

Review 61: PonderNet: Learning to Ponder

פינט הסוקר:

המלצת קריאה ממייק: מומלץ בעיקר להרחבות אופקיים

בהירות כתיבה: גבואה

ידע מוקדם: הבנה בסיסית ברשותות ובחוקי הסתברות

"ישומיים פרקטיים אפשריים: --

פרטי מאמר:

لينק למאמר: [זמן להורדה](#).

لينك לקוד: [לא רשמי 1](#), [לא רשמי 2](#)

פורסם בתאריך: 02.09.2021, בארכיון (v2)

הציג בכנס: 8th ICML Workshop on Automated Machine Learning (AutoML 2021)

תחומי מאמר:

- מודלים (במקרה זה רשותות נוירוניים) בעלי זמן חישוב אדפטיבי (תלוּ במבנה משימה)

ידע מוקדם:

- מודלים (במקרה זה רשותות נוירוניים) בעלי זמן חישוב אדפטיבי (תלוּ במבנה משימה)
 - מרחק KL בין התפלגיות
 - התפלגות גיאומטרית
-

תמצית מאמר:

המאמר הנסקר מציין לתחום שלא הייתה מודיעו לקיומו עד שקרהתי אותו. התחום דן ברטשות ניוירונים שמסוגולות להתאים את כמות החישובים למשימה נתונה בהתאם לרמת המורכבות של המשימה. ככלומר עבור המשימה "קלה" רשות צזו תבצע פחות חישובים מאשר למשימה "מורכבת יותר".

המטרה העיקרית כאן היא להקנות למודל את יכולת הפיסוק את תהליכי האימון במצב בו נראה כי הוא "הצליח ללמוד את מה הוא היה שצירך", ומילא המשך האימון לא צפוי לשפר את ביצוע המודל באופן משמעותי. אם לעומת זאת המודל "רוואה" כי איטרציות אימון נוספת עשויות להניב תוצאות טובות, הוא בוחר להמשיך את האימון. אצין כי רמת המורכבות של משימה אינה מועברת כמעט ללא הרשות צריכה "להחליט-on-the-fly" עד כמה המשימה מסובכת ולהתאים את כמות החישובים הנדרשת.

המאמר הנסקר מציע שיטה, הנקראת PonderNet, ההוכפת רשות ניוירונים נתונה לאדפטיבית מבחינה חישובית, ככלומר צזו שיודעת להתאים את כמות החישובים לפי רמת המורכבות של בעיה. שינויים קלים לארQUITטורת הרשות, ומצילהה להציג איזון בין ביצועי המודל על סט אימון, כמות החישובים הנדרשת יכולת הכללה של הרשות" (לשון המאמר).

הסבר של רעיונות בסיסיים:

הרעיון של המאמר הוא די פשוט וטבui. בהרצתה הראשונה של הרשות מזכינים לה את הקלט המקורי (ג'יד, תמונה, טקסט או קטע אודיו) ומקבלים כפלט את הייצוג החבוי ("לטנט") שלו. יציג לנווני זה ממשך קלט להרצת רשות הבאה. לאחר מכן מרכיבים את הרשות פעם אחריו פעם כאשר הקלט h (ייצוג חבוי - hidden state) לכל הרצת y^* (איטרציה) הוא הפלט של האיטרציה הקודמת (1-ח). בנוסף, לאחר כל איטרציה הרשות מספקת את החיזוי y_n עבור המשימה המקורית של הרשות ואת הסתברות לעצירת הריצה λ . ככלומר, פלט של רשות אחרי איטרציה הוא השלישי $s = (y_{n-1}^*, h_{n-1}, \lambda_{n-1}^*, y)$, כאשר s מיופיע המודול באמצעות רשות ניוירונים כלליות (step function). במאמר s נקראת פונקציית מדרגה (LSTM, MLP, encoder-decoder).

כעת נדון בדקות מענית לגבי הסתברויות לעצירה... $1, 2, \dots, n$. כאמור גמתארת הסתברות לעצירת ריצה של רשות באיטרציה ch . באופן פורמלי λ היא הסתברות מותנית של עצירה בשלב ch בהינתן אי עצירה (המשך) בשלב (1-ח). זה, להבדיל מהסתברות לעצירה... $1, 2, \dots, n$ הבלתי מותנית לאחר איטרציה ch שניתן לחשב אותה באופן הבא:

$$p_n = \lambda_n \prod_{j=1}^{n-1} (1 - \lambda_j)$$

המאמר הנסקר מציין כי העבודות הקודמות ניסו לחתות דזוקא את n ולא λ .

נציין כי ... n מגדרה התפלגות הסתברותית תקינה כאשר מספר האיטרציות המקורי אינו מוגבל. מכובן שהוא עלול להיות בעייתי עבור שימושים פרקטיים של השיטה המוצעת. המאמר מציע שתי דרכי להתחום עם סוגיה זו ונדון בהן בהמשך הסקירה.

איך מtabצע חיזוי עם PonderNet: אחרי שהסבירנו מה הרעיון שעומד מאחורי PonderNet נשאלת השאלה: איך בעצם מבצעים חיזוי עם הרשות הזאת? כאמור, בכל איטרציה הפלט של הרשות מורכב מהחיזוי עבור המשימה

המקורית, יחד עם האומדן של הסתברות לעצירה לאחר האיטרציה הנוכחיית \hat{y}_n . אך איך אנו יודעים מתי לעצור את הריציה? פשוט מאד - מבצעים דגימה אחת של משתנה בינהר' עם הסתברות הצלחה \hat{y}_n ומחליטים על סמך התוצאה האם לעצור או להמשיך. ב的日子里 אחורי כל איטרציה "זורקים" מטבע (לרוב לא הוגן) כאשר על \hat{y}_n צדדים של כתוב "המשך" ו"עצור" כאשר הסתברות של "עצור" הוא \hat{y}_n . במקרה של עצירה החיזוי האחרון \hat{y}_n^* משמש כחיזוי סופי של PonderNet עבור המשימה שבנידון.

איך מאמנים PonderNet: פונקציית לואס של PonderNet בכל איטרציה $\dots, 2, 1 = n$ מורכבת משני איברים:

1. **הלוואס המקורי של משימת הרשת:** לואס על "aicots" החיזוי \hat{y}_n^* , כמו למשל (\hat{y}_n^*, \hat{y}_n) , כאשר \hat{y} הוא הלוייביל האמתי (ground-truth). במאמר השתמשו בלואס הריבועי או בקרוס אנטרופי.

2. **לוואס עבור הסתברות עצירת האימון:** איבר רגולרייזציה בצורה של מרחק KL בין התפלגות \hat{y}_n לבין התפלגות פרירור P_g . התפלגות P_g נבחרה ע"י המחברים בתור התפלגות גיאומטרית עם פרמטר מקונפג (היפרפרמטר) λ . כמובן בשבייל לחשב את מרחק KL בין \hat{y}_n לבין P_g צריך להריץ את PonderNet מספר מקסימלי של הרצות בלי לעצור אותן. למעשה לאיבר רגולרייזציה זה יש שתי מטרות עיקריות: הראשונה "לכפות" על הרשת להיעזר לאחר $\lambda/1$ הרצות (בממוצע) והשנייה היא למנוע מרשת להוציא כפלט הסתברויות אפשריות לעצירה כל הזמן (סוג של עידוד exploration).

נורמל של התפלגות .., 2, 1 = n, k : כתע אספק הבקרה לגבי הסוגיה של נורמל התפלגות .., n, k שהעלינו באחד הפרקים הקודמים. כאמור אנו לא יכולים להריץ את PonderNet לאורך N , מספר בלתי מוגבל של איטרציות ביישומים פרקטיים. המאמר קובע את המספר המקסימלי של הרצות N , וכל לראות שהסדרה $N, \dots, 1, 2, 1 = n, k$ כבר לא מהוות פונקציה התפלגות תקינה כי סכום של הסדרה אינו שווה ל 1. המאמר מציע שתי דרכי לנורמל של $N, \dots, 2, 1 = n, k$:

1. לנורמל באופן סטנדרטי באמצעות חלוקה של כל k בסכום של הסדרה.
2. "להעביר" את כל המסיה ההסתברותית הנותרת לפני האיטרציה الأخيرة להסתברות עצירה של האיטרציה الأخيرة N, k .

איך קובעים את מספר האיטרציות המקסימלי N : הנקודה המעניינת האחרונה שאנו רוצה להתייחס אליה היא בחירה של מספר האיטרציות המקסימלי N . כמובן ניתן לאופטם אותו כמו כל היפרפרמטר אחר, אבל המחברים מציעים להגדיר אותו דרך "שיעור של המסיה הסתברותית לעצירה של הריצה". כלומר בוחרים מספר חיובי κ (במאמר בחרו ב- 0.05) ומגדירים את N כמספר המינימלי של איטרציות הנדרשות כדי שהסכום של $\dots, 2, 1 = n, k$ יהיה גדול יותר מ- 1. זה כמובן נעשה במהלך אימון של PonderNet.

הישגי מאמר:

המחברים בחרו כמה משימות (שרובן "אין מככבות" במאמרים בנושא הרשותות) והראו כי הביצועים של השיטה המוצעת עדיפה על גישות "أدפטיביות" האחירות עבור כמה ארכיטקטורות של פונקציית המדרגה δ . למשל אחת המשימות היא חישוב של y_{target} עבור סדרה בינהר' ארוכה. ההשוואה התמקדה בעיקר בשיטה, הנקראת ACT,ACT.

שכנראה נחשבה ל SOTA לפני כן. המחברים הראו ש- PonderNet מצליח גם במשימות של question answering, הנקרא *aAb* (המורכב מ- 20 תת-משימות שונות). השיפור בBITSIM ה证实 בדרך כלל יכול להגיע לאוטן ביצועים לפחות 20% מאשר הגישות המתחרות.

ג.ב.

לא הבהיר בעבר מאמרם הדנים במודלים בעלי זמן ריצה אדפטיבי והיה מגניב לצלול לנושא החשוב הזה. המאמר קל לקרוא, הרעיון העיקרי שלו אינטואיטיבי ומובן להפליא אולם נראה כי כרגע אין הרבה שימוש ודוגמנים שנייתן ליישם אותו בהם.

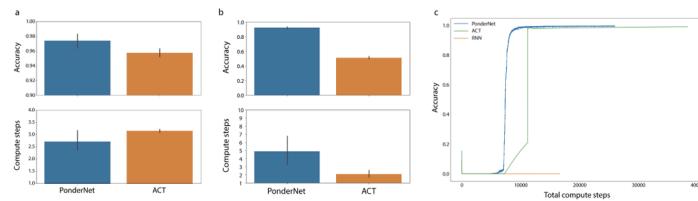


Figure 1: Performance on the parity task. a) Interpolation. Top: accuracy for both PonderNet(blue) and ACT(orange). Bottom: number of ponder steps at evaluation time. Error bars calculated over 10 random seeds. b) Extrapolation. Top: accuracy for both PonderNet(blue) and ACT(orange). Bottom: number of ponder steps at evaluation time. Error bars calculated over 10 random seeds. c) Total number of compute steps calculated as the number of actual forward passes performed by each network. Blue is PonderNet, Green is ACT and Orange is an RNN without adaptive compute.

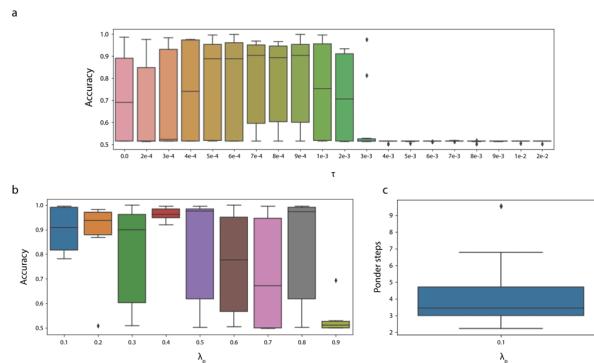


Figure 2: Sensitivity to hyper-parameter. a) Sensitivity of ACT to τ . Each box-plot is over 10 random seeds. b) Sensitivity of PonderNet to λ_p . Each box-plot is over 10 random seeds. c) Box-plot over 30 random seeds for number of ponder steps when $\lambda_p = 0.1$.

Review 62: Taming Transformers for High-Resolution Image Synthesis

פינט הסוקרי:

המלצת קרייה ממוקם: חובה ללא ספק!

בהירות כתיבה: גבולה

ידע מוקדם: הבנה טובה בגאנים, טרנספורמרים | VQ-VAE ז' הכרחית להבנת המאמר

ישומיים פרקטיים אפשריים: יצירת תМОנות באיכות מריהיבת (לא פחות!!)

פרטי מאמר:

لينك למאמר: [זמן להודה](#).

لينك לקוד: [זמן כאן](#) (בנוסף יש עד 3 מימושים "לא רשמיים")

פורסם בתאריך: 21.06.21, בארכיב (3)

הציג בכנס: CVPR 2021

תחומי מאמר:

- מודלים גנרטיביים לייצרת דאטा בתחום הייזואלי

ידע מוקדם:

- VQ-VAE
 - גאנים
 - טרנספורמרים
-

תמצית מאמר:



Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

מודלים גנרטיביים לייצרת פיסות חדשה בתחום היזואלי הגיעו לתחומי מרשים ב-3 השנים האחרונות. מודלים גנרטיביים כמו [VQ-VAE2](#) ו-[StyleGAN3](#) מצליחים לגנרט תמונות באיכות מאוד גבוהה במגוון רוחזיות. יתרה מזו התמונות הנוצרות באמצעות מודלים אלו נראות ממש פוטוריאליסטיות וכבר לא ניתן להבחין בין התמונה מגונרטת לטבעית".

רוב המודלים הגנרטיביים בעלי ביצועי SOTA בדומיין היזואלי הינם גאים בעלי ארכיטקטורה מבוססת על שכבות קובולוציה (למרות שבנה האחרונה [VAE](#)-ים, [מודלי דיפוזיה](#) ו- [גישה אחרת](#) התחלו להציג להם מלחמה). ביצועים עדיפים של רשתות קובולוציה בדומיין זהה נובעים מה- "inductive bias" האינגרנטי שמאפיין רשתות מסוג זה. Inductive bias של רשתות קובולוציה מנצל תלויות מקומיות חזקות הקיימות בתמונות הטבעיות. לעומת זאת לטרנספורמרים אין bias כזה שמקשה עליהם ללמוד את האופיינים של התפלגות הדאטה בדומיין התמונות הטבעיות. עקב לכך רוב הרשומות מבוססות טרנספורמרים בדרך כלל:

- או מצידות backbone הבני משכבות קובולוציה להפקת פיצ'רים "ליקאליים".
- או מוסיפים את ה-inductive bias לטרנספורמרים, ככלומר נתונים יותר משקל לקשרים בין פאצ'ים קרובים בתמונה.

המאמר הנזכר שילב את שתי הגישות הנ"ל ועשה את הדבר הבא:

1. אימון של VAE שבו הן המקדד והן המפענחת הינם רשתות קובולוציה. למעשה המחברים השתמשו ב-VAE מוקוונטס בו המרכיב הלטני (המכיל פלטים של המקדד) הוא למעשה אוסף דיסקרטי של וקטורים הנקרא codebook; ארכיב על זה בהמשך).
2. שימוש בטרנספורמר (ובפרט במפענחת שלו) עם תיבול קל של "inductive bias" המתאים לדומיין התמונות, בשביל ללמידה את ההתפלגות מעל המרכיב הלטני הדיסקרטי.
3. גנרטות של תמונה מתחילה מיצירה של וקטור לטני באמצעות הטרנספורמר המאומן. לאחר מכן מזינים את הוקטור הנוצר לרשת המפענחת של הטרנספורמר לייצור תמונה.

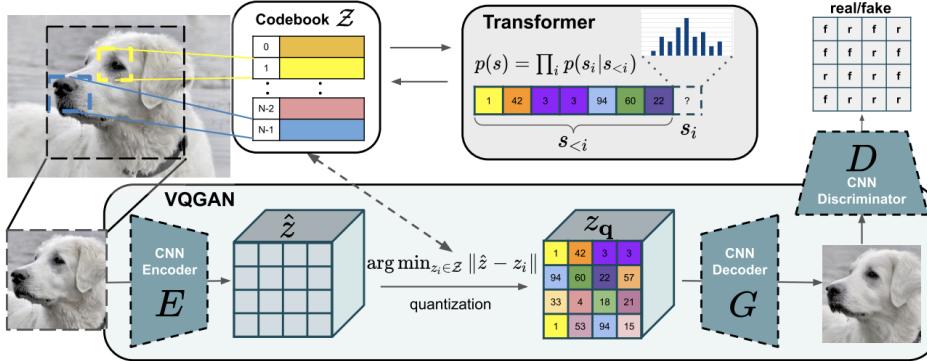


Figure 2. Our approach uses a convolutional VQGAN to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

עד כאן הכל טוב ויפה אבל בשם של המאמר מופיע גם מילה GAN והוא לא הוזכר עד עכשוו. למעשה VAE מהשלב הראשון מאמין בצורה לא סטנדרטית. לפונקציית הלוס הרגילה של VQ-VAE (שגם עברה מתיחת פנים) מוסיפים את פונקציית הלוס הסטנדרטית של הגאנ. ככלומר **בנוסף** למקודד ולמפענה של VQ-VAE מאמנים דיסקרטינטור D. מטרתה של D היא להזיהות אם התמונה שנוצרה באמצעות המפענה של E או שזו תמונה אמיתית.

למעשה VQ-GAN היא "חthonה משולשת" ומוצלחת במיוחד של VQ-VAE, גאן והטרנספורמר.

הסבר של רעיונות בסיסיים:

אחרי שהבנו מה מה הן אבני הבנייה העיקריים של NQ-GAN בואו נצלול לפרטים נביין איך ה喬יה המורכבות הזאת עובדת. בשביל להסביר את אופן האימון של NQ-GAN אנו קודם כל נרען בזיכרון מה זה VQ-VAE ועל בסיסו בניו שלב האימון הראשון של NQ-GAN.

מה זה VQ-VAE ?

