לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners (לא הספקתי לסיים אתמול בלילה 🙂) עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא**:**

INFOBERT: IMPROVING ROBUSTNESS OF LANGUAGE MODELS FROM AN INFORMATION THEORETIC PERSPECTIVE

שיצא לפני שבוע וחצי.

תחומי מאמר: BERT/transformers, adversarial training, representation learning

-

כלים ומושגים מתמטיים במאמר: צוואר בקבוק למידע ברשתות נוירונים, מידע הדדי, InfoNCE(noise contrastive estimation)

מאמר יוצג בכנס: ICLR 2021

תמצית מאמר: המאמר מציע שיטה להתמודד עם התקפות אדוורסריות נגד מודלי שפה גדולים בסגנון BERT (עם אינפוט טקסטואלי) ומציע גישה חדשה לכיולם (fine-tuning) של מודלים מסוג זה.

הגישה שלהם מתבסס על העיקרון של צוואר בקבוק למידע ברשתות נוירונים. עקרון זה מגדיר את מטרת האימון של רשת נוירונים כמיקסום של פונקציית מטרה L_ib. למעשה L_ib זה הפרש בין שני איברים (כל אחד מהם הינו מידע הדדי), שהראשון מתאר את יכולת חיזוי של רשת והשני מודד את מידת דחיסת אינפוט עי" רשת. המאמר מציע להוסיף L_ib עוד איבר הממקסם את המידע ההדדי של ייצוג אינפוט (של המשפט) לבין הייצוגים של הטוקנים רובסטיים שמצד אחד ומועילים למשימת downstream מאידך (טוקנים אלו הנקראים localized anchored tokens).

הטענה המרכזית של המאמר (מוכחת בחלקה תיאורטית ובחלקה אמפירית) שאימון מודל שפה עם פונקציית מטרה זו משפרת את הרובסטיות של הרשת נגד דוגמאות אדוורסריות. מעניין שהם מראים (אמפירית) שטענה זו נכונה גם עבור אימון רשת עם דאטה סט רגיל ללא דוגמאות אדוורסריות וגם באימון על דאטה סט המכיל דוגמאות כאלו.

רעיון בסיסי: מחברי המאמר טוענים (ומוכיחים ריגורוזית) שאימון רשת נוירונים (כללית) עם פונקציית מטרה המקורית L_ib מקטין את ההפרש בין המידע ההדדי של ייצוג האינפוט הנקי (הלא אדוורסרי) והחיזוי (לייבלים) המסומן כ- I(T, Y), לבין המידע ההדדי של ייצוג האניