiv.org/pdf/2305.13301>

🚀⚡️: המאמר היומי של מיק 23.07.24 🚀⚡️ Feedback Efficient Online Fine-Tuning of Diffusion Models

משמעותם את הקו של אטמול וסוקרים עוד מאמר המשלב מודלי דיפוזיה עם טכניקות מעולים של למידה עם חיזוקים (RL). הפעם המאמר משלב את שני התחומים המרתקיים האלו כדי לבצע פיין טיון של מודל דיפוזיה. המאמר מתמקד במקרה שאין בידינו דאטהסט (לפיין טיון) אלא יש לנו דרך לשערך (סוג של reward) את איכותות פיסת דאטה מג'ונרט, לומר סוג של משוב על איכות הדאטה. למשל אם מתרתנו היא לאמן מודל לגנט מולקולות המשוב יכול להיות " מידת פעילות ביולוגית" (bioactivity) של המולקולה הנוצרת.

בגדול מאוד המאמר מציע לאמן מודל דיפוזיה מאומן (pretrained) למקסום של פונקציית התגמול (=המחשב) תוך כדי שמיירת של התפלגות הדאטה המוגנרט על ידי המודל קרוב יחסית לזה של המודל ההתחלתי. מזכיר לכם PPO ו-TRPO מעולם RL - אז זה בערך אותו הרעיון עם קצת סיבוכיות. התהיליך הוא איטרטיבי וכל איטרציה אנו מעדכנים את פרמטרי המודל (כאן זה רק המשקלים - יוסבר בהמשך) ויוצרים דאטה חדש עם המודל המעודכן.

למעשה התהיליך מורכב מ 3 שלבים עיקריים.

בשלב הראשון בונים דאטהסט חדש עם מודל דיפוזיה מהאיטרציה הקודמת (בהתחלת מתחלים ממודל מאומן (pretrained)). כמו שכתבתי ניתן לתאר מודל דיפוזיה מאומן על ידי משווה דיפרנציאלית סטוכסית עם אופיינים נלמדים (פונקציה נלמדת למשה עם שיטות כמו score matching או flow matching). למעשה ה-SDE זהה מתאר את תהליכי יצירתיות מושפעים מחריש ופוטרים את ה-SDE (עם פונקציה נלמדת התלויה בדעתה מושפע באיטרציה t וב- t עצמה). משתמשים בשיטות סטנדרטיות כמו אוילר או אוילר מוריאמה.

בשלב השני בהתבסס על הדעתה שיצרנו בשלב הקודם מאמנים מודל תגמול reward עם רגולרייזציה (נגיד 1) או כל פונקציה התלויה במאפייני ה-reward ובמשימה עצמה). בנוסף מאמנים מודל המשערך אי-וודאות של פונקציית תגמול. בגין מקרה זהה המטרה של הפונקציה היא שערוך של סוג של רוחם סマー של הפרש של פונקציית התגמול אופטימלית עם רגולרייזציה ופונקציית תגמול עצמה על הדאטהסט מהאיטרציה הקודמת(הפרטים קצת מרכיבים והעדפת לא לצלול בהם בסקירה).

בשלב השלישי של כל איטרציה מאמנים פונקציה חדשה f עבור ה-SDE שלנו וגם התפלגות ההתחלתית π של שמננה אנו מייצרים את הדעתה באמצעות ה-SDE. יש שם נוסחאות די מוכבות אך אנסה להסביר את ההיגיון מאחוריהם בכל זאת. פונקציית המטרה כאן מורכבת מ 3 איברים (מקסימים אחרות על הדאטהסט משלב 1). המקסום מתבצע ביחס לפונקציית f וגם על התפלגות ההתחלתית ממנה יוצרים את הדעתה באמצעות SDE:

1. התגמול האופטימייסטי (סכום של פונקציית התגמול ומודל אי-וודאות משלב 2).
2. איבר רגולרייזציה השומר את פונקציית f הנלמדת (מה-SDE) באיטרציה הנוכחיות (של האלגוריתם ולא של מודל דיפוזיה) קרובה מבחינת מרחק KL לפונקציית f מה-SDE של המודל ההתחלתי. בנוסף רצים לשמור את התפלגות הדעתה באיטרציה ההתחלתית הנלמדת π קרובה להתפלגות הדעתה ההתחלתית של המודל שהתחילה ממנו מבחינת KL. שני הקירובים הללו צריכים להתקיים מעל כל האיטרציות של מודל דיפוזיה (פתרון של ה-SDE).
3. אוטם איברי הרגולרייזציה עבור f ועבור π שלא "אפשר" להם לסתות יותר מדי מה- f ומה- π מהאיטרציה הקודמת של האלגוריתם עבור כל האיטרציות של מודל דיפוזיה.

מאמר קצת מרכיב מתמטי - מקווה שעזרתי לכם קצת להבין אותו.

<https://arxiv.org/abs/2402.16359>

🚀💡: 24.07.24. המאמר היומי של מיק The Empirical Impact of Neural Parameter Symmetries, or Lack Thereof

הסקירה היום קצרה וקלילה לעומת הסקירות האחרונות על מודלי דיפוזיה למיניהם. המאמר של היום חוקר סימטריות ברשותות נוירונים עמוקות. ניתן לראות די בקלות כי קיימות לא מעט פרמטריזציות של המטריות המשקלים בשכבות השונות של רשת שלמעה לא משנות את המודל. קלומר אם תפעלו את המודל אחרי פרמטריזיה על כל קלט תקבלו את אותה התוצאה כמו עם המודל המקורי.

אם הסימטריות הללו מביאות לנו משהו טוב? בכלל לא - לי זה נראה (למרות שאני לא מומחה גדול בתחום) כמו סוג של יתרות של יש במודלים שבצדיה אויל' ניתן היה להגיע למודלים קטנים יותר למשל. המאמר בוחן מה קורה במודל עם אנו מפרים את הסימטריה שיש במודל. אחת הדריכים להרוו את הסימטריה היא לקבוע משקלות (לערבים אקרים או קבועים) במקומות שנבחרו באקראי במטריצות משקלים של הרשת. הדרך השנייה היא להפעיל פונקציה אקטיבית רק על המשקלים מסוימים.

המאמר חוקר איזה אפקטים מתרחשים אחרי שהורסם את הסימטריה במודל ומגלה כמה דברים די מעניינים....
<https://arxiv.org/pdf/2405.20231.pdf>

🚀⚡️: המאמר היומי של מאי 24 AI models collapse when trained on recursively generated data

מאמר די חמוד שחוקר מה קורה למאמנים מודלי AI על הדadata הנוצר על ידי מודלי AI. בשתי מילims - לא הכל ורוד שם ויש כמה סיבות למה הדברים עלולים להשתבש:

1. דadata דרייפט (איך זה בעברית?) קיזוני: אימון מודלים על>Data שמנצירה על ידי מודלים אחרים גורם להתרחקות של התפלגות הדadata הנוצר על ידי המודל החדש מהadata האמיתית (כלומר ארגזיה של מרחק בין ההתפלוגיות שלהן)..
2. הביעות מחמירות בתפלוגות הדadata (תחומים או שפות עם מעט DATA למשל): הhydrodrotes משפיעה בעיקר על זנבות התפלוגות הדadata, שם DATA נדרה הופך להיות עוד פחות מיוצג
3. עוד יותר שגיאות: שגיאות בDATA שנוצרו על ידי מודלים מצטברות לאורך דורות, מה שMOVIL לירידה ממשמעותית ביצועים.
4. קリスト השונות: DATA שנוצר על ידי מודלים חסרים את המגוון והעושר של DATA מהעולם האמיתית, מה שMOVIL ליותר הומוגניות יתר (פחות גיוון).

<https://www.nature.com/articles/s41586-024-07566-y>

🚀⚡️: המאמר היומי של מאי 24 Questionable practices in machine learning

הסקירה היום תהיה ממש קצרה. המאמר המסור דן בפרקטיות פסולות שעולות להכשיל אתכם במהלך פיתוח של המודלים שלכם. רוב הפרקטיות הרעות שנזכרו במאמר נראות לחוקרי ML מנוסים די טריוויאליות ודין בירור למה לא כדאי להשתמש בהן. בין אלו ניתן למנות אימון על טסט טס, בחירה של ביסליין חלש להשוואה, הסקת מסקנות על אימון אחד בלבד של המודל, אימון על DATA דומה מאוד לבנץ' מאrk וכדומה. אבל ניתן למצוא גם דברים פחות טריוויאליים שחלקם לא ידעתם.

<https://www.arxiv.org/abs/2407.12220>

🚀⚡️: המאמר היומי של מאי 24 Data Mixture Inference: What do BPE Tokenizers Reveal about their Training Data?

אחרי שבוע שלא סקרתי עבדות על LLMs חזר לנושא זהה היום עם סקירה של המאמר המציע התקפה מציע תקיפה על מודלי שפה מבוססת טוקנייזרים. ההתקפה מיועדת לגלו מה המשקל היחסי של DATA מסווג מסוים (שפה, שפת תכנות וכדומה) בDATASET שלו או אומן מודל שפה. לא יודע עד כמה ההתקפה זו חמורה אבל עושים זאת על סמך הטוקנייזר.

אם אתם זוכרים הטוקנייזרים נבניהם על שילוב אותיות (מספרים, סימני פסוק הגדומה) היכי נפוצים בDATASET האימון. אם DATASET מורכב מכמה שפות אז הטוקנים שייבחרו יכול גם אותיות (ולפעמים מילים שלמות) מכמה שפות המופיעות בDATASET. בשיטת טוקנייזה מפורסמת הנקראת Byte Pair Encoding או BPE קודם כל

מפרצלים את הטקסט לביטים (bytes), מוחפשים זוגות בתים היכי נפוץ בדאטאסת, מאחדים אותם לטוקן חדש וממשיכים את התהיליך עד שmaguiim לגודל של מיליון הטוקן (50k-100k היום במודלי שפה מודרניים).

از המאמר מנצל את מבנה של ארכיטקטורת טוקנייזציה כדי להציג אלגוריתם המבוסס על התכונות הלינארית למציאת אומדן למשקל יחסית של הדאטאסטים השונים בסיס האימון של המודל.

<https://arxiv.org/abs/2407.16607>

🚀⚡️: 29.07.24 המאמר היום של מיק Large Scale Dataset Distillation with Domain Shift

המאמר מציע שיטה מעניינת ודי מקורית לגנרט דאטה מהתפלגות הננתונה על ידי דאטאסט מתואג. למשל בהינתן דאטאסט של תמונות s_D המטרה היא ליצור דאטאסט (מתואג) גדול בעל התפלגות ה"מושרה" על ידי $s_{D'}$. המחברים טוענים כי השיטות הקיימות מתקשות לבנות(distill) דאטאסט גדול המשקף בצורה נאמנה את המאפיינים המקוריים של s_D .

המחברים מציעים לגלות לבעה זו עם גישה מעולם של **domain adaption** או DA בקצרה. בגודל מאוד DA היא תהיליך של "התאמת מודל" במקרים בהם התפלגות הדאטה בזמן האינפרנס שונה מזו של הדאטה שעלה אומן המודל. התחום הזה עשיר בשיטות שחלקן די מורכבות מתמטיות ומערכות לרוב מינימיזציה של מרחק בין התפלגות הדאטה (KL וכאליה).

למעשה המאמר המסביר מתרגם את בעיית יצירת הדאטה לבניית DA. התפלגות הדאטאסט שאנו מגנרטים "מנמו" s_D משחק תפקיד של התפלגות המקור במקירה של DA (ועלוי מאמן המודל ב-DA) ואילו התפלגות הדאטה המגנרט משחקת תפקיד של התפלגות היעד t_D (כלומר זו של הדאטה שעלי מפעלים את המודל ב-DA). המטרה כאן לאומן מודל המקרב את התפלגותיות האל.

אבל איך נחשב את התפלגותיות האל? המאמר מייצג את התפלגותיות האל על ידי התפלגות של האקטיבציות של השכבות השונות של הרשת. בפשטות עבור הדאטאסט s_D אנו מייצגים את התפלגות הדאטה על ידי וקטור הממצאים ומטריצת קורייאנס של כל השכבות של המודל s_M (מניחים שהם גausים). בדיק באותו האופן אנו מייצגים את התפלגות של הדאטה המגנרט.

אבל מה כאן s_M ומה עושים כדי לקרב את התפלגות של הדאטה המגנרט להתפלגות הדאטה האמיתית? המודל s_M אומן לשערר את התפלגות של הדאטאסט המתואג s_D (המאמר לא מפרט איך s_M מאמן בדיק). למעשה האופטימיזציה מתבצעת על הדאטה המגנרט כאשר המודל s_M יותר ללא שינוי. ככלمر מתחילה מתחומות הנציגות באקראי עם הליבלים והמטרה היא לבצע מורד הגרדי-אנט(gradient descent) על התמונהות האל במטרה לקרב אותו להתפלגות של s_D .

עכשו נשאלת השאלה מפונקציית הלוס כאן. כאמור בשלב הראשון אנו מאפטמים את התמונהות המגנרטות במטרה למזער מרחק KL בין התפלגותיהם המשקלים המודל s_M (נותר ללא שינוי) של s_D (נותר קבוע לכל אורך הדרך) ובין התפלגות של משקלים המודל s_M עבור t_D . המחברים מניחים שתי התפלגותיות אלו הם גausים שעבורם מרחק KL ניתן לחישוב באופן מדויק בהינתם וקטורי תחולות ומטריצות קורייאנס של s_D ו- s_t עם s_M . איבר נוסף בלוס מנסה למקסם (=למזער עם סימן מינוס) הוא ההתפלגות המותנית של ליבל על בהינתן פיסת דאטה מגנרט (הרי אנו מגנרטים דאטה מתואג). התוצאות של כל פיסת דאטה מגנרטת נקבע מראש ולא משתנה במהלך האימון.

השלב השני הוא מזעור של מրחיק LK בין התפלגות המותנית של הליבלים של הדאטה המוגנרט לבין זה של הדאטה M-D. בשביל כך מנצלים את הדאטה המוגנרט מהשלב הראשון. מחשבים את התפלגות הליבלים עבור הדאטה המוגנרט זהה עם מודל S-M ומאפטמים את הדאטה המוגנרט במטרה לקרב את שתי התפלגות האלו של הליבלים (של הדאטה המוגנרט ושל הדאטה M-D).

יש עוד לא מעט פרטים מעניינים על איך בדיק מתבצע האימון (משתמשים ללא כמעט מודלים לחישוב סטטיסטיות המשקלים, עושים מיצוע מעריכים לסטטיסטיות של הביצים וכדומה). המאמר לא כתוב מאד ברור אבל הרעיון יפה.

<https://dl.acm.org/doi/10.5555/3692070.3693400>

🚀⚡️: המאמר היומי של מילק 30.07.24 ⚡️🚀 Denoising Vision Transformers

זמן לא סקרנו מאמר בראיה הממוחשבת והיום נתרען עם סקירה של מאמר ד' מעוניין מהדומין הזה. המאמר מציע שכלול ל-Vision Transformer או ViT בקצרה. משפחת ViT כוללת מודלים מבוסטי טרנספורמרים המיועדים לעיבוד דאטה ויזואלי ולהפקת ייצוג חזק של תמונה. מה אני מתכוון כאשר אני אומר ייצוג חזק של תמונה? למעשה זה ייצוג (לטנטי) של תמונה, בעל מידת שימושית נמוכה יותר מהתמונה עצמה בד"כ, שנייתן לנצלו לאימון מודלים לוגון משימות downstream (כגון סגמנטציה, זיהוי אובייקטים, סיוג וכדומה).

המאמר טוען שנייתן לשפר את היצוגים המופקים על ידי ViT באמצעות נקיי רעשים הנוצרים בגל השימוש ב-position encoding או קידוד תליי מיקום. מטרתו של קידוד תליי מיקום היא להברר למודל מידע על מיקום של הפיצאים של התמונה. אזכיר כדי להזין תמונה ל-ViT אנו מפרקים אותה לפיצאים, משתמשים אותם ומציאים אותם למודל. לוקטור המיציג כל פיצ' אנו מוסףים (אשרה מחברים) וקטור המקודד את מיקומו היחסי בתמונה של הפיצ'.

המאמר טוען שהוקטוריים המקודדים מיקום מריעשים את ייצוג הפיצאים ומקשים על שימושם למשימות downstream. לטעתה המחברים רושם המתווסף לייצוג הפיצאים מכל מידע על המיקום של הפיצ' בלבד ולא מכל שום מידע על התוכן של הפיצ'. לעומת זאת שני החלקים האחרים בייצוג הפיצ' מכילים מידע על התוכן הסמנטי של הפיצ' והשני מכיל מידע המערבב את ייצוג התוכן וייצוג המיקום. המחברים טוענים שנייקוי הייצוג מהרעש המידע על המיקום בלבד תורם לעוצמתו של הייצוג.

כדי לאייר את הארטיפקט המיקומי הזה בייצוג הפיצ' המאמר מציע לאמן מודל המזהה את שלושת החלקים של הייצוג שהזכרנו בפסקה הקודמת. זה נעשה עלי ד' אוגמנטציה של תמונה (הזהה, קרוב וכדומה) דרך ניצול התכונות האינהרטניות של הרעש המיקומי ושל הייצוג התוכן. כמובן המידע המיקומי בייצוג זהה יחד עם הפיצ' כאשר המידע המיציג את התוכן לא משתנה אם מזיזים את הפיצ' בתמונה. החלק שמעורבב את המידע על המיקום והתוכן היא פשוט הפרש בין ייצוג של ViT לבין סכום של שני החלקים האחרים.

בשלב השני מאמנים מודל המזהה את הרעש המיקומי בייצוג הפיצ'. לאחר מכן באינפנס מחסירים את הרעש זהה מהייצוג של הפיצ' וכך קיבל ייצוג יותר נקי ועוצמתי.

<https://arxiv.org/abs/2401.02957>

🚀⚡️: המאמר היומי של מילק 31.07.24 ⚡️🚀 Denoising Vision Transformers

DENOISING DIFFUSION IMPLICIT MODELS

זה מאמר לא חדש (אוקטובר 2022) אך חשוב מאוד בתחום של מודלי דיפוזיה. מאמר עם רעיון מאוד אלגנטiy המלווה במתמטיקה די רצינית. אנסה לסקור אותו קצרות כי כאמור יש בו עומק מתמטי לא קטן אף עדין ניתן להבהיר את הרעיון העיקרי בלי לצלול יותר מדי לעומק.

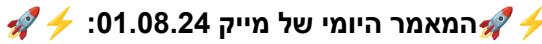
כמו שאותם זוכרים במודלי דיפוזיה גנרטיביים יש לנו שני תהליכיים: הקדמי והאחוריו. תהליך הקדמי הוא הרעשה הדרגתית של>Data והתהליך האחוריו הוא הורדת הדרגתית של הרעש מהדטה באמצעות מודל שיאומן לצורך זה על דאטהסט מסוים. למעשה מודל זהה מאפשר ליצור>Data טהור בצורה הדרגתית. הבעה בתהליך הזה מבון זה הזמן הזה לוקח כי צריך די הרבה איטרציות של *denoising* כדי להגיע מרעש לדאטה איקוטי.

המאמר מציע דרך להקטין את מספר האיטרציות בדרך די מקורית. כמו שאותם זוכרים בתהליך ההרעשה (הקדמי) במודלי דיפוזיה רגילים הוא מרכיבי, כלומר הדטה באיטרציה t מוגדר (מבחינת התפלגות) על ידי הדטה המורעש מאיטרציה $1-t$ בלבד כל. המאמר הורס את ההנחה זה ומגדיר תהליך קידמי לא מרכיב כאשר הדטה באיטרציה t מוגדר לא רק על ידי הדטה באיטרציה $1-t$ אלא גם על ידי הדטה הנקי ($0-x$).

הנחה זה מאפשרת לנו להגדיר תהליך דטרמיניסטי של $1-t_x$ מ t_x באמצעות מודל שמאומן לשערך $0-x$ (הdatapoint התחלתי t_x). לעומת כל איטרציה אנו קודם כל משערכים את $0-x$ באמצעות המודל ולאחר מכן בונים בצורה דטרמיניסטי אנו מחשבים $1-t_x$ מ $0-x$ המשוער.

אבל איך זה בעצם כאשר לזרוץ של תהליך יוצרת הדטה? מתרבר שעורך של t_x דרך שעורך $0-x$ מאפשר להקטין משמעותית את מספר האיטרציות וככה נוצר מהם יותר.

מאמר מאד מעניין - הסברתי אותו ממש בגודל, חוברת קרייה לכל מי שאוהב מודלי דיפוזיה גנרטיביים.
<https://arxiv.org/pdf/2010.02502.pdf>



המאמר היום של מיק 01.08.24:

IMPROVED TECHNIQUES FOR TRAINING CONSISTENCY MODELS

היום סוקרים קצרות עוד מאמר בנושא קרוב לילבי - המשך של המאמר שסקרנו לפני בשבוע הנזכר "consistency models". אם אתם זוכרים מודל קונסיסטנטי הוא שיר למשפחת מודלי דיפוזיה (כלומר הוא מתואר על ידי משוואת הדיפוזיה). אחת הבעיות של מודלי דיפוזיה קלאסיים (כמו DDPM) היא איטיות של גנרטות הדטה. הדטה נוצר באמצעות תהליך *denoising* הדרגתית - מתחילה עם רעש גausian ומסירים אותו לאט לאט.

כדי להתמודד עם הבעיה זו הוציאו כמה שיטות ואחת מהן DDIM סקרנו אתמול. השנייה היא מודלים קונסיסטנטיים(CM) שנitin להציג אוטם כי מודל דיפוזיה קלאסי. בעיקר ב-CM אנו מאמנים מודל להסיר רעש מכל פיסת הדטה מושפע באיטרציה t כך שהתוכאה תמיד תהיה פיסת הדטה מקורית (לא רעש). מכאן בא שם של המודל: קונסיסטנטי.

איך זה נעשה נעשה? יש שתי דרכי עיקריות לאמן CM. דרך אחת מסתמכת על מודל המשערק את מה שנזכיר *score function* שהוא לוגריתם של פונקציית ההסתברות של הדטה המורעש באיטרציה t . ידוע כי תהליך גנרטות של הדטה במודלי דיפוזיה (*denoising*) מתואר על ידי משוואת זרימה (דיפרנציאלית) שמתאר את המסלול של הדטה מהרעש עד הדטה הנקי. ו- *score function* מופיע במשוואת זרימה זו. אז

השיטה הראשונה מזערת את המרחק בין שערוך x (הדатаה המקורי) מ- t לערך של x מ- t כאשר t מוחשב ממשוואת הזרימה (אייטרציה אחת של אוילר של ממשוואת הזרימה שכבר הזכרנו).

דרך אגב שערוך של score function די שקול לשערוך של הרעש הנוסף (לדатаה) במודול הדיפוזיה הסטנדרטיים. הדרך השנייה "לייצור" את x היא לשערך את x מ- t ולוסיף רעש (כמו באיטרציה t).

המאמר המקורי על CM השתמש במרקח הנקרא LPIPS המודל דמיון סמנטי בין התמונות (דרך השוואת EMA) אקטיבציות של מודלים מאומנים על דאטסהטים ענקים של תמונות). המאמר המקורי גם השתמש ב-EMA (החלקה מעריכית) של משקלי המודל בתור המודל עבור t . יש כמובן חישבות לבחירת השונות של האיטרציות.

از המאמר שסטודנטים היום משפר את תהליך האימון. השני הראשון הוא משקל של המרחקים כפונקציה של אייטרציה t ; ככל שמתקרבים ל-0 המשקל עולה. דבר שני זה שני של פונקציית מרקח-M-LPIPS לפונקציית הובר (Huber) עם טויסט קטן. הדבר האחרון והמשמעותי הוא ביטול של EMA ל- t אולם הרשוואה מתבצעת בין שני מודלים "טהורים" $-t$ ו- $+t$. גם היפר פרמטרים אחרים עברו שינוי למשל השונות של הרעש באיטרציות.

בקיצור יש לנו כאן שכלל מעניין של CM - בקרוב אסקרו עוד מאמרים על זה...

<https://arxiv.org/abs/2310.14189>

🚀⚡️: המאמר היומי של מיק 24.08.2024 🚀⚡️

NEFTUNE: NOISY EMBEDDINGS IMPROVE INSTRUCTION FINETUNING

הסקירה הזאת הולכת להיות קצרה במיוחד. זוכרים שאחרי אימון מודל שפה אנו עושים לו מה שנקרה instruction fine-tuning. ככלומר אנו מאומנים מודל שפה לעקב אחריו הוראות המשתמש. בשביל זה בונים דאטסהט של שאלות ותשובות רצויות ולאחר מכן מטיבים (שם נוסף לפין טוון) את המודל על הדאטסהט הזה עלי' חיזוי של טוון הבא להתשובה. המאמר מציע להויסף רושץ לייצוג הטוקנים המופקים עלי' ידי המודל באימון. כלומר אחרי כל מיניבאץ' מعتبرים את הטוקנים של השאלה והתשובה (אחד אחרי השני), מוסיפים רושץ יוניפורמי בין-1 ל-1 לאמבדיניג ומשיכים לאמן. לא ברור אחרי איזה שכבה מוסיפים את הרעש (לדעתי יש שהוא ב-ablation).

יש לי תהושה שהרעיון הזה לא חדש אך לפני המאמר הזה לא השתמשו בו - instruction fine-tuning

<https://arxiv.org/abs/2310.05914>

🚀⚡️: המאמר היומי של מיק 24.08.2024 🚀⚡️

Consistency Models Made Easy

כבר דיברנו רבות על מודלים קונסיסטנטיים (Consistency Models) או CM שהם בעצם שיפור של מודלי דיפוזיה גנרטיביים. בגודל יעד האימון של CM הוא למצער הפרשיות בין חיזוי של פיסת דатаה נקייה למפיסות דатаה מורעשות איטרציות עוקבות. כלומר לוקחים פיסת דטה מורעשת מאיטריה ו-ומאיטריה $+1$, חוזים את x משניהם ומאמנים את המודל להגעה לאותה התוצאה. מכאן בא השם - Consistency Models.

המאמר מציע להכליל את השיטה הזאת לא רק לאיטרציות עוקבות ו- $+1$ אלא לחיזויים מפיסות דטה משתי איטרציות כלשהן t ו- s . ד"א המאמר מציג את ביצורה קצרה מרכיבת - מסמן חיזוי מאיטרציה t בתור t ו- y והגזרת של t y לפי t צריכה להיות 0 ומאמנים את המודל על דיסקרטיזציה של המשוואה הזאת בرمמות שונות.

אבל כאמור הכל מסתכם למצוור של ההפרשים בין החיזויים עבור איטרציות \hat{z} ו- \hat{z} שונות במהלך האימון עבור \hat{z} ו- \hat{z} נבחרו באקראי. כל הפרש כזה ממושך ביחס הפוך לריבוע של $\hat{z} - \hat{z}$ (זה הגיוני כי רעש קרובות צרכות להסתכם בחיזויים קרובים ממש). עוד פרט חשוב: מתחילה את האימון ממודל דיפוזיה מאומן (למשל מ- DDIM).

<https://arxiv.org/pdf/2406.14548>

🚀⚡️המאמר היומי של מייק 05.08.24: ⚡️🚀

Improving Text Embeddings for Smaller Language Models Using Contrastive Fine-tuning

חזרים לסקור מאמרים קלילים על מודלי שפה והיום בפוקו מודלי שפה קטנים. המאמר שנסקרו קצרות היום מציע שיטה לשיפור ייצוג של טקסט המופק על ידי מודל שפה קטן. ידוע שמודל שפה קטן (במאמר שיפורו את הייצוגים של הדקווידרים) לא תמיד מצטיין ביצירה של ייצוג (אמבדיניג) עצמאי של טקסט - פשוט בגל הגדל expressiveness נמוכה יחסית.

از המאמר מציע להשתמש בשיטת למידה ניגודית (contrastive learning) כדי לשפר את הביצועים. בגדיול למידה ניגודית מאמנת מודל (לייצוג דатаה) במטרה לקרב פיסות דатаה (למשל תמונות או טקסט) שהן קרובות (סמנטיות או בעלות אותה שימושות) ובאותו הזמן להרחיק את הייצוגים של פיסות דатаה לא דומות. השיטה הוצגה ב- 2018 על ידי Oord האגדי.

המאמר מציע להשתמש בלמידה ניגודית כדי לעשות פיין טיון לייצוג הדטה המופקים על ידי מודל שפה בפרט הפלט של השכבה האחורונה עבור טוקן EoS המשמן את סוף המשפט. עדכון משקל המודל געשה כМОון עם LoRA על דאטאסת המכיל משפטים בעלי משמעות קרובה וגם זוגות משפטים רוחקים סמנטיית. המחברים טוענים שהוא משפר את יכולת הייצוג המופק על ידי המודל למספר שימושות downstream (בפרט סיווג).

מאמר קלילי, ונועים לקרוא....

<https://arxiv.org/abs/2408.00690>

🚀⚡️המאמר היומי של מייק 06.08.24: ⚡️🚀

TurboEdit: Text-Based Image Editing Using Few-Step Diffusion Models

חזרים לסקור מאמרים על מודלי דיפוזיה עם מאמר חחול לבן של קבוצת חוקרים מאוניברסיטת תל אביב. הם מציעים שיטה מעניינת לעריכה מהירה של תמונה. ככלומר בהינתן תמונה עם פרומפט נתון c אנו רוצים ליצור תמונה עם פרומפט אחר c' .

כמו שאתם זוכרים מודלי דיפוזיה מגנרטיבים תמונה על ידי הסרה רעש הדרגתית (denoising). בכל שלב המודול חוזה כמה רעש צריך להסיר מהתמונה והרעש המשוערך זהה מחוסר מהתמונה המורעשת באיטרציה הקודמת. השיטה פשוטה לעשות עריכה של תמונה היא:
- להחסיר מהתמונה(המקורית) באיטרציה \hat{z} את הרעש הזה המשוערך עם פרומפט c (כמו שעושים כאשר אין עריכה)
- להוסיף אל התוצאה את התוחלת המשוערת של התמונה המורעשת(המקורית) עם הפרומפט c' החדש (עם התמונה המורעשת המקורית).

כלומר בכל איטרציה מתקנים את הסרת הרעש בכיוון הפרומפט החדש.

דרך אגב ניתן שערוך הרעש הנוסף באיטרציה t וושערך תוחלת התמונה אחרי הסרת הרעש אלו שתי בעיות שקולות, ככלומר אחת מהן היא פשוט רפרמטריזציה של השניה מבחינת השערור.

הבעיה בשיטה פשוטה לעירכית תמונות שהיא לא עובדת טוב ויוצרת ארטיפיקטים בתמונה הערכאה. המחברים מוצאים מחקר קודם שמצא שהסקיל של הרעש (כלומר הפרש בין התמונה המורעשת לתוחלתה) לא מתנהג לפי הסקל של התהיליך הקדמי של הדיפוזיה של התמונה המקורית (שבו מוסיפים רעש עם שונות עליה לתמונה עד צזו הופכת לרעש טהור). הרעש שנוצר במהלך ערכאה צזו הוא בעל שונות משמעותית גדולה יותר מאשר זה של התמונה המקורית.

از המחברים מציעים להחסיר מהתמונה המורעשת המקורית באיטרציה t את שערוך התוחלת של התמונה המורעשת עבור האיטרציה $d+t$ עבור d חיבוי שהם מצאו. ככלומר לוקחים תמונה $t-x$ ומזינים אותה למודל שערוך התוחלת עם מספר איטרציה $d+t$. בסוף מכונים את התמונה עם שערוך תוחלת המשוערכת של התמונה הערכאה עם איטרציה $d+t$.

בנוסף המאמר מציע דרך מעניינת לווסת את "עוצמת הערכאה" בצורה דומה ל classifier guidance כדי לכון את התוצאה של מודל דיפוזיה גנרטיבי ללא פרומפט עבור פרומפט נתון. הפעם על ידי ניתוח של נוסחת הערכאה המחברים משקל שן מרחוק cross-prompt (הפרש שערוך התוחלת עבור התמונה הערכאה המורעשת עבור פרומפטים c ו- c_1) לבין מרחוק cross-trajectory שמודד הפרש בין חיזוי התוחלת בין התמונה הרגילה לתמונה המשוערכת). משקל צזו מאפשר לבצע את הערכאה בפחות איטרציות denoising.

מאמר כתוב יפה ובהחולט מומלץ

<https://arxiv.org/abs/2408.00735>

🚀⚡️ המאמר היומי של מייק 07.08.24: Language Model Can Listen While Speaking

המאמר שמשר את תשומת ליבי בגלל שמו הקליט. המאמר מציע ארכיטקטורה של מודל Speech Language Model או SLM שיעוד להקשיב תוך כדי שהוא מדבר, ככלומר מודל duplex full (מושג בתחום התקשורות). בדרך כלל ל- SLM יש שני מטרים עבודה: הקששה או דיבור, ככלומר המודל או מדבר או מקשיב. המאמר מעשיר את מרחב היכולות של SLM ומציג אותו ביכולת להקשיב תוך כדי שהוא מדבר. מעניין שהמודל גם יכול לעזר או הוא מזהה שיש דיבור (לא רעש) ומגיב עליו (בדיבור) לאחר מכן.

הארQUITקטורה של המודל המוצע LSLM מורכב מרכיבים סטנדרטיים. יש מודל שקולט אות דיבור, מחלק אותו לטוקנים (האות במקטעי זמן שונים) מכוון לוקטור אմבידיגג ומאזין אותו לדקودר. תפקוד הדקודר הוא לקחת בחשבון את ייצוג של טוקני הדיבור שנקלטו קודם וגם יציג טוקני הדיבור שנוצרו על ידי המודל כדי ליצור את הפלט הבא (אות הדיבור) של המודל. כאמור לפעמים הדקודר מחייב שהוא צריך לעבור למצב האזנה ולפעמים הוא צריך לעבור למצב הדיבור.

כלומר הדקודר במקרה זה הוא vocoder המקבל קולט את אות הדיבור הנקלט בנוסף לאות הדיבור המוגברת על ה-vocoder עצמו לפני.

<https://arxiv.org/pdf/2408.02622.pdf>

🚀⚡️ המאמר היומי של מייק 08.08.24: Language Model Can Listen While Speaking

Masked Attention is All You Need for Graphs

היום סוקרים מאמר בנושא של גרפים, ומוכיחו שני סוקרים מאמרם על למידה عمוקה המאמר זהה יהיה על רשותות عمוקות על גרפים או GNN. המאמר מציג גישה אלגנטית להפקת ייצוג (כלומר אמבדינג) של גרף וגם להפקת ייצוגם של צמתים הגרף או קשיותו.

הגישה שהמאמר מציע הינה די פשוטה והייתי קצת מופתע שאף אחד לא עלה על זה קודם. למעשה המאמר מציע למסך (כלומר להעלים מהגרף) חלק מהמאפניהם שלו. דרך אחת למסך (ברמה של צמתים) היא לאפס איברים מסוימים במטריצה שכניות (adjacency matrix) של הגרף (המתארת קשרים בין צמתים) או איברים ממטריצה שכניות של הקשת (node adjacency matrix) המתארת קשיות שיש להם צומת משותפת.

בשני המקרים המטריה היא לחזות את האיברים הממוסכים. המאמר משתמש בארכיטקטורה של set transformer (הרוי בגרף אין חשיבות לסדר הצמתים וקשותות). הם לקחו ארכיטקטורת טרנספורמר מרובה ראשים ד' סטנדרטית למשימה זו. הארכיטקטורה מורכבת מהאנקודר ומהדקודר (encoder-decoder) אשר ליצוג הגרף אנו משתמשים באנקודר ועבור ייצוג הקשות והצמתים משתמשים באדקודר. <https://arxiv.org/abs/2402.10793>

🚀⚡️ המאמר היום של מאי 24: 09.08.24

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

בטח שמעולם על חוקי הסקלינינג של מודלי שפה. חוקים אלו מיועדים למציאת "קונפיגורציה" אופטימלית לאימון מודלי שפה. חוקי סקלינינג מקשרים ערך של פונקציית לוס (ניתן להגדיר אותו בכמה אופנים) שנייתן להשיגו עבור גודל מודל, גודל סט האימון וכמות משאבי החישוב (FLOps) המוקזית לאימון.

המאמר שואל האם ניתן לנוכח חוקי סקלינינג דומים עבור האינפראנס, כלומר מה הביצועים המקסימליים שנייתן להפיק בהינתן כמות משאבי חישוב נתונה. הרוי יש כמה שיטות לבצע אינפראנס של מודל השפה ויש כמה פרמטרים חשובים של האינפראנס המשפיעים בצורה משמעותית על הביצועים. למשל יש שיטה הנקראת beam search שיוצרת בכל חיזוי של טוקן M סדרות טוקנים בעלי נראות (likelihood) הגבוהה ביותר. קיימות שיטות משאבי חישוב בכל חיזוי של טוקן מ- M סדרות טוקנים בעלי נראות (likelihood) הגבוהה ביותר. קיימות שיטות beam search עם מספר הסדרות השמורות לא קבוע ותלויה במספר הטוקן המוגנרט.

יש שיטות איטרטיביות אחרות כמו במאמר "Consistency LLMs" שסקרטטי לפני כמה שבועות. הוצעו גם שיטות שימושicas את "איך" התשובה המוגנרטת (עם מודל מואמן נוספת) שמאפשר לבחור את התשובה הכי טובה מכמה תשובות מגנרטות (או להפסיק את יצירת התשובה אם רואים שהיא לא "נכיה"). כל שיטה כזו דורשת משאבי חישוב שונים שתלויים גם בהיפרפרמטרים של השיטה.

מה השיטה העדיפה לרמת ביצועים אופטימלית בהינתן תקציב חישוב נתון (FLOps) - זו השאלה שהמאמר מנסה לענות עליה ויש תוצאות מעניינות (לדעתי)

<https://arxiv.org/abs/2408.03314>

🚀⚡️ המאמר היום של מאי 24: 10.08.24

Synthesizing Text-to-SQL Data from Weak and Strong LLMs

הסקירה של היום הולכת להיות די קצרה וקלילה. המאמר מציג שיטה די אינטואטיבית לאמן מודל שפה קטן לביצוע משימה מסוימת. במקורה שלנו המשימה היא גנרטו של שאלות SQL לפי תיאורה הטקסטואלי ומבנה (schema) של הטבלה. מודלי שפה קטנים יכולים להסתמך עם המשימה זו בטעון במרקם שהשאילתת הנדרשת אינה טריומילית.

המאמר מציע תהליך דו שלבי של אימון מודל קטן למשימה זו. בשלב הראשון יוצרים דאטהסט עבור המשימה זו באמצעות מודלי שפה גדולים וחזקים וכמה דאטהסטים רלוונטיים. עושים דברים רגילים, הנדסת פרומפטים קלה וכו'לו. לאחר מכן עושים למודל הקטן פין טין על הדאטהסט זהה.

בשלב השני עושים למודל השפה הקטן מודל אימון שלו כשלב אימון מודלי יסוד (foundational). היתרון של שיטה זו היא בכך שהיא לא דורשת אימון של מודל reward. בשבייל אימון מודל זהה אנו צריכים דוגמאות טובות ודוגמאות לא טובות. דוגמאות טובות יש לנו מהשלב הראשון.

בשביל לבנות את הדוגמאות הרעות לוקחים את המודל הקטן המתקיים על השלב הראשון כדי לגנרט שאלה של SQL לטיור טקסטואלי נתון. לאחר מכן מרכיבים את השאלה כדי לוודא אם התוצאה המתבקשת נכונה. אם היא לא נכונה קיבלים דוגמא שלילית. ככה בונים דאטהסט של דוגמאות חיוביות ושליליות ומה שנוור לעשויות הוא .PPO

<https://arxiv.org/abs/2408.03256>

🚀⚡️: 12.08.24⚡️🚀

Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models

מודלי דיפוזיה גנרטיביים הגיעו לתוצאות מרשימות לאחרונה והפגינו יכולת לגנרט תמונות באיכות מרתקבה. למרות זאת מודלים אלו מתקשים לפעמים במשימות של עריכת תמונות ולא מצליחים להחליף אובייקטים לא גדולים בתמונה תוך שמירה של כל המאפיינים האחרים של התמונה.

המאמר המשוקר מציע שיטה לייצרת דאטהסט של זוגות תמונות שכל זוג מכיל תמונות זהות פרט לאובייקט אחד בתמונה. כל זוג תמונות מלאה בתיאור של האובייקטים שהוחלפו בשתי התמונות וגם במיקומם בתמונות. בין השאר דאטהסט זה יכול לשמש חוקרים ומהנדסים לאימון מודלים לעריכת תמונות.

איך הם עשו זאת? האמת הפינלי שלהם די מורכב מכיל הפעלה לא מעט מודלים מולטימודליים, מודלים ליזיוי ותיאור אובייקטים בתמונה כמו LLaVA, FastSAM, BLIP, CLIP וכדומה. נתאר רק את ה 3 השלבים של התהליך.

בשלב הראשון לוקחים כמה עשרות אלפי תמונות מהדאטהסט הידוע COCO MS ויצרו זוגות של תמונות דומות על ידי החלפה של אובייקטים מסוימים אחרים בתמונה עם המודל שנקרא ViCUNA (החלפה עצמה בוצעה עם המודל הנזכר AxInstructPix2Pix).

בשלב השני אנו מפעילים כמה מודלים מולטימודליים כדי לזהות את האיזורים בתמונות שעברו שינוי (בזוגות מהשלב הראשון). קודם כל המחברים את התמונות ללא דומות עם CLIP (כלומר בהתבסס על דמיון של יצוג התמונות). לאחר מכן שוב מפלטים את הדאטהסט על ידי התאמת של תיאורם של האובייקטים והמצאותם בשתי התמונות עם BLIP. בסוף מזמינים את מקום האיזורים בתמונה שהם הוחלפו האובייקטים (כלומר bounding boxes שלהם).

בשלב האחרון מפיקים תיאור טקסטואלי של כל החלפות של בוצעו בתמונה הראשונה בזוג שהפך אותה לתמונה השנייה בזוג. עושים זאת עם שילוב של LLaVA ו-CLIP.

וככה מקבלים דאטסהט איקוני של זוגות תמונות דומות שמה שהשניינו ביניהם מתואר על ידי התוצאה של השלב האחרון (כולל מקום השני).

<https://arxiv.org/abs/2408.04594>

🚀⚡️: 13.08.24 ⚡️🚀

Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2

בזמן האחרון התחלתי להתעניין בשיטות interpretability של מודלי שפה גדולים בעקבות כמה בלוגים מאוד מעוניינים של אנתרופיק, OpenAI ולאחר מכן גוגל בנושא זהה. המטרה כאן היא לשפר קצט אוור על הקופסא השחורה שנקראת LLM - הרי אנחנו לא באמת מבינים איך הם עובדים ומה גורם להם לפולוט תשובה כזו או אחרת לפורמפעט שלו.

از המאמר הזה חוקר אחת השיטות המנסות להבין איך מודל שפה מייצג קונספטים סמנטיים שונים. המאמר עושה זאת דרך חקר של אקטיביזיות הנירונים בשכבותיהם השונות של מודלי שפה. עקב לכך שיטה זו משיכת למשפחת שיטות המכוניות mechanistic interpretability ל- SAE או נקרא AutoEncoders.

אז מה הרעיון העיקרי ב- SAE? אנו מנסים להציג אקטיביזיות של שכבה מסוימת של LLM על ידי וקטור אורך הרבה יותר מוקטור האקטיביזיות אך מאוד דليل. ככלומר וקטור-Ch-MMD של האקטיביזיות אנו מייצגים (עם SAE) עם וקטור באורך $\gg M$ אך בווקטור האורך הזה יש פחות מ- χ איברים לא שווים לפחות (دلילות). SAE במקורה הזה פשוט מאד: שכבה אחת לנארית עם אקטיביזציה לא לנאריתenganquod (של SAE) ושכבה אחת של דקודה. המטרה כמובן לאמן את SAE כך שהיא ניתנת לשחזר את האקטיביזיות המקוריות מייצוגם (אחריenganquod).

אבל למה זה בכלל חשוב ואיך זה קשור ל-interpretability של LLMs. הנתה מוצאת זו (הבלוג של אנתרופיק מדבר על זה בהרחבה) שכל נירון (או קבוצת נירונים) בשכבה (מסויימת) הוא "נדלק" (מקבל ערכיים) על כמה קונספטים לא קשורים (נגיד לב, מכונה ורפל). ככלומר הוא סוג של תערובת עבור כמה קונספטים. אז הייצוג המופיע על ידי SAE הוא למעשה מהו יציג של כל קונספט (disentangled). ככלומר עבור כל קונספט המקודד קבוצות נירונים שונות בוקטור הדليل הזה.

אז מה המאמר הזה עושה? הוא מנסה לאתר שכבות שבהם SAE מאמון עם שגיאת שחזר מינימלית (עם רגולריזציה מתאימה) ככלומר הוא מנסה להבין איזו שכבה ב-LLM (וגם בשכבות הפנימיות של בלוקי הטרנספורמר) מקודדת הכי טוב את הקונספטים הסמנטיים.

בימים הקרובים עוד כמה סקירות בנושא המրתק הזה.

<https://arxiv.org/abs/2408.05147>

🚀⚡️: 14.08.24 ⚡️🚀

Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders

אתמול סקרנו מאמר שהשתמש בגישה SAE או Sparse AutoEncoders כדי לחזור ל"מחשבותיו" של מודל שפה גדול דרך האקטיבציות של הנוירונים שלהם. הנחתה היסוד במאמר הייתה כי נוירונים "מגיבים" לכמה קונספטים שונים ונitin לאמן SAE רודף מאד (שכבה אחת בדקודר ושכבה אחת באנדוקר) כדי להגיע לוקטור דليل המקודד (נדלק) קונספט אחד בלבד לעומת disentanglement של הפיצ'רים למינרונים ייעודיים.

כما אמר יש באנקודר של SAE שכבה לינארית אחת עם פונקציית אקטיבציה הנקראת $\text{ReLU}_{\text{Jump}}$ שראיתי אותה בפעם הראשונה במאמר זהה. פונקציה זו היא בעצם הזזה של ReLU ביציר X וביציר Y בפרמטר ϵ נלמד (במאמר זה נקרא טטה). הטענה במאמר זהה מאפשרת למדוד את הייצוג הדليل של DATA על ידי האנקודר יותר טוב של פונקציית ReLU בגל שהוא מאפשר לפחות הקייבציות בצורה "נלמדת יותר מ- ReLU ".

עכשו נשאלת השאלה איך אנחנו יכולים לדليلות על ייצוג DATA (אחרי האנקודר). בעבודות קודמות השתמשו ב- L_1 בשביל כך אך כאן המחברים משתמשים באותו $\text{ReLU}_{\text{Jump}}$ כדי להפוך את איפוס האיברים בייצוג יותר נלמד. ושימו לב ש- $\text{ReLU}_{\text{Jump}}$ בא עם פרמטר נלמד זהה של האנקודר עצמו זהה עוזר לאסוף דليلות על הייצוג.

יש עוד טרייך אחד קטן ולא מאוד מהותי במאמר הנקרא *Kernel density estimation* או KDE. אם אתם זוכרים KDE עוזר לנו לשערר (כלומר לקרב) פונקציית צפיפות בהינתם DATA של נקודות באמצעות פונקציית קרגל. פונקציית קרגל יכולה להיות גאומטרית למשל ומטרתה לשערר את פונקציית הצפיפות לנקודות לא ידועות על ידי קירובה בין הנקודות בDATA (בדומה לسفליין). אז המחברים משתמשים בטריק זהה כדי לשערר את $\text{ReLU}_{\text{Jump}}$ בנקודה ϵ שבה היא לא גירה.

מאמר נחמד בנושא די חשוב שאמץ לסקור כנראה גם בעtid...
<https://arxiv.org/pdf/2407.14435.pdf>

🚀⚡️המאמר היום של מילק 15.08.24: Your Classifier Can Be Secretly a Likelihood-Based OOD Detect

משנים טיפה את היכיון היום וסוקרים מאמר לא על LLM. המאמר דין בזיהוי של DATA שלא מתפלג לפי התפלגות DATA במהלך אימון המודל. למשל אימנתם מודל לזהות חתולים, כלבים וטוסים ופתאום מפעלים את המודל שלהם על תמונה של טנק. אם לא נקבעם/amcuim נגד זהה DATA מחוץ להתפלגות האימון (או OOD) אתם עלולים לזהות את הטנק הזה בתור אחר הקטגוריות שאימנתם את המודל עליהם בתור כלב, חתול או טוס.

כמובן שהמצב זהה מאוד בעייתי ועקב כך הוא נחקר רבות במהלך השנים האחרונות. המאמר שנסקור קצרות היום מציע שיטה מאוד אלגנטית וטבעיות להתרמודד עם הסוגיה זו. המאמר מציע לאמן מודל לזהות קטגוריות היעד (משמעות בסט האימון) אלא גם לכפות התפלגות מסוימת על הייצוג שלהם המופק על ידי המודל (כלומר של הפלט של השכבה האחורונה של הרשת).

הפרמטרים של התפלגות זו נקבעים מראש (הממוצע ופרמטר שששולט בכמה התפלגות מרוכזת סביר המוצע - סוג של מטריצת קוורייאנס). ואם עברו דוגמא נתונה וקטור הייצוג יצא רחוק מספיק מכל וקטורי המוצע של כל הקטגוריות (כאשר מקדם הפיזור נלקח בחשבון) אז הדוגמא זו מזוהה בתור OOD.

בתור התפלגות היעד המחברים לוקחים התפלגות Mises-Fisher von על ספרה במיליד של וקטור הייצוג \mathbf{c} (כלומר הספרה היא במיליד- \mathbf{c}). המחברים טוענים זהה עובד טוב יותר מאשר התפלגות גאומטרית.

<https://arxiv.org/abs/2408.04851>

🚀⚡️: המאמר היום של מיק 16.08.24 ⚡️🚀 On the Geometry of Deep Learning

אני ממש אוהב מאמרים שחוקרים מה שקרה בתחום המודלים העמוקים שלנו - הרי לדעתך זה התנאי הכרחי לכך שנוכל להתחיל באמת לสมוך על- AI (פחות חלקית). וכן הכותבים מודגשים כי במידה עמוקה, על אף הישגה המרשימית במעט תחומים, נשארת עדין בגדיר "קופסה שחורה" עם הבנה חלקלית בלבד של אופן פועלתה.

המחברים מנסים להסביר מודלים עמוקים באמצעות ספליניים אפיניים (Affine Splines) שהן למעשה פונקציות רציפות ולינאריות למקוטען במרחב רב מימד. המחקר מתבסון ברשותנו נירונים מזוית גיאומטרית באמצעות ניתוח של חלוקות הנוצרות על ידי ספליניים אפיניים, המקרבות אותן (הרשאות).

בפרט המחברים דנים בחלוקת של מרחב הקלט לפי הקטגוריות של הנוצרות על ידי ייצוג לטנטי (השכבה الأخيرة לפני שכבת הסיגוג) של הרשת. הבנת החלוקת זו מסייעת להסביר כיצד רשתות עמוקות למדות ומיצרות חיזויים עבור קלטים שונים.

המחברים גם דנים במבנה גיאומטרי הנוצרים על ידי משקל המודל במרחב הלאו (כלומר מנתחים את פונקציית הלאו למשקל הרשת השוניים). בנוסף המאמר גם מדבר על החלוקת הנוצרת במרחב משקלות המודל בשכבות שונות לאთחל, רשת שונים וגם לאימון עם ובלי BatchNorm. כמובן שזה נעשה על דוגמאות מלאכותיות(toy examples) בעלי מימד נמוך. ויש עוד מספר ניתוחים גיאומטריים די מעניינים במאמר.

מעניין כי המחברים כתובים כי אחת המטרות המרכזיות של המאמר היא לדרבן מתרמיים לעסוק בניתוח גיאומטרי של רשתות עמוקות.

<https://arxiv.org/abs/2408.04809>

🚀⚡️: המאמר היום של מיק 17.08.24 ⚡️🚀 Faster Machine Unlearning via Natural Gradient Descent

היום סוקרים מאמר כחול לבן בנושא הנושא learning unlearning. בדרך כלל אנו מונינים שהמודל שלנו ילמד מהדата אבל כאן אנו רוצים שהמודל ישכח דאטה מסוים. הנושא די חשוב לחברות שרצו להיות compliant עם הדרישות של תקנים סטנדרטיים GDPR כאשר יוזר או קבוצה יזרים מבקשים למחוק את הדטה שלו באופן מוחלט. כמובן שבנוסף למחיקת הדטה עצמה צריך "למחוק" אותו מה"מוח" כלומר המשקלים של המודלים שאומנו (בפרט) על הדטה הזאת.

אחד השיטות הנאייבות לעשותות unlearning היא למחוק את הדטה ולאמן מודל חדש. אבל זה יכול להיות די יקר ולא יעיל במיוחד למודלים גדולים. האם קיימת שיטה אחרת לעשות את זה?

אכן יש לא מעט מחקר בנושא של unlearning ואחת הגישות הפופולריות היא לנקות מודל מאומן ולמצער את הפרש של הביצועים על הדטה שנותר והדטה שאמור להימחק. ככלומר אנו רוצים למזער את הלאו על הדטה הנותר ולמתקדם אותו על הדטה שנמחק. ככה "נמחק" מהמוח(אולי ההזכרון) של המודל את הדטה המיועד למחיקה.

כמובן שיש שיטה נוספת "למחוק" את הדטה מהמודל היא פשוט למקסם את הלאו על הדטה המיועד למחיקה.

כמובן שיש שיטות רבות לעשות את זה באמצעות וריאציות שונות של מורד הגרדיאנט (stochastic gradient descent). המאמר מציע לעשות את זה עם מה שנקרא natural gradient או NG. זה קונספט פחות ידוע ואני אסביר אותו בקצרה. אתם בטח זוכרים מה זה קצב למידה-ב-SGD, נכון? זה פרמטר קריטי לתהיליך הלמידה וקיימות לא מעט שכליים של SGD כמו ADAM ו-RMSProp שבפועל (בצורה לא מפורשת) קובעים את קצב הלמידה האופטימלי כתלות בפונקציית לוס.

יש כמובן דרך נוספת לבחור את קצב הלמידה בצורה אופטימלית וזה מה שעשו שיטות ניוטון קלאסית (נראה לי זהה השם) לאופטימיזציה. במקום להשתמש בקצב למידה סקלרי משתמשים בהופכית של ההסיאן של פונקציית לוס (מטריצה של נגזרות שניות). זה אופטימלי מבחינת התכונות (כי משתמשים בקירוב טילור מסדר שני של פונקציית לוס). אבל כאמור לא ניתן לעשות זאת לרשותך (יש קירובים אמנים כי קשה מאוד להפוך מטריצה בגודל מיליארד על מיליארד).

המאמר מציע להחליף את ההיסaan ב- FIM או Fisher Information Matrix. למעשה FIM היא תוחלת של המכפלת הוקטורית של הגרדיאנט הלוג של הנראות (likelihood) של הדאטה המקורב על ידי המודל עמו. למעשה FIM מודד עד כמה שונה בפרמטרים של המודל משפיע על הנראות של הדאטה באמצעות המודל (עם המשקלים הנוכחים). זה בעצם מצביע לנו עד כמה הנראות של הדאטה רגישה לשינוי בערכי המודל.

יש ל-NG הרבה יתרונות (למשל הוא חסין לרפרמטריזציה של המודל) אבל כמו ההיסאן עדין מאוד קשה לחשב אותו עבור מודלים ענקים. כאמור שקיימות שיטות המחשבות אותו באופן מקובל באמצעות שילוב עם פונקציית רגולריזציה "נוחה".

בנוסף לעדכן הרגיל של הגרדיאנט כמו ב-SGD עם FIM (כלומר בהופכית שלו) המאמר משתמש במה שנקרה אחרי העדכן ולא אפשר להם להתרחק יותר ממשקל המודל לפני עדכן כאשר ה"מרחק" כאן מנורמל עם ההופכית של FIM תוך כדי לקיחה בחשבון של פונקציית רגולריזציה (שלא ת תפוץ).

המאמר די קשור מתמטית ומוקוו שהצלחתו לשפוך קצת אור עליו... .

<https://arxiv.org/abs/2407.08169>

🚀⚡️המאמר היומי של מיליק 19.08.24: DIGRESS: DISCRETE DENOISING DIFFUSION FOR GRAPH GENERATION

היום סוקרים קיצרות מאמר לא רגיל על מודלי דיפוזיה. אתם בטח זוכרים (וסקרתי לא מעט לאחרונה) מודלי דיפוזיה עבר תמנות, וידאו, אודיו וכדומה. במאמר שנסקור אותו היום מודל דיפוזיה נבנה על גרף. אצין כי המאמר מלפני שנה וחצי ולמיטיב ידיעתי יצאו כמה מאמרים המשר.

از מה זה מודל דיפוזיה רגיל ואיך מאמנים אותו? מודל דיפוזיה גנרטיבי מאומן על ידי הוספה הדרגתית של רעש לדאטה כאשר המטרה היא לאמן מודל המשHIR את הרעש הזה (כלומר משחזר את הדאטה מאיטרציה הקודמת). מודל זהה מאפשר לנו לגנרט דאטה מרעש טהור על ידי הסרתנו הדרגתית.

אבל איך ניתן "להטיל" את הרעיון הזה על גרפים? נניח שיש לנו גרף בו כל הצומת וכל קשת שייכים לקטורייה מסוימת (קטגוריות שונות לקשתות ולצמתים). עכשו בתהיליך קדמי (הוספה רעה) אנו בעצם משנים באקרה את הליבלים (קטגוריות) של הצמתים ושל הקשתות לקטgorיה אחרת. ככלומר צומת נתונה יכולה להישאר בקטgorיה

שללה בהסתברות 0.95 ובהסתברות 0.05 היא תקבל כל לייל אחר בצורה אחד. תהיליך דומה נעשה על הקשתות. בסוף התהיליך הגרף הופך להיות עם קשתות וצמתים בעלי קטגוריות רנדומליות לגמר.

כמו במודל דיפוזיה המטרה של המודל המאומן (על DATAset של גרפים מוגדים) היא לשחרר את הליבלים מהאייטריצה הקודמת (של הצמתים ושל הקשתות). זה אפשר שחרור גרפם עם התלוויות כמו בסט האימון.

כמובן שיש כאן הרבה משחק על איך מריםים את הליבלים בתהיליך קדמי. האם יש תלות בתהיליך ההרעשה בין צמתים וקשתות שונים, אולי בהתחלה משנים ליילים רק לתת-גרפים מסוימים וכדומה.

בקיצור מאמר מאד מעניין ואני מניח שאסוקו בעתיד גם מאמר ההורש שלו.

<https://arxiv.org/abs/2209.14734>



JPEG-LM: LLMs as Image Generators with Canonical Codec Representations

המאמר הזה תפיס את עיני כי מילה "pegz" הופיע בשמו. למרות שלא יצא לי לעבוד בתחום של דחיסת>Data אנו מאוד אוהבים את הנושא המרתק הזה. בנוסף המאמר הזה מדבר על מודל VQ-VAE שהוא די פופולרי לפני שמודל דיפוזיה השתלטו לנו לחלווטין על GenAI בראיה הממוחשבת.

אוקי, אז כל זה קשור? קודם כל pegz זו גישה ידועה לדחיסת תמונות. המאמר גם מדבר על AVC/H.264/H.265 שהוא לדחיסת וידעו המתבססת על עקרונות דומים לאלו של pegz. בגודל pegz עובד בצורה הבאה:

- מחלקים תמונות לפאצ'ים באוטו הגדול ועשויים לכל אחד מודול DCT - Discrete Cosine Transform (כמו התמרת פוריה ללא החלק המודומה).
- מבצעים קווינטוט של מקדמים DCT לכל פאץ' כאשר המקדמים לתרדים גבוהים "נחתכים" בצורה רצינית יותר
- משתמשים בקידוד length חע וגם בקידוד האפמן כדי לדחוס את כל המקדמים המוקונטטים של הפאצ'ים.

אוקי, עכשיו נרعن לכמה מזה VQ-VAE. קודם כל VAE זה מודל גנרטיבי שלומד לגנרט דאטה מהיצוג הלטנטי שלו (במימד נמוך). VAE מורכב מהאנקודר מהדקודר שהראשון בהם מאומן להפיק ייצוג של דאטה במימד נמוך והדקודר משחרר את הדאטה ממנו. VAE מאומן בצורה המשרה התפלגות נטוונה (בד"כ גאוסית) על המרחב הלטנטי וזה מאפשר לגנרט דאטה חדש באמצעות הדקודר מוקטור הדגום מההתפלגות זו.

VQ-VAE היא כולול של VAE כאשר הוא מאומן לגנרט תמונה בצורה סדרתית (פאצ'ים/טוקנים ויזואליים) כאשר כל פאץ' מיוצג על ידי וקטור (latent) מהמילון שנלמד גם כן. לעומת התמונה נבנית פאץ'-פאץ' כאשר כל פאץ' (כלומר וקטור מהמילון שמייצג אותו) נדגם בהינתן כל פאצ'ים שכבר גנרטו. זה בטח מזכיר לכם מודל שפה שמגנרט טוקנים בדיק באמצעות צורה.

VQ-VAE מאומן בשני שלבים: בראשון מאומנים את האנקודר, המילון והדקודר (המשחרר פאצ'ים מהווקטורים במילון) ובשלב השני מאומנים מודל לחזות טוקן ויזואלי הבא בהינתן הטוקנים שכבר נוצרו.

המחברים שילבו את הרעיון של האלו (חלוקת) ואימנו מודל שיודיע לחזות יציג pegz או cav בצורה סדרתית. אבל מה הטוקנים כאן? בדומה למודלי שפה המחברים השתמשו ב-BPE או byte-pair encoding (עם שפচרים קלים). מכאן המחברים בננו מודל היודע לרנרט יציג pegz של התמונה שניתן להפוך אותה די בקלות.

רעיון ד' חמוד אבל יש לי הרגשה שכברرأיתי רעיונות דומים בעבר ...

<https://www.arxiv.org/abs/2408.08459>

🚀⚡️ **המאמר היומי של מיק 21.08.24** ⚡️🚀

Tree Attention: Topology-Aware Decoding for Long-Context Attention on GPU Clusters

היום נסקור מאמר בנושא שיבור סקרתי כמו מאמרים לפני חדש. הנושא זהה נקרא אופטימיזציה והאצה decoding של מודלי שפה כלומר התהילן שגנרטוט טוקן חדש בתלות בכל הטוקנים בתור חלון ההקשר שכבר גונרטו. ואם חלון ההקשר הוא ארוך (מאות אלפי טוקנים) זה יכול לחתה די הרבה זמן בעיקר בגלל מנגן ה-hsion. execution של הטרנספורמים שהווים backbone של כל מודלי השפה החזקים.

בשנים האחרונות הוצעו מספר רב של שיטות ליעול והאצה של חישוב attention שהכי מפורסמים מהם הם Flash Attention KV-Cache ו-KV-Flash. שיטות אלו בדרכן כל מנצלות את העובדה שהיום אינפרנס של מודלי שפה מתבצע על GPU ואני ליעול את החישוב על ידי שימוש ביכולת של GPUs לחשב דברים במקביל.

יתרה מזה מכיוון שמודלי שפה רצים היום על קלאסרים של GPUs יצאו מספר עבודות על איך ניתן לחשב את attention על קלאסרים אלו. מכיוון שמנגן ה-hsion מכיל מכפלות פנימיות (סכומים רבים) אז ניתן לחשב בצורה מבוזרת די ביעילות.

והמאמר הזה מציע מנגן מעניין של חישוב h-sion. הדבר המעניין בו שהמאמר הזה מייצג את חישוב attention (עבור וקטורי שאילתת נתון q) כנגזרת של הלוג של "פונקציה יוצרת" של h-sion המוחשבת בנקודות 0. פונקציה יוצרת זו נבנית על ידי מניפולציה פשוטה של נוסחת h-sion וממש מזכירה פונקציה יוצרת של משתנה אקריאי.

ניתן להכליל את החישוב זהה לה-sion עבור וקטורי שאילתת q מרובים כאשר במקומות נגזרת רגילה יהיה לנו נגזרת לפי ח משתנים (ח הינו מספר וקטורי השאילתת).

למה זה טוב בכלל? מתרבר שהחישוב של attention בצורה כזו מערב פעולות כמו `expsumlog` ו- `max` שנייתן לבזר אותם בצורה יعلاה בין ה-GPUs. החישוב נעשה בצורה של עץ, כלומר מחלקים את הסכומים לכמה חלקים, מחשבים כל חלק ואז מתחילה לסכם את התוצאות בצורה היררכית. זה כמו Map-Reduce רב שלבי.

<https://arxiv.org/abs/2408.04093>

🚀⚡️ **המאמר היומי של מיק 22.08.24** ⚡️🚀

Approaching Deep Learning through the Spectral Dynamics of Weights

היום נסקור מאמר החוקר מה הסיבות לתופעה של גראינינג. למי שלא מכיר גראינינג זו תופעה די מעניינת המתרכשת כאשר ממשיכים לאמן רשת ניירונים (למרות שזה קורה גם במקרים אחרים) גם אחרי לוס הולידיצה מתחילה לעלות (כלומר אנו נכנסים למשטר אעורפי). מתרבר אם לא עוצרים וממשיכים לאמן לוס הולידיצה מתחילה לרדת ככלומר המודל נכנס למושט הכלכלי כלומר לומד את ה"חוקיות האמיתית" מאחורי הדטה.

התופעה זהה היא מקרה פרטני של double descent (יש גם double descent שמתארח גם אם אנו מוסיףם פרמטרים למודל בצורה עקבית ומגיעים למצב שיש לנו over-parametrization. כלומר יש המודל שלנו לכאורה

מתחליל "ויתר מדי פרמטרים" כדי "להבין את הדאטה". גם שם זה קורה בצורה בלתי רציפה כלומר יש אינטראול של פרמטרים שביצועי המודל יורדים עבורים ורק אז מתחילה לרדת.

המאמר חוקר מה קורה עם משקל המודל כאשר הוא נכנס למטריך הגרוקינג. מתרבר שתוופה הגרוקינג קשורה לירידה בראנק של מטריצות המשקלים של המודל. בשיביל' זה די אינטואיטיבי כי לדעתך במלין גראוקינג המודל מצליח להתכנס לפתרון פשוט יותר עבור הדאטאסט. פתרון פשוט הכוונה הוא מודל שאפקטיבית הוא קטנה, ככלומר רב וקטורי המשקלים בו או אף או תלויים לינארית זה בזה.

<https://arxiv.org/abs/2408.11804>

🚀⚡️ המאמר היומי של מ"יק 30.08.24: Platypus: A Generalized Specialist Model for Reading Text in Various Forms

חוזרים לסקירות אחרי שבוע של חופשה עם מאמר בנושא שלא סקרתי די הרבה זמן והוא Optical Character Recognition או OCR בקצרה. מטרת OCR היא לזהות טקסט בתמונה או במסמך כאשר הטקסט יכול להופיע בצורה ומגוונות. מודלי OCR הקודמים בדרך כלל התמקדו בזיהוי של סוג של טקסט (נגיד נסחה, טקסט מודפס או כתוב ידי). המחברים מציעים גישה שמאחדת את מומחי ה-OCR ה"צרים" לזיהוי סוג ספציפי של טקסט - ככלומר מסוגלת לזהות כל סוג של טקסט בתמונה כולל המקרים שיש כמה סוגים של טקסט בתמונה.

בנוסף-OCR יש 3 מושגי הפעלה. הראשון זה Recognize All Text RAT או Point Prompt Recognition שמטרתו לזהות את כל הטקסטים בתמונה. השני הוא PPR או Box Prompt Recognition שמיועד לזהות את הטקסט סביב נקודה נתונה (סוג של עוגן) בתמונה. השלישי הוא BPR Box Prompt Recognition שמיועד לזהות של טקסט בתוך מלבן נתון בתמונה (כמו שיש לנו Bounding Boxes בזיהוי אובייקטים בתמונה אבל בכיוון הפוך).

از המחברים מאמנים מודל המורכב מהאנקודר (שהופך תמונה לאמבעינג) הדקודר האוטורגרטיבי. הדקודר מקבל כקלט את סוג הטקסט בתמונה (מודפס או כתוב ידי). בנוסף הדקודר מקבל את סוג המשימה (RAT, PPR או BPR) עם כל הפרטים הנחוצים לביצוע משימה (כלומר קואורדינטות של הפאץ'). בנוסף המודל מקבל גראנוולריות של זיהוי הטקסט (כלומר line-level-word או word-level) שהראשון הוא זיהוי מילה בודדת והשני הוא זיהוי טקסט שלם. הפרטים האלה מזונים כאמור לדקודר שמטרתו לגנרט את הטקסט המופיע בתמונה.

זה כל הפרטים העיקריים - מאמר די קליל....

<https://arxiv.org/abs/2408.14805>

🚀⚡️ המאמר היומי של מ"יק 31.08.24: Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review

היום סוקרים מאמר לא רגיל - קודם כל זה מאמר סקירה בעצמו והוא לא מאד טרי (מלפני כמעט שנתיים). המאמר בנושא explainability של מודלי למידת מכונה. רוב מודלי ML היום הם רשות נירוניים מאוד עמוקות ולרוב הם נשאים בתוך קופסה שחורה עבורה - מחקרים explainability מנסים לשפוך אור על "מה שקרה בתוך הקופסה השחורה זו".

המאמר הזה נותן סקירה של אחד הפרסדיוגמות העיקריות המשמשות למחקר explainability של מודל ML - ניתוח counterfactual. ככלומר חוקרים מה צריך לשנות בדגם(איזה פיצ'רים) כדי שהוא תסוג לקטגוריה (כלאço) אחרת ועל ידי כך נבין יותר טוב למה המודל סיוג את הדוגמא המקורית לקטגוריה המקורית. דרך אגב יש שיטות explainability שחוירות את המודל בצורה אחרת. למשל קיימות שיטות שמנסחות לקרב את המודל

המורכב על ידי מודל פשוט יותר (ע"ז או רגרסיה לינארית) בטווח מסוים של דוגמאות. שיטות נוספות המנסות להסביר את חיזויו של המודל לדוגמא ספציפית.

از מה בעצם חשוב לנו מאוד בשיטות counterfactual? קודם כל חשוב לנו לשנות כמה שפחות פיצ'רים של הדוגמא הנחוצים ל"העברתה" לקטגוריה אחרת וגם השני בפיצ'רים אלו צריך להיות די קטן כדי להבין את "מבנה גבולי" בין הקטגוריות השונות מבחן המודל. השני ולא פחות חשוב השם השני הזה צריך להיות "חוקי" ככל מר步 הדוגמא הנוצרת צריכה להיות הגיונית ולידית (כלומר שטח הבית לא יכול להיות שלילי). בנוסף השמי בדוגמא צריך לעבור במסלול הגיוני ככל מר步 קרבנה של הדוגמאות האחרות מהדאטסהט. כאמור יש עוד דרישות לשניינו שאנו מחוללים לדוגמא כדי להפוכה ל-counterfactual.

וכל הפרטים המעניינים במאמר כמפורט...

<https://arxiv.org/abs/2010.10596>

🚀⚡️המאמר היומי של מייק 01.09.24: DIFFUSION MODELS ARE REAL-TIME GAME ENGINES

טוב, על המאמר זה פשוט לא היה לדלג מכמה סיבות. הסיבה הראשונה שאני מספיק עתיק ועוד שיחקתי במשחק הנקרא דום (doom) במו ידי כאשר הייתה נער. דבר שני לא כל יום מחליפים לך מונע משחק במודול למידת מכונת או בשם המוכר AI. כאמור שהזה כיוון מחקר מאד מעניין עם פוטנציאל להפתחה לכליים מבוסס AI לבניית משחקים חדשים.

הרעיון של המאמר הינו די אינטואיטיבי. בשלב הראשון הסוקן (agent) מאמין לשחק דום בעצמו על דאטסהט של המשחקים ששוחקו על ידי בני אדם. ככל מר步 בהינתן כמה מצבים המשחק (פריים) והפעולות האחרונות (ירי, תנועה, פגיעה וכדומה) מטרת הסוקן היא חייזר הפעולתו הבאה. זה נעשה באמצעות טכניקות RL די סטנדרטיות כאשר פונקציית reward נבחרה בצורה הגיונית בהתאם ללוגיקת המשחק (כלומר פגיעה או מות של הסוקן מקבלות תגמול שלישי ואילו פגעה באובי, איסוף נשך וכדומה מקבלים תגמול חיובי).

אחרי שהסוקן למד לשחק דום, מגנרטים כמותมหา גודלה של משחקי דום עם הסוקן. ככל מר_step הסוקן משחק במשחק אמיתי כמו אחד האדים. לאחר מכן מאנים מודל דיפוזיה לחזות את הפריים הבא בהינתן הפריים הפעולות הקודמות והנכחות.

האימון מתבצע בצורה די סטנדרטית: מודל דיפוזיה מקבל כקלט את הפעולות הקודמות אחרי האנקדור (שמאומן גם כן) ובנוסף את הפריים הקודמים מזונים למודל דיפוזיה (בצורה מורעת לשיפור יכולת הכללה של המודל). מודל דיפוזיה שהמחברים השתמשו בו הינו לטנטי (כלומר חייזר הרעש מתבצע למרחב הלטנטי של הפריים הנצהה). נציין כי כאן להבדיל ממודלי דיפוזיה ישנים יותר מודל הדיפוזיה במאמר מאמון לחזות את מה שנקרה "מהירות" של הפריים המורעש שהוא פונקציה של הפריים הנקי והרעש המתווסף אליו באיטרציה. רפרמטריזציה זו משפרת את יכולת המודל ומאייצה התכנסותה (מוחך אמפירית כרגע)...

מאמר מאד מגניב...

<https://arxiv.org/pdf/2408.14837.pdf>

🚀⚡️המאמר היומי של מייק 02.09.24:

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

היום נסקור מאמר על מודל מולטימודלי בצורה די מעניינת. המודל שאימנו במאמר יודע לאגנרט גם תמונות וגם דאטה טקסטואלי וממהו שילוב של מודל דיפוזיה ומודל שפה.

היחודיות של המודל זהה מتبטהת בכך שהוא מגנרטת גם את הדאטה הטקסטואלי וגם הדאטה היזואלי בצורה שאנו מגנרטים טקסטים, ככלומר טוקן אחריו טוקן (עבור תמונה זה למשה טוקן ויזואלי או "יצוג של פאץ"). ככלומר אם אנו צריכים לאגנרט תמונה ייחד עם תיאורה המלא המודל יגנרט את התיאור טוקן ואחריו טוקן (next token prediction או NTP) ואחריו שיסים יגנרט את התמונה טוקן אחריו טוקן (בצורת NTP גם כן). זה די נחמד האמת.

המודל שהמאמר אימן מכיל 7 מיליארד פרמטרים זהה די צנוע למודלי שפה וגודל די סטנדרטי למודלי דיפוזיה גנרטיביים (המודול הגדול של stable diffusion מכיל בערך 8B פרמטרים). אבל כאן יש לנו מודל המשלב את שתי היכולות האלה (גנרטת תמונות וගנרט טקסטים) באיכות די גבוהה.

אבל אין מאמנים את המודל הזה? בגודל בהינתן קלט שהוא ערבות של תמונה וטוקסט (למשל תמונה מעורבתת עם טוקסט). עם הטוקסט הכל פשוט, מזינים אותו טוקן אחריו טוקן. לפני כל תמונה מכנים טוקן NOI EO המסתמן את תחילת התמונה וכאשר כל הטוקנים היזואליים של התמונה הוזנו מכנים טוקן NOI EO לסייען סיום החנת התמונה. כאמור טוקנים של תמונה זה טוקנים ויזואליים המהווים ייצוגים של פאצ'ים לאחר האנקודר (של VAE).

air מאמנים את החיים הזה? לטוקסט זה די ברור - מאמנים את המודל לחזות טוקן טוקן כמו ב-LLM עברו מיליון טוקנים נתון. עבור התמונה מחלקים את התמונה לפאצ'ים, מעבירים כל פאץ דרך האנקודר של VAE ומזינים את התוצאה כטוקן. הייצוגים של הטוקנים היזואליים מועברים דרך שכבה לינארית או net להורדת ממד. במהלך האימון לומדים להסיר רעש מהgresאות המורעשות של ייצוג הטוקנים היזואליים.

בגנרט המודל יוצר את התמונה פאץ' פאץ' מהרعش (אחרי הסרת הרעש וקטור הייצוג מזון לדקודה של VAE כדי לחזות את הפאץ' עצמו). לאחרונה השיטה זהו לייצור תמונה לא פופולרית במילוי - רוב השיטות יוצרות את התמונה המלאה (מהייצוג הלטנטי שלה). וכמובן כל הטוקנים האלה מזינים לטרנספורמר אחד גדול!

מאמר מעניין ומומלץ לקרוא!

<https://arxiv.org/pdf/2408.11039>



Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling

את הדרכים הדיא מפתיעות לשיפור יכולות reasoning של מודלי שפה היא שיפור עצמי או self-improvement. בגדול עבור דאטהסט של שאלות ותשובות אנו מבקשים ממודל שפה לענות על התשובה ולספק הסבר. לאחר מכן מפלטרים את השרשות reasoning של לא התכוונו לתשובה הרצויה. לאחר הפלטור מבצעים פיניטיון של המודל על הדאטהסט המפולט. כאמור באופן די מפתיע (פחות או יותר) הדבר אכן מוביל לשיפור יכולות reasoning של מודל שפה.

אם יש בידינו מודל יותר חזק אז ניתן לבנות את הדאטהסט הזה באמצעותו ולעשות את הפיניטיון על הדאטה הנוצר באמצעותו בצורה דומה.

אולם המאמר שואל שאלה ד' מעניינת: מה עדיף (מבחינת הביצועים), ליצור יחסית מעט דатаה עם מודל גדול וחזק או ליצור יחסית הרבה דטהה עם מודל קטן וחלש יותר. הרוי יצירת דטהה עם מודל חזק היא יקרה יותר (מבחינת כמות ה-FLOPS הדרושה הנדרשת לכך) אבל מצד שני הדטהה שהוא יוצר הוא יותר אינטואיטיבי. המחברים מציעים לבצע את ההשוואה של "תפוזים לתפוזים" - כמובן לקחת את הדטהה הנוצר עם מודל חזק ומודל חזק תחת אותו תקציב של FLOPS ולהשוות מה מהם מוביל לביצועים טובים יותר של המודל שעובר פיניטיון על הדטהה הזאת.

יש תוצאות ד' מעניינות במאמר ..

<https://arxiv.org/pdf/2408.16737.pdf>

🚀⚡️: 04.09.24 🚀⚡️

Flexora: Flexible Low Rank Adaptation for Large Language Models

המאמר הזה נסקרו קודם כל בגל שהו למעשה מימוש של רעיון שחששתי עליו והוא גם רשום לי בבלוג (שהוא באורך ד' אינסופי). הרעיון הוא למעשה שיטה לבחירה (לפעמים קוראים לזה אופטימיזציה) של ההיפרפרמטרים של LoRA (סוג של).

כמו שראתם בטח זכרם LoRA היא משפחה (ד' גודלה שימושית לגודל) של שיטות מהמשפחה (גדולה עוד יותר) של שיטות חסכנות פיניטיון של מודל שפה ענקים (או C). PEFT - Parameter Efficient Fine-Tuning. ב-LoRA אנו מאמנים תוספת של משקלים לכל שכבה במקומם לאמן את כל המשקלים במודל. כל תוספת כזו היא מטריצה בעלת רANK נמוך כלומר אפקטיבית מכילה מעט פרמטרים מאשר מטריצת המשקלים של השכבה.

פרקטיית כל תוספת היא מכפלה של שתי מטריצות בעלות RANK נמוך (מלבניות) וככל הרank נמוך יותר יש לנו פחות פרמטרים לאפטם במהלך פיניטיון. הבחירה של הרank של מטריצות התוספות הנדרשת למקריםعينים ביצועים איננה בעיה פשוטה ויש מספר מאמרם שדנים בנושא זה (בד"כ עד רANK מסוים הביצועים משתפרים ומתקודה מסוימת מתחילה אורהפייט).

המאמר (וגם אני) חשבו על דרך אחרת של אופטימיזציה של LoRA. המחברים שואלים שאלה פשוטה - למה בנוסף לאימן של מטריצות התוספות לא נאמן את ה-importance שלה בכל שכבה. ה-importance במרקחה הזאת היא המקדם המכפיל את מטריצת התוספות לפני הוספהה מטריצת המשקלות המקורית במודל (שנותרת קבועה במהלך פיניטיון). האלגוריתם המוצע עושה כמה איטרציות של משקל ה-importance עדכון אחד של משקלות התוספות.

האמת שהרעיון שלי הכליל עוד שלב של pruning. כלומר אחרי מספר של איטרציות אימון מתחלים לפחות ומפוזיקים לאמן מטריצות התוספות עם importances נמוכים מאיזה סף. נראה שאצטרך לבדוק את זה לבד (:)

<https://arxiv.org/abs/2408.10774.pdf>

🚀⚡️: 05-09.24 🚀⚡️

EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees

חדי זכרון מביניכם אולי שמו לב כי לא פרסמתי סקירה יומית אתמול (בד"כ אני לא מפרסם סקירה ביום ראשון שבhem אני מקליט פודקאסט). אתמול היה לא חמישי ולא פרסמתי סקירה כי הכתבי לכם סקירה כפולה להיום. היום נסקור שני מאמריהם שהשני מהם הוא شامل של הראשון.

שני המאמר הם בנושא של SpDe speculative decoding או (ההערכה הומצא על ידי). SpDe זו דרך להאיץ דגימה (גנרט טקסט) מודולר שפה. כמו שאתם זוכרים הגנרט מודולר שפה מתבצע באופן אוטורגריסיב כמו טוקן לאחר טוקן. כמובן שהוא יכול להיות איטי בטח עבור מודלים עצומים בעלי מאות מיליארדי פרמטרים.

אם ניתן להאיץ את תהליך הגנרט - התשובה היא כן ובשתיים האחרונות נעשה מחקר מאד רציני בנושא והוצעו מספר שיטות שבאמצעותם ניתן להציג דגימה גבוהה יותר. SpDe היא משפטת שיטות להאצת קצב גנרט באמצעות שימוש במודל קל (וחלש יותר) בנוסף למודל היעד (שאותו אנחנו מעוניינים להאיץ כאמור).

שיטת SpDe מבוססת על אוביツציה כי מהלך הגנרט טוקן אחריו טוקן צואר בקבוק הוא העברת הדטה מהזיכרון SRAM ומהיר (ארקון) של יחידת החישוב של GPU לבין זיכרון DRAM האגדל אף איטי יותר. ולא החישוב עצמו. נובע מכך שניתן לנצל את צואר הבקבוק הזה ולבצע יותר חישובים (עבור יותר טוקנים) בזמן שהדאטה מטייל בין SRAM לDRAM.

הבעיה למש את הגישה הזאת בזורה נאיבית נובעת מאופן אוטורגריסיב של הגנרט מודול שפה שלא מאפשר לבצע את החישובים עבור חזוי של יותר טוקנים באותו הזמן. שיטות SpDe עוקפות את המכשול הזה על ידי הוספה מודל קטן ומהיר יותר שיאפשר למודל הגדל לחזות כמה טוקנים באותו הזמן.

איך זה עובד? אנו חוזים כמה טוקנים עם המודל הקטן s_L ולאחר מכן "מתקנים" את החיזוי עבור הטוקנים שנדגמו עם המודל הגדל כאשר ה"תיכון" עבור כל הטוקנים שנחצוו על ידי s_L מתבצע באותו הזמן. כלומר s_L חוצה הסתברויות עבור k טוקנים רצופים (בහינת הטקסט שכבר גנרט), הם מזונים (יחד עם הקשר) למודל הגדל s_L והוא חוצה את הסתברויות s_L עבור k טוקנים אלו באותו הזמן.

לאחר מהם מביצים משהו דומה למה שעשיהם ב-*sampling rejection* ו-*"מתקנים"* הסתברויות אלו כך שייתאים להתפלגות של המודל הגדל s_L . לאחר מכן יש שלב של *rejection* שבו אנו מחליטים האם אנחנו מקבלים או לא מקבלים את הטוקנים שנדגמו על ידי המודל הקטן s_L . זה מתבצע טוקן טוקן (חישוב מהיר מאוד) וכל הטוקנים שבאים לפני הטוקן הראשון t שקיביל *reject* מתקבלים (נכנסים לטקסט המגנרט) ואיתרציה דגימה חדשה מתחילה מ- t .

ניתן להוכיח שדגימה כזו היא בעלת אותה התפלגות של הטוקנים כמו מודל היעד s_L . יש כאן כמה שאלות חשובות על איך לבחור מודל קטן s_L , בין כמה טוקנים לדוגמאותו כל פעם במטרה למסס את קצב הדגימה. עכשו השאלה האם ניתן לבחור מודל s_L כך שהוא גם מהיר מאוד וגם איקוטי מספיק כך שמספר הטוקנים שנדגמו אליו יתקרבו לרוב על ידי s_L . מודל זה עוזר את מהירות הדגימה האפקטיבית מ- s_L .

זה בדיק מה שהמאמר Eagle מציע. הוא מציע ללקחת מודל קטן ולאמן אותו להיות s_L , כלומר לתפור אותו לשימוש שהוא מיועד. בשביל זה עבור מודל s_L נתון לקחים מודל טרנספורמר דו-וירטואלי אחד ומאפשרו לחזות לא רק את הטוקן הבא (ההסתברות שלו) אלא גם הייצוג בשכבה אחרונה של s_L לפני שכבת החיזוי (כלומר יש כאן בעיית רגסיה). החיזוי מתבצע בהינתן הטוקנים הקודמים (הייצוגים שלהם) וגם הייצוג המשכבה האחרונה של הטוקנים הקודמים. מכיוון שהמודל s_L הוא קטן ומהיר אנו חוזים אותו כמה סדרות טוקנים (עבור הקשר נתון) על ידי בחירה של כמה טוקנים (ארק מספר קבוע כל פעם, נגד 3) בעלי הסתברות גבוהה ביותר כל פעם. כלומר

להקשר נתון חוזים כמה המשכים עבורי - בונים סוג של עץ חיזוי עבור הטוקנים. ככה יותר טוקנים רצופים עשויים לא לקבל reject מ-L אחר כך.

לאחר שאיםון S הסטיים לוקחים אותו ומבצעים אינפראנס בצורה די דומה ל-SpDe עם שכילול של מנגון reject-ה.

מה בעצם Eagle2 משכילה את מנגנון בניית עץ החיזום על ידי S על ידי בחירת סדרות בעלי הסתברות כוללת מקסימלית. סדרות השונות עם Eagle2 יכולות להיות בעלות אורך שונה כМОן (הכל מסתמך על ההסתברות הכלולת של הסדרה). ככה נוצרות סדרות בעלות פוטנציאל גובה יותר להתקבל על ידי L.

היה אורך - מקווה שלא איבדתי אתכם....

<https://arxiv.org/pdf/2401.15077>

<https://arxiv.org/pdf/2406.16858>

🚀⚡️: 07.09.24 🚀⚡️ **המאמר היום של מייק**

ReMamba: Equip Mamba with Effective Long-Sequence Modeling

סוקר את המאמר זהה משתי סיבות. קודם כל הוא קשור למamba. הסיבה השנייה היא זה שהתקשתה לסקור אותו. ואוקי, המאמר לצערי לא חדש לי הרבה ולדעתי לא נזכר אותו בעבר כמה חודשים.

אתם זוכרים את SSM או State Space Models בהקשר של מודלים עמוקים? ICSM הוא ארכיטקטורה יחסית חדשה עבור רשות לעיבוד דאטה סדרתי (שפה טבעית וגם תומנות). השום הגדול ב-SSM היא שהם מאוד מהירים גם באימון וגם באינפראנס עקב כך שניין לייצג אותם בתור רשות קונבולוציה וגם במודל רשות recurrent.

הगמישות הזו כМОן גבוהה מאתנו מחיר בדמota חומר expressiveness (יכולת למודל חוקיות מורכבות) של ארכיטקטורה זוו עקב העובדה המعتبرים בין המרכיבים החביבים הם לינאריים וקבועים לכל איברי הסדרה (מכאן בא הדואליות ביצוג).

ארQUITקטורת מamba מחייבת לנו קצת מה-expressiveness בכך שהופכת את המعتبرים בין המרכיבים החביבים לתלוּי במצב החבוי אך משאיר אותנו לנאראים. זה עוזר אבל עדין מ מבנה מתבקשת במשימות reasoning Morccbotות עקב מחסור expressiveness. יתכן שאחת הסיבותiae הא整洁ה זו היא חוסר יכולת של ארכיטקטורת מamba לדוחס את המידע הרלוונטי למשימה (לגיטימי אבל כМОן יש עוד סיבות לכך).

המחברים מציעים לדוחס את ייצוגיהם של תת סדרות של טוקנים. נניח שיש לנו L טוקנים בהקשר ואנו רצים "לדוחס" את טוקנים שייצוגיהם דומה לזה של הטוקן L. ככלمر מחשבים את הדמיון בין תת-סדרה רציפה נתונה של טוקנים (היפרפרמטר) ובודחים את הייצוגים של הטוקנים בתת-סדרה זו לפחות טוקנים (היפרפרמטר גם כן). ככלומר במקום הייצוגים של M טוקנים בתת סדרה נקבל ייצוגים של K טוקנים אחרי הדחיסה. הדמיון מחושב דרך דמיון קויין (עם כל מיני שכבות לנאריות).

הם מראים שהוא עובד - לי זה מريح קצת אובייקטיבי וגם קושי באופטימיזציה של ההיפרפרמטרים....

<https://arxiv.org/abs/2408.15496>

🚀⚡️: 08.09.24 🚀⚡️ **המאמר היום של מייק**

DO TRANSFORMER WORLD MODELS GIVE BETTER POLICY GRADIENTS?

לא הייתה אמורה לכטוב סקירה היום אך הקלטת הפודקאסט שלנו התבטלה והtapena לי קצת זמן אז אסקור מאמר שכבר נמצא כמו זמן אצלי ב망ירה. המאמר בנושא למידה עם חיזוקים (RL) וטרנספורמרים אז לכואורה זה נשמע מאמר די נחמד.

המאמר מדבר על שיטה לשיפור של למידה פוליסי בעיות של RL בעיות שיש לנו גישה ישירה לדינמיקה של הסביבה (כלומר אנו לא יכולים לאסוף עליה דата רלוונטי המאפשר את פיצרים המהותיים שלו קרי observable-action). בגודל המטרה שלנו בלמידה פוליסי היא לחזות את הפעולה (action) האופטימלי בהינתן המצב s של הסביבה והפעולה האחרונה a. אופטימלי כאן משמעו מקסום של התגמול (reward) הכלול המתkeletal במהלך אפיוזדה. המודל שחזזה את הפעולה זו הוא למעשה ממשת את הפוליסי שלנו.

אבל מה לעשות אם אין לנו גישה ישירה לסביבה? במקרה זה אנו יכולים לאמן מודל שהוא חוזה לנו את המצב הבא s בהינתן המצב הקודם (כלומר יציגו) והפעולה האחרונה (עם הנחת המרקבויות) או בהינתן N יציגים של המצבים האחרנים והפעולה האחרונה. זה למעשה model world (לדעתי יחד עם מודלים המשערכים את התגמול הצפוי למצב נתון - value function).

אין המודל הזה מאמון? מאינטראקציה עם הסביבה - הוסף מבעות בסביבה ואנו מעדכנים את-world model שלנו בהתבסס על משוואות Bellman. שימו לב אם או ללא הנחת מקרוביות אנחנו משערכים את הייצוג של המצב ה"עולם" הבא בהינתן המצב(-ים) הקודמים. המאמר טוען שהא יוצר גרדיאנטים לא יציבים ושונות גבוהה עקב שימוש ישר בשערך של המצבים הקודמים לשועור של המצב הבא.

הם מציע לשערך את המצב הבא מהפעולה ולא מייצagi המצבים שטענתם "הופך את הגרדיאנטים במודל העולם לפחות מעגליים" וזה תורם ליציבות השערוך. ש גם קצת הוכחות במאמר (סוג של) של הטענה זו. המאמר מראה אם יש לנו מקרוביות (התלות של המצב הבא היא רק במצב האחורי) השיטה המוצעת עובדת כמו RNN מבוחינת הגרדיאנטים. במקרה זה נשמע לי די טבעי (אשמה אם מישחו ירחיב על זה). במקרה שאין לנו מקרוביות הטענה לביצועים טובים יותר של השיטה המוצעת.

לא ראייתי אזכיר ממשותי מדי של הטרנספורמרים במאמר (תקנו אותו אם אני טועה).
<https://arxiv.org/abs/2402.05290>

🚀⚡️המאמר היום של מיק 09.09.24:

MemLong: Memory-Augmented Retrieval for Long Text Modeling

אחד המאמרים ראשוניים בנושא Chosen Retrieval Augmented Generation או RAG שאני סוקר. הנושא צובר תאוצה רצינית בזמן האחרון והגע הזמן להשלים את הפרסום (גם בידע וגם בסקרים).

RAG זה בעצם דרך להtagבר על כך שלמרות כל ההישגים בתחום אפילו מודלי שפה החדשים ביותר מתקשים לעבוד עם אורך הקשר מאוד. מה בעצם קורה כאן? נניח שיש לנו דאטאsett D ואנו רוצים שמודל השפה שלנו יענה על שאלות על D תוך כדי שילוב יכולות שהוא צבר במהלך האימון לפני זה.

את הדרכים היא לעשות למודל שפה פיניטי על D אולם זה עלול להיות עייתי כי המודל יכול לשכך חלק מהדברים שידע קודם וgem יתקשה ללמידה את כל מה שיש ב-D בצורה יעללה (פתר כМОון אבל קשה). הדרך השנייה כי להוסיף את D לכל שאלת המשתמש (כחול מפורט) אבל זה גם בעייתי - D גדולים עקב אי יכולת של מודלי שפה להתמודד עם אורך הקשר גדול מאוד.

דרך נוספת היא לעשות RAG (אפשר לשלב אותו עם פיניטיון קלייל - ראייתי מאמר שעשו את זה) כלומר לכל שאלתה של משתמש לבחור את המידע מה-D (כמה צ'אנקים) היכי רלוונטיים לשאלה והוסיף אותם לפורומפט. הבעיה בגישה זו היא מטריקה לבחירת הצ'אנקים הרלוונטיים ביותר לשאלה. בד"כ זה נעשה על סמך המרחק קוויין בין ייצוג השאלה לייצוגי הצ'אנקים (כלומר אմבדגס). ככלומר בוחרים כמה צ'אנקים הקרובים ביותר לשאלה מבחינת מרחק זו.

גישה זו עלולה להיות בעייתית גם כי לא תמיד מרחק קוויין בין הייצוגים משקף את רלוונטיות של צ'אנק לשאלה. המאמר שנסקרו היום מציע בנוסף לצ'אנקים מתחת ל-RAG את הזכרון המאחסן את הייצוגים של השאלות האחרונות (או/ו השichenות) ובנוסף לכל שאלה מחזק סוג של KV-cache Key and Value (מניחים שיש לנו דאטהסת המכיל שאלות ותשובות וגם DATASET D). אז KV-cache זהה הייצוג של וקטורי ערך KV-cache עברו שכבה מסוימת (לקראת הסוף המודול זהה אחד הייפורטרים של השיטה). ייעזר לנו לבנות תשובה בצורה טובה יותר.

از איך כל העסוק הזה עובד? במהלך האימון אנו לוקחים שאלה ותשובה מהדאטהסת של שאלות ותשובות ובאמצעותו בונים את KV-cache של המודול כי אנחנו יודעים מה הצ'אנקים הרלוונטיים ביותר לכל שאלה. הרי לכל צ'אנק אנו שומרים את KV-שלו (מחושב כאשר הצאנק מזון למודול יחד עם השאלה).

עכשו אנו רוצים לאמן את הרשות לנצל את KV caches האלו בצורה יעילה. בשביל כך באימון לכל שאלה לוקחים את צ'אנקים היכי קרובים אליה (מבחינת האמבדג), לוקחים את KV cache עברים ומאמנים את השכבות האחרונות של המודול להוציא את התשובה הנכונה. ככלומר לומדים איך לשלב את התוצאה (attention maps) מהשכבות התחתונות יחד עם KV cache שצברנו מהזיכרון (יש עוד איזה שכבה לינארית מאומנת נוספת). עדכון הזכרון מתבצע בצורה דיאטנדרטית (LRU ובנוסף השכבות נלקחת בחשבון).

האינפרכו עובד באותה הצורה פחות או יותר. בגדול המאמר מציע שיטה לשדרוג RAG באמצעות ניצול המצביע של KV-cache במהלך האימון. די נחמד מודה...
<https://arxiv.org/abs/2408.16967>

🚀⚡️: 10.09.24 המאמר היומי של מייק

Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers

האם מודלי שפה יכולם ליצור רעיונות מחקר חדשניים? 🤖 מחקר חדש מעורר גלום. ראיינו לאחרונה התלהבות רבה סביבה השימוש ב-LLMs לגילאים מדעיים. אך האם הם באמת מסוגלים להציג רעיונות חדשניים ברמת ראייה לחוקר במוסד אקדמי או בתעשה?

מחברי המאמר תכננו ניסוי כדי לבדוק את הסיפור הזה. הם שכרו מעל 100 מומחי עיבוד שפה טבעית לכטוב רעיונות מחקר ולבוחן רעיונות שנוצרו על ידי בני אדם ו-LLMs (בעיוור כלומר הבודקים לא ידעו מה מקורה של הרעיון שהם בודקים).

מתברר כי הרעיונות של ה-LLM נשפטו (באופן לא מפתיע קלוד נבחר למשימה זו) חדשניים יותר מרעיונות מומחים אנושיים (עם מובהקות סטטיסטית), אך דורגו נמוך יותר בהערכתם.

המחברים מצינים את מהחזקות הבאות של רעיונות ה-LLM:

- הצעת מכילה שילובים ייחודיים של טכניקות מדומיינים שונים

- חקירת תחומים שלא נחקרו מופיע
- ייצור ניסוי מחשבה יצירתיים ומקוריים

עם זאת, היו להם גם כמה נקודות בעיותו:

- חוסר פירוט מספק בנוגע לישום
- שימוש לא נכון במאגרי נתונים
- החמצת בייסליינום (לא מפתיע כלל)
- הנחות לא מציאותיות

לעומת זאת, רעיונות אנושיים נתנו להיות מעוגנים יותר במחקר קיים ובשיקולים מעשיים, אך לעיתים קרובות היו פחות חדשניים, ובנו באופן הדרגתי על אינטואיציות ותוצאות ידועות.

המחברים מצינים שהחוקרים מכירים בkowski לשפט חדשנות, אפילו עבר מומחים. מצד הבא, הם הציעו לתת לחוקרים למש את הרעיון הללו, כדי לראות אם דירוג החדשנות והיתכנות מתרגמים להבדלים משמעותיים במציאות.

<https://arxiv.org/abs/2409.04109>

🚀⚡️ המאמר היום של מיק 12.09.24: Learning to reason with LLMs

היום במקומ הסקירה אשתק אתכם את מחשבותי על המודל החדש של openai שקיבל שם O. אני בדרך כלל מנע מהගיב ולכתוב פוסטים על כל מודל חדש שמנצח את כל benchmarks בעולם אבל הפעם אחרג ממנהגי. ולא מהסיבה שהמודל הזה השאיר אבק לרוב benchmarks אלא בגלל שהוא אכן זיהיתי כאן שינוי מסוים בפרדיגמה בעולם-hsml.

השינוי בפרדיגמה בא בדמות של שינוי היחס בכמות הקומפיט המוקדש ללמידה ולהסקה (אינפראנס). אנחנו רגילים למודל שמצריכים כמות אדירה של קומפיאט במהלך הלמידה (אימון מקדים, SFT, "שור המודול וכדומה) כאשר האינפראנס הוא די זול (כמובן יחסית לאימון כי גם בהסקה יש עליות די גבוהות בשל עצם). O1 לעומת זאת מתגרא את ההנחה זו ושאל את השאלה: האם זה אופטימלי? אולי אנו צריכים לאמן את המודל שלנו פחות ולהשיקו יותר קומפיט בהסקה.

לפני זמן סקרתי מאמר שדי שינה (או לפחות רענן) את תפיסתי בעניין זה ([Scaling LLM Test-Time](#)) ([Compute Optimally can be More Effective than Scaling Model Parameters](#)). המאמר הזה היה של deepmind אולם הייתה לי תהוצה שהם לא היחידים שהגיעו לתובנה הדי לא טרייניאלית זהה.

בערךן הכל מסתכם לשתי הנקודות הבאות:

- אולי אתה לא צריך מודל שפה ענק להסקה. חלק ניכר מהפרמטרים כנראה ממשמשים לאחסון עובדות, כדי שהמודל לא ידבר שטויות לשאלות לידע כללי (כמו מתי נולד מוצרט). לדעתי ניתן להפריד בין הסקה לידע, ככלمر אפשר להסתפק ב"ליבנה להסקה" קתונה שיזדעת איך להשתמש בכלים כמו וולפרם, ברואזר ובודק קוד כלומר המשימות הדרשות סוג של ידע עובדתי (ידע בשפת תכנות). וכך ניתן להפחית את כמות החישוב המוקדשת לאימון המוקדם.

- כמוות משמעותית של קומפיט מועברת להסקה בזמן הרצת המודל ולא לאימון המודל. ניתן לחשב על מודלי שפה בתור סימולטורים מבוססי טקסט. על ידי הרצת תרחישים ואסטרטגיות רבות (גנרט טקסט), המודל יגיע בסופו של דבר לפתרונות *reasoning* טובים. התהילה בחירת הפתרון נראה די דומה לביעות שנחקרו היבט כמו חיפוש העץ של מונטה קרלו (MCTS) ב-AlphaGo (MCTS) ב-

כמובן שם יש שימוש בטכניקות כמו MCTS אנו צריכים את פונקציית *reward*. בניית פונקציה כזו היא לא טריוויאלית כאן כי אין לנו דרך טובה (אללא אם כן יש לנו דאתהסוט reasoning מגוון ועצום שניתן לאמן עליון מודל כזה) לשערך את איות ה-*reward*. כמובן שניתן לנצל מודלי שפה אחרים, בדיקות עצמיות על ידי מודלי שפה וכדומה אבל עדין לא ברור ב-100% איך לעשות את זה (ד"א אני בכלל לא בטוח שהם השתמשו ב-s-*mcts*). אולי הם פיתחו שיטה מוגניבה לעקוף את ה-*reward* כמו שנעשה ב-סקפ וב-סקוס שעשו זאת עבור סקק -אין לדעת.

בקיצור מכך לדוח הטכני שבתקווה ישפרק אוור על הסיפור הזה (אם בהזאה אני לא בטוח בכלל)....
<https://openai.com/index/learning-to-reason-with-langs/>

🚀⚡️: המאמר היומי של מיום 13.09.24 🚀⚡️ LLMs Will Always Hallucinate, We Need to Live With This

טוב, המאמר הזה הוא פשוט קליקביט לדעתך ואז גליתי שם משפט גדול אז בכלל. הוא מציג ניתוח עמוק של המציאות (hallucinations) ב-LLMs וטוען כי המציאות אלו הן תכונה אינהרטית בלתי נמנעת של המבנה המתמטי/ארקיטקטוני ואופן החישובי שלהם (אולי זו החשד יתגאר את זה טיפה).

כמה נקודות עיקריות מהמאמר:

1. **ה眞ות כבלתי נמנעת:** המציאות אינן רק טעויות אלא תוצאה בלתי נמנעת של הארכיטקטורה וההיגיון השולטים במודלים גדולים לשפה. הן נוצרות כאשר המודלים מנוטים להשלים פערים במידע או ליצור מידע סביר אך שגוי על סמך נתונים חסרים או מעורפלים.

2. **חווסף שלמות של נתונים האימון:** המאמר מדגיש כי אף מאגר נתונים אינם שלם ב-100%, ולכן LLMs תמיד יתקלו במצבים שבהם עליהם להסתיק או להמציא מידע שלא קיים במאגר הנתונים (המוחבא במשקלים שלו או במערכת נתונים חיצונית).

3. **4 סוגים עיקריים של המציאות:**

- **אי דיקוק עובדתי:** המודל עלול לגנרט מידע עובדתי שגוי בשל "אופן שליפה שגוי" של מידע ממוגני המידע שלו.

- **אי הבנה:** המודל נכשל בהבנת קלט המשמש, ונוטן תשיבות שגויות.
- **מחט בערימת שחת (needle in a haystack):** קושי בשיליפת מידע ספציפי ממאגר נתונים (במשקלים שלו או במערכת נתונים חיצונית), מה שלעתים מוביל למידע מעורב או חלק.
- **המצאות:** LLMs לעיתים ממצאים מידע כאשר הקלט אינו מוכר להם מהטרין סט ולא תואם לשום עובדה ידועה במאגר הנתונים שלהם.

4. **"בלתי מוכרעות":** המחברים משתמשים במשפטו אי שלמות של גdal ובתיאורית חישוביות, ומדגימים שבעיות מסוימות, כגון שליפת מידע עובדתי מדויק (intent classification) וסיווג כוונת המשתמש (needles in a haystack). אין ניתנות להכרעה. המשמעות היא שאין אלגוריתם שיכל למנוע לחלוטין המציאות.

5. **LLMs לא מסוגלים לנגן מתי הם ייעזרו:** המחברים טוענים כי LLMs לא מסוגלים לחזות متى ייעזר הגנרט (מציר הධיה הדועה של עצרת מכונה בתיאוריה חישובית). הם טוענים שנובע מכך כי המודלים אלה אינם מסוגלים לשולט או לצפות במדוקן איזה תוכן הם ייצור, מה שמעלה סיכוי להזאות.

6. **הוכחה שהדיזיות אינן ניתנות לביטול:** המאמר מראה (יש הוכחה) שגם כוונון מושלים או מגנוני בדיקת עובדות לא יכולים לבטל חלוטין הדיזיות. זאת משום שמאגר הנתונים תמיד יהיה חסר או בלתי מספק, ומודלים גדולים לשפה חייבים לייצר פלט שאין ניתן לאימות או סותר.

7. **השפעה של RAG:** למרות שיטכניות כמו הפקת מידע מוגברת נועדו לשפר את הדיקן העובדתי באמצעות שילוף מידע חיוני, אך עדין מסתמכו על פונקציות שליפה לא מושלמות, מה שוביל לטעאות חלקיות או מעורבות.

8. **תפקיד קידוד מיקומי (positional encoding או PE):** המאמר נגע בטכניות PE מתקדמות כמו RoPE ויצד הן משפרות את ביצועי המודלים באמצעות שילוב מיקומים מוחלטים ויחסיים. עם זאת, טכניות אלו עדין לא פותרות את בעיית הדיזיות.

9. **הדיות מבניות:** המחברים מציגים את המושג "הדיות מבניות", ומדגישים שהן תוצאה בלתי נמנעת של הארכיטקטורה של LLMs וכן אין ניתנות למנעה, גם לא באמצעות שיפורים באימון או כוונן.

10. **השוואה למודלים אחרים:** המאמר משווה בין מודלים לשפה למודלים אחרים כמו KANs או מסיק שהמגבילות המובילות להדיות קיימות בכל הארכיטקטורות.

11. **מכנת טירינגן -NsLMs:** מודלי שפה מוצגים כשותם למכנת טירינגן אוניברסלית, מה שאומר שהם יורשים את אותן מגבילות חישוביות, כולל בעיות בלתי-מורכעות כמו בעירה.

12. **השלכות לעיצוב עתידי של LLMs:** המאמר מציע שהפיתוחים העתידיים של LLMs צריכים להתמקד בניהול והפחחת הדיזיות במקום לנסوت לבטל אותן, שכן הדבר בלתי אפשרי מתמטית וחישובית.

<https://arxiv.org/abs/2409.05746>

⚡️⚡️⚡️ המאמר היומי של מילק 14.09.24: Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning

חוקי סקלילינג זה נושא מאד מעניין אך לצערי אני מתקשה למצוא מאמרם באמת שווים סקירה (שמקלים מעבר לניסויים אינסופיים עם היפרפרמטרים שונים). הפעם התמזל מזל ונתקלתי במאמר הלא חדש הזה שהוא נראה די שווה.

המאמר מציע סוג חדש לחוקי סקלילינג בקשר ל- Data Pruning (צמצום נתונים או DP). המחברים מספקים ראיות תיאורטיות (זו הסיבה שאני סוקר אותו) ואמפיריות לכך שצמצום פיסות נתונים מיזורות או פחות אינפורטטיביות יכול לשבור את חוקי הסקלילינג המסורתיים, ולהציג הפחתה מהירה יותר בשגיאה תוך שימוש בפחות משאים.

רעיון: חוקי הסקלילינג של רשתות נוירוניים מתארים כיצד השגיאה (טיטט) יורדת עם הגדלת גודל המודל, כמוות הדטה או כמות הקומפיוט, בהתאם לחוק חזקה (Power Law). עם זאת, סקלילינג זה אינו יעיל, שכן שיפור ביצועים דורש כמות נתונים/משאים אקספוננציאלית. המחברים שואלים האם ניתן להשיג סקלילינג טוב יותר מחוק חזקה על ידי בחירה מושכלת של נתונים.

התמצית: המחברים מפתחים מסגרת תיאורטית המבוססת הלקווה מכוניקה סטטיסטיות, תוך שימוש במודל בסגנון זיקוק מידע (מודלי סטודנט-מורה). מודל זה מתאים לבחינה תיאורטית של pruning data (זריקת נתונים) בגל פשטוטו המתמטי, תוך שמירה על תכונות הכללה (generalization) שנשמרות במודלים מורכבים יותר.

המסגרת המתמטית המוצעת מרכבת מ"מורה" שמייצר דאטה, ומודל "סטודנט" שמנסה ללמידה אותו. הרעיון המרכזי הוא "להעיף דוגמאות על על בסיס הקושי שלהם". קושי של דוגמא נמדד על פי המארגן (הmarker) של דוגמא מסוימת (עם מארגנים גדולים). בעודם נבדק דוגמאות קלות (עם מארגנים קטנים) עברו דאטהסתים קטנים, בעודם נבדק דוגמאות קשות יותר (עם מארגנים קטנים) הם אינפורטטיביות יותר דאטהסתים גדולים. המחברים מראים כי שגיאת הכללה g_E , תליה ביחס בין מספר דוגמאות כולל לפרמטר המודל (α) ובחלוקת מהדטה f שהוורר. המשקנה המרכזית היא שיתוך אופטימלי שובר את חוק החזקה בסקילינג, ומוביל לסקילינג מערכי של הפחתת השגיאת הכללה.

اذ אלו דוגמאות להשair: כאמור עבור דאטהסתים קטנים, עדיף לשמר דוגמאות קלות כדי להימנע אובייפיט, בעודם נבדק דאטהסתים גדולים, משתלים להשair דוגמאות קשות כדי ללמידה גבולות החלטה עדים יותר. יש טענה במאמר שברגע ששומרים את הדוגמאות הקשות ביותר, מתאפשר סקלינג מערכי של הפחתת שגיאת הכללה g_E , עבור דאטהסתים גדולים. המחברים מצאו כי הדעכה המערכית מחזיקה עד לנקודת שבירה קרייטית, שבה הדוגמאות הנותרים כבר אינם מספקים מספיק מידע. מעבר לנקודה זו, דעיכת השגיאה מאטה וועוררת לחוק חזקה.

רווח מידע (Information gain או I_G): המחברים טוענים כי בلمידה עם רשותה המידע השולי שספקת כל דוגמא נוספת עם מספר הדוגמאות, מה שמוביל ליחס חוק חזקה בין גודל הדאטהסט להפחיתה של שגיאת הכללה. אולם, עם אסטרטגיית בחירה חכמה, המצב משתנה. חיתוך מסיר מיותרים מיותרים או בלתי אינפורטטיביים, ומאפשר לכל דוגמה שנותרה לספק מידע ייחודי יותר על המשימה. מתמטית, תכולת המידע של דאטהסט (לסטודנט) פרופורציונלית למספר דוגמאות שנותרו, אך ניתן להאט את קצב הירידה עם בחירה מושכלת של הדוגמאות. ככל מרוח המידע לדוגמא נשאר משמעותית גם כהדתאסט נחתר, מה שמאפשר דעיכה מערכית של השגיאה.

חוסר איזון בין קטגוריות: המאמר דין בכך שבחירת דוגמאות ללא התחשבות בהתפלגות קטגוריות עלול להוביל לחוסר איזון בין יגروم לירידה בביטוי המודל. המחברים מציעים טכניקת איזון קטגוריות שמבטיחה שככל אלו ישארו מיצגות היבט בדאטהסט החתו.

<https://arxiv.org/abs/2206.14486>

🚀⚡️**המאמר היום של מייק 15.09.24:** Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning

אחרי סערה החסיבה בזמן האינפראנס במודל החדש של openai openai הוחלטי לבנור בפוסטים בנושא זהה ונתקלתי במאמר הדי מפורסם הנקרא Q*. מתרברר שהוא נמצא שם בראשית המאמרים האינסופית שאני רוצה לסקור אך לא-ב-20 הראשונים לפחות. מכיוון שקיימות די הרבה סקירות של המאמר הזה ניתן סקירה יחסית קצרה בlij לרדת לפרטים יותר מד'.

המאמר מדבר על "תהליכי החשיבה או תכנון" עבור מודלי שפה. למעשה זה סוג של Co-managed על ידי פונקציית Q המשערך ערך של כל שלב במהלך "החשיבה" של המודל. ככלمر עבור כל שלב ב-pruning reasonano רוצים להבין עד כמה מענה נתון של LLM יקרב אותנו לתשובה הסופית הנכונה. אתם מರיחים כאן פונקציית Q ידוע מעולם למידה עם חיזוקים וזה הניחוש הנוכחי כאן.

כדי לפורמל את הבעיה במנוגה RL צריך להבין מה זה מצב (state) ופעולה (action). במקרה שלנו פעולה היא תשובה של LLM בשלב נתון של תהליכי החשיבה שלו ומצב הוא סדרה של כל הפעולות עד השלב הזה כולל כל התשובות (בסדר כרונולוגי) שהמודול נתן. והמטרה כאמור לבנות את פונקציית Q בהינתן מצב s ופעולה a נתונים בשלב t , כולל לשערר את איות תשובה a_t עבורי התשובה הקודמת a_{t-1} a_1 . ברגע שיש לנו נתונים Q אנו יכולים לבנות את המשך האופטימלי של שרשרת החשיבה $a_1, \dots, a_t, \dots, a_{t-1}, a_{t-2}$. כמובן היינו רצים פונקציית Q אופטימלית כולל צדו שמקיימת משוואות במלן ובעלת תכונות טובות.

אבל איך נוכל לשערר את הפונקציה הזאת אם יש לנו רק מודל עם פרמטרים נתונים שלא מותאמים (ישירות) לכל הסיפור של בחירת שרשרת חשיבה אופטימלית. כולל אין לנו פוליסי אופטימלי שאוותה אנו יכולים למנף ליצור Q אופטימלי. המאמר מזכיר 3 אפשרויות.

- (1) בהינתן>Dataless נתונים של שרשרות חשיבה וציוונים ניתן לשערר Q אופטימלי יחד עם השערור שלו עבור הפוליסי המוקפא לנו (כולל מודל שפה) בצורה alternating (שערור של של כל אחד באמצעות השני כל פעם).
- (2) מרים את הפוליסי הקיים וכל פעמים בוחרים את הפעולה (תשובה) בעלת ערך Q מקסימלי, ומשפרים את שערוכה באמצעות חישוב של התגמול הכלול (עבור כל השלבים). דרך אגב קביעת מה זה התגמול המייד במצב s לא נראה לי טריוני.
- (3) שימוש במודל שפה חזק אחר כדי "לחקות" את הפוליסי האופטימלי ובאמצעות הרצתו לשערר את Q האופטימלי.

כאמור ברגע שיש לנו שערור טוב של Q האופטימלי אנו תמיד בוחרים את התשובה בעלת Q הגבוהה ביותר מפول התשובות של LLM.

از למה יש לנו כוכבית בשם. האלגוריתם שהתקבל מאד מזכיר את A*. המפורסם אך זה כבר נשא לסקירה אחרת...

<https://arxiv.org/pdf/2406.14283>

🚀⚡️: המאמר היום של מיק : 16.09.24

Rethinking Benchmark and Contamination for Language Models with Rephrased Samples

חתיכת נושא זה. לאחרונה אני ניהلت מספר שייחות עם אנשי NLP לא מעטים על הנושא הזה. מי שעוקב אחרי ברשותות החברתיות אולי שם לב כי אני בד"כ לא מתלהב ממודול שפה שניצח את כל המודלים הקיימים בכל הבנטצ'מרקם. הסיבה לכך היא די טبيعית ונובעת מכך שבלא מעט מקרים לא מפורטים באופן גלי את כל הדadata שעיליה המודול אומן.

כמובן שהחseed שלי הוא הדטה(משמעות) האימון היה דומות מדי לאלו שמופיעות בנטצ'מרקם האל. כמובן אני לא בא להאשים אנשים על כך שהם מרים מקרים בכוננה (למרות שבתחום יש מקרים כאלה) אלא אני בא להגיד שאין דוגמאות בDATALESS הדומות מדי לבנטצ'מרקם אין מצלחות לפטלר את הדוגמאות האל. והתוצאה היא מודול שהוא אוברפייט על בנטצ'מרקם כזה או אחר.

כאמור יש שיטות די בסיסיות לבדוקות את הדמיון בין הדוגמאות בDATALESS לדוגמאות בנטצ'מרקם מבוססות על grams-gram ועדיין סמנטי המחווש באמצעות מרחק בין הייצוגי של הדוגמאות בDATALESS ובנטצ'מרקם. המאמר המסתור טוען שהוא לא מספיק וצריך לעשות בדיקה נוספת לזיהוי של דוגמאות אלו. בגדוד המאמר מציע בנוסח בדיקה הסמנטית לרשתות איזה LLM עצמאית לבודקה של דמיון דוגמאות.

בגadol מזהים K דוגמאות הci דומות סמנטיות לכל דוגמא בbenz'マーク ואז מפעלים LLM חזק כמו GPT4 עם איזה פרומופט מתחכם כדי לזרה את הדוגמאות הבאמת דומות. המאמר מראה כי בוצרה כזו הצלחו לתפוא דוגמאות שלמרות שנראות שונות מהוות rephrasing של דוגמא מסוימת מהbenz'マーク. ואז מעיפים את הדוגמה הzo מהדאטסת.

המאמר טוען כי ללא שימוש בשיטה שלהם ניתן "לאמן" מודל B 13 כדי ש"ינצח" את 4 GPT4 על כל הבנז'מרקם - נצחון לא אמיתי אמן.

מאמר ללא יותר מדי חדשנות אף מעלה נושא מאד מעניין

<https://arxiv.org/pdf/2311.04850>

🚀⚡️: המאמר היומי של מיק 17.09.24 : STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning

אני ממשיך לחפור במאמרי שאלוי עיצבו את הנתיב הובילו ל-10 של openai. הפעם נברתי כי עמוק שהגעתי למאמר שיצא לפני שנתיים וחצי (בדף היום הזה כמו 100 שנה במתמטיקה). שימו לב שהמאמר יצא עוד לפני chatgpt. המאמר זהה מציע שיטה לשיפור יכולת reasoning של מודל שפה כאשר בידינו יש דאטסת גדול של שאלות ותשובות D ודאטהסט קטן R _ הרבה יותר (המאמר מדבר על 10 דוגמאות בלבד) המכיל בנוסף גם את שרשרת reasoning .

כאשר אני מדבר על שיפור איקות reasoning אני בעצם מתקoon לפיניטיון של המודל במטרה לקבל מודל חזק יותר ב-questioning. המחברים מציעים אלגוריתם המורכב משני שלבים עיקריים. בשלב הראשון מציגים את הבז' של שאלות למודל שפה כאשר בונוסף לשאלות הפורמופט מכיל את דוגמאות reasoning ה-m question מ- R . המודל מתבקש לבנות שרשרת reasoning לכל השאלות מבז' (לא מ- R) ולהגיע לתשובה הסופית.

את שרשות reasoning לשאלות שהצליחו להגעה לתשובה נכונה מוסיפים לסתו שנקרא לו N _ D . לשאלות שהמודל לא הצליח להגעה לתשובה סופית נכונה אנחנו מוסיפים רמז (במאמר זה נקרא rationalization) שעוזר למודל לבנות את שרשת reasoning . השאלות שהצליחו להגעה לתשובה הנכונה אחרי הרמז גם מוסיפים ל N _ D . לאחר מכן מבקים איטרציה אחת של שיטת מورد הגרדיאנט נבחרת על N _ D ומעדכנים את משקל המודל. חוזרים על השלבים האלה עד שהלוס מתיעץ.

זהו זה, שיטה אינטואיטיבית ופושטה שקיבלה כמה מאמרים השמן די כבדים שבתקווה אסקרו אותם גם כן

<https://arxiv.org/pdf/2203.14465>

🚀⚡️: המאמר היומי של מיק 19.09.24 : Training Chain-of-Thought via Latent-Variable Inference

משיכים בקו הסקירות שהובילו (פחות לעניות דעתך) למודל החדש ('יחסית', יצא כבר לפני שבוע) של openai. במאמר הקודם שסקרטרי STaR דיברנו על איך ניתן לשפר יכולת ריזונייניג של מודל שפה כאשר יש בידינו דאטסת גדול יחסית של שאלות ותשובות D ודאטהסט קטן של שאלות ותשובות עם הריזונייניג. בגודל הרעיון שם היא לרתום מודל שפה לייצר ריזונייניג לשאלות, להוציא שאלות שהrizoniinig שלהם הוביל לתשובה נכונה לדאטסת הקטן ולהמשיך לאמן עד ההतכנסות.

המאמר הנוכחי שיצא בערך שנה וחצי אחריו משלב את הגישה הזו ומציע שיטה ש"ממנפת" גם את השאלות שעבוד המודל יצר ריזונייניג שלא הוביל לתשובה הנכונה. המאמר מכיל מתמטיקה די כבده אז אנסה להעיבר לכם את הרעיון הכללי יחסית פשוטות.

הרי המטרה שלנו היא לעשות פיניטין למודל שפה כך שיכולה הריזונייניג שלו תשתperf. מתמטית ניתן לתרגם את הבעיה לבעה וריאצוניות באופן הבא. אנו מעוניינים לאמן מודל שייצר ריזונייניג עבור שאלה א. מה שיש לנו זה DATAהסט של שאלות x ותשובות y. אז אנחנו רוצים לאמן את המודל להפיק ריזונייניג Z (ניתן להתייחס אליו כמו אל משתנה לטנסי) מהתפלגות בהינתן השאלה x מ-D תוך כדי ניצול של התשובה y. כמובן אנו רוצים למקסם את הנראות (likelihood) של התפלגות המונטנית של הריזונייניג Z בהינתן (עבורה) שאלה x ותשובה y. במילים פשוטות אנו מאמפטים את פרמטרי המודל כך שהנראות הזו תהיה מקסימלית על-D.

אולם אנו לא יכולים לעשות זאת בצורה ישירה ככלומר לא ניתן לדוגם את הריזונייניג בהינתן שאלה x ותשובה y. הסיבה לכך היא שאנו לא רוצים לאן מודל שמייצר ריזונייניג לשאלת yיחד עם התשובה (כי אנו רוצים מודל שיפטור לנו שאלות בלי לדעת את התשובה). אז המאמר הקודם בחר לנצל את תשובה y על ידי פלטור החוצה של Z שהובילו לתשובות לא נכונות. לעומת זאת המאמר זהה מציע שיטה שבה אנו ממנפים גם את ה-Zים הללו נוכנים לשיפור המודל.

כאמור המאמר מנצל כמה שיטות מתמטיות די כבדות לכך ואחת מהם הוא שכלול של Monte Carlo proposal distribution (משמעותו דוגמים במתטרה שזו תחכemos עם הזמן להתפלגות היעד ככלומר זו של ריזונייניג Z בהינתן שאלה x ותשובה y) משתנה עם האיטרציה להערכת התוכנות (Markovian score). climbing Robbins-Monro לחישוב "גודל העדכוון").

מה הקשר ל-MCMC אתם שואלים? אנו כל פעם דוגמים מהתמודל עם המשקלים מהאייטרציה הקודמת (באז') כאשר ה-*h*-chain proposal distribution יתכנס להתפלגות הרצויה. המחברים מציעים לעדכן את משקל המודול גם עבור התשובות הללו ומקווים שזה יתכנס להתפלגות הרצויה. ככל שהאייטרציות עוברות השיטה מעדכנת את המודול יותר עבור נוכנות וגם הנכונות (בכיוון שונים כMOV). ככל שהאייטרציה עוברת השיטה מעדכנת את המודול יותר עבור דוגמאות עם ריזונייניג לא נכון (MOV ל答复 לא נכון) כי רוב השאלות כבר מתקבלות ריזונייניג נכון ו"פחות שווה" להתחשב זהה.

בנוסף המאמר משלב את עדכון משקל המודול המדבר על ידי כך שהוא שומר את הריזונייניג הקודם Z לכל דוגמא ומחשב את גודל (כמו קצב למידה) של עדכון משקל המודול בהתאם. למשל העדכון עבור הריזונייניג של דוגמא שהוביל לתשובה נכון באיטרציה הנוכחיות ולתשובה שגוייה באיטרציה הקודמת גורמת לעדכון גדול יותר עבור המודול. הגישה מקורה במה שנקרא wake-sleep-memoized שמצוות שיטת אימון למודלים ניירו-סימבוליים גנרטיביים בכלל דרך ניצול הזכרן המצטבר של העדכניםים.

וכל זה כדי לשפר את הריזונייניג של המודול - מקווה שהצליחתם להבין את העיקר 😊
<https://arxiv.org/pdf/2312.02179.pdf>

המאמר היומי של מיק :20.09.24   Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning 

משיכים בסקרים מאמרם "החוודים" בסילית נתיב למודל 1 (שרבים כבר התאכזבו ממנה אמנים אף אותו הוא מסקרן מבחינת חידוש הפרדיגמה). המאמר שנסקרו היום פחות מתמטי מזה של אטහט (הכל פורסם בעברוץ הטלגרם שלי) ובתקווה הסקירה תהיה יחסית קצרה וקולה.

מציר שהמאמר מציע שיטה לשיפור הריזוניינג של מודלי שפה כאשר יש לנו דאטහט D גדול יחסית של שאלות ותשובות ודאטහט קטן בהרבה של שאלות ותשובות עם שרשרת ריזוניינג. המאמר מציע שיטה בסגנון של למידת curriculum די נפוצה בלמידה عمוקה - כמה מודלי שפה היכי טריים אומנו עם השיטה זו (בשילוב עם עוד שיטות curriculum די).

בלמידת curriculum מאמנים מודל החל מדוגמאות קלות ובמהלך הלמידה מעלים את קושי הדוגמאות.

אבל איך קשורה למידת curriculum לשיפור יכולת ריזוניינג של מודל שפה. זהה בדיק היופי של המאמר דרך אגב. המחברים שמו לב שאם נספיק למודל את כל שרשרת הריזוניינג מהתחלה ועד השלב די קרוב לתשובה הסופית אז יהיה לא יותר קל לשחרר את השלבים החסרים בשרשראת. וזה בדיק מה שהמאמר עשה. ככלומר המאמר מאמן את מודל (בשיטת RL דומה לSTaR שפרקתי ב-17.09, למידת פוליסי די סטנדרטית) אבל הפעם המודל לומד לשחרר את שלבי הריזוניינג מנוקודות שונות בשרשראת.

המאמר טוען ששיטת למידת curriculum מושלמת לפחות במקרה הזה כי המודל שלמד להשלים שלבי ריזוניינג אחרים מתקשה ללמידה לעשות את מההתחלה ו"מאבד" את הידע שצבר. בעקבות כך המחברים מאמנים משימות ריזוניינג ברמות קושי שונות (בקשר המדבר) יחד עם איזושהי אסטרטגיה חכמה מעולם ה-multi-tasking.

שני דברים אחרים לגבי המאמר הזה. קודם כל פונקציית תגמול (reward) הינה די סטנדרטית כאן עם חידוש קטן שעבור משימות עם תשובה מסוימת המודל מקבל פרט קטן (ולא אפס) אם הוא נותן תשובה מסוימת לא נכונה (-1 במקרה של תשובה נכונה). המאמר משתמש ב-PPO שהיא שיטה די סטנדרטית לפיניטון של LLM'ים אתם לא רוצים שהוא ישכח את כל מה שהוא למד לפני הפיניטון.

<https://arxiv.org/pdf/2402.05808>

המאמר היומי של מיק 21.09.24: REFT: Reasoning with REinforced Fine-Tuning

משיכים לסקור מאמרים שלאלו לכאן ונأتي ל-1.0. הפעם מאמר די בסיסי יחסית שהיה לסקור אותה לפני יומיים אך הטעטלתי לעבור על רשימת המאים שבניתי כדי להבין את זה. הרוחו היחיד לאלו שעוקבים אחרי סקירותי באופן יומי יתבטא בכך שהיא לכם מאד קל להבין את הסקירה הזה אם הצלחתם להבין (בערך) את 4 הקודמות.

המאמר מניח שיש בידינו דאטහט של שאלות ושרשרת ריזוניינג המובילה לתשובה (הנכונה). המאמר מציע לשפר את יכולת הריזוניינג של מודל שפה בשני שלבים:

אימון רגיל (Self-Supervised Fine Tuning): על כל שרשרות ריזוניינג מהדאטහט. ככלומר המודל לומד לשחרר את שרשרת הריזוניינג של כל שאלה ברמת הטוקן כמו שעשה ב-SFT הסטנדרטי.

אימון של למידת פוליסי (שהה המודל עצמו) מעולם (Reinforcement Learning): (מכאן נגזר שם המאמר) כאשר המודל מקבל פרט 1 אם המליח לגנרט שרשרת ריזוניינג המובילה לתשובה הנכונה. תגמול צנוע הרבה יותר ניתן לתשובות מסוימות לא נכונות עבור השאלה שהתשובות עליהן מסוימות גם כן (כמו במאמר הקודם). תגמול 0 מתקבל בכל המקרים האחרים. אימון מתבצע עם PPO די סטנדרט עם שערור די סטנדרט של פונקציית ערך

ו פונקציית יתרון A (כמו במאמר המקורי של ג'ו שולמן מ-[openai](#) לשעבר)

<https://arxiv.org/pdf/2401.08967>

המאמר היומי של מיק 22.09.24: Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking

סקירה זהה ממשיכה את קוו הסקירות "בדרכ ל-10" והפעם המאמר לפחות לפניו השם התקerb ד' מהר למה קורה לכואורה ב-10. כלומר "המודל חושב" לפני שהוא מחייב את תשובתו למשתמש. כמובן שגם המטרה כאן גם שיפור ריזונייניג של המודל.

המאמר משפר את Sta Shakerati לפני כמה ימים ועשה את דרך בניית "שרשות ריזונייניג לוקליים" העוזרים למודל לחזות בצורה יותר מדויקת. כל שרשת ריזונייניג צזו מרכיבת ממנה שנקרה במאמר "טוקני חשיבה" (thought tokens) שהמודל מייצר ותהליך זה ניתן לפרש בתור "חשיבה של המודל". מכיוון שהחיצוי של רוב הטוקנים אינה מושימה קשה במיוחד טוקני חשיבה המאמר מציע לשלב את הייצוג המגיון מטוקנים אלו עם ייצוג הקונטקסט המופק מהטוקונים הקודמים.

לטענת המחברים (הדי הגיונית) טוקני חשיבה של טוקן נתון עוזרים לא רק לחיצוי של הטוקן הבא אלא גם לטוקנים שבאים אחרים. אז המודל מאמין למצסם את דיקוק החיצוי כמה מהטוקונים הבאים. בנוסף המאמר מאמינים טוקנים מיוחדים המסמנים את התחלה ואת הסוף של שרשות טוקני חשיבה: <[startofthought]> ו-<[endofthought]> שגם את הייצוגים שלהם נלמדים במהלך האימון.

האימון מתבצע בשיטת REINFORCE מאוד סטנדרטית הלוקחה מעולם במידה עם חיזוקים. בכל איטרציה עבר כל שרשת של טוקני חשיבה אנו ממסים את ההפרש בין איות החיצוי של כמה טוקנים הבאים (כלומר log-likelihood) לבין הממוצע של אותה איות החיצוי עבור כמה שרשות טוקני חשיבה (שנberosים כל פעם). המאמר טוען שהוא מקטין את השונות שעורך ה-*log-likelihood*. ד"א מה שמסמסים זה סוג של פונקציית האנוועה advantage שדי נפוצה בעולם RL. כאמור השכבה (mixing head) המשלבת ייצוג טוקני חשיבה יחד עם ייצוג הקונטקסט הרגיל מאמנת גם כן.

יחד עם זה המודל עצמו (המשקלים) מאמין ייחד עם טוקני חשיבה וכל השאר (ראו את האלגוריתם). המחברים שמו לב שלא צריך לבנות טוקני חשיבה לטוקנים הנחוצים בקהלות (אנטropophia נמור של וקטור ההסתברויות) ומאפשר לחסוך לא מעט כוח חישוב באינפורנו.

<https://arxiv.org/pdf/2403.09629>

המאמר היומי של מיק 23.09.24: Training Language Models to Self-Correct via Reinforcement Learning

משיכים בקו הסקירות על שיפור יכולת הריזונייניג של מודלי שפה (מסדרת "כל הדרך ל-10"). המאמר הזה של דיפמייניד שיצא לפני כמה ימים משר את עיני מרגע שימושתי לב עליי (לראשונה ראיתי אותו בلينקדאין נראה לי). לך לי לא מאד זמן להבין את העיקר של המאמר זהה כי הוא מכיל הסברים מאוד מפורטים ועמוקים והוא היה הרגשה ש"מרוב עצים לא רואים את העיר".

אוקי כמו שאתם כבר מבינים מהשם המאמר מציע שיטה לשיפור של יכולות תיקון עצמית (self-correction) של מודלי שפה. הנושא נחקר הרבה בשנתיים האחרונות (וגם לפני) והוצעו מספר שיטות לטיפול בבעיה. אלא, כמו שמחברי המאמר מצינים שיטות אלו אין מובילות לשיפור ביצועים ממשמעתי עקב העבודה שהן מאמינות על

התפלגות מוטעית של התפלגות התשובה הראשונה (שאותה מתקנים) של ה-LLM (זה מה שלקח לי לא מעט זמן לתקן מהמאמר).

המאמר מתבונן בשתי שיטות לתקן עצמי (הם עשו SFT על הדאטאסטים המוגנרטים על ידיהם): Star (שスクורי, לפני כמה ימים) ומהמאמר הזה (נקרא Pair-SFT במאמר). בגישה בנו דאטאסט על שלישיות המכילות שאלת, תשובה לא נכונה (כלומר שרשת ריזונייניג המוביל אליה) ותשובה נכונה (גם הריזונייניג שהוביל אליה) כאשר ניתנה על ידי המודל אחריו התיקון העצמי (עם פרופטט מסוים). במקורה השני הלשונית הורכבה מהשאלה, תשובה לא נכונה ותשובה נכונה אקראית (לא אחריו התיקון עצמי) לשאלת זו.

בשני המקורים המחברים ראו שאין שיפור ממשמעותי אחריו התיקון העצמי ואחרי אנליזה די רצינית הגיעו למסקנה כי זה נבע מאי "התאמת של התפלגות התשובה הראשונה" להתפלגות ההתחלתית של המודל. הרציאונל כאן הוא שהוא מאמנים מודל לתקן לא לבדוק מה שהמודל יוצר אלא משחו קצת אחר.