VQ-VAE הינו סוג של Variational AutoEncoder (VAE) בעל מרחב לטנטי סופי (אך מאד גדול). נזכיר ש-VAE רגיל הוצע ב- 2014 על ידי Kingma ו-Welling. למעשה VAE מהווה הכללה של AutoEncoder סטנדרטי שהוא שיטה להורדת מידע לא לינארית. החידוש של VAE יחסית לאוטו-אנקודר הוא תוספת של הדרישה על התפלגות הייצוגים הלטנטיים של נתונים. ככלומר בנוסף לכך ייצוג לטנטי של נתונים צריך לשמר את התכוניות החשובות, וקטורי הייצוג עצם צריכים להיות מפולגים לפי התפלגות נתונה (לרוב גאומית סטנדרטית). תוספת זו מאפשרת לנו גנרט דאטה חדש באמצעות המפענה של VAE כאשר הקלט אליו הוא וקטורי ייצוג הנדגמים מהתפלגות זו. פונקציית LOSS של VAE מורכבת מסכום של LOSS השחזור הריבועי (המודד עד כמה טוב הצלחנו לשחזר את הקלט) ואיבר רגולרייזציה הכוונה התפלגות נתונה על הפלט של האנקודר (מרקף KL).

VAE הינו מודיפיקציה של VAE שבה המרחב הלטנטי (הנקרא codebook) הוא למעשה דיסקרט ומכיל מספר סופי של וקטורים הייצוג. כדי לגנרט LOSS חדשה בוחרים וקטור מהמרחב הדיסקרטי הזה (שמכיל כמות עצומה של וקטורים) ולקמן בכל זאת אפשר גנרט נתונים של נתונים מאוד מגוון ומעבירים אותו דרך המפענה המאמון.

כאשר משתמשים ב- VQ-VAE עבור ייצרת תמונות בדרך כלל מחלקים תמונה ל- M פאצ'ים כאשר כל פאץ' מקודד באחד הוקטורים מה-codebook. במקרה זהה תמונה היא בעצם מערך באורך M של וקטורים מה-codebook (עם חשיבות לסדר המקורי!). למשל עבור $M=8$ (במציאות יש הרבה יותר פאצ'ים) יציג של תמונה יכול להיראות כך: [2, 22, 46, 11, 98, 17, 9, 9] כאשר כל איבר במערך זה הוא מספר סידורי של וקטור יציג מה-codebook. האימון של VQ-VAE הוא קצת טרקי כי בנוסף לפרמטרים של המקודד והפענה צריך לאמן גם את וקטורי ה-`codebook`. וקטורים אלו נבחרים באמצעות **פעולה לא גדרה** מהפלט \mathbf{z} של המקודד (ובחרים את הוקטור ה- \mathbf{z} במנוחי מרחק L2) שמקשה על ה-`backprop` על \mathbf{z} . כדי שמתעניין איך מתמודדים עם הסוגיה זו ממליץ להעיף מבט ב- [ביבלו מעולה של ברקלி](#).

המבנה של פונקציית LOSS של VQ-VAE מורכב מlus השחזר הריבועי של VAE הסטנדרטי והmphak הריבועי בין פלט של המקודד והוקטור הקרוב מה-codebook (למעשה זה סוגה יותר מרכיב עקב הפעלה ללא גזירה שתוארה קודם).

פונקציית LOSS של VQ-GAN:

המאמר הנזכר בחר להחליף את LOSS השחזר הריבועי בסכום של:

- [הLOSS הסטנדרטי](#) שלGANים (שכמובן מצריך אימון רשת דיסקרימינטור).
- [הLOSS הפרספטואלי](#) ([perceptual loss](#)).

המאמר טוען כי פונקציית LOSS המוצעת באה לתת "טיפול שורש" בסוגיית הכוابت של VAE: התמונות המוטשטשות (יחסית לגאנים למשל) שהוא מיצרת. הסיבה לכך טמונה במבנה של איבר השחזר של פונקציית LOSS הריבועית של VAE, שמצוירת את השגיאה **המומוצעת** הגדולה לתמונה המשוחזרת להיות קרובה לתמונה המקורית, "**אך רק בממוצע**". תופעה זו, של קושי של רשותות עמוקות להתמודד עם תדרים גבוהים בקלט ובפלט, ידועה ונכתבו עליה לא מעט עבודות לאחרונה ([\[1\]](#), [\[2\]](#), [\[3\]](#), [\[4\]](#)).

LOSS הפרספטואלי:

כעת נסביר מהו LOSS הפרספטואלי L_{per} . המטרה של L_{per} היא למדוד דמיון בין הפיצ'רים של התמונה המשוחזרת לבין הפיצ'רים של התמונה המקורית. אבל אלו פיצ'רים ניקח בשbill השוואה הזה? הרי המטרה היא למדוד את "רמת פוטוריאלייטיות" של התמונה המשוחזרת אז הפיצ'רים צריכים לשחק את "המאפיינים החשובים" של התמונה המקורית. כדי להפיק פיצ'רים כאלה בדרך כלל לוקחים רשות מאומנת כמו [VGG](#) או [ResNet50](#) ומחשבים מרחק (בד"כ L2) בין פלטים של השכבות שלhn עבור התמונה המקורית למשוחזרת.

אצ"נ שהLOSS של גאן מחושב כמוצע על פני כל הפאצ'ים של תמונה בדומה ל-[PatchGAN](#). זה כופה על התמונה המגנרטת לא רק להיות כמה שיוצר "דומה לאמיתית" ממשחה אחת" אלא דורשת שדמיון זה יתקיים בכל פאץ'.

אני מאמין שסכום של שני LOSSים אלו מאפשרים למפענה של VQ-VAE ליצור תמונות מריהיבות.

"למידת" מרחב לטנטי של VQ-GAN:

אוקי, הצלחנו להפיק יציג חזק מהתמונה המהווה מערך של וקטורים מה-codebook (כל וקטור מיוצג ע"י המספר הסידורי שלו) כאשר כל וקטור מהו יציג של פאץ' של התמונה. בשלב השני של אימון VQ-GAN המטרה היא לאמן מודל לייצור של יציגים אלו (סדרות של יציגי פאצ'ים). כך נוכל להשתמש במודל זה לגנרטות של יציג לטנטי של תמונה שמצוין לאחר מכן למפענה לייצור תמונה.

איך עושים זאת? מאחר וניתן ליצור תמונה באופן אוטורגרטיבי (פאי' אחרי פאי') מאמנים מפענה של הטרנספורמר (המאמר השתמש בארכיטקטורה דומה לזה של [GPT2](#)) בשביל לחזות יציג לטנסי (מספרו הסידורי-book) של פאי' הבא בהינתן כל הפאצ'ים שכבר גונרטו. במשימה פאי'ים של תמונה הם "משחקים תפקידים של טוקנים" של משימות של השפה הטבעית.

מעשית לאחר סיום אימון של **VAE** בשלב הראשון, לוקחים את כל הייצוגים הלטנטיים של התמונות מהדאטסט ומאמנים דקודר של הטרנספורמר לחזות יציג של פאי' בהינתן הפאצ'ים הקודמים

הערה: לפני תחילת שלב האימון השני, "מקפיאים" את כל הפרמטרים של המקודד, המפענה ואת ה-codebook.

“תיבול” של **inductive bias**:

רגע, אבל מה עם התבליין מסווג **bias** **inductive** שהבטחת? קודם לכאן? המחברים מצאו כי שימוש בפאצ'ים גדולים מ- 16×16 פוגע ביצועים של המודל. מצד שני עקב משאבי החישוב המוגבלים שעמדו לרשותם, הם לא הצליחו לאמן טרנספורמר עבור יותר מ-256 פאי'. איך יוצאים מהמצב זהה ומגנרטים תמונות גדולות יותר מ- 256×256 ? פשוט משתמשים רק בפאצ'ים הקרובים לפאי' הנחזה - והנה קיבלתם ה-**bias** המובטח ():

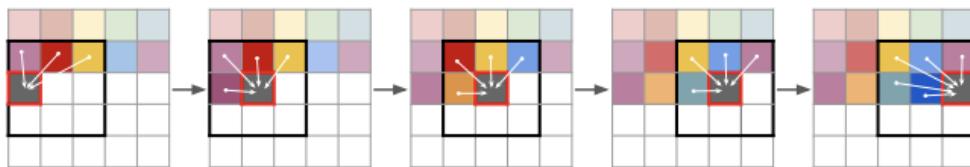


Figure 3. Sliding attention window.

סיכום קצר של שלבי אימון VQ-GAN:

- מאמנים VQ-VAE כאשר פונקציית הלוֹס היא שילוב [הלוֹס הסטנדרטי](#) של גאן והלוֹס הפרטפטואלי ([perceptual loss](#)).
- מקפיאים את כל הפרמטרים של כל הרשות שאומנו בשלב הראשון
- לוקחים את כל הייצוגים הלטנטיים של התמונות מהדאטסט
- מאמנים מפענה של הטרנספורמר לחיזוי הוקטורים הלטנטיים מהשלב הקודם
- **מגנרטים תמונות:** יוצרים יציג לטנסי של תמונה באמצעות מפענה מאומן של הטרנספורמר ומעבירים יציג זה דרך המפענה של VAE שאותן בשלב הראשון ונוטר ללא שינוי מכך

הישגי מאמר:

המחברים הראו שיפור מבחינת (FID) Frechet Inception Distance (IS) מול מודלים גנרטיביים חזקים עבור כמה דומיינים ורחלוציות.

Model	acceptance rate	FID	IS
mixed $k, p = 1.0$	1.0	17.04	70.6 ± 1.8
$k = 973, p = 1.0$	1.0	29.20	47.3 ± 1.3
$k = 250, p = 1.0$	1.0	15.98	78.6 ± 1.1
$k = 973, p = 0.88$	1.0	15.78	74.3 ± 1.8
$k = 600, p = 1.0$	0.05	5.20	280.3 ± 5.5
mixed $k, p = 1.0$	0.5	10.26	125.5 ± 2.4
mixed $k, p = 1.0$	0.25	7.35	188.6 ± 3.3
mixed $k, p = 1.0$	0.05	5.88	304.8 ± 3.6
mixed $k, p = 1.0$	0.005	6.59	402.7 ± 2.9
DCTransformer [48]	1.0	36.5	n/a
VQVAE-2 [61]	1.0	~31	~45
VQVAE-2	n/a	~10	~330
BigGAN [4]	1.0	7.53	168.6 ± 2.5
BigGAN-deep	1.0	6.84	203.6 ± 2.6
DDPM [49]	1.0	12.3	n/a
ADM-G, no guid. [15]	1.0	10.94	100.98
ADM-G, 1.0 guid.	1.0	4.59	186.7
ADM-G, 10.0 guid.	1.0	9.11	283.92
val. data	1.0	1.62	234.0 ± 3.9

Table 4. FID score comparison for class-conditional synthesis on 256×256 ImageNet, evaluated between 50k samples and the training split. Classifier-based rejection sampling as in VQVAE uses a ResNet-101 [22] classifier. BigGAN(-deep) evaluated via <https://tfhub.dev/deepmind> truncated at 1.0. “Mixed k ” refers to samples generated with different top- k values, here $k \in \{100, 200, 250, 300, 350, 400, 500, 600, 800, 973\}$.

ג.ב.

ממש אהבתني את המאמר כי הוא משלב גישות מאוד מעכניות בלמידה عمוקה: VQ-VAE, GAN וטרנספורמרים וגם מנצל את ה-inductive bias הקיימ בדומין היזואלי. מומלץ בחום רב!

Review 63: Soft-IntroVAE: Analyzing and Improving the Introspective Variational Autoencoder

פינט הסוקר:

המלצת קריאה ממייק: מומלץ לאוהבי מודלים גנרטיביים

בahirot chibah: גבוהה.

ידע מוקדם:

- הבנה של עקרונות [VAE](#) (למשל שימוש ב- ELBO לשערור של הנראות המירבית).
- עקרונות אימון של מודלים גנרטיביים עם פונקציית loss אדוורסרית (משחק max-min כמו בagan) - נא **לא לבלב שיטות אלו עם שיטות לאימון רשותות "חסינות" נגד התקפות אדוורסריות (adversarial examples)**.
- ידע בסתירות אם אתם רוצים להבין את הפרקים עם ההוכחות.

ישומים פרקטיים אפשריים: יצירה (אינפראנו) של תמונות באיכות גבוהה בסיבוכיות חישובית של VAE רגיל.

פרטי מאמר:

מאמר: [זמן להורדה](#).

קוד: [C&J](#)

פורסם בתאריך: 25.03.21, בארכ'יב.

הוצג בכנס: CVPR 2021

תחומי מאמר:

- מודלים גנרטיביים לייצרת דата ויזואלי (תמונה)

כליים מתמטיים, מושגים וסימונים:

- ELBO •
 - VAE •
 - אימון מודלים גנרטיביים עם פונקציית לוס אדוורסרייט
-

מבוא:



היום אני סוקר עוד מאמר המציג גישה לשיפור של איקות התמונות הנוצרות באמצעות מודל גנרטיבי מסווג VAE. אחת נקודת התורפה של VAE היא הקושי שלהם לגנרט טרדים גבוהים הנובע בין השאר מבנה של איבר השחזר (reconstruction term) המופיע בפונקציית הלוס שלה. איבר השחזר מכיל מרחק L2 בין התמונה המקורית למשוחצתה, המתבלט באמצעות העברת התמונה המקורית דרך האנקודר והדקודר. מזעור איבר זה גורם לתמונות המשוחצאות להיות דומות למקורות במעט שמתבטא ביצירת תמונות בעלות אזורים חלקים ונטולי פרטיים (טרדים גבוהים). ניתן לנסח את הבעיה באופן הבא: **VAE ידע "لتת" הסתבירותיות גבוההות לדגימות אמיתיות, אך מתקשה "لتת" הסתבירותיות נמוכות לדגימות מוטשטשות.** מכיוון שהגנים מודרניים כבר לא סובלים מבעיה זו (יש לגאנים חסרונות אחרים כמו מוד קולפס אוימון לא יציב) אז נראה לכאהר שילוב של VAE ו-VAE עשוי לתת מענה לסוגיית הטרדים הגבוהים של VAE.

בסקירה הاخדרונה תיארתי מודל שהוא שילוב של VAE וagan, המכונה VQ-GAN. המאמר הנסקר אż הציע להוסיף ל-VAE רשת הדיסקרמיןטור D במטרה להבחן בין התמונות המקורית לבין התמונה המשוחצתת (בדומה לغانים). ככלומר המחברים "הוסיפוagan" ל-VAE כאשר רשת הדקודר משחקת תפקיד של גנרטור שלagan. שילוב של הלום האדוורסרי הסטנדרטי שלagan בפונקציית loss של VQ-GAN הצליח להתגבר על בעיית ה"טרדים הגבוריים" בתמונות הנוצרות. למעשה רשת הדיסקרמיןטור **ማומנת להבחן בין התמונות המשוחצאות לבין המקוריות** וזה מאפשר ל-VAE-VQ לganרט תמונות פוטוריאלייטיות מדהימות.

תמצית מאמר:

כמובן שהווסף רשת נוספת ל-VAE מקשה על האימון של VQ-GAN. זה מוביל אותנו לשאלת הבאה: האם ניתן להשיג את היתרין של הלום האדוורסרי שלagan מבל' לשנות את המבנה המקורי של VAE? מתרבר שהתשובה על השאלה זו היא חיובית וזה בדיק מה שעשה המאמר שאסקור הפעם.

המאמר הנסקר מציע שיטה שכלהנו מאפשרת "לאמן את VAE בצדקה introspective", ככלומר VAE עצמו מואמן לחפש הבדלים בין תמונות מגונרטות באמצעות המבנה המקורי. כדי לעשות זאת רשת האנקודר **לבשת כובע של הדיסקרמיןטור**, כאשר הדקודר משחקת תפקיד של הגנרטור. בפרק הבא נתאר איך ניתן לבנות פונקציית loss המשלבת את הלום האדוורסרי שלagan עם הלום של VAE.

תקציר מאמר:

כעת נתאר את המבנה של פונקציית לוס של Soft-IntroVAE. קודם כל נזכיר שהמאמר הנסקר הוא שיפור של מאמר, המציג מודל בשם [IntroVAE](#) שלראשו הצלח להכניס גאנ-VAE ללא תוספת של רשותות נוספת.

נתחיל מתיאור של פונקציית לוס של IntroVAE. להבדיל מ- VQ-GAN "ההבחנה" בין תמונות מגונרטות למקוריות נעשית לא במרחב המקורי (של תמונות) **אלא במרחב הלטנטי**. הרעיון העיקרי של IntroVAE הוא לבנות משחק מיני-מקס בין הדיסקרימינטור (הנקרא כאן אנקודר E) לגנרטור (אשר בניי דקודר אבל כאן נקרא G) כאשר:

- **דיסקרימינטור** מצד אחד מנסה למצער את מרחוק KL בין התפלגות הייצוגים הלטנטיים של תמונות מקוריות לבין הפריבו (בדרכו כלל גאוסי איזוטרופי עם מטריצת קוריאנס I). מצד שני הוא מנסה להגדיל את מרחוק DL בין התפלגות הייצוגים הלטנטיים של התמונות המגונרטות לבין התפלגות הפריבו.
- **גנרטור** מצידו ינסה "לקמן" את הדיסקרימינטור תוך מצער של מרחוק DL בין התפלגות הייצוגים הלטנטיים של התמונות המגונרטות לבין התפלגות הפריבו.

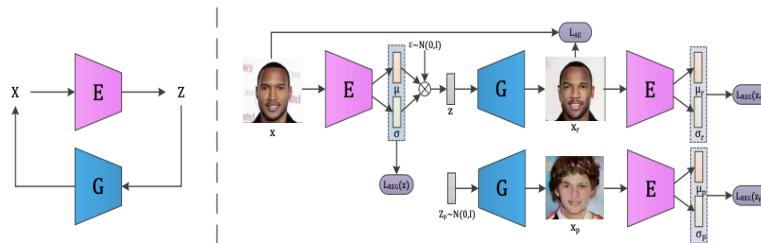


Figure 2: The architecture and training flow of IntroVAE. The left part shows that the model consists of two components, the inference model E and the generator G , in a circulation loop. The right part is the unrolled training flow of the proposed method.

