לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא "Learning to summarize from human feedback" שפורסם לראשונה לפני כחודשיים

תחומי מאמר: תמצות אבסטרקטיבי (abstractive summarization) של טקסטים, למידת חיזוק (reinforcement learning)

מאמר יוצג בכנס: NeurIPS 2020

כלים מתמטיים, מושגים וסימונים: טרנספורמרים, proximal policy optimization (PPO), מרחק KL, מבחן ROUGE, פונקציית גמול (reward function - F_rew) למודלים למידת חיזוק, שיטות אזור אימון (trust region TR), פונקצית מטרה סרוגייט (surrogate objective - F_sur)

תמצית מאמר: המאמר מציע שיטה לשימוש יעיל בתיוג אנושי של דאטה במשימת תמצות אבסטרקטיבי של טקסטים. תמצות אבסטרקטיבי של טקסט/פוסט זה סיכום קצר (עד 48 טוקנים במאמר זה) של עלילתו שלא בנוי מהמשפטים מוכלים בטקסט המקורי (כמו תמצות אקסטרקטיבי).

כמו שאתם אולי יודעים רוב המודלים לתמצות אבסטרקטיבי היום מאומנות לחקות את התמציות שנכתבו ע״י אדם וביצועיהם נמדדים לרוב בשיטות השוואה בין קטעי טקסט כמו ROUGE. המאמר מציין שלא שיטות האימון ולא מטריקה לשערוך הביצועים הנ״ל לא מספקות אינדיקציה מספיק טובה על איכות התמצית שזה המדד הכי חשוב לביצועי מודלים מסוג זה. בעקבות זאת בשנים האחרונות נעשו מאמצים לשלב פידבק אנושי בתהליך אימון של מודלים לתמצות אבסטרקטיבי. הגישות האלו מבוססות לרוב על דירוג של תמציות, שגונרטו ע״י המודל, ע״י בני אדם. הבעיה בגישה זו היא סקלביליות - רוב המודלים המודרניים לתמצות מכילים מיליארדים של פרמטרים ונדרשים דאטה סטים מאוד גדולים בשביל לאמן אותם.

המאמר מציע דרך יותר יעילה לניצול של פידבק אנושי על תמציות שמקטין באופן ניכר את כמות התמציות שצריך לתייג (לדרג בעצם). בגדול המאמר מציע לדגום זוגות של תמציות של אותו טקסט/פוסט מכמה מודלים מאומנים לתמצות. לאחר מכן המתייגים (בני אדם) מחליטים מה מה התמצית היותר טובה מכל זוג - ככה הם בונים את הדאטה סט שלהם. לאחר מכן הם מאמנים מודל M_abs המשערך את איכות התמצית על סמך התיוגים האלו(ככל שהתמצית יותר טובה, היא מקבלת ציון גבוה יותר). בשלב האחרון הם מריצים שיטת מעולם של למידת חיזוק, הנקראת PPO, כאשר המטרה הינה לאמן מודל הבונה תמציות אובסטרקטיביות תוך כדי מקסום הציון ניתן עי" M_abs.

הסבר של רעיונות בסיסיים: כמו שכבר אמרנו התהליך המוצע במאמר מכיל 3 שלבים עיקריים:

- בניית דאטה סט D_pair:

השלב הזה הוא היחיד שבו נרדשת ההתערבות האנושית. נותנים לכל מתייג את הטקסט ושני תמציות של טקסט זה שנדגמו מאחת המודלים שאומנו לגנרט תמציות. המתייג צריך לסמן את התמצית הטובה מבין השתיים. מה זה אומר טוב - תמצית צריכה להוות סיכום טוב של עלילת הטקסט ובנוסף עליה להיות מספיק קצרה (עד 48 טוקנים). נציין את המתייגים לא נותנים שום ציון רך לתמציות, רק נותנים לייבל ״0״ לתמצית פחות טובה ולייבל ״1״ לתמצית טובה יותר מהשתיים.

- אימון מודל המשערך את איכות התמצית M_score (בהינתן הטקסט כמובן)

כאן לוקחים קטע טקסט ושתי תמציות ומעבירים אותם למודל (רשת נוירונים כמובן) המחשב את ״איכותן״. המודל פולט שני ציונים (אחד לכל תמצית) כאשר פונקצית לוס מנסה למקסם את הפרש בין הציונים של תמצית טובה יותר לבין הפחות טובה מהזוג. בדרך זו התמציות היותר איכותיות יקבלו ציון גבוה כאשר הפחות טובים יקבלו ציון יותר נמוך. לאחר אימון מודל זה מקפיאים אותו ועוברים לשלב הבא. שימו

לב שבשלב זה לא מאמנים שום מודל לבניית תמציות - רק את המודל שמחשב את ציון התמצית בהינתן טקסט. פונקציית לוס כאן היא לוגריתם של הסיגמואיד של הפרש הציונים.

- <u>אימון מודל לתמצות אבסטרקטיבי על סמך M_score</u>

מריצים אלגוריתם PPO מעולם למידת החיזוק בשביל לאימון מודל תמצות אבסטרקטיבי M_abs כאשר פונקצית גמול F_rew זה הציון שניתן לתמצית עי״ M_score (היא קבועה בשלב הזה). זאת אומרת הם מנסים לאמן מודל תמצות לגנרט תמציות בעלי ציון גבוה. המטרה כאן הינה לאמן M_abs (שהוא בעצם הפוליסי במקרה הזה) כך שזה ימקסם את F_rew. בעצם הם לוקחים מודל מאומן לתמצות ועושים לו כיול בדרך זו.

