# Review 53: Learning to summarize from human feedback

**Paper: https://arxiv.org/abs/2410.13116v2**

פינת הסוקר:

המלצת קריאה ממייק: מאוד מומלץ.

בהירות כתיבה: גבוהה מינוס

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: הבנה טובה בשיטות הקיימות של abstractive summarization , בטרנספורמרים וידע בסיסי ב-reinforcement learning.

יישומים פרקטיים אפשריים: אימון של מודלים לתמצות אבסטרקטיבי עם עם פחות דאטה מתויג

פרטי מאמר:

פורסם בתאריך: 27.10.20, בארקיב.

הוצג בכנס: NeurIPS 2020.

תחומי מאמר:

תמצות אבסטרקטיבי (abstractive summarization) של טקסטים

למידה באמצעות חיזוקים (RL - reinforcement learning)

כלים מתמטיים, מושגים וסימונים:

טרנספורמרים

פונקצית מטרה סרוגייט (surrogate objective - F\_sur)

(proximal policy optimization (PPO

שיטות אזור אימון (trust region TR)

פונקציית גמול (reward function)

מרחק KL

מבחן ROUGE

תמצית מאמר:

המאמר מציע שיטה לשימוש יעיל בתיוג אנושי של דאטה עבור משימות תמצות אבסטרקטיבי של טקסטים. תמצות אבסטרקטיבי של טקסט/פוסט הינו סיכום קצר (עד 48 טוקנים במאמר זה) של עלילתו, שלא בנוי מהמשפטים מהטקסט המקורי (אם תנאי זה לא מתקיים המשימה נקראת תמצות אקסטרקטיבי).

כמו שאתם אולי יודעים, רוב המודלים לתמצות אבסטרקטיבי היום מאומנים לחקות את התמציות שנכתבו עי״ בני אדם וביצועיהם נמדדים לרוב בשיטות השוואה בין קטעי טקסט כמו ROUGE. המאמר מציין כי הן שיטות האימון והן מטריקה למדידת ביצועים אלו לא מספקות אינדיקציה מספיק טובה על איכות התמצית, המהווה המדד הטבעי החשוב ביותר לביצועי מודלים לתמצות אבסטקטיבי.

בעקבות זאת בשנים האחרונות נעשו מאמצים לשלב משוב (פידבק) אנושי בתהליך אימון של מודלים לתמצות אבסטרקטיבי. גישות אלו מבוססות לרוב על דירוג של תמציות, שגונרטו באמצעות המודל, ע"י בני אדם. הבעיה העיקרית עם גישה זו היא סקלביליות - רוב המודלים המודרניים לתמצות מכילים מיליארדים של פרמטרים ונדרשים דאטהסטים מאוד גדולים בשביל לאמן אותם בצורה מספיק טובה.

המאמר הנסקר מציע גישה יותר יעילה לניצול של פידבק אנושי על תמציות. השיטה המוצעת מקטינה באופן ניכר את כמות התמציות שצריך לתייג. בגדול המאמר מציע לדגום זוגות של תמציות של אותו טקסט/פוסט מכמה מודלים מאומנים לתמצות. לאחר מכן המתייגים (בני אדם) מחליטים מה התמצית היותר טובה מכל זוג של תמציות - ככה המחברים בונים את הדאטהסט שלהם. לאחר מכן הם מאמנים מודל (נסמן אותו ב-M\_abs), המשערך את איכות התמצית על סמך תיוגים אלו (ככל שהתמצית יותר טובה, היא מקבלת ציון גבוה יותר). בשלב האחרון מריצים שיטת מעולם של למידת באמצעות חיזוקים, 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 דאטהסט D\_pair:

שלב זה הוא היחיד שבו נרדשת התערבות אנושית. נותנים לכל מתייג טקסט ושתי תמציותיו שנדגמו מאחד המודלים שאומנו לגנרט תמציות. המתייג צריך לסמן את התמצית הטובה מבין השתיים. "טיב" התמצית מוגדרת לפי שני הקריטריונים הבאים: התמצית צריכה להוות סיכום טוב של עלילת הטקסט ועליה להיות מספיק קצרה (עד 48 טוקנים). נציין כי 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 טקסט ושתי תמציותיו ומעבירים אותם למודל (רשת נוירונים כמובן), המשערכת 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 ציונים גבוהים ואילו הפחות טובים יקבלו ציונים נמוכים יותר. לאחר אימון המודל מקפיאים את המשקליו 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.