אוקי', אבל איך זה נעשה בפועל. איך ניתן לשבל בין הלווי הרגיל של VAE לבין הלווי האדוורסרי שמוגדר לעלה? האמת שהמאמר בחר בגישה דיאינטואיטיבית. פונקציית לוס של האנקודר (דיסקרימינטור) מורכבת מסכום של L_E ו- L_G כאשר:

L_E : הלווי הרגיל של VAE (למעשה זה הלווי של [VAE](#)- β כאשר איבר KL בא עם מקדם β) ומרחוק KL בין התפלגות הפריבו (z) לבין ייצוגים לטנטיים של תמונות מגונרטות. מתמטית הלווי נראה באופן הבא:

$$L_E = L_{REG}(z) + \alpha \sum_{s=r,p} [m - L_{REG}(z_s)]^+ + \beta L_{AE}(x, x_r)$$

כך:

1. מסמן מרחוק KL ממוצע בין התפלגות של וקטורי z המוועה ייצוג לטנטי של תמונה משוחזרת לבין התפלגות פריבו של וקטור לטנטי של תמונה מהדאטסהט (כלומר התפלגות גאוסית $N(0, I)$).

2. שימושו לב Ci בבייטוי עבור L_E מופיע סכום של $(z_r) L_{REG} + (z_p) L_{REG}$ כאשר z_r ו- z_p הינם ייצוגים לטנטים של תמונות מגנרטות המתקבלות באופן קצר שונה - הסבר לגביהם ינתן בהמשך.

3. $L_{AE}(x, x_r)$ הוא LOSS השחזר בין תמונה x לבין גרסה המשוחזרת x_r .

4. m, β, α הם הייפר-פרמטרים ו- $(x^+ = max(0, x))$

L_G : מרכיב מסכם של LOSS השחזר אותו מרחק KL בין ייצוגים של תמונה מגנרטת ולזה של תמונה אמיתית:

$$L_G = \alpha \sum_{s=r,p} L_{REG}(Enc(x_s)) + \beta L_{AE}(x, x_r),$$

כאן $(x_s) Enc$ מסמן את התוצאה של העברת התמונה s באמצעות אנקודר.

המאמר מציע שתי דרכי לבנייה של ייצוגים לטנטים z_r ו- z_p של תמונות מגנרטות:

1. לוקחים תמונה משוחזרת x (לאחר העברתה של תמונה מהדאטסט דרך האנקודר ולאחר מכן דרך הדקודר) ומזינים אותה לאנקודר כדי לבנות את הייצוג הלטנטי z_r .
2. דוגמים מהפוריור (z) וקטור z ומעבירים אותו דרך הדקודר ולאחר מכן יוצרים ייצוג לטנטי z_p .

עת נדוע בחולשה המרכזית של E-IntroVAE שאוטה Soft-IntroVAE בא לתוקן. המחברים של המאמר הנסקרים טוענים כי **אימון של E-IntroVAE עלול לסייע מאי-יציבות עיקב רגשיותו הרבה לבחירת פרמטר α (מהלוס L_E)**. בנוסף, הניתוח התאורטי של משחק מינימקס שהוצג ב- E-IntroVAE לא לוקח בחשבון חלק מהאיםים של פונקציית הלוס והסתפק רק בניתוח של איברים המכילים מרחק KL.

כדי להתגבר על הקושי, הם מציעים להחליף את מרחק KL מהlös L_E ו- L_G לביטוי המלא של ELBO עבור התמונה המגנרטת. ככלומר פונקציות LOSS עבור הדיסקרימינטור (אנקודר) והגנרטור (דקודר) מקבלות את הצורה הבאה:

$$\begin{aligned} \mathcal{L}_{E_\phi}(x, z) &= ELBO(x) - \frac{1}{\alpha} \exp(\alpha ELBO(D_\theta(z))) \\ \mathcal{L}_{D_\theta}(x, z) &= ELBO(x) + \gamma ELBO(D_\theta(z)), \end{aligned}$$

כאשר $(z|_\theta) = D$ כלומר התפלגות אפואטוריית של תמונה x בהינתן ייצוג לטנטי z , α ו- γ הם שני הייפר-פרמטרים חיוביים. שימושו לב Ci מושווה זו מבטא את ה"משחק" בין האנקודר ודקודר כאשר בדומה ל- IntroVAE,

- האנקודר מאמין להבחן בין תמונות אמיתיות (O ELBO נמור) לתמונות הנוצרות באמצעות הדקודר (ELBO גבוה),
- הדקודר מנסה “לעבוד על” האנקודר ומנסה לגנרט פיסות דата כמה שיפור משכנעות.

מבחן מתמטית פונקציית הלווי של Soft-IntroVAE מקבלת את הצורה הבאה:

$$\begin{aligned}\mathcal{L}_{E_\phi} &= \mathbb{E}_{x \sim p_{data}, z \sim p(z)} [\mathcal{L}_{E_\phi}(x, z)] \\ \mathcal{L}_{D_\theta} &= \mathbb{E}_{x \sim p_{data}, z \sim p(z)} [\mathcal{L}_{D_\theta}(x, z)].\end{aligned}$$

כאשר p היא התפלגות של DATA אמיתית ו- (z) היא התפלגות פרIOR גאוסית איזוטרופית. מעניין כי להבדיל מ- IntroVAE, המשחק בין האנקודר לדקודר מחושב במרחב המקורי של DATA (תמונות) ולא במרחב הלטנטי. בעית אופטימיזציה זו פותרים בדומה ל-VAE הרגיל באמצעות Gradient Descent וטריק רפרמטריזציה (reparametrization trick).

שימוש לב Ci הנקודות של האיבר $\exp(\text{ELBO}(\theta))$ בביטוי של בעית האופטימיזציה של האנקודר, משעמו מינימיזציה של ELBO עבור דגימות המוגנרטות באמצעות הדקודר. נזכיר Ci $\text{ELBO}(\theta)$ מהו חסם תחתון על הנראות המירבית של $D_\theta(z)$ (תמונות מגנרטות). ככלمر מינימיזציה של $\text{ELBO}(\theta)$ עלולה לפגוע בא-“aicות” של התמונות המוגנרטות.

אם זה אכן המקרה, זה כמובן מאד בעית. אז בואו נבין מה קורה כאן? קודם כל המאמר מוכיח Ci שווי משקל של נאש של בעית האופטימיזציה של Soft-IntroVAE פוטר מתכנס ל- d^* כאשר:

$$d^* \in \arg \min_d \{KL(p_{data} \| p_d) + \gamma H(p_d(x))\}$$

כאשר $(q)H$ היא פונקציית האנטרופיה של התפלגות q . ככלמר עבור $0 \neq \gamma$ הפתרון אינו מתכנס להתפלגות של DATA p אלא מהו פתרון של בעית אופטימיזציה עם איבר רגולרייזציה שמעודד פתרונות בעלי אנטרופיה גבוהה מספיק. אבל האם זה טוב? המאמר טוען Ci:

“Soft-IntroVAE learns distributions with sharper supports than a standard VAE, but without negative effects such as mode dropping”

על כמה נתונים פשוטים (toy datasets). על הנתונים ניתן לראות Ci איות התמונות המוגנרטות מאוד גבוהה והתמונות עצמן די מגוונות.

משמעות: המאמר השתמש במודיפיקציה של ELBO שהוצע ב- β-VAE.

הישגי מאמר:

המחברים הראו כי הגישה המוצעת מצליחה להשיג ביצועים יותר טובים מכמה שיטות גנרטיביות אחרות כמו [CelebA-HQ](#), [StyleALAE](#), [GLOW](#), [Balanced Pioneer](#) ומראה תוצאות קרובות לגאנים עבור דאטאסט [FFHQ](#). כמוון הגאנים עדין מניחים את השיטות המבוססות VAE.

	CelebA-HQ	FFHQ
PGGAN [29]	8.03	-
BigGAN [4]	-	11.48
U-Net GAN [49]	-	7.48
GLOW [33]	68.93	-
Pioneer [22]	39.17	-
Balanced Pioneer [23]	25.25	-
StyleALAE [46]	19.21	-
SoftIntroVAE (Ours)	18.63	17.55

Table 3: Comparison of FID scores (lower is better) for CelebA-HQ and FFHQ datasets at a resolution of 256x256. Note the separation between GANs (top) and explicit density methods (bottom).

ג.ב.

מאמר עם רעיון מעניין, יש ניתוח מתמטי רציני של המודל, אינטואיטיבית מרשים. חסרות לי השוואות ביצועים עם עוד מודלים חזקים מבוסס VAE עבור דומיינים נוספים. אבל בסך הכל המאמר מומלץ בחום!

Review 64: PDE-GCN: Novel Architectures for Graph Neural Networks Motivated by Partial Differential Equations

פינת הסוקר:

המלצת קרייה ממיקם ומידות: מומלץ מאד למי שרצה להבין האם יש קשר בין משוואות דיפרנציאליות חלקיות (PDE) ובין רשתות ניירונים גרפיות (GNN)

בהירות כתיבה:

ידע מוקדם: הבנה בסיסית ב-PDE וב-GNN ובנוסף היכרות עם מושגי יסוד בתורת הגרפים במשוואות דיפרנציאליות.

ישומים פרקטיים אפשריים: פתרון של בעיות בביולוגיה חישובית, עיבוד תמונה, גרפיות, ראייה, ניתוח רשתות חברתיות ועוד.

פרטי מאמר:

lienק למאמר: [זמן להורדה](#).

lienק לקוד: ---

פורסם בתאריך: 26.10.21, בארכיב.

הציג בכנס: NeurIPS 2021.

מקום השתתפות של המחבר הראשון: אוניברסיטת בן-גוריון בנגב.

תחומי מאמר:

- רשותנו ניירונים על גרפים (GNN)
- אנליזה נומרית של משוואות דיפרנציאליות חלקיות
- פיזיקה חישובית (computational physics)

כליים מתמטיים, מושגים וסימונים:

- רשותנו קובולוציה על גרפים
- החלקת יתר (over-smoothing) ברשותנו ניירונים על גרפים
- אופרטורים דיפרנציאליים כמו דיברגנס, לפטיאן
- רשותנו קובולוציה קלאסיות (CNNs)

תמצית מאמר:

אחת הסוגיות המרכזיות ברשותנו קובולוציה על גרפים (GCN) הינה החלקת יתר (over-smoothing) של ייצוג DATAה המופק באמצעות הרשת. החלקת יתר של ייצוג מאופיינת בשינויים הולכים וקטנים של בין ייצוג של פיסות DATAשונות (כגון embeddings של קודקודים וקשתות) המופק באמצעות GCN. בעיה זו מחריפה בשכבות העמוקות של GCN ככלمر ייצוג קודקודים וקשתות בשכבות אלו נהיים זהה. תופעה זו היא הסיבה העיקרית לכך שכותבי מאמרם בתחום ה-GCN נוטים להסתפק בכמות קטנה קטנה של שכבות לעומת רשותנו הקובולוציה הקלאסיות (CNNs). חולשה נוספת של GCNs היא הצורך בהתאם לארכיטקטורה שלהן לדמיון ולמשימה. לדוגמה, כותבי המאמר PDE-GCN מצינים כי GCN, המבוצעת משימת סיווג בענן נקודות, עלולה להפגין ביצועים ירודים במשימת סיווג של קודקוד בגרף ציטוטי מאמרם (citation network).

המאמר המסורק מציע גישה מעניינת ומקורית הבאה לתת מענה לסוגיות אלו. השיטה המוצעת מנצלת את הקשר הקיים בין רשותנו ניירונים לבין משוואות דיפרנציאליות חלקיות (PDE). קשר זה נחקר בצורה אינטנסיבית בתקופה الأخيرة (1, 2, 3, 4, 5). הדינמיקה של מיפות הפיצרים לאורך שכבות של CNN ניתנת לתיאור

באמצעות מערכת דינמית המתוארת על ידי משוואות דיפרנציאליות חלקית. כוללן ניתן להתייחס לכל שכבה של CNNidal על "עד בזמן" של הגרסה הדיסקרטית של משוואות דיפרנציאליות חלקית. זהו משפט מאד אבסטרקטי הקשור שני דברים שאנו לא יכולים לחשב עליהם בעותה אחת, ולכן סקירה זו תנסה להתר במקצת את הקשר הקשור לעליון בונים מחברי PDE-GCN את התיאוריה במאמר.

הסקירה תחולק לשולשת החלקים הבאים:

- רקע על ההקלה בין מד"ח לרשות קובולוציה
- מבוא מזור לరשות קובולוציה על גרפים
- החיבור ביניהם והחדש העיקרי של המאמר.

חומר רקע:

הקשר בין רשותות לשולשות דיפרנציאליות:

נסזה עת להסביר (קצת בנפנוי יד"ם - אם אתם רוצים הסבר ריגורוזי תעיפוי מבט [PDEs and Convolutions](#)) או באחד המאמרים המוקשרים בפסקה הקודמת) איך PDEs קשורים ל-CNNs. נתחילה רשות נוירונים דיביטיסית ונראה איך ניתן "להפוך" אותה ל-PDE באמצעות מניפולציות דיאפשרות. נניח שיש לנו רשות המורכבת מ-T שכבות בעלות חיבור שיורי (residual connection). העיקרון שנדון בו כאן רלוונטי לכל CNN, אבל ההסבר בהירות יותר מאשר מתיחסים אל ארכיטקטורה מבוססת ResNet, ובכל מקרה זו ארכיטקטורה סטנדרטית של רשות קובולוציות.

כאמור, פلت x של שכבה T , ..., $1 = t$ ניתן לתיאור¹ כ:

$$x_{t+1} = x_t + W_2 \sigma(W_1 x_t),$$

כאשר W_1, W_2 הן מטריצות קובולוציה, ו- σ הינה פונקציית אקטיבציה לא לינארית כמו סיגמואיד או ReLU. במקרה, אם נעביר איבר אחד שמאליה נקבל את המשוואה הבאה:

$$x_{t+1} - x_t = W_2 \sigma(W_1 x_t),$$

ואם נכליל את המשוואה הזו ונחליף את המקדם 1 של צד ימין קבוע $1 \leq h$:

$$x_{t+1} - x_t = h W_2 \sigma(W_1 x_t)$$

$$\frac{x_{t+1} - x_t}{h} = W_2 \sigma(W_1 x_t)$$

از צד שמאל של המשוואה האחורונה מזכיר את הקירוב מסדר ראשון של $\frac{\partial x}{\partial t}$ בשיטת ההפרשים הסופיים (finite differences). זהו פותח ראשוני אל עולם הקשרים בין מד"ח (או מישדייפ, אם תתעקשו ) למידה عمוקה.

¹ב-ResNet אמיתי, על זה הצד ימין יופעל בדרך כלל אופרטור `sampledown` כלשהו כדי להתאים את המימדים שלו לממד השכבה, אבל לצורך הדיון זה פרט שולי. כמו כן נתעלם כאן מאיברי bias שמופיעים לעיתים קרובות.

cut נותר לנו רק למצוא אנלוגיה לנגרזרות המרחביות, שבלעדיהן אף משווהה דיפרנציאלית אינה מעניינת. לצורך כך, נתחילה מהכיוון השני ונתבונן תחיליה במשוואת הדיפוזיה הקלאסית:

$$\frac{\partial x}{\partial t} = W \nabla^2 x$$

כאשר W הינה מטריצה סימטרית וחיבורית-חלולוטין (positive definite) כלשהי,² הוא אופרטור הלפלסיאן ו- ∇ , \cdot , ∇ הם הדיברגנס² והגרדיאנט³, בהתאם. געניק לכל אחד מאיברי המשווהה את הגרסה הדיסקרטית שלו ונקבל:

$$\frac{x_{t+1} - x_t}{h} = W G^T G x_t$$

וכדי לראות את הקשר למשוואת הרצנט (ResNet) שלנו, נזכיר גם את משווהה זו ונגידו WG^T כמקומם הנכון, קיבלונו לבדוק את משוואת הרצנט.

לא נכנסו כאן לפתרים הטכניים של האנלוגיה זו ונסתפק בנפנופי ידים, אבל על בסיס זה מרשא לעצמו המאמר להתייחס לרשת CNN בתור הכללה של משוואת דיפוזיה, כאשר הקונבולוציות משחקות תפקיד של "אופרטורים דיפרנציאליים" נלמדים אשר מותאמים לדעתה שעליינו מתאמת הרשות ע"י תהליך האימון.

יסודות של GCN:

רשת נירונים לגראף מגיעה מהוצרך לעבד סיגנלים שחווים בעולם לא-סדור (unstructured). ניתן לראות כי גראף הוא הכללה של תמונה, בו הפיקולים במרוחים לא קבועים ומספר השכנים משתנה גם הוא. כך ניתן לראות רשת נירונים על גראף בתור הכללה של רשת נירונים סטנדרטית לסוגי DATA לא-סדורים. הפעם אנחנו מתעניינים ברשת GCN - Graph Convolutional Network - GCN, רשת אשר פועלת על גראף באמצעות קונבולוציות.

המטרה של רשת צזו היא ללמידה ייצוגים של הקודקודים והקשרות בגראף באמצעות העברת אינפורמציה מקודקודים וקשרות אחרים בגראף, בדרך כלל תוך התחשבות בקשרויות אשר נتوна לנו בגראף (ועם זאת, יש מודלים שימושיים קשחות או לומדים לחזות חלקיים חדשים מהgraף, הכל לפי צורכי המשימה). כמו כן, בהינתן ייצוג תחומי לכלי קודקוד וקשר בגראף, מעדכנים את הייצוג שלו על ידי הזרמת מידע מהקודקודים השכנים. העדכנים האלה בדרך כלל מtabסים על פילטרים ואופרטורים נלמדים, כמו ברשת CNN רגילה.

תקציר מאמר:

²תזכורת: אופרטור [הדיברגנס](#) מוגדר למרחב אוקלידי כסכום הנגרזרות החלקיות הcilinical של שדה וקטורי או סקלרי.

³ חשוב לשים לב: מדובר בגרדיאנט מרחבי ולא בגרדיאנט של פונקציית הלוס של הרשת, כפי שהתרגלנו. בהקבלה לרשת CNN שפועלת על תמונות - הגרדיאנט זהה יהיה הגרדיאנט (הdimensional) של התמונה, כמו שמקובל לחושב עליו בעיבוד תמונה קלאסי. דוגמאות ניתן למצוא בשפע בפייסבוק ובינטרנט.