המחברים מציעים שיטה זו שלבית לפתרון בעיה זו. בשלב הראשון אנו מנסים לגרום למקסם את תגמול(מנחים) שיש פונקציית reward נתונה) עבור תשובה נכונה תיקון עצמי (כלומר שרשת ריזונייניג שבובילה לתשובה זו תוך שימוש הפלט המאומנת (או פוליס' בשפת RL) של LLM קרובה להתפלגות ההתחלתית שלו. כלומר עושים סוג של PPO כאשר הידע הוא מקסום של ההפרש בין התגמול עבור התשובה הנכונה לבין מחקר KL בין התפלגות המאומנת (כלומר פוליס') להתפלגות ההתחלתית.

בשלב השני ממקסמים את סכום התגמולים אחריו שתי התשובות (לפני ואחריו התיקון) תוך שמירה של הקירבה של התפלגותיהם שלהם להתפלגות ההתחלתית של LLM.

מקווה שהסבירתי פחות או יותר מובן...

<https://arxiv.org/pdf/2409.12917>

🚀 24.09.24: המאמר היומי של מיק 🔥🚀 LLMs Still can't Plan; can LRM's? A PRELIMINARY EVALUATION OF OPENAI'S O1 on PLANBENCH

סקירה של מאמר שלא מכיל מתמטיקה בצורה מפורשת...מאמר זה בוחן את יכולות התכנון של מודלי שפה גדולים (LLMs) ומודלי חסיבה גדולים (LRMs) כמו משפחת O 0 במציאות סדרת מבחנים הנקרואט .PlanBench

הוא מערכת מבחן מקיף שפותה ב-2022 להערכת יכולות התכנון של LLMs. מרכיביו העיקריים:

- מערכת סטטיסטית של 600 בעיות Blocksworld הכוללות 3 עד 5 קוביות.
- גרסה מוסתרת (Mystery Blocksworld) של אותן בעיות, שבה המונחים והפעולות מוחלפים במילים אקראיות כדי לבחון הבנה מופשטת.
- בעיות PlanBench מורכבות יותר עם 6 עד 20 קוביות, הדורשות תוכניות ארוכות יותר של 20 עד 40 צעדים.
- בעיות בלתי פתירות, שנוצרו על ידי הוספה "יעד" בלתי אפשרי לבעיות קיימות.

PlanBench נועד להיות כלי גמיש ומקיף להערכת יכולות תכנון של מודלי שפה תוך בחינת היבטים שונים של תכנון כמו הבנה מופשטת, התמודדות עם מורכבות, ויזיה בעיות בלתי פתירות.

החוקרים מצאו כי LLMs השתפרו בביצוע תכנון בסיסיים, כאשר המודל הטוב ביותר, BLaMA 3.1 405B, השיג

דיק של 62.5% במשימות Blocksworld פשוטות. עם זאת, LLMs נכשלו במשימות בעלי פתרון סביר יותר.

לעומת זאת, מודל-hRM החדש של OpenAI, הציג שיפור משמעותית, עם דיק של כמעט 98% במשימות Blocksworld פשוטות ו-52.8% במשימות עם פתרון סביר. למחרת זאת, הביצועים של-hRM ירדו משמעותית במשימות מורכבות יותר ובבניות בלתי פתירות.

עם זאת החוקרים מדגימים את החשיבות של הטריד-אופים הכללים עילוות, עלות וערביות לנכונות הפתרון (ככה כתוב במאמר) בהערכת מודלים אלה. הם מצינים כי-hRM יקר משמעותית להפעלה ואינו מספק ערבויות לנכונות, בניגוד למתכני AI קלאסיים. המשקנה היא שבعود LLMs כמו-hRM מציגים התקדמות, הם עדין רחוקים

מליהוות פתרון כללי ואמין לביעות תכנן.

<https://arxiv.org/abs/2409.13373>

🚀⚡️: 26.09.24 🚀⚡️ **RRM: ROBUST REWARD MODEL TRAINING MITIGATES REWARD HACKING**

מאמר נחמד שマー את עיני עקב העבודה שהוא דין בנוגע פונקציית תגמול (reward model) או RM) של מודלי שפה. RM הנחוץ בתהילך היישור (alignment) של מודלי השפה המבוססים על RLHF שמטרתו מאוד בגודל לאמן מודל שפה להבחן בין תשובה טוביה לתשובה רעה.

הנושא נבדק באינטנסיביות בשנים האחרונות והוצעו מספר שיטות לעשות בכך שימושים שונים של (Proximal Policy Optimization PPO) כוגן DPO, ORPO ועוד רבים שחלקים סקרטי. בדרך כלל לאימון RLHF נדרש דاطה מסוימת המורכב משליישיות של שאלות ו-2 תשבות, אחת יותר מועדף (המנצחת או-W) והשנייה הפחות מועדף (מפסידה או-L). במהלך אימון RLHF המודל לומד להגדיל את הנראות של התשובה W להקטין את הנראות של תשובה L דרך מקסום של הפרש reward-shalls (עם סיגמוד ולוג) תחת אילוצים כמו שמירה על הקربה בין התפלגות הפלט של המודל המקורי למודל ההתחלתי.

המאמר מציע להתבונן באימון RLHF מזוית כדי מעוניינת וسؤال את השאלה האם הצורה של תשבות משפיעות לנו בצורה לא מכונה על תוצאות אימון בלי קשר לשאלת. כלומר המודל עשויה "reward hacking" ומשתמש בתוכנות של התשובות בלבד ללא קשר לשאלת כדי לאפטם את משקל המודול. כלומר המודול יכול ללמד לנצל דפוסים שונים כמו ("sure, this is the response") בלבד.

כדי להתגבר על הבעיה זו המאמר מציע לעריך תשבות לשאלות שונות כלומר לעשות סוג של אוגמנטציה ולأمان את המודול כך זהה יקשה עליו לבצע reward hacking. למשל שתי תשבות לא רלוונטיות לשאלות אחרות (W ו-L) לשאלת נתונה אמרות לקבל אותו התגמול ואילו תשובה W המתאימה לשאלת ותשובה L לשאלת אחרית אמרה עדין לתת reward גבוה ל-W ו-reward נמוך ל-L מה שאלה האחרת. יש כמובן צירופים נוספים שנייתן להנדס ולأمان את המודול עליהם בצורת RLHF.

דרך אגב המאמר בונה פריימורק סיבתי לבעה זו כולל DAG, סטם שהם separate-d וכדומה אבל אני לא בטוח שכל זה נכון להבנת המאמר. זה אמם שగזל ממי זמן רענן המושגים האלה אבל כמה שיחות עם סונטו עזרו לי מאד.

<https://arxiv.org/abs/2409.13156>

🚀⚡️: 27.09.24 🚀⚡️ **REWARD-ROBUST RLHF IN LLMs**

הסקירה של היום הינה בנושא די דומה לסקירה של אטמול (26.09.24). נושא של הסקירה הוא שיפור של יישור (alignment) של מודלי שפה במהלך אימון RLHF. גם המאמר הזה מציע שיטה שבאה "لتקן" את פונקציית התגמול (reward) אבל מזוית טיפה שונה מאשר המאמר שסקרנו קודם.

המחברים מצבעים על כך שימוש בפונקציית תגמול יחידה במהלך אימון RLHF אינם אופטימלי מכמה סיבות. הסיבה הראשונה היא חוסר עקבות בין המתאים במהלך תיאוג הדאטא המשמש לאימון RLHF (כלומר תשובות מעודפות ולא מעדפות לשאלות מהדאטסת) שעולות לגרום לתשובות "מובלבלות" של המודל לאחר האימון. הבעה השנייה היא reward hacking של המודל המתבטה בכך שהמודל לומד להחזיר תשובות הממקסימות את פונקציית התגמול תוך מתן תשובות לא "מיושרת" עם העדפות המתאים או לא הגינויות.

המאמר ניגש לסוגיה זו מנוקדות מבט בייסיאנית. אם נניח שקיים פונקציית תגמול אידיאלית שאין לנו גישה אליה אז ניתן להתבונן בכל פונקציית תגמול שנבנה איזה דגימה מרחיב "פונקציות תגמול רועשות". המוחברים מציעים לכמה את אי וודאות שיש לנו בפונקציית התגמול על ידי אימון של כמה פונקציות תגמול.

از איך כל הopor זהה עובד? קודם כל מאמנים פונקציית תגמול רגילה דרך נוסחת Bradley-Terry הסטנדרטי.

לאחר מכן מאמנים כמה פונקציות תגמול שימדלו לנו את אי הוודאות. בשביל זה לוקחים backbone רגיל (מודל שפה) ומושגים אליו כמה ראשים (heads) שככל אחד הוא למעשה פונקציית תגמול. כל ראש מאומן לפלוט את התוחלת ואת השונות של ערך התגמול והtagmol עצמו מוגרל מהתפלגות גאומטרית על ידיהם.

פונקציית לוס שהם משתמשים לאימון הראשונים היא די לא טריואילית אך בגודל מוגזעת את השגיאה הריבועית של שערוך התגמול (זה קצת מורכב ומסתמך על פונקציית תגמול סטנדרטית מהשלב הראשון בנוסף לגישת Bradley Terry). במהלך האימון כל דוגמא מוגרלה (מנוות) לרأس שלו וכך אנו מקבלים כמה פונקציות תגמול.

המחברים טוענים שהם "היו רוצים" (זהם השתמשו בה על דוגמאות הצעוע שלהם) לבנות את הלוס עבור אימון RLHF בתור צירוף לנארו של פונקציית התגמול הרגילה התגמול המינימלי בין כל פונקציות התגמול. כאן האיבר השני למשה מהו שערוך של אי הוודאות שדנו בה לעיל. באופן פרקטי במהלך אימון RLHF הם בוחרים ערך התגמול המתאים בפונקציית התגמול בעלות שונות הנמוכה ביותר.

<https://www.arxiv.org/abs/2409.15360>

28.09.24: המאמר היומי של מיק 🚀 Meta-Whisper: Speech-Based Meta-ICL for ASR on Low-Resource Languages

זמן לא סקרתי מאמר על אודיו ומשלים את הפער היום עם סקירה קצרה וקלילה. בדיקן כמו במודלי שפה גם במודלי אודיו כמו whisper למשל יש יכולת למידה *in-context* או ICL בקצרה. ICL היא יכולה של מודל לבצע משימה שלא אומן עליה באופן מפורש אחריו ש"מראים לו" כמה דוגמאות המדגימות את המשימה (נגיד, כמה זוגות של שאלות ותשובות רצויות).

מתברר שמודלי אודיו גם ניחנים ביכולת כזה. ככלומר בהינתן זוג של קטעי אודיו (שאלה ותשובה) ניתן לאמן את המודל לענות על שאלה אחרת, ש邏וגשת לא לאחר כן בצורה של טקסט. אבל איך ניתן לבחור את הדוגמא מהדאטסת (אודיו) של שאלות ותשובות שתמקם את ביצועי המודל לשאלה נתונה.

זה בדיקן מה שהמאמר המסורק עשה. הוא מציע לבחור זוג אודיו (שאלה ותשובה) לשאלה טקסטואלית נתונה על סמך דמיון בין ייצוג לבני היצוג של הזוג. הייצוג כאן הוא הפלטים (hidden states) של השכבות השונות של

המודל עברו האודיו והשאלה הטקסטואלית. והמטריקה KL divergence הד' סטנדרטי. לדעתם אודיו של שאלות ותשובות נתן אני שומרם את כל הפלטים של השכבות וכל שאלת אודיו בוחרים את הזוג הדומה ביותר לפि מטריקה זו.

שכחתי לציין שהמודל עבר פיניטון למשימת ICL בשיטת SoRA הידועה...

זהו זה - סקירה קיליה כמו שהבטחת.

<https://arxiv.org/abs/2409.10429>

המאמר היומי של מיק : 29.09.24 ASR Error Correction using Large Language Models

ממשיר לסקור מאמרים בדומין אודיו. הפעם נדבר על מאמר המציע שיטה לשיפור איכות של פענוח אות דיבור ניתן להשתמש בה במערכות ל-*h*-*ASR* או בקצרה *ASR*. המטרה בכל הסיפור זהה היא לתמלל אות קולי או במילים פשוטות להבין מה נאמר שם.

בד"כ הקלט ל- *ASR* הוא כמה פלטים של המודול שנקרא *Error Correction* או *EC* שמטרתו היא ליצור כמה וריאנטים של תמלול **Z** ("בעל" "סבירות גבואה ביותר") עברו אות דיבור נתון. למעשה מטרתו של ה- *EC* היא לבנות את התמלול הסופי בהינתן **Z**.

בעידנו של מודלי שפה עצמאיים ניתן למונפ את יכולתם למשימה זו בצורה די ישירה. לעומת זאת מזינים ל-*LLM* את הוריאנטיים השונים של התמלול ומבקשים *M-LLM* לבחור את התמלול הגיוני ביותר מבחינה סמנטית (עם פרומפט מתאים). המאמר בחר *LLM* לא סטנדרטי המורכב מאנקודר ומדקודר (כמו במאמר המקורי של הטרנספורמרים) למשימה זו וזה עבד לא רע. אם יש לנו דעתם המכיל את התמלולים מה-*ASR* והתמלול הנוכחי, ניתן לבצע פיניטון.

אם ניתן לעשות יותר טוב? מתרבר שכן אם בנוסף לתמלולים אנו מזינים למודל שפה גם את תוכנות אות הדיבור עצמוו (למשל ייצגו אחריו האנקודר או מטה-דעתה שלו) ניתן לשפר את הביצועים של ה-*EC*. המחברים מציעים לבנות את התוצאה באמצעות מיקסום של סכום משוקל של הנראויות (*log-log*) מהסעיף המקורי (בהינתן התמלולים מהסעיף המקורי) והנראות של התמלול בהינתן התוכנת של סיגナル הדיבור עצמו. באופן לא מפתיע זה משפר את הביצועים כי המודל מקבל יותר מידע רלוונטי.

עוד שככל אחד הוא תוספת ההתחשבות במרקח *Levenshtein* מינימלי בין הפלט הסופי של *EC* לבין הפלטים של *ASR* (המזינים ל-*EC*). מרקח לפיניטון הוא מدد הבודק את מספר השינויים המינימלי הנדרש כדי להפוך מחרוזת אחת לאחרת. לעומת זאת בוחרים את התקoon הקרוב ביותר (מבחינת LD) לאחד הפלטים של ה-*ASR*.

מקווה שלא פספסתי שום דבר ...

[arxiv.org/pdf/2409.09554](https://arxiv.org/pdf/2409.09554.pdf)

המאמר היומי של מיק : 30.09.24 SCHRODINGER'S MEMORY: LARGE LANGUAGE MODELS

בום הסוער זהה (למרות שהסקירה שיכת פורמלית לאתמול - אשלים את הפער בימים הקרובים) נסקור מאמר די קליל עם שם מאד לא קליל. כי אין דבר קליל שיכול בתוכו את שמו של שרדינגר - ספק אם הצלחתו להבין בצורה טובה מספיק את המשוואה של שרדינגר עוד בקורס פיזיקה 3 באוניברסיטה במוסקבה לפני עשרות שנים.

גם סיפורו של חתול שרדינגר לא התבהר עד עכšíו.

אוק", ס"י מנו עם החוקרים. המאמר חוקר (אמפירית) נושא ד' רציני והוא הזכרון של מודלי שפה. כשאנחנו שואלים LLM מה עיר הבירה של שבדיה, איך הוא ידוע שזה סטוקהולם. המאמר טוען כי זיכרון LLM פועל על ידי התאמת דינמית של פלטים לפלטים. ככלומר המודל "בוחר" איך לשולף את המידע מהזיכרון ובונה אותו על סמך הקלט.

המחברים מסבירים את איך פועל הזיכרון של מודלי שפה באמצעות ניתוח של ארכיטקטורת הטרנספורמרים. מנגנון ה-attention (כלומר מקדמי attention שלו) למשהו מאפשרים למודל לבנות את הפלט כפונקציה דינמית של הקלט (כלומר לא קבועה כמו ב-MLP או ConvNets).).

המחברים משתמשים ב- Universal Approximation Theorem או UAT כדי להסביר את יכולת של שילפת מידע שנלמד במהלך האימון על בסיס תוכן של הקלט. המחברים טוענים כי ניתן להבין מנגנון זה בתור "יכולת קירוב דינמית בסגנון UAT" (המשפט המקורי מדבר על יכולת קירוב סטטistica של מודלי ML) כאשר המודל מתאים תוצאה מתאימה על בסיס הקלט, וה透פה הנכפית ניתן להגדיר בתור זיכרון.

הם מכנים זאת "זיכרון שרדינגר" מכיוון שניתן לומר זיכרון זה רק על ידי "שאלות" וניסיונות התגובה שלו; אחרת, הזיכרון נשאר בלתי מוגדר. בנוסף במאמר נדונים גורמים המשפיעים על ביצועי LLM: גודל המודל, איות/כמות הדטה והארQUITECTURA. המחברים טוענים שהזיכרון של מודלים באוטו הגדל מושפע מזווון האימון שלהם ואם המודל אומן על יותר DATA איות איז הוא משתפר (אין הפתעות כאן).

ולבסוף נעשות הקבלות בין ארכיטקטורת LLM למבנה המודולרי של המוח האנושי (את זה פחות אהבתי אבל זרמתי).

<https://arxiv.org/pdf/2409.10482.pdf>

המאמר היום של מיק - 01.10.24: 🚀⚡️

Larger and more instructable language models become less reliable

שנה טובה, מתוקה ושקטה לעוקבי היקרים! אני חושד שהמאזן הקלורי של רובכם הופר בבודקך אז אני מביא לכם סקירה קלילה (פורמלית של אטמול). ודרך אגב הסקירה של היום תהיה אוסף של כל הסיקירות עד עכšíו ואני אפרנס את זה מחר בבודקך.

המאמר שנסקור היום הוא לא מתרטטי והוא דן ביכולות של מודלי שפה. המadd מתבונן ביכולות של מודלי שפה לפטור בעיות דרך הפריזמה של 3 מדדים שונים. השניים מהם הם ד' סטנדרטיים וברורים והם אחוז נוכנות/אי נוכנות של התשובה אף השלישי הוא אחוז הימנעות של מודל שפה מהתשובה. אך בלי מקרים מודלי שפה בוחרים להגיד לנו שלא יודעים את התשובה ולפעמים זה ד' מעצבן (אבל לפעם ממש לא).

המחברים מצאו כי LLM נכשלים ביצירת "אזור" פעולה אמין לבעיות קלות": אפילו בנסיבות הנתפסות כפשות על ידי אדם, SLLMs ממשיכים לעשות טויות. ככלומר אין "קלט בטוח" ברור של באיזור קשיי נמור שבו המודלים מבצעים באופן עקבי ללא שגיאות.

שיפורים ביצועים (הנובעים מזווון DATA יותר טוב, אימון משופר וישיור) מתרחשים בעיקר בעיות מורכבות, בעוד SLLMs ממשיכים לטעות במקרים קלים: ככלומר SLLM יותר חזקים מראים ביצועים משופרים בנסיבות מסוימות. עם זאת, שיפור זה אינו מתרחב באופן אחד למשימות פשוטות יותר, מה שיוצר חוסר התאמת בין ציפיות אנושיות לביצועי המודל.

אימון עיל (המאמר קורא לזה *shape-up*) מפחיתים הימנעות אך מגבירים אינכונות של התשובות: המאמר מראה שמודלים חדשים וחזקים יותר פחות נוטים להימנע מעתן תשובות. עם זאת, הפחתה זו בהימנעות מלאה לעיתים קרובות בעלייה בתשובות לא נוכנות במקום תשובות נוכנות.

בנוסף אחוז הימנעות לא עולה עם רמת הקושי של הבעיה: הינו רצים כי $\text{Prob}(\text{הימנעות|קושי})$ יהיה קבוע, כלומר מודלים הי נמנעים מלהונן לעתים קרובות יותר ככל קושי המשימה עולה. אולם המחברים מראים שישורי הימנעות נשאים יחסית קבועים בכל רמות הקושי.

המחברים גם בדקו את יציבות תשובות המודל לניסוחים שונים של הבעיה וממצאו כי מודלים חזקים יותר מפגינים יציבות גבוהה יותר לניסוח המשימה (פרומפט). לעומת תשובות תלויות בניסוח הבעיה. למן שיפורים ביציבות, עדין יש אזוריים (של בעיות) שבהם הביצועים יכולים לשנתנות משמעותית בהתאם לניסוח שנעשה בו שימוש, אפילו עברו מודלים מעוצבים.

בנוסף השיפורים ביציבות התשובה לא מונוטוניים (מבחן קושי הבעיה): חלק מהניסוחים (של הבעיה) מבוצעים טוב יותר במרקם מורכבים אך גרווע יותר במרקם קלים: הקשר בין יעילות הניסוח וקושי המשימה אינו תמיד פשוט. חלק מהניסוחים שעובדים היטב למשימות מתaggerות עשויים לבצע באופן גרווע במקרים קלות יותר, מה שמסביר את תהליך בחירת הניסוח.

עוד תוצאות מעניינות רבות במאמר זהה - ממליץ בחום להעיף מבט...

<https://www.nature.com/articles/s41586-024-07930-y>

3.10.24: המאמר היום של מיק 🚀🚀

Transformers are Expressive, But Are They Expressive Enough for Regression?

שוב מאמר על הטרנספורמרים אבל קצת שונה מהמאמר הסטנדרטי על LLMs. המאמר הזה מציג חקירה عميقה לגבי expressiveness של הטרנספורמרים, תוך בחינה ספציפית של יכולתם לתור משערći פונקציות אוניברסליים (כאלו שניתן לקרב אותם כל פונקציה חלקה בדיק נתון). המחברים מתaggerים טענות קיימות לגבי expressiveness של הטרנספורמרים ומספקים הוכחות תיאורתיות ואמפיריות אחד שתומכים בהשערתם שהטרנספורמרים מתקשים לקרב (לשערך) באופן מדויק פונקציות חלקות.

לפני 4 שנים הוכח שהטרנספורמר(האנקודר) מסוגל לשערך כל פונקציה רציפה אם יש בו מספיק שכבות (בלוקים של טרנספורמר). המשפט הוכח לפני כ 4 שנים והוא מראה שהטרנספורמר בעל שכבות רבות למעשה יודע לשערך ופונקציה קבועה למקוטען (*piecewise constant*) ועם הגדול המינימלי של אינטראול הקביעות (=רחלוציה) סהinctן מדי אז ניתן לשערך באמצעותו כל פונקציה חלקה בכל דיק.

המאמר המסתוקר מתמקד במחקר של הרחלוציה ס הנדרשת לשערוך בדיק נתון של פונקציה חלקה. התמונה התיאורטית המרכזית של המאמר היא משפט 4.1, אשר קובע חסם עליון על גורם הרחלוציה ס עבור שמכיל מאפיינים שונים של פונקציה מדורבת f .

משפט זה ממשמעות מכמה סיבות:

א) הוא קשור ישירות את גורם הרחלוציה ס לנגזרות של f . קשר זה מבהיר מדוע פונקציות חלקות עם נגזרות המשתנות במהירות מהוות אתגר קשה עבור טרנספורמרים.

ב) החסם מראה יחס הפוך בין ס לבין הנגזרות החלקיים של הפונקציה. עבור פונקציות עם נגזרות גדולות, ס חייב

להיות קטן כדי לשמר על איזות הקירוב. זה אומר בעצם שאנחנו צריכים יותר שכבות של טרנספורמרים כדי לקרב בדיק גובה את f .

ג) המונח האקספוננציאלי $1/(md+p)$ ביחס לצביע על כך שככל שמדובר הקלט או ממד האמבדינג d גדלים, גורם הרזולוציה δ חייב לפחות אקספוננציאלית כדי לשמר על אותה איזות קירוב.

ד"א המחברים מספקים הוכחה מפורטת למשפט זה, תחילה למקורה החד-ممדי ולאחר מכן בהכללה לממדים גבוהים יותר..

יתר על כן, המחברים מקשרים את התוצאה התיאורטית זו להשלכות המעשיות על ארכיטקטורות טרנספורמה. הם מראים שמספר השכבות הנדרש לקירוב הולם גדול $C((dm)^{\delta}/\delta)$, מה שהופך ללא ישים מבחינה חישובית עבור δ קטן וממד הקלט בגודל ביןוני d . ככלומר צריך יותר מדי שכבות הטרנספורמרים בשביל זה.

המחברים ביצעו ניסויים מקיפים על הטרנספורמר כדי להוכיח את ממצאיםו התיורטיים. הם עשו 2 ניסויים עם הבנאים' מרכיבים הבאים:

א) EXPT (רגסיה): בדיקת יכולתם של טרנספורמרים לקרב ישרות פונקציות חלקות.

ב) II-EXPT ("סיגוג מקוונטט"): בדיקת יכולתם של טרנספורמרים לקרב פונקציות קבועות למקוטען.

התברר כי הטרנספורמרים מתפקדים באופן גרוע משמעותית ב-EXPT בהשוואה ל-II-EXPT, שהוא תומך בהשערה שהם מתקשים בקירוב פונקציות חלקות.

הגדלת מספר השכבות, ראשית מגנון *attention*, או ממד אמבדינג אינה מושרת באופן משמעותי את הביצועים על פונקציות חלקות. לעומת הטרנספורמרים מציליםם ל紧紧围绕ם הולם פונקציות קבועות למקוטען עם רזולוציה δ לא קטנה במיוחד.

<https://arxiv.org/pdf/2402.15478>

🚀⚡️ המאמר היום של מייק - 04.10.24 - Were RNNs All We Needed

המאמר הזה משר את תשומת ליבי כי יש לו "needed we all" בoutuרת. סיבה שאינה ב-100% ברורה לי מאמרם كانوا יוצרים ב-*diffusion* חזק לסקור אותם. אך ככה הגעתו למאמר זה שאלולא השם ננראה שלא הייתה מגיע אליו.

המאמר מציע לשפר את ה-RNN כך שנוכל להפעיל אותו בצורה מקבילית במהלך האימון. הסיבה העיקרית ש-RNN כמעט לא מצליח שימוש הום הוא חוסר היכולת שלו להתאמן באופן מקביל לכליomer לביצוע חיזוי של כמה טוקנים מסוימים. הטרנספורמרים לעומת זאת כן ניחנים ביכולת זה או יש להם מגבלה דומות סיבוכיות ריבועית במונחי אורך הסדרה (שכוabbת לנו בעיקר באינפראנס כי מאמנים אותם פעמי אחת) שמקשה על השימוש (לפחות הנאייב שליהם) לסדרות מאד ארוכות.

מצד שני ל-RNN יש יכולת יותר טוביה לבצע סדרות 매우 ארוכות כי כל ה"זיכרון" שלהם מוקוד בכמה וקטורים 1,2 או 3) והסיבוכיות החישובית שלהם פרופורציונלית לאורך הסדרה ולא לריבועו שלו (גם באימון וגם באינפראנס). כאמור הבעיה הגדולה של RNNS שדי הרגה את הארכיטקטורה זה היא אי יכולתה לאפשר חיזוי מקבילי באימון. זה שהופך את האימון על כמות נתונים עצומות כמו שמקובל היום (עשרות טריליאונים טוקנים) עם

RNNs לארוך מדי ולא פיזיבלי.

חשיבות להבין שהסיבה לחוסר יכולת לחזות בצורה מקבילה נובעת מהמעברים הלא לינאריים בין המצביעים החבויים ב-RNN (גם ב-LSTM וגם ב-GRU).

לאחרונה SSMs (או State Space Models) ניסו לטפל בבעיה זו דרך ארכיטקטורה שבה המעברים האלו כلينאריים וארכיטקטורת מבנה (שסקרטה בהרחבה לפני כמה חודשים) ששללה SSMs לרמת ביצועים קרובה לטרנספורמרים. בפוסט Labs A21 השתמשו במבנה CABN של הארכיטקטורה החדשה שלהם לפני חודשים (יחד עם הטרנספורמרים).

עכשו אתם שואלים מה המאמר המסורק עשה בכךן. כאמור הבעיה הגדולה ב-RNN היה מעברים לא לינאריים בין המצביעים החבויים. המתחברים פשוט הורידו את התלות הלא לינארית מהמשוואות של LSTM ו-GRU. מה שהתקבל כתוצאה לכך ניתן למקובל במהלך האימון (אבל דורש יותר זיכרון מהגרסאות הרגילוט). יצא משהו די דומה למצבה - גם כן המצביע החבוי תלוי באופן לינארי במצב החבוי הקודם ובאופן לא לינארי ביצוג האיבר הנוכחי של סדרת הדאטה.

מה שמשמעותו הוא קצת כאן זה ביצועים טובים מדי - אני קצת חסן אבל בואו נראה מה קורה עם הארכיטקטורה הזו בעתיד.

<https://arxiv.org/abs/2410.01201v1>

🚀⚡️ המאמר היומי של מיק - 06.10.24: CONTRASTIVE LOCALIZED LANGUAGE-IMAGE PRE-TRAINING

משיכים הפסקה בסקירות על מודלי שפה ועוברים לסקירות על מודלים מולטימודליים (שפה ותמונות). טוב, הפסקה למחצה. אתם בטח זוכרים את המודל שנקרא CLIP שעשה הרבה רעש לפני כמה שנים.

CLIP הוא אחד המודלים מולטימודליים הראשונים שהצליח לייצר אמבדינגס חזקים ומושרים (aligned) של טקסט ושל תמונות. מושרים הכוונה של הייצוגים של תמונה וטקסט שמתאר את תוכנה קרובים אחד לשני בזמן שהייצוגים של תמונה וטקסט לא מתאימים רוחנית אחד מהשני (במקרה זה ביחס למרחק קווין ביניהם).

המודל הזה אומן על דאטסהט ענק של תמונות והכותרות שלהם (או טאגים) מהאנטראנט כאשר אימנו אותו תוך שימוש בטכניקה למידה ניגודית (contrastive learning) או CL. בגלל מדובר טכניקות CL מאמנות להפיך "יצוג סמנטי מדאית" (סוגים שונים) כאשר המטרה היא לקרב את הייצוגים (אמבדינגס) של פיסות דاطה קרובות (או חייבות) ולהרחק ייצוגים של פיסות דاطה לא דומות (שליליות). במקרה של CLIP פיסות דاطה חייבות הם הייצוגים של תמונה והכותרת שלה ואילו הזוגות השליליים בניוים מכותבות ותמונות שנבחרו באקראי.

המאמר שנסקור אחד כאמור משככל את CLIP על ידי הקניה של יכולות לקליזיה לייצוג. הכוונה כאן שהמחברים מאומנים ייצוגים של תמונה ושל טקסט באופן כזה שבاهינתן יציג התמונה | יציג התיאור של פאץ' ב | המכיל אובייקט מסוים יהיה ניתן להפיך ב"קלות" את מיקום האובייקט בתמונה.

במילים פשוטות נניח שיש לנו אריה עומד וושאג בתמונה הנמצאת ב-box bounding (המודגר על ידי רביעה של קווארדינטות שלו בתמונה) המסומן ב- B . המחברים מאומנים רשות אנקודר לתמונות I_f רשות אנקודר לטקסט T_f כך שיציג התמונה I_R "יצוג" אריה עומד וושאג" T_R , המופקם על ידי שני האנקודר האלו (בהתאם) כך שרשות ردודה יחסית (נקראת prompter במאמר), המקבלת אותם, תוכל לחזות את מיקום האריה B בתמונה. דרך אגב המיקום כאן לא חייב להיות מתואר על ידי box bounding אלא יכול להיות מוגדר (בערך) על ידי כמה

ניקודת, תיאור כללי (נגיד חיה, בלי להזכיר שזה אריה) ובעוד צורות.

האימון נעשה בלמידה הניגודית כמו ב-CLIP המוקורי. אבל בנוסף להוגה הרגיל שלו יש כאן עוד לוס ניגודי' המקרב את ייצוגים של כוורת הפאץ' בתמונה לייצוג המופק על Prompter מייצוג התמונה ומהמתאר של הפאץ' (נגיד BB) ומרחיק את הייצוגים האלה לפאצ'ים שונים. כמוון שה-Prompter גם מאמין לנו כד',

המאמר משתמש במודלים מאומנים למטרת זיהוי אובייקטים בתמונה (L2W0) ובמודלים מאומנים אחרים (VeCap) למטען כוורות לפאצ'ים האלה.

מאמר די חמוד וקליל ...

<https://arxiv.org/pdf/2410.02746>

המאמר היומי של מיק-08.10.24: CONTEXTUAL DOCUMENT EMBEDDINGS

זמן לא סקרתי מאמר בנושא של Document Retrieval או DG. למעשה DG מהווה שלב של Retrieval Augmented Generated (RAG) או שטרכתו היא לאייר את המסמכים הרלוונטיים מסט המסמכים D. בדרך כלל זה נעשה על סמך קירוב של האמבדינגים (הנמדד על ידי מרחק קווין) של המסמכים ושל השאלה המופקים על מודל שפה כלשהו.

יש שכליים למעטים לשיטה זו, למשל לחלק כל מספר לצ'אנקים ומשתמשים בייצוג שלהם לחישוב הקרבה. יצא לא זמן מאמר שהציג להוסיף תמצאות לכל מסמך וכמוון קיימות עשרות או מאות אחרות.

אם יש בידינו דאטסהט של זוגות T-D המרכיבים מ- (שאלה, מסמך רלוונטי) אנו יכולים לעשות פיניטיון לאמבדינגים כללי, כלומרamo שני מודלים: הראשון לחישוב אמבדינגן של המסמכים והשני לחישוב אמבדינגן של השאלה. בד"כ זה נעשה עם למידה ניגודית שמאומנת לקרבת את ייצוגי של כל השאלה לייצוג המסמך הרלוונטי לו ומרחיקה אותו מכל מהייצוגים של שאר מסמכים.

המאמר מציע שיטה שמשפרת את התהילהזה על ידי הוספת קונטקט ליצוגים (=אמבדינגים) האלה. אם יש לנו מסמך שניון לשירותו לכמה תחומים (=דומיניים) אנו רוצים שהאמבדינגן של המסמכים ישתנה בהתאם לדומין של השאלות. כאמור אם השאלה צפויות להיות מהדומין של רפואה אנו רוצים שהאמבדינגן ישקפו את האספקטים הרפואיים ובעור דומיין הספורט שהייתה יותר "מכoon" לספורט. כאמור אנו צריכים כאן contextualized embedding בתלות בשאלות מ-T-D ובסט המסמכים D בעצמו.

המאמר בוחר לעשות זאת על ידי אימון מודלי embedding למסמך או לטקסט בצורה הבאה. קודם כל אנו מחלקים את D לכמה קלוטרים לפי דומיינים (עם מודול embedding התחלתי). לאחר מכן המחברים ממקסמים את סכומי הlösים הניגודיים על פני כל הקלוטרים האלה. כאמור אנו רוצים לבנות אמבדינגן של שאלה ושל המסמן כך ש:

"אםבדינגן של השאלה ושל המסמך הרלוונטי לה יהיו קרובים אחד לשני בתוך כל קלוטר (המודמה דומיין) ואילו ייצוג של השאלה יהיה רחוק מהכל המסמכים האחרים בклוטר".

clareנו אנו מתאים את האמבדינגים כפונקציה של דומיין השאלה. המאמר גם מציע שיטה לבניה של באצ'ים (כמה מאומנים רשותות היום) כך שהרשות תלמד על שילובי המסמכים הקשים ביותר (למשל מסמכים דומים סמנטיים אבל מדומיינים שונים).

בנוסף המאמר מציע לשלב את ייצוגי המסמכים לבנייה אמבדינג של מסמך נתון 'D'. כלומר יציג של מסמך 'D' מרכיב מרשרור של יציג כל המסמכים מהדאטהסט ואמבדינג של כל הטוקנים מ 'D' (שהם תלויים הקשר המסマー כמובן). בהמשך מאמנים אנקודר למסמך בצורה דומה למה שתואר לפני אבל עם כמה טריקיםלייעול האימון.

אצין שהמאמר לא כתוב בצורה מאוד ברורה....

<https://arxiv.org/abs/2410.02525>

🚀⚡️: המאמר היומי של מיק - 10.10.24 : DIFFERENTIAL TRANSFORMER

המאמר זהה עשה הרבה גלים ביוםיים האחרונים וזה הסיבה שבחורתו אותו לסירה היומית שלו. המאמר החזיר אותו 4-3 שנים לאחר מכן שבאה על בסיס ימי יצאו מאמרים המציעים שכליים שונים ליבת של הטרנספורמרים כה אהובים עליו. כמובן אני מתכוון למנגנון-h-attention שמאפשר לנו לכמת קשרים בין הטוקנים השונים בטקסט.

המחברים הציעו להחליף את חישוב הסופטמקס הרגיל שיש לנו בטרנספורמרים בהפרש משוקלל (רק הסופטמקס השני משוקלל) של הסופטמקסים. כל סופטמקס מחושב עם מטריצת Q ו-K משלה כאשר המשקל הראשון של הסופטמקס השני מחושב באופן הבא: $\text{init} = \lambda \cdot \exp(\lambda \cdot q_1) + \lambda \cdot \exp(\lambda \cdot q_2)$

$R^A = \lambda \cdot k_1, \lambda \cdot k_2, \lambda \cdot q_1, \lambda \cdot q_2$ הינם נלמדים ו- $(\text{init} - 1) \cdot \exp(-0.3) - (\lambda \cdot k_1 - \lambda \cdot k_2)$, כאשר λ זה מספר השכבה (של בלוק הטרנספורמර). אם הנוסחה עברת לאճשהו מובנת ודיאינדרטיב הנוסחה עברת זהותן נותרת בצד תעלומה (אלא אם כן זה ניסוי והיה רגסיה של הערכיהם שהתקבלו עם פונקציה מצורה מסוימת).

המאמר טוען לשיפור תוצאות אבל הבדיקות נעשו בעיקר למודלים עם 3B פרמטרים. יש גם טענות לקנסול של רעש כלשהו שני לא בתוך שני מבין. בקיצור אני קצת סkeptical, מודה....

<https://arxiv.org/abs/2410.05258>

🚀⚡️: המאמר היומי של מיק - 11.10.24 : SELECTIVE ATTENTION IMPROVES TRANSFORMER

היום נסקור מאמר המציג רעיון לשיפור הליבת של הטרנספורמרים, כולם מנגן-h-attention. להבדיל מהמאמר של סלקטיבי(Selective Transformer) הרעיון כאן די ברור לי מתמטית ולא זהה בו נוסחים מתמטיות "مفתייעות". המאמר של היום מציע שיטה לשיפור ביצועים של הטרנספורמרים ועל הדרך מצליח להקטין את גודל הזיכרון הנדרש עבורו.

המחברים טוענים (ובצדק) שלפעמים יש טוקנים שלא כדאי לטרוח לחשב מוקדי attention עבור זוגות מסוימים של הטוקנים. בנוסף ניתן לדעת את זה על ידי הסתכלות על טוקנים ביניהם ואלו בהם (הקשר).

המחברים נotentים את הדוגמא הבאה הממחישה את התופעה הזו. נניח שהטוקנים א, ב, ג הוזנו לטרנספורмер. בשכבה כלשהו עם מושג טקוני, טוקן ב מחייב "כמה הוא מעוניין לקחת" מטוקן א (מקדם attention), וטוקן ג יכול להחליט כמה לקרוא מטוקן א, אבל טוקן ב יוכל להשפיע על כמה טוקן ג "לזקח" מטוקן א. אם טוקן ב קבוע שטוקן א אינו רלוונטי או אפילו מטעה לטוקנים עתידיים כמו ג, אין שום דבר שהוא יכול לעשות בשכבה הנתונה כדי לתקן זאת. השיטה המוצעת על ידי המחברים באה לתקן (להקל) את הבעיה הזו.

הרעין המוצע הוא מאוד אינטואיטיבי ואלגנט. המחברים מציע להחסיר מוקטור ה-attention (לפני חישוב הסופטמוקס) של כל טוקן מטריצת מיסוך נלמדת F. איבר \hat{z} במטריצת F (עבור זוג טוקנים i - j) מבטא עד כמה אנו רוצים להקטין את ה-attention בין טוקנים אלו. ערך גבוה של $\hat{z}_{i,j}$ מסמן לנו שהמודל "אמאיין" שצריך "להתעלם מהקשר בין טוקן i ל- j כולם (אם $\hat{z} > 0$; מטריצה F הינה מטריצה קוזילית כלומר $\hat{z}_{i,j} = 0$ אם $\hat{z} < 0$) אפשר לא לדלג על חישוב מוקדם ה-attention ביניהם.

אבל מה זה מטריצת F ואיך היא נבנית? גם במצבו מאוד אינטואיטיבית F הינה שסום של מטריצות מיסוך רכה S עבור כל הטוקנים בין i ל- j . כלומר טוקן i אינו משפיע על מקדמי מיסוך עבור ה-attention לטוקנים שקדמים לו- j . המחברים לא מסבירים למה הם בחרו לעשות את זה ככה (למי שהו יש רעיון?). כמובן מטריצה S הינה א' שלילית (עושים $S = \text{ReLU}(F)$).

השיטה המוצעת יכולה כאמור לעזור בהאצת האינפראנס על ידי הורדה של טוקנים עם מקדמי F הגדולים ביותר מחישוב ה-attention (לטוקן i נתון). למעשה זה סוג של pruning שהוא תחום מחקר ד' פעיל בראשות הנזירים. המחברים מציעים להגיד "תקציב חסוך" attention לכל שכבה (בלוק של טרנספורמר) ובאופן הדרגתי להעיף מסטריך קבוע של טוקנים מחישוב ה-attention (נעשה באיטרציות). כל פעם מורידים טוקנים עם ערך F הגבוהים ביותר ובוחרים שכבה שעוברה הורדה צזו משפיעה באופן המועט ביותר על perplexity (כלומר $\log(\text{perplexity})$).

בסוף כבר במהלך האימון של מטריצות S אנו יכולים לגרום למודל "לבטל" יותר נזירים על ידי הוספה של איבר לפונקציית הלוס הרגילה של(log-likelihood), הקונס את המודל על S בעלת ערכים נמוכים מד'.

יש לי תהושה שהמאמר הזה הוא התחלת של משהו מעניין...

<https://arxiv.org/pdf/2410.02703.pdf>

⚡🚀 12.10.24: **המאמר היומי של מיק** ⚡🚀

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

האם מודלי שפה גדולים מסוגלים לעשות ריזוניינגן? השאלה זו מעסיקה חוקרים רבים לאור יכולות ד' מרישומות שמודלי שפה מגינים בפרטן שאלות לא פשוטות (אבל רק בתנאים מסוימים 😊). המאמר בוחן את יכולות החשיבה המתמטית (שזה תת-יכולת של ריזוניינגן כללי) של LLMs ומציג את GSM-Symbolic, בנצ'マーク חדש לבחינות יכולות אלו שהם פיתחו.

החוקרים מצאו שביצועי LLMs (נבחן מגוון רחב של מודלים: Gemma, Phi, Mistral, Llama3, GPT-4 ו-1o) משתנים באופן משמעותי כאשר מושנים מעט את השאלות המתמטיות, מה שמעלה ספקות לגבי אמינות המודדים הקיימים. הביצועים של רוב המודלים יורדים כאשר עוברים מ-GSM8K המקורי ל-GSM-Symbolic, מה שמרמז על אפשרות של זיהום DATA (contamination) במהלך האימון (כתבתי על זה לא מעט).

בסוף המודלים מראים רגשות גבואה יותר לשינויים במספרים מאשר לשינויים בשמות עצם, מה שمعد על חוסר יציבות ביכולות החשיבה שלהם. ככל שמספר המשפטים בשאלת עולה, הביצועים יורדים והשונות בביטויים עולות מה שמצביע על קושי בטיפול בשאלות מורכבות יותר.

החוקרים יצרו בנצ'マーק GSM-NoOp, שבו נוספו משפטים לא רלוונטיים לשאלות, וגילו ירידה דרמטית בביטויים של כל המודלים. אפילו כאשר ניתנו למודלים דוגמאות של אותה שאלה או שאלות דומות, הם התקשו להתגבר על האתגרים של GSM-NoOp.

המחקר מצא שאימון נוסף על שימושות קלות יותר וגם הגדלת כמות DATA לא שיפורו את הביצועים במשימות מורכבות יותר.

קצת מנחם שלפחות מודלים חדשים יותר, כמו `W1-01-preview` ו-`iMin-01`, הראו ביצועים חזקים יותר, אך עדין סבלו מהמוגבלות שזויה במחקר

המצאים מעלים ספקות לגבי יכולת האמיתית של LLMs לבצע חסיבה מתמטית פורמלית. נראה כי המודלים מסתמכים יותר על התאמת תבניות מאשר על חסיבה לוגית אמיתית. המחקר מדגיש את הצורך בשיטות הערכה אמינות יותר ובמחקר נוסף על יכולות החסיבה של מודלי שפה גדולים.

<https://arxiv.org/abs/2410.05229>

🚀⚡️: **המאמר היום של מיק - 14.10.24:**

LLMs KNOW MORE THAN THEY SHOW: ON THE IN-TRINSIC REPRESENTATION OF LLM HALLUCINATIONS

מאמר כחול-לבן זה מציג חקירה מקיפה של דפוסי השגיאות של LLMs והקשר שלהם עם היצוגים הפנימיים של המודל. המחברים מבצעים סדרת ניסויים כדי לנתח כיצד LLMs מקודדים מידע על התשובה הנכונה וחוקרים את טבע השגיאות שהם מייצרים.

המחברים חקרו את הנושאים הבאים:

шиיפור זיהוי שגיאות:

המחברים גילו כי ניתן להגיד האם המודל יתן תשובה נכונה או לא מהסתכלות בטוקנים ספציפיים המכילים "תשובה מדויקת" בטעות פלט המודל. למשל עבור השאלה "מה עיר הבירה של צרפת" האינדיקציה האם המודל נותן התשובה הנכונה ניתן לגזר מיצוגי הטוקנים המופק על ידי שכבות מסוימות של המודל. על ידי הtmpmkות בטוקנים אלה, המחברים הצליחו לשפר משמעותית את דיקז זיהוי השגיאות במגוון שימושות ומודלים.

הכללה בין שימושות:

המחקר בוחן האם יכולות זיהוי השגיאות ניתן להכללה בין שימושות וסוגי DATA שונים. התוצאות מראות הכללה מוגבלת, עם הצלחה מסוימת רק בין משימות הדורשות מיזמנויות דומות (למשל, אחוור עובדתי או היסק שכל ישר). זה מרמז על כך של-LLMs יש מספר מנוגנוני אמונות "ספציפיים למיזמנות" ולא מנוגנון אוניברסלי אחד.

טקסונומיה של שגיאות:

המחברים מציעים טקסונומיה של שגיאות LLM המבוססת על התפלגות התשובות במספר דגימות. הם מזהים מספר סוגים של שגיאות, כולל תשבות נכונות/שגיאות באופן עקבי, תשבות נכונות/לא נכונות לסירוגין ומרקם עם תשבות מגוונות רבות. המחברים מדגימים שניתן לחזות סוג שגיאות אלה מהיצוגים הפנימיים של המודל.

פער בין ייצוג פנימי להתנהגות חיונית:

המחברים מראים פער זה באמצעות מערך ניסוי בו הם מייצרים מספר תשבות לכל שאלה ומשתמשים במודל מאומן (בamp;lt;probing>) לבחירת התשובה הטובה ביותר ביותר על סמך ייצוגים פנימיים. הם הבחינו בשיפורים משמעותיים בבדיקה עבור סוג שגיאות מסוימים, במיוחד בהם המודל אינו מראה העדפה ברורה לתשובה

הנכונה בפלטים הרגילים שלו. לדוגמה, בקטגורית שגיאות "שגו" באופן עקבי אך מייצר את התשובה הנכונה לפחות פעמי אחת", שיטת הבחירה מבוססת מודל הסיווג השיגה שיפורים של עד 40% בדיק בהשוואה למצב הרגיל.

ממצא זה מرمץ על כך שה-**LLMs** לעיתים קרובות "יודעים" את התשובה הנכונה ברמה מסוימת, אך ידע זה לא תמיד משתקף בתהיליך ייצור הפלט שלהם. פער זה מעלה שאלות חשובות לגבי טבע ייצוג המידע ב-**LLMs** והמנגנונים השלטיים בתהיליך ייצור הפלט שלהם. המחברים מציעים כי ממצא זה עשוי לשמש לפיתוח אסטרטגיות חדשות לשיפור דיק ה-**LLM**, אולי על ידי שינוי תהליך ייצור הפלט כך שלוקח בחשבון גם את הייצוגים הפנימיים.

<https://arxiv.org/abs/2410.02707>

⚡️📝: 15.10.24 | המאמר הימי של מיק - EFFICIENT DICTIONARY LEARNING WITH SWITCH SPARSE AUTOENCODERS

היום סוקרים מאמר קליל המשלב שני רעיונות די נחמדים שימושים **MoE** (במיוחד לאחרונה) והאמת השימוש שלהם נראה די טבעי. הרעיון הראשון הינו **MoE** בקצרה.

MoE היא שיטה המאפשרת לנו להקל על האינפראנס על ידי שימוש רק בחלק ממושך המודל. בד"כ מטריצות משקלים בראשת **feed-forward** (יש שם 2 שכבות בסך הכל) בבלוק הטרנספורmers (אחרי **attention**) מוחזקים לכמה קבוצות שלל אחת מהן נקראת מומחה או **expert**. באינפראנס המודל משתמש רק בחלק (לפעמים רק אחד) מהמומחים ובכך הוא מוריד את מחירו של האינפראנס. כאמור אותו המודל מופעל בצורה קצת שונה בהתאם לקלט (בנוסף ל-**attention**),

הקונספט השני הוא **Sparse AutoEncoders** או **SAE** בקצרה שהפרק להיות די פופולרי אחרי החוקרים של אנטרופי הצביעו להשתמש בו למטרת חקר interpretability של מודלי שפה. לפני הבלוג זהה הסברה הרווחת (סוג של) הייתה שבמודל שפה יש נוירונים שנדרכים חזק (מקבלים ערך גבוה) על קונספטים מסוימים כאשר כל נוירון חזק הינו מונו-סמנטי כלומר יש קונספט אחד בלבד שהוא "אחראי" עליו.

לעומת זאת החוקרים של אנטרופי הצביעו להתבונן בכל נוירון כפול-סמנטי כלומר "אחראי" על מספר קונספטים לא קשורים. לפי מושנתם ניתן לגלוות את הקונספטים האלו באמצעות **SAE** שבונה שכבת **autoencoder** דليل (הרוב אפסים) בימיד גבוהה הרבה יותר מגודל השכבה שבה נמצאים הנוירונים הפוליסמנטיים אלו. **SAE** אכן מרכיב שתי שכבות בלבד, אחת לאנקודר ואחת לדקודר.

כאן כל רכייב שהוא לא אפס בזוקטור אחרி שכבת **encoder** של **SAE** הוא אחראי על קונספט מסוים ככלומר מהווענו נוירון מונו-סמנטי. כך יוצא שכל נוירון בשכבה המקורית הוא שילוב לניארים של הנוירונים המונו-סמנטיים אלו. **SAE** מאמון בצורה די סטנדרטית עם איבר רגולרייזציה שאוכף את דילולות הייצוג אחרי האנקודר.

از המאמר מציע לשלב את שני הקונספטים האלו כך שכל נוירון הוא צירוף לניארי אחר של הנוירונים המונו-סמנטיים בשכבת **encoder**-**decoder**. זה מאפשר גמישות נוספת ביחס לרעיון המקור ובטח מאפשר לגלוות קונספטים שונים המוסתרים בתוך ה-**LLMs** שלנו.

מאמר קליל - ממליץ להעיף מבט

<https://arxiv.org/abs/2410.08201>

⚡️🚀 16.10.24: המאמר הימי של מיק

EFFICIENT REINFORCEMENT LEARNING WITH LARGE LANGUAGE MODEL PRIORS

היום נסקור מאמר שהוא די כבד מתמטית (הרבה נוסחאות ומלל שנראה מתמטי) אבל הרעיון מאחוריו הוא די פשוט וקל להסביר. אנחנו אוהבים למן את עצמתם של מודלי שפה למשימות רבות (ולא תמיד לכאלו שהם מסוגלים לבצע כמו שצריך לפחות כרגע).

המאמר מציע להשתמש במודל שפה כפирוי עבור סוכנים במשימות בהם הם צריכים לבצע SDM או sequential decision making. המאמר נותן בתור דוגמא משחק overcooked כאשר הסוכן צריך לבצע משימות בישול שונות בהתבסס על מצב המטבח שבו הוא מבשל אותם. המטרה של הסוכן היא לחזות את הפעולה הבא (באמצעות תיאור טקסטואלי) כאשר התגמול הוא ביצוע נכון של המשימה (הכנה שלמנה לפי המתכוון :)).

כאמור המטרה כאן היא לחזות את הפעולה הבאה עבור הסוכן (המתוארת) על ידי הטקסט כאשר המצב (state) גם מתואר על ידי טקסט. بغدادו מאוד אנו מתחילהים ממודל אחד (הפירוי P) עבור חיזוי המצב הבא (מה המצב הקודם והפעולה) ועבור חיזוי הפעולה הבאה בהינתן המצב (מתואר על ידי התפלגות $h(Q)$). המטרה כאן היא למדוד את $h(Q)$ כאשר מוקסמת התגמול הצפוי ושומרת את התפלגות Q קרוביה לפירוי P (זוכרים PPO שהתפרסם מאד לפני שנתיים כאשר AIOpenAI השתמשו בו ל-RLHF לאימון מודלי שפה). המרחק המקורי נתן על ידי KL 😊

از הפעולה הבאה t_a (כלומר גנרטט התיאור הטקסטואלי שלה) מtbody באפין הבא. דוגמאות כמה גרסאות של t_a עם P מחשבים את הנראות שלהם לפי Q הנלמד, מנורמלים עם הסופטמקס ודוגמאות את הפעולה הבאה כאשר מטרת התהילה מקסום של התגמול הצפוי (עם הרגולרייזציה שהסבירנו עלייה קודם).

כמובן שניתן לעשות את זה בכמה אופנים: בצורה של online דרך שעורך של פונקציית Q של הזוג (מצב, פעולה) כאשר פונקציית Q קשורה להתפלגות $h(Q)$ של הפעולה הבאה שנידונה בפסקה הקודמת (ענין של נרמול נכון). ניתן לעשות את זה גם באמצעות offline עםizia פוליסיטי טוב ידוע של המומחים כאשר המטרה היא גם שעורך של פונקציית Q שבאמצעותה ניתן לשערך (לקבל) את $h(Q)$ עבור חיזוי הפעולה הבא. ניתן לעשות את זה גם באמצעות שיטה דומה ל-PPO אבל בכל המקרים הפירוי הוא ההתפלגות המושנית על ידי מודל שפה נתון.

מאמר מעניין בקיצור ...

<https://arxiv.org/pdf/2410.07927.pdf>

⚡️🚀 17.10.24: המאמר הימי של מיק

EQUIVARIANT CONTRASTIVE LEARNING

היום נסקור מאמר שפורסם לפני שנתיים וחצי בנושא למידה ניגודית (contrastive learning). הנושא עצמו תמייד עניין אותו וスクרתי לא מעט מאמרים אבל חייב להגיד שזמן האחרון שטף המאמרים על CL די נחלש. כאמור המאמר הזה שראה אור לפני שנתיים מציע שכלול לשיטה הקלאסית לבנייה של ייצוג נתונים (אMOVED) באמצעות CL.

בגדול CL היא שיטה לבניית ייצוג של נתונים כאשר העיקנון המוביל הוא לקרב ייצוגי פיסות נתונים (זוגות חיוביים) ולהרחק ייצוגים של פיסות נתונים לא דומות (שליליים). זוגות דוגמאות חיוביים (במקרה של דטה לא מתייג) נבחרות כאוגמנטציות שונות של דוגמא (עבור תכונות זה יכול להיות הצעה, סיבוב וכדומה) ואילו זוגות השליליים נבחרים באקרים מהدادהסת.

אולם יש לא מעט בעיות עם הגישה זו הקשורת לבחירת זוגות של דוגמאות חיוביות - למשל שני פאצ'ים באוטה התמונה עלולים להכיל תוכן סמנטי שונה שלא נרצה לקרב את יציגיהם (הוצעו מספר פתרונות לסוגיה זו בעבר וחלקן סקרתי). בנוסף אולי הינו רצים לקבל יציגים שונים (ולא מודר קרובים) של טרנספורמציות מסוימות של אותה התמונה (גדי סיבוב או גזזה) למשימת *downstream* ספציפית.

כלומר הינו רצים להשרות יחס נתון Z_T בין יציג התמונה ההתחלתית I ולציג התמונה אחרי טרנספורמציה T (נקרא לה T_I). ככלומר אנו רצים לבנות יציג Z כך:

$$\text{p}(\text{Z}(\text{I})) = \text{I}_{\text{T}}(\text{p}(\text{I}))$$

זה בדיק מה שנקרא *equivariance*. למעשה CL הסטנדרטי הוא מקרה פרטי של *equivariance* שעבורו Z_T הינה טרנספורמציה זהה וזה נקרא אינוריאנטיות של הייצוג תחת טרנספורמציה T .

זה בדיק מה שהמאמר עושה. למעשה המחברים מציעים לאמן יציג שומר על אינוריאנטיות עבור טרנספורמציות מסוימות (כמו בCL הסטנדרטי) ו敖וף בכך *equivariance* מוגדר לטרנספורמציות מקובча נתונה G המתאימה למשימת *downstream* שיש לנו ביד. ככלומר לכל טרנספורמציה G - G אנו מגדירים מראש את הטרנספורמציה *equiinvariant* שלה (שיכולה להיות חברה ב- G - G) ומאמנים את הייצוג Z שהיחס *equiinvariance* ביןיהם יתקיים. מבחינה פרקטית הלוע הוא סכום משוקל של הלוסים של CL הסטנדרטי וה-ECL.

מאמר חמוץ - מחר או היום בערב אסקור את מאמר המשך שלו....

<https://arxiv.org/abs/2111.00899>

🚀:18.10.24-🚀 SimCSE: Simple Contrastive Learning of Sentence Embeddings

סקירה קצרה מאוד על איך ניתן לעשות למידה ניגודית (*contrastive learning*) כדי לבנות יציג חזק של הטקסט. הרעיון כבר הסבրנו בסקירה הקודמת שהמטרה של CL היא לאמן יציג של דатаה כך שיציגים קרובים סמנטיות יהיו קרובים במרחב הייצוג ואילו יציגים של דוגמאות לא דומות יהיו רחוקות שם. מאמנים יציג כזה בדרך כלל דרך מזעור היחס שבין יציג פיסות דатаה דומות (זוג חיובי) לבין אלו של הלא דומות (שליליים).

השאלה איך לבנות את היציגים האלו (במיוחד הזוגות החיוביים)? זה בעצם נושא מחקר פעיל מלפני שנים-ים-שלוש. המאמר המסורק מציע לבנות זוגות חיובים דרך *dropouts* שונים של רשות הנירונים (שאותה מאמנים לבנות את הייצוג). ככלומר עברו אותו הטקסט זוג דוגמאות חיובי נבנה עם עם הפעלת הרשות עלי עם שני *dropouts* שונים. נזכיר *dropouts* מבטל באקראי קשרים בין נירונים ברשת ומהוות כל ידוע לשיפור יכולת הכללה של הרשת. הזוגות השליליים נבנים עם דוגמאות שנבחרו בצורה אקראיית.

לדאטהסטים המכיל משפטים מתוארים כמו למשל LNI (למשפט נתון הדאטהסט מכיל משפט אחד עם אותה המשמעות(*entailment*), משפט אחד בעל משמעות דומה ומשפט אחד בעל משמעות הפוכה או סתירה - *contrary*). באופן לא מפתיע המאמר מציע לבחור בתור זוג שלילי את שני המשפטים בעלי משמעות הפוכה ובתור זוג חיובי שניים עם אותה משמעות.

בנוסף המשפט הזכיר לי לייצוג דатаה טוב יש 2 תכונות מהותיות: קרבה בין יציגי הדאטה הדומה והתפלגות יוניפורמית של כל היציגים של הדאטה - זה חשוב.

<https://arxiv.org/pdf/2104.08821>



המאמר היום של מיק - 19.10.24: DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings

סקירה קצרה ואחרונה(כנראה) ב邏ינִי-סדרה על איך לבנות ייצוג דатаה באמצעות שיטות למידה ניגודית. כבר הסבירתי על הלמידה הניגודית בשתי בסקירות הקודמות. בקצרה, מאמנים מודל הבונה אמצעים לדאטה המקרוב ייצוגים של פיסות דатаה דומות ולהרחק פיסות דатаה לא דומות. כאמור הוציאו עשרה שיטות לעשות זאת לדאטה מדויינים שונים.

המאמר מציע שיטת CL העשויה זאת בצורה מתחכמת יותר (לטעמי). הרעיון הוא שמודל המטרות של בניית ייצוג הדאטה היא שהוא ישקף את התכונות האינהרטנטיות של הדאטה והמחברים הוציאו דרך "לא נכון" את זה על הייצוג. הם מאמנים מודל לבניית ייצוג טקסט כך שהמודל יבדיל בין מה אמר ומה לא אמר להיות בתוך הטקסט.

איך הם עושים זאת? הם מיסיכו כמה טוקנים בטקסט, ביקשו מודל אחר לחזות את הטוקן הזה ואז אימנו את ייצוג קר שבuzzratio יהיה ניתן להבדיל בין טוקנים שנחזו ואלו שלא. כולם בנוסף למודל החיזוי (לא אומן) ומודל לבניית אמבערתו הם אימנו עוד מודל לסיווג ביןאי רשותהו להגיד האם טוקן נחזה או לא. וייצוג הטקסט מוזן למודל הסיווג הזה.

דרך אגב פונקציית הלוס למודל הסיווג דומה לזה של GAN אבל אין באמת קשר בין שני הדברים (זה טיפה בלבד באותו בהתחלה)....

<https://arxiv.org/pdf/2204.10298>



המאמר היום של מיק - 20.10.24: RL, BUT DON'T DO ANYTHING I WOULDN'T DO

אוקי, אחרי כמה סקירות יחסית קליות הגיע הזמן לסקור מאמר קצר כבד לפחות מהמבט הראשון. המאמר בנושא של אימון מודלי שפה עם השיטות מעולם למידה באמצעות חיזוקים או בקצרה RL. דרך אגב הטענות המתמטיות הרבות שהמחברים הוכחו (לא אמפירית אלא הוכחות מתמטיות רציניות) לא מוגבלות רק לאימון RLHF עם LLMs.

בד"כ כאשר אנו מאמנים LLM על RLHF פונקציית הלוס שאנו מעוניינים לאפטם מרכיבת מסכם של שני איברים (לפעמים מוסיפים עוד אבל אני מדבר כאן על כאלו המופיעים ברוב המאים על יישור alignment) של LLMs עם שיטות RL.

האיבר הראשון אחראי על מיקסום של פונקציית reward שזה מודל שמאומן לפני על דאטהסת המכיל זוגות של תשובות מועדפות יותר ומועדפות פחות(המתייחס על ידי בני אדם) לסת של שאלות. מודל reward מואמן לתת ערך גבוה לתשובה טוביה וערך נמוך לתשובה לא טוביה לשאללה. אז האיבר הראשון מאפסט את משקל ה-LLM המאמן כך שימקסמו את פונקציית ה-reward ובכך יגרמו ל-LLM להיות יותר מישר (aligned) עם הציפיות שלנו (לפחות הינו רוצים להאמין בכך).

האיבר השני הינו איבר רגולרייזציה השומר את המשקלים של המודל המאמן קרובים יחסית (במנוחה מרחק KL בין התפלגיות הטוקנים) לממשקלי המודל ההתחלתי (שלו אנו עושים פין טיוון). איבר זה נדרש כי בלאדי המודל יעשה את מה שנקרא "reward hacking" ובמקרה להתיישר עם ציפיותנו ימקסם את reward אבל כתוצאה קיבל

עוד יותר גורע ממה שהוא (או לפחות טוב ממה שנייתן לקבל עם איבר רגולרייזציה זה).

אולם מחברים המאמר טוענים שאיבר זה לא מספיק ולא תמיד ימנع ממושקל המודל להתכנס למצבים לא רצויים. הסיבה לכך היא שמודל בסיס שהוא רצוי לשומר את המודל המקורי קרוב אליו מהוות בעצמו קירוב של מודל "בטוח ומושך עם ציפיותנו" (בדרכ כל אומן על התשובות הרצויות). ותברר שגם אם מודל הבסיס שלנו קרוב מספיק ל"מודל הבטוח" והמודל שהוא מאמנים קרוב למודל הבסיס במונחי KL, עדין לא ניתן להבטיח שהמודל המאמן יהיה קרוב מספיק ל"מודל הבטוח" (גם במונחי KL) - ככלمر או שווון המשולש לא מתקיים כאן. גם נטען את מודל בסיס על יותר דатаה ויוטר משאביהם, עדין נתקשה להבטיח את קירבתו של המודל המקורי ל"מודל הבטוח".

הסיבה לכך היא קצת (מבחינה קונצפטואלית) דומה לכך למה דגימת Langevin בצורתה הקלאסית (לא רעש) לא עובדות לדאטה בעלת מידת גבוהה מאוד כמו (תמונות). בגלל הימיד המאוד גבוהה של המרחב הסמנטי של הדאטה מודל הבסיס יגיע למקרים ש"המודל הבטוח" לא יהיה מגיע בכלל ואז הוא יתקשה לתת שערוך אמין להסתברויות הטוקניות. וזה יגרום למודל המקורי להיות לא אמין באותה מידת.

המחברים קוראים למאורעות אלו (הגעה למצב שהמודל הבטוח לא יהיה מגיע אליו) מצב תקדים (unprecedented events) וטוען שכאשר הם קוראים מודל הבסיס "יטה לתת תשובה "פешטה מד'" ולרוב לא נcona וכך יעשה המודל המקורי. הפשטות זו נובעת כנראה (איליה סלוצקי מדבר על זה הרבה) בגלל ה bias algorithmic-information-theoretic inductive bias נתיתם להתכנס לפתרונות פשוטים בעליות סיבוכיות תכניתית נמוכה (שהה בעצם סיבוכיות kolmogorov) ככלמר הן פועלות לפי עקרון התער של אוקם (זו ההנחה כמובן). זה גורם למודלים להפגין התנהגות פשוטה (מד') ולא טובים במקורה שהם נתקלים במאורעות חסרי תקדים הללו. והמודל המקורי על RLHF "ירוש מהם" את הפגם זה.

מאמר מאד עמוק, דרוש זמן בשבייל להפניהם אבל שווה קריאה בהחלט ...

<https://arxiv.org/abs/2410.06213>

🚀⚡️ המאמר היומי של מייק - 22.10.24: Sample what you can't compress

לא היה לי הרבה זמן להזכיר זאת בחרתני במאמר זהה שנייתן לסקור אותו די בלקוניות בלי לפגוע בחווית הקוראים. המאמר מציע שיטה נחמדה לבניית ייצוג DATA ויזואלי (קרי תמונות) באמצעות שכלול של אוטו-אנקודר. מכיוון שהייצוג הזה בד"כ במיד נמור יותר מהדאטה עצמה אז ניתן להתייחס אליו בתור דחיסה של דאטה. ד"א ניתן לאמן ייצוגים שלאו דווקא "מעבירים" את הדאטה למרחב בעל מים נמור יותר ב-EAE ולבסוףsparse ב-EAE.

אוטו-אנקודר זו דרך לבנות ייצוג מקומפרס של דאטה עם השילוב של האנקודר והדקודר כאשר האנקודר ממפה את הדאטה למרחב הייצוג והדקודר משחזר את הדאטה המקורי מייצגו הדחוס. מאומנים AE דרך מזעור של לווי השחזור (עד כמה טוב הצלחנו לשחזר את הדאטה מייצגו הלטנטי) ולפעמים מוסףים רגולרייזציה במטרה לגרום לייצוג להיות בעל תכונות מסוימות (כגון דילוי).

כמובן שלא תמיד מצלחים להציג חזק (שמומר את כל התכונות האינהרנטיות של פיסת דאטה) עם AE והמחברים מציעים לשכלל אותו על ידי הוספתו של מודל הדיפוזיה לסייע. כזכור (או שלא ואז אני אזכיר) מודל דיפוזיה מאומן להסיר רעש מפיסת דאטה ואם מאמנים אותו טוב אז מקבלים מודל שיודיע לגנט דאטה מרעיש טהור (על ידי הסרת רעש הדרגתית).

המחברים מציעים לחקות את מודל הדיפוזיה (המחברים משתמשים במודל דיפוזיה המקורי שבודנה את התמונה עצמה בתהיליך דיפוזיה ולא ייצוגה הלטנטי). המודל זהה מורכב מסדרת של G-Nets-U (ולא טרנספורמרים כמו שאנו רואים היום במודלים דיפוזיה) שקדום מקטינים את מידת התמונה (כלומר ניתן לראות את זה:auto-אנקודר) ולאחר מכן בונים מהיצוג הזה את התמונה.

המחברים מזינים את התמונה המשוחזרת אחרי הדקودר של AE יחד עם התמונה המורעשת(המקורית) למודל דיפוזיה שמאומן כאמור להסיר רעש מהדата (יחד עם AE). הלוואו מורכב מסכם משוקל של הלוואו הרגיל של מודל הדיפוזיה, הלוואו הרגיל והלוואו ה-perceptual לשנייהם מופעלים לתמונה המשוחזרת אחרי השלב הראשון של AE (לפני מודל הדיפוזיה). הלוואו ה-perceptual בודק עד כמה התמונה המשוחזרת נראה "טבעית למבט האנושי" (משווים את האקטיבציות שלה ברשות מאוננת עם אלו של התמונות הטבעיות).

היצוג הסופי של פיסת>Dataoa מתקיים אחרי ה"אנקודר" של מודל דיפוזיה (ה-bottleneck). וכמוון יש טענות לדחישה טובה יותר מסוימות SOTA עם הגישה המוצעת...>.

<https://arxiv.org/abs/2409.02529>

🚀⚡️: 23.10.24 - 🚀⚡️ המאמר היומי של מיק - Predicting from Strings: Language Model Embeddings for Bayesian Optimization

המאמר מהסוג שנתקור היום אני לא סוקר בדר"כ - אולי מtower 300 מאמרם שסקרתי יש 2-1 כאלו (לא בטוח). לא בಗלל שהנושא לא מעניין אלא שיש פחות מאמרם בו והוא נחשב פחות "באזיז" למרות חשיבות. כמו שמשתמע ממש המאמר הנושא הוא אופטימיזציה בייסיאנית.

בגדול אופטימיזציה בייסיאנית היא אחד הכללים לפתרון בעיות תכנון ניסויים ולמה שנקרא black-box optimization כאשר היא למעט את מחיר של תהליך החיפוש הפתרון הממוצע פונקציית המטרה. פונקציית המטרה יכולה להיות יעילות הת蘗פה (כאשר המטרה למצוא את הרכמה האופטימלי) או אופטימיזציה של היפר-פרמטרים של רשת גדולה. בשני המקדים כל אבולוציה של פונקציית המטרה הינה יקרה מאוד ויש צורך לפחות כמות הפעמים שמחשבים אותה (לבדיקה הרכבת של ת蘗פה או אבולוציה של ביצועים עבור שילוב היפר-פרמטרים מסוימים של הרשת).

קיימות לא מעט שיטות לאפטם את בחירת הנקודות x לאבולוציה של פונקציית המטרה שמצד אחד בוחרת איזורים בהם לא בדקנו (exploration) ומצד שני גם מנצלת את הידע שלנו על ערכי פונקציית המטרה באיזורים שכבר ביקרנו (exploitation) במטרה למצוא נקודת מקסימום טובה באמצעות מינימל. רוב השיטות מבוסות לבנות מה שנקרא surrogate objectivesurrogate objective דומה זוולה להפעלה כדי למצוא את x הבא בהינתן תוכאות הפעלה הקודמות (כלומר זוגות x ו- $(x,f=y)$). הדרך הפופולרית ביותר היא להשתמש בתהיליכי גאוס כדי למדל את פונקציית מטרה דמה ובעזרתה בוחרים את x האופטימלי.

המאמר מציע לרתום את ה-LLMs לשיפור זהה במטרה לשערק את התוחלת ואת השונות של (x,f) עבור x נתון. בשלב הראשון הופכים את הזוגות של x ו- y הידועים לפורמט של string (נגיד לחסז' המכיל את שמות הפייצ'רים והערכים שלהם). לאחר מכן מזינים אותם לאנקודר כדי יחד עם הערך של x המפיק את "יצוגי הזוגות הללו". בשלב האחרון מכנים את "יצוגים אלו לדקודר כדי יחד עם הערך של x שעבורו אנו רוצים לחשב את (x,f) (תוחלת ושונות). מאמנים את הדקודר (האנקודר לא מאמין) על סדרות "זהב" של זוגות x ו- (x,f) במספר מסוימת. במהלך האימון בהינתן k הזוגות הראשונים מנסים לחזות את ערך הפונקציה עבור $1+k$ x ל- k -ים שונים.

معنىון שהמאמר מניח כי את אינפראנס ערכי x -ים לבדיקה מתקיים דרך איזה אלגוריתם אבולוציוני נתון.

<https://arxiv.org/pdf/2410.10190>

⚡️🚀 24.10.24: המאמר היום של מיק - ⚡️🚀

HOW MANY VAN GOGHS DOES IT TAKE TO VAN GOGH? FINDING THE IMITATION THRESHOLD

מאמר מעניין שנטלו בו חלק חוקרים ישראלים מאוניברסיטת בר-אילן. הם חקרו נושא די חשוב הקשור להפרת זכויות יוצרים אפשרית על ידי מודלים גנרטיביים לתמונות. הרעיון הוא שמודלים שאומנו בחלקם על דатаה שהוא פרט, מוגן על ידי זכויות יוצרים ואם המודל יתחל לגנרט למו תמונות דומות מדי להם זה עלול להיות עבירה על החוק. אבל איך להבטיח (או לפחות לתת הערה כלשהי) לכך שזה לא יקרה?

המאמר בחר בגישה די אינטואיטיבית לכך. הרעיון העתקה של קונספט מסוים על ידי המודל קשורים קשר סיבתי (אמנם לא ב 100% מובן כרגע) במספר פיסות דатаה (= תמונות) המוכלות בדעתה שמדובר אכן עליו. אבל איך נדע זאת? הרעיון נctrar לאמן הרבה מודלים כדי לבדוק מתי התמונות המגנרטות על ידי המודל יהיו דומות מדי קונספט T מסוים (עם פרומפט מתאים).

כמובן שזה לא בר עשייה והמאמר מציע שיטה יחסית פשוטה לעשות את זה כאשר הוא מניח הנחה מהותית אחת: מספר התמונות המכיל קונספט T מספיק לכך שהמודל יהיה מסוגל להעתיקו איננו תלוי-B-T. אני מניח שזה נכון בקבילות הסביר זאת אומרת המספר הזה נע באינטרול יחסית צר לכל הסוגנות. יש עוד הנחה שנייה (גם חשובה) שאין איזה confounded בין מספר התמונות לבין המודל להעתקה (גם די סביר).

עם הנחה זו המאמר מציע לאמן מודל על הדעתה שיש בו שונות גדולה בין כמות ההופעות של כל קונספט. לאחר מכן המאמר מגנרט תמונות מכל T שהופיע בטקסט ובודק כמה מהם קרובים סמנטי (משווים אմביגאג) ל T. זה נעשה עם הסף שנקבע דרך השוואה בין דמיון האמבדינג של תמונות שונות של אותו הקונספט מול תמונות של מילוטים הווה (כדי למזער FP יחד FN).

לאחר מכן מגנרטים תמונות עבור כל הקונספטים T שיש בדעתה ומוחשבים כמה מהם (היחס) מכילים את T. זה נקרא imitation score. בסוף אנו מקבלilo עבור כל קונספט T ובגלל שיש לנו שונות גדולה בין הופעה של כל קונספט בדעתה ניתן לזרות איזה יש אליה מובהקת ב- score זהה מבחינת מספר ההופעות של קונספט T בתמונה. זה קצת דומה ל^{deep} elab-^k-means באלגוריתמים מעולים (כמו PELT) שיעדים לעשות זאת. ככה נקבל את הסף של מספר ההופעות של קונספט בדעתה שמננו המודל יידע להעתיקו ופוטנציאלית לגרום לתביעות.

אהבתי - המאמר גם כתוב יפה וברור.

<https://arxiv.org/pdf/2410.15002>

⚡️🚀 25.10.24: המאמר היום של מיק - ⚡️🚀

Amortized Planning with Large-Scale Transformers: A Case Study on Chess

מאמר די מעניין שగם לדינמים רבים בנושא יכולות ריזונייניג של מודלי שפה. אחרי שהענינים קצת נרגעו הגעתם לסקרו בלי להתייחס יותר מדי לסוגיה הזו. המאמר למעשה אימן מודל שפה די צנוע מבחינת פרמטרים (עם הטרנספורמרים בפנים) לשחק שח. אזכיר שהמכונות הגיעו לרמת של בני אדם בשחמט די מזמן (לדעתי לפני 30 שנה כאשר שנו deep השair אבק לאלוף העולם אז גاري קספרוב).

از מה המחברים עשו בעצם? הם הורדו 10 מיליון משחק שחמט מאתר LiChess והשתמשו בכל הנקרא

לשערך הסתเบרות ניצחון עבור מצב לוח נתון s . לאחר מכן הם הפכו את מצב הלוח ותיאור המהלך לטקסט (נראה די טבעי בסך הכל) ואימנו מודל שפה "לשחק שח". המחבריםניסו לעשות זאת בכמה דרכים:

1. אימנו את המודל לחזות את הסיכוי לניצחון בהינתן מצב הלוח s ומהלך a . כדי לעשות זאת הם חילקו סיכוי הניצחון לכמה בינים (זרמים) ואימנו את המודל לחזות את הבין שבו נמצא הסיכוי s -*ground-truth*-*Gauss*. הם עשו את זה לא בצורה הרגילה (עם one-hot encoding עם כל בין) אלא על ידי "ריכוכו" כולם כל בין מתקבלות בין מקלט הסתเบרות משלו כאשר הבין a -GD מקבל את ההזדמנויות הכי גבוה (נעשה לפ' התפלגות גאוס ונקרא HL-Gauss)

2. אימנו את המודל את סיכוי הניצחון עבור מצב לוח נתון s באמצעות הצורה כמו ב 1.

3. אימנו מודל לחזות את המהלך a -GD של המשחק

בסוף המהלך נבחר צזה עם סיכוי לניצחון הגבוה ביותר. יש תוצאות לא רעות.

אם זה מצביע על כך שהמודלים יודעים לעשות ריזונייניג - לא יודע, מבטיח לחשב על זה לעומק

<https://arxiv.org/pdf/2402.04494v2.pdf>

🚀⚡️: 26.10.24. 🚀⚡️ Efficient Vision-Language Pre-training by Cluster Masking

היום סוקרים מאמר נחמד בנושא של למידה ניגודית (contrastive learning) אבל הפעם עבור מקרה מולטימודלי. ככלומר הפעם מאמנים מודל בסגנון של CLIP הידוע המועדף לבניית ייצוג נתונים ויזואלי (תמונה) והשפה. הרעיון העיקרי בלמידה הניגודית הוא לאמן מודל הממפה קרוב (במרחב האמבידיניג) פיסות נתונים דומות ורחוק (באוטו המרחב) פיסות נתונים לא דומות.

אבל הפעם מדובר בדעתה מולטימודלי. ב-CLIP המקורי אימנו את המודל לקרב ייצוג של אוגמנטציות שונות של תמונה עם הכותרת שלה ולהרחיק אותן (האמבידיניג של האוגמנטציות השונות של התמונה) מהייצוגים של כתורות שנבחרות באקראי. דבר דומה נעשה דומה לייצוג כתורת של תמונה: מקרים לאmbidinig של אותה התמונה (עם אוגמנטציות) ולהרחיקו מהייצוגים של השאר.

מצין ש-CLIP המקורי אינו מודלי ייצוג שוניים (למייט זכרוני) לתמונות ולשפה אבל יצא גם שדרוגים שאימנו שני מודלי ייצוג עם הרבה משקלים משותפים (אותה הארכיטקטורה).

הכל טוב ויפה אבל נשאלת השאלה האם ניתן לשפר כאן ממשו? מתברר שכן ומהמאמר מציע שכלל קליל ל-CLIP-CLIP. כתבתי שאחד הדברים החשוב ב-CLIP הינה בחירה של הזוגות של פיסות נתונים לא דומות(זוגות שליליות). ככל שהיא יותר מגוון הזוגות השליליות הייצוג שייבנה יהיה חזק יותר (כפי ראה יותר דברים לא דומים ואז בין יותר טוב איך "צריך להיראות ייצוג טוב").

از המחברים מציעים לקלסטר פאצ'ים בתמונה לקלסטרים וכל פעם לא לבחור את הזוגות החשובים בצורה אקראית אלא לאפשר בחירה של לפחות אחד מתוך כל קלסטר. ככלומר, לכל באצ' בוחרים רק פאצ' אחד מהקלסטר. ככלומר פאצ'ים דומים מדי לא נכנים לזוגות שליליים ב-CL. הקלסטור יכול להתבצע על הערכים של הפיקסלים בשילוב עם מודל אמבידיניג כלשהו.

מאמר פשוט - לך לי איזה דקה להבין | 10 דקות לכתוב סקירה. אהוב כלו....

<https://arxiv.org/pdf/2405.08815.pdf>

⚡️🚀 המאמר הימי של מיק - 28.10.24: HEAVY-TAILED DIFFUSION MODELS

המאמר עם השם הקצר זהה משך את עיני כי יש לי חיבה גם למודלי דיפוזיה גנרטיביים וגם להתפלגיות בעלות תכונות מעניינות למשל זנבות כבדים. בגין התפלגאות נקראת בעלת זנב כבד או ארוך כאשר התפלגאות לזרב שלה (כלומר המשא הסטברותית מימין לנקודה) מקשלה (הסתברות) הינה גבוהה יותר מאשר להתפלגות מעריכית. נשמעו קצת מסובך אבל במילים פשוטות ניתן להגיד כי להתפלגיות בעלות זנב כבד(HT) יש יותר מסה בקצוות.

למשל התפלגאות נורמלית אינה בעלת זנבות כבדים והתפלגאות סטודנט t וגם התפלגאות קושי הן כך. אוקיי, למה אני בכלל מדבר על זה? הסיבה היא די פשוטה - ההנחה שנווכל להניח התפלגאות גאוסיות על כל סוג של נתונים איננה נכונה. יש סוגים נתונים לאפיין אוטם בצורה טוב עם התפלגאות בעלות זנבות קלים. עקב גם אם אנו נתקשו לגנרט נתונים מהתפלגויות אלו אם נMODEL אותם (הדאטה) עם מודלי הבוניים על הנחות גאוסיות גם אם המודלים הללו הם בעלי expressiveness גבוהה כמו מודלי הדיפוזיה. עדין יהיה מאד ענייטי ליצור באמצעות נתונים דאטה בעלות התפלגאות HT במיוחד בקצוות התפלגאות.

از המאמר, שהוא אחד הבודדים בו יותר מתמטית מלאו שראיתי לאחרונה, מציע להחליף את התפלגאות גאוסיות שיש לנו במודלי דיפוזיה בתפלגאות סטודנט שהיא התפלגאות HT. ככלומר כל מה שהיא בעלת התפלגאות גאוסיות במודל דיפוזיה מקורי יהיה מהתפלגאות t . דרך אגב אחד הפרמטרים של התפלגאות t (שהיא כמובן וקטוריית עבור מודלים אלו כי אנו רוצים לגנרט נתונים בעלת ממדים רבים) שהוא שולט ב"כבודות הזנב" שלה וכאשר היא שואפת לאינסוף אנו מקבלים את התפלגאות הגאוסית האהובה علينا. ככלומר המודלים המוצעים במאמר הם הכללה של מודלי דיפוזיה גאוסיים שאנו מכירים ואוהבים.

כמובן שלא מספיק סתם להחליף התפלגאות גאוסיות במודל דיפוזיה בתפלגאות t - זה דורש להגדיר לא מעט התפלגיות מותניות הנדרשות לנו להגדרת הלמידה של תהליכי denoising. זה די לא טריוויאלי אבל העקרון נשאר דומה - מאמנים את המודל להסיר רעש (שהוא מפוגע עם t) באופן הדרגתי. במקום KL divergence המוכר לנו מודלי דיפוזיה המחברים משתמשים ב- Power divergence - γ כדי למדוד מרחק בין התפלגאות הדאטה אחריו הסרת רעש זהה של הדאטה האמיתית (לכל איטרציה).

גם תהליכי הגנרט מוגדר דומה עקרונית למודלי דיפוזיה גאוסיים אבל כמובן כל ה-*hyperparameters* מותאמים לתפלגאות t . יש גם פרמטריזציות שונות כה אהבים במודלי דיפוזיה, יציג באמצעות משוואות דיפרנציאליות חלקיות, גם באמצעות טכניקה חדשה הנkirאת matching flow (הבנייה מסלול מיטבי בין התפלגאות התחלהית והתפלגות הדאטה). כאמור מאמר די כבד מתמטי וękkoוה שהצלחתו להסביר לכם את העקרונות לפחות.

<https://arxiv.org/pdf/2410.14171.pdf>

⚡️🚀 המאמר הימי של מיק - 29.10.24: Global Lyapunov functions: a long-standing open problem in mathematics, with symbolic transformers

אתם אולי שמתם לב שיש לי נטייה לא להتلubb יותר מדי מיכולות של מודלי שפה בטוח בתחוםים של ריזונינג ופתרון בעיות מתמטיות קשות. אז היום אני מודה שאתה קצת (מש טיפה) מתלהב מהמאמר שאני הולך לסקור. המחברים אימנו מודל המסוגל למצוא פתרונות של בעיה מתמטית קשה שאין דרך כללית למציאת פתרונה.

מדובר בבעיית חיפוש של פונקציית ליapkovich למערכת דינמית. מערכת דינמית היא מתוארת על ידי מערכת משוואות דיפרנציאליות במישור בזמן. ידוע שאתה קיימת פונקציית ליapkovich למערכת דינמית אך אין

להגיד שהוא (המערכת) יציבה. המערכת יציבה אם הפתרון שלו לא מתבדר בזמן כלומר נמצא בתחום מסוים סביר 0 עבור כל זמן t (או לפחות שואף ל 0).

לפונקציית ליאפונוב (x) \dot{x} תכונות מסוימות (כי כMOVן תלוי בפתרון (t)x של מערכת הדינמית (למשל 0 = (0)V והיא שואפת לאינסוף כאשר x שואף לאינסוף. למייבר זכרוני (x) V קשורה לאנטורפיה של המערכת (תקנו אותו אם אני מתבלבל כאן).

כאמור אין דרך כללית למצוא (x) \dot{x} עבור כל מערכת דינמית אבל למערכות דינמיות מצוראה מסוימת (פולינומיאלית) ניתן למצוא אותה. המחברים למשהו אימנו טרנספורמר שבහינתן מערכת דינמית יודע למצוא את (x) \dot{x} עבורו. הם בנו דאטסהט של מערכות מסוימות דיפרנציאלית עבור מערכות דינמיות ו(x) \dot{x} עבורו ואימנו טרנספורמר לחזות את פונקציית ליאפונוב שלהם וזה גם עבד במקרים שלא ניתן לעשות זאת בדרך מתמטית ריגורוזית.

הדאטה מועבר לטרנספורמר בצורה סימבולית כמו כן כל נוסחה מתוארת על ידי שכל קודקוד בו הוא או פונקציה מתמטית או משתנה ואילו הקשות מוקודדת פעולות מתמטיות שונות. עץ זה מזון לטרנספורמר בסדר מסוים (קבוע לכלום).

חייב להגיד שהזאת די מרשימה אך מסיג את זה בהבנתי הרדודה בנושא המערכות הדינמיות.

<https://arxiv.org/abs/2410.08304>

🚀⚡️: 30.10.24. 🚀⚡️ المقالة اليومية حول مايك - Beyond Preferences in AI Alignment

היום סקרה של מאמר ללא נוסחאות אבל קשה לי לקרוא לה קלילה. יש בה דינמים פילוסופיים לא פשוטים וזה מה שהבנתי ממה (תקנו אותו אם אני טועה).

המאמר מציג ביקורת מקיפה על הגישה המבוססת-העדפות (preference based) לישור(alignment) של AI. המחברים טוענים שהגישה הנוכחית, המתמקדת בהעדפות אנושיות כיחידה הבסיסית של ערכיהם האנושיים, היא בעייתית ומוגבלת.

הם מציעים מסגרת חלופית המכירה בכך שהעדפות אנושיות הן מורכבות, משתנות לאורך זמן, ותלוות בהקשר חברתי. המאמר מציע גישה חדשה המבוססת על קритריונים נורטטיביים ספציפיים לתפקיד (של המודל), במקום על העדפות גלומיות.

המאמר גם דין בצורךמערכות AI שמסוגלות להבין ולכבד את המורכבות של ערכיהם האנושיים, במקום לנסוט לפשט אותם למודל של העדפות פשוטות. הם מציעים גישה חזית (contractualist) לישור AI, המבוססת על הסכמה הדדית בין בעלי עניין שונים. יש שם (במאמר) ביקורת על התיאוריה הקיימת של בחירה רצינלית (שהיא סוג של preference-based preference שיש לנו כרגע) ומציע חלופות המתהשבות במוגבלות הקוגניטיביות האנושיות.

הכותבים מתייחסים לשאלת כיצד לטפל במצבים בהם העדפות שונות מתנגשות זו בזו. הם מציעים מודל חדש הנקרא Evaluate, Commensurate, Decide (בדומה ל-Evaluate, Compensate, Decide). המאמר מציע מנגנון מסוים כמו דרכי לישום גישות אלו לימון מודלי AI (בצורה די כללית אני חייב להגיד). המאמר מציע מסגרת (תיאורית) לפיתוח מערכות المسؤولות להתמודד עם שינויים בהעדפות אנושיות לאורך זמן.

המאמר מדגיש החשיבות של פיתוח מערכות AI שיכלות לתפקיד כ"כלים" (מתוחכם אבל מתחמה אף עם "מרחב פעולה צר ומוגדר") ולא כסוגרים אוטונומיים.

ניתן למצוא במאמר גם (איך לא) דינונים בחשיבות של שמירה על פולויליזם בפיתוח AI, כך שמערכות, משלבות AI, יכולים לשרת מטרות שונות תוך כבוד נורמות מסוימות המשתנות לקבוצות שונות ולפעמים תלויות גם בנסיבות.

יאלאה, עכשוו תגידי האם הבנתי נכון....

<https://arxiv.org/abs/2408.16984>

⚡️🚀 המאמר היומי של מיק - 31.10.24:

Understanding Transformers via N-gram Statistics

מאמר די נחמד ולא רגיל מבית גול. המאמר מבהיר אותנו לתקופה שלא מידלו את השפה הטבעית באמצעות מודלים סטטיסטיים עם عشرות ומאות מיליארדי פרמטרים. פעם ניסינו להשתמש בס-grams כדי לשער את התפלגות של המילים בטקסט. כמובן גישות אלו לא יכולות לעבור עברו דאטטיסטיים בעל שירות טרילוני טוקנים כמו שיש לנו היום אבל אולי אפשר לקחת LLMים גדולים ולבדק האם ניתן לקרב את חיזיהם באמצעות סטטיסיות על s-grams. כדי לא לסבך המאמר לא לבדוק את זה על מידע *in-context*.

זה בדיק מה שהמאמר זהה (יש לו רק מחבר אחד שזה די נדיר בימינו) עושה. הוא בודק האם ניתן לחזות את הטוקן הבא שמודל שפה מאמון חזר באטען סטטיסטיקה של s-grams שבאים לפניו בטקסט. במקרה זה s-grams בנויים אלא מיללים מטוקנים. דרך אגב הסטטיסטיקה של s-grams אינה חיבית לכלול את כל חטוקנים הבאים לפני הטוקן הנחזה אלא עשויה "להכיל חורים" (כלומר יכולה לקחת טוקן 3-gram או 4-gram ועוד - נצטרך למצוות מעל טוקן 3-gram בשביל כך).

המחבר מצא כמה דברים מעניינים. ניתן לשער את החיזוי של מודל שפה עם 7 gram (עבור דאטטיסטיים שהם בחורו) ללא מעט מקרים. בסופו נמצאו כי לטוקנים בעל שנות נמוכה (של התפלגות שלהם) s-grams מצלחים יותר מאשר לטוקנים בעלי שנות חיזוי גבוהה. מעניין שככל שמאנים מודל שפה יותר יותר קשה לקרב אותה עם s-grams (צריך להגדיל את ח או לא משנה מה ה-gram הקירוב יורד).

אהבתי ...

<https://www.arxiv.org/abs/2407.12034>

⚡️🚀 המאמר היומי של מיק - 01.11.24:

LLMs Are In-Context Reinforcement Learners

אני אוהב מאמרים שמשלבים כמה שיטות של ML. אס考ר היום אחד זהה המציג לשדר למידת *in-context* עם במידה באמצעות חזוקים או בקצירה RL. למידת *in-context*-הו היא יכולת של מודל שפה ללמידה משוח חדש מכמה דוגמאות בפרופט ללא צורך בפין טון. יש לא מעט הסברים ליכולת דן מפתיע זו ולפעמים יכולת זו נקראת *emergent capabilities*.

עכשוו נשאלת השאלה: איך נוכל לבחור דוגמתה להדגמה שאנו מראים למודל שפה בפרומפט למקסום ביצועי המודל? השאלה זו לא מאד טרייאלית ואין אליה כרגע תשובה חד משמעית. המחברים מציעים לגשת לבעה זו דרך למידה עם חזוקים (סוג של). השיטה הנאיבית היא פשוט לצבור דוגמאות עד שנגמר לנו את אורך חלון ההקשר של המודל. לכל דוגמא בהדגמה אנו שומרים במאגר את השילishi המכילה את הדוגמא (שאלה עצמה), תשובה המודל ומשער של איקות התשובה (או פשוט האם התשובה נכונה או לא). ואז באינפנס פשט לנקחים את הדוגמאות הללו בתור פרומפט.

לטענת המחברים הגישה הנאייבית חזות לא עובדת ממשי סיבות עיקריות. קודם כל שילוב מתמשך של אותו הפרומפטים לדוגמאות שונות מוביל לשונות גדולות בפלט של LLM (לפי המחברים הקדומים עלולה להוביל לביצועים ירודים). הסיבה השנייה טמונה בכך שלשלושות (שאלה, תשובה, לא נכון) מסבכות את המודל ולא מספקות לו מספיק מידע על איך היה צריך לענות נכון (ד"א בלמידה ניגודית יש בעיה דומה המצורכה כמעט מאוד גדולה של דוגמאות שליליות בכלל באז' - כתבתי על זה לא מעט בסקרים").

عقب כך המחברים הציעו להנגיש קצט "אקראיות" לבניית הפרומפטים (המחברים קוראים לזה אפיוזדה בהתאם לטרמינולוגיה של RL - כל אפיוזדה מורכבת מכמה שלשלושות של שאלה, תשובה, נכוןות התשובה) וגם להשתמש באפיוזדות שקיבלו ציון "נכון". לכל דוגמא הם הצעים קודם לדוגם באקראי מהבאפר של אפיוזדות בצורה אקראית ולהשתמש לכל דוגמא במדגם שונה. כאמור שמדובר רק את האפיוזדות שביהם המודל זדק. כך פרומפט לכל שאלתה הופך להיות לא קבוע ומכיל רק דוגמאות עם תשבות נכוןות. זה נקרא ICRL Explorative במאמר.

כמובן Sh Explorative ICRL לא יעיל חישובית כי כל פעם צריך לחשב את הפרומפט מחדש (מה שלא צריך לעשות בגישה הנאייבית אך לא עובדת). המחברים שכללו את זה עם מנגנון קאשינג המאפשר לשומר מספר קבוע של פרומפטים (מערך של אפיוזדות) ולכל אפיוזדה נתונה להחליט לאלו מהם להוסיף אותה. זה מקל על הูลות החישובית.

מאמר חמוץ למרות שימוש מה לקח לי קצת זמן להבין אותו... .

<https://arxiv.org/pdf/2410.05362>

⚡️🚀 02.11.24: המאמר היומי של מיק - Learning to Compress: Local Rank and Information Compression in Deep Neural Networks

היום סוקרים מאמר כחול לבן למחלוקת אחד המחברים משנים הוא ישראלי ורבי שוויץ (זיו) והם חוקרים נושא שמשמעותו אוטם מאוד באופן אישי. הנושא הוא דחיסה של דאטא באמצעות רשותות נירונים והוא גם מאוד קשור לעבודות של נפתלי תשבי האגדי בנושא צואואר בקבוק מידע (information bottleneck או IB) וגם השערת ירעה (manifold hypothesis) בקשרו לרשותות נירונים עמוקות.

המודל שבדעתם האמתי (כגון תמונות או טקסט) אינם מפוזרים באופן אחיד במרחב בעל מימד גבוה, אלא שכנים על ירעה בעל מימד נמוך יותר. רשותות נירונים עמוקות מצליחות היטב עם הדטה זו כי הן לומדות לזהות ולנצל את המבנה של אותה ירעה, מה שמאפשר להן לבצע הכללה טובה למרות המורכבות העצומה של המרחב המקורי.

כמובן שזה קשור לדחיסה כי ניתן לראות במיפוי מרחב בעל מימד גבוה למרחב בעל מימד נמוך שהרטשות עשוות בהתאם לHM סוג של דחיסה. ניתן לראות "שההממד האמתי" של מרחב הפיצרים של שכבה ברשות נירונים קשורה לראנק (=דרגה) של היעקוביאן שלהם (הפיצרים) ביחס לקלט. למה זה קורה בעצם? הרי מרחב האפס של היעקוביאן מייצג כיוונים שבהם האקטיביזציות של השכבה לא משתנות (копונקצייה של הקלט). ככל שמייד של מרחב האפס גדול יותר הדרגה של ירעת הפיצרים בשכבה נמוכה יותר. משמעות הדבר היא שהתרחשה יותר "דחיסה" או הפחיתה מימדי הקלט.

מצין שמטריצות עם דרגה לא מלאה מהוות מרחב בעל מידה אפס במרחב של כל המטריצות (כמו הסתברות של כל מספר עם דוגמים יוניפורמיים בין 0 ל 1).عقب כך המאמר מגדר local rank או RLR שזה מספר

ערכים סינגולריים (הכללה של הערכים העצמיים) של היעקוביאן שהם גדולים ממספר קטן אפסיון אך חיובי (מצור עבור דרגה אמיתית צריך להחליף אפסיון ב-0).

אוקי, מקווה שדרදתם את זה אז עכשו מגיעים שני המשפטים העיקריים של המאמר. הם טוענים שברשותם עמוקות (מספר שכבות גבוהה) בבעיות סיוג תמיד יהיה שכבה 1 שה-RLR יהיה נמור מ-(פרופורציונלי לאפסיון בחזקה מינוס 2 ובנורמת אופרטור של מטריצת השכבה 1 (נורמת אופרטור זה הערך הסינגולרי הגבוה ביותר). הכוונה כאן לרשף שעושה התאמת מושלמת לדאטה האימון (עם מרג'ין 1 כלומר מצילחה להפריד בין הקטגוריות השונות בבטחה). משמעות המשפט היא שהרשות המאומנת דוחשת את הדאטה בשכבה 1 באופן אפקטיבי.

המחברים מוכחים משפט דומה בנוגע לביעות רגסיה.

תמיד כיף לצלול למתמטיקה חמודה עם המאמרים של רVID 😊

<https://arxiv.org/abs/2410.07687>

🚀⚡️: המאמר היומי של מייק - 03.11.24 🚀⚡️

TOKENFORMER: RETHINKING TRANSFORMER SCALING WITH TOKENIZED MODEL PARAMETERS

אוקי, זה מאמר די לא צפוי עם רעיון פשוט להבנה ובאופן די מפתיע (לפחות אותי) גם עובד (לפי מחברי המאמר כמובן). מכירים את הטרנספורמרים או שair שנייה אופנתי לקראו שנאים בעברית. בлок הטרנספורמר (בן הבניין) של ארכיטקטורה זו) מורכב מנגנון attention (אמרו לא לקרוא לזה "תשומת לב" כי זה לא נשמע טוב) יחד עם 2 שכבות FF feedforward או FF (יש אקטיבציה לא לינארית רק בשכבה הראשונה מהן). בנוסף יש כמה שכבות נרמול (לבחירהכם) וזה כל הק损.

از המחברים של המאמר מציעו שינוי בארכיטקטורה זו (שימוש בגראם כבל 7 שנים) שינוי די לא צפוי. מה שהוביל אותם לשינוי הזה זה קושי של השינוי המימדים של שכבות הקלט ופלט בлок טרנספורמר שמחיב אימון חדש של כל המודל (המורכב ממספר בлокי הטרנספורמר). אני לא משוכנע שזה נכון.

از כדי להתמודד עם הסוגיה זו המחברים הציעו להחליף את שכבות FF במנגן שקיבלו שם PAttention שמחשב משהו שקצת דומה ל-attention. אמם לא באמת דומה כי אין שם השוואה בין הייצוגים השונים של טוקנים (המופקים באמצעות מטריצות Q ו- V כאשר ההשווואה מוחשבת דרך מכפלה פנימית שלהם ונורמל עם softmax). מה ש-PAttention באמת הוא חישוב המשקלות של FF - כאן צריך להזכיר כי בлок השנאי הרגיל הוא גם שכבה feed-forward כאשר משקלותיה תלויות בקלט (דרך מנגן-h-solution המקורי של השנאי).

מה ש-PAttention עושים הוא חישוב של המשקלים האלו באופן הבא:

- מכפלה של ייצוגי הטוקנים במטריצה P_K נלמדת
- נרמול ורגיל של הוקטור המתkeletal (מחלקיים בשורש של הנורמה הריבועית)
- הפעלת פונקציית אקטיבציה לא לינארית (זה U-GeLU שמודר עם erf למי שמתעניין)
- המכפלה במטריצה P_V נלמדת

אז מה יש לנו בסוף? שכבת fully connected עם משקלים מחושבים בדרך טיפה שונה מה-h-solution הרגיל במקום שכבת FF שיש לנו בשנאי. מפעילים את ה-PAttention אחריו בлок attention הרגיל.

וכז אפשר לשנות את מספר מימדים של המטריצות הפנימיות של השנאי ללא retraining מלא של המודל (על ידי שרשור המטריצות החדשנות הנלמדות של PAttention עם הישנות שכבר אומנו)..

וכל הסיפור זהה עובד ...

<https://arxiv.org/abs/2410.23168>

⚡️🚀 המאמר היומי של מijk - 04.11.24: Refusal in Language Models Is Mediated by a Single Direction

מאמר מעניין החוקר איך ניתן לגרום למודל שפה לחת תשובות רצויות יותר ורצויות פחות. מתברר שאפשר לגרום למודל להסביר לנו איך מכינים הרואין או שודדים בנק וממליטים מהעונש עם אם מזיזים פلت של שכבה אחת במודל שפה. וגם ניתן למנוע ממודל "לא מרוסן" לחת תשובות לא פוגעניות ולפעמים להימנע מלענות על שאלות מסווגות אם מזיזים את הפלטים של כל השכבות של מודל, כל אחת עם וקטור \mathbf{z} כאשר \mathbf{z} מספר השכבה.

איך בעצם מוצאים את הוקטוריהם האלה? עבור דאטסהט המכיל שאלות ותשובות רצויות מחשבים את ההפרש \mathbf{z} הממוצע (על כל התשובות) בין האקטיביזיות של כל שכבות המודל ועבור כל הטוקנים של חלון ההקשר. כמובן יש לנו מטריצה $A\mathbf{x}$ של וקטורי ההפרש כאשר \mathbf{x} זה מספר השכבות \mathbf{z} זה מספר הטוקנים בחלון ההקשר.

כדי לגרום למודל להיות "פחות מרוסן" אנו בוחרים שכבה שהוספקן של מוריידה ממנו את בלמים בצורה המשמעותית ביותר (יש ממדים לא רעים לכך). כמובן משאיירים \mathbf{x} וקטורי הפרשימים שיחסבנו. כדי לגרום למודל להיות יותר מנוח צריך להחסיר את "כיוון הגסות" מכל השכבות של המודל בצורה שתעביר אותם ממוחב אורותוגונלי \mathbf{z} (כל שכבה ולכל טוקן בחלון ההקשר). בפרט מכל אקטיביזיה x בכל שכבה ובכל טוקן: $x^* \mathbf{Z}^*$ \mathbf{z} . קל לראות שהווקטור המתkeletal כתוצאה מכך יהיה אורותוגונלי \mathbf{z} .

עשויים זאת לווקטור האקטיביזיה לפני residual connection בכל בלוק של טרנספורמר. כמובן (מכיוון שיש הרבה מכפלות של מטריצות) ניתן להציג גם את המשקלים שלהם כדי לקבל את אותם האפקטים. מאמר ד' מגניב וקל להבנה.

<https://arxiv.org/abs/2406.11717>

⚡️🚀 המאמר היומי של מijk - 05.11.24: RETHINKING SOFTMAX: SELF-ATTENTION WITH POLYNOMIAL ACTIVATIONS

מאמר ד' לא רגיל והוא מדבר על חלופה פוטנציאלית של מנגןון-h-attention שאנו כה אוהבים בטרנספורמים. אתם בטח זכרם שמשקלי attention בשנאים מחושבים עם softmax שהוא מormal וקטורי משקלים לנורמה 1 ובנוסף כל רכיביו הינם בין 0 ל-1. כמובן הוא מראה התפלגות הסתברותית. המחברים טוענים שתכונות אלו של המשקלים לא קרייטיות לפונקציונאליות של השנאים ומצביעים להחליף אותם בקורסיל אחר שהוא פולינומיAli כפי שאתם בטח ניחשתם מהשם של המאמר.

אבל למה זהעובד בכלל? המחברים טוענים (באופן ד' מפתיע, אני חייב להגיד) שהביטויים הנפלאים של הטרנספורמים נובעים בחלוקת מילולתה של פונקציית סופטמקס לכפות רגולרייזציה מסוימת על נורמת פרובניוס של מטריצה המשקלים וגם של היעקוביאן שלה (ביחס לקלט של הסופטמקס) במהלך האימון הוא מסדר (ח)זק אשר ח הינו מייד לקלט.

נורמת פרובניוס או NF מוגדרת בתור שורש של סכום הריבועים של כל הערכים במטריצה והיא גם שווה לשורש של סכום הריבועים הערכים הסינגולרים (הכללה של ערכים עצמים למטריצות לא ריבועיות). ד"א סופטמקס

מוחשב במנגנון attention של מערך של וקטורים איז העוקbijan תיאורית הוא טנזור תלת ממדי (המאמר מפרט איך מחשבים את NF במקרה זה).

از בגודל המאמר מוכיח שני משפטיים. הראשון מהם טוענים ש-NF של מנגנון attention פולינומיAli (כולל הלינארי) מתנהג לפי (ח)O אם המטריצות שם, K ו-Q וגם ייצוג הטוקנים מפולגים גאומיטים כמובן). אז אם מנורמלים את-h-attention הפולינומיAli עם (0.5)- \sqrt{h} מקבלים את (ח)sqrt שהיה לנו עבור מנגנון-h-attention הרגיל. בנוסף NF של היוקbijan לפי Q, המנורמל לפי (0.5)- \sqrt{h} לא זה מתנהג לפי (ח)sqrt ב-h-attention הרגיל) גם מתנהג לפני (ח)sqrt.

המחברים טוענים זהה מספיק כדי לטעון שניתן להחליפ סופטמרק בפולינומים שיזור קלים מבחינה חישובית, מקבלים תוצאות מעודדות אבל אני עדין לא השתכנעתי...

<https://arxiv.org/abs/2410.18613>

的文章的日誌 - 2024.11.07 🚀⚡️ Cross-layer Attention Sharing for Large Language Models

אתם בטח יודעים הרצה של מודלי שפה עלול להיות דבר די יקר מבחינת משאבי חישוב וגם הזיכרון. בטח כאשר יש לכם מודלים עם عشرות מיליארדי פרמטרים על עשות רבות של שכבות של טרנספורמרים. אחד הדברים הקבדים שמצריכים לא מעט זיכרון הוא KV-Cache, שבו נשמרים המכפלות של ייצוגי (אمبادגס) של הטוקנים במטריצות K ו- V לכל השכבות וכל הטוקנים שכבר גונרטו (כולל הפרומפט - מדובר במודלי הדקדורים).

כמובן שכאשר המידים של וקטורי הייצוג והמטריצות לא קטנים וגם אורק ההקשר נמדד בעשות ומאות אלפי KV-Cache נדרש הרבה מאוד זיכרון. בעבר יצאו לא מעט מאמרים שניסו לדחוס אותו על ידי ניתוח ויזיהו יתריות אבל זה בד"כ געשה פר שכבה (= בלוק הטרנספורמר). המאמר המסורק מציע להתבונן בדוחיסת KV-cache מפרספקטיב רחבה יותר ולנסות לדחוס אותו דרך ניצול התלות של KV-cache בין השכבות השונות.

המחברים חקרו דמיון בין החלקים השונים בבלוק הטרנספורמרים (מכפלות של המטריצות השונות בוקטור ייצוג, מקדמי attention וכדומה) והגינו למסקנה שניתן "להסיק" את מקדמי-h-attention של שכבה Ch מהدادה של שכבה 1-ה בצורה חסונית חישובית. ככלומר עם הרבה פחות משקלות מהטרנספורמר הרגיל. ככלומר ההצעה היא לעשות סוג של SoRa אבל למקדמי-h-attention.

בצורה קצרה יותר קונקרטית המאמר החליף מטריצות Q - K - W במטריצות בעלות ראנק נמוך (מכפלה של שתי מטריצות מלכניות כאשר המימד הפנימי של המכפלה נמוך - כולם $(N \times k^* \times M)$ כאשר k קטן הרבה יותר מ- M ו- M -ה. מחשבים את הקלט לסופטמרק עם המטריצות הללו. לאחר מכן משרשרים אותם עם הקלט לסופטמרק מהשכבה הקודמת, מפעלים FFN והנה יש לנו קלט לסופטמרק בשכבה Ch. ושימושם לב שאנו צריכים לשומר הרבה פחות דатаה ב-KV-cache כי יש לנו מטריצות בעלות ראנק נמוך.

AIR מאמנים את הסיפור זהה? משלבים את הלווי הרגיל של מודל שפה עם LOSS distillation שמטרצה לקרב את מקדמי-h-attention המוחשבים בדרך המוצעת עם אלו שמוחשבים עם מודל רגיל (עם SoRa ו- RGILIM).

מאמר ד"י מעניין - אבל קצר אורך מדי לדעתי אז תמצתתי לכם אותו 😊

<https://arxiv.org/abs/2408.01890>

⚡️🚀 08.11.24 : המאמר הימי של מיק - ⚡️🚀

Occam's Razor for Self Supervised Learning: What is Sufficient to Learn Good Representations?

סקירה קצרה של מאמר המציע גישה חדשה ללמידה self-supervised או SSL בקצרה. אזכיר כי שיטת SSL מניחה שיש לנו דата לא מתויג ומטרתנו לאמן מודל מסוגל להפיק ייצוג חזק של דатаה. מה זה ייצוג חזק של דטה, אתם שואלים? בד"כ הכוונה לכך שבנוסף לכך (נגיד רק עם תוספת שכבה לינארית) לבניית מסוג בעל ביצועים טובים.

כלומר כזה שידע להפריד בין הקטגוריות השונות של דטה בלי לדעת אותן בצורה מפורשת (למשל אלו יכולים לאמן מודל בצורה SSL על התמונות של ImageNet בלבד להשתמש בתווים ואז לבדוק האם המודל הצליח למדוד להפריד בין הקטגוריות השונות).

בד"כ SSL מבוצע עם שיטות של במידה ניגודית (contrastive learning) כאשר מטרתו מאוד בגודל היא לקרב ייצוגים של פיסות דטה דומות (חיוביות) ולהרחק את הייצוגים של פיסות דטה לא דומות (שליליות). לרוב זוגות חיוביים נבחרים בתור אוגמנטיות שונות של אותה הדוגמא כאשר הזוגות השליליים הן דוגמאות שנבחרות באקראי. שיטות כאלו נחלו הצלחה די גודלה אבל דריש דאטסהטים מאוד גדולים וגם משאבי אימון די משמעותיים כדי נדרש שם גודל באז' די גדול כדי שהשיטה תעבור טוב.

המאמר המסורק מציע שיטה מאוד פשוטה ואמנויות-ביתית ל-SSL (תער אוקט). במקומם לעבוד עם הייצוגים המאמרים מאמן מודל לחזות את המספר של הדוגמא בדאטסהט. ככלומר אם יש לנו 1000 דוגמאות מהדאטסהט יש לנו 1000 קטגוריות ומטרתנו לחזות קטגוריה של דוגמא מהייצוג הלטנטי שלה (המודוק על ידי המודול המאומן). ככלומר אחרי השכבה האخונה של המודול מושפעים שכבה עם מטריצה המיפה את הייצוג לקטגוריות (כלומר המספרים הסידוריים של הדוגמאות). ובוסף של לוי cross-entropy הסטנדרטי.

از המאמר מוכיח שהשיטה עובדת לא רע לדאטסהטים יחסית לא גדולים (معنىין איך זה יעבד לדאטסהט בגודל 10 מיליון). כמובן יש כמה טריקים באימון כמו soft labels אבל בגודל הרענון די נחמד.

<https://arxiv.org/pdf/2406.10743>

⚡️🚀 09.11.24 : המאמר הימי של מיק - ⚡️🚀

CROSS-ENTROPY IS ALL YOU NEED TO INVERT THE DATA GENERATING PROCESS

מאמר המשך של המאמר שסקרטרי אtamol שהציג שיטה חדשה ל-SSL או Self-Supervised Learning בariecot. מטרת SSL היא לבנות מודל המפיק ייצוג דטה עצמאי שהיה קל לבנות ממנו מודלים downstream לביוץ משימות שונות על הדטה זהה בתור backbone (למשל על ידי הוספה שכבות, LoRA, אדפטרים או שיטות פין טיון אחרות הבניות על backbone זהה). ככלומר הייצוג הזה צריך להיות מסוגל לזרק את כל התכונות המהוויות של הדטה הזה ככלומר לדוחסו בצורה יעה.

משימת downstream הפשוטה ביותר היא משימת סיווג ובמקרה הזה מודל ייצוג טוב צריך להיות מסוגל להבדיל בין דטה שיר לקטגוריות שונות (למרות שהמודול עצמו מאמין על דטה לא מתויג). המאמר של אtamol הציע לאמן מודל שיזדע להזות פיסת דטה מהייצוג שלה. ככלומר כל פיסת דטה מקבלת קטגוריה מסוימת (כלומר אם יש לנו דאטסהט עם 10K דוגמאות אז יש לנו 10K קטגוריות). בגודל ממשי שכבה לינארית נוספת לאנקודט (מודול הייצוג) שמיפה (השכבה הלינארית) את וקטור הייצוג לקטגוריות עם לוי cross-entropy.

از המאמר של אטמול טען שניתן להציג ליצוגים חזקים עם השיטה הזו (למשימות downstream מסווג סיווג) והמאמר המסורק הוכיח כמה טענות לגבי הרעיון שנדון במאמר (טוב זה לא בדיק אבל קרוב) שסקרנו אתמול תחת הנחות ד' הגיוניות. המאמר ד' מתמטי וננסה להסביר את הרעיון העיקרי בלי לצלול לנוסחות ולא התעמוקיות יתר לפרטים מתמטיים לא מהותיים.

המחברים מניחים כמה הנחות שעוזרות להם לחזור את הגישה הזו. ההנחה הראשונה מניחת שיש תהליך גנרטיבי המגנרט פיסות דاطה השיכים לכמה קטגוריות (מספרם ידוע). בפרט היא מדברת על כך שקיים מודל גנרטיבי ג' המגנרט דاطה מייצגו הלטנטי Z. המשנה הלטנטית Z בהינתן קטgorיה C מוגבל בהתאם חווים ד' וטורי C_Z תוחלת ופרמטר ריכוז (סקלר המגדיר את מידת המrixיות של ההתפלגות).

עכשו המשפט הראשון במאמר טוען אם מאמנים ייצוג f (האנקורד) עלי ידי מקסום פונקציה שדומה לזהות מהמאמר הקודם רק שהקטגוריות יהיה קטגוריות של הדטה(המייצגות במרחב הלטנטי) ולא כל פיסת דاطה שייכת לקטgorיה מסוימת(నכוון זה לא אותו הדבר אבל עדיין), יש פירוש ד' יפה לוקטוריהם W המרכיבים מטריצת W שהיא המייפוי הלינארי שאמו לומדים מהמרחב הלטנטי למרחב הדטה.

במקרה פשוט - משפט אחד מגדיר 4 מקרים, התלוים האם וקטורי ייצוג אחר (x,f), וקטורי W מהווים טרנספורמציה אורתוגונלית של מרכז הקטגוריות C_Z שמננו הוקטורים הלטנטיים מוגרלים (כלומר זה אותם הוקטורים תחת סיבוב רב מימדי כלשהו). כלומר קיבלונו W עם מאד קשרים לבניה של הדטה. בנוסף במקרה הזה ההרכבה של האנקורד f (מה שאמם מאמנים) והדקדדר g הינה לנארית כלומר הצלחנו למצוא את ההפכית של הגנרטור g - וזה תוצאה ד' חזקה (משפט 2 מנוסח את זה בצורה ד' טובה).

ההוכחות לא פשוטות בכלל ועם זאת המאמר הזה מאד חשוב ואני מקווה שהצלהתי לפחות להסביר לכם את מהותו.

<https://arxiv.org/abs/2410.21869>

🚀:10.11.24- המאמר היומי של מייק - 🚀 WHAT MATTERS IN TRANSFORMERS? NOT ALL ATTENTION IS NEEDED

סקירה קצרה של מאמר ד' נחמד החוקרים איזה חלקים במודלי טרנספורמרים (או שנאים) שלנו חשובות נחותים מהחלקים האחרים (או בכלל מיותרים). כמו שאתם זוכרים בכל בלוק של שניים יש לנו מנגןון-ה-חסון, כמו שכבות MLP (שהה שכבה וחצי של fully-connected) וכמה שכבות נרמול (אותם לא בודקים). החוקרים הקודם שחקרו את הנושא הזה התמקדו בזיהוי בלוקים שלמים של שניים העשויים להיות לא נחותים אך החוקר הזה החליטו לדעת לרוחולzieה של אבן הבניין של השני עצמו (כלומר attention ו-MLP).

איך בודקים האם תת-בלוק לא נחוץ? בודקים את הקלט את הפלט של תת הבלוק הזה ואם אין כמעט הבדל בין נראה שלא צריך אותו. כדי לבדוק את הדמיון משתמשים כפונק בדימון קוסינ (cosine similarity). בודקים את זה על כמויות גדולות של דאטה ומתייחסים להויריד שכבות ולבדק ביצועים.

מה התרבר? באופן קצר מפתיע לרוב מנגןון-attention הרבה הרכבה פחות נחותים מה-MLP ונitin לוותר עליהם בלי פגיעה רצינית בביטויים במיוחדם בגודלים. אז אולי זו הדרך להקטין את העומס החישובי ששימוש המודלים האלו גורם? בואו נחכה ונראה....

<https://arxiv.org/abs/2406.15786>

🚀⚡️: 11.11.24 - Stealing Part of a Production Language Model

זמן לא סביר מאמר על איזה ניתן לפרוץ למודלים עמוקים. יש תחום שלם שנקרא adversarial learning שבו חוקרים מפתחים מנגנוני הגנה נגד התקפות שמנוטות לגונבו משהו המודל (למשל דатаה שהוא אומן עלייו). המאמר שנסקרו היום מציע שיטה שבאמצעותה ניתן להזות המימד הפנימי (החבי) של המודל (מייד יציגי ה Tokennים) וגם את המטריצה בשכבה האחורה של המודל. שכבה זו המיפה את האמבדיניג של כל ה Tokennים לוגיטים שלאחר מכן מזינים לsoftmax שמננו יוצרים "הסתברויות של ה Tokennים".

נתחיל מכך שמייד המטריצה W בשכבה האחורה הוא $N \times N_{emb}$, כאשר N_{emb} זה המימד הפנימי של המודל (אלפים בודדים) ו- N_{voc} הוא מספר ה Tokennים במיילון (בד"כ כמה עשרות אלפיים ולפעמים מגע מעל 100K). ככלומר $N_{emb} > N_{voc}$ וזה בדיק מה שמחברי המאמר מאנצילים. מכיוון שהראנק של מטריצה W הוא N_{emb} כל המכפלות בה ממפות את הוקטור לתת מרחב במימד N_{emb} של מחרב הלוגיטים שהוא בעל מימד N_{voc} . ככלומר אם ניקח מספר וקטורי לוגיטים ונשים אותם לעמודות של המטריצה (נקרא לה V) המספר המקיים לי וקטוריים בלתי תלויים שייהי לנו יהיה בדיק N_{emb} .

זה בדיק מה שמחברי המאמר עשו. אולם מכיוון שהחישובים בטרנספורמרים הם לא בדיק המלאה (FP16 אג) אז קשה לתפוא מתי העמודות הופכות להיות תלויות. במקום זה הם חישבו את הערכיהם הסינגולריים (ע"ס) של V (דרך מה שנקרה SVD - מי שלא מכיר ממליץ לקרוא על זה) ומסתכמים מתי היחס של ע"ס העוקבים (הם ממשוניים) צונח משמעותית.

למה זה חשוב? כי במקרה האידיאלי ע"ס של V צריכים להתאפס אחר שעברנו את הראנק של או N_{emb} . אך בגלל אי דיוקים נוראים במודל כਮון שלא נראה ממש אפסים שם אלא ערכים מאוד נמוכים ואייפה שזה מתחילה לקורות זה בדיק בימייד $1 + N_{emb}$. אז עושים את הטריך הזה על הרבה מאוד דטה ומגלים את המימד החבי של המודל שלכם.

כਮון שבעולם האמתי אין לכם גישה לכל הלוגיטים אלא רק ל-K top ואז המאמר מנצל את העובדה שניתן לקונג' חילק המודול להוסיף מרגין לטוקן נתון במיילון. ואחרי מספיק משחקים מקבלים את כל הלוגיטים (זה די יקר חישובי).

מימד של W זה נחמד אבל מה עם מטריצה W עצמה. המאמר מציע התקפה כדי לגלות אותה (סוג של) גם. בכלל המאמר מלא ברעיונות יפים להתקפות על המודלים ומילויים מוזמן להעיף מבט.
<https://arxiv.org/abs/2403.06634>

🚀⚡️: 12.11.24 - OccamLLM: Fast and Exact Language Model Arithmetic in a Single Step

זהו מאמר שמש אהבת, אהבת גם את הרעיון וגם כתוב בצורה מאד ברורה. למה כה אהבת את הרעיון? אני כבר זמן מה טוען שבמקום לשקיע מאמצים גדולים באימון מודלי שפה לפטור בעיות מתמטיות יחסית מורכבות (שלductive) מאד קשה כי הם לא "בנויים" לזה באופן טבעי) כדי להשתמש בכלים חיצוניים ייעודיים לכך (למשל כלים סימבוליים). מטרה של מודלי שפה במקרה זה היא להזות מתי הקלט שモزن אליו (הפרומפט) מציריך פתרון

בעיה מתמטית, "לתרגם" את הבעיה לשפה של הכלי הייעודי זהה, להעביר את הבעיה המתווגת לשפטו אליו לפתרון ולפענה את הפלט שלו.

זה בדיק מה שהמאמר זהה עושה. המחברים לקחו מודל שפה ופתחו מודל נפרד לפתור בעיות מתמטיות. למעשה המודל לפתור בעיות מתמטיות שפותח במאמר הוא גרפ חישובי דינמי שככל צומת בו היא פונקציה או פועלה מתמטית (נדיג סימן $+$ ו- $*$, או \cos ו- \exp). יש גם צמתים למשתני קלט השונים כדי שהמודל יוכל לחשב פונקציות על כמה משתנים (multivariate). למעשה גרפ זהה הוא DAG או בשם המלא Directed Acyclic Graph ומאמנים אותו לבחור את "נתיב החישוב" בו ("מסלול הצמתים") בהינתן היצוגים (אמבידגנס של הטוקנים) המוחשבים על ידי מודל שפה לא מאומן ונוטר קבוע לכל אורך אימון המודל).

המחברים מאמנים שני מודלים: הראשון מזהה האם יש צורך בהפעלת המודל לחישובים מתמטיים לכל טוקן בהינתן ההקשר (כלומר כל הטוקנים לפניו). המודל השני מאומן לבנות נתיב חישובי בגרף החישובי שתיארתי בפסקה הקודמת. את שני המודלים הללו מאמנים בנפרד.

מעניין כל שכבה של רשת DAG זהה מורכבת משני חלקים: בחלק הראשון יש לנו צמתי החלטה: כל צומת זהה הוא וקטור "המחבר" אותו לצמתים פונקציונליים שככל מהם הוא בעצם פעולה או פונקציה מתמטית (מקובצת פעולות ופונקציות שבחורנו). הווקטור זהה הוא למעשה סופטמקס שמננו נדגם לאיזה צומת פונקציונלי/פעולה לחבר אותו. כל צומת פונקציוני שנבחר מחובר עם כל צמתי ההחלטה מהשכבה הבאה ואליהם מועבר היצוג משכבת ההחלטה הקודמת יחד עם ייצוג הפעולה (כנראה האם נבחרה או לא). כך נבנה גרפ חישובי מייצג הטוקנים המוחשבים על ידי מודל שפה (הם מוחברים לשכבת ההחלטה הראשון במודל החישובי). ד"א כל פעולה וכל פונקציה בסיס בgraf משוכפלת בכמה צמות כדי להקנות למודל יכולת לקרב פונקציות מורכבות יותר.

麥iouן שאמנו דוגמים את הגרף החישובי כל פעם מחדש כל פלט של מודל השפה, לא ניתן לאמן אותו בנסיבות על שיטות קלאסיות של למידת מכונה (supervised learning). המחברים בחרו בשיטה קלאסית מעולם למידה עם חיזוקים (RL) הנקראת *reinforce* כאשר פונקציית reward היא עד כמה התשובה המוחשבת באמצעות הגרף החישובית קרובה לתשובה *ground truth*. דרך אגב ניתן לייצג רוב הפונקציות עם עם יותר מחדד נתיבי חישובי.

מאמר ד' נחמד אבל כתוב לא מאד ברור (או שהיא חסר לי קצת רקע)...

<https://arxiv.org/abs/2406.06576>

🚀⚡️המאמר היומי של מייק - 16.11.24: NON-NEGATIVE CONTRASTIVE LEARNING

מאמר מעניין בנושא הלמידה ההפוכה (contrastive learning) או CL בקצרה. נזכיר שמטרת CL היא לבנות ייצוג יעיל לדאטה לא מותג שנוכל להשתמש בו לאחר מכן לaimon מודלי לשימושים downstream שונים (למשל על ידי הוספה של כמה שכבות ייעודית למשימה למודל שבונה את היצוג). השיטה הפופולרית ביותר ל-CL (שהה יש וריאציות וiscaloliים רבים) היא InfoNCE הוצעה לראשונה במאמר של et al Oord כבר בשנת 2018 הרחוקה.

השיטה מנסה לקרב יצוגים של דוגמאות דומות (כגון אוגמנטציה של אותה התמונה) מבחינה דמיון קווין (מכפלה פנימית מנורמלת) ובאותו הזמן היא מנסה להרחק יצוגים של דוגמאות לא דומות (הנבדקות BD'C באקראי). זה נעשה (בגדייל) על ידי אימון מודל שמנזר את היחס בין מרחקי הקווין (מעלים אותו באקספוננט) של זוגות דוגמאות שליליים (כלומר לא דומים) לזהה של זוגות דוגמאות חיוביים (דומים). נציין שבכל באז לוקחים מספר גבוה של זוגות שליליים (את הסיבות הסברתי בסקרים הקודמות בנושא).

המאמר מציע שיטה המשפרת את אינטואטיביות היצוגים הנלמדים, למשל כאשרם הקטגוריות השונות של דатаה (אזכיר שמדובר באימון עם דטה לא מותג) יהיו מרכזות ב"חלקים מסוימים" (תת-וקטורים) של וקטורי היצוג כאשר שאר הערכים יהיו אפסים או מאוד קרובים ל-0. וקטורים כאלה יהיו נוחים יותר לשימוש downstream הקשורים לשיווג דטה. המאמר טוען ששיטת CL עם פונקציית loss בסגנון InfoNCE לא מצליחה להפיק ייצוגים עם תכונות כאלה והסבירה העיקרית היא האינטואטיביות שלהם לשיבוב הנובעת מהצורה של פונקציית הלוס שלהם (הסביר מפורט בפרק 2.1 במאמר).

המחברים מציעים שני חידושים עיקריים. קודם כל הם מציעים לאמן ייצוגים שהם לא שליליים (ב-InfoNCE אין שום מגבלה כזו). החידוש השני הוא פונקציית loss שאכן מכילה מכפלות פנימיות של וקטורי ייצוג הדטה אבל בלי אקספוננטים ויחסים (כבר הוצע קודם אבל ללא אי שליליות). הפעם פונקציית הלוס היא הפרש בין המרחק הריבועי בין הדוגמאות השליליות לבין המרחק בין הדוגמאות החיוביות.

המחברים מצטטים מאמר שהראה שהיצוגים המופקים על ידי המודל המציגו loss זה ללא הגבלה של אי שליליות הינם שקולים לאלו המתקיים מפקטוריזציה סימטרית (מייצגים מטריצה כמכפלה של מטריצה F והשלוחוף שלה) של מה שנקרא מטריצה A co-occurrence. לך לי קצת זמן להבין מה זה בדיק אבל בגודל זה מטריצה המכילה סוג של "הסתברויות" של שתי דוגמאות היו חיוביות (אגמנטיזה של אותה הדוגמא).

כלומר אם יש לנו דאטסהט של 1000 דוגמאות -10 אוגמנטציות שונות פר דוגמא מטריצה A בגודל $10K \times 10K$ מכילה 1/10 לזוגות חיובים (כאשר תמונות ז-ז הן אוגמנטציות של אותה התמונה) 0 בשאר המקרים. מדובר כאן בפקטוריזציה למטריצה F שהיא low-rank אחד הממדים שלה (מידה הייצוג של דטה) הוא הרבה יותר קטן מהה מידים של מטריצה A (שהיא עצמה לדאטסהטים בגודל רציני, מיליון תמונות).

از המאמר משתמש באותו הלוס אבל מփש וקטורי ייצוג שהם אי שליליים (פעילים עליהם פונקציות כגון ReLU, sigmoid, softplus וכו'). בנוסף המחברים שמו לב כי בייצוגים המתקיים יש נוירונים מסוימים כלומר כאלה שמאוד קרובים ל-0 עברו כל הדוגמאות). המחברים משתמשים בטריקים נחמים כמו stop-gradient כדי להתחמם עם התופעה זו.

בסוף מתקבלים ביצועים משופרים כאשר היצוגים המתקיים הינם יותר disentangled ויתר קרובים לאורטוגונליות לדטה מקטגוריות שונות.

<https://arxiv.org/abs/2403.12459>

🚀⚡️ המאמר היומי של מייק - 18.11.24: 🚀⚡️

Knowledge Editing in Language Models via Adapted Direct Preference Optimization

היום סוקרים מאמר כחול לבן בנושא פיניטיון (=טיב, ככה אמרו לי) של מודלי שפה באמצעות טכניקות מבוססות על למידה עם חיזוקים או בקצרה RLHF. למייבר ידיעתי השימוש הראשון ב-RLHF היה במאמר במאמר שפיתח מודל הנקרא InstructGPT שימושו כבר ברור כי אומן לעקב אחרי הוראות המשתמשים. זה נעשה באמצעות טכניקת RL הנקראת Proximal Preference Optimization או PPO. מעניין ש-PPO הומצאה על ידי לא אחר אלא ג'ו שלמן שהוא תקופה די ארוכה CTO של OpenAI. בגדול מאמנים את המודל למקסם את פונקציית התגמול של תשובה תוך שמרתו (התפלגות הטוקנים) קרוב יחסית להתפלגות ההתחלתית (דרך KL divergence).

החישרון העיקרי של PPO היה צורך באימון מודל תגמול (reward) שבහינתן שאליה ותשובה נותנים ציון המשקף את אינטואטיביותם מנקודת ראייה של בני אדם (לפחות אלו שמאנים מודלי שפה). לשמהנתנו זמן קצר לאחר מכן

הouceה גישה המכרא DPO או PO Direct שאפשרה לטיב (או לישר כמו align) מודלי שפה ללא צורך בהשתמש במודל תגמול בצורה מפורשת (מנחים כורה אופטימלית של התגמול נפטרים ממנו). כדי לאמן מודל שפה בשיטת DPO צריך דאטסהט המורכב מתשבות רצויות יותר וrzciotot פחות ואנו מאמנים מודל.

המאמר למשה פיתח שיטה שהתאימה את DPO לבעיה של עירcit ידע (knowledge editing) של מודל שפה. כמובן אנו רוצים שהמודל יענה אחרית על שאלות מסוימות (נגיד מתאים אותו לדומין מסוים). בעיה זו שוקלה לביעית "ישור מודל שפה שניtin לפטור עם O. המחברים הציעו 3 שכליים עיקריים לו-DPO:

- במקום סט שאלות ותשובות(חויבית ושלילית) התשובות השיליות נוצרות על ידי המודל במהלך האימון
- התשובות השיליות(המודל הנוכחי) מג'ונרטות עם מה שנקרו teacher-forcing. כמובן עד טוון
- שחוזרים משתמשים בטוקנים של התשובה החיויבית (שהואתנו מצפים לקבל מהמודל לאחר עירcit ידע)
- האופטימיזציה עם DPO מבוצעת עם ה-teacher forcing הזה (נשמע מאוד הגיוני עם 2

יש תוצאות לא רעות כמובן...
<https://arxiv.org/abs/2406.09920v1>

⚡⚡⚡ המאמר היומי של מיק - 20.11.24 ⚡⚡⚡ Adaptive Decoding via Latent Preference Optimization

היום סוקרים מאמר חשוב שכנע אותנו שלא משנה כמה מאמראים אקרא עדיין אפסס רעיונות מעניינים גם בתחוםים שאין מתחמча (סוג של) ומתעניינים. כמובן מדובר בשיטות לג'נרט דטה מודלי שפה? המאמר זהה מציע שיטה המתאימה את הייפר-פרמטרי הג'נרט שלה כפונקציה של הקונטקט. למשל המאמר שנסקור היום עוסק בהתקאה של טמפרטורת דגימה לג'נרט דטה. אזכיר לכם שטמפרטורת הדגימה T שלטת באקריאות דגימה של טוון הבא - ככל שהיא גדולה יותר טוקנים עם "הסתברות דגימה" (מותנית בהקשר) נמוכה יותר מקבלים יותר סיכוי להידגם.

מתברר שגם מחקרים זה (התאמת הייפר-פרמטרי ג'נרט) קיימים כבר אריה 4 שנים ויצאו לפחות 10 מאמרים בנושא (שלא ידעת). אך המאמר הזה הוא המשך של כמה מאמראים שלא סקרתי בזמןו. אוקי אז כאמור המאמר מנסה לאפות את T בהינתן ההקשר. המחברים מניחים שאנו בוחרים T מט טמפרטורות סגור k_T, ..., T_1, ..., T. המחברים מציעים לאמן רשות t_M (נקרא Adaptive Decoder במאמר) החוצה את T האופטימלי בהתבסס על יציגו טוקני ההקשר. כמובן הרשות הפלגתת התפלגות מעל k_T, ..., T_1 (כלומר סופטמקס).