אימון מודל לתמצות אבסטרקטיבי על סמך M\_score.

מריצים אלגוריתם PPO מעולם למידת באצעות חיזוקים בשביל לאימון מודל תמצות אבסטרקטיבי M\_abs, כאשר פונקצית גמול F\_rew היא הציון שניתן לתמצית ע״י המודל M\_score. זאת אומרת מנסים לאמן מודל לגנרט תמציות בעלי ציון גבוה. המטרה כאן היא לאמן המודל M\_abs (שהוא בעצם הפוליסי במקרה הזה) כך שזה ימקסם את F\_rew. בעצם המאמר לוקח מודל מאומן לתמצות ועושים לו כיול בדרך זו.

אם נסתכל על הנוסחה של פונקצית גמול F\_rew, נגלה כי יש בה עוד איבר, המכיל מרחק KL (עם מינוס) בין התפלגות הפלטים (מותנה בטקסט הקלט) בין המודל הנלמד באמצעות PPO לבין המודל הנלמד 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. כאן יש הנחה סמויה שהמודל המקורי הוא לא כזה גרוע וצריך לשפר אותו רק ״בקטנה״ בשביל להשיג ביצועים טובים טובים.

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

הסבר על מושגים חשובים במאמר:

עקרונות של אלגוריתם PPO: אלגוריתם זה שייך למשפחת שיטות policy gradient שהיא בעצם הכללה של שיטת TR הקלאסית. Trust Region (TR) מנסה לאמן מודל פונקציית פוליסי (policy function - P\_i) שבמקום למקסם פונקצית גמול F\_rew באופן ישיר, ממקסם פונקצית גמול חלופית (סרוגייט) F\_sur. פונקציה חלופית זו מנסה לשפר את הפוליסי P\_i ע״י מקסום התוחלת (מעל 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. 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 בבאטצ'ים עלולה להיות גבוהה. קיימים כמה צורות של פונקציית גמול חלופית F\_sur שאחת מהן, למשל, משנה את ערך המקסימלי של מרחק KL כפונקציה של ממוצע של מרחקי KL בין P\_i החדש לישן בכמה באטצ'ים אחרונים.

PPO מאמצת גישה דומה לבניית פונקציית מטרה של מוסיפה אליה שתי תוספות:

מוסיפה לפונקצית מטרה את השגיאה הריבועית הממוצעת של שערוך פונקצית ערך (value function) על הבאטץ'.

מנסה לשפר את יכולת גילוי (exploration) של P\_i עי" מקסום של האנטרופיה שלה.

נציין כי המאמר בחר להשתמש בשתי רשתות שונות לשערוך של P\_i ושל פונקציית ערך.

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

הישגי מאמר:

המאמר משווה את איכות התמצות של המודלים שאומנו באמצעות הגישה המוצעת, מול המודלים שאומנו ללא התערבות אנושית כאשר מספר פרמטרים במודלים שווה (כאן הם לוקחים בחשבון גם את המודל מהשלב השני). המאמר מראה כי עבור אותו מספר פרמטרים המודל שלהם מוציאה תמציות יותר איכותיות (ההשוואה מתבצעת עי" אדם שמחליט איזה מהתמציות יותר טובה). בנוסף הם מראים שיכולת ההכללה של השיטה המוצעת יותר טובה מאשר מודלי SOTA (מאמנים על דאטהסט מדומיין טקסטואלי מסוים ומריצים בדומיין אחר). המחברים 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: המאמר מציין (בצירוף דוגמאות) כי איכות התמצית מגיע למקסימום כאשר מאמנים את המודל M\_abs מספיק זמן (היא לא עולה אם מזרימים אליה דוגמאות נוספות וממשיכים לאמן) ולא מספקים הסבר לכך. אני חושב שזה נובע מהצורה של פונקציית המטרה המוצעת, 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.

נ.ב.

מאמר מרשים עם רעיון מקורי המשלב טכניקות מלמידת באמצעות חיזוקים 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.