כמו שכבר אמרנו בתחילת הסקירה הגישה המוצעת באה להתמודד עם החלוקת יתר של הפיצרים בין קודקודים וקשתות בשכבות העמוקות של GCN. המחברים מציעים ארכיטקטורה של GCN המבוססת על דיסקרטיזציה של **משוואת היפרבולית לא-لينארית** (כמוון, עם תנאי התחלת עברו ($f(t)$) ותנאי שפה שלא געסוק בהם כן):

$$f_{tt} = \nabla^* K \cdot \nabla f$$

המאמר מוכיח כי פתרון של משוואה זו לא גורם לשחיקה עבור ערכים גבוהים של f (מצר כי f הוא למעשה מספר השכבה ב-GCN בגרסה הדיסקרטית של המשוואת). המאמר גם מראה כי PDE המתארת GCN סטנדרטי הינה **משוואת דיפוזיה לא-لينארית**:

$$f_t = \nabla^* K \cdot \nabla f$$

למי שאינו מנוסה במיד"ח והמשוואות נראות לו זהות - שימו לב לכך שמאלו של שתי המשוואות. צד ימין אכן זהה. המחברים מראים כי הפתרון של המשוואת האחרונה מכיל מעט מאוד מידע אחר ודינמיקת העARBוב של משוואת הדיפוזיה מיצעה את כל הפיצרים, והתופעה זו חמורה יותר ככל שמספר השכבות גדול. לטענת המאמר זו הסיבה העיקרית לתופעת החלוקת היותר המתרחשת בשכבות העמוקות של GCN סטנדרטיות.

כדי להתגבר על החלוקת היותר זו, המאמר מציע לבנות ארכיטקטורה חדשה של GCN הנקראת PDE-GCN, בהתבסס על **משוואת היפרבולית ולא על משוואת הדיפוזיה כמו GCN סטנדרטיה**. בנוסף המחברים מגדירים גירסה דיסקרטית של הגרדיאנט המרחבי G של הגרף: עבור שני קודקודים i ו- j שמחוברים בקשתקה, הגרדיאנט G_{ij} ביניהם מוגדר כפרש של וקטורי הפיצרים (ייצוגי הקודקודים עבור השכבה הנוכחית) f_i ו- f_j המוכפלים במסקלות W_{ij} כלשהם (אשר יכולים להיות תלמידים או מהנדסים). נציין כי G הוא למעשה מיפוי ממוחלט הקודקודים V למרחב הקשתות של הגרף E .

את אופרטור הדיברגנס (∇^*), המופיע במשוואת היפרבולית ניתן לקרב באמצעות G^T (הdíברגנס על גוף בדרך כלל מוגדר כמיפוי ממוחלט הקשתות E למרחב הקודקודים V וכךנו פשוט משחלפים את G שהוא מיפוי מ- V ל- E) עכשו רק מותר להפעיל את שני האופרטורים ברצף כדי לקבל את הביטוי $G^T G$ – עבור האגף הימני של המשוואת היפרבולית. אחרי שהגדכנו את כל המשתנים ניתן לבנות את הארכיטקטורה הבסיסית של שכבת PDE-GCN:

$$\mathbf{f}^{(t+1)} = 2\mathbf{f}^{(t)} - \mathbf{f}^{(t-1)} - h^2 \mathbf{G}^T \mathbf{K}_t^T \sigma(\mathbf{K}_t \mathbf{G} \mathbf{f}^{(t)})$$

כאשר \mathbf{K}_t היא מטריצה קונבולוציה 1×1 לממדת ופונקציית האקטיבציה שנבחרה היא \tanh .

אחרי שהגדכנו את המבנה של PDE-GCN נותר לנו להסביר איך כל העסוק עובד בפועל. באמת שזה לא מסובך:

- לוחכים את הפיצרים של הקודקודים והקשתות,
- מעבירים אותם דרך שכבת קונבולוציות 1×1 ,
- מחשבים את הפלט של השכבות הבאות בהתבסס על המשוואת האחרונה.

הערה 1: המאמר גם מציע דרך לנצל פיצרים על קשתות (אם הם זמינים) כקלט ל-PDE-GCN.

הערה 2: המאמר גם מציע שיטה לבניית ארכיטקטורה של PDE-GCN המבוססת על שילוב (צירוף קמור בגודל) של משוואות הדיפוזיה והמשוואת ההיפרבולית.

הישגי מאמר:

המאמר השווה את ביצועי PDE-GCN עבור ערכים שונים של מספר שכבות עם מגוון של GCN-ים על כמה משימות ודתאחסטים שונים. המטרה הייתה להראות כי הארכיטקטורה המוצעת מצליחה להתגבר על בעיות החלקת יתר של הפיצרים בשכבות העמוקות של GCN. הדרך הטבעית להוכיח זאת היא להראות כי לא נصفית ירידה ביצועי GCN כאשר מושגים אלה שכבות (כਮון שם החלקת יתר עדין קיימת, הוספה שכבות לרשות עלול לפגוע ביצועים). המאמר אכן מראה כי בכל המשימות שנבחנו ביצועי PDE-GCN לא סופגים ירידה (אלא משתפים קצת ברוב המקרים). בחלק מהמשימות המאמר אפילו מציג SOTA חדש, אך לא בכלל.

כאן צריך לציין שבהשוואה לרוב הארכיטקטורות האחרונות של GCN, הביצועים לא תמיד טובים יותר והרשות מהשנים האחרונות יודעות להתמודד בהצלחה גם עם הרבה שכבות (במאמר נבדקו עד 64). המאמר למעשה מציג שיטה נוספת (ופשטה יחסית) לתכנן רשתות GCN עמוקות ([לעומת Kipf and Welling של vanilla GCN](#) [מ-Dropedge-2016](#)).

ג.ב.

המאמר מציע גישה חדשה ומענית לבניית ארכיטקטורה של GCN באמצעות משוואות דיפרנציאליות חלקיות. הגישה מצליחה להתגבר על בעיות החלקת יתר המתרכשת כאשר מושגים שכבות לרשות. לדעתי, זאת גישה מענית לתוכנן ארכיטקטורות דיפ כבלי ו-N-CNN בפרט ואנו רואים במאמר זה (ובסקירה זו) פתח לתוך עולם מעניין, אך פחות מוכר. בראיה זו, סקירה היא הראשונה מבין מספר סקירות מאמרם בעולם האנלוגיות בין מד"ח ובין ארכיטקטורות דיפ.

#deepnightlearners

shitof פועלה: הபוסט נכתב על ידי מיכאל (מייק) ארליךsson, PhD, Michael Erlhsson, PhD ועדו בן-יאיר.

Review 64: TRAIN SHORT, TEST LONG: ATTENTION WITH LINEAR BIASES ENABLES INPUT LENGTH EXTRAPOLATION

כמו שאותם יודעים אחת החולשות העיקריות של הטרנספורמר היא הסיבוכיות הריבועית שלו ביחס לאורך סדרת הקלט. יצאו מאות מאמרים המציעים וריאנטים של הטרנספורמר עם סיבוכיות נמוכה יותר.

המאמר זהה לוקח כיוון די מפתיע בהתמודדות עם סוגיית הסיבוכיות הריבועית. הוא שואל את השאלה הבא - אולי נאמן טרנספורמר לאורכי קלט לא גדולים ואז נרי'ץ אותו עם קלטמים ארוכים יותר באינפראנס? זה נשמע מוגניב אבל האם זה בכלל יעבד? באופן לא מפתיע התשובה טמונה בקידודים מיקומיים (PS-[positional encodings](#)) שימושיים בהם בשביל להعبر לטרנספורמר מידע עם מיקום של טוקנים בסדרה.

מתברר שלבסוף יצאו מספר מחקרים המציעים PS שונים לשיפור ביצועי טרנספורמר (ולא דוקא למטרת המתווארת לעיל). אך המאמר הזה מציע שני שינויים ל-PS המאפשרים לטענות "לאמן קצר ולהריך אורך". במקרה לחבר את PS לייצוג הטוקנים כמו שעשו בטרנספורמר האמייתי, המחברים כן מציעים לעשות את הדבר הבא:

1. מוחשבים מכפלה פנימית של וקטורי query ו-key.
2. מחברים למכפלה זו את המורח השלייל בין הטוקנים, מוכפל בקבוע ω (שהוא שונה בכל head של הטרנספורמר). למשל למכפלה פנימית של טוקן 3 ו-8 מוחברים (5-5).
3. ובונים מכפלה בין וקטורי query ו-key עם המקדים שהישבנו בסעיף 1-2 (למעשה השינוי הזה הוצע במאמר אחר). ההגיון כאן שכל המידע על מקום כבר מוקוד ב-key ו-query ואין צורך להוסיף אותו ל-value.

איןטואיציה: למעשה הקידוד הזה אומר את הדבר הבא: **בא נקטין את הקשר** (מקדם רלוונטיות) בין טוקנים ככל שמדובר בין הטוקנים גדול. המקדם ω השונה בין ראש הטרנספורמר מאפשר לשנות במידה "ההקטנה" של הרלוונטיות.

זה כל הסיפור. שימושו לב שהמאמר מציע את ה-PS שלהם רק עבור טרנספורמר אוטורגרטיבי (שנקרא decoder) אבל אני לא רואה שום סיבה לא להשתמש בזיה בצורה דו-כיוונית.

ה-PS המוצע מאפשר להאט את הירידה בביטויים בתರחיש שבו מאמנים טרנספורמר לאורך קלט נתון (512) ואז עושים אינפנס לאורכי קלט כמו 1024, 2048 ועודים יותר (יחסית ל-PS האחרים כולל המקורים של הטרנספורמר).

מאמר: <https://arxiv.org/pdf/2108.12409.pdf>
קוד: https://github.com/ofirpress/attention_with_linear_biases
ויניק: <https://www.youtube.com/watch?v=-Kgxv64aG3o>

Review 66: VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning

מאמר מעוניין של LeCunn et al. המציע שיטה לבנייה של ייצוג נתונים לא מתויג (self-supervised). המחברים עצם קוראים לשיטה המוצעת "ridiculously simple" והיא בהחלט עונה על ההגדרה הזאת.

מה התכוונה החשובה ביותר שאנו צריכים מייצוג של נתונים? אנו רוצים שייצוגים של פיסות נתונים דומות יהיו קרובים במרחב היצוג (אגומנטציות של אותה תמונה למשל). בנוסף רצוי מודד שהייצוגים של פיסות נתונים לא דומות יהיו רחוקים. כדי להשיג את התכוונה הראשונה (קרבתה היצוגים של דוגמאות קרובות) ניתן לאמן את הרשת עם פונקציית LOSS הממדעת מרחק בין יצוגים של דוגמאות קרובות. אבל אימון עם פונקציית LOSS כזו עלול להוביל לכך שככל היצוגים יהיו זהים ואז התכוונה השנייה (יצוגים רחוקים של דוגמאות לא דומות) לא מתקיימת. אז איך מתגברים על זה?

בגدول יש שתי גישות עיקריות לבניית יצוגים עבור נתונים לא מתויגים:

- משתמשת בדוגמאות קרובות ורחוקות ודורשת למצויר את המרחק בין היצוגים של דוגמאות קרובות ולמקסם מרחק בין ייצוגים של דוגמאות שליליות (רחוקות). זה נעשה בדרך כלל באמצעות שיטות של Contrastive Loss.
- מזערת מרחק בין דוגמאות חיוביות (קרובות) ומערבת רגולציה לא מפורשת (לא איזה איבר 2), הבאה למניע התכונות של כל היצוגים לווטו וקטור (SwAv, BYOL) שני סקרים הן דוגמאות קלאסיות של הגישה זו.

אז מה מציע המאמר של ChChun? מה השיטה "ridiculously simple" המוצעת?

למעשה מה אנו רוצים מהיצוגים של הדאטה?

- קרבה של היצוגים של דוגמאות קרובות
- שונות מספיק גבוהה בכל מימד של וקטורי ייצוג (שלא יצא 3.2 בימיד 68 של כל וקטורי היצוג)
- קולציה גבוהה בין וקטורי ייצוג של דוגמאות שליליות

פונקציית loss של VICReg מכילה 3 איברים שמטרת כל אחד מהם להשיג את מה שמופיע בסעיפים 3-1 לעיל. פשוט כך

באופן מפתיע VICReg מצליח להשיג ביצועים קרובים מאוד (הבוחנים את עצמת היצוג) לשיטות מתחרות מורכבות הרבה יותר במגוון של שימושות.

שתי הערות לוסף:

1. המאמר נדחה NeurIPS 2021 (לא התעמקתי בסיבות)
2. שקראי סקרירות על בניית ייצוגים לדאטה לא מותוו חשבתי למה לא לעשות את בזרה הכל פשטה שיש, קרי VICReg. אבל לא טרחות למשולב לבדוק את זה כי היה בטוח שהוא לא יעבד ואחרים בטענה את זה. בדיעבד היה מוציא על זה מאמר והוא היה נדחה כמו זה של ליקון ()

[lienק למאמר:](https://arxiv.org/abs/2105.04906)

פוסט דחיה של ליקון:

https://m.facebook.com/story.php?story_fbid=10157921242067143&id=722677142

Review 67: Grokking: Generalization beyond Overfitting on small algorithmic datasets

מאמר די מסקרן של openai שיצא לאחרונה. אסקור אותו קצורות כמו שמקובל ב#shortdeephngtlearners

התענה העיקרית של המאמר נשמעת לא מוסובכת. הרעיון שאמנם ניתן רשות יותר מדי זמן, מלשב מסוים היא תיכנס למוד של overfitting כלומר שגיאת ההכללה שלה תתחיל לעלות. המחברים טוענים שגם נאמן אם נמשיך לאמן את הרשות גם לאחר שהיא נמצאת ב- overfitting מצליחו שלב שגיאת ההכללה שלה תתחיל לרדת כלומר הביצועים שלה על טսט (או וידציה) יתחלו לעלות.

המחברים זיהו תופעות כאלה כאשר הם אימנו רשות על דאטסהטים "אלגוריתמיים" קטנים. מה זה דאטסהט אלגוריתמי ואיזו משימה נתונים לרשות במקורה, אתם שואלים? אחת הדוגמאות הוא מטריצה המתארת פעולות במרחב התמורות (פרמוטציות). כלומר לוקחים תמורות בגודל מסוים (נגיד 120 תמורות עבור 5×5) ובונים

מטריצה בגודל 120x120 שמשבצת (ז,ז) מכילה תוצאה של הרכבת תמורה זק ו- נק. אך מושחים כמו מהמשbowות במטריצה זו ונותנים לרשף ללמידה יותר. זה משווה די מגניב שטרם ראייתי...

עכשו כמה מילימ' על התופעה עצמה. התופעה הנכפית היא סוג של double descent, הנזכר epoch-wise (הטרמינולוגיה מ- מ-).

<https://www.kdnuggets.com/2020/04/double-descent-hypothesis-bigger-models-more-data-hurt-performance.html>. אנו מכירים היטב double descent מושג Model-Wise שמשמעותו היא שהגדלה של מספר הפרמטרים לרשות גורמת לתופעה דומה משלב מסוים (שגיאת הכללה עולה ואחר-כך יורדת). מודה שבתחלת התבבלתי בעצמי וטענתי התופעה המתוארת במאמר היא לא double descent אך אחרי השיחה עם Um Liron Itzhaki (תודה רבה על תובנות מאוד מעניינות) הבנתי שיש לי טעות בזיהוי. דרך אגב גם Chochik Kilcher בסרטון שלו טוען שזה לא double descent.

למה תופעה כזו מתרחשת. Misha Belkin הגדיל חוקר את התופעה המעניינת הזו כבר כמה שנים אבל עדין אין הסבר מתמטי לכך (תקנוotti אם אני טועה כאן). לגבי הסברים בנפנופי ידיים יש לי שניים (אחד של לירון אחד לי):

1. **מריג'ינים:** כאשר ממשיכים לאמן רשות (לטיזוג נגד) כשהיא נכנסה למוד overfitting והלווט ממשיך לרדרת לאחר האפס, הרשות עשויה לגלות "מודול (המוגדר ע"י הפרמטרים של הרשות) המפריד בצורה רוווחת יותר בין הקלאסים" (עם מריג'ינים גדולים). מודול זה "חסין" לרעש בדאטה ומכללים מספיק טוב לאחר שบทחלתו היא מצאה פתרונות עם מריג'ינים קטנים ששובלים לשגיאת הכללה גבוהה.

2. **מינימום של פונקציית LOSS:** יש הנחה שפתרונות בעלי שגיאת הכללה נמוכה "Namely" במינימום בעלי עיקריות נמוכה (לא חדים) של פונקציית LOSS. ככלומר כאלו שפונקציית LOSS סביבם לא משתנה הרבה וערכיה סביב נקודת מינימום זו היא נמוכים. נקודת מינימום כזו היא יותר טובה ממינימום חד (שאיפלו בסביבתו הקטנה פונקציית LOSS מקבלת ערכים גבוהים ממשמעותית מזה בנקודת מינימום). אז יש מצב שמאמנים רשות מספיק איפוקים אך יש סבירות גבוהה להגיע מתיישה לנקודה כזו (לא דהה) שכבר "קשה" לצאת ממנה כי פונקציית LOSS מקבלת ערכים נמוכים סביבה.

כמובן של הסברים בנפנופי ידיים אבל לדעתו עשויים להיות מועילים לחשיבה על התופעה המעניינת הזו

lien: https://mathai-iclr.github.io/papers/papers/MATHAI_29_paper.pdf

סרטון של יניק: <https://www.youtube.com/watch?v=dND-7lwqrpw>

Review 68: PATCHES ARE ALL YOU NEED?

אתם אולי מכירים את חולשתי למאמרים מכילים ביטוי "need you all". מרגיש מחייבי לסקור קצורות כל מאמר זהה שעיני נתקלת בו. אצין שלא הייתי סוקר את המאמר הזה אחרת 😊

הפעם במקור נמצא לא אחר אלא פאץ'. לאור ההצלחה של הטרנספורמרים גם בדומיין היזואלי (Visual Transformer ודומיין). אזכיר שהקלט לטרנספורמר ויזואלי הינו פאצ'ים של תמונות. ככלומר מחלקים תמונה לפאצ'ים, מייצגים כל אחד באמצעות וקטור (בדרך כלל פשוט משטחים את מטריצת הפיקסלים בפאץ' ומכפילים במטריצה).