אבל אם נסתכל על הנוסחה של פונקצית המטרה R(x,y) נגלה שיש בה עוד איבר המכיל מרחק KL (עם מינוס) בין התפלגות הפלטים (מותנה בטקסט הקלט) בין המודל הנלמד עי" PPO לבין המודל שנלמד בתהליך אימון רגיל (ללא שימוש בלייבלים על זוגות תמציות -נקרא לו מודל בייסליין). יש לזה שתי מטרות: המטרה הראשונה היא למנוע ״מוד קולפס״ של מודל מבוסס PPO. המטרה השנייה היא מניעת ״התרחקות יתר״ של מודל PPO מהמודל הבייסליין. <u>כאן יש הנחה סמויה שהמודל המקורי הוא לא כזה גרוע וצריך לשפר אותו רק ״בקטנה״ בשביל.</u>

צריך לציין שהם השתמשו בארכיטקטורה של הטרנספורמר (בסגנון 3-GPT) לגינרוט תמציות בכל המודלים שלהם.

הסבר על מושגים חשובים במאמר:

<u>עקרונות של אלגוריתם PPO</u>: אלגוריתם זה שייך למשפחת שיטות policy gradient שהיא בעצם הכללה של שיטת TR הקלאסית. TR מנסה לאמן מודל פונקציית פוליסי (policy function - P_i) שבמקום למקסם ישירות את F_rew באופן ישיר, ממקסם פונקצית חלופית (סרוגייט) F_sur. פונקציה חלופית זו מנסה לשפר את P_i עי״ מקסום התוחלת (על מרחב המצבים) של פונקצית היתרון F_adv המוכפלת ביחס של P_i החדשה ל- P_i הישנה. בדרך זו P_i החדשה לומדת לתת הסתברויות גבוהות למצבים שבהם פונקצית היתרון מקבלת ערכים גבוהים כלומר הגמול אחרי עדכון של P_i הינו מקסימלי. דרך אגב השם של השיטה (אזור אימון) נובע מהעובדה שבעיית אופטימיזציה זו פותרים תחת אילוץ שבכל עדכון של P_i מרחק KL בין P_i החדשה לישנה חסום עי״ קבוע קטן. אילוץ זה נדרש בשביל לא לתת ל P_i ״להתפרע״ כי השונות בבאצ'ים עלולה להיות גבוהה. קיימים כמה סוגים של F_sur שאחד מהם, למשל, משנה את ערך המקסימלי של מרחק KL כפונקציה של ממוצע של מרחקי KL בין P_i החדש לישן בכמה באצ'ים אחרונים.

אז PPO מאמצת גישה דומה לבניית פונקציית מטרה של מוסיפה אליה שתי תוספות: היא מוסיפה לפונקצית מטרה את השגיאה הריבועית הממוצעת של שערוך פונקצית ערך (value function) על הבאטץ' ומנסה לשפר את יכולת גילוי (exploration) של P_i עי" מקסום של האנטרופיה שלה. נציין שהמאמר בחר להשתמש בשתי רשתות שונות לשערוך של P_i ושל פונקציית ערך.

מדד ROUGE: משווה בין שני קטעי טקסט עי" השוואה של סטטיסטיקות על n-גרמים בין הקטעים.

הישגי מאמר: המאמר משווה את איכות התמצות של המודלים שלהם מול המודלים שאומנו ללא התערבות אנושית כאשר מספר הפרמטרים במודלים שווה (כאן הם לוקחים בחשבון גם את המודל מהשלב השני). הם מראים שעבור אותו מספר פרמטרים המודל שלהם מוציאה תמציתים יותר איכותיים (ההשוואה מתבצעת עי" אדם שמחליט איזה מהתמציות יותר טובה). בנוסף הם מראים שיכולת ההכללה של השיטה שלהם יותר טובה מאשר מודלי SOTA (מאמנים על דאטה סט מדומיין טקסטואלי מסוים ומריצים בדומיין אחר). הם גם משווים את איכות התמצית בכמה פרמטרים שונים כמו קוהרנטיות, דיוק וכיסוי וגם כאן הם משאירים את המתחרים מאחור (לאותו מספר של פרמטרים).

נקודה מעניינת: הם מציינים במאמר (הם נותנים גם דוגמאות)שאיכות התמצית מגיע למקסימום כאשר מאמנים את המודל M_abs מספיק זמן (לא עולה עם אנו מזרימים אליה דוגמאות נוספות וממשיכים לאמן) ולא מספקים הסבר לכך. אני חושב שזה נובע מהצורה של פונקציית המטרה שלהם המשלבת מקסום של ציון התמצית תוך כדי שמירת מרחק KL קטן בין המפלגות התמצית המגונרט עי" לבין ההתפלגות המגונרט עי" מודל ללא התערבות אנושית. אני חושב שזה גורם ל PPO "ליצור דוגמאות אדוורסריות" קרי ללא שינוי משמעותי בהתפלגות הפלט לגרום לשינוי גדול בציון שלו. זו

לינק למאמר:https://arxiv.org/pdf/1707.06347.pdf

לינק לקוד: לא מצאתי

נ.ב. מאמר מרשים עם רעיון מקורי המשלב טכניקות מלמידת החיזוק שאנו לא מרבים לראות במאמרי NLP. השיטה שלהם מעלה את יעילות הניצול של הפידבק האנושי אך עדיין יקרה מדי (ל OPEN AI אין בעיות תקציביות) כדי לבנות מודלים לתמצות אבסטרקטיבי בדומיינים אחרים.