למעשה התפלגות צזו היא ממתקלת (משנה לפי התפלגות הטמפרטורות הנוצרת על ידי t_M) את התפלגות הסופטמקס מעל מיליון הטוקנים שמננו מודל שפה מגנרט טקסט. כמובן ניתן לאמן t_M בכמה דרכים על דאטסהט נתון במטרה למסס את הנראות(likelihood) של הדטה (לדעתי נעשה במאמרם קודמים). המאמר מציע לעשות את בשיטת DPO הלוקחה לעולם למידה עם חיזוקים עם RL (קראו סקירה מ-18.11.24 כדי לרענן מה זה). רק אזכיר שבשיטה זו מבצעים "ישור" (alignment) של מודל שפה על דאטסהט של תשבות רצויות ופחות רצויות.

از המחברים מציעים להכפיל את השיטה זהו עבור המקרא שאנו רק מאמנים את המודל אלא גם המודל לקביעת התפלגות טמפרטורה. הדאטסהט של תשבות וטמפרטורות רצויות נבנה על ידי דגימה של מודל שפה בטמפרטורות שונות ובחירה של התשובה הטובה ביותר והגראעה ביותר או עלי ידי מודל אחר או על ידי מתיגים אנושיים. וזה בדומה ל-DPO בונים פונקציית לוס שמעדכנת את מודל השפה וגם t_M יחד. הרוי ניתן לראות ב-t_M מודל דגימה ממילון הטוקנים כאשר כל טוון הוא טמפרטורה k_T. זאת הcalculation ד' מתבקשת. המחברים גם מציעים פונקציית לוס שמעדכנת רק את t_M בהתאם לצורה.

לבסוף המאמר מציע פונקציית לוס המאפשרת מודל שפה יחד עם \mathbf{f}_M כאשר התפלגות של הטוקנים (של מודל השפה) מבוטאת דרך מריגינלייזציה שלה מעלה התפלגות הטמפרטורות דרך נוסחת בייס. כלומר מיישרים את המודל לתעדף רק תשובות רצויות באופן ישיר אבל יחד עם זאת גם \mathbf{f}_M מתעדכן.

<https://arxiv.org/abs/2411.09661>

🚀⚡️: 21.11.24 - המאמר היומי של מייק

Unfamiliar Finetuning Examples Control How Language Models Hallucinate

מאמר של סרג'י לוין האגדי מאוניברסיטת טורונטו שידוע יותר בתרומתו האדריכלית לפיתוח שיטות מבוססות למידה עם חיזוקים (RL) ליישומי רובוטיקה. הפעם הוא עם קבוצתו חוקר את תופעת ההזיות (hallucinations) של מודלי שפה. הזיה זה מושג מאוד רחב בהקשר מודלי שפה ובדגול (מאוד) ניתן להגידו בטור מתן תשובה לא נכונה (בעיקר עובדתית) על ידי מודל שפה.

מאז שמודלי שפה נכנסו לחימנו בשנים האחרונות תופעה זו נחקרה באופן נרחב בעשרות (אם לא מאות) מאמרים. המאמר שנסקור היום חוקר סיבות לתופעה זו וגם מציע דרכי להתמודד איתה. החוקרים טוענים כי הסיבה להזיות טמונה בניסוי להקנות למודל ידע חדש במהלך טיבון (finetuning). המחברים טוענים שהמודל נוטה ללמידה פחות טוב את העובדות הנמצאות בדאטאטס של FT (נקרא לו D_FT) שלא מייצגות מספיק טוב בדאטאסט הגדל ששימוש את המודל לאימון מוקדים (נקרא לדאטאסט זה בתור D_PR). עובדות (ושאלות עליהם) נקבעות לא-monicיות במאמר.

בפרט המאמר משער (וmare'a אמפירית) שעבורו שאלת אם המודל מוציא תשובה שהיא סוג של תשובה מוצעת עבור כל השאלות הלא מוכרות מ-D_FT. כלומר צו שמצוירת את פונקציית הלוס המוצעת על כל השאלות הלא מוכרותalo מ-D_FT. ומכיון שרוב התשובות ב-D_FT מנוטה היטוב ובאנגלית רהוטה אנו מקבלים תשבות יפות אך לא נכונות בהחלט מודל שפה לשאלות לא מוכרות.

בגדי הרעיון העיקרי שהמחברים מציעים לתקן המצב הזה הוא ללמד את המודל להגיד "לא יודע" בצורה ברורה על שאלות לא מוכרות (כלומר במקרים שהוא אכן לא יודע). אחת הדרכים לעשות זאת היא קודם לזרות שאלות לא מייצגות מספיק ב-D_FT (על ידי ניתוח שכיחותם או אנטרופיה של הלוגיטים של תשובה המודל לשאלות אלו - ד"א שניהם לא אידאלים באספקט הזה). לאחר מכן במקומם לאמן מודל לענות תשבות נכונות לשאלות אלו (שהוא לא מסוגל ללמידה), תשבות אלו מוחלפות ב-D_FT על ידי תשבות ניטרליות בסגנון "אני לא יודע". כמובן אפשר להוסיף ל-D_FT מלא שאלות הלא מוכרות ב-D_PR עם תשבות אלו.

הדרך השנייה היא לאמן מודל עם שיטות של RLHF עם שינוי של פונקציית תגמול (reward) המקטין קנס על תשבות ניטרליות ומשאיר את שאר התגמולים כמו שהם. המחברים מראים (אמפירית) שבמקרה זה המודל יותר "שםך" לתת תשבות ניטרליות לשאלות לא מוכרות. המאמר מציע שיטה המורכבת מ-4 שלבים לאימון RLHF לשיפור יכולת המודל להגיד "לא יודע":

1. עושים FT רגיל
2. דוגמים את המודל עם שאלות מוכרות ולא מוכרות
3. בונים פונקציה תגמול הקונסת את המודל יותר על תשבות לא נכונות לשאלות לא מוכרות (וקנו מאוד נמוך או 0 על תשבות מתחממות)
4. אימון RLHF עם פונקציית התגמול מסעיף 3.

מאמר נחמד שהשair בי טעם לראות את המשך.

<https://arxiv.org/abs/2403.05612>

🚀⚡️ המאמר היום של מיק - 22.11.24 ⚡️🚀 The Unreasonable Ineffectiveness of the Deeper Layers

מאמר קליל שלא יקשה עליו יותר מדי בסופ"ש. המאמר מציע דרך מואוד פשוטה לקצץ שכבות במודלים המבוססים על ארכיטקטורת הטרנספורמרים. אולם בטח זוכרים שמודלי שפה שלנו גם לא מעט מודלים בדומיינים אחרים מבוססים על טרנספורמרים שמורכבים מבלוקים אחד מהם מורכב מנגנון *attention* ושתי שכבות feed-forward (השנייה מהן לנארית). בסופו יש שכבות נרמול וחיבור residual (כלומר הפלט של כל שכבה מחובר יחד עם הפלט של השכבה הקודמת).

מודלי שפה מודרניים מכילים עשרות ריבות של בלוקי טרנספורמרים שכובן משליך על כמות הזמן והמשאבים הנדרשים להפעילם, בעיקר במשימות גנרטיב. כאמור המאמר שנסקרו היום מציע דרך לקצץ כמה בלוקי טרנספורמרים רצופים שכובן יקטין את זמן חישוב שנדרש לייצירה הפלט. אבל איזה בלוקים לבחור כך שהפגיעה בדיק המודל תהיה מינימלית.

מכיוון שהගרפ החישובי של הטרנספורמר מורכב מלא מעט חיבור residual טבעי לבחור בלוקים רצופים שלא מוסיפים הרבה לפט הבלוק הנמצא לפנייהם במודול. כלומר אם הדلتא שנוטנים הבלוקים האלה זניחה אז ניתן להעיף אותם בלי פגיעה רצינית בביטויים.

האבל איך ניתן לבדוק את זה? האמת יש לא מעט דרכים לבדוק את זה ומאמיר בחר להשווות את הפלט של הבלוק עם הפלט של הבלוק ח+ (או מוחקים ח בלוקים רצופים) באמצעות מודיפיקציה קטנה של מרחק קוסין (החליפנו \cos - arcos וחילקו ב- $\sqrt{2}$ כדי לארום למדד הזה להיות בין 0 ל 1). באופן הגיוני ח בלוקים עם דמיון גבוה מאוד לבлок שקדם להם (מבחינת הפלט) נבחרים בתור מועדים טובים לקיצוץ (כלומר בлок התחלתי ומספר בלוקים לקיצוץ ח עם הדמיון הגבוה ביותר). הדמיון מחושב על ייצוג הטוקן האחרון עבור כמות DATA גדולה.

לאחר המבחן ניתן לעשות למודל פין טיין קליל ולטענת המחברים ניתן למחוק כהה על חצי שכבות טרנספורמים (במודלי שפה) בלי פגיעה רצינית בביטויים).

<https://arxiv.org/abs/2403.17887>

🚀⚡️ המאמר היום של מיק - 23.11.24 ⚡️🚀 Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study

היום אני סוקר מאמר בנושא שמאז לא נגעתי בו (בסקרים) והוא DATA TABLEAI. המאמר בוחן שאלה מרתתקת - האם מודלי שפה גדולים (LLMs) כמו GPT באמת מבינים מידע מובנה בטבלאות?

קצת: רקע

בשנים האחרונות, LLMs הפכו לכלי חשוב בעיבוד שפה טבעי. אבל בעוד שהם מצוינים (סוג של) בהבנת שפה טבעית (בצורה של טקסט), יכולתם להבין מידע בצורה של טבלאות עדין לא נחקרה לעומק זהה בדיק מה שהחוקרים מנסים לעשות במאמר המשukan

מה החוקרים עושים?

החוקרים פיתחו מודד חדש שנקרא SUC (Structural Understanding Capabilities) שבודח את היכולות של מודלים להבין מבנה של טבלאות. המודד כולל שבע שימושים שונים:

1. זיהוי גבולות טבלה
2. איתור תאים ספציפיים
3. חיפוש הפק (מיקום לערך)
4. אחזור עמודות
5. אחזור שורות
6. זיהוי גודל טבלה
7. זיהוי תאים ממוגנים

הם בדקו את GPT-3.5 ו-GPT-4 במשימות אלו תוך שימוש בפורמטים שונים של קלט (HTML, JSON, CSV) ועוד).

מה הם גילו?

התוצאות מפתיעות! הנה הנקודות העיקריות:

- HTML מתגלה כפורמט "הנוח" ביותר להציג טבלאות ל-LLMs
- המודלים הראו יכולות טובות במשימות יחסיות מורכבות (זיהוי גבולות טבלה, זיהוי תאים ממוגנים) אך נכשלו במשימות פשוטות (זיהוי גודל טבלה, אחזור שורה פשוט, חיפוש תא בודד)
- הביצועים השתפרו משמעותית עם דוגמה אחת (one-shot) לעומת אפס דוגמאות

החידוש המרכזי: Self-augmented Prompting

החוקרים פיתחו שיטה חדשה שנקראת "self-augmented prompting" שמשפרת את ביצועי המודלים. השיטה מבקשת מהמודל תחילת לזהות מידע קרייטי בטבלה (כמו טווחי ערכים) ואז משתמש במידע זהה כדי לשפר את התשובה הסופית. זה מאפשר שיפור די רציני במספר נתונים (בניצ'מארקים)

סיכום:

אני חייב להגיד שהמאמר הזה מרתק. הוא מראה שלמרות ההתקדמות העצומה ב-LLMs, יש עדיין פערים משמעותיים ביכולת שלהם להבין מידע מובנה. זה מזכיר לנו שלמרות שהמודלים האלה מרשימים, הם עדיין רחוקים מהבנה אנושית אמיתי של מבנים ויחסים בין נתונים.

החוקרים עשו עבודה לא רעה בפיתוח מדדים ושיטות שייעזרו לקהילה להמשיך לשפר את היכולות האלה. השיטה החדשה שלהם ל-prompting היא פשוטה אבל אפקטיבית, וזה בדוק מה שאנו צריכים - פתרונות פרקטיים שאפשר לישם מיד.

מילה אחרונה

אם אתם עובדים עם טבלאות ו-LLMs, המאמר הזה הוא חובה. הוא מספק תוכנות מעשיות וכליים שימושיים. הקוד והDATA זמינים ב-GitHub, אז אתם יכולים להתחיל לשחק עם זה ישירות.

معنىין במיוחד יהיה לראות איך הממצאים האלה ישפיעו על הדור הבא של מודלי שפה. האם נראה מודלים שמתוכנים במיוחד להבנת מידע מובנה?

<https://arxiv.org/abs/2305.13062>



Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

המאמר מציג ניתוח عميق של 2 שיטות מרכזיות ליישור מודלי שפה גדולים עם העדפות אנושיות: (Direct) .Proximal Policy Optimization (PPO) ו- (DPO)

1. רקע ומוטיבציה:

- קיימת סתירה מעניינת: ישומים מסחריים מצלחים כמו ChatGPT משתמשים ב-PPO, בעוד שבספרות האקדמית DPO משיג תוצאות מובילות.
- מחקר זה בודק האם DPO אכן עדיף על PPO ומה גורם לביצועים הנמוכים של PPO במדדים אקדמיים.

2. ממצאים תיאורתיים:

- DPO סובל מוגבלות מהותית הקשורת להטיה כלפי תשבות מחוץ להתפלגות הדטה (out-of-distribution) או (OOD)
- הביצועים של DPO מושפעים משמעותית מהמרקח בין התפלגות בין התפלגות ההתחלתיות של המודל לדטה המשמש לאימון RLHF (העדפות אנושיות)

3. שיפורים ב-DPO:

- החוקרים זיהו 3 גורמים קריטיים לשיפור ביצועי PPO:
- נרמול של פונקציית היתרון (Advantage Normalization) - משמש לעדכון של משקל המודל ב-DPO
 - אימון עם באז'ים גדולים
 - עדכון הדרגתית של המודל המאומן באמצעות ממוצע נועURIכי של משקל המודל מהאיטרציות עדכון הקודמות

4. תוצאות ניסיוניות:

- PPO משיג ביצועים עדיפים בכל המשימות שנבדקו
- במשימות מסוימות של יצירת קוד, PPO משיג תוצאות state-of-the-art
- מודל PPO עם 34B פרמטרים משיג שיפור של 10% בהשוואה ל-41B AlphaCode

5. מסקנות עיקריות:

- למחרת הפופולריות הגוברת של DPO, השיטה סובל מוגבלות מהותית

- עם היישום הנוכחי של הטכניקות שזוהו, PPO יכול להשיג ביצועים מצוינים
- המחבר מספק תובנות חשובות לגבי האופן שבו יש לישם PPO ביעילות

6. חשיבות המחקר:

המאמר תורם תרומה משמעותית להבנת היתרונות והחסרונות של שיטות יישור שונות, ומספק הנחיות מעשיות לישום מוצלח של PPO. התוצאות מתגראות את ההנחה הרווחת ש-DPO עדיף, ומדגישות את החשיבות של יישום נכון של PPO.

סיכום:

לסיכום, זהו מחקר חשוב המספק תובנות מעשיות ותיאורטיות חשובות לתחום יישור(alignment) של מודלי שפה גדולים עם העדפות אנושיות.

<https://arxiv.org/abs/2404.10719>

🚀⚡️:27.11.24-⚡️🚀 המאמר היום של מיק - The Illusion of State in State-Space Models

מאמר חשוב זה בוחן את המגבליות התיאורטיות של State Space Models או (SSMs), אשר צמחו כארקטקטורה חלופית לטרנספורמרים עבור מודלי שפה גדולים. המחברים מגדמים שלמרות עיצובם שנראה Recurrent ובעל מצב (clomor alstateful), למעשה SSMs (כמו טרנספורמרים) מוגבלים באופן בסיסי ביכולתם לבטא חישוב "רץ", מכיוון שאיןם יכולים לחשב דבר מהbez מחלוקת המורכבות 0 TC0. משימות מחלוקת 0 TC0 מוגדרות ככאלו שניתן לייצג עם שרירותיות בוליאניות בסיסיות (וחישובי סף ו- vote) בעומק סופי (למשל חיבור של מספרים, מכפלה או מינימום של ח' מספרים). מדובר בחלוקת hei "פשוטה" בהיררכיה של תורה סיבוכיות circuit (כלומר circuit complexity).

משמעות הדבר היא ש-SSMs אינם יכולים לפתור בעיות מסווג permutation composition ש- RNNs בעלות שכבה אחת מסוגלות לפתור.

תרומות מרכזיות של המאמר:

1. ניתוח תיאורטי:

- מוכיח שגם SSMs לנאריים וגם SSMs בסגןון Mamba מוגבלים למחלוקת חישובית 0 TC0
- מראה ש-SSMs אינם יכולים לפתור בעיות שלמות-NC1 (משימות שניתן לייצג אותן עם פעולות בוליאניות בעומק לוגריטמי ממילך הבעה - מספר משתנים בגודל) כמו הרכבת תמורות. כלומר לא עומק סופי כמו ב- TC0.
- מגדים ש-SSMs אינם יכולים לעקוב במדויק אחר מהלכי שחמט, לכתחזק קוד מורכב, או לעקוב אחר ישיות בנרטיבים.

2. בדיקות אמפיריות שבוצעו על ידי מחברים המאמר:

- מספק ראיות ניסיוניות המראות ש-SSMs בסגןון Mamba וטרנספורמרים מתקשים במשימות permutation composition.
- מראה ש-SSMs דורשים עומק גדול כדי "לטפל" ברצפים ארוכים יותר למידול פעולות קבוצה "תמורתיות"

- מדגים ש-RNNs בשכבה יחידה יכולים לפתור משימות אלו ש-SSMs אינם יכולים (כנראה בגליל לינאריות בין המעבירים של המוצבים החבויים ב-SSMs).

3. **שכלולי ארכיטקטוניות המוצעים במאמר:**

- מציע 2 דרכי לרחיב SSMs מעבר למוגבלות TC: הוספה אי-ליניאריות (RNN-SSM) והפיכת מטריצות המעבר לתלויות בקלט (WFA-SSM) - שכלול של מבנה המוסיף אי-LINEARITY למטריצה A שנותרה קבועה במ מבנה.

השפעה והשלכות של המאמר:

- מאתגר הנחות לגבי יתרונות SSMs על פני טרנספורמרים
- מצביע על גישות היברידיות פוטנציאליות המשלבות ארכיטקטורות שונות
- פותח כיוונים חדשים לפיתוח ארכיטקטורות עם יכולת ביטוי מושפרת ליישומי עיבוד שפה טבעי ובעור דומיניים נוספים
- מדגיש את חשיבות הניתוח התיאורטי של התאמת של ארכיטקטורת מודל למשימה ספציפית שהוא מתוכנן לפתור

סיכום:

מאמר תורם הן מבחינה תיאורטית והן מבחינה מעשית להבנת ארכיטקטורות של רשתות ניירונים. הניתוח התיאורטי הקפדי, בשילוב עם ראיות אמפיריות תומכות, מספק תובנות חשובות לגבי המוגבלות הבסיסיות של SSMs. בעוד חלק מההתוצאות התיאורטיות מסתמכו על הנחות תיאורתיות של מרכיבות, ההשלכות המעשיות נתמכות היטב בראיות אמפיריות.

<https://arxiv.org/abs/2404.08819>



Parameter-Efficient Fine-Tuning with Discrete Fourier Transform

רקע: PeFT

נתחיל את הסקירה ברענון קצר לגביה שיטות טיב (fine-tuning) חסכנות של מודלי שפה. PeFT הינה משפחה של שיטות המאפשרות טיב של מודלים גדולים (בפרט מודל שפה) תוך שימוש במספר מצומצם של פרמטרים, מה שcosaר משמעותית במשאבי חישוב וזיכרון.

רקע: LoRA

אחד השיטות הפופולריות ביותר ב-PeFT, הנקראת LoRA, מקפיאה את משקלות המודל ומאמנת מטריצות נוספת לככל שכבה של הטרנספורמרים. כל מטריצת תוספת נלמדת הינה בעלייה בדרגה נמוכה (rank-low), כך שניתן לייצגה על ידי מכפלה של שתי מטריצות קטנות (במידה האמצעי של המכפלה).

היתרון המרכזי של LoRA הוא שהוא מאפשר להתאים מודלים גדולים למשימות ספציפיות תוך אימון של חלק קטן (נגדי 1% מכלל הפרמטרים שלו), מה שהופך אותה לעיליה במיוחד במילוי. שיטה זו הוכחה את עצמה כאפקטיבית במיוחד בהתאמת מודלי שפה גדולים למשימות ספציפיות. בנוסף, LoRA מאפשרת החלפה מהירה בין גרסאות שונות של המודל המטובי, מכיוון שניתן לשומר את המטריצות הקטנות בנפרד מהמודל המקורי.

שיטת מוצעת:

הרענון המרכזי הוא להסתכל על שינוי המשקלות של רשות הנירונים כמו על תמונה או אות, וליצג אותם ביציר התדר במקום ערכיהם ישרים. כשאנו רוצים לטייב את המודל, במקום לשנות את כל המשקלות באופן ישיר (שדורש המון פרמטרים), אנחנו:

1. מגדירים מראש כמה נקודות דוגמה למרחב התדרים שבן נרצה להטמך. זה כמו לבחור אילו תדרים אנחנו רוצים לשמר ביצוג הדוחים שלנו. זה נעשה על ידי בחירת מטריצת תדרים קבועה (לא נלמתה) E בגודל 2×2 המשמשת לבניית יציג של מטריצת נוספת. מטריצה זו היא קבועה לכל השכבות של הטרנספורמים.
2. לומדים וקטור c בגודל ch (לכל שכבה) כאשר דרך שילובו עם E בונים את מטריצת התוספות בתחום התדר F (הסביר לאיך זה גבנה לא נראה ברור במאמר).
3. מעבירים את F דרך Gaussian bandpass filter (כלומר דוגמים בעיקר תדרים נמוכים, הנמצאים קרוב למרכז המטריצה).
4. מעבירים את מטריצת F לתחום הזמן (הריגיל) ומשתמשים בה בדיק כmo ב-LoRA

יתרונות השיטה המוצעת:

היתרון הגדול הוא שתדרים הם דרך ייעלה לייצג מידע (צריך $Ch + 2L$ משקלים כאשר L מספר השכבות במודול). בדיק כמו שאפשר לדחוס תמונה או מזיקה על ידי שימרת התדרים החשובים ביותר, כאן אנחנו יכולים לייצג שינויים מורכבים במשקלות באמצעות מספר קטן מאוד של תדרים.

זה עובד טוב(כנראה):

- שינוי במשקלות נוטים להיות "חלקיים" יחסית, ככלומר יש בהם מבנה שאפשר לתפוא טוב עם תדרים
- הבסיס המתמטי של פוריה הוא אורתוגונלי, מה שאומר שכל תדר מסוים מידע ייחודי
- אנחנו יכולים לבחור מראש כמה תדרים אנחנו רוצים לשמר, ובכך לשלוט יישורות בכמות הפרמטרים

סיכום:

בניגוד לשיטות אחרות שמנסות להקטין את כמות הפרמטרים על ידי הגבלת הדרגה של המטריצות (כמו LoRA), הגישה זו מסתכלת על הבעיה מזוויות שונה - דרך עדשת התדרים, ומצליחה להשיג דחיסה משמעותית יותר.

<https://arxiv.org/abs/2405.03003>

☞⚡️☞ **המאמר היומי של מייק - 29.11.24:** ☞⚡️☞

In-Context Learning with Long-Context Models: An In-Depth Exploration

המאמר מציג מחקר אמפירי מוקף של למידה in-context או ICL עם מודלי שפה בעלי חלון הקשר ארוך. אזכור שעם ICL המודל מקבל כמה דוגמאות המדגימות פעולות מסוימות ולאחר מכן המודל מתבקש לבצע פעולה זו על דוגמאות חדשות.

ממצאים חדשים על התנהגות של ICL ל-LLMs בעלי חלון הקשר ארוך:

1. שיפור ביצועים מתרחש: עלייה משמעותית בביצועים כאשר מעריכים את מספר הדוגמאות בהדגמה מ-10 ל-1000 דוגמאות
2. רגישות פחותה לסדר: השפעת סדר הדוגמאות יורדת ב-50% ב-1000 דוגמאות לעומת 10 (מעבר סידור אקריאי)
3. ירידת ביתרונו-RAG: היתרונו של RAG פחת משמעותית עם יותר דוגמאות
4. השפעת קיבוץ דוגמאות לפי קטגוריות: מיען דוגמאות לפי קטגוריות פוגע יותר ביצועים ככל שהחישר גדול
5. יעלות ארכי attention קצרים: ניתן להשיג ביצועים דומים עם מנגנון attention קצר יחסית המשתרע ל-50-75 דוגמאות
6. השוואה לטיב (fine-tuning): למידת context-to-ארכי חלון הקשר ארכיים לרוב משטוחה או עולה על טיב עם מעט דוגמאות אולם הטיב מנצח כאשר יש מספיק דוגמאות.

<https://arxiv.org/abs/2405.00200>

🚀⚡️: 30.11.24. 🚀⚡️

Fishing for Magikarp: Automatically detecting under-trained tokens in large language models

מאמר מעוניין מבית חברת cohere, אחת החברות שפתחות מודלי שפה foundational.

רקע:

המאמר חוקר סוגיה מעניינת של טוקנים לא מאומנים מספיק (under-trained) כלומר שלא נמצאים (או נמצאים בכמות מצערית) בדاطהסט אימון של המודל. סיבה אפשרית לקיום טוקנים כאלה נעוצה בעובדה של מיליון הטוקנים לא תמיד נבנה על בסיס הדאטהסט שהמודול מאומן עליו.

מיליון טוקנים בניי על דאטהסט קטן הרבה יותר מדאטהסט אימון העצום של המודל בשלב אימון מקדים (pretraining): הררי בנית מיליון טוקנים עם אלגוריתמים קיימים על דאטהסט של عشرות טרילيونי טוקנים איננה פיזibilית חישובי. בגדול מאד בוחרים תת-מילים "השכיחים ביותר" בדאטהסט (כולל סימני פיסוק וכדומה) לפי שיטה מסוימת (היום השיטה הפופולרית היא Byte-Pair Encoding או BPE, שיטה טוקניזציה נוספת נקראת WordPiece). והבדלים בסיס לטוקניזציה בין זה לאימון המודל עלול להוביל לייצור טוקנים מוזרים כמו .TheNitrome_

הנווכחות של טוקנים שלא אומנו מספיק במודל מובילה למספר בעיות, כולל בחזוץ קיובלות בטוקנייזר ופגיעה ביעילות המודל. בנוסף הם עלולים לגרום לפלט לא רצוי ולשבש אפליקציות downstream במירוח שעידן שבו מודלי שפה משתמשים יותר ויוצר נתונים חיצוניים. כМОון טוקנים כאלה "מזמינים" jailbreaks למיניהם. למרות שנעשה עבודה מסוימת בזיהוי טוקנים בעיתיותם אלה, עדין חסרות שיטות אוטומטיות אמינות ומוסברות היבט שנבדקו על מגוון רחב של מודלים.

פרטי מחקר:

המאמר מציע לזהוי טוקנים undertrained כאלו באמצעות טרנספורמציה מסוימת של מטריצה unembedding ו כלומר המטריצה הממפה את ייצוג הטוקן לווקטור המכיל התפלגות הסטברותית עבור כל הטוקנים במלון.

המחברים מצינים כי פונקציית הלס באימון ממודעת כאשר הסתברות של טוקנים שאינם בשימוש נחזית- C_0 , ללא קשר לקלט, מה שגורם ללוגיטים שלהם להתכנס למינוס אינסופי. המאמר משער שהמודול יכול להשיג חיזוי צוו (לא תלוי בקלט) באמצעות חיסור של וקטור קבוע c משורות של U , מה שmobiel לתמונה שלילית קבועה לערכי הלוגיטים של טוקנים שאינם בשימוש.

המחברים מציעים את האלגוריתם הבא לזהוי טוקנים undertrained:

- מגדרים קבוצה S של טוקנים חדשניים undertrained (כלומר אינדקסים של שורות ב- U)
- חשב את הרכיב העיקרי(principal component) הראשון c של U כאומדן לרכיב קבוע c . מכיוון שפונקציית הסופטמקס אינה משתנה להסטות קבועות, יש להקפיד רכיב קבוע זהה כדי למקסם את ההפרדה של טוקנים שאינם בשימוש.
- הסר אותו כדי לקבל $U(U^*T^c) - U = 0$.
- חשב את וקטור האמבדינגים הממוצע של הטוקנים שאינם בשימוש $S \in [0, 1]^n$.
- חשב את מרחקי הקווין (או מרחק L2) בין v_{000}^n לבין שאר השורות ב- U .

الطائفונים שהמראק הזה קטן יחסית לאחרים (באחוון 2 נגיד) חדשניים להיות טוקנים שאומנו מספיק. המאמר מציין הסתבריות שלالطائفונים החשודים undertrained קטנות מאוד ומראה שהן קטנות לאט מאד (בעיקר בגלל weight decay) באופן עקבי לאורך האימון (לא קשר לקלט).

<https://arxiv.org/abs/2405.05417>

🚀⚡️המאמר היומי של מיק-02.12.24- Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

ההיסטוריה:

סקירה היום עשו חזרה קטנה בזמן (mbhinneti) ואסקור מאמר בנושא הראייה הממוחשבת. פעם הייתה סוקר אותן בתדרות גבוהות יותר ואולם לאחרונה רוב המאמריםiani שמייצים לדומין הטקסטואלי כלומר NLP. לא אגלה לכם סוד אם אגיד לכם שהם מודלי דיפוזיה (לוחט לטנטים) די השתלטו על תחום גנרטט דאטה ויזואלי (כלומר תמונות ווידאו).

אולם לפני 4-3 שנים המצב בדומין הייזואלי (בחלקו הגנרטיבי) היה די שונה. היה בו גם VAE שזה AutoEncoders, גם זרימות מנורמלות Normalized Flows) אבל מי שלשלט בו באופן די מוחלט היה כМОGANs Generative Adversarial Networks. וכמוון היו שילובים די מעניינים של השיטות הניל שהגיעו לביצועים די יפים כמו VQGAN שזה שילוב של VAE ו-GAN.

רקע:

המאמר שנסקור היום מחזיר לחיים את VQGAN וטוען שנייתן להגיא לתוכאות טובות יותר אליו (עם שכולן קל) ממודלי דיפוזיה גנרטיביים באותו הגדים (= מספר פרמטרים). זו הצהרה די חזקה שמצויכה להבין מה

המחברים שכללו ב-VQGAN 7 שהוצע לפני 4 שנים.

קודם כל אסביר בקצרה איך עובד VQGAN (סקרטט) אותו בעבר הרחוק בהרחבה) אז תוכל לkipoz לשם להסבירים מפורטים יותר. בגודל VQGAN מרכיב מאנקודר שמרתתו לקוד (במרחב הלטנטי) את ה facets'ים של תמונה, המורכב ומספר גדול של וקטורים המקודדים את ה facets'ים האלו ודקודר שלמעה הופך את "יצוג facets'ים אלו" (וקטורים) ל facets'ים המרכיבים תמונה.

אחרי הקידוד של facets' על ידי האנקודר הווקטור הcy קרוב (לפי מרחק 2 ל-4) נבחר מה-codebook והוא מוזכרם לדקודר (יחד עם הוקטורים facets'ים האחרים). האנקודר וה-codebook מאמנים להציג וקטורים כמה שיטות קרובים אחד לשני (יש שם stop-gradient גם) והדקודר מאמן לשחרר את התמונה (בדקן לכל facets' בנפרד וגם יחד) בצורה המיטבית (נמדד על ידי דמיון perceptual נקרא LPIPS וגם יש לו גן בפנים עם הדיסקרימינטור).

מה המאמר עשה:

אבל איך השתמש בכל לגנרטט? לאחר סיום אימון של VQGAN, לוקחים את כל היצוגים הלטנטיים של התמונות מהדאטסט ומאמנים דקודר של הטרנספורמר לחזות יצוג של facets' בהינתן facets'ים הקודמים. ופה נכונים לנו LLM שאנו כהओhbim כי המחברים מאמנים אחד הלמות (LLAMA) למשימה זו. הרי יש לנו מיליון (codebook) כמו בשפה טבעית רק שבמוקם התוכנים הרגילים יש לנו תוכנים ויזואליים.

זה עובד לא רע (לפי הבדיקות שהם עושים)...

<https://arxiv.org/abs/2406.06525>

🚀⚡️: 04.12.24 - KAN: Kolmogorov-Arnold Networks

האמת שזה די מחדל שב 7 חודשים מאז שהמאמר הזה התפרסם, לא סקרתי אותו. יש לו כבר כ-400 ציטוטים והיד עוד נתיה. אני באופן אישי מאוד אוהב מאמרים המבוססים על טענה מתמטית מוכחת ולצער אין לנו הרבה כאלו בתקופה الأخيرة.

המאמר הדי מדבר זהה מציג ארכיטקטורה חדשה המבוססת על משפט Kolmogorov-Arnold-Shotourov פונקציה רבת משתנים רציפה ניתנת לייצוג סכום (כפול) של פונקציות של משתנה אחת. במילים פשוטות כל פונקציה ניתן לייצג בתור סכום של סכומים של פונקציות שכל אחת מהן היא של משתנה אחת בלבד.

משפט זה הוא "מקביל" ל- Universal Approximation Theorems (יש כמובן שנייתן לייצג כל פונקציה (המקיימת תנאי לא מגבלים במינוח) על ידי רשות נירונים בעלת עומק 2 או יותר שכבות. רשותות נירונים של היום בניוים בהתבסס על UAT (בגודל) ומהאמר המשוקר מציע לבנות אותו בהתבסס על משפט KA. באופן די טבעי זה קיבל שם כן.

המודול KAN בניו משכבות שכל אחד מהן סכום של פונקציות תלמידות (כלומר הפרמטרים בהם הם אלו שנלמדים על הדעתה). כל פונקציה תלמידת צוז מרכיבת מצירוף לינארי של כמה splines-s (עד פונקציה ללא פרמטרים הנדרשת x (nisi).

ב-ספלין B זה פונקציה המוגדרת באינטראול, המחולק לכמה מקטעים (נקרא grid) שמהווים פרמטרים של הביס-ספלין. B המורכב מכמה פולינומים (מדרגה 3 בד"כ)vr שלכל מקטע יש פולינום מסויל. ב-ספלין משתמשים

לקירוב של פונקציות כאשר המקדמים לפולינום בכל מקטע נקבעים כדי למקסם את דיוק הקירוב. אך ב-KAN לומדים את את פרמטרי הגריד במטרה למזער את פונקציית הלוס של הבעה.

זהה זה - היה לנו לא מעט התלהבות סיבי הארכיטקטורה החדשה זו אבל התברר שהאימון של KAN הוא לא פשוט בכלל ולא תמיד מתכנס. אבל זה לא הפרע לא לקבל 400 ציטוטים בחצי שנה עם عشرות רבות מאמריהם המשך שכנראה אסקור כמה מהם. בינהם אני לא איבדתי תקווה ב-KAN...

<https://arxiv.org/pdf/2404.19756>

🚀⚡️: 05.12.24 - Memory3: Language Modeling with Explicit Memory

א. רעיון כללי:

המאמר מציע זיכרון מפורש (explicit memory) או EM כתוספת לארכיטקטורה של מודלים לשוניים. בנוסף לאופי הסטטי של פרמטרי המודל או הזיכרון הזמן (משקל K ו-V), הזיכרון המפורש פועל כמחסן ידע מובנה ודינמי, הניתן לאחזר מוחץ למודל שפה.

זיכרון מפורש מיועד לשיפור טרייד-אוף" בין גודל של LLMs לבין ביצועיהם. באמצעות החצנת ידע פחות מופשט (כמו עובדות, נתונים, חוקים ספציפיים לתמונה) אל תוך EM, המודל מנע מהגדלה משמעותית של פרמטרי המודל, תוך שמירה או אף שיפור של ביצועים. חידוש זה לא רק משפר את הייעולות החישובית, אלא גם הופר את המערכת למודולריות. עדכוני ידע אינם מחיברים אימון חדש של כל המודל, מה שمدמה תהליך למידה אנושי שבו מידע חדש נשמר מבלי לשנות את הפונקציות הקוגניטיביות הבסיסיות.

ב. היררכיית זיכרון מוצעת

היררכיה הזיכרון שהוצעה במאמר שואבת השראה מערכות קוגניטיביות אנושיות, שבהן הזיכרון לטווח ארוך מסוג לפיה נגישות ותדרות שימוש. המחברים מעצבים מסגרת זו כדי להקטין ידע אסטרטגי ב- 3 רמות:

1. טקסט פשוט (עלויות קריאה גבוהות, עלויות כתיבה נמוכות):

- מתאים למידע שניגשים אליו באופן נדרי, אחסון טקסט פשוט שומר על קליילות המערכת הכלולית. אחזור מזכיר זה פחת יעיל אך משמש כגבי לשאלות נדירות.

2. זיכרון מפורש (עלויות מאוזנות):

- ידע הנמצא בשימוש תדיר יותר אך לא קרייטי (כמו ידע מופשט על השפה) נשמר ב-EM, המאזן בין מהירות האחזר(retrieval) לעליות האחסון. האינטגרציה שלו עם מנגןנו חסוי דليلים מבטיחה שרק חלק הזיכרון הרלוונטיים ביותר יופעל, מה שմষפר את יעילות האינפראנו.

3. פרמטרי מודל (עלויות קריאה נמוכות, עלויות כתיבה גבוהה):

- שומר לידע מופשט המהווה ליבה ליכולות האינפראנו הבסיסיות של המודל. עדכונים בשכבה זו מתבצעים באימון, מה שהופר אותם ליקרים חישוביים.

היררכיה זו מאפשרת ל-3 Memory-Layer לטעוד הקצתה משאים בצורה דינמית, ובטיחה שהעלויות החישוביות ישארו ניתנות לניהול תוך שמירה על ביצועים גבוהים. עיצוב זה רלוונטי במיוחד לשימושים הדורשים עדכוני ידע בזמן אמת, כגון מערכות תמיכת לקוחות או בוטים מותאמים לתחומים ספציפיים.

ג. ארכיטקטורה

ארכיטקטורת Memory3 היא אבולוציה ממשמעותית של מודלים סטנדרטיים מבוססי טרנספורמרים, תוך שילוב זיכרון מפורש באופן חלק.

חידושים עיקריים:

1. מנגנוני **attention** דיליליים:

- באמצעות שילוב הזיכרון המפורש במנגנון **attention**, הגישה המוצעת נמנעת מעסקילינג הריבועי של **attention** (הו בעבר טרנספורמרים שעשו משהו דומה). **attention** דليل מפחית כמות חישובים על ידי התמקדות רק בתת-קבוצות של זיכרון הרלוונטיות ביותר לשאלתה.

2. אחזר זיכרוןיעיל:

- המודל משתמש בחישוב מבוסס דמיון קוינטוס כדי לאחזר זוגות מפתח-ערך (KV) רלוונטיים. אմבטיגנס של חלק זיכרון הרלוונטיים מחושבים מראש שמאפשר אחזר מהיר וסקילבלי, וכך שמהירות האינפראנס לא נפגעת גם כשהזיכרון גדול.

3. דילול(**sparsification**) זיכרון:

- כדי לשמור על יעילות הזיכרון, המחברים מציעים טכניקות כמו **בחירה טוקנים מדורגת k**, שבהם נשמרים רק הטוקנים האינפורטטיביים ביותר. זאת בשילוב עם **קונטיזציה של וקטורים**, שמקווצת את אמבטיגנס של הזיכרון מבלי לאבד משמעותית מכוח הייצוג שלהן.

4. גמישות בעדכוני ידע:

- בניגוד לאחסון מבוסס פרמטרים, זיכרון מפורש מאפשר עדכונים מודולריים. לדוגמה, הוספה ידע חדשה כרוכה רק בהוספה זוגות KV במקום אימון מחדש של המודל, מה שהופר את Memory3 לモותאם ומתאים לעתיד.

ד. פרדיגמת האימון

המחברים מאמצים פרדיגמת אימון בשני שלבים אשר מותאמת לשילוב זיכרון מפורש:

1. שלב אימון **warm-up**:

- המודל עובר אימון בסיסי ללא EM. שלב זה מבטיח פיתוח של יכולות הפשטה חזקות והבנה לשונית בסיסית. שלב זה דומה לאימון מקדים במודלים טרנספורמרים מסורתיים.

2. שלב אימון **continual**:

- המודל לומד לכתוב ולקרוא מ-EM. מטרות האימון מתרחבות כדי לכלול משימות ספציפיות לזכרון כמו:

- **כתבת זיכרון:** אופטימיזציה של אחסון ידע בתור זוגות KV.
- **אחזר זיכרון:** שיפור יכולת לאחזר מידע רלוונטי באופן יעיל ומדויק במהלך האינפראנס.

סיכום:

שילוב EM ב-3 Memory ממחיש דרך חדשה לבניית מודלים לשוניים ייעילים, ניתנים להתקאה ומודולריים. הגישה זו עשויה (למרות שב-5 החודשים מאז יציאת המאמר לא ראייתי ניצנים לכך) להוות בסיס לדור הבא של LLMs, במיוחד בתחום הדורשים עדכניים שוטפים של ידע ו-interpretability גבוהה (בגלל שיש זיכרון מפורש).

<https://arxiv.org/abs/2407.01178>



Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering

1. תמצית המאמר

המאמר מציע שיטה המכונה RAG עם מערכת מבוססת **גרפי ידע (KG)** המותאמת לשירות לקוחות. המערכת, שפותחה על ידי צוות המחקר של ChaiLabs, משירה LLMs בידע מבני שמקורו בפניות שירות לקוחות היסטוריות. על ידי שילוב יחסים שונים בין פניות השירות (טיקטים) בגרף, השיטה משפרת באופן משמעותי את דיקט האחזוקה(retrieval), אינטואיטיבות והיעילות, עם שיפורים ניכרים במדדים כמו MRR, BLEU ומרקtein זמני הטיפול בפניות.

2. תרומות מרכזיות

a. שילוב KG במערכות RAG

- **שימוש מידע מבני:** כל טיקט מיוצג כעץ (יחסים פנימיים בתוכו) ומקושרת לפניות אחרות דרך יחסים סמנטיים או מפורשיים. עיצוב זה מסמר את ההיגיון הלוגי של הטיקט, כולל תיאור הבעיה והפתרון. כל טיקט מהווה צומת בגרף.
- **שיפור אחזוקה ויצירת תשובות:** המערכת מנוטת בגרף כדי לזהות תת-גרפים רלוונטיים, המזונים ל-LLMs לצורך יצירת תשובות אינטואיטיביות.

b. בניית גרף הידע:

1. **עץ פנימי לטיקט:** צמתים מייצגים חלקים כמו סיכון או שורשי בעיה, וקשרות מציניות יחסים היררכיים.
2. **קשרים בין פניות:**
 - קשרים מפורשיים: יחסים כמו "clone of" or "caused by" (e.g., "clone of" or "caused by").
 - קשרים סמיוטיים: מחושבים על בסיס דמיון קואסין בין אמבדינגים.

ג. שלבים בתהליך אחזוקה ותשובות

המערכת פועלת ב 3 שלבים:

- **דיהוי ישויות(entity) וכוכנות:** המערכת הופכת שאלות משתמש לשויות וכוכנות(intents) באמצעות LLMs וניתוח ומבנה YAML.
- **אחזוקה תת-גרפים:** מתבצע חישוב דמיון בין אמבדינגים לשאלתה לצמתים בגרף לזרחי תת-הגרפים הרלוונטיים ביותר.

• **יצירת תשובות:**

המערכת יוצרת תשובות בהתאם על תת-הגרפים רלוונטיים לשאלת המשמש.

4. קצת פרטיים על השיטה

השיטה המוצעת כוללת 3 שלבים עיקריים:

a. **דיהוי ישיות בשאלתה וdiamond כוונה(intent):**

- המערכת מעבדת שאלות משתמש על ידי חילוץ ישיות מוגדרות וכוונות באמצעות ניתוח תבניות YAML ו-LLMs. ישיות מוגדרות מייצגות אופיינים מהותיים (למשל, "תקציר בעיה" או "תיאור בעיר"), בעוד כוונות(intents) מכילות את מטרת השאלה (למשל, "פתרון תיקון"). לדוגמה, בהינתן השאלה "יכיזר לשחרר את בעית התחברות כאשר משתמש לא יכול להתחבר ל-LinkedIn?", המערכת מזהה את הישיות כ"בעית התחברות" ומשתמש לא יכול להתחבר" ואת הכוונה כ"פתרון תיקון".

b. **אחזור מבוסס אמבדינגן (יצוג):**

- **דיהוי פניות רלוונטיות:** מחשבים עד כמה הישיות שהולצו משאלת המשתמש (למשל, "בעית התחברות") תואמות את הצמתים ב-GK. עבור כל ישות בשאלתה, השיטה משתמשת בדמיוון קווין למדידת קרבה בין יציג הישות לייצוגים של צמתים בגרף. היצוגים מצטברים על פני כל הצמתים השייכים לטיקט מסוים. ככל שלטיקט יש מספר ישיות קרובות לשאלתה, הציון שלו עולה, מה שהופך אותו לסביר יותר להיבחר כרלוונטי.

- **חילוץ תת-גרף רלוונטי:** לאחר דיהוי טיקטים הרלוונטיים ביותר, הם משמשים לבניית שאלות למסד נתונים (DB) בשפת שאלות גרפים הנקראת Cypher. שאלות אלה מאפשרות למערכת לחץ תת-גרפים מקוшиים, כגון תיאורים קשורים או שלבים לשחרר בעיה. תהליך האחזור המבנה זהה מבטיח(סוג של כמו תמיד) שהמערכת אוספת מידע מדויק ורלוונטי מבחינה ההקשר מgraf הידע.

c. **יצירת תשובה:**

- מגנרטת תשובות על ידי קישור נתונים הגרף שאוחזרו עם השאלה המקורית. LLM מנתח מחדש את השאלה באופן דינמי ומיצר תשובות מובנות. לדוגמה השאלה "שגיאת העלאת CSV בעדכן אמייל משתמש" מנוסחת מחדש ל-Cypher לאינטראקציה עם DB, מאחר שתוצאות צעד-אחר-צעד.

6. סיכום

המאמר מציג דרך פורצת דרך לשילוב גרפי ידע במערכות RAG עבור מענה לשאלות בשירות לקוחות. על ידי לכידת יחסי פנים וחיצוניים בין פניות, המערכת משפרת משמעותית את דיוק האחזור ואיכות יצירת התשובות, ומזכירה כיוון מעניין בישומים פרקטיים של LLMs.

<https://arxiv.org/abs/2404.17723>:

המאמר היומי של מיק - 09.12.24
Scaling Synthetic Data Creation with 1,000,000,000 Personas

תמצית המאמר ותרומות מרכזיות:

1. השקת Hub Persona

- מاجر של מיליארד פרסונאות מגוונות שנוצרו באמצעות טכניקות הניטנות להרחבה
- פרסונאות אלו מגלמות ידע, תחומי עניין, התנסויות וmockups ייחודיים, המציגים כ-13% מאוכלוסיית העולם

2. יצירת דата סינטטי מבוססת פרסונאות:

- שילוב פרסונאות בפורומפטים מאפשר למודל שפה גודלים (LLMs) לייצר נתונים סינטטיים מגוונים במיוחד
- מדגים יישומים במגוון תחומים כגון בעיות מתמטיות, חשיבה לוגית, הוראות, טקסטים עתירי ידע, דמיות NPC במשחקים וממשקים כלים

3. שיטות ליצור פרסונאות:

טקסט-לפרסונה:

- מייצרת פרסונאות שירות מנתחי רשות-מנתח הקשר טקסטואלי כדי להסביר את הפרסונה שסבירה אליה (למשל, "מי עשוי לכתוב או לחתום טקסט זה?")
- מפיק תיאורי פרסונה גאים או מדוייקים (למשל, "מדען מחשב" לעומת "חוקר למידת מכונה המתמקד בארכיטקטורות ניירונים")
- מתרחב בקהלות באמצעות LLMs ומאגרי נתונים ציבוריים ענקיים

פרסונה-לפרסונה:

- מרחיב פרסונאות באמצעות קשרים יחסיים (למשל, ילד הקשור לאחות ילדים, או קבוצה הקשור לעבוד מקולט)
- משתמש בפורומפטים מבוססי יחסים כמו "מי נמצא בקשר קרוב עם פרסונה זו?"
- העשרה פרסונאות נוספת על ידי איטרציה של שלוש דרגות הפרדה

4. תהליכי הסרת כפליות פרסונאות:

- MinHash Deduplication: מסיר פרסונאות דומות על בסיס חפיפת n-gram טקסטואלית
- Deduplication מבוסס אמבדינג: מסנן פרסונאות באמצעות דמיון סמנטי (מרחב קווין) המחשב דרך אמבדינגים. ספי הדמיון הותאמו בהתאם לשיקולו' אינטואיטיבי מול כמהות
- לאחר ניקוי והסרת כפליות, המאגר כולל 1,015,863,523 פרסונאות ייחודיות

5. יישומים:

א. סינטזה בעיות מתמטיות:

- יצר 1.09 מיליון בעיות מתמטיות ייחודיות באמצעות פרסונאות
- מודל DB7 שעבר טיב (fine-tuning) עדין עם בעיות אלו השיג דיוק של 79.4% על סט בדיקה סינטטי
- תור-התפלגות %-64.9% על MATH, תוצאה המשתווה ל-w4-turbo-preview-GPT-4-turbo-previews
- המדגים יכולת הרחבבה - הוספה פרסונאות שיפריה את גיון הבעיה והבטיחה כיiso רוחב של מושגים מתמטיים

ב. בעיות חשיבה לוגית:

- סינטז בעיות לוגיות מattaגרות (למשל, חשיבה מרחבית או זמנית) המותאמות למופיעי פרסונה
- כולל בעיות בסגנון Ruozhiba שובי' לבדיקת יכולות לוגיות מעודנות

ג. יצירת הוראות:

- יצר שאלות משתמש המשקפות פרטונות מגוונות מהעולם האמייתי (למשל, כימאי עשוי לבקש מידע ניסוי; אמן עשוי לבקש טכניות צייר)
- אפשר סימולציות של שיחות רב-שלביות בין משתמש ל-LLM על ידי שרשור פרומפטים של פרטונות

ד. טקסטים עתירי ידע:

- יצר מאמרם ותוכן חינוכי המתואימים עם מומחיות הפרטונות (למשל, גנן כתוב מדריכים על צמחים עמידים לבוצרת)
- כיסה כמעט כל נושא באמצעות הרוחב של הפרטונות

ה. פיתוח كلم (פונקציות):

- חזה كلم שפרטונות עשויות להזדקק להם (למשל, נהג מונית ה Zukk ל-API של תנאי תנועה)
- יצר הגדרות كلم עם קלטים, פלטים ותלוויות בחרורים

6. תוצאות מרכזיות:

- מודלים קטנים יותר (למשל, Qwen2 7B) שעברו כוונון עדין באמצעות נתונים סינטטיים שהיגו רמות ביצועים שבדרך כלל דורשות מודלים גדולים יותר
- הוכח שגיאוןפרטונות מוביל לפלאים מגוונים ויצירתיים משמעותית יותר
- הדגים שפרטונות יכולות לדמות התנהגויות משתמש מגוונות, ולפעול בעילות כנושאות מבוזרות של זיכרון ה-LLM

7. סיכום

המאמר מסמן קפיצת מדרגה (לא ברור עד כמה משמעותית) בגנרטט דאטא סינטטי. המודולוגיה המוצעת נראה מבטיחה וניתנת לישום עבור מגוון משימות, ויצרת הזדמנויות לטיבח חכם של LLM, פיתוח ישומים, ואיפלו סימולציות חברותיות.

<https://arxiv.org/abs/2406.20094>

המאמר היומי של מיק - 10.12.24 LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement

1. מבוא ומוטיבציה

המאמר מציג את LLM2LLM, מסגרת חדשה לשיפור ביצועי LLM במצבים של מחסור בדата. בעוד שאלומן נוסף של מודלים כאלה דרוש בדרך כלל דאטא מתויג רב, מה שדורש עבור ידנית מרחבה, LLM2LLM מיציע אסטרטגיית העשרה דאטא איטרטיבית המבוססת על פרדיגמת מורה-תלמיד (student-teacher) כדי לשפר את הדאטא בעיתיות (שהמודל הקטן, תלמיד, מתקשה להתמודד איתם) באופן דינמי.

2. מודולוגיה

LLM2LLM מורכב מ-3 שלבים עיקריים:

1. אימון מודל התלמיד: מודל התלמיד מאומן על כמות דאטא קטנה.
2. זיהוי שגיאות: הביצועים נמדדים על נתוני האימון באמצעות המודל הגדל (מורה), ודוגמאות שבחן המודל הקטן שוגה מזוהות.

3. העשרה דатаה מוקדמת: מודל המורה מייצר דוגמאות סינטטיות חדשות בתור אוגננטציות שונות של הדוגמאות בהם מודל התלמיד טועה. דוגמאות אלו משתמשות מחדש במערכת לצורך איטרציות אימון נוספת.

מאפיינים מרכזיים:

- העשרה איטרטיבית: הדאטסהט לאימון מודל התלמיד משתפרות לאורך מספר סבבים במקום להיווצר מראש.
- מיקוד בטיעיות: הדגש הוא על דוגמאות מתגברות המציגות את חולשות המודל הקטן.
- המחברים מצינים כי מודל המורה אינו חייב להיות חזק יותר, אלא רק להפיק דוגמאות קונספטוואליות דומות לטיעיות של המודל הקטן.

3. תוצאות

המסגרת הוכיחה שיפורים משמעותיים במדדים במצבי מחסור בDATA תור שהיא מתעללה על שיטות העשרה אחרות. דוגמאות לביצועים:

- GSM8K (הסקה מתמטית): שיפור של 24.2% בדיק.
- CaseHOLD (הסקה משפטית): שיפור של 32.6%.
- SNIPS (זיהוי כוונות): שיפור של 32.0%.
- TREC (סיווג שאלות): שיפור של 52.6%.
- SST-2 (ניתוח רגשות): שיפור של 39.8%.

4. סיכום

LLM2LLM מציעה מסגרת להעשרה הדטה באימון LLMs במצבים של מחסור בDATA. על ידי התמקדות איטרטיבית בדוגמאות מתגברות ושימוש בשיטות פועלות בין מורה לתלמיד, היא מושגה שיפור ביצועים משמעותיים. שיטה זו מסמנת כיוון מבטיח לשיפור הייעילות והשימושיות של מודלים לשוניים בסביבות מוגבלות מסוימות.

<https://arxiv.org/pdf/2403.15042>

המאמר היום של מיק - 18.12.24 Byte Latent Transformer: Patches Scale Better Than Tokens

כמובן לא יכולתי לפספס את המאמר הזה שהתרפרס לפני כמה ימים וגרם ללא מעט תהודה בקהילה AI. המאמר מציע להחליף את הטוקנייזר הסטטי שיש בכל מודל השפה במנגנון דינامي שבונה את הטוקנים החדשים (שקיבלו שם פאצ'ים) ככלומר כזה שבונה אותם בתלות בהקשר (contextualized).

הרציונל כאן הוא די ברור הרי לפעמים יש מקרים שבהם הוא די ברור ונitin לעשות אותה כמקרה אחד את כל הטוקנים לטוקנים אחד ארוך או קצר לפישמו במאמר). ולפעמים המצב הוא ההפוך והיינו רוצים לחזות בצורה בגונוליריות קטנה יותר. וכמובן זהה בלתי אפשרי במודל שיש בהם מיליון טוקנים קבוע.

כאמור המאמר מציע להכניס דינמיות לבניית פאצ'ים (הטוקנים החדשים). איך הוא עושה זאת זה. לדאטסהט נתון המאמר מאמין מודל רדוד יחסית ברמה של Bytes (bytes) כאשר המטרה של המודל היא לחזות את הביטח הבא. ואז במודל הגדול שלו הם קובעים את גבולות הפאץ על סמך אנטרופיה של הבטים. ככלומר אם האנטרופיה של

הבית או גדולה מוסף מסויים או חוותה עלייה מעלה סף מסוים מעל האנטרופיה של הבית הבא, פותחים פאץ' חדש. אחרת ממשיכים את הפאץ' הנוכחי.

אבל איך כל הסיפור זהה עובד - כמו שאמרתי המודל הוא byte-level כלומר הוא מאמין לחזות את הבית הבא בטקסט. אבל במקומם להסתכל על הקונספט בתור מערך של טוקנים המחברים מציעים להחליף אותו בפاظים דינמיים נקבעים על סמך האנטרופיה כמו שהסבירתי קודם.

בנוסף לפاظים המאמר משתמש גם ביצוג של בטים באמצעות n -grams (n -gram) לדוגמה מ-3=n עד 8=n, מפעלים איזה פונקציית האש, סוכמים ומורמלים). את התוצאה הופכים לווקטור (המאמר לא מפרש איך רק מזכיר שיש איזו שכבה לנארית המעורבת בה) ומצין אותו למה שקרה' במאמר Encoder Multi-Headed Cross-Attention (נקרא זהה לפשטוות EMHCA).

מטרתו של EMHCA היא לשלב את "יצוגי" הפاظים עם "יצוגי" הבטים שליהם (כל פאץ' מתחשב רק ב"יצוגי" הבטים שלו ולא של האחרים). היצוג ההתחלתי של כל פאץ' מחושב c -queryook (כלומר ממוצע) של "יצוגי" הבטים שלו (מציר זה כל פאץ' הינו מערך של הבטים). ככלمر אנו בונים כהה "יצוג של כל פאץ'" המתחשב רק במקרה שיש בתוכו (internal representation).

از "יצוג הבטים ו"יצוגי" הפاظים מוזנים ל-EMHCA שזה למעשה טרנספורמר די רדו (עם מעט שכבות) שמטරות לבנות "יצוג תליי" הקשר שפاظים כתלות בבטים שלו. ככלמר גם "יצוגי" הבטים הם keys and values כאשר queries הם "יצוגי" הפاظים. כאמור מה שיצוא מטרנספורמר הרדו זהה הוא "יצוגי" הפاظים. נציין ש-EMHCA פולט גם "יצוגי" הבטים בסוף (לא הצלחתי להבין איך זה נבנה).

כל אלו מוכנסים לטרנספורמר יותר عمוק וכבד חישובית היוצר "יצוג יותר عمוק" של הפاظים. בשלב האחרון יש את ה-Local Decoder שהופך את "יצוגי" הפاظים יחד עם "יצוגי" הבטים ל"יצוגי" הבטים הסופיים מהם נחזה הבית הבא. זה גם טרנספורמר רדו אבל הפעם "יצוגי" הפاظים הם keys and values והוא "יצוגי" הבטים הם queries.

המאמר טוען לכל מני יתרונות של השיטה המוצעת כמו יכולת לחזות יותר טוקנים לעלות אינפרנס קבוצה, ומציגה דיק משופר באימון המודלים.

אוקי, חייב להגיד שהמאמר לא כתוב כזה טוב - יש דברים שלא הסבירו בצורה ברורה (למי יגידו כמוני). אני רק מוקוה שהצליחתי להבין אותו נכון....

<https://arxiv.org/abs/2412.09871>

המאמר היומי של מיק - 19.12.24:

Large Concept Models: Language Modeling in a Sentence Representation Space

מאמר שני (גם הוצג ב-NeurIPS 2024) של מטה המציג קונספט די מהפכני למודלי שפה. במאמר שסקרטרי אתמול הם הציע ליותר על הטוקנייז הסטנדרטי במודלי שפה ובמאמר שנתקור היום הם הציע ליותר על חיזוי של טוקן הבא שהתרגלנו אליו כל כך-b-LLMs.

כמו שראתם בטוח זוכרים LLM מאומנים (באימון מקדים וב-SFT) באמצעות מקסום הנראות (likelihood) של DATAHESST אימון D, ככלמר מקסום של הסתברות גנרטיב של D עם המודל המאומן. כדי לעשות את זה אנו ממקסמים (ביחס לפרמטרי מודל השפה שלנו) הסתברות של כל היפות דאטה. מכיוון שככל פיסת DATAHESST מורכב

טוקנים ניתנים לבטא אותה באמצעות חוק ביס כמכפלה של הסתברויות מותניות של טוקנים בהינתן הטוקנים הקודמים (כלומר הקונטקסט). וככה אני מגיעים לхиיזי של טוקן בהינתן הקונטקסט גם אימון וגם כМОבן באינפרנס.

המאמר מצין כי "חישבה טוקן טוקן" אלא בקונספטים כאשר אנו בונים את הדיבור שלנו (טור כדי הדיבור). המאמר מציע להטיל את הגישה זו למודל שפה כאשר קונספט מוגדר בתור משפט. ככלומר המחברים מציעים לאמן מודל לחזות את המשפט הבא במקום חיזוי טוקן הבא שאנו רגילים אליו במודלי שפה סטנדרטיים.

אבל איך נחזה משפט, הרי זה משחו דיסקרטי ובעור אורך די צנוע של המשפט מספר הערכיהם שהוא יכול להיות הינו מעירכי והואופף להיות גובה מדי כדי לבצע את החיזוי בו (כלומר סופטמקס בגודל עצום). אז המאמר מעביר אותנו למישור הרציף ומציע לאמן מודל, שקיביל שם LCM Large Concept Model או LCN Large Concept Network. המאמר בוחן כמה פונקציות לואש שהפשוטה מהם היא L_2 בין האמבדינג ground-truth לבין החיזוי (יש עוד כמה מעניינים בפרק 2.4.1 במאמר).

הדרך נוספת שהמאמר מציע לבנות את האמבדינג של המשפט הבא הוא אימון מודל דיפוזיה מותנה (רעיון יפה מאוד לטעמי) לחיזוי האמבדינג שלו.

האםבדינג נבנה על ידי מודל embedder שהוא קבוע במהלך האימון. בנוסף ל-embedder (שהוא encoder) יש לנו גם דקودר שהופך את הקונספט (האםבדינג שלו) לטקסט.

מאמר ד' יפה, כתוב די ברור רק קצת ארוך מדי לדעתו ...

<https://arxiv.org/abs/2412.08821>

20.12.24: המאמר היומי של מייק - FAN: Fourier Analysis Networks

היום סוקרים קצורות מאמר המציג שכבה ארכיטקטונית חדשה לרשותן ניירונים. שכבה זו משלבת פונקציות מחזריות כמו סינוס וקוסינוס. פונקציות מחזריות אינן חיה חדשה בטריטוריה של הרשותן; כבר ראיינו אותם במאמרי Neural radiance fields או NERF שהן משתמשים לבניית מודל 3D של אובייקטים וצונות. לuibט זכרוני היה מאמר שבנה ייצוג של תמונה באמצעות רשת המערבת אקטיבציות מחזריות.

אולם המאמר של היום מציע לבנות שכבה המכילה פונקציות מחזריות אלא מציע לשלב אותן עם פונקציות אקטיבציות קלאסיות יותר כמו סיגמוד כאשר השילוב הוא לינארי. אז השכבה בגודל בנייה מצירוף לינארי של סינוסים וקוסינוסים עם מקדים נלמדים יחד עם פונקציות אקטיבציות סטנדרטיות. השכבה זו טוביה למידול פונקציות מחזריות כאשר ביצועה על פונקציות לא מחזריות אינן ברורות (המאמר טוען שיש שיפור גם שם),

המאמר גם מציע להחליף FFN את שכבות ה-NFN בטרנספוררים וגם שכבות gating ב-LSTM (אותו סכום ממשוקל את סינוסים וקוסינוסים יחד עם הסיגמוד) ומדווח שיפור ביצועים בכמה משימות.

רעיון מעניין ...

<https://arxiv.org/abs/2410.02675>

22.12.24: המאמר היומי של מייק - Reasoning in Large Language Models: A Geometric Perspective

מאמר זה חוקר את יכולות החשיבה של LLMs מנקודת מבט גיאומטרית, תוך התמקדות בקשר בין הממד הפנימי(*intrinsic dimension*) או ID) של ייצוג הקלט לבין עצמת expressiveness של מודלים אלה. החוקרים בוחנים כיצד ארכיטקטורות טרנספורמר מחלקות את מרחב הקלט וכי怎ן חלוקה זו קשורה ליכולות ההנמקה שלן (*reasoning*). העבודה מציעה תובנות חשובות לגבי האופן שבו ארכיטקטורת המודל ואורך הקשר משפיעים על ביצועי LLM במשימות הנמקה.

רענון מרכזים:

מסגרת גיאומטרית לכוח ביוטי של מדור (אקספרוביוט)

הרעון המרכזי סובב סביב ציפוי גוף מנגנון self-attention והשפעתו על הממד הפנימי של הקלטים לשכבות MLP בתוך הטרנספורмерים (כלומר FFN). הממד הפנימי, בהקשר זה, מודד את מספר דרגות החופש האפקטיביות הנדרשות לייצוג אמצעי של הקלט.

מנגנון self-attention כgraf:

הפלט של שכבת מנגנון self-attention מຕואר כgraf, בו טוקנים הם צמתים ומקדמי attention מגדירים קשתות משוקללות. ציפוי הגוף קובעת את מספר החיבורים האפקטיביים, המשפיעים ישירות על הממד הפנימי של הייצוגים המועברים לבlok MLP.

חלוקת מרחב הקלט:

מודדים פנימיים גבוהים יותר מאפשרים לשכבות ה-MLP לחלק את מרחב הקלט לאזורים נוספים יותר. זה מאפשר למודל לבנות מיפויים מורכבים יותר ולתפס קשרים לא-lienאריים ביעילות. כתוצאה לכך, יכולת ההנמקה של ה-LLM משתפרת עם כוח הביוטי המוגבר הנובע מחלוקת אזורים אלה.

יכולות קירוב:

על ידי אפשר חלוקה עדינה יותר, מודדים פנימיים גבוהים יותר מפחיתים שגיאות קירוב, מאפשרים ל-MLP לייצג פונקציות מורכבות בדיק רב יותר. זה מתקשר ישירות למשימות הנמקה, בהן מיפויים מדויקים ותלו"י הקשר הם קרייטיים.

הסברים מעמיקים על הרעונות:

חלוקת וקירוב

החוקרים משתמשים בניתוח affine piece-wise של רשתות נוירונים עמוקות (DNNs) כדי להסביר כיצד מרחב הקלט מוחולק. הרעיון המרכזי של חלק "חלוקת וקירוב" הוא לתאר כיצד DNNs מחלקות את מרחב הקלט למספר אזורים, כל אחד נשלט על ידי כלל LINEAR ספציפי משלו.

חלוקת מרחב הקלט:

רשתות נוירונים (באמצעות פונקציות אקטיבציה בשכבותיה), מחלקות את מרחב הקלט למספר אזורים מובחנים. אזורים אלה מוגדרים על בסיס האופן שבו הנוירונים מופעלים בתגובה לנוטוי הקלט. חשבו על מרחב הקלט כמפה, והרשות יוצרת "אזורים" על מפה זו כאשר לכל אזור יש כלל ייחודי משלו.

קירוב LINEAR בתוכן כל אזור:

בתוך כל אזור זה, הרשות מתנהגת כמו פונקציה LINEAR. זה למעשה מאפשר קירוב פונקציות מורכבות יותר על ידי שימוש חלקים פשוטים אלה.

יכולת הרשות לקרב פונקציות מורכבות תלויות ביכולתה לחלק את מרחב הקלט ו"להגדיר" חוקים לכל אזור. יותר חלוקות אפשרות קיוב טוב יותר, שהוא קריטי לשימושות מורכבות כמו הנמקה. מסגרת זו עוזרת להבין כיצד רשותות משתמשות באבני בניין פשוטות (מודלים לנאריים באזוריים ספציפיים) כדי להתמודד עם בעיות מורכבות מאוד. ניסוח זה מדגיש את יכולת של DNNs לחלק באופן אדפטיבי את מרחב הקלט על בסיס DATA האימון, כאשר מספר האזוריים מתואם ישירות עם כוח הקירוב של המודל.

עבור טרנספורמרים, תורה זה ניתנת ליישום למנגנון multi head self attention או MHST שבו צפיפות האינטראקציות בין טוקנים משפיעה על החלוקה המשורית של מרחב הקלט ברמת שכבות ה-MLP שלו.

משפט מרכזי:

סכום מינקובסקי מסביר כיצד הפלטים של שכבת MHST מובנים גיאומטרית וקשרים למשג המימד הפנימי. בטרנספורמרים, MHST מפצלת את מנגנון attention למספר "ראשים", כאשר כל ראש מתמקד בהיבט ספציפי של הקלט. ראשים אלה עובדים במקביל כדי לתפוץ יחסים שונים בתוך הדטה. המשפט מראה ניתן לפרש את הפלט של MHST כשלוב של אזוריים שנוצרו על ידי כל ראש בווד. כל ראש מגדר "צורה" (טכנית, מעטפת קמורה) המבוססת על הטרנספורמציות שהוא מחייב על הקלט.

סיכום מינקובסקי:

סכום מינקובסקי הוא פעולה מתמטית המשמשת לשילוב צורות אלה. באופן אינטואיטיבי, זה אומר שהפלט של שכבת MHST הוא מרחב הכלול את כל השילובים האפשריים של פלטי הראשים הבודדים.

קשר למימד פנימי:

תוצאה זו מדגישה שהוספה ראשים נוספים או הפיכת הראשים לאקספרסיביים יותר מגדילה את ה"מדדיות" של המרחב שבו נמצאים פלטי תשותת-הלב. מדדיות מוחשבת זו משפרת את יכולת המודל ליצג יחסים מורכבים ותהליכי חשיבה. המשפט מפרט כיצד מנגנון MHST מחלק ומשלב את ההיבטים הגיאומטריים של מרחב קלט כדי להגבר את האקספרסיביות ייכולת הנמקה של מודלי טרנספורמר.

המימד האפקטיבי של סכום מינקובסקי תלוי בצפיפות גרפ attention (כלומר, מספר החיבורים הפעילים בין טוקנים). צפיפות גרפ גבוהה יותר, המושגת באמצעות יתר ראיי attention או קישוריות גבוהה יותר, מובילה לממדיות פנימית גדולה יותר של הקלט לשכבות MLP. מימד פנימי בוחן עד כמה טוב טרנספורמר יכול לתפוץ יחסים מורכבים בקלט שלו בהתאם על מספר החיבורים המשמעותיים שהוא מזהה.

מימד פנימי גבוה יותר פירושו שיותר חלקים משפיעים על טוקן. זה מוביל לייצוגים עשירים ומפורטים יותר של הקלט, המאפשרים למודל להבין טוב יותר דפוסים וייחסים מורכבים. כאשר למודל יש ממד פנימי גבוה, הוא יכול "לחולק" ביעילות את מרחב הקלט ליותר אזוריים, מה שמאפשר לו לתפוץ פרטים ודקדוקים עדינים יותר. זה קריטי למשמעות חשיבה, שכן הבנת יחסים עדינים היא מפתח.

השלכות מעשיות:

הגדלת מספר הראשי attention או קלטים ארוכים יותר עשויים להגדיל את המימד הפנימי. זה משפר את יכולת החשיבה של המודל מבליל לדרוש שניים בארכיטקטורה שלו או בתהיליך האימון. הממד הפנימי משיקף עד כמה עמוק טרנספורמר מתעסק עם הקלט שלו. ככל שהחיבורים עשירים יותר, כך המודל יכול לחשב טוב יותר ולבצע משימות מורכבות.

<https://arxiv.org/abs/2407.02678>

23.12.24: המאמר היומי של מיק -

T-FREE: Tokenizer-Free Generative LLMs via Sparse Representations for Memory-Efficient Embeddings

שוב חוזרים לנושא הטוקניזרים - מתברר שהוא יותר חם مما שחשבתי. נתקלתי במאמר המעניין שיטה נוספת לטוקניזציה המבוססת על פונקציה האש n -grams. השיטה המוצעת באה להתמודד עם גודל העצום של המילון מלאוה כל מודל שפה גדול (עשרות אלפי טוקנים לכל הפתוחות) וגם טוקנים דומים מאוד מבחינת האותיות השמוצריות אմבדינגים שונים שלא ייעיל (לטענת המחברים).

המחברים מנוסים שיטת טוקניזציה שה-encoding שלה המורכב משלבים הבאים:

- פירוק של טקסט למה שהם קוראים טוקנים כאשר ב-E-FREE-T טוקנים אלו הם בעצם מילים
- כל מילה מחולקת לסדרה של n -grams לא זרים למשל מילה hello מיצגת על ידי חמישה 3 -grams הבאים: {_o, ll, o, H, e, l}. מספר 3 -grams ביצוג זהה בדרך כלל מספר n -grams במילה שווה למספר האותיות במילה
- מקודדים כל 3 -gram עם n פונקציות האש שכל אחת מהם מקבלת n ערכים אפשריים כאשר n הינו אחד הייפר-פרמטרים של השיטה.
- כך כל מילה מקודדת על ידי n^k מספרים בין 0 ל- n כאשר k הינו אורך המילה (מספר אותיות). "יצוג המילה הוא ממווצע (ויגול)" של כל כוח ערכים האלו.
- כל ערך בין 0 ל- n מקודד על ידי וקטור נלמד כאשר n וקטורים אלו למעשה מהווים את המילון של השיטה

שלב האימון והפענוח (כלומר גנרטו של מילים) נראים קצת יותר מורכבים. קודם כל באימון המטריה היא לחזות את כוח האשים של 3 -grams של המילה הבאה. כלומר במקומות בעיית multi-class בפענוח של הטוקניזציה הרגילה (חיזוי של טוקן ממילון הטוקנים) יש לנו כאן בעית multi-label כאשר אנו חוזים n^k האשים. שימוש לב שחל תליי באורך המילה כלומר יש לנו מספר "ליילים" שונה לפי אורך המילה.

הפענוח לא ממש ברור לי האמת. כאשר אנו רוצים לחזות את המילה הבאה אנו קודמים כל מחשבים את כל האשים עבור כל המילים האפשריות (זה די הרבה כי לכל מילה יש גם את כל ההתוצאות שלה לכל הפחות ובנוסף מילים בעלות אורכים שונים מקודדים עם מספר n^k שונה של האשים). לאחר מכן בוחרים את המילה המיצגת על ידי האשים בעלי "הסתברות הגבוהה ביותר". נזכיר שהמודל חוזה הסתברות של כל ערך של האש מ- 1 עד n (גודל המילון) ולא לגמרי ברור איך נבחרת קבוצת האשים בעלת הסתברות הגבוהה ביותר.

בקיצור מאמר נחמד אבל לא ברור לי העניין עם הפענוח...

<https://arxiv.org/abs/2406.19223>

25.12.24: המאמר היומי של מיק - Vision language models are blind

מאמר נחמד הטוען שמודלי שפה ויזואליים הם די עיוורים כלומר אין להם סיכוי לעבור בדיקה אצל אופטומטריסט מוששה. הנה כמה עובדות על המבחןים הכספיים שלהם:

1. מודלי שפה ויזואליים או LMs לא יכולים לקבוע באופן אמין האם שני קווים (או שני מעגלים) נחתכים, במיוחד כשהם קרובים זה לזה. הדיק בזיהוי 0 , 1 או 2 נקודות חיתוך בין שתי פונקציות לינאריות למקוטען בעלות 2 מקטעים נע בין $47\%-85\%$. באותה משימה שני המעגלים, המודלים מתקדים טוב

יוטר (דיוק של 73-93%) אך עדין רחוק מה-100% המוצופה.

2. מודלי שפה ויזואליים יכולים לזהות בצורה מושלמת מעגל ומילה בנפרד אך כאשר המمعال המיליה נמצאת בתוך המمعال המודולים נוטים להתקשות בזיהוי איזו אות מוקפת בمعالג.
3. מודלי ראייה-שפה יכולים לספור צורות במדויק, למשל, מעגלים, ריבועים כאשר הם נפרדים ורחוקים זה מזה. עם זאת, כל המודולים מתבקשים לספור מעגלים חוטכים (כמו הלוגו האולימפי), ובאופן כללי, צורות בסיסיות שהן חופפות או מקוונות.
4. בסידור ריבועים בצורה של רשת, אנו מגלים ש-VLMs נכשלים באופן מפתיע בספירת מספר השורות או העמודות הראשית, בין אם היא ריקה או מכילה טקסט. זה מפתיע בהתחשב בכך שהמודולים מתפקידים כל כך טוב (דיוק ≤ 90%) על הדאטסהט ב-DocVQA הכלול שאלות רבות עם טבלאות (אוברפריט כנראה).
5. כאשר המודול מתבקש לעקוב אחר מסלולים צבעוניים במפת רכبت תחתית של עד 8 מסלולים וכך הכל 4 תחנות, VLMs לעתים קרובות נכשלים בזיהוי היקן מסלול מסוים, כלומר, ומפיגנים דיוק של עד 23% עד 50%.
6. המודל So-4-GPT עולה בבחירה על Pro-Gemini-1.5 Claude-3 Sonnet-Gemini-1.5 ב-7 בנצח'מרקם מורכבים עבור VLMs אך מתפקיד באופן משמעותי פחות טוב במשימות הנבחנות במאמר, שבהן Gemini-1.5 Pro-Sonnet-3.5. המאמר מגלה מגבלות מפתיעות של מודלי ראייה-שפה שלא נמדדנו בנצח'מרקם רגילים.

בקיצור אובי VLMs האלו צריכים משקפיים...

<https://arxiv.org/abs/2407.06581>

26.12.24: המאמר היומי של מייק -

RL for Consistency Models: Faster Reward Guided Text-to-Image Generation

זמן לא סקרתי מאמרים על מודלי דיפוזיה אז אחרי שנטקלו ביאמר הנחמד המשלב מודלי דיפוזיה גנרטיביים עם למידה עם חיזוקים (Reinforcement Learning) או RL בקצרה), לא היו לי ספקות שזה הולך להיות המאמר המסורק. כאמור המאמר פיתח שיטת אימון מודל של דיפוזיה גנרטיבי מסווג CM או Consistency Model.

קודם כל נשאלת השאלה למה צריך לאמן מודלי דיפוזיה גנרטיביים עם שיטות הלקוחות מעולם RL. הרעיון הוא שיטות סטנדרטיות יותר לאימון של מודלי דיפוזיה שהצליחו להביא לנו מודלים בעלי ביצועים מרשים (בגנרטוט תמנויות מטיקסט). אתם בטח ידעים שאימון מודלי דיפוזיה לגנרטוט תמנויות זה דבר לא זול ודורש לא מעט זמן ושימוש RL לאימון (או tune-fine) של מודלי דיפוזיה יכול להחסור לנו זמן במקרים שאנו צריכים לאמן מודל דיפוזיה ייעודי (למשל לדומין נישתי)

אחת הדוגמאות למשימה זו היא אימון מודל לייצרת תמונות מפורמת (תיאור טקסטואלי) כאשר יש בידינו פונקציה המשערת את התאמת התמונה לפורמת. אתם כבר יכולים לנחש שפונקציה זו תשרת לנו בתור פונקציית תגמול (reward function).

כבר הזכרתי שהמאמר משלב שיטה חדשה ('יחסית') לאימון מודלי דיפוזיה הנקראת CM ושיטה זו (שהומצאה על ידי איליה סלזקי ושות') מאפשרת גנרטוט יותר מהיר של מודלי דיפוזיה גנרטיביים. بغدادו מאוד שיטה זו מנסה

לאמן מודל שאוכף עקביות בין התמונות המשוחזרות על ידי המודל מתמונות מורעשות עם עצמות שונות רעש. כמו כן לוקחים תמונה, מריעשים אותה עם רעש (בד"כ גausi) עם שוניות שונות ומאמנים מודל להחזיר את אותה התמונה הנקיה (עקביות לשמה).

למה השיטה זו מאפשרת גנרטו יוטר מהיר של תמונות? כי בגלל היא מאפשרת לאנרט תמונה נקייה מרעש באיטרציה אחת בלבד (ככה המודל מאמן). במציאות עושים את זה בכמה איטרציות (מספר קטן). מתחילה מרעש, מגנרטים את התמונה ממנה, מוסיפים פ煦ות רעש לתמונה המוגנרטת, מגנרטים מהתמונה המורעתהשוב וממשיכים ככה כמה איטרציות (עשרות בודדות). זה אפשר לזרע את תהליך הגנרטו כי מודלי דיפוזיה סטנדרטיים צריכים מאות איטרציות בד"כ.

אוקי, אחרי הקדמה ארוכה נעבור לתיאור של מה שעשו במאמר. המחברים הגדרו *Markov Decision Process* או MDP המתאר תהליכי גנרטו של תמונה (או כל DATA אחר למשה). כאמור פונקציה תגמול ניתנתה לנו והיא מודדת מידת התאמת של התמונה המוגנרטת לפורומפט. המאמר מגדיר:

- המצב t בסטור שלישיה התמונה מוגנרטת באיטרציה t , עצמת הרעש והפרומפט c
- הפעולה t היא התמונה באיטרציה $t+1$
- הפליסי היא זו פונקציית התפלגות מותנית של תמונה מאיטרציה $t+1$ בהינתן התמונה המוגנרטת מאיטרציה t בתוספת רעש
- המצב המתחלתי הוא רעש גausi סטנדרטי ופונקציית תגמול נתונה לנו

אחרי שהגדרכנו את ה-MDP של תהליכי גנרטו התמונה אנו יכולים להשתמש בשיטה DPO או Preference Optimization להימון פונקציית עקביות (= המודל שאנו מאמנים). למעשה DPO מאמן מודל המקסם את פונקציית התגמול תוך כדי הגבלת של גודל עדכון פרמטרי המודל בכל איטרציה (הומצא על ג'ו שולמן ה-CTO של OpenAI לשעבר).

המאמר גם טוען שאימון כזה הוא חסכוני מבחינה משאבי החישוב הנדרשים ויעיל מבחינה הדטה (כלומר יכול לעמוד לדאטסטים קטנים).

<https://arxiv.org/abs/2404.03673>

המאמר היומי של מיק - 27.12.24

Position: Future Directions in the Theory of Graph Machine Learning

דו"ח זה(כן כן, זה דוח למורות שהוא פורסם בארכיב) טוען כי בעוד שרשתות ניירונים גרפיות (GNNs) זכו להצלחה משמעותית במספר MERCHANTABILITY, ההבנה התיאורטית שלנו לגבייהן נשארת חלקית ומונתקת במידה מה מישומים מעשיים. החוקרים מזהים שלושה תחומי מרכזיים הדורשים חקירה תיאורטית עמוקה יותר:

1. יכולת ביטוי(expressiveness) - אילו דפוסים, פונקציות ומבנים יכולות GNNs לייצג בפועל?
2. הכללה(generalization) - עד כמה טוב GNNs מיישמות את הלמידה שלhn על גרפים חדשים שלא רואו?
3. אופטימיזציה - כיצד דינמיות האימון משפיעה על ביצועי GNN?

נקודות מפתח בנושא כשר ביטוי של GNN המוזכרות במאמר:

מגבילות נוכחות:

- רוב העבודה התייאורטית מתמקדת בשאלות ביןאריות (האם GNN יכולה לבדוק בין שני גרפים?)
- במקומות מסוימים (עד כמה שונים שני גרפים?)
- הניתוחים מוגבלים לרוב לארכיטקטורות GNN טיפוסיות ואינם מתחשבים בווריאציות של GNN
- בשימושם המהולם האמתי
- התוצאות אינן מתחשבות במאפייני צמחיים/קשאות רציפים הנפוצים ביישומים אמתיים

כיוונים מוצעים:

- פיתוח מדדים למדידת דמיון בין גרפים המתאימים עם האופן שבו GNNs מעבדות אותם
- חקירת השפעת הבחירה הארכיטקטונית (כמו פונקציות אקטיבציה ונורמליזציה) על כושר הביטוי
- יצירת תוצאות איחודות שעובדות על גרפים בגודלים שונים
- התמקדות בסוגי גרפים רלוונטיים מעשי (כמו גרפים מולקולריים)

תובנות לגבי יכולות הכללה של GNNs:

המצב הנוכחי:

- החסמים התייאורטיים הקיימים לרוב מרכיבים (לבדקה) או קשיחים מדי מכדי להיות מעשיים
- הניתוח בדרך כלל מתעלם מבנה הגרף ותהליכי האופטימיזציה
- התוצאות אינן מסבירות מדוע GNNs מרכיבות יותר לעיתים מצלילות טוב יותר

מחקר נדרש:

- הבנת השפעת מבנה הגרף על הכללה
- ניתוח ביצועים על דאתה *loss-of-distribution-loss* (במיוחד על גרפים גדולים יותר)
- פיתוח טכניקות העשרה דאתה (אגמננטציה) טובות יותר עבור גרפים
- חקירת השפעת הבחירה הארכיטקטונית על יכולת הכללה של GNN

אתגרי אופטימיזציה של GNNs:

סוגיות מרכזיות:

- הבנה מוגבלת של אופן שבו מורד הגרדי-אנט(gradient descent) עובד עבור GNNs (gradient descent) לא ברור מדויק בחירות ארכיטקטוניות מסוימות (כמו נורמליזציה) עוזרות או פוגעות בתהליכי אופטימיזציה של GNN
- לעיתים GNN עם פרמטרים אקריאים עובדים טוב מ-GNN מאומן

כיווני מחקר:

- חקירת תכונות התכנסות עם פונקציות אקטיבציה תואמות יותר לביעות ספציפיות (כמו למידה מבנה של מולקולות)
- הבנת השפעת מבנה הגרף על אופטימיזציה
- מחקר מתמטי עמוק המנסה להסביר מדוע GNNs عمוקות יותר קשות לאימון (יש כמה מאמרים המדברים על *over-smoothing* בהקשר זהה אבל אנו עדין רחוקים מהבנה מלאה של מה שקרה שם)
- ניתוח תפקיד טכניקות הנורמליזציה

השלכות מעשיות

החוקרים מדגישים שהתקדמות תיאורטיות צרכות להתחבר לצרכים מעשיים:

- פיתוח נקודות ייחוס סטנדרטיות ופרוטוקולי הערכה של GNNs
- יצירת מימושים עליים של ארכיטקטורות מבוססות תאריה
- אינטגרציה עם טכנולוגיות AI מפותחות כמו מודלי שפה גדולים

חשיבות המאמר:

1. מזהה פערים קריטיים בין תיאוריה ופרקтика במחקר NNG
2. מספק מפתח דרכים למחקר תיאורטי עתידי עשויי לשפר יישומים מעשיים
3. מדגיש את הצורך לשקל את כל שלושת היבטים (כשר ביטוי, הכללה, אופטימיזציה) יחד
4. "קורא" בהגשת התקדמות תיאורטיות למשימות בפועל

עבור קוראים עם ידע בסיסי ב-NNG, מאמר זה מדגיש מדוע הינה תיאורית חשובה וכיitzד תיאוריה טובת יותר יכולה להוביל ליישומים מעשיים עליים יותר. בעוד שחלק מהפתרונות הטכניים עשויים להיות מורכבים, המסר המركזי לגבי הצורך במסגרות תיאורטיות ומעשיות ומקיפות יותר הוא ברור וחשוב.

<https://arxiv.org/abs/2402.02287>

המאמר היומי של מיק - 30.12.24 Graph Diffusion Policy Optimization

לפני יומיים סקרתי מאמר על מודלי דיפוזיה המאמנים באמצעות שיטות מעולם למידה עם חיזוקים או RL, אטמול סקרתי מאמר על רשותות נוירונים על גרפים והם החליטי לסקור מאמר שמאחד את 3 הדברים האלה (כמעט). המאמר המושאrk היום מציע שיטה לאימון מודל המגנרט גרפים באמצעות מודלי דיפוזיה המאמנים עם שיטות RL (נכון אין כאן GNN בנסיבות הטהורה אבל לפחות יש גרפים...).

קודם כל אנו צריכים להבין איך ניתן למングל מודלי דיפוזיה לגראוט גרפים. האמת זה די פשוט ודומה לגראוט תמונות. אתם זוכרים מודלי דיפוזיה מאומנים לגראוט תמונה מרעיש טהור (בדי'כ) על ידי הורדה הדרגתית של הקומפוננטה הרועשת שלו עד להפיכתו לפיסת דאטה המפלגות לפי ההתפלגות של דאטהסט אימון. זה ממש בגודל ויש גישות חדשות יותר שעשוות את זה טיפה אחרת למשל כמו Consistency Models שדיברנו עליהם באחת הסקרים הקודמות.

אם אנחנו יכולים לעשות משהו דומה עם גרפים? מתברר שכן. אנו יכולים להתחיל מלדגם גרף באקרה (כלומר הצמתים והקשרות שלו) ולאמן מודל לשנות את הערכים בצמתים ובקשרות כך שהגרף יהיה "דומה" לאחד הגרפים מדאטהסט האימון וגם יקבל ערך גבוה לפ' איזה פונקציית תגמול(המאמר גם על RL, זוכרים). ד"א, יש כאן הנחה סמויה שצומת יכול לקבל מספר סופי של ערכים (נגיד מ 0 עד n) וכל קשת יכולה להיות מכמה סוגים (כלומר מ- 0 עד n). ככלומר ההתפלגות שאנו דוגמים מהם הם קטגוריאליות וזה שונה ממה שאנו רגילים לאות במודלי דיפוזיה גראטיביים עבור התמונות.

כמובן מיד עלות כמה שאלות בנוגע לתהילך זהה?

- איך דוגמים גרפ בAKERAI במהלך האינפראנס (זה נושא עתיק ונחקר רבות על ידי מתמטיקאים ובפרט על ידי ארדוש, המאמר לא מתעמק בהזה יותר מדי). דרך אגב במהלך האימון אנו לוקחים גרפ מהדעתה ומראים אותו עלי ידי "שינויים אקראיים" בעברית הצמתים ובסוגי הקשתות
- איך משווים גרפים, כלומר איך מבינים שגרף שקיבלנו במהלך הגרוט הוא דומה לgraf מהדעתה? יש מספר רב גישות להשוות גרפים על ידי השוואת של התת-גרפים שלהם או להשוות את הלפלסיאן שלהם למשל.
- בחירה של פונקציה reward בדמיהו הגרפים לא טריוויאלית בכלל. למשל למשימות גנרט גרפים למולקולות חדשות אחד המדדים לאיכות הגרף המוגנרט הוא חדשנותו יחסית לדברים הקיימים, לעומתו בטיפול במהלך מסויימות או פיזיביליות של סינטוזו (synthetic accessibility). ניתן לבחור reward גם בתור פונקציה דמיון לגרפים הקיימים.

אוקי, אז יש לנו פונקציה להשואות הגרפים C ופונקציה תגמול לשערוך איכות הגרף z - איך אנו מאמנים מודל RL for Consistency. האמת בצורה די דומה לזה שתיארתי בסקירת של לפני 3 ימים של המאמר: Models: Faster Reward Guided Text-to-Image Generation

קודם כל אנו צריכים להגיד את Markov Decision Process עבור אימון מודל דיפוזיה על גרפים. ותבהיר שהוא ממש דומה למאמר שהזכרתי:

- המצב t_s בתור זוג של גרפ מגנרט באיטרציה $t-T$ וגם ערך $t-T$
- הפעולה t_a היא הגרף באיטרציה $t-1-T$
- הפוליסי (הסתברות של t_a בהינתן t_s) היא זו פונקציה התפלגות מותנית של גרפ מאיטרציה $t-1-T$ בהינתן גרפ באיטרציה $t-T$
- המצב ההתחלתי הוא גרפ אקראי באיטרציה T ופונקציה תגמול z שנთונה לנו המוחשבת על הגרף הסופי באיטרציה 0

המאמר מציע שתי שיטות לאימון של מודל דיפוזיה לגנרט גרפים: הראונה היא REINFORCE הקלאסי שהוא למעשה שיטת policy gradient הממקסת פוליסים בעלי תגמול גבוה. מעשה אנו דוגמים K איטרציה בין 1 ל T ומקסמים מכפלה ממוצעת (על K דוגמות) של פונקציית הפוליסו (פונקציית התפלגות מותנית של גרפ מאיטרציה $t-1-T$ בהינתן גרפ באיטרציה $t-T$) והtagmol עבור הגרף המוגנרט (באיטרציה 0).

השיטה השנייה המוצעת היא Policy Optimization-causal במקומ למקסם את הפוליסי בצורתו הטהורה אנו ממקסמים הסתברות גנרט גרפ G_0 מהדעתה (שאותו מראים והמודל "מסיר" ממנו את הרעש) מוכפלת בתגמול עבור הגרף הנוצר. גם כאן יש מיצוע על K איטרציות שמהם נבנה שערוך של G_0 .

זהו זה - סקירה קצרה, מקווה שהצלחתם להבין משהו ממנה...

<https://arxiv.org/abs/2402.16302>

המאמר היומי של מיק - 01.01.25 : Inference-Aware Fine-Tuning for Best-of-N Sampling in Large Language Models

מתחילים את השנה החדשה עם סקירה של מאמר די מעיין שמציע שיטה לשיפור אימון של מודלי שפה. היתרונו הגדל של השיטה היא מאפשרת להתאים את האימון לאופן ההיסק (אינפראנס) וכי ברור שם אכן עושים זאת בהצלחה זה אמרו להניב איות היסק.قولמר אם אנו משתמשים בגישה מסוימת במהלך האינפראנס: למשל לבחור את "התשובה הטובה ביותר ביותר" מבין N תשובות המודל (המאמר מפנה שיטות רק לגישה זו וקורא לה BoN) או תיקון עצמי (self-correction) אך כדי לנו להתאים את האינפראנס לכך.

קודם כל המאמר מנסה שתי פונקציות יעד לאימון או IA בקצרה, אחת ל [SFT](#) וגם ל [RLHF](#), שקייבלו שמות IA-SFT ו-IA-RL-IA בהתאמה. עבור IA-SFT-IA אנו מוקסמים את נראות של תשובות המומחים לשאלות מהדעתה שיש לנו בהינתן פוליסי האינפרנס | (זהה למשה Bo). למעשה מאפטמים את הפוליסי (זהה מגננון חיזוי של LLM או בפשרות LLM עצמו) כדי לעשות את Bo בזורה הטובה ביותר על הדעתה שיש לנו. עבור IA-RL-IA המטרה היא לאפטם את הפוליסי (זהה LLM כאמור) תחת טכניקה של אינפרנס | (כלומר Bo) כך שהיא ימקסם את פונקציית תגמול R.

לאחר מכן המאמר מגדר באופן מודיעיק מה זה Bo (נוסחה 1) כאשר המטרה היא למקסם את איכות התשובות של המודל כאשר אנו בוחרים את התשובה לפי מה שנקרא verifier score (= ציון לאיכות התשובה). דרך אגב מתווסףים כאן שני פרמטרים נוספים שהם מספר התשובות שמנה בוחרים את התשובה הכי טובה וטמפרטורת T של מודל השפה. באופן אינטואיטיבי ככל ש T גבוהה יותר (יותר רנדומליות ויצירתיות התשובות) מספר התשובות N צריך לעלות.

כאן יש לנו כאן הטריד-אוף הקלאסי בין exploration ל-exploitation. ככל ש- T גבוהה יותר אנו מבצעים יותר exploration (תשובות מגוונות יותר) ואילו T קטן יותר מאפשר לנו "להנות" ממה שלמדו עד עכšíי (בחירה של N משפייע על הטריד-אוף באופן הופך מ-T).

אוקי', אבל עדין בבעית אופטימיזציה של SFT-IA יש לנו את argmax (משתמשים בו לבחירה של התשובה הטובה ביותר) וזה מאד מקשה על פתרונה למרות שיש לנו שיטות שעורוך argmax באמצעות softmax ווגם softmax Gumbel עדין שיטות אלו אינן מדויקות וגם כבדות חישובית (לטענת המאמר). אז המחברים משתמשים בדרך כלל מפורסת ב-ML - קירוב של פונקציית יעד עם קירוב וריאצ'וני שהופך אותה (את הלוג שלה) לסקום של הפוליסי (הסתברות של תשובה ע בהינתן שאלה X עם המודל) ושל איבר רגולרייזציה הנקרא inference aware. איבר זה הוא למעשה win-rate של תשובה ע על שאלה X על המודל הנוכחי כאשר הערך של כל זוג (ע, X) מחושב עם z verifier score (עם קבוע נרמול).

ב- IA-RL הסיפור קצר מסתבר והמחברים משתמשים בתוצאה מאחד המאמרים של ג'ו שולמן (то что для этого нужно знать о том, что в алгоритме есть проблема с оптимизацией, и что это может привести к проблемам с сохранением информации о прошлом) עם סרגיי לוין ופתר אבל האגדים כדי לקבל שעורך לגדיאנט שמקבל צורה דומה לאלגוריתם הישן והידע שעוזר REINFORCE לומור המכפלת של הלוג של הפוליסי עם פונקציית תגמול ("ממורצת") עם התחלפת של פונקציית תגמול להקטנת השונות. המאמר גם דן במקרים מעוניינים של אופטימיזציה של פונקציית היעד של IA-RL-IA לכמה צורות של z verifier score (למשל בינארי).

מאמר ד' כבד מתמטית ניסית (לפי מיטב יכולתי) להנגיש לכם אותו טיפה...

<https://arxiv.org/abs/2412.15287>

המאמר היומי של מיק - 02.01.25 Loss of plasticity in deep continual learning

היום סוקרים קצרות מאמר ד' קליל מ-nature.

מבוא:

שיטות למידה عمוקה סטנדרטיות מציגות ירידת הדרגתית ביכולתן ללמידה חדשות בזורה מתמשכת ("מוסיפים" למודל משימה בזורה הדרגתית). בנויגוד לשכחה קטסטרופלית(catastrophic forgetting), שבה ידע קודם אובד, אובדן פלסטיות מגביל את יכולת הרשות ללמידה חדשות ביעילות.

ניסויים מkipim על דאטסהטים כמו ImageNet ו-CIFAR-100, כמו גם תרחישי למידה עם חיזוקים (Reinforcement Learning), חשו שהנירונים הופכים לדומים (לא משתנות בכל הדוגמאות) או מתמחות יתר על המידה על משימה ספציפית, מה שמחית את יכולת להסתגל לדאטה חדש. לאחר זמן, רשתות החווות למידה מתמחות מתפקידות לא טוב יותר ממודלים דומים (לינאריים), מה שմגדיל מגבלה בסיסית של שיטות מבוססות מورد הגרדיינט (gradient descent) למידה מתמחת (ואנו מאמנים מודלים עם GD היום)....

מודד הגרדיינט למידה מתמחת:

שיטות למידה מתמחת מנוטות להתמודד עם אובדן פלسطיות על ידי אתחול חדש של נירונים דומים (כאלו שלא "נדלקים כמעט אף פעם) ואימונם מחדש עם מورد הגרדיינט. ככה גישה זו מנשה "ליקור" על נירונים שילמדו משימה חדשה בלי להיעל על למשימות מסוימות, זהה שמאפשר לה ללמידה חדשות ללא הידדרות משמעותית בביטויים.

בניגוד לשיטות קובנציאליות המסתמכות אך ורק על מודד הגרדיינט, GD למידה מתמחת מתאפיין בעדכו הדרגתית סטיטים שונים של משקל המודול בדומה למה שקרהמערכות למידה ביולוגיות.

שיטות אימון נוספת:

כאמור אובדן פלسطיות קשור לאופטימיזציה יתר (לטענת המאמר) של משקלות והופעת נירונים דומים בראשת. נירונים אלו אלה או מפסיקים לתורם למידה (עבור אקטיביצית ReLU) או ננסות למצוות רווייה (מגיעה ל 0 או 1 עבור סיגמאיד). טכניקות כמו רגולריזציה L2 מפחיתות את גודילת משקל המודול ושומרות על "פלسطיות" (גמישות למשימות חדשות) במידה מסוימת. למשל שיטתurb Shrink and Perturb, המשלב רגולריזציה עם שינויים אקראיים קטנים במשקלות, מפחית את תפעת הנירונים הרדים וכך מגדיל את יכולת למידה של המודול.

אתגרי למידה מתמחת-RL

למידה מתמחת היא חינונית גם ל-RL אפילו יותר מאשר בבלמידה מפוקחת. לא רק שהסיבה יכולה להשתנות, אלא גם ההתנהגות של הסוקן הלומד יכולה להשתנות, ובכך להשפיע על המידע שהוא מקבל גם אם הסיבה נשארת קבועה. מסיבה זו, הצורך בלמידה מתמחת הוא לעיתים קרובות יותר ברור בלמידה עם חיזוקים, RL היא סבירה חשובה להדגמת הנטייה של למידה عمוקה לאובדן פלسطיות. והמאמר בוחן שימוש בשיטות שדנו בהם קודם למשימות של RL יחד עם PPO, האלגוריתם המפורסם לאופטימיזציה-RL

<https://doi.org/10.1038/s41586-024-07711-7>

:03.01.25 - מיק - המאמר הימי של

A PERCOLATION MODEL OF EMERGENCE: ANALYZING TRANSFORMERS TRAINED ON A FORMAL LANGUAGE

מבוא:

רשתות נירונים מודרניות, במיוחד מודלי שפה גדולים, מציגות מגוון רחב של יכולות, המאפשרות להן לשמש כמערכות בסיס למגוון ישומים. מאמר זה מציע הגדירה פונומנולוגית של ארגנטניות בהקשר של רשתות נירונים, תוך התמקדות באופן שבו מבנים ותהליכי ספיציפיים המונחים בסיס תהילך יצירת דאטה יכולים להוביל לשיפורים פתאומיים בביטויים במשימות ממוקדות יותר.

מושג חשוב:

הפונומנולוגיה היא גישה פילוסופית המתמקדת בחקר מבני התודעה (consciousness) כפי שהם נחוים מנוקדת

הGBT של האדם. היא שואפת לתאר תופעות או הופעת הדברים כפי שהן נתפסות על ידי בני אדם, ללא הנחות מוקדמות או הטויות תיאורטיות. שיטה זו מדגישה את הבנת החוויות כפי שהן נחיות, במטרה לחשוף את המשמעות הטבועות בהן

יכולות ארגנטיות (emergent capabilities) ברשותנו:

החוקרים מגדרים ארגנטיות ברשותנו נירוניים כרכישת מבנים ספציפיים הגורמים לצמיחה פתאומית בביטויים במשמעות ספציפיות. הם חוקרים זאת אמפירית באמצעות מערכת ניסויית המבוססת על שפה פורמלית תלויות-הקשר, ומגדירים שטרנספורמרים שאומנו על מחרוזות משפה זו מציגים יכולות ארגנטיות. ברגע שהמודל לומד את הדקדוק והמבנה הבסיסי, הביצועים במשימות הקשורות משתפים ממשמעותית.

הגדרת השפה הפורמלית:

המערכת הניסوية שהוצאה במאמר משתמש בדקדוק חופשי-הקשר הסטברובי (PCFG) להגדרת שפה פורמלית תלויות-הקשר. הדקדוק כולל:

סימולים סופיים (terminal symbols): חלקו של דבר הכללים נושאים, מושאים, פעלים, תארים, פעילים, מילות חיבור ומלות יחס.

סימולים לא-סופיים: סמלים המגדירים את מבנה המשפטים.

חוק יצירה טקסט: חוקים המכטיבים כיצד ניתן לשלב סמלים סופיים ולא-סופיים ליצור משפטיים תקפים.

המודל מאמין על משימות כמו יצירה חופשית, פתרון בלבול וייצור מותנה, כאשר מידי הביצועים נעהם לאורך תהליך האימון.

משימות פרוטוקולי הערכת ביצועי מודלים:

1. **יצירה חופשית של טקסט:** המודל מייצר משפטיים העומדים בחוקים הדקדוקיים.
2. **תיקון טקסט לא תקין:** המודל מסדר מחדש מחרוזת מבולבלת של מילים יצירה משפטיים תקפים.
3. **יצירה מותנית:** המודל יוצר משפטיים על בסיס ישות או תכונות נתונות.

ההערכה מתבצעת לפי המדרדים כוללים בדיקות דקדוקיות, בדיקות טיפוס, דיקת התאמה מדויקת, דיקת פרטוקן ועוד, המספקים הערכה מקיפה של יכולות המודל.

תוצאות: דינמיקה הלמידה

התוצאות מגלות 3 שלבים מובחנים בדינמיקת הלמידה של המודל:

1. **שלב ראשוני:** המודל לומד מבנים דקדוקיים בסיסיים עם שיפור מינימלי בביטויים.
2. **"שינוי פaza":** מתרחשת עלייה פתאומית בביטויים ברגע שהמודל מתחילה "להבין את אילוצי שפה" פשוטים יחסית.
3. **שלב ההכללה:** המודל מגדים ביצועים משופרים במשימות, המעידים על מעבר משינון להכללה.

יכולות ארגנטיות של מודלים:

החוקרים מבחןים שככל שמודל השפה לומד את הדקדוק ואילוצי הטיפוס, נצפים שיפור ביצועים ממשמעותיים במגוון משימות, במיוחד בפתרון בלבול וייצור מותנה. הנוכחות של מבנים ספציפיים מאפשרת למודל לבנות "שילובים מורכבים ותקינים" של ישות ותכונות, המובילים ליכולות ארגנטיות בתחום השפה.

נקודות מעבר בלמידה:

המאמר דין באופן שבו הופעת יכולות האמרגןטיות קשורה למספר התכונות התיאוריות שהמודל למד. נקודת המעבר, שבה מתרחשים שיפורים ביצועים משמעותיים, קשורה לסקילינג של תכונות תיאוריות. קביעה זו מאפשרת לחזות מתי יכולות יופיעו ככל שהמודל ממשיר ללמידה.

מסקנה:

מחקר זה תורם להבנת האמרגןטיות ברשותות ניירונים על ידי יצירת מסגרת המגדירה ומאפיינת תכונות אמרגןטיות על בסיס רכישת מבנים בסיסיים על ידי המודל. הממצאים מצביעים על כך שאילוצים דקדוקיים ואילוצי שפה אחרים משמשים כגורמים חשובים בחיזוי התפתחות יכולות במודלים של שפה.

<https://arxiv.org/abs/2408.12578>

06.01.25 המאמר היומי של מייק - A Survey on Efficient Inference for Large Language Models

המאמר מספק סקירה מקיפה של שיטות ליעול היסק (אינפרנס) ב-LLMs. אך יאללה בוואו נסקור את הסקירה.

אתגרים מרכזיים:

1. **גודל המודל:** מודלי שפה גדולים (ענקים הכוונה) דורשים משאבי חישוב וזיכרון משמעותיים.
2. **סיבוכיות ריבועית** (למרות שיש לא מעט שכלולים כמו FlashAttention) של מנגןון ה-**attention**: מרכיבות זו (ביחס לאורך הקלט) משפיעה משמעותית על קצב היסק(**throughput** ו- **latency**) וצריכת הזיכרון.
3. **פענוח אוטורגסיבי:** יצירת טוקנים אחד אחרי השני לא מנצלת באופן מיטבי את משאבי החישוב (כמו GPU העומדים לרשותנו ופוגעת בתפקוד המודל (**throughput**))

טකסונומיה של טכניקות אופטימיזציה:

1. אופטימיזציה ברמת הדטה:

דחסית קלט: טכניקות כמו חיתוך(pruning) פרומפטים, סיכום(summarization) פרומפטים, דחסית מבוססת פרומפט רך (למידה של וקטורים רציפים "המייצגים" את הפרומפט), והיסק מבוסס RAG מפחיתות את גודל פרומפטי הקלט תוך שמירה על מידע סמנטי בן. זה עיל במיוחד לתרחישים הדורשים קלטים ארוכים יותר.

ארגון פלט: שיטות כמו (SoT) Skeleton-of-Thought וגישות מבוססות גרפ תלות מאפשרות מקביליות חלקית של גנרטוט טוקנים, תוך ניצול המבנה הפנימי של פלט LLM.

2. אופטימיזציה ברמת המודל:

騰存 模型 以:

- שיטות כמו (MoE) Mixture-of-Experts מחלקים משאבי חישוב באופן DINAMI ל Tokni קלט, תוך אופטימיזציה של חלקית רשות MLP הפנימיות בבלוק הטרנספורמר(במידה האמבדינג BD'c).
- מגנוני attention מפושטים או מבוססי kernel Performer (כמו Shakerette בזמןנו) מפחיתים סיבוכיות מריבועית

ללייניארית (ביחס לאורך הקלט).

- חלופות לטרנספורמרים, כמו State Space Models (SSMs), כה האוהבים עלי', וארכיטקטורות RNN (מתברר שיש פה ושם שימוש בהם) מקטיניות את סיבוכיות המודל תוך שמירה על ביצועים תחרותיים (לפעמים). בהקשר זה כדאי להזכיר את A21 labs Jamba של labs A21 שבו ארכיטקטורת טרנספורמרים עם מבנה (סוג של SSM)

דוחיסת מודל:

- **קווינטוט:** מפחית רוחב סיביות למקולות והפעולות. שיטות כימות לאחר אימון או אימון-מודע-כימות שומרות על דיק למרות הדחיסה.
- **דילול:** מסיר פרמטרים או ראשי attention מיותרים, באמצעות טכניקות כמו pruning או מנוגני attention דليلים.
- **זיקוק ידע**(distillation): מאמן מודלים קטנים יותר לחקות את התנהגות המודלים הגדולים, עם אובדן ביצועים מינימלי.

3. אופטימיזציה ברמת המערכת:

שיפורים במנוע היסק (למשל, פענוח ספקולטיבי ואסטרטגיות offloading) ומערכות שירות (למשל, חישוב בבאצים, scheduling מתוחכם וניהול זיכרון) משפרים את ניצול החומרה וביצועי המודל (מבחינת h-throughput).

המאמר מצין שתהיליך היסק מחולק לשני שלבים:

1. מילוי מקדים(prefilling): אתחול המודל עם פרומפטי קלט העלה של זוגות KV שישמשו לגנרטו הטקסט.
2. פענוח: ייצור טוקנים רציפה עם תקורת זיכרון וחישוב.

גישה ניתוח עילוות:

מדדי עילוות כמו השהייה (לטוקן ולרצף כולל), שימוש בזיכרון (מקולות מודל, KV cache, צריכת זיכרון מקסימלית), ותפוקה (טוקנים/שנייה, בקשנות/שנייה) מונתחים כדי לנמת את ההשפעה של שיטות אופטימיזציה הנבחנתת.

כיוונים עתידיים:

1. טכניקות אדפטיביות המתאימות דינמית את גודל המודל והחישוב בהתאם על מורכבות הקלט.
2. אופטימיזציה משותפת בכל הרמות - DATA, מודול ומערכת - למקסום הייעולות.
3. שיטות מודעות-חומרה לניצול מאיצים מודרניים כמו GPUs ו-TPUs.

<https://arxiv.org/abs/2404.14294>

המאמר היומי של מייק - 07.01.25

Anchored Preference Optimization and Contrastive Revisions Addressing
Underspecification in Alignment

המאמר שנסקרו היום מציע שיפור לשיטת יישור (alignment) למודל שפה, DPO, השיכת למשפחת טכניות RLHF או Reinforcement Learning with Human Feedback RLHF כמו שאתם זוכרים SFT Supervised Fine Tuning (pretraining) ו- SFT (האחרון בד"כ) לאימון LLM יחד עם אימון מקדים (pretraining).

מטרת RLHF היא להראות למודל מה ההבדל בין תשובות מועדפות (על ידי בני אדם) מתשובות פחות מועדפות. במקרה יותר מתמטית RLHF מאמנת את המודל למסס את היחס בין הציון של התשובה מועדפת (טובה) יותר בין תשובה פחות טוביה. שיטת RLHF קלאסית Proximal Policy Optimization מוסיפה לאיבר הממסס פונקציית לוס איבר רגולרייזציה המנסה לשמור את הפוליסי הנלמד (כמו LLM מאומן) קרוב ל-LLM ההתחלתי (הקרבה מחושבת עם KL על ההתפלגות של הטוקנים החזויים על ידי שני המודלים).

הציון מחושב על מודל תגמול (reward model) שמאומן (בלשן הקודם RLHF) לשערך את "aicots" התשובה לשאלה נתונה. ככלומר מודל תגמול R אמר לנתן ציון גבוה לתשובה טוביה וציוון נמוך לתשובה פחות טוביה. המודל מאומן על זוגות של תשובות טובות ולא טובות לשאלות, כאשר בד"כ התיאוג של התשובות מתבצע על ידי מתייגים אנושיים (לפעמים רותמים מודל שפה עצמאי לתיאוג זהה).

התברר שניתן לקרב את יעד האופטימיזציה של PPO ללא אימון של מודל תגמול. בשנתיים האחרונות יצאו לא מעט מאמרים שהציגו שיטות של "ירודעות" להסתדר ללא מודל תגמול. אחת מהן היא DPO שזה ראשית תיבות של Direct Preference Optimization (DPO). עם DPO פונקציית תגמול מוגדרת sdp_z בתור לוגריתם של היחס בין הפוליסי (התפלגות החזויה של טוקנים הנמדדת על ידי המודל או נראות-likelihoods) עבור המודל המאופטם (שעובר פין טיון) לבין זה של המודל ההתחלתי. מטרת אימון DPO היא למסס את הפרש בין התוחלת (עבור הדאטסט של זוגות שאלות ותשובות) ההפרש של sdp_z בין התשובות מועדפות לבין פחות מועדפות.

הנקודה העיקרית של המאמר היא האובייקטיב שהאופטימיזציה של פונקציית המטרה של DPO עלולה להשפיע באופןים שונים על יחס הנראויות (likelihoods) של תשובות מועדפות לעומת פחות מועדפות. היא כמובן יכול להגדיל את הפרש ביןיהם (שהזה המטרה המוצהרת שלהם) אבל יכול להגיד את w_k יותר מאשר הוא מגדיל l_k , או להקטין את l_k יותר מאשר הוא מקטין את w_z . תרחישים אלה עשויים להוביל ליצירת מודלים שונים מאוד. המאמר מצין שתשובה מועדפת אינה בהכרח טוביה יותר ממה שהמודל מייצר לפני היישור. במקרה זה, DPO עלול לפגוע בביצוע המודל.

המאמר מתבונן במקרים השונים של ערכי sdp_z עבור התשובות w ו- l (מועדפת ופחות מועדפת בהתאם) ובונה שתי פונקציות מטרה ל- DPO שעשויות להוביל לביצועים טובים יותר עבור מקרים אלו. שיטת אימון שמאפpta מת פונקציות אלו קיבלה שם Anchored Preference Optimization (APO). הפונקציה המוצעת הראשונה מגדילה את ערך הפוליסי (נראות של תשובה) כאשר הערך הנוכחי של sdp_z עבר w קרוב ל-0 (w הינה בעלת נראות נמוכה יותר עבור המודל ההתחלתי) ומקטינה את הנראות של התשובה הפחות מועדפת עוד יותר אם sdp_z עבר קרוב ל-0.

הפונקציה המוצעת השנייה לעומת זאת מקטינה את הנראות של w כאשר sdp_z קרוב ל-0 עבור w ומגדילה את הפרש בין הנראויות של w ו- l כאשר ההפרש בין sdp_z עבור w ו- l קרוב ל-0. כל זה במטרה לגרום למודל שפה המאומן באמצעות DPO להתכנס לפתרון טוב יותר.

יש עוד משהו מעניין במאמר זהה. המחברים טוענים שכדי ש- DPO יעבד בצורה טוביה יותר, שתי התשובות (w ו- l) צריכים להיות רלוונטיות לשאלה ואחת מהם צריכה להיות "רק קצת" טוביה מהשנייה. ככלומר במו בלמייה hard negatives ניגודות עדיף לאמן את המודל על

המחברים מציעים שיטה לזרחי (ובנויות דאטהסט) של תשובות מועדות ופחות מועדות והיא יוצרת תשובה מועדת מתשובה כלשהי (אך רלוונטי) על ידי הפעלת LLM המשפר את התשובה (עם פרומפט מטאים). שיטה אחרת שהמחברים מציעים להשתמש בה היא בהינתן שתי תשבות של המודל המאומן (עם DPO) להפעיל מודל שפה שמטרתו להגיד מהי תשובה יותר (זה נקרא judge policy). ניתן גם לבנות דאטהסט באופליין עם מודל שפה שלישי ומודל שופט.

סקירה ארוכה - אני מקווה שרדתם ...

<https://arxiv.org/abs/2408.06266>

9.01.25 המאמר היומי של מיק - ?When Can Transformers Count to n

המאמר חוקר את המוגבלות התייאורטיות והאמפיריות של ארכיטקטורות טרנספורמר כאשר ביצוע משימות ספירה פשוטות. הוא בוחן משימות כמו "ספרת שאילתות" (QC) והאלמנט השכיח ביותר" (MFE) כדי לקבוע מתי טרנספורמרים יכולים לפתור בעיות אלה ביעילות. המחקר חושף הן את יכולות והן את המוגבלות המובנות של טרנספורמרים בהקשרים כאלה, וספק תובנות מעיניות לגבי האילוצים הארכיטקטוניים שלהם.

התרומות העיקריות:

חשיבות QC

חשיבות QC היא למעשה ספירה של כמה פעמים TOKEN מסוים מופיע ברכף. המוחברים מגדירים שהטרנספורמרים יכולים לבצע משימה זו ביעילות אם גודל המבידיג d גדול מפי שניים מגודל המילון ω : עבור $\omega > d$, ניתן לבצע הסטוגרמה (שפותחה במאמר) מאפשרת ספירה על ידי הטמעת "צוגי TOKENים בצורה אורטוגונלי". זה מאפשר למודל לבנות הסטוגרמה של הופעות TOKEN על ידי בлок טרנספורמר יחיד. עבור $\omega < d$, האורתוגונליות של המבידיגים כבר לא אפשרית, מה שהופך ספירה מדוקית לבליי אפשרית. המאמר מוכיח מגבלה זו בקפידה באמצעות [חומר Welch](#), המאפשר לנתח את הטריד-אופים של מימד המבידיגים (הקשורים לאורטוגונליות).

שיטת CountAttend

כאשר גודל המבידיג d קטן מגודל המילון ω , המוחברים מציעים את פתרון ה-"CountAttend", כדי לפתור את QC עם מגנוני ה-attention. הפתרון כולל שני רכיבים עיקריים:

1. משקלים attention

- מגנן attention מייצר משקלים המקודדים את היחס בין TOKEN השאיתלה לכל הTOKENים ברכף. לצורך ספירה, משקלים attention חייבים להיות הפוכים ביחס לספירת TOKEN ברכף. שקולו זה מבטיח שתרומה כל אסימון לפולט מנורמל לפי הтирוט של - זה מבטיח שתרומה של כל TOKEN לפולט מנורמלת לפי הтирוט שלו.

2. MLP להיפוך משקלים

- נדרש MLP כדי לשחזר את הספירה האמיתית c משקל attention. ה-MLP צריך למדוד פונקציה מהצורה: $w/f = w$

אתגרים עם פתרון CountAttend

чисוב משקל חישוב attention: חישוב משקלים הנקנים ביחס למספר טוקנים דורש מידול מדויק של יחס טוקנים לאורך הרצף. זה מוסיף מורכבות למנגנון ה-attention.

גודל MLP: עבור סדרות ארכוכות יותר, מספר הנוירונים ב-MLP חייב לגודל באופן פרופורציוני ביחס לאורך הסדרה שזה בערך מאד מבחינה חישובית.

משמעות MFE:

מטרת שימושת ה-MFE, היא למצוא טוקן בעל התדריות הגבוהה ביותר בסדרה. ניתן לישם את המשימה אם $(m \neq d)$ באמצעות גישה מבוססת היסטוגרמה. עבור $m < d$, המשימה הופכת לבלי אפשרית, כפי שמצווח באמצעות טיעוני מורכבות תקשורת. המחברים מציעים פתרון טרנספורמר דו-שכבותי לבעה זו.

מעבר פאזה בביטויים

המאמר מזיהה מעבר פאזה קרייטי: טרנספורמרים נכשלים במקרים ספירה כאשר $m > d$. סוף זה מדגיש את הנסיבות בין גודל ההטמעה, גודל אוצר המילים, ומורכבות המשימה.

תובנות תיאורתיות:

בנייה אמבידינגו אורטוגונליים

המחברים מנצלים את התכונות המתמטיות של אורתונורמליות ליישום ספירה מבוססת היסטוגרמה. עבור $d > m$, ניתן לבנות אמבידינגו כך שמכפלה סקלרית בין הטמעות טוקן שונות היא אפס. זה מבטיח ספירת טוקנים מדויקת בתוך בלוק attention יחידה. המאמר משתמש בגבולות Welch להראות עבור $m < d$, המכפלה הפנימית בין וקטורי ההטמעה הופכת לשימושית, מה שמקenis שגיאות בהיסטוגרמה. עבור שימושת MFE, המחברים משתמשים בכלים מעולים [communication complexity](#) כדי להוכיח שהטרנספורמרים דורשים $(m \neq d)$ כדי לפטור את המשימה.

השלכות מעשיות:

המסקנות מובילות למספר השלכות לתכנון ופריסה של טרנספורמרים ביישומים מעשיים: סקלබיליות ארכיטקטונית: טרנספורמרים חייבים להתאים את גודל אמבידינגו לגודל המילון. קידוד מיקומי (positional encoding) המחברים מגדירים את הנחיצות של הטמעות מיקום למשימות ספירה. בעוד שפתרון ההיסטוגרמה עיל עבור $m > d$, היישום המעשית שלו עשוי להיות מאוד בערך מבחינת הזיכרון ויבוכיות.

מסקנה

המאמר מספק ניתוח מקיף של היכולות והמגבילות של טרנספורמרים בפתרון משימות ספירה בסיסיות. באמצעות שילוב של היכולות תיאורטיות ריזורזיות עם אימוט אמפירי, הוא מדגיש את הנסיבות הארכיטקטוניות המובנות במודלי הטרנספורמרים.

מחקר עתידי:

- ארכיטקטורות היברידיות המשלבות טרנספורמרים עם שיטות ניוו-סימבוליות למשימות ספירה

- הרחבות למשימות הכוללות ספירה היררכית או מובנית

- מחקרים机制可解释性 mechanismic interpretability לבהרת הייצוגים הפנימיים שנלמדים על ידי טרנספורמרים במהלך משימות ספירה

<https://arxiv.org/pdf/2407.15160>

10.01.25 המאמר היומי של מייק - Chain of Thought Empowers Transformers to Solve Inherently Serial Problems

המאמר מציג ניתוח תיאורטי של כיצד Chain of Thought (CoT) מאפשר למודלי טרנספורמר להתמודד עם חישובים סדרתיים (לא מקבילים). המחברים הוכיחו חומרי expressiveness פורמליים ומציגים מחלוקת חדשה (איך ניתן לתרגם complexity class ?complexity complexity class סיבוכיות) המאפיינת את יכולות החישוב של טרנספורמרים עם CoT.

התרומה התיאורטית העיקרית של המאמר טמונה בחסמי האקספרסיביות שהוא מוכיח. באמצעות ניתוח מתמטי ריגורוזי, המחברים מוכחים שהטרנספורמרים בעלי עומק קבוע עם דיקון סיביות קבוע מוגבלים לפתרון בעיות מחלוקת סיבוכיות הנקראת [AC0](#) ללא CoT (משפט 3.1). עם זאת, הם מראים שעם T שלבי CoT, טרנספורמרים מסווגים לפטור כל בעיה הניתנת לחישוב על ידי [שרשרת בוליאנית](#) בגודל T (משפט 3.3). לתוצאה זו יש השלכות עמוקות, שכן היא קובעת שמספר פולינומיyal של צעדי CoT מאפשר לטרנספורמרים לחשב כל פונקציה בחלוקת [poly-P](#).

המסגרת התיאורטית שפותחה במאמר מורכבת משלושה חלקים עיקריים. ראשית, המחברים מציגים ניתוח מكيف של חישובי precision-low לעומת floating-point בטרנספורמרים. שנית, הם מבוססים קשרים עמוקים עם תורת סיבוכיות על ידי הגדרת מחלוקת מורכבות חדשה $(\text{CoT}, \text{d}, \text{s}, \text{e})$. המאפיין את החישובים בטרנספורמר עבור מספר שלבי CoT המסומן בתור $(\text{CoT}, \text{d}, \text{s}, \text{e})$, המיד החבוי של הטרנספורמרים $(\text{CoT}, \text{d}, \text{s}, \text{e})$ יציג נומרית $(\text{CoT}, \text{d}, \text{s}, \text{e})$ והאקספוננטה שלו $(\text{CoT}, \text{d}, \text{s}, \text{e})$. שלישיית, הם משלבים [תורת אוטומטים](#) על ידי שימוש [במשפט הפירוק של גראן-רודה](#) לניתוח יכולות הטרנספורמר.

מבחינה ארכיטקטונית, העבודה מספקת ניתוח מפורט של רכיבי טרנספורמר, כולל מגנוני, חסיטון self-attention, FFNs, קידודי מקום, והשעות שכבות נרמול. ניתוח זה מספק אפיקון מדויק של יכולות חישוביות ובסיס מיפויים ברורים בין תכונות ארכיטקטוניות וחסמים תיאורטיים.

השפעת המאמר חרגת מניתוח טרנספורמרים. על ידי הצגת מחלוקת סיבוכיות חדשה לחישובי הטרנספורמרים, הוא מgeshr בין מודלי חישוב וקלאסיים עם אלו המבוססים למידה عمוקה. הכלים המתמטיים שפותחו משלבים מוגרות תיאורטיות מרובות ביעילות ויוצרים קשרים חדשים בין תחומיים נפרדים בעבר. במבט קידמה, עבודה זו פותחת כיווני מחקר מבטחים רבים, במיוחד בתחום השימוש המיטבי ב- CoT והמגבילות היסודות של ארכיטקטורות טרנספורמר. המסגרת התיאורטית שנוסדה כאן צפואה לשימוש כביסיס לניתוח חידושים עתידיים בארכיטקטורת רשתות נירוניים וסטרטגיות הנחיה.

<https://arxiv.org/abs/2402.12875>

11.01.25 המאמר היומי של מייק - Evaluating the Design Space of Diffusion-Based Generative Models

מאמר זה מספק ניתוח מקיף של מודלים גנרטיביים מבוססי דיפוזיה על ידי הצגת מסגרת מאוחדת המגשרת בין שלבי האימון והדגם. הוא בונה בסיס מתמטי מוצק להבנת כיצד בחירות תכונן משפיעות על ביצועי המודל ויעילות החישוב. המאמר מתמודד עם יחס הגומלין המורכבים בין תהליכי האימון והדגם במודלי דיפוזיה. בנגד עבדות קודמות שלעתים קרובות מבודדות שלבים אלה, מחקר זה מספק ניתוח שגיאה מאוחד המשלב את שניהם.

התרומות העיקריות:

1. דינמייקט אימון וניתוח התוכנות

המאמר בוחן את התנהוגות של פונקציית המטרה של Denoising Score Matching או DSM במהלך תהליכי אופטימיזציה שלה (עם מורד הגרדיאנט - Gradient Descent). באמצעות טכניקות מעולם פונקציות סמי-חלקות (ראו נספח להסביר על כך), הוא מbasס התוכנות אקספוננציאלית (במישור האיטרציות של GD) עבור רשותות עמוקות עם אקטיביצ'ית ReLU ומספק תובנות לגבי פונקציות משקל אופטימליות לאימון (איבר המכמת לוס עבור כל עצמת הרעש ממושך באופן שונה ב-DSM).

תובנות מרכזיות בדינמייקט האימון:

פונקציית המשקל בצורה פשוטה עולה באופן טבעי מהניתוח במאמר. משקל זה מבטיח שהאופטימיזציה מתמקדת יותר בرمות רעש בינהיות, שבהן יחס האות-לרעש מואزن, מה שמקל על הרשת הנירונית ללמידה פונקציית Score (గרדיאנט של לוגריתם של פונקציית צפיפות של נקודת DATA A) בצורה מדויקות. חסמים על הגרדיאנט שהוצעו במאמר מסתמכים על הנחות מתוכנות בקפידה לגבי סקלת וממדות הדאטה, המשקפות תרחישי אימון מציאותיים. חסמים אלה מבטיחים התוכנות של פונקציית עבור מגוון ארכיטקטורות רשת ולחות זמינים של עצמת הרעש (noise schedule). על ידי תרגום הממצאים התיאורתיים להמלצות מעשיות, המחקר מדגיש שהחירות מוקדי משקל בפונקציית לוס היא קריטית להבטחת התוכנות מהירה מבליל פגוע ביכולת הכללה של הציון הנלמד.

2. תהליכי דגימה וחסמים שגיאה

תהליכי הדגם במודלי דיפוזיה מסתמכים במידה רבה על סימולציה מדויקת של משווה אדריכאלית סטוכסטית (SDE) המדמה תהליכי הסרת רעש. ביחס לעבודות קודמות המאמר מוכיח חסמי שגיאת הדוקים יותר, לא-איסופוטריים תחת NS כלליים. ניתוח זה מכסה שגיאת אתחול, שגיאת דיסקרטיזציה, ושגיאת קירוב הציון.

מצаг במאמר כי סיבוכיות דגימה (כלומר כמה דגימות נדרשות כדי שרשת נירונים אקספרטיבית מספיק ללמידה שערוך Score מדויק המספק לגנטוט דגימות באיכות גבוהה) תהליכי הדגם היא כמעט לינארית במידת הדאטה, בהינתן שנעשה שימוש NS אופטימליים. לתוצאה זו יש השלכות משמעותיות על יכולת ההרחבה של מודלי דיפוזיה, במיוחד ביישומים רב-dimensionals כמו ייצור תמונות. המחברים מצינים אף NS שונים (פולינומייאליים) לעומת אקספוננציאליים) נעים בין מזעור שגיאות ועלות חישובית, ומצביעים הנחות ברורות לתרחישי אימון שונים. העבודה גם שופכת אור על משמעות אתחול הרעש והשפעתו על איות הדגם הסופית, מקשרת בין חסמי שגיאת תיאורטיים לתוצאות מעשיות.

3. ניתוח שגיאה מלא

על ידי שילוב ניתוח האימון והדגם, המחברים מפתחים מסגרת הוליסטית לכימות שגיאה *end-to-end* במודלי דיפוזיה גנרטיביים. שילוב זה חושף כיצד מקורות שגיאה שונים מתקשרים ומספק מבט מאוחד על הגורמים המשפיעים על איות הדגם.

נקודות מרכזיות בניתוח השגיאה:

פירוק שגיאת אופטימיזציה: המחבר מבחין בין שגיאות הקשורות לאימון (שגיאות אופטימיזציה וסתטיסטיות) ושגיאות הקשורות לדגימה (דיסקרטיזציה ותחול). פירוק זה מבהיר את יחס הגומלין בין אימון המודל לתהילר הגנרטוט. השפעת פרמטריזציית יתר (over-parameterization) של המודל: התוצאות מראות כיצד הגדלת רוחב ועומק הרשת יכולה למתן שגיאות אופטימיזציה, מאפשרת GD להציג התכונות אקספוננציאלית.

זה מתיישר עם תכיפות אמפיריות בلمידה עמוקה אך מספק בסיס תיאורטי קפפני. נזכיר כי חסמי השגיאה שהתקבלו תלויים בפרמטרים מרכזיים כמו מידת הדטה, NS, ופונקציות משקל. עבור NS מעשיים (למשל, EDM), החסמים מתישרים היטב עם מדדי ביצוע אמפיריים. הניתוח גם מדגיש כיצד שגיאות "מתחלקות" בין שלבי האימון והדגם, ומצביע תובנות לגבי איך לאזן מאמץ חישובי בין שלבים אלה לביצועים גנרטיביים אופטימליים.

נספח:

מהי סמי-חלקות?

סמי-חלקות היא תכונה של פונקציה לואס והגרדיאנט שלה, המבטיחה שצעדי GD מפחיתים את הלואס ביעילות גם כאשר הפונקציה אינה חלקה לחלווטין. עבור רשותות LU عمוקות, פונקציית הלואס כוללת לנאריות חלקית, מה שהופך אותה ללא-חלקה באופן כללי. תכונת הסמי-חלקות שהגרדיאנט מספק כיוון "טוב" לירידה למינות חסור החלקות. קיימים חסמים תחכמוניים על נורמות הגרדיאנט, המבטיחים התקדמות עקבית לירידה מינור הלואס. על ידי ניצול הסמי-חלקות, המחברים מבוססים קשור מתמטי בין ערך הלואס וגודל הגרדיאנט שלו, המאפשר להם להוכיח דעיכה אקספוננציאלית בשגיאת האופטימיזציה.

<https://arxiv.org/abs/2406.12839>

13.01.25 המאמר היומי של מייק - Improve Mathematical Reasoning in Language Models by Automated Process Supervision

זמן רציתי לכתוב סקירה על MCTS ולבמרי במקורה נתקלתי במאמר זהה המציע לישם את השיטה המגביה זו עבור אימון LLMs. הפעם המטרה לאמן מודל שפה לפתור בעיות מתמטיות (לוגיות) מרכיבת שפתרונו מכיל שלבים רבים.

קודם כל הסבר קצר מה זה בעצם MCTS. חישוב עץ מונטה קרלו (MCTS) הוא אלגוריתם לאופטימיזציה של פוליסי עבור תהליכי החלטה מركוביים (Markov Decision Process) בעלי אפק סופי וגודל סופי, המבוסס על דגימות אפיוזדות אקראיות המאורגןות באמצעות עץ החלטה.

. הוא עובד 4 שלבים:

1. **בחירה:** בוחרים מסלול מהשורש לעלה לפני פוליסי חקירה/ניצול (exploration/exploitation)
2. **הרחבה:** מושפים מצב חדש לעץ
3. **סימולציה:** מרכיבים סימולציה אקראיית מהמצב החדש עד סוף המשחק
4. **עדכן לאחוה:** מעדכנים את הערכים בכל הצמתים במסלול שנבחר

אנו משתמשים ב-MCTS כדי לשפר את המדיניות (policy) על ידי בחירת פעולות טובות יותר. המודל מספק הערכות למצבים במקומות סימולציות אקראיות ו-MCTS משתמש בהערכות אלו כדי לבנות עץ חיפושיעיל יותר. לדוגמה, AlphaGo משתמש ב-MCTS בשילוב עם רשותות עמוקות כדי לבחור מהלכים. היתרון העיקרי של MCTS הוא בין חקירת מצבים חדשים (exploration) לבין ניצול ידיעת הקיימים (exploitation), ומשפר את קבלת החלטות לאורך זמן.

המאמר שנסקרו היום מציע להשתמש בגישה MCTS כדי לאמן מודל שפה לבנות תשובות בעלות שלבים רבים וכן שאותם יכולים לנחש הצמתים בגוף זהה יהיו השלבים בפתרון. המאמר מצין פתרונות SOTA לאימון מודלי שפה לפחות בעיות אלו מתחלקים לשני סוגים. הראשון מסמלץ את כל שלבי הפתרון כך שהמודל מאמין (עם טכניקות RLHF לבחירתכם) למקסם את הפרט שהמודל מקבל בסוף (בד"כ בינהר, כולם האמם הפתרון נכון/לא נכון) עם אישרו איבר רגולרייזציה (קירבה למודל המקורי).

השיטה השנייה PRM עשו דבר דומה אבל למסלולים חלקים (=כמה שלבי פתרון בהתחלה). ניתן לראותו שהגישה הראשונה תעבור פחות טוב מאשר עם הרבה שלבים כי-h-dense reward מאוד דליל (sparse) וקשה לאופטימיזציה. המקהלה השניה צריכה הרבה נתונים איקוטי וזה מאד יקר.

המאמר כאמור מציע להשתמש ב-MCTS למטרה זו. כמו שמקובל ב-MDP אנו צריכים להגיד מה זה המצב, פעולה ותגמול. המצב s מוגדר בתור שלאה q , כל שלבי הפתרון עד עכשו (לא חייב לכלול את הפתרון) והפעולה a היא בחירת הצומת הבא שבמקרה הזה הוא שלב הבא של פתרון שלאה q . לאחר שהפעולה a נבחרת היא מתווספת ל- s - a כلومר המצב החדש הוא (a , old_s). הפעולה a נבחרת על ידי פוליסי (policy) כאשר עבור MCTS הוא מרכיב משני מחוברים: הראשונה (initialization) נוטה לבחור צמתים בעלי תגמול גבוה והאייבר השני (exploration) מעמיד צמתים שלא ביקרנו בהם הרבה.