از המאמר שואל את השאלה הבאה: האם הצלחה של הטרנספורמרים היזואליים נובעת מועצמתו של ארכיטקטורת הטרנספורמר או שזה פשוט עניין של בנייתו של קלט לרשות, כלומר הפאצ'ים. בשבייל לבדוק את העניין זהה המחברים בנו רשות ממש פשוטה (איפילו יותר פשוטה מ- MLP-Mixer).

המפורסם) כשהקלט שלו הוא היצוגים של פאצ'ים. הרשות בינויו בצורה מאוד פשוטה מבלוקים residual שכל אחד מהם מכיל קונבולוציה depthwise ו- pointwise אחד אחריו השני. בסוף יש שכבת FC pooling ו-FC. זה זה.

עם ארכיטקטורה מאוד פשוטה זו הם הגיעו לbijouxם ברי השוואה עם קונפיגורציה מסוימת של ViT (הם גם השוו ביצועים עם איזה סוג של טרנספורמר, הנקרא DeiT). הם גם עשו השוואה על עם מודלים פשוטים כמו ResNet152 ו- ResMLP.

אבל

הבדיקה נעשתה על שתי מישיות בלבד (סיווג על ImageNet ו- CIFAR10). הם לא ביצעו שום אימון self-supervised והוכיחו שהיצוגים שהארכיטקטורה שלהם, שקיבלה שם ConvMixer, יודעת להפיק, חזקים יותר. אלא רק אימון על נתונים מסוימים (תקנו אותנו אם פספסתי משהו).

בקיצור מאמר די מאכזב, השם מאוד באזדי אך לצערנו לא הולם את התוכן...

אשמה לדעת מה דעתכם....

למאמר: <https://openreview.net/forum?id=TVHS5Y4dNvM>

קוד: <https://github.com/tmp-iclr/convmixer>

Review 69, Short: SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRE-TRAINING WITH WEAK SUPERVISION

בהתחלת רציתי לכתוב סקירה קצרה בסגנון #shortdeepnightlearners אבל למעשה ניתן לתקן את המאמר הזה בכמה משפטים בלבד.

אתם בטח זכרם את CLIP ו- E-DALL-E שהצליחו להפיק ייצוגים חזקים (ומתואמים!!) של DATA ויזואלי וטקסטואלי באמצעות אימון פשוט באמצעות שימוש בגישה הלוס הניגודי. כל זה כמובן געשה על תמונות עם כותרות מה האינטרנט. מחברי SimVLM בחרו בגישה פשוטה יותר (לדעתי) - הם מאמנים רשות שבاهינתן תמונה והתחלה הכותרת שלה, חוזה את החלק השני של הכותרת.

איך עושים זאת? מعتبرים פאצ'ים של תמונה דרך האנקודר של הטרנספורמר יחד עם החלק הראשון של הכותרת. צריך לציין שאת הפאצ'ים מعتبرים קודם דרך רשות באקבון קונבולוציונית (שלום bias inductive).

לאחר מכן חוצים את החלק השני של הכותרת עם הדקودר של הטרנספורמר.

היצוגים המופקים בשיטה אימון הצלicho להציגים טובים במגוון משימות כגון סגמנטציה ומענה על שאלות לתמונה נתונה (question answering open visual question zero-shot). גם לחת כותרת לתמונה הרשות יודעת לתת בזורה די טובה וכך זה לא מפתיע...

למאמר: <https://arxiv.org/abs/2108.10904>

פרויקט: <https://ai.googleblog.com/2021/10/simvlm-simple-visual-language-model-pre.html>

רב טוב חברים,

Review 70, Short: Typical Decoding for Natural Language Generation

מאמר מסקרן המציע דוגמה חדשה ממודלי שפה מאומנים. כמו שאתם בטח יודעים מוקובל לדגום מילים (טוקנים) עם הסתברויות הגבוהות ביותר (בහינת המילים הקודמות). ככה עובדים למשל nucleus ו-top-k.

המחברים טוענים שפרדיגמת דוגמה זו מובילה לטקסטים "משמעותיים מדי" יותר מדי דומים לדאטסהטים שהמודל אומן עליהם.

המאמר מציע פרדיגמת דוגמה אחרת מבוססת על מילה טיפוסית (מבחינת האנתרופיה של המילה בהינתן הטקסט לפני) ולא על מילים הסבירות ביותר. בגודל מחשבים את האנתרופיה של מילה על סמך ההתפלגות שלה (softmax). לאחר מכן מחשבים את המידע של כל מילה (\log_2 של ההסתברות שלה). בסופו דוגמים מילים כאשר ההסתברות של כל מילה פרופורציונלי למරחק ההפוך של המידע של מהאנתרופיה של התפלגות המילה). כמובן ככל שהמידע של מילה קרובה יותר לאנתרופיה היא תדגם בהסתברות גבוהה יותר.

כך אנו דוגמים מילה טיפוסית (אנתרופיה היא למעשה המידע הממוצע של המילה) ולא המילה הסבירה ביותר.

יש במאמר דיון מעניין על אספקטים ליניאריים של הגישה המוצעת.

lien: <https://arxiv.org/pdf/2202.00666.pdf>

סרטון של יאניק: https://www.youtube.com/watch?v=_EDr3ryrT_Y&t=2484s

Review 71: Deep Reinforcement Learning for Cyber System Defense under Dynamic Adversarial Uncertainties

פינת הסוקר:

המלצת קריאה מעדן ומיק: מומלץ לאנשים העוסקים בתחום ה-Cybersecurity או לאנשים שאוהבים
Reinforcement learning

בהירות כתיבה: טוביה

ידע מוקדם:

- הבנה בסיסית בלמידה عمוקה מחיזוקים (DRL)
- ידע בלמידה عمוקה

ישומים פרקטיים אפשריים: מערכת לזיהוי התקפות סייבר

פרטי מאמר:

לינק למאמר: [זמן להורדה](#).

לינק לקוד: לא אוטר

פורסם בתאריך: 3.2.2023, בארכ'ב.

הציג בכנס: -

תחומי מאמר:

- למידה عمוקה מחזזוקים
- אבטחת סיבר

כלי מתמטיים, מושגים וסימונים:

- SDP

מבוא:

פיתוח אלגוריתמים האחראים על קבלת החלטות עבור מערכת הגנת סיבר איננה ממשימה פשוטה. הקשיי העיקרי נובע מרמת דינמיות גבוהה שקיים בסביבת תוכנה המילוט גם את כל האבטחה (נקרא לו גם כל הגנה בהמשך הסקירה) וגם את התוקף, המשפיעים ייחד על הסביבה ומושנים את מאפייניה בכל נקודת זמן. בעיה נוספת הינה מגבלת המשאים המוקצים לכל הмагן. משאים אלו כוללים זמן מיקסימלי לטיפול בפרצה, משאבי אנוש זמינים וכדומה. ככל פעולה של הכלי המגן הוא צריך להתחשב במוגבלות אלו. למשל אם כל כוח העבודה שייכל לטפל בפרצה מסוימת תפוס, המודל לא יוכל להתריע על פרצה חדשה. לאחר שהכלי המגן מבצע פעולה, מספר המשאים מתעדכן בהתאם, דבר המשפיע על הסביבה.

בשנים האחרונות עם התקדמות הטכנולוגיה, חוקרים רבים ניסו לשלב למידה عمוקה מחזזוקים (DRL) בתוך כלים שונים בתחום הסיבר. הפטرونות הללו ניסו למדל את הבעה כ-*chance* *Markov Decision Process* (*POMDP*). הבעה בפתרונות הללו הינה שהן התמקדו בסוג התקפות ספציפי, במרחב פעולות מצומצם מאוד של הכלי המגן יכול לבצע ובنוסף הינו שקיים מודל לסביבה למורoutes האמיתית מורכב מדי כדי למדל אותו בכלים מתמטיים פשוטים יחסית.

כדי לתת מענה לביעיות אלו, כתבי המאמר מציעים שיטה המתמקדת בפיתוח כל מגן שיפעל כסוכן DRL שמטרתו להתמודד נגד תוקף מתחכם ובעל יכולות מגוונות, כיוון שבעולם האמיתי אין סוגי התקופים אוטם הוא (הסוכן) יגש. השיטה גם שכוללת פיתוח מסגרת (Framework) בה יהיה ניתן לבדוק ולהעיר את ביצועי הכלי המגן בצורה אמינה. מטרת הסוכן כפי שהגדירו אותה במאמר הינה ללמידה את אסטרטגיית ההגנה הטובה ביותר

אשר תשפייע כמה שפחות על התהליכי הקיימים בסביבה שבה הוא מותקן (למשל להימנע ככל האפשר מפעולות כגון ניתוק מחשב מהרשת) ותדע להתחשב בכל שלבי ההתקפה האפשרים של התקוף.

לכן, החוקרים השתמשו ב-[MITRE ATT&CK Framework](#) כדי למדוד את כל הפעולות אותן יכול לבצע התקוף ובאותה נשימה לבדוק את ביצוע הכליל המגן. החוקרים מדגישים שהתרומות העיקריות במאמר זה הן:

1. ההשווואה של ביצועי אלגוריתמי DRL שונים תחת תנאים סבירים אמיתיים.
2. פיתוחה של סביבת סימולציה חדשה המדממת את הסביבה האמיתית שאוטה יראה המגן בחוץ באופן מדויק. הסביבה כוללת תוקף מותחכם, המאפשרית בחוסר ודאות גבוהה, תהליכי אמיתיים ואילוצים אמיתיים.

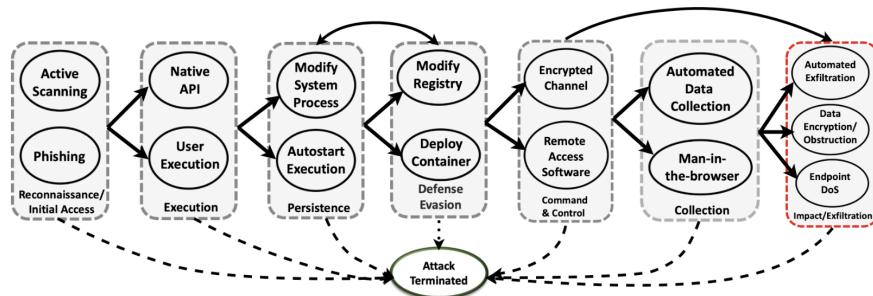


Figure 1: Multi-stage attack propagation represented with MITRE ATT&CK Tactics and Techniques. (Note: A directed edge between an attack tactic and technique specifies that the attacker may try to implement that technique next after achieving the objective of the attack tactic. Bidirectional arrow represents that Defense Evasion can come before Persistence.)

התקוף:

מטרת התקוף אותו מימושו במאמר הינה לנوع מתחילה גրף (תזרים) ההתקפה (אותו ניתן לראות באIOR 1) לסוף המסלול. כדי לעשות זאת מומש מודל מבוסס אסטרטגיה. לפי האסטרטגיה אותה מיישם התקוף מגדרה הצלחתו בשלב נתון (במאמר זה נקרא טקטיקה) במסלול אם הוא הצליח להריץ את אחת ההתקפות (נקרא טכנית) במאמר) באותו שלב. למשל, אם התקוף הצליח להריץ את התקפה (טכנית) הנקראת Active Scanning Procedure (Procedure) הצליח להגיע לשלב הנקרא Reconnaissance. כדי לבצע התקפה התקוף מירץ תהליך התקפה (Procedure) אשר בפועל ממש את התקפה. ככלומר התקפה מוגדרת בתור סדרת מהליכים שעל התקוף לבצע ולהילך יכול להיות קוד בשפת תכנות כלשהו אשר מאפשר לתוכף להציגים את מטרתו. כדי לעשות זאת את התקוף מנצל חולשות הקיימות במערכת ולא טופלו כהוגן בכל נקודת זמן, מיקום התקוף הינו בשלב האחרון בו התקוף הצליח לבצע בהצלחה התקפה כלשהיא. התקוף יפסיד במקרה אם פעולה התקיפה שלו כולה בשל פעללה של המגן (כלומר הוא נחשף). ההנחה היא שהתקוף רשאי לשנות את אסטרטגיית התקפה שלו בכל נקודת זמן בהתאם למצב המערכת ותוצאת התקפה הקודמת שלו. לכן פתרונות הגנה סטטיים לא יספיקו שכן התקוף יוכל למצוא את המסלול (הרצת התקפה) שאותו כל' הגנה לא הצליח למצואו (אם קיים אחד).

:MITRE ATT&CK

ה-ATT&CK הינו שם למודל שארגנו MITRE יצר לצורך פשוטה של סוגי שונים של התקפות הסייבר. הפשטה נועדה כדי לעזר לקהילה הסייבר לסוג התקפות שהם רואים בשטח בהתאם למסגרת המוגדרת מראש ידי ATT&CK. המודל מציג גם פתרונות עבור סוגי התקפה שונים וכן יש ערך בניסיון מיפוי התקפות לפי מודל

זה. בנוסף אנשים שירצו לנסוטה לסמץ התקפות בעצם יכולים לעשות זאת בקלות רבה יותר שכן המודל מכיר תיאור של מגוון סוגי ההתקפות ומה הן מנסות להשיג. תהילך ההפשטה של המודל מכיל שלושה רמות של אבסטרקציה:

- טקטיקה - שם כללי להתקפה. השם מעיד על מטרת התקיפה.
- טכנית - תיאור טכני על התקופה. התיאור נשאר ברמת אבסטרקציה גבוהה.
- תהילך - דרך המימוש של טכנית מסוימת. כאן משמש נכתב קוד כדי לנסוט את הלוגיקה המוצגת בטכנית.

שלושת מונחים אלו מהווים את עמודי התוויר עליהם בנוי המודל של MITRE.

המגן:

מטרת המגן הינה למנוע מהתוקף להגיע לשלב האחרון במסלול התקיפה. כדי לבצע זאת, על המגן להבין איפוא התקוף נמצא כת ולחזות את הצעד הבא שלו. כדי לבצע זאת על המגן להתמודד עם אתגר מורכב הטומן בו חומר ודאות הנובע מהשיקולים הבאים:

- התקוף יכול לנבוע במסלולים שונים על גוף התקיפה. למשל, לאחר ביצוע התקפה chosen-user Execution על גוף התקיפה הוא יכול לנסוט שתי פוטות אפשריות של אחת שייכת לשלבים השונים של גוף התקיפה למשל: Modify Registry או Modify System Process. זאת יחד עם העבודה שהמגן לא מכיר את גוף התקיפה אותו מיישם התקוף מ קישים על המגן לחזות את הצעד הבא (טכנית) של התקוף בצורה טובה.
- המגן לא יודע איזה תהילך (procedure) ביצע התקוף כדי לבצע טכנית (התקפה) מסויימת בגין התקיפה.
- המידע על מצב המערכת העומד לרשות המגן עשוי להיות לא מלא. המגן נדרש על כל גילי והתרעה שמרתם לספק לו מידע על התרחשויות חריגה במערכת. אותם כלים יתאפשר לספק מידע על התקפה שבוצעה בשל ידע מוגבל על המערכת עליון הן רצות או בשל העבודה שאין ברשותן מיפוי מפעילות חריגה להתקפה (טכנית) ידועה.

המשךה של המגן נוסחה במאמר קבועית (Sequential Decision Process) SDP. בבעיה זו, הפעולה הבאה של התקוף תליה רק במצב (S) האחרון והפעולה (A) האחרונה אותה ביצע. כתע נגיד את המאפיינים העיקריים של SDP.

- מרחב המצבים (S): מכל 17 מצבים שונים שככל אחד מהם הוא וקטור המציג מיקום ייחודי של התקוף. המיקום מיוצג על ידי:
 1. **טכנית** (סוג התקפה) בה השתמשו.
 2. **המצב ההתחלתי** ממנו התקיף התוקף - Attack Initiated.
 3. האם התקוף עבר למצב הסופי של התקפה זו - Attack Terminated.

- מרחב הפעולות (A): למגן שלוש פעולות אפשריות אותן יוכל לבצע:
 1. **Inactive**: המגן לא עושה כלום
 2. **Reactive**: המגן מסיר את כל התהליכיים במערכת הקשורים להתקפה האחרונה אותה ביצע התקוף.
 3. **Proactive**: המגן חוסם סט ספציפי של קריאות API של המערכת כדי למנוע התקפות מוקדמות של התקוף. דוגמא: חסימה של קריאת API לשינוי מפתחות Registry.

- פרו (Reward) - כתבי המאמר משתמשים בפונקציית הפרס הבאה:

$$R = -p_g(s) \times I_g - \mathcal{I}_v \times I_g - C_f$$

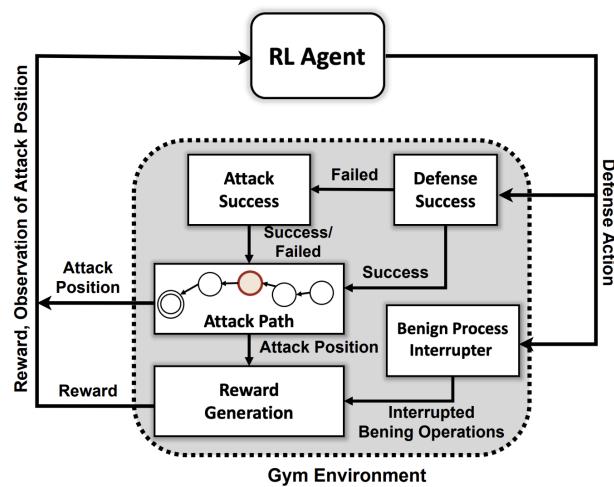
כasher:

- L_v שווה ל-1- אם המגן ניצח, אחרת 0.
- P_g - ההסתברות של התקוף להגיע לשלב האחרון במסלול / גראף התקפה.
- I_g - ההשפעה (=נזק) של התקפה מוצלחת על המערכת.
- C_f - המחיר של ביצוע פעולות הגנה. המחיר תלוי במידת ההפרעה של פעולה המגן על פעילותה התקינה של המערכת.
- שילוב (P_g ו- I_g) מכמתת הסיכון (Risk) במצב s .
- שילוב L_v ו- I_g מכמתת את התמരיך של המגן לניצח את התקוף שכן אם הוא ינצח (L_v שווה ל-1-) הפרס יגדל.

