סקירה זו היא חלק מפינה קבועה בה אני סוקר מאמרים חשובים בתחום ה-ML/DL, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמתי, אתם מוזמנים לבדוק את העמוד שמרכז אותם תחת השם deepnightlearners.

לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא:

InfoBERT: Improving Robustness of Language Models from an Information Theoretic Perspective

---

פינת הסוקר:

המלצת קריאה ממייק: חובה בהחלט לאוהבי נושא של אימון אדוורסרי ותורת המידע. לאחרים מומלץ מאוד

בהירות כתיבה: בינונית פלוס

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: היכרות עם עקרונות של התקפות אדוורסריות לרשתות נוירונים (בדגש על NLP), הבנה טובה במושגי יסוד של תורת המידע כמו מידע הדדי של משתנים אקראיים.

יישומים פרקטיים אפשריים: אימון מודלי NLP, עמידים להתקפות אדוורסריות.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: רשמי, לא רשמי

פורסם בתאריך: 22.03.21, בארקיב.

הוצג בכנס: ICLR 2021

תחום מאמר:

- טרנספורמרים, BERT
- אימון אדוורסארי - adversarial training
- למידת ייצוג - representation learning

כלים ומושגים מתמטיים במאמר:

- צוואר בקבוק מידעי (information bottleneck) ברשתות נוירונים
- מידע הדדי (mutual information)
- InfoNCE (noise contrastive estimation)

תמצית מאמר:

המאמר הנסקר מציע שיטה להתמודדות עם התקפות אדוורסריות כנגד מודלי שפה גדולים בסגנון BERT (עם קלט טקסטואלי). הגישה המוצעת מבוססת על העיקרון של צוואר בקבוק מידעי (information bottleneck) עבור רשתות נוירונים. עקרון זה מגדיר את מטרת האימון של רשת נוירונים כמיקסום של פונקציית מטרה L_ib כאשר L_ib היא הפרש בין שני איברים (כל אחד מהם הינו מידע הדדי). האיבר הראשון משערך יכולת חיזוי של רשת והאיבר השני מודד את מידת דחיסת קלט ע"י רשת.

המאמר מציע להוסיף ל-L_ib איבר נוסף, הממקסם את המידע ההדדי של ייצוג הקלט (משפט או מקטע של טקסט עבור מודלי שפה) לבין ייצוגי טוקנים, שנקראים במאמר localized anchored. טוקנים localized anchored הינם טוקנים רובסטיים (חסינים) להתקפות אדוורסריות וגם מועילים למשימת downstream.

הטענה המרכזית של המאמר (מוכחת בחלקה תיאורטית ובחלקה אמפירית) שאימון מודל שפה עם פונקציית המטרה המוצעת משפר את הרובסטיות (חסינות) של הרשת נגד

דוגמאות (התקפות) אדוורסריות. מעניין כי המאמר מראה (אמפירית) שטענה זו נכונה גם עבור אימון רשת על דאטהסטים רגילים ללא דוגמאות אדוורסריות וגם באימון על דאטהסטים, המכילים דוגמאות כאלו.

רעיון בסיסי:

מחברי המאמר טוענים (ומוכיחים ריגורוזית) שאימון רשת נוירונים כללית עם פונקציית המטרה המוצעות מקטין את ההפרש בין:

- מידע הדדי, המסומן ב-I(T, Y), של ייצוג קלט נקי (לא אדוורסרי) והחיזוי של רשת עבור קלט זה (לייבל)
- מידע הדדי בין ייצוג קלט מורעש (אדוורסרי) לבין אותו החיזוי של רשת, המסומן ב-I(T',Y).

בנוסף אימון עם פונקצית מטרה כזו ממקסם את מידע הדדי בין ייצוג הקלט לבין הלייבלים הנחזים באמצעות רשת, המתורגם לביצועי מודל במשימת downstream.

למה זה טוב, אתם שואלים? שימו לב שבסופו של דבר המטרה של האימון האדוורסרי הינה הפיכת הרשת לעמידה נגד דוגמאות אדוורסריות. כלומר חיזוי של רשת לא אמור להשתנות כאשר הופכים דוגמא רגילה לדוגמא אדוורסרית (צריך לזכור כי בדרך כלל משנים דוגמא בצורה מינורית כדי להפוך אותה לאדוורסרית). עקב העובדה ש-I(T',Y) ו-I(T,Y) מהווים מדד לביצועי רשת עם קלט אדוורסרי ורגיל בהתאמה, מזעור ההפרש ביניהם מתורגם (לטענת המאמר) לביצועים טובים יותר של מודל בתרחיש אדוורסרי.