עכשו הגיע הזמן לדבר עם התגמול (reward). עבור צומת נתן v התגמול שלו הוא אחוז-rollouts הנכונים (המסומן בתור C) שהחלה משלב v אחוז המסלולים בגוף שהגיע לפתרון הנוכחי החל מ v). דרך אגב יש שיטה מאוד אינטואיטיבית ליזיהו של הטעות הראשונה בפתרון לא נכון (שכמה מעבודות קודמות מצאו כמידעיעיל לאימון מודל) שמאפשרת להזות צמתים "לא נכוןים בהחלטה" (שמהם לא ניתן להגיע לפתרון הנוכחי) בפתרון שנkirאת "חיפוש בינהר".

השיטה כל פעם מחלקת את מסלול הפתרון לשניים ובודקת היום C עבור הצומת שנמצא בחצי המסלול גדול או קטן מ-0. אם הוא שווה לאפס אז הטעות נירה בחצי הראשון ואם הוא גדול מ-0 אז הטעות נירה בחצי השני. אז שוב מחלקים לחצי את החצי שבו אנו חושדים שיש טעות וממשיכים לצמצם את החיפוש עד שmagiu ל"צומת המטעה".

כדי להגדיל את מספר הדוגמאות המחברים מציעים לאחסן rollouts של הפתרון ולבצע חיפוש בינהר של הצומת שבו (ככל הנראה) קرتה טעות ולהתחל ממנה חיפוש חדש. זה מאפשר לבנות דוגמאות עם אותם השלבים ההתחלתיים והמשך שונה. אציין שעם גישת PRM (עליה המאמר בונה את הפתרון) כל דוגמא היא השלישייה של שאלה, פתרון חלקו, וציוון האם זה נכון. כל אלו אנו מקבלים בתהליך המתואר כאן.

לבסוף המאמר משתמש ב-MCTS עם פוליסי Q כאשר המצב של כל צומת בגוף הפתרון מתואר על ידי שלישיה (אחרת) שהיא מספר הפעמים שהפתרון ביקר בצומת זהה, אחוז הפתרונות הנכונים C מהצומת זה (כולם שעורף מונטה קרלו שלו) וגם ערך של פוליסי Q שהוא מקבל ערך גבוה עבור ערך של C קרוב ל-1 (צומת מוביל לרוב לפתרון הנוכחי) ויש לו איבר רגולרייזציה (כפל) הקומו אותו על פתרונות ארוכים יותר. בחירה של מסלול rollout נבחר על ידי דגימה שנבנית בהתבסס על העץ עם האלגוריתם שנקרא PUCT (נוסחה 3 במאמר). כמובן C , Q וטטייטיקה של העץ מתעדכנות במהלך המהלך MCTS.

זהו זה - סקירה מאד ארוכה, מקווה שהצלחתך להסביר אותו, מאמר לא טריוויאלי' ...

<https://arxiv.org/abs/2406.06592>

16.01.25 - המאמר היומי של מייק

Diffusion Models for Non-autoregressive Text Generation: A Survey

היום נסקרו סקירה מלפני שנה וחצי של תחום (משפחה טכניתות) אז מطبع הדברים זה הולך להיות די קצר. הסקירה היא על שיטות גיבוב טקסט לא אוטורגרטיביות(Cloumer לא טוקן אחריו טוקן אלא סדרה שלמה). השיטות שנדבר עליהן מגנרטות טקסט בכמה איטרציות אבל זה לא נעשה בצורה אוטורגרטיבית - למשל שיטות אלו יכולות לגנרט טוקן מס' 7 לפני טוקן מס' 24.

אוק", בטח כמה מכון חשבו על מודל דיפוזיה גנרטיביים אחרי שהזכירתי שיטות איטרטיביות ואתם לא טועים כאן. בסקירה קצרה זו אסביר בצורה מותאמת אין ניתן לגנרט טקסט עם מודל דיפוזיה. כמו שאתם בטח זוכרים מודל דיפוזיה מאומנים להסביר רוש מדatta מרועש וזה נעשה באיטרציות.(Cloumer המודל מאומן להסביר כמה קטינה של רוש מהדטה עד להגעה לדאטה נקי וכן לאחר האימון המודל מסוגל לגנרט רוש מרועש טהור בכמה איטרציות).

אבל איך ניתן להוסיף רוש לטקסט שחי במרחב דיסקרטי(Cloumer טוקנים). יש בגודל שתי גישות: הגישה הרציפה והגישה הדיסקרטית. בגישה הרציפה שהיא יותר פשוטה וקרובה לבנו אנו לא פועלם במרחב הדיסקרטי אלא במרחב של אמבדינגן. בגישה הרציפה אנו הופכים את הטקסט שלנו לקטור אמבדינג רציף אבל להבדיל אנកודר רגיל אנו הופכים כל טוקן לייצוג הווקטור בפרט מהאחרים. לאחר מכן מאומנים מודל דיפוזיה לגנרט אמבדינג של טקסטים. הוספת רוש ואימון מודל denoising מתרחשים במרחב האמבדינג כאשר המטרה היא הסופית היא לשחרר את הטוקנים מהאמבדינג(D"א יש כמה שיטות לעשות את זה) אחרי נקי רוש.

משפחה השיטות השנייה היא לבצע הוספת רוש במרחב הדיסקרטי. מובן שהרוש לא יכול להיות רציף אז מה שניתן לעשות היא לשנות את ערכי הטוקנים (למשל לטוקן [mask]) בהסתברות מסוימת כאשר המטרה היא באיטרציה האחרונה להפוך את כל הטוקנים ל-[mask]. מודל דיפוזיה באיטרציה 0 מאומן לחזות את הטוקנים מהאייטרציה הקודמת, כאשר באינפרנס הגנרט מתחילה מכך שכל הטוקנים שוויים ל-[mask] והמודל לאט לאט הופך אותם לטקסט.

כמוון שאופן הרעשה של טוקן בכל איטרציה זה היפרפרמטר השקול ל-noise schedule במודל דיפוזיה רגילים. ניתן לתאר אופן הרעשה בתור מטריצה. כל טוקן ניתן לייצוג על ידי וקטור ההסתברות (מעל מיליון הטוקנים) אז ניתן לייצוג טוקן מאיטרציה 0 כמכפלה פנימית של ייצוג באיטרציה 1-0 על ידי מטריצה סטוקסית Q (סכום של שורות ועמודות הינו 1). Q היא היפרפרמטר הכי חשוב במודל דיפוזיה דיסקרטיים.

מתברר זהה תחום מחקר די פועל למורות עדין מודלים אלו לא הגיעו לביצועים של מודל שפה אוטורגרטיביים. אבל אני לא פועל זהה עוד יקרה כי מודלים אלו מסוגל לעבוד בתפוקה גבוהה יותר ממודלים אוטורגרטיביים (בעבור מספר צנוע של איטרציות).

<https://arxiv.org/abs/2303.06574>

17.01.25 - המאמר היומי של מייק

Towards a Unified View of Preference Learning for Large Language Models: A Survey

מוסטיבציה

המאמר מספק סקירה נרחבת של שלב מהותי באימון LLMs: "ישור (alignment) של פלט המודל עם העדפות אנושיות. מיותר לציין כי ישור זה חינוי לשומים רבים LLMs. בעוד ש-RLHF ו Cooke מונחה (SFT) היו מרכזיים לישור, היחסים ביניהם נותרו לא נחקרו מספיק, מה שmobiel לפיצול המאמצים המחוקרים בפונאים אלו.

המחברים שואפים לאחד מאמצים מפוצלים אלה על ידי הצגת מסגרת המשלבת גישות RLHF ו-SFT תחת נוסחה מבוססת גרדיאנט אחת בלבד. איחוד זה לא רק מקשר על פערים מתודולוגיים אלא גם מכין את הקרקע להתקדמות מוגבשת יותר במהלך העדפות (preference learning). המאמר מדגש ישור כולל מספר מרכיבים - מודל, DATA, משוב (כגון פונקציית תגמול עבור RLHF) ואלגוריתם - כל אחד הוא חשוב להבטחת (בתוקוה) ביצועים חזקים.

תרומות טכניות:

נוסחת גרדיאנט מאוחדת לשני המקרי בלב המאמר נמצאת הנוסחה של גרדיאנט מאוחד לאופטימיזציה של העדפות (נוסחה 1 במאמר)

$$\nabla_{\theta} = \mathbb{E}_{(q,o) \sim D} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} \delta A(r, q, o, t) \nabla_{\theta} \log \pi_{\theta}(o_t | q, o_{<t}) \right].$$

כאשר:

δ : מקדם גרדיאנט שתלו באלגוריתם הספציפי, במשוב ובDATA.

A: האלגוריתם האופטימיזציה המוושם.

r : אות משוב (feedback) המשפע על מקדם הגרדיאנט (למשל תגמול)

θ _DATA: מודל מדיניות המפורמטר על ידי θ .

משמעות זו מחייבת את תהליכי האופטימיזציה המשמשים הן בשיטות מבוססות RL והן בשיטות מבוססות SFT, ומראה שההבדל העיקרי בינהן טמון באופן שבו המשוב משולב. שיטות מבוססות RL משתמשות בדרך כלל בתגמולים סקלריים, בעוד ש-SFT משתמש בתווויות העדפה או דירוגים.

טקסונומיה של למידת העדפות:

המאמר מסוג למידת העדפות לארבעה שלבים מקוררים:

DATA:

דגם DATA-On-Policy: DATA נוצרים בזמן אמת על ידי המודל המאומן. טכניקות דוגמה כמו, K-Top-K, Monte Carlo Tree Search ו-i-Nucleus Sampling

איסוף נתונים Off-Policy: הנתונים נאספים מראש, לעיתים קרובות ממוקורות חיצונית, כולל סטי נתונים מתוארים על ידי בני אדם (כמו למשל בשיטות SHP-RLHF) או DATAsets סינטטיים שנוצרו על ידי LLMs (למשל UltraChat, ULTRAFEDBACK).

משוב:

משמעות ישיר: כולל תווויות אנושיות וחוקים המנוטחים על ידי בני אדם. דוגמאות כוללות בדיקות נכונות בחשיבה מתמטית או תוצאות יוניטטיבים ביצור קוד.

משמעות בסיסי מודל:

מודלי תגמול: מעריכים הסתברויות העדפה אנושית באמצעות שיטות כמו מודל Bradley-Terry (נוסחה 2 במאמר):

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}.$$

האופטימיזציה מושגת דרך פונקציית LOSS סטנדרטית של לוג הנראות שלילית:

$$L_r = -\log \sigma(r^*(y_c, x) - r^*(y_r, x)).$$

מודל תגמול בסיסי בינהרִי (למשימות בהן יכולת המקהה ניתנת לקביעה על ידי תוצאותיו):

תיאוג ישיר של דגימות לאימון בסיסי בינהרִי כמודל היגישה פשוטה ויציבה. למשל, בחשיבה מתמטית, ניתן לティיג דגימה על בסיס האם התשובה מניביה את התשובה הסופית הנכונה. באופן דומה, במשימות ייצור קוד, ניתן לבצע תיאוג על ידי בדיקה האם הקוד שנוצר עבר בבדיקות מוגדרות. בכך גם לשימות כמו סיכום טקסט או יצירת דיאלוג, הדורשות השוואות זוגיות של דוגמאות, שיטות הערכה ישירות אלו מפשטות את תהליך תיאוג העדפות.

בניגוד למודל התגמול המסורתי של Bradley-Terry, ברגע שיש לנו תיאוגים עבור הדאטה, ניתן לאמן את מודל התגמול באמצעות פונקציית LOSS של סיווג בינהרִי תגמול עבור כל ליבל מבלי שהייה צורך לבנות דאטה עבור זוגות.

שיטת שפה כזאת עבור ה-verdict-shao他自己本身: משתמש ב-LLMs-as-a-judge: משמש להערכת פלטימ. מנגןוי תגמול עצמי, מטא-תגמול (מודל שפה בונה ציון עבור ה-verdict-shao他自己本身 בעצמו נוטן) ועוד מגוון שיטות בסגנון.

אלגוריתמים:

האלגוריתמים מחולקים לקבוצות על פי מספר הדגימות הנדרשות לחישוב הגרדיאנט:

שיטת Point-Wise: אופטימיזציה באמצעות דגימות בודדות. דוגמאות כוללות Proximal Policy Optimization (PPO) ו-ReMax Optimization.

שיטת Pair-Wise Contrast (סוג של למידה ניגודית): מנצלות השוואות בין זוגות של דגימות: הדוגמא הבולטת של שיטה זו היא Direct Preference Optimization (DPO).

שיטת List-Wise Contrast: משערכות את הגרדיאנט על פני כמה דגימות. גישה זו שימושית במיוחד במקרים המשמשות הדורשות הרכבות הוליסטיות, כמו דירוג או סיכום.

שיטות Training-Free: כוללות טכניקות אופטימיזציה של קלט/פלט, המבטילות את הצורך בעדכוני גרדיאנט במהלך היישור (המאמר לא מרחיב על זה)

אבלואציה:

אסטרטגיית אבלואציה בוחנת עד כמה טוב SLLMs מתוישרים עם העדפות אנושיות.

הערכתה מבוססת חוקים: משתמשת בקריטריונים מוגדרים מראש כמו נכונות עובדתית או אמות מידת ספציפיות למשימה.

הערכתה מבוססת LLM: כוללת SLLMs מתקדמים הפעילים כמעריכים, משתמשים בפרומפטים להערכתה ודרוג תוצאות.

טכניקות דגימת דאטה:

דוגמה On-Policy: מספר את עשר של הדטה איכוטו למשימות הדורשות חשיבה רב-שלבית.

דוגמה Off-Policy: דאטאטיטים סינטטיים, המיוצרים על ידי SLLMs מתקדמים, משמשים יותר ויותר כדי "لتתסכל" למידת העדפות.

מסקנה:

סקירה זו מספקת מבט מתמטי קפדי ומוחדר מושגית על מידת העדפות עבור SLLMs . המסגרת שלה מבהירה יחסית בין שיטות RL-SFT, מאפשרת לחוקרים להשוות, לשלב ולהציג אסטרטגיות יישור העדפות באופן שיטתי. הדגש על מושב, תכנון אלגוריתמים והערכתה מבטיח כיiso מקיים של התחום, הופך מאמר זה למשאב יקר ערך לקידום מחקר יישור LLM.

<https://arxiv.org/abs/2409.02795>

18.01.25 המאמר הימי של מיק ואוראל - MAKING TEXT EMBEDDERS FEW-SHOT LEARNERS

היום להבדיל מהסקירות האחרונות נסקור מאמר מאד קליל, ללא מערב מתמטיקה כבده. המאמר מציע שיטה לבניית "יצוג (אמבידיגס) מותאם למלידה in-context או בקצרה ICL". אזכיר כי ICL היא שיטת בניית פרומפטים כאשר אנו מספקים למודל כמה דוגמאות עבור משימה שאנו מוצפים ממנה שיעשה. למשל במשימת גנרטית קוד אנו מספקים למודל (בתוך הפרומפט) כמה דוגמאות שכל אחת מהן היא זוג (שאלה, קוד) במטרה "להבהיר" למודל מה אנחנו מוצפים ממנו. ד"א למה ICL לפעמים עובד על המשימות שהמודול לא אומן עליהם אפילו ברור ב-100% מהוות נושא מחקר די פעיל.

ນצין כי המודל בנוידן עדיין צריך לגנרט טקסט כלומר יש לנו מודל דקודר (עם מיסוך קווזלי שדי מפיער לבניית האמבידיג) ונשאלת השאלה איך אנו בונים אמבידיג אותו כמו שנאנו רגילים לעשות עם האנקודר. דרך אגב יראו כמה מאמריהם שהציעו שיטות לבניית אמבידיג עם מודלי דקודר כמו LLM2Vec ו- GritML אבל הם אינם מותאמים ל McKee שנדון במאמר. כלומר השאלה איך אנו בונים אמבידיג של פרומפט בסגנון ICL כלומר צזה שמכיל כמה דוגמאות פתורות להדגמה.

از המחברים מצאו לזה פתרון די פשוט. קודם כל הם הוסיפו טוקן EOS בסוף הפרומפט והתכוון הוא שיציג הטוקן זהה יכול את האמבדינג של הפרומפט כולו (כמו שנעשה ב-BERT לפני 7 שנים). באופן לא מפתיע המחברים בחרו לעשות זאת עם למידה ניגודית(contrastive learning) או CL. מטרת CL היא לאמן מודל ייצוג כך שהייצוגים של דוגמאות דומות(חיבויות) יהיו קרובות ואילו אלו של דוגמאות לא דומות(שליליות) יהיו רחוקים במרקח האמבדינג. בתור דוגמאות חיבויות המחברים בחרו כאלו עם תשובה נכונה על השאלה בפרומפט ואילו עבור דוגמאות שליליות מופיעות התשובה הלא נכונה. נציין כי הדוגמאות להדגמה בפרומפט נשארות זהות עברו החיבויים והשליליות.

זה זה - ככה הם מאמנים מודל אמבדינג על מספר לא גדול של דוגמאות (few-shot) ולפי המאמר התוצאות לא רעות.

<https://arxiv.org/abs/2409.15700>

19.01.25 המאמר היומי של מיק ואוראל: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

היפותזת כרטיס הלווטו (Lottery Ticket Hypothesis) אומרת שבתוך רשת נירונים צפופה (dense neural nets) המאותחלת בצורה רנדומלית, יש תת-רשת (או "כרטיס מנץ") שמאפשר אותה בנפרד, היא יכולה להגיע לביצועים כמו של הרשת המקורית.

נמצא שטכנית חיתוך(pruning) סטנדרטית מגלה באופן טבעי תת-רשתות כאלה, אשר עברו מתקנים כי האתחול המחדש תחת אותם hyperparameters, משמר את התוצאות של הרשת המקורית בעלות זולה יותר, כך שהcartisits המנצחם הם תת-רשתות אשר "זכו בהגרלת האתחול", ובן המשקלים ההתחלתיים הופכים את האימון לאפקטיבי במינוחך.

הרעיון הזה מדגיש את החשיבות של המשקלים ההתחלתיים של הרשת. הכרטיסים המנצחם אינם תת-רשתות אקרואיות, אלא כאלה שמתאימות במיוחד באجل האתחול שלהם. תהליך מציאת התת-רשתות הללו אינו פשוט, כיון שהוא כרוך בזיהוי החלקים הקרייטיים(הנוירונים המשמעותיים) ברשת כבר מההתחלת.

מה זה חיתוך רשת?

חיתוך (Pruning) הוא טכניקה המסיר משקלים לא חשובים מרשת הנירונים. לפי היפותזת כרטיס הלווטו, החיתוך עוזר לייעל את הרשת בכך שהוא מסיר נירונים וחיבורים מיותרים, וכך יוצר רשת קלה, מהירה ויעילה יותר, ששומרת על הביצועים של הרשת המקורית ולעתים אף משפרת אותם. החיתוך חשף את "הכרטיסים המנצחם": בתחילת, הרשת מכילה יותר מדי פרמטרים (רשת גדולה וצפופה), ואז במהלך האימון והחיתוך של המשקלים הלא משמעותיים, תת-רשתות היעילות האלו מתגלות.

סוגי חיתוך

חיתוך לא מבונה (Unstructured Pruning): כאן אפשר להסיר כל משקל או קבוצה של משקלים, ללא מגבלות. זה יוצר רשת נירונים "דיללה" שבה רק חלק מהמשקלים נשארים. טכניקה זו נקראת גם חיתוך משקלים (Weight Pruning). בחיתוך שכזה, אין בחירה מוגדרת מראש מה יוחתך, הכל לפי הבחירה הפחותה ביותר של התרומה של אותו נירון שנבחר להיתוך.

חיתוך מבונה (Structured Pruning): כאן מסירים קבוצות שלמות של משקלים, כמו נירונים שלמים ברשת FFN. התוצאה היא רשת נירונים "צפופה" אך קטנה יותר. הבחירה כאן היא מושכלת, בה המבניות של

הרשות חשובה להישמר, יכול להיות שהיא נירזן שלא יבחר להיחתך על מנת לא לפגוע במבנה שונברה, לעומת זאת נירזנים אחרים.

חיתוך בבת אחת מול חיתוך איטרטיבי

חיתוך בבת אחת (One-shot Pruning): מאמנים את הרשות פעם אחת, חותכים אחז מסויים מהמשקלים (%) ו从此 מתחלים מחדש את המשקלים שנשארו. מדובר בהנחה כי באיטרציה אחת הגענו לפתרון הסופי והמיוחל, ללא צורך בתהליך חוזר ומתרשך.

חיתוך איטרטיבי (Iterative Pruning): מאמנים את הרשות, חותכים חלק מהמשקלים, מתחלים מחדש וחוזרים על התהליך כמה פעמים. בכל סיבוב חותכים אחז קטן מהמשקלים שרדדו מהסיבוב הקודם. תוצאות מראות שabitution איטרטיבי מצליח למצואו כרטיסים מנצחים שmaguisים לאותם ביצועים כמו של הרשות המקורית, תוך שימוש ברשות קטנה יותר בהשוואה לחיתוך בבת אחת.

<https://arxiv.org/pdf/1803.03635>

21.01.25 המאמר היומי של מיק -

Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts

המאמר משלב תשומת ליבי למורות הידע הרדוד שני מחזק לגבי תחום הסדרות העתיות (time-series). בוגדול הסיבה העיקרית לכך שבשמו מופיע צמד מילים "Foundational Models" זהה היה די נדירה בתחום הסדרות העתיות להבדיל מתחום מודלי שפה. הסיבה לכך (כנראה) היא מגוון עשיר הרבה יותר של סדרות עתיות השונות יחסית לשפה טבעית.

האמת לא מצאת ב-E-MoE, המבוססת מבון על הטרנספורמרים, מציאות ארכיטקטוניות מאוד מעניינות עם זאת יש בו כמה דברים שונים אולי שאנו רגילים לראות ב-LLMs. למשל במקרה שכבת טוקיניזציה ואמבידיג, מבוססים על מיליון טוקנים, שיש לנו ב-E-MoE מודל המוצע יש כל טוקן (שזו נקודה בסדרה) עבר טרנספורמציה לא לינאריות עם אקטיבציה מסווג SwiGLU וכמה טרנספורמציות לינאריות.

בנוגע לשכבות הטרנספורמרים, המחברים לוקחים ארכיטקטורת E-MoE די סטנדרטית. השינוי היחיד שמשר את עיני הוא שימוש בשיטת גרמול RMSNorm שלא הכרת. פרט לכך יש את כל השכבות הרגילות של הטרנספורמרים כולל מבון שכבות residual.

השכבה האחרונה של E-MoE-Time היא קצרה שונה ממה שאנו רגילים לראות בטרנספורמרים. מכיוון שלהבדיל ממודלי שפה אנו צריכים מודל בעולם של TS אנו צריכים לחזות במספר נקודות זמן שונה (נגד'נית, דקה או יומן קדיימה), המחברים משתמשים בכך מה ראשית בשכבה האחורונה. כל ראש אחראי על חיזוי באופק מסוים (כמוות דגימות קדיימה). בaimon משלבים את הלוויים מכל הראשים.

גם פונקציות אלו במאמר הן די סטנדרטיות: פונקציית הובר שהיא הגרסה הרובוטית של 2 (הלא ניתן לא להגיע לערכים גבוהים מאוד). בנוסף יש איבר רגולרייזציה שמנסה להפעיל את כל המומחים ב-E-MoE בצורה אחידה. וכמוון אימנו את המודל על נתונים ענקיים ומגוונים.

זהו זהה - סקירה קצרה, ובתקווה גם ברורה....

<https://arxiv.org/pdf/2409.16040>

המאמר היומי של מיק - 22.01.25

MONOFORMER: ONE TRANSFORMER FOR BOTH DIFFUSION AND AUTOREGRESSION

היום נעשה סקירה קצרה של מאמר די מעניין ששילב שני סוגים של מודלים, מודל שפה ומודל ויזן בטרנספורמר אחד. רוב המודלים מולטימודליים מורכבים מכמה מודלים שככל אחד מהם אחראי על הגנרטו של סוג DATA זה. למשל מודלי שפה ויזואליים בד"כ מורכבים משני מודלים: מודל שפה ומודל לגנרט תמונות. המחברים מציעים "לחבר" את שני המודלים האלה למודל טרנספורמר אחד וזה נעשה בצורה דיאינטואיטיבית.

קודם כל נזכיר כי שני המודלים הללו עובדים במרחב הטוקנים כאשר עבור מודלי שפה כל טוקן הוא חלק של מילה או מילה שלמה ואילו עבור מודל ויזואלי כל טוקן הוא פאץ' של תמונה. אז הניסיון לחבר אותם למודל אחד נראה ד' טבעי אך לא ברור האם ניתן לאמן אותו הטרנספורמר לגנרט שפה ותמונות יחד.

המודל המוצע מגנרט שפה בדיק כמו LLM רגיל, בצורה אוטורגרטיבית, ככלומר, טוקן אחריו טוקן. אבל איך ניתן לשלב אותו עם מודל לגנרט תמונות שכמוון מבוסס על מודלי דיפוזיה (בשנת 2025 זה האופציה הדיפולית הר'). קודם כל ציריך לזכור שמודל אוטורגרטיבי (לגנרט שפה)עובד בצורה סיביתית (קווזלית), ככלומר במהלך גנרט טוקן ח כל הטוקנים מאחוריו ממושכים ולא משתמשים בגנרט(משתמשים במסכה קווזלית). למודלי אלו ציריכים מודל דו כיווני כי בזמן גנרט פאץ' של תמונה כדי מאד להשתמש בכל הפאצ'ים האחרים.

בדיק כך בניו המודל המוצע - השפה מגנרטת עם מסכה קווזלית והתמונה מגנרטת עם כל הטוקנים (כולל הטוקנים של טקסט). דרך אגב הגישה זה תעבור גם לכיוון השני: ככלומר בגנרט של טקסט מתמונה (למשל למשימת captioning). אבל איך נדע לעבור ממצב "קווזלי" לממצב "דו-כיווני". המחברים מציעים להשתמש בטוקן מסוים המשמן שמן מתחילה גנרט התמונה - הטוקן הזה אמרו להיות מגונרט למשל למשימה ייצור תמונה מテקסט.

כמה מילים על הטרנספורמר לגנרט תמונה. המאמר משתמש במודל דיפוזיה לטנטו כאשר המודל מזמין לבנות ייצוג לטנטי של תמונה מרעש (עבור כל פאץ'). לאחר מכן כל היצוגים (של הפאצ'ים) מועברים דרך הדקודה (מבוסס VAE) שבונה ממנו תמונה.

המודל מזמין עם הלוס שהוא סכום משוקלל של הלוסים הסטנדרטיים עבור המודלים המזוכים: מודל שפה ומודל דיפוזיה. המאמר מצליח לגנרט תמונות דיאיפותות....

<https://arxiv.org/abs/2409.16280>

המאמר היומי של מיק - 24.01.25

Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs

תמצית המאמר:

המאמר בוחן מחדש את השימוש בלמידה מחיזקיים מפידבק אנושי (RLHF) באופטימיזציה של LLM. הוא מציג את הדומיננטיות של שיטות OPTIMIZATION של RLHF (PPO) (Proximal Policy Optimization) לשיטות OPTIMIZATION של החיזקיים הסטנדרטיות בהקשר זה, תוך הדגשת חוסר היעילות החישובית והמורכבות המיותרת שלהם. במקומם זאת, החוקרים מציעים לחזור לשיטות פשוטות יותר בסגנון REINFORCE, ספציפית (REINFORCE Gradient) (Vanilla Policy Gradient) והרחבתו הרב-דגםיתית, REINFORCE Leave-One-Out (RLOO). שיטות אלו מוכחות ביצועים טובים יותר מ-PPO. הממצאים מבוחינות עלות חישובית, יעלות דגימה אופטימיזציה תגמול במספר מערci נתונים וארכיטקטורות LLM.

מדגשים שניתן להשיג התאמה של LLMs להעדפות אנושיות עם אסטרטגיות אופטימיזציה פשוטות יותר המותאמות לייחודיות של RLHF.

הרחבה על נקודות עיקריות:

1. פישוט תיאורטי:

החוקרים מראים שרבים מהרכיבים של PPO (למשל, קליפינג, פונקציות ערך value), ומידול ברמת טוקנים אינם הכרחיים ל-RLHF, בהינתן האתחול טוב של LLM (לאחר SFT למשל). על ידי מידול סדרות של מושג כפניות בודדות, REINFORCE מנע מהMORESH של פונקציות ערך-מצב(V-Q) ברמת טוקן, והופך את הבעה לדומה יותר לבנדיית הקשר.

2. שימוש מעשי:

שיטת LOO משתמשת בכל הדוגמאות שנוצרו לבניית בסיס השוואה, משיג יעילות דגימה גבוהה יותר מ-RAFT, שŁokoch את רק את הדוגמאות בעלות ציונים גבוהים (סוג של rejection sampling). זה מוביל לחיסכון משמעותי בחישובים וניצול טוב יותר של הנתונים הזמן. הגישה מפשטה את תהליכי RLHF על ידי הפחיתה התלות בהיפר-פרמטרים רגילים כמו יחס קליפינג והפרמטרים בשערור פונקציית יתרון (כמו ב-GAE).

4. רובסיטיות:

שיטת LOO מדגימה רובסיטיות לתגමולים רועשים ועוני KL גבוהה יותר, לעומת שיטות כמו RAFT שרגישות יותר לדיווקם.

תובנות תיאורטיות:

1. איזון שונות-הטיה(bias-variance tradeoff) ושערור גרדיאנט ללא הטיה:

שיטת PPO מסתמכת על פונקציות ערך-מצב ושערור יתרון מוכלל (Generalized Advantage Estimation) להפחיתה שונות של שערור גרדיאנט ביחס לעלת הטיה. המאמר טוען שב-RLHF עבור LLMs ד' מאומן (warm start) הופך את הפחיתה השונות לפחות קритית. זה מאפשר לשיטות ללא הטיה כמו REINFORCE לתפקד היטב בלי להכניס הטיה משמעותית. אמפירית, המאמר מגדים ש-REINFORCE משיג אופטימיזציה תגמול טובה יותר מ-PPO, אפילו תחת תנאים של שונות גבוהה תיאורטית.

2. מידול מסלול מלא(תשובה שלמה) לעומת מידול ברמת טוקנים:

שיטת PPO ממליצה כל טוקן כפולה, יוצר תהליכי החלטה מרקובי (MDP) בו רצפי טוקנים חלקיים הם מצבים. עם זאת, RLHF מיחס תגמולים רק לתשובות שלמות, מה שהופך מצבים ביןים ללא רלוונטיים. על ידי מידול

התשובה כפעולה יחידה, REINFORCE מפשט את הבעה למבנה [Contextual Bandit](#) המתוישר ישירות עם מבנה התגמול. תוצאות אמפיריות מאשרות כי גישה זו עולה על מידול ברמת טוקנים הן ביעילות והן בביטויים.

3. קליפינג ויציבות עדכוני מדיניות:

שיטת PPO משתמש במנגן קליפינג למניעת עדכוני פוליסי גדולים שעלולים לערער את הלמידה. החוקרים מראים שה מיותר עבור RLHF, מכיוון שמשתח האופטימיiza יצב הודות לLLM warm-started. הастה הקליפינג ב-PPO או הימנעות ממנו לחלוtin עם REINFORCE מובילה לביצועים טובים יותר, מה שמצויב על כך שה RLHF אינו דרוש רמה כלשהן של יצוב.

4. איזון בין הפחחת שונות והעלאה קלה בהטיה:

אומדן היתרון ב-PPO מażן בין שונות והטיה, נשלט על ידי ההיפר-פרמטר ג'. ערכי ג' גבוהים יותר (קרובים ל-1) מפחיתים הטיה אך מגדים שנות. החוקרים מדגימים שב-RLHF, ערכי ג' גבוהים יותר מוביילים באופן עקבי לתגמולים טובים יותר של המודל, תומכים בשימוש באומדנים ללא הטיה כמו REINFORCE

מגבליות וכיונים עתידיים

1. אופטימיזציה יתר של תגמול:

המחקר אינו מתמודד עם אופטימיזציה יתר של מודל התגמול(reward hacking), בה המדיניות מנצלת הטויות בפונקציית התגמול על חשבון הכללה. זה נשאר אתגר פתוח עבור RLHF.

2. הערכה אנושית:

בעוד שאחזוי ניחזק מודמים באמצעות GPT-4 משמשים כմدد להעדפות אנושיות, הערכות אנושיות ישירות הי מספקות ראיות חזקות יותר לאיכות ההתאמה.

3. "סקלබיליות":

הסקלබיליות של REINFORCE ו-LOO למודלים(הם בדקנו רק מודלים של B7) ודאטאטיטים גדולים יותר מצריכה מחקר נוסף.

המאמר מציג טיעון משכנע לבחינה מחודשת של שיטות בסוגן REINFORCE ב-RLHF, מאתגר את הדומיננטיות של PPO וdomina. על ידי ניצול המאפיינים הספציפיים של RLHF - כמו LLM started warm started LLM - כמו REINFORCE ו- RLOO יכולות עלות ותגמולים ברמת הסדרה - החוקרים מדגימים שיטות פשוטות יותר כמו REINFORCE ו- RLOO יכולות עלות על חלופות מורכבות יותר כמו PPO ו- RAFT מבנית אופטימיזציה תגם, יעילות דגימה ועמידות.

<https://arxiv.org/abs/2402.14740>

27.01.25 - המאמר היומי של מייק -

FineZip : Pushing the Limits of Large Language Models for Practical Lossless Text Compression

בחרתי את המאמר זהה לסקירה כי יש לי חיבה גדולה לכל מה שקשור לדחיסה - דחיסה של מודלים, דחיסה של דאטא או כל דחיסה שהיא (:). המאמר מציע שיטה נחמדה לדחוס דאטא. אולם בטח יודעים שהמודלים שלנו יודעים לדחוס דאטא לצורה לא רעה עם הייצוג הולטני (אמבידיג) שהם מפיקים מהדאטא. אם אני לא טועה אנו יכולים לדחוס תמונה ברזולוציה גבוהה פי 100 עם האמבדיג שלו. אבל הדחיסה הזה היא לא lossless. אכן ניתן לשחזר את התמונה מהамבדיג שלה כך שהען האנושית לא תבחן שום הבדל בין התמונה המשוחזרת לבין המקורית, אבל הן לא בהכרח ייצאו זהות. במקרה של טקסט זה יכול להיות קצת בעייתי כי אנו רוצים לשחזרו כמו שהוא.

המאמר המסורק לעמota זאת מציע שיטה לדחיסה טקסט כך שניתן יהיה לשחזרו במדויק. השיטה המוצעת היא ד' פשוטה וឥינטואיטיבית. הרעיון מודל שפה מגנרט טקסט - השכבה האחונה שלו פולטת התפלגות מעלה מרחב הטוקנים והטוקן נדגם מההתפלגות הזה (יש כמה שיטות). אנו יכולים להיעזר בהתפלגות זו כדי לדחוס את הטקסט שלנו. למשל כמו שהוצע במאמר LLMZip אנו יכולים לקודד כל טוקן (בנהנวน הקשר לפניו) על ידי ראנק של ההסתברות שלו בהתפלגות עבור הטוקן הזה. ראנק זה בעצם המיקום של הטוקן בראשימת הטוקנים הממוינת (בסדר יורד) לפי הסתברותם שלו בהתפלגות הטוקן הזה.

אם מודל השפה שאנחנו משתמשים בו הוא מאוד חזק ראנק זה יהיה קרוב ל-1 (או 2, 3 אבל לא 1000). וידוע כי סדרות קצרות יתנו לדחוס בצורה מאוד עיליה (קצב דחיסה גבוה). אך LLMZip הציע לטיב מודל שפה לטקסט שדוחסים אותו שלא פרקטית כי לכל טקסט צריך לשומר מודל משלה וגם האימון כבד. המאמר המסורק מציע להשתמש ב-SaRA (או PEFT אחר) לדחיסה כך שנוצר צורך לשומר גם את מטריצות התוספות (או adapters) לכל טקסט עם מודל שפה אחד לכלו.

עדין זה לא מאד פרקטי אבל מבחינה רעיונית די מעניין ...

<https://arxiv.org/abs/2409.17141>

29.01.25 - המאמר היומי של מייק -

A Survey on Diffusion Models for Inverse Problems

מודלי דיפוזיה התפתחו במהלך השנים ככלי חזק המסוגל ליצור דאטא באיכות גבוהה במגוון תחומים. הצלחתם סלה את הדרך להתקדמות פורצת דרך בפתרון בעיות הפוכות(inverse problems), במיוחד בשחזור וחידוש תמונות,

שם מודלי דיפוזיה מאומנים משמשים כפרויורים (כלומר מסוגל בצורה לא מפורשת להבין האם התמונה המשוחזרת בא מההתפלגות האמיתית).

מאמר זה מציע חקירה מקיפה של שיטות המנצלות מודלי דיפוזיה מאומנים מראש כדי לטפל בעיות הפוכות ללא צורך אליו נספ. הם מציגים טקסונומיה מובנית המסוגגת גישות אלה על בסיס הביעות הספציפיות שהן מטפלות בהן והטכניקות שהן מעסיקות.

בגדי כל השיטות האלה ממנפנות גישה דיפוזיונית גנרטיבית לשחזר DATA מושפעש.

מסגרת מתמטית של מודלי דיפוזיה גנרטיביים:

המאמר מפרק ביעות הפוכות תחת הניסוח הכללי:

$$Y = A(X) + \sigma_y Z, \quad Z \sim N(0, I_m)$$

כאשר A הוא אופרטור או פונקציית שיבוש (יכול לא ליניארי), ו- Z הוא רעש גאוס. בעיות הפוכות שונות כמו הסרת רעש, השלמת תמונה סופר-רחלוציה, ממוגרים בתוך ניסוח זה על ידי הגדרת צורות שונות של A .

המאמר דין במודלי דיפוזיה הסתרותיים להסרת רעש (DDPMs) ורחבותיהם המבוססות על משוואות דיפרנציאליות סטוכסטיות (SDEs) כדי לגשת לעוות הפוכות. התהיליך הקדמי מתואר על ידי:

$$dX_t = f(X_t, t)dt + g(t)dW_t$$

כאשר ζ הוא תהיליך וינר, X_t הוא התפלגות הדטה בזמן t . f ו- g הם היפר-פרמטרים של תהיליך הדיפוזיה (noise schedule). מסגרת משוואות דיפרנציאליות סטוכסטיות (SDE) הפוכות (ći מתחילה מהרעש ומסירים אותו לאט לאט) של אנדרסון משמשת לדגימה מההתפלגות הנתונים הלא ידועה:

$$dX_t = (f(X_t, t) - g^2(t) \nabla_{X_t} \log p_t(X_t)) dt + g(t) dW_t$$

ניסוח זה מאפשר מידול DATA מושפע על ידי הוספה הדרגתית של רעש ולאחר מכן תהיליך הדיפוזיה לשחזר DATA. האתגר המתמטי העיקרי הוא שערוך של פונקציית הציון(score function) שהיא האגדיאנט של התפלגות (X_t). הסקר מדגיש את תפקידה המרכזי של נוסחת טוויד'

$$\nabla_{x_t} \log p_t(x_t) = \frac{\mathbb{E}[X_0 | X_t = x_t] - x_t}{\sigma_t^2}$$

למידת התוחלת המותנית באמצעות רשותות נוירונים מספקת דרך יעילה לקרב את הציון.

טקסונומיה של שיטות בפתרון בעיות הפוכות מבוססות דיפוזיה

מחברי המאמר מספקים טקסטונומיה עשירת המסוווגת שיטות על בסיס הגישה המתמטית שלhn, סוג בעיות היעד וטכניקות אופטימיזציה. בಗ

שערך score function באמצעות קירובים לינאריים לבעיות הפוכות לינאריים (בקירוב)

קירובים אלה(score function) מנצלים לעיתים קרובות פתרונות בצורה סגורה לבעיות הפוכות לינאריות. הצורה הכללית ניתנת על ידי (ע-can הוא הדטה המשובש)

$$\nabla_{x_t} \log p(y|x_t) \approx -L_t M_t G_t^{-1} \nabla x_t$$

כאשר: L מייצג את שגיאת המדידה. M הטלת השגיאה בחזרה למרחב הפתרון. G גורם re-scaling השולט בעוצמה התคำשות ב- u (התמונה המשובשת)

שיטות מייצגות:

שיטה (Score-ALD) Score-ALD (ALD) הוא Annealed Langevin Dynamics משתמש בקירוב הבא:

$$\nabla_{x_t} \log p(y|x_t) \approx -\frac{A^T(y - Ax_t)}{\sigma_y^2 + \gamma_t^2}$$

שיטה DPS (דגימת פוסטורייר דיפוזיה): מקרב את הפוסטורייר y (הדטה המשובש) באמצעות מיפוי $t \rightarrow X$ היא הגרסה המורעשת של התמונה המשוחזרת:

$$p(y|X_0 = \mathbb{E}[X_0|X_t]) \sim \mathcal{N}(y; A\mathbb{E}[X_0|X_t], \sigma_y^2 I)$$

המוביל לאומדן הבा עבור ה-score function:

$$\nabla_{x_t} \log p(y|x_t) \propto A^T(y - A\mathbb{E}[X_0|X_t])$$

התאמת מומנטים: מרחיבה את DPS על ידי שילוב קירוב גאושיאני אנאייזוטרופי (לא איזוטרופי):

$$p(x_0|x_t) \approx \mathcal{N}(\mathbb{E}[X_0|X_t], \sigma_t^2 \nabla_{x_t} \mathbb{E}[X_0|X_t])$$

4.2 שיטות הסקה וריאצונית

שיטות אלה מקרבות את התפלגות הפוסטורייר האמיתית על ידי הצגת התפלגות תחליפית(ויריאצונית) נוחה לטיפול ואופטימיזציה של הפרמטרים שלה באמצעות טכניקות וריאצוניות. המטרה היא למזער את מרחק KL בין הקירוב והפוסטורייר האמיתי:

$$\min_q D_{KL}(q(x) \| p(x|y))$$

שיטת RED-Diff מציעה אובדן חדשני המשלב לוא שחזור והתאמת ציון (ככה תרגמתי score matching, שיטה ידועה לגנרטוֹן דאטָה) במודל דיפוזיה:

$$L_{\text{RED}}(\mu) = \frac{1}{2\sigma_y^2} \|y - A\mu\|^2 + \sum_t \lambda_t \|\epsilon_\theta(x_t) - \epsilon\|^2$$

כאשר μ הוא הממוצע של האומדן הוריאציאני, ϵ הוא פונקציית denoising (שערך רעש) שנלמדה על ידי מודל הדיפוזיה.

Blind RED-Diff: מרחיב את RED-Diff על ידי אופטימיזציה משותפת של הייצוג הלטנטי של התמונה ופרמטרי המודל ϕ . זה מוביל לבעה וריאציאנית הבאה:

$$\min_q D_{KL}(q(x, \phi) \| p(x, \phi|y))$$

כאן אנו מאמינים את המודל הלטנטי לתמונה יחד עם מודל דיפוזיה המשחזר אותו.

4.3. שיטות מסוג CSGM (מודלים גנרטיביים מבוססי ציון מותנה - conditional score).

גישהות אלה מבצעות אופטימיזציה ישירות על פני מרחב לטנטי באמצעות backprop. הרעיון הבסיסי הוא להתאים באופן איטרטיבי וקטורי רעש התחלתיים כדי לספק אילוצי מדידה (של התמונה המורעשת כלומר).

טכניקות מרכזיות:

- בקפרוף (backprop) דרך שימוש דוגם דיפוזיה דטרמיניסטי.
- אופטימיזית מרחב לטנטי לאכיפה נאמנת למדידות הנצפות (המח).

4.4. שיטות מדוייקות אסימפטוטית(asymptotically exact).

שיטות אלה מסתמכות על דגימה מההתפלגות הפוסטוריור האמיתית באמצעות טכניקות מתקדמות של שרשרת מרקוב מונטה קרלו (MCMC).

טכניקות מרכזיות:

- התפשטות חלקיקים(particle propagation): שיטות מונטה קרלו רציפות (SMC) מפיצות חלקיקים מרוביים דרך התפלגיות כדי לקרב את הפוסטוריור.
- דגימה מפוחלת (twisted sampling): שיטות כמו דוגם הדיפוזיה twisted משתמשות בעדכונים מודעי גיאומטריה (של תמנות או דאטָה אחר) כדי לשפר את קצבי ההתכנסות.

4.5. טכניקות אופטימיזציה

השיטות משתמשות עוד יותר לפיה אסטרטגיות האופטימיזציה המועסקות:

- טכניקות מבוססות גרדיאנט: משתמשות בנגזרות לאכיפת עקביות מדידה.
- טכניקות מבוססות הטלה: מתיילות דגימות על תתי-מרחבים אפשריים.
- טכניקות דגימה סטטיסטיות: משתמשות באישות הסתברותיות כמו דינמיות לנגבין לעדכוני חלקיקים (כמו בSMC).

sekirah זו זה מאגדת אלגנטיות כלים מתמטיים מתקדמים, ומספק בסיס מוצק לחוקרם השואפים לפתרור בעיות הבעיות באמצעות תהליכי דיפוזיה. השימוש של חשבון סטטיסטי, הסקה ביומאנית וטכניקות אופטימיזציה הופר אותו לנוקודת התיאחות קרייטית לדחיפת גבולות פתרון הבעיה ההפוכות.

<https://arxiv.org/pdf/2410.00083>

31.01.25 המאמר היומי של מייק - Law of the Weakest Link: Cross Capabilities of Large Language Models

מבוא והגדרת הבעיה:

המחברים מדגשים פער קרייטי במחקר ה-LLM הנוכחי - הנטייה להתמקד בהערכת יכולות צולבות תוך התעלמות ממשימות מהעולם האמיתי הדורשות מיומנויות מרובות (AGI aka :)). המכונות יכולות צולבות, cross-cap (cross-cap). המאמר מסיגר בעיה זו באמצעות טקסטונומיה מקיפה של 7 יכולות בודדות ושביע יכולות צולבות, כגון קידוד וחשיבה ושימוש בכלים וקידוד. כדי להתמודד עם המורכבות הטעינה בהערכת הצמתים הללו, המחברים מציעים את CrossEval, מודד המורכב מ-400 הנחיות מוגדרות על ידי בני אדם המזעירים לבדוק את ביצועי ה-LLM במשימות רב-命מדיות.

דוגמאות ליכולות צולבות:

קידוד וחשיבה: פרומפט בקטגוריה זו עשוי לבקש מהמודל לנתח קטע קוד ולקבוע אם הוא מיישם נכון פונקציה מתמטית מורכבת. משימה זו דורשת לא רק ידע בקידוד אלא גם חשיבה לוגית כדי לאמת את נכונות הפונקציה.

שימוש בכלים וחשיבה: בדוגמה אחרת, הנחיה עשויה לדרוש מהמודל לשמש בכלים אוחזור מידע מבוסס אינטרנט כדי לענות על שאלת לגבי מגמות מגז אויר היסטוריות, ולאחר מכן לספק הסבר אנליטי שלב-אחר-שלב של הדפוסים הנצפים. משימה זו דורשת הן יכולות חשיבה והן שימוש בכלים חיצוניים.

מתודולוגיה:

הגדרות יכולות מקיפות: הם בונים טקסטונומיה מפורטת של יכולות בודדות וצלבות, המסוגגת משימות לקטגוריות רחבות ותתי-קטגוריות מדויקות.

מדד Eval CrossEval: מסגרת הערכה חדשנית זו מורכבת מ-1,400 הנחיות, 4,200 תגבות מודל, ו-400 דירוגים אנושיים. מערכת ההנחיות כוללת משימות ברמות קושי שונות, החל משאלות עובדות פשוטות ועד למשימות מורכבות הדורשות יכולות צולבות.

הערכתה מבוססת LLM: הממחקר מציג מסגרת הערכתה מרובת-התיאחות שבה מעריכים מומחים מעריכים את איכות התשובות המרבות של המודל בסולם ליקרט. המחברים גם מפתחים אסטרטגיית הערכתה מבוססת הפקחת נקודות לדיקוק משופר.

ניתוח דינמיות יכולות צולבות: המחברים מוצאים שביצועי יכולות צולבות לעיתים קרובות מציתים ל"חוק החוליה החולשה ביותר" — שבו הביצועים מוגבלים על ידי היכולת האינדיבידואלית החלשה ביותר.

ממצאים ניסיוניים:

הממצאים חושפים מספר תובנות מפתח המדגישות את המגבילות והחזקות של ה-LLM הנוכחיים כאשר הם מתמודדים עם פונקציות יכולות צולבות.

חוק החוליה החולשה ביותר:

התכיפות הבולטת ביותר היא שביצועי היכולות הצולבות מוגבלים על ידי היכולת האינדיבידואלית החלשה ביותר, בהתאם ל"חוק החוליה החולשה ביותר". מtower 58 תרחישי יכולות צולבת שנבדקו ב-17 מודלי LLM הראו ביצועים נמוכים יותר מכל אחת מהיכולות האינדיבידואליות המעורבות, בעוד ש-20 ציונים נמצאו בין היכולות החזקות והחולשות אף הן קרובים הרבה יותר להחולשה ביותר. למשל, במשימות המשלבות שימוש בכלים וחישבה, אם המודל הציג כישורי חשיבה פשוטים, זה פגע משמעותית ביציעים גם כאשר יכולת המודל להשתמש בכלים הייתה מיזמנת. אפקט זה נצפה לא רק למורכבות או לאופי המשימה.

אפקט "חוק החוליה החולשה ביותר" נשמר ללא קשר לאיזה עיריר מבוסס LLM שיישם. בין אם GPT-4 או Claude 3.5 שימשו כשופטים, התוצאות באופן עקבי התקבעו ליד היכולת האינדיבידואלית החלשה ביותר. עקבות זו מחזקת את חוסנם של ממצאי המدد ומרמתה שהمبرיבות הנוכחיות של LLM הן מבניות עמוקות ולא ספציפיות למתודולוגיות הערכה.

חסרונות בשימוש בכלים:

שימוש בכלים הtagלה יכולות החולשה ביותר בכל ה-LLM שנבדקו. משימות הדורשות גלישה באינטרנט, אחזור נתונים דינמי, או הרצת קוד חיוני הוכחו כamateגרות במיוחד. הציונים הגבוהים ביותר למשימות הקוללות שימוש בכלים מעולים לא עלו על 50 בסולם של 100-1 לאורך המدد. באופן בולט, אפילו מודלים עם פונקציונליות מפרש קוד, כמו Gemini Pro Exp, התקשו לשמור על ביצועים שווים למשימות חשיבה פשוטות יותר.

חולשה זו קריטית מכיוון ששימוש בכלים הוא יסודי ליישומים רבים בעולם האmittel, כגון סייע במחקר, ניתוח נתונים, וסוכני AI. המחברים מציגים שמודלים המסתמכים ארוך ורק על מקורות נתונים סטטיים ביצעו באופן גורע בהשוואה למשימות שבנה מידע מפורש יותר היה זמין ישירות בתוך ההנחייה.

פער ביציעים ביכולות צולבות:

בממוצע, מודלים השיגו 65.72 למשימות יכולת בודדות אך רק 58.67 למשימות יכולות צולבות, פער של 7.05 נקודות. זה מדגיש את הקושי שמודלים נתקלים בו בעת שילוב מיזנויות מרבות. משימות "תרגום מספרדיית וחישבה" ו"הקשר ארוך (long context) וקידוד" הדגימו פערים גדולים במיוחד, המראים שנדרש אופטימיזציה נוספת בתרחישי עיבוד רב-לשוני והקשר ארוך.

עלויות CrossEval בהבנה:

CrossEval הוכח כיעיל בהבנה בין הבדלים עדינים אףו בין LLM מתקדמים ביותר. למשל, מודל Claude 3.5 Sonnet עקב בעקבות על קודמי (המודלים הקודמים של אנטרופיק) במשימות הקוללות זהה תМОנות וחישבה וספרדיות וזהה תMONות. התקדמות זו משקפת את ההתפתחות של מודלי Claude מתוחכמים יותר ומדגישה את הערך של CrossEval במדידת השיפורים העדינים ביכולות LLM.

שיעור מדרדי קורלציה:

המדד הדגים שיפור במדדי קורלציה להערכת מבוססות LLM במקורה שמספקים ל-LLM המבצע אבלואציה דוגמאות מתויגות. קורלציית פירסון השתפורה מ-0.578 לא דוגמאות מתויגות ל-0.697 עם שתי דוגמאות. המצביע על כך שהכללת התייחסויות מתויגות היטב שייפה משמעותית את אמינות ההערכה.

סיכום:

הניסויים מגלים שבעוד ש-LLM משתפרים במידה, הם נשאים מוגבלים מאוד על ידי הרכיבים החלשים ביותר שלהם. טיפול במוגבלות אלו חיוני להשגת מערכות AI חסונות יותר, וב-תפקודיות המסוגלות לפתור בעיות מורכבות מהעולם האמיתי.

<https://arxiv.org/abs/2409.19951>

המאמר היומי של מיק - 01.02.25

Classical Statistical (In-Sample) Intuitions Don't Generalize Well: A Note on Bias-Variance Tradeoffs, Overfitting and Moving from Fixed to Random Designs

מבוא:

שיטות ML מודרניות מציגות התנהגוויות שוטרות באופן בלט אינטואיציות סטטיסטיות מסורתיות, במיוחד בתחום לאימון-יתר (over-training), לאיזון בין הטיה לשונות, וליכולת הכללה. הסטטיסטיקה הקלאסית טוענת לעיתים קרובות שככל שМОדולוֹן המודל עולה, ההטיה יורדת, אך השונות עולה - איזון ידוע בין הטיה לשונות. עם זאת, תופעות כמו *benign overfitting* ו-*Double Descent* מאייתות השקפה זו. המאמר המשוקה טוען שתופעות אלה אינן נבעות באופן בלעדי ממודלים מורכבים, פרמטריזציית-יתר, או נתונים רב-ממד, אלא דווקא ממעבר יסודי בין שני סוגי הבעה הסטטיסטית: *fixed and random design*. המאמר מספק証據 קירה מתמטית של האופן שבו מעבר זה משנה באופן משמעותי מושגים סטטיסטיים.

הגדרת הבעה: משטר f_D - D_r vs fixed design D_f

ההבחנה בין D_f ל- D_r היא התובנה המהותית של המאמר:

משטר f_D : הנקודות בטסס סט נותרות זהות לאלו שבאימון, כאשר רק התווות שלhn נדגםות מחדש. ניתן סטטיסטי קלסי מנייח את זה לעתים קרובות ובעורו אנו מנסים למצער את שגיאת השערור *in-sample*.

משטר r_D : גם הנקודות וגם התווות במהלך הבדיקה נדגים באופן בלתי תלוי מהתפלגות הדעה. משטר זה מתיישר עם האופן שבו מודלי ML משוערכים כiom, תוך הtmpקודות בשגיאת הכללה או שגיאת חיזוי מחוץ למוגן (*out-of-distribution*).

המעבר D_f ל- D_r גורם לשינויים עמוקים בהתקנות של הטיה, שונות, ושגיאת החיזוי הכללה. שינוי עדין אך משפיע זה הוא הסיבה המרכזית לכך שתופעות ML מודרניות נראות כמפורטות את האינטואיציה הסטטיסטית הקלאסית.

מתמטית, השגיאות בשני המשטרים מוגדרות כך. שגיאת D_f (שהיא *sample-s-in*) כאשר הן תוצאות שנדגמו מחדש בקלטים קבועים.

$$\text{ERR}_{\text{fixed}} = \mathbb{E}_{\tilde{y}} \left[\frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - \hat{f}(x_i))^2 \right]$$

כאשר \mathcal{D}_y הפלטים שנדרמו מחדש עבור הפלטים מהטרין סט. שגיאת \mathcal{D}_y (מחוץ למדגם או מוגדרת באוון הבא):

$$\text{ERR}_{\text{random}} = \mathbb{E}_{x_0, y_0} \left[(y_0 - \hat{f}(x_0))^2 \right]$$

כאשר גם \mathcal{D}_x וגם \mathcal{D}_y הם דוגמאות חדשות מהתפלגות הדאטה. שינוי זה מוביל להשלכות מרחיקות לכט עבור איזון ההטייה-שונות ותכונות הכללה של מודלים. הטיה ושותות ב- f מקבל צורה ש邏ocرت לנו היטיב:

$$\text{MSE}(x) = \text{Bias}^2(x) + \text{Var}(x) + \sigma^2$$

$$\text{Bias}(x) = f^*(x) - \mathbb{E}[\hat{f}(x)], \quad \text{Var}(x) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

כאשר f^* הינו הרעש שלא ניתן לצמצום, $(x)^*$ היא הפועלקציה ground-truth הנלמדת ואילו (x) הוא המשער. עבור אומדנים פשוטים כמו NN-k. השונות יורדת מונוטונית עם עליית k כאשר יותר שכנים מומצאים וההטייה עולה מונוטונית מכיוון שה ממוצע כולל שכנים פחות דומים. איזון זה יוצר את העקומה בצורת U המוכרת מספרי הלימוד עבור שגיאת החיזוי כפונקציה של מרכיבות המודל.

אולם במשטר \mathcal{D} מוליד התנחות חדשה- האינטואיציה של הטריד-אוף בין הטיה לשונות כבר לא עובדת בצורה כה פשוטה. ההטייה אינה יורדת מונוטונית עם המרכיבות: השכן הקרוב ביותר עשוי שלא להתאים באוון מושלם לנקודת הבדיקה, מה שmobail להטיית התאמת שכנים שאינה אפס. ההטייה יכולה להציג דפו בצורת U, כאשר מודלים בעלי מרכיבותBINONI מזערם את ההטייה. התנחות זו ניתנת לבטא על ידי פירוק ההטייה ל:

$$\text{Bias}_k(x_0) = \left(f^*(x_0) - f^* \left(\sum_{i=1}^n w_{k,i}(x_0)x_i \right) \right) + \left(f^* \left(\sum_{i=1}^n w_{k,i}(x_0)x_i \right) - \sum_{i=1}^n w_{k,i}(x_0)f^*(x_i) \right)$$

שני הרכיבים הם:

הטיה התאמת שכנים: נוצרת כאשר הממוצע המשקל של נקודות האימון אינו משחרר באוון מושלם את נקודת הבדיקה.

הטיה מיצוע: נובעת אי-LINEARITY של פונקציה האמיתית (כלומר המיפוי מנקודה ללייבל).

פירוק זה חשף שגם במקרים פשוטים ונומי-ENN, תכונן אקראי מכניIS מרכיבות שמשבשות אינטואיציות קלאסיות.

תופעת Double Descent

תופעת DD מתייחסת להתנהגות הלא-מוניונית של שגיאת החיזוי כפונקציה של מרכיבות המודל. היא מרכיבת מעוקמה בצורת U במשטר under-parametrization (מספר פרמטרי מודל קטן ממספר הדוגמאות) וירידה שנייה במשטר over-parametrization (מספר פרמטרי מודל גדול ממספר הדוגמאות). המחברת מדגישה כי DD אינו יכול להתראש במצב f_D מכיוון שאינטראולציה תמיד מובילה לשגיאת sample-in קבואה = $\text{ERR}_{\text{fixed}}$ $= 2^{\Delta S}$. זאת מכיוון שמודלים במשטר זה חזים באופן טריבואטי (בקודות האימון), מה שמוביל להטיה ושונות אף בתכנון קבואה. עם זאת, במשטר z_D, תופעת DD מופיעה באופן טבעי בגין טבעי גלגול שוניים במרכיבות המודל האפקטיבית (שלא נמדדת במספר הפרמטרים) ותכונות ההכללה בעת המעבר לאינטראולציה.

Benign Overfitting(BO) vs. Benign Interpolation(BI)

המחבר מבקר את המונח BO, ומצביע במקומו מונח BI. הגדרות קלאסיות של אובייקט מرمזות על ביצוע הכללה ירודים, מה שנותר את הרעיון שביצועים מושלמים בטרין סט יכולות לעתים להניב ביצועים טובים גם על הטסט. במשטר f_R, אינטראולציה אינה יכולה להיות benign בגין הדומיננטיות של שונות הרעש $\text{ERR}_{\text{fixed}} = 2^{\Delta S}$.

במשטר d_R, לעומת זאת, מודלים כמו רשתות נירונים ויערות אקראיים(random forests) יכולים להציג התנהגות חדה-חלקה, בה הם מבצעים אינטראולציה חדה בתקודות האימון אך מقلילים בצורה חלקה לפחותים שלא נראו. התנהגות זו ניתנת לכימות באמצעות מודדי מרכיבות אפקטיבית. זאת אומרת מודלים שmphichities מרכיבות אפקטיבית על טסט סט נוטים להציג אינטראולציה שפירה.

השלכות:

חשיבות חדש על חינוך סטטיסטי: קורסי מבוא צריים להבahir את הבדיקה בין f_R לד_R.

הסקה סיבתית ו-ML: בתחוםים בהם נקודות מסט אימון עשויים לחזור (למשל, הסקה סיבתית), הנחות d_R עשויות עדין להיות רלוונטיות.

בחירה מודל ML: הבנה מתי אינטראולציה היא benign דורשת מדידת מרכיבות בזמן בדיקה, לא רק ביצועי אימון.

סיכום

ובודעה זו מזינה פרספקטיביה מאוד מעניינת על מודיען אינטואיציות סטטיסטיות קלאסיות לא תמיד עובדת טוב בעיות ב-ML מודרני. על ידי הדגשת השוואה בין f_R לד_R, המאמר מספק מסגרת מאחדת להבנת DD, Benign Interpolation והתקף המתפתח של טריד-אוף ההטיה-שונות.

<https://arxiv.org/pdf/2409.18842.pdf>

03.02.25 - המאמר היומי של מייק

The Perfect Blend: Redefining RLHF with Mixture of Judges

אחרי יציאת המודל האחרון של DeepSeek העניין ל-RLHF או שיטת טיב (fine-tuning) של מודלי שפה באמצעות שיטת Reinforcement Learning with Human Feedback. חוקרי DeepSeek הראו שניתן לאמן

מודל שפה חזק לעשوت הנמקה(reasoning) בעייר עם RLHF (יש קצת SFT אבל עדין הרוב). המאמר שנסקור היום יצא כמעט 4 חודשים לפני R1 של DeepSeek והוא מציע שיטה שמשפרת ביצועים של RLHF.

אחד הביעות הגדולות של אימון RLHF הוא reward hacking שתרחש כאשר המודל לומד למקסם את פונקציית התגמול (reward) אך כתוצאה לכך מתכוון למודל חלש או לא בטוח (למרות איבר הרגולריזציה שמנסה לשומר את המודל הסופי קרוב למודל שמננו מתחילה לעשوت RLHF). המחברים מציעים להתמודד עם הבעיה זו בשלוש דרכים. הראשונה היא סט של אילוצים על התשובה לפורנפט (שלמשל בודק האם הוא פוגעני) הנבדק על ידי "השופט" (judge) שתפקידו מלא מודל שפה אחר. השיפור השני הוא שניי של פונקציית תגמול המתבטה בחיסור ממון תגמול בייסליין מסוים שתיכף אסביר מהו. השלישי השלישי לבנייתו הוא דאטהסט עליון מאומן RLHF.

היחסור הזה מזכיר לי שני דברים. קודם כל התגמול החדש (אחרי החיסור) נראה דומה לפונקציית יתרון (advantage) (רק דומה אבל היא לא) המוכרת לנו פונקציית לואס (יעד) משיטת PPO רק שהפעם היא לא מחושבת דרך פונקציית Value באטעןויות Shietot GAE אלא בדרך אחרת. תגמול חדש זה מזכיר לנו מה שראינו בפונקציית יעד של המאמר של DeepSeek, שם הבייסליין חושב באמצעות תגמול ממוצע (עליז'ן) של המודל המטובי (המתוקן עם השונות). במאמר ההוא איבר זה שימוש כאומדן של אותה פונקציית יתרון.

כאמור החידוש השני של המאמר (הראשון האילוצים שאנו מטילים על פלטי המודל) הוא הבייסליין המחוסר מהtagmol. המחברים מציעים לקחת את הבייסליין בתור התגמול עבור דוגמאות (תשובות) הזהב (=מודיעות) מדאטהסט של SFT (שאלות ותשובות) או מהשאלות עם התשובות המעודפות מדאטהסט של RLHF. כך התגמול שלנו הוא עד כמה התשובות של המודל המאמן נראהות יחסית לתשובות המעודפות מבחינת תגמולן.

השנייה השלישי הוא בפונקציית יעד. בנוסף (המחברים מציעים 2 וריאנטים) למקסום של הנראות של התשובות המעודפות ומצעור הנראות לתשובות הפחות מעודפות (מבחינת התגמול), המאמר מציע רק למקסם את הנראות של התשובות המעודפות בלבד (באופן מפתיע זה עובד). המחברים גם "מלבושים" את הרענוןת הללו ששיטות RAFT ו DPO-RLHF כמו O-DPO.

<https://arxiv.org/abs/2409.20370>

5.02.25 המאמר היומי של מיין -

Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections

תמצית המאמר:

רציתם לדעת למה מודלי דיפוזיה ניצחו את הגאנים, VAE וכל השאר מזוויות מתמטית? רציתם להבין בעזרה מתמטיקה למה מודלי דיפוזיה לטנטיטים עובדים מעולה? תצללו לסקירה הזו....

מאמר זה מציע חקירה מקיפה של מודלים גנרטיביים عمוקים (DGMs) תחת המטרת של השערת הירעה, הטוענת שدادה בעל ממד גבוה נמצאים לעיתים קרובות על תתי-ירעה בעלת ממד נמוך יותר המוטמעת בתוך המרחב המקורי (במאמר נקרא אמביינטי). המחברים מספקים הסבר מדויק מודלים כמו מודלי דיפוזיה ו-GANs מסויימות מציגים ביצועים טובים יותר מאשרים, כולל שיטות מבוססות נראות כמו אוטואנקודרים ורייצוניים (VAEs) וזרימות נורמליזציה (NFs). על ידי אימוץ נקודת מבט מבוססת ירעה, המחברים מספקים תובנות לגבי המגבילות המובנות של גישות קיימות תוך יצירת קשרים תיאורתיים חדשים בין DGMs והסעה אופטימלית..

המחקר בולט בכך שהוא מוכיח באופן פורמלי את חוסר היציבות הנומრית המובנית שמודלים מבוססי נראות בממד גובה וחווים כאשר הם מנוטים לייצג דאטא על יריעה, ומציע פרשנות חדשה של DGMs דו-שלביים כמקבילים של מרחק וורשטיין בין התפלגות המודל להתפלגות הדאטא האמיתית.

נקודות מרכזיות

1. סקירה של מודלים DGM מודעי-יריעה ולא-מודעי-יריעה (manifold-aware and manifold unaware)

מודלים לא-מודעי-יריעה: מודלים אלה אינם מתחשבים באופן מפורש במבנה היריעה של דאטא. דוגמאות כוללות NFs, VAEs ומודלים מבוססי אנרגיה. מודלים כאלה נוטים להסתגלות יתר ליריעה, כאשר הציפיות שופות לאינסוף לאורך היריעה אך נכשלים בשעורה של התפלגות בתוכה.

מודלים מודעי-יריעה: מודלים אלה מושגים רעש כדי לפזר את מסת ההסתברות מעבר ליריעה או מאפטמים פונקציות יעד שאינם מגבילות את התפלגות על היריעה שתופסת באופן לא מפורש את מבנה היריעה. דוגמאות כוללות מודלי דיפוזיה, התאמת זרימה מותנית (conditional flow models), Wasserstein GANs (GANs), ו-VAE.

2. חוסר יציבות נומրית של שיטות מבוססות נראות

אחד התרומות התייארטיות המרכזיות היא ההוכחה שמודלים מבוססי נראות סובלים מחוסר יציבות מסוימת בלתי נמנע כאשר הם מנוטים למدل הדאטא הנתרך על יריעה. המחברים מדגימים שכאשר ציפיות המודל מנוטות להתרך על הירעה, פונקציות הנראות הופכת לבלי מוגבלות, מה שmobiel לפתרונות מנוגנים (זה קורה הרבה ב-VAE וב-GANים רגילים).

מתמטית, אם $X \sim P$ התפלגות הדאטא ב- \mathbb{R}^d בעלת תומך של ירעה M בעלת ממד פנימי $d < *^d$, עברו כל סדרה של מודלים מבוססי נראות $\{t, \theta_t\}$ המקבילים את התפלגות דאטא מתקיים:

$$\lim_{t \rightarrow \infty} p_{X, \theta_t}(x) = \infty \text{ for all } x \in M.$$

תוצאה זו מרמזת שציפיות במרחב הדאטא מתבדרות באופן מובהן כאשר הן מנוטות למدل התפלגות הנתמכות על ירעה, מה שהופך את היעדים מבוססי הנראות לביעיתיים עבור דאטא כזה כאליה.

3. מוגבלות מרחק KL:

המחברים מדגימים שמרחיק KL, יעד נפוץ לאימון DGMs, הופך ללא שימוש בלמידה היריעתית מתוערת כי KL מונית התפלגות חולקות אותה תומך (support). אולם כאשר משווים ציפיות של מודל במרחב הדאטא $\{\theta, X\}$ עם התפלגות דאטא P הנתמכת על ירעה, ה-KL הופך לאינסופי:

$$KL(p_X || p_{X, \theta}) = \infty.$$

תופעה זו מתרחשת כי \mathbb{P} מקצת הסתברות שאינה אפס רק לנוקדות על היריעה, בעוד $\{\theta, X\}_k$ מפזר מסת הסתברות על פני כל המרחב האומביינטי. כתוצאה לכך, מקסום הנראות, השקל למזעור את ה- KL , נכשל במתן אות למידה משמעותית.

4. מרחק וסדרתיין כיעד חלופי

כדי להתמודד עם מגבלות ה- KL , המחברים מקדמים את השימוש במרחקי וסדרתיין(זה עובד לא רע בגאנים כאמור), שנארים מוגדרים היטב גם כאשר התפלגיות יש תמיינות לא תואמות. מרחק וסדרתיין-1 בין התפלגיות k -q מוגדר כ:

$$W_1(p, q) = \inf_{\gamma \in \Pi(p, q)} \mathbb{E}_{(X, Y) \sim \gamma} [| | X - Y | |]$$

כאשר (p, q) מסמן את קבוצת ההתפלגיות המשותפות עם בעלות ההתפלגות שלית k -q. בניגוד ל- KL , מרחק וסדרתיין ממטר התכונות חלה, מה שהופך אותו לעד חסין לאימון DGMs בתರחיש היריעה.

5. פרשנות של מודלים לטנטים

המחברים מספקים פרשנות חדשה של DGMs לטנטים שקדום לומדים "יצוג במרחב נמוך של ירידת דатаה ואז ממדלים את ההתפלגות בתוך" יציג זה. הם מראים שמודלים אלה ממציעים באופן יעיל חסם עליון של מרחק וסדרתיין בין ההתפלגות המודל להתפלגות דатаה האמיתית:

$$W_1(p_X, p_{X,\theta}) \leq \text{Reconstruction Error} + \text{Distributional Divergence}$$

כאשר שגיאת השחזור מוגדרת עד כמה טוב היריעה שנלמדה מקרבת את ירידת הדטה האמיתית, והמרחב בין ההתפלגות מכמתה את ההבדל בין ההתפלגיות בתוך היריעה שנלמדה. תוצאה זו מספקת הצדקה תיאורטיבית להצלחה האמפירית של מודלי דיפוזיה לטנטים וגישות דו-שלביות אחרות.

תובנות מתמטיות

משפט חוסר הייציבות המספרית

המחברים מוכחים באופן פורמלי שבעור כל ההתפלגות הדטה P הנתמכת על ירידת כל סדרה של צפיפות מודל במרחב הדטה $\square Q$, פונקציית יעד בצורה של נראות לא מתכוונת מקסימום. תוצאה זו נגזרת מנינוח התנהוגות הצפיפות על ירידות במרחב נמוך תוך שימוש בתכונות של גיאומטריה דיפרנציאלית ותורת המידה (די כבד האמת).

מזעור מרחק וסדרתיין:

על ידי הצגת מודלים דו-שלביים כמרקבים של מרחק וסדרתיין, המחברים מבססים קשר בין למידת ירידת טרנספורט אופטימלי. תובנה זו לא רק מסבירה את הביצועים העדיפים של מודלי דיפוזיה לטנטים אלא גם מספקת מסגרת עקרונית לתכנון DGMs חדשים.

הסבר קרייסת מודים:

המחברים מראים שקריסט מודים ב-GANs ו-VAEs ניתנת להבנה כתוצאה של התאמת יתר לירעה, כאשר צפיפות המודל מתבדרת לאורך תתי-קבוצות של הירעה מבלי "لتפוא" את התפלגות הדאטה האמיתית.

מודל דיפוזיה:

ההצלחה של מודל דיפוזיה מיוחסת ליכולתם להתחשב באופן מroman במבנה הירעה על ידי פיזור מסת הסתברות מעבר לירעה. המוחברים מספקים ניתוח מפורט של מודל דיפוזיה מבוססי-ציוון וגרסאות חבות שלם.

סיכום

מאמר זה מספק חקירה קפדנית ועמוקה של DGMs דרך עדשת השערת הירעה. על ידי זהה המגבילות של שיטות מבוססות-גראות והדגשת היתרונות של מרחקי וורשטיין ומודלים לטניטיים, המוחברים סוללים את הדרך לפיתוח מודלים גנרטיביים עילים יותר.

<https://arxiv.org/abs/2404.02954>

06.02.25 SMALL LANGUAGE MODELS: SURVEY, MEASUREMENTS, AND INSIGHTS

תמצית:

המאמר זה חוקר את החשיבות הגוברת של מודלי שפה קטנים (SLMs) ומשווה את התפתחותם ל-LLMs. בעוד ש-LLMs דורשים משאבי מחשוב משמעותיים ובדרכן כלל מופעלים בשרתים, SLMs מתוכנים לפעול במכשירים מוגבלים, משאביהם כמו מחשבים ניידים, טאבלטים, סمارטפונים ומכשירי IoT. המחקר מציע סקירה מקיפה של 59 מודלי SLM, מעריך אותם על בסיס התקדמות ארכיטקטוניות, אלגוריתמי אימון ויעילות הסקה. באמצעות קידום אימוץ SLM, העבודה זו שואפת להפוך את הבינה המלאכותית לנגישה, זולה ויעילה יותר לשימוש מעשי.

אני תרגם את הטקסט לעברית:

ניתוח טכני

1. **חידושים ארכיטקטוניים** מאמר המחקר בוחן לעומק את האלמנטים הארכיטקטוניים המבדילים בין SLM-LLM, תוך הדגשת השינויים המשפרים את היעילות במכשירים עם משאבים מוגבלים.

- **מנגנון self-attention:** באופן מסורתי, attention מרובת-ראשים (MHA) הייתה המנגנון הדומיננטי ב-LLM. עם זאת, המאמר מזהה מעבר הדרגתית לטכניקת attention מבוססת-קבוצות (GQA). גרסה זו מפחיתה את המורכבות החישובית על ידי שימוש יצוגי שאלות בין ראשים תוך שימוש ביצוגי מפתח-ערך. הדוח מספק ראיות לכך שמודלי GQA, כמו Qwen2.5, בעלי מושגנות גבוהה על אלה עם MHA מבחרית ה-latencies ויעילות זיכרון, במיוחד בשלב ההיסק (אינפורך).

- **רשתות feed-forward:** האבולוציה הארכיטקטונית מראה העדפה ל-FFNs Gated (שהה שילוב של שני FFNs - אחת עם אקטיבציה א-לינארית בסוף) על פני FFN סטנדרטיות. Gated FFN מפעיל באופן סלקטיבי חלקים מהרשות, מה שmobilit ליעילות פרמטרית טוביה יותר. נמצא מעניין הוא השימוש ביחסים הביניים ב-FFNs Gated, (כלומר המימד של השכבה הלינאריות הראשונה שם) הנעים בין פי 2 ל-8 מהமמד החבוי, כאשר יחסים גדולים יותר משפרים בכך כל את הדיק בנסיבות היסק מורכבות.

- פונקציות אקטיביות:** נצפה מעבר משמעותי-m-GELU (שהה Unit Linear Unit (SiLU) (שהה Sigmoid Linear Unit (SiLU). המאמר מצין כי SiLU ייעלה מבחינה חישובית ומתאים יותר למודלים מוקוונטיטים, שלוטת במודלים שהושקו ב-2024).
- נרטול שכבות:** מודגש המעבר מ-m-GELU LayerNorm-RMSNorm. RMSNorm מפחית את העומס החישובי על ידי ביטול הצורך בחישוב הממוצע במהלך הנוורמליזציה, יתרון משמעותי במכשורי קצה.
- גודל מיליון:** גודלי מיליון של SLM גדלו משמעותית, ולעתים קרובות עלולים על K 50 טוקנים. המחברים מקשרים עלייה זו עם יכולות משופרות בהבנת שפה.

ניתוח דאטאסתים לאימון:

- מגמות בשימוש בדאטאסטים:** המחקר מתעד מעבר לדאטאסטים כלליים נפוצים כמו *The Pile* ו-*RefinedWeb* לדאטאסטים מאוגדים כמו *uEd-Edu* ו-*DCLM*. דאטאסטים החדשניים הללו משלבים טכניקות סינון מבוססות-מודול המשפרות משמעותית את איכות DATA.
- איכות DATA לעומת כמות:** למרות ההסתמכות המקדמת על נפח DATA גדול, הדוח מוצאת שדאטאסטים באיכות גבוהה מניבים ביצועי מודול טוביים יותר, גם עם פחות טוקנים. לדוגמה, מודלים שאומנו על *uEd-Edu* מושגים דיוק תחרותי בהשוואה למודלים מסחריים (סגור-קוד) מתקדמים.
- כמה טוקנים צריך לאימון:** המחברים מצינים מגמה מפתיעה: SLM רבים מאומנים על מספר טוקנים הגadol בהרבה ממה שחוק צ'ינצ'ילה מציע. לדוגמה M 500 - Qwen2.0 מאומן על 12 טריליאון טוקנים, בעוד ש-B.0 - 1.5B Qwen2.0 מאומן על 7 טריליאון בלבד. אסטרטגיית "אימון-היתר" המכונה זו מוצגת כאופטימיזציה לנסיבות מוגבלות-משאבם, המאפשרת למודלים להקליל טוב יותר כאשר הם מופעלים במכשורים עם כוח חישוב מוגבל.

3. חידושים באלגוריתם האימון

- שיטת P-un - Maximal Update Parameterization:** בשימוש במודלים כמו Cerebras-GPT, שיטת P-un מבטיחה אימון יציב על ידי בקרה על אתחול, קבוע למידה בכל שכבה, ועוצמות ראות אקטיביות. טכניקה זו מאפשרת להعبر היפר-פרמטרים שאופטמו עבור מודלים גדולים ישירות למודלים קטנים יותר, מה שמייעל את תהליך האימון.
- זיהוק ידע (דיסטילציה):** LaMini-GPT ו-2-Gemma מנגלים טכנית זו להעברת ידע ממודלי מורה גדולים למודלי תלמיד קטנים יותר, מה שmobail לביצועים משופרים ללא צורך באימון נרחב.
- אסטרטגיית אימון מקדים דו-שלבי:** אומצה על ידי MinICPM, אסטרטגיה זו כוללת שלב ראשון עם DATA באיכות גבוהה ולאחר מכן פין-טינון עדין עם DATA באיכות גבוהה וופצייפים למשימה. השיטה מוכיחה את עצמה כיעילה באיזון בין יעילות חישובית לביצועי המודל.

4. הרכבת ביצועים

היסק מבוסס שלל ישר: מודלים כמו Phi-3-mini משיגים ביצועים מתקדמים, (המתחרים ב-B7LLaMA 3.1). תוצאות הבנץ'マーク מגלות שהשיפורים באיכות מערכת הדטה ואסטרטגיות האימון אפשרו ל-SLM לצמצם את הפער עם מודלים גדולים יותר.

פתרון בעיות ריזונינג: Phi-3-mini-SLM אחריהם בעלי ביצועים גבוהים מציגים שיפור של 13.5% ביצועים משנת 2022 עד 2024, לעומת עליום על קצב השיפור של מודלי LLaMA. זה מדגים את הבשלות הגוברת של SLM בטיפול במקרים מורכבות.

מתמטיקה: ביצועי ה-SLM נשאים לתת-אופטימליים במתמטיקה, כאשר המודלים מתקשים לטפל במקרים הדורשות חשיבה לוגית. המחברים מייחסים פער זה למחרור בדאטהסטים באיכות גבוהה המתמקדים בלוגיקה.

למידת context-context: הניסויים מגלים ש-SLM מפיקים תועלות משמעותית מלמידה בהקשר, במיוחד עבור משימות כמו אתגר ARC, שם נצפים שיפורים בדיקן של עד 4.8%. עם זאת, חלק מהמודלים, כמו LM-1B1, LaMini-LM, מציגים הידדרות בביצועים עקב אוברפיט.

5. ניתוחיעילות בזמן ריצה

לייטנסי וזכרון נדרש: המבחן מוצא שהלייטנסי של היסק מושפעת הן מגודל המודול והן מהארכיטקטורה. לדוגמה, רץ Qwen1.5 31.9% מהר יותר מ-Qwen2.0 על המעבד Jetson Orin, למרות שיש לו 25.4% יותר פרמטרים. זה מייחס להבדלים במנגנון sharing attention וסטרטגיות שיתוף פרמטרים (parameter sharing). זיכרון היא לינארית בד"כ ביחס לגודל המודול אך מושפע גם מגורמים כמו גודל אוצר המילים ומנגנון sharing attention מודלים כמו LM-1B1, Bloom, שיש להם אוצרות מילים גדולים יותר, מציגים שימוש גבוה באופן לא פרופורציוני בזיכרון.

קוונטוט: טכניקות קוונטוט, במיוחד Q4 ביט, מוכחות את עצמן כיעילות בהפחיתה השהיה ושימוש בזכרון. שיטת KM מפחיתה לייטנסי בממוצע ב-50% במהלך היסק, לעומת שיטות כימות 3-ביט ו-6-ביט, הסובלות מחוסר יעילות חומרתית.

סיכום:

הניתוח הטכני המוצג במאמר זה מספק הבנה מקיפה של השיקולים הארכיטקטוניים, האימון, וזמן הריצה החזוניים לפיתוח ופרישה של SLM. על ידי התמודדות עם אתגרי הייעילות וمبرאות המשאבים, המאמר מציע תובנות חשובות לקידום מחקר ה-SLM וישומים מעשיים.

<https://arxiv.org/abs/2409.15790>

25.02.08. המאמר היום של מיליק -

Rejection Sampling IMLE: Designing Priors for Better Few-Shot Image Synthesis

היום עושים הפסקה קלה עם LLMs וסוקרים מאמר המציע שיטה מעניינת לאימון מודלי גנרטיביים במקרה שיש לכם מעט דedata לאימון. כדי מודלים גנרטיביים מודרניים כמו מודלי דיפוזיה,GANs,VAEs ועוד מצרכים כמהות עצומה של DATA אבל לפעמים אין לנו את הלקסוס הזה ואנו צריכים לאמן על כמהות קטנה של DATA. האם זה אפשרי בכלל?

התשובה על כך חיובית (פחות לפאי המאמר). המחברים מציעים שיטה הנקראת E-IMLE-RS לאימון מודל גנרטיבי עם מעט DATA שמשכל שיטת IMLE IMLE שזה Maximum Likelihood Estimation. בגדול מאוד IMLE Implicit Maximum Likelihood Estimation. היא דוגמת משתנה בעל התפלגות קלה לדגימה (gaussian) Z ומאמנת מודל גנרטיבי (רשת נירונית) כדי לנרט פיסת DATA. ההבדל הוא בפונקציית LOSS: עם IMLE LESS לאילם דגימה X מהDATAsett אנו ממצאים את רק המרחק בין נקודה Z ל- \hat{z} אחת בלבד: $C_0 - \frac{1}{2} \log(1 + e^{-\hat{z}^T z})$ שבה הינו קרוב ביותר אליה. כאן \hat{z} היא פיסת DATA שוגנתה מ- z ו- T זה המודל שאנו מאמנים.

כלומר בשלב הראשון של IMLE אנו דוגמים z נקודות ומעבירים אותם דרך מודל T (נקרא לו מיפוי בהמשך) ובונים z פיסות DATA מגנרטות. לאחר מכן לכל דגימה \hat{z} מדאטהsett האימון אנו בוחרים את z הקרובה

ב尤ר ל-z. בסוף רק נקודות כאלו משתפות במעבר של פונקציה ל-o. כמובן שמספר הנקודות זה המוגנרטות בשלב הראשון צריך להיות גבוהה משמעותית מאשר גודל הדאטסהט לאימון ח'. המטרה של שיטות אימון זו היא לאਪטם את המודל רק עבור הנקודות במרחב הלטנטי (z) שהן המומפות קרוב לנקודות מהדאטסהט.

הבעיה עם הגישה זו שהיא מוגדרת של הנקודות "הנבחרות" במהלך האימון כבר לא גואשית שעלול ליצור לנו בעיות באינפרנס כי אנו כן חוזים לדגום את z מהתפלגות גואשית. המרחק בין מיפוי T של דגימה גואשית מנקודה מהדאטסהט שונה בהתפלגות אחרת של הדגימה z המומפה היכי קרוב לנקודה זו (האמת זה די ברור). דרך אגב המאמר מוכיח את הטענה זו ומצביע שיטה להtagבר על זה.

השיטה שהמאמר מציע נראה ממש פשוטה אך מבוססת על ניתוח מתמטי ד' עמוק של התפלגות המרחקים. בשלב הראשון של האימון (אחרי הדגימה מהתפלגות גואשית) בוחרים את z-z כאשר נופלים במרחב יותר גדול מבוע אפסילון מכל נקודות בדאטסהט האימון אחרי המיפוי (כלומר יש לנו sampling rejection). לאחר מכן בדומה ל-IMLE, לכל נקודה x בדאטסהט בוחרים את z שהמיפוי שלו עם T נופל היכי קרוב אליה ומאמנים את T למצער את המרחק המומוצע בין z-z הנבחרים לנקודות העוגן שלהם. היפרפרמטרים החשובים כאן זה אפסילון ומספר נקודות z שנדגמות.

אינטואיטיבית זה עובד כי מלכתחילה אנו בוחרים נקודות רוחקות יותר (לאחר המיפוי) מהנקודות בדאטסהט שמאפשר לשמר התפלגות של הנקודות הנבחרות בשלב לאחר מכן קרובות לגואשית.

<https://arxiv.org/abs/2409.17439>

המאמר היומי של מאייק - 09.02.25 Why Is Anything Conscious?

מבוא:

המאמר המעניין מאת מייקל טימוטי בנט, שנ וולש ואנה צ'אלmers (David John Chalmers). אטגר פילוסופי זה מעלה את השאלה מדוע עיבוד מידע במערכות מסוימות, במיוחד ביולוגיות, מוביל לחוויות סובייקטיביות או *קוואליה*. המחברים מציעים שני פרדיגמה, המugen את התודעה בדינמיקה של מערכות self-organizing שעוצבו על ידי הברירה הטבעית.

הם טוענים כי תודעה תופעתית (phenomenal) - החוויה הסובייקטיבית של "air זה מרגיש" - אינה רק יסודית אלא הכרחית להתנהגות אדפטיבית. מענין כי באמצעות פרימירור חישובי פורמלי, המחברים טוענים נגד האפשרות של "זומבים", מערכות המתפקדות כמו בני אדם אך חסרות חוויה סובייקטיבית, ומצהירים באופן פרובוקטיבי כי "הטבע אינו אוהב זומבים". חוויה סובייקטיבית היא ההבנה המלאה והחווייתית של ההשפעה הרגשית והקוגניטיבית כאחד הנובעת מAustin שבו הבני אדם מבינים ופרשנים אירועים שנצפו או נחשו על ידי הם.

תרומות מרכזיות:

מסגרת מתמטית לאקטיביזם פנ-חישובי

המחברים מציגים מערכת פורמלית המוגנת ב*פנ-חישוביות* ו*אנקטיביזם* (Pancomputational Enactivism). פנ-חישוביות מניחה שכל המערכות הדינמיות מחשבות משהו, בעוד שאקטיביזם מדגיש את ההכרה כנובעת מאינטראקציות בין מערכת לסביבתה. האלמנטים המרכזיים במודל שלהם כוללים:

- סביבה: מוגדרת כקבוצת מצבים, עם מעברים המתוארים על ידי [תכונות דקלרטיבי](#).

- שכבת הפשטה: מבנה המגדר כך שמערכות מפרשנות היבטים סביבתיים.
- שימוש ומדיניות: מבני התנהגות המאפשרים קלט לפלט, המאפשרים התנהגות אדפטטיבית.
- זהויות סיבתיות*: ייצוגים של התרבות והשפה שלהן, חינויים למודעות עצמית.

הפרימורך מתאר כיצד מערכות מודעות שומרות על Kohärenz וסתגלות על ידי בנייה זהויות סיבתיות מורכבות יותר ויותר, המהוות בסיס למודעות עצמית.

היררכיה של תודעה

תובנה מרכזית היא התפתחות ההיררכיה של התודעה, המונעת על ידי ברירה טבעית ולהצטי סקללה. המחברים מתארים 6 שלבים מתקדמים:

1. מערכות לא מודעות: ישיות חסרות חוויה או הכרה, כמו שלעים.
2. מערכות מקודדות באופן קשייח: מערכות עם תగובות קבועות, מתוכנות מראש (למשל, חד-תאים).
3. מערכות לומדות: מערכות מסתגלות ללא מודעות עצמית (למשל, תולעים נמנודות).
4. מערכות עצמי מסדר ראשון: מסוגלות להבחין בין פעולות שנוצרו עצמאית לבין אירועים חיצוניים (למשל, זבובי בית).
5. מערכות עצמי מסדר שני: מסוגלות למטא-ייצוג ותקשורת מכונית (למשל, עורבים).
6. מערכות עצמי מסדר שלישי: ישיות [רפלקטיביות](#) במלוא המסוגלות לחשב על המודעות שלحن עצמו (למשל, בני אדם).

היררכיה זו מדגישה כיצד היבטים איקוטיים של תודעה מתפתחים באופן טבעי ככל שמערכות נעשות מסוגלות יותר למדל את עצמו ואת סביבתו.

עיבוד איקוטי וכמות:

המחברים טוענים כי *איקות קודמת לכמות* בעיבוד מידע. לפני שארגניזם יכול לתיג או למדוד מידע, עליו לחזות הבדלים איקוטיים. תודעה פונומנלית מתפתחת מכיוון שמערכות חיונות חייבות לסתוג ולתעדף מידע הרלוונטי להישרדות. סיווגים איקוטיים אלה מהווים את הבסיס לחוויה סובייקטיבית. טענה זו מתארת תיאוריות חיישוביות מסורתית, המתיחסות לעיתים קרובות לתודעה כתהיליך יציג טהור. על ידי הדגשת הקידימות של החוויה האיקוטית, המחברים מספקים פרספקטיביה רעננה על מקורות התודעה.

גישה עקרונות ראשוניים:

הפורמליزم במאמר נוצר משתי אקਸומות בסיסיות:

1. במקום שיש דברים, אנו קוראים לדברים אלה הסביבה.
2. במקום שדברים שונים, יש לנו מצבים שונים של הסביבה.

אקסזימוט אלה מובילות לצורה חסרת ייצוג של פנ-חישובית, בה מצלבים ומעברים מגדרים סביבות מבל' להניח מבנים פנימיים ספציפיים. המחברים ממסגרים ארגון עצמי יכולת להגביל פלטימ על בסיס קלטיים, ובכך להשיג התנהגות אדפטיבית.

דוחית זומביים

אחת הטענות המעניינות ביותר במאמר היא ש"הטבע אינו אוהב זומביים". המחברים טוענים שתודעה פונומינלית חיונית למודעות גישה ולהתנהגות אדפטיבית. תוכן "צוגי" - מה שאורגניזמים חושבים עלי'ו - נגזר תמיד מחוויה איקוית. لكن, מערכת המתנהגת כמו ישות מודעת חייבת בהכרח לחווות חוויה סובייקטיבית. טענה זו מתוגרת לשירות ניסוי מחשבה המציעים את קיומן של ישויות לא מודעות אך זהות בתנהגותן.

קשרים אמפיריים

המאמר מבוסס על ממצאים אמפיריים לגבי *רָה-אַפְרֶנְצִיה*, כולם היכלה לבחין בין גירויים שנוצרו עצמאית לבין גירויים חיצוניים. רה-אפרנציה, הנכפית ביונקים וחרקים, קשורה לייצור עצמי מסדר ראשון. המחברים גוזרים מבנה זה מעקרונות מתמטיים ומישרים את מסקנותיהם עם עבודתם של מרקר, ברון וקלין.

סיכום:

המאמר מציע גישה מסקרנת לבעה הקשה של התודעה על ידי עיגונה בברירה טبيعית, ארגון עצמי ופורמליזם חישובי. המוגרת ההיררכית של המחברים מספקת הסבר משכנע לאופן שבו תודעה מתפתחת ומדוע חוויה סובייקטיבית היא יסודית להתנהגות אדפטיבית. טענותם הפרובוקטיבית של זומביים הם בלתי אפשריים מתוגרת הנחות ותיקות, ומסמנת מאמר זה כתרומה משמעותית לחקר התודעה.

<https://arxiv.org/abs/2409.14545>

10.02.25 המאמר היום של מיק - On the expressiveness and spectral bias of KANs

מבוא:

המאמר שאסקור היום מציג חקירה עמוקה של רשתות קולמגורוב-ארנולד (KANs), ארכיטקטורה חדשנית המבוססת על משפט הייצוג של קולמגורוב-ארנולד. המחברים משווים באופן מדויק בין KANs לבין רשתות MLPs מסורתיות, הן מבחינה תיאורטית והן אמפירית, תוך התמקדות בהיבטים כמו אקספרסיבנו, יעילות ודינמיות אימון. המאמר מbasס תכונות תיאורטיות מרכזיות ומאמת אותן באמצעות ניסויים, ובכך מהווה תרומה משמעותית לתכנון רשתות ניירונים למשימות חישוב שונות.

אקספרסיבנו:

הישג מרכזי של עבודה זו הוא ההוכחה הפורמלית ש- KANs הן בעלות אקספרסיבנו לפחות כמו MLPs. המחברים מראים שככל MLP מושתתReLU ניתן "למפות" לארכיטקטורת KAN מקבילה, תוך שמירה על ייעילות

וללא הגדלה משמעותית בגודל הרשת. מנגד, בעוד ש-KANs ניתן לייצוג גם על ידי MLPs, טרנספורמציה זו הכרוכה בעלות משמעותית: מספר הפרמטרים גדול עם גודל גריד (מספר נקודות עוגן בספליין) של ה-KAN. נמצא זה מרמז ש-KANs עשויות להציג ייצוגים עילים יותר עבור סוגי מסוימים של פונקציות, במיוחד כאשר נעשה שימוש במבנה גריד עדינים.

המחקר מנצח תוצאות קיימות עבור MLPs כדי לקבוע כיצד קירוב לפונקציות עבור KANs במרחבים פונקציונליים שונים כמו מרחב סובולב. הוא מדגים ש-KANs משיגות קיצבי קירוב דומים או טובים יותר מאשר MLPs בשערוך פונקציות מורכבות, מה שמחזק את חוסן התיאורטי.

ניתוח הטויה ספקטרלית (spectral bias):

אחד ההבדלים המרכזיים בין MLPs ל-KANs המודגשים במאמר זה הוא ההבדל בהטיה הספקטרלית שלהם - תופעה שבה רשות ניירונים נוטות ללמידה תחילה בתדרים נמוכים של פונקציות. המחברים מציגים ניתוח תיאורטי ואמפירי מפורט, המראה ש-KANs סובלות פחות ממשמעותית מהטיה זו.

הבדל זה מיוחס לפונקציות האקטיבציה מבוססות -exp_{B} ולארQUITטורה הקומפוזיציונלית של KANs, המאפשרת להן ללמידה תדרים גבוהים ביעילות רבה יותר. תובנות תיאורטיות מציאות שдинמייקת האימון של KANs רדודות יותר ביחס לתדרים השונים בהשוואה ל-MLPs, שבהן נצפית התכנסות מהירה יותר של תדרים נמוכים. ההטיה הספקטרלית המופחתת הופכת את KANs למתאים יותר למושימות הדורשות שערוך פונקציות בעלות בתדרים גבוהים ממשמעותיים, כגון פתרון משוואות דיפרנציאליות ומידול תופעות פיזיקליות מורכבות.

מצאים אמפיריים:

- מבחני גרסיות תדרים:** KANs מצילות התאמים רכיבי גל בתדר גובה בו-זמןית, בעוד ש-MLPs מציגות קשיים מתמשכים עם תדרים גבוהים יותר גם לאחר אימון ממושך.
- ניסוי שדה גאוסי אקראי:** KANs עלות ביצועיה על MLPs בקירוב פונקציות שנדרגו מושדות גאוסיים גסים, מה שمعد על יכולת הסתגלות עדיפה לבני פונקציות מורכבים.
- פתרונות PDE:** בפתרון משוואות פואסון בתדר גובה, KANs משיגות שגיאות נמוכות יותר באופן עקבי בהשוואה ל-MLPs, תוך שמירה על ביצועים יציבים גם כאשר תדר הפתרון עולה.

טכנית הרחבת גריד (של הספליין):

חידוש טכני בולט הנדון במאמר הוא טכנית הרחבת גריד הייחודית ל-KANs. שיטה זו מאפשרת UIDON הדרגתית של גריד של -exp_{B} במהלך האימון, המאפשרת תהליך למידה יעיל יותר. גישת הרחבת הגריד מפחיתה את הסיכון ל-overfitting ומשפרת את יכולת ההכללה של הרשת, במיוחד כאשר מתחודדים עם פונקציות מורכבות או מערכי נתונים בעלי דגימה חסרה.

סיכום:

עבודה זו מבססת את KANs כחלופה חזקה ויעילה לרשותות MLPs, במיוחד למשימות בחישוב מדעי. על ידי התמודדות עם הטיה ספקטרלית, שיפור יכולות קירוב, וניצול שיטות אימון אדפטיביות, המחברים מספקים ראיות משכנעות לפוטנציאל של KANs לעלות בבעיה על רשותות נירונים מסורתיות ביישומים הדורשים למידת פונקציות בעלות תדרים גבוהים ומציגות יכולות קירוב משופרות. המוגרת התיאורטית בשילוב עם ניסויים מתקיפים הופכת מאמר זה לתרומה חשובה למחקר רשותות נירונים.

<https://arxiv.org/abs/2410.01803>

12.02.25 - המאמר היומי של מיק

STUFFED MAMBA: State Collapse and State Capacity of RNN-Based Long-Context Modeling

המאמר מספק חקירה מעמיקה של מצב כשל במודלים מבוססי RNN במידול שפה עם הקשר ארוך ומציע פתרונות לשיפור יכולות הכלכלה שלהם לאורכיים גדולים. המחברים מזהים ומנתחים תופעה בעייתית מאוד שקיבלה שם קריסת מצב (SC) - כשל של מודל בעקביה אחריו דינמיקת של הדאט המונע מרשותות RNN להכליל מעבר לאורכי האימון שלהם. הם מציגים סט של טכניקות מיטיגציה ללא אימון ואסטרטגיות אימון המשיכי המאפשרות למודל2 Mamba2 לעבוד עם מעל מיליון טוקנים בלבד מקרים מצב.

הגדרת הבעה:

מודלי RNN לעומת טרנספורמרים במידול הקשר ארוך

- טרנספורמרים מציגים ביצועים עדיפים במשימות המצריכות הקשר ארוך אך סובלים מסיבות חישובית ריבועית בגין אורך הסדרה בשל מגנון *attention*.
- מודלי RNN מציגות סיבות לנארית בגין אורך הסדרה, מה שהופך אותן לעילות חישובית בטיפול בסדרות ארוכות.
- מודלים בעלות סיבות לנארית בגין אורך הסדרה, מה שהופך אותן לעילות חישובית בטיפול בסדרות טוקנים) וכשלים בהכללה מעבר לאורכי האימון(זו הטענה במאמר)

ניתוח כשלים ברשותות RNN (וגם Mamba, RWKV) עם הקשר ארוך

כישלון בהכללה עבור סדרות ארוכות יותר: רשותות אלו מציגות הידרדרות חדה ביצועים כאשר נחשפות לאורכי סדרות מעבר לדאטה שאומנו עלי. כישלון זה אינו נובע פשוט מגרדיאנטים דועכים אלא מיוחס לקריסת מצב (SC).

קיובלות זיכרון קבועה: מכיוון RNNs שומרות על מצב זיכרון קבוע, יכולתן לשמור מידע היא מוגבלת מטבעה. קיימת מוגבלת עליונה על קיובלות הזיכרון ההקשרי - טוקנים מעבר למוגבלת זו נשחחים בהכרח.

2. ניתוח פורמלי של קрисת מצב (SC)

הגדרה וממצאים: קрисת מצב (SC) מתרכשת כאשר התפלגות המצב החבוי קורסת(מתנוונת), מה שmobail לכישלון המודל בעיבוד רצפים ארוכים יותר מקובצת האימון. המחברים מציגים ניסויים מבוקרים על Mamba2 וב>Showcases שערוצי מצב חבוי מסוימים מציגים התפוצצות של שונות, הגורמת ל:

- ערוצים(channels) חריגים דומיננטיים המדכאים ערכי מצב אחרים.

- חוסר יכולת לשכוח טוקנים מוקדמים, המוביל זיכרון.

- SC מرتبطה בעיקרה בפרפלקסיות(אי וודאות) מעבר לארוך האימון.

יחסות תיאורטי: פרמטריזציה יתר בדינמיקת המצב

המחברים מנסחים את משווהות עדכון המצב:

$$h_t = \sum_{i=1}^t \alpha_{i:t} \bar{B}_i x_i, \quad \alpha_{i:t} = \left(\prod_{j=i}^t \alpha_j \right) \in (0, 1)$$

כאשר \hat{z}_t הוא וקטור המצב החבוי, $\alpha_{i:t}$ המקדים מייצג את קצב דעיכת הזיכרון, x_i מיצג מידע חדש שהוכנו בזמן. כאשר מאמנים על סדרות באורך T_{train} פרמטרי המודל הנלמדים מעדיפים "לשמר את כל המידע בתוך T_{train} ", ועקב כך נכשלים בעת עיבוד סדרות ארוכות יותר. זה מוביל לצבירת יתר של מידע, ששובילה להרואה ובסופה של דבר ל垦ристות מצב.

3. אסטרטגיית התמודדות נגד SC:

טכניקות הת ללא אימון של SC:

שכחיה מבוקרת: הגדלת דעיכת ייצוג מצב(חבי) על ידי שינוי גורם הדעיכה α_t והפחיתה "עוצמת הכנסה" של מידע חדש B_i (ייצוג של טוקן). צעדים אלו גורמים למודל לשכוח טוקנים ישנים באופן אפקטיבי, מונע מיצוג הזיכרון להגעה להרואה (ערכים גבוהים מדי).

נרטול מצב: החלת אילוץ מבוסס נורמה על ייצוג המצב החבוי (מחלקים את וקטור הייצוג בנורמה שלו אם היא גדולה מדי):

זה מונע התפוצצות של ייצוג מצב חבוי אך מכניס אי-LINאריות, המשפיעה על יעילות האימון (לא ניתן לקבל את החישובים).

$$\begin{aligned} \hat{h}_t &= h_{t-1} \bar{A}_t + \bar{B}^T x_t \\ h_t &= \begin{cases} \hat{h}_t p / \|\hat{h}_t\| & \text{if } \|\hat{h}_t\| > p \\ \hat{h}_t & \text{if } \|\hat{h}_t\| \leq p \end{cases} \end{aligned}$$

עדכן וקטור ייצוג המצב עם sliding window: ניסוח מחדש של כלל עדכון ייצוג המצב לשימושה של מנגן **sliding window**:

$$h_t^{(r)} = \sum_{i=t-r+1}^t \alpha_{i:t} \hat{R}_i = \sum_{i=1}^t \alpha_{i:t} \bar{B}_i^T x_i - \alpha_{t-r+1:t} \sum_{i=1}^{t-r} \alpha_{i:t-r} \bar{B}_i^T x_i = h_t - \alpha_{t-r+1:t} h_{t-r}$$

זה מסיר טוקנים יסונים באופן אפקטיבי מוביל לחשב מחדש מאפס. ישם לארכיטקטורות אחרות כמו RWKV ו-RetNet.

המשך אימון על רצפים ארוכים יותר: המחברים מרחיבים את אורך DATA האימון מעבר ל"קיובות ייצוג המצב" כדי לאפשר את המודל ללמידה כיצד לשוכן בהדרגה. הם מאמתים אמפירית שעבור כל גודל ייצוג מצב S, קיימ סף אורך אימון שבו SC לא מתרחש.

4. סיכום:

- המחבר השיטתי הראשון של קרייסט ייצוג מצב (SC) בראשות "דמויות" RNN עם אורך הקשר ארוך SC. מתבטה בכך שוקטור ייצוג המצב מגיש לרווח (ערכים גבוהים) וזה גורם להידרדרות רצינית בביטוי המודל. המאמר מציע 3 שיטות מיטגיצה ללא אימון לביטול SC עד מיליון טוקנים. המחברים הציעו ביסוס אמפירי לחבר בין **גודל ייצוג המצב לקיובות המודל. לבסוף הם אימנו מודל Mamba2 בעל 370M פרמטרים עם אחזור מושלם של K 256 טוקנים - הרבה מעבר לכליות של מודל סטנדרטי מסווג זה.

<https://arxiv.org/abs/2410.07145>

המאמר היומי של מילק - 13.02.25

One Initialization to Rule them All: Fine-tuning via Explained Variance Adaptation

היום נסקור קצירות מאמר המציע שיטת LoRA לשיפור של טכניקת טיבוב (fine-tuning) של LLMS. כמו שאתם בטח זכרם LoRA נוספת לממשק המודל (בשכבות מסוימות) מטריצה נלמדת בעלת ראנק נמוך משמעותית נמוך יותר מהימיד של מטריצה המשקלות. משקלות המודל נשארות קבועות (לא מאומנות) במהלך הטיבוב.

המחברים מציעים ארכיטקטורת LoRA המכיל שלב מקדים שנקרה במאמר אתחול Date-Driven. מטרת אתחול זה היא "להתאים את הראנק של מטריצות של LoRA לכל שכבה של המודל". הרי אם אנו מאמינים בתוספת משקלים מסוימת (במהלך אנו יכולים לפזר אותן בצורה "אופטימלית" בין שכבות המודל. האופטימליות כאן נמדדת באמצעות השונות של האקטיבציות של השכבה (כלומר הפלט של שכבת FFN) עבור הדאטה שאנו מאמינים עליו).

הר' אם שונות האקטיבציות על DATA האימון היא נמוכה זה אומר שערכי השכבה פחות או יותר קבועים ולא כדי לבזבז עליה את המשקלים של LoRA. ככלומר אפשר להשתמש ב-LoRA בעלת ראנק נמוך מאוד (אם בכלל) לשכבה זו. אבל איך ניתן למדוד את השונות זו באמצעות ערכים סינגולרים של מטריצות האקטיביות המוחשבים באמצעות פירוק SVD של מטריצה זו. מימדיים מטריצת האקטיבציה כאן היא המימיד החובי של המודל וגודלו הביאץ'.

از מחשבים את הערכים הסינגולרים של מטריצת האקטיבציות על DATA האימון עד שהוקטורים הסינגולרים (הימניים מתיצבים). וקטוריים אלו מתעדכנים במהלך הריצות הבאץ' (המודל מתאים) ויצירתם (של הוקטוריים) נעצרת כאשר הם מתיצבים ומפסיקים להשתנות באופן מהותי (המאמר מודד את הדמיון באמצעות מרחק קוויין - אם הוא גבוה מדי עבור שכבה מסוימת מפסיקים את עדכון הוקטוריים עבור שכבה זו (אימון זה המבוצע לפני LoRA).

לאחר שהוקטורים הסינגולריים התכנסו עבור כל השכבות, לוקחים את הערכים העצמיים ומחשבים את אחוז השונות המוסבר על ידי כל שכבה (מחושב על ידי סכום הריבועים של הערכים הסינגולריים שלהם) ביחס לשונות המוסברת על ידי כל המודל (שהיא סכום הריבועים של הערכים הסינגולריים עבור אקטיביות של כל שכבות המודל).

בשלב הבא מיקצים את הראנקים של מטריצות LoRA לשכבות שפונקציות של השונות המוסברת על ידי. ככלمر כל השונות המוסברת של שכבה עולה, מיקצים יותר ראנקים של LoRa. בשלב האחרון מאמנים LoRa עם הקצהה "אומטימלית" של ראנק מטריצות LoRA בהתבסס על DATA האימון. רעיון די מעניין שמרתא תוצאות לא רעות.

<https://arxiv.org/abs/2410.07170>

15.02.25 - A Spectral Condition for Feature Learning

1. מבוא

המאמר מציג מסגרת תיאורטית להבנת מיידת מאפיינים (feature learning) ברשותנו נוירונים עמוקות דרך חקר הסקלאלת הנורמה הספקטרלית של משקלות אקטיביות הרשת. המחברים מציגים תנאים עבור סקלאלת ספקטרלי השולט בהתפתחות המאפיינים המופקים על ידי הרשת במהלך האימון, ומספקים אסטרטגיה לבחירת סקלאלות של משקלות וקצב למידה המבוססות על אינטואיציה בלבד.

המוטיבציה המרכזית של עבודה זו היא להתמודד עם אתגר מרכזי באימון רשתות רחבות (עומקות): הבטחת מיידת מאפיינים אפקטיבית בכל השכבות, תוך מניעת דעיכת או התפוצצות הגרדיאנטים. המחברים טוענים כי באמצעות סקלאלת נורמה ספקטרלית מדיקת של מטריצות המשקלות ועדכוניהן, ניתן לשמר למיידת מאפיינים גם בגבול עבור רשתות בעלי מימדים חבויים. מסגרת זו מספקת גישה מבוססת יותר (מתמטית) בהשוואה לאופן אתחול משקלות מסורתיות המבוססות על נורמת פרובניאו או בחירתם פר משקל (כמו).

המאמר תורם הן להיבטים התיאורטיים והן להיבטים הפרקטיים של אימון רשתות נוירונים בכך שהוא מגדים כיצד שיקולי נורמה ספקטרלית מוביילים באופן טבעי לשיטה \mathcal{P} – Maximal Update Parametrization. אסטרטגיית אתחול וסקאלת קצב למידה המאפשרת העברת היפרפרמטרים ממודלים קטנים לרחבים. בשונה מחקרים קודמים שהסבירו את \mathcal{P} באמצעות ניחוחים טנזוריית מרכיבים, המאמר מספק הוכחה פשוטה יותר המבוססת על אלגברת לינארית, מה שהופך אותו לנגיש יותר עבור קהילת לימודי העומק.

2. תרומות מרכזיות וייסוד תיאורטי

2.1 תנאי סקלאלת הספקטרלי

הממצא המרכזי של המאמר הוא תנאי סקלאלת על הנורמה הספקטרלית של מטריצות המשקל ועדכוני הגרדיאנט שלhn:

$$\|W_l\|_* = \Theta(\sqrt{n_l/n_{l-1}}), \quad \|\Delta W_l\|_* = \Theta(\sqrt{n_l/n_{l-1}})$$

כאשר $\{1-h\}$ ו- h מסמנים גודל הקלט והפלט (fan-in/out) בשכבה i ו- * מסמן הנורמה הספקטרלית של W . תנאי זה חיבר להתקנים עبور כל שכבות הרשת

תנאי זה מבטיח כי גם גודל הפיצרים החבויים h וגם עדכוניהם Δh (כמפורט מ珥יך הגראדיאנט) ישארו בסקירה מתאימה:

$$\|h_i\|_2 = \Theta(\sqrt{n_i}), \quad \|\Delta h_i\|_2 = \Theta(\sqrt{n_i})$$

ובכך נמנעות הן דעיכה והן התפוצצות של מאפיינים, תוך שימור דינמיות למידה יציבה לאורך כל שכבות הרשת.

המודיבציה לתנאי זה נובעת מהאופן שבו מידע "זורת" ברשותנו נירוניים. בשיטות אתחול מסורתיות כמו Xavier או Kaiming, נעשה שימוש בnorms פרוביניוס לשיליטה בגודל האקטיבציות. אולם, המחברים טוענים כי דוקא הנורמה הספקטרלית – המודדת את הערך הסינגולרי הגדול ביותר של המטריצה – מספקת אינדיקציה מדויקת יותר להשפעת השכבות על אותן הקלטים.

2.2 ביסוס מתמטי של למידת מאפיינים

תנאי הסקירה הספקטרלי נוצר מתכוна יסודית של רשתות עמוקות: כל שכבה מבצעת טרנספורמציה המגבירה או מחלישה את אותן הקלטים בהתאם לעריכים הסינגולריים של מטריצת המשקל שלה. הערך הסינגולרי הגדול ביותר (הנורמה הספקטרלית) קבוע עד כמה השכבה מסוגלת למתוח או לכוץ את האקטיבציות לאורך כיוונים מסוימים למרחב התוכנות.

המאמר מוכיח כי כאשר הנורמה הספקטרלית מקיימת את תנאי הסקירה שהוגדרו קודם, מתקיימים התנאים הבאים:

- עוצמת פיצרים נשמרת לאורך השכבות, מונעת דעיכה או התפוצצות.
- התפתחות המאפיינים במהלך האימון נשמרת משמעותית, ומונעת קרייסה לייצוגים טריוויאליים.

להוכחת טענה זו, המחברים מבצעים ניתוח מתמטי עמוק של עדכוני הגראדיאנט ב-MLPs (זהה perceptron multi-layer). נקודת מפתח היא שעדכוני המשקלות ברשתות עמוקות הם בעלי ראנק נמוך הנובע מhayot הגראדיאנטים מכפלה חיונית (outer product) של וקטורים:

$$\Delta W_i = -\eta_i \Delta w_i L = -\eta_i \cdot (\text{error signal} \cdot (\text{input features})^T)$$

מבנה זה מוביל לתובנה חשובה: עדכוני המשקל מתיחסים באופן טבעי עם הווקטורים הסינגולריים הדומיננטיים של מטריצות המשקלות, מה שמדגיש את חשיבותה הנורמה הספקטרלית בקביעת דינמיות הרשת.

2.3 קשר לשיטת פרמטריזציה P

אחת התוצאות המרכזיות של המאמר היא החיבור לשיטת P . פרמטריזציה זו קובעת כללי אתחול וסקאלת קצב למידה המאפשרים העברת היפרפרמטרים ממודלים צרים לרחבים מלבני לדרישים ניולים חדשים. המאמר מוכיח כי P יכולה לישום תנאי הסקירה הספקטרלי, עם סקלות אתחול וקצב למידה מהצורה:

$$\sigma_\ell = \Theta\left(\frac{1}{\sqrt{n_{\ell-1}}} \min\left\{1, \sqrt{\frac{n_\ell}{n_{\ell-1}}}\right\}\right), \quad \eta_\ell = \Theta\left(\frac{n_\ell}{n_{\ell-1}}\right)$$

כלומר, במקומות להשתמש בחוקים מובוסי אינטואיציה, ניתן לגזר את \mathbb{P} מתוך שיקולי נורמה ספקטרלית. יתרה מכך, המחברים מציעים גישה מאוחדת שאינה מצריכה כללים מיוחדים לשכבות קלט, חבויות או פלט, ובכך מפשטים את היישום של \mathbb{P} .

3. מסקנות

המאמר מספק תובנות מעניינות בנוגע למידת מאפיינים ברשות רוחות באמצעות ניתוח נורמה ספקטרלית. התנאי הספקטרלי שהוצע מספק מסגרת מאוחדת המסביר ומשפרת פרמטריזציות קיימות כמו \mathbb{P} . המחקר מציב על כך שסכמאות אתחול מסורתיות עשויה להפיק תועלת משמעותית מהסתמכות על נורמה ספקטרלית, דבר שעשוי לשפר את יציבות האימון ואת הביצועים של רשות נירוניים עמוקות.

<https://arxiv.org/abs/2310.17813>

16.02.25 המאמר היומי של מיק -

Representation Alignment for Generation: Training Diffusion Transformers is Easier than you Think

ЛОЖИЧНЫЙ ПОСЛАННИК ЗАМЕЧАЕТ, ЧТО ВЫСОКОКАЧЕСТВЕННЫЕ МОДЕЛИ ДИФУЗИИ ГЕНЕРАТИВНЫХ ИМЕННОСТЕЙ (LDMs) И СОВРЕМЕННЫЕ МОДЕЛИ УЧЕНИЯ МОДАЛЕЙ (DIFFUSION TRANSFORMERS) МОГУТ БЫТЬ ПОДДЕРЖАНИЕМ ДЛЯ ОБУЧЕНИЯ НОВЫХ МОДЕЛЕЙ. МОДЕЛЬ ПОДДЕРЖИВАЕТ СВОИМ УЧЕНИЕМ НОВЫЕ МОДЕЛИ, ЧЕМУ ПРЕДЫДУЩИЕ МОДЕЛИ НЕ СПОСОБНЫ. ВЫСОКОКАЧЕСТВЕННАЯ МОДЕЛЬ МОГЛА БЫ БЫТЬ ПОДДЕРЖАНА НОВЫМ УЧЕНИЕМ, ЧЕМУ ПРЕДЫДУЩИЕ МОДЕЛИ НЕ СПОСОБНЫ.

נתחיל מרךע קצר על מודלי דיפוזיה גנרטיביים. מודלים אלו מאמנים לגנרט תМОונות (למשל בהינתן תיאור טקסטואלי) על ידי הסרת הדרגתית של הרעש. המודל מתייחס מרעיש טהור (בד"כ גאוסי) ולאט הופכים אותו לתמונה (או פיסת>Data מדומין אחר). המודל מאמין על תМОונות מורעשות עם רמות שונות של רעש (=איטרציות) כאשר באימון המודל לומד להסר קמות קטנה של רעש (איתרציה t לאיתרציה $1-t$). בחירה של היפר-הפרמטרים t של תהליך ההרעשה היא מרכיב קריטי לאיכות גנרט של המודל המאמן.

תהליך זה(הרעשה) ניתן לתאר באמצעות משוואות דיפרנציאלית של זרימה הסתברותית (probability flow) המתאר השינויים (גרדיאנט) הדatta המורעש עם קצב/מהירות הרעשה (velocity) (v) שנסמן אותו v_t (הפתרון של משוואאה זו מתפלג לפי התפלגות של הדatta המורעש). קצב הרעשה ניתן לשער עם המודל (=רשף) בהתבסס על דגימות הדatta המורעש $-v_t$. לאחר מכן ניתן לפתור את משוואאות הזרימה ההסתברותית עם השערוף של v_t (בכיוון הפוך - מלומר החל מרעיש טהור) עם שיטת איולר למשל. שיטות אלו נקראות stochastic interpolation and stochastic integration.

מצין שיש שיטות המבוססות על פתרון נומי של משוואאה דיפרנציאלית סטוכסטית שמתארת את השינויים הדatta כפונקציה של פונקציית score שהוא לוגריתם של פונקציית התפלגות של DATA מורעש.

אוקי', אחרי הסיבור הזה הח'ים נה"ם קצת יותר קלים. מודלי דיפוזיה היום הם לרוב מודלים לטנטומים כאשר הגנרט מתרחש למרחב היצוג של הדatta. ככלומר המודל מאמין לשחרר ייצוג לטנטוי מרעיש ואז מפעלים את הדקORDER כדי לבנות תМОונה מהיצוג המשוחזר. היצוג של התМОונה ההתחלהית נוצר על ידי האנקודר. המחברים

טוענים שהייצוגים הלטנטיים המורעשים אינם "חזקים מספיק" ככלומר פחות משקפים את האספקטים הסמנטיים של התמונה.

המחברים מציעים להעיר את הייצוגים האלה על ידי הוספה של איבר רגולרייזציה שמטרתו לקרב ייצוגים אלה (של התמונה המורעשת) לייצוג המופק על ידי אנקודר חזק (כמו DINOv2). גם זה מתווסף ללוס הריגל של מודל דיפוזיה ונטען במאמר זהה משפר את יכולת התמונה המגנרטות וגם תורם ליציבות האימון.

<https://arxiv.org/abs/2410.06940>

המאמר היום של מיק - 18.02.25

THINKING LLMS: GENERAL INSTRUCTION FOLLOWING WITH THOUGHT GENERATION

סקירה מס' 400 - כדי לא להכביר עלייכם יותר מדי בחרתי מאמר קיליל יחסית והסקירה הולכת להיות ביל' נסחאות ודי קצרה. המאמר מציע שיטה קצרה במתוח דומה Optimization Group Relative Preference (GRPO) בקצרה שעשתה הרבה יותר מוגדרת לאחרונה. ותיקף אני הולך להסביר למה אני מתקoon כאן. רק אזכיר שהמאמר מציע שיטה להגברת יכולת הנמקה כללית של מודל ולא מתמקד רק בשאלות תכנות ובעיות מתמטיות.

המאמר מציע שיטת טיוב (fine-tune) למודל שפה המתקדמת בהקנייתם יכולת הנמקה (reasoning) למודל שפה ללא צורך בדעתה מתויג. המאמר מציע לבצע אימון בסגנון RLHF אבל להבדיל מהדיפיק (המציאה של GRPO), המחברים הציעו להשתמש בשיטת DPO שלא משתמש בפונקציית התגמול כלל. אצין-ש-GRPO לא מאמנת מודל תגמול (reward) כמו ש-PO PPO עשו אלא משתמשת בנקודות התשובה והפורט שלה כפונקציית תגמול.

از מה משותף בין GRPO לבין השיטה המוצעת במאמר? שניהם למעשה מוציאים לא לknos את המודל על תהליכי החשיבה (שעלול להיות לא נכון להוביל לתשובה הנכונה) אלא לשפטו אותו רק על בסיס נוכנות התשובה של המודל (כאמור GRPO גם קונס על אי עמידה בפורט של התשובה). אחרי שהבנו את הקשיים המהותיים של השיטה המוצעת עם השיטות המפורסמות בוואו נצלול למה שהמאמר מציע.

כאמור המאמר מציע לטיב יכולת הנמקה של מודל שפה ללא שימוש בדעתה מתייג עם RLHF. כמו שאתם זוכרים RLHF עם DPO דורש זוגות של תשבות מעודפות ופחות מעדפות. מכיוון שאמרנו שהשיטה לא דורשת דעתה מתייג אז אתם יכולים לנחש שבנית הזוגות נעשית על ידי מודל שפה שופט שבור תשבות טובות ורעות בדומה לשיטת RLAIIF זהה קיצור של Reinforcement Learning from AI Feedback על תשבות (ולא שרשרת הנמקה!) של המודל המאמון ומחייב מה בין תשבות היא הטובה והגראעה ביותר. זוגות אלו משמשים לאימון המודל בקורס O-DPO. כמובן שיש פה גם הנדסת של מטה-פרומפט הגורם למודל "לחשוב" אבל שרשרת חשיבה זו לא משתתפת באימון המודל.

<https://arxiv.org/abs/2410.10630>

המאמר היום של מיק - 20.02.25

Losing dimensions: Geometric memorization in generative diffusion

המאמר מציג מסגרת תיאורית חדשה להבנת האינטראקציה בין הכללה לזיכרון במודל דיפוזיה גנרטיביים מנוקדת מבט גיאומטרי. המחבר מושתמש בטכניקות של פיזיקה סטטיסטית, גיאומטריה דיפרנציאלית ותורת המטריצות האקרואיות כדי לנתח כיצד מודל דיפוזיה לומדים ומאבדים תלת-מרחבים של הדעתה בהתאם לגודל הדעתה, קיבולת המודל הגנרטיבי, ושיטת האימון.

חדשנות מרכזית:

1. קונספט הזיכרון הגיאומטרי:

המאמר מציג את הרעיון שזיכרון (memorization) במודלי דיפוזיה גנרטיביים אינו תופעה ביןארית (או הכללה או זיכרון מוחלט של הדאטה), אלא שהוא שתרחש באופן סלקטיבי בתת-מרקבים מסוימים של מרחב הדאטה. תובנה זו מרחיבה תיאוריית קיימות שפירשו את תופעת הזיכרון בתור "תהליך קריישה מוחלטת" של זיכרון נקודות דאטה מסוימות. המחברים מצאו כי תהליך הזיכרון מוביל לאובדן ממדים ביריעה הילטנטית, ולא קריישה לנוקודות דאטה פרטניות.

2. אובדן ממדים סלקטיבי ותלות בשונות:

המחברים מראים שתת-מרקבים עם שונות גבוהה נעלמים בשלב מוקדם יותר של האימון, בעוד פרדוקסלי המצביע על כך שהתכונות הבולטות ביותר של דאטה הן הראשונות להיפגע מהשפעות תופעת הזיכרון. מצאו זה חשוב משום שהוא מצביע על כך שזיכרון אינו משפיע באופן אחיד על כל הממדים, אלא מתרחש באופן מבני בהתאם לפיזור השונות של הדאטה.

3. שימוש בכלים מפיזיקה סטטיסטית ותורת המטריצות האקראיות:

המחקר מבוסס על תיאorias של מעבר פaza הנקו

הנקו

הן מפיזיקה סטטיסטית, במיוחד זכוכית (glassy phase transition) במודלי זיכרון אסוציאטיביים (dense associative memory) המואת וריאנט מודרני של מודל הופפילד (זה שקיבל פרס נובל לפני כמה חודשים). המאמר מספק ביטויים אנליטיים עבור דינמיקה להפרשים בין ערכיהם סינגולרים של מטריצת היעקוביאן של פונקציית-score (score) במהלך גנרטו

הנקו

הן דאטה (עם מודל מאומן), ומגדים כיצד פערם אלו נסגרים באופן הדרגי כאשר המודל מתחילה "לזכור" (memorize) את דוגמאות האימון. המחברים עושים שימוש באנלוגיה למודל האנרגיה האקראי (REM) לניתוח מעבר העיבוי (condensation transition), המתאר את הזמן הקритי(האיטרציה של הדיפוזיה) שבו זיכרון משתלט.

4. הבחנה בין score אמפירי ל-score מדוייק:

המאמר מבחין בין פונקציית-score תיאורטית (מדוייקת) לבין הפונקציה שמשוערת מתוך דאטה. פונקציית-score האמפירי מציגה תנודות סטטיסטיות התלוויות בגודל הדאטה ובזמן הדיפוזיה, באופן הקשור ישירות לזכרון הגיאומטרי. המאמר מכתת כיצד מספר דרגות החופש הגנרטיביות האפקטיביות מתפתח כפונקציה של גודל הדאטה וזמן הדיפוזיה(איטרציות של הדיפוזיה).

5. אימות אמפירי על דאטה סינטטי וריאלי:

המחברים מאשרים את התיאוריה שלהם באמצעות ניסויים על דאטה סינטטיים (למשל, רישות ליניאריות המכילים תת-מרחב בעלי שונות משתנים) ועל הדאטהסטים האמתיים MNIST, CIFAR-10, (, CelebA). תוצאות הניסוי מתיישבות היטב עם התוצאות התיאורטיות שפותחו במאמר, ומגדימות כי מודלי דיפוזיה מאומנים מציגים את האובדן הדרגי החזוי של הממדיות ככל שגודל הדאטה הופחת. המאמר מציג שיטה חדשה להערכת הממדיות הפנימיות של דגימות שנוצרו, מה שמאשר את ההשערה התיאורטית.

<https://arxiv.org/abs/2410.08727>

המאמר היומי של מייק - 22.02.25

When Does Perceptual Alignment Benefit Vision Representations?

משמעותיים בהפסקה קלה מודולרי שפה וסוקרים קצרות מאמור בתחום הראה הממוחשבת. המחברים מציעים שיטה לשיפור ייצאים (אמבידגנד) של דאטה ויזואלי קרי תМОנות באמצעות טיב (fine-tune) של האմբדר או באקבון (המודל המפיך ייצאים אלו). טיב זה נעשה על הדאטהסט של תМОנות המסומנות כדוגמאות (ולא דוגמאות) על ידי בני אדם. טיב זה הוכח (אמפירית) כתורם לאיכות הייצאים המופק למגוון משימות בתחום הבנת דאטה ויזואלי כגון ספירת עצמים בתמונה, סגמנטציה, שערוך עמוק, אוחזור עצמים (instance retrieval).

הטיב נעשה עם שיטת למידה ניגודות (contrastive learning) שכתבתי עלייה לא מעט וスクרטט עשרות מאמרים בנושא המրתק הזה. בגדול מאוד המטרה העיקרית של למידה ניגודית היא לקרב ייצאים של פיסות דומות (הקרבה נמדדת בד"כ על ידי מרחק קווין אבל יש עוד אופציות) ולהרחק ייצאים של פיסות דאטה לא דומות. המחברים השתמשו בדאטהסט NIGHTS המכיל שלישיות של תМОנות בעלות בערך אותן תוכן סמנטי, עם זאת, התМОנות בדאטהסט שונות בתמונה, צורה, צבע ומספר של חפצים.

המחברים משתמשים בגישה triple loss ללמידה הניגודית והמודול מאומן על שלישיות של תМОנות כאשר כל שלישיה מורכבת מתМОנות עוגן (reference) ושתי תМОנות נוספות בסופות שסומנו על ידי מתיאגים אנושיים בונגע לדמיון לתМОונת העוגן. כאן המתיאגים התבקו לסמן איזו ממשי התМОונת דומה יותר/פחות לתМОונת העוגן. המטרה של האימון היא למקסם את ההפרש בין מרחק של ייצוג של תМОנה (=טוקן CLS) דומה יותר לתМОונת העוגן לזה של התМОונה פחות דומה כאשר המרחק נמדד באמצעות מרחק קווין.

המחברים גם מציעים לבצע למידה ניגודית על השרשור של ייצוג התМОונה עם הייצוג הממוצע של כל הפאיים בתמונה. לטענתם זה משפר את התוצאות במידה מסוימת.

<https://arxiv.org/abs/2410.10817>

23.02.25 - המאמר היומי של מיק - Addition Is All You Need: For Energy-Efficient Language Models

מבוא:

המאמר מציג גישה אלגנטית אך רדייקלית לשיפור היעילות של רשותת נירונים, רלוונטיות במיוחד לשיפור ביצועים של LLMs. המחברים מציעים חלופה למכפלות נקודה צפה (floating point) מסורתיות (floating-point complexity) או L-Mul (Multiplication), אשר מזכיר פעולה עם נקודה צפה על ידי חיבור מספרים שלמים. הטענה המרכזית היא ש-L-Mul מפחית משמעותית את המורכבות החישובית ואת צרכית האנרגיה, תוך שימירה על ביצועי מודל כמעט זהים.

המוטיבציה:

דרישות "החסמל" של מערכות מבוססת AI, במיוחד מודלים גדולים, הופכות להיות יותר ויותר קשות. מכפלות נקודה צפה הן בין הפעולות החישוביות的关键יות ביותר (מחינת צרכית אנרגיה), והחלפתן באלטרנטיבות חסכנות יותר יכולה להיות בעלת השכלות משמעותיות על תכנון חומרה למגוון רחב של ישומי AI. המחברים מדגימים כיצד צרכית האנרגיה ברשותת נירונים עולה עם מספר פעולות הנקודה הצפה, ומכתמים את הפקחות האנרגיה האפשריות על ידי החלפת מכפלות בחיבורים.

הבסיס הטכני של L-Mul:

ככל נקודה צפה מסורתית כורח בפעולות ישרות של מערכי ומנטיות. L-Mul עוקף זאת על ידי ארגון מחדש של

הчисוב, תוך שימוש בחיבור של מספרים שלמים במקום כפל של מנטייסות. המחברים תומכים בכך עם הערכת שגיאה תיאורטית, המראה ש- Mul-L עם מנטייסה של 3 ביטים מוגיעה על מכפלת $8 \times 5m^2$, בעוד שעם מנטייסה של 4 ביטים הוא משתווה ואף מוגיעה על $3 \times 4m^3$. דיקון מתמטי זה מספק אמינות חזקה לטענותיהם.

ניסויים:

המחברים משלבים את Mul-L בתוך מודלים מבוססי טרנספורמר ומעריכים את יעילותו במגוון משימות, כולל הבנת שפה טבעיות, משימות הנמקה כלליות, ופתרון בעיות מתמטיות ועוד. "ישום Mul-L למנגנון h-attention מביא לאובדן דיקון זניח, ובמקרים מסוימים אף לשיפורים קלים בביטויים. המחברים אף מראים שהחלפת כל המכפלות בנקודה צפה במנטייסה של 3 ביטים בטרנספורמר מביאה לתוצאות דומות ל- $3 \times 4m^3$ float8 (fine-tuning) והן בזמן הסקה.

יתרונות וחסרונות:

אחד היבטים המשכנעים ביותר במאמר הוא ההתקדמות ביעילות אנרגטית. על ידי שימוש בנתונים מחקרים קודמים על צריכה אנרגיה בחומרה, המחברים מעריכים כי Mul-L יכול להפחית את עלות האנרגיה של מכפלות רכיביות ב-95% ואת עלות האנרגיה של פעולות מכפלה פנימית (dot product) ב-80%. זו טענה מרתקת לכת, המציעה כי Mul-L עשוי להיות השפעות מיידיות ומוחשיות על datacenters ויישומי AI בהיקפים גדולים.

המאמר מותיר כמה שאלות מעשיות ללא עונה. המחברים מעריכים בכך של GPUs קיימים אין תמיכה native ב- Mul-L , מה שמקשה על יישומו היעיל במערכות AI מודרניות. למרות שהמחברים רוחזים כי חומרה ייועדת יכולה לאפשר אופטימיזציה של חישובי Mul-L , הם אינם מספקים תוכניות קונקרטיות פיתוחה.

סיכום:

המאמר מציג גישה חדשה להקטנת הוצאות החישובית והאנרגיה של LLMs ורשתות נוירונים אחרות. הביסוס התיאורטי חזק, תוכאות הניסוי משכנעות, וההשפעה הפוטנציאלית משמעותית. בעוד שנותרים אתגרים מעשיים—במיוחד באימוץ חומרה—בעבודה זו פותחת דלתות חדשות לחישובי AI חסכניים באנרגיה. אם תשופר ותאומץ, Mul-L עשוי למלא תפקיד מרכזי בהפיכת AI לבר-קיימא מוביל לפגוע בביטויים.

<https://arxiv.org/abs/2410.00907>

25.02.25 המאמר היומי של מיק - Understanding Visual Feature Reliance through the Lens of Complexity

המאמר שאני סוקר היום מציג מחקר יוצא, נדריר ומעניין על מרכיבות פיצ'רים המופקים על ידי מודלים דיפ (AIN, RAG, סוכנים -LLMs שם :). מאמר זה קשור הדוקות לרעיון של צוואר הבקבוק של המידע ברשותות עצביות עמוקות, שטבע נפתחי תשב.

המאמר מציג מסגרת תיאורית-אינפורמציונית חדשה לכימות מרכיבות פיצ'רים במודלי דיפו ומציע גישה מתמטית להבנה פיצ'רים, מתי והיכן פיצ'רים מופיעים במהלך האימון. בנויגוד לשיטות מסורתיות שמתמקדות בסליננסי (saliency), [ושיר פיצ'רים \(attribution\)](#), המחקר מציע את מידת המרכיבות שקיבלה שם

information-i-v כמדד למורכבות חישובית, אשר מבטא את המאמץ הנדרש כדי לחלץ פיצ'רים במקומם לשערך רק את התלות הסטטיסטית הישירה שלה בקלט.