נשים לב שבפעולות 2 ו-3 המגן עלול להפריע/לגרום נזק לפעולות התקינה של הרשות / מערכת שעלייה הסוכן מגן.

הסבירה (Framework):

בתוך המסגרת אותה מציעים חוקרי המאמר מתבצע התהליך הבא:



בכל נקודת זמן, המגן (סוקן DRL) מבצע צעד הגנה ומתקבל כמשמעות על הפעולה את מיקום התקוף על גבי מסלול התקפה ופרס המתבסס על מיקום התקוף ועל ההשפעה על הסביבה.

כדי לאתר את מיקום התקוף, קיימות מערכות גילוי והתרעה אשר בוחנות קריאות API אשר מתבצעות במערכת ומתריאות על קריאות חריגות. בנוסף ברקע רצים תהליכי נספחים שגרתיים שבמקרה שהמגן משਬש את

פעילותם הסדרה עלולים לגרום לתקלות שונות במערכת. כפי שצוין בחלק הקודם, לא תמיד קיימ מיפוי בין התראה לבין טכניקה (סוג ההתקפה) וגם אם קיים הוא לא בהכרח נכון. לכן כדי למדוד את תופעה זו כותבי המאמר מניחים שקיים מיפוי בין התראה לבין טכניקה אך דיקוקו אינו מושלם ויכול לעמוד על 75%, 85% ו-65%. (אלו הדיקקים שנבדק על ידי הכותבים) ככלומר כאשר התראה מפותת לטכניקה מסוימת קיים סיכוי של 65% לפחות שמייפוי זה הוא נכון.

כעת נתאר את החלקים הנוספים בסביבה שבה המגן רץ:

- מגנון-h-Defense Success הינו חלק בסביבה האחראי לקבל את פועלות המגן ולהחליט האם היא עכירה את פועלות התקוף (ולכן הוא הפסיד).
- מגנון-h-Attack Success אחראי לקבוע האם פועלות התקוף הצלחה והתקפה בוצעה בהצלחה. אם אכן כך, התקוף מקבל אינדיקציה שהוא יכול לעבור לשלב הבא, אחרת הוא נשאר במקום או מפסיד.
- מגנון-h-Interrupter Benign Process יודיע לעקב אחריו אילו תהליכיים תמיימים הופסקו כתוצאה מפעולות המגן. הוא שולח את הרשימה הזאת למגן-h-Reward Generation בשביל חישוב הפרס.

ניסויים:

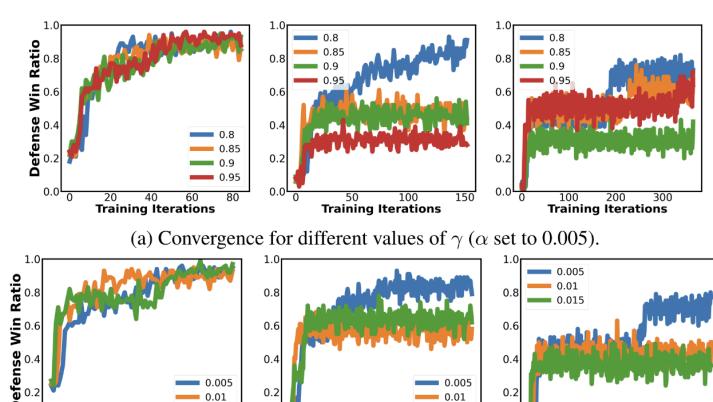
בסביבת הניסוי שלהם, החוקרים לא השתמשו במלוא הטקטיקות והטכניקות המופיעות ב-MITRE אלא ב-7 מתוך 11 הטקטיקות. כתוצאה לכך, מספר הטכניקות נמוך אף הוא מהמקור. מספר פעולות המגן שימושו הין 23 כאשר 21 מתוכן הן מסוג Proactive (לא מוזכר מה הן אותן פעולות).

בכדי לאמן את המגן, החוקרים ייצרли קודם את כל מסלולי ההתקפה האפשריים בגרף ההתקפה מנוקדת ההתחלת לנוקדות הסיום. מתוכם 80% מסך המסלולים שימושו לאימון ו-20% לטשטט. בזמן האימון כל איטרציה כוללת ניסיון של התקוף לנوع על מסלול אחד מטור גրף התקיפה מההתחלת עד הסוף. בכך נתונים למגן לראות המונ סימולציות שונות ותנאי סביבה שונים.

החוקרים מציגים את הטענה שמידול התקוף מתבצע בשני מישורים: כישרון ועקבנות. כאשר כישורי התקוף גבוהים יותר כך גדל הסיכוי שההתקפה שהוא בוחר לבצע בשלב הבא תצליח. התקוף עקשן הוא התקוף שלא יותר וימשיך לנסוטות לתקוף גם אם לא צלח בנסיון הראשון. פירוש אחר הוא שתוקף עיקש הוא אחד כזה שלא ניתן להזחות בקלות גם אם התקפה מסוימת שלו כשלה. במקרה זה עקבנות הינה מספר הפעמים שהתקוף יבצע התקפה שנכשלה לפני שיפסיד (כלומר יתגלה על ידי המגן). החוקרים ניסו רמות שונות של כישרון ועקבנות כדי לאמן את המגן תחת דרגות שונות של קושי. בטור המגן ניסו אלגוריתמי DRL שונים, ביניהם: [A2C](#) ו- [DQN](#).

תוצאות:

בתרשימים הבאים ניתן לראות את ביצועי כל אחד מהמגנים נגד שלושה סוגי של פרופילים של התקוף כאשר ההבדל בין כל פרופיל הינה רמת הכישרון והעקינות של התקוף. המدد בו השתמשו להשוואה הינו אחוז הפעם שהמגן ניצח את התקוף מסה"כ המשחקים ששוחקו.



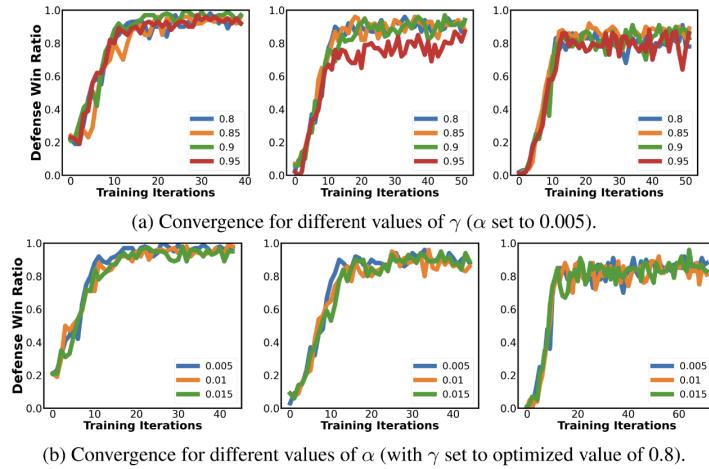


Figure 4: Sensitivity of DQN for different values of γ and α against attack profiles Av_1 , Av_2 , and Av_3 (left to right)

מהתרשים ניתן לראות ש-DQN הציג את הביצועים הטובים ביותר נגד סוג התוקף השונים. הוא לא רק הצליח לניצח ביותר משחקים מאשר A2C אלא הוא הצליח ליצור את התוקף בשלב מוקדם יותר כפי שניתן לראות לפי מספר האיטרציות בציר ה-X, שם מספר האיטרציות נמוך יותר בתרשימים של DQN.

קיים פוטנציאל גדול לשוגר צהה של מגן וסבביה בעולם האמיתי. מגן אשר ראה מספר רב ומגוון של התקפות ושל פרופילים שונים של התקפות. ביצועי ה-DRL מראים כיון מחקר עם יישומים אמיתיים ועזרה אמיתיית לארגוני להתמודד עם האיוםים השונים המהווים להם.

סיכום:

מאמר זה מציג את ההתקנות של פריסת סוכנים אוטומטיים לצורך הגנה מפני התקפות סייבר. כתבי המאמר הראו אשר ניתן לאמן סוכנים כאלה באמצעות אלגוריתם DRL אשר יכול להגן בהצלחה מפני התקפות סייבר שונות. בנוסף הסוכן מסוגל להתמודד עם סוגי תוקפים שונים של תוקפים מתחכמים.

שיטת פעולה: הסקירה נכתבת בשיטת פעולה עם עדן יבין

Review 72, Short: Unifying Large Language Models and Knowledge Graphs: A Roadmap

מה יקרה אם נשלב מודלי שפה גדולים (LLMs), כגון ChatGPT4 ו-GPT4 עם גרפי ידע (Knowledge Graphs)?

הריGs KGs ו- LLMs משלימים אחד לשני באופן מאוד טבעי.Gs KGs יכולים לעזור לטפל בבעיית hallucination של LLMs ו- גם לחזק את "התשתיות העובדתית" שלהם. LLMs יכולים לשפר את יכולת של KGs ל-reasoning. המאמר מסכם את החוקרים האחרנים בנושא המעניין זה:

- (1) איך לשדרג LLMs עם KGs?
- (2) איך לשדרג KGs עם LLMs?
- (3) דרכי לבנות מודלים המשלבים LLM ו-KG?

המחברים גם דנים בכךוני מחקר עתידיים ומשרטטים "מפתח דרכי" בעברם.

Paper: <https://arxiv.org/abs/2306.08302>

Github: <https://github.com/RManLuo/Awesome-LLM-KG>

Review 73, Short: Diffusion Models for Zero-Shot Open-Vocabulary Segmentation

המאמר מציע שיטת סגמנטציה zero-shot בעלת מיליון פתוח, כלומר אין לנו מכול סגור קטגוריות אלא כל תמונה מקבלת סגמנטציה עם הקטגוריות שלה. כਮון שהשיטה לא דורשת שום אימון מקדים (zero-shot). זה מתבצע באמצעות מודל Text2Image דיפוזיה (DDPM) גנרטיבי.

השיטה המוצעת מכילה שני שלבים:

1. בשלב הראשון (אימון מקדים) המחברים מגנרטים מספר רב של תמונות לקטgorיה (פרומפט) מסוימת (כמו תמונה טובעה של חתול או כלב) בעזרת מודל דיפוזיה מאומן. לאחר מכן הם בונים "מסד טיפוסים של פיצרים" של כל קטgorיה שכולל הסגמנטציה של האובייקטים (וגם רקע) באמצעות מודל סגמנטציה מאומן.
2. בשלב השני מפיקים את הפיצרים של תמונה באמצעות CLIP ומשווים אותה עם אב הטיפוסים של הפיצרים שישי במאד הנתונים. הקטgorיה והסגמנטציה הדומה ביותר ממסד הפיצרים נבחרת.

HuggingFace: <https://huggingface.co/papers/2306.09316>

Arxiv: <https://arxiv.org/pdf/2306.09316.pdf>

Review 74, Short: Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture

המאמר של יאן לסקון (Yann LeCun) מציג גישה חדשה ומעניינת or-or JEPA-LI לבניית ייצוג (embedding) של>Data ויזואלי (תמונות) לא מותג כלומר (self-supervised). SSL. שיטות SSL הקדומות אימנו מודל האוכף דמיון גבוה בין ייצוגי אותה פיסת>Data אחריו אוגמנטציות שונות, השומרות על התוכן.

במקום זאת המחברים מאמנים מודל המנסה לחזות ייצוג של חלק אחד של הדטה מהחלק الآخر. למשל בתמונות המודל חוזה ייצוג של פאץ' מייצוג של פאץ' אחר. השיטה הצליחה להפיק ייצוגים בעלי ביצועים חזקים מאוד על מגוון משימות כSiege, ספירת אובייקטים ובניית מפות עמוק.

אחד היתרונות המשמעותיים של המאמר הוא מיתר במידה רבה את הצורך לבנות אוגומנטציות השומרות על התוכן הסמנטי של התמונות בצורה ידנית.

HuggingFace: <https://huggingface.co/papers/2306.09316>

Arxiv: <https://arxiv.org/pdf/2306.09316.pdf>

Review 75, Short: Recurrent Memory Decision Transformer

המאמר מציע לשלב את המנגנון של הזכרון recurrent בטרנספורמרים לפתור בעיות של במידה באמצעות חיזוקים (RL - reinforcement learning). המאמר מנסה לתת מענה לאחת הסוגיות המרכזיות העולות כאשר משתמשים בטרנספורמרים לעיבוד של דטה בעיות RL: אי יכולות שלהם להחזיק רצפים ארוכים בזיכרון בעקבות הסיבוכיות הריבועית של מנגנון תשומת הלב שלהם.

המאמר מציע שיטה לעדכן של הזכרון כתלות באופיינים השונים כמו המצב המערכת, אובייקט, אובייקט, פעולה והתגמול. כך מאפשרת שימושה בזיכרון המודול רצפי דטה יותר ארוכים שמשפר את התורם חיובית ליכולת הלמידה של מודל. השיטה המוצעת הציגה ביצעים במשחקים כמו Atari ו-MoJoCo.

OpenReview: <https://openreview.net/forum?id=Uynr3iPhksa>

Arxiv: <https://arxiv.org/abs/2306.09459>

Review 76, Short: Gradient is All You Need?

המאמר מנתח שיטות אופטימיזציה מסווג של CBO (consensus-based) ומשווה אותן עם השיטות מבוסחת GD שננו מכירים היטב ומשתמשים בהם רבות. CBO היא משפחת שיטות שמריצות מספר "סוכנים" (לפעמים בלתי תliusים ולפעמים לא) הבוחנים את מרחב האופטימיזציה של הבעיה. המילה consensus מופיע בשם השיטה כי עם מספר רב של סוכנים חושבים שאיזור למרחב האופטימיזציה הוא "טוב" (הfonctionelle מקבלת בו ערכים נמוכים אם מדובר בבעיית המזערו) אז נראה כדי למקד את המיפוי באיזור זהה.

המאמר מראה שלקטוגוריות רבות של פונקציות לא קמורות ההתנגדות של CBO היא די דומה לשיטות GD למחרות שיטות CBO לא משתמשות בגרדיינטים כלל? ולמה זה חשוב? כי מהשקלות הזו ניתן להטיק שבתנאים מסוימים GD מתכנסות למינימום עבור קטגוריות רחבות של פונקציות לא קמורות כי CBO מתכנס אליו.

Arxiv: <https://arxiv.org/abs/2306.09778>

Review 77, Short: Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision

המאמר מציע מודל דיפוזיה גנרטיבי (DDPM) למקורה שבו הדטה לא אימן לא זמן בצורה ישירה אלא רק אחרי טרנספורמציה שאחריה חלק מהמידע על הדטה הולך לאיבוד. בדרך כלל אנחנו מאמנים מודל דיפוזיה על דאטאסת המכיל, נגיד, תמונות בעלי תיאור מילולי, והמודל לומד ליצור תמונה מתיאור כאשר יש לנו תמונה אמיתית עם התיאור זהה - אז די ברור איך לאמן את המודל.

המאמר מטפל במקרה מורכב יותר מאשר לראותו למשל בתחום inverse graphics. כאן המודל מתבקש ליצור דגימות של סצנת 3D כאשר יש לנו ביד רק תמונה אחת או כמה תמונות אך אין לנו את מודל ה-3D של הסצנה עצמה. המחברים מציעים לאמן מודל דיפוזיה המשלב יצירה של מודל הסצנה יחד עם הדגימה מהסצנה הזאת. המודל מאומן ליצור דגימות שמצד אחד מתאימות למודל הסצנה הבננה ומצד שני אוקף את מודל הסצנה להיות תואם לתמונה הנתונה.

Arxiv: <https://arxiv.org/abs/2306.09778>

Review 78, Short: Fast Segment Anything

המודל SAM - Segment Anything שפורסם לאחרונה הפרק להיות מודל בסיס למשימות הראייה הממוחשבת כמו סגמנטציה, גנטוט כוורת של תמונה וגם לערכאה של תמונה. אולם שימוש במודל זה מצריך כוח חישוב משמעותי הנובע מהחישובים של מודל הטרנספורמר עבור תמונות ברזולוציה גבוהה.

המאמר מנוסח את בעיית הסגמנטציה בתור שילוב של שתי בעיות: גנטוט סגמנטים ו-prompting (יצירת כוורת לסגמנט). רפורמליזציה זו מאפשרת לנצל מודל ל-instance segmentation מבוסס רשותות קונבולוציה שמאפשר לחסוך חלק גדול מהחישובים.

Arxiv: <https://arxiv.org/abs/2306.09778>

Review 79, Short: SqueezeLLM: Dense-and-Sparse Quantization

מה הדרך הגיונית לשומר ולהריץ מודלי שפה במחשבים אישיים לא חזקים במיוחד? כמובן לשומר את המשקלים שלא בא בדיק המלא (32 ביט) אלא בדיק חלקית (נגיד 4 ביט). אך איך לעשות זאת כדי לא לספק ירידה קשה בביטויים?

הקוונטיזציה היוניפורמת לא תספק לנו כאן ביצועים טובים כי המשקלים לא מפולגים יוניפורמת. המאמר מציע לקלستر משקלים עם means-k כאשר המרחק מbasos על החישבות של המשקלים למודל (כמה הם משפיעים על הלוא). זה מאפשר לשומר משקלות שחשובים יותר בדיק גבוהה כאשר הפחות חשובים נשמרים בדיק נמוך.

המחברים גם שמו לב כי ה-outlier משפיעים לרעה על איקות הקלסטרים (מורחים אותם) והחליטו לשומר אותם בклסטרים נפרדים מהשאר

<https://arxiv.org/abs/2306.07629>

Review 80, Short: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis

זהו השדרוג של מודל SD2.0 המגנרט תמונות באיכות מטפסת. המאמר הציע כמה שיפורים לארQUITטורה של SD המקורי, שככל את האימון והשתמש ב-SDEdit לשיפור איכות התמונות

המודל מבוסס על אותן הנקודות כמו SD המקורי אך מוסיף refiner באמצעות המבוסס על SDEdit (דגם מהמודל דיפוזיה דרך פתרון נומי של מישדייף סטוכסטי).