תקציר מאמר:

בשביל להבין את רעיון המאמר אנו צריכים להבין מה זה בעצם עיקרון צוואר בקבוק מידעי ברשתות נוירונים:

עקרון צוואר בקבוק מידעי ברשתות נוירונים:

עיקרון צוואר בקבוק מידעי (שהומצא ע״י פרופ' תשבי ב-2015) מגדיר את מטרת למידה עמוקה (כלומר אימון של רשת נוירונים) כטרייד-אוף בין דחיסת מידע ע״י רשת (בניית ייצוג דחוס של קלט) לבין יכולת החיזוי שלה. עקרון זה מתורגם למיקסום מידע הדדי בין

ייצוג קלט T לבין חיזוי של רשת Y, המסומן I(T, Y), ובאותו זמן למינימיזציה של מידע הדדי בין קלט X לייצוג עצמו, המסומן כ- I(X, T). שימו לב כי I(T, Y) מהווה אינדיקציה לביצועי רשת על סט האימון. לעומת זאת I(X, T) אפשר לפרש כאיבר רגולריזציה למזעור אוברפיטינג (overfitting).

המאמר הנסקר מציע לאמן מודל שפה על משימת downstream עם פונקציית מטרה שבליבה עקרון צוואר הבקבוק של מידע. כמו שכבר ראינו קודם הפונקציה המוצעת מכילה את מידע הדדי בין ייצוג של קלט T לבין קלט X (שבמקרה שלנו הינו משפט). T מכיל את האמבדינגס (ייצוגים) של טוקנים, המרביסים משפט X, כאשר מימד של ייצוג של טוקן T_i הוא 768 (עבור BERT Base). המימד הגבוה של T אינו מאפשר לחשב/לשערך את I(X,T) בצורה ישירה. המאמר מציע (ומוכיח ריגורוזית) שניתן לחסום I(X,T) מלמטה ע״י סכום של I(X,T_i) המוכפל במספר הטוקנים במשפט. שימוש בחסם זה הופך את בעיית אופטימיזציה זו לקלה יותר מבחינה חישובית.

דוגמא אדוורסרית בעולם NLP:

בשביל להמשיך את ניתוח המאמר בואו נבין מה זה דוגמא אדוורסרית בדומיין של NLP. נזכיר שדוגמא אדוורסרית נוצרת באמצעות הוספת רעש קטן לדוגמא רגילה כדי לעוות את הלייבל הנחזה עבורה באמצעות הרשת. בדומיין טקסטואלי משפט אדוורסרי נוצר עי״ שינוי של משפט רגיל השומר מרחקים בין האמבדינגס של המילים במשפט המקורית לבין המילים של ״המשפט האדוורסרי״ קטנים. שינוי כזה לא גורם לשינוי הלייבל של המשפט (ז"א מתייג אנושי היה מעניק למשפט את אותו לייבל כמו למשפט המקורי) אך הוא כן ״מבלבל את הרשת״ שמשנה את החיזוי שלה עבור המשפט המורעש (האדוורסרי).

כבר אמרנו כי המאמר הנסקר מציע להוסיף לפונקציית המטרה המקורית L_ib איבר רגולריזציה נוסף, הממקסם את סכום של המידעים ההדדיים של ייצוג משפט Z והייצוגים של הטוקנים הנקראים במאמר local anchored (LA).

טוקני LA:

ייצוגים של טוקנים LA הם בעלי שתי התכונות הבאות:

- רובסטיים (חסינים) בתרחישים אדוורסריים.

- מכילים מידע מועיל למשימת downstream.

המאמר מציע לאתר טוקנים בעלי תכונות אלו באמצעות איתור של טוקנים שדווקא לא מקיימים את הדרישות האלו (!!). כדי לאתר את טוקנים לא חסינים נגד התקפות, המאמר מציע "לבצע" התקפות אדוורסריות על הטוקנים. המטרה כאן היא לזהות טוקנים ששינוי קטן בייצוגם מביא לעלייה משמעותית בלוס של ממשימה downstream. טוקנים כאלו מהווים מועמדים נוחים לבנייה של דוגמא אדוורסרית על גביהם. מצד שני יש לנו טוקנים כמו stopwords או סימני פונקטואציה שאפילו שינוי גדול באמבדינג שלהם לא גורם לעלייה גדולה בלוס של המשימה. עם זאת טוקנים אלו כלל לא מועילים למשימה. בהקשר זה המאמר מציע לאתר טוקנים ששינוי מתון בהם גורם לשינוי מתון בלוס עבור משימת downstream ולנצל אותם כ״עוגני אמבדינג של המשפט״.

מכיוון שאנחנו רוצים לנצל כמה שיותר את המידע מטוקני LA בשביל לבנות ייצוג משפט עמיד נגד דוגמאות אדוורסריות. זו הסיבה שמוסיפים סכום של כל המידעים ההדדיים בין ייצוג המשפט Z וטוקני LA לפונקצית מטרה.