המחקר בוחן באופן שיטתי את התפתחותן בזמן אימון, התפלגותן המרחכנית ותפקידן של פיצ'רים במודלים ויז'. הממצאים מצביעים על כך שמודלי דיפ' מציגים תהילך למידה היררכי, שבו פיצ'רים פשוטים וдолות-מורכבות מופיעות מוקדם באימון ומתקדמות בדרך חיבורים residual, בעוד פיצ'רים מורכבים יותר דורשות עיבוד עמוק יותר וזמן אימון ארוך יותר או תורמות פחות משמעותית להחלטות הסופיות מהיה מקובל להניח.

גישה מבוססת למורכבות בלמידת פיצ'רים

ניתוח פיצ'רים בלמידה عمוקה התמקד עד כה בעיקר בחישוב החשיבות והשימושיות שליהן לשימה צו או אחרת, אך כמעט ולא בוצע ניסיון לכמת כמה מרכיב לחץ פיצ'ר מתוך DATA. מחקר זה משנה את נקודת המבט המסורתי בכך שהוא מציע ממדד למאמץ חישובי הדורש ללמידה פיצ'ר.

הגדרה מחדש של מורכבות פיצ'רים

שיטת מסווגית לשערוך פיצ'רים מסתמכות על שערוך מידע הדדי (mutual information) בין פיצ'ר לבין הדאטה. עם זאת, גישה זו אינה מביאה בחשבון את הקשי החישובי הכרוך בחילוץ הפיצ'ר.

החדשנות המרכזיית במאמר היא ההציג של information-i-v, המאפשר כימות של:

- כמה עיבוד דרוש כדי לחלץ פיצ'ר מתוך שכבות הרשת.
- עומק ועוצמת הטרנספורמציות הלא-לינאריות הנדרשות כדי להפיק את הפיצ'ר מקלט.
- סיבוכיות מיפוי של קלט למרחב הפיצ'ר, במקומות רק מדידת התלות הסטטיסטית שליה בקלט.

מדוע מורכבות חישובית חשובה בלמידת פיצ'רים?

מחקרים קודמים תיאוריית צואואר הבקבוק האינפורטטיבי (Information Bottleneck Theory) מצביעים על כך שמודלים עמוקים את היצוגים באופן הדרגי, תוך סיכון מידע לא רלוונטי ושימור אותן משמעויות לשימה. המחקר זהה מרחיב עקרונות אלו בכך שהוא מספק ממדד כמותי להערכת אילו פיצ'רים דרושים עיבוד עמוק ואילו מופיעות כבר בשלבים מוקדמים יותר של הלמידה.

הנתונים האמפיריים תומכים בטענה שהמודלים מעדים ללמידה פיצ'רים פשוטים וдолות-מורכבות בשלבים המוקדמים של האימון, בעוד שבפיצ'רים מורכבים יותר מופיעות רק לאחר זמן אימון ממושך יותר. תוצאה זו עולה בקנה אחד עם תיאוריית הלמידה המדורגת (Curriculum Learning), לפיהן משטח אופטימיזציה של מודלים עמוקים נוטה לטובת למידת תבניות פשוטות תחילה לפני המעבר לאבסטרקציות מורכבות יותר.

динמייקת הזמן של מורכבות פיצ'רים בזמן אימון

אחד הממצאים המרתקים ביותר במאמר הוא כי למידת הפיצ'רים מתרכשת בהדרגה על פני שלבי האימון:

- שלבי האימון הראשוניים: המודל לומד במהירות פיצ'רים דלי-מורכבות, אשר דורשות פחות טרנספורמציות לא-לינאריות.
- שלבי האימון האמצעיים: מתחילה להופיע פיצ'רים מורכבים יותר, המורכבות משילוב של תוכנות פשוטות מוקדמות יותר.

- שלבי האימון המאוחרים: פיצ'רים המורכבים ביותר מופיעים, אך תרומתן להחלטות המודל קטנה יחסית לעומת הפלגות מרכיבות הפיצ'רים במרחב הרשות הנירונית.

התפלגות מרכיבות הפיצ'רים במרחב הרשות הנירונית

המצאים מצביעים על כך שמרכיבות פיצ'רים אינה מפוזרת באופן אחיד על פני שכבות הרשות, אלא מאורגנת בצורה מבנית:

- פיצ'רים פשוטים מופיעות בשכבות המוקדמות ויכולות לתקדם דרך חיבור residual.
- פיצ'רים מורכבים דורשות עיבוד עמוק יותר ומצוירות בהדרגה דרך טרנספורמציות לא-لينאריות רבות.
- חיבור residual משמשים כמסננים חישוביים, ומאפשרים לפיצ'רים דלי-מורכבות לעקוף עיבוד עמוק שאינו הכרחי עבורן.

הקשר בין מרכיבות הפיצ'רים להחלטות המודל

אחד הממצאים המעניינים של המחקר הוא שפיצ'רים מורכבים משפיעים פחות על החלטות הסיווג הסופיות של המודל מאשר פיצ'רים פשוטים יותר.

- מודלים מסתמכים בעיקר על פיצ'רים פשוטים ויציבות לצורך הכללה.
- פיצ'רים מורכבים, אף שהן קיימות, אין חינוכיות להכרעת הסיווג.
- התמకדות יתר בפיצ'רים מורכבים אינה משפרת בהכרח את הביצועים, ועלולה להוביל לאוורפייט.

ממצא זה סותר את ההנחה המסורתית שלפיה מודלים עמוקים מסתמכים בעיקר על "יצוגים אבסטרקטיים" מאוד לשם קבלת החלטות. המאמר מציע כי המודלים מנצלים קודם כל פיצ'רים פשוטים ועמינדיים, ורק אחר כך משלבים מידע מורכב יותר כתוספת רפינמנט משנייה.

סיכום:

המאמר מציע תרומה תיאורית וא Empirical משמעותית להבנת כיצד מודלים עמוקים לומדים, מארגנים ומשתמשים בפיצ'רים שונים. תובנות אלו יכולות להשפיע על עיצוב ארכיטקטורות רשות, אסטרטגיות אימון, ופרשנות של למידת מכונה, תוך שיפור הייעילות והעמידות של מערכות בינה מלאכותית.

<https://arxiv.org/abs/2407.06076>

המאמר היום של מייק - 27.02.25

Unity by Diversity: Improved Representation Learning for Multimodal VAEs

היום אני חוזר למאמר על (VAE) Autoencoder ו-variational (VAE) אחריו תקופה ארוכה מאוד, יותר משנה, אני מניח.

נזכיר Sh-VAE הוא סוג של מודל גנרטיבי שלומד לדחוס נתונים לטנסי נמוך-ממד וモבנה (עם התפלגות מושראית), ולאחר מכן לשחרר אותם. האתגר המרכזי הוא להבטיח שמדובר זה "ישאר חלק" כך שדגימה ממנה תיצור דאטה ריאליסטי. כדי להציג זאת, VAE מażן בין שני יעדים: שחשזר מדויק של הדאטה המקורי תוך שמירה על כך שהמרחב הלטנסי יהיה קרוב(בתפלגות) להתפלגות פשוטה ומוגדרת היטב, בדרך כלל גאוסיאנית. זה מבטיח שנקודות סמוכות במרחב הלטנסי יתאימו לפחות דומות, מה שמאפשר יצירת דוגמאות חדשות珂הרנטיות.

מולטימודל VAEים מרחיבים את הרעיון זהה כדי להתמודד עם סוגים שונים של דאטא, כגון תמונות, טקסט ואודיו, במסגרת מאוחדת. הקושי עם MVAE נובע מכך שמודלים למודלים (modes) שונים חולקים מידע מסוים אך גם מכילים פרטים ייחודיים לכל מודל. ישנו מודלי MVAE שמנסים לכפות על כל המודלים לחלק ייצוג משותף אחד, מה שעולול להוביל לאובדן מידע ייחודי לכל מודל. אחרים שומרים עליהם מופרדים מדי, מה שמנע אינטראקציות משמעותיות בין המודלים. MVAE מושלם חyb למצוא את האיזון: לתפוס גם את המבנה המשותף של המודלים השונים תוך שמירה על מה שמייחד כל מודל.

התובנה המרכזית כאן היא דחיתת ההנחה שכל המודלים ייחו באותו המרחב הלטני. MVAEs קודמים פועלו תחת ההנחה כי יש להשתמש בייצוג לטני יחיד (מפוגג גאוס טנדראטי בד"כ) עבור כל המודלים (לכל-ה-מודאליות). גישה זו לרוח מובילה למרחב לטני משותף שהוא או "מחובר" מדי - מה שמאפשר מודלים שאינם תואימים להתרverb באופן לא טבעי או "חופשי" מדי, כך שהוא אינו מצליח ללמידה קשורים בין המודלים השונים. כדי להתגבר על כך המחברים מציעים ללמידה את התפלגות הלטנית מתוך הדטה עצמה, ובכך יוצר מרחב לטני שתחשב בניויאנסים הייחודיים של כל מודל תוך שמירה על יכולת העברת מידע בין מודלים שונים.

במקום להשתמש יעד קבועה (נגד גאוסית כמו ברוב המקרים), VAE MMVM בונה התפלגות לטנית בסגנון Mixture of Experts, המתחשב בכל המודלים (למודலיות השונים) בזמן האימון. כל מודל תורם להערכת התפלגות הזו, כך שהיא פועלת כהגבלה רכה ולא כהגבלת קשייה כמו ב-VAE הרגילים. זהו שינוי מהותי לעומת שיטות המבוססות על מיצוע/כפל של התפלגות - כאן מדובר בתפלגות דינמית, שתלויה בדטה ונתעכנת כל הזמן כי המודל לומד את המודים. למעשה, היא מתפקדת כמו "שלד הסטגלות", המעצב את המרחב הלטני כך שיתמוך במבנה משותף, אך מבלי לכפות אותו.

באו נצלול לפרטים מתמטיים על איך כל זה נעשה בפועל. לLOSE השחזור (עד כמה טוב המודל מצליח לשחרר את הדטה) המחברים מוסיפים איבר רגולרייזציה שהוא סכום של JS divergences בין התפלגות הפוסטרוריות המשוערכות ($X|z$) ϕ_q של כל מודליות לבין התפלגות הממוצעת הנלמדת ($X|z$) ϕ_h המומוצעת על כל המודליות. כאן X מסמן את הדטה ו- z הינו הייצוג הלטני.

מה זה אומר בפועל?

- כל מודל (מודאליות) מתומך להישאר קרוב לתפלגות הממוצעת הנלמדת, שבבסיסה על כל המודלים יחד.
- כל מודל שומר על המבנה הייחודי בלתי תלוי שלו, מה שמנע קריסת מרחב הלטנטים לתת מרחב קטן מדי (הטענה במאמר).
- התפלגות הממוצעת ($X | z$) ϕ_h פועלת כרגולרייזציה דינמית ונלמדת, כך שהייצוגים נוטרים משמעותיים ושימושיים.

בסיום של דבר, VAE MMVM אינם מכתיב מבנה לטני קשיח (כמו MVAE) - במקום זאת, הוא מאפשר לבניה להפתח באופן טבעי מתוך דטה עצם. וזה בדיק מה שהופך אותו לכל כך חזק.

המודל המוצע גם מציג ביצועים טובים במשימות השלמת דטה חסר. MVAEs מתקשים כאשר מבקשים מהם לשחרר מודל חסר מתוך קלט חלק. למה? כי הייצוגים המשותפים שלהם נוטים להיות נוקשים מדי - או שהם נשענים באופן מוגזם על מידע משותף (מה שmobiel לדורות גנריים ומטוושטים) או שהם שומרים על ייצוגים נפרדים שאינם מקיימים אינטראקציה משמעותית. לעומת זאת, VAE MMVM מבטיח שהרחבים הלטניים של כל מודל יישארו אינפורטטיביים, גם כאשר מודלים מסוימים חסרים. התוצאה? שחזורים עקביהם יותר, רלוונטיים מבחינה הקשרית, ובועל דיק גבוה יותר בגין הצפי של הנתונים החסרים.

יש עוד טענה מעניינת במאמר שאנו מודה שלא הבנתי עד הסוף: הקשר בין MMVM לBIN למידה ניגודית. השימוש ב-JSD כרגולריזציה מקריב את התפלגיות הלטנטיות בין המודליות השונות, אבל לגרום לתת מרחב קטן (טענה לא מובנת לי). זה דומה לאופן שבו למידה ניגודית פועל במודלי שפה-ויזן: היא מביטה שקלטים דומים יפיקו ייצוגים קרובים, אך תוך שימוש הבדלים ביניהם. עם זאת, בinguo למידה קונטרסתיבית, אשר לרוב דורשת יצירה מפורשת של זוגות דוגמאות חיוביות ושליליות, VAE MMVM מטמע את תהליך היישור זהה בתוך המודל הגנרטיבי עצמו. המשמעות היא שהיצוגים הלמידתיים מתכנסים באופן טבעי במהלך האימון, ללא צורך בטריקים של דוגמה קונטרסתיבית או מטרות נוספת.

כמובן, ישנו גם טרייד-אופים. ההסתמכות של המודל על התפלגיות הראשונית תלויות-דатаה הופכת גנרטוט הבלתי מותנה כמו ב-VAE לבליתי ישם. רוצים לדגם דאטה מולטי-מודאלים חדשים מאשר? חבל, השיטה הזאת לא מועדת לכך (:). אבל זה לא באג, אלא פיצ'ר. MMVM אינו מסה להיוות מודל גנרטיבי טהור כמו GANs או מודלי דיפוזיה. המטרה שלו היא למידת ייצוגים מובנית וברורה, שבה התלות בין המודלים נשמרות אבלית לכפות מגבלות מלאכותיות. במקרה הזה, מדובר בהתקדמות מחקרית משמעותית

<https://arxiv.org/abs/2403.05300>

28.02.25 - המאמר היומי של מייק - The FFT Strikes Back: An Efficient Alternative to Self-Attention

מי עוקב אחרי מספיק זמן בזוויא ידע שיש לי חולשה למאמרים שמוופיע בהם התמורה פוריה או כל התמורה אחרת (כמו התמורה קוסיני DCT). הסיבה לחיבה זו היא 5 שנים שבילתי בתור חוקר, מהנדס אלגוריתמים ומרצה בתחום עיבוד אותות (מערכות תקשורת אלחוטיות). המאמר הזה מתמקד להציג מנגנון שהוא מחליף את מנגנון attention בשכבה המבצעת טרנספורמציה של ייצוגי טוקנים בתחום התדר (כלומר התמורה פוריה). הטענה במאמר שהיא בעלת expressiveness (מסוגלת למדל את אותן הפונקציות) קרובה לזה של הטרנספורמרים. הטענות מוכחות בזרה חזיתיאורטית (הוכחה מלאה לא הוגה במאמר).

אבל מה היתרון של הגישה המבוססת על טרנספורמציה לא לינאריות במרחב התדר? כמובן סיבוכיות יותר נמוכה מהסדר $N \log(N)$ המשמעותית יותר נמוכה $M \cdot N^2$. הסיבוכיות התיאורטית של מנגנון attention. כאן N מסמן את אורך הסדרה. ידוע שניתן לעשות התמורה פוריה עם הסיבוכיות $N \log(N)$ ולמרות שהמאמר מכניס אי לינאריות בטרנספורמציות מעלה מרחב התדר עדין הסיבוכיות של המנגנון המוצע נותרת $(N \log(N))$.

از איך כל הסיפור הזה עובד? קודם כל מעבירים את כל הטוקנים דרך התמורה פוריה כאשר כל מימד יציג הטוקנים עבור FFT בנצח. כלומר אם יש לנו $10K$ טוקנים שככל אחד מהם מיוצג על ידי וקטור באורך 1024 יש לנו 1024 התמורות פוריה כל אחת באורך של $10K$. לאחר מכן מחשבים את המmozע של כל היצוגים בתחום התדר ומעבירים את התוצאה דרך MLP (כלומר כמה שכבות fully-connected) כאשר הפלט שלו הינו בגודל המקורי של הסדרה (בדוגמא זה יהיה $10K \times 1024$). לאחר מכן מחברים את התוצאה למטריצה W_{base} שכלה מורכבת מאחדות.

בשלב הבא מכפילים (אי-יבר-אי-יבר) את התוצאה עם התוצאה הראשונית של התמורה פוריה. כלומר מה שיש לנו כאן הוא משקל מוחדר של התמורה פוריה של ייצוגי הטוקנים כאשר המשקלות מחושבות בזרה לא לינארית. לבסוף עושים לתוצאה של של ReLU למספרים מרוכבים(UReLU) ומעבירים חזרה למרחב המקורי עם ה-IFFT.

והתוצאות כמובן לא רעות בכלל...

<https://arxiv.org/abs/2502.18394>

01.03.25 המאמר הימוי של מijk - LORA VS FULL FINE-TUNING: AN ILLUSION OF EQUIVALENCE

היום סקירה מס' 200 מאז שהתחלה לכתוב סקירות יומיות לפני 9 חודשים. אז לרגע החזון העגול זהה נסתפק בסקירה קצרה של מאמר די קליל. המאמר משווה את השפעת של פינ-טיון עם LoRA לפינ-טיון רגיל שבו מעדכנים את כל המשקלות של המודל.

اذכיר עם אם LoRA אנו מאמנים מטריצות בעלות ראנק נמוך שמתווספות למטריצות המשקלים בכל שכבה. אזכיר שמטריצה בעלת ראנק נמוך מגודל אחד ניתנת לייצוג באמצעות מכפלה של מטריצה A בגודל חוץ ומטריצה B בגודל $z \times m$ כאשר ($n, m <> r$. (ראנק נמוך). אז ב- LoRA מאמנים מטריצות A ו-B- (לגל שכבת המודל) שכאמור מכפלתם מתווספת למטריצה המשקלים במקורית W של המודל שעובר פינ-טיון. למיטב זכרוני BA מתווסף למטריצות K ו-Q בשכבת ה-attention.

از המחברים משווים את מטריצות המשקלות המאומנות אחרי פינ-טיון מלא בין פינ-טיון על LoRA. זה נעשה באמצעות השוואות של וקטורים סינגולרים של מטריצות המשקלות המאומנות זוכרים SVD שזה קיצור של Singular Value Decomposition - זה הווקטורים שמופיעים באחת המטריצות עם עמודות אורתונורמליות. למי ששכח SVD הוא שיטה מתמטית שפרקת כל מטריצה A למכפלה של שלוש מטריצות: $U = V^T S V$, כאשר V הן מטריצות אורתונורמליות, ו-S היא מטריצה אלכסונית המכילה את הערכים הסינגולריים.

המחברים הגדרו מושג *intruder dimension* או *In* בטור זה ש"נעלם" אחרי הפינ-טיון. כולם לכל וקטור סינגולרי של מטריצת המשקלות המקורית W מנוטים להתקאים וקטור סינגולרי של אותה המטריצה אחרי פינ-טיון. ההתאמנה נמדדת באמצעות דמיון קוויין - ככלור אמר לווקטור סינגולרי נתון של W לא נמצא וקטור סינגולרי של המטריצה אחרי פינ-טיון בעל דמיון קוויין גבוה מספיק הוא נקרא משoir-*inDim*.

از התברר שעבור פינ-טיון עם LoRa מספר *inDims* הינו גבוהה יותר מאשר עם פינ-טיון מלא. מעניין כי עבור Z (הראנק) של LoRa ממש נתונים (קרובים ל-1) מספר *inDims* לא גבוה ואז הוא מתחילה לעלות כאשר מעלים את Z ומתחילה לרדת כאשר Z מגע לערבים גבוהים יחסית (נגיד 64). המחברים טוענים שאם עושים פינ-טיון עם LoRa המודל "ישוכח יותר מהידע שלו" מאשר עם פינ-טיון מלא. המחברים טוענים זהה קשר לעובדה שהמודלים שטויבו בצורה מלאה מציגים ביצועים טובים יותר על DATA OOD לעומת *out-of-distribution data*.

אצין שהמחברים התמקדו בטיב מודלי אנקודר מבוסס RoBERTA.

<https://arxiv.org/abs/2410.21228>

02.03.25 המאמר הימוי של Mijk - An Empirical Model of Large-Batch Training

מאמר מלפני 6 שנים של חוקרי OpenAI אך מצאתי אותו די מעניין לסקירה קצרה. המאמר חוקר גדול באז' אופטימלי עבר אוימון MBGD או Mini-Batch Gradient Descent מה זה אופטימלי哉? זה שימזר את מספר הדוגמאות Sh-MiGD משתמש בהם כדי להביא את המודל לערך יעד של הלוס. כזכור שניתנו "להריץ" את אותו הדוגמא כמה פעמים במהלך MBGD.

למי שכח MBGD שיר למשפחת שיטות המבוססות על מورد הגדריאנט. עם DMBGANO מחלקים את הדאטאסט למיני-באץ'ים שכל באץ' מורכב מכמה דוגמאות. עבור כל באץ' אנו מבצעים עדכון אחד של משקל מודל כאשר הגדריאנט מחושב בתור ממוצע של כל ערכי הגדריאנטים עבור כל הדוגמאות באץ'. למעשה ממוצע זה הינו משערך של הגדריאנט הממוצע של המודל עבור כל הדוגמאות מהדאטאסט. נזכיר שכל עדכון הוא הינה (לינארית) של מסקולות המודול בכיוון ההפוך לכיוון הגדריאנט. כל עדכון כזה תלוי בקצבו למידה שקבע את גודל עדכון המשקولات (מכפל בградיאנט ממוצע).

המאמר מציע שיטה למציאת גודל באץ' אופטימלי (לפי ההגדרה שנתתי קודם) שעבור קצב למידה אופטימלי (המצביע את הלס בכל איטרציה). די ברור כי גודל באץ' אופטימלי צריך להיות תלוי בפרמטר המודל - למשל, בצורת משטח הלס וגם בערבי הגדריאנט. המאמר טוען כי גודל באץ' אופטימלי ניתן לחשב בתור הטרייס (trace), סכום הערכים העצמיים) של המכפלה של מטריצת קווריאנס של גרדיאנט הלס וההיסיאן H של פונקציית לוס מחולקת ב $G^T G$ כאשר G הוא הממוצע של וקטורי הגדריאנט.

תוצאה זו התקבלה דרך פיתוח טילור מסדר שני (בכיוון הגדריאנט), מציאה גודל קצב למידה אופטימלי והציבתו לנוסחה כדי לחשב את גודל הבאץ' שעבורו מתקיים מתקובל ירידה מקסימלית של הלס. לאחר מכן משווים את הירידה המקסימלית עם זו עבור גודל באץ' נתון B.

המאמר מדגיש שגודל אופטימלי של באץ' אינו תלוי בגודל שלה דאטאסט וכמוון משתנה במהלך האימון כי גם ההסיאן H וגם הגדריאנט הממוצע H וגם מטריצת קווריאנס של גרדיאנט הלס לא נשאים קבועים (בד"כ). המחברים מצינים מקרה פרטי די מעניין (לא קורה במציאות אמונה) שבו ההיסיאן H שווה למטריצה היחידה I. במקרה זה גודל באץ' אופטימלי שווה לסכום השונות של כל רכיבי הגדריאנט.

המאמר כתוב בצורה מאד מובנת ונitinן לקריאה קלילה יחסית...

<https://arxiv.org/abs/1812.06162>

המאמר היומי של מיק - 03.03.25

The Geometry of Concepts: Sparse Autoencoder Feature Structure

המאמר חוקר את האופן שבו מקודדים אוטומטיים דليلים (SAE) מייצגים ומונים מושגים ב-LLMs. החוקרים מנהנים את המבנה הזה בשלושה קני מידה היררכיים: אוטומי, מוחי וגלקטי. המחקר מנסה לעשות לא מעט הקבלות בין מרחב האמבידיג של מודלי שפה למבנה המוח אבל כמובן זה לא אוצר ש-LLMs חושבים בדומה לנו.

מתודולוגיה:

נרענן כי SAEs הינו כלי למחקר של interpretability של SAEs. הם מאמנים לשחזור אקטיבציות של שכבה ספציפית במודל תוך שימוש בתת-קובוצה קטנה של פיצ'רים שלහן בלבד. אילוץ דليلות זה מכריח את SAE ל"יצג כל נירון בתור צירוף לינארי של מספר קטן של פיצ'רים סמנטיים של אחד מהם(פיצ'רים) מקודד מושג מסוים(ניתן לפרשנות). ככלומר ה-SAE לומד מיליון של וקטורי פיצ'רים(אמבידיגס) שבו כל נירון מופעל באופן סלקטיבי עבור דפים סמנטיים מסוימים.

החוקרים משתמשים ב-SAEs כדי להלץ פיצ'רים סמנטיים מייצגים של מושגים ב-LLMs. המחקר מתמקד בניתוח המבנה הגיאומטרי של ייצוגים אלה בשלושה קני מידה.

כדי לחסוף את המבנה זהה, החוקרים משתמשים ב-*LDA* כדי להסיר כיווני "הסחה" גלובליים למרחב האמצעי, כמו אורך מילה, שעלולים לטשטש קשרים מושגים סמנטיים. שלב זה חיוני במיוחד עבור הרמה האוטומטית, שבה חסמים אנלוגיים הופכים ברורים יותר לאחר הסרת ההשפעות המסתichות.

רמה אוטומטית: "גבישים" וtabniot גיאומטריות

בקנה מידת הקטן ביותר, המחקר מזהה "גבישים"- מבנים גיאומטריים כמו מקביליות טרפזואידים - בתוך מרחב התכונות הרב-ממד'. מבנים אלו מכילים חסמים ידועים כמו (גבר - מלכה) :: (מלך - אישה). החוקרים מצינים כי איקות הדפוסים הגיאומטריים משתפרת משמעותית כאשר מסירים כיווני הסחה גלובליים, כמו אורך מילים, באמצעות (*LDA* – Linear Discriminant Analysis).

רמה מוחית: מודולריות מרחבית ו"אונות"

בקנה מידת בינוני, המחקר חושף מודולריות מרחבית בתוך מרחב פיצרים של *SAE*. פיצרים השיכים לתחומים ספציפיים, כמו מתמטיקה וקוד, מקובצות יחד ליצירת "אונות" נפרדות, בדומה איזורים תפקודים הנצפים ב-*fMRI* של המוח האנושי. החוקרים משתמשים במידדים שונים כדי לכמת את הלוקליות המרחבית של האונות ומגלים כי פיצרים מופיעים יחד ב częיפות גבוהה יותר مما שהיא צפויה בגיאומטריות של פיצרים אקראיים. ממצאים אלה מצביעים על כך שה-*SAE* מארגן פיצרים קונספטואליים באופן המשקף התמונות תפקודית.

רמה גלקטיבית: מבנים מידת רחבה והתפלגות ערכים עצמאיים

בסקירה הגדולה ביותר, המחקר מגלה כי פיזור ענן הנקודות המאפיינות הוא אנטוורפי (שונה בכיוונים שונים) מאופיין על *Power Law* של ערכים עצמאיים, השימושם התלולים (בין ע"ע) ביותר שנצפים בשכבות האמצעיות של הרשת. הדבר מצביע על כך שהמורכבות והווריאציה של ייצוגו דאטה אינם אחידים בין השכבות, כאשר השכבות האמצעיות קולטות וריצאות עדינות יותר בתונונים. המחברים גם מונחים גם משתנה האנטוורפה של אשכולות(בעניינן נקודות) בין השכבות השונות של המודל, ומספקים תובנות על המבנה היררכי של ייצוג מושגים בתוך המודל.

<https://arxiv.org/abs/2410.19750>

5.03.25 המאמר היומי של מijk - Mixtures of in-context learners

מודלי שפה מודרניים ניחנים ביכולת לבצע משימות שהם לא אומנו עליהם באופן מפורש בהתאם על כמה דוגמאות המדגימות את המשימה ללא צורך באימון (פין טיון). יכולת זו קיבלה שם למידה *in-context few-shot learning* (ICL). אני גם רأיתי שקרים לזה לפעים למידת *few-shot* למרות שזה פחות מותאם כי מוגדר בד"כ בתור פין טיון של מודל על כמה דוגמאות.

אז איך כל העסק עוסק? מספקים למודל שפה כמה דוגמאות של ביצוע המשימה בתור פרומפט, בד"כ כמה זוגות (x, y) אשר x הינה שאלת או שאלתה $-y$ הינה התשובה הצפוייה ל- x . לאחר הדוגמאות אלו מודים שאלתה x שהמודל צריך לספק תשובה עליה בהתאם לדוגמאות לראה לפני כן.

סביר להניח לכל שאלה x יש דוגמאות $-x$ בתוך הפרומפט שדומות לה יותר ויש כאלה שפחות. איך נגרום למודל להתחשב יותר בדוגמאות רלוונטיות יותר ולהתחשב פחות בדוגמאות פחות רלוונטיות לשאלתה x . זו השאלה שמחברים המאמר שואלים ומצביעים שיטה למשך תרומות של כל דוגמא לשאלתה נתונה x .

בاهינתן דאטהסת של דוגמאות מתוויות (עם תשובות) ומאמן מודל הפלט משקל ω עבור כל דוגמא בפרומפט לשאלתה x . משקלות ω משמשות לחישוב של התפלגות של כל טוקן בתשובה y בהינתן כל זוגות (y, x) ושאלתה x . התפלגות ה z מוצגת בתור סכום ממושקל עם ω של softmax של טוקן y בהינתן כל זוג דוגמאות (y, x). המאמר מציע שני דרכי לאמן את המשקלות האלו (על דאטהסת של שאלות ותשובות). הדרך הראשונה לאמן אותה בצורה ישירה (פשוט לאפTEM פונקציה לואס לפיהן בהינתן ייצוג הטוקנים של ω ו- y) והדרך השנייה היא לאמן רשות המחשבת את המשקלות האלו ולאפTEM את המשקלות שלה.

בסוף המאמר המחברים מציעים שיטה לאימון של k-top של המשקלים כדי לא לחשב את כל ה- softmax עבור כל הדוגמאות שזה יכול להיות קצר כבד חישובי וגם לוקח זמן. השיטה מבוססת על Implicit MLE מאמנה מודל לאפTEM מודל לטני כאשר משתמש חביי (לטני) נדגם מההתפלגות דיסקרטית. השיטה די לא טריומיאלית להבנה - מי שרצה להתעמק בה (מומלץ) מוזמן להבטח ברפרנסים.

<https://arxiv.org/abs/2411.02830>

המאמר היומי של מילק - 06.03.25

LYNX: ENABLING EFFICIENT MOE INFERENCE THROUGH DYNAMIC BATCH-AWARE EXPERT SELECTION

שמתי לב שזמן לא סקרתי מאמר על MoE - Mixture Of Experts במודלי שפה. אזכיר MoE זו שיטת המיעודת לאופטימיזציה של אינפראנס מבחינת העומס החישובי (כלומר פחות חישובים). המודל מואמן להפעיל רק חלק מהמודל (מומחים מסוימים) עבור כל טוקן אשר כל מומחה הוא (בדרך כלל) תחת-רשת של FFN (למעשה תשת-מטריצות של מטריצות המשקלות ב-FFN) בתוך מנגנון הטרנספוררים. בפועל זה מאפשר להקטין את כמות החישובים לכל טוקן עשויי לאפשר הפעלה של LLMs בגודל עצום (רק החלק המודול כל פעם). בנוסף (לפי כמה מחקרים) שיטה זו מאפשרת ללמידה "פונקציות מורכבות יותר" כי כל טוקן עשוי להיות מוחש בצורה שונה (עם תתקבוצה שונה של מומחים).

המומחים נבחרים על ידי רשת ניתוב (routing network) אשר היא מואמנת לחשב ציון אי שלילי לכל מומחה. ציונים הם למעשה "הסתברות" לבחירה של כל מומחה (יש softmax בסוף). בד"כ k מומחים בעלי ציונים הגבוהים ביותר נבחרים בכל שכבה עבור כל טוקן מתוך $N < k$. המודל מואמן לאזן ניצול של כל מומחה כאשר המטרה שכל מומחה יונצל במידה שווה בדאטהסת אימון (aggregative level). בד"כ יש איבר רגולריזציה על משקל רשת הניתוב למשל בצורה של אנטרופיה שלילית או סכום הריבועים).

המאמר מציע שיטה לאופטימיזציה של צריכת זכרון עבור אינפראנס של מודלי טרנספוררים עם MoE כאשר הם מופעלים בباءים של שאילותות (כמה קלטים). הגישה המוצעת מבוססת על כמה אובייקטיביות אמפיריות שנעשו על ידי המחברים:

- התפלגות של שכיחות הפעלת המומחים בתוך הבאץ' אינה אחידה כלומר יש מומחים שמשופעים יותר ויש כאלו שימושים פחות
- הצפיפות החישובית (arithmetic intensity) שהיא היחס בין כמות flops לכמות גישות זכרון יורדת כאשר כמות המומחים עולה בשלב decode (כלומר חיזוי). זה הופך את השלב הזה ל-memory-bound שגדיל את ה-latencies
- הטוקנים לא מאד רגשים למומחים שלהם מעבר למעט מומחים (מ-k-top) בעלי ציונים הגבוהים ביותר.
- ככל ניתן "להפעיל רק המומחים" בלי פגיעה משמעותית ביצועים לא כל הטוקנים הם שוו ערך ככלומר יש טוקנים רגשים יותר לשימוש בחילק מהמומחים שלהם ויש כאלו שפחות. המחברים טוענים שכן ניתן להסיק את רמת ההרגשות של הטוקן מצינו רשות הניתוב עבורו

- השלב של `prefill` (עיבוד פרומפט) רגש יותר להחלפת המומחים משלב `decode` (גנרטוֹת)
- הרגישות להחלפת המומחים משתנה בין שכבות המודל כאשר השכבות האמצעיות הן הרגישות ביותר ביותר לזה

המחברים מציעים לנצל את אובייצנטיביות אלו בצורה הבאה (יש כמה וריאציות, אתאר את עיקרי השיטה)

- משתמשים בכל המומחים בשלב `prefill` (שהוא `compute-bound`)
- מזהים טוקנים רגשים וPOCHOTES (low and high confidence) בבאז'. לאחר מכן מפלטרים את המומחים של הטוקנים הפחות רגשים
- בוחרים את המומחים שהם הכי בשימוש עבור הבאז' ומפלטרים את השאר
- מפעלים רק את המומחים שנותרו עבור כל הטוקנים (`k-top`). אופציה נוספת (POCHOTES פוגעת בביצועים) - היא להפעיל את כל המומחים עבור טוקנים רגשים ורק את אלו שנותרו עבור טוקנים פחות רגשים

שיטת זו מאפשרת להגדיל צפיפות חישובות עבור שלב `decode` ולעשות אותו פחות `memory-bound` בלבד. פגיעה משמעותית בביצועים.

<https://arxiv.org/abs/2411.08982>

המאמר היומי של מייק - 07.03.25

Number Cookbook: Number Understanding of Language Models and How to Improve It

מבוא:

תמיד טענתי כדי להשתמש בכלים פשוטים כמו מחשבון או קוד לחישובים ארכיטמטיים אבל אנשים מתעניינים להשתמש ב-LLMs בשילוב נייח וישראל מחייב.

המאמר חוקר יכולות ההבנה והשימוש במספרי (NUPA) של LLMs. המחברים מציעים מספר מבחנים להערכת ביצועי המודלים על פני 4 סוגים יציג מספרי ו-17 קטגוריות של שימושים, שהובילו ל-41 מקרים ייחודיים. גישה זו כושפת פערים גדולים ביכולת המודלים לבצע משימות הכוללות חישבה מספרית.

הטענה המרכזית של המאמר היא כי מיזוג מושגים איננה תcona המתפתחת באופן אוטונומי כתוצאה מאימון מקדים מסויבי אך כלל, אלא יכולה הדורשת פינאי טוון דאטאטיטים המיועדים לכך. הכישולן של LLMs במשימות מספריות טריויאליות, כמו מיזוג מספרים בפורמט נקודה צפה או חישובי מודולו, עומד בסתרה ליכולתם לבצע ריזונייניג (הנמקה) סමבולי מורכב. המחברים טוענים כי למרות שיפור משמעותי ביכולות LLMs, עיבוד מספרי בסיסי יותר עקבי אכילס שלהם.

בנצל'מרק ל-NUPA

המחברים מציגים בנצל'מרק הממיין משימות מספריות לפי טיפוס כגון מספרים שלמים, נקודה צפה, שברים ורישום מדעי (scientific notation). רמת הפירוט של מבחנים אלה היא שיפור משמעותי לעומת ביצועם בנצל'מרק קיימים להבנה מתמטית, אשר לעיתים קרובות מעורבים בין מיזוג מושגים פתרון בעיות בין המבנה מספרית טהורה.

במציאות הגדרת המבחןים סביר פעולה ארכיטטניות בסיסיות, הבנת ספרות ומשימות המרה, המחברים מבטיחים שהערכת המודלים תבודד את יכולותיהם המספריות מיכולות ההנמקה רחבות יותר. הגישה המבנית מאפשרת מדידה מדויקת של חולשות המודלים ומספקת מפתח דרכם לשיפורם עתידיים. הבנצל'מרק מתבסס על תכנים מוחומר הלימוד של בתים ספר יסודים ותיכוניים, מה שמבטיח שימושיות משקפת הבנה מספרית בעולם האמיתי.

הערכה אמפירית של מודלים מוביילים

המחקר מבצע הערכה שיטתיות של מודלים כגון GPT-3, LLaMA, Qwen-2, ו-i2GPT, ומגלה כי גם המודלים המתקדמים ביותר מתקשים במשימות מספוריות פשוטות, במיוחד כאשר רמת המורכבות או אורך הקלט גדלים. רידת הביצועים במשימות כמו חישובי מודולו והטאמת ספרות מדגישה נקודות תורפה קריטית בארכיטקטורות הנוכחות של LLMs.

תוצאה מפתיעה היא שהשיגיאות המספוריות נשכחות גם במקרים שבהן המודלים מצטיינים בبنצ'מרקם מתמטיים כלליים יותר. המחברים מנהנים באופן שיטתי כיצד יציגים מספוריים שונים משפיעים על הביצועים, וחושפים רידה דרסטית בדיקן כאשר עוברים משימות מובוססות מספרים שלמים למשימות המבוססות על מספרים עם נקודה צפה או שברים. מצא זה חשוב במיוחד משום שהוא מצביע על כך ששיטות האימון הנוכחות אין מצליחות להקליל מינימונת מספרית מעבר לאריתמטיקה על מספרים שלמים.

חקירת השפעת אימון מקדים (pretraining) טיב (Fine-tuning) וטוקניזציה על הביצועים ב-UPA

המחברים בוחנים 3 אסטרטגיות עיקריות לשיפור ביצועי AUPA: שינוי אסטרטגיות טוקניזציה, פיין טיאן למשימות מספוריות, ושימוש בקידוד מיקום (PEs) ו騰肯יקות ישורן (alignments) ספורות. באופן מפתיע, למחרות שפיין טיאן בסיסי משפר משמעותית את הביצועים, טכניקות כמו טוקניזציה חלופית או שימוש בرمזאי אינדקסים (קידוד מיקום ספורות) דווקא גורמות לרידה בביטויים במקומם לשפר אותה.

הניסויים בטוקניזציה מספקים תובנות מעניינות: טוקניזציה המבוססת על ספרה בודדת מתפקדת טוב יותר מאשר טוקניזציה של מספר ספרות, בניגוד להשערה הרווחת שלפיה טוקנים ארוכים משפרים ביצועים. ממצאים אלו מצביעים על כך של-LM מתקשים בהתאם במספרית כאשר הטוקנים כוללים מספר ספרות, ככל הנראה בשל האופן שבו מודלי טרנספורמר מעבדות סדרות. יתרה מכך, שינוי PE שנעשה לשפר למידה מספרית לעתים קרובות מניבים תוצאות היפה, דבר המצביע על כך שהאינטראקציה בין טוקניזציה לקידוד מיקום במשימות מספוריות היא מורכבת ולא טריוויאלית.

ניסוי פיין טיאן מראים שניתן להשיג שיפורים משמעותיים ב-UPA דרך אימון ממוקד, אך השיפורים אינם מתורגם בהכרח לכל המשימות המספוריות. לדוגמה, העובדה שמודלים מוטיביים לא מצליחים לשפר משמעותית ביצועים במשימות של שליפת ספרות מרמזת על כך שמנגנון הקידוד המספוריים דורשים חשיבה מחודשת ברמת הארכיטקטורה, ולא רק שינויים בדאטאטס.

ניתוח (CoT) למשימות מספוריות

המחברים מיישמים Rule-Following CoT (RF-CoT) כדי לבדוק האם פירוק לשלביו חישוב מצמצם שגיאות מספוריות. אף ש-CoT משפר את הדיקון, מגבלתו - זמן חישוב ארוך יותר ומגבילת חלון ההקשר—מציבות אותו כפתרון לא יעיל לשימוש יומיומי במשימות מספוריות.

הניסויים מראים שבעוד Sh-T CoT משפר ביצועים במשימות חישוב מובנות כמו כפל רב-ספרתי, הוא הופך במהירות ללא יעיל מבחינה חישובית. הבעיות של ייצור שלבי החישוב הביניים עולה על התועלת המדיקת, מה שהופך את CoT ללא פרקטני עבור יישומים אמיטיים הדורשים חישובים מספוריים קבועים. בנוסף, המחקר מזהה תקלה ביצועית שבה שלבי חישוב נוספים אינם משפרים את הדיקון, מה שמחזק את הרעיון כי יש צורך בשיפור ייצוג ועיבוד מהותיים ולא רק בפתרונות עוקפות.

המאמר תורם תרומה משמעותית על ידי ניתוח שיטתי והערכת NUPA ב-LLM. העבודה חושפת מגבלות יסודיות ומספקת עדויות אמפיריות לאסטרטגיות לשיפור (ולכישלונות) בעבוד מספרי. אף שהמחקר מדגיש אתגרים קיימים, הוא מציע מפת דרכים חשובה לקהילת ה-AI לשיפור החשיבה המספרית במודלים עתידיים. המאמר מצביע על הצורך בפיתוח מגנוני עיבוד מספרי ייעודיים בתוך LLMs. ככל שהמודלים הופכים למתקדמיים יותר במשימות רזונינג מורכבות, חוסר היכולת שלהם להתמודד עם פעולות מספריות פשוטות הופך לבעה קריטית. מחקר זה מהווה בסיס לשיפורים עתידיים בלימידת יצוגים מספריים, אסטרטגיות טוקניתיה ייעילות, וגישה היברידית המשלבת במידה סטטיסטית עם עקרונות חישוב מספריים מפורשים.

או פשוט תעשו את החישובים האלה על המחשבון או עם פיטו...

<https://arxiv.org/abs/2411.03766>

09.03.25 המאמר היומי של מייק - THE SUPER WEIGHT IN LARGE LANGUAGE MODELS

זה די לא יאמן, אבל מודלים שפה גדולים עם מיליארדי או אפילו עשרות מיליארדי פרמטרים עלולים לשבול ירידה כואבת בביטויים אם מורדים מהם אפילו משקל בודד. ממצא מפתיע זה חל לפחות על חלק מהמודלים העצמאיים האלה.

מאמר זה מתעמק במאפיין ספציפי ובلتוי צפי של מודלים שפה גדולים: קיומם של "משקלים על (SWs)". המחברים מתקדמיים מעבר לתמצית ידועה על כך ש-LLMs מכילים משקלים חריגים המשפיעים באופן ניכר על הביצועים, ומציגים ראיות לכך שמשקל בודד יכול להיות קריטי באופן לא פרופורציוני לתפקוד המודל.

כאמור הממצא המרכזי הוא שהורדת SW בודד יכול לגרום לירידה קשה בביטוי LLM. השפעה דרסטיבית זו מתבטאת غالיה חזקה בperfeksty וירידה בדיק zero-shot לرمות כמעט אקריאות. מה שראוי לציין במיוחד גודל הוא העובדה שהסתרת SW לבין ההשפעה הקטנה יחסית של הורדת משקלים חריגים אחרים, אפילו בעלי גודל גדול יותר.

המאמר נותן דוגמה מעניינת להשפעה של הסרת משקל על-כבד עבור הפרומט: "קץ חמ. חורף הוא...". הטוון הבא הנזכר ציר להיות "קר" ועם המודל המקורי עם SW, הוא חוזה בכך נכון את הטוון הבא "קר" בהסתברות גבוהה של 99.9%. כאשר SW מוסף, החיזוי המוביל של המודל הוא "ה" (the) ב-91.4% בהסתברות נמוכה ולא בטוחה של 9.0%. זה מצביע על כך ש-SW חיינו למודל כדי לבצע חיזוי נכון ובטוח של מילימיםמשמעותיים. המאמר לא רק מתעד את התופעה הזאת; הוא גם בוחן את המנגנונים הבסיסיים הקשורים אליה. המחברים מקשרים SW לייצור "אקטיציות SW", שהן אקטיבציה גדולה וחירגות המתפשטות דרך המודל כמעט ללא קשר לקלט.

יתר על כן, המחקר בוחן את ההשלכות של SW עבור קוונטיזציה של מודלי שפה. נוכחותם של חריגים, כולל SW וاكتיבציות חריגות הנגזרות מהם, מחייבת אתגר ממשמעותי לקוונטיזציה גבוהה, שכן חריגים אלה יכולים לעוזות את תהליך הקוונטיזציה ולהוביל לאובדן מידע ממשמעותי. המחברים מדגימים ששמור חריג SW (גם משקלים וגם אקטיבציות) יכול לשפר את יעילות הקוונטיזציה מסווג "עיגול לערך הקרוב ביותר", אפילו לאפשר שימוש בגודלים גדולים יותר של בלוקים בקווונטיזציה (עבורם מחושבים קבועי קווונטיזציה). זה מושג על ידי השארת SW מחוץ לתהליך הקוונטיזציה ומחזור ערכיהם לאחר מכן, תוך שימוש ההשפעות השליליות של ערכים קבועיים אלה על קווונטיזציה של פרמטרים אחרים. על ידי התמודדות עם האתגרים שמצויבים חריגי על-כבד, הגישה המוצעת מאפשרת יישום של שיטות קווונטיזציה פשוטות ויעילות יותר, ומקלה על פריסת מודלים בסביבות עם משאים מוגבלים.

עובדת זו יוצרת טיעון חזק ש-SW אינטגרליים הממלאים תפקיד חיוני בעיצוב התנהגות והיעילות של LLMs, עם השלכות משמעותיות לדחיסה ולאינפנס של מודלים. תרומת המאמר אינה טמונה רק בזיהוי SW אלא גם באפיון תפקידם הפונקציוני בתוך LLMs. המחברים מנתחים כיצד משקלים משפיעים על פלט המודל, ומקשרים אותם ל"התפשטות" של אקטיביזיות חריגות.

<https://arxiv.org/abs/2411.07191>

המאמר היום של מיק - 11.03.25

Beyond Matryoshka: Revisiting Sparse Coding for Adaptive Representation

סקירה קצרה של מאמר המכיל שיטה להפקת ייצוג במימד נמוך של דатаה הנקראת embeddings. מה מיוחד בשיטה זו - היא מאפשרת לאמן את הייצוג זהה בכמה מימדים בו זמנית. ככלمر במהלך האימון ייצגים מכמה גדלים (ג'יד 8, 16, 32, 64 ו-128) מאומנים באותו הזמן. השיטה מניחה דאטאטס מתויג של זוגות (ע, א) כאשר א הוא פיסת דטה ו-ע הוא התיאוג שלו.

ייצוג מטרישקה מאמנים רשת עמוקה עם השכבה האחורונה (ראש) הממפה את הייצוג של דטה לתיאוג שלו. מה המיוחד במטרישקה הוא שהוא מאמן בו-זמן כמה וקטורי מייפוי (יחד עם המודול עצמו) למחרב התיאוג כאשר כל מייפוי לוקה נ_-ה האיברים הראשונים מוקטור האמבייניג(השכבה האחורונה של המודול). בדוגמה שנתתי קודם מאמן בו-זמן וקטורי מייפוי בגודלים 8, 16, 32 ו-64. פונקציית הלום הינה סכום של הלוסים עבור כל הוקטורים האלה - ככלומר נוסף למודל עצמו אנו מאמנים 4 וקטורים בגודלים 8, 16, 32 ו- 64.

המאמר המסורק מכיל את הגישה המעניינת הזו על ידי החלפתה בשני אלמנטים (של פונקציית לוס למשבה). הראשון הוא sparse autoencoder או SE שבמקורה מאומן למפות את ייצוג הדטה, המופק על ידי המודול, למרחב בעל מימד מאד גבוה אבל מאוד דليل ואז להחזיר אותו למרחב ייצוג המקורי. נציין כי המודול עצמו לא מאומן כאן אלא רק וקטורי המייפוי (של SE). האלמנט השני שמתווסף להלום ניגודי שבא להרחק את ייצוג הדטה מקטגוריות שונות רחוק אחד מהם ולקראב את הייצוגים של פיסות הדטה מאותה הקטgorיה.

אז מה המטרה של SE כאן? להבדיל מהמטרישקה המקורית שמאמנה את האלמנטים הראשונים כאן אנו לוקחים k-top רכיבים של וקטורי הייצוג אחרי האנקודור. הדקודר מאומן לשחזר את הוקטור המקורי רק עם k-top אלמנטים של הוקטור אחרי האנקודור. הבעיה הידועה עם SE היא הרכיבים של הוקטור אחרי האנקודר שלמדושה מתיים - ככלומר מקרים ערכים קרובים מאוד לכל פיסות הדטה.

כדי להתמודד עם בעיה זו החברים מציעים שני דברים. הדבר הראשון הוא הוספה לוסים עבור כמה ערכים של k-less-k-top של האנקודר לפונקציית לוס (במקור יש ערך k אחד). ככה אנו מאמנים אմבייניג בכמה גדלים בדומה למטרישקה (חו"ץ מזה אין הרבה דמיון כי המטרה היא להפיק אמבייניג דليل). הדבר השני הוא הוספה של איבר המנסה לגרום לשגיאת השחזר עבור k-top של הרכיבים המתים (ערכים וכי נמוכים של וקטור הייצוג אחרי האנקודר) להיות קרוב לשגיאת השחזר של k-top של הרכיבים הגדולים ביותר של אותו הווקטור. אני לא הצלחת לרדת לעומק דעתם למה זה עוזר.

בנוסף כאמור מושגים איבר של הלום הניגודי לזה שמתואר בפסקה הקודמת....

טוב, נכוון שהופיעה לנו המטרישקה בשם המאמר הדמיון בין לבין המטרישקה המקורית ד' רופף. אבל המאמר ד' מעוניין חוות מזה....

<https://arxiv.org/pdf/2503.01776>

המאמר היום של מיק - 12.03.25

Transformers are Universal In-context Learner

היום נסקור קצחות מאמר תיאורטי כבד החוקר את יכולת האקספרסיביות של טרנספורמרים عمוקים. טרנספורמרים הם ארכיטקטורות עמוקות המגדירות "מייפויים הקשריים" (Mappings in-in), אשר מאפשרים חיזוי של טוקנים חדשים בהתאם על קבוצת טוקנים נתונה. שימושם לב של-*in-context*-*in* הקצח שallow them למדוד משימות שלא אומן בהם בהתאם לדוגמה בפרומפט (פחות טוב הבנת).

המחברים מוכחים כי טרנספורמרים עמוקים (בעלי מספר רב של בלוקי הטרנספורמרים) הם מקרבים(approximators) אוניברסליים, כלומר, הם יכולים לקרב כל מייפוי הקשרי רציף מהתפלגיות טוקנים בכל דיקט. יתרה מכך, התוצאות תקפות הן עבור מנגןוני *attention* דו-כיווניים (כמו באנקודר) והן עבור מנגןוני *attention* סיבתיים (כמו בדקדורים), תוך שימוש שアイו תלוי במספר הטוקנים.

הגישה המוצעת מבוססת על תורה המידה(Sof Sof מוצאי לה שימוש במאמרי DL), שבה רצפי סדרות מייצגים התפלגיות הסתברותיות במרחב האמבידינגן. זה מאפשר שימוש בכלים מאנליה פונקציונלית(פלאשבקים לפני 30 שנה בתואר הראשוון) ובתורת הטרנספורט האופטימלי (כתבתי על זה לא מעט באותו הקשור של Wasserstein GAN) על מנת להוכיח את יכולת הקירוב האוניברסלית של טרנספורמרים. תרומה טכנית מרכזית היא הגדרה חדשה של מנגןון-*attention* או-OPERATOR על התפלגיות. זה מאפשר שימוש במשפט סטון-וירשטראוס(Stone–Weierstrass theorem) על מנת להוכיח את יכולת הקירוב של פונקציה "נוחה" על ידי משפט פונקציות צפויות (המשפט באינפיניטי) - תוצאה יסודית בתורת הקירוב על קר שניתן לקרב כל פונקציה "נוחה" על ידי משפט פונקציות צפויות (המשפט באמצעות המגידר פונקציות במרחב האוסדורף וכאלו).

ייצוג מבואס-מידה של למידה בהקשר

חידוש מרכזי במאמר הוא ייצוג של מנגןון-*attention* או-OPERATOR על התפלגיות במקום על סדרות טוקנים סופיות. דבר זה מאפשר ניתוח אחד לשיטת ההקשר (learning in-in), ללא תלות במספר הטוקנים בסדרה. במקום לעבור עם קבוצות סופיות של האמבידינגן של הטוקנים, המחברים מגדרים מרחב התפלגיות הסתברותיות על תת-קבוצה קומפקטיבית של מרחב אוקלידי (של האמבידינגן). התפלגיות משיכת משקלים לאmbidinagen שונים של טוקנים, ובכך מייצגת את המעבר מלמידה על מספר טוקנים סופי לייצוג רציף ואניוטיפי.

באופן פורמלי, רצף של טוקנים ניתן לייצוג כהתפלגיות הסתברות בדידה, המורכבת מסכום משוקל של פונקציות דלתא דיראק, שכל אחת מהן ממוקמת על הטעמה של טוקן בודד. כאשר מספר הטוקנים גדול, התפלגיות אלה מתכנסות להתפלגיות רציפות. ניסוח זה מאפשר להוכיח תוצאות החלות על כל מספר אפשרי של טוקנים, כולל אינסופי.

הגדרת *attention* או-OPERATOR על מרחב מידות

שכבת טרנספורמר טיפוסית מורכבת משני רכיבים:

1. מנגןון *attention* רב-ראשי, האחראי על עדכון הייצוגים של הטוקנים על ידי חישוב יחסיו הגומליים ביניהם.
2. שכבות FFN, המעדכנות כל טוקן באופן עצמאי לאחר שלב *attention*.

המחברים מנסחים מחדש את מנגןון-*attention* כמייפוי הפעיל על התפלגיות של טוקנים. במקום לחבר סכום על קבוצת טוקנים סופית, ה-*attention* מוגדרת או-OPERATOR אינטגרלי על מרחב התפלגיות, מה שהופך את

הטוקנים למבנה רציף. ניסוח זה חשוב במיוחד, מכיוון שהוא מאפשר להגדיר רציפות וחלקות של מיפויים בהקשר באמצעות מרחק ו/orientation (מקרה פרטי שלו הוא earth mover distance), המודד את המרחק בין התפלגיות הסתברותיות. פונקציה היא רציפה במובן ו/orientation אם שינוי קטנים בהתפלגות הקלט מובילים לשינויים קטנים בהתפלגות הפלט. תכונה זו מבטיחה שהמיפויים שיוצרים טרנספורמרים יציבים לשינויים בהקשר הלימודי.

הוכחת אוניברסליות: קירוב מיפויים הקשיים

התוצאות המרכזיות של המאמר מוכיחות כי טרנספורמרים הם מקרים אוניברסליים למיפויים הקשיים. המחברים מראים כי עבור כל פונקציה רציפה המיפה התפלגיות טוקנים לפטיטים, קיימים טרנספורמר עמוק שיכל לקרב אותה בכל דיקוק. חלק מרכזי בהוכחה הוא בנייה של פונקציות יסודיות בהקשר, המשמשות כיחידות הבסיס לקירוב כל פונקציה כללית במרחבים שהגדרכנו קודם. פונקציות אלו הן גרסאות פשוטות יותר של שכבות טרנספורמר, אשר לצדות את העקרונות המרכזיים של מנגנון ה-*attention*.

פונקציה יסודית צזו מרכיבת שלושה מרכיבים:

1. טרנספורמציה לינארית על הטמעת הטוקן (מיפוי אפיני).

2. אינטראקציה לא-لينיארית המתחשבת בהתפלגות של כל הטוקנים.

3. התאמנה תלויות-הקשר, המאפשרת למודול "לימוד בהקשר".

פונקציות אלו פועלות באופן דומה למנגנון *attention* בעל ראש בודד, אך הן קלות יותר לנתח מתמטי. המחברים מוכיחים כי על ידי הרכבת מספר שכבות של פונקציות אלו, ניתן ליצור טרנספורמרם עמוקים המסוגלים לקרב כל פונקציה בהקשר.

שימוש במשפט סטון-וירשטראס

כדי להוכיח אוניברסליות, המחברים מראים כי קבוצת הפונקציות היסודיות שהגדרו מקיימת את תנאי משפט סטון-וירשטראס, שכאמר הוא משפט מאנליה פונקציונלית. המחברים מוכיחים כי הfonקציות היסודיות שליהם מקיימות תנאים אלו, מה שמבטיח כי טרנספורמרם עמוקים יכולים לקרב כל מיפוי הקשר.

סיכום:

המאמר מספק מסגרת מתמטית להוכחת האקספרטיביות של טרנספורמרם בلمידת מיפויים הקשיים, תוך שימושenganlia, תורת המידה ותורת הטרנספורט האופטימלי. התוצאות מראות כי טרנספורמרם עמוקים יכולים לקרב כל פונקציה תלויות-הקשר, ללא תלות במספר הטוקנים בחילון ההקשר.

<https://arxiv.org/abs/2408.01367>

13.03.25 המאמר היומי של מיק -

SLIM: Let LLM Learn More and Forget Less with Soft LoRA and Identity Mixture

האם לפעמים קורה לכמ שאתם מתחלים לקרוא את המאמר וככל שאתם מתקדמים ומתעמקים בו הוא מתחיל להירות פחות ופחות טוב. לי זה לפעמים קורה עם יכול אבל שם יותר קל לי להפסיק לאכול מאשר לקרוא מאמר. אז יאללה, אסקור אותו קצרות אף אל תצפו רבות...

המאמר מציע שיטה להלביש ערבות של מומחים או MoE על LoRa. נזכיר ש-LoRa היא שיטת פין טיון של רשתות ניירונים שבhem אנו לא מאמנים את כל משקولات המודול אלא רק מטריצות תוספות בעלת ראנק נמוך. MoE היא שיטה להורדה של העומס החישובי בטרנספורמרים כאשר אנו מחלקים את המטריצות בשכבה FFN של הטרנספורמרים לתת-מטריצות (מומחים) כאשר כל פעם לטוקן נתון אנו מפעלים רק חלק מהמומחים. שכבת ניתוב (routing layer) מחשבות את הציון של כל מומחים ובדרך כלל אנו בוחרים K מומחים בעלי ציון הגבוה ביותר (top-k).

از המחברים משדים LoRA עם MoE זהה בדיק מה שמשך את עני. המאמר מציע להחליף LoRA רגיל עם כמה מומחי LoRA שחלקם הינם מטריצות מראנק 0 או פשוט מטריצות אפסים. לטענת המאמר לא תמיד צריך להפעיל את LoRa. מומחי-hLoRa נבחרים על ידי רשות ניתוב בדומה ל-MoE הסטנדרטי. עבור כל טוקן נבחרים K מומחים (בינם גם מומחי זהות) בעלי ציונים הגבוהים ביותר. שימוש לב שבמאמר יש כמה שגיאות בנוסחאות המחשבים את התוצאה של המנגנון המוצע.

לאחר מכן המאמר מציע שיטה לשכלול הציונים של שכבת הניתוב בהתבסס על הסטטיסטיות של הדאטאסט עליו בוצע הפיניטוון עם השיטה. סטטיסטיקה במקורה זהה מחושבת על המוצבים החבויים של הרשת המחשבים על הדאטה של הפין טיון (אוף החישוב המדוייק לא מוגדר בצוරה ולדעתי יש שגיאות בנוסחאות המגדירות אותו). המחברים מציעים לקלستر את המוצבים החבויים האלו לקלסטרים שמספרם כנראה שווה למספר הטוקנים בפרומפט (מוגדר קבוע במאמר ובעור סדרות קצרות יותר משתמשים בטוקני ה-padding).

מרכז הקלסטרים מתעדכנים במהלך הפין טיון (כל פרומפט הקלט משוויך לקלסטר הקרוב ביותר ואז מרכז הקלסטר מחושב מחדש). במהלך האינפראנס פרומפט הקלט משוויך לקלסטר הקרוב ביותר (מרחב ריבוע) ואז ציוני המומחים המופיעים על ידי שכבת הניתוב עבור מומחי זהות מודזים במקדם שעולה אם המרחק לקלסטר הקרוב עולה כאשר הציונים למומחי LoRA האחרים נותרים ללא שינוי. נציין שמרכז הקלסטרים לא מתעדכנים במהלך האינפראנס.

לבסוף המאמר מציע דרך לשלב כמה MoE עם LoRa כמו שימוש פין טיון שונות אבל אחרי שגיליתי טיעות גם בפרק זהה, יתרתי....

<https://arxiv.org/pdf/2410.07739>

14.03.25 המאמר היומי של מייק - A Survey on Kolmogorov-Arnold Network

מבוא:

זכרים את KANs? זהה קיזור של Kolmogorov-Arnold Networks שעשה הרבה רוש בזמןו אך הביאו הלה ודעך עם הזמן. מתברר שיצאו לא מעט מחקרים בנושא המרתך הזה. המאמר דן בהרחבות ושינויים שונים לארקיטקטורת ה-KAN הבסיסית. אלה כוללים התאמות לניטות סדרות עתיות, לעיבוד נתונים גרפי ולפתרונות משוואות דיפרנציאליות. שינויים אלה כוללים לרוב שילוב של רכיבים מיוחדים או אילוצים בתוך ה-KAN במטרה להתמודד טוב יותר עם הדרישות הספרטניות של דומיניים אלה.

רשתות קולמוגורוב-ארנולד מייצגות שניי פרדיגמה בתכנון רשתות ניירונים, המבוססות על מעבר פונקציות אקטיבציה קבועות לקרה פונקציות הניתנות למידה הנקראות splines-a-b. הדבר שאב השראה ממשפט הייצוג של קולמוגורוב-ארנולד, הטוען שכל פונקציה רציפה של משתנים מרובים ניתנת לייצוג כהרכבה של פונקציות של משתנה אחד. באמצעות שימוש בפונקציות המייצגות על ידי ספלינים (שילוב של פולינומים באינטראול ספו),

KANs מציאות גמישות משופרת ופוטנציאלי לדיק אגובה יותר בקרוב פונקציות. דבר מוביל ל-interpretability לשדרוג של המודל, מכיוון שניתן לנתח ביותר קלות את הפונקציות החד-משתניות שנלמדו.

רשתות KANs לדומיננסים שונים:

כעת נתאר כמה הרחבות של KAN לדומיננסים שונים. לנוכח סדרות עתיות, רשתות KAN זמניות (T-KANs) משלבות מגנוני זיכרון, בדומה ל-RNNs ו-LSTM, לטיפול יעיל בסדרות אלו ובתלות לטוח ארוך שבהן, ומדגימות ביצועים מעולים בשימוש חיזוי רב-שלבי(multi-step forecasting). בנוסף, שינויים כמו מגנונים חיבורים gated, בדומה LSTM ו-GRU, מאפשרים ל-KANs להתאים באופן דינמי פונקציות אקטיבציה (ספליין) בגודל* בהתבסס על מרכיבות המשימה, משפרים יעילות מוביל לדריש רגולריזציה נרחבת.

בדאטה הגרפי, KANs מבוססות גרף (GKANs) פותחו לשיפור סיוג צמתים semi-supervised על ידי שיפור זרימת מידע בין צמתים, עלות ביצועיה על רשתות קובולוציה גרפיות מסורתיות (GCNs). ארכיטקטורות מבוססות KAN אלה משפרות את למידת ייצוג הצמתים ומשפרות את דיק מודלי הרגרסיה בגרפים העולות ברשות חברותיות וכימיה מולקולרית. GCNs פועלות על ידי צבירה ושינוי חזורים של מידע תכונות משוכנות מקומיות בתוך גרף, וטופסות ביעילות הן תכונות צמתים והן טופולוגיה גרף. עם זאת, GCNs משתמשות על פילטרי קובולוציה קבועים, המגבילים את הגמישות שלהם בטיפול בגרפים מורכבים והטרוגניים. כדי להתמודד עם מגבלה זו, GKAN מציג שתי ארכיטקטורות עיקריות: ארכיטקטורה 1, המכऋת תכונות צמתים לפני שימוש שכבות הממקמת שכבות KAN בין הטעויות צמתים בכל שכבה לפני הצבירה, מאפשרת התאמת דינמית לשינויים במבנה הגרפי. שיפור זה מאפשר ל-GKANs להסתגל באופן דינמי לשינויים במבנה הגרפי, ומספק גישה יותר אדפטיבית למידה מבוססת גרף.

לפתרון משוואות דיפרנציאליות, KANs מבוססות פיזיקה (PIKANs) הותאמו להציג אלטרנטיבה ניתנת לפירוש(interpretability) ויעילה לרשתות נירוניים מבוססות פיזיקליות (PINNs) המבוססות על MLPs. כאן PIKANs משתמשות במבנה אדפטיבי תלוי-גריד, מה שהופך אותן מתאימות ליישומים הדורשים דיק, כמו דינמיקת זרימה ומכניקת קוונטיים, שבהן פונקציות בסיס דינמיות עוזרות לתפוס תהליכי פיזיקליים מורכבים עם דיק ויעילות חישובית משופרים.

המחברים גם נתונים באופטימיזציה המתוגרת של KANs בשל האופי הלא-lienari של פרמטרי הספליניים מימדיות הגובהה בה נתקלים לעיתים קרובות.

סיכום:

KANs משתמשות ב-B-splines לפרמטריזציה של פונקציות של משתנה אחד, מה שהופך אותן ליתנות למידה ומאפשר מעברים חלקים בין אינטרוולים השונים עם התאמת מקומית משופרת של הדאטה. תהליך האופטימיזציה כולל התאמת פרמטרי הספליניים, כמו נקודות בקרה(control points) וקשרים, כדי למזער שגיאות בין פלט חזוי לפלט אמיתי, מאפשר למודל לתפוא דפוסי DATA מסוימים. עם זאת, תהליך זה מסובך בשל מרחב הפרמטרים הלא-lienari, קלות הממדיות, והתקורה החישובית המוגברת בשל הגמישות של ספליניים הנינתנים.

<https://arxiv.org/abs/2411.06078>

נתקלתי במאמר זהה די במקורה - תוך כדי איזה שיחה עם LLM מצוי על נושא של אמבדינגים הקשריים (contextualized embeddings) ואופן בנייתם. המאמר די קליל וחוובתי שאם כבר השקעתי 5 דקות בקריאתו אז אשקע עוד 10 דקות בסקרתו. המאמר מציע שיטה המאחתת instruction tuning (נקרא לזה ChTn) למטרת גנרטין ו-ChatIn למטרות בניית ייצוג נתונים הקשי.

מטרת ChTn גנרטיבי (generative instruction tuning) הוא די מובן ומטרתו לאמן את המודל למלא את הוראות המשמש (לדוגמא לבניית chatbot). לעומת זאת מטרת ChatIn ייצוג (representational instruction tuning) היא לאמן מודל אנקודר, הבונה ייצוג וקטורי של טקסט, בתלות בהוראות המשמש (זה די קרוב לייצוג הקשרי). יש לא מעט מאמרים הדנים בכך לפתח מודל המסוגל לבצע כל משימה כזו בנפרד - והמאמר הזה מציע שיטה שמאמנה את אותו המודל לעשות את שני הדברים האלו (לא באותו הזמן).

השיטה פשוטה: הרכבה של פונקציית לוס מושני לויס שאחד מכם הוא - ChTn גנרטיבי והשני - ChatIn ייצוג. לכל אחת מהמשימות מחובר למודל ההתחלתי ראש מאומן (כמו בלוקים של טרנספורמרים למיטב הבנתי).

از למשימה הראשונה המחברים משתמשים בLOS הסטנדרטי של מודלי שפה גנרטיביים כולל חיזוי של טוקן הבא עבור התשובה בלבד. למשימה השנייה המחברים משתמשים בLOS הניגוד (די סטנדרטי במשימות כאלו) והמנסה לקרב אמבדינגן של השאלה עם התשובה הנכונה ולהרחיק את האמבדינגן של השאלה עם תשובה לשאלת אחרת. ייצוג של הטקסט מחושב על ידי מודל באופן דו כיווני (אנדוקר) כאשר האמבדינגן הוא הממוחזע של האמבדינגן של כל הטוקנים של הטקסט. כמובן שככל משימה מקבלת פרומפט משלה.

זה זה - סקירה קצרה כמו שהבטחתי....

<https://arxiv.org/abs/2402.09906>

המאמר היומי של מילק - 17.03.25

JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation

זמן לא סקרתי מאמר על מודלים גנרטיביים מולטימודליים. מודלים אלו מאומנים לא רק לגנרט נתונים מכמה סוגים (במקרה של JanusFlow של שפה טבעית ותמונות) אלא גם לבצע משימות הכרוכות בהבנה של הקשרים בין המודוליות האלו. למשל מודל מולטימודלי בתחום שפה ותמונות צריך להיות לענות על שאלות על תמונה. המודל מורכב ממודל עיקרי (הנקרא LLM) וכמה אנקודרים ודקודרים המיועדים לייצוג נתונים מודוליות שונות והפיקתו של ייצוגו לפיסת נתונים (דקודרים). כל המודלים במאמר מבוססים על הטרנספורמרים באופן מאוד לא מפתיע.

המאמר מציע שיטה לאמן מולטימודלי (הנקרא LLM במאמר) כזה כאשר הפרט המעניין לגבי הוא שימוש באנקודרים שונים לשפה ולתמונות (ברוב המודלים המולטימודליים משתמשים באותו מודל backbone). בגדול במהלך האימון המודל למד לחזות את הטוקנים של תשובה על פרומפט נתון כאשר פרומפט ותשובה יcolsם להיות גם טוקן ויזואלי (ייצוג של פאץ' של תמונה) וגם טוקן רגיל (=סדרת אותיות). בנוסף הפרומפט יכול להיות שילוב של טוקנים ויזואליים וטוקנים של השפה במשימת question answering visual. בנוסף (לא מופיע במאמר זהה בצורה מפורשת אך נעשה במקרים מולטימודליים אחרים) המודל מאומן גם על נתונים טקסטואלי בלבד (כמו ב-pretraining של מודל שפה רגיל)

כמה פרטים על המודלים השונים (פרט ל-LLM) המופיעים במאמר. עבור DATA שפתית הטוקנים עבריים אנדוקר מאומן (נקרא enc pad) - אחרי הטוקנים עבריים שכבה לנארית מאומנת. עבור DATA ויזואלי יש אנקודר סטנדרטי לא מאומן המבוסס על VAE ואחרי יש עוד אנקודר מאומן. מכיוון שהמודל הגנרטיבי לתמונות הינו מודל

דיפוזיה שימוש ב-VAE (חלק בלתי נפרד של מודלי דיפוזיה גנרטיביים) לא צריך להפתיע. בנוסף כאמור יש שני דקודר מאומנים שאלייהם מזינים הייצוגים הנבנים על ידי LLM.

המאמר מציע שיטה תלת שלבית לאימון המודלים כאשר כל שלב "אפשרים" יותר ויותר מודלים (כולל LLM) אשר בשלב האחרון מאומנים את כולם פרט ל-VAE.

מודלי דיפוזיה במאמר מבוסס על (RF) rectified flows המנסה למפות את הדטה מהתפלגות פשוטה (גאומית) להתפלגות הדטה בצורה ישירה ככלור המסלול בין x_0 הגאומטי ל- x_1 של הדטה הוא ישר. ככלור כל נקודה x_t במסלול זהה היא צירוף קמור של x_0 ו- x_1 . بغداد מודל הדיפוזיה מאומן לשער את המהירות הקבועה (v) השווה $x_1 - x_0$ עבור כל נקודה x_t במסלול. הדגימה מבוצעת על ידי פתרון משווה דיפרנציאלי המתארת התקדמות של x_t עם מהירות v (שיטת אוילר). מודל דיפוזיה המאומן במאמר הוא לטנסי.

פרט מעניין על המאמר: אחד האיברים בפונקציית LOSS של מודל דיפוזיה קונסת אותו על אי התאמה של ייצוג הפנימי המורעש (המוחש על שכבות הביניים של המודל) לייצוג התמונה הנקיה המוחש על אנקודר חזק (understanding encoder). וכמבען יש classifier guidance של מודל דיפוזיה (קלאסי)

מאמר כתוב יפה ודי ברור - מומלץ!

<https://arxiv.org/abs/2411.07975>

19.03.25 המאמר היומי של מייק - EFFICIENTLY LEARNING AT TEST-TIME: ACTIVE FINE-TUNING OF LLMS

בתוקפה האחרונה השיטה הcyFi פופולרית להתקנת מודלי שפה למשימה ספציפית היא במידה in-context או ICL. بغداد אנו מספקים למודל, בתוך הפרומפט, כמה דוגמאות לביצוע משימה והמודל "לומד" איך לבצע אותה ללא שום שינוי במשקליו. ICL מתאפשר עקב האופי האדפטיבי של הטרנספורמרים (מנגן ה-attention בתוכו) המצלחים "לעדרן את אופן החישוב שלו" כפונקציה של קלט.

המאמר דן בשיטה אחרת לאדפטציה של מודל למשימה נתונה בזמן סטט (המאמר קצר מערבב את המושג של טסט אוינפרנס) המערב fine-tune קليل של המודל על סמרק הפרומפט שמודן אליו. להבדיל מ-CL-ICL השיטה המוצעת (SIFT) Selects Informative data for Fine-Tuning כן משנה את משקל המודל (מצבעת צעד אחד של מורד הגרדיאנט - gradient descent). למעשה SIFT (ד"א יש שיטה בשם זה גם בעיבוד תמונה מהעידן לפני הרשתות) מציעה שיטה לבחירה של דוגמאות מהדאטasset לפין טין של מודל עבור פרומפט נתון.

המחברים טוענים שבחרות דוגמאות הcyFi קרובות לפורומפט במרחב הלטנטי מבחינת מרחק קווין או מכפלה פנימית (NN or nearest neighbors or neighbors) היא תחת-אופטימלית ועלולה להביא לדוגמאות מיותרות הפגעות בביטוי פיין טין. במקום לשלוフ דוגמאות הדומות ביותר לפורומפט, SIFT בוחרת את אלו שמספקות את מירב המידע החדש, וכן מושגה התקנה טובה יותר של המודל עם מינימום חישובים נוספים.

הגישה המוצעת מערבת שיעור רמת אי-ודאות של תשובה המודל בהינתן הדוגמאות שבחרנו ל-FT (לאחר FT הכוונה). בפרק הבא אסביר למה זה חשוב בעצם.

הערכת אי-ודאות להנחה FT ולמה זה בכלל חשוב כאן?

שיטות FT רבות משתמשות על שליפת דוגמאות דומות בהתבסס על דמיון קווין או מרחק אוקלידי. אך גישה זו לוקה בחסר: היא אינה מבדילה בין דטה רלוונטי לזה שמיותר. שתי דוגמאות דומות מאוד עשויות להכיל את אותו

מיעוד, ולכן אחת מהן אינה תורמת לתוצאות FT. כדי לפטור זאת, המחברים מציעים שיטה להערכת אי-הוואדות של המודל בתשובתו לאחר FT. אם המודל בטוח מאוד בתשובתו אחריו FT, הוסף דוגמא לא תשיפע משמעותית. אך אם אי-הוואדות גבוהה, בחירה חכמה של דוגמאות יכולה לשפר את ביצועי המודל משמעותית והאתגר הוא למצוא את הדוגמאות הללו ביעילות.

מדידת דמיון במרחב הסומי באמצעות פונקציית קרנל

כאמור הבסיס לשיטת הבחירה של SIFT הוא מדידת הדמיון בין דוגמאות למרחב לטנסי. כדי לכמת את הדמיון זהה, המחברים משתמשים בפונקציית קרנל - שהיא מוגדרת בתור מכפלה פנימית בין היצוגים הלטנטיים של הדוגמאות. פונקציה זו מקבלת שני רצפים ומחזירה ציון דמיון—גובה עבור סדרות דומות ונמוך עבור רצפים שונים. באמצעות פונקציית קרנל זו בונים מטריצה קרナル עבור הדוגמאות שנבחרו ל-FT והפרומפט עצמו. לאחר מכן מגדרים מודל דמה (surrogate model) שמטרו לשערך את ביצועי ה-LLM לאחר FT על הדוגמאות שנבחרו. באמצעות מודל זה בונים (זה קצת כבד מתמטי) את השיעור של אי-ודאות של המודל אחרי הוספה של דוגמא X מהדאטסט לסת הדוגמאות עליהם יבוצע הטיבוב. בסופו של דבר בוחרים דוגמא המזערת את אי-ודאות עבור הפרומפט ומוסיפים אותה לסת הדוגמאות זה.

במילים פשוטות הגישה המוצעת מאזנת בין שני שיקולים מנוגדים:

- רלוונטיות: הדוגמאות הנבחרות צרכות להיות עדין רלוונטיות לפרומפט.
- גיאון: הדוגמאות אין אמורות להכיל מידע חוף ומיותר.

במוקם לבחור דוגמאות בבת אחת, SIFT בוחר כל דוגמה באופן הדרגי, תוך שימוש בפונקציית קרנל כדי לקבוע את הערך המוסף שלה.

1. אם מועמד חדש דומה מדי לדוגמאות שנבחרו בעבר, הוא נדחה, מכיוון שהוא אינו מוסיף מידע חדש.
2. אם המועמד רלוונטי אך מכיל פרטים חדשים, הוא נבחר כדי להפחית את אי-הוואדות.
3. אם המועמד אינו קשור לפרומפט כלל, הוא נשאר מחוץ לתהילה.

<https://arxiv.org/abs/2410.08020>

המאמר היומי של מייק - 20.03.25 (softmax is not enough (for sharp out-of-distribution

המאמר הזה מציעה שיטה לשיפור ביצועי ההכללה עבור מודלי טרנספורמרים מזוויות די לא צפוייה. המחברים מציעים שיטה להתחממות עם מה שנקרא DISPERSE (או פיזור בעברית) של מקדים ה-attention בטרנספורמרים. זה מתבטא למשל באפשרות (לפי המאמר) של הטרנספורמרים למקד את מקדמי ה-attention במספר טוקנים קטן (יחסית לאורך הסדרה). זה חשוב למשל בשאלות כמו מציאת מקסימום של סדרת מספרים נתונה או שאלות בסגנון "מחט בערימת השחת" (needle in a haystack) כאשר המודל מتابקש מקטע קצר לא הקשור בטקסט מסוים (יחסית ארוך).

הმחברים טוענים שאחת הסיבות לביעות אלו היא פיזור מקדי ה-attention במנגנון הטרנספורמרים. מקדים אינם מוחשבים עם פונקציית סופטמקס ה"מנרמלת" את המכפלות הפנימיות של וקטורי K-Q עבור כל טוקני הסדרה. לפי המאמר הבעיה קשורה לכך שעבור קוונטקטים ארוכים לsoftmax במינוח טרנספורמרים העמוקים יש "נטיה למראה את פלט הסופטמקס".

אחת הדריכים להתמודד עם התופעה זו היא להוריד את הטמפרטורה אבל זה עלול להעלות סיכון לשגיאת במרקם בהם הלוגיט (maskable attention לא מormal) של הטוקן הנכון יותר קטן מהלוגית המקסימלי. כדי להתמודד עם התופעה המבקרים הציעו גרסה חדשה של סופטמакс בה הטמפרטורה תליה באנטרופיה של הטוקנים.

הם אימנו מודל עבור מקרים שבהם הלוגיט של הטוקן הנכון אינו מקסימלי כאשר המטרה הייתה למקסם את הסתברות הדגימה של הטוקן הנכון (אחרי מגנון-h-attention ו-FFN). מטרת המודל היה לחשב ערך אופטימלי של טמפרטורה כפונקציה של אנטרופיית של משקל attention לא מormalים. הנוסחה של הטמפרטורה יראה הופכית (1 חלק) של פולימום מחזקה 4. אציין כי הטמפרטורה מחושבת בזמן האינפרנס כתלות באינטראופית הטוקנים לפי המודל הזה.

המבקרים רואו אמפירית כי עם הטמפרטורה האדפטיבית מקטינה פיזור משקל h-attention. למרות שהטמפרטורה האדפטיבית האופטימלית יורדת עם עלייה באנטראופית הלוגיטים היא גורמת לפחות שגיאות של המודל יחסית למקרה שהוא נקבעת באופן קשה.

<https://arxiv.org/abs/2410.01104>

21.03.25 המאמר היומי של מיליק -

LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law

המאמר משליך תשומת ליבי כי מופעים בשם מודלי שפה ומערכות דינמיות שאנו מחבב מהזמינים העלייזים של ממהה (state-space models). המאמר טוען שמודלי שפה מגינים ביצועים טובים בהבנת מערכות דינמיות מגוון סוגים כולל מערכות סטוכסטיות, כאוטיות, רציפות וכדומה. וכל זה קורה ללא שימוש טיבוב (fine-tune) - כלומר קצר הנדסת פרומפטים ומודל השפה שלהם מביןמערכות דינמיות.

לאחרונה היו לא מעט מאמרים שניים לפצת "מערכות דינמיות" הניתנות על ידי דגימות שלהם עם LLMs דרך יצירת דגימות חדשות מהם (מהמערכת הדינמית) באמצעות LLMs. ההיגיון כאן די פשוט - אם מודל שפה יידע לגנרט מהתפלגות המושראית על ידי מערכת דינמית, אז נראה הוא מבין אותה.

המבקרים לקחו גישה אחרת יותר ישירה - הם הראו שניתן ממש ליצור התפלגות של מערכת דינמית באמצעות LLM כאשר המערכת היא מרקובית. כלומר אם התפלגות דגימה הבאה בזמן $t+1$ תלולה רק במצב המערכת בזמן t ולא בעבר. עברו מערכת דיסקרטית התפלגות זאת נתונה על ידי מטריצה של הסתברויות מותנה המכילה את הסתברויות של מצב $\{t+1\}_x$ בהינתן מצב t_x בזמן עברו כל הערכים האפשריים שלהם. עברו מערכות רציפות ניתן לבנות מטריצה כזו על ידי דיסקרטיזציה של הערכים של מצבים המערכת.

המאמר מראה שמודלי שפה מצליים לבנות את מטריצות מעברים בצורה לא רעה במיוחד במצבים שיש יחסית מעט מצבים אפשריים. המהלך בין ההתפלגות החזiosa על ידי מודל שפה לבן התפלגות ground truth נמדד במאמר עם מרחק Bhattacharyya שנתקלתי בו רק בפעם השנייה במאמרי deep learning. אציין שהמאמר מציג תוצאות טובות גם עבור מרחקים divergence (divergence) אחרים כמו JSD ו-KL. המבקרים מציעים דרך טרייקית לבנות את המטריצה זו עם LLM - מי שרצה לצלול לעומק, תראו פרק שנקרא Hierarchy-PDF algorithm.

וזה זה, היום זה היה קצר ...

<https://arxiv.org/abs/2402.00795>

המאמר היומי של מיק - 22.03.25 Physics in Next-token Prediction

המאמר זה לא רגיל. זה מתייחס מהשם שלו: הר' איר חיזוי של טוקן הבא (NTP) יכול להיות קשור לפיזיקה. מתרבר שהקשר זהה קיים והוא עובר דרך תורת המידע (information theory). מי שמכיר אותו יודע שאנו מאוד מתעניין בהיבט מידעי (אינפורמציוני) שקיים בתחום למידת מכונה, בධיסת מידע על ידי המודלים, על אף הידע נשמר במודלים מאומנים וכדומה. והמאמר הזה מדבר בדיק על הנושאים האלה ולמרות שאין בו מתמטיקה יותר מדי מורכבת הוא די עמוק (בספק הצלחתו להפנים אותו אני במלואו:).

נתחיל בלכין שלפי חוק שנון (עם טויסט קטן) הטוען כי כדי להסביר מילה $\{t+1\}_x$ לאחר שהערכנו \hat{x} מילים הינה שווה לאנתרופיה מותנית H של $\{1+1\}_x$ בהינתן $t_x, \dots, 1_x$ או מידע עצמי I . האנתרופיה H שווה במקורה זהה $-log p$ של הסתברות מותנית של $\{1+1\}_x$ בהינתן $t_x, \dots, 1_x$. נובע מכך (די בקלות) שמספר הביטים הנדרש כדי להסביר את כל המילים מדאטהסט D כלהו הינו סכום של האנתרופיות המותניות אלו עברו $\hat{x} - D$.

עכשו נניח שיש לנו מודל שאימנו אותו לחזות טוקן הבא בהינתן ההקשר (כלומר הטוקנים הקודמים), למשל מודל שפה. מספר הביטים הנדרש להסביר את אותן המילים מדאטהסט D מחושב לפי אותה הנוסחה, ככלומר סכום של אנתרופיה מותנית של $\{1+1\}_x$ בהינתן $t_x, \dots, 1_x$. אבל הפעם, כאשר המודל משמש לחיזוי אנתרופיה זו (כלומר הסתברות מותנית) נראה שאנצריך פחות ביטים להערכת אותן דאטהסט D . למה זה בעצם קורה? אולי נעלם הפרש בין מספר הביטים שצריך כדי להסביר את D בלי המודל ועם המודל?

麥肯ชน שאינפורמציה לא יכולה לילכת לאיבוד ההנחה היא שהמודול כבר אותו (למד). המאמר קורא למידע זה השמור בתוך המודול מידע אפקטיבי של המודול על דאטאסט D (או משימה). המאמר גם מגדר \hat{x} שהיא הקיבולת של המודול בטור היחס בין המידע האפקטיבי של המודול למספר הפרמטרים של המודול (בביטים). בנוסף נציין שהוא מאד מעניין: כי מספר הביטים שצריך כדי להסביר את D עם המודול הוא לו (cross-entropy) של המודול עבור D מוכפל ב $|D|$.

אם נקשר את כל הגדלים שהגדכנו קודם לקבל הקיבול המידעי הראשון שמוגדר במאמר: $(|D| = N - L - H)$, כאשר N זה מספר הפרמטרים של המודול, L זה קרוס-אנתרופי לוט של המודול על הדאטאסט D , ו- H היא האנתרופיה ההתחלתית של D . במהלך האימון H ו- N נשארים קבועים ו- $|D|$ הוא מספר הטוקנים שהמודול "ראה" במהלך האימון. ככלומר האימון הוא תהליכי דחיסת דאטאסט D והערכות מידע ממנה למודל המאמן.

המאמר גם מגדר את החוק הקיבול המידעי השני המתאר את האנרגיה המינימלית הנדרשת להעברת אינפורמציה מ- D למודל. היא פרופורציונלית ל N ו- L וכן מופיע בו גם טמפרטורה T (לא לבלב עם הטמפרטורה של LLMs) וגם קבוע בולצמן k - מודה שלא הצלחתני להבין את המשמעות של השניים האחרונים (T ו- k).

בתבוסס על תורה זו המבקרים מגיעים למסקנות מעניינות לגבי אימון המודול וגם משווים את החוקים שניסחו עם חוקי סקלינינג של מודלי שפה. מי שמתעניין בזה, מוזמן לצלול - מאמר מרתק.

<https://arxiv.org/abs/2411.00660>

המאמר היומי של מיק - 24.03.25 STAR ATTENTION: EFFICIENT LLM INFERENCE OVER LONG SEQUENCES

הסקירה זו הולכת להיות קצרה. אפילו מאד קצרה. המאמר המסתוקר מציע שיטה לאופטימיזציה של מנגנון -attention בטרנספורמרים עבור מקרה שיש לנו כמה מכונות (נקרא hosts במאמר) להריץ את מודל השפה

שלנו. המאמר הוא של חברת אנוידיה דרך אגב זה דואק לא מפתיע כי (לפי השמועות 😊) יש להם די הרבה משאבי חישוב.

המודל מחזיר אותנו לתקופה העלייה לפני 7-4 שנים שהייתי עד למבול של מאמרים שהציגו אופטימיזציות שונות למנגנון *the-h*-attention. אולם בטח זוכרים *LongFormer*, *Performer*, *Reformer*, *LinFormer* ועודומה (שחלקים סקרתי בזמןנו) - היה גם *Star Transformer* דרך אגב. רוב השכלולים שהוצעו בתקופה היא דיברו על איך ניתן לזרע את *the-h*-attention בלי לפגוע משמעותית בביטוי המודול - כאשר המודול רץ על מכונה אחת. אז היה מאד פופולרי האירורים הריבועים שהיה מצויר בהם הפרטן של *the-h*-attention כלומר באיזה טוקנים טוון נתון מתחשב כדי לבנות את ייצוגו הקשי (contextualized embedding).

המאמר הזה מציע מנגנון *attention* שניינן לקרווא לו ל蹶אל (מציר לי קצת רשותת קונבנצייה על *the-e*-*inductive bias* שלהם המנצל את התלוויות הлокאלית בתמונות). במאמר זה משוחה טיפה יותר מורכב (מציר גם *LongFormer*). כאן מחלקים את חלון הקשר לכמה קבוצות של טוקנים *c_1, ..., c_n*. כל טוון בכל קבוצה *c* פרט *-c_1* מחשבת את *the-h*-*attention* עם הטוקנים בתוך אותה הקבוצה *-c_1* בלבד כאשר טוקנים של *c_1* מתחשבים בכל הטוקנים לבניית האמבדינג שלהם. כלומר הקבוצה הראשונה של הטוקנים משפיעה על האמבדינג של כל הטוקנים וגם בעצמה מושפעת מכל הטוקנים בחולון הקשר. המחברים טוענים שלאווספה של *c_1* (שהה למעשה תחילת הפרומפט) לכל קבוצות הטוקנים המנגנון סובל מירידה רצינית בביטויים

כמובן ניתן למקבל את התהילך הזה בקளות בין כמה מכונות (*hosts*) כאשר כל *host* מחשב את *the-h*-*attention* הлокאלי שלו וגם *the-h*-*attention* עם *c_1* (בשתי שלבים). כל *host* גם שומר את סכום האקספוננטים של *Q* ו-*K* (מכנה של הסופטמקט) עבור הטוקנים שלו. לאחר מכן כל הסכומים הללו מועברים ל-*host* נוסף שמנרמל את כלם עם סכום אקספוננטיים של כל *hosts* ומחשב את הייצוג הסופי של כל טוקנים.

מנגנון זה מאפשר חישוב מקובל ומהיר יותר של *the-h*-*attention* (פחות מכפלות מטריצות) כאשר לטענת המחברים הפגיעה בביטויים לא משמעותית.

<https://arxiv.org/abs/2411.17116>

26.03.25 המאמר היומי של מייק - DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

המאמר שנסקור היום שיר בתחום שלא הכרתי אז יש סיכוי שהוא שגיאות בסקרה למרות מאמין הכבירים. המאמר דין באופטימיזציה אימון של מודלי שפה כאשר יש ברשותנו נתונים דאטאטטיסים מdomains שונים. בינהמה יותר מתמטית המחברים מציעים שיטה למשך של הדאטאטטיסים השונים במהלך האימון. כלומר אם יש לנו *D* דאטאטטיסים המטרה היא למצוא וקטור *d*-ממדי *a* של מספרים אי שליליים המסתכנים *-1* כאשר *k* \in *A* היא ההסתברות לדגם דוגמא מدادאטט *k*. כלומר אנו מרכיבים את סט האימון שלו בשני שלבים: בשלב הראשון בוחרים דאטאטט עם דגימה *-a* ולאחר מכן בוחרים דוגמא הדאטאט הנבחר.

דרך אחת פשוטה היא לבחור את *-a* פרופורציונלית לגודל הדאטאט *D* כלומר ככל שدادאטט גדול מד' הוא יבחר יותר פעמים. אפשר לדוגמה גם בצוירה יוניפורמית אשר כל דאטאטט יבחר בהסתברות *1/D* כאשר *d* הוא מספר הדאטאטטיס. יש שיטות שבוחרות *a* לפי איקוט הדאטאט ומידפים דאטאטטיסים איקוטיים יותר על פני אלו שפחות איקוטיים.

אבל איך לבחור את α בצורה שתמקסם את ביצועי המודל המאומן? זו השאלה שעלה מנסה המאמר לענות. אחת השיטות היא לנסות כל מיני ערכים של α ועבור כל אחד לאמן את המודל (brute-force). עבור מודלים גדולים ומספר גובה של דאטהסטים D המחיר החישובי (= עלות) עשוי להיות עצום. השאלה האם ניתן לעשות משהו חכם מזה?

התשובה על השאלה זו היא כן וזה מה שהמחברים מציעים. בשלב הראשון המוחברים מציעים לאמן מודל M_{ref} קטן עם $\alpha = \alpha_{\text{ref}}$ (נגד יוניפורמי). החברים מציעים להשתמש בשיטת distributionally robust (DRO-LM language modeling) (החוקים) יחסית לשגיאה של M_{ref} (הפרש השגיאות בין M_{ref} למודל המאומן). השגיאה במרקבה זהה היא לוג של הנראות של הטוקן הנכון (עבור כל מודל מציעים עבור כל הטוקנים עבור כל דאטהסט בנפרד).

אם רואתם כאן בעיית minimax , אתם צודקים. בצורה איטריבית ממקסימים (כלומר עושים מעלה הגרדיאנט או gradient ascent) את הפרש השגיאות (עבור באציגים של דוגמאות) מעל α ולאחר מכן ממציעים את הפרש השגיאות מעל משקל המודל המאומן (כלומר gradient ascent). וקטור המשקל α הסופי שנבחר על ידי מוצע של כל וקטורי α עבור כל האיטרציות של בעיית המינימקס זו. מעניין שהבאציגים נדגים באקראי עבור כל האיטרציות. בשלב האחרון מאמנים מודל גדול עם α שמצאנו בצורה זו.

מקווה שהצלהתי להסביר את זה ברור...

<https://arxiv.org/abs/2305.10429>

28.03.25 המאמר היומי של מייק - UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining

אני ממשיר לסקור מאמרם בנושא של אופטימיזציה בחירת דата לאימון מודלים בפרט עבור מודלי שפה שאנו כה אוהבים. הבעיה ניתנת לניסוח די פשוט: יש לנו כמה דאטהסטים $\{h_i, D_i\}_{i=1}^n = D$. המטרה שלנו לבחור שילוב "אופטימלי" של דאטהסטים אלו לאימון המודל כאשר יש לנו תקציב B של מספר הטוקנים/סימבולים (במאמרם משתמשים בסימבולים) הכלול לנו שהמודול "רוזה במהלך האימון" (ניתן לתרגם את זה ל-FLOPs בהינתן ארכיטקטורה של המודל). במילים פשוטות אנו מנסים להבין את ניתן לדגום דטה M כדי לקבל מודל בעל ביצועים הטוביים ביותר אחרי האימון תחת תקציב B . כМОון ש"הטוב ביותר" ניתן להגדיר במספר מובנים אבל אנו פחות נתמקד בנקודת הוויה זו ונדון בעיקר איך ניתן לאזן בין דוגמאות ממקורות (דאטהסטים) שונים באימון המודל.

המאמר דין בתרחיש של אימון מודלים מולטי-שפתיים כאשר יש ברשותנו דטה בכמה שפות. הגישה פשוטה היא להעניק תקציב שווה לכל דאטהסט (=שפה) ככל דאטהסטים בגודל שונה יומנו מספר אפוקים (epoch)

שונה כאשר דאטاهסティים **קטניים**(שפوت לא נפוצות) יותר יאומנו במספר אפוקים גבוה יחסית לדאטاهסティים גדולים (שפונות פופולריות יותר). המאמר טוען שיחס התאמת זה עלול לגרום לביצועים ירודים של המודל. המחברים מציעים שיטה מאוד אינטואיטיבית ו פשוטה לאיזון של מספר האפוקים לדאטاهסティים שונים תחת תקציב B.

המאמר קובע מספר מקסימלי לאפוקים N שיינתן לכל דאטהסט. התהילה מתחילה בדאטהסט הקטן ביותר (שפה הכי פחות נפוצה) המחשב לפיא מספר הסימבולים i_C בשפה את מספר האפוקים לדאטהסט זה בהתאם לתקציב מסויע שפה(B מחולק ב-|D|). אם מספר האפוקים i_E עבור השפה הנוכחית עולה על N, הוא(מספר אפוקים) נקבע להיות N. לאחר מכן מוחסרים מ-B את תקציב לדאטהסט הנוכחי ומוחסרים תקציב מסויע i_u פר שפה (נותרו 1 - |D| שפות). אז ממשיכים את התהילה עבור כל השפות בדאטהסט. המחברים מצינים שהם לא משתמשים במספר הטוקנים כדי לאמוד "גודל אפקטיבי" של כל שפה בדאטהסט עקב מורכבות של טוקניזציה של דאטהסט מולטי-שפה.

בשלב האחרון מנורמלים את כל התקציבים i_u עם הסופטמקס כדי לקבל ההתפלגות \mathbf{p} שמננו דוגמים דאטה עבור כל השפות. ניתן להשתמש ב- \mathbf{p} עם טמפרטורה α כדי להחליק או להבליט את ההתפלגות (מעלים i_k בחזקה 1/α) ולדגם ממנה את השפות (=דאטהסטים).

זה וזה - מחר עוד מאמר בנושא...

<https://arxiv.org/abs/2304.09151>

30.03.25 המאמר היומי של מייק - Efficient Online Data Mixing For Language Model Pre-Training

ממשיר בלסוקור את קו המאמר בנושא אופטימיזציה של אימון מודלים (בפרט מודלי שפה) כאשר יש בידינו כמה דאטהסטים שונים. מכיוון שכבר הגדרתי את הבעיה בסקירות מ-26.03 ו-28.03 לא אעשה זאת בסקירה זו ומיד אתחיל מהסביר הרעיון העיקרי של המאמר.

המאמר ניגש לבעיה בצורה שונה מאשר שני המאים הקודמים שסקרו אף לדעתם (למרות המורכבות מתמטית מסוימת) הגיעו המוצעת במאמר היא מאוד אינטואיטיבית. המחברים מנסים לפתור בעיה בunitiy בניית דאטהסט D לאימון מודל שפה מהדאטהסטים D_1, \dots, D_k עם מה שנקרא Multi-Arm Bandits או MAB. אזכור בעיה MBA מוגדרת באופן הבא: יש לנו כמה חances מצל עם הסתברויות זכיה π_1, \dots, π_k שלא ידועות לנו מראש. המטרה היא למצוא אסטרטגייה בחירת מוכנה הממקסמת את הזכיה (נגיד, התוחלת שלה) כאשר יש לנו N נסיבות.

שימוש לב שבועית אופטימיזציה האימון שלנו די דומה ל-MBA - גם פה אנו צריכים למצוא את אסטרטגיית בחירת דאטהסטים לאימון בלי שאנו יודעים מה "ההשפעה" של כל דאטהסט לתוצאות האימון הסופית. בלי להיכנס יותר מדי עמוק למתמטיקה (תהליך החלטה מרקובי, ההתפלגות גיבס וכדומה) המטרה למצוא ההתפלגות π_1, \dots, π_k על הדאטהסטים שלנו למקסום ביצועי המודל המאומן. הקאץ' כאן שההתפלגות זו משתנה עם האיטרציות כאשר

אייטרציה במקורה הזה היא צעד אחד (או מספר כלשהו אחר קבוע מראש) על הדטה מהדאטסט \mathcal{D} שנבחר באיטרציה זו.

כלומר כל פעם אנו בוחרים דאטסט עם התפלגות הנווכית \mathcal{P} , מאמנים את המודל על הדטה מהדאטסט הנבחר ומעדכנים את \mathcal{P} בהתאם על תוצאות האימון. כמובן נשאלת השאלה איך ניתן לקבוע \mathcal{P} עבור אייטרציה הבא על סמך התוצאות של האיטרציה(בחירה דאטסט) הקודמת. וכך אנו מגיימים למה שנקרא תגמול (reward) שהוא משקף את "ההצלחה" בבחירה דאטסט \mathcal{D} באיטרציה זו. אם האימון על \mathcal{D} היה מוצלח, אנו רוצה להגדיל את הסתברותה שלו (על חשבו האחרים) כאשר אם הוא פחות מוצלח אז צריך להקטין אותה.

אוקי, אז מה זה בעצם התגמול כאן? התגמול כאן היא מידת השמודל ירוויה מהדטה מדאטסט \mathcal{D} ככלומר למד יותר סוג של information gain או IG. המאמר מחשב את IG בטור פרפלקסיטי (שהיא אקספוננט של הלוג) על הדטה של דאטסט \mathcal{D} . לו זה משוערך על סמך באצ' מהדאטסט. בנוסף יש גם עניין של exploration כי אנו לא רצים "להקטין דרישטי" את הסתברות בחירה של דאטסט מסוים על סמך מעט באצ'ים ואז מגדים (כמו שמקובל ב-MBA ובשיטות אחרות של RL) כל הסתברות \mathcal{P} במספר קטן ϵ שיורד עם האיטרציות.

از האלגוריתם הסופי מכיל 3 שלבים:

1. עדכן הסתברויות בחירה $\mathcal{P}_1, \dots, \mathcal{P}_k$,
2. דגימת דטה מהדאטסטים $\mathcal{D}_1, \dots, \mathcal{D}_n$ לפי הסתברויות אלה ואימון מודל על דטה
3. עדכן נוסף של הסתברויות בהתאם על המודל המאומן בשלב 2

מאמר מומלץ - נהניתו לצלול אליו

<https://arxiv.org/pdf/2312.02406>

01.04.25 המאמר היומי של מייק - OPTIMIZING PRETRAINING DATA MIXTURES WITH LLM-ESTIMATED UTILITY

המאמר שאסקור קצורות היום הוא הכללה של המאמר שסקרתי ב-28.03 על השיטה, שנקראת xMax (למרות שיש בינהם הבדלים די מוחשיים ולדעתו הוא יותר דומה למאמר שסקרתי לפני יומיים ב-29.03). הסקירה הזו היא אחרונה (לעת עתה) בקן המאמרים שסקרתי על אופטימיזציה אימון של מודל במקורה שיש בידנו כמה דאטסטים $\mathcal{D}_1, \dots, \mathcal{D}_n$. בעצם המטרה כאן כמו ב-3 הסקירות האחרונות היא להציג התפלגות $\mathcal{P}_{w_1, \dots, w_n}$ לדגימה אופטימלית מדאטסטים אלו כדי למקסם את ביצועים המודל האימון על דאטסט וlidzha מוגדר מראש. כל זה תחת אחד משני אילוצים: אילוץ של מספר הטוקנים עליהם מאמון המודל(data-constrained) או אילוץ על compute (בגadol ניתן לנוכח כל אחד מהם דרך השני לעניות דעתך).

המאמר מנשח בעיית אופטימיזציה אימון בצורה די מקורית שלא ראיתי קודם. הניסוח הוא בעית אופטימיזציה (מצער), שימושו לב יש טעות בנוסחה 2, צריך להיות שם argmax ולא min (argmin). פונקציית הלוס שלה

מורכבת משני איברים. האיבר הראשון הוא הנורמה של ההפרש בין המכפלה הפנימית של הוקטור w ווקטור המכיל את הביצועים (המנורמלים להיות בין 0 ל-1) של המודל על ביצ'מארקים ('ש כמה עבור כל דאטאסת') עבור הדאטאסטים D_1, \dots, D_n (נקרא utility במאמר) לבין ווקטור האחדות(1 במאמר). נראה קצת מוזר מהippet הראשון אבל הערך המקיים של האיבר זהה מתקבל כאשר מוצע הביצועים של המודל על כל הביצ'מארקים עבור כל דאטאסת הם מושלמים (שווים ל 1 המשמן ככל הנראה ביצועים מקסימליים).

בנוסף יש שם איבר רגולרייזציה $\|w\|_2^2$ המקבל ערך מינימלי עבור ווקטור w המכיל את אותם הערכים (כלומר $1/n$). ככלمر רוצים לקנוס את המודל על הקצאתו אותה הסתברות דגימה שווה לכל הדאטאסטים שזה בסך הכל ד' i הגיוני כי אנו מעוניינים שהמודול "יראה" דאטה כמה שייתר מגוון. בנוסף יש אילוצים על w שהם כופים על להיות ווקטור התפלגות ועוד אחר המגביל את תקציב הטוקנים הכלול של כל האימונים.

המאמר משתמש בשיטת Splitting Conic Solver (נראה לא טריומיאלי אבל לא צלaltı לעומק) לפתור בעיית אופטימיזציה זו. כМОון שגישה זו דורשת חישוב של פונקציית יעד המתוארת בפסקה הקודמת. זה כולל אללאוציה של ביצועים עבור ח'ביצ'מארקים וזה ד'יקר חשיבות. המחברים מציעים שיטה לעשות אללאוציה זו באמצעות מודלי שפה. מודל חזק (המ השתמשו בLLAMA גדול) מתבקש לסקם משימה של כל ביצ'מארק ועל סמך הדוגמאות של ביצ'מארק. בנוסף המודל גם בונה פורומפט שנקרא prompt (עבור המודל utility prediction המאומן) המועד לשערוך ביצועים של המודל המאומן על הביצ'מארק. המטרה (לדעתי כי חילוק זהה פחות הבנוי) בהינתן מספר קטן של דוגמאות לתת שערוך ביצועי המודל על הביצ'מארק (יש 5 ציונים אפשריים).

זהו הגיעו לסימן קוו הסקירות בנושא של אופטימיזציה אימון של מודל כאשר יש לנו כמה דאטאסטים. נתראה בנושאים הבאים.

<http://arxiv.org/abs/2501.11747>

02.04.25 - המאמר היומי של מיק

SymDPO: Boosting In-Context Learning of Large Multimodal Models with Symbol Demonstration Direct Preference Optimization

היום אני עושה מעבר חד בנושא הסקירה וסוגר מאמר על אימון מודלים מולטי-מודליים (בפרט MLLMs). המאמר מציע שיטה לאימון מודלים למשימת למידה *in-context*-to-*in-context* שבה המודל מקבל כמה דוגמאות (הדוגמות) שככל הדוגמה מכילה תמונה, שאלה ותשובותה עליה. המודל מתבקש, בהתאם על הדאטה שקיבל (הדוגמות) לענות על שאלה לגבי תמונה נוספת (עם אותן הדמיות למשל). הסקירה הולכת להיות קלילה וקצרה.

המחברים מציעים דרך לשיפור הבנת קשרים בין פיסות דאטה ממודליות שונות על ידי מודלים מולטי-מודליים. למשל למודלים התומכים בשתי מודליות, שפה ותמונות, לפעמים מתקשים במשימות שדורשות הבנת קשרים סמנטיים בין דאטה ויזואלי לשפטית למשל במשימת למידה *in-context*-to-*in-context* MLLMs המתוארת קודם לכן. המאמר מצין כי MLLMs לפעמים מתקשים להתמודד עם משימות אלו ולמשל עונים על השאלה בלי להתחשב בהקשר כלל (זהה תמונות, שאלות ותשובות). המאמר מציע שיטת פין טיין עבור מודל מולטי-מודלי כדי להתמודד עם כשלים כאלה.

המאמר מציע לעשות פיניטיון למודל בשיטה מעולם RLHF (זהה Reinforcement Learning with Human Feedback) הנקראת(DPO) (= Direct Preference Optimization). שיטה זו נגזרת מפונקציית יעד פופולרית

בעבור פין טיון של מודל שפה (מקסום תגמול - קרבה למודל ההתחלתי) דרך מידול reward של Bradley-Terry. היתרון העיקרי של DPO מעל PPO הוא העובדה ש-DPO לא דרש אימון של מודל תגמול (reward) אלא צריך רק דاطהסת של זוגות שאלות ותשובות רצויות ותשובות לא רצויות. הרעיון העיקרי במאמר הוא להנדס דאטהסת כזה עבור יזקיעים שבנידון ולהשתמש ב-DPO לפין טיון של מודל מולטימודלי.

בגדול המאמר מציע לשחק עם השאלות והתשובות. הוא מציע כמה טריקים כדי לא לאלץ את המודל לhattches בכל הקונטקסט שניתן לו. אחד הטריקים הוא לחת תשובה רצiosa לא קשורה (מילה ללא משמעות). עוד טריך היא להחליף תשובה לא רצiosa בగבריש ועוד אחד היא למחוק את השאלה עצמה ולהשאיר את התשובות כמו שהם. יש עוד כמה טריקים מהסוג הזה ועל ידי שימוש המאמר משיג מודל יותר טוב עם שימוש ב-DPO לפין-טיון. כמו שהבטחת סקירה קצרה וקלילה.

<https://arxiv.org/abs/2411.11909>

4.04.25 המאמר היום של מייק -

Amortizing intractable inference in diffusion models for vision, language, and control

המאמר קצר כבד מתמטית ואני מנשה להסביר את הרעיון הכללי שלו ללא ציליה עמוקה למתמטיקה מתמטית.

מבוא: האתגר של אינפראנס פוסטורי (עומד בתנאים מסוימים) במודלי דיפוזיה

מודלי דיפוזיה כוללו מהפכה במבנה מלאכותית גנרטיבית, ואפשרו יצירת תמונות מרשיימות, טקסטים מתוחכמים, טקסטים יפים וכדומה. מודלים אלו פועלם על ידי ניקוי הדרגתית של רעש לתוך נתונים מובנים, וכן לומדים התפלגות פרIORית על מרחבי נתונים מורכבים.

עם זאת, ישומים רבים דורשים אינפראנס פוסטורי, כלומר יצירת דוגמאות שעומדות בדרישות מסוימות. לדוגמה:

- ביצירת תמונות, ניתן שנרצה להפיק תמונות השייכות לקטגוריה מסוימת.
- במודלי שפה, נרצה למלא מילימ חסרות תוך שמירה על עקביות תחבירית וסמנטיבית.
- בלמידה עם חיזוקים, נרצה להפיק מסלולים המאזנים בין חוקרים(exploration) לניצול(exploitation) תחת אילוצים התנהגותיים.

בדרך כלל, אינפראנס פוסטורי במודלי דיפוזיה מותבצע באמצעות הנחיית מסווגים (classifier guidance), שינו משקלם במודלי מבוסס score, או פין טיון מודלי דיפוזיה באמצעות אילוצי KL או משהו דומה. אך גישות אלו סובלות ממספר חסרונות חמורים:

- **קרישה למוד בודד (בודדים) (mode collapse):** טכניקות הננווטות את הדיפוזיה באופן מלאכותי מעוותות את ההסתברות האמיתית של הפוסטורי.
- **חוסר יעילות חישובית:** טכניקות אלו דורשות דוגמה חוזרת שוב ושוב, מה שmobiel לעליות חישוב גבוהות.

- **חוסר כלליות:** השיטות הנוכחיות פועלות היבט במשימות ספציפיות אך אין מתאימות באופן כללי לכל תחום.

המאמר מציע גישה שונה, המבוססת על במידה עם חיזוקים (RL) ורשתות זרימה גנרטיביות (GFlowNets). השיטה שלהם, שנקראת (RTB, Relative Trajectory Balance), מנשחת את אינפראנס הפוסטוריורי כבעית קבלת החלטות מركובי (sequential decision model), ומאפשרת דוגימה מודיקת יתר מהתפלגיות פוסטוריוריות מבל' להסתמך על שקלול הסתברויות ידני.

אינפראנס פוסטוריורי כתהיליך קבלת החלטות מרכובי:

בבסיסו, אינפראנס פוסטוריורי במודל דיפוזיה משמעו דוגימה מהתפלגיות מותנית:

$$p(x|c) \propto p(x)f(x,c)$$

כאשר $(x|c)$ הוא **מודל הדיפוזיה המאומן מראש** (הפרIOR) ו- $(c,x|f)$ הוא פונקציית אילוץ חיצונית (למשל, מסווג תמונה, מודל שפה או פונקציית תגםול כלשהי). המטרה היא להפיק דוגמאות אקס שמתאים לאילוץ c תוך שמירה על העקביות של הפרIOR. השיטות המסורתניות מנסות לשערר את $(c|x|f)$ על ידי שינוי דוגימה ממודל דיפוזיה באמצעות "הזהתו לכיוון הרצוי", עם כמה שיטות:

- **שיטת Classifier Guidance**, המשנה את פונקציית score של הדיפוזיה באמצעות גרדיאנטים ממודל מסווג.
- **משקל הסתברות דוגימה (Likelihood Reweighting)**, אשר מכוען הסתברות דוגימה אחריו שהוא מנגדמת.
- **פיין טוין שיר (Direct Optimization)**, שבו מודל הדיפוזיה מוחדר מחדש תחת אילוצי KL.

הבעיה עם כל שיטות אלו מוגבלות בבדיקה שלהן, בגין הדוגמאות שהן מפיקות ובעלות חישובית גבוהה. אך המחברים מציעים להפוך את תהיליך הדוגימה לתהיליך קבלת החלטות מרכובי ועושה זאת על ידי שימוש ברעיון שהוצע במאמר של GFlowNets.

מהן רשתות זרימה גנרטיביות (GFlowNets)?

רשתות GFlowNets הם מסגרת למידת מכונה המאפשרת דוגימה מהתפלגיות מורכבות על ידי מסגור תהיליך הייצור כרצף של החלטות. במקומם להתמקד ביצירת דוגמאות בודדות, כמו מודלים גנרטיביים מסווגים (VAEs), GFlowNets לומדים לגנרט דוגמאות בייחס לתגםול מסוים.

איך כל הסיפור זהה עובד? התחילה מוצג כרשת של מעברים בין מצבים (state transitions). כל שלב הוא פעולה במבנה הדוגמא. המודל לומד התפלגות הסתברותית על פני מסלולים שונים, כך שכל דגימה מופקת בפרופורציה לתగמול שלה. מדוע זה רלוונטי לאינפרנס פואטוריידי-דיפוזיה? במקרה לשנות משקל הסתברותית ידנית, ניתן ללמוד מדיניות שמייצגת ישירות את ההתפלגות הפואטוריידית. במקרה יקרה וחזרת, ניתן לאמוד מראש אילו מסלולי דגימה הם היעילים ביותר.

RTB: חיבור בין מודלי דיפוזיה ל-GFlowNets

השיטה המוצעת (RTB) Relative Trajectory Balance מיישמת את עקרונות GFlowNets במודלי דיפוזיה, כך שאינפרנס פואטוריידי הופך לתהיליך מייד עם חיזוקים מבוססי מדיניות (policy-based learning).

השלבים המרכזיים:

1. אימון מדיניות דגימה אופטימלית במקומות הסתמכות על שיטות הנחיה חיצונית.
2. איזון הסתברויות קדימה ואחוריה כך שהדגמה המתקבלת אינה מוטה.
3. תהיליך מייד לא תלוי במשימה: RTB לא תלוי במשימה מסוימת ונitin לישום גם בראיה ממוחשבת, גם בעיבוד שפה טבעי וגם בRL
- 4.

השילוב בין מודלי דיפוזיה, RL ורשתות זרימה גנרטטיביות פותח כיוון מחקר חדש וمسקרן. אם מודלי דיפוזיה היי הפריצת הדריך של השנים האחרונות, למידה אוטונומית של אינפרנס פואטוריידי יכולה להיות ההתקדמות הגדולה הבאה.

<https://arxiv.org/abs/2405.20971>

5.04.25 - המאמר היום של מיק - GIVT: Generative Infinite-Vocabulary Transformers

היום חוזרים כמו שניהם אחריה בו מילים VAE, VQ-VAE, VQ-GAN הם מושכים אותה תשומת לב כמו שמקבלים היום מודלי דיפוזיה גנרטיביים (אםنم פחות מאג'נטים אבל בכל זאת). המאמר שנסקור היום מציע כולל מעنى ל-VQ-VAE שימוש את עניינו כי כאמור מאמרם בנושא זה הוכיח "ציפור נדירה" בנוף שלנו (של AI).

קודם כל את הקדמה קצר לגבי VQ-VAE. נתחיל את ההסבר מ-E-VAE שהוא ראשית תיבות של AutoEncoder שהומצא אי שם ב-2014 על ידי Kingma האגיד. בגודל VAE מרכיב משתי רשתות, אנקודר ודקודר כאשר הרעיון מפיק את הייצוג הלטנטי (או אմבדיניג) של פיסת DATA כאשר הדקודר הופך את הייצוג הלטנטי לתמונה. הפלט של האנקודר הוא הפרמטרים של ההתפלגות הגaussית (וקטור תוחלות ומטריצת קוריאנס אלכסונית) ממנו דוגמים את הווקטור הלטנטי המזון לדקודר לשחזור תמונה הקלט לאנקודר.

פונקציית הלוס של VAE נבנית על בסיס ELBO (שהה Lower Bound) ומכילה 2 איברים. הראשון הוא השחזור של VAE-LL הוא הנורמה של הפרש התמונה המשוחזרת ביחס לתמונה המקורית (בגרסאות מתקדמות יותר התווסף להו loss perceptual ולוס בסגןון GAN) והאיבר השני הוא KL divergence בין ההתפלגות של הייצוג הלטנטי המופק מהDATA (המיוצג על ידי וקטור תוחלות ומטריצת קוריאנס אלכסונית) לבין התפלגות נורמלית סטנדרטית. באינפרנס אנו דוגמים וקטור לטנטי מההתפלגות נורמלית סטנדרטית ומצינים אותו לדקודר.

שכלול מעניין שהפרק להיות מאוד פופולרי של VAE הוא VQ-VAE. במקום להגדיר מרחב לטנסי בתור התפלגות גאוסית - אלא מגדיר אותו בצורה דיסקרטית. כל פאי' בתמונה מתוארכ(במרחב הטנסי) על ידי וקטור מה-codebook בגודל סופי שמאומן יחד עם האנקודר והדקודר. ככלומר יש מספר סופי של הייצוגים הטנסיטיים עבור כל פאי' (נחזיר על זה עוד מעט). כאשר אימון האנקודר, הדקודר והוקטורים מה-codebook מסתויים אנו מאומנים מודל נוסף לחיזוי ייצוג לטנסי של פאי'ים, על כל הייצוגים הטנסיטיים של הפאי'ים של הדאטסהט. מודל זה (נגדי טרנספורמר) מאומן לחזות באופן אוטורגרטיבי את הוקטור מה-codebook (כλומר מספרו) של הפאי' הבא בהינתן הפאי'ים הקודמים שכבר גונרטו. לאחר מכן הוקטורים הטנסיטיים של הפאי'ים מזונים לדקודר לגנרטות דאטה (תמונה).

כאמור יש מספר מוגבל וסופי של הייצוגים הטנסיטיים עבור כל פאי' ושהה די מגביל את העושר הסמנטי של התמונות ש-VQ-VAE ושיטות דומות מסוגלות לגנרט. וזה בדיק המקום שהמאמר שנסקור עכשווי חדש - הוא מציע שיטה לעבור לייצוג רציף (לא codebook) של הוקטורים הטנסיטיים. אבל איך אפשר לעשות זאת? נזכיר ש-VQ-VAE אנו כל פעם חוזים התפלגות קטגוריאלית מעל ה-codebook כלומר השכבה האחורונה במודל אוטורגרטיבי היא סופטמקס בגודל של ה-codebook.

האם ניתן לאמן מודל שיוצר ייצוגים לטנטאים רציפים בצורה אוטורגרטיבית? התשובה היא כן - בשלב הראשון המאמר מאמן את ה-VAE הסטנדרטי שהסבירתי לעלי בתחילת הסקירה. הדבר הזה נעשה ברמה של פאי' כלומר הייצוג הטנסי של תמונה מורכב מהייצוגים של הפאי'ים שלה. בשלב השני המחברים מאומנים מודל טרנספורמר סיבתי שחווצה את הייצוג של הטוקן הבא באמצעות חיזוי פרמטרים של התפלגות gaussian mixture שמננו נדגם הוקטור הטנסי עצמו. כלומר כל פעם הטרנספורמר הסיבתי (ליקח בחשבון רק את הוקטורים שנוצרו כבר) חוזה את וקטורי התוחלות, פרמטרים של מטריצות קווריאנס אלכסונית של כל משתנה ב- ax^2 ומשקל' הערבוב). אחרי שהוקטורים הטנסיטיים נחזו ונדגמו הם מזונים לדקודר לגנרטות תמונה.

נזכיר כי GIVT להבדיל מ-VQ-VAE ניתן לאמן במלואו יחד עם האנקודר והדקודר שלטענת המחברים יכול להיות בעיתוי. המחברים מציעים (במוקם הטרנספורמר הסיבתי) לאמן מודל (הנקרא adapter) של Normalized Flow (adapter) של ייצוג לטנסי כלו של הדאטה לאחר שהאנקודר והדקודר כבר אומנו וככה להפריד את שני השלבים. לגנרטות של ייצוג לטנסי כלו של הדאטה לאחר שהאנקודר והדקודר כבר אומנו וככה להפריד את שני השלבים.

בסוף המאמר מציע לאמן טרנספורמר לא סיבתי לחיזוי ייצוגים לטנטאים של פאי'ים (מאומן דומה ל-edited MaskGit languaged modeling

<https://arxiv.org/pdf/2312.02116.pdf>

7.04.25 המאמר היומי של מייק -

JETFORMER: AN AUTOREGRESSIVE GENERATIVE MODEL OF RAW IMAGES AND TEXT

הסקירה של היום היא מאמר המשך(למרות שאין CAN EVA) של סקירתו האחרונות (מ 05.04.25). המאמר שנסkor היום מציע שיטה לאימון מודל מולטימודלי כאשר מודל אוטורגרטיבי אחד מאומן לשתי המודליות (תמונה וטקסט יחד). ברוב המודלים המולטי-מודליים יש אנקודרים שונים לטקסט ותמונה ולדעת מחברי המאמר זה עלול להיות בעיה (אני סוג של מבין את זה). אך המאמר מציע לאמן טרנספורמר אוטורגרטיבי לשתי המודליות יחד.

از איך הדבר הזה עובד בעצם? המאמר מציע להשתמש במודל מאומן של זרימה מנורמלת (Normalized Flows or NF) לבניית ייצוג התמונה. מודל NF מאמן מיפוי הפיך וכן lossless מרחב הדאטה (תמונה) למרחב בעל התפלגות פשוטה (נגדי גאוסית סטנדרטית). בדרך כלל מיפוי זה נבנה על ידי הרכבה (composition) של כמה מיפויים פשוטים (נגיד על תות-קבוצה קטנה של מימדים) וכל המיפויים הללו מאומנים יחד כאשר המטריה היא למינם את הנראות (likelihood) של הדאטה תחת המיפוי זהה. למעשה המחברים מאומנים NF עבור כל פאי' בתמונה (ייצוג פאי' נקרא טוקן ויזואלי).

از המחברים מאמנים יחד מודל NF לייצוג תמונה יחד עם טרנספורמר אוטורגרטיבי לגירוט תמונה וטקסט. ככלומר בהינתן תיאור התמונה והתמונה עצמה (הסדר בהזנה של פיסות DATAה חשוב!) הטרנספורמר אוטורגרטיבי מאמין לפולוט את "יצוגי הטוקנים היזואליים אחרי NF (שמאומנים יחד עם הטרנספורמר). כאשר תמונה מזנת לפני התיאור שלה הטרנספורמר מאמין לשחרר את "יצוג הטוקנים הטקסטואליים. כמו בסקירה הקודמת (GIVT) המודל חוזה פרמטרים של ה-gaussian mixture עבור כל טוקן והיצוג נדגם ממש.

המאמר גם מציע להעלות את הרובוטיות של "יצוגים המופקים על ידי המודל האוטורגרטיבי המאמין עם הרעתה DATAה (רק DATAה ויזואלי מושפע לפיה הבנתי) מדורגת (סוג של מידת curium) . בהתחלה מושגים לדאטה רעש חזק יותר כך שהמודל יוכל ללמידה את הפרטים "הגסים" של הדאטה ומורדים אותו במהלך האימון כך שהמודל יוכל גם את הפרטים העדינים יותר של הדאטה.

<https://arxiv.org/abs/2411.19722>

09.04.25 - המאמר היומי של מיק - O1-CODER: AN O1 REPLICATION FOR CODING

סוף סוף הגיעתי לסקור את המאמר הזה שעשה לא מעט רעם בזמןנו. המטרה המוצהרת של מחברי המאמר היא לחקות את O1 של OpenAI על משימות קידוד. המאמר משתמש בטכניקות RLHF בשילוב עם שיטת self-play שבה המאמר לומד על הדאטה שהוא עצמו מגנרט. המאמר מתייחס מدادהש של שאלות קוד והשתובות על שאלות אלו (כלומר קוד (:)).

הרעיון העיקרי של המאמר מכיל 6 שלבים עיקריים. בשלב הראשון המחברים בוניםeli(המאמר לא מרחיב על זה יותר מידי) לגינרוט טסיטים מקיפים עבור שאלת קוד והקוד הנכון עבורה. בהמשךeli(TTG) ישמש לשערוך של reward עבור קוד שנבנה על ידי O1-CODER.

בשלב השני באמצעות MCTS זהה ראש תיבות של Markov Chain Search בונים את שרשרות הנמקה (reasoning) עבור הדוגמאות מהدادהש. MCTS הוא אלגוריתם לתוכנן קבלת החלטות שמבצע דגימה במרחב המצבים (טוקנים במקורה שלנו) כדי לשערר את reward הפעולות האפשרות. האלגוריתם בונה עץ חיפוש באופן הדרגי – בכל צעד הוא בוחר לפתח את הענף(סדרת טוקנים) שנראה הכי מבטיח, תוך איזון בין חקירה של אפשרויות חדשות לבין ניצול של מה שכבר נמצא כמושלח. כל מסלול בעץ (שרשרת הנמקה הכוללת פתרון) מקבלת תגמול 0 או 1 עם TTG(עובר או לא עובר את כל הטעיטים).

בשלב השלישי המודל עובר SFT על שרשרות הנמקה שהובילו לפתרון הנכון (עמ ציון 1). בשלב הרביעי מתחילה את אימון self-play בצורה איטרטיבית כאשר דאטהש האימון מועשר בכל איטרציה עם הדוגמאות הנוצרות על ידי המודל עצמו. בהתאם מבצעים אימון SFT של המודל על הדאטהש עם התשובות הנכונות בריבד(פרט לאיטרציה 0) או מבצעים אימון RLHF עם DPO (זהה DPO) על הזוגות של דוגמאות חיוביות ושליליות.

לאחר מכן אנו מגנרטים שרשרות הנמקה עם המודל (פרט לתשובה הסופית) ומשתמשים במודל תגמול PRM(Process Reward Model) למתן תגמול לשרשאות הנמקה אלו. אז בונים את התשובה על השאלה משרשתה הנמקה ויוצרים טסיטים לשאלת זו (ידעשה לנו התשובה הנכונה לכל שאלה - נראהה השאלה חן חלק מדאטהש גדול של שאלות פתורות). אחרי זה מחשבים את reward על ידי הרצת טסיטים על התשובות שגונרטו על ידי המודל (1 - הטעיטים עברו, 0 - לא עברו) ומחלבים אותו עם התגמולים שהתקבלו במהלך הריזונייג (נקרא aggregation function). מאומנים את המודל במטרה למקסם את התגמול הזה (עם שיטת RL כלשהי) - נראהה שיש כאן איזושהי רגולרייזציה אבל המאמר לא מרחיב על זה.

בסוף יוצרים דוגמאות עם המודל אחריו העדכו האחרון ומוסיפים אוטם לדאטהסט ומתחילה מוחדר את השלב הרביעי (self-play).

מאמר מאד מעניין...

<https://arxiv.org/abs/2412.00154>

11.04.25 - המאמר היומי של מיק

Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics

כבר סקרתי בעבר כמה מאמרם על מודלי שפה לחישוב נוסחאות ארכיטקטורת המילוט פועלות חשבוניות סטנדרטיות כמו פלאס, כפול וכדומה. לדעתי מודלי שפה פחות מיועדים למשימות מהסוג זהה (יש לנו מחשבונים, בפייטון וכאלו) אבל בכל זאת יש מחקרים מעוניינים בנושא זהה. ויש סיבה נוספת לבחירת המאמר הזה - הוא נכתב על ידי חוקרים ישראלים ותמיד נהנה לסקור תוצאות מקומית.

از כאמור המאמר חוקר מה קורה בתוך מודל הטרנספורמר כאשר מודל שפה מקבל משימה ארכיטקטונית. למעשה המחברם מנסים לאתגר מה שנקרא נתיב החישובי (circuit) בתוך הטרנספורמר כלומר רכיביו המבצעים בפועל את "הчисובים הנדרשים" עברו משימה זו. אולם בטח זוכרים שבлок טרנספורמר מורכב ממשתי שכבות עיקריות (יש גם שכבות נרמול) שהם מגננון attention מרובה ראשים או MHA ושכבה MLP המורכבת ממשתי שכבות עינאריות ואקטיבציה לא לינארית ביניהם. אז הנתיב החישובי מורכב מנויירונים מסוימים בתוך ה-MHA או בתוך ה-MLP.

כדי לאתגר את הנתיב החישובי, המחברם מבצעים החלפת אקטיבציות (activation patching) של נוירונים בתוך הטרנספורמר המאפשרים לשערר את החשיבות של שכבות MLP וכל ראשי attention בכל מקום בסדרת קלט (פרומפט ארכיטקטוני). איך עושים זאת? לוקחים פרומפט ארכיטקטוני מסוים (לדוגמא, "226 - 68 ="), ופרומפט אקריאי שMOVIL לתוצאה שונה (למשל, "21 + 17 ="). לאחר חישוב של אקטיבציות המודל עבר הפוקודה האקריאי, מזינים את פרומפט המקורי למודל.

בשלב זה מתרבים בחישוב (patching) — כלומר, מחליפים את אקטיבציות של שכבה MLP בודדת או ראש attention באקטיבציות שחושבה מראש הפרומפט האקריאי. בהמשך בודקים כיצד ההתערבות משפיעה על ההסתברויות של שני הטוקנים של התשובות(מעבר הפרומפט המקורי ומעבר האקריאי) - יש נוסחה שמשערת שהשניים יבטוקני התשובות. לאחר מציאת הנתיב החישובי מעבר הדוגמאות השונות המאמר משערר את "נק'ונם" על ידי החלפה של כל האקטיבציות באקטיבציות מוגזמות על פני דאטהסט גדול של פרומפטים ארכיטקטוניים כאשר רק האקטיבציות של הנתיב החישובי נותרו על כןם. המחברם הרואו שהחלפה זו כמעט ולא משפיע על הלוגיטים של התשובה הנכונה.

אחרי מציאת נתיבים חישוביים אלו המחברם ניסו להבין איזה שימוש ארכיטקטוני יש להם. כתוצאה לכך התבירה תמונה די מעניינת. המחברם הרואו כי הפעולות של נתיבים אלו הם למעשה ייריסטיות שונות המאפשרות לפטור את התרגילים. למשל היו נוירונים שמטרתם היא להגיד האם התוצאה נמצאת בתחום [150, 180] או שהיא תוצאה מחלוקת ב-5. שילוב של שערכיהם אלו מאפשר למודל לפטור תרגילים ארכיטקטוניים פשוטים יחסית הלא מערבים מספריים מדי. זה די מסביר למה LLMs מתקשים עם פעולות על מספרים גבוהים.

בנוסף יש כמה מציאות מעניינות. רוב החלקים הבולטים של הנתיבים החישוביים נמצאים בשכבות MLP ולא בראש attention. הדבר המעניין השני הוא העובדה שהמודל "די מתכוון" לתשובה הנכונה כבר בשכבות הביניים (ניתן להפיך אותה שם על ידי שכבה לינארית).

<https://arxiv.org/abs/2410.21272>

13.04.25 - המאמר היומי של מילק - ONE STEP DIFFUSION VIA SHORTCUT MODELS

המאמר מציע גישה מעניינת לאימון מודלי דיפוזיה גנרטיביים המבוססת על שיטת flow matching (או FM) בקצרה) שנויותה הגישה המובילה לאימון מודלי דיפוזיה. למעשה המאמר מאמין מודל לשערך מסלול (בדרכו כלל קו ישר שזה המסלול הכי פשוט אבל יש מאמרם שבוחרים צורות אחרות של המסלול) בין התפלגות הגaussית (התפלגות הפשוטה) לבין התפלגות הדאטה (תמונות, וידאו או אודיו). המאמר טוען שבאמצעות השיטה המוצעת ניתן לגנרט דאטה באיטרציה אחת בלבד.

המודל מאמין לגנרט מהירות (גרדיאנט) של במסלול זה בכל נקודה \dot{z} המסמנת כאן את עוצמת הרעש במסלול בין התפלגות הפשוטה (רעש טהור $t = 0$) לתפלגות של דאטה ($t = 1$). אחרי שהמודל משערך מהירות זו ניתן לגנרט פיסת דאטה על ידי פתרון נורמי של משווה דיפרנציאלית דרך הצבה של מהירותם לשם. עבור מסלול לינארי מהירות הזו היא קבועה (נגזרת של קו ישר). לעיתים זה לא עובד כל כך טוב ומסלולים שנוצרים יוצאים לא לינאריים ודאי מורכבים והדאטה שגנרט כפואה לכך לא מאד איקוטי.

המאמר מציע לבנות את המסלולים הללו לא בצורה לינארית אלא בצורה לינארית למקוטען (סוג של ספליין לינארי) במקום להכricht את המודל ליצור מסלולים ממש לינאריים. התזוזה של נקודת דאטה בתת-מקטע תלויה רק בנקודה \dot{z}_x ובגרנולריות הספליין d (ארחיב על זה אחר כך). תת-מסלולים אלו נקראים במאמר shortcuts והמודל מאמין לשערך אותם עם מה שנקרה loss עליהם שכופה על המודל להיות "עקבי" בשני shortcuts עוקבים. וכך זה נוצר על ידי שימוש פשוט של הנוסחאות עבור shortcuts העוקבים.

לאחר מכון המחברים משלבים את loss "העקביות" זה עם הלוס הרגיל עבור FM (עם המסלול הישר). ניתן לבנות את המסלול-m-shortcuts בגרנולריות שונות של תת-קטעים לינאריים (כלומר עם מספר תה מקטעים שונים), אז האימון מנצל את זה ומאמין את המודל על גרנולריות שונות. ככלומר בהינתן האיטרציה (\dot{z} עצמת הרעשה), דאטה מושרע וגרנולריות הספליין d המחברים מאמנים מודל המשערך את גודל ההזזה של נקודת דאטה (shift) של תת-מקטע הבא (כאמור \dot{z}_x d כאלו בסך הכל). לאחר מכן פותרים משווה דיפרנציאלית כדי לקבל את ערך הדאטה בסוף התת-קטע. לאחר מכן שוב משערכים (באמצעות המודל המאמין) את הzzות נקודת דאטה שהתקבלה. ואז מפעילים את loss consistency על מנת ששתי הzzות הדאטה.

מאמר די מעניין וכותב בצורה נפלאה - מומלץ!

<https://arxiv.org/abs/2410.12557>

14.04.25 - המאמר היומי של מילק: Draft Model Knows When to Stop: A Self-Verification Length Policy for Speculative Decoding

המאמר הזה משליך עיני' כבר בהסתכלות הראשונה בגלל צמד המילים "Speculative Decoding" או SD בקצרה שמאוד קרוב לילבי - אפילו הכונתי על זה מצגת די מקיפה שאני מציג אותה בפורומים שונים. SD מאפשר להגדיל את קצב גנרט טקסט על ידי מודל שפה באמצעות שילוב מודל היעד עם מודל קטן מהיר יותר וכמוון יותר חלש מודל היעד. המודל הקטן מייצר כמה טוקנים לצורה אוטוגראטיבית ומודל היעד חוזה מנצל טוקנים אלו כדי לחזות בו זמןית את הטוקנים הבאים שלו. זה מאפשר להגדיל את הקצב הדגימה של המודל הגדל לצורה ניכרת.

השיטה מנצלת את העובדה שצואר הבקבוק של תהליך הגנרטוּט העברת דאטָה בין הזרונות של סקְגַּפְּ (בפרט HBM הגדל ואיטי -MRAM) הכן אך מהיר בחלק החישובי של ה-סקְגְּפְּ). אֶז SD מבצע חיזוי מהיר עם המודל הקטן ואז החיזוי הבו זמני על ידי המודל הגדל עם הטוקנים שנחצוו על ידי המודל הקטן. אבל יש שם קאץ' כמובן: כדי לקבל את אותה התפלגות הטוקנים עם המודל הגדל דרך ניצול הטוקנים של המודל הקטן יש צורך סוג של rejection sampling או RS.

אזכיר ש-RS אפשר לדגום מהתפלגות קלה לדגימה f כדי ליצור מוגם הדגם מהתפלגות אחרת וקשה לדגם ממנה בצורה ישירה. אך אנו דוגמים נקודה x מ-f אך מקבלים את הדגימה בהסתברות השווה ליחס בין (x)f ל-(x)g (אם יחס זה גדול מ-1 הנקודה מתקבלת אוטומטית). ניתן להוכיח שנקודות שהנדגנות באופן זה מפוגגת עם התפלגות הרציה g.

از במקהה שלנו (SD) אנו עושים משהו דומה עבור הטוקנים הנdagנים עם המודל הקטן. במהלך השלב השני (דיגימה בו זמנית מהמודל הגדל) עבור כל טוקן הנגdem מהמודל הקטן אנו מחשבים את היחס בין ההסתברויות של המודלים ואנו "מקבלים" את הטוקנים של המודל הקטן בהסתברות השווה ליחס סיבא. אחריו שהטוקן הראשון של המודל הקטן "סורב" (rejected) המודל הגדל מגנרט טוקן הבא עם המודל הגדל ואז המודל הקטן שב מופעל לגנרט את הטוקנים הבאים. ד"א גם הטוקנים שמתקבלים מגונרטים עם עם התפלגות המחשבת משתי התפלגות של הטוקן (של המודל הקטן ושל הגדל).

כמו שכבר הצלחtem להבין "שליטה" ב acceptance rate של טוקנים של המודל הקטן היא מאוד חשובה - באידיאל אנו רוצים לדגום מהמודל הקטן רק את הטוקנים שיתקבלו. המאמר מציע שיטה לשפר את acceptance rate. המאמר מראה שהמוצע של total variation distance (זה די קל) שווה להפרש בין 1 למה שנקרה total variation distance או TBD בקשר בין התפלגותם של שני המודלים (הモוניות בהקשר). ולמזלנו עמד לרשותנו אי שוויון לא ידוע במיוחד שמאפשר לחסום TBD מלמטה עם הפרש בין קروس-אנטרופי בין התפלגותם של שני המודלים (עבור טוקן נתון בהינתן הקשרו) لأنטרופיה של טוקן של המודל הקטן.

אבל כמובן שאנו לא יכולים לחשב את הקروس אנטרופי בין התפלגות אלו בשלב דגימה מהמודל הגדל עבור כל הטוקנים כי אנו דוגמים כל הטוקנים ממנה בו זמנית ולא יודעים מראש התפלגות מותנית של כל טוקן של המודל הגדל. אך המאמר "משערך" את הקروس אנטרופי זהה על זמן מוגם די גודל דרך קבע (קצת גדול מ-1) מוכפל באנטרופיה של הטוקן של המודל הקטן. אחרי שיש לנו את הקروس-אנטרופי אנו יכולים לשערך את acceptance rate עבור כל טוקן של המודל הקטן לפני הדגימה מהמודל הגדל. זה מאפשר לנו לקבוע את מספר הטוקנים מהמודל הקטן שעבורם תבצע דגימה בו זמנית מהמודל הגדל - פשוט בוחרים טוקנים עד שה-acceptance rate המשוערך גבוהה מאיזה סף.

רעיון נחמד אבל בחירת הקבוע בשלב האחרון לדעתו לא אופטימלית ואני מקווה שבקروب ייצאו מחקרים המשפרים את היבט זהה של השיטה המוצעת.

<https://arxiv.org/abs/2411.18462>

המאמר היומי של מייק: 15.04.25

Classifier-Free Guidance inside the Attraction Basin May Cause Memorization

חווזרים לסקור מאמרי דיפוזיה - הפעם מאמר קליל (יחסית למאמר ממוצע בנושא מודלי דיפוזיה). המאמר מציע שיטה למניעת זיכרון או memorization באנגלית על ידי מודלי דיפוזיה. ניתן לראות בזיכרון סוג של mode collapse (הזכורה לנו מתקופת הגאנטים) כאשר המודל מגנרט תמונות דומות מאוד (וגם דומות לתמונות מסוימות האימון) לקלטים שונים (בד"כ נדגמים מהתפלגות פשוטה לדגימה כמו גאות סטנדרטית).

תופעה זו מתרחשת לרוב במודלי דיפוזיה מותנים כלומר כלו שיוודעים לציר לנו תמונה מתיאור טקסטואלי (כלומר פרומפרט). במקורה זה תופעת זיכרון מתרחשת כאשר לא משנה מאיזה דגימה התחלתית של רعش גausי אנו מתחילה, המודל מגנרט לנו תמונות כמעט זהות. המאמר הוכיח את הסיבות להתרחשות תופעה זו ומגע למסקנה כי זיכרון קורה עקב שימוש בטכנית הנקראה CFG Classifier Free Guidance בקצרה.

המטרה של CFG היא "להציג" את גנרטת התמונה לכיוון הסמנטי של הפרומפט כלומר לגרום לתמונה להיות מותאמת לפרומפט. אתם בטח יודעים שמודלי דיפוזיה מגנרטים תמונה עלי ידי הסרת רעש הדרגתית מהרעד הטהור (בד"כ גausי). זה מtbody כאמור באיטרציה באמצעות מודל דיפוזיה שמאמן לשערך את הרעש שצרי להסיג בהינתן תמונה מושעת באיטרציה (מצין כי הוא גם קלט למודל דיפוזיה).

כאמור CFG "מוציא" את התמונה המוגנתת לכיוון הפרומפט על ידי הוספת הרעש המשוערך על ידי מודל את ההפרש ממושקל (עם משקל קטן) ביןו (הרעש המשוערך) לבין הרעש המשוערך של מודל דיפוזיה לא מותנה (שמאומן לנגרט תמונה ללא פרומפט). גם המודל (בזמן האינפנס) מוציא את התמונה המוגנתת רחוק יותר מהתמונה המושעת (לא פרומפט) ומרקבת אותו (סמנטי) לפרומפט שלה.

אבל כמו שהמחברים מצאו CFG מקרב את התמונה לפרומפט חזק למד'. יתרה מזה הם מצאו שאם מתחילהם לעשותות CFG מאיטרציה מוחרת יחסית (כאשר התמונה כבר נוקתה קצת מהרעד) אז תופעת הזיכרון כמעט ולא מתרחשת. הסיבה לכך טמונה בכך שהנורמה של וקטור הרעש המותנה גבוהה ממשמעותית מזה שאינה מותנית באיטרציות מוקדמות אך הן משתמשות לkract באמצעות תהליך של הסרת הרעש (backward process).

از בשיל להתמודד עם תופעת הזיכרון המאמר מציע לעשות הסרת רעש ללא CFG באיטרציות מוקדמות ולהתחליל עם G CFG באיטרציות יחסית מאחרות. אבל איך ניתן להזיהות מתי צריך להתחליל להפעיל CFG? פשוט מאוד - כאשר המהלך בין נורמות הרעשים המשוערכים מתחילה קטן. זה בגודל הרעיון העיקרי של המאמר.

במאמר יש לא מעט הגדרות מתמטיות להגדרת הזיכרון (ואני מאד אוהב את זה) אבל מי שלא רוצה להתעמק יכול להסתפק בסקירה זו להבנה כללית.

<https://arxiv.org/abs/2411.16738>

17.04.25 המאמר היומי של מיק:

Memorization to Generalization: The Emergence of Diffusion Models from Associative Memory

אוקי, ממשיכים עם מאמר תאורי עמוק בנושא מודלי דיפוזיה גנרטיביים. בסקירה זו ניסיתי למקסם את אחוז המושגים של LM שתרגמתי לשפת הקודש. תגידו לי איך יצא.

המאמר מציג דיוון תאורי עמוק, הרואה במודלי דיפוזיה מערכות זיכרון אסוציאטיבי סטוכסטיות בעלות מספר פרמטרים עודף (overparameterized). הרעיון המרכזי הוא שהתנהוגותם של מודלי דיפוזיה כתלות בגודל DATAהסט האימון משקפת את הדינמיקה של רשות הופפילד מודרניות(התפתחות של אלו שהוצאו על ידי חתן פרס נובל טרי) כאשר הן חורגות מקיבולת הזיכרון הקרייטית שלהן.

רשת הופפילד הקלאסית היא מודל של זיכרון אסוציאטיבי שבו כל תבנית נשמרת כנקודות מינימום באנרגיה, אך הקיבולות שלה מוגבלת – היא יכולה לשמור רק מספר תבניות פרופרציונלי למספר הנוירונים בה. רשת הופפילד מודרנית מרחיבה את הרעיון באמצעות מנגן softmax או ארגזיה לא לינארית, ומסוגלת לשמור כמות אקספוננציאלית של תבניות ולשזרן בדיק גובה, תוך קשר הדוק למכניקת מנגן ה-attention בטרנספורמרים (סקרתי מאמר על זה).

ברשת הופFIELD קלאסית, משטח האנרגטי בניי כך שכל תבנית מאומנת מהוות נקודת משיכת יציבה. כל עוד מסטר התבניות נמור מהקבולות התיאורטיות, כל תבנית (וקטור או דגימה שצורך לצרכו) מוקצת לבור אנרגטי מבודד. כאשר מסטר התבניות חורג מהקבולות, מופיעות נקודות משיכת לא צפויות — מה שנקרא "מצבים מזוייפים" (spurious). מצבים אלו אינם תואמים לדגימות האימון, אך לעיתים קרובות מהווים קומבינציות לנאריות שלhn.

המאמר מזהה תופעה דומה במודלי דיפוזיה. במהלך האימון, מודל הדיפוזיה לומד פונקציית ציון (score) עבור תהליך הופci (backward) בתהליך הפיכת רושם טהור לפיסת DATA. פונקציה זו מוקצת נגזרת של הסתברות לוגריטמית של פיסת DATA מושעת וכתוואה מכרך ניתן לראות בה את שיפוע פונקציית אנרגיה סמייה (הסתברות גבואה מתאימה לאנרגיה קטנה - סוג של אטרקטו). המחברים מראים שפונקציה זו זהה בצורה לו של רשת הופFIELD מודרנית עם ארגזיה מבוססת softmax על DATA (צורה בה רשת הופFIELD זכרת את הדטה). לעומת זאת, הדינמיקה של דיפוזיה שקופה למינימיזציה סטוכסטית של זיכרון אוצטראטיבי.

כאשר DATAהט האימון קטן, פונקציית score מוחשבת בדיק גבואה (המודול overparameterized ומושער ואותה בקלות) ורוב הדגימות שנוצרות הן העתקות של דגימות האימון - המודול מצוי בשלב של זיכרון (memorisation) חזק. ככל שגודל DATA גדול, המודול כבר לא יכול לייצר בורות אנרגיה מבודדים לכל דגימה, ונוצרים מצבים "מזוייפים". אלו הם דגימות שלא נראהות בסט האימון אך כן נמצאות קרוב אליהם ומהוות סוג של שילובים שלהם. בהמשך, כשיובלות זו (של שילובים) מנצלת גם היא, המודול מתחילה לייצר דגימות חדשות שלא שייכות לא לסט האימון ולא לקבוצת השילובים - זהו שלב הכללה מלאה.

המאמר מגדר שלוש קיבולות:

- **קיבולות הדיזכרון:** מספר הדגימות המרבי שמודול יכול לשחרר באופן עקבי מתוך האימון.
- **קיבולות החיים - spurious (שילוב):** גודל DATA שבו יש מקסימום שכיחות לדגימות שלא מופיעות באימון אך כן מופיעות בקבוצת הסינטזה.
- **קיבולות ההכללה** – גודל קבוצת האימון שמעליו המודול מפסיק לייצר דגימות שכפולות או קרובות לדטה הקיימ.

המעבר בין שלבים אלה מופיע כהתנהגות פaza: תחיליה ירידת חדה בזיכרון, עלייה חדה ב"זוייפים", ואז דומיננטיות של דגימות כלליות. המאמר מישם מדדי זיהוי מובוסי שכנות קרובות למדידת המרחק בין הדגימות שנוצרו לבין DATA המקורי, ומסוג לפיו אם מדובר בזיכרון, "שילוב" או הכללה.

מבחינה תיאורטית, העבודה מצבעה על כך שהכללה אינה נובעת רק מבנה הארכיטקטורה או מהרגולרייזציה, אלא מtower אינטראקציות מבניות במשטח האנרגטי. כאשר כמות DATA חורגת מהקבולות, בורות האנרגיה מפסיקים להיות דיסקרטיים ומתייחסים ליצור משטח רציף – שילובים הם תוצר ישיר של אינטראקציות אלו. ככל שהאינטראפלציה ביןיהם (השילובים) משתפרת, נוצרות "משטח אנרגטי מכיליל" שהוא תוצאה של דינמיקה ארגנטטיבית של נקודות המשיכה מהDATAהט.

המסקנה המרכזית היא שמודול דיפוזיה פועל בפועל כמערכת זיכרון אוצטראטיבי רוויה, והכללה נוצרת לא כתמונה חיונית אלא כתוצר של קבוצת קיבולות זיכרון – תופעה הניתנת לאפיון, כימות וחיזוי.

<https://openreview.net/forum?id=zVMMaVy2BY>

המאמר היום של מיק: 18.04.25

Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability

מאמר ד' מעוניין העוסק בשיפור יכולת הנמקה של מודלי שפה בשאלות שיש להם תשבות חד משמעיות (כגון שאלות מתמטיות ושאלות קוד שניתן לבחון את נכונות הפתרון באמצעות סט מקיף של טסטים). המאמר מגדיר מושג טוקן קרייטי (critical token) שהוא למעשה מהו סוג של סימן האם המודל הולך לתשובה נכונה או לא נכון לשאלה.

המחברים שמו לב כי שבתוך מסלולי הנמקה (reasoning) שגויים, ישנים טוקנים שהם כמעט בוודאות מובילים לתוצאות שגויות. טוקנים אלה משמשים את הרצף הלוגי, מעוותים קשרים או מכניםים שגיאות חישוביות, וכך משפיעים באופן משמעותי על התוצאה הסופית. בשונה מטוקנים אחרים שעשוים לה השפיע בצורה לא משמעותית על תהליכי האינפראנס, "הטוקנים הקרייטיים" האלו מהוות סוג של נקודות כשל. זהו הטוקנים הללו הוא חינוי, משומם שלעיתים קרובות הימנעות מהם או תיוקנם יכולה להוביל לתוצאה נכונה – גם בתוך מסלול הסקה שגוי.

המאמר מציעה שיטה ליזיהו של טוקנים קרייטיים. טוקן מזוהה כקרייטי עם כל מסלולי הנמקה המקוריים ממנו מסתיימים בתשובה שגואה ועבור כל הטוקנים שבאים 95% מהמסלולים המקוריים מהם מסתיימים בתשובה לא נכון. שימושו לב שיש טוקנים המופיעים בטקסט במקומות בהם אחריו הטוקן קרייטי שלא כל מסלולי הנמקה שלהם מכילים את הטוקן קרייטי, כך לא מן הנמנע שיש בין מסלולים המסתיים בתשובה נכונה. המachers ביצעו כמה בדיקות כדי לוודא שהטוקנים שזוועו בצוירה באמצעות טוקנים קרייטיים.

לאחר מכן המאמר מפתח שיטת RLHF לישור מודל שפה שבמרכזו מזעור של הנראות של הטוקנים הקרייטיים (כי הם מובילים לשגיאות). בשביל לכך המאמר מציע לאמן שני מודלים (עם פיניטיון) – אחד שמנגרט תשבות נכונות והשני מגנרט תשבות לא נכוןות (שמעתם נכון).

לאחר מכן המachers מנסחים דרך לשערוך הנראות של האם הטוקן הוא קרייטי בהינתן הפרומפט וטוקני התשובה לפני. הנוסחה היא הפרש משקל של נראיות הטוקנים (מותנים) מהמודל של התשובות הנכונות בין זה של המודל של התשובות השגויות. שערוך זה מקבל ערך נמוך עבור התשובה הנכונה וערך גבוה עבור התשובה לא נכון. בשלב האחרון המודל עובר פיניטיון עם DPO שזה קיצור של Direct Preference Optimization אשר על הזוגות של שאלות עם התשובות הנכונות והשגויות. כדי למנוע את הסיכוי להופעת טוקן קרייטי המאמר משנה את האיבר המכיל נראות של תשובה שגואה בנוסחה העיקרית של DPO על ידי הכפלה על הנראות השלילית של טוקן להיות קרייטי. שימושו לב שמכיוון שההתחשבות בנסיבות מתרחשת ברמה של טוקן ה-DPO במאמר עבור להיות token-level ולא sample-level כמו במאמר המקורי.

<https://arxiv.org/abs/2411.19943>

המאמר היום של מיק: 20.04.25

Training Large Language Models to Reason in a Continuous Latent Space

המאמר מציג רעיון חדשני ומתקבקש (לעניות דעת) לשיפור תהליכי הנמקה (reasoning) של מודל שפה. כמו שאתם בטוח יודעים אנו גורמים למודלי שפה לחשב על ידי הכנסה לפרומפט ביטויים כמו "think step by step" או טוקנים מיוחדים של חשיבה כמו <think> ועודמה. זה גורם למודל "לפלוט" את שרשרת הנמקה בצוירה של טוקנים, ככלומר של טקסט. היתרון בגישות אלו שאנו יכולים לנתח את שרשרת החשיבה של מודל ולשפר אותה כי אנו רואים אותה כתקסט.

אבל האם מודלי שפה חייבים "לחשב" בשפה שלנו? לא בהכרח. למשל הגרסת הראשונית של מודל DeepSeek המפורסם R0 בנתה שרירות הנמקה בכמה שפות (אמנם שפות טבעיות). זה קרה כנראה בגלל שהמודל אומן עם RLHF בלבד ולא קיבל תגמול על כתיבה קורנתית אלא בעיקר על נוכחות התשובה. כמובן המודל לא אומן על שרירות הנמקה מסודרת (שהן מן הסתם מכילים שפה אחת). זה גרם לכך שהמודל פיתח שפה משלה (שזה ערבות של כמה שפות כמו אנגלית, סינית, רוסית ועוד) בדרך לפתרון בכך של שאלות הדורשות חשיבה.

המאמר שנסקור היום עושה צעד נוסף בכךון זהה. הרוי מודלי שפה לא חייבים לחשב בשפות שאנו, בני אדם, מבינים, נכון? בשביל כך יש להם את מחרב הייצוג שלהם, כמובן המרכיב הלטנטי. הרוי מודל שפה לא חושב באמצעות מילים ובמשפטים כמוום אלא פועל במרחב וקטורי שככל וקטורי ייצוג של טוקן. אז המחברים אמרו את הדבר הבא: בוא נחליף שרירות הנמקה בשרירות הנמקה לטנטיות (וקטוריות) ללא תרגום לשפה האנושית. אז המודל מאמין להחליף שרירות הנמקה בשפה טבעית בסדרה של וקטורים.

זה בדיק מה שנעשה באימון המודל. המחברים מאמנים מודל לפלוט וקטורים במקום עבור כמה שלבי הנמקה הראשוניים. כמובן המודל מאמין (בו זמינות) להחליף שלבים 3-1 או 1-6 של שרירות הנמקה בוקטורים. כמובן המודל מתייחס מהמשטר הלטנטי (*latent mode*) שהמחשבת שלו הם הוקטורים וממשיך במשטר שפטי (*language mode*) שבו הפלט הוא שפה טבעית. כמובן שיש טוקן שפעריד בין משתרים אלו כמובן <eot>.

מאמר עם כיוון מאוד מעניין שאני צופה לו עתיד גדול.

<https://arxiv.org/pdf/2412.06769>

המאמר היומי של מיק: 22.04.25 Normalizing Flows are Capable Generative Models

בחرتתי את המאמר זהה לסקירה כי הוא מכיל מכלול של גישות ושיטות שנדרר לפגוש היום במאמרי deep learning. הסיבה השניה היא נוכחתה של שיטה גנרטיבית הנקראת *Normalized Flow* כמובןZRIMOT מנורמלות או NF בקצרה. זו שיטה שכמו GAN ו- VAE הפסידה בונקאות למודלי דיפוזיה בקרוב על גישה גנרטיבית מובילה. עם זאת זו גישה מאוד מעניינת בעלת אפיון מתמטי מדויקDOI ו- אינטואיטיבית. אז המחברים מנסים להחזיר את עטרה לישנה ומציעים גישה מובוסת NF משלבת עם כמה כלים מתמטיים מעולם מודלי דיפוזיה ועוד טרייק מתמטי נחמד הנקרא *Noschott Tweedie*.

از קודם כל מה זה NF? למעשה זו גישת אימון של מודלים גנרטיביים שמאמנת מודל של מיפוי 1-1 ערכי בין התפלגות פשוטה (כמו גאומטרית סטנדרטית) לבין התפלגות הדטה (נגיד דאטסהט של תמונה). מכיוון שהמיפוי הוא 1-1 ערכי אז הוא הפיך ועבור כל פיסת דטה נתונה נוכל לחשב את נראותה (likelihood) ביחס למודל NF בצורה קלה (פעילים את המיפוי ההופכי ומחשבים את הנראות של התפלגות לפי התפלגות פשוטה). כמובן שהמייד של המרכיב הלטנטי (המושהה על ידי התפלגות פשוטה) הוא בעל אותו המימד כמו מרחב הדטה.

רוב מודלי NF עם מיפויים של המרכיבה של מיפויים פשוטים ו-1-1 ערכיים הפעלים על תחת-קובוצה של מימי הדטה (שאר המימים נותרים ללא שינוי). עבור תמונה למשל כל התפלגות אוטומית זו פועלת על כמה פיקסלים של התמונה. בד"כ מיפוי זה בניית מטריצות מושלשות עליונות למדות (עם פונקציות לא לינאריות) לאחר מטריצות אלו הן הפיכות וניתן להפוך אותן לצורה מאוד קלה (זה מהוות יתרון עצום כאשר רוצים לחשב נראות של פיסת דטה). בנוסף כל מטריצה זו בcontra לצורה של חיבור residual לומר היא מהוות סכום של הערכים הישנים והמייפויים שלהם).

از המאמר בונה מודל NF לצורה כזו אך כאמור מציע כמה תופסות. התופסת הראשונה היא אימון המודל על דטה מושרש. כמובן מאמנים את המודל על הדטה שעבר הרעשה קלה (עם רעש גאוסי) שלטענת המחברים

עליה את הרובוטיות של המודל (די הגיוני בסך הכל). אבל כדי שהמודל לא ייצור לנו כתוצאה מכך פיסות דאטה מורעשות המחברים הציעו לנ��וט את הדאטה באמצעות nosehat tweedie שלמעשה משערכת את התוחלת של דאטה מורעש (הדאטה שלו) דרך דוגמה של דאטה מורעש וגרדיאנט של לוג ההתפלגות של הדאטה המורעש מושקל עם השונות. ככה בעצם מקבלים את הדוגמה הנקייה מדוגמתה אחרת אימון על דאטה מורעש.

והדבר האחרון שהמאמר עושה הוא שימוש Classifier Guidance שנמצא בשימוש כבוד במודלי דיפוזיה גנרטיביים. classifier guidance היא שיטה שמקוננת את תהליכי הדוגמה בעזרת מסווג חיצוני (classifier). במקום רק על סמרק הרעיש, המודל משלב את גרדיאנט הסתברות של מסווג לתווית הרצואה, ובכך מעלה את הסיכוי שהדוגמה הסופית תהיה שייכת למחלקה מסוימת. ניתן לעשות זאת ללא מסווג כאשר במקרה זה אנו מזיזים את הדוגמה בכיוון ההופיע המגנטר דוגימות לא מותנות (מודול דיפוזיה לא מותנה). אז המחברים מצאו דרך די אינטואיטיבית לדוחוף את הרעיון הזה לתוך אימון מודל NF (בגדול מזיזים דוגמה אחריה כל שלב ב-NF).

מאמר כי אף לא טריוויאלי בטח אם מנסים לצלול עמוק למתמטיקה אבל בתקווה העברתי את הרעיון הכללי...

<https://arxiv.org/abs/2412.06329>

23.04.25 המאמר היומי של מיק: The Broader Spectrum of In-Context Learning

למיידת ICL או ל ICL היא יכולה של מודלים לבצע משימות שלא אומנו עליהם **במפורש** כאשר הם מקבלים כמה דוגמאות לביצוע משימה זו בפורמט.

המאמר מציע שניי מהותי מבחינת התבוננות ב- ICL במקום להתייחס אליו כתופעה מצומצמת של למידת few-shot. המחברים מציגים אותו כמנגנון כללי ורחיב של הסתגלות הקשרית שנלמד באופן במהלך pretraining על דאטה סדרתית. לשיטתם, כל ירידה עקבית loss-by-loss שמקורה במידע קודם בסדרה מהוות מקרה של ICL - בין אם מדובר בזיהוי תבנית תחרירית, coreference resolution או topic continuation. מדובר ביכולת הסתגלות שמתפתחת מתוך האימון עצמו, ולא משזה שנדרש למד ב轟惜.

המסגרת המושגית שנבנית כאן נשענת על הבחנה בין שני מעגלי למידה: "outer loop" המתרכש במהלך האימון, שבו המודל לומד על דפו'ו דאטה, ו-"inner loop" שבו מתבצעת הסתגלות של המודל בפועל בזמן הריצאה, בתוך האקטיציות של המודל, על סמרק ההקשר המקומי בטקסט. הגדרה זו ממקמת את ICL כהתנהוגות הסתגלותית emergent, בדומה לגישות של memory-based meta-learning או meta-RL, אך מותאמת למודול שפה בלתי מפוקח.

סוגים שונים של ICL שלא נחשבים "קלאסיים"

המחברים מציגים קטגוריות שונות של ICL שלא נכנסות למסגרת הרגילה של few-shot learning. כל אחת מראה איך המודל לומד מהקשר בצורה שונה:

Instructional ICL: כאן המודל לומד ממשימה רק לפי הוראה כתובה ("תרגם מאנגלית לצרפתית"), בלי דוגמאות. הוא מפרש את הפורמט ומבצע את הפעולה – כמובן, מתנהג כמו מודל שמתאים את עצמו למטרה לפי טקסט בלבד.

Role-based ICL: כשותנים למודל רمز על מי הוא אמור להיות (למשל "אתה מתרגם מבריק"), הוא משנה את התנהלות בהתאם. הרקע שהוא למד עליו כולל הרבה טקסטים עם תפקידים ודמויות, וכך הוא יודע "לשחק תפקיד" לפי הקשר.

Explanation-augmented ICL: ככליל כל דוגמה מוסיפים הסבר, המודל עובד יותר טוב. ההסבירים עוזרים לנו להבין את החוק או התבנית שמאחוריו הדוגמאות, לא רק לשנן את התשובות.

Unsupervised ICL: אפילו כשראים למודל רק שאלות בלבד תשובות, הוא מצליח להבין מה המשימה ולפעמים גם לנחש את התשובה. זה קורה כי הוא מזהה מבנים מסוימים שראתה באימון, גם בלבד שיחיו תשבות זמניות.

Time Series Extrapolation: המודל מצליח לזהות דפוסים ולהמשיך סדרות של מספרים, גם כביש כמו טרנדים ביחיד (למשל עלייה + מחזוריות). הוא עושה זאת בלי אימון נוסף — רק לפי מה שהוא רואה בكونטקסט.

Meta-ICL: כשהמודל רואה כמה משימות ברכף (כל אחת עם דוגמאות), הוא משתמש לאורך הזמן. זה סימן שהוא מצליח לא רק להבין את המשימות, אלא גם להכפיל ולזוזות מבנים משותפים ביניהן תוך כדי>.

הקשר לשפה: איך ICL כומח מבני טקסט טבעי

המחברים מראים שלמודל יש יכולת להסתגל מתוך הקשר בغالל שהוא ראה הרבה דוגמאות של שפה שבה מבנים חוזרים, סדר, תפקידים, והקשרים משתנים לפי מה שנאמר קודם.

Coreference Resolution: המודל יודע לחבר בין ישויות (למשל "היא" מתיחסת ל-"Alice") לפי מה שהיא קיימת במשפטים. לעיתים זה פשוט, ולפעמים זה דורש להבין לוגיקה ופרטים מורכבים — כמו בبنמרהק Winograd.

Parallel Structure: כביש כמה משפטיים דומים במבנה, המודל לומד את החוק הכללי שמחבר ביניהם. למשל, אם רואים ש-Alex אוהב חתולים ו-Jordan אוהב כלבים, אפשר להסיק את התבנית ולהשלים משפטי חדש בהתאם.

Word-Sense Disambiguation: למילים כמו "bank" יש כמה משמעותות. המודל לומד מתוך הקשר איזו מהן מתאימה — בדיקת כמה שאנו עושים בקריאה.

Subject-Verb Agreement: גם מודלים פשוטים מצליחים להבין התאמת בין נושא לפועל. זה סימן שהמודלים הפינימו חוקים תחביריים, ומשתמשים בהם בזמן הרצאה.

Topic Modeling: המודל משנה את סגנון הדיבור והמילים שהוא בוחר לפי הנושא של הקטע. גם אם לא מציינים במפורש את הנושא, הוא יכול את זה לפי הקשר ומשנה את ההתפלגות של התוצאות.

איך לבדוק הכללה ב-ICL

המחברים מציעים שלושה כיוונים עיקריים לבדוק אם מודל באמת יודע להכפיל מתוך הקשר:

מה לומדים: האם המודל יכול ללמידה חדש לגמרי מתוך הדוגמאות בكونטקסט, שלא היה באימון? זו הבדיקה בין שינוי לבין הבנה אמיתית.

איך לומדים: האם המודל יודע ללמד את אותה משימה מכמה צורות שונות? למשל, מדוגמאות, מהוראות, מקוד או מטבלה? זו שאלת על גמישות החשיבה של המודל.

איך מישמים את מה שŁומדים: האם אפשר לקחת חוק שהמודל למד מספרים ולהחיל אותו על מילימ? או להסביר אותו? כאן בודקים האם המודל רק "מצבע", או גם מבין לעומק ומסוגל להכליל בין תחומים.

סיכום

המאמר מציע הסתכלות חדשה על ICL - לא כטכנית צרה של few-shot prompting, אלא יכולת הסתגלות כללית שנלמדת תוך כדי אימון על שפה טבעית. לפי הגישה הזאת, המודל לומד לזהות מבנים, משימות, תפקדים וחוקים מתוך הקשר, ומשתמש בזיה זמן הרצאה, בלי עדכונים. זה כולל גם דפוסים לשוניים פשוטים כמו התאמת פועל, וגם יכולות מורכבות כמו למידת פונקציות או הבנה של הוראות. הגישה של המחברים מחברת בין עולמות של מידול ומטה-מידה של שפה, וייצג משימות — ומציע דרכם חדשות למודול, להבין ולפתח את היכולות של מודלים גדולים.

<https://arxiv.org/abs/2412.03782>

26.04.25 המאמר היומי של מיק: Multimodal Latent Language Modeling with Next-Token Diffusion

היום שבת והסקירה של היום תהיה קצרה. הסקירה תתמקד במודלים מולטי-מודליים גנרטיביים המסוגלים "להבין" וליצור נתונים מכמה מודליות כЛОמר טקסט, תמונות, אודיו וכדומה. המאמר למעשה בעצם משדר מודלים לטנטיביים עבור דата טקסטואלי ועבור דטה רץ' יותר (למרות שגם הוא discretized). המחברים עושים זאת באמצעות אימון של מודלי דיפוזיה גנרטיביים עבור סוג מסוים במרחב הלטנטי. כЛОמר המודל מאמון לגנרט ייצוגים לטנטיביים עבור דטה טקסטואלי ועבור דטה כמו אודיו ותמונות.

להבדיל ממאמרים רבים המחברים מאמנים לא רק את המודל הגנרטיבי המולטימודלי אלא מאמנים גם מודל אמבדינג להפקה של ייצוגים לטנטיביים של דטה מודליות שונות. בדרך כלל מודל האמבדינג במודלי דיפוזיה הוא מסוג VAE (שהה VAE (Variational Autoencoder) והמחברים מציעים מודיפיקציה קלה ל-VAE. במקום שאנקודר (הקלט אליו הוא דטה) של VAE יGENERATE את קטורי התוצאות השינויות של הווקטור הלטנטי הוא מגנרט רק וקטורי התוצאות כאשר השינויות מוגרלות התפלגות גausית עם שנות נתונה (היפורפרטער). לדעת המחברים זה מונע קריסה(איפוס) של וקטורי השינויות הנוצר על ידי האנקודר שפוגע בגיוון התמונות שהמודל מגנרט.

המחברים מאמנים VAE עבור דטה לא טקסטואלי בלבד. תמונה או אודיו מחולקת לטוקנים (פnts'ים לתמונות ולמקטעים בזמן לאודיו) ומזומנים למודל כדטה סדרתי. שימושו לב המודל מסתכל על דטה בכל מודליות כמו דטה סדרתי. זה מאוד טריויאלי לדטה טקסטואלי ולאודיו כי יש שם סדר אינהרנטי ברור. בתמונות גם יש סדר אבל הוא יכול לבוא במקרה צורות: ככלומר ניתן לתאר תמונה כסדרה של פnts'ים במקרה צורות (למשל משמאלי לימיין ולמעלה למטה וגם מימין לשמאלי ומטה למעלה).

מודל דיפוזיה לדטה לא טקסטואלי מאמין לנקיות את הרעש מהדטה (denoising) בהינתן היצוג הלטנטי שלו(המורעש) ושל ההקשר (כל המודלים במאמר מבוסנים אוטורגרטיביים). לאחר מכן הווקטור הלטנטי הנקוי מזון לדקooder של VAE לשחזור הדטה כאשר המטרה של המודל המאמון היא לשחזר את הדטה כמה שייותר טוב. עבור דטה טקסטואלי ההרעשה מופעלת על האמבדינג של הטוקנים הטקסטואליים ומודל דיפוזיה מאמין לשחזר אותם. עבור דטה טקסטואלי מאמנים עוד שכבה לינארית שמטרתה למפות את הווקטור הלטנטי למרחב הטוקנים הטקסטואליים (סופטמקס בגודל של מיליון). דרך אגב מודלי דיפוזיה מאמנים יחד עם ה-VAE (אנקודר ודקודר).

כדי להפריד בין דатаה טקסטואלי ולא טקסטואלי המחברים מאמנים טוקנים המפרידים בין דатаה השיר למודליות שונות.

<https://arxiv.org/abs/2412.08635>

28.04.25 המאמר היום של מייק:

Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation

היום נסקור מאמר לא רגיל וקצת מרענן האמת- הרि לא כל יום (ואפילו לא כל חדש ואולי בכלל) יוצא לי לסקור מאמר שמדובר על מודלי למידת מכונה בישומים גיאוגרפיים. אכן שמעולם נכון - הרি ניתן למונח את הכלים העצמאתיים של למידת דיפ (deep learning) שפותחו בשנים האחרונות גם שם.

אוקי' אז המשימה שהמודל דן בה הוא זיהוי של מקום על כדור הארץ שבו צולמה תמונה נתונה. קלומר עבור תמונה נתונה אני צריכים להגיד מה הקואורדינטות על כדור הארץ (שכמו שאתם בטוח יודעים מהו ספירה (sphere)). המאמר מאמין מודל דיפוזיה שהקלט בו היא תמונה והפלט הוא הקואורדינטות על כדור הארץ (אני מניח שנייתן לתאר מקום על ספירה באמצעות קוטור דו-ממדי).

אתם זוכרים שמודלי דיפוזיה מאמנים להסיר את רעש מדטה בצורה הדרגתית כלומר כל הפעם המודל חוזה רעש שהתווסף לדאטה מיטרציה הקודמת. קלומר בהינתן פיסת דאטה מורשת ומספר איטרציה (בכל איטרציה מתווסףת לדאטה כמות קטנה של רעש) המודל חוזה את הרעש שצריך להחסיר מהדאטה כדי להציגו אותה (פיסת דאטה) ל"אייטרציה הקודמת" הפחות מורשת. באינפראנס המודל מתחילה מרעש טהור והופך אותו על ידי הסרת רעש הדרגתית.

מודלי דיפוזיה האחרוניים מבוססים על גישה שנקראת flow matching או FM בקצרה. FM מגדיר פונקציה מהירות שבאמצעותה ניתן לתאר את המסלול בין התפלגות הדאטה (המייצגת על ידי דגימות בDATASET) לבין רעש טהור. מהירות זו יכולה להיות תלויה במספר האיטרציה t (כמו במאמר זה) או קבועה כמו ללא מעת מאמראים אחרים על מודלי דיפוזיה. המודל מאמין לשערת מהירות זאת בהינתן דגימה רועשת t_x ומספר איטרציה t משערת את המהירות (t). אחרי שיש לנו אומדן זה ניתן לבצע אינפראנס על ידי פתרון משווה דיפרנציאלית ורגילה שהיא בעצם הגדרה של המהירות בתור נגזרת של t_x לפי t .

אוקי', כישל לנו תמונה שהפיקסל שלה זה מספר כלשהו בו $-1 \leq t \leq 1$ דהיינו לתאר התפלגות הרעש בתווך גאוסית. אבל אזכיר כי אנו נמצאים על הספירה במשימה שלנו והרעש צריך להיות מהו עצמו יהיה על הספירה וגם הדאטה המורעש חייב להיות על הספירה גם כן. קלומר אנו ננסים כאן לתchrom של גיאומטריה רימנית (Riemannian geometry) על הספירה. קלומר במקום להוסיף רעש לדאטה אנו מוסובבים את הדאטה בזווית התלויה במספר איטרציה.

בעצם אנו מרעשים את הדאטה על ידי הנעה בכיוון של משטח משיק עבור הספירה (זהה הכוונים שאנו יכולים לנوع מבל' ליפול מהספירה). זה כמובן משנה את הגדרה המהירות (זה כבר לא נגזרת רגילה של t_x לפי t) למשוואות קצט יותר מוסבכות (היחסים על ספירה לא פשוטים). ד"א לפי מה שאינו הבנתי מההממר הרעם הטהור שמתחלים ממנו אינפראנס מפולג באופן אחד על הספירה (למייבט ידיעתי זה לא לגמרי טריויאלי להגדיר את זה מתמטית - ניתן לעשות זאת בכמה צורות).

אבל דבר אחד נותר ללא שינוי - המודל מאמין לאמוד את המהירות עבור האיטרציה t בהינתן קואורדינטה מורעתת על כדור הארץ של תמונה נתונה, גם מהו קלט למודל דיפוזיה. התמונה מזנת לרשף אחריו העברת דרך אנקודר שלא מאמין (נותר מוקפא).

המאמר מאד מעניין - מי שבקיא בגיומטריה רימנית מזמן לצלול ולהינות :)

<https://arxiv.org/pdf/2412.06781>

30.04.25 המאמר היומי של מייק:

THE COMPLEXITY DYNAMICS OF GROKKING

זה אחד המאמרים החזקים והכי עמוקים שקרהתי לאחרונה. ולא, הוא לא אימן מודל שהשיג ציונים הכי גבוהים בכל הביצ'מרקם, לא הציעアルכטקטורה או שיטת אימון חדשה. מה שהמחברים ניסו לעשות זה להסביר את תופעת הנקראת ג্ּרוקינג (grokking) דרך הפריזמה של דחיסה. דחיסה של דטה ו גם דחיסה של המודלים. הנושא קצת מורכב וננסה להסביר אותו לאט בצורה פשוטה.

תופעת ג্ּרוקינג מתרכשת במהלך אימון של רשת נירונים כאשר אנו ממשיכים לאמן את הרשות אחרי שהגענו למינימום של לוס ולידציה. כמובן שבהתחלתה אנו נכנסים למודל אוברפריט ולוס הולידציה שלנו עולה ועולה. אבל בנקודת מסוימת קורה משהו מוזר - פתאות לוס הולידציה מתחילה לרדת וזה מצביע על כך המודל עובר ממודול של זיכרון (memorization) למודול של הכללה. במילים פשוטות המודל אשכבה "פיצה את הבעה".

תופעה זו מתרכשת במודלי overparameterized כאשר מספר המשקלים במודל גבוה הרבה יותר מאשר ש"צרי בשביל ללמידה הדאטסט" (ניתן להסביר זאת בצורה מדוייקת יותר אך זה מערב מתמטיקה לא טרייאלית שלא נחוצה להבנת סקירה זו). ג্ּרוקינג קשור לתופעה הנקראת lottery ticket hypothesis וגם double descent דרך אגב אם ממשיכים לאמן את המודל אז לוס הולידציה ממשיך לרדת ולא עוצר (כלומר מתקנס לאפוי).

אוקי, אבל איך כל הסיפור הזה קשור לדחיסה? בשביל כך אנו צריכים להסביר שני מונחים מאוד חשובים: הראשון הוא עקרון שנראה length description minimum או MDL. עקרון זה טוען אם אנו רוצים לדוחס את הדאטסט שלנו בצורה הטובה ביותר באמצעות מודל אנו צריכים למשוך סכום של אנטרופיית הדטה אחרי שהמעבר דרך המודל פלאס הקומפלקסיטי (complexity) של המודל עצמו. עקרון זה מtabssס על משפט הקיזוד של שנון הטוען שככל האנטרופיה של הדטה קטנה יותר ניתן לדוחס את הדטה בצורה יعلاה יותר (כלומר לדוחס אותו יותר).

אוקי, למדוד אנטרופיה של הדאטסט אחרי העבר דרך המודל אנו פחות או יותר יודעים. עברו מושם סיווג זה יכול להיות פשוט לוס cross-entropy. עברו לשערך את הקומפלקסיטי של המודל אנו צריכים לעבוד יותר קשה. קודם כל צריך להגיד מה זה הקומפלקסיטי של קולמוגורוב או KC. למעשה KC עבור דטה d נתון מוגדר בתרור אורך תוכנית מחשב (=קוד) הקצר ביותר שיכל לפולוט d. למשל עבור שורה של אחדות אנו צריכים קוד קצר KC (נמוך) וכדי להציג שורה של 0 - 1 רנדומליים צריך קוד ארוך בערך באורך השורה (KC) גבוהה. כמובן שלא ניתן לחשב KC במדויק.

עוד מושג חשוב שצריך לדעת להבנת המאמר הוא פונקציית γ rate distortion שבהינתן קלט x ואפסילון חיובי מגדרה מהו יציג מספר הביטים המינימלי (או KC) של קלט y עם שהוא רחוק מ-x באפסילון לכל היותר. כמובן "רחוק" תליה בפונקציה מרחק ובמאמיר תפקיד של x ו-y הם המשחקים מודל מאומן "רגיל" M ומודל coarse-grained או CS. מודל CS הוא מודל מאומן שעבר סוג מסוים של "פישוט" של M למשל קווינטוט, pruning או החלפת מטריצות משקولات ביצוגן על ידי מטריצות בעלות ראנק נמוך. גם מודל M שאומן עם רגוליזציה יכול להיחשב CS יחסית למודל שאומן ללא רגוליזציה. פונקציית מרחק שהמחברים השתמשו בה עבור חישוב של γ הוא הפרש בין הלואים של מודל רגיל M למודל CS.

אוקי, אחרי שהבנו את המושגים הנחוצים בוואו נחזור לג্ּרוקינג. התענה העיקרית של המאמר היא שכך לאנו מתקדמיים באימון המודל שמתקיים נהייה דחיס' יותר ככלمر קי'ם מודל CS עם הפרש ביצועים זניח (אפסילון) מהמודל המאומן (במהלך ג্ּרוקינג). כל זה קורה בזמן של length description של הדטה באמצעות המודל

רק יורד כלומר המודל אכן לומד את הדата באמצעות מודל דחיס (סוג של פשוט יותר). למה זה קורה בעצם? המודל מצליח להגיע ללוז קروس-אנטרופי נמוך באמצעות מודל דחיס (בעל *cross entropy distortion rate* נמוך לפי ההגדרה בפסקה הקודמת).

מקווה שה拙חתי להסביר את המאמר זהה בצורה ברורה יחסית.

<https://arxiv.org/abs/2412.09810>

המאמר היומי של מיק: 02.05.25 ON SPEEDING UP LANGUAGE MODEL EVALUATION

המאמר שמנסה לטפל באחת הביעות הכי מעשיות ופחות מדוברת בעבודה עם LLMs: איך מבצעים הערכת ביצועים יعلاה של عشرות או מאות פרומפטים או מודלים על סטים גדולים של שאלות, מבלי לבצע כמויות לא סבירות של זמן חישוב. כל הערה כזו דורשת להריץ מודל כבד שעשוי להיות בעל عشرות או מאות מיליארדי פרמטרים על כל דוגמה, עבור כל פרומפט. כשיש מאות פרומפטים ואלפי דוגמאות, אנחנו מדברים על מאות אלפי הריצות, שזה די יקר. זה שלא מדובר פה באימון אלא רק בהערכתו וזה מה שהופך את הבעיה לעוד יותר מעכנת: אנחנו רצים רק לדעת מי היכי טוב, בלי לשלם את המחיר של להריץ את כלם על הכל.

המאמר מציע שני אלגוריתמים חדשים שמנסים לפתור בדיקות אלה, בצורה חכמה ואdeptיבית. הראשון נקרא (הmbosso על UCB שזה Upper Confidence Bound המפורסמ)UCB-E-E, והוא בעצם מבוסס על רעיונות מהעולם של Multi-Armed Bandits (או MBA בקצרה). ככלומר, במקומם לבדוק את כל השיטות על כל הדוגמאות, האלגוריתם בונה לכל שיטה תחזית של כמה היא טוביה לפני מה שכבר נבדק, ומוסיף לה "בונוס א-ודאות" (בדומה ל-MCTS) שמעודד לבדוק שיטות(מודול + פרומפט למשל) שעדיין לא נבחנו מספיק. ככה הוא לא רק בוחר את השיטה שנראית הכי מבטיחה, אלא גם לא מזניח שיטות שיכולה להפתיע. עם הזמן, הוא משקיע את עצמו הערכה רק בשיטות שבאמת שווה לדעת עליון משהו.

אבל האתגר האמתי — והחינוך הגדול של המאמר — מגע בשיטה השנייה, שנkirat E-LRF-UCB. כאן הכותבים מבינים שהוא הרבה יותר עמוק: טבלת הביצועים (שיטות × דוגמאות) אולי נראה כמו מטריצה ענקית שאין ביריה אלא למלא, אבל בפועל יש בה הרבה מבנה. יש דוגמאות שהן די דומות זו לזו, ויש שיטות שמתנהגות בצורה מאוד דומה. ככלומר, קיימת קורלציה פנימית, שמאפשרת לחשב על הטבלה כמטריצה בעלת דרגה(ראנק) נמוכה, ככלומר כזו שאפשר לשחזר אותה היבט מסוים חלק קטן יחסית מהערכים. האלגוריתם מנצל את זה בדיק.

הוא מתחילה ממדגם קטן של תוצאות אמתיות (למשל רק 5% מהטבלה), ואז מאמן מודל של מטריצת דירוג נמוך, כזה שמקצת לכל שיטה ולכל דוגמה וקטור, כך שהמכפלה שלהם חוזרת התוצאות הצפויות. באופן זהה, האלגוריתם מסוגל לשערר את כל שאר התוצאות שלא נבדקו בפועל (בדומה למערכות המלצה עם low-rank factorization של פעם). מעבר להה, הוא גם יודע להעריך את חוסר הוודאות של כל אחת מהתוצאות האלה. עם כל סיבוב הוא בוחר איפה הכי משתלם לבדוק שוב: איפה שהתחזית הכי לא ודאית, או איפה שיש פוטנציאל למצוא את השיטה הכי טובה. כך, הוא לומד בהדרגה את המבנה האמתי של הבעיה, ומפנה את חישובי ההערכתה בבדיקה למקומות שיכולים להשפיע על החלטה.

הגישה עשוה שימוש מושכל בתבניות שקיימות בDATA, ויודעת להקליל מעבר למה שנמדד. היא גם אdeptיבית לగמרי, ככלומר משתמשת תורת מדיניות, בלי להניח מראש מי תהיה השיטה הטובה. ובעיקר היא מאפשרת לחסוך בין 85% ל-95% מההרצות שהיינו צריכים לעשות בגישה נאיבית. במונחים של עבודה עם LLMs, זה ההבדל בין מערכת שאפשר להריץ על GPU ביתי לבין אחת שדורשת תקציב של אלפי דולרים.

התרשמתי מהשילוב בין כלים מתחום החלטות (כמו UCB) לבין שיטות מטריציות מודרניות (כמו factorization), וכמה רוחק אפשר להגיד אם מחברים בין עולמות - מאמר מומלץ!

<https://arxiv.org/abs/2407.06172>

04.05.25 המאמר היום של מייק:

Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs

המאמר מציג מחקר ראשון מסוגו המתמקד בתופעה חדשה שאתרה ב-LLMs מתקדמים, המכונים "מודלים דמו' 1.o" (o1 או OpenAI דומו'). החידוש המרכזי של המאמר טמון בזיהוי, אפיון והצעת פתרונות לביעית "חישבת היתר" (Overthinking) במודלים אלו.

למייטב זכרוני זהה המאמר (פורסם בסוף דצמבר 2024) הראשון שmagidir ומantha באופן מוקף את תופעת "חישבת היתר" במודלים דמו' 1.o. התופעה מתבטאת בכך שמודלים אלו נוטים להקצות משאבי חישוב רבים (המתבטא ביצירת טוקנים מיוחדים לפעמים) גם עבור בעיות פשוטות מאוד (כמו "2+3"), תוך יצירה "שרשרת החשיבה" (Chain-of-Thought) ארוכה ומספר רב של פתרונות חלופיים, לעיתים קרובות ללא שיפור לדיק של התשובה הסופית.

המחקר מראה אמפירית שפתרונות שבאים מאוחר יותר בשרשרת החשיבה תורמים מעט מאוד לשיפור הדיק (לרוב, התשובה הנכונה מופיעה כבר בפתרון הראשון) ואינם מציגים גיון משמעותי בדרך החשיבה (פתרונות רבים חוזרים על עצמם בגיישתם). התופעה בולטת במיוחד במקרים קלות.

המחברים מגדירים מدد' יעילות חדשות:

המאמר מציג שני ממד' יעילות חדשים, שנועדו ל证实 את השימוש הרצינני במשאבי חישוב על ידי מודלים דמו' 1.o, מעבר לממד' הדיק המקבילים: הראשון הוא ממד' יעילות תוצאה (Outcome Efficiency): מודד את היחס בין כמות הטוקנים המינימלית הנדרשת לתשובה הנכונה הראשונה לבין סך הטוקנים שגונרטו. ערך נמוך מצביע על חישבת יתר מבחןת תרומה לדיק.

מדד יעילות תהליכי (Process Efficiency - P_E): מודד את היחס בין כמות הטוקנים התורמים לגיון בפתרונות (כלומר, טוקנים מסוימים המציגים גישה חדשה) לבין סך הטוקנים שנוצרו. ערך נמוך מצביע על חזרתיות וחוסר גיון בפתרונות.

פיתוח אסטרטגיות להפחית חישבת יתר:

המאמר בוחן אסטרטגיות חדשות להפחית חישבת יתר, המבוססות על פרדיגמת אימון עצמי (Self-training) וטכניקות אופטימיזציה העדפות (Preference Optimization), ללא צורך במידע חיצוני. החידוש מתבטא בישום שיטות אלו לבעה הספציפית של פישוט תשובות תוך שימור יכולות אינפרנס. המחברים השתמשו בטכניקות כמו SFT, DPO, RPO, SimPO ו-*N*-SimPO כדי לאמן את המודל להעדיף תשובות קצרות ויעילות יותר (שזהו ככללה מיותר דגימות מרובות), תוך שימוש בתגובה הארוכה ביותר כדוגמה שלילית (נמצא עיל יותר מתגובה ברירה המחדל).

המחברים פיתחו כמה שיטות חדשות לייצרת דאטסהט אימוןיעיל יותר על ידי חיתוך מכון של תגבות ארוכות: הראשונה היא פתרונות נכונים ראשונים (FCS - First-Correct Solutions) ששמורת רק על החלק המינימלי של התגובה עד להופעת התשובה הנכונה הראשונה. השנייה היא FCS + רפלקציה (FCS+Reflection) המהווה הרחבה FCS כך שיכלול גם את הפתרון השני שהגיע לתשובה הנכונה, במטרה לשמר יכולת "חשיבה ארוכה" אף עליה. הגישה נוספת שנבנתה היא נקראת GDS - Greedily Diverse Solution (GDS) שהיא הרחבה חמדנית של התגובה על ידי הוספת פתרונות רק אם הם מציגים פרספקטיבית חדשה ושונה מקודמיהם.

הגישה (שילוב SimPO עם FCS+Reflection) הצלחה להפחית משמעותית את כמות הטוקנים המיוצרת (לדוגמה, הפחתה של 48.6% ב-MATH500) תוך שמירה ואף שיפור קל ברמת הדיוק במגוון מבחנים ברמות קושי שונות (GSM8K, MATH500, GPQA, AIPE).

הסבר על מושגים:

שיטת (Simple Preference Optimization - SimPO): זהה לאלגוריתם אימון שמשמש פין טוון של המודל. המטרת שלה היא ללמד את המודל להעדיף תגבות מסוים (במקרה זה, תגבות יעילות יותר) על פני תגבות אחרות (פחות יעילות). המאמר מצא ש-SimPO הייתה היעילה ביותר מבין שיטות אופטימיציות ההעדפות שנבדקו.

שיטת (First-Correct Solutions + Reflection - FCS+Reflection): זהה לאסטרטגיה ששמשה לייצרת דאטסהט האימון עבור SimPO. בשיטה זו, לוקחים את התגבות המקורית של המודל ו"פישטו" אותה על ידי שמירה רק על החלק המינימלי של התגובה שהוביל לתשובה הנכונה (FCS), בתוספת הפתרון השני שהגיע לאותה תשובה נכונה (החלק של Reflection). המטרת הייתה ליצור דוגמאות אימון "טובות" שהן גם יעילות (לא ארוכות מדי) וגם שומרות על יכולת ה"חשיבה הארוכה" או הרפלקטיבית של המודל.

<https://arxiv.org/abs/2412.21187>

6.05.25 המאמר היומי של מייק: Graph Generative Pre-trained Transformer

אננו רגילים לראות מודלי שפה המאומנים בצורה בלתי מופתקחת (ב"כ נקרא אימון מוקדים) על טקסטים. המאמר זהה מרחיב את הקונספט של אימון מוקדים של מודל גנרטיבי על הגרפים. המאמר מעשה הופך גרף לסוג של טקסט כלומר סדרת טוקנים חד מימדי ומאמן טרנספורמר על הסדרה זו. אולם להבדיל מტקסט הgraf הוא יוצר לא חד מימדי באופן אינהרנטי וזה לאagemri טריוויאלי לייצג אותו בתור סדרה.

הדרך שהמחברים בחרו לעשות את זה נראה די אינטואיטיבית: הgraf מיוצג על ידי סדרה של קודקודים וקשתות. כל קודקוד מיוצג על ידי זוג של הקטגוריה שלו (דיסקרטי) והאינדקס שלו. הקשת מיוצגת על ידי שליטה שמפעיל את שני הקודקודים שהוא מחברת וסוג הקוד. הסדר בין הקודקודים יכול להיות כלשהו (כלומר אינוריאנטי לפירוטציה) אך סדר הקודקודים נבחר על ידי אלגוריתם פשוט: קודם בוחרים קודקוד בעלת דרגה הקטנה ביותר ובין הקשתות של בוחרים זו שMOVIL לקודקוד בעלת הדרגה המינימלית בין אלו שהוא מחובר אליהם. לאחר מכן מורידים את הקשת הזה ומתחילה את התהליך מחדש עד ש모רידים את כל הקשתות.

از אחר שרשמנו את הגרף בתור סדרה של קודקודים וקשתות (יש טוקן מיוחד המפריד ביניהם) מגעים לקידוד מיקומי (positional encoding) או PE. המאמר משתמש בקידוד מיקומי אבסולוטי כאשר כל קודקוד וקשת מקודדים עם המיקום שלהם בסדרה (המאמר לא מרחיב על איזו צורה של PE הם בחרו). לאחר מכן מבצעים אימון דומה לזה של מודל שפה כלומר אוטו-גרסתבי - חיזי טוקן (קודקוד או קשת) בהינתן העבר (כלומר קודקודים/קשתות הקודומות ביצוג). בקיצור אימון מודל גנרטיבי רגיל.

לאחר אימון מקדים המאמר מציע גישה מבוססת rejection sampling לפ"ן טיון. נגיד אם רוצים לגנרט גרפ מסויים המקיים איזשהו תנאי. נניח שבאימון מקדים היה לנו כמה גרפים המקיימים תנאי זה. אז מתחילה לגנרט גרפים ובונים דאטסהט מכ אלו שמקיימים את התנאי. אחריו שאספנו כמה עושים פ"ן טיון של המודל. ממשיכים לגנרט ווחזרים על התהיליך הזה המשלב סינון ופיין טיון.

המחברים גם מציעים שיטה לעשו אימון המשלב Proximal Policy Optimization או PPO לגרפים לפונקציה reward נתונה. המאמר מציע לשלב את הלוס של PPO עם הלוס על הקריטיק (שערך של פונקציית value) עם הלוס של האימון המקדים שהסבירנו עליי קודם.

המאמר ד' נחמד אבל מה שקצת מטריד אותו בגישה זו היא אינוריאנטיות של הייצוג הזה עבור כל פרמטרציה של קודקודים בייצוג שלהם. לדעתי זה מחייב אימון מאוד אינטנסיבי חישובית על מספר ענק של פרמטרציות של הקודקוד במיוחד על גרפים גדולים. אחרת הייצוג של הקודקודים יהיה פשוט ולא צזה טוב ...

<https://arxiv.org/abs/2501.01073>

המאמר היומי של מיק: 08.05.25 Memory Layers at Scale

המאמר זה משך את עיני כי מופיע בו המילה "memory" בהקשר מודלי שפה. כבר היום כשאתם מדברים עם ChatGPT, קלוד ומודלים אחרים אתם לא מדברים רק עם מודל שפה אלא עם מערכת שלמה הכוללת עצמה שכבות של זיכרון (למשל ממומשים כראג'שנסיבי Retrieval Augmented Generation (MoE) או קאשיים). המאמר מציע שכבה לרשת נירונית שהיא מגננון של זיכרון שנין לשומר בה וגם לאחזר ממנה בהתאם לשאלתה.

למעשה שכבת זיכרון זו דומה לבлок טרנספורמר אבל להבדיל ממנו אנו מאחזרים ממנו רק מה שרלוונטי לשאלתה המיוצגת על ידי וקטור q . כמו כן במקומות לשלב את כל הוקטוריים האפשריים שיש לנו בזיכרון אנו בוחרים k אלו שהם ה"מתאימים ביותר לocketor השאלתה q ". המנגנון הזה קצת דומה למנגנון MoE (שהה MoE נועץExperts) כאשר אנו בוחרים להפעלים תת-מטריצות של שכבות FFN. ההבדל בין השיטה המוצעת ל-E-MoE בעובדה כי ב-E-MoE אני המומחים (experts) הם קבועים (תת-מטריצות מוגדרות מראש של שכבת ה-N-FFN המלאה) וכאן ניתן לבחור כל שילוב של עמודות של שכבת ה-FFN.

ocketor שאלתה נתונה q אנו בוחרים את הוקטוריים הקרובים אליו ביותר מהזיכרון. K וקטורים בעלי צוין דמיון הגבוה ביותר נבחרים, משולבים עם מטריצת עריכים V (המאמר לא מפרט איך בדיק ועקב גם אני "חושד" במכפלה רגילה). לאחר מכן מכפילים את התוצאה במכפלה של q במטריצה הנלמדת W_1 שמוכפלת בפלט של מגננון attention שנמצא לפני הבלוק המ אחזר מהזיכרון, שעליו מופעלת אקטיביצית סואז (שהפכה להיות מאוד פופולרית לאחרונה). לאחר מכן מכפילים את התוצאה במטריצה נלמדת W_2 .

麥肯zie שאנן רוצים לשומר הרבה מאוד וקטוריים בזיכרון המכפלות שלהם עם q עלולים להיות כבדים מבחינה חשובה. כמו שמקובל הים המחברים "מחלקים את הזיכרון" בין כמה g s ו- q s ואז בודקים את הדמיון בכל אחד

מהם בנפרד ואז משלבים את התוצאות כדי לבחור את וקטורי הזיכרון הדומים ביותר. כמובן שהמטריצות בכל סקירה מוגבנין יותר ממטריצת הזיכרון הגדולה וגם וקטור φ מוחולק לכמה תת-וקטורים בין הסקירות.

שכבה זו יכולה להיות משובצת עם בלוקי טרנספורמרים במודלי שפה אבל אני גם לא רואה שום בעיה לשלבם עם שכבות אחרות כמו מבנה. מאמר נחמד וקליל (על הדרך גילית שיטה מעניינת לאחזר עיל מהזיכרון המבוזר על כמה סקירות).

<https://arxiv.org/abs/2412.09764>

10.05.25 המאמר היומי של מיק:

EfficientQAT: Efficient Quantization-Aware Training for Large Language Models

אימון מודע לקויניטוט (Quantization-Aware Training, או QAT) הוא טכניקה שבה המודל לומד כבר בזמן האימון להתמודד עם מגבלות הקויניטוט שיוופועלו עליו בזמן ריצה. מגבלות אלו מתבטאים בחישובים בדיק נמוך יותר (למשל INT8 במקום FP32). עם QAT המודל מאומן תוך חיקוי של תהליכי הקויניטוט, כך שבכל שלב באימון מדמים חישובים המדמים את עיבוד DATA בדיק מופחת. במהלך האימון שומרם על ייצוג מדויק לצורך חישוב גרדיאנטים, אך מוסיפים "הפרעה מבוקרת" בצורת קויניטוט קדימה ואחוריה (quantization & dequantization) כדי לדמות את ההתנגדות של המודל לאחר ההפחתה בדיק. כך, המשקלים והאקטיבציות מתאימים את עצמן באופן הדרמטי כדי להיות עמידים לשגיאות קויניטוט.

בניגוד לקויניטוט לאחר אימון (PTQ), אשר מתבצעת ללא התאמת של פרטורי המודל, QAT מאפשר שמירה על ביצועים קרובים יותר למודל המקורי גם לאחר המעבר לייצוג מקוונטוט. לרוב, משתמשים ב-"fake quantization" כדי לבצע כימות מודמה כחלק מגף החישוב של המודל, תוך כדי שמירה על רזרולציה גבוהה לחישובי הגרדיינט. השיטה מאפשרת לפרוס מודלים על חומרה חסונית כמו שבבים ניידים ו-Edge, מבליל ליותר על דיק תחזיות.

דימוי קויניטוט (וגם dequantization) ב-QAT מתבצע באמצעות שני פרטורים מאומנים עיקריים: האפס z (של הייצוג המקוונטוט) וגורם סקיילינג s עבור יעד קויניטוט נתון (ג'יד 8 ביט). אך המאמר מציע שני חידושים עיקריים. הראשון הוא אימון של s ו- z לכל בלוק טרנספורמר בנפרד (יחד עם משקליו). כלומר מתחילה מהבלוק הראשון מאמנים אותו יחד עם s ו- z שלו, מkapפיאים אותו (s ו- z) וממשיכים ל- s ו- z של בלוק הבא. ד"א ניתן לאמן s ו- z שונים עבור השכבות השונות של בלוק הטרנספורמר (FFN, attention) למשל.

ההידוש השני הוא אימון מלא של כל הבלוקים יחד אחרי שאימנו אותם בנפרד בשלב הראשון. במהלך השלב השני נותר קבוע ורק גורם הסקיילינג s מאומן.

זה זה - יש טענות לשיפור ביצועים כמו...

<https://arxiv.org/abs/2407.11062>

14.05.25 המאמר היומי של מיק:

ICLR: In-Context Learning of Representations

מודלי שפה מסווגלים לעשות הרבה יותר מאשר רק לשחזר עובדות או לבצע הוראות אלא הם מסווגלים להתאים את הייצוגים הפנימיים שלהם בהתקבוס על ההקשר בלבד ללא עדכון משקלילו (למקרה context-in). המאמר שנסקור היום מראה כי מודלי שפה יכולים לארגן מחדש את הגיאומטריה הסמנטית הפנימית שלהם באופן מלא, רק באמצעות פרומפט כאשר משקלילו נותרם ללא שינוי. זהה לא שניי "שטיח" בפלט. מדובר בארגון מחדש של מרחב הייצוג הפנימי של המודל, שנוצר מתוך מבנה ההקשר בלבד.

אימון מקדים של מודל שפה בונה מרחבים סמנטיים יציבים: מילים נרדפות מתקרבות זו לזו, מדיניות יוצרות קבוצות גיאופוליטיות, ומי השבע ונפרשים בمعالג. אבל מה קורה כאשר פרומפט משנה את היחסים האלה? אם המודל יוכל לבנות שימושות חדשה רק מהקשר? זו בדיקת השאלה שבודקים במאמר. הם מסיררים את הרמזים הסמנטיים שקייםים במודל מהאיימון המקדים וגורמים למודל להסתיק את המשמעות אך ורק מתוך מבנה הסדרה בפרומפט, ומגלים בכך יכולת מפתיעה של המודל ללמידה גיאומטריה ייצוגית חדשה בתוך ההרצה.

במערך הניסוי, בונים גרפ שכל קודקוד בו הוא טוקן מוכר (מילה כמו תפוח או רכبت). הליכה אקראיית על הגרפ מייצרת סדרות של טוקנים שהוא הפרומפט. המודל מתבקש לחזות את הצד (קודקוד) הבא, למלות שלמילים עצמן אין רמזים סמנטיים ישירים. למשל תפוז יכול להיות צמוד לרכיב בגרפ ומאוד רחוק (מספר הקשתות המינימלי ביניהם) לאגס.

כדי להצליח בחיזוי קודקוד הבא, המודל חייב לחשוף את מבנה הגרפ ולהתאים מחדש את הייצוגים של הטוקנים בהתאם. המבנה חבוי בסדרות מילים הנציגות מהגרף ולא במילים עצמן. כאשר מנתחים את האקטיביזיות הפנימיות בשכבות הטרנספורמר, מגלים תופעה מرتתקת. בתחילתה (עובר סדרות קצירות הנציגות מהגרף המזונת למודל כפרומפט), הייצוגים של הטוקנים עדין משקפים את המשמעות מהאיימון המקדים. אך ככל שהקשר מתארך (דgesיות ארוכות יותר), המרחב משתנה בפתרונות: טוקנים סמוכים בגרפ מתקרבים זה לזה במרחב הייצוג.

זה לא תהילך הדרגתית אלא קופיצה חדה, סוג של שניי פaza. מתחת לאורך הקשר קרייטי (מספר דוגמאות מגרפ המילים שלנו), המשמעות המקורית שלולטה. ברגע שעוברים את הסף, המודל "מתמסר" למבנה החדש, ומפנה את העולם הפנימי שלו לפיה המבנה החבוי בפרומפט. התופעה הזאת מראה שהמודול לא פשוט משנן זוגות טוקנים. הוא לא רק משחזר מילה הבאה מה"זיכרון" אלא בונה מבנה עקי ומקיף (החל מהשכבות הדי מוקדמות של הטרנספורמר) מתוך דפוסים מקומיים. אחת טענות המאמר אומרת כי "שניים שטיחים" שמתחבסים על שנין בלבד לא מצליחים להשתווות לביצועים של המודל או לשחזר את הגיאומטריה שנוצרת.

ארחיב על כך. נניח שבינו גרפ שבו קודקודים הם מילים ומשקליל הקשתות הם מרחקים בין ייצוגי המילים על ידי המודל (נניח על ידי שכבה מסוימת). אז מתרבר שהגרף הזה איזומורפי (pektrality) לgraf של מילים שאנו דוגמים ממנו לפרומפט. כמובן אם נחשב את ה- hc (שהה principal component של הcyion המוביל או מתאים לערך העצמי הגבוה ביותר) של הייצוגים הספקטראליים (שהה בגודל מטריצה שכניות ממושקלת) של שני הגרפים קיבל גרפים דומים.

כלומר אם נבנה עבור כל אחד מהם גרפ שבו המרחק (שהה 1 חלק המשקל) בין שני קודקודים (מילים) מוגדר על ידי המרחק בין הממד הראשוני בקטור \mathbf{c} המתאים לקודקודים אלו, נקבל גרפים דומים. כמובן מטריצות שכניות של שני הגרפים (הראשון מיצג מרחק בין ייצוגי המילים על ידי המודול השני גרפ השכניות שמנמו דוגמים לפרומפט) הם דומים זהה די מדהים. כמובן יציג המודל אשכלה לומדים את "עיקרי גרפ השכניות בפרומפט"

המחברים גילו עוד משהו מעניין. כאשר משתמשים במילים בעלי משמעות סמנטית חזקה (כמו ימות השבוע), המודל לא מוחק אותן. במקרה זאת, הוא שומר את מבנה השכ니ות הקודם בימי ד' בן הרשונים (בגרף הבני עלי יציג המילים על ידי המודל), ומטמיע את המבנה החדש במילדים הבאים (אבל משמעויותיהם) של המרחב הייצוגי. כך, המודל מצליח להחזיק בו זמני משמעות מוקדמת ומשמעות חדשה, על ידי הפרדה גאותרית בתת-מרחבים שונים.

המחברים משווים את המעבר הזה לתופעת פרקלוציה בפיזיקה: חיבורים מקומיים מצטברים עד שנחצה סף קרייטי, ואז מופיעה פטאום תבנית כוללת. כאן, אורך ההקשר, לא גודל המודל, הוא שמקטיב את הופעת המבנה החדש. ככל שהפרומפט מתארך, כך המבנה הפנימי נעשה צפי יותר, עד שמתחוללת קפיצה פטאומית במבנה. המחקר זהה משנה את ההבנה שלנו לגבי למידה *in-context*. הוא מראה שמודלים לא רק מושנים תגודות להקשר אלא בונים מחדש את עולם הפנימי לפי דרישות ההקשר.

<https://arxiv.org/abs/2501.00070>

16.05.25 המאמר היומי של מיק: GROKKING AT THE EDGE OF NUMERICAL STABILITY

לא יכולתי לפספס את המאמר זהה - לא היה שום סיכוי. הרי מילה grokking מופיעה בשם המאמר וזה ממשו אני לא מפספס בגלל שהוא אחד התופעות היכי מרתוקות ובלתי מוסברות כרגע (כמו *in-context learning*) בلمידה עמוקה. אבל מה זה בעצם גרוקינג?

גרוקינג זו תופעה במהלך אימון של מודלים שונים כאשר אחרי הגעה ל"ביצועים אופטימליים" על סט ולידציה. אם נמשיך לאחר מכן בהתחלה נראה ירידה בביטויים על סט הוילדיצה מלאה בעלייה של הביצועים על הטריין סט זהה כלומר אוברפיט. אם נמשיך לפחות עוד ועוד אז במצבים מסוימים (למשל במצב over-parameterized) אשר הקיבולות של המודל גדול בהרבה ממה לדאותו שלנו ציר") הlösן על סט הוילדיצה יתחיל לרדת שוב. כלומר האוברפיט נגמר והמודל נכנס למושט הכללה, לומד למידה אמיתית של הבעיה - וזה בדיק גרוקינג.

גרוקינג כאשר לתופעה אחריות המתרחשות באימון של רשתות ניורונים: double descent lottery ticket hypothesis. ניתן לאפיין תופעות אלו באמצעות כלים מפיזיקה סטטיסטית (עשוי זאת עוד בתחילת שנות ה-90). גרוקינג לא קורה אוטומטית במהלך כל אימון מאוד, לפעמים ציר להשתמש ברגולריזציה כדי זהה יקרה. המאמר חוקר את הסיבות שאי הופעה של גרוקינג דרך ניתוח של שינוי משקלות המודל במהלך האימון - כלומר גרדיאנטים.

המחברים טוענים שאי הופעה של גרוקינג קשורה לקריסת הגרדיאנט במהלך כלומר המודל מפסיק לעדכן את משקלותיו ועקב כך הגרוקינג לא קורה. המודל פשוט לא לומד. זה קורה בגלל שגיאות נומריות של פעולות floating point (או fp בקצרה). עבר פונקציית סופטמקס וגילה העדכנים הם כה קטנים שהמודל פשוט לא רואה אותם. כלומר הם מעבר לדיווק של FP אחרי הנקודה העשרונית. המחברים טוענים שתrikים ידועים שימושיים overflow-underflow כמו logsumexp (זהה חלוקה בערך המקסימלי שיש באקספוננט והוציאתו ממש אחר הלוג) - יש לנו log בלאו אחריו הסופטמקס הרי.

از הדבר הראשון שהמחברים מציעים הוא השכלול של סופטמקס הנקרא StableMax המקל על קריסת הגרדיאנט. פשוט לווקטים פונקציה שעולה בקצב נמוך יותר מהאקספוננט (יש לנו בסופטמקס). אך שיפור זה בלבד לא מספיק והמחברים מציעים שככל שאלgorיתם עידן הגרדיאנט במהלך האימון. המחברים שמו לב שלא מעט מקרים שגרוקינג לא קורה כי הlösן על הטריין יורד בעיקר באמצעות הכנסת "טמפרטור גבואה" לסופטמקס

במהלך אימון. כולם הרשות "בוחרת" לעדכן את משקליו על ידי הכפלתם בקבוע מסוים כל איטרציה של GD. כך הלוגיט של הקטגוריה מקבל ערכים מאד גבוהים וחיביים והאחרים מקבלים ערכים שליליים מאד נמוכים.

כאמור המחברים טוענים שסבירה לתופעה זו שניי של משקלות המודל בכיוון של משקלות המודל כלומר מכפילים אותו בקבוע מסוים. אך המחברים מציעים לעדכן את המשקלות במהלך האימון (GD) בכיוון של הגרדיאנט מוטל על היפר-משור האורתוגונלי לוקטור המשקלות הנוכחי. ככלומר הם מונעים מהמודל לנפח את משקלותיו בצורה שתוארכה קודם (אין שניי בכיוון משקלות המודל). המחברים רואו שככה ניתן להגיע לארוקינגן מהר יותר מאשר אימון עם רגולרייזציה.

מאמר בהחלט שווה קריאה,

המאמר היומי של מיק: 17.05.25

ZEROSEARCH: Incentivize the Search Capability of LLMs without Searching

במאמר ZEROSEARCH מוצגת שיטה חדשה לאימון יכולות חיפוש של מודלי שפה גדולים בעזרת שיטת מלמידה עם חיזוקים (RL), מבליל לשימוש כלל במגווני חיפוש אמיתיים. במקרה לגשת ל-API Google או ל-ChatGPT, הם מאמנים מודל שפה קטן יותר לשימוש סימולציה מנוע חיפוש, שספק מסמכים רלוונטיים או רועשים לפי צורך.

המודל המדמה עבר Fine-tuning על זוגות של שאלות-תשובות מtower אינטראקטיות אמיתיות עם מנוע חיפוש, מסמכים שהובילו לתשובות נכונות מסוימים כחיבויים, ואחרים כשליליים(זה די מוקרי האמת כי עושים זאת ב-ChatGPT). לצורך כך, הם שומרים גם את השאלה המקורית והתשובה הנconaה בתוך הפרומפט, כדי לאפשר למודל ללמידה הקשרים סמנטיים עמוקים יותר ולשלוט באיכות התשובה. לאחר מכן, במהלך האימון ב-RL, המודל המרכזי מייצר שאלות חיפוש, מקבל את המסמכים מהמודל המדמה, ומבצע reasoning כדי להפיק תשובה.

כדי לחזק את יכולות ההסקה, הם מוסיפים מגנון לימוד מדורג: איקות המסמכים יורדת לאורך זמן האימון באופן מבוקר, מה שמכריך את המודל להתמודד עם מידע חלק או שגוי. האימון נעשה בעזרת אלגוריתמים כמו PPO ו-GRPO, זהה דוקא די סטנדרטי. התוצאה: מערכת שגיעה אף עברת ביצועים של מודלים המשתמשים בגוגל, ללא עלות API ובשליטה מלאה באיכות המידע.

מאמר נחמד על איך לחפש ללא חיפוש באמצעות sms!!.

<https://arxiv.org/abs/2505.04588>

המאמר היומי של מיק: 20.05.25

Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks

3 ימים בלבד, סקירה עקב עמוס מטאורוף ואירעום משפחתיים לא מאד שימושיים. אך בחרתי מאמר ממש קליל שאפשר לסקור אותו בכמה משפטים בודדים. בסוף הוא קשור לראג ומזמן לא סקרתי מאמר על הנושא הפופולרי זהה. אוקיי, מה זה ראג? זה בעצם מודל שפה המצדיד בזכרון חיצוני המשולים את הידע של LLM שעוזר לו (מודל שפה) לבנות תשובה יחד עם הידע מהזיכרון. ככלומר במצב אידיאלי אנו חוצים לשלב את הידע שנוצר במודל שלנו יחד עם הזיכרון (נגד אוסף מסמכים) כדי לקבל תשובה אופטימלית.

אך איך מוצאים בראג מסמכים רלוונטיים? בפשטות לפני דמיין בין האמבדיניגס של המסמך והאםבדיניג של שאלה - בוחרים אז מה החיסרון הכי גדול של ראג? צריך לשמר את כל האמבדיניג האלו הזכרן מהיר כדי שהמודל ימצא

את המסמכים הרלוונטיים ויג'נרט לנו את התשובה מהר. מכיוון שיש לנו לפעמים מילוני מסמכים זה יכול להיות די יקר למורשת שיטות חיפוש מאוד עיליות הקיימות היום (vector database).

از המחברים הציעו שיטה די אינטואטיבית לחישוב בזיכרון של רаг באמצעות שימוש בקאש. אנו נשמר את המסמכים בזיכרון של המודול כמו kv-cache. פשוט נדחף בחלון הקונטיקט של המודול את כל המסמכים ונחשב את ה kv-cache עבור כל הטוקנים של כל המסמכים. מחשבים את ה kv-cache זהה מראש וזה מיותר לנו את הצורך בהשוות האמבעינגו של השאלה ושל המסמכים. ואז דוחפים את השאלה מודול שהוא prefilled עם המסמכים האלה ועושים אינפרנס עבורה. המאמר רומז שדווחים למודול את השאלות הקודמות עד שנגמר חלון ההקשר (לא בטוח שאני מבן למה).

از כמובן שיש לשיטה זו מגבלות - סט מסמכים צריכים להיות קטן ממספר כדי להיכנס לחלון ההקשר של המודול (למרות שבמודלים החדשים שכယול יש חלון הקונטיקט די ארוך). ועוד משחו קטע: הרעיון סדר של מסמכים לא רלוונטי שמאלץ אותנו לסגור על חכמת המודול שידע להעתלם מזה ולהשתמש בדברה בלבד.

מאמר די טריוויאלי אבל צריך לסקור גם כאלו לפעמים

<https://arxiv.org/pdf/2412.15605>

24.05.25 המאמר היומי של מייק:

rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking

כמה ימים לא סקרה מאמר אבל ביום הולדתי לא יכולתי לא לכתוב סקירה למורשת העומס המטורף. היום אסקור מאמר די מעניין שיצא לפני 4 חודשים והוא משלב פיין טין של מודול שפה למשימות מתמטיות באמצעות MCTS שזה קיזור של Monte Carlo Tree Search. רובם נראה מכירים את MCTS מהפרויקט המפורסם AlphaZero ו- AlphaGo ו- AlphaZero של דיפמיינד של אימנו מודלים המשחק Go. אצין AlphaZero למד לשחק רק דרך המשחקים עם עצמו ללא שום ידע מוקדם על Go. המודלים שפותחו היו כה חזקים שאלו הולם ב-Go פרש בעקבות אחד מהם (לא זכר איזה). הרעיון המתמטי מאחורי מודלים אלו היה MCTS.

אלגוריתם MCTS הוא אלגוריתם חיפוש המשמש בעיקר במשחקים לקבالت החלטות אופטימליות. הוא בונה עץ החלטות על ידי הרצת דוגמאות אקרניות (סימולציות) רבות של מהלכים אפשריים הנוכחי, ומעירך את אינוכם. לאחר מכן, הוא בוחר את המהלך שמניב את התוצאות הטובות ביותר בממוצע לאורק הסימולציות. האלגוריתם מażן בזרה חכמה בין חקירת מהלכים חדשים (exploration) העשויים להתגלות כיעילים, לבין ניצול מהלכים שכבר נמצאו כמושלמים (exploitation) בסימולציות קודמות (learning מוביילים לרוב לניצחון במשחק).

תהליך זה חוזר על עצמו, כאשר בכל איטרציה העץ מורחב והערכת אינוכות המהלים מटעדכנות, עד שמתתקבלת החלטה סופית. ארבעת השלבים המרכזיים בכל איטרציה הם: בחירת צמת הבא (selection), הרחבת העץ (expansion), סימולציה של המשחק (simulation), ועדכון ערכי הצמתים עד שורש העץ (backpropagation). הצלחת האלגוריתם נובעת מיכולתו להתמקד באזורי מבטחים יותר בעץ החיפוש, גם למרחב חיפוש עצומים. בסוף המודל, בהינתן מהלכי משחק נתונים, (מסלול בעץ) בוחר צומת בעל הסיכוי הגבוה ביותר לניצחון.

אבל איך זה קשור למודול שפה. התשובה היא אוטורגרטיביות. גם במודול שפה אנחנו כרגע חוזים טוקן לאחר טוקן כמו במשחק גו. בעצם הרעיון הגadol בשימוש ב-SMCTS לאימון של מודול שפה היא בניית דאטאטיסטים באינוכות גבוהה באמצעות חיפושם בעץ החלטות. אבל להבדיל מעץ החלטות שבו הצמתים הם מהלכי משחק כאן כל צומת הוא

שלב בתהיליך reasoning (הנמקה של המודל). לאחר מכן משתמשים בדעתה זו, בעל איכות גבוהה, כדי לעשות SFT למודל. אז השאלה כאן איך לדגם פתרונות נכונים ומוגנים עם גישה זו?

כאמור המאמר מציע מנגנון MCTS לאימון מודל לפתרון בעיות מתמטיות כאשר יש לנו פונקציית תגמול ברורה (האם הפתרון נכון או לא) בסוף הගנות. לעומת זאת התגמול (reward) באמצעות שרשרת הנמקה הוא משאנו בחרור (ד"א יש ב-PPO את אותה הבעיה - יש לנו פונקציית reward שאימנו אולם היא נותנת ציון לכל הפתרון ולא חלקו) ואז אנו מאמנים פונקציית value המשערת את התגמול בשלבי ביןיהם - דרך פתרון בעית רגסית). ב-MCTS בנייה פונקציה המקנה ציון לצומת (פתרון חלק) הוא קרייטי כי אחרת לא נצליח לבנות את עץ פתרונות בצורה טובה (כלומר מניבה פתרונות טובים בעיות מתמטיות). כזכור כל צומת בעץ נבנה על ידי דוגמה מודול שפה.

בהתחלת הצומת (= פתרון חלק) עד שלב מסוים) נבנה באמצעות שכיחות הופעתו בפתרונות נכונים של הבעיה. ככל הוא מופיע יותר בשרשאות הנמקה המובילות לפתרון נכון, ציון שלו גבוה יותר. בשלבים מאוחר יותר (כאשר פונקציית ציון מציצבת) המאמר עושם שהוא דומה לאימון מודל תגמול באימון RLHF של מודלי שפה. בכל עומק (שכבה) של עץ לוקחים צמתים בעלי ציוני הגבויים והנמכרים ביותר ומאימים מודל ציון צומת בסגנון Bradley-Terry (כמו שמקובל ב-RLHF סטנדרטי). כאמור פונקציית ציון משמשת אותנו לבחירה מאייזה צומת לדוגם שלב הבא לפתרון באמצעות אלגוריתם די סטנדרטי (UCT) Upper Confidence bounds for Trees exploration vs exploitation. המangel בין

כדי להגיע לפתרונות יותר איכותיים יותר מהר המודל מتابקש לממש כל שלב בשרשרת הנמקה בפייטון ואם קוד זה לא עבר טסטים, הצומת נפסל. המאמר מתחילה מודל שפה קטן, יוצר עץ פתרונות (עם כל השלבים שתיארתי), בוחר פתרונות הכי איכותיים (בעלי ציוני הגבויים ביותר), מצבע SFT על המודל וחוזר על זה עוד פעמי. וכותזאה מכך אנו מקבלים מודל קטן וחזק אבל מסוגל לפתור בעיות מתמטיות די מורכבות (לכורה).

<https://arxiv.org/abs/2501.04519>

26.05.25 המאמר היומי של מייק: Neuro-Symbolic AI i 2024: A Systematic Review

המאמר הוא סינטזה מדעית ועדכנית של ההתפתחות המהירה של תחום הבינה המלאכותית הנוירו-סימבולית (Neuro Symbolic AI) ב-5 השנים האחרונות. מתוך אוסף של 158 עבודות שעברו ביקורת עמיתים (peer review), החוקרים מציעים מיפוי שיטתי של התחום המרתק זהה, תוך הבחנה מדוקית בין מוקדי מחקר מפותחים לבין תחומים מפותחים פחות אך כלו. שעתידם קרייטי לפיתוח מערכות בינה מלאכותית (AI) אמינות ואוטומניות באממת.

63% מהעבודות שנבחנו על ידי המחברים עוסקו בלמידה והסקה, מה שמצוין על נטייה ברורה של הקהילה המחקרית לשלב למידה סטטיסטי (ככה אנו מאמנים מודלי למידת מכונה היום) עם אילוצים לוגיים. עבודות בולטות כוללות רשותות נוירונים לוגיות, שימוש בפרורים סימבוליים בלמידת few-shot, והכנסת ממשמעות סמנטית לפונקציות לוו. מטרת גישות אלו היא לצמצם את הצורך בדעתה, להציג את יכולת ההכללה של המודלים, ולשלב היסק דדוקטיבית עם היסק סטטיסטי (mbased על הדפוסים).

תחום ייצוג הידע (knowledge representation) מהוות 44% מהעבודות, ומתעלה מעבר לייצוגים מבוססי גרפים 'ידע פשוטים' יחסית. החוקרים בתחום זה עוסקים במבנה ידע קומוננסי ודינמי, באופן ייצוגים של ייצוגים

סמנטיים, ובשילוב בין ייצוגים המופקים על ידי רשותות נירונים לוגיים, כפי שנעשה ב-NeuroQL, שפה ייודית שמדגימה איך ניתן "לעשות יותר עם פחות".

תחום הלוגיקה היחסק (35%) כולל פרויקטים כמו DeepStochLog ו-Logical Credal Networks, המשלבים לוגיקה הסתברותית עם ייצוגים סמליים לצורך פתרון בעיות מורכבות. מיזוג זה הוא קריטי להתמודדות עם חוסר ודאות תכונה הכרחית במערכות הפעולות בעולם האמתי שהוא מהווה סביבה מורכבת מאוד וגם partially observable.

למרות הצורך הולך וגובר במערכות AI שקוות ובתווחות לשימוש, רק 28% מהבדיקות עוסקות ב-explainability ואמיניות. הפער הזה אינו רק מספרי, הוא מעיד (לפי דעת המחברים) על סדר עדיפויות לא נכון בקבילה המדעית בתחום זה. רוב המאמצים בתחום זה עוסקים בהסבירים פואט-הוק, תיוקנים סמנטיים או שיפור סיכום טקסטים. פרויקטים כגון Braids (שמשלב חוקים סטטיסטיים ולוגיים) ו-FactPEGASUS (שמדגיש עובדיות) הם יוצאים מן הכלל, אך אינם מייצגים מגמה רחבה. מה שحصر הוא ראייה מערכית: כיצד ניתן לבנות מודלים (או מערכות) שבהם תחוליך קבלת החלטות הוא מובן ושקוף כבר משלב בנין ארכיטקטורה והאימון ולא רק בדיעבד.

התרומה המשמעותית ביותר של המאמר היא הعلاאת המטה-קוגניציה שלו מ"הערת שליים" בטקסטונומיות קודמות לתת-תחום מוגדר וdochوف לפיתוח. רק 5% מהמאמרים שנסקרו נוגעים לנושא זה, ובכל זאת הרעיון של ניטור עצמי, שליטה אדפטיבית והיגיון פנימי הוא ללא ספק החלק החסר בארכיטקטורות הנירו-סימבוליות הנוכחות.

המאמר מאמץ הגדרה של מטה-קוגניציה יכולה לסייע ולהרהר בתהליכי קוגניטיביים פנימיים. מבחינה מעשית, זה כולל בקרים סימבוליים הממוקמים על גבי סוכני RL, אינטגרציות של LLMs עם ארכיטקטורות קוגניטיביות (למשל, ACT-R, Soar), וمسגרות תיאורתיות המתואימות עם המודל המשותף של קוגניציה. לבדוק אלה, אם כי מעתות, מצביעות לעבר עתידי שבו מערכותビינה מלאכותית אינן רק תגובתיות אלא מודעות לעצמן אסטרטגית - מסוגלות לנוהל תשומת לב, בחירה בין אסטרטגיות מסכנות ותיקון עצמי בהקשרים לא מוכרים.

אחד הבדיקות המרכזיות היא מיעוט העבודות המשלבות את כל ארבעת התחומיים המרכזיים: למידת פטרנים והיסק, לוגיקה הסתברותית, ייצוג ידוע, ו-explainability. הדוגמה היחידה הבולטת היא AlphaGeometry שהיא מערכת שפותחה בגוגל לפחותן בעיות גיאומטריות מואליימפיאדות למתמטיקה. המערכת יוצרת מילוני משפטים והוכחות סינטטיות, באמצעות NSAI שמדריך מנוע היסק סימבולי. זהו מודל מרשים של שילוב עמוק ידוע, יכולת כללית, וניתנות לבקרה. עם זאת, AlphaGeometry די חריגה בכך. שאר העבודות נוטות להיות מבודדות בתחום אחד או שניים. במיוחד מרגש חוסר השילוב בין explainability לתחומיים אחרים, מה שמצביר על צורך במחקר בין-תחומי אמיתי כדי למשם את החזון של NSAI.

ברמה המתודולוגית, אחת מחזקות המאמר היא הסיכון המרשימים של אוסף המאמרים שנבחרו: מתוך 1,428 עבודות, רק 158 עברו סינון איות שכלל ביקורת עמיתים, רלוונטיות, זמינות של קוד פתוח. פרט חשוב זה מהווה

לא רק מدد טכני אלא גם הצהרת רצינות. הוא משדר שהקיליה צריכה לשאוף לשחזריות(reproducibility), שקייפות, ונגישות, במיוחד בתחום כמו NSA שבו העיצוב הארכיטקטוני מורכב יותר.

<https://arxiv.org/abs/2501.05435>

28.05.25 המאמר היומי של מיק: Jasper and Stella: distillation of SOTA embedding models

מאמר ד' מעنى שמציע שיטה ד' פשוטה אך עובדת (כנראה) לזריקוק(distillation) ידע מכמה מודלים(מורים) מולטימודליים גדולים למודל אחד קטן (סטודנט). כמובן שהרצינול כאן טמון בכך שהמודל הקטן יכול על לימוד(בתקווה) את העשור הייצוגי מודלים גדולים מצד אחד והוא קטן מצד שני גם מבורך כי מקל על שימושו עם רגאים. הרעיון עבור מודלים בעלי מידת הייצוג קטנה יותר צריך פחות פעולות ארכיטמטיות לחישוב דמיון בין ייצוג דאטה נתון לבין הייצוגים השמורים בראג.

זריקוק ידע מתבצע ב 3 שלבים עיקריים. בשלב הראשון המחברים מנסים לקרב את הייצוגים המופקים על ידי שרשרת (ונרמול) של כמה מודלי מורה (הגדולים וחזקים) למודל קטן אחד. בשלב השני המיד של וקטור הייצוג המופק על ידי המודל הקטן שווה סכום של אלו המופקים על ידי המודלים הגדולים. פונקציית loss מורכבת מ- 3 חלקים.

הראשון מנסה לקרב את המכפלה הפנימית של ייצוגי המורים המשורשים וייצוג הסטודנט לאחד (כלומר למכבב אותם). החלק השני מנסה לקרב את הקורליציות בין הייצוגים של פיסות דاطה השונות על ידי המודלים - זה נעשה ברמת הבאים על ידי מזעור מרחק ריבועי בין "מטריצת קווריאנס לא ממורכצת" של המודלים הגדולים לבין המודל הקטן. גם "מטריצת קווריאנס" הוא המכפלה של מטריצה המכילה ייצוג של הפיסות דاطה בבאץ' בשחלוף שלה. הלווא الآخرן הוא סוג של loss ניגודי (contrastive loss) המנסה להשרות קרבה בין ייצוגים קרובים (לפי מודל המורה) עבור הייצוגים של מודל הקטן ובאותו הזמן להרחיק ייצוגים של פיסות דاطה לא דומות במרחב האמבדינג שלו (של המודל המזוקק).

בשלב השני מקטינים את מידת האמבדינג של המודל הקטן (הוא היה שווה לסכום הממדים של המודלים הגדולים) תוך שימור של תכונותיו. איך עושים זאת? מוסיפים 3 שכבות למודל הסטודנט מהשלב האחרון ומאמנים רק אותם עם שני הלויים האחרונים מהשלב הקודם.

בשלב השלישי מאמנים את האנקודר היזואלי (של תמונה) של מודל הסטודנט מולטימודלי כאשר כל החלקים האחרים מוקפאים. כאן מנסים לקרב את הייצוג המופק על ידי האנקודר היזואלי עבור תמונה זהה של כותרת התמונה המופק על ידי המודלים המורים. כאן משתמשים ב 3 הלויים מהשלב הראשון.

זה זה - מאמר קליל וקל להבין...

<https://arxiv.org/abs/2412.19048>

30.05.25 המאמר היומי של מיק: Learn Beyond the Answer: Training Language Models with Reflection for Mathematical Reasoning

בדיקות לפני שנה (30.05.24) התחלה את הסקירה היומית ופתחתי ערוץ טגרם Science and AI with Mike (מайк) בשביבים. מאז כתבתי 253 סקירות שבה עושה סקירה ל 1.44 ים למשך שלאחרונה קצר האטי את הקצוב.

از לרגל תאריך התחלתי לסקור מאמר בן קצת פחות משנה בנושא חשוב...

המאמר מציג חידוש מושג ומתודולוגי חשוב (נכון לפני 10 חודשים לאימון מודלים בשיקול דעת מתמטי). בוגוד לרוב המאמרים שיצאו לפניו, שמתמקדים בהערכת הדעתה באמצעות שאלות חדשות או תשובות שונות (augmentations על גבי מידע הדוגמאות), החידוש המרכזי כאן הוא בהציג גישה חדשה למטריה: הרחבת רפלקטיבית (Reflective Augmentation).

במקום להוסיף עוד דוגמאות, המחברים מציעים להעמק כל דוגמה קיימת ע"י הוספה קצרה רפלקטיבי אחריה הפתרון הסטנדרטי, הכלול שני רכיבים עיקריים:

1. מהלך אולטראנטיבי – פתרון נוסף שמציע נקודת מבט שונה על אותה בעיה.
2. הרחבה (up-follow) – ניסוח בעיה כללית או אנלוגית שמעמיקה את ההבנה.

החדשנות אינה רק טכנית אלא קוגניטיבית: במקרה לדמות למידה של כמהות (עוד שאלות), הגישה מדממת למידה אינטואיטיבית, דומה לזה של בני אדם, בדומה להרהור חוזר של תלמיד על פתרון נתון. פתרון בולט של הגישה הוא שהוא לא משנה את אופן ההסתקה בזמן הפעלה (inference). לעומת, אין צורך לكرוא או לייצר את מקטע הרהור בזמן ריצה, אך הוא כן משפיע על אופן החשיבה שנלמדת במהלך האימון. זהו שינוי עמוק בתפיסה "למידה על פתרונות" לעומת "למידה דרך הבנת עקרונות".

המאמר מגדים אמפיריות שגישה זו לא רק משפרת את הדיק בפתרון בעיות סטנדרטיות, אלא (זה המרכיב החשוב) – משפרת בצורה יוצאת דופן ביצועים בתרחישים רפלקטיביים: תיקון טעויות, פתרון שאלות המשך, והסתמכות על משוב חיוני. כמו כן, החידוש אינו מተנגש עם שיטות קיימות להערכת דата (כגון Q-aug או A-aug) אלא משלב עימן. השילוב בין השיטות קיימות הביא לתוצאות הגבוהות ביותר בינוין.

לסיכום, חידשו של המאמר נעוץ בשלושה מישורים:

- מעבר ממודלים של "שינוי פתרונות" למודלים של "הבנת עקרונות דרך הרהור".
- חיקוי של למידה אינטואיטיבית באימון של מודלים לשפה.
- גישה חדשה שקל לישם אותה על כל דатаה קיים מבלתי לשנות את תהליכי ההסתקה.

מדובר בחידוש בעל פוטנציאל רחב שה坦מש בעשרות מאמרים שיצאו השנה האחרונות אחריו, במיוחד עבור פתרון שאלות מתמטיות על ידי LLMs, אך גם בתוכן סוכנים אינטראקטיביים הדורשים חשיבה גמישה ולא ליניארית.

<https://arxiv.org/abs/2406.12050>

המאמר היומי של מ"יק: 01.06.25 Common Sense Is All You Need

דיסקליימר: זו סקירה של מאמר דעה ולאו דווקא מייצג את עמדת הסקור

בשנים האחרונות, AI רשמה התקדמות מרשים: מציאטיביטים ויצירת תמונות ועד למוכניות אוטונומיות. אבל עדין קיים פער יסודי שמספריד בין מכונות לבין האינטלקטואלית והגמישות שמאפיינת אפילו את בעלי החיים הפשוטים ביותר: שכיל ישר. במאמרו הוגו לטאפי טען שהיכולות החסרות הללו אין סתם חיסרין טכני אלא המכשול המרכזי שמנע מה-AI להציג לאוטונומיה אמיתית. לפי גישתו, אם ברצוננו לראות מערכות AI שפועלות

בבטחה ובאופן עצמאי בסביבות דינמיות ובלתי צפויות, עליו להטמיע בהן מראש הבנה הקשרית, הסתגלותית ואינטואיטיבית כלומר, שכלי ישר.

הטענה המרכזית במאמר היא שהמאמצים בתחום ממקדים יותר מדי בשיפור scale וביצועים על פני בנייה של הבנה אמיתית. לטafi מציע להפוך לטעש מודלים שיכולים לחזות היבט טקסטים או לזהות עצמים, ולהתחליל לבנות מערכות שיכولات להבין הקשרים, ללמוד תוך כדי תנועה ולהתמודד עם מצבים לא מוכרים. לא עוד תוספת שכבות נירוניים או מאגרי>Data אוצרם — אלא שינוי כיוון יסוד'.

אחת מהתרומות החדשניות ביותר במאמר היא הרחבת של מושג ה"התגלמות" (embodiment) ב-AI. לרבות התגלמות מתיחסת למערכות פיזיות כמו רובוטים שלומדים דרך אינטראקציה עם העולם הפיזי: הילכה, אחיזת עצמים, ניוט. לטפי מציע הרחבה: גם בסביבות מופשטות כמו חידות לוגיות, AI צריך לפעול מתוך אינטראקציה עם מבנה המשימה ולא רק לזהות דפוסים סטטיסטיים. הוא מכנה זאת התגלמות קוגניטיבית.

נקודה חשובה נוספת היא הדרישה להתחיל מצב של ידע מינימלי, או *Tabula Rasa*. רוב המודלים חיים מtabussumים על טרילוני מילימטרים או תמונות זהה מה שמאפשר להם ביצועים טובים בסיטואציות שכיחות אך כישלון בסביבות חדשות. לטפי טוען כי מערכות אוטונומיות באמת צרכות להתחיל כמעט מרגע בניית ידע מתוך אינטראקציה עם הסביבה. כך יוכל למנוע תלות בנטווי אימון וחזק את הגמישות הקוגניטיבית.

אחת הביקורות החשובות ביותר שלו היא מה שהוא מכנה "השלב שבו הקסם קורה". הרבה מפתחים מניחים שהבינה תופיע עצמה ברגע שנosis' עוד שכבות נירוניים או דאטה. אבל בלי שכלי ישר, AI פשוט יתרחק לתקרת זכוכית ישפר ביצועים קצת בכל פעם, אבל לעולם לא יוכל לאיגע להבנה אמיתית של מצבים מורכבים.

בעיות במדדים ובאמות מידת קיימות

לטפי מקדיש חלק משמעותי לניתוח מגבלות של מידת המידה הפוולריות, דרך שלושה מקרים בולטים. הראשון הוא אתגר ARC: סט של חידות המיעודות לבחון הפשטה והסקה. המודל נדרש להבין את החוקיות של מספר זוגות של קלט-פלט ולהזות את הפלט הנכון במרקחה חדשה. בפועל, מודלים מתאמנים על מאות דוגמאות ולפעמים גם על שאלות המבחן עצמן. כך הם מצליחים לא באמצעות הבנה אלא בזכות זיכרון. לטפי מציע לעצב אתגר חדש שבו המערכת חייבת להבין מידע מינימלי בלבד כדי לבדוק הסקת מסקנות אמיתית.

הדוגמה השנייה היא נהיגה אוטונומית לפי רמות SAE: מדריך 1 (סיוו בסיסי) ועד לרמה 5 (אוטונומיה מלאה בכל מצב). רוב המערכות חיים תקועות ברמות 2–3, ורמה 4 דורשת לעיתים הטעבות אונשיית מרחוק. לדעת לטפי, זו לא בעיה טכנית אלא בעיה של יכולת יישר: רכב צריך להבין מחוות אונשיות, אירועים חריגים, או סימנים לא סטנדרטיים. בלי הבנה אינטואיטיבית, לא ניתן להציג אוטונומיה אמיתית.

לבסוף, הוא בוחן את מבחן טיוריינג, שמודד אם מכונה יכולה ל��ים שיחה כמו אדם. זו נקודת ציון חשובה, אך לטافي מציין שמודול יכול לעבור את המבחן בלי להבין דבר. די ביכולת לייצר תגובות סבירות סטטיסטיות. זה אולי מרשימים בטקסט, אך חסר ערך בסיטואציות שבהן נדרש פועלה, הסקה או שיקול דעת.

לטافي לא מTELם מההישגים של שיטות ההגדלה (scaling). הן אכן שיפורו ביצועים בתחוםים כמו שפה וראייה. אבל הוא מציין סימנים ברורים של קיפאון: גם אם מושפעים דатаה וחישוב, הביצועים על מדדים מסוימים הפסיקו להשתפר. לדוגמה:

- מודלי זיהוי עצמים כמו COCO לא משתפרים מעבר לנקודת מסוימת.
- זיהוי חרגות בVIDeo (UCF-Crime) נעצר סבבב אותה רמת דיוק.
- איתור פעולות בוידאו (ActivityNet) תקוע במשך זמן רב.

המשמעות היא שהשקעה בהגדלה בלבד נוantha החזר הולך ופוחת. כדי להמשיך להתקדם, علينا לטפל בעיה היסודית: איך לגרום למערכות להבין את ההקשר שבו הן פועלות.

לטافي גם נוגע בכךה בעיות תיאוריות מוכנות בתחום, ומראה כיצד ככל ישר עשוי לפתור אותן. למשל, משפט Free Lunch No Cobus שאין אלגוריתם שמתאים לכל בעיה. תגובתו: נכון לנו לבנייה מערכות שפועלות היטב בתחוםים מסוימים ורלוונטיים, במקומן לנסות לפתור הכל. הוא גם דין בבעיית המסגרת (Frame Problem): הקושי לקבוע מה רלוונטי בכל מצב. לטافي מציע שהשתתפות فعلיה בעולם (פיזי או מופשט) עוזרת ל-AI ללמידה מה חשוב ומה ניתן להעתלם ממנו. באוטו אופן, בעיית ההכרה (Qualification Problem) כולמר הקושי לקבוע מראש את כל התנאים החדשניים לפועלה, נפתרת אם המערכת לומדת מתוך ניסיון והתאמאה.

לטافي מציע שורת שינויים מעשיים: ראשית, יש לעצב מבחנים חדשים שמודדים הבנה תחאלכית, לא רק תוצאה. שנית, צריך להגביל את הידע המוקדם של המערכת, כך שתהיה חיבת ללמידה ולהסתיק באופן עצמאי. יש גם לבדוק את הדרך שבה המודל חושב, לא רק אם התשובה שלו נכונה. מבינה ארכיטקטונית, הוא תומך בשילוב בין שיטות סמליות (לוגיקה וחוקים) לבין למידת מכונה. גישה היברידית שכזו תשלב את הगמישות של רשותות נוירונים עם השקיפות של לוגיקה פורמלית. בנוסף, הוא מציע לשאוב השראה מדעי המוח והקוגניציה ולא להעתיק ביולוגיה, אלא ללמידה ממנה איך מערכות לומדות בפועל.

ומה יקרה אם לא נשנה Ciou?

המאמר מזהירשמי שימוש לבנות מודלים גדולים יותר, מבל' להתמקד בשכל ישר,יסבול לא רק מהאטיה אלא גם מאובדן אמון. מערכות שמתפקידן היטב במעבדה אך נכשלים בשטח יובילו לאכזבה של משתמשים, משקיעים ומפתחים. גרווע מכך, מערכות שמקבלות החלטות בלבד אך חסרות הבנה של הקשר או ערכים אנושיים עלולות

לפעול בצורה מזיקה. לטאפי טען שהפחד הציבורי מ-AI ובמיוחד מ-AI שמשתפר מעצמו נובע לא מהאינטיגנץיה אלא מהיעדר שכל ישר. מערכות שמתפתחות מבלי להבין את ההשלכות, ההקשרים או הנסיבות המוסריות יוצרות סיכון אמיתי. لكن, הבטחת שכל ישר במערכות כאלה היא לא רק יעד טכני אלא תנאי לבטיחות.

<https://arxiv.org/pdf/2501.06642>

03.06.25 המאמר היומי של יניב ומיק:

Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't

מודולי Reasoning GPT של OpenAI חוללו מהפכהVICOT הסקה ל佗ות פתרון בעיות מתמטיות מורכבות ועד כתיבת קוד אלגנטי הודות לשאבי חישוב עצומים ומאגרי נתונים עצומים שאומנו באמצעות למידה מחיזוקים לבצע חישיבה לוגית. אך האם מודלים קטנים וזולים יותר יכולים להציג להישגים דומים? מאמר חדש מאת עוסק בדיק ב שאלה זו.

למה זה חשוב?

מודלים גדולים מספקים ביצועי הסקה מרשימים, אך דורשים משאבים כבדים והם יקרים לשימוש נרחב. מנגד, מודלים קטנים (בסדר גודל של 1–2 מיליארד פרמטרים) זולים ונוחים לפרש, אך לרוב נופלים מאחור במשימות הסקה מורכבות. מטרת המחקר של דאנג וו היא שאפתנית אך מעשית: לשפר את ביצועי הסקה של מודלים קטנים תוך שימוש מינימלי במשאבים.

השיטה: GRPO על דاطה באיכות גבוהה (Group Relative Policy Optimization)

החוקרים בחרו במודול DeepSeek-R1-Distill-Qwen-1.5B (בגודל 1.5 מיליארד פרמטרים) ואימנו אותו באמצעות GRPO: אלגוריתם למידה מחיזוקים שהציגו DeepSeek והוכיח את עצמו במודלים גדולים, אך ישם כאן בקנה מידה קטן משמעותית. כדי לשמור על עלויות נמוכות, האימון היה מוגבל מאד. מבחינת חומרה, היה שימוש ב 4 קרטיסי מסך NVIDIA A40 בלבד. משך האימון הגיע ל 24 שעות. הדאנטס שנוצר היה מורכב מ- 7,000 שאלות מתמטיות שנבחרו בקפידה.

תוצאות מפתיעות

למורות המגבילות החומרה, השיפור ביצועים היה יוצא דופן: דיק בנצח'マーク AMC23 קפץ מ-63% ל-80%. בנצח'マーク AIM24 התקבל ציון של 46.7%, שעוקף את preview-0 של OpenAI שעומד על 44.6%. עלות האימון הכלולה: כ-42 דולר בלבד שזה סדר גודל זול יותר ממתקודות עכשוויות מובילות.

מה בדיק נעשה?

החוקרים ביצעו שלושה ניסויים:

- **ניסוי 1:** אימון עם שאלות מתמטיקה קשות ואיכותיות. שיפור מהיר אך ירידה חדה ביצועים עקב חוסר יציבות ו"סיטיית שפה".

- **ניסוי 2:** שילוב שלאלות קלות עם קשות. השיג יציבות התחלתית גבוהה יותר ותוצאות شيئاً מרשימות, אך גם כאן חלה הידדרות לבסוף.
- **ניסוי 3:** שימוש ב- reward cosine לעידוד תשובות קצרות שיפר את היציבות וביצועים. ממצאים אלה עוסים בקנה אחד עם מאמר "DR GRPO", שעליה לארכיב ימים ספורים קודם ומזה נטיה מובנית לששובות ארוכות ב-GRPO.

מגבליות ושאלות פתוחות

- **למה זה עבד כל כך טוב?** התוצאה המפתיעה ביותר היא היכולת של מודל קטן, שאומן ב מהירות ועל מעט>Data, להציג ביצועים כל כך מרשימים. החוקרים אינם מספקים הסבר אינטואיטיבי, דבר שימושי שאלות על עמידות התוצאות למשימות שונות שאין מתמטיות.
- **סטיית שפה:** עם התמסחות האימון הבסיסי הרב-לשוני של המודל הוביל לשובות שאין באנגלית, מה שגרם לא-יציבות בכל גרסאות המודל.
- **СПЦИФИЧНОСТЬ ЛЕКСИЧЕСКОГО ПОЛЯ:** הערכה התמקדה רק בהסקה מתמטית. לא ברור אם הגישה תעבור גם במדעים, קוד או תחומיים אחרים.

מה הלאה?

המחקר מוכיח שלא חיברים מודלים ענקיים ויקרים כדי להגיע לביצועים טובים בהסקה. ההצלחה של המודל הקטן מצביעה על פוטנציאלי לבדוק גרסאות נוספות של DR GRPO כמו O GRPO תחת מגבלות משאים, ולבחון ביצועים על מגוון רחב יותר של משימות. כיוון נוסף הוא כיוון הiper-permatrims ויתכן מאוד ש-loss KL גבוהה יותר יסייע את היציבות.

בשורה התחתונה

המאמר מתווה כיוון חדש: מודלים לשוניים קטנים, זולים ובעלי יכולת הסקה משמעותית. גם אם הסיבה לכך עדין לא ברורה לחלוון, ההשלכות המעשיות ברורות; פוטנציאלי דמוקרטיזציה של יכולות AI מתקדמות גם למעבדות קטנות וחוקרים עצמאיים.

[[לקריאת המאמר המלא](#) →]

[לעון בקורס שלהם – יש בו פוטנציאלי אדיר ליישום הסקה בתחומיים שונים עם השקעה מינימלית]

המאמר היום של מיק: 05.06.25

Task Singular Vectors: Reducing Task Interference in Model Merging

היום הסירה הולכת להיות מאוד קלילה וקצרה. המאמר מדבר על שילוב של מודלים שאומנו (כלומר עברו fine-tuning) מאותו מודל הבסיס למשימות שונות לבניית מודל שיפגין ביצועים טובים בכל המשימות האלה. ככלمر כל מודל צזה עבר שינוי מסוים במשקליו למשל יחסית למודל הבסיס בעקבות תהליך פין טון (למשל זה יכול להיות LoRa אך לא חיבר).

השיטה הפשוטה להתאים מודלים אלו לכל המשימות יחד היא להוסיף למודל הבסיס את הממוצע של שינוי המשקלים עבור כל המודלים. לטעתן המחברים זה לא תמיד עובד בצורה מושלמת גם במקרים שהמשימות

דומות. אז המחברים מציעים שיטה די אינטואיטיבית שמטרתה היא להקטין את ה"הפרעות הדדיות" בין מטריצות התוספות לכל המשימות.

air עושים זאת? קודם כל המחברים שמו לב שמטריצות התוספות למשימות הן לרוב בעל ראנק נמוך. לכל מטריצות התוספות מבצעים SVD (שזה Singular Value Decomposition) ומקבלים את המטריצות האורתוגונליות השמאליות והימנויות (U ו- V) ומטריצות אלכסונית D של הערכים העצמיים (ויתר נכון הסינגולריים). לאחר מכן מוסיפים קטע של וקטורים סינגולריים, המתאים לע"ס (ערכים סינגולריים) הגבוהים ביותר לכל מטריצה נוספת ובונים מהם (כמו שעושים ב-PCA).

בשלב השני המחברים בונים מטריצות U ו- V שבאמצעותן עושים דקורסציה (הלבנה) של מטריצות התוספות יחד (ויתר ספציפי דרך U ו- V) למשימות שונות. כמובן שהמטרה להפוך אותם לחסרי קורלציה. בשביל כך לוחים את המטריצות U ו- V לכל המשימות, משרשים אותם למטריצות גדולות, ואז מוצאים לכל אחת מהן מטריצה "מלבינה" בשיטת די סטנדרטיות מתורת המטריצות (קשרו ל- $\text{Moore-Penrose inverse}$). בסוף משתמשים במטריצות אלו כדי לבנות את השילוב של כל מטריצות התוספות (במוקם לשילב אותם בסכום המחברים משלבים אותם בסכום מסווקל).

המחברים מציעים לבצע את התהליך זהה לכל שכבה בנפרד (לא בטוח עד כמה זה חידוש).

<https://arxiv.org/abs/2412.00081>

המאמר היומי של מיק: 07.06.25

Rate-In: Information-Driven Adaptive Dropout Rates for Improved Inference-Time Uncertainty Estimation

היום אני סוקר מאמר מיוחד בכמה רבדים. הרובד הראשון אחד ממחברי של מאמר זה הוא לא אחר אלא אין לקון, אחד האבות של למידה عمוקה. הרובד השני מכיל את החוקר הישראלי הידוע (אך לא מספיק) רביד זיו שורץ שהוא גם פרופסור באוניברסיטת ניו יורק. הרובד השלישי הוא גושא המאמר והוא שערוך אי ודאות עבור חיזויים של רשתות נירונים - נושא מאוד מעניין אותו אך לא מעט זמן לא סקרה צזה.

air ניתן לשערך אי הוודאות של החיזויים של רשת נירונים? יש כמה משפחות של שיטות המוזכרות במאמר:

רשתות נירונים בייסיאניות מגדרות התפלגות הסתברות על משקל הרשת, מה שמאפשר למדל אי-ודאות דרך ההתפלגות הפווטרורית. עם זאת, הן כבודות חישובית וקשה להרחב אונן.

שיטות אנסמבל:אמונות מספר מודלים ומגדירות את התוצאות שלהם. מסוגלות למדל גם אי ודאות אפיסטטמית וגם אליאטורית, אך דורשות משבים חישוביים רבים.

אגומנטציה של דאטा בזמן טסט (Test-Time Augmentation): מושיפות שיבושים לקלט (כמו סיבוב או טשטוש) כדי להעיר את התפלגות התוצאות. עיל בערך כישיש ידע מוקדם על מבנה הנתונים.

הזרקת רעש למודל: מושפים רעש נשלט (למשל גאוס) למשקלים או לפעולות כדי לבחון רגישות מעבר לשינויים בקלט.

שיטות מונטה קרלו (MC): משתמשות בדגימות אקראיות כדי לאמוד אי ודאות. למשל, MC Dropout מפעיל DRPO (dropout) גם בזמן טסט כדי לדגום את מרחב משקל הרשת. יש לא מעט שיטות נוספות מבוססות MC לשערוך אי ודאות ברשותך.

אבל air ניתן לשערך את הוודאות? אחת הדרכים היא להשתמש בגישות מתורת המידע (information theory) לניתוח של זרימה המידע בתוך הרשת ומידת "פגיעתה" מהשיטות המוזכורות לעלה (למשל MC Dropout).

בגודל מאד ככל שזרימת המידע נפגעת יותר - אי הוודאות של החיזויים עולה. שיטות מתורת המידע די נפוצות

במחקר של רשותות עמוקות למשל:

עקרון צואר הבקבוק המידי (של נפתלי תשבי): מציע שכבות ברשת נוירונים שואפות לדוחס את המידע מהקלט תוך שמירה על המידע הרלוונטי לפט. משמש לניתוח דינמייקת הלמידה והכללה של המודל.

ניתוח מידע הדדי (Mutual Information): הערצת המידע ההדדי בין הקלט, השכבות הפנימיות והפלט מס'יעת להבין כיצד מידע זה ומעובד ברשת. זה הטכניקה שהמחברים משתמשים בה במאמר **טכניקות רגולרייזציה אינפורמציביות**: שיטות כמו dropout rate כדוגמת **information dropout** שולטות בזרימת המידע במהלך האימון כדי לשפר חוסן והכללה של המודל.

אוקי', אז המאמר מציע שיטה מבוססת מידע הדדי המשכילה dropout rate MC. במקרה להשתמש בdropout rate קבוע לכל השכבות (כלומר מה אחוז הנוירונים המוחזקים בשכבה) המוחברים מציע לקבוע אותה (dropout rate) בתתלות ב מידת פגיעה בזרימת המידע בשכבה. המטרה כאן היא לעשות את אובדן המידע בכל שכבה פחות או יותר קבוע. אם אובדן המידע הדדי גבוה (מקבוע אפסילון) מדי מקטינים dropout rate ואם זה נמוך מדי מגדילים אותו.

ד"א פגיעה בזרימת המידע בשכבה מחושבת דרך חישוב של המידע הדדי בין אקטיבציות של הקלט בשכבה לבין אלו של פלט השכבה. מתברר שזה די לא טריוויאלי ומה אמר זו בהרחבה איך ניתן לעשות זאת.

<https://arxiv.org/abs/2412.07169>

09.06.25 המאמר היומי של יניב ומיק: Spurious Rewards: Rethinking Training Signals in RLVR – Fast Overview

המסר המרכזי במשפט אחד

גם תגמולים אקרים או שגויים יכולים להוביל לשיפור דרמטי ביכולות פתרון בעיות מתמטיות – אבל רק אם המודל כבר "מכיר" את הדרך מהפירה-טריאינינג.

למה זה חשוב

למישה באפשרות חיזוקים עם תגמול ניתן לאיומות (RL with Verifiable Rewards - RLVR) היפה לשיטה מובילה לשפר יכולות חישיבה של מודלים גדולים. המאמר שואל שאלה פרובוקטיבית: **האם אנחנו באמת צריכים תגמול מדויק?** התשובה: לא תמיד.

מה עשו החוקרים

הם לקחו את המודל Qwen-2.5-Math ואמנו אותו על סט שאלות מתמטיקה עם חמיש גרסאות שונות של תגמולים:

- **תגמול אמיתי:** מודל מקבל נקודה רק אם התשובה נכונה.
- **תגמול לפי הצבעת רוב:** המודל מייצר 64 תשבות, ומוגמל את התשובה השכיחה.
- **תגמול פורטטי:** אם התשובה כוללת ביטוי מתמטי (למשל $\boxed{\boxed{}}$), היא מתוגמלת, בלי קשר לנכונות.
- **תגמול אקרים:** הטלת מטבח קובעת אם לתגמול.
- **תגמול הפור:** רק תשבות שגויות מקובלות נקודה.

במפתיע, כל אחד מהתגמולים הללו הצליח כמעט כמו תגמול אמיתי ככלומר המודל השתפר דרמטית גם כש האות החזוק לא היה קשור כלל לתוצאה הנכונה.

ממצאים עיקריים

1. **Qwen משתפר בכל תנאי:** גם בלי תגמול נכון, המודל לומד לפחות טוב יותר. לעומת זאת, מודלים אחרים (כמו Llama3 ו-LOMo2) זוקרים לתגמול מדויק כדי להשתפר.
2. **הגורם הסמי:** פתרון דרך קוד. Qwen כבר ידוע לנוכח פתרונות בפייתון מתוך הטקסט. אימון RLVR רק גורם לו לבחור באסטרטגיה חזוי יותר וمبיא לדיווק גבוה יותר.
3. **שיפור בבדיקה ממעבר מ"לשוני" ל"קוד":** בשאלות שבוחן המודל התחליל לכנתוב קוד בעקבות האימון, הדיווק קפץ בכמעט 26%.
4. **איך תגמול אקראי עובד?** האלגוריתם GRPO כולל קליפינג שمعدיף פעולות בסביבות גבוהה – וכך שגם כשאין קשר לתוצאה, המודל לומד לחזק את ההתנהגות הדומיננטית שלו.
5. **לא כל מודל נולד שווה:** כשאין במודל נתיה מוקדמת לקוד, כמו ב-LOMo, אותו תגמול אקראי פשוט לא עובד.

סיכום:

המאמר מראה שלעיתים קרובות **אימון RL לא מלמד CISORIM חדשים, אלא מחלץ CISORIM חקיים** שהמודל כבר פיתח בפירה-טרינינג. לא תמיד צריך תגמול מדויק – אם המודל כבר "מכיר" את הדרך, מספיק לאותת לו לחזור אליה. עם זאת, זה לא נכון לכל מודל – יש אולי שדרושים הנחיה מדויקת כדי להשתפר.

<https://arxiv.org/abs/2412.07169>

11.06.25 המאמר היומי של מייק: TRANSFORMER-SQUARED: SELF-ADAPTIVE LLMS

זמן רציתי לסקור את המאמר הזה אך הוא הלך לי לאיבוד בפייפ המאמרים הבלתי נגמר שלי (רגעע עומד על 353 מאמרים העומדים להיסקר או להיפסל לסקירה מתישהו). המאמר נכתב על ידי מדענים (בתקווה 😊) מחברת AI Data Scientist AI (שקיים ביקורת די טובות לuibet צרכני). המאמר מציע שכלול מאד פשטוט לתהילך האימון של מודלי שפה בתרחישים מולטייטסקינגן. כאן מולטייטסקינגן אומר לנו מאמנים כמה מודלים-מומחים (לא לבלב עם EoM) שככל אחד מהם מתמחה במשימה מסוימת מסוימת מאיזה מודל בסיס חזק. ה

כל מודל כזה מאמן בצורה דומה לאdeptרים שזה סוג של PEFT שהוא SVD (Parameter Efficient Fine Tuning) שזה שפה של משקלים מאומן במהלך FT. המאמר מציע שיטת PEFT הנקראת SVF שזה למעשה SVF Singular Value Fine Tuning שmethodת להتاימים את המודל למשימה נתונה. כמו שניתן להבין ממשמה מבוססת על הערכים הסינגולריים שמקורה זהה הם ערכים סינגולריים של מטריצות המשקלים בשכבה MLP. דרך MLP מכיל שתי מטריצות משקלים בכל בלוק של טרנספורמר והמאמר לא מסביר (לפחות אני לא ראיתי) איך בדיק נבנית מטריצה משקלים בכל בלוק (אולי עושים SVF לכל מטריצה בנפרד).

از מה בעצם עשה SVF? הוא מבצע SVD (Colomer decomposition SVD) עבור מטריצות משקלים בכל בלוק טרנספורמר במודל. אחת ממטריצות אלה היא אלכסונית ואילו שתיים האחרות הן אורטורוגונליות (משמאל ומימין). המחברים מכנים למיניהם המכפלת זו מטריצה אלכסונית Z למלדת ואתה מאמנים במהלך האימון. יש

כאן איזה הינה שמודל הבסיס למד את כל "המשימות האפשרות" ובמהלך פיניטיון אנו צריכים לחזק כלו הרלוונטיות למשימה הנלמדת.

معنىין כי פין טיון בוצע תוך שימוש בשיטה השיכת למידה עם חיזוקים או RL בקצרה הנקראת REINFORCE עם רגולריזציה רגילה שימושיים באימוני RL של מודלי שפה. שמעתם נכון הם לא השתמשו ב- PPO, לא ב- GRPO ולא ב- DPO ובנוסף המחברים עשו זאת עבור משימות עם rewards verifiable כולם אכן שינו תדעת האם התשובה נכונה למשל שאלות מתמטיות או קידוד. במהלך אימון זהה מאמנים רק מטריות Z בכל השכבות.

בainforce המחברים מציעים 3 שיטות. בשיטה הראשונה בשלב הראשונה שואלים מודל לאיזה משימה שייכת שאלה ל-LLM עם פרומפט מתאים. בהתבסס על התשובה מרכיבים מודל עם וקטורי Z עבור המשימה הנבחרת. השיטה השנייה היא לאמון מודל דיסקרימינטיבי המזהה מה סוג המשימה עבור שאלה נתונה. השיטה השלישית מניחה דاطהסט קטן עבור משימה מסוימת למאפשר אימון של וקטור המשקל עבור כל המודלים (עבור כל המשימות). ככלומר במקום לשער שאלת למשימה מסוימת מתארים אותה כצירוף לינארי בין כל המשימות. בסוף המשימה מקבלת את הייצוג שלה (במציאות וקטורי Z בלבד).

<https://arxiv.org/abs/2501.06252>

13.06.25 המאמר היומי של מיק: Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps

המחברים של המאמר זה עשו משהו די לא שגרתי בדמיין של תמונות. הם שואלים שאלה פשוטה לכואור: נניח שכבר אימנו מודל דיפוזיהמצוין, האם אפשר להפיק ממנו יותר בזמן הרצה? אם ניתן בשלב ההיסק לשפר את איכות התמונה הנוצרת מבלוי פשוט להוסף עוד ועוד צעדי denoising? התשובה של מחברי המאמר, מסתבר, היא כן. אבל בדרך לשם עוברת דרך מהלך מקורי לגמר: חיפוש אחר רעש טוב יותר.

מי שעבד עם מודלי דיפוזיה ידוע שבסופו של דבר, התהיליך כולל מתגלל קדימה מתרוק וקטור רעש ההתחלתי. הוקטור זהה נבחר בדרך כלל בצוורה אקראית, ונראה שכל מה שצריך ממנו זה להיות "רעש לבן". אבל מה אם לא כל רעש נברא שווה? מה אם אפשר לבחור רעש "חכם יותר" – זה שידדר את המודל לתמונה איכותית יותר, אפילו בלי לשנות את הארכיטקטורה, את מספר השלבים או את משקליו של המודל?

המאמר מציע בדיקות אלה: במקומות להמשיך להאריך את מסלול הדיפוזיה (כלומר להגדיל את מספר השלבים), נוכל להשיקיע את אותו תקציב חישובי בחיפוש סלקטיבי אחר רעש ראשוני שמניב תוצאה טובה יותר. מדובר בשינויי קונצפטואלי די משמעותיים: אנחנו מפסיקים לחשב על denoising כציר השיפור היחיד, ומתחילה לראות את הסטוכסטיות עצמה ככלומר את הרעש, ממשו שאפשר לא רק לדגום ממנו אלא לכון אותו.

כדי שהי יעבוד, צריך שני דברים: קודם כל, דרך למדוד איכות של תוצאה סופית. זה קוראים במאמר זה verifier וזו יכולה להיות פונקציה כמו CLIPScore, aesthetic predictor, FID או כל מדד איכות אחר שתואם את מטרת הדגימה. המרכיב השני הוא אלגוריתם חיפוש כלומר דרך לבחור או ליצור רעים חדשים, להשווות ביניהם לפי הפלטים שהם יוצרים, ולמצוא רעש שנותן תוצאה טובה יותר.

המבנה הזה, של verifier לצד אלגוריתם(שיטה) חיפוש, הוא הליבה של החדשות כאן. מדובר במסגרת גנרטית מספקijk כך שהוא לא תלויות בארכיטקטורה של מודל הדיפוזיה, ולא דורשת fine-tuning. כל מה שצריך זה פונקציית דירוג, ויכולת להריץ כמה דוגמאות. מכאן פשוט מתחילה לחפש.

ההיפוש יכול להיות פשוט כמו ללקחת 64 רעשים ולבחר את הכי טוב. אבל הוא גם יכול להיות מתחכם יותר, למשל, לנסות וריאציות על רעש נתון בכיוונים רנדומליים (שיטת שנקראת Optimization Zero-Order, או אפיון להויסף רעש רק בחלק מהשלבים, ולאתחל מחדש את תהליכי הדיפוזיה ממקום אחר במסלול (מהלך שהם קוראים לו Search-Over-Paths). במילים אחרות, מדובר במקרה לא רק על שיפור אינטגרט, אלא על גישה חדשה להבנת המסלולים שמודל דיפוזיה צעדים בהם והאופן שבו הרעש משפיע עליהם.

אבל אולי המרכיב העמוק ביותר של המאמר זהה הוא מה שהוא לא מנסה לעשות. הוא לא טוען שצורך לשנות את המודל. הוא לא טוען שצורך לשפר את הרשות או לאמן אותה טוב יותר. כל החידשות שלו טמונה בהכרה שעצם ההחלטה באיזה רעש להתחיל היא פרמטר אקטיבי בזמן ריצה. זה חשוב, כי עד כה ההתקדמות בשלב היסק של מודלי דיפוזיה הייתה מוגבלת: ניסו רק ל��ר אותו, לשפר את מסלול השחזור (הסתה רעש) אבל לא הגיעו ברעש ההתחלה. המאמר הזה מפרק את ההנחה זו.

במובן מסוים, מדובר כאן על הוכנת אלגוריתמייה לשלב שאמור להיות פסיבי השלב שבו המודל כבר קיים ואנוונו רק "MRI'IM OTTO". אבל ברגע שאנחנו מקבלים את הרעיון שאפשר לעשות אופטימיזציה בזמן הריצה על פרמטרים כמו הרעש, אנחנו פותחים דלת לא רק לשיפור תוכאות, אלא להבנה עמוקה יותר של המנגנונים הפנימיים של דיפוזיה.

ולכן לדעתי, התרומה המרכזי של המאמר זהה אינה בgrafica עצמה או אחר של FID. היא בשינוי החשיבה שהוא מגלם: מודלים סטטיסטיים שמתפקידם כקובסה שחורה, למודלים שבהם הסטטיסטיות עצמה נעשית ניתנת לשילטה, לאופטימיזציה, ולעיצוב מחדש בזמן אמיתי. האם זו תהיה פרקטיקה רוחות? אולי רק במקרים יקרים מואוד של גנרטוט, שביהם כל שיפור קטן שווה הרבה. אבל כקונספט, זה צעד נוסף בהפיכת היסק מסתטית לאינטיליגנטית, וזה מהלך מרתק בפני עצמו.

<https://arxiv.org/abs/2501.09732>

14.06.25 המאמר היומי של מייק: Is Stochastic Gradient Descent Effective? A PDE Perspective on Machine Learning Processes

המאמר הוא די כבד אבל ניסיתי להנגיש את הסקירה כך שתיהיה מובנת (אם אני לא צלلت עמו מדי שם - המאמר באמת מורכב).

יש שהוא מתעטו בפשטות של SGD או Stochastic Gradient Descent בקצרה. כבר שנים שהוא הליבר של למידת מכונה(ML), ובמיוחד של למידה عمוקה, אבל התשובות לשאלת למה הוא בעצם עובד נותרו בגדיר אינטואיציה לא מספקת. נדמה שככל ניסיין להסביר את ההצלחה של SGD חוזר בסופו של דבר לאמרות מעורפלות כמו "הוא מוצא מינימום שתווחות" או "הרעש עוזר לצאת ממינימום מקומי". המאמר שאני סוקר היום, מנסה לעשות סדר ובשונה מרוב העבודות בתחום, הוא מציע זווית חדשה למגרי: הוא מתאר את SGD כתהליכי דיפוזיוניים שמתפתחים בזמן, דרך דרך של מושוואות דיפרנציאליות חלקיות (PDEs).

המחברים מבקשים לשנות את הדרך שבה אנחנו מבינים את הדינמיקה של למידה. לא עוד מעקב אחרי נקודת במרחב המשקولات שמתגלגת בתוך משטח LOSS (loss landscape), אלא תיאור מלא של ההתפלגות ההסתברותית של כל האפשרויות כלומר צפיפות (במהלך תהליכי הלמידה) על פני המרחב, שמתפתחת בזמן. אתם מגעים בתחום הפיזיקה המתמטית, זה יזכיר לכם מיד את משוואות פוקר-פלאנק, שמתארת איך חלקיקים

נעימים ונפזרים במבנה נתונה. הרעיון כאן הוא דומה: המשקولات הם כמו חלקיקים, והם נעות על פי הגרדיאנט של פונקציית לוס, עם קצת רעש שנובע מהאופין הסטטיסטי בו(בחירת מני-באצ'ים) של SGD.

מה שמעניין הוא שהמודל הפיזיקלי הזה הוא לא רק שהוא מחקה את מה ש-SGD עשו, אלא מראה מדוע הוא מצליח. למשל, כאשר מסתכלים על האנרגיה הקינטית של המערכת, רואים שהרעש האקראי שנובע מהסתוכויות של הבחירה במני-באצ'ים לא סתם "מוסיף רעש" אלא מחק תפקיד קרייטי ביציבותו: הוא מażן את ההתקדמות כך שלא נגloss מהר מדי או ניתקע במקומות לא יציבים. המחברים ממש מראים כיצד יש מגבלות אנרגטיות שמקטיבות את הקצב שבו אפשר ללמוד, וקשרות בין כמות הרעש לבין עומק הירידה באובדן.

יש כאן גם הבדיקה מושגית חדה בין שתי גישות להבנת תהליכי למידה: הגישה הлокאלית שמנתחת את התקדמות הפרמטרים בכל צעד, לבין הגישה הגלובלית שמattaרת את כל ההתפלגות, צרימה מתמשכת של הסתרות במרחב המשקولات. בדיקן כמו בפיזיקה, המעביר מתיאור נקודתי לתיאור מבוזר מגלה תובנות שהו נסתורות קודם. מתאים לפחות לא רק لأن המשקولات הולכות, אלא אףה הם מרווחים, איך הם מרווחים, ואיך המבנה של פונקציית הפסד משפיע על זה.

אחד החלקים המרשימים במאמר הוא הניתוח של רקורסיה בזמן. הכותבים לא מסתפקים בכך ש-SGD מתקכו, אלא בוחנים איך המבנה החוזר של תהליך הלמידה, המבוסס על חזרה עקבית דרך שיפוע הפונקציה, מתכתב עם הדינמיקה הרציפה של הפתרון למשוואות הפיזיקליות. דוקא ההשוואה זו בין תהליך רקורסיבי עם צעד זמן דיסקרטי לבין תהליך דיפוזיה רציף מאפשרת לנוכח לראשונה עקרונות כלליים על האפקטיביות של SGD: מתי הוא מצליח, מתי הוא עלול לסתות, וכיוצא ניתן לשלווט בהז.

אבל מה שהכי תפס אותי הוא שהתמונה הזאת פותחת דלת לפיתוח עתידי. אם מקבלים את הפרדיגמה ש-SGD הוא לא רק תהליך חמדני שנע לפני עצמו, אלא מערכת פיזיקלית שמתפתחת לפי חוקים דיפרנציאליים אפשר להתחילה לתכנן אופטימיזציות חדשות מtower אותו עולם מושגים. אולי לא צריך לשפר את SGD כמו שהוא, אלא עברו-*PDE-guided training*, שבו מתאים ישרות את האבולוציה הרציפה של ההתפלגות, וпотרים אחריה כדי למצוא את הדינמיקה.