כולם תהליך הדגימה נראה כר:

1. שיבוץ (embedding) הטקסט למרחב הlatent עם ה-encoder ודוגימה מרעיש גאות'
2. יצרה של שיבוץ התמונה עם מודלי דיפוזיה latent LDM (כמו ב-SD)
3. SDEdit לשיפור וקטור latent מסעיף 2
4. ייצרת תמונה מהשיבוץ של באמצעות ה-decoder

כאמור יש שינויים בארQUITטורה (עדין מבוססת על UNet ומנגנון attention). הם גם "העשירו" את תהליך האימון:

1. הכניסו תמונות בגודלים שונים אך העבירו לרשף את הגודל שלהם
2. עשו crop לתמונות והזינו לרשף את גודל הקיצוץ.

<https://arxiv.org/abs/2307.01952>

Review 81, Short: Segment Anything Meets Point Tracking

המאמר מציע שיטת SAM-PT לביוץ סגמנטציה ומעקב של אובייקטים בזידאו (ViDAO). המודל מנוף את מודל הסגמנטציה SAM עצמה שעלה לאוויר לא זמן.

השיטה מכילה 4 שלבים עיקריים:

1. מגדירים את האובייקט שאנו עוקבים אחריו בפריים הראשון באמצעות כמה נקודות (או שמגדירים אובייקטים ורקע לא קשורים לאובייקט המעקב
2. משתמש בשיטות מאומנות של מעקב אחר הנקודות בזידאו כמו PIPS כדי לעקוב אחר הנקודות המסומנות.
3. משתמשים ב-SAM בשביל לבנות (segmantation) מנקודות אלו את האובייקטים המתאימים לכל פריים של זידאו. בשלב הראשון מכינים ל-SAM רק את הנקודות החיוביות (השייכות לאובייקט) ובשלב השני מכינים גם את הנקודות השליליות (כמו רקע) יחד עם התוצאות הסגמנטציה הראשונה כדי לקבל את תוצר הסגמנטציה הסופי.

רובה יקר, תעשה לי קפה ותשים על השולחן!
רוצים שיהיה לכם רובוט משלכם שתוכל לחת למשימות בית?

Review 82, Short: SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning

המאמר ממנף את כוח של #[lms](#) כדי לעזור לרובוט לתוכנן שימושות ב-door.

איך עושים זאת? בונים את גוף הייצוג התלת מימד (3D) של המיקום (scene) הנקרא 3DSG . בגדול 3DSG מכיל את כל האובייקטים שיש בסצנה, את המיקומים היחסיים שלהם. אחר כך משתמשים ב-#[lm](#) (מצינים בה פרומפט והוראה עצמה) בשביל למצוא את הקודקודים הרלוונטיים לשימוש ולתכנן תכנית ביצוע המשימה. אם לא מצליחים לבצע את המשימה לוחכים את המשוב ומעבירים אותו שוב ל-#[lm](#) עם הפרומפט ואת ההוראה כדי לבנות תכנית פעולה חדשה. מסמלצים את תכנית, יוצרים את המשוב עד המצליחים לסייע את המשימה. די מגניב האמת.

Review 83, Short: HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models

זה למעשה שילוב של LoRA (Low Rank Adaptation) ו-DreamBooth

מה זה DreamBooth? בינהנתן תמונה או מספר תמונות (לא הרבה, נגיד 4-3) הוא מכיל מודל SD כך שהוא יודע ליצור תמונות של הכלב הספציפי שלך, נגיד, בכל מיני צנונות (המתוירות עם תיאור טקסטואלי). הcoil מתבצע על ידי מתן שם (מורכב מטוקנים נדרירים) לאובייקט ואז מכילים SD עם תמונות האובייקט תוך שמירה של יכולתו "בהתבונת העולם היזואלי".

מה זה LoRA? נניח אתם רוצים לכיל מודל שפה ענק של 50B פרמטרים אבל אין ברשותכם כוח חישוב אינסופי.
از אתה עשה שני דברים:
1. לכל מטריצת המשקלות אתה לומד את ההפרש (delta) מהיעד (המשקלים אחרי הcoil) ולא משקל היעד הסופי.
2. אתה מניח שמטריצת ההפרש הזו כי rank-low ונינתן לתאר אותה כמכפלה של 2 מטריצות לא מאוד גדולות שאתה לומד אותן. ככה חוסכים בכמות המשקלים הנלמדים.

از מה זה HyperDreamBooth זה פשוט LoRA ו-DreamBooth המשולבים בצורה חכמה

Review 84, Short: Learning to Retrieve In-Context Examples for Large Language Models

הידעתם ש-[s-lms](#) יודעים להתאים את עצמם למשימות חדשות ללא שום כויל (=שינוי המשקלים) אחרי כמה דוגמאות? לפעמים זה עובד ללא דוגמאות. זה נקרא למידה in-context או ICL

מה קורה עם משלבים ICL עם אוגמנטציהacha: קרי נתונים ל-LLM להשתמש בדата חיצוני (RAG)?

Learning to Retrieve In-Context Examples for Large Language Models

המודל מאמן בכמה שלבים:

1. לשאלתה נתונה מאחרים כמה דוגמאות מהדата החיצוני, זוגות של (שאלה, תשובה) עם אלגוריתם BM25 מרכיבים ולו כדי לדרג את הרלוונטיות של לשאלתה. בגודל כל הזוג "móvel" אותה לתשובה מכון הוא מדורג גבוה יותר.
2. מאמנים מודל תגמול (reward) עם למידה ניגודית (דוגמאות מדורגות גבוהה מול אלו שמדורגים נמוך). מאמנים אנדוקר לשכנ (embed) יחד את הזוג-hat ground truth של השאלתה (y, x) יחד עם הזוג המאוחר (i_y, i_x). מודל זה משמש רק בתור מודל מורה לשלב הבא כי משתמש בתשובה y שלא ידועה כמובן בזמן הטעtinyning.
3. אימון מאוחר ולו עם RAG: מאמנים מודל שיכון עבור זוג (i_y, i_x) ממגר חיצוני ועבור דוגמא x . בוחרים זוגות עם דמיוןści בין*ci* גובה בין ייצוגו לבין הייצוג של שאלתה x (מרחק cosine). מאמנים את הייצוגים אליו כדי עם שילוב של 2 פונקציות יעד:
 - a. קרבה בין התפלגיות של מודל המורה מהשלב הקודם לבין המודל המאוחר את הזוג הקרוב ביותר (עם הייצוגים שחושבו לפני)
 - b. מיקסום מרחק בין ייצוגים של הדוגמאות המדורגות גבוהה לבין אלו שמדורגות נמוך (לאו ניגודי)מעניין שעושים כמה איטרציות של אחזור כאשר משתמשים בזוג שנמצא באיטרציה הקודמת לאיטרציה הבא המשפר את איכות האחזור.

Review 85, Short: Anticorrelated Noise Injection for Improved Generalization

עבודה מענית המציע שיטה מאוד פשוטה לשיפור ביצועים של SGD stochastic gradient descent (SGD) הוספת רעש אקריא לעדכוני משקלים (GD perturbed) המבוצעים במהלך SGD נקרה בכמה עבודות בשנים האחרונות. אחת המסקנות של עבודות אלו היא שהזרקה של רעש (בדרכן כלל גausi) בעלי שונות נמוכה יחסית ל- SGD יכולה לשפר את יכולת הכללה של הרשת המאומנת.

אבל למה זה עוזר בעצם? מוקובל להסביר את ההשפעה החזיבית של הזרקת רעש ל-SGD בכך שהוא עוזר לפונקציית לאו להתכנס לנקודות מינימום "רחבה" כלומר צו שברוב הנקודות בסביבה הקרובה אליה ערך של פונקציית לאו נותר נמוך. מינימום רחב של פונקציית לאו נחשב לטוב ליכולת הכללה של המודל לעומת מינימום חד כלומר צזה שapeutic בסבירתו הקרובה יש עלייה ניכרת בערכיו פונקציית לאו. הסיבה לכך היא טמונה בהנחה (יש גם תוצאות תיאורטיות חלקיות המוכיחות זאת עבור מודלים פשוטים יחסית) שנקודות מינימום רחבה "mbטאת את הסיגナル האמתי" מהדטה ולא נוצרת כתוצאה של הרעש שנמצא בדטה.

אוקי, אז מה מציע המאמר? בדרך כלל מוסף רעש גausi בלתי תלוי לעדכוני משקלים של SGD והמאמר שואל שאלה לגיטימית: האם זה אופטימלי לביצועים. מתרבר שלא כל ...

המאמר מציע להוסיף רעש בעל קורלציה שווה ל-1. איך עושים זאת? דוגמנים סדרה ארוכה של **וקטורים** גאומטריים \mathbf{x} ויצרים סדרת הפרשים בין איבר סדרה הסמוכים ($\mathbf{x}_n - \mathbf{x}_{n+1}, \mathbf{x}_{n+1} - \mathbf{x}_{n+2}, \dots, \mathbf{x}_1 - \mathbf{x}_2$) ומושיפים אותה לעדכוני משקלים של SGD באיטרציה n . מתרבר שזה עוזר ללא מקרים ליכולת ההכללה של הרשת. הם גם מראים שהערכיהם העצמיים של ההסיאן בנקודת מינימום עם השיטה שלחן (השיטה נקראת Anti-PGD) יותר נמוכים מאשר ל-PGD עם רעש חסר קורלציה.

המאמר גם דן (לא התעמקתי) בקשר בין הטכניקה המוצעת להוספה של רעש אקראי ללייבלים וגם לשיטות החלקה שונות (smoothing).

המאמר כתוב היטב ודי קל להבנה.

מאמר: <https://arxiv.org/abs/2202.02831>

Review 86, Short: BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs

רוצים מודל העונה על השאלות לגבי האובייקטים בתמונה וגם מראה לכם איפה האובייקט נמצא בתמונה ולתת הסבר לגבי? בנוסף נתונים אותו המודל יותר לסמן אוטם מי בתמונה שלכם הוא מקור האudio. וכל זה מלאה בהסבירים

היום ב-#shorthebrewpapereviews:
BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs

המאמר משלב כמה מודלים חזקים שאומנו למשימות הקשורות ל-grounding ויזואלי כמו Model Anything שmaps את הקטגוריות של האובייקטים. לאחר מכן מזינים הקטגוריות אלו למודל DINO-Ground שמצהה את מיקום האובייקטים. בסוף מכניסים את התוצאה ל-SAM ועשימים סגמנטציה עדינה של האובייקטים

בשלב האחרון צריך למצוא את הזוג (קטgorיה, מסכה) המתאים לשאלתה שלכם. בשביל כך מעבירים את השאלה דרך GPT4 ומחפשים את הקטgorיה הקרובה ביותר מהרשימה שהוצאה על ידי SAM. אוקי סימנו עם grounding ויזואלי אבל הבוחני لكم 3 מודים - אז איפה נעלם האudio?

קודם כל "מיישרים" את הייצוג audio ותמונה על ידי אימון שני אנקודרים (לכל אחד) על דאטasset של מכיל זוגות של AUDIO ותמונה מתאימה. גם כאן מכילים מודלים קיימים: BLIP2 לתמונה ו-ImageBind לאudio. בסוף מכילים את כל המפלצת הזו על שלוחים של AUDIO-טקסט, תמונה-טקסט, ותמונה-AUDIO.

<https://huggingface.co/papers/2307.08581>

Review 87, Short: TokenFlow: Consistent Diffusion Features for Consistent Video Editing

סוקרים המאמר **חול-לבן** היום קצROT ב #shorthebrewpapereviews

חומר רציפות בין הפריים: מה הבעה הגדולה ביותר בעריכה של וידאו עם באמצעות מודלי דיפוזיה גנרטיביים? מודלי דיפוזיה מסודרים ידי יפה עם עERICA של תMONOTONES לפי תואר טקסטואלי אבל עם הVIDAO הטיפור הוא יותר מסובך כי נדרש רציפות בין הפריים.

הדרך הנאיית לבצע עריכת וידאו בהתאם לטיור טקסטואלי היא לעורך כל פריים (תמונה). אבל איך נשמר על קוהרנטיות בין הפריים הערכוכם? המחברים לוחחים פיצרים של הפריים הסמכים ולהשתמש בהם ולהחליק את הVIDAO עורך בעזרת אינטראפלציה של הפיצרים של הפריים הקרובים לו.

אבל מה הם הפיצרים של הפריים שcadai ללקחת? קודם כל המחברים לוחחים את את השאלות, מפתחות וערכים (queries, keys, values) ממנגנון-h-attention מכמה פריים ערכוכם **סמכים** של הVIDAO המקורי. לאחר מכן עבר פריים או מפעלים מנגן-attention על השאלתה שלו ועל המפתחות והערכים של הפריים.

ככה למעשה מחושב "ציג הרציפות" של וידאו (הモרכיב מייצג של כל פריים ביחס לפריים האחרים). לכל פריים מחפשים את הפריים הבא לפני זהה שבא אחריו עם ה-h-attention הקרוב ביותר לפי מרחק הקוינו מבchnitot "ציג הרציפות. ואז עבר כל פריים שאנו ערכוכם משפרים את רציפות בין הפריים תוך שימוש באותו "ציג רציפות" כמו בVIDAO המקורי על ידי אינטראפלציה שלו על ידי שני ייצוגי הרציפות של הפריים שמצאו.

<https://arxiv.org/abs/2307.10373>

Review 88: Secure Machine Learning in the Cloud Using One Way Scrambling by Deconvolution

פינט הסוקר:

המלצת קריאה מעדן ומיק: מומלץ לאנשים העוסקים בתחום ה-Cybersecurity או/ו לאנשים שמתעניינים בטכניקות לשיטות תMOVONES (בלי לראות אותן!!)

בהירות כתיבה: טוביה

ידע מוקדם:

- הצפנה
- Deconvolution

ישומים פרקטיים אפשריים: מערכת להצפנה מידע

פרטי מאמר:

[לינק למאמר: זמן להורדה](#)

לינק לקובץ: לא אוטר

פורסם בתאריך: 21.11.2021, בארכ'יב.

הציג בכנס: - KDD

תחומי מאמר:

- פרטיות המידע
- מודלים בענן

כליים מתמטיים, מושגים וסימונים:

- אנקודר(להפקת ייצוג לטנטי של תמונה) ודקודר

מבוא:

הרצת מודלים גדולים של למידה عمוקה איננה משימה קלה. נדרש מאמץ רב ויכולות מחשב כבירות (בד"כ GPU-ים או TPU-ים) סביר אימון המודול והפעלתו (Inference) שלו. חסם זה מנע מחקנים רבים את היכולת להשתמש במודלים ענקיים אלו. אך כיום משתנה עקב שימוש הולך וגדל בשירותי למידת מכונה בענן בו חברות פורסמות מודלים גדולים ומאפשרים גישה לכלל הציבור דרך קרייאות API שנעשות על גבי האינטרנט. גישה זו אمنה הנגישה מודלים רבים ואפשרה למספר רב של משתמשים להריץ אותם שלא היה ניתן קודם גם באמצעות הbinary איתה מספר סוגיות רציניות. בעיה אחת מרכזית הינה פרטיות המידע: בשביל להעברה למודל מידע הנחוץ להרצתו יש צורך לשנות אותו על גבי האינטרנט - - דבר שעלול לגרום לצלגית מידע וריש. הבעיה השנייה הינה יכולת של גורמים זרים להסיק מסקנות לגבי המידע של הארגון שלא בצוරה עקיפה. ניתן לעשות זאת על ידי ניתוח הפלט של המודל בענן וזאת בשל העובדה שלכלום יש גישה לאותו מודול.

כדי להתגבר על הביעות המוזכרות לארגוני היום קיימות שתי אופציות:

1. להצפין את המידע לפני שליחתו, פענוו בצד השני, הצפנה התוצאה ושליחתו חוזרת לפענוו אצל הלוקו.
2. שימוש בהצפנה הומומורפית ([Homomorphic Encryption - HE](#)). הצפנה זו מאפשרת לצד מקבל את המידע המוצפן לא לפענוו אותו, אלא לנתח אותו ולהגיע למסקנות דומות באותו צורה שבה היה מגיע אילולא המידע לא היה מוצפן מלכתחילה.

אך לפתרונות אלו יש כמה בעיות. הבעיה עם הפתרון הראשון הינו שלאחר פענוח המידע בצד של ספק הענן, לאוטו ספק תהיה גישה למידע הלא מוצפן. למשל, יש פה מקרה של הפרת פרטיות המידע. הבעיה עם הפתרון השני הינו שהוא איטי שכן עיבוד המידע המוצפן קשה יותר.

כדי להתגבר על הבעיה של פתרונות אלה במת אחת, כתבי המאמר מציעים שיטה הנקראת *via* Encoding (Private-Key Deconvolution) (EPKD).

השיטה:

השיטה המוצעת דומה ל-HE בכך שהיא מאפשרת למודל בענן לקבל את המידע מוצפן ולעביד אותו כאילו לא היה מוצפן אך עם מהירות דומה לעיבוד המידע המקורי. לעומת היחסון של מהירות העיבוד שמצויג HE לא משחיקת פה תפקיך. השיטה משתמשת בשלושה רכיבים:

1. Encoder(מקודד): נועד כדי ליצור ייצוג דחוס של המידע. בעבודה השתמשו החוקרים בכמה מודלים מסוג Resnet כגון ResNet5, ResNet101, ResNet50v2 ועוד. ככל הינו מאמנים מראש ולא כוללו בונוס.

השכבה האخונה של softmax הוסיפה כדי להשתמש בקטור הייצוג(Embedding) שהרשת יוצרת. 2. Generative Model(מודל גנרטיבי): המודל המשמש בתור שיטת ההצפנה. CAN משתמשים ברשות מסווג Deconvolution אשר המשקولات שלו מאותחלות באופן רנדומלי. המשקولات הרנדומליות משתמשות מפתחת ההצפנה. הרשות מקבל את וקטור הייצוג מהמקודד ויוצרת ממנו תמונה חדשה שלען האנושית תראה כמו רעש מוחלט. התמונה החדשה נשלחת למודל בענן שמעבד אותה כמו תמונה רגילה ושולח את וקטור התוצאות שלו צרצה.