איך משערכים את פונקצית המטרה בפועל:

אז הכל טוב ויפה אבל נשאלת השאלה איך אנחנו נאמן רשת כאשר פונקצית מטרה שלה כוללת כל מיני מידעים הדדיים בין וקטורים אקראיים שונים? הרי ידוע ששיערוך של מידע הדדי הינו untractable ובדרך כלל משתמשים בחסמים בשביל לבנות פונקצית מטרה שהיא יותר נוחה לאימון רשת. במאמר הנסקר נעזרים ב-InfoNCE עם פונקציית מרחק נתונה g בין הייצוגים ובונים פונקצית מטרה ש״מקרבת״ את הייצוגים שאנו רוצים למקסם את המידע ההדדי ביניהם (כמו ייצוג המשפט והטוקנים LA), ״ומרחיקה״ את הייצוגים של טוקנים ומשפטים הנבחרים בצורה רנדומלית. פונקצית מרחק g יכולה להיות מרחק cosine או שממודלת באמצעות רשת MLP עם שתיים-שלוש שכבות.

Algorithm 1 - Local Anchored Feature Extraction. This algorithm takes in the word local features and returns the index of local anchored features.

1: **Input:** Word local features t, upper and lower threshold c_h and c_l
2: $\delta \leftarrow 0$ *// Initialize the perturbation vector δ*
3: $g(\delta) = \nabla_\delta \mathcal{L}_{\text{task}}(q_\psi(t+\delta), y)$ *// Perform adversarial attack on the embedding space*
4: Sort the magnitude of the gradient of the perturbation vector from $||g(\delta)_1||_2, ||g(\delta)_2||_2, ..., ||g(\delta)_n||_2$ into $||g(\delta)_{k_1}||_2, ||g(\delta)_{k_2}||_2, ..., ||g(\delta)_{k_n}||_2$ in ascending order, where z_i corresponds to its original index.
5: **Return:** $k_i, k_{i+1}, ..., k_j$, where $c_l \leq \frac{i}{n} \leq \frac{j}{n} \leq c_h$.

איך זה נעשה? בונים מיני-באטץ', המורכב מזוג אחד של ייצוג משפט וטוקן LA ממנו (ייצוגים קרובים), כאשר שאר הזוגות מורכבים ממשפט וטוקנים שנבחרו רנדומלית. פונקצית מטרה היא יחס כאשר המונה מכיל אקספוננט של מרחק בין ייצוגים של "הזוג הקרוב" והמכנה מכיל את סכום אקספוננטים של המרחקים בין כל הזוגות. Oord et al הוכיח ב-2018 שמיקסום של פונקציה מטרה מצורה זו מגדיל את מידע הדדי בין ייצוגים של זוגות קרובים כלומר משיג את המטרה שלנו כאן.

הישגי מאמר:

המאמר מראה את עליונותה של שיטת האימון (כיול) InfoBert בהתמודדת נגד דוגמאות אדוורסריות עבור BERT ו-ROBERTA עבור כמה דאטהסטים אדוורסריים בעלי דרגות קושי שונות. כמו שכבר הזכרתי גם אימון של InfoBERT על דאטהסט ללא דוגמאות אדוורסריות וגם עם דוגמאות אדוורסריות גורם למודל המאומן להיות יותר עמיד לרעש.

Training	Model	Method	Dev				Test			
			A1	A2	A3	ANLI	A1	A2	A3	ANLI
Standard Training	RoBERTa	Vanilla	49.1	26.5	27.2	33.8	49.2	27.6	24.8	33.2
		InfoBERT	47.8	31.2	31.8	**36.6**	47.3	31.2	31.1	**36.2**
	BERT	Vanilla	20.7	26.9	31.2	26.6	21.8	28.3	28.8	26.5
		InfoBERT	26.0	30.1	31.2	**29.2**	26.4	29.7	29.8	**28.7**
Adversarial Training	RoBERTa	FreeLB	50.4	28.0	28.5	35.2	48.1	30.4	26.3	34.4
		InfoBERT	48.4	29.3	31.3	**36.0**	50.0	30.6	29.3	**36.2**
	BERT	FreeLB	23.0	29.0	32.2	28.3	22.2	28.5	30.8	27.4
		InfoBERT	28.3	30.2	33.8	**30.9**	25.9	28.1	30.3	**28.2**

Table 1: Robust accuracy on the ANLI dataset. Models are trained on the benign datasets (MNLI + SNLI) only. 'A1-A3' refers to the rounds with increasing difficulty. 'ANLI' refers to A1+A2+A3.

נ.ב.

מאמר מציע רעיון מאוד מעניין המתבסס על עקרון צוואר בקבוק מידעי עבור רשתות נוירונים לשיפור את עמידות רשת נגד התקפות אדוורסריות. אהבתי את ההוכחות הריגורוזיות ויפות שיש במאמר (במיוחד משפט 3.2).

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson.

מיכאל עובד בחברת הסייבר Salt Security בתור Principal Data Scientist. מיכאל חוקר ופועל בתחום הלמידה העמוקה, ולצד זאת מרצה ומנגיש את החומרים המדעיים לקהל הרחב.