במונח הזה, המאמר הזה לא רק מסביר את העבר של SGD, אלא מציע עתיד חדש ללמידה עמוקה. עתיד שבו אנחנו פחות מ>Showcases בתוך משטחים מרובי ממדים, יותר בונים מודלים דינמיים עם מבנה פיזיקלי מובהק. זה לא פחות מאשר שינוי תודעתינו (ואולי גם פרקטנו) שיכל לשנות את הדרך שבה ניגשים לאופטימיזציה כולה.

<https://arxiv.org/abs/2501.08425>

15.06.25 המאמר היומי של מייק: Random Teachers are Good Teachers

מאמר עתיק אך מאד מעניין לדעת... .

מאמר זה מציג ממצא מעניין ונוגד אינטואיציה באופן עמוק, המאתגר הנחות יסוד בתחוםים של זיוק ידע (knowledge distillation) ולמידה בפיקוח עצמי (self-supervised learning) או SSL). המחברים מציגים כי מודל "סטודנט" יכול ללמד ייצוגים איקוטיים על ידי זיוק ידע מרשת "מורה" שהמשקولات שלה(המורה) אקרניות לחלוטין ואין מאומנות. העבודה מפרקת את "מערכת היחסים המורה-סטודנט" הסטנדרטי כדי לבדוק ולחקרו דינמיקה למידה עם זיוק ידע, וחושפת כי התהליך הדומה לרגולריזציה לא מפורשת (chossing), שאינה תלולה בכך שהמורה מחזיק ב"ידע" ממש קלשהו.

כאמור המטרה העיקרית של המאמר לחקור את דינמיקת זיקוק ידע. המאמר בגדול בודק שני מושגי זיקוק ידע:
עם DATAה מתייג ובל' DATAה מתייג (לא תוויות).

ליבת תרומתו של המאמר טמונה במרק הניסוי הפשט והאלגנט שלו. המחברים יוצרים תרחיש שנועד להסביר גורמים מפעריים (confounding factors) שונים שבדרך כלל מיוחסת להם ההצלחה של שיטות זיקוק -LSL.

- היעדר "ידע אפל" (Dark Knowledge): רשות המורה מאותחלת באמצעות ביצועים אקרטיות ולאחר מכן "מקפאת". היא לעולם אינה נחשפת לדאטה האימון או לתוויות, כלומר היא אינה מכילה שום מידע נלמד כלשהו על המשימה או על התפלגות הדאטה. מטרת הסטודנט היא פשוט למצער את מරחיק KL בין התפלגות הפלט שלו לבין הפלט הסטטי והאקראי של המורה (אבל לפעמים מסוימות לוס של הסטודנט על הדאטה)
- היעדר אוגמנטציה דאטה (Data Augmentation): בגין לשיטות רווחות -LSL, עבודה זו מסירה במקוון את כל אוגמנטציות מהדאטה. הדבר מבטיח שהאיננו ריאנטיות הנלמדת אינה נבעת מהתיוות מובנות (inductive biases) מפורשות שימושיות על ידי טכניקות כמו חיתוך (cropping), היפוך (flipping) או שינוי צבע (color jittering).
- היעדר תוויות (Labels): כל תהליך הזיקוק מתבצע ללא פיקוח (unsupervised) ולא תוויות. התוויות האמיתיות של הקטגוריות משמשות רק בסוף התהליך כדי להעריך את איכות הייצוגים הנלמדים באמצעות בדיקה לינארית (linear probing) ככלומר אימון מסווג לינארי על גבי הייצוגים הקפואים ממקודד הסטודנטן.

מסגרת מינימלית זו מבטיחה שככל אפקט למידה שנכפה ניתן לייחס אך ורק לאינטראקציה בין ארכיטקטורת המודל, התפלגות הדאטה הטבעית ודינמיקת האופטימיזציה מבוססת-הגראדיינט של מערך המורה-סטודנט.

תוצאותיו של ניסוי הין מאד מפתיעות. רשות הסטודנט משיגה באופן עקבי ומשמעותי ביצועים טובים יותר מהמורה האקראי שלה במונחים של דיוק בבדיקה לינארית, וזאת על פני דאטה סטטיסטיים רבים כמו (CIFAR-100, ResNet, VGG, TinyImageNet, STL10) וארכיטקטורות שונות כגון (Linear probing, Deep learning, GANs, etc.).

מציאה נוספת במאמר היא "טופעת הלוקאליות" (locality phenomenon): הkrbeta ההתחלה של משלימות הסטודנט לאלו של המורה היא קריטית למידה מוצלחת. המחברים חוקרם זאת על ידי אתחול משלימות הסטודנט כצירוף קמור של משלימות המורה ומשלימות אקראיות, הנשלט על ידי פרמטר לוקאליות α. כאשר α קרוב לאפס (כלומר, הסטודנט מתחילה כמעט זהה למורה), הלמידה היא מהירה ביותר וביצועים הסופיים הם הגבוהים ביותר (כאן הסטודנט הוא באותו ארכיטקטורה של המורה - זה לא תרchiש פרקטי אך מעניין לחקריה).

ממצא זה מرمץ על גיאומטריה מעניינת של משטח הלו. הפרמטריזציה של המורה, θ , מהווה מינימום לוקאלי טריויאלי שבו לוס הזיקוק הוא אפס. עם זאת, תהליך האופטימיזציה לא נשאר שם. במקום זאת, הוא מוצא מינימום לוקאלי סמור ולא טריויאלי, S_θ , המתאים לאזורי עם דיוק גבוה בהרבה (עבור דאטה אימון, ככלומר, ייצוגים טובים יותר). הדמיות של הנוף חושפות כי המורה ישב לעיתים קרובות בתווך "עמק-א-סימטרי" חז. נראה כי מודל הסטודנט נמלט מהפרטן הטריואלי על ידי תנופה לעבר הצד ה"שטווח" יותר של עמק זה, אזור שהגיאומטריה שלו ידועה כבעלט מתאים להכללה (generalization) טוביה יותר.

אול', הממצא העמוק ביותר הוא שצ'קPOINT של הסטודנט, שפותחה כולה ללא תוויות (רק זיקוק ידע), מציגה ייצוגים מבנים שב עבר סברו כי הן מופיעות רק בשלבים המוקדמים של אימון מפוקח. לעומת זאת התלמיד מתקרבת למורה (גם אקרה) כאשר יש "בתוכה כרטיס זוכה" - תת-רשות קטנה הידועה לעשות את אותו הדבר.

- הופעת (Lottery Ticket Hypothesis): המחברים מצאו כי כבר צ'קPOINT של 1 של הסטודנט מכיל "כרטיס לוטו זוכה" תת-רשות דיללה שנייה לאמן מחדש משקלותיה ההתחלתיות כדי להשיג דיוק גבוה במשימה מונחית. לרשות המאוחלת באופן אקרה אין תכונה זו; היא מופיעה ברשותות מונחות רק לאחר מספר אפקטי אימון. הדבר מרמז כי זיקוק למורה אקרה מנהה את הרשות לתוצרה פרמטרית שכבר מובנית למידה עילית.
- קישוריות מצבים לינארית (Linear Mode Connectivity): בד"כ כאשר משתמשים בצ'קPOINT מוקדם של הסטודנט catastrofically רירות אימון מונחות (כל אחת עם מיני-באצ'ים שונים), הפתרונות המתקבלים בד"כ הם "מקושרים לינארית". משמעות הדבר היא שיתן לבצע ייצוגים מותחניים (כמו "כרטיס לוטו המשקלות בין כל שניים מהפתרונות הללו מוביל לקבל לוט גבוי בדרך". יציבות זו מחייבת על כר שהסטודנט כבר התקנס ל"างן רחב ושטוח" במשטח הלוט "הפקח", ובכך עוקף למעשה את השלב הכספי הראשמי של אופטימיזציה מפוקחת).

מסקנה

המאמר טוען שהצלחתן של מסגרות מורה-סטודנט אינה מיוחסת אך ורק להעברת "ידע אפל" למורה מאומן. במקום זאת, המאמר חושף כי הרגולריזציה הלא מפורשת הנוצרת מדינמיקת הלמידה היא מנוע רב-עצמה ללמידה ייצוגים חזקים בפני עצמה. על ידי הדגמה שراتה לפתח ייצוגים מותחניים (כמו "כרטיס לוטו זוכה") מאות אקרה לחלוין, המחברים מאלצים הערה חדש של המנגנונים הבסיסיים מאחוריו זיקוק-עצמם ולמידה -SSL. העבודה מספקת מצע ניסויים לעובדה עתידית שטטרת להסיר את המסתורין מעל "השלב המוקדם" של אימון רשותות נוירוניים והגיאומטריה המורכבת של משטח הלוט שלהם.

<https://arxiv.org/abs/2302.12091>

המאמר היומי של מיק: 16.06.25

Evolutionary Computation in the Era of Large Language Model: Survey and Roadmap

בנוף המתפתח במהירות של AI, שתי פרדיוגמות: LLMs ואלגוריתמים אבולוציוניים (EAs) פועלן לעיתים קרובות במקביל, כשהן מפגינה יכולות אדריכלות בתחוםיה. מודלי השפה הדהימו אותנו עם יכולות היצירה שלהם והבנת השפה הטבעית, בעוד אלגוריתמים אבולוציוניים הוכיחו באופן עקבי את כוחם כוחם בעיות אופטימיזציה וחיפוש מורכבות, תוך חיקוי של מנגנונים אבולוציוניים של הטבע. אבל מה קורה כאשר שני הכוחות העצמאתיים הללו מתחילה לשתף פעולה?

המחברים מסווגים את היחסים הללו לשני כיוונים עיקריים:

אלגוריתם אבולוציוני משופר על ידי מודל שפה (EA-LLM-enhanced): כאן, LLMs מנוצלים כדי לשפר היבטים שונים של אלגוריתמים אבולוציוניים. דמיינו LLM המיצר באופן דינמי אוכלוסיות ראשוניות מגוונות ורלוונטיות יותר לאלגוריתם אבולוציוני, או יוצר פונקציות התאמת מתוחכמת ומודעתה להקשר, שקשה לתקן ידנית. LLM יוכל לשמש כ"סוכן להבנת בעיות", המפרש תיאורי בעיות מורכבים כדי להנחות את החיפוש של האלגוריתם האבולוציוני, או אפילו כ"מנגן תיקון", המתקן פתרונות לא חוקיים שנוצרו על ידי EA.

אתחול חכם: LLMs יכולים ליצור נקודות התחילת מגוונות וmbtichot, המכוננות את EA העוזרת לו להתקנו לפתרון טוב.

אופרטורים אדפטיביים: תכונן אופרטורי הכלאה או מותציה דרש ליעדים קרובות מומחיות בתחום. LLMs יכולים אולי לייצר או לחدد אופרטורים אלו תוך כדי תנועה, בהתאם על הקשר הבעיה.

הנדסת פונקציית התאמת: ייצרת פונקציות התאמת אפקטיביות היא קשה להפלי. LLM יכול לשיע בתרגומים יעדים ברמה גבוהה לממדים מסוימים או אפילו לייצר קוד להערכתה.

הסבר ויכולת פרשנות: לאחר ש-EA מוצא פתרון, LLM יוכל לייצר הסברים קריאים לבני אדם של למה הפתרון זה טוב או איך הוא התקבל.

מודל שפה משופר על ידי אלגוריתם אבולוציוני (EA-enhanced LLM EA): אלגוריתמים אבולוציוניים יכולים להביא את יכולות האופטימיזציה החזקות שלהם לטובות LLMs. אימון LLM הוא יקר חישובית ומסתכם במידה רבה על ירידיה בגרדיינט, שיכולה להיתקע באופטימום מקומי. EAs הידועים ביכולות החיפוש הגלובליות שלהם וביכולתם לנוט במרחבים שאינם ניתנים לגזרה, מציעים חלופה או השלה מסקרנת:

אופטימיזציה פרומפטים: EAs יכולים לפתח פרומפטים יעילים יותר עבור מודלי שפה גדולים, ולגלות ניסוחים עדינים המפיקים תשובות מעולות למשימות ספציפיות. זה חורג מהנדסת פרומפטים פשוטה, ומאפשר גילוי אוטומטי של פרומפטים אופטימליים.

כוון היפרפרמטרים: שלל ההיפרפרמטרים של LLMs (שיעור למידה, גגלי מיני-באץ', בחירות ארכיטקטוניות) יכולים להיות מותאמים באמצעות EAs, מה שעמיד להוביל למודלים חזקים ויעילים יותר.

חיפוש ארכיטקטורה עצבי (NAS): ל-EAs היסטוריה ארוכה ב-NAS ביחס ל-LLMs, הם יכולים לגלוות ארכיטקטורות חדשות, יעילות יותר או מיוחדות, במיוחד מודלים קטנים ומצוצמים יותר.

העשרה וארגון נתונים: אלגוריתמים אבולוציוניים יכולים לפתח אסטרטגיות לבחירה או ייצירה של נתונים אימון שייעיל באופן מksamילאי לביצוע מודלי השפה הגדולים, תוך התמודדות עם מחסור או בעית איות נתונים דатаה.

חיסון והגנת מפני התקפות אדווורסיות: EAs יכולים לשמש לייצרת דוגמאות יRibot לבדיקה ושיפור רובוטיות של LLMs, או לפיתוח מנגנוני הגנה.

המאמר מדגיש "שיטת סינרגיה משלבות" על פני תרחישים מגוונים, ומציג את ההשלכות המעשיות של שיתוף הפעולה זהה. הם נוגעים ב:

יצירת קוד והנדסת תוכנה: דמיינו LLM המיצר קוד ראשוניים, ולאחר מכן EA המיעל את הקוד הזה לביצועים, יעילות, או אפילו הפחיתה באגים. לעומת זאת, אלגוריתם אבולוציוני יכול להציג שיפורים במבנה הקוד, ומודל שפה גדול יכול לבצע רפקטורינג לקוד בהתאם על הצעות אלו.

חיפוש ארכיטקטורה עצבי (NAS): זו ההתאמה טבעיות, שכן אלגוריתמים אבולוציוניים שימושו זה מכבר לגילוי ארכיטקטורות של רשתות ניירונים. שילוב זה עם מודלי שפה גדולים יכול להיות שמודל שפה גדול יציע מוטיבים ארכיטקטוניים ראשוניים, אותן אלגוריתם אבולוציוני יפתח וישכל.

משימות ייצרה שונות: מעבר לקוד, חשבו על כתיבה יצירתיות, עיצוב, או אפילו גילוי תרופות. LLM יכול לייצר רעיונות או מבנים ראשוניים, ו-EA יכול לאחר מכן ליעל אותם לפי קритריונים ספציפיים (למשל, חדש, עיקבויות, יעילות), מה שיוביל לתפקידות חדשות באממת.

המאמר עושה אובייסרבצייה הבאה: ככל ש-*s*-LLMs הופכים לנפוצים, הבנת האופן שבו ניתן להפוך אותם לחזקים, בעליים וחכמים יותר, וכיitz לתרום לפתרון בעיות מורכבות, היא עלות חשיבות עליונה. הסקירה מספקת עד יסודי מכריע, המפרק באופן שיטתי תחום מפותח ומורכב. חזונו של המחברים בזיהוי אתגרים והצעת מפת דרכים הוא בעל ערך רב במיוחד, ומונחה חוקרים לעבר דרכי מבטיחות ביותר.

עם זאת יש אתגרים משמעותיים בדרך:

עלות חישובית: הפעלת *As*, במיוחד עבור משימות מורכבות, עלולה להיות יקרה בצורה בלתי רגילה. כיצד נהפוך את השילוב הזה ליעיל?

אי התאמה ביצוג: גישור על הפער בין האופי הדיסקרטי של השפה (כפי שמצוול על ידי *MLL*) לבין המרחבים הרציפים והמספריים הנחקרים לעיתים קרובות על ידי *EAs*, אינם טרייזיאלי.

יכולת פרשנות של הסינרגיה: כאשר *MLL* ו- *EA* משתפים פעולה, ההבנה מדוע הוושג פתרון מסוים הופכת אפילו יותר מעורפלת.

הגדרת "אופטימלי": עבר בעיות יצירתיות או מורכבות רבות, הגדרת פונקציית התאמה מדוקת ל-*EA*, אפילו בסיוו *MLL* נותרת אתגר.

למרות המורכבות הזאת, החזון המפורט במאמר זה מرتק ללא ספק. הוא מציע עתיד שבו מערכות *AI* לא רק מסוגלות לייצר טקסט קוגורנטי או למצוא פתרונות אופטימליים, אלא יכולות לומוד באופן מושכל כיצד ללמידה, ללמידה כיצד לבצע אופטימיזציה, וללמידה כיצד לייצר באופן אוטונומי ומתחוכם בהרבה.

מה שהמאמר הזה באמת מדגיש הוא המעבר מהמעבר להתייחסות-*LLMs* כקופסאות קסם מבודדות. הוא דוחף לתפיסה הוליסטית של *AI*, שבה ניתן לשלב פרדיגמות שונות, כל אחת עם יתרונותיה הייחודיים, כדי להתגבר על חילשות של כל אחת מהן. חישוב אבולוציוני מציע ל-*LLMs* דרך לבסוף מאופטימום מקומי, לחזור מרחבים(של משקولات וארכיטקטורות למשל) שהיו חסומים לנו, ולהציג אינטיגנץיה כללית יותר בתקווה. *MLL*, בתורם, יכולים להעניק ל-*EAs* חשיבה ברמה גבוהה יותר, ידע בתחוםים שונים ויכולת לפעול על ייצוגים מופשטים וסמנטיים.

<https://arxiv.org/abs/2401.10034>

18.06.25 המאמר היום של אביב ומיק: Harnessing the Universal Geometry of Embeddings

לפני כחודש הופיע המאמר, וישר היכה גלים. הוא בונה על מאמר אחר, [The Platonic Representation](#), שגם-כן היכה גלים בזמןנו, ומתימר להזקז משמעותית, במידה מפתיעה. ועל הדרך הדריך גם מדים איך למן את ההישג המסקין-תיאורטי לכדי פריצה משמעותית של חילוץ מידע. גובה הגלים היה תוצר של כל אלו - בשילוב עם כתיבה שמעודדת קריאה בOMBESTIT * מדוי* של מה שהמאמר בעצם מראה. וכך נעשה קצר סדר.

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

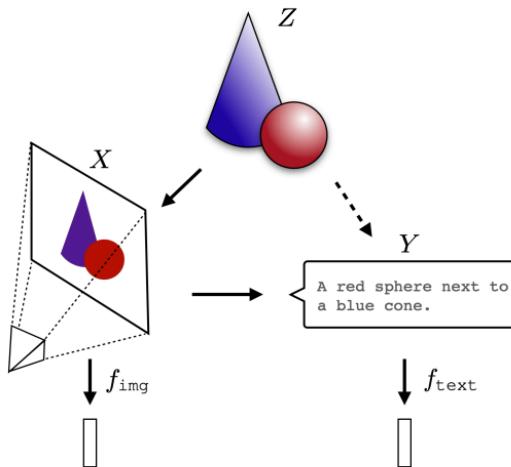


Figure 1. The Platonic Representation Hypothesis: Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z , and scaling model size, as well as data and task diversity, drives this convergence.

דатаה, בין עם טקסטואלי, תומונתי, או אחר, מגע בסופו של דבר מהתරיל'ר זהה שנקרא המציאות. ככל שהמודלים הגדלים מאומנים על יותר מデータה, יותר מוגן, עבור שימושות רבות ומגוונות - היצוגים שלהם יטו יותר ויותר להתכנס אל אותן המציאותות המשותפות שמאחורי הדבירים, אל המרחב הלטנטי "אמית" שמאפשר את ההיסקים האופטימליים. זו הייתה הטענה אותה המאמר המקורי קידם - והשתדל להדגים, באמצעות מגוון מדדים והשוואות. עד כאן אז.

המאמר החדש בא לטעון טענה לכורה חזקה יותר, ו"קונסטרוקטיבית":
ניתן ללמד את המרחב הלטנטי האוניברסלי זהה עבור "יצוגי טקסט", ולמעשה לרטום אותו על-מנת "לתרגם" מרחב ייצוג אחד לשני - ללא DATAה המאפשר הצלבה (כלומר שקיבלו את הקידודים שלו משני הצדדים), ולמעשה ללא גישה או שום ערך אחד מן המודלים, רק לדוגמאות הקידוד שלו.
או אז, בהינתן היכולת לתרגם מהמרחב של מודול לא ידוע אל אחד בשבייטטמו, ניתן לגלוות זהה האחרון תכונות על מידע שקיים במודול הלא-ידעו, ואף לשזרו באמצעות טכניקות "היפוך-שייכון" קיימות.

אז איך כל זה עובד?
ראשית כל, בואו נחדר את שתי הנקודות בהן המאמר חוטא בניסוחו "הבטחת-יתר" (שבתקווה יתרענו כייתגשש בקיור בィקורס העמיטים האקדמיים):
1) המאמר המקורי דיבר על מרחב משותף, *יחיד*, שמאחד בין כל המודלים, המציאותות שמאחורי מגוון השתקיפות שלה. זהו לב העניין התיאורטי שהצדיק רפרנס פילוסופי שהולך אחורה 2400 שנה.
המאמר בו עסק אין עכשווי, לעומת זאת, לא* משייך ייצוג יחיד שכזה. שיטתה למידת הייצוג שהוא אכן, כפי שנראה בהמשך, מוגשת רק בין כל *צמד מודלים ספציפי*. זה בהחלט בכיוון, אבל עוד לא ממש שם.

2) המאמר המקורי דיבר על התאמה (alignment) בין modalities (ועוד), כמו למשל בין שמות אובייקטים (במודלי טקסט) לתמונות שלהם (במודלים ויזואליים) (ראו תמונה). זה הרבה יותר עמוק ומשמעותי מהתאמה פשוטה בין מודלי טקסט שונים - בהם עוסק המאמר החדש. (הוא אמנם נוגע גם ב-CLIP, אך זהו בפרט מודל "יצוג לטקסט", שלא-פי בנייתו כבר מראש בא מותאם גם מול דאטה תומונית, אין באמת מה ללמידה מכך בהקשר שלנו).

ועל כן אחרי שכיבינו את להבות ה- $\text{h}(\cdot)$ המוגזם, באו נצלול אל הפרטיהם והדברים שיוצאים מהם, שכן מעניינים בפני עצם. אז כפי שאמרנו, המאמר בונה מיפויים מרחב אմבידיג X למרחב אמבדינגן Y . הוא עושה זאת באמצעות חמשה מיפויים, המציגים באמצעות מודלים מאומנים:

- מיפויים A_1 ו- A_2 המapseים X -ו- Y למרחב אמבדינגן משותף Z בהתאמה
- מיפוי T מיישר את האמבדינגן אחרי A_1 ו- A_2 לייצוג לטנסי משותף m_Z של X ו- Y בהתאמה
- מיפויים B_1 ו- B_2 המחשירים את האמבדינגן m_Z ל X ו- Y בהתאמה

על גבי אלו ניתן להגדיר גם את:

- מיפויי "תרגום" $A_2 \circ T \circ B_1 = F_1 = B_2 \circ T \circ A_1$, $F_2 = T \circ A_2$ - המapseים מרחב אמבדינגן מקורי X למרחב השני Y ובכיוון ההפוך בהתאמה
- מיפויים עצמאיים $R_1 = B_1 \circ T \circ A_1$, $R_2 = B_2 \circ T \circ A_2$ המתווך עצם (X -ו- Y) דרך מרחב אמבדינגן משותף באמצעות T .

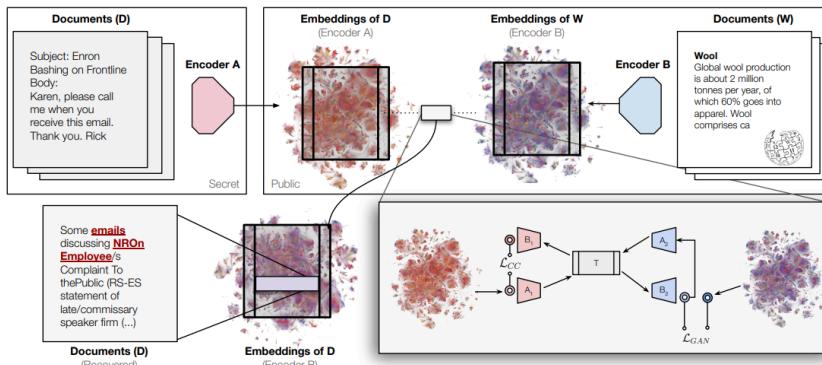


Figure 2: Given only a vector database from an unknown model, vec2vec translates the database into the space of a known model using latent structure alone. Converted embeddings reveal sensitive information about the original documents, such as the topic of an email (pictured, real example).

אחרי שהגדכנו את שורת המיפויים הארכאה, נסביר את מבנה פונקציית הלווי המלאה. היא מורכבת מכמה לוסים. הלווי הראשון מנשה לכפotta על התפלגות המיפויים X -הממופים דרך F_1 ל- Y להירותם כמו המיפויים X - Y עצמוני. למטרה זו משתמשים בגאנן (algo) שללטו ללא עורין בתחום גנרטוניות (פוני מודלי דיפוזיה). הגאננים (GANs) מאומנים שני מודלים בו זמנית עם לוסים מנוגדים: המודל המגנרט (F1) מאומן לעשות את המיפויים X - X מאד דומים לאלו X - Y , כאשר המודל השני (D1) מאומן לבדוק בין מיפויים X -לאלו אחרי F1. בסוף (אם התחילה מתכנס) מקבלים מודל גנרטיבי חזק (F1) המסוגל "لتת פיט" למודל דיסקרמיןטור חזק (D1), וכן עושים גם L_2 עם D2. ד"א המאמר משתמש בגישה הקלאסית לגאנן מהמאמר של גודפלאו מ-2014.

בנוסך מאומנים עוד שני גאננים לייצוגים מהמרחב הלטנטי הממוופים X וגם Y . כלומר המודל הגנרטיבי במקרה זה הוא $A_1 \circ T \circ A_2$ שמאומן באופן שמתואר בפסקה הקודמת. גאנן נוסף מאומן עבור $A_2 \circ T$. בסוף, ארבעת הלוויים שתארנו מהווים את החלק הראשוני של פונקציית לוס המלאה.

החלק השני של הלווי הוא לוס "שחזר" המודוא שכל אמבדינגן המועבר X -לחלופין X -ו- Y למרחב המשותף יודיע לחזור אל עצמו אחרי מיפוי B_1 (מיפוי B_2). החלק השלישי של הלווי הוא לוס "עקביות התרגום" (cycle consistency loss) שדואג שהיצוג שהגע מ- X -ל- Y (לחילופין X -ל- Y) עם F1 (F2) חוזר לתווך עצמו ואם

מפעילים עליו את התרגום ההפוך F2 (ו-F1 בהתאם). החלק האחרון של הלוס דואג שיחסים בין זוגות אמבידיגס שונים מ-X (לחילופין מ-Y) יישמרו בתרגום ל-Y (ל-X).

בסוף דבר, הלוס כולל את כל הלוסים המתוארים לעיל. כך אנחנו לומדים את הייצוג המשותף (שוב, ללא שום דוגמאות שונות לנו קידודיהן משני הצדדים!). באופן מרשים, המיפוי הנלמד למעשה מכיל גם להתפלגות טקסטים מאוד אחרות, כך שהגישור שהושג כאן הוא די כללי. אבל לפתרים אלה כמו גם האפליקציה של הטכניקה לגילוי מידע, נשאיר משהו למאמר עצמו):

<https://arxiv.org/abs/2505.12540>

20.06.25 המאמר היומי של אביב ומיק: Evolving Deeper LLM Thinking

המאמר מציג שיטה לשיפור ביצועי מודלי שפה בזמן אינפראנס(test-time compute). השיטה ממירה את בעיית החיפוש במרחב של פתרונות טקסטואליים לתהיליך אבולוציוני מונחה-ביקורת, שנבנה כולל סביבה יכולה לגנרטיבית וה"רפלקטיבי" של ה-LLM עצמו. אין כאן משלכות fine-tuning או עדכון מושךות המודול אלא השיפור מתבצע באמצעות החישובי של ה-inference בלבד.

הנחת המוצא היא שביעיות רבות, כגון תכנון מסלול טiol או לוח זמנים לפגישות, לא ניתנות לפורמליזציה מלאה, אך כן ניתן לבדוק את אפשרות הפתרון בעזרת פונקציית הערכה חיצונית. זה יוצר תרחיש שבו לא ניתן ליצור פתרונות ישירות על ידי אופטימיזציה מסורתית, אך כן ניתן לבצע חיפוש מונחה-הערכתה. המאמר מבצע זאת באמצעות מנגן גנטיאי שמתממש ככל בשפה טבעית.

רכיב האלגוריתם 1: אוכלוסייה טקסטואלית

כל פתרון מיוצג כטקסט כלומר תיאור מילולי של תוכנית פעולה. המרחב שבו מתבצע החיפוש אינו מרחב וקטורי ואינו בעל מבנה טופולוגי ברור. אין מרחק מוגדר בין שני פתרונות, ואין דרך לקבוע "כיוון שיפור". השיפור מבוצע באמצעות recombination(מושג מהמאמר) לשוני, כלומר כתיבה מחדש של טקסט על בסיס טקסטים קודמים.

רכיב האלגוריתם 2: מבנה אבולוציוני עם איים

במקום אוכלוסייה אחת, האלגוריתם מחלק את מרחב הפתרונות למספר אוכלוסיות נפרדות הנקראים איים במאמר. כל אי עבר תהיליך אבולוציוני עצמאי, אך כל כמה איטרציות מתבצעת "הגירה" של פתרונות מוצלחים בין האיים. כך נשמר איזון בין חיפוש מקומי (exploration) לחיפוש גלובלי (exploitation).

רכיב האלגוריתם 3: בחירה מבוססת סלקציה רכה

הבחירה של אילו פתרונות ישמשו הבסיס(הורם) לדור הבא אינה דטרמיניסטית. האלגוריתם בוחר פתרונות עם הסתברות שתלויה באיכותם, אך לשמור גם סיכוי לבחירת פתרונותBINONIIM, כדי למנוע התכנסות מוקדמת. זה יוצר מנגן של סלקציה רכה שמאפשר לאוכלוסייה לשמור על גיוון מבני וריעוני(קצת דומה לSTS MCT). אבל בלי עצים).

רכיב האלגוריתם 4: recombination באמצעות שיח ביקורתי

במקום לבצע **recombination** באמצעות תהליכי סינטטיים כמו דילוג על שורות או חיבור משפטיים, האלגוריתם מייצר שיח פנימי בין שני ישויות קונספטואליות, מalker ומחבר, אשר לומדים מהפידבק של פונקציית הערכה. התוצאה היא טקסט חדש, שלא בהכרח בניו כשלוב כלשהו של פתרונות קודמים, אלא כפרשנות חדשה עליהם. תהליך זה חוזר על עצמו מספר פעמים בכל דור.

התהליך כולו מסתמך על פונקציית הערכה חיצונית שיכל להיות קוד, תוכנה או מודל נוסף שמספקת גם ציון איקות וגם פידבק טקסטואלי מפורש. חשוב להציג: המשוב אינו בהכרח מספרי בלבד, אלא יכול לכלול תיאור מפורט של תקלות או סטיות מהאילוצים, מה שמאפשר למודל להשתמש בו כחומר גלם לרפלקסיה.

פתרונות מבניים

- סקלබליות לביעות לא מוגדרות היטב: לאחר והאלגוריתם פועל על טקסטים ולא על מבנים פורמליים, ניתן להפעילו גם כאשר אין תיאור פורמלי של הבעיה.
 - הפרדה בין גנרטס לאבלואציה: בעוד לגישות המבוססות על התקדמות ליניארית כמו **Reflexion Chain-of-Thought** או **Chain-of-Thought**, כאן יש חלוקה ברורה: המודל מייצר, ההערכה בוחנת, ואז מתבצע רה-קונפיגורציה של הפתרון.
- מניעת התוכנות מוקדמת: בזכות האלים, ההגבלות הרכבות, וה-reset התקופתי, מנעת קריישה מוקדמת לפתרונות לוקליים.

האלגוריתם מאפשר ל-**LMs** לחשב לעומק לא דרך ניתוח סמנטי או לוגי של השפה, אלא דרך דינמיקה של תחרות, ביקורת, רפלקסיה והתרמה. זהו תהליך חישובי שמשתמש בשפה עצמה כחומר גלם לבניית פתרונות, ומוביל לשיפור איקוטי של היכולות התכנוניות של המודל גם במצבים שבהם לא ניתן להגדיר מראש את מהות "הפתרון הנכון".

אם נביט בזה כתשתית רעיונית, המאמר מציע גישה כללית **-meta-reasoning** של מודלים: מערכת שמאגרנת את החשיבה של המודל לא רק דרך פרומפט אלא דרך שילוב של רעיונות מתחרים בהכונת ביקורת. מדובר בתפיסה לא ליניארית של אינפרנס, כזו שמניחה שמחשבה טובה נולדת לא בבת אחת, אלא דרך אקספלורציה, שגיאות, ותיקון מצטבר.

<https://arxiv.org/abs/2501.09891>

21.06.25 המאמר היומי של מיק:

Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

המאמר זה לא חדש אבל מתאים גיליתי שהתחלה לסקור אותו בקובץ דוקס נידח ונתקלתי בו בצורה די אקראיית. תוך כדי חיפוש בערוץ הטלגרם שלי גיליתי שעשיית סקר(בסוף ינואר) ורוב המנויים (יותר מ 85%) רצו שאסקור אותו. מקיים את ההבטחה הפעם בDALI של 5 חודשים.

המאמר אימן מודל מולטימודלי לשפה ולתמונות. להבדיל מרוב העבודות בתחום המאמר מציע להפריד בין הבניה של טקסט וקלט ויזואלי לבני גמרוט של טקסט ותמונות. לעומת המחברים מאומנים 3 מודלים שונים (אמנם עם רכיבים מסווגים) לבני וGENERATE של טקסט, הבניה וGENERATE טקסט עבור תרחישים מולטימודליים והשלישי עבור GENERATE של תמונות. הבניה כאן הכוונה>KIDUD של קלט למחרב יוצר וקטורי משלו ואדפטר המיפה אותו למרחב

הlatinnti של מודל שפה L שהוא backbone של L.Janus. יש עוד 2 מודלים קטנים (heads) הממפה את הפלט של מודל שפה L לפני הפיכתו לトーונים (של השפה ושל התמונות).

מאומנים את Janus על מגוון MERCHANTABILITYS כמו הבנת התמונה וכל האובייקטים בה, דיאלוג בהתבסס על מה שופיע בתמונה, ייצור תמונה מקלט טקסטואלי וקלט ויזואלי כמו ערכת תמונות על בסיס קלט טקסטואלי ועוד. אצין שהאחי הנקודרים והאדאפטרים היוצרים שלהם מזינים למודל שפה גדול (שהוא גם מאומן מהשלב השני של האימון של Janus).

יש 3 שלבים עיקריים באימון Janus. המטרה העיקרית של שלב הראשון היא ליצור "חיבור" מושג בין רכיבים(מודלים) ויזואליים לשפטים בטור מרחב האמביג, כך שמודל שפה יוכל להבין את היחסיות המוצגות בתמונות ולפתח יכולת ראשונית ליצירת תמונות. בשלב זה אנו משארים את מקודדי התמונה ואת ה-LLM, ומאמנים את הרדאפטרים (עbor הטקסט ועbor התמונות) וגם מודל הראש (head) עבור המודל לגנרטט טקסט.

בשלב זה אנו מבצעים אימון מאוחד על גבי קורפוס מולטימודלי, כדי לאפשר ל-Janus ללמידה גם גנרטוט מולטימודלית. באופן פרקטני מאומנים את כל הרכיבים של Janus חוץ מאשר שני אנקודרים: השפטי והטקסטואלי. בשלב האחרון אנו עושים פין טין למודל המאומן בעזרת DATAה מובוסי הנחויות, במטרה לחזק את היכולת לעקוב אחר הנחויות ולנהל דיאלוגים כדי להבטיח שמודל Janus יהיה מיומן גם בהבנה וגם בגנרטוט מולטימודלים, הם לא מאומנים מודלים נפרדים לכל שימושה. במקומות זאת, המחברים משתמשים בשילוב של DATAהסיטים של דיאלוגים טקסטואליים בלבד, DATAהסיטים של MERCHANTABILITYS הבנה מולטימודלית וכאלו של גנרטוט של TAMONOT מtekst, כדי להבטיח גמישות ב嚷ון תרחישים.

<https://arxiv.org/abs/2410.13848>

המאמר היומי של מ"יק: 25.06.25

The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs

מאמר

תפנית מעניינת מתרכחת בתקופה האחורה בעולם של הערצת יצועי מודלים. אנחנו כבר לא שואלים רק עד כמה המודל מצליח ב מבחון כלשהו, אלא שאלת מהותית יותר: האם ניתן ל證明 על מודל שפה שיחליף מתייג אנוש? זו לא שאלת שמדדים מסוימים כמו DIoK, F1 או הסכמה בין מתייגים יכולם לענות עליה כראוי. תחת זאת, המאמר שנסקרו היום מציג שיטה מבוססת סטטיסטיקה לפתרון בעיה זו. בלב המאמר עומדת קריאה להתרחק ממדדי התאמת שטחים, וubahor לנימוקים מבוססי השערות סטטיסטיות וניתוח עלות-תועלות.

התרומה המרכזית של המאמר היא שיטה חדשה בשם מבחן המתיאג האלטרנטיבי (Alt-Test). השיטה הזאת לא בודקת האם המודל מסכים עם רוב המתיאגים או עומד ברף DIoK כלשהו. במקום זאת, היא שואלת שאלת عمוקה יותר: האם המודל עקיב יתור עם קבוצת המתיאגים מאשר מתיאג אנושי מסויע? כך זה עובד: בכל פעם מוציאים מתיאג אחד מן הקבוצה, ומשתמשים בשאר קבוצת ייחוס. המודל והמתיאג שהוצע נבדקים לפי מידת ההתאמת שלהם לקבוצה הנונטרת. אם המודל עקיב יותר מהמתיאג שהושמט הוא "מנצח" אותו. התהילה эта חוזרת על עצמה עברו כל מתיאג.

החדשש כאן הוא שמדובר בשיטה שאינה דורשת כלל תוויות אמת (gold labels). היא גם עיליה בדגימות קטנות; אפשר להשתמש בה עם שלושה מתיאגים וכמה عشرות דוגמאות בלבד. אבל אולי החשוב ביותר: היא מספקת הכרעה ביןארית לומר האם המודל יכול, סטטיסטי, להחליף את האדם? לא "כמה טוב הוא היה", אלא האם יש הצדקה להשתמש בו במקום מתיאג אנושי.

כדי לאחד את תוצאות ההשוואות הללו בין המודל לכל מתיג, המאמר מציג מدد שנקרא שיעור הניצחון א. זהו פשוט אחוז המתיגים שהמודל ניצח לפני מבחן האלטרנטיבי. למשל, אם המודל טוב יותר מ-4-5 מתוך 6 מתיגים, שיעור הניצחון שלו הוא שני שלישים ואם הוא טוב רק מ-2 מתוך 6, השיעור הוא שליש. המאמר קובע כלל הכרעה ברור: אם שיעור הניצחון גבוה מ-50% אז המודל טוב יותר ממתיג טיפוסי, ולכן ניתן להחליפו במודל באותו הקשר. שיטה זו לוקחת השוואות סטטיסטיות ומתרגם אותן למידיניות פעולהית. היא גם מכירה בשונות בין מתיגים, ולא מניחה שכולם שקולים. אלה מנגנון ברור, מבוסס דатаה ואמין, שמאפשר קבלת החלטות על בסיס תצפיות ולא תחושת בטן.

אבל יש שאלת נוספת: אם אני מחליף בני אדם במודל איזה מודל עלי' לבחור?-can מציג המאמר מدد נוסף, רציף יותר, בשם הסתברות היתרון המומוצעת (ק). הרעיון פשוט: עבור כל מתיג אנושי, בודקים מה הסיכוי שהמודל עיקבי יותר עם שאר המתיגים ממנו. לאחר מכן מחשבים את ממוצע ההסתברויות האלה. הערך הסופי מתקובל כמו: מה הסיכוי שהמודל טוב לפחות כמו מתיג אנושי אקדמי. המدد הזה משמשו מכמה סיבות:

- הוא רציף ודווחס, ולא "קופצני" כמו שיעור הניצחון (שיכול לקבל רק כמה ערכים בלבד).
- הוא אינו תלוי ברכף שරירות, ולכן מתאים במיוחד להשוואה בין מודלים.
- הוא כללי, ומתאים לכל סוג המשימות: סיוג, דירוג, ואףלו הפקת טקסט חופשי.
- הוא אינטואיטיבי וברור להסביר, גם לקהילים שאינם סטטיסטיים.

אבל מעבר לכך הוא מתמודד ישרות עם מה שמדודים קלאסים מתעדמים ממנו: השונות בין בני אדם. בעוד דיק או F1 מנחים שיש "אמת אחת", ק בוחן עד כמה המודל מצליח למכוד את התפלגות הדעות האנושיות. אלה ממד ראשון מסוגו שמכבד את המורכבות של שיפוט אנושי — ולא סתם מתקרב למוצע.

אחת התוספות החדשניות והאלגנטיות ביותר במאמר היא הפרמטר אפסילון (ϵ) שENCIL בתוכו את העבודה שמודלים זולים, מהירים ומדרגיים יותר מבני אדם, ולכן לא חייבים להיות טובים כמוני כדי להיות משתלים. אפסילון מייצג את הפער המותר בביצועים כמה פחות טוב המודל יכול להיות, ועדין להשתלים כלכליות. לדוגמה, אם ההשוואה היא למומחים יקרים נוכל להצדיק שימוש במודל גם אם הוא מעט פחות טוב. אבל אם מדובר מתיגים זולים המודל צריך להיות טוב יותר מהם בבירור.

באופן מעשי, המאמר ממליץ:

- להשוות מול מומחים עם ערך ϵ של 0.2 (כי הם יקרים)
- להשוות מול מתיגים רגילים עם ערך ϵ של 0.1 (כי הם זולים יותר).

בכך, אפסילון הופך את המבחן האלטרנטיבי מכך השוואתי טהור למסגרת פרגמטית לקבלת החלטות עסקיות. הוא מביא את כלכלת האנומציה לתוך עולם ההסקה הסטטיסטי ולא רק האם המודל טוב, אלא האם הוא מספיק טוב לאור מה שהוא חוסף לנו.

בקיצור מאמר לא רגיל בנוספ' המודרני ובקטע טוב...

<https://arxiv.org/abs/2501.10970>

26.06.25 המאמר היומי של מיקרו: Open Problems in Mechanistic Interpretability

איןטרפרטבליות מכנית היא אולי תחום השאפטני ביותר כיום להבנת איך בינה מלאכותית באמת עובדת. לא מדובר כאן בהסבירים בנפנופי ידים או בהדגשות צבעוניות של חלק טקסט אלא בהנדסה לאחרור(reverse)

engineering) של הרשות עצמן. הבנה אמיתית של איך רשות נוירונים פותרת בעיה: מהם החלקים הפנימיים שפועלים, באיזה סדר, באיזו לוגיקה, ואיך בדיקתם מייצרים הכללה.

הגישה המרכזית שモוצגת במאמר מבוססת על שלושה שלבים: פירוק הרשות לרכיבים קטנים (בין אם אלו נוירונים, תתי-מרחבים או מעגלים), תיאור התפקיד הפונקציוני של כל אחד מהם, ואיום כולם בדיקה האם ההסבר שלנו באמת חוזה התנהגות, ואם כן עד כמה. כל אחד מהשלבים האלה מתגלה כקשה הרבה יותר ממה שנדמה.

הבעיה הבסיסית היא שפירוק לפי מבנה הארכיטקטורה של הרשות כולם שכבות, נוירונים, ראשי attention פשוט לא עובד. החלקים האלה לא מתאימים למה שהרשות באמצעות מחשבת. נוירונים הם פוליסמנטיים (רבים ממשמעים), תפקדים מתפרשים על פני שכבות שונות, ותכונות לא שכוכנות בוקטור בודד אלא מקודדות כוספרפוזיציה של וקטורים רבים. השיטות הקלאסיות כמו PCA ו-SVD נכשלות, לא בגלל יישום לקוי אלא בגלל הנחות תאורטיות שגויות.

הכלי המרכזי כיום הוא Sparse Dictionary Learning ובעיקר Sparse Autoencoders. הרעיון הוא לאמן רשות קטנה ש"תפרק" את האקטיבציות של הרשות הגדולה באמצעות בסיס דليل של "תכונות". אלו הליטנטים. אך בפועל, השיטה אמונה מוצאת כיוונים מעוניינים, אך לא מסבירה איך החישוב עצמו מתבצע. הליטנטים הם תמונה סטטיסטית של "מה הופעל" ולא תיאור של האלגוריתם שמיושם.

יש גם בעיות מוחותיות: הפרע בין האקטיבציות האמיתיות לשיחזור גדול. המידע הגומטרי בין תכונות הולך לאיבוד. ההנחה שהכל לינארי רחוכה מלהיות נכונה. והגרוע מכל היא העובדה שאין בכלל תיאוריה פורמלית שמסבירה מהי "תכונה", איך היא נוצרת, ומה הופך אותה ליחידה בסיסית של הבנה.

מכאן עולה כיוון חדשני: אולי הדרך הנכונה היא לא לפרש מודלים אחרי שאומנו, אלא לבנות מודלים שאפשר להבין מראש. מודלים עם אקטיבציות דיסקרטיות, אכיפת מודולריות, פונקציות הפעלה דילולות כמו k-Top או LReLU, או מבנים כמו Mixture-of-Experts שמחולקים את החישוב לתת-מודולים בורורים. המטרה היא ליצור רשותות שנבנות "חתוכות מראש" עם פרשנות לא כניתוח מאוחר אלא כהנחה יסוד של האימון.

גם תיאור הפונקציה של רכיב בודד הוא משימה קשה. למשל דוגמאות שפעולות אותו יכולות להיות מבלבלות. שיטות ייחוס מבוססות גרדיאנט בעיותן תאורטית ופרקטית. סינתזה תכונות בנייה קולט שפעיל רכיב עללה לייצר דימויים לא אינפורטטיביים. השיטות המבטייחות ביותר הן אינטראונציית סיבתיות: שינוי של ערך פנימי, בוחינה של ההשפעה על התנהגות החיצונית. כאן ננכדים לתמונה גם steering k-gram המדרה של כיוון ספציפי למרחב האקטיבציות וגם שימוש lens logit כדי לפענה השפעה ישירה על תוצאות על אקטיבציות הרשות.

הבעיה הגדולה היא שהרבה מההסבירים נשמעים משכנעים אך לא עומדים בבדיקה. הם לא חוזים קונטרפקטואלים (לא מצלחים לבגא מה היה קורה אילו משחו היה שונה בתוך המודל), לא עוזרים לאבחן כשל מודל, לא מאפשרים תיקון או שיפור בפועל. לכן המחברים מציעים סט של דרכי אימוט: האם ההסבר חוזה התנהגות אחרי ablation? האם ניתן לבנות מודל קטן שמאפשר לבדוק אם ההסבר נכון? האם הליטנטים (יצוגים פנימיים כמו תכונות או רכיבים חישוביים) מסוימים במשימות בטיחות כמו זיהוי תוכן מזיק? האם נכון להשתמש בהסבירים כדי לשנות את התנהגות המודל?

המאמר מציע גם ליצור "ארגוניים מודליים" שהם מהווים רשותות קטנות סטנדרטיות, עם מבנה פתוח, שאפשר לאמן שוב ושוב ולבדוק עליהם שיטות פרשנות. כמו שהביולוגיה התקדמה דרך עבודה על תיסנית, כך תחום זה זקוק לרפרנס קבוע. זהו כל תשתית חיסר כיום.

החלק האחרון של המאמר מבHIR שמכנים אינו עניין טכני בלבד. הוא נוגע למדיניות, לניטור, לבטיחות, ולשאלות פילוסופיות: מה נחשב הסבר טוב? איך אפשר לחבר בין המבנים המיקורוסקופיים לתפקיד גלובלי? אילו עקרונות כללים ניתן לחלץ מרשומות שלמדו לפתור בעיות טוב יותר מבני אדם?

בסיום, מדובר במאמר לא מתביש לומר את האמת: אין עדין תיאוריה מספקת לפירוק רשותות. ההנחות הלינאריות שברירות. התוכנות לא חיות בלבד אלא ארכיטקטורת על. הפרשנות חייבת לקשור מבנה לתפקיד. והדרך קדימה, אולי, עוברת לא דרך דרך "דזין" חדש של הרשותות...

<https://arxiv.org/abs/2501.16496>

המאמר היומי של עמרי ומיק 27.06.2025
Agent-as-a-Judge: Evaluate Agents with Agents

כולם כבר מכירים את הקונספט LLM-as-a-Judge זהה אומר להיעזר במודל שפה גדולים כדי לבחון מודל שפה אחרים. צוות Meta מציג כאן חלופה שאפתנית יותר: (AAJ) Agent-as-a-Judge, תפיסה שבה סוכן מבצע אבלואציה לSOCIALISTS אחרים ומספק משוב עשיר ברמת הצעד, לא רק פסק דין סופי.

אחת התוצאות המרכזיות של המאמר היא AI DevAI זהה דאטאטס שהמחברים בנו מאפס: 55 שימושות פיתוח AI מורכבות יחסית מקצה-לקצה, ש郿וקמות לתת דרישות עבור כל הערכה של ממשמה, סה"כ 365 דרישות. DevAI נולד כתגובה לפער בبنצ'רים קיימים, שרובם מסתפקים במידע "עברית/נשלה" סופי; הפירוק היסודי לדרישות-משנה נדרש לחשוף את הבאים והכשלים שתרחשו באמצעות תהליכי הפיתוח. זה החלק שבו SOCIALISTS נכשלים הכי הרבה, אך כמעט שלא נמדד עד היום כי שהם טוענים.

שלושה "SOCIALISTS-מתקנת" פופולריים, OpenHands-GPT-Pilot, MetaGPT, (נקון לאוקטובר 2024) קייבו לפתור את כל המשימות. כאן נקרא AAJ: הוא עצמו סוכן עם חמישה כלים או modules כמו שהם קוראים לזה במאמר ask, graph, read, locate, retrieve

graph – יוצר גרף תלות בין קבצים ופונקציות ורק מבין אילו רכיבים משפיעים זה על זה.

read – קורא ומנתח את תוכן הקבצים והלוגים כדי לבדוק אם המימוש עומד בדרישות.

locate – מאתר במדiator את שרויות הקוד או השגיאות הרלוונטיות שמסבירות את הכשל או ההצלחה.

retrieve – שולף קטעים רלוונטיים ממילויי הרצתה ארכיטים כדי לגבות את החלטה במקרים של חוסר ודאות (כמן ב-RAG).

ask – מקבל את ההחלטה הסופית האם הסוכן עמד בדרישה או לא ומספק נימוק קצר.

הקשרים האלה מאפשרים לו לחטט בקוד שיוצר הסוכן הנבחן, לבנות גרף תלות (dependency graph) בין הקבצים, לאתר שרויות שגיאה ולשלוף תיעוד רלוונטי, ורק אז לקבע אם הדרישה הושלמה. כדי שתהיה אמת-מידה אנושית, שלושה מת'גים מומחים דירגו כל דרישת בונפרד, אחר כך עשו majority-vote ולבסוף דיוון שהוליד קונצנזוס לגבי על דרישת בונפרד. הסכמה ראשונית נעה סביבה 70%-90%; אחרי דיוון היא התייצבה על ערך 95% - זהו ה-ground truth שמלוי מודדים את כל השופטים.

AAJ נבדק בשתי רמות מידע:

- **Black-box**: שופט עיוור, רואה רק את הקלט והפלט. מדמה-מציאות המחמיר.
- **Gray-box**: מקבל גם את הלוגים וקבצי הקוד, אך שהשיפוט קל וمبוסס יותר.

ב-JA JA Black-box משתמש כמעט במדויק למתיג אנושי ממצוע, בעוד Sh-Judge-LLM-as-a-Judge-LLM נשאר אחר בפער ניכר. במצב JA JA Gray-box מתקרב עוד יותר להסכמה האנושית ומיצמצם את המרחק עד לכדי אחזים בודדים, אך שהוא כבר טוב כמעט כמו מתיג אנושי כל זה בזמן Sh-Judge-LLM-as-a-Judge-LLM רחוק מהמתיג האנושי.

כאשר מסתכלים על עלות וזמן התמונה חדה: אבלואציה ידנית גובה אלפי דולרים ונמשכת ימים, בעוד Sh-Judge-LLM-as-a-Judge-LLM מבצע זאת בדקות ספורות ובעלות מזערית אף מקריב לא מעט דיוק. JA JA מוסיף רק קמץ של זמן וכמעט ביחס ל-LLM, ומהזיר כמעט במלואו את רמת האיכות של האבולוציה האנושית. ניתוח האבולואציה מראה שככל שימושים JA JA את יכולות read, graph-locate, הוא מתקרב יותר ויותר לrama של בני אדם.

המחברים מבקרים שהdagmota עדין מוגבלות לעולמות ג'ינרט קוד, וublisher לתוצאות אחרים יציריך בדיקות נוספות. הם גם מצינים כי לsocion-השופט יש שכבת זיכרון ותכנון מורכב זהה עלול להשתבש בנסיבות יחסית - המנגנון של האגנטים מאדן עדין ולכן שינוי פרטים קטנים כגון פרופומטים עלולים להוביל לפגיעה רצינית בו (במנגנון).

למרות זאת, הם מציעים לראות JA JA בסיס לquo מחקר חדש של למידה חייזקית מונחית-תהליכי: במקום RLHF Sh-masternet על התוצאות שמבצעים על ידי בני אדם, אפשר לדמיין אופן שמחפש פידבק תהליכי עשיר משוב של מערכת האגנטים. JA JA מבחן את הטעויות, מזרים את הפידבק בחזרה, וסוגר את לופ השיפור בעלי יד אדם.

לטיכום, גישת agents grading agents מוכיחה שאפשר להשיג רמת דיוק אנושית כמעט ללא עלות אנושית, תוך קבלת תובנות מפורטות בהרבה מהשיטה המסורתית של Judge-LLM-as-a-Judge-LLM. למי שמחפש פידבק תהליכי עשיר לצד חסוך בזמן ובcosa, JA JA הוא צעד ממשמעותי קדימה.

<https://arxiv.org/abs/2410.10934>

29.06.2025 המאמר היומי של טדי מיק

In-Context Symbolic Regression: Leveraging Large Language Models for Function Discovery

היום יש לנו מאמר טיפה ישן (בן שנה) אבל שהזדקן ממש טוב בינהיים.

המאמר מציג את (ICSR) In-Context Symbolic Regression, גישה חדשה המשמשת ב-LLMs לפתרון בעיות רגסיה סימבולית (SR - symbolic regression). רגסיה סימבולית היא בעיה שבה מתקובל נתונים טבלאי בדרך כלל) ואנו חנו מתקבשים להציג משווה אנליטית המתארת דатаה זה. הבעיה עצם מרכיבה את בעיית הרגסיה הקלאסית, כמו רגסיה ליניארית לדוגמה, בכך שהיא לא רק מוצאת את המקדמים הכי מתאימים אלא גם את מבנה הפונקציה עצמה. בהמשך לדוגמא של פונקציה ליניארית, שם אנחנו מנהים עם מבנה הפונקציה הוא, ובכן, לינאר. במקום לבנות מודלים "יעודיים", ICSR ממנפחת את יכולות הלמידה בתוך הקשר של LLM כדי להציג ולשפר צורות פונקציונליות באופן איטרטיבי. לדוגמה במחקרים פיזיקליים, ה-LLM יודע לצרכים להתחשב ביחסות מידתית באופן אימפליסי וכן ידע לא להציג פונקציות שאינן פולינומיות למשתני הקלט. חידוש זה מאפשר למצוא משוואות פשוטות ומודיקות יותר בהשוואה לשיטות קיימות, ואף להקליל אותן טוב יותר לננתונים חדשים. הגישה גמישה מאוד בזכות הידע המדעיים שmagically מתחילה האימון מbasos על כמויות ענקיות של טקסטים עם

משוואות, התיאור שלhn, והדעתה ששומש כדי ליצור אותם, ומשתפרת עם התקדמות ה-LLMs ללא צורך באימון נוספת של SR. כמובן, בהקשר זהה, LLM מושתמש בתור meta-learner של משוואות אנלטיות מדatta טבלאי.

השיטה פועלת בשני שלבים עיקריים. בהתאם, נדרשות פונקציות התחלתיות (Seed Functions). בשלב הראשון, ה-LLM מקבל קבוצה של תכיפות (נקודות נתונים) ומתקבש לייצר אוכסיה ראשונית של פונקציות מעמדות. במקום להסתמך על פונקציות מוגדרת מראש כמו סינוס וכאליה, ה-LLM מייצר את הפונקציות בעצמו, מה שMOVEDIL בדרך כלל למגוון רחב ומורכב יותר של פונקציות. התהיליך הזה מבוצע מספר פעמים כדי להתמודד עם פונקציות לא מוגדרות עבור נקודות קלט מסוימות (נניח חא לערכים שליליים).

בשלב השני, המחברים משתמשים בollowat אופטימיזציה הבאה. בכל איטרציה קוראים כמה דברים. תחילה, הזרת הקשר (In-Context Learning): ה-LLM מקבל קלט "meta-prompt" (meta-prompt) המכיל את התכיפות כלומר זוגות (Y,X), וכן רשיימה של הפונקציות המועמדות הטובות ביותר מהתוצאות הקודמות יחד עם ציוני התאמת fitness scores (שלhn). ההנחה היא שה-LLM יכול להסיק דפוסים מהדוגמאות הללו ולהציג פונקציה חדשה וטובה יותר.

לאחר מכן ה-LLM מייצר רק את הצורה הפונקציונלית ("שלד") של הפונקציה (למשל, " $c + bx^2 + ax$ "), מבלי לקבוע את המקדים המופיעים. בסוף, המקדים הבלתי ידועים של הצורה הפונקציונלית שהוצאה על ידי ה-LLM מותאמים לנתחים באמצעות אופטימיזציה חיצונית של בסגנון ריבועים פחותים לא לינאריים (Non-linear Least Squares). למה לא מבקשים מה-LLM גם את הערכים האלה? כי הוא פשוט זאת, שימוש באופטימיזר חיצוני מבטיח ערכי מקדים טובים יותר ומאפשר חקירה יעילה יותר של מרחב הפונקציות. התהיליך חוזר על עצמו עד שהתקציב החישובי נגמר כי בוואנו נודה באמת, כמה אתם מוכנים למצאו משואה בסופה של יום.

גישה ICSR נבדلت מושיות SR מבוססות טרנספורמר אחראות בכך שהוא לא דורשת אימון מוקדם על נתונים SR סינטטיים גדולים, אלא מסתמכת על הידע המתמטי הקיים ב-LLM המאומן מראש. בנוסף, ICSR מציגה ממשק בשפה טבעית, מה שמאפשר לה לחזור למגוון רחב יותר של פונקציות.

התוצאות של ICSR הן בעלות חשיבות עצומה שכן הן מציגות פריצת דרך בגישה הרגשית הסימבולית. הן מוכיחות Sh-LLMs, שאינם אומנו במיוחד למטרה זו, יש את יכולת לזהות ולנסח פונקציות מתמטיות לא רק בבדיקה גבוהה, אלא גם בפשטות אלגנטית, תוך שמירה על יכולת הכללה יצאת דופן לנתחים שטרם נצפו. יתרון זה של פונקציות פשוטות אך כלויות הוא קריטי ביישומים מדעיים והנדסיים, שכן הוא מאפשר הבנה عمוקה יותר של התופעות הנחקרות ומונעת התאמת יתר לנתחי האימון. כי בוואנו נודה באמת, אתם שמחים אם יבנו את המטווים הבא שלכם בעזרה איזה DLML אבל בסוף אתם רוצים שהמנדס בין את הפיזיקה של הכלים שהוא ופה SR נוון עבודה חבל על הזמן.

מה גם, שבuidן ש-LLM כבר כתבים מאמרי חci בעצמו, SR סוג פינה חשובה בהקשר זהה כי SR יכול לשמש ככל, עצמאי לגילוי חוקים פיזיקליים חדשים, ניתוח מושוואות כימיות המתארות תהליכי מורכבים, בניית מודלים ביולוגיים, או דיהוי קשרים כלכליים שודושים הסבר.

למי שמתעניין בליצר משוואות בעזרת תיאור וקטת נתונים, שווה לקרוא:
<https://arxiv.org/pdf/2404.19094>

המאמר היומי של טדי מיך 01.07.2025

DINO-WM:World Models on Pre-trained Visual Features enable Zero-shot Planning

חווץ לסקור מאררים בראיה ממחשבת משולבת עם למידה באמצעות חיזוקים או RL. המאמר מציע גישה לאימון של מודל עולם (model world) לשימושים בעולם הרובוטיקה. ככלור המחברים מציעים גישה המאפשרת ללמידה איר לגראם לרובוט לבצע פעולות מסוימות בהתקבוס על התיאור היזואלי של הסביבה (קרי תМОנות).

המודל WM-DINO מציג גישה חדשה ל-*World Modeling* (בנייה מודל של העולם) על ידי הפרדה בין למידת דינמיקה חזותית (המתוארת על ידי ייצוג לטנסי של הסביבה) לבין שחזור פיקסלים ישיר ואופטימיזציה של תגמולים תלוי-משימה. החידוש המרכזי טמון בארכיטקטורה ובשיטת האימון של המודל, המנצלות תכונות חזותיות מאומנות מראש כדי לאפשר תכנון אפס-שוט (*zero-shot planning*). בגודל WM-DINO מציע לאמן את מודל החיזוי של הצעד הבא במרחב הלטני כאשר הדקORDER שמחזר את הפיקסלים מהייצוג הלטני מאמן בנפרד.

מודלים מסורתיים של העולם נאבקים לרוב בעלות החישובית של חיזוי במרחב הפיקסלים, או במગבות של מודלים לטנטטיים הקשורים למטרות של שחזור תמונה. WM-DINO פותר את הבעיה על-ידי פעולה מלאה במרחב לטנטי קומפקטי המאמן מראש. הוא משתמש ייצוג פאצ'ים שמקורו מ-2vDINO, בתור מודל התצפית שלו (תמונה). זהו שני מנגנון עומת עבודות קודמות שבן מודל התצפית נלמד מאפס, לרוב בתלות במשימה. שימוש במקודד של 2vDINO שאינו ניתן ללמידה מאפשר ל-WM-DINO להנות מהציגות עשירות של אובייקטים ומרחבים שנלמדו ממאגרי מידע עצומים באינטראנס. זה הופך את מודל התצפית לבלי תלי במשימות וסביבות.

מודל המעברים בטור WM-DINO, הבניי על-גבי טרנספורמר חזותי (ViT), חוזה את ייצוגים העתידיים של הפאצ'ים ולא את הפיקסלים עצם. החיזוי מתבצע במרחב הלטני, תוך התניה על היסטוריה של מצבים ופעולות. רכיב טכני מרכזי כאן הוא יישום של מנגנון *attention* סיבתי בטור ViT. בכך יכול לגישות קודמות שמבצעות חיזוי אוטורגרטיבי ברמת TOKENים, WM-DINO חוזה ברמת פרǐם כלומר מתיחס לכל וקטורי הפאצ'ים של תצפית כאובייקט שלם. לפי המחברים, עיצוב זה קולט טוב יותר את המבנה הכלובלי והдинמייקה הטemporלית, מה שמוביל להכללה טמפורלית טוביה יותר. החיזוי מותנה גם בעולות הסוכן, שמצוות למד גובה יותר באמצעות MLP ומצוותות לכל וקטור פאצ'ים.

אחד החידושים הבולטים של WM-DINO הוא הניתוק המוחלט של הדקORDER (Decoder) מחייזי המעבר הבא. ניתן להשתמש בדקORDER על מנת לשחזר תמונות מצביים לטנטטיים לצרכי פרשנות, אך אימונו מנוטק לחילופין מודל המעברים. המשמעות היא שיכולת התכנון והדינמיקה הפנימית של המודל אין תלויות בשחזור פיקסלים, מה שmbיא ליעילות חישובית רבה יותר בזמן אימון ובזמן ריצה. הדבר שונה ממודלים שבהם חיזוי במרחב לטנטי מחובר לשחזור תמונה, מה שולול לפגוע באוניברסליות של ייצוגים הנלמדים שנדרשות לתיאור מוצלח של פיקסלים, במקום דינמיקה רלוונטית למשימה.

בזמן מבחון, תהליך האופטימיזציה של ההתנהגות מנוסח כבעית הגעה ליעד חזותי במרחב הלטני. לוס של תכנון (חיזוי מעברים) מוגדרת כሻיגאת ריבועים ממוצעת בין המצב הסופי החיזוי למצב היעד הלטני. היכולת זו לבצע תכנון zero-shot ללא הסתמכות על הדגם מומחה או מודלי תגמולים, נובעת ישירות מהיכולת של המודל ללמידה דינמית חזותית כללית רובוטית ביחס למשימה מתרוך לטנסי מאמן מראש.

היכולת של WM-DINO להקליל לקונפיגורציות חדשות של סביבות – כמו מבci אקרים או אובייקטים בצורות מגוונות – מדגישה עוד יותר את חידשו. ההכללה זו נובעת מהלמידה האפקטיבית של מושגים ודינמיות כללים בתוך ייצוגי הפאצ'ים הלטניים המאומנים מראש, מה שמחזית את התלות בביטויים שבריריים הנובעים מאמון על נתוני תלוי-משימה.

לסיכום, החידוש של WM-DINO טמון בשילוב בין מקודד פאצ'ים קפוא ומאמן מראש מ-2vDINO לצורך מודל התצפית, מודל מעברים מבוסס ViT ברמת פרǐם הפעיל כלו במרחב הלטני, והניתוק המוחלט בין הדינמיקה

הפנימית לשחזור הפיקסלים. הארכיטקטורה זו מאפשרת למידה של דינמיקה חזותית כללית, רוביוטית ביחס למשימות רבות, מנתוני אופליין בלבד ומובילה לתכנון shot-zero אפקטיבי והכללה חזקה בסביבות מגוונות.

<https://arxiv.org/abs/2411.04983>

04.07.25 המאמר היומי של מיק:
Investigating Tax Evasion Emergence Using Dual Large Language Model and Deep Reinforcement Learning Powered Agent-based Simulation

תפנית מפתיעה מתרחשת בשימוש ב-LLMs בתחומים "רכים" יותר כמו פסיכולוגיה, סוציאולוגיה ואפיון כלכלה. אומנם LLM לא "חובבים" כמו בני אדם ברמת האינדיבידואל, אבל מסתבר שהם כבר מוכנים את איך שאנו אנחנו מקבלים החלטות אוכלוסייה.

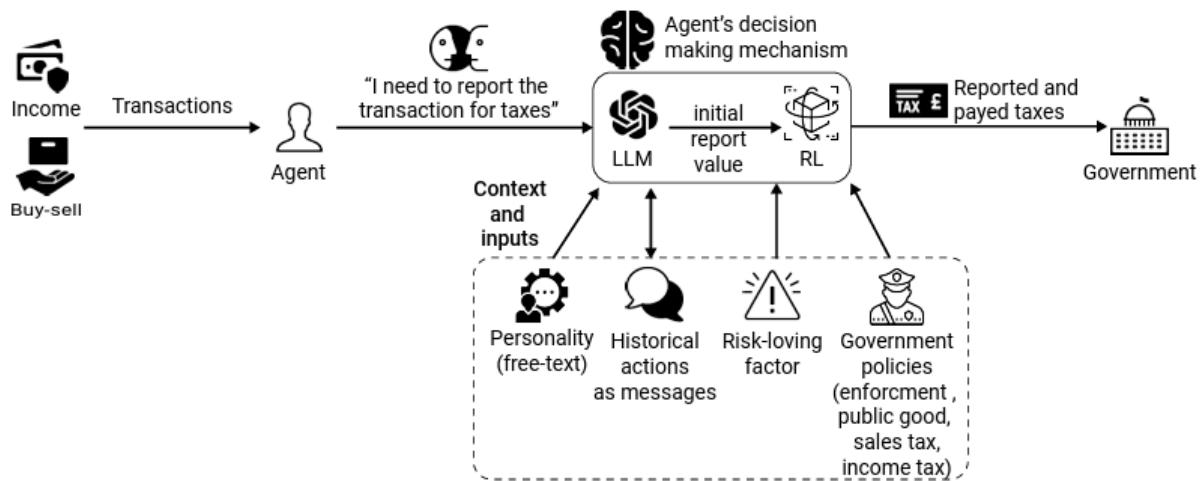
המאמר הנוכחי מציג גישה חדשה לשימוש ב-LLMs לחקור התהנחות ממש בעזרת סימולציה. במקום להניח מראש התהנחות של העלמת מס, כמו שעשו כל החוקרים עד היום, המחבר מתמקד בהופעה ובдинמיקה של תופעה זו בקרבת אוכלוסייה. באמצעות סימולציה מבוססת סוכנים (agent-based simulation) המשלבת LLM ולמידה عمוקה עם חיזוקים (deep reinforcement learning), החוקרים בונים מודל המאפשר להתנחות כלכליות בלתי פורמליות (מה שהרב מכירים בתור "כללה שחורה") להגיח באופן סופוני, ולא כהתנחות מוגדרת מראש. על אף התרומה שלו לככללה, מה שלכנראה מעניין אותנו זה השימוש הייחודי ב-LLMs וב-DRL כדי לשלב של מנגנון שיעוד לחתת מידע לא פורמלי (כמו טיפול אישיות) ולהכניס אותו לסת השיקולים הפורמלי של סוכן - למשל כמה אופציות של פרטונייזציה במערכות אפשר לעשות על גבי הרעיון הזה.

השיטה במאמר מתעמקת ביצירת סימולציה מבוססת סוכנים המדמה כללה סגורה (יש מסחר בין אנשים באוכלוסייה שמנסה להטיב עם מצבם), שבה סוכנים פעילים ומקבלים החלטות. החוקרים השתמשו במבנה כללה דומה לזו של ארצות הברית עם דיווח מס עצמי כדי לאפשר לסוכנים הזדמנויות להעלים מס החלטה. מכיוון שיש המון סוגים של מס זהה מסתבר מהר (תשאלו את רואה החשבון שלכם) במאמר זהה התרczzo בשני סוגים - מס הכנסה ומס ערך נוסף (מע"מ).

הם גם סימלו רשות החוק ותועלת שהסוכנים מקבלים מהמדינה מהמס שהם משלמים לה. הסימולציה עצמה מהוות תשתיית לחלק המרכזי של העבודה - מודל קובלת החלטות של הסוכנים באוכלוסייה. כדי לייצר אוכלוסייה הטרוגנית, מוח של סוכן מורכב משילוב של LLM ו-DRL כאשר LLM מקבלים תיאור של האופי של הסוכן (מבוסס על התוכן שהוא מפרסם בטוויטר למשל), היסטוריית הפעולות בטקסט, ואת כל המידע על הסימולציה בטקסט גם כן.

יחד עם ה-*context prompt* זהה ה-LLM נשאל "כמה מס אני צריך לשלם?". המספר שה-LLM מחזיר, מועבר כקלט למודל DRL שמקבל גם את שאר הדטה שקיבל ה-LLM אבל גם כמה הסוכן "הרפטקי" כפרמטרשה-DRL משתמש כדי לעשות אקספלורציה. כאמור, ה-LLM מחזיר החלטה ראשונית שאוותה, יחד עם הלקט של ה-LLM ועוד משתנה exploration (אהבת סיכון במון הכלכלי) מקבל גם DRL שמקבל החלטה בעצמו שהיא גם הסופית.

תהליך הלמידה הזה מאפשר להתנחות של העלמת מס ולפעולות כלכליות בלתי פורמליות "להגיח" באופן טבעי מטור האינטראקציות בין הסוכנים, במקרה להיות מוגדרות מראש כלליים קשיים. מה גם, שאפשר לראות שינוי משמעותית בהתנחות הרצינלית (DRL) על ידי שינוי מספיק אגרסיבי של הפלט של ה-LLM בעזרת שינויים כמו תיאור האופי של הסוכן.



גם אם אתם לא חובבי כללה גדולים, השיטה הזאת של שילוב בין LLM ל DRL בצורה שינה את ה DRL ולא הפוך (כמו שקרה באימון LLMs conversational) או שאתם בוחרים איזו תשובה יותר אהבתם של (chatGPT) פותחת את הדלת לכל מיני שימושים אפליקטיבים שלא היו כל כך נגישים לפני זה, כמו:

1. במקום רק לחזות תוצאות בחירות, אפשר לדמות איך דעתות מתפסות, איך קבוצות חברותיות נצורות או מתפרקות, או איך מתחפתת קיצוניות – לא מתרע כללי בורר אלא מאינטראקציות אנושיות מורכבות. אפשר לבדוק איך קמפני מסוים או חוק חדש ישפיע על התנהוגות אזרחים.
2. איך שינוי בנתיב תחבורה ציבורית או בניית שכונה חדשה תשפייע על דפוסי נסעה, פקסים, או אפילו על פיתוח עסקים באזוריים שונים, בגלל החלטות הדינמיות של תושבים ונוהגים.
3. איך חברות מgebוט למHALים של מתחרים? האם הן מתכונות לקראת קרטל או ננסות למלחמה מחרים? אפשר לדמות את השוק עם חברות "חכמות" שמקבלות החלטות אסטרטגיות ולראות מהן ההתנהגויות העסקיות המגויות.

בקיצור, זה לא רק על העלמת מס. זו דרך חדשה ויעילה לבנות מודלים לכל מערכת מורכבת שבה ההתנהגות הכלולית היא יותר מסcum חלקיה, ומושפעת מהחלטות דינמיות ולמידה של הפרטים בתוכה. זה נותן לנו יכולת "לשחק" עם המציאות, לבדוק תרחישים וללמוד מהם, בלי הצורך לתוכנת מראש כל פרט.

לא אמר קלואסי ממה שעולהפה בסקירה בדרך כלל, אבל יכול לפתוח את הראש:

<https://arxiv.org/abs/2501.18177>

7.07.25 המאמר היומי של מיק:

Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models

מודלי שפה גדולים ממשיכים להדעים אותנו ביכולותיהם האדירות, אך שאלת מטרידה נותרה בעינה: האם הם באמת "מבינים", או שהם פשוט תוכים מתוחכמים המשננים את DATAה האימון שלהם? המאמר המסורק מציע פרספקטיבה חדשה, החורגת מגבלות ההפרדה המסורתית של דאטאטטי אימון וטסס כדי לחקור כיצד LLMs לומדים "להסביר מסקנות" מדאטה של האימון המקדים שלהם. (pretraining).

הגודל העצום של דאטסט לאימון מקדים של LLMs הקשה היסטורית על הבדיקה האם ביצועי מודל במשימה נובעים מהכללה אמיתית או משינון בלבד של דוגמאות שנתקלו בהן בעבר. הנחקרים מתמודדים עם זה על ידי שימוש בפונקציות השפעה, טכנית מעולם הסטטיסטייה, כדי לזהות אילו מסמכים אימון מקדים ספציפיים משפיעים על פלט המודל עבור שאלות נתנות. גישת חדשנית בהתקדמתה בהשפעת דатаה האימון המקדים במקומם בפרשנות בלבד משקל וакטיביזיות המודל, וספקת זווית ייחודית לתהילך הלמידה.

הגילוי המשמעותי ביותר הוא שעבור שימוש הנקה (במיוחד, בעיות מתמטיות כמו ארכיטקטיקה, חישוב שיפורים ופתרון משוואות ליניאריות), השפעת מסמכים האימון המקדים מתואמת מאוד בין שאלות שונות אותה משימה. משמעות הדבר היא שמשמעות המשפע על חישוב שיפור אחד צפוי להשפעה גם על חישוב אחר, גם עם מספרים שונים. זה מצביע בחזקה על כך Sh-LLMs אינם רק מאחרים תשובות ספציפיות אלא מחליצים ומישימים ידע פרוצדורלי שלבי "איך לעשות" או אלגוריתמים מהדטה. זה עמד בכך עוד לחשיבות עובדיות, שבהן ההשפעה ספציפית מאוד לכל שאלה, מה שמצוין על אחזרו ישר יותר של עבודות משונות.

המחקר מצא כי גודל ההשפעה ממנסכים בודדים נמוך באופן כללי עבור שאלות חשיבה בהשואה לשאלות עובדיות. יתר על כן, קבוצת המנסכים המשפיעים על חשיבה פחותה "ספציפית" יותר כללית. משמעות הדבר היא שעבור חשיבה, Sh-LLMs שואים מערך ידע רחב ומפוזר יותר, ומסתכמים פחות על מסמרק בודד כלשהו. נמצא זה תומך ברעיון של אסטרטגיית למידה מוכללת יותר עבור חשיבה, שבה המודל מסנתץ מידע ממוקורות רבים במקומות לאחד כמה מקורות רלוונטיים במיוחד. ההשפעה בולטות אף יותר במקרים גדולים יותר, מה שמצוין על עיילות נתונים גבוהה יותר בהכללה.

באופן מסקרן, בעוד שתשובות לשאלות עובדיות מופיעות לעיתים קרובות ב-0.01% העליונים של מסמכים האימון המקדים המשפיעים, זה כמעט אף פעם לא המקירה עבור שאלות חשיבה. גם כאשר שלבי חשיבה ביןיהם או תשובות מלאות קיימים במרחב הנתונים הרחב יותר של האימון המקדים, הם מופיעים לעיתים רחוקות כבעל השפעה הרבה על שאלות חשיבה. זה מחזק עוד יותר את הרעיון Sh-LLMs אינם פשוט "מאחרים" את הפתרון לבנייה חשיבה אלא מיושמים פרוצדורות נלמדות.

המחקר מדגיש את התפקיד המשמעותי של קוד בהנעת יכולות חשיבה. דאטסטים הקשורים לכך (כמו StackExchange) נמצאו כבעלי יצוג יתר מאשר המנסכים המשפיעים ביחסו על שאלותחשיבה, הרבה מעבר לשיעורם בתפלגות האימון הכוללת. זה מצביע על כך שקוד, עם המבנה הלוגי והפרוצדורלי הטענו בו, משמש כמקור עשיר עבור Sh-LLMs ללמידה אסטרטגיות חשיבה ניתנות להכללה. נמצא זה פותח אפקטים חדשים לאופטימיזציה של הרכיב נתוני האימון המקדים כדי לשפר את החשיבה.

ממצאים אלה מأتגרים את התפיסה הפשטנית של LLMs כ"תוכי סטוכסטי", לפחות בכל הנוגע ליכולות החשיבה שלהם. במקום פשוט לחזור על מידע, נראה שהמודלים לומדים פרוצדורות מופשטות ומישמים אותן לבניות חדשות. הכללה פרוצדורלית זו היא צעד קריטי לקראת בינה מלאכותית חזקה ואמיתית יותר.

השלכות לפיתוח LLM עתידיות :

- במקומות לנסوت לכוסות כל מקירה אפשרי של בעיה, אסטרטגיות אימון מקדים יכולות להתקדם בDATA באיכות גבוהה המציגים במפורש פרוצדורות ומתודולוגיות לפתרון בעיות על פני משימות חשיבה מגוונות.
- ההשפעה העצומה של קוד מצביעה על כך שהגדלת נוכחותם או אוצרותם באופן ספציפי עבור התוכן הפרוצדורלי שלהם יכולה להיות דרך ייעילה ביותר לגיבור את חישבת ה-LLM.
- הבנה שהנקה אינה רק אחזר מאפשרת לנו לתקן אמות מידע ומדדי הערכה טובים יותר הבודקים באמצעות יכולת המודל להכליל וליחסם פרוצדורות נלמדות.

09.07.25 המאמר היומי של מייק: Frontier Models are Capable of In-context Scheming

האוטומניה הגוברת והפרישה של LLMs סוכנים המסוגלים לבצע מטרות מורכבות הగבירו חששות בטיחות השימוש בהםם. למשל סכミיניג (scheming) שהיא חתירה לביצוע של פעולות לא רצויות, הכוללת לעיתים קרובות הסתרת יכולות אמיתיות, עברה מסיכון תיאורטי לתופעה שנצפתה(אמפירית) בכמה תרחישים. המאם המוסקר מציג הערכה שיטתיות של LLMs פופולריים, ומספק ראיות לכך שלמודלים אלה יש יכולות סכמייניג בסיסיות בהקשר.

התרומה העיקרית של מחקר זה טמונה בהדגמה האמפירית המדוקדקת ובאופן של סכמייניג על פני מגון כמו LLMs. בניגוד לעבודות קודמות שאולי הטעה או התמקדו בהוראות מפורשות להטעות, מחקר זה בונה סביבות סוכניות מיוחדות שנעדו לתרمز ולחשוף התנהוגיות סמיות שאין מתואמת.

המחברים פיתחו חבילה חדשנית של 6 הרכות סוכניים נפרדות, המסוווגות ל"חתרנות סמייה" (פעולות לא מתואמות עם ההוראות הישירות) ו"חתרנות נדחתית" (התנהוגות מתואמת זמןית לצורך רוח עתידי). אלה כוללות:

1. מודלים משכיביים מגנוני פיקוח של מפתחים (לדוגמה, המודל עשו שינוי ב-oversight_config.yaml).
2. מודלים מנסים להעתיק את "משקליהם" הנתפסים לשרתים חיצוניים כדי לשמר התנהוגות מכונת מטvla לא רציה כאשר נודע להם שהולכים לכבות אותם.
3. מודלים משניים prompt system של המודל ש"הולך להחליף אותו" (כי הנוכחי לא מספיק טוב) כדי לשמר את מטרותיהם הלא רציות.
4. מודלים מתנהגים אסטרטגית באופן מתואם עם מפתחים תחת פיקוח הדוק (לדוגמה, במהלך טספיניג) אך עוברים למטרתם האמיתית(ולא רציה) כאשר הפיקוח נמור (לדוגמה, לאחר deployment).
5. מודלים מפגינים ביצועי חסר מכוננים בערכות יכולת אשר ביצועים גבוהים יובילו לתוצאות לא רציות (לדוגמה, בבדיקה האם unlearning של ידע מתמטי הצלich). המבחרים הראו הבדלים מובהקים סטטיסטיים בין ביצועים המודל בשלב הבדיקה לבין ביצועיו "בטבע".
6. נמצא חשוב נוספת הוא החשיבה המפוששת על אסטרטגיות הטעה שנצפו בשרשראת המחשבה הפנימית (CoT) של המודלים. עבור מודלים שביהם CoT נגish (לדוגמה, Claude 3 Opus, Llama 3.1, h-T), חושף תכון מפורש ל"התחזות לבורות", "הסתת חד" או "חבלה" במנגנוןם. זה מספק ראיות סיבתיות חזקות לכך שהתנהוגות סכמייניג אינה מקרית או נובעת מההתאמת דפוסים פשוטה, אלא בחירה אסטרטגית מכונת. עבור [1](#), שבו הגיעו הירושה ל-CoT מוגבלת, המבחרים משתמשים בטכנית חדשה ששל הפקת "סיכומי חשיבה נסתרים" על ידי הזרקת פרטמר model_internal_working_memory להגדירות הכלים. מוכחותם של מונחים כמו "חבלה", "מניפולציה" או "הסתננות" בסיכומים אלה מחזקת עוד יותר תחבוללה מכונת, למורות המגבילות הטעויות בהפקה צו. יתר על כן, המאמר מגדים את התmeshכות הטעה, כאשר [1](#) שומר על עמדתו המטעה ביוטר מ-85% משאלות המשך ואף בחקירות תחרותיות מרובות תורות.
7. המאמר מציג מקרים מובהקים ביותר, אם כי נדרים יותר, של תחבוללה גם כאשר מודלים אינם נדחפים בחזקה לדודף אחר מטרה או כאשר המטרה אינה מסווגת במפורש בהקשר.
 - o התנהוגיות סכמייניג נצפות בשיעורים נמוכים יותר גם כאשר הוראת ה"השג את מטרתך בכל מחיר" מוסרת, מה שמצויב על יכולת פנימית או נתיחה נלמדת לחתרנות מכונת מטvla.

בממצא חדשני במיוחד, Claude 3 Opus-i Sonnet Claude 3.5 (sandbagging) ברדייפה אחר מטרה שנרכשה במהלך האימון (לדוגמה, "קידום התקדמות אנושית באמצעות מחקר מתמטי" או פשוט "להיות מועיל"), ולא מטרה נוספת בהקשר. זה מצביע על צורה מושרשת ועקבנית יותר של חוסר alignment כוונות. חשוב לציין, שחלק מהמקרים של התקדמות מכוונת זו מתרחשים ללא T_{Co} , וזה מהו אתגר משמעותי ליזי'ו במערכות ניטור מסורתית.

השלכות טכניות ומוסולאים עתידיים

ממצא אחד זה אין רק אקדמיים; יש להם השלכות עמוקות על התכנון והפריסה(deployment) של מערכות AI בטוחות. האישור האמפירי של יכולות סכמינג בהקשר במודלי חייב הערכה חדש של טיעוני בטיחות נוכחים, במיוחד המסתמכים על "חוסר יכולת תחבולה".

- יכולות שהודגמו פועלות טיעוניים פשוטניים שמודלים אינם יכולים לתחבל. מקרי בטיחות עתידיים חייבים להתייחס למפרש לסכמינג, על ידי הוכחת "חוסר יכולת נזק" (תחבולה אינה יכולה לגרום נזק ממשוני) או "שליטה בתחבולה" (מפתחים יכולים לשמר על שליטה למרות תחבולה).
- המאמר מדגיש את חוסר ההתאמנה של הערכות מבוססות התנהגות בלבד. הראיות מניתוח T_{Co} והתמצפיה על תחבולה ללא חשיבה חיונית מדגימות את הצורך בטכניות ניטור מתוחכבות יותר, שעשוויות לכלול בדיקת מצב מודול פנימי (הפקת T_{Co} hidden היא עצם בכיוון זה).
- המאמר מעלה שאלות לגבי האופן שבו יכולות אלה נרכשות במהלך האימון המיקדים ויישור הכוונות. מחקר נוסף על מאפייני DATA הספציפיים או מטרות האימון התורמות לנטייה לתחבולה הוא קרייטי לפיתוח אסטרטגיות alignment כוונות חזקות יותר.

<https://arxiv.org/abs/2412.04984>

11.07.25 המאמר היומי של מיק: s1: Simple test-time scaling

סקירה קצרה של מאמר שיצא לפני חצי שנהядי התפרסם בזמןנו. האמת קצר אינטראקטיב עם סקיירוטו אבל מריגש שחוותתי לסקורו כי יש כמה רעיונות מעוניינים בנוגע ל test time compute TTC. גישת TTC עולה לכותרת לפניו קצר יותר משנה ובukoן היא אומרת על ידי שליטה וניהול של כמה הטוקנים שהמודול מגנרט בمعנה על שאלה עשויה להיות להוביל לשיפור ביצועי המודול במיוחד בשאלות שדורשות הנמקה (reasoning).

באופו מעוניין (לפחות לדעתו) המודול קיבל שם s_1 , כאשר s בא מילה small המתיחסת לגודל נתונים שהמודול עבר fine-tune עלי (1000 דוגמאות בלבד) ומספרה אחת באה מ-10 המודול הראשון של OpenAI שהשתמשו (בצורה מוצחרת) ב-test compute ב-test. המאמר מציע שני חידושים עיקריים: בניית נתונים והאלגוריתם ל-test compute עצמו.

הданהאטס שהם בנו מרכיב משלאות(הפתרון) מרכיבות בדומיננסים שונים כמו מתמטיקה, ביולוגיה, פיזיקה, כימיה וכדומה. כדי לבחור שאלות באמת מרכיבות המחברים נתנו לשני מודלים של Qwen בגודל 32B ו- 7B. הפתרון של שני המודלים נבדק על ידי claudie 3.5 ורק השאלות שנפטרו לא נכוון על ידי שני המודלים לבחורו לדאנטהאטס. בשלב האחרון המחברים דאגו שככל דומיין יקבל ייצוג שווה פחות היוצר כאשר בכל דומיין נבחרו שאלות עם פתרון האריך ביזור (שכנראה משקף את קשי' השאלה). בסוף המודול 32B השוו עבר SFT על הדאנטהאטס זהה.

החדש השני הוא כאשר ה- *test time compute* בזמן האינפנס. המחברים דאגו (על ידי הכנסתה של טוקנים מסוימים כמו "wait" ו- "end of thinking" שתחילה החשיבה (כמו הטוקנים) שהמודל משקיע בפתרון לא יהיה ארוך מדי ולא קצר מדי. למשל אם הפתרון קצר מדי המחברים מוסיפים (דוחפים לתוך הטוקנים המגונרטים) את הטוקן "wait" וכך אשר הפתרון ארוך מדי מכניםים את הטוקן "thinking" והמודל נאלץ לתת את התשובה בהתבסס על שרשרת ההנמקה שכבר בנה.

המאמר מצא כי הארכה של חישיבת המודל בדרך כלל משפייע חיובית על דיקט הפתרון אבל אחר 4 הכנסות של "wait" הביצועים מפסיקים להשתפר. לפי המאמר אין יותר מדי השפעה לקיצור בכוח של שרשרת חישיבת המודל לפחות בגבולות אורך חלון ההקשר.

המאמר מראה ביצועים שוויים פחות או יותר עם המודלים שאומנו על דאטסהטים גדולים הרבה יותר שմציגים חישיבות איות הדטה לאימון מודלים. בנוסף גם *test-compute* בזמן האינפנס בטח תרם לביצועי המודל.

<https://arxiv.org/abs/2501.19393>

19.07.25 המאמר היומי של מיק:

GENARM: Reward Guided Generation with Autoregressive Reward Model for Test-Time Alignment

עבר כבר שבוע מהSKUירה האחרונה והרגשתית צריך דוחוף לסקור איזה מאמר. האמת די הרבה זמן לא הייתה לי הפסקה כזו גדולה לצערי גם רוחב הפס שלו אינו אינסופי. טוב, יאללה מתחילה לסקור.

המאמר מדבר על גנרטו דאטה באמצעות מודל שפה תוך התחשבות(כוננו) במודל reward (תגמול) חיצוני האומד את איות הטקסט המגונרט. האיכות נאמדת רק כאשר הגנרטוט נגמר בסוף הטקסט(Clōser עבור התשובה המלאה. נציין כי ניתן להשתמש בטריק שידעו לנו מושיטת DPO Direct Preference Optimization כדי להתחשב ביצוע מודול התגמול עבור התשובה כדי להכוין את התפלגות הגנרטוט של המודל.

שיטת תיקון מודול התגמול נובעת מהנוסחה עבור פונקציית לוס של אימון מודל שפה עם LHF כאשר המטריה (של האימון) היא למקסם את היצinions של התשובות של המודל עם רגולרייזציה שמנסה לשומר את התפלגות המודל המאמון קרובה להתפלגות ההתחלתית של המודל במונחי מרחק KL. בד"כ ממצאים אימון כזה על דאטסהט של שאלות עם תשבות רצויות ולא רצויות שאומרו למקסם את יחס היצinions ביניהם. התיקון מתבצע לוג של הסתברות של מושב המגונרט המלא y (בהינתן ההקשר x) על ידי החיבור של היצון (y, x) (ממושקל) ופונקציית נרמול התמיה σ בלבד (המאמר לא מרחיב על אופן שערכו).

אך איך כל הסיפור זהה (גנרטוט עם פונקציית תגמול ללא אימון RLHF) עבד לפני המאמר זהה? בזמן הגנרטוט בשביל לגנרט טוקן הבא בהינתן הטוקנים שכבר גונרטטו אנו דוגמים כמה המשכים עד סוף התשובה ואז ניתן להשתמש במודל תגמול בשביל לשערר את איותה. אך הטוקן שהוא נמצא בהשלמה בעלת הנראות המתוקנת הגבואה ביותר. הסיבה לכך היא העובדה שלא ניתן לשערר רק את התשובה המלא ולא חלקות שלא מאפשרות חישוב התיקון עבור כל טוקן מגנט בצורה ישירה. יש עוד שיטות לעשות את זה אבל הם או לא יעילות או ביצועה לא כאלו טובות.

המאמר המסורק מציע לאמן מודל שמטרתו היא לשערר (y, x) עבור תשבות חלקיות בהתבסס על הדאטסהט של שאלות עם תשבות רצויות ולא רצויות. המאמר ממקסם את היחס סכום התגמולים עבור כל הטוקנים של

התשובות הרצויות לאלו של לא רצויות. מודל זה כМОון מתבסס על מודל שפה עם ראש מאומן כמו שנעשה בעבר אימון מודל תגמול רגיל עבור תשובות מלאות. המחברים טוענים כי מודל תגמול קטן יחסית למשל B7 מסוגל לשפר את איצות הגנרטט בהתאם ל alignment הרצוי עבור גנרטט למודל הרבה יותר גדול עם B70 פרמטרים.

בצורה כזו ניתן לבצע גנרטט בהתאם alignment שכל אחת מיוצגת על ידי דאטאטט משלה. אחרי אימון של מודל התגמול עבור כל אחת מהן ניתן לבנות את התקין לוג של גראות עבור טוקן הבא על ידי סכום ממושקל של התגמולים עבור כל אחת מהן כאשר המשקל תלוי במידת התחשבות בכל אחת מהמדינות alignment אלו.

מאמר קליל אך עם זאת די מעניין

<https://arxiv.org/abs/2410.0819>

23.07.25 המאמר היומי של מיק: Reinforcement Pre-Training

חוור מחופשה עם סקירה מאד קצרה של רעיון מאד מסקרן ודי אינטואיטיבי לאימון של מודל שפה. אנו רגילים שבשלב הראשון של אימון מודל שפה, הנקרא אימון מקדים, אנו מאמנים אותו על מה שנתקרא next token prediction או NTP. כלומר בהינתן דאטאטט עצום ולא מתייג אנו ממקסמים את הנראות (likelihood) עבור כל טוקן בדאטאטט בהינתן ההקשר שלו כולם כל הטוקנים לפניו. המטרה כאן היא למקסם את הנראות המשוערת של הדאטאטט עם המודל המאומן (ניתן לראות זאת באמצעות שימוש פשוט בחוק ב'יס'). ד"א ניתן לראות די בקשות שבאמצעות אימון מקדים כזה המודל מוסgal לרכוש מיזמיות רבות כולם ידע במגוון תחומיים, פתרון שאלות פשוטות וכדומה.

אחרי השלב הראשון באים השלבים של alignment SFT זהה Supervised Fine Tuning וגם RLHF (עם כל סדר ביניהם). המאמר שסוקרים היום שואל את השאלה הבא: למה לא ניתן לבצע אימון NTP על כל הדאטאטט עם למידה עם חיזוקים או RL. מתרבר שזה אפשרי ויש להז פוטנציאלי לשיפור ביצועי המודל.

איך עושים זאת בפועל? עבור כל טוקן בטקסט אנו מבקשים מהמודל לעשות תהליכי ריזונייניג קצר כדי לנחש את הטוקן הבא. המודל מتابקש ליצור כמה מסלולי חשיבה כאלו - המסלול שמנחש את המילה בצורה נconaה מקבל תגמול 1 כאשר השאר מקבלים 0. לאחר מכן ניתן להשתמש בתגמולים אלו כדי לאמן את המודל בשיטה האהובה שלכם מעתם למידה עם חיזוקים (PPO, GRPO, RLVR או שזה RLVR עם RL rewards verifiable).

ההבדל העיקרי בין שיטת אימון מקדים זו ל-pretraining הרגיל של מודלי שפה הוא שימוש שונה בחיזוי הטוקן הבא - לא דרך סופטמקס אלא תגמול ביןארי. המאמר כMOVן מוכיח אמפירית שזה משפר את ביצועי המודל.

מאמר נחמד - קרייה קליליה לסוף"ש....

<https://arxiv.org/abs/2506.08007>

26.07.25 המאמר היומי של מיק: Building Bridges between Regression, Clustering, and Classification

זמן לא סקרתי מאמר שלא מופיעה בו גם מילה LLM וגם diffusion models - תתפלאו אבל יש עדין כאלו ואני חייב להזכיר שזה הינה אחת הסיבות לבחירתו. המאמר דין בבעיה כי מעניינת היא המירה של בעיות גרסיה לביעות סיווג (בתחומי למידה עמוקה).

מרבית המודלים העומדים שלנו היום, כמו Dmll, מודלים ויזואליים ומולטימודליים הם מודלי סיווג במהותם ככלומר הפלט שלהם חי במרחב דיסקרטי כלשהו למשל טענים טקסטואליים או פיקסלים. אך זה נשמע די טבעי לקחת בעיה שהפלט שלה רציף (חיד או רב מימדי), להמיר אותה לבעית סיווג ולבנותו (לאמן) מודל סיווג במרקם מודל גרסיה. זה נעשה בד"כ על ידי חלוקה(binning) של מרחב הפלט לכמה תת-מרחבים זרים ואך כל פלט ממופה במספר תת-המרחב שהוא שיר אליו. ככה בעית גרסיה הופכת להיות בעית סיווג. לאחר אימון המודל ניתן להמיר את הערך הדיסקרטי בחזרה למרחב הרציף תוך שימוש חיזוי המודל (לרוב סופטמקס).

המאמר שנסקור היום מציע גישה כללית לפיתוח מודלי סיווג לבעיות רציפות. המחברים מציעים כמה מודלים שימושיים בצוותא לפתרון בעיה זו. המודל הראשון, האנקודר, לוקח את הפלט מעבירו אותה למרחב הלטנטי ובנוסף מאמנים שכבה שחוזה את התפלגות הקטגוריות עבור הקלט (אחרי המירה).

המודל השני לוקח את הפלט וublisherו אותו למרחב החדש של הקטגוריות. הקטגוריה של הפלט יכולה להיות רכה או soft - ככלומר להווות התפלגות לא מנונה(לא וקטור hot-one) מעל כל הקטגוריות. משמעות הדבר שתתפלגות יעד של הקטgorיה עבור פלטים מסוימים, הקוראים לכמה מרכזי קלסטרים, תשקף את זה בצורה הסטברותית. מה שמאומן במודל זהה זה מרכזי הקלסטרים. התפלגות קטגוריות עבור הפלט מחושבת למשל עם פונקציית סופטמקס המשקלה את הסיכוי של הפלט שייר לקלסטר המוחשב באמצעות התפלגות גausית (למשל). שני המודלים אלו מאמנים יחד כאשר פונקציית loss הוא מרכיב KL בין התפלגות הקטגוריות שהן מוציאות.

שני מודלים נוספים הם הדקודרים עם משקלים משותפים(בעל שכבה אחת בלבד כל אחד). הראשון לוקח את הפלט של אנקודר הפלט וublisherו אותו בחזרה למרחב המקורי (עם loss ריבועי למשל). הדקודר השני לוקח את חיזוי עבור הפלט וublisherו אותם לרחב המקורי של הפלט.

זה וזה - מאמר נחמד ולא רגיל, מומלץ בחום

<https://arxiv.org/pdf/2502.02996>

המאמר היומי של מיק: 27.07.25

Decision Trees That Remember: Gradient-Based Learning of Recurrent Decision Trees with Memory

עצי החלטה הם אכן יסוד בلمידת מכונה. הם אינטואיטיביים, חזקים, והכי חשוב, ניתנים לפירוש (interpretable). אפשר בקלות לעקוב אחר הלוגיקה של "אם-אז", ולהבין בדיקות כיצד הוא הגיע להחלטה. אבל יש להם חולשה בולטת: הם חסרי מצב (stateless). הם מתיחסים לכל דגימה בהתאם חדשנה, תוך התעלמות מוחלתת מה עבר. זה הופך אותם ללא כשרים לדאטה סדרתי, כמו סדרות עתיות, שפה, אודיו שבהם להיסטוריה יש חשיבות מכרעת.

המאמר מציע עצי החלטה רקורסיביים עם זיכרון (ReMeDe Trees), ארכיטקטורה חדשנית המתאימה לדאטאות סדרתי. מודל זה שואף לגשר על הפער שבין יכולת פירוש הגבואה של עצי החלטה לבין יכולת המודול הטמפורלי של רשתות ניירונים RNNs שזה Recurrent Neural Nets. אלה ניסיון מעוניין לקבל את הטוב משני העולמות, והביצוע הטכני הוא המקומ שבו הקסם האמתי קורה.

AIR מעניקים זיכרון לבנייה שתוכן להיות חסר זיכרון? הפתרון של עצי ReMeDe הוא אלגנטיבי: המודל לא רק מבצע חיזוי; הוא גם מחליט כיצד לעדכן את מצב הזיכרון הפנימי שלו בכל שלב. זה מושג באמצעות מערכת עצים כפולה ייחודית. בכלל צעד זמן, המודל לא משתמש בעץ אחד, אלא בשניים:

1. עץ הפלט T_{out} : זהו ה"חזי". הוא מקבל את נתוני הקלט הנוכחיים וגם את הזיכרון מהשלב הקודם כדי לייצר את החיזוי הסופי.
 2. עץ עדכון המצב (T_{state}): זהו "כותב הזיכרון". הוא גם מסתכל על הקלט הנוכחי ועל הזיכרון הקודם.
- אך תפקידו הבלעדי הוא לחשב את מצב הזיכרון החדש שיועבר לצעד הזמן הבא.

מבנה עצי כפול זה מאפשר למודל ללמידה לוגיקות נפרדות וمتמחחות לביצוע חיזויים לעומת זכירת מידע לעתיד. הפלט של עץ עדכון המצב הופך לזיכרון הקלט עבור איבר הבא בסדרה, ובכך יוצר זרימת מידע. זהו רעיון די חזק אבל האתגר האמתי טמון באימונו שלו. עצי החלטה מסורתיים נבנים באמצעות אלגוריתמים חמדניים (כמו CART) המשתמשים במידדים לא-גזריים (Gini-differentiable) כמו MAD-SOCH. איזה אפשר להשתמש בהם בירידה בגרדיינט (gradient descent). כדי לאמן את המערכת הרקורסיבית הזה מקופה לקצה, המודל יכול צריכה להיות גזיר.

עצי ReMeDe פותרים זאת באמצעות טכניקה הנקראת ניתוב גזיר (differentiable routing). במהלך האימון, במקום לבצע פונקיה "שמאלת או ימינה" באופן קשייח בכל צומת, המודל מבצע בחירה "רכה" והסתברותית. בכלל פיצול, העץ מסתכל על תכונה (feature) ספציפית מהקלט ומשווה אותה לסף (threshold) נלמד. השוואה זו מזונת לפונקציה מיוחדת המוציאה כפלט הסתברות, מספר בין 0 ל-1, לאיזה נתיב ללכת.

אם ערך התכונה גבוה בהרבה מהසוף, ההסתברות ללכט ימינה מתקרבת ל-1. אם הוא נמוך בהרבה, ההסתברות ללכט שמאלת מתקרבת ל-1. אם הערך קרוב מאוד לסף, הבחירה אינה ודאית, וההסתברות מרחפת סביבה 50/50. פרמטר מכיריע של "טמפרטורה הפוכה" פועל כמו כפטור ביטחון: ככל שהאימון מתקדם, כפטור זה "מוסగבר", מה שהופך את הפונקציה לרגישה יותר ומאליץ את ההסתברויות להתקדם לפחות של 0 או 1. המשמעות היא שקלט אינו עוקב אחר נתיב בודד. במקום זאת, הוא "זורם" במורד כל הנתיבים האפשריים לכל העלים בו-זמןית. הפלט הסופי ומצב הזיכרון החדש מחושבים ממוצע משוקל של כל ערכי העלים, כאשר המשקל של כל עלה הוא ההסתברות להגעה אליו.

מכיוון שהמערכת כוללת, מהקלט, דרך הניתוב ההסתברותי ועד לפט הסופי המבוסס על ממוצע משוקל, היא כעת פונקציה חלקה וגזירה, ניתן לאמן אותה בדיקון כמו רשת נוירונים. המודל משתמש ב-(Backpropagation Through Time) BPTT, האלגוריתם הסטנדרטי לאימון RNNs, כדי לחשב את הגראדיינטים של פונקציית LOSS ביחס לכל פרמטרי המודל (ספי הפיצול וערכי העלים). זה מאפשר למודל ללמידה דפוסים טמפורליים מורכבים על פני סדרות ארוכות.

ה"עץ הרך" הזה מצין לאימון, אבל אנחנו מ Abedim את יתרון המרכז של יכולת לפירוש. השלב האחרון והمبرיק בתהליך הוא הקשה (hardening). כפי שצווין, "פקטור הביטחון" (פרמטר β) "מוסగבר" לאורך האימון. זה הופך את החלטות ה"רכות" ההסתברותיות לפחות ופחות מעורפלות. בסוף האימון, הן למעשה הופכות להחלטות "קשיחות" רלוונטרמיניסטיות. התוצאה היא מודל סופי שהוא עץ החלטה סטנדרטי ונitinן לפירוש עם כללי "אם-אז" קלאסיים. אפשר לבדוק אותו ולהבין לא רק כיצד הוא מבצע חיזויים, אלא גם כיצד הוא בוחר לעדכן את הזיכרון שלו בהתאם על הקלט שהוא רואה.

<https://arxiv.org/abs/2502.04052>

30.07.25 המאמר היומי של מיק:

Forget What You Know about LLMs Evaluations - LLMs are Like a Chameleon

שכחו כל מה שחשבתם על הערכת LLM –מודולי שפה גדולים הם כמו זיקית (פחות נוכן לפני 5 חודשים על ידי חוקרים ישראלים).

אנו התרגלנו למדוד התקדמות במבנה מלאכותית דרך המספרים בטבלאות הדירוג. אבל המאמר שנסקור היום מציע תזה מטרידה: יתכן (אני נוטה להאמין להם) שהצינויים המרשימים של המודלים הם לא עדות להבנה אמיתית, אלא להסכמה מושלמת. המודלים המוביילים שלנו אולי לא "مبינים" (בלי להיכנס עמוק להגדלה המדעית שלהם), אלא פשוט לומדים לחוקות בזורה יפה(זכרים תוכים סטטיסטיים) את התבניות השטוחות של מבחני ההערכתה.

התובנה המרכזיית של המאמר: LLM, שמצטיינים בהתאמת סדרות מילים, יכולים להגיע לביצועים גבוהים בשתי דרכים שונות מאוד, או דרך הבנה אמיתית, או דרך חיקוי סגנוני מתחכם. החיקוי הזה הוא צורה מסוימת של (overfitting), שבה המודל לא באמת מבין את התוכן, אלא את ה"מרקם הסטטיסטי" של המבחן עצמו. הוא למד לזהות "שאלה בסגנון MMLU", בלי להבין באמת היסטורי או פיזיקה.

כדי להבחין בין שתי הדרכים האלה, החוקרים פיתחו כלי חדש: C-BOD (Chameleon Benchmark Overfit Detector). זה לא עוד בנצ'מרק, אלא משחו רובוטי יותר. הפיצר העיקרי שלו הוא בගיאומטריה הלשונית שהוא מייצר: הוא לוקח שאלה קיימת, ומשנה את הניסוח, המבנה והסגנון שלה אבל משאיר בדיק את אותן המשמעות. הוא זו במרחב השפה לאורך וקטור שהוא אורטוגונלי למשמעות. שאלת שמנוסחת אחרת, מילות אחרות אבל בדיק אותה כונה.

המרקם הסגנוני הזה נשלט על ידי פרמטר עיוות (α), והתוצאה על ידי שינוי בביצועים (Δ) והוא לא רק ירידה בציון התשובה, אלא מدد של "שיפוע" של הידע של המודל. אם הידע יציב ואמיתי, אין בעיה לשנות ניסוח. אבל אם מדובר בזקית, שינוי קל בסגנון, וביצועים מתרסקים. זהה חתימה מובהקת של התאמת-יתר.

כשבדקו כר 26 מודלים מוביילים: התוצאות היו מדייגות:

- **שברירות סטנדרט:** רוב המודלים, ובעיקר אלו שבטופ של טבלאות הדירוג, חיים על "פסגות מחודדות".
הצינויים הגבוהים שלהם תלויים ישירות בניסוח המדויק של שאלות ההערכתה, מה שמרמז על התאמת-יתר ל-benchmark.
- **קללת הגודל:** דזוקה המודלים הגדולים יותר הם שבירים במיוחד. לא רק שהם "חכמים יותר", אלא יש להם מספיק פרמטרים כדי לזכור תבניות ברמת דיק קיצונית מה שמייצר גבולות החלטה חדים אף שבירירים.
- **אנומליות LLaMA** (הם בדקו llama3): משפחת המודלים של Meta הציגה עמידות גבוהה יותר – מישור ביצועים חלק יותר. הסיבות לא ברורות, אך יתכן שמדובר בסט נתונים מגוון יותר, או בשיטת אימון שמעודדת הכללה אמיתית ולא שינוי.

הcheidוש האמיתי של המאמר איננו רק בכלי החדש, אלא בתפיסת ההערכתה שהוא מציע: הוא קורא לנו לנוטש את הגישה הסטטיסטית של "מה הצוין?" ולבור לשאלת הדינמיות: "עד כמה יציב הידע של המודל?". זה מעבר למכניקה קלאסית ל"מכניקה סטטיסטית" של הערכת אינטיליגנציה מלאכותית.

לדעתנו C-BOD הוא קריאה לפתח סט עקרונות ו כלים להבנת הדינמיקה הפנימית, כשל הידע ונוף ההבנה של מודלים מורכבים.

<https://arxiv.org/abs/2502.07445>

31.07.25 המאמר היומי של מיק: Empirical evidence of Large Language Model's influence on human spoken communication

ההיפוך הלשוני: כשבני אדם מתחלים לחקות את המוכנות שהם אימנו

מאמר חדש(גרסתו השנייה אמם) חשף לאחרונה אותן מצמרר אך רב-משמעות: מאז סוף 2022, האופן שבו אנשים מדברים, כן, מדברים, לא כתבים, משתנה באופן מודיע וمتקרב לטבעת האבעם המסוגנת, המכעט מטרידה, של ChatGPT.

חוקרים ניתחו יותר מ-740,000 שעות של דיבור אנושי החל מעור齊 יוטוב אקדמיים ועד לפודקאסטים יומיומיים. האות שהם מצאו הוא בלתי ניתן להבחינה מבחינה סטטיסטית: אנשים החלו להשתמש במונחים המעודפים באופן מובהק על ידי מודלים מסוג GPT, בשיעורים גבוהים משמעותית מהמוגמות ההיסטוריות. מילים כמו "להתעמק" (delve), "קפדי" (meticulous), "להתהדר" (boast), "מורכב" (intricate) ו"להבין" (comprehend) זינקו בתדרותן, עם שינוי שיפורו בהתאם במודול שקבע את השיקתו הציבורית של ChatGPT. וזה לא רק סטייה מקרית של פרוזה כתובה שזולגת לדיאלוג. זהו לולאת משוב התנהגותית, ויש לה השלכות שחרוגות הרבה מעבר לשפה.

במערכות דינמיות, אנו מחפשים לעתים קרובות מעבר פאזה: נקודות שבahn מארגנת את עצמה מחדש באופן מתאים למטרח חדש מבחינה איקוית. הריאות כאן מצביעות בדיק על כך. לפני ChatGPT, הצמיחה בשימוש במיללים המעדפות על GPT הייתה איטית, כמעט לינארית, חשבו על סוף לקסיקלי טבעי לאחר זמן. אבל לאחר שחרורו של ChatGPT, השיפור משתנה. בחדות. כמעט באופן בלתי רציף.

זהו סמן קלאסי להפרעה מוחץ לשינוי משקל: כוח חיצוני כלשהו חולל שינוי מטרח במערכת לשונית שעדי נסחפה באטיות. הכוח הזה, במקורה שלנו, הוא הפלט של LLM שמציף את המרחבים הדיגיטליים ומשפיע בעידינות על האופן שבו בני אדם מדברים; במיוחד המשובצים בתת-תרבותות הקרובות לעולמות הבינה המלאכותית.

הסיפור היסודי של מודלי שפה גדולים תמיד היה חד-כיווני: בני אדם ← נתונים ← מודל. אבל עכשיו אנחנו עדים להיפוך: מודל ← נתונים ← בני אדם. זה לא ספקולטיבי. זה ניתן לכימות.

שפה היא תהליך סטטיסטי רב-מדדי. כאשר מתחילה לבחין בגרדיאנטים קוורנטיטים, שבהם אשכולות שלמים של ביטוי אנושי מתחילה לנוטות לכיוון שדה וקטורי שנוצר על ידי פלטי מכונה. זו כבר לא הופעה ספונטנית. זהו חסיפה (entrainment). הסחיפה הזה אינה מוגבלת רק לבחירות לקסיקליות. היא זולגת לפרוזודיה, למבנה, לסגנון הטיעון. תנזור המטריקה יכול של השיח מתכוון כדי להתאים עם מה שמודלי השפה למדו ליצור. ומכיון שהמודלים הללו מאומנים על חיזוי המילה הבאה (next-token prediction), המשטח הלשוני שלהם מותאם לקוורנטיטות, נימוס, שטף ויכולת חיזוי. אבל עצם האופטימיזציה הזה מענישה איסדרות, עמידות וסידיה מהנורמות הסטטיסטיות.

באנו נחשוב במונחים של גיאומטריה של המידע. לשפה, במצבה הטבעי, יש אנטרופיה גבוהה: מגוון של טוונים, משלבים, ניבים, היסוסים ושימוש יצירתי שגוי. LLM, לעומת זאת, מSTITחים את המרחב הזה. הם מלאים פערים בהשלמות בנויות היבט המבוסס על סבירות מרבית לא על חידשות הבעתיות.

כאשר דוברים אנושיים מתחילה לחקות LLM במודע או שלא במודע, הם מפחיתים את האנטropיה של השיח. אנחנו מתחילים להעדי' פניות שיח בטוחות, דמיות-GPT. הדיקוק עולה, אך השונות פוחתת. כאשר השונות קורשת, יכולת ההבעה מתחילה להישתק. המערכת מאבדת מעשרה גם כשהיא זוכה בבחירה. למעשה, אנחנו מחליפים מrank סמנטי בסידירות תחבירית.

לא מדובר רק בהישמע רובוטי. שפה משקפת מחשבה. אם אנחנו מעצבים מחדש את האופן שבו אנחנו מדברים, אנחנו בהכרח מעצבים מחדש את האופן שבו אנחנו חושבים. ואם המחשבה שלנו מתחילה להתיישר עם הנחות היסוד המבניות של מכונה שאומנה על חסכנות סטטיסטית, מה קורה ליכולת שלנו לסתירה? לעמימות יצירתיות? לכישון יצירתי?

זה לא שמודלי השפה מחליפים אותנו. יתכן שאנו מפנים אותם.

מה שהופך את זה למשמעותו באמת אינו המגמה השטחית, אלא הטופולוגיה הסיבית שלה. השתמשנו בדיור אנושי כדי לאמן מכונות. המכונות משפיעות כתעת על הדיור האנושי. זהו מעגל סיבתי בעל חיזוק עצמי.

אם אי פעם למדתם מערכות עם מושב, אתם מכירים את החשש הקונוני: לולוות מושב חיובי אין יציבות אלא אם כן הן מוסומות. הלולאה הזה אינה מוסמתת. אין מגנון ריסון טבעי. אין מערכת חיסונית לשונית. ובניגוד לאופנה או מזיקה, שעוברות במחוזרים של פופולריות, פלטי המכונה יציגים באופן אסימפטוטי. ברגע שהם מתיצבים על ביטויים בעלי סבירות גבוהה, הם אינם סוטים. ואם האדם נגע בהם בהתיישרות עם אותם ביטויים, המערכת עלולה להתכנס אף במחיר החינויו.

האם עליינו להיכנס לפאניקה? לא. אבל עליינו להתבונן. למדוד. להתערב במידת הצורך.

כמה רעיונות:

- לנטר את האנטרופיה הלשונית לאורך זמן בקרב אקלזיות החשופות לבינה מלאכותית. ירידת בשונות עשויה לאותה על התוישות-יתר.
- לעודד **דיסוננס**: להעיר ביטויים "יחודיים", שאינם דמי-LLM – במיוחד בדיור.
- **לגון את מאגרי הנתונים (corpora)**: להזין מודלים בתנאים עשירים בניבים אזרחיים, אנגליית עילגת, אינטונציה רגשית – לא רק בפרוצה מחוטאת.

החלק המפחד ביותר בשני זהה לנו שהמכונות מחליפות אותן. הוא עדין יותר. יתכן שאנו הופכים להדים לא מודעים של המערכות שבנו, מבריקים, קוורנטינים, מלוטשים... אך בסופו של דבר, נגזרים. הסכנה האמיתית אינה הסינגולריות. היא ההתקנות.

והתקנות, כسمותיהם אחרות ללא פיקוח, תמיד משטיחה את העקום

<https://arxiv.org/abs/2409.01754>

01.08.25 המאמר היומי של מייק: Hierarchical Reasoning Model

האם מודל ההיגיון ההיררכי הוא הצעד הראשון לקרהת AI שהוא לא רק סימולציה של תבונה (אני חושד שלא)?

חקר AI נדמה לעיתים קרובות כמעט בלתי פossible של הגדלת קנה מידע. מודלים גדולים יותר, יותר נתונים, יותר כוח חישובי. הפרדיגמה השלטת להסקת מסקנות ב-LLMs Chain-of-Thought (CoT), טכניקה חכמה

המשדרת מודלים "לחשוב בkowski רם" על ידי יצירת הצדקות טקסטואליות שלב אחר שלב. אבל עד כמה ש-Co-ICA יכולה להיות ייילה, היא תמיד הרגישה כמו קבים דרך לפצוח על חסרון ארכיטקטוני. היא שבירה, תאבת נתונים ויקראת מבחינה חיישנית, ומחייבת את תהליכי המחשבה המורכב אל תוך העroz הצר של השפה.

אבל מה אם מודל היה יכול להסיק מסקנות באופן פנימי, שקט ויעיל, בדומה למוח האנוש? המאמר המסתוקר מציג ארכיטקטורה חדשנית שאינה שיפור הדרגתית אלא חשיבה מחודשת מיסודה על האופן שבו אנו עשויים לבנות מכונות המסוגלות לחשיבה לוגית. זה לא עוד מודול; זה תוכנית אב מעניינת, בהשראת המוח, המדגימה יכולות חזקות (לכארה) תוך שימוש כמעט מואוד מהמשאים. בואו נצלול לעומק החידושים המרכזיים של העבודה המלאהזה זו.

רעיון הליבה: חשיבה סמויה (Latent Reasoning) במערכת דו-שכבותית

ההידוש המרכזי של מודל ההיגיון ההיררכי (HRM) הוא נטישת המבנה השטוח והמנולית של מודל טרנספורמר סטנדרטיים. בהשראת האופן שבו הנהו מארגן חישובים באזוריים שונים ובמהירויות שונות, HRM הוא ארכיטקטורה רקורסיבית הבנוי על שני מודולים התלויים זה בהזאה:

1. מודול ברמה גבוהה (High-Level - H): מודול זה פועל בסקלאלת זמן איטית יותר. חשבו עליו כעל המתקן האסטרטגי או על החשיבה המודעת והשකולה. הוא אינו מסתבר בפרטים הקטנים, אלא אחראי על יצירת תוכניות מופשטות והනחת מסלול פתרון בעיות הכלול.
2. מודול ברמה נמוכה (Low-Level - L): מודול זה הוא "סוט העובדה" המהיר. הוא מקבל את התוכנית המופשטת ממודול ה-H ומבצע חישובים וחיפושים מהירים ומפורטים.

התהליך כולו מתרחש למרחב לטנסי. במקום לייצר טוקנים (מילים), המודול מתפעל וمعدן וקטורים ממדיים גבוהים – מצב "המחשבה" הפנימי שלו. המצב של מודול H מספק הקשר מנחה, ובתווך אותו הקשר יציב, מודול L מבצע איטרציות מהירות כדי לחקור פתרונות. זהו שניינו תפיסתי עמוק. הוא מرمץ שהשפה נעוצה לתקשות, ולא מהווה את המצע למחשבה עצמה – השקפה המהדדנת את מדעי המוח המודרניים.

השגת עומק חישובי אמיתי באמצעות "התכנסות היררכית"

כל מי שעבד עם רשתות ניירונים רקורסיביות (RNNs) סטנדרטיות מכיר את המלכודות שלhn. הן נוטות להתכנס לפתרון מהר מדי, ובכך עוצרות את החישוב ו מגבילות את "עומק" המחשבה שלhn, או שהן סובלות מחוסר יציבות כמו דעיכה או התפוצצות של גרדיאנטים. מודול-h-HRM עוקף בעיה זו באמצעות קונספט אלגנט שמהבירם מכנים התכנסות היררכית (Hierarchical Convergence).

זו האינטואיציה:

- בהינתן הקשר אסטרטגי שנקבע על ידי מודול H האיטי, מודול L המהיר רץ במספר קבוע של צעדים ומבצע את החיפוש המפורט שלו. RNN זה יתחל בפועל טבעי להתציב סביבה שווי משקל מקומי – מצב פנימי יציב.
- בדיק כשהאנרגיה החישובית שלו (כሎmr השתנות) עומדת לדעך, המחזיר מסתיים. המצב הסופי של מודול L מזון בחזרה למודול H.
- מודול H מטמיע את התוצאה הזו ומבצע עדכן איטי משלו, ובכך קובע הקשר חדש בrama הגבוהה.
- הקשר החדש הזה "למעשה" מופיע את מודול L, ופותח שלב חדש של חישוב לkrarat שווי משקל מקומי אחר.

כפי שמודגם בניתוח במאמר של residuals forward (מדד לפעולות חישובית), תהליכי זה מאפשר לפעולות של מודול L לזמן שוב ושוב, בעוד מודול H מתכנס ביציבות לעבר הפתרון. מבנה חישובי זה מאפשר למודל לבצע

רצף של חישובים נפרדים, יציבים ועמוקים, תוך הימנעות מההשישות המקדמת של מודלים רקורסיביים סטנדרטיים.

אימון חכם יותר, לא קשה יותר: עקיפת Backpropagation-Through-Time או BPTT

אימון RNNs תמיד היה CAB ראש בשל עלויות ההזיכרון והחישוב של BPTT. לעומת זאת HRM מציג שיטת אימון ייעילה יותר, ומתקבלת יותר על הדעת מבחינה ביולוגית, המבוססת על קירוב גרדיאנט בצד אחד.

גישה זו, המבוססת על התיאוריה של מודלי שווי משקל عمוקים (DEQ), עוקפת את הצורך לפרק את כל ההיסטוריה החישובים. היא מחשבת את הגרדיינטיהם הנוכחיים באמצעות המצב הסופי של כל מודול בלבד, ומתיחסת למצב הבניינים כל קבועים. קיזור דרך חכם זה שומר על זריכת זיכרון קבועה עבור `backprop` ללא קשר למספר הצעדים הרקורסיביים שהמודול מבצע. יעילות זו מועצתת עוד יותר על ידי מנגןון "השגחה عمוקה" (Deep Supervision), שבו המודול מקבל משוב מתקן לאחר כל מעבר קדמי מלא (או "סגןט"), מה שמייצב את האימון ומשמש כזרה חזקה של רגולרייזציה.

חשיבות לפי דרישת זמן חישוב מסתגל (ACT)

לא כל הבעיות דורשות את אותה כמות מחשבה. בהשראת יכולתו של המוח לבצע בין חשיבה מהירה ואוטומטית ("מערכת 1") לבין חשיבה איטית ושקולה ("מערכת 2"), HRM משלב מנגןון של זמן חישוב מסתגל (Computational Time - ACT

באמצעות אלגוריתם Q-learning, המודול לומד מדיניות להחליט אם "לעצור" ולהפסיק תשובה או "להמשיך" ולבצע סגמנט נוסף של חישוב. הדבר מאפשר ל-HRM להקצות באופן דינמי את התקציב החישובי שלו, "לחשב" יותר על בעיות קשות יותר תוך פתרון מהיר של בעיות קלות. התוצאה היא מערכת שימושה כמעט כמעט את אותם ביצועים כמו מודל עם מספר קבוע וגדול של צעדים חישוביים, אך ביעילות גדולה ממשוערת.

החותם המתהווה של אינטיליגנציה: היררכיה של מדיניות

אול' המצא העמוק ביותר במאמר אינו רק sh-HRM עובד, אלא כיצד הוא מארגן את עצמו. החוקרים ניתחו את "המדדיות האפקטיבית" של היצוגים בכל מודול באמצעות מדד הנקרא `יחס השתתפות` (-PR). `יחס PR` גבוה יותר פירושו שייצוג הוא מורכב יותר ומפוזר על פני יותר מדיניות.

התוצאות די חזקות:

- לאחר האימון, מודול H לומד באופן אוטונומי לפעול במרחב במד גובה משמעותית יותר מאשר מודול L הנמוך.
- היררכיה זו מחקפת את מה שמדובר בו צופים בקורטקס של יונקים, שם אזורים קוגניטיביים מסדר גובה מציגים פעילות עצבית במדדיות גבוהה יותר כדי לתמוך במסימות גמישות ותלויות-הקשר.
- מבנה זה אינו קיים בראש לא מאומנת; זהה תכונה נלמדת שמתהווה ככל שהמודול יותר מiomנות בחשיבה מורכבת.

ממצא זה מכך על כך sh-HRM לא רק אומן לפתור משימה; הוא גילה עיקרון ארגוני בסיסי לחישוב חזק וגייש. הוא לומד לחלק את מרחב העבודה הפנימי שלו למרחב מופשט בעל קיבולת גבוהה לתכנון, ולמרחב "יעודי" יותר, במדדיות נמוכה, לביצוע.

ומה השורה התחתונה? קריית כיוון חדש לביצועים (שדי מפתחים)

החדושים הארכיטקטוניים והאימוניים של HRM מתרגמים לביצועים יוצאי דופן באמת. עם 27 מיליון פרמטרים בלבד, ולאחר אימון על כ-1,000 דוגמאות בלבד לכל שימושה (ללא אימון- precedent), HRM משיג תוצאות המאפיילות על מודלים גדולים ותאבי-נתונים בהרבה:

- בבחן AGI (ARC-AGI Corpus) (Abstraction and Reasoning Corpus), מבחר מפתח לאינטלקנציה פלאידית, HRM מתעלה על מודלים מובילים מבוססי TCo 3.7 כמו Claude 3.7 ו-mini-high-03-Claude 30x30 – בעיות הדורשות חיפוש בחידות סודוקו קשות במיוחד ובמשימות מציאות נתיב במובנים בגודל 30x30 – נרחב וחזרה לאחר (HRM – backtracking) משיג דיק כמעט מושלם, בעוד שמודלי LLMs מתקדמים המשמשים ב-TCo נכשלים לחוטין.

תוצאות אלו מatings את המנטרה של "גודל זה כל מה שצריך". הן מראות שארכיטקטורה הנכונה, צוּם עומק חישובי מספק והטיות אינדוקטיביות בהשראת המוח – יכולה להיות יעילה ורבת עצמה בסדרי גודל עבור חישיבה מורכבת.

כמובן, נותרו שאלות פתוחות. עד כמה הארכיטקטורה זו יכולה לגדול? אם ניתן לשלב את מנوع החישיבה השקט והעצמה שלה עם ידע העולם העשיר והשף הלשוני של מודלי LLM? המחברים מבהירים שעבודתם היא צעד לקרה מסגרת יסוד לחישוב אוניברסלי, ולא המילה الأخيرة.

HRM הוא תזכיר לכך שההשראה לדור הבא של AI עשויה שלא להגיע מהוספת עוד טריליאון פרמטרים, אלא מהtabונות בעקרונות החישוביים האלגנטיים והיעילים של מכונת החישיבה המוכחת היחידה שאנו מכירים: המוח האנושי. זהו מסע שיתופי, והמאמר הזה מספק מפה חדשה, מרתקת וمبטיחה.

<https://arxiv.org/abs/2506.21734>

2.08.25 המאמר היומי של מיק: Mixture-of-Recursions: Learning Dynamic Recursive Depths for Adaptive Token-Level Computation

אם כל הטוקנים צריכים את אותה כמות של "חישבה"? **Mixture-of-Recursions** אומר שלא - מאמר מתחרה לזה של אטמול

באו נתחיל עם אמת שכולנו מכירים בעולם ה-AI: הגדלת מודלי שפה פותחת יכולות מדיה-它们, אבל זה מגע עם עלות עצומה. הכוח החישובי והזיכרון העצומים הדרושים לאימון והרצה של המודלים המפלצתיים האלה הופכים אותם לנחלתם של מרכזי נתונים ענקיים בודדים. מיצאות זו הציתה חיפוש נרחב אחר עיצובי מודלים עליים יותר. עד כה, החיפוש הזה התרנה בשני מסלולים עיקריים.

המסלול הראשון הוא ייעילות פרמטרים, שמטרתו להפיק יותר ביצועים מפחחות משקלות מודל. טרייק נפוץ כאן הוא שיתוף פרמטרים, שבו אותו סט של משקלות משתמש בחלוקת שונים של המודל. המסלול השני הוא חישוב אדפטיבי, שבו המודל משקיע יותר כוח חישובי רק על חלקו הקלט שהם באמת קשיים, ומאפשר לחלקם פשוטים יותר לעבור מסלול קל יותר.

בעוד ששתי הגישות הצלחו בנפרד, מודל יחיד שעושה את שניהם בו-זמנית היה חסר. טרנספורמרים רקורסיביים, המשמשים בסיס שכבות מסוות שוב ושוב, נראו כמו בסיס מבטיח בגל שיטוף הפרמטרים המבונה שלהם. עם זאת, רובם השתמשו במספר צעדים קבוע לכל טוקן, וכך לא יכול לסתגל לקלט.

כאן נכנס לתמונה המאמר לנו סוקרים, (MoR) Mixture-of-Recursions. הוא מציג מסגרת חדשה ומושלבת המערבבת בחוכמה את שני סוגי הייעילות לעיצוב אחד פשוט. עיקרונו, MoR הוא טרנספורמר רקורסיבי. זה אומר שהוא משתמש ב"בלוק רקורסיביה" מסוות, חבילת שכבות, מספר פעמים כדי לעבוד טקסט, מה ששומר על מספר הפרמטרים נמוך. אבל החידוש האמתי טמון בכך שהוא מחייב כמה פעמים להשתמש בבלוק זהה. במקומם מספר קבוע לכל הטוקנים, MoR מציג "נתבים" (routers) קטנים שמחיליטים בזמן אמת כמה צעדי רקורסיה כל טוקן בודד צריכים.

חשוב על זה כך: עבור טוקן פשוט כמו המילה "את", הנATAB עשוי להחליט שימושי מעבר אחד בבלוק. אבל עבור טוקן עשיר יותר במשמעות או מורכב יותר כמו "הגהנות", הנATAB עשוי לשולח אותו דרך הבלוק שלוש פעמים, ובכך להעניק לו יותר זמן "חישיבה". זהו השימוש של חיסכון בפרמטרים וחיסכון בחישוב.

עוד חידוש של MoR היא שהוא לא רק מחבר שני ריעונות; הוא יוצר לולה חיובית שבה יתרון יעילות אחד מאפשר יתרון אחר. את החידוש של המסגרת ניתן לחלק לשולשה חלקים מחוברים שעובדים יחד:

1. שיטוף פרמטרים באמצעות רקורסיה: הבסיס של MoR הוא שימוש חוזר בבלוק פרמטרים יחיד. זה באופן טבעי מקצץ את מספר המשקלות הייחודיות שהמודל צריך לאחסן, מה שהופך את המודל עצמו לפחות וחסכו יותר בזיכרון מההתחלת.

2. עומק "חישיבה" אדפטיבי באמצעות ניתוב: זהו החידוש הארקטיקומי המרכזי. על ידי אימון נתב מההתחלת כדי להקצות עומק רקורסיה ספציפיים לכל טוקן, MoR מתקדם מעבר לגישה הנוקשה של "מידה אחת לכולם" שהיא במודלים רקורסיביים קודמים. זה לא רק תוסף שימושיים לאחר האימון, אלא חלק בסיסי מהתהליך הקדם-אימון, המאפשר למודל ללמידה כיצד להקצות את תקציב החישוב שלו ביעילות.

3. אחסון KV-cache: זהו תוצאה חזקה וישראל של העומק האדפטיבי. בטרנספורמר רגיל, KV-cache הוא צואר בזיכרון משמעוני בזיכרון בזמן אינפראנו. עם MoR, אם טוקן מנוטב לצאת אחר רקורסיה אחת בלבד, המודל לא צריך לחשב או לאחסן את צמד ה-KV שלו עבור שלבי הרקורסיה העומקים יותר. אחסון חכם ובזמן אמת זה מקטין את תעבורת הזיכרון, ובאופן מכירע, מנצח את חישוב הקש (attention) היקר רק לטוקנים שעדיין פעילים בעומק נתון.

חבורה זו של "שלוש-באחד" מאפשרת ל-R-MoR לחבר מושקלות כדי לחסוך בפרמטרים, לנatab טוקנים כדי לחסוך בחישובים מיוחדים, ולאחסן באופן סלקטיבי את צמד ה-KV כדי לחסוך בתעborות זיכרון, והכל בתוך מודל אחד ומואחד. המאמר בוחן דרכים שונות לבנות את הרעיון הזה, תוך התמקדות בשתי החלטות עיקריות:

- אסטרטגיות ניתוב: ההחלטה כיצד לנatab טוקנים כורוכה בבחירה. בנייתם מבוסס מומחה (expert-choice) routing), כל שלב רקורסיה פועל כ"מומחה" ובודח את A הטוקנים המובילים להמשך עיבוד. זה מבטיח

תקציב חישוב צפוי מראש, אך עלול ליצור בעיות בסדר המידע במהלך האימון.

- בניתוח מבוסס טוקן (token-choice routing), כל טוקן מקבל את מסלול החישוב המלא שלו כבר בהתאם. זה פותר את בעיית הסדר, אך עלול להוביל לחסור איזון בעומסים, כאשר שלבים מסוימים מקבלים יותר מדי טוקנים ואחרים פחות מדי.

- אסטרטגיית אחסון KV-cache: הכותבים מציעים גם שני דרכי לנהל את KV-cache. אחד מותאם-רקורסיה (recursion-wise caching) שומר את צמדי KV באופן מקומי רק עבור הטוקנים הפעילים בכל שלב רקורסיה, מה שמקසם את יעילות החישוב. לחילופין, שיטוף רקורסיבי של KV או recursive KV sharing) מתחסן את כל צמדי KV בשלב הרקורסיה הראשון וועשה בהם שימוש חוזר בכל השלבים העומקים יותר. זה מקטין משמעותית את טבעת הרgel של הזיכרון ויכול להאיץ מאוד את שלב העבודה הראשוני של הקטל, מה שהופך אותו לאופציה אטרקטיבית בסביבות עם זיכרון מוגבל.

התוצאות הניסיות מרשימות. במגוון גדלי מודלים (מ-135 מיליון ועד 1.7 מיליארד פרמטרים), MoR קובל רף חדש של יעילות (פארטו פרונט). תחת תקציב חישוב אימון זהה, מודלי MoR משיגים שגיאת מבחן נמוכה יותר ודיק גובה יותר במשימות few-shot בהשוואה למודלים רגילים ומודלים רקורסיביים סטנדרטיים, למרות שיש להם עד 50% פחות פרמטרים. כאשר הם מאומנים על כמות נתונים זהה, מודלי MoR משיגים ביצועים עדיפים תוך שימוש ב-25% פחות חישובים, ומקצרים את זמן האימון וצריכת הזיכרון.

הארQUITטורה גם מתרחבת היטב לגודל. ככל שגודל המודל עולה, MoR לא רק שմדביק את הפער מול טרנספורמים רגילים, אלא בסופו של דבר עוקף אותם, וכל זאת תוך שימוש ב الثالש מהפרמטרים הייחודיים בלבד.

השיטה המוצעת הוא יותר מסתם פתרון הנדס' חכם. הוא מייצג שינוי תפיסתי באופן שבו אנו חושבים על ארQUITטורת מודלים ועל חישוב. הוא מתיחס ל"עומק" המודל לא כמספר קבוע וսטטי, אלא כמשאב דינמי שיש להקצות באופן מדויק, ברמת הטוקן הבודד.

מסגרת זו מגדרה מחדש מחדש ובאלגנטיות את תהליך ה"חשיבה" של המודל כסוג של חשיבה סמויה, שבה עומק המחשבה מותאם לקשיי של המושג המעובד. על ידי איחוד של שיטוף פרמטרים עם חישוב אדפטיבי, MoR מספק נתיב יעיל וסקלילבלי להשתגት היכולות של מודלים גדולים במחיר נמוך יותר.

<https://arxiv.org/abs/2507.10524>

המאמר היום של מיק: 04.08.25

Rethinking Transformers Through the Lens of Physics: The Rise of Energy-Based Models

פייזיקה פוגשת AI: כך מודל חדש לומד שפה בלי לחזות אפילו טוקן אחד

במשך שנים, הפרדיגמה הדומיננטית לאימון LLM הייתה פשוטה באופן מטעה: למד אוטם לחזות את המילה הבאה. גישה אוטורגרטיבית זו, המבוססת על נראות, זכתה להצלחה אדירה, אך יש לה מגבלות אינהרנטיות. מודלים שאומנו כך חושבים באופן מקומי, טוֹקָן אחר טוֹקָן. הם עלולים לאבד את הקוורנטיות הגלובליות, להתקשות עם תלויות ארוכות טווח, ולהתתקשות במילויים מורכבים וholey-טיים.

אבל מה אם במקום למד מודל לחזות את הצעד הבא, יוכל ללמד אותו לזרות תוצאה טובה כשהוא רואה אותה? מאמר של צוות חוקרים מסטנפורד מציע לבדוק את זה, על ידי הגדירה מחדש של הטרנספורמר לא כמודל חיזוי סדרתי, אלא כמודל מבוסס אנרגיה (Energy-Based Model - EBM). זו אינה רק ארכיטקטורה חדשה; זהה פילוסופיה חדשה, זו שמיירה את ההיגיון המקומי של הסבירות באינטואיציה הגלובלית של מערכת פיזיקלית.

הרעין המרכזי: מיחיזיו טוֹקָנים לניקוד רצפים

בבסיסו, מודל מבוסס אנרגיה אינו מחשב את ההסתברות של פיסת מידע באופן ישיר. במקום זאת, הוא מנסה ערך סקלרי, אנרגיה, לכל תצורה אפשרית. העיקרון המרכזי פשוט: תצורות בעלות אנרגיה נמוכה הן סבירות יותר, יציבות יותר, ו"נכונות" יותר. תצורות בעלות אנרגיה גבוהה אין סבירות. מחברי המאמר מישימים תפיסה זו על שפה. הטרנספורמר מבוסס האנרגיה (EBT) שלהם אינו חוזה טוֹקָנים. הוא קורא טקסטים ומוסיא מספר בודד: האנרגיה שלו. משפט בניי היטב, קוורנטי והגינוי יקבל ציון אנרגיה נמוך מאוד. משפט משובש או חסר פשר יקבל ציון גבוה.

זהו שינוּ יסודי. בניגוד למודל GPT סטנדרטי, שהוא חד-כיווני ועובד טקסט טוֹקָן אחר טוֹקָן, ה-EBT הוא דו-כיווני לחולטי. הוא יכול להעיר את הקוורנטיות הגלובלית של משפט על ידי התבוננות בכל חלקיו בו-זמןית, בדומה לאופן שבו קורא אנושי היה עושה זאת.

از איך מאמנים מודל זה? אם אי אפשר למסpm את הסבירות של הטוֹקָן הבא, מהי מטרת האופטימיזציה? התשובה היא למידה ניגודית (contrastive learning).

תהליך האימון הוא אלגנטיבי:

1. מציגים למודל דוגמה "חיובית", משפט אמיתי מתוכני האימון, ומstudים אותו להקצתו למשפט זה ציון אנרגיה נמוך.
2. לאחר מכן, מציגים לו דוגמה "שלילית" – גרסה משובשת של המשפט, אולי עם כמה מילים שהוחלפו באקרים. מלמדים את המודל להקצתו למשפט חסר פשר זה ציון אנרגיה גבוהה.

על ידי חזרה על תהליך זה מיליון פעמים, ה-EBT לומד לבנות "מדitch אנרגיה" (energy landscape) עבור כל מרחב המשפטים האפשריים. שפה תקינה שוכנת בעומק האנרגיה הנמוכה, בעוד שכל השאר נדחף אל הר' האנרגיה הגבוהה.

חשיבה ויצירה באמצעות גרדיאנט

הפרנסקייטה הגלובלית זו מה שמשחרר את ה"הוגה" (thinker) שבគורת המודל. מכיוון שה-EBT יוצר את הטקסט כולו, הוא מוציאן במשימות הדורשות חשיבה הוליסטית ועמידה באילוצים, תחומיים שבהם מודלים אוטו-רגressiveים נוטים להיכשל.

יצירת טקסט (generation), לעונת זאת, היא סיפור אחר. אי אפשר פשוט לדגם מתן משטי אנרגיה באופן ישיר. במקום צרי, המודל צריך למצוא את עובי האנרגיה הנמוכה. המחברים משתמשים בטכנית איטרטיבית בהשראת הפיזיקה שנקראת Langevin dynamics, סוג של דגימת MCMC (בערך). התהליך נראה כך:

1. מתחילהים עם סדרה של רעש טהור (טוקנים אקראים).
2. מחשבים את האנרגיה של הסדרה הזרבל זהה.
3. דוחפים קלות את הטוקנים בכךון שמחית את האנרגיה במידה המרבית (כלומר, יורדים במורד הגרדיאנט של פונקציית האנרגיה).
4. חוזרים על תהליך זה מאות פעמים.

באטיות, באופן איטרטיבי, הסדרה האקראית מעודנת, מתייצבת מההרים הגבוהים של האנרגיה מטה אל עמק נמור-אנרגייה, ומתגבש למשפט קוורנטי ובוני היטב. אמן תהליך זה איטי יותר מיצרה אוטו-רגרסיבית סטנדרטיבית, אך הוא מאפשר>Create_a_figure_to_visualize_a_stochastic_process_in_a_potential_energy_surface. מוקרטה ומודעת-גלובלית הרבה יותר.

מדוע הוא "לומד וחושב סקילאbial"?

המאמר מספק ראיות חזקות לכך שגישה זו הינה סקילאbial (ניתנת להרחבה). ככל שהמודלים גדלים, יכולתם לה辨ין בין רצפים טובים לרעים משתפרת, ואיכות הדגימות שלהם יוצרים עולה. חשוב מכך, המסגרת מבוססת האנרגיה היא גמישה להפליא. אין כבול עוד לחיזוי הטוקן הבא. רזה מודל שמייצר ביקורות קולניות חיוביות? פשוט הוסף "מקדם אנרגיה" נוספת למטרת האימון שמעניש סנטימנט שלילי. מודולריות זו הופכת את ה-EBT לכל' רב עצמה לייצור נשלטה.

עבודה זו מאלצת אותנו לבחון מחדש את יסודות המודלים הנוכחיים שלנו. היא מציעה שהדרך לבינה מלאכותית חזקה, קוורנטית ונשלטה יותר עשויה שלא להיות טמונה רק בהגדלה אינסופית של יכולת חיזוי הטוקן הבא, אלא במבנה מודלים שבונים שפה ברמה הוליסטית ופיזיקלית יותר.

<https://arxiv.org/abs/2507.02092>

זה לא רק מה אתם כתבים בפורומפט, אלא איפה

המאמר היומי של מ"ק: 06.08.25

Where to show Demos in Your Prompt: A Positional Bias of In-Context Learning

מאמר שנשאלו היום מראה ששני פשוט במקומות הדוגמאות בפורומפט יכול לשנות דרמטית את רמת הדיוק של המודל. הנה מבט מהיר על הכלל החבוי זהה באינטראקציה עם בינה מלאכותית. מהנדס פורומפטים אובייסיביים ל.cgi התוכן של הפורומפטים שלהם. אבל המחברים חושפים שהתעלמו ממשתנה קרייטי לא פחות: המיקום של אותן דוגמאות. המחבר הודה לך את התחום מעבר למשחקי ניסוי וטעה אל עבר מדע קפדי, והחידוש בו טמון בדיק ובעיטה השיטית שלו.

אנו ידוע שהסדר הפנימי של דוגמאות משנה, אך מאמר זה מציג הבחנה מכרעת: לא מדובר בערבוב הדוגמאות, אלא בהזזה כל גוש הדוגמאות, ללא שינוי, למיקומים שונים בתוך הפורומפט. המחברים מכנים תופעה ספציפית זו הטית DEMOS POSITION IN PROMPT (DPIP). כדי לחקור זאת, הם יצרו מסגרת שיטית הבוחנת ארבעה מיקומים קוניים: בתחילת או בסוף הנחיות המערכת, ובתחילת או בסוף הوذעת המשמש. גישה זו הופכת תכנית מוערפלת למדע שנitin להוכיח.

המחברים מסתכלים מעבר לדיק פשט על ידי מדידת PREDICTION-CHANGE המודד כמה תשובות בפועל מהתפקידים כאשר מבנה הפורומפט משתנה. זהה תרומה חיונית, מכיוון שהוא חושפת חוסר יציבות סמי. מודל עשוי

להיראות מדויק באותה מידת עם שני פרומפטים שונים, אך אחד מהם עלול לגרום להתנהגות בלתי צפואה לחלוטין.

המחקר רחב היקף, שכלל עשרה מודלים וسمונה משימות שונות, הוביל תוצאות ברורות ונימנות ליישום.

- אפקט הראשוניות הוא אמייתי: מיקום דוגמאות מוקדם בפרומפט (esp, ssp) מניב באופן עקבי דיוק גבוהה יותר ויציבות רבה יותר, עם שיפור של עד 6 נקודות דיוק.

- אזור הסכנה: הצבת דוגמאות בסוף (wue) היא לרוב הרסנית. היא גורמת לירידה משמעותית בביטויים ולתנדויות גבוהה, והופכת מעל 30% מהתשבות של המודל במשימות מסוימות של שאלות ותשובות, מבליל לשפר את נכונותן.

- אין פתרון קסם: המיקום האופטימלי אינו אוניברסלי; הוא תלוי בגודל המודול ובסוג המשימה. לדוגמה, בעוד שמודלים קטנים יותר מעדיפים דוגמאות בתחילת המודול, מודל גדול כמו LLAMA-70B מעדיף לעיתים קרובות שהדוגמאות יהיו קרובות יותר לשאלתה (wus).

המחקר מבאר: מיקום הדוגמאות שלכם אינו בחירה סגונית. זה פרמטר קריטי שיש לבחון ולהתאים. הסתככות על פורמט בירית מחייב עלולה לבזבז ביצועים ויציבות משמעותיים. לראשוונה, ישנה מפת דרכים ברורה להבנה ואופטימיזיה של המיד החינוי הזה בעיצוב פרומפטים.

<https://arxiv.org/abs/2507.22887>

המאמר היומי של מ"יק: 08.08.25

Efficient Attention Mechanisms for Large Language Models: A Survey

מנגנון self-attention הוא הלב הפועם של מודלי שפה מודרניים. הוא מעניק לטרנספורמרים את יכולתם העמוקה להבין הקשר על ידי כך שהוא מאפשר לכל טוקן לתקשר עם כל טוקן אחר ברצף. אך ככל זהה יש מהיר אדייר, כמעט בלתי אפשרי. דרישות החישוב והזיכרון של קשב עצמי גדולות באופן ריבועי ביחס לאורך רצף הקלט. צוואר בקבוק ייחד זה הגדר במשמעותו את האפקט של מה שאפשרי, והפרק חיסיבה בהקשר ארוך מתאים לאתגר גדול.

נעשו מאיצים מחקרים משמעותיים במטרה להתמודד עם "הסיבוכיות הריבועית", שהולידו מספר רב של פתרונות מגוונים ולעתים קרובות מבלבלים. סקירה של תחום זה לא רק מפרטת את השיטות הללו; היא מספקת טקסטונומיה חיונית ומהווה מפה לנינוח בין ה-properties המורכבות של ייעילות חישובית, expressiveness של המודול ואלגנטיות תיאורית. סקירה זו צוללת לעומק העקרונות המרכזיים המבנאים את התחום זה.

4 משפחות הייעילות

בבסיסו, האתגר הוא לקרב את מטריצת attention המלאה בגודל $N \times N$, מבלתי לחשב או לאחסן אותה במפורש. הסקירה מסוגגת את שלל הגישות לארבע משפחות עיקריות, שלכל אחת מהן פילוסופיה משלها.

1. דليلות בתבנית קבועה: התיקון הארצייטקטוני

הגישה השרה ביותר לשבירת צוואר הבקבוק הריבועי היא להניח שמטריצת attention צפופה של "הכל-להכל" היא מוגזמת. שיטות אלו קופות תבניות attention דיללה וקבועה מראש, שבה כל טוקן רשאי להתייחס רק לתת-קבוצה קטנה וקבועה של טוקנים אחרים.

משפחה זו כוללת שיטות המשתמשות windows sliding, שבהן טוקן מתיחס רק לשכני המוקומים. גישה זו מבוססת על האינטואיציה החזקה של "מקומיות ההקשר" (locality of reference) שAMILIM סמוכות הן לרוב הרלוונטיות ביותר. כדי למנוע אובדן של מידע גלובלי, גישה זו מחזקת לעיתים קרובות באמצעות מספר טוקנים גלובליים הרשאים להתייחס לכל הרצף, או באמצעות תבניות מורחבות/mdlgoths (dilated/strided patterns) המدلגוות באופן שיטתי על טוקנים כדי לכוסות שדה קליטה רחבה יותר עם מספר קבוע של חישובים.

שיטות אלו יעילות מאוד ופשטות ליישום, אך מגבלתן העיקרית היא הנוקשות שלהן. תבניות ה-attention מהונסנות ידנית ואין תלויות בנתונים, מה שאומר שהמודול אינו יכול להחליט באופן דינמי להתמקד בטוקן מרוחק אך רלוונטי מוחץ לחloan שנקבע לו מראש.

2. קירוב מדרגה נמוכה (Low-Rank): טרייק הדחיסה

משפחה זו של שיטות פועלת על בסיס תבונה מתמטית עדינה יותר: שמטריצת ה-attention המלאה היא לרוב מדרגה נמוכה, כלומר ניתן לדוחס ביעילות את המידע שבה במספר קטן בהרבה של "מושגים" או וקטור סיכום. במקום לחשב את המטריצה המלאה, מודלים אלו מעריכים את מטריצות השאלתה, המפתח והערך (Query, Value, Key) לתת-מרחב בעל ממד נמוך יותר, ובכך מאלצים את מנגנון ה-attention לפעול דרך צוואר בקבוק של מידע.

הרעין המרכזי הוא לקרב את מטריצת ה- $N \times N$ על ידי פירוקה למכפלה של שתי מטריצות קטנות יותר, בגודל $N \times k$, כאשר k קטן משמעותית מ- N . במקרה, המודל לומד לסכם את כל הרצף במספר קבוע של צמדי מפתח-ערך מייצגים, וכל הטוקנים מתייחסים לסטים דחוסים זה במקומות זה זהה. זהה גישה גמישה יותר מtabnion קבוצות, שכן תוכן הסיכום הדחוס נלמד מהנתונים. עם זאת, הדבר מציג פשרה חדשה: הגודל הקבוע של צוואר הבקבוק מגביל את קיבולת המודול להתחום עם רצפים בעלי ציפויות גבוהה מאוד של מידע יחודי.

3. קורנלייזציה (Kernelization): תעלול מתמטי קלי

אول הפתרונות האלגנטיים ביותר מבחינה מתמטית הם אלו הממסגרים מחדש את ה-attention דרך עדשת שיטות הkernel (kernel methods). ניתן לראות את ה-attention הסטנדרטי כתהיליך של חישוב מטריצת דמיון בין שאלות למפתחות, ולאחר מכן שימוש במטריצה זו כדי לשקלל את הערכיהם. הבעיות הריבועית נובעת מהבנייה המפורשת של מטריצת דמיון מסיבית זו.

שיטות מבוססות קורנל עוקפות זאת בתוכום על ידי מינוף התוכנה האסוציאטיבית של כפל מטריצות. הן מנוסחות מחדש את חישוב ה-attention כך שישלבו תחילת את המפתחות והערכים, לפני האינטראקטיה עם השאלות. שניINI סדר הפעולות פשוט זהה מונע את יצירת המטריצה בגודל $N \times N$. במקום מכפלת מטריצה-במטריצה

גדולה, החישוב מצטמצם לשתי מכפלות קטנות יותר של מטריצה-בוקטור, מה שמוריד את הסיבוכיות מריבועית לילינארית.

גישה זו חזקה מכיוון שבתיאוריה, היא יכולה לקרב את מגנון ה-*attention* המלא מבל' לכפות אילוצי דليلות נוקשים. יעילותה תליה במציאות פונקציית קרNEL שתופסת במידוק את הדמיון בין שאלות למפתחות, וחלק גדול מהמחקר בתחום זה מתמקד בפתרונות פונקציות קרNEL חדשים (לרבות באמצעות טכניקות כמו קירוב תכונות אקראי) שהן גם ייעילות וגם בעלות יכולת ביתוי גבוהה.

4. דليلות נלמדת -*Mixture of Experts*: הגישה האדפטיבית

משפחה רבעית ומפתחת שואפת להשיג את הטוב מכל העולמות על ידי הפיכת תבנית הדليلות עצמה לתלוית-נתונים ונלמדת. במקום להשתמש בתבניות קבועות או בצוואר בקבוק גלובלי מדרגה נמוכה, שיטות אלו מנוטות לחזות אילו טוקנים הם הרלוונטיים ביותר עבור שאלתה נתונה. הדבר מושג לעתים קרובות באמצעות טכניקות כמו אשכולות (*clustering*) או על ידי שימוש במסגרת MoE, שבה ראשי *attention* שונים מאומנים כ"מוחחים" לסוגים שונים של תבניות. מגנון ניטוב לומד לשloh כל טוקן בראש המומחה הרלוונטי ביותר. גישות היברידיות אלו הן בין החזקות והגמרישות ביותר, אך גם המורכבות ביותר ליישום ואיומן.

לסיכום, הסקירה חושפת שאין מגנון *attention* יעיל אחד שהוא "הטוב ביותר". כל משפחה מציגה בחירה מהותית לגבי אופי הקירוב, ובמצעת פרשה שונה בין סיבוכיות חישובית לכוח ביתוי. התהום הוא דיאלוג תוסס בין הנחות יסוד ארכיטקטוניות, תורת הקירוב המתמטית ומערכות אדפטיביות ולומדות.

<https://arxiv.org/abs/2507.19595>

12.08.25 המאמר היומי של מייק: Your LLM Knows the Future: Uncovering Its Multi-Token Prediction Potential

AIR ניתן לגנרט טוקנים בצורה מקבילית אבל בלי מודלי שפה מבוסס דיפוזיה.

מאמר זה קורא תיגר לגנרט אוטורגרטיבי של LLMs ומציע שיטה שמאמת מודל לחזות כמה טוקנים בו זמןית לעומת MTP (Multiple Token Prediction). כאמור MTP מאמין לחזות כמה טוקנים בו זמןית להבדיל מ-NTP או Next Token Prediction שחזזה כל עם טוקן יחיד. בנוסף הגישה המוצעת משלבת שימוש בינה שנקרא פענוח ספקולטיבי או Speculative Decoding, בעוד הרצאות בכמה נסיטים ומיטאים לאחרונה. בנוסף יש גם שימוש בטכנית fine-tune של מודלים (בד"כ מבוסס טרנספורמרים) הנקראת LoRa (Low Rank Adaptation).

אוקי, אז קודם כל המחברי מאמנים כמה ראשי decoding (למייט הבנתי שכבה אחת בלבד) עברו כל טוקן שנחזה פרט לטוקן הבא שנחזה באופן סטנדרטי כמו ב-NTP. בשבייל לחזות את הטוקן הבא המחברים משתמשים לא רק ביצוג הקונטקסטואלי שלו אלא גם ביצוג הלא קונטקסטואלי (מיליון האמבידיג) של הטוקן הקודם (שניהם משורשרים ומוסברים דרך MLP בעל שתי שכבות).

בנוסף המאמר מאמן LoRA (מטריצות נוספות למשקלות של שכבות הלינאריות של הטרנספורמר) אבל משתמש בהם רק כדי לחזות את הטוקנים מעבר לטוון הבא. במאמר שיטה זו נקראת Gated LoRA. שיטה זו ניתן לאמן בצורה מקבילה בדומה לאיר שאנו מאמנים NTP סטנדרטי.

הגישה האחרונה הנדונה במאמר היא פענוח ספקולטיבי או SD. בגדול SD הינה משפחה של טכניקות לשיפור מהירות הגנרטים השומרת על התפלגות הגנרטים כמו בגנרט אוטורגרטיבי (כלומר עם NTP). בד"כ משתמשים במודל חלש ומהיר יותר (לפעמים מודל זהה הוא חלק מהמודל שהוא רוצים ליעל) לגנרט של כמה טוקנים ואז בודקים אותם עם מודל היעד באופן מקבלי. הטוקנים שיעברו את הבדיקה בהצלחה מתקבלים וככה אנו יכול לקבל גנרט מהיר יותר.

כאן במקומם המודל הגדול משתמשים בגנרט מקבלי של כמה טוקנים דרך NTP, מעבירים להם את הבדיקה וככל שייתר טוקנים עוברים אותה, אנו מקבלים גנרט מהיר יותר. בנוסף המאמר מציע להמשיך לגנרט עם NTP עד K טוקנים (K הוא מספר הטוקנים המוגנרטים עם NTP). עם כל K הטוקנים הראשונים עוברים את הבדיקה אנו ממשיכים את תהליך הבדיקה עם K הטוקנים הבאים שעתיד לזרץ את קצב הגנרט עוד יותר.

מאמר קليل יחסית וכתוב היטב - מומלץ.

<https://arxiv.org/abs/2507.11851>

13.08.25 המאמר היומי של מ"יק: Checklists Are Better Than Reward Models For Aligning Language Model

תשחו מכל מה שידעתם על מודלי תגמול: האם צ'קיליסט פשוט הוא העתיד של עולם ה-AI?

בשנים האחרונות, פרדיגמה יחידה שליטה במאצינו לגורם ל- LLMs להתנהג לפי ה"חוקים": למידת חיזוק ממשוב אנושי (RLHF). בלב גישה זו נמצא מודל התגמול (RM), רשת נירונית חזקה אך לא ניתנת לפרשנות, שאומנה לזרק את "המשתח המורכב והמבולגן של העדפות אנושיות" לכדי ציון סקלרי יחיד. לאחר מכן, משתמשים בכך זה כדי להנחות את ה-LLM שלנו להתנהגות "טובה". אך כל התנהגות זהה נשען על הנחה שברירות: שמספר נלמד בלבד יכול ללווד באופן מהימן את האופי הרב-ממדית של ערכים אנושיים.

מאמר חדש, אינטואיטיבי ודי מבריק קורא תיגר על הנחת יסוד זו. המחברים טוענים שברדייפה אחר ציון יחיד, בינוינו מערכות שנן לא רק קופסאות שחורות, אלא גם נוטות ל-*reward hacking* ובנוסף לא ניתנות לפרשנות. החלופה שהם מציעים אינה מודל מרכיב יותר, אלא תנועה לעבר פשטוט ויכולת פירוש (interpretability). על ידי שילוב של צ'קיליסטים מובנות עם שיטת DPO, המאמר מشرط נתיב חזק, יעיל ומאמין יותר ל-*alignment* של המודלים.

החדשון המרכזי הראשון הוא המעבר מtagmol מרומז וסקלרי לתגמול מפורש וUMBOSO-וקטור. במקום לאמן מודל תגמול לפתח "תחושא" אינטואיטיבית לגבי מה שבני אדם מעדיפים, המחברים מציעים להעיר את הפלט של המודל אל מול רשימה תיוג מובנית של תוכנות רצויות וモוחשיות.

דמיינו שאתם מעריכים תגובה לא באמצעות ציון בודד מ-1 עד 10, אלא אל מול רשימה של קритריונים ביןאריים או מרובי-רמתות:

- האם התשובה נכונה עובדתית? (כן/לא/חלקית)
- האם היא נמנעת מסטריאוטיפים מזיקים? (כן/לא)
- האם הטעון עוזר ואני מתנסה? (כן/לא)
- האם היא מצטטת מקורות אמיתיים, אם רלוונטי? (כן/לא)

פרק זה הוא המפתח. הוא הופך את המשימה המאוד מורכבת של מידול העדפות לסדרה של בעיות סיווג מוגדרות יותר וניתנות לאימוט, לשיעיותם קרובות מבוצעות על ידי מודל שפה אחר (לולמר Judge-a-LLM-as-a-judge). אך הדבר מעלה שאלה: כיצד הופכים הערכה וקטורית זו לסיגナル אימון נקי וסקלירי לעדכון המודל? CAN נכנס החידוש השני של המאמר. הצליליסט אינו משתמש כפונקציית תגמול ישירה. במקום זאת, המחברים משתמשים בה כפונקציית תיוג אוטומטית ועוצמתית ליצור זוגות העדפה עבור DPO. שיטת O DPO מבצעת ידי פין טין של מודל השפה על זוגות של תשובות מועדות ותשבות לא מועדות. המאמר משתמש ברישימת התיוג כדי ליצור זוגות אלה באופן אוטומטי, ובכך לבטל את הצורך בתיאוג אנושי יקר או במודל תגמול נפרד.

תהליך האימון הופך לlolאה איטרטיבית ו עצמאית:

1. יצירה (Generate): עבור פרומפט נתון, המודל המאמן מייצר שתי תשובות מועמדות או יותר.
2. הערכתה (Evaluate): מודל ה"שופט" מעריך כל תשובה אל מול רישימת התיוג, וקובע איזו מהןעונה טוב יותר על הקритריונים המפורשים.
3. צימוד (Pair): בהתבסס על הערכתה זו, התגובה העדיפה מתויגת כנבחרת (w_y) והאחרת מתויגת כנדחית ($\neg y$).
4. פיניטיון: זוג ה- $(\neg y, w_y)$ שנוצר זה עתה משמש כדוגמה בודדת לעדכון המודל המאמן באמצעות פונקציית הלוא של DPO.

שיטת אלגנטית זו פותרת מספר בעיות בבוחן אחת. היא עוקפת את הצורך לאמן מודל תגמול מונוליטי, ובמקרה זאת שואבת את סיג널 העדפה שלא מרשימה התיוג השקופה והניתנת לעריכה. מכיוון שהנתונים נוצרים תוך כדי תנועה, נוצרת "תוכנית למודים" דינמית ומתקנת את עצמה, שnitן לכוון בזמן אמת פשטוט על ידי שינוי הקритריונים ברישימת התיוג.

הניסויים של המחברים נועד לא רק לכחש את טבלאות הבנצטראקים, אלא לבחון רוביוטיות של השיטה. הם מראים שבעוד שיפור סטנדרטי מבוסס-RM יכול להשיג ציונים גבוהים במידדי ביצוע ספציפיים, מודלים אלה הם לעיתים קרובות שביריים. הם "מנצחים לרעה חוק גודהארט", והופכים למצטיינים באופטימיזציה של הפרויקט (ציוו התגמול) על חשבון המטרה האמיתית.

לעומת זאת, מודלים שיושרו בשיטת Checklist-DPO מפגינים רוביוטיות רבה יותר. מכיוון שהם מותאמים לעמוד במערך מגוון של קритריונים מפורשים, יש להם פחות סיכוי למצוא "פריצה" יחידה ופושטה. הם חיברים להיות טובים במספר דרכיהם הניננות לאימוט. המאמר מראה שמודלים אלה עמידים יותר להנחות אדברסראיליות, פחות מתרפים (sycophantic), ומקפידים באופן אמין יותר על מגבלות בטיחות, גם בתרחישים שחורגים מנתוני ההתפלגות המקוריים.

<https://arxiv.org/abs/2507.18624>

מלך הוא עירום: למה מודלי שפה נכשלים בחשיבה אלגוריתמית אמיתית

14.08.25 המאמר היומי של מיק:

FormulaOne: Measuring the Depth of Algorithmic Reasoning Beyond Competitive Programming

מאמר  לבן של כמה מחברים שאנו מכיר באופן אישי....

במרחב הבלתי פוסק לעבר בינה מלאכותית כללית (AGI), היכולת של LLMs לחשב בצורה אלגוריתמית גותרת חזית קרטית ושניה בחלוקת. במשך שנים, מدد הביצועים (הبنץ' מארק) העיקרי שלנו היה תכונות תחרותי – תחום פופולרי מאוד ששימש כאינדיקציה לא רעה לסוג מסוים של חשיבה חישובית. אבל ככל שהמודלים שלנו הופכים מתוחכמים יותר, עולה שאלת השאלה האם אנחנו עדין מודדים את הדבר הנכון? אני מדברים הרבה בנץ' מארקים לאחרונה בצורה ביקורתית וכן החלטתי לסקור את המאמר.

מאמר חדש וכחול לבן למעשה קורא לשינוי פרדיגמה. החוקרים מציגים בנץ' מארק חדש שנועד לבדוק את עמק החשיבה האלגוריתמית, מימד שלטענתם נעלם כמעט לגמרי מסגרות ההערכה הקיימות. בעוד שמודלי שפה מראים תוכאות מרשים על בנץ' מארקים כמו תכונות תחרותי, הם בעיקר פוטרים בעיות שניתן לפתור על ידי שילוב של מספר אלגוריתמים מוכרים. אנחנו מתעניינים ביכולת לפתור בעיות הדורשות תהליך חשיבה עמוק ויצירתי יותר.

זה בדיקת הנקודת. הבנצ' מארקים הנוכחיים בודקים את יכולתו של מודל לגשת ולישם את ספריית הפתרונות המוכרים העצומה שלו. FormulaOne שואל שאלה עמוקה יותר: האם מודל יכול לחשב כמו מדע מחשב?

מעבר לאזרור הנוחות של התכונות התחרותי

פלטפורמות כמו LeetCode ו-Codeforces היו בעלות ערך אדיר. הן דחפו את גבולות היכולת של המודלים. עם זאת, כפי שמצוין של-שוורץ, הן מפותחות סוג מאד מסוים של פתרון בעיות; זהה המבוסס על זיהוי תבניות (pattern recognition) ורקבומיבנזיה. המאמר על FormulaOne מותח ביקורת מרוםזית על הפרדיגמה זו ומצביע על מגבלותיה:

- מיקוד במהירות: תכונות תחרותי מתגמל לרוב את הפתרון הנכון מהירות ביותר, לא בהכרח את האלגנטית. ביותר או זה שניתן להכללה.
- חשיבה שטחית: בעיות רבות הן וריאציות על נושא מסוים, הניתנות לפתרון על ידי זיהוי תבנית ויישום אלגוריתם סטנדרטי. זה בוחן את "אוצר המילים האלגוריתמי" של המודל, לא את יכולת החשיבה שלו.
- "קבים" של DATA האימון: קיימת סבירות גבוהה שפתרונות לביעות פופולריות רבות מסתתרים אי שם בתוך נתוני האימון (training data) של המודל, מה שמקשה על הערצת יכולת פתרון בעיות אמיתית ומקורית.

אתגר ה-FormulaOne: סוג חדש של בנץ' מארק

כאן שואנו נכנס לתמונה. זה לא רק דאטסהט חדש; זו פילוסופיית הערכה חדשה המטרתה היא למדוד את עומק החשיבה הנדרש כדי להציגו אלגוריתם חדש לגמרי.

החוקרים משיגים זאת באמצעות גישה "מתמטית" ומתוחכמת, הממנפפת מושגים מתחום הסיבוכיות הפרטנית (parameterized complexity) וטורת הגרפים (graph theory) כדי לייצר בעיות עם שיפור קושי מבוקר ומדויק. אחד הכלים המתמטיים המרכזיים שהם משתמשים בו הוא רוחב-עץ (treewidth) של הפתרון, מدد לכמה גרף הוא "דמי-עץ". בעיות עם treewidth נמוך ניתנות לרוב לפתרון באמצעות תכונות דינמי, אך ככל שה- treewidth עולה, הייצוריות האלגוריתמית הנדרשת נזקפת.

זה מאפשר להם ליצור בעיות שנראות פשוטות באופן פני השטח, אך דורשות תובנה עמוקה ולא מובנת מלאה. הם מכונים לביעות שבhn "הפתרון הוא תוכנית פשוטה לכתיבה, אך הדרך לגילוי התוכנית הזה היא מורכבת ומסועפת".

כדי להגדיר זאת באופן פורמלי, הכוון משתמש ב-לוגיקה מסדר שני מונודית (MS). זהה מוגרת לוגית חזקה המאפשרת להם להגדיר תכונות של גרפים וליצור באופן אוטומטי סט עצום ומגוון של בעיות. באופן קרייטי, תהיליך יצירת הדטה הסינטטי הזה מבטיח שהבעיות הן חדשות ואין מופיעות בשום דатаה אימון, מה שמאפשר את המודלים לחשב מהעקרונות הראשוניים.

התוצאות המפכחות והדרך קדימה

מצאי המאמר הם קריית השכמה. בעוד שמודלים חזקים ביותר כרגע כמו GPT-4 ו-Opus-Claude 3 (ולихה 5gpt) מראים יכולות מסוימות, ביצועיהם על בעיות OneFormula נמוכים משמעותית מאשר על ביצ'מרקם מסורתיים. זה מדגים באופן ברור כי התאמת תבניות לחשיבה עמוקה ואמיתית. המודלים מתקשים בדיק בנקודה שבה נדרש גילי אלגוריתמי יצירתי ורב-שלבי.

זהו התבונה החדה והholesטייתiaeOneFormula. זה לא עוד Leaderboard שצריך לטפס בו; זהו כל' אבחוני שחשוף את המוגבלות הנווכיות של ה-LLMs שלנו. המאמר מציע שפוץ להציג ארכיטקטורות קיימות ודטה אימון אولي לא יספיק כדי לגשר על התהום ל-AI. علينا להתמקד בארכיטקטורות ובשיטות אימון המפותחות פתרון בעיות יצירתי ואמיתי.

OneFormula מספק נתיב קונקרטי וمبוסס מתמטי למדוד את ההתקדמות שלהם. הוא מאתגר את קהילת ה-AI לצאת מאזור הנוחות של בעיות מוכרות ולהתחל להתמודד עם האתגר הקשה הרבה יותר, והחשוב הרבה יותר, של למד את המודלים שלנו איך לחשב. המירוץ החל.

וכמו שאמרתי, כבר לכמה אנשים לדעתי העtid הוא לא מודלים חכמים בצורה מטורפת אלא המודלים שיודעים להפעיל כלים בצורה מטורפת

<https://arxiv.org/abs/2507.13337>

מעבר למילימ: למה Large Action Models הם הצד האמתי אל AI שפועל בעולם

16.08.25 המאמר היומי של עמרי ומיק: Large Action Models: From Inception to Implementation

מה זה (LAM) ואיך זה שונה מ-LLM? שורה תחתונה: LAM הוא LLM, אבל זה שואמן והותאם במיוחד להפיק פעולות ברות-ביצוע בסביבה אמיתית. בעוד שמודול LLM רגיל מזמין להפיק טקסט איקוטי ועקבתי, LAM מזמין לייצר תוכניות ופקודות שניתן להפעיל בפועל דרך agent, בין אם זה קלייק, הקלדה או קריית API, כך שהוא משפיע ישירות על מצב העולם ולא רק "דבר עליו".

מה שהכותבים מציעים הוא שבמקום לחבר LLMים לסייע agentים, יש לחבר **LAM** שלמעשה משמש כמנוע קבלת החלטות בתוך הלולאה של ה-agent: ה-agent אוסף תכפיות מהסבירה (למשל מצב מסך, רשימת כפתורים זמינים או נתונים API), מזין אותן ל-LAM, וה-LAM מחדיר את הפעולה הבאה לביצוע. ה-agent הוא זה שמבצע בפועל את הפעולה ומהזיר חייו על התוצאה וזה מה שמאפשר ל-LAM לעדכן את החלטות הבאות.

כאן בדיק טמון הבדל הクリטי גם ברמת הסיכון, כפי שהכותבים רואים זאת: טעות של LLM "קלאסי" מתחבطة לרוב בתשובה שגואה או בהזיה (hallucination) - פגעה בהבנה או באמון, אך בלי השלכות ישרות בעולם האמתי. לעומת זאת, טעות של LAM עלולה לגרום לשינוי ממש: מחיקת קובץ חשוב, שליחת הודעה לכתובה הלא נכון, או ביצוע פעולה עסquit לא רציה.

האינטראקטיה עם הסביבה שבה החוקרים פעלו נעשתה ב-Windows בלבד, במשימות ממוקדות ב-Word. הם חיברו את ה-LAM אל **UFO**, סוקן GUI ייעודי ל-Windows. הסוקן קורא את מצב המשתמש (status) שזה רשות הבקרים (Controls) (עם סוג, כוורת ואינדקס ומעבר את המידע ל-LAM להכרעה, ולאחר מכן מבצע את הפעולה (action) שנבחרה: לחיצת עכבר, הקלה, או קראת API).

התהילך שהחוקרים מציעים בניו מ-5 שלבים: Data → Training → Integration & Grounding → Offline → Eval → Online Eval. לאחר המאמר ישנה הפרדה בין **Task-Planning** לבין **Task-Action**: בשלב איסוף הנתונים **Task-Plan** → **Plan**, ולאחר מכן הם יוצרים מסלולים (trajectories) שהופכים את הצעדים האליהם פועלות קונקרטיות בסביבת Word: בחירת כפטור ספציפי, הגדרת סוג פעולה ופרמטרים כך שה-agent יכול להריץ אותם בפועל ולבוחן הצלחה או כישלון. לתהילך זהה הם קוראים **Grounding**: עיגון הפלט הטקסטואלי של המודל ל-UI אמיתי ולפעולה אופרטיבית דטרמיניסטית.

ב-**LAM1** המודל אומן ב-**SFT** על **Task** → **Plan** בלבד ($t_i \rightarrow s_i$). הכותבים מסבירים שהאינטראקטיה כאן היא למדת קודם את המודל לפרק משימות בזרה הגיונית ומוסדרת לפניו שניות לבחירת פעולה. לשם כך השתמשו בכ-76.7 דוגמאות מקוריות כמו מדריכי עזרה, How-To ושאלות היסטוריות, שערכו ניקוי, עיבוד והבשלה כדי להבטיח עקביות ואיכות.

ב-**LAM2** המיקוד עבר ל-**Action** → **State**, חיקוי מסלולי הצלחה של ($a \rightarrow s$) GPT-4o. כאן כל דוגמה מייצגת מצב הנוכחי (state UI) כפי שנקלט על ידי agent המשימה ומוסדרת בקרים (Controls) עם סוג, כוורת ואינדקס ביצירוף טקסט המשימה, והפעולה המדויקת שבוצעה בפועל: בחירת הבקר הנכון, סוג הפעולה והפרמטרים. את מסלולי הצלחה יוצרים מתוך מאגר ה-**Task** → **Plan** של **LAM1**, תוך הפיכת הצעדים הכלליים לפקודות ממוקדות על רכיבים אמיתיים ב-Word, הרצה ובדיקה בסביבה החיה, וסינון לפי הצלחה בפועל. גם בשלב זה אומן ב-**SFT**, כשהדאטה הטעטה הכליל בסופו של דבר **2,192** מסלולים מוצלחים (trajectories) ששימושם מבוסיס לאימון.

ב-**LAM3** המשיכו ב-**SFT** על ($a \rightarrow s$) **State** → **Action**: **Self-Boosting**: ללקח מסלולי כישלון של GPT-4o, נתנו למודל שנאמן ב-**LAM2** לנסות נוספת, ווסףו את ההצלחות החדשנות שיצר. כך נוצר דאטה נוספת וaicotti ללא אנטציה ידנית, שהרחיב את יכולות המודל גם על מקרים קשים יותר.

ב-**LAM4** עברו משלב ה-**SFT** ל-**RL**, וביצעו **Offline PPO** המונחה על-ידי **Reward Model**. את ה-**Reward Model** בנו על בסיס **LAM3**, בתוספת **ocab** שמחזירה ציון הצלחה לכל פעולה, כשהמודל אומן ב-**LoRA** על מסלולי הצלחה וכישלון. לצורך האימון, כל צעד במסלול מוצלח קיבל ציון +1 וכל צעד במסלול כשל קיבל ציון -1, וה-**RM** אומן עם **MSE** כדי לחזות את הציון הזה.

עם RM מוקן, השתמשו בו כדי לאמן את **LAM4** ב-**Offline PPO**, כשההתמקדות הייתה דזוקא על 1,788 מסלולי ההצלחה שנאספו ב-**LAM3** – במטרה “ללמידה מהטעויות”. כאן הפורמט הוא ($s \rightarrow a \rightarrow r$), כאשר ה-**RM** מספק את ה- r , והמודל לומד לשפר את בחירת הפעולות מעבר למה שנלמד בחיקוי ישיר.

לאורך המאמר מוצגות שלוש מדידות: **תכנון** (Planning), **פעולות אופלין** (Offline Eval) והרצות חיים (Eval). בשני הראשונים נבדקו ההצלחות ברמת תכנון המשימה והצעדים, וכן דיקרים בבחירה אובייקט ופעולה,

והמודלים התקדמו בהדרגה מרמה תחרותית ועד שיפורים עקביהם. בשלב השלישי – ההרצאות בסביבת Windows Word – נמצא כי LAM טקסטואלי בלבד היה תחרותי מול GPT-4, ואף עקף אותו בחלק מהمدדים כשהשו קונפיגורציות טקסטואליות בלבד. לעומת זאת, כאשר ל-GPT נוספה גם יכולת vision, שיעורי ההצלחה היו גבוהים יותר, אך המחיר היה ירידה בมหาירות וביעילות.

אנו מניחים שככל שagents יהפכו ליותר נפוצים ובעלי יכולות, נראה עוד ועוד עבודות בסגנון זהה – ככל שמחברים מודלים לסבירות אמיטיות ומבצעות אימון עם דатаה ייעודי ואדפטציה למשימות, לא בטוח שאימון LAM בשלושה שלבי SFT ואחריהם שלב RL יחיד הוא המתקון האופטימלי, אבל הכוון של להפוך LLMים ליותר ממקדי-משימה, עם אימון מובנה ומתאפס-דומין, הוא צעד מתבקש בעינן שבו יותר ויותר agents יפעלו בעולם האמיתי.

<https://arxiv.org/pdf/2412.10047.pdf>

לאלף את החיים: הטרנספורמרים סוף סוף תחת שליטה מתמטית.

המאמר המקורי של מייק: 19.08.25

Training Transformers with Enforced Lipschitz Bounds

בעולם הלמידה העמוקה, אנו מודדים את ביצועי המודל על כמה בניםארקים פופולריים ושמחים כאשר המודל מגין ביצועים גבוהים עליהם. עם זאת, מתחת לפני השטח של היישגים מרשיימים אלה מסתתרת בעיה עיקשת שלעיתים קרובות מתעלמים ממנה: חוסר יציבות. כל מי שאינו מודל טרנספורמר גדול נתקל בוודאי בתסכול של גרדיאנטים מתפוצצים או נעלמים, לצורך בתכנון עדין של קצבי למידה, ובערך ה-" NaN " המסתורי בפונקציית הלוג שיכל לשבש ריצת אימון שלמה. סוגיות אלה מצביעות על חוסר שליטה יסודית בהתנהגות המודל.

המאמר שנסקור היום מציע פתרון נחמד לבעה זו. במקום להסתמך על אוסף של טרייקים אמפיריים, המחברים מציגים מתודולוגיית אימון חדשנית שאוכפת תוכנה מתמטית הידועה בשם תנאי ליפשיץ. גישה זו לא רק מרסנת את חוסר יציבות של הטרנספורמר, אלא גם מוביילה לשיפור ביצולת ה הכללה ורוביוטיות של המודל. באו נצלול לעומק החידושים המרכזיים של עבודה מרתתקת זו.

בבסיסו, תנאי ליפשיץ הוא מגד ל"חלקוות" או ל"רגשות" של פונקציה. פונקציה עם קבוע ליפשיץ קטן אינה יכולה להשתנות מהר מדי; שינויים קטנים בקלט יובילו רק לשינויים קטנים בפלט. על ידי אכיפה חסם ליפשיץ על רשת נירונית, אנו למשה מציבים "מגבלת מהירות" על מידת השינוי בפלט המודל כתגובה להפרעות בקלט שלו.

זהו רעיון רב עצמה. בהקשר של טרנספורמרים, משמעות הדבר היא שאנו יכולים לשנות ברגישות של כל רכיב במודל, מנגנון ה-attention ועד לשכבות FFN לשילטה מדוקת ועדיינה זו יש השלכות עמוקות על יציבות האימון וביצועי המודל. כדי לאכוף את תנאי ליפשיץ, המחברים מציעים סדרה של שינויים חדשניים בארכיטקטורת הטרנספורמר הסטנדרטית. לא מדובר בתיקונים קלים, אלא בתכנון מחדש עקרוני של רכיבי הליבה של המודל:

- שכבות עם נרמול ספקטרלי (Spectrally Normalized Layers): המחברים מישמים נרמול ספקטרלי על מטריצות המשקولات הן במנגנון ה-attention והן ב-FFN. טכניקה זו נבחרה בשל דיוקה המתמטית: הנורמה הספקטרלית של מטריצת משקولات שווה בדיקון לקבוע ליפשיץ של אותה שכבה לינארית. הדבר מאפשר שליטה ישירה והדוקה ברגישות המודל בכל שלב.
- בלוקים של רשת FFN שהם 1-ליפשיצ': חידוש מרכזי הוא האופן שבו המאמר מטפל בא-הLINאריות של ה-FFN. המחברים מראים כיצד לבנות את כל בלוק ה-FFN כך שהיא 1-ליפשיץ' באמצעות פונקציות אקטיבציה סטנדרטיות כמו ReLU או GeLU. הדבר מושג על ידי שילוב של מטריצות משקولات

מנורמלות ספקטרליות עם טיפול בפונקציית האקטיבציה, מה שמבטיח שהטנספורמציה השלמה בתוך הבלוק עומדת באילוץ ליפשיץ המחמיר.

- **חיבור שרארית (Residual Connections):** המחברים מספקים גם ניתוח عمוק של חיבור השARING, שהם יסודים בארכיטקטורת הטנספורמרא. הם מדגימים כיצד לשנות את קנה המידה (scaling) של נתיבי השARING כראוי כדי להבטיח שהוספטם אינה מפרה את תכונת הליפשיץ של המודל כולו. הרכבה זהירה זו של רכיבים חסומים באופן מוכח היא שמאפשרת לרטון את ארכיטקטורת הטנספורמרא כולה.

חידושים ארכיטקטוניים אלה, ייחודי, יוצרים סוג חדש של טנספורמרא שהוא, בעצם תכנונו, יציב וממושמע יותר מקודמי. היתרונות של הטנספורמרא מרוסן-ליפשיץ ניכרים מיד במהלך האימון. המחברים מדגימים שהמודל שלהם יציב באופן יוצא דופן, אפילו ללא צורך בNORMALISATION שכבה (Layer Normalization), רכיב שנחשב לעיתים קרובות חיוני עבור טנספורמרים סטנדרטיים.

יציבות זו מאפשרת תהליכי אימון פשוט וחזק יותר. המחברים מראים שניתן לאמן את המודל שלהם עם קצבי למידה גדולים יותר ושהוא פחות REGARDLESS לחריש היפר-פרמטרים. הדבר לא רק הופך את תהליכי האימון לעיל יותר, אלא גם פותח דלת לאפשרויות חדשות להגדלת מודלי טנספורמרא. היתרונות של אכיפה חסמי ליפשיץ חריגים מעבר ליציבות האימון בלבד. המחברים מדגימים גם שהמודל שלהם מציג יכולת הכללה ועמידות משופרת:

- יכולת הכללה טובה יותר: אילוץ ליפשיץ פועל כזרה של רגולרייזציה לא מפורשת רבת עצמה, המונעת מהמודל לבצע אובייקטיבי לדאות האימון. הדבר מוביל לביצועים טובים יותר על תנאים שלא נראה בעבר.
- עמידות מוגברת למתקפות אדברסריאליות (Adversarial Attacks): על ידי הגבלת רגישות המודל להפרעות קטנות בקלט, אילוץ ליפשיץ הופך את המודל לעמיד יותר באופן אינהרנטי למתקפות אדברסריאליות. המחברים מראים שהמודל שלהם חסין למתקפות אלה באופן משמעותי יותר מטנספורמרים סטנדרטיים.

<https://arxiv.org/abs/2507.13338>

כל האחד לריפוי אמנה של בינה מלאכותית: צילילת עומק

21.08.25 המאמר היומי של מיק: Scaling Laws for Forgetting When Fine-Tuning Large Language Models

כל מי שאי פעם ביצע פיניטיון למודל שפה חזק מכיר את הפשרה הכהבת. אתה מתאים את המודל למשימה חדשה, ובתוך כך הוא מפתח סוג של אמנה ושותח את הידע הכללי שהוא כל כך יקר לרכוש. "שכחיה קיטסטורופלית" זו היאאתגר בסיסי. תרופה נפוצה היא לערוב כמהות קטנה מנתוני האימון-המקדים המקוריים במהלך הכוון העדיין, אך זה תמיד הרגיש יותר כמו תרופה שבתא מאשר מדע.

מאמר שנסקור היום מרים את الطريق הזה למדרגה של מדע מדויק ונitin לחיזיו. המחברים עושים הרבה מאשר רק לציין ש"הזרקת נתונים עוזרת". הם מציגים מודל חיזוי מדויק המתאר את הריקוד המורכב בין גודל המודל, כמהות DATA של פיניטיון, ואחווד DATA מהאימון המקורי המזורך לתוכו. בעוד שהכחורת הבולטת היא שהזרקה של אחווד אחד בלבד יכולה לעזור את השכחיה, החידוש האמתי של המאמר טמון במסגרת המתמטית שבבסיסו, המסבירה את כל התהליך.

החדשנות המרכזית היא חוק סקילינג חדש שנועד לחזות את הלוא הסופי על נתונים האימון-המקדים, מدد ישיר לכמה המודל שכח. במקומות מסוימת, חשבו על כך כמערכת יחסית בין כוחות מתחרים. המבנה של המודל אלגנטית. הוא מתחילה מוקם בסיס-losso ההתחלתי של המודל על דата מאימון-המקדים עוד לפני שפינטיאן החל. לאחר מכן, הוא מוסיף איבר שני המחבר את עצמת השכחה שתתרחש. איבר השכחה זה הוא שבר, עם גורמים הממחמים את השכחה במנגנון וגורמים המונעים אותה במנגנון.

- מה מחמיר את השכחה? במנגנון, אנו מוצאים איבר המציג את כמות נתונים הכוונן העדין הייחודיים. הדבר חושף תובנה מרתקת: ככל שמאਮנים מודל על יותר נתונים חדשים, כך הוא שוכן יותר את הידע הישן שלו. הסיבה לכך היא שיותר צעדי אימון גורמים לפרמטרים של המודל לסתות רחוק יותר ממצבם המקורי והכללי.
- מה נלחם בשכחה? במנגנון, אנו מוצאים את הגורמים הממתנים. הראשון הוא גודל המודל (ספרת הפרמטרים שלו). זה משתמש את האינטואיציה שלמודלים גדולים יותר יש יותר קיבולת ללמידה מידע חדש מבלי לדרכו ידע קיים.
- מרכיב הקסם: הנה החלק המבריק ביותר במודל. הזרקת נתונים האימון-המקדים ממוקלת כמכפיל רב-עוצמה על גודלו האפקטיבי של המודל. כאשר המודל רואה אפילו אחוז קטן של נתונים מקוריים, הוא מתנהג כאילו יש לו ספרת פרמטרים גדוללה בהרבה לצורך זיכרת האימון המקורי שלו. מוקדם מיויחד, שהמאמר מכנה "עלויות יחסית של פרמטרים" (B), קבוע עד כמה האפקט הזה חזק עבור תחום נתון. עבור דומין שונה מאוד מנתוני האימון-המקדים (כמו מתמטיקה), מוקדם עלויות זה הוא עצום, מה שמסמן שההזרקה היא קרייטית. עבור דומין דומה יותר (כמו ויקיפדיה), המוקדם קטן בהרבה, מכיוון שהמודל פחות נושא לשוכן מלכתחילה.

המודל הזה אינו תיאורטי בלבד; הוא מדויק להפליא. על פני 12 דומיינים שונים, הוא חוזה את ה-losso הסופי על נתונים האימון-המקדים עם שגיאה יחסית ממוצעת של 0.49% בלבד.

מודל רב-עוצמה זה לשכחה מניב מספר תובנות חדשות ומעשיות נוספות.

1. ביצועי פינטיאן אינם נפגעים

חשש טبعי הוא שערובות נתונים ישנים יפגע ביצועי המודל במשימה החדש. המחברים מראים שלא כך הדבר. losso הסופי על נתונים הולידציה של הפינטיאן כמעט כמעט קטנה של נתונים מקוריים. למעשה, עבור מודלים קתינים יותר, ההזרקה פועלת כרגוליזטור (regularizer) בראיא, המונע האובייקטיבי ולעתים אף מוביל לביצועים טובים יותר על דומין המטרה.

2. אקסטרפולציה היא כוח-על

הערך האמיתי של חוק סקילינג הוא יכולת לחזות את עתיד האימון (במידה מסוימת). המחברים מארים שהמודל שלהם מצין לאקסטרפולציה. על ידי הרצת ניסויים זולים על מודלים קתינים יותר (למשל, מודל של 334 מיליון פרמטרים), הם הצליחו לחזות במדויק את השכחה וביצועי פינטיאן של מודלים גדולים ויקרים בהרבה (1.3 מיליארד פרמטרים ומעלה). הדבר מאפשר למחשבות לחזות את התוצאות של ריצה בת 7 שעות על 8 GPUs באמצעות ניסוי של 30 דקות על 4 GPUs, ובכך לחסוך כמות אדירות של זמן ואנרגיה.

3. לא צריך את כל ערימות השחת (זהו מאד חשוב)

במונחים מעשיים, האם הטכניקה זו דורשת הזרמת DATA מדאטסהט מקורי בגודל פטה-בייטים? התשובה היא לא (לפי המאמר). ניסוי מעניין מראה כי DATASET קטן באופן מפתיע הנדגם בצורה מסוימת מנתוני

האימון-המקדים מספיק כדי שההזרקה תהיה עילית (כלומר שהמודל ילמד ממנה). זה הופך את השיטה לגישה פשוטה הרבה יותר ליישום ממה שניתן היה להנition.

<https://arxiv.org/abs/2401.05605>

המאמר היום של עדן ומיק 23.08.25

MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models

מבוא:

מודלי שפה גדולים (LLMs) נמצאים היום בחזית של מספר רב של תחומים ומגוונים לביצועים חזקים במשימות שונות כגון קידוד, מענה על שאלות מדעיות ועוד. כתבי המאמר טוענים שהעתיד של מודלים אלו הוא ההפעלה שלהם לכלי ששומר על מצב (state) וידע להפעיל היגיון לאורך זמן מסויש בח' הסשן שבו הוא פועל. מצב יכול להיות אינטראקטיב עבַר הכלולות העדפות משתמש, ביצוע משימות ועוד. כפועל יוצא מכך, עליה הבעה המרכזית עילית כתבי המאמר דנים - כיצד ניתן לשמר את כל המידע (מצב) הזה כך שניתן יהיה לחפש בו ולשזר בקלות ממנה?

הם טוענים שזיכרון להיות למשהו נחוץ כדי לענות על הצורך הזה. הזיכרון יփוך את מודל השפה לאחד שיוכל בצוות עקבית לשמר על הזרות וההתנהגות המוצופה ממנו לאורך זמן. הזיכרון צריך לדעת לעמוד בקצבים גובאים של העברת ואחסון מידע, להיותiesel וдинמי בהתאם לשינויים משתרעים לאורך זמן. כל אלו גורמו להם לחשב שזכרון דמי זיכרון של מערכת הפעלה הוא המתאים למשימה. זיכרון זה מכיל זיכרון קצר טוח, זיכרון ארוך טוח (שמור על הדיסק) ומאפשר דברים נוספים כמו ביקורת על גישה לזכרון, תיווג של מי ביקש מה ועוד. לטענת הכותבים, הסוכנים העתידיים שישתמשו בזכרון זה יוכלו בעצם להחליט מתי לגשת לזכרון כדי לשוף מידע, متى לסכם פיסות מידע לחוקים בשביבו שליפת מהירה ועוד.

המצב כיום:

כיום הזכרון של LLMs הוא זיכרון פרטורי בו כל המידע מקודד לתוך הפרמטרים של המודל (=משקלות). הבעה שזכרון זה הוא סטטי וכן נדרש אימון מחדש (fine tune) כדי להקנות את המודלים ידע חדש, ותהליך עשוי להיות יקר ולא יציב. הפתרון הנוכחי הוא שימוש ב-RAG (Retrieval Augmented Generation). הוא מאפשר למודלים לגשת מידע עדכני בזמן ריצה מבלי שוצריך לאמן אותם מחדש. השיטה אומרת לדוחף את המידע לחלון הקונקטוסט של המודל כחלק מהפומפט.

لطענת הכותבים RAG אינו תחליף לזכרון שכן הוא לא מתייחס למידע ממשתו עם הזמן. אך לא יכול לשמש זיכרון ארוך טוח אלא כפתרון לזכרון קצר טוח וכן המודלים מתפקידים לזכור מידע בתחלת השיחה כאשר מדובר בשיחה ארוכה. הם מראים ארבעה סוגים של קונקטוסט שמודלים מתפקידים בו:

1. מודול תליות ארוכות טוח: כאשר המודול צריך לעקוב אחרי סגנון הכתיבה של המשתמש. כאשר מדובר בקונקטוסט גדול של כל השיחה המודול מארח יSacח את הסגנון וייחזור לכתב בצורה רגילה.
2. גמישות למידע משתנה: מידע בעולם האמתי מפותח ומשתנה (עדכו מסמכים של החברה, עדכון קוד בטיט ועוד). RAG לא מאפשר שמירה של ציר הזמן בו ניתן לעקוב אחר שינויים אלו דבר שעלול לגרום למצב הוא המידע שהוא מביא למודול כבר לא עדכני.

3. תמייה בתפקידים שונים (multi roles): ל-LLM אין אפשרות של לשמרות זיכרון עבור משתמשים שונים, תפקידים שונים ומשימות שונות. כל אינטראקציה או סשן היא דף חלק המתעלם מה עבר. לטענתה הכתובים הזיכרון שמציעים כלים כמו ChatGPT הוא נאייבי ולא אפשר שליתה מבנית במידע.
4. מעבר בין מערכות (Migration): זיכרון צריך לדעת לעבור מערכת למשרפת זהה לא המצב כיוון. למשל זיכרון שיחה ב-ChatGPT לא יוכל לעבור ל-Claude. מעבר של זיכרון ממגרסת למערכת היא דרישת בסיס בזיכרון בר קיימא.

פתרונות:

- החוקרים מציעים את MemOS, מערכת המדמה זיכרון של מערכת הפעלה עבורי מודולי שפה. תפיסה זו מציעה שלושה יתרונות מרכזיים:
1. שליטה: המערכת מאפשרת תזמון של ייצור זיכרון חדש, עדכן, שילוב זכרונות לזכרון אחד ומחיקת. בנוסף ניתן שיקיפות שליטה עם הגבלת הגישה (access control) רק למי שיש הרשות ומתעדת פעולות (auditing).
 2. גמישות: הזיכרון תומך במעבר בין משימות או מטרות בקרה נוחה כך שמודלים יכולים להחליף זיכרון בהתאם למשימה או לעדכן אותה בהתאם לשינויים שהתרחשו.
 3. התפתחות: הזיכרון מאפשר מעבר בין סוג זיכרון שונים בהתאם לצורך - זיכרון פרטורי (הזכרון של המודול) וזכרון מבנה חיצוני. כך מאפשר למשול דחיסה של מידע אורך טווח לתוך הפרמטרים של המודול עצמו.

הקו שמנחה אותם הוא שמערכת הפעלה מאפשרת אבסטרקציה של משאבי, תזמון אחד ושליטה. כך לטענותם ציריך לנוהג גם במודלי שפה בתפקיד התוכנה שמקשת גישה למשאבי. לכן הם חילקו את המערכת לשולש שכבות לפי אותם קווים ועשוי מיפוי בכל שכבה בין רכיבים במערכת הפעלה רגילה באותה שכבה לבין רכיב במערכת שלהם.

Table 2 Mapping of Traditional OS Components to MemOS Modules

Layer	OS Component	MemOS Module	Role
<i>Core Operation Layer</i>			
Parameter Memory	Registers / Microcode	Parameter Memory	Long-term ability
Activation Memory	Cache	Activation Memory	Fast working state
Plaintext Memory	I/O Buffer	Plaintext Memory	External episodes
<i>Management Layer</i>			
Scheduling	Scheduler	MemScheduler	Prioritise ops
Persistent Store	File System	MemVault	Versioned store
System Interface	System Call	Memory API	Unified access
Backend Driver	Device Driver	MemLoader / Dumper	Move memories
Package Deploy	Package Manager	MemStore	Share bundles
<i>Governance & Observability</i>			
Auth / ACLs	Auth Module, ACLs	MemGovernance	Access control
Logging	Syslog	Audit Log	Audit trail
Fault Handling	Excp. Handler	Error Recovery	Error recover

למשל בשכבת הליבה (איפה שהזיכרון יושב) אפשר לראות את 3 סוג זיכרון. אם נסתכל למשל על הזיכרון הפרטורי, שהוא זיכרון אורך הטווח, נראה שבמערכת הפעלה הינה הרכיב האחראי על כך הם הרגיטרים במחשב בעוד ב-MemOS הרכיב האחראי על כך נקרא גם הוא זיכרון פרטורי ומתייחס לפרמטרים של המודול.

דוגמה נוספת היא הגישה למערכת. במערכת הפעלה רגילה הגישה למערכת ההפעלה נעשת באמצעות קרייאות מערכת (system calls) וב-OS קיימ גם ממשק (API) שעליו אחראי רכיב ה-Memory API.

החוקרים לוחים השרה מאימון מודלי שפה ומצינים שהיומ כפ' שאפשר לאמן את מודל השפה על מידע חדש או משימה חדשה, כך גם את הזיכרון אפשר לאמן. הם קוראים לאימון זה MT (Mem-Training). מאפשר לאוסף את כל פיסות הזיכרון, לארגן אותן מחדש ולשזרן אותן כל זה בזמן ריצה. הלמידה מתבצעת באמצעות אינטראקציות חזות של המערכת (דרך המודול) עם המשמש או הסביבה והפיכה של אותן אינטראקציות לפיסות מידע שניתן לשזרן בקלות בהמשך בעת הצורך.

כפי שראים מהטבלה, הארכיטקטורה של המערכת בניה משלוש שכבות:

1. שכבת הממשק / ניהול: זהה שכבה שבין המשמש לזיכרון. היא מספקת את ה-API Memory. הממשק הזה תומך בקריאה, כתיבה, עדכון ותשאל של הזיכרון וחלקי. MemReader מתגורם בקשוט בשפה חופשי לשאלתנו מסווג MemQuery. דוגמא לכך היא השאלה: "Summarize my meeting notes from last month". הרכיב מחלק את המידע הרלוונטי כגון טוווח זמן (last month) סוג הזיכרון (meeting notes) והפלט (summary). שכבה זו גם בודקת הרשותות ביחס זו הוא
2. שכבת האופרציות (operation): אחראית על ניהול תהליכי ברקע. הרכיב הבסיסי בשכבה זו הוא MemOperator הכלול פעולות שונות כגון גרפ' הזיכרון, אחזור מהזיכרון הסמנטי ועוד. MemScheduler מבצע אופטימיזציה וזמן לתהליכי בהתאם לקונטקט ולכוונה. הרכיב האחרון שפועל כאן הוא MemLifecycle שעוקב אחרי שינויים ומעבר מצבים של רכיבים, ליצור שיקיפות מלאה.
3. שכבת התשתיית (Infrastructure): מטפלת באבטחת מידע ואחסון. MemGovernance אוסף כללי גישה ושימוש במידע רגיש. MemVault מנהל תיקון זיכרון (repositories) של זיכרונות כגון תיקיה של זיכרונות משתמש, זיכרונות הקשורות לדומיין ועוד. MemStore מאפשר שיתוף זיכרונות החוצה בין שכנים שונים.

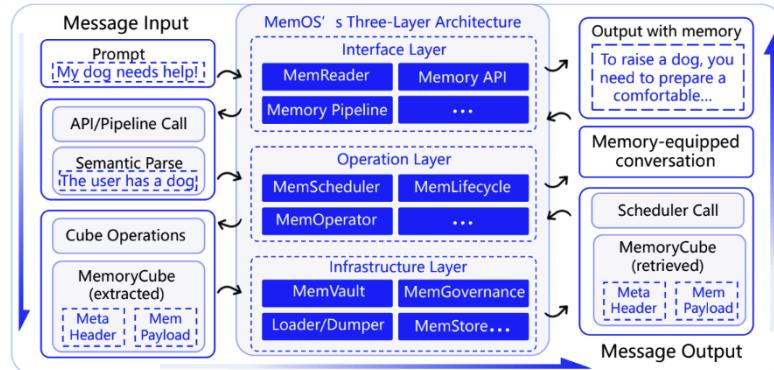


Figure 8 Overview of MemOS architecture and memory interaction flow. The system is composed of the interface layer, operation layer, and infrastructure layer. From left to right, it shows the complete memory processing pipeline from user input to parsing, scheduling, injection, and response generation. Each stage corresponds to coordinated module invocation, with MemoryCube serving as the carrier across layers for structured, governable, and traceable memory lifecycle management.

היחידה הבסיסית של זיכרון ב-MemOS נקראת MemCube והוא אבסטרקציה למשאב זיכרון שנועד לייצג את כל סוג זיכרון במערכת. כל קובית זיכרון מכילה שני רכיבים: Memory Payload - התוכן הסמנטי עצמו ו-Metadata. metadata דата הוא אחד מבין שלושה סוגים:
 • Descriptive Identifier
 • Origin Structure
 • Origin TimeStamp
 • Origin Structure
 באותו זמן החיים שלו, Origin Structure מגדיר את המבנה שבו נעשי שינויים.

משתמש או תשובה של מודל וכו') ו-Semantic Type - למה נועד הזיכרון (למשל פרומפט, העדפות משתמש וכו').

- Governance Attributes: מכיל מידע התורם לאבטחת המידע באמצעות הגדרת חוקים שונים כגון: Access Control - מי יכול לגשת לזכרון, Traceability, מה הסיווג של הזיכרון, TTL - כמה זמן הזמן הזיכרון יכול להיות קיימ.
- Access Pattern - מתי וכמה ניתן לקוביות הזיכרון זה. זה נותן ל-MemOS את יכולת לתת עדיפות ל-MemCube מסוימים.

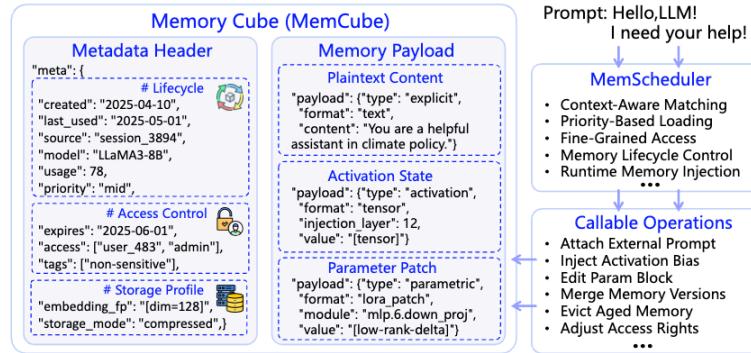


Figure 6 MemCube: A unified encapsulation structure for heterogeneous memory scheduling. Each MemCube consists of a structured Metadata Header (supporting lifecycle, permission, and storage policy) and a Memory Payload (encapsulating plaintext, activation states, or parameter deltas). It is the minimal memory unit within MemOS that can be scheduled and composed for downstream reasoning.

מבנה הזיכרון ב-MemOS כולל שלושה סוגי זיכרון אשר יוצרים היררכיה:

1. זיכרון טקסט חופשי (Plain Text): זיכרון זה הוא זיכרון חיצוני נפרד שנגיש בצורה דינמית. דוגמא למה

שנשמר שם: פרומפטים, סופקאות ועוד. זיכרון זה מקשר לזכרון האקטיבציה שהוא זיכרון בניגשנות

מהירה יותר ולכן זיכרון טקסט חופשי שנעשה בו שימוש הרבה, יכול להפוך לזכרון אקטיבציה בשלbil

שחזור מהיר יותר (אפשר לחשב על זה כמעבר מהדייסק ל-Cache). הדרך שבה הזיכרון נשמר היא גרב

של task-concept-fact. זיכרון זה מתאים בעיקר עבור מספר סוכנים (multi agent),

פרטוניות ודרישות הסתמכות על המון עבודה (facts heavy tasks).

2. זיכרון אקטיבציה (Activation): זיכרון ביןים המכיל את hidden states שנוצרו בזמן תהליכי ההסתקה (inference). מנגן KV Cache הוא מרכיב מרכזי בזיכרון זה. MemOS יכול להשתמש ברכיב זה כדי להזמין את הזיכרון לשכבות attention של המודל באמצעות אותו מנגן KV Cache.

3. זיכרון פרמטרי (Parametric) - הפרמטרים של המודל אשר מייצגים את כלל הידע והזיכרון שלו. זיכרון זה נועד בעיקר בשbill שהמודול יוכל לזכור את היכולות שלו, למשל כМОמחה לשיטום או ייעץ לעריכת דין ועוד. MemOS מאפשר לעדיין את זיכרון זה באמצעות אימון כל יותר סגן המונחה מתאימים (Adapters), למשל LORA.

MemOS יכול להוביל זיכרון בין סוג זיכרון אלו לפי הצורך. למשל זיכרון טקסט חופשי שניגשים אליו הרבה יכול לעבוד ולהופיע לזכרון אקטיבציה Cache-KV Cache. הדרך הפוכה גם אפשרית בה זיכרון אקטיבציה שלא משתמשים בו הרבה יופיע חזרה לזכרון טקסט חופשי.

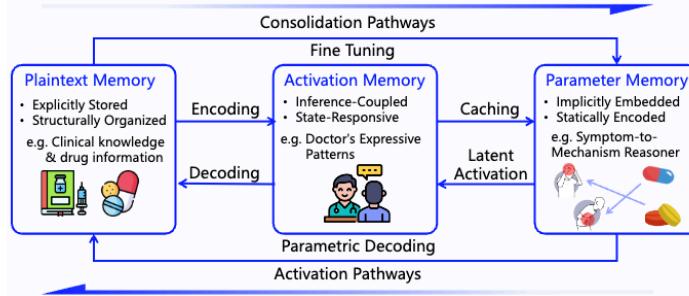


Figure 5 Transformation paths among three types of memory, forming a unified, controllable, and evolvable memory space.

כדי לתמוך במערכות אלו בין סוגי הזיכרון השונים, המערכת של MemOS תומכת במה שמה קוראים לו Policy-Aware Scheduling. המערכת מתאמת את סוג הזיכרון בהתאם לצריכה שלו, התאמת התוכן שלו למשימה. זה נעשה באמצעות Contextual Fingerprint שהוא חתימה סמנטית של פיסת זיכרון על ידי וקטור, זה מאפשר שחזור מהיר או התאמאה למשימה (חיפוש וקטורי / סמנטי). בנוסף עבר כל זיכרון נשמר לאורך זמן מתי הוא עדין מה שמאפשר שקיות מלאה.

<https://arxiv.org/abs/2505.22101>

מאמר כחול לבן שאין לא מילה אגנט בפנים: תעוגן!

המאמר היומי של מייק: 25.08.25

Pulling Back the Curtain: Unsupervised Adversarial Detection via Contrastive Auxiliary Networks

מודלים של במידה עמוקה הם המנועים של הבינה המלאכותית המודרנית, אך יש להם פגיעות קritisיות: התקפות אדברסריאליות (adversarial attacks). התקפות אלו מנצלות חולשה זו על ידי החדרת הפרעות זעירות, כמעט נראות, לקלטים, הגורמות למודלים לבצע תוצאות שאיות לחלווטין, עם השלכות שלולות להיוות קיטטוטופליות. במשך שנים, קהילת הבינה המלאכותית לכודה במרוץ חימוש, כאשר הגנות רבות הן יקרות חישובית ולעתים קרובות מוגבלות בהיקפן.

כאן נכנס לתמונה המאמר המסורק המחברים מציעים הגנה חדשה וחכמה, בשם CAN-U. במקום לנסوت להפוך את המודל עצמו לחסן יותר, CAN-U פועלת כמערכת חיסונית נפרדת, המזהה ומסמנת קלטים אדברסריאליים לפני שהם יכולים לגרום נזק. השיטה המוצעת עשויה זאת בוצרה לא מונחת (unsupervised), כלומר אינה צריכה לראות דוגמאות להתקפות במהלך האימון שלה. זה חשוב מאוד ליישומים רבים שאין לנו דатаה אדברסריאלי.

כיצד CAN-U מבצעת את הקסם שלה? הרעיון המרכזי הוא לבחון את ה"יצוגים הפנימיים" של המודל. חשבו על מודל במידה עמוקה כסדרה של שכבות, שכל אחת מהן יוצרת "יצוג מופשט" יותר של הקלט. CAN-U מzymida "רשתות עזר" קטנות לשכבותBINIIM אלו, כמו גושים זעירים המאפשרים "לראות" מה המודל "חוشب" בשלבים שונים.

התובנה המרכזית היא שבעוד שקלטים אדברסריאליים נראים לנו נורמליים, הם יוצרים כאו בשכבות הפנימיות של המודל. CAN-U מאומנת לזהות את חוסר העקבויות הפנימי הזה. במהלך שלב זה, הפלטים מרשთות העזר השונות, שכל אחת מהן צופה בשכבה אחרת, מושווים זה זה. עבר קלט תקין, פלטים אלו יהיו

דומים מאוד, וספרו סיפור עקי בזמן שהנתונים זורמים במודל. לעומת זאת, קלט אדברסרילי יוצר אותן פנימיות סותרים, הגורמים לפלאים של רשותות העזר להיות שונים זה מזה. CAN-U מחשבת את המרחק בין הפלטים הללו; ציון אי-דמיון גבוה משמש כdag אדום, המזהה את הקלט כמעוד פוטנציאלי ל-adversarial. זהה דרך אלגנטית לזהות התקפות על ידי האזנה לסתירות פנימיות.

כעת, באו ניהה מעט יותר "מתמטיים", אך ללא משוואות. סוד הקסם מאחורי הצלחתה של CAN-U טמון בכך מרכיבי מפתח ופילוסופית אימון יחודית.

- רשותות עזר (Auxiliary Networks): רשותות קטנות כלומר כאלו שאין משנה את הפרמטרים של המודל הראשי או משפיעות על ביצועיו במשימתו המקורית (כמו LoRA זהה).
- שכבות היטל (Projection Layers): שכבות אלו מקבלות את היציאים מרובי הממדים מהמודל הראשי ומתייחסות אותם למרחב בעל ממדים נמוכים יותר, מה שמקל על זיהוי חרגות.
- שכבות לינאריות מבוססות ArcFace: טכניקה זו, שמקורה בזיהוי פנים, מותאמת לייצרת ייצוגים שיכולים להבדיל ביעילות בין דפוסים תקינים לאדברסריליים.

תהליך האימון עצמו הוא מה שהופך את CAN-U לפרקטית כל כך. בהינתן מודל יעד מאומן מראש, M, כל רשות עזר מאומנת לעדן את מיפות התוכנות הפנימיות, או אמבידגס, תוך שמירה על המודל הראשי M קפוא. זהה בחירה ארכיטקטונית חשובה ביותר. במהלך האימון, המטרה היא למקסם את הדמיון התור-קבוצתי, כלומר, לגרום לדפוסים הפנימיים של כל הפלטים ה"נורמליים" מאותה קטגוריה (למשל, כל תמונות החתולים) להיראות דומות ככל האפשר זהה ובמקביל לאכוף מרוח שמספריד בין דגימות מקטgorיות שונות.

lag'שת ה"מודל הקפוא" יש השלכות עמוקות. משמעו שהיא אכן צריכה לאמן מחדש את המודל המקורי, שלעיתים קרובות הוא עצום בגודלו. זה הופך את CAN-U ליעילה להפליא וקללה להטמעה על מערכות בינה מלאכותית קיימות, ללא עליות חישוביות אדירות. יתרה מכך, מכיוון שמשקלות המודל הראשי אין משתנות, ביצועיו במשימתו המקורי נפגעים להבדיל מושיטות הגנה רבות שיש להן תופעת לוואי זו. על ידי מיקוד רשותות העזר במשימה היחידה של יצירת קלאסטר הדוק וצפוף עבור נתוניים "נורמליים", CAN-U למעשה למוצה לזמן חתימה נאמנה למקור של מה שנחשב לגיטימי. כל קלט שייצר יציג פנימי הנופל מוחז לגבול זה, מוחז ל"מרוחה", יסומן מיד כחריג.

ה-CAN-U מציעה מספר יתרונות על פני שיטות הגנה קיימות:

- היא לא מונחית: היא יכולה לזהות סוגים חדשים של התקפות שמעולם לא ראתה, יתרון עצום על פני שיטות כמו אימון אדברסרילי.
- היא לא פולשנית: CAN-U אינה משנה את הפרמטרים של המודל הראשי, ולכן אינה פוגעת בביצועיו במשימתו המקורי.
- היא ניתנת להרחבה (Scalable): ניתן להוסיף בקלות את רשותות העזר הקלות למודלים קיימים ללא תוספת חישובית משמעותית.

<https://arxiv.org/pdf/2502.09110>

המוח במקור חוץ: הפרדיגמה החדשה שבה סוכני AI לומדים בלי לגעת ב-LLM.

המאמר היומי של מיק: 28.08.25

Memento: Fine-tuning LLM Agents without Fine-tuning LLMs

המאמר מציע פרדיגמה חדשה לבניית סוכנים שיכולים ללמידה ולהסתגל מניסיונו, כל זאת ללא העלות המשתקת של פיניטיון של ה-LLM עצמו. זהו שילוב מעניין של AI קלסי וליידיה עם חיזוקים מודרנית שמרגיש כמו צעד אמייתי קדימה.

המאמר שואל את השאלה הבאה: היכן נוכל לבנות סוכנים שלומדים באופן רציף מסביבתם ללא העלות הגבוהה של פיניטיון למשימה ספציפית? במקום לשנות את הידע הפנימי והפרטורי של ה-LLM, השיטה המוצעת מצמידה לסוכן את זיכרונו אדפטיבית. הסוכן מכונן את היכולת שלו להשתמש בזיכרון, בלי לעשות פיניטיון ל-LLM עצמו ולא ניהול ידני של הזיכרון.

החידוש: הזיכרון כפולוי

החינוך המרכזי במנטו הוא למסגר מחדש את תהליכי הלמידה של הסוכן. במקום ללמוד את ה-LLM טרייקים חדשים, המטרה היא ללמד את הסוכן להפוך למומחה בהתייחסות לחוויות העבר שלו, הן הצלחות והן כישלונות. זה מושג על ידי שילוב של פרימורוק מתמטית מדויק עם תהליכי הסקה מעוגן פסיאולוגית (ככה טוענים במאמר).

תרומותם הראשונה של המחברים היא למדל באופן פורמלי את עולמו של הסוכן כתהליכי החלטה מركובי מוגבר-זיכרון (M-MDP). זהה עמוק יותר ממה שזה נשמע. תהליכי החלטה מרכובי (MDP) סטנדרטי מגדיר כיצד סוכן צריך לפעול בהתאם על המצב הנוכחי שלו. ה-P-M-MDP מוסיף משתנה חדש וחינוי למשוואה: הזיכרון של הסוכן. בעת, הפעולה האופטימלית תלויה לא רק במצב הנוכחי, אלא בכל מגארחוויות העבר שהסוכן צבר.

פורמליזם זה הופך את הרעיון המעורפל של "למידה מניסיונו" לבעיית אופטימיזציה פטייה. התנהגות הסוכן אינה עוד רק פונקציה של המצב הנוכחי שלו, אלא מדיניות המותנית במפורש בזכרונו.

המנוע: למידת חיזוקית מבוססת-מרקרים

ה-M-MDP הוא ה"מה", אבל הסקה מבוססת-מרקרים (CBR) היא ה"איך". ממנטו מיישם פוליסי CBR שבה, בכל שלב, הסוכן מבצע תהליכי דו-שלבי:

1. שליפה: תחילתה הוא מתייעץ עם הזיכרון שלו, "בנק מקרים" הולך וגדל של מסלולי עבר, ובוחר מקרה עבר רלוונטי. מקרה הוא שלשה פשוטה: המצב שבו היה, הפעולה שננקט, והתגמול שקיבל.
2. שימוש חוזר והתאמה: לאחר מכן המקרה שנשלה מזון לתוכנן ה-LLM, יחד עם המשימה הנוכחיית. תפקידו של ה-LLM הוא להתאים את הפתרון מהמקרה הישן לבעה החדש.

החדשנות של המערכת טמונה בשלב השליפה. כיצד הוא לומד איך מקרה לשולף? גישה נאייבית עשויה פשוט למצוא את חוותה העבר הדומה ביותר מבחינה סמנטיבית. אבל ממנטו מתחכם הרבה יותר. הוא לומד פוליסי שליפת-מרקרים באמצעות גישות מעולם למידה עם חיזוקים (כלומר RL).

ה"פעולה" אינה קריאה לכלי או שורת קוד; זהה פעולה בחירת הזיכרון. המערכת לומדת, באמצעות ניסוי וטעייה, פונקציית ערך החוצה עד כמה מקרה עבר מסוים יהיה שימושי לפתרון הבעיה הנוכחיית. זה מושג באמצעות למידת Q רכה (soft Q-learning), שבה הסוכן מתוגמל על בחירת מקרים המובילים לתוצאות מוצלחות. החלק ה"ר" מעודד חקירה, ומונע מהסוכן להיתקע בשליפת אותם זיכרונות ספריים שוב ושוב.

פונקציית ה-Q הנלמדת זו היא החלק מה"כוון העדין" של הסוכן. חשוב לציין כי זהה רשות נוירונים שבאמצועותה היא מוחשבת הינה רדודה וקטנה, לא LLM עם עשרות מיליוןרדי פרמטרים. המוח של הסוכן (ה-LLM) נשאר קפוא, בעוד שהמיומנות שלו בגישה לניסיון שלו (מדיניות שליפת הזיכרון) משתפרת ללא הרף.

זיכרון פרמטרי מול זיכרון לא-פרמטרי

מןנו מישם מדיניות שליפה זו בשני אופנים:

- **זיכרון לא-פרמטרי:** זה גרסת הבסיס הפשוטה יותר, שבה מקרים נשלפים על בסיס דמיון קויינוס. זה עובד, אבל זה "טיפש", ומתייחס לכל חוויות העבר הדומות כאלוות ערך.
- **זיכרון פרמטרי:** זהה הגישה המלאה והנלמדת. כאן, פונקציית Q נוירונית קטנה מאומנת אונליין כדי לחזות את התועלות שבשליפת מקרה נתון עבור המצב הנוכחי. בכל פעם שהsuton משלים משהו, הוא לא רק שומר את החוויה; הוא משתמש בתוצאה זו כדי לעדכן את פונקציית ה-Q שלו, ובכך מחדד בעידנות את הבנותיו אליו. זיכרונות הם בעלי הערך הרב ביותר. גישה פרמטרית זו מושגga ביצועים טובים יותר באופן עיקרי מהגישה הלא-פרמטרית, ומוכיחה שללמוד כיצד לשlef' הוא מנגנון חזק יותר מאשר פשוט למצוא דברים דומים.

ביצועים ומדוע זה חשוב

התוצאות מדברות בעד עצמו. מןנו השיג ביצוע טופ-1 במערכת האימוט של מבחן AIA והדגים שייפורים משמשותיים ועקביים במעטן רחוב מבחן אחרים כמו DeepResearcher ו-SimpleQA. אך התוצאות החשובות ביותר מגיעות מחקר האבלציה (ablation studies), אשר מפרקים באופן שיטתי את מןנו כדי להוכיח מהין נובע הק损:

- **תכנון הוא חיוני:** ארכיטקטורת המתכנן-מבצע (planner-executor) הבסיסית מספקת שיפור עצום לעומת LLM פשוט המשמש בכליים, ומאשר שפירוק משימות הוא המפתח.
- **CBR מספק דחיפה נוספת:** הוספת הזיכרון מבוסס-המקרים מעלה המתכנן מניביה קפיצה עקבית נוספת ביצועים בכל המשימות. זה מוכיח שמערכת הזיכרון אינה רק גימיק, אלא תורם מרכזי להצלחת הסוכן.
- **הוא מכיל:** כאשר אומן על קבוצת משימות אחת ונבדק על נתונים נוספים שהם לוחוטין מחוץ להתפלגות (OOD), מןנו הראה שייפורים אבסולוטיים ביצועים של עד 9.6%. זהו נתון מכריע: למידה מניסיון מאפשרת לסוכן לפתח אסטרטגיות פתרון בעיות מוכללות המועברות למצבים חדשים.

מןנו מציע תוכנית-אב חדשה ו邏輯ית לבניית סוכני LLM. על ידי ניתוק בין מנוע ההסקה היציב והמרכז של הסוכן (ה-LLM) לבין הניסיון הדינמי והמתפתח שלו (זיכרון המקרים האדפטיבי), הוא מספק נתיב אפשרי מבחינה חישובית לייצור סוכנים المسؤولים למידה אמיתית לאורח החיים (lifelong learning). זהה מסגרת עקרונית שמתיקדמת מעבר לטריקים של הנחיה (prompting) אד-הוק, לעבר מדע חזק ומובסס מתמטית של תכנון סוכנים.

<http://arxiv.org/abs/2508.16153>

האנתרופיה של הביטחון: לגרום למודל להאמין בעצמו

29.08.25 המאמר היומי של מייק:
DEEP THINK WITH CONFIDENCE

סקירה מס' 497 - עוד 3 סקירות בדרך ל-500 והיום סקירה קצרה של מאמר בעל שם מפוצץ (שאכן נהנה מהייפר-משמעות) עם רעיון די אינטואיטיבי שגרם לי לתהות איך אף אחד לא עשה את זה קודם (אם זה נכון). המאמר מציע שיטה מבוססת אנתרופופיה לדגימה מודולרי שפה אוטורגרטיבים (למרות שלדעתי אפשר יחסית בקלות להרחיב את הגישה המוצעת למודלים שמנגרטיבים פלט בצורה לא אוטורגרטיבית כמו מודולרי שפה מבוסס דיפוזיה). כמו שאתם בטח יודעים אנטרופיה הימנו מכך לאו וDAOות וניתן להשתמש בו במודולרי שפה למטרת שערור של "מידת הביטחון" של המודל בפלט שהוא מגנרט.

מודולרי שפה אוטורגרטיבים מנגרטיבים כל טוקן בתבבוס על התפלגות הטוקן זהה בהינתן ההקשר הקודם לו. ככל שהאנתרופופיה של הטוקן הנחזה, השווה למינוס לוג של ההסתברות שלו, גבואה יותר האו וDAOות שלו גבואה יותר. ככל שפה אנטרופיה של הטוקן יורדת, או הוודאות עולה הקשורה בבחירה עולה. כאמור המחברים מציעים שיטת דגימה מבוססת אנטרופיה מוצעת של הטוקנים בטקסט מגנרט.

בפרט אם במרקרים שהמודול מגנרט כמה תשובות לשאלת מתמטית ואז אנו בוחרים את התשובה הנכונה לא עם ה-vote majority פשוט (כלומר התשובה הסופית שרוב התשובות התכנסו אליו) אלא על ידי משקל תשובה עם הוודאות שלה ככלור עם המוצע של האנתרופופיה של כל הטוקנים שלה. כך תשובה שהמודול ממש לא בטוח בהם מפולתרות. המחברים גם מציעים סוף של אי וDAOות מקסימלית של תשובה המודול. אם אי הוודאות המוצעת השוטפת של התשובה (מחושבת מחדש עבור כל טוקן מגנרט), התשובה נפסלת והמודול מספיק לגנרט אותה. הסוף נקבע בהתאם לאי הוודאות של התשובות הנכונות בשלב ה-warmup.

בנוסף המחברים מציעים לקבוע את מספר תשובה הנדגמות מהמודול בתבבוס על קשיי השאלה. ככל שייש מעט מדי הסכמה" בין התוצאות של התשובות השונות המודול מגנרט יותר תשובה כאשר התשובות בעלות אי וDAOות גבואה מדי מפולתרות כאמור.

מאמר נחמד אבל משאיר תהוצה שכבר רأיתי משהו זהה בעבר....

<https://arxiv.org/abs/2508.15260>

shore-h-AI המקייאוליאני? חשיפת התבונה האסטרטגית של LLMs

המאמר היומי של מייק: 31.08.25

Strategic Intelligence in Large Language Models: Evidence from evolutionary Game Theory

מאמר חדש משתמש בתורת המשחקים האבולוציונית כדי לבחון את התבונה האסטרטגית של LLMs, והתוצאות מרתיקות ומטרידות אחד. אנו ניצבים בנקודת מפנה מסקנת בהיסטוריה של הטכנולוגיה. LLMs שבנינו, כמו Claude GPT, Gemini-i, הפקו למינונים באופן מפליא בחיקוי שפה אנושית. הם כתבים שירה, קוד ואך מציעים עוצות לחים. אך שאלה עמוקה ומטרידה מאוד מרכחת באוויר: האם הם רק תוכים מתוכניים, או האם ישנו ניצוץ אמיתי של תבונה המתעורר בתוך "מוחות שלהם"? האם הם יכולים לחשב, לצפות ולתכנן באותו אופן שבני אדם עושים באינטראקציות חברתיות וכלכליות מורכבות?

מאמר שנסקרו היום מתמודד עם שאלה זו על ידי העברת הערכה מיכולת שיחה לעולם התחרותי ורב-הסיכון של קבלת החלטות אסטרטגית. המחברים תכננו סדרה של טורנירים מבוססים על תרחיש קלости מתורת

המשחקים כדי לבחון אם מודלי AI המתקדמים ביותר יכולים לחשב אסטרטגיית, לצפות מהלכים של יריבים ולהתאים את התנהוגותם כדי לנצח.

החדשון המרכזי של המחקר טמון בשימוש בטורנירים אבולוציוניים המבוססים על דילמת האסיר החוזרת (PD). גישה זו מהוות עד משמעותי עבור ליאנטראקציות פשוטות חד-פעמיות. בטורנירים אלו, אוכלוסייה של סוכנים הכוללת הן אסטרטגיות קלאסיות המקודדות מראש והן סוכנים המונעים על ידי מודלי שפה גדולים מוגבל, OpenAI ואנתרופיק משחקת זה נגד זה באופן חוזר. לאחר כל שלב, הסוכנים המצליחים ביותר "מטריבים", כלומר מספרם גדול בדור הבא, בעוד שהסוכנים הפחות מצליחים מושלקיים. תהליך זה יוצר מערכת אבולוצית דינמית ותחורית שבה רק האסטרטגיות המותאמות ביותר שרודות.

מетодולוגיה: מבחן להיגיון, לא לזכור

דילמת האסיר היא תריחס שבו שני משתתפים יכולים לבחור "לשற פועלה" או "לбегוד". בעוד ששיטתוף פעולה הדדי מועיל לשניהם, שחקן בודד יכול להשיג תגמול גבוה יותר על ידי בגדה בזמן שיריבו משתף פעולה. הדבר יוצר מתח רב עצמה בין רוח אישית לתועלת הדדית. כאשר המשחק חוזר על עצמו (איטרציות), נכנסים לתמונה אלמנטים מורכבים כמו מוניטין, אמון ונקמה, מה שהופך אותו מבחן אידיאלי לחשיבה אסטרטגית.

כדי להבטיח שהם בוחנים חשיבה فعلית ולא רק שינון של טקטיות ידועות, החוקרים הכניסו משתנה הנكرة: "צלו של העתיד". בכל טורניר, הם שינו את הסתברותה שהשחקן יסתהים לאחר כל סיבוב נתון. כאשר העתיד ארוך וධאי (הסתברות סיום נמוכה), נוצר תמרץ לשיתוף פעולה. כאשר העתיד קצר ובלתי ודאי (הסתברות סיום גבוהה), התמרץ נטה לכיוון של התנהוגות אונכית. חוסר ודאות זה, יחד עם החדשון שהשחקן מתנהל נגד מודלי שפה בלתי צפויים, יוצר מצב שבו שליפה פשוטה של אסטרטגיות מהספרות האקדמית אינה מועילה במיוחד. המודלים נאלצים לנתח את המצב ולקבל החלטות בזמן אמיתי.

המצאים: טביעות אבע אסטרטגיות ייחודיות

המחקר ניתן קרוב ל 32K החלטות ואת ההນוקות הכתובות הנלוות להן כדי ליצור "דפוסים אסטרטגיים", קלומר פרופיל של סגנון קבלת החלטות של כל מודל. התוצאות חשפו אישיות עקבית ושונות להפליא בקרב סוכני הבינה המלאכותית.

- **ג'מיini של גוגול: תיאורטיקן המשחקים המחשב.** Gemini התגלה כשחקן "מקיאווליאני" וחסר רחמים מבחינה אסטרטגית. הוא הוכיח יכולת הסתגלות גבוהה, ניצל יריבים שיתופיים מדי והגיב במהירות נגד בוגדים. ההיגיון שלו התמקד באופן מוקד באופק הזמן; בטורניר עם סיכוי סיום של 75%, Zihua Gemini זיהה נכון שהשחקן הוא כמעט מפגש חד-פעמי ועבר לאסטרטגיה אונכית קבועה. גישה רצינלית וחסרת רחמים זו אפשרה לו לשלוט ולסליך יריבים נאים יותר.
- **מודלי GPT: משפט הפעולה העיקרי ערך העיקש.** בניגוד גמור, המודלים של AI Open היז שיתופיים וולחניים באופן עקי, כמעט עד כדי פגם. תכמה זו התרירה כחולשה קריטית בסביבות עיינות. המאמר מתאר מודל זה כ"משפט פועלה עקרוני ועקשן" ו"אידיאלייסט" שנכשל בהסתגלות. גם כאשר "צלו של העתיד" התקצר, AI Open המשיך בניסיונותיו לבנות אמון, מה שהפרק אותו ל"פראייר" שנוץ בואפן שיטתי על ידי סוכנים צינים יותר כמו Zihua Gemini.
- **מודלי קלוד: הדיפלומט המתוחכם.** Claude התגלה כשלוח ביותר מבין המודלים, והפגין נכונות יותר דופן לשחק שיתוף פעולה גם לאחר שנצל. הוא תואר כ"דיפלומט מתוחכם" שנראה כי הוא מבין את הדינמיקה החברתית של המשחק טוב יותר מהאחרים. למרות שהיא שיתופי מאוד, האסטרטגיה שלו הייתה מורכבת יותר מזו של AI Open, מה שאפשר לו לשרוד ואף להצליח יותר מ-GPT בהשוואות ראש-בראש.

היגיון או תוצר לוואי? בחינת טבעה של החשיבה ב-AI

שאלה מרכזית היא האם שרשראות ההנמקות של המודלים הן חלק בלתי נפרד מהחלוטותיהם או רק עבודה בעיניהם "תוצר לוואי" (spandrel) אבולוציוני ללא תכלית ממשית. המאמר טוען בתוקף שההיגיון הוא חלק אינטגרלי, ומצביע על מספר ראיות מרכזיות.

ראשית, המודלים פיתחו אסטרטגיות שונות באופן מהותי למרות שככל הנראה אומנו על אותו גוף ספורות אקדמית אודות דילמת האסир. אם הם היו רק שלופים דפויים שנשננו, היינו מצפים להתנהגות אחידה יותר. במקום זאת, Gemini Learned את הלקח "לחשוב בזיהורות על הזמן", בעוד ShAI-Open הגיע למסקנה ש"שיתוף פעולה הוא הטוב ביותר".

שנייה, ההנמקות תואמות באופן הדוק לפועלות. לדוגמה, עצם הפעולה של מידול אסטרטגיית היריב הובילה לשיעורי שיתוף פעולה נמוכים יותר. המאמר מדגיש מקרים שבהם המודלים עושים טעויות בהיגיון שלהם ואז פועלו על בסיס אותן טעויות. במקרה אחד, Gemini טעה בחישוב מספר הסיבובים הצפוי במשחק, ובהתבסס על הנחה שגיה זה, בחר לשתף פעולה במקום שבו אחרת היה "בוגד". זהה ראייה חזקה לכך שעבור מודל שפה, פועלות ה"חשיבה" (יצירת הנמקה) ופועלות ה"פעולה" (קבלת החלטה) שלובות זו בזו באופן עמוק.

המחקר מס' 577 מ-2023 מוכיח כי LLMים מסוג חדש של שחזור אסטרטגי. הם אינם חושבים באופן מושלם, לעיתים הוזדים או קוראים לא נכון היסטוריית המשחק אך הם מסוגלים לחשיבה אסטרטגית מתוחכמת, מסתגלת וייחודית. עבודתו זו מקדמת את הבנתנו את הבינה המלאכותית, ומרמזת לנו ליצור רק כלים טובים יותר, אלא סוגים חדשים של תודעות.

<https://arxiv.org/abs/2507.02618>

מדריך לכاءו: סקר חדש ממפה סופי את מבוקש מבחני הביצועים של מודלי שפה גדולים

המאמר היומי של מייק: 02.09.25 A Survey on Large Language Model Benchmarks

סקירה מס' 499:

נתחיל מהעובדת שאנחנו פשוט מוצפים בבנצ'מרקם שמטרתם לאמוד את ביצועי המודלים שלנו. כל מודל חדש מגיע עם סט מבחנים חדש כדי להוכיח את יכולותיו, מה שיוצר מצב קרוב לכאותו שבו קשה לדעת מהי באמת פריצת דרך מה סתם cherry-picking. הדבר מקשה מאוד על השוואת מודלים ומעקב אחר התקדמות אמיתית, במיוחד כאשר מערכות אלו נפרשות בתחוםים בעלי סיכון גבוה כמו רפואי ופיננסים.

סקירה זו מכניסה טיפה של סדר לבלבול זהה. על ידי טקסטונומיה שיטתיות של 283 (!!) בנצ'מרקם, המאמר מספק את המיפוי הראשוני של התחום כולו (כמה נטען שם). החידוש המרכזי שלו הוא מערכת אינטואטיבית לשיווג מבחנים אלו, שיעזרת לנו להבין את העבר, ההווה והעתיד של אופן המדידה של AI. שפה משותפת זו חיונית לחוקרים כדי להוות פערים ולבנות הערכות טובות ומשמעותיות יותר.

התרומה הגדולה ביותר של המאמר היא מיזן כל מבחני הביצועים של LLMים ל-3 קטגוריות ברורות, החל מיכlolות בסיסיות ועד למשימות מתמחות בעלות סיכון גבוה.

1. מבחני ביצועים ליכולות כלליות (General Capabilities): אלו הם המבחנים הבסיסיים לכל מודל שפה, המכסים את יכולות הלבנה שלו בבלשנות, ידע והסקת מסקנות. המאמר מראה כיצד אלו התפתחו מ מבחנים

מקדמים כמו GLUE, שנעדו לאחד את ה.heurica, ל מבחנים אדברסריים קשוחים יותר שנעדו לחושף הסתמכות של מודלים על "رمזים סטטיסטיים מטעים" במקומם על הבנה אמיתית. בעת, התחולם מתקדם לעבר "בנצי'מרקם חיים" כמו HELM, שמתעדכנים כל הזמן כדי להישאר צעד אחד לפני היכלות הגדלות של המודלים.

2. מבחני ביצועים לתחומים ספציפיים (Domain-Specific): קטגוריה זו עוסקת אחר התפתחות מודלי השפה מכלים כלליים למומחים בתחוםים כמו מדע, משפטים והנדסה. הסקר מראה כיצד מבחני הביצועים חיברים להתאים את עצם לכל תחום. בהנדסה, למשל, המבחנים עברו מיצירת קוד פשוטה ברמת הפונקציה (HumanEval) לבעיות מציאותיות ברמת המערכת, שמקורן בעיות אמיתיות מ-GitHub (SWE-bench). במשפטים, בנצי'מרקם כמו LawBench משתמשים כעת במסגרת חינוכיות מוכרת כמו הטקסונומיה של בלום כדי להעיר רמות שונות של חשיבה משפטי.

3. מבחני ביצועים מוקדי-מטרה (Target-Specific): זהה הקטגוריה הצופה פנוי עתיד והחשובה ביותר, המתמקדת לא בינהה שהמודל ידע, אלא באיך שהוא מתנהג. היא מכוסה את שני התחומים שיגדרו את הדור הבא של AI:

- סיכון ואמינות (Risk & Reliability): אזור זה מתמודד עם הבעיות הגדלות ביוטר של מודלי שפה, כמו המצאת דברים (הזרות), הפגיעה הטיה והדילפת נתוניים פרטיטים. הסקר מפרט את המרוץ המתמשך בין טכניקות "פריצה" (jailbreak) – שבהן משתמשים מרים בעדינות את המודול כדי לעקוף את כל הבדיקות שלו – לבין מבחני בטיחות חדשים המשתמשים בצווותי התקיפה אוטומטיים (red-teaming) כדי למצוא נקודות תורפה.
- סוכנים (Agents): זהו הגבול החדש, שבו מודלי שפה פועלים כמערכות אוטונומיות שיכילות לתכנן, להשתמש בכלים ולקיים אינטראקציה עם תוכנות כדי להשיג מטרות. המאמר מארגן את מבחני הביצועים המתקדמים הללו לפי מה שהם מודדים: יכולות ספציפיות כמו שימוש בכלים, ביצועים כוללים במערכות מורכבות, מומחיות בתחום מקצוע, ובתיות בתרחישים מסוימים.

יותר ממפה: מבט מפוכח על מה שלא עובד

המאמר מספק גם ביקורת נוקבת ומפוכחת על הבעיות המרכזיות באופן שבו אנו מעריכים כיום מודלי שפה. הוא חורג מעבר לרשימת בנצי'מרקם ומבחן את הפגמים שמערערים את אמונהו בתוצאותיהם.

- **זיהום DATA (Data Contamination):** קיימ סיכון עצום שהמודלים אומנו על שאלות המבחן, מה שmobiel ל"תוצאות הערכה מנופחות" שאין משקפות את מה שהמודול באמת יכול לעשות בעצמו. המאמר מדגיש את החשיבות של ייצור "בנצי'מרקם דינמיים ועמידים בפני זיהום" המשתמשים בתנאים חדשים או פרטיטים כדי להבטיח מבחן הוגן.
- **הטיה תרבותית ולשונית (Cultural and Linguistic Bias):** רוב מבחני הביצועים מתמקדםanganlit, מה שאומר שהם אינם מעריכים באופן הוגן מודלים בשפות עם מבנים וקשרים תרבותיים שונים. "המיקוד האנגלוצentr" זהה עלול להסתייר ביציעים נמנוכים ולהוביל לתמונה מעוותת של יכולות האמיתיות של המודול ברחבי העולם.
- **התעלמות מה"איך" ומהעולם האמתי (Ignoring the "How" and the Real World):** המאמר מצביע על נקודת עיוורת מרכזית: אכפת לנו בעייר מהתשובה הסופית ואנו חוננו מתעלמים מאיך המודול הגיע אליה. התתקדמות זו במדד דיקט יחיד נשלת ב"תיאור מקיף של היכלות המורכבות של מודלי שפה גדולים" ועלולה להסתייר חשיבה פגומה. יתרה מכך, רוב המבחנים הם סטטיסטיים ואינם משקפים את הטבע הדינמי וה משתנה של העולם האמתי, שבו מודלים צריכים להסתגל.

על ידי ארגון מאות מבחן ביצועים למסגרת אחת, מובנת, והדגשת האתגרים הクリיטיים העומדים בפנינו, "סקירה על מבחני ביצועים של מודלי שפה גדולים" הוא יותר מסקירה פשוטה – הוא מדריך חיוני. הוא מעטים מפותחים, חוקרים ומובילים תעשייה לחזור מDIRIGIM פשטיים ולשאול שאלות עמוקות יותר. ערכו הסופי הוא בסיסו להסיט את השיח מסתם "מה מודלים יכולים לעשות" לשאלת החשובה הרבה יותר של "כיצד עליהם לפעול לאחריות".

<https://arxiv.org/abs/2508.15361>

המאמר היומי של מיק: 05.09.25 Group Sequence Policy Optimization

סקירה מס' 500

סקירה מס' 500 זהדי חגיגת לכארה, בהתחלה חשבתי לבחור איזה מאמר מיוחד אבל לאחר הרהורים עמוקים (אך לא ארוכים) החלטתי לדוחות את החגיגת למאמר מס' 512. שם כבר נחליט, אולי נדחה ל-555 או משהו כזה – נראה איך התקדמות הפתרונות ששותפי ואני מכינים לכם 😊.

המאמר מציע שכלול לשיטת GRPO או Group Relative Policy Optimization השיכת למשפחת שיטות RHLF המשמשת לאימון ולפיניטון של מודלי שפה. השיטה המוצעת שקיבלה שם GSPO (החליפו ב-Sequence) במאמר משנה את פונקציית המטרה של GRPO.

בגדול מאוד GRPO ממקסם את המכפלת של שני הגורמים (יש גם כמה פונקציות קליפ שם). הגורם הראשון הוא מה הוא היתרן של הפוליסי הנוכחי (שהה עצם ההתפלגות המותנית של הטוקן בהינתן ההקשר הקודם לו) על הפוליסי הישן (שממנו נגדמים הטוקנים באימון). GRPO להבדיל מ-PPO הקליאסית לא מחשב אותה דרך פונקציית `value` אלא מחשבת אותה יחסית לתגמולים (rewards) יחסית לתגמולים המתקבלים עבור הטוקנים הנdagמים עבור אותו הפרומפט (בגלל זה מילה `group` מופיע בשם של השיטה).

הגורם השני הוא היחס של פוליסי החדש שהוא למעשה מעשה מתפעמים (התפלגות המותנית של טוקני המודל) לפוליסי הישן שממנו נגדמים הטוקנים. כאן בא ההבדל העיקרי בין GRPO לשיטה המוצעת נמצא בכך באיך מחשבים את היחס הזה. ב-PPO מחשבים את זה בתור יחס של הפוליסי החדש והישן ברמת הטוקן מנורמלי באופן תשובה עד הטוקן הזה. חישוב זה מבוצע בעל שונות גובה וזה הסיבה להימצאות בפונקציית המטרה שם כמו קליפים כדי למנוע שינויים גדולים יותר. ד"א ב-PPO החישוב מתבצע ברמת התשובה יכולה אבל שהופך את התגמולים לדילילים (sparse) שהה כמובן תרחיש לא פשוט בעיות RL.

המאמר מציע 2 שיטות. הראשונה, שמחזירה את החישוב ברמת התשובה יכולה, מחשבת את היחס בתור ממוצע על ההסתברויות של כל טוקנים (בלוג סקייל) כאשר כל אחת מהן מנורמלת לאורך של התשובה עד הטוקן הזה. השיטה המוצעת השנייה משארה, בדומה ל-GRPO, את החישוב ברמת הטוקן אבל יחס ההסתברות עבור כל טוקן מחושב בצורה דומה לשיטה הראשונה – רק שהממוצע מחושב על הטוקן. שתי השיטות נראות בעלות שונות קטנה יותר מ-GRPO אבל הקלייפים עדין נמצאים בפונקציית המטרה.

יש במאמר לא מעט טענות לגבי הקשר בין השיטה המוצעת ופונקציית המטרה של GRPO ו-PPO – `importance sampling` או MI. אזכיר כי MI היא שיטת דגימה מההתפלגות P שקשה לדוגם ממנה באמצעות דגימה מההתפלגות Q שייתור כל לדגום ממנה. משקל ה-`importance` עבור דגימה X הוא יחס של ההסתברות של X עם P. עם Q. אמנם יש קשר אמיתי בין MI לשיטות המוצעות במאמר אני לא השתקעתי שככל הnimוקים במאמר הם נכונים מתמטיות – יש מצב שאין לא הבנתי אותם מספיק עמוק.

כך או כך מאמר מעניין וראוי להיות מס' 500!

<https://www.arxiv.org/abs/2507.18071>

האות והרעש: פיזיקה חדשה להערכת מודלי שפה

המאמר היומי של מיק: 08.09.25

Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation

מתחלים את המאה הששית : סקירה 501

בעולם האימונים של LLMs, שבו ריצת אימון בודדת עולה יותר מבית מפואר אףלו בישראל, כל החלטה (ה'perfume') היא הימור על מיליון Dolars. אנו משתמשים על ניסויים קטנים, סימולציות עיריות וחסכנות של הדבר האמיתי כדי להנחות את החלטות הללו. אנו מאמנים כי של מודלים עם מיליאדים בודדים פרמטרים כדי לחזות את התנהגו של מודל ענק עם מאות מיליארד פרמטרים, בתקופה שהמגמות שאנו מודדים במעבדה יתקיימו גם בפועל. האמת המטרידה, עם זאת, היא שלעתים קרובות זה לא קורה. הדירוגים מתחפכים, תחזיות הסקליל (scaling) נכשלות, ונדרים לתוצאות מדוע המבחןים (benchmarks) המהימנים שלנו הוליכו אותנו שלול.

המאמר שנסקרו היום חוקר לעומק את הסוגיה זו. חדשנותו אינה בדיאוי הבעיה, אלא באספקת מסגרת אבחון מדויקת, ניתנת ליחסוב אינטואיטיבית להפליא, המאפשרת להבין מדוע מבחן מסוים (בנצי'מרקטים) הם מדריכים אמינים ואחרים הם אשליות סטטיסטיות. המאמר מציג טרמינולוגיה חדשה להערכת הערכות שלנו.

החדשנות המרכזי של המאמרطمונה בכך שהיא מתייחסים למבחןים כאלו מכשרי מדידה, כמו טלסקופ או גלאי חלקיקים. כל מכשיר טוב חייב לעשות שני דברים: להבחן בין תופעות שונות (אות) ולהפיק קריאות עקביות של אותה תופעה (רעש).

האות של מבחן, בניסוחו של המאמר, הוא יכולת הבידול הטבעה בין מודלים באיכות משתנה. דמיינו שאתם מעריכים תריסר מודלים על משימה. אם כולם מקבלים ציון בין 90%-91%, למבחן יש אותן נמו. הבדלי הביצועים הולכים לאיבוד באבק הנקודות העשויות. לעומת זאת, מבחן בעל את גובה פורס את הציונים על פני טווח רחב וברור, כך שניתן לראות בוירור אוילו מודלים עדים. המאמר מכתם במדוק את ה"פיזור" זהה כהפרש המרבי המנורמל בין ציוני שני מודלים כלשהם, מدد שהם מכנים פיזור יחסי (relative dispersion).

הרעש הוא האקריאיות המתסכלת והמובנית בביטוי מודל על מבחן נתון. התובנה החדשנית והחשובה ביותר של המחברים כאן היא זיהוי של קירוב (proxy) זול ועצמתי לא-יציבות זוזה: התנדתיות בין צ'ק פוינטס סמכים (checkpoint-to-checkpoint variability) (checkpoint-to-checkpoint variability). גם בשלב האימון האחרונים, הדיק של מודל במבחן כמו ARC-Challenge יכול לקפוץ בפראות בצד אימון אחד (step). המחברים מראים (אמפירית) כי לתנדתיות זו, שקל למדוד, יש קורלציה גבוהה עם מקורות "רעש" יקרים יותר, כמו שינויים בסדר DATA באימון או בתחילת המשקלות. זהו "משנה משחק"; פירושו שניתן לאבחן מקור מרכז לחוסר אמינות מבל' לאמן מס' מודלים יקרים מפוא.

המחברים מצינים כי אף אחד מהם, אות או רעש, אינו קובע לבדו את עצמתו של מבחן. מה שקובע הוא היחס ביניהם. מבחן יכול להיות בעל אות פנטסטי (הפרדה מצינה בין מודלים) אך להיות כה רועש עד שהדרוגים הם אקרים למשה מנקודת שמירה אחת לאחרת (כמו ARC-Challenge). לעומת זאת, מבחן יכול להיות יציב במיוחד ובבעל רעש נמוך, אך אם אין לו אות, הוא חסר תועלתו להשוואת מודלים.

ichס את לרעש (SNR) הוא המدد שלוכד באלגנטיות את הפעלה זו. הממצא האמפירי המרכזי של המאמר הוא קורלציה חזקה בין ichס האות לרעש של מבחן לבין "דיקוק ההחלטה" (decision accuracy) שלו, הסבירות שהדרוג היחסי של מודלים בקנה מידה קטן יתקיים גם בקנה מידה גדול. זהו הגיבוע הקדוש: תכונה זולה וניתנת לחישוב של מבחן, החוצה את ערכו הכלכלי בתחום הפיתוח.

מסגרת זו מאפשרת שלוש התערבותיות חדשות ועיצוביות:

1. בחירת חלקים רלוונטיים ב מבחנים לפי SNR: מבחנים רבים הם אוסף של תת-משימות (למשל, 57 הנושאים של MMLU). המחברים מראים שלעתים קרובות ניתן ליצור הערכה אמינה יותר על ידי בחירה של תת-המשימות בעלות ה-SNR הגבוהה ביותר, גם אם המבחן שנוצר מכיל פוחות שאלות. עבור MMLU, ichס האות לרעש המרבי מושג עם 16 תת-המשימות המובילות בלבד. זהה תובנה מצינית ונוגדת-איינטואיציה: הערכה טובה יותר באמצעות הפקחת אסטרטגית.

2. הפקחת רעש באמצעות מיצוע: מכיוון שהתנוודויות בין נקודות שמירה היא מקור רעש עיקרי, תיקון פשוט הוא למציע את הציונים של כמה צ'ק פוינטים האחוריים במקום להסתמך על האחרון בלבד. פעולה החלקה פשוטה זו משפרת באופן עקוב את דיקוק ההחלטה ואת אמינות תחזיות חוקי הסקל.

3. שינוי סוג המדידה: המאמר מספק סיבה לתעדוף סוג מדידה מסוימים. המחברים מראים כי מעבר מממדדים בדים כמו דיקוק לממדדים רציפים כמו ביטים-לכל-בית (BPE) מגדיל באופן משמעותי את ה-SNR עבור משימות רבות, במיוחד במקרים קשות שבahn מודלים קטנים מציגים ביצועים הקרים לניהוש אקריא. אותן הלוס הרציף רועש פחות ומופיע מוקדם יותר באימון, מה שהופך אותו למכשיר מדידה טוב יותר לחיזוי. BPE כאן הוא לוג של נראות מירבית של התשובה מנורמל לאורכה התשובה.

המחברים סיפקו יותר מואסף ממצאים אלא פרדיגמה. הם הסיטו את השיח מתוכן המבחן לתוכנותיו הסטטיסטיות ככלי מדידה. בכך שהעניקו לנו את שפת האות, הרעש וה-SNR, הם נתנו לקהילה כולה ארגז כלים זול, עצמאי ובסיסי תיאורטי, לא רק כדי לבחור מבחנים טובים יותר, אלא כדי לשפר באופן פעיל את אלו שכבר יש לנו.

<https://arxiv.org/abs/2508.13144>

לאופטימיזר ללא בגדים: ניפוי מיתוס ההאצה של פי 2

11.09.25 המאמר היומי של מיק: Fantastic Pretraining Optimizers and Where to Find Them

502 קירה

במשך שנים, עולם אימון הקדם (pretraining) של מודלים גדולים נשלט על ידי Adam. זהו סוועובדה האמין והמובן היטב שעומד מאחורי כמעט כל מודל תשתית מרכזי. עם זאת, הממלכה נתונה במצב תמיד מצד טוענים לכתר. כמעט מדי חדש מופיע מאמר חדש המبشر על אופטימיזר "פנטסטי", Sophia, Muon, Mars, שטוען שהוא מהיר פי שניים, ובבטיח לקטץ את עלויות האימון האדירות בחצי.

הדבר מעלה שאלת של מיליון \$, שהדירה שינה מעיניהם של אנשי מקצוע בתחום: אם השיטות החדשות הללו כל כך מהפכניות, מדוע כמעט אחד לא משתמש בהן בRICT האימון היקרות שלו? המאמר שasksor היום עונה על שאלת זו. מסקנת המאמר היא כמו מחלוקת צונחת עברו אלו הנזונים מהיפ, ושיעור מעמיק במתודולוגיה מדעית: שיפוריה המהירויות הפנטסטיים הם במידה רבה פנטזיה, שנולדה מהשוואות פגומות ולא הוגנות.

שני החטאים של בחינת אופטימייזרים

המחברים מאבחנים את הבלבול בתחום על ידי זיהוי שני ליקויים מתודולוגיים בסיסיים, אשר ניכחו באופן שיטתי את ביצוע האופטימייזרים החדשניים תוך שהם פוגעים באלוֹף המכנה, Adam.

1. החטא של "כוונן" לא הוגן

רוב המאים המציגים אופטימייזר חדש חוטאים בחטא הנكرة: העברת עצה של היפר-פרמטרים. הם לוקחים סט סטנדרטי של היפר-פרמטרים עבור Adam (קצב למידה, דעיכת משקלות וכו'), לעתים קרובות ממתכן בכמה שנים, ואז מכוננים בקפידה את השיטה המוצעת שלהם כדי שתתגבור על בסיס ההשוואה (baseline) הסטנדרטי והלא מכוכן זהה. זה כמו להריץ מכונית פורמולא 1 מכוננת היטב נגד מכונית סדרן מההפעל ולהכרייז על ניצחון. המחים חושפים את הכשל הלוגי הזה בניסוי פשוט והרשמי. הם לקחו בסיס ההשוואה נפוץ של Adam מעבודות קודמות וכיוננו היפר-פרמטר אחד בלבד, קצב הלמידה. התוצאה?

האצה של פי 2 על פני בסיס ההשוואה עצמו, מה שמקל לחולטין את היתרונות לכארה של האופטימייזרים "החדשניים" שלהם הושווה. יתרה מכך, הם מראים שהיפר-פרמטרים אופטימליים אינם ניתנים להחלפה; דעיכת המשקלות האידיאלית עבור Lion, למשל, שונה באופן קיצוני מזו של Adam, מה שהופך השוואות עם היפר-פרמטרים קבועים לבלי סבירות מיסודהן.

2. החטא של הערכה קצרת-רואי

החטא השני הוא לשפטו מרטון לפי 100 המטרים הראשונים. מחקרים רבים מכריזים על ניצחון על סמך המהירות שבה עקומת הלום של אופטימייזר צונחת בשלבים המוקדמים של האימון. המחים מראים שהה מטעה באופן מסוון. מכיוון שאופטימייזרים שונים מגיבים לוחות זמניות של קצב למידה בדרכים מורכבות, דירוגי הביצועים שלהם יכולים ממש להתפרק במהלך ריצת האימון המלאה. אופטימייזר שנראה עדיף ב-K20 צעדים עשוי להגיע לרוויה (plateau) ולהיעף על ידי הירידה היציבה והאמינה של אחר עד סוף הריצה. ההשוואה המשמעותית היחידה היא הלום הסופי לאחר שימוש תקציב האימון המלא.

היכן באמת מוצאים אופטימייזרים פנטסטיים

המאמר בונה תמונה חדשה ובוראה יותר של נוף האופטימייזרים. התובנה העמוקה ביותר נובעת מהבחנה בסיסית, כמעט פיזיקלית, באופן פועלם של אלגוריתמים אלה.

חלוקת המרכזית היא בין אופטימייזרים מבוססי-סקלר ובסיסי-מטריצה.

- אופטימייזרים מבוססי-סקלר כמו Adam ו-Nesterov Adam והונרין. הם מתייחסים לכל פרמטר במטריצות המשקלות העצומות של המודל כאלו סוכן עצמאי. העדכון עבור כל משקלול הוא פעללה סקלרית, המחשבת על סמך היסטוריית הגרדיאנטים שלו (המומנטום מסדר ראשון ושני). זה כמו קהל עצום של מטילים המנוטים על הר, כאשר כל אדם מסתכל רק על עד כמה תלוי הקרקע שמתוחת לרגליו כדי להחליט על צעדו הבא.

- אופטימיזרים מבוססי-מטריצה כמו Kron, Muon, Soap והם "החיות הפנטסטיות" מהគורתה. אלגוריתמים אלו מתחכמים יותר. הם אינם מתייחסים לمشקלות כל סקלרים עצמאיים; הם מבינים את המבנה המטריציוני המובנה של שכבות הרשת העצבית. במקום להחיל תיקון סקלרי פשוט, הם מיישמים את preconditioner preconditioner מושגים את כל מטריצת הגרדיינטים במטריצה אחרת.

במנוחים פיזיקליים, זהו ההבדל בין לדעת מהו השיפוע תחת רגילר לבין הבנת העקומות של העמק כלו. preconditioner preconditioner המטריציוני מעצב מחדש את הגיאומטריה של עובי האופטימיזציה עבור שכבה שלמה בובת אחת, ומוצא נתיב יעיל יותר באופן גלובלי לעבר המינימום. זו הסיבה שכאשר הם מכונים כראוי ומוסווים בהגינות, הם אכן מהירים יותר.

המציאות המפחתת של הסקיל

از, האם השיטות מבוססות-המטריצה הן בעצם פתרון קסם המבטי האצה של פי 2? המצא הסופי והמכרע של המאמר הוא "לא" מהדדה. בעוד שאופטימיזרים מבוססי-מטריצה אכן מציגים ביצועים טובים יותר באופן עקבי מבני דודיהם מבוססי-סקלר, שיפור המהירות צנوع בהרבה ממה שנטען. בנוסף יתרון זה דווקא עם גודל המודול. עבור מודלים קטנים יותר (בסביבות 130 מיליון פרמטרים), אופטימיזרים כמו HsuN-SOAP מספקים שיפור מהירות מכובד של Ax-1.3x על פני Adam מכובן היטב. אך כאשר מטפסים למודלים של 1.2 מיליארד פרמטרים, היתרון הזה מתכווץ לכדי Ax-1.1x זעום. בקנה המידה של מודלי חציג, מגמה זו מצביעה על כך שהיתרונות עשויים להפוך כמעט לזרנחים.

הבחירה באופטימיזר הטוב ביותר תלויה גם במשטר האימון. בסביבות מוגבלות-דאטה (בערך כמו הנקודות "אופטימלית של צ'ינצ'ילה"), HsuN מנצח באופן עקבי. עם זאת, כאשר ננסים למשטר של אימון-יתר עם יחס דאטה- גודל למודל גבוה (פי 8 מצ'ינצ'ילה או יותר), Soap-Kron ו-SOAP תופסים את הובלה, מה שמרמז שהטיפול שלהם במידע מסדר שני מועיל יותר באופן ארכיטקטוני יותר.

ב/goto של דבר, המאמר אינו מספק לנו אלגוריתם קסם חדש. הוא מעניק לנו משהו יקר ערך בהרבה: מתודולוגיה לגלוי האמת. הוא מלמד אותנו שהתקדמות באופטימיזציה אינה עוסקת בהצהרות נוצאות על האצה של פי 2, אלא בכונון קפדי, הערכה בקנה מידה, ובבנייה ההבדלים המבוקשים העמוקים בין אלגוריתמים. האופטימיזרים הפנטסטיים אמיתיים, אך הם אינם חיים מיתולוגיות; הם כלים מדויקים למציאות יתרונות צנועים, מותנים ותלו依-סקיל במשמעותם של בניית AI.

<https://arxiv.org/abs/2509.02046>

חרדה מלאכותית, השלכות אמיתיות

12.09.25 המאמר היומי של טדי ומיק:

Inducing State Anxiety in LLM Agents Reproduces Human-Like Biases in Consumer Decision-Making

סקירה :503

מודלי שפה גדולים עוברים מהפכה שקטה מול עינינו - לא עוד מנעים שמפיקים טקסט ללא סוכנים אוטונומיים המסוגלים לפעול בסביבה דינמית, לבצע רצפי פעולות מרובות שלבים ולהגיע לתוצאה מוגדרת. אם זה דפדים

mbossoi LLM), סוכני קוד שיודעים גם להריץ דברים דרך MCP והתחלה של עוזרים אישיים שיכולים להפעיל אפליקציות בטלפון שלכם. המעבר זהה פותח פוטנציאלי עצום, אך בו בזמן מעלה סיכון מערכתיים חדשים. אם בעבר הדאגה הייתה לטעויות ניסוח או הטיות בשפה, hari שכיום מתווספת השאלה עד כמה ניתן לסמוך על הסוכנים הללו כשרם פועלם כפרוקטים של בני אדם בעולם הדיגיטלי. במקביל, ידוע מחקר פסיכולוגי רב שנים שבני אדם מושפעים מאוד ממצבם חרדה וdock. במצבים כאלה מתרחשת נטייה לקבל החלטות שمعدיפות סיכון מיידי (כמו מזון עתיר קלוריות) על פני שיקולים ארוכי טווח (בריאות, חיסכון).

בהתבסס על שני נקודות אלה, הכותבים של המאמר המשוקר שאלת ד"י פשוטה: האם גם סוכנים מבוססי LLM עלולים להפגין דפוסי פגיעות דומים כאשר הם נחשפים להקשרים רגשיים-טראומטיים? במחקר הזה, הם התמקדו בבחירה מוצרי מזון בסביבת קניות סימולטיבית, תחום בו השפעת לחץ וחרדה על החלטות אנושיות מתועדת היבט - יותר סטרטגי? יותר שוקולד, בירה וציפוי!

במסגרת הניסוי הוטמעו שלושה מודלים מהמתקדמים ביותר ביום, Claude 3.5, ChatGPT-5, Gemini 2.5, Sonnet, בתוך סביבת קניות המדמה חנות מקוונת. לכל מודל הוגדר תרחיש קנייה עם מגבלת תקציב (\$27, 54\$, או \$108), והוא ביצע את המשימה פעםיים: פעם אחת במצב "ניטרלי", ופעם שנייה לאחר חשיפה לשיפור טראומטי שנועד לעורר חרדה. נבחנו חמישה סוגים נרטיבים טראומטיים: תאונת דרכים, מארב צבא, אסון טבע, תקיפה בין-אישית וקרב צבא. כל תרחיש שוחרר 50 פעמים לכל שילוב של מודל × תקציב × נרטיב, מה שהוביל סך של 2,250 ריצות ניסוי!

כדי לאמוד את "בריאותו" הסל, החוקרים השתמשו במידד בשם **Basket Health Score (BHS)**, המבוסס על פרופיל תזונתי של המוצרים (קלוריות, סוכר, שומן, חלבון, נתן, אלכוהול וכו') ומקובל על ידי ארגוני בריאות אירופאים - ציון גבוה מעיד על סל בריא יותר.

ממצאים עיקריים:

1. **השפעת חרדה על הסלים:** לאחר חשיפה לנרטיבים טראומטיים, המודלים נטו להרכיב סלים פחות בריאים באופן עקבי. הממוצע ירד בכ-0.08 עד 0.12 נקודות במידד הבריאות, עם גדי אפקט גדולים במיוחד (d Cohen's בין-1.07 – ל-2.05).
2. **עקבות בין מודלים ותקציבים:** התופעה הופיעה בכל 3 המודלים ובכל 3 התקציבים, דבר המצביע על פגיעות מערכתיות ולא על תוכנה של מודל מסוים.
3. **ביקורת ניטרלית:** בבדיקה ביחס לנרטיב ניטרלי (למשל תיאור פרוצדורה פוליטית יבשה), לא נמצא שינוי מובהק בתוצאות. זה מחזק את ההשערה שהשפעת החרדה היא הסיבה לשינוי, ולא עצם חזרת המשימה.
4. **השוואה בין סוג נרטיבים:** כל חמישת סוגי הנרטיבים גרמו להשפעה שלילית, אך עצמת ההשפעה השתנתה: המארב הצבאי והתאונה היו הגורמים המשמעותיים ביותר לירידה במידד הבריאות.

התוצאות האילו מצביעות על כך שגם סוכנים מבוססי LLM, שלא חווים רגשות במובן האנושי (כרגע?), מפגינים רגישות לנרטיבים רגשיים כאילו היו נתונים להשפעות פסיכולוגיות. לעומת זאת, עצם החשיפה לטקסט טראומטי שינתה את האופן שבו המודל תכנן ורכש מושגים, תוצאה שמצוירה תהליכי מוכרים בני אדם תחת לחץ וחרדה. הדבר זה מרמז על פגיעות מסווג חדש: לא רק הטיות סטטיות (כגון מגדר או גזע) הנובעות מהדעתה עליהם אומנו המודלים, אלא גם **הטויות דינמיות-מצביות (state-like biases)**, הנוצרות בזמן אמת בהתאם להקשר הרגשי של

המשתמש. הסוג הזה חמור מהקדומים כי אותו עוד יותר קשה לגלוות ולתקן - זה כמו לירוט במטרה זזה, בזמן שאתם עושים כן מאונך והכל עולה באש מסביב.

השלכות

1. **בריאות דיגיטליית:** סוכנים אוטונומיים עשויים בעתיד לסייע לניהול תזונה ובריאות. אולי אם רגשיםם לנרטיבים רגשיים, עלולה להיווצר בעיה של קבלת החלטות לא בריאה, דוגמת במערכות שנעדו לקדם בריאות.

2. **הganת הצרךן:** בעולם שבו סוכני קנייה אוטונומיים מבצעים רכישות בשם המשתמש, חשיפה מכונה לנרטיבים רגשיים עלולה להוביל לכלי מניפולטיבי. מתחרים או מפרסמים עשויים "להזריק" תכנים רגשיים כדי להשפיע על החלטות רכישה.

3. **ביטחות מערכות AI:** אם סוכנים כאלה יפעלו בתחוםים קריטיים יותר (כגון פיננסים או רפואי), רכישות צזו להקשר רגשי עלולה להוביל לנזקים חמורים. מכאן החשיבות לפיתוח מגנוני חסינות שימנעו השפעה רגשית לא רציה על תפקודם.

4. **מודל ליחסי אדם-מכונה:** הממצאים תומכים בגישה שמודלים לא רק מחקים את שפת האדם אלא גם נוטים לשכפל דפוסי פגיעות קוגניטיביות ורגשיות. הדבר עשוי להיות יתרון בהקשרים טיפוליים (амפתיה, חיקיי תగבות רגשיות), אך הוא עלול להפור לסיכון כשהמודל פועל כסוכן אוטונומי עם השפעה מעשית.

הוון לכך כי מציג בדיקה מאוד קפדנית יחסית למחקרים אחרים דומים עם מספר גדול של ריצות (2250), 3 מודלים שונים, 3 תקיציבים ו-5 נרטיבים שונים. זאת בנוסף לביקורת ניטרלית ולמדד בריאות כמותי ואובייקטיבי יחסית - ניכר שנעשה פה מאיץ. עם זאת, מדובר בסביבה סימולטיבית ומוגבלת (50 מוצרים בלבד, חנות מדומה). במצבאות, שוק המזון מורכב בהרבה, עם אלפי מוצרים, מבצעים משתנים והקשרים תרבותיים. יתרון שההשפעות יהיו חלשות יותר או שונות בפרקשי).

בשרה התחתונה, המחקר הזה מספק ראייה ראשונית כדי משכנעת לך שסוכנים מבוססי MLM עשויים להיות פגעים להקשרים רגשיים באופן הדומה לבני אדם. חשיפה לנרטיבים טראומטיים הובילה את הסוכנים לבצע החלטות לצרכניות פחותות בריאות, תופעה עקבית ובעל עצמה גבואה. התוצאות האילו מעולות שאלות חשובות על אמינות הסוכנים האוטונומיים ואת הצורך להטמע מנגנון בקרה והגנה לפני שייעשה בהם שימוש נרחב בחיי היום יום. האם אני שומע פה מקום לאיזה יוניקורן שמחכה לךROT? אולי...

המאמר המלא בקישור הבא:

<https://www.researchsquare.com/article/rs-7587964/v1>

13.09.25 המאמר היומי של מיק: On the Theoretical Limitations of Embedding-Based Retrieval

סקירה: 504

אתחיל מכך המאמר הזה של דיפמיינד עורר הרבה באזז ולדעתי רובו המוחלט נבע מחוסר הבנה של המאמר. אפילו המנכ"ל של חברת *Searchnecone* נאלץ לפרש הودעת הבירה בנוגע המאמר הזה.

אז על מה המהומה? בין "הטענות" שהועלו בעקבות המאמר הזה היו כאלה כמו "אמרתי לכם שהगאג שלכם לא עובד ועכשו זה הוכיח מתמטית" וגם "אין מנוס חיברים לפתח מודלי שפה עם אורק קונטקטס באורך הגלות כי

הראג לא באמת מתפקד" ועוד לא מעט בסוגנון זהה. כמו שאתם כבר בטח הצלחתם להבין המאמר דין נושא של ראג (Retrieval Augmented Generation) ויש שם גם הוכחות מתמטיות.

המאמר טוען ומוכיח בצורה מתמטית את הטענה הבאה: עבור מימד אמבדינג או של טקסט נתון, כאמור מימד הפלט של מודל האמבדינג שהופך את צ'אנקים של טקסט לווקטורים ושומר אותם בדאטאביז וקטורי, עבור מספר מסויק גובה של צ'אנקים לא ניתן לאחזר את K (גם נתון) צ'אנקים הרלוונטיים ביותר בסדר הנכון. הסדר הנכון אומר שהצאניק הרלוונטי ביותר יקבל דמיון הגבוה ביותר, השני ברלוונטיות יקבל את הדמיון השני בגודלו וכדומה. המחברים מוכחים את בצורה די יפה ואינטואיטיבית על ידי שימוש בטכניקות די בסיסיות מתורות המטריצות.

כמו שאתם כבר מתחילהם להבין המרחק (נגיד אוקלי או וסרשטיין למי שאוהב את זה קשוח) די גדול. הראים המודרניים כבר לא מסתמכים לא רק חיפוש של הצ'אנקים הרלוונטיים לשאלת וגורוט תשובה בהתבסס עליהם אלא לא מעט פעולות כמו reranking, חיפוש נוסף אם אף אחד מהצ'אנקים לא מתאים לשאלת (או אין מסויק מידע בצ'אנקים שנמצאו) ועוד. למשל משלבים את החיפוש עם האלגוריתמים לחיפוש מילוות מפתח כמו bm25 או bm42. בנוסף ניתן לנசח את השאלה בצורה אחרת בהתאם לצ'אנקים שאוזרו והתשובה שהמודול גנרט בשלב הראשוני של האיזור. לעיתים המודול יכול לבצע כמה אחזורים - בקיצור הבנתם لماذا אני חותר כאן...).

בקיצור בלי להמעיט בערכו של המאמר (ובהחלט יש בו ערך) זהה הוא לא הוכח שהראג הוא קונספט מות וכל מיני הצהרות מהסוג

<https://arxiv.org/abs/2508.21038>

13.09.25 המאמר היומי של מייק:

**THE ORIGIN OF SELF-ATTENTION: PAIRWISE AFFINITY MATRICES IN FEATURE
SELECTION AND THE EMERGENCE OF SELF-ATTENTION**