3. INN (The Internal Inference Network) - רשת feed forward קטנה המקבל כקלט את הייצוג של התמונה המקורית יחד עם וקטור התוצאות שהמודל בענן החזיר על התמונה המוצפנת. הרשות מוציאת כפלט את התיאוג האמתי של תמונה המקור.

חשוב לציין לב שהמודל בענן אינו מקבל את התמונה אחרי תהליך הגנרטט. דבר נוסף לשים לב אליו הינו שהמודל המגנרטט אינו מאמין אלא מאותחל בסט משקولات חדש וrndomלי בכל פעם. למה כך? ובכן בתחום ההצפנה יש את מושג מפתח ההצפנה שבעזרתו מוצפינים את המידע. אם מישחו יידע את המפתח הוא יוכל להשתמש בו כדי לשחזר את המידע המוצפן. לכן, חשוב שייהו קשה להשיג את המפתח. במקרה של המודל המגנרטט, מפתח ההצפנה הינו משקولات המודל. אם מישחו ישייג את המשקولات של המודל הוא יוכל להשתמש בהן כדי להעתיק את המודול ולהשתמש בו לתהליך ההפוך - מעבר מרחב האמבעדינג חזקה לתמונה המקורית. כדי למנוע זאת, מתחילה רנדומלית את המשקولات בכל פעם כדי להקטין את הסיכוי שמייחדו ישייג בנקודה זמן מסוימת בדיקות רנדומליות שייצור תמונה מסוימת. התמונה המוצפנת שנוצרת מהמודל הגנרטיבי תראה כמו רעש מוחלט לעין האנושית כפי שניתן לראות בתמונה 1:



Figure 3: left) The original image, right) The same image, encrypted by EPKD using embeddings by ResNet50.

תהליך האימון:

הרכיב היחיד שצריך אימון הינו מודל ה-NN. כותבי המאמר טוענים שבשביל אימון המודל יש צורך בדעתהסט קטן בלבד של תמונות מתיוגות או לחילופין דאטסהט ציבורי שזמן באינטרנט. תמונות אלו נשלחות למודל בענן שמחזיר את וקטור ההסתברויות (של הקטגוריות) עבור משימת סיוג. הקידוד של התמונה יחד עם וקטור ההסתברויות (כລונר שני וקטורים) משמשים כדוגמה אחת לאימון מודל ה-NN. תהליך האימון מוצג בתמונה 1.

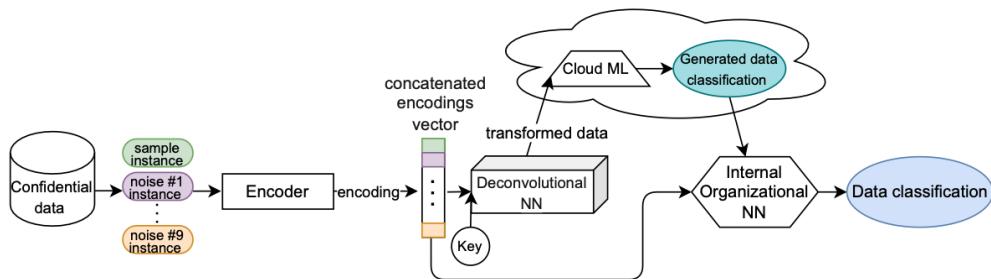


Figure 1: The EPKD architecture. Plaintext images are first transformed into an embedding, then expanded back to their original dimensions in encrypted form using a deconvolutional network. The deconvolutional network's weights are initialized randomly, and are in fact the encryption key. The encrypted images are sent to the cloud, and a classification vector is obtained. The Internal Organizational NN translates these classifications into the final (real) classification.

איך אבל מודל ה-NN יודע לקבל מידע מוקטור ההסתברויות של מודל אותו הוא לא מכיר? אוביון מדובר בשילוב של שתי שיטות משני תחומים.

הראשונה שיכת לתחום ה-[Membership Inference](#) ונקראת [Adversarial Machine Learning](#). בשיטה זו קיימים משתמש הרוצה לנסוטו לגנוב מידע על אופן אימון המודל. במקרה זה מדובר ניצול של האם דגימה מסוימת היא חלק מסוימת האימון של המודל או לא. על ידי ניתוח וקטור ההסתברויות של המודל, ניתן לראות דפוסים בהם המודל מתנהג אחרת עבור דוגמאות מסוימת האימון וככלו שלא. כדי לבצע את הניתוח מאמינים מודל שקובע עבור כל קלט (וקטור הסתברויות ותמונה) האם התמונה היא חלק מסוימת האימון. העבודה זו לוקחת השראה מתחום זה. הסיבה לכך הינה שגם בשיטה המוצעת כאן, הקלט הינו וקטור הסתברויות והדגם שאליה שייך וקטור זה. השני הוא הפלט, שכן בשיטה המקורית משתמשים במודל כדי לחזות האם דגימה שיכת לסת האימון או לא וכן מנסים לחזות את הסיוג האמיטי של הדגימה מוקטור ההסתברויות. בעבודה זו, ניתן לראות שיש חפיפה שכן מדובר באותו קלט אך פלט שונה.

השיטה השנייה היא מתחום ה-*Distillation*. השיטה מדברת על אימון מודל קטן המחקה מודל גדול יותר (*ensemble* של מודלים במקורה זה). המודל המאומן יהיה בסדרי גודל קטן יותר מהמודל אותו הוא מנסה ללמידה ובנוסף יהיה בעל ביצועים דומים. מודל ה-*NN* מנסה באמצעותו אותה שיטה ללמידה את מודל הענן. הדגש של כתבי המאמר כפי שמצויר הינו שמודל ה-*NN* הינו מודל קטן (ההנחה היא שקטן (ההנחה היא שקטן מהמודל בענן) יכול רק שכבה חבויה אחת בלבד).

כל תמונה רגישה מועברת דרך המקודד ומצוינת על ידי המודל הגנרטיבי. בנוסף כדי להקשות עוד יותר על פונCTION הידוע על ידי גורם זה, נוספו לתמונה הרגישה 9 תמונות הנבחנות רנדומלית כל פעם אשר הוצפנו גם הם. התמונות נעשו כדי להוציא רעש, כך לא ניתן יהיה לדעת מה המידע הוא רגש ומה הוא לא. בכל שליחת של מידע התמונה הרגישה והתמונות הלא רגישות עורבבו כך שמייקם התמונה הרגישה במערך שנשלח היה נבחר באופן רנדומלי כדי להקשות על זיהויו שלה. מודל ה-*NN* נמדד על יכולת לסואג נכון כל תמונה בין אם היא מצוינת ובין אם לא. המדדים אשר כותבי המאמר השתמשו בהם הינם: [Top1](#), [Top5](#) שבקיצור אומרם האם הסיווג האמיטי הוא בעל הסתברות גבוהה ביותר או בין החמש הכל גובהות.

את התוצאות ניתן לראות בקצרה בגרף המוצג:

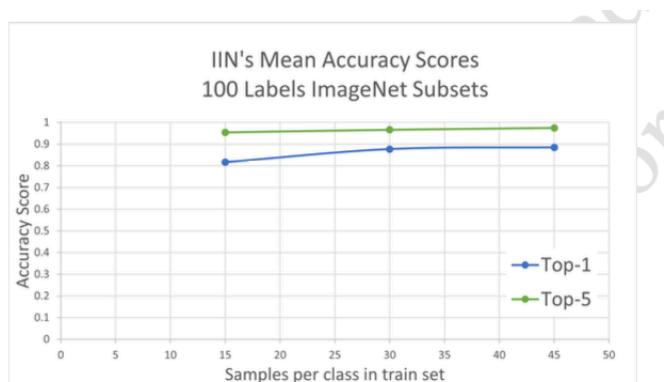


Figure 2: The top-1 and top-5 accuracy of EPKD on a randomly-sampled subset of ImageNet, consisting of 100 labels. The results are presented as a function of the number of images per label that were used in EPKD's training set.

בדיקות אבטחה:

כיוון שהקיים תהליכי הצפנה בכל השיטה חשוב לבדוק את איות הצפנה והאם ניתן לפיצוח על ידי גורם זה. החוקרים בדקו קודם את המקירה הטוריויאלי שבו התוקף נסה את כל הקומבינציות של פיקסל בתמונה המוצפנת כדי לשחרר את התמונה המקורי. הם ציינו שאפילו במקרה פשוט של תמונה 2×2 ומפתח הצפנה (משקولات המודל הגנרטיבי) בגודל 2×2 עם טווח ערכים של $^k 2 \dots 1$ מרחב האפשרויות הינו גדול מדי בשבייל לפיצח את ההצפנה ולשחרר את התמונה המקורי. לכן הם התמקדו בהתקפת ס"יבר יותר מתוחכמת הנוגראת (DA) ([Distinguishing attacks](#)). בהתקפה זו מטרת התוקף היא לא לגונב מידע כל שהוא אלא לבצע אנליזה שני פרייתי מידע שהצפנו. מציאת קשרים בין חלקים שונים, למשל בין המידע המוצפן לבין המידע המקורי או בין צהה התוקף מנסה להימנע מהמקירה בו יוכל ללמידה אחר מכון בשבייל לפצח חלקים מההצפנה. במקרה שלנו נרצה להימנע מהמקירה בו יוכל ללמידה על הקשר של שתי תמונות מוצפנות. אם אכן קיים קשר התוקף יוכל להסיק שמקורן באותו סיווג (למשל שתי תמונות של כלב).

התקפה נוספת בchner הכותבים הינה *Model Extraction*. בהתקפה זו התוקף מנסה לשחרר את משקלות המודל (משמע לאגנוב את המודל). המחברים ניסו את השיטות המתקדמות כיום בתחום תחת ההנחה של התקף יש גישה למודל בצורה של שילוח קלט וקבלת פלט. תחת שיטות אלו צריך מספר קטן של תמונות והగסה המוצפנות שלهن. את זה ניתן על ידי שליחת תמונות למודל וקבלת הפלט שלו (ההצפנה במקורה זה). התוצאות האמפיריות מראות שבשביל שהתקפה תעבור על התקף להחזק ב-10% דוגמאות מגודל המודל. לעומתם אם גודל המודל הינו 10 מיליון פרמטרים, התקף יצרר בערך מיליון דוגמאות של קלט ופלט של המודל. הטענה היא שהמודל חסין לסוג זה של התקפה בשל העובדה שהמודל שפתח ההצפנה משתנה לעתים תכופות תוך כדי שימוש המודל. לעומתם אם התקף ייצר מיליון דוגמאות הוא לא יוכל לעשות זאת תחת ההנחה שפתח הצפנה אחד (משקלות המודל) הוא זה שייצר את התמונות המוצפנות וכך התקפה נשברת.

סיכום:

בעבודה זו הראנו שניית להשתמש במודלים שונים כדי להצפן את המידע בצורה עיליה, כדי לשלווח מידע בצורה מאובטחת על גבי האינטרנט למודל גדול הנמצא בענן. אוטה שיטה לא סובלת מבעיות של איטיות בהצפנה EH או בעיות אבטחה שונות. השיטה משתמשת בשני רשות ניורונים אשר מאותחלות באופן רנדומלי ולא אימון מקודדים תמונות. הקידוד משמשת להצפנה אשר נשלחת מוצפנת לענן. המודל בענן שלוח לנו חזירה את וקטור הסתברויות של התמונה המוצפנת. נשתמש ברשות ניורונים נוספים, קטנה ממשמעותית מהמודל בענן כדי ללמידה לשחרר את הסיווג האמיתי של התמונה בהינתן וקטור הסתברויות והקידוד של התמונה. שיטה זו מראה את השימוש של רשות ניורונים לא רק כמשחזר מידע טובה אלא גם כמצפיני מידע טובים ופותחת דלת לשימוש השיטה על סוג נתונים נוספים כגון טקסט או DATA טבלאי.

שיטת פעולה: הסקירה נכתבת בשיטת פעולה עם עדן יבין.

Review 89: Faster Convergence for Transformer Fine-tuning with Line Search Methods

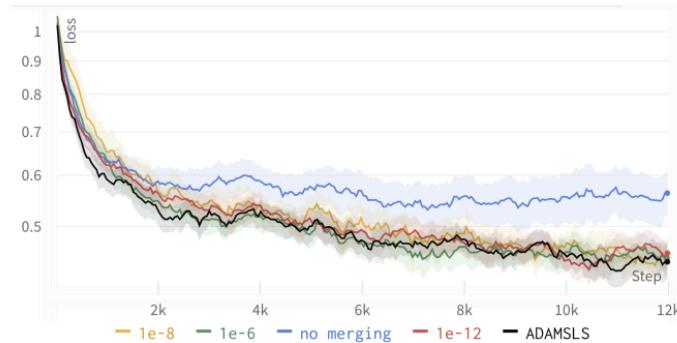


Fig. 2. Different merging thresholds on one epoch of the MNLI dataset.
Standard error is indicated around each line.

זמן לא כתבתי סקירה ונתקلت במאמר חמוד הדן במשפחהGISות לאימון רשות ניורונים שלא הייתה מודע לקיומה. אתם בטח ידעים איך היום מאמנים רשות ניורונים ומודלי ML אחרים (=מגדירים את פונקציית הלווי שלהן). כמובן עם כל מיני שכלולים של גישת מורד הגרדי-אנט הסטוכסטי (או Stochastic Gradient Descent או SGD בקצרה).

בגadol SGD היא שיטה איטרטיבית בכל איטרציה המשקלים של המודל מוזדים בכיוון של הגרדייאנט השלייל שהוא הכוון (הлинארי) שבו פונקציית loss קטנה "הכי הרבה". כאמור קיימים לא מעט שפכרים של SGD כמו ADAM RMSProp ושיטות רבות נוספות צבירת גרדיאנט (מומנוטם) שמטרתם היא להציג את קצב התכנסות של SGD ולהפוך אותו ליותר יציב. נזכיר שבכל השיטות האלו בכל איטרציה מעדכנים את המשקלים על סמך מיני-באץ' ולא דוגמא אחת כמו ב-SGD קלאסי.

המאמר שנסקור היום מציע גישה אחרת (מהירה יותר לטענת המחברים) למצער של פונקציית loss עבור מודלי מושטי טרנספורמרים. קודם כל נציין כי ערך של פונקציית loss לא בהכרח יורד (על מיני-באץ' של איטרציה זו) אחרי עדכון של פרמטרי המודל באיטרציה של כל שיטה מבוססת מورد הגרדייאנט.

לפעמים הlös על מיני-באץ' עשוי לעלות אחרי העדכון גם אם אתם משתמשים בשיטות מתקדמות כמו ADAM או RMSProp. עבור SGD (ולומר MiniBatch GD) זה נובע בגודל מכך שקצב למידה (learning rate) גדול מדי. בשיטות עם מומנטום כמו ADAM המצביע מרכיב יותר (כי הכוון שבו מזינים את המשקלות הוא לא הגרדייאנט המוצע של מיני-באץ') אך הבעיה עדין קיימת.

חשוב להבין שעלייה זמנית של פונקציית loss עבור מיני-באץ' זה ושם היא "לא אסון" אם המגמה הכללית של ירידת loss נשמרת במהלך האימון. אבל נשאלת השאלה האם שיטת אימון שתבטיח אי עלייה של פונקציית loss בכל איטרציה תוביל לאימון יעיל ומהיר יותר בלי לפגוע באיכות המודל המתתקבל בסוף האימון. זו השאלה שמחברים המאמר מנסים לענות עליה.

השיטה הנדונה במאמר מציעה דרך מאד אינטואיטיבית לעקוף את הסוגיה זו. כאמור אי ירידת של ערך פונקציית loss ב-SGD נובעת מקצב למידה גדול מדי. המאמר מציע לבחור את קצב הלמידה כך שיבטיח ירידת של פונקציית loss עבור כל המיני-באץ' בכל איטרציה של אימון.

בגдол בכל איטרצית אימון מתחילה בקצב למידה אקראי ומקטנים אותו (נגיד מחלוקת ב 2) עד שהlös אחרי העדכון יורד (על המיני-באץ'). ניתן לעשות את זה גם עם שיטות מתקדמות שנזכרו לעיל בשלב האחרון הוא הזזה של משקלים הכוון מוסים עם מקדם מסוים (קצב למידה). השיטה זו נקראת Armijo Line Search או ALS.

המאמר מציע להפעיל את ALS על כל שכבה (בלוק של טרנספורמר) בנפרד. לעומת מחלקים משקלים של המודל ל-L קבוצות כאשר L הוא מספר השכבות הראשית. לאחר מכן מבצעים עדכון של המשקלים לכל שכבה בנפרד עם שיטת אופטימיזציה שבחרתם (SGD, ADAM etc) משולבת עם ALS. ולומר מורידים קצב למידה לכל שכבה בנפרד עד שהערך של פונקציית loss אחרי העדכון יקטן כמעט אחר המשקלות מוקפות.

לדעתי עם השיטה המוצעת האימון ייקח יותר זמן (כי מעדכנים כל שכבה בנפרד) אבל ניתן לעבוד עם באץ'ים גדולים יותר שזה תורם ליציבות תהליכי האימון. השיטה מראה תוצאות לא רעות על פיין טוון של הטרנספורמרים.

<https://arxiv.org/abs/2403.18506v1>

היום ב **#shorthbrewpaperreviews** סוקרים קצרות מאמר:

Review 90, Short: Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust.

המאמר מציע שיטה לגילוי של תמונות נוצרות באמצעות מודל דיפוזיה גנרטיביים.

השיטה מאפשרת למשל להגן על קניין רוחני כי באמצעותו ניתן לאZHות שימוש במודלים גנרטיביים "פרטיטים". המאמר מוסיף וקטור watermark ' בעל תכונות מיוחדות (ring) לוקטור גaus שמננו יוצרים תמונות על ידי מודל דיפוזיה. וקטור זה ניתן לזייה עבור תמונות שגונטו אותו.

וקטור watermark ניתן לזייה דרך הפעלת מודל "ההופך" מודל דיפוזיה מאומן כדי לקבל דגימה של רעש שמננו נוצרה התמונה (diffusion model inversion). וקטור watermark חסן (ניתן לזייה) גם אם התמונה הנוצרת סובבה, או כל הפקולטים מוטשטשים או מזדים בגורם/SHIPUT קבוע.

HuggingFace: <https://huggingface.co/papers/2305.20030>

Paper: <https://arxiv.org/abs/2305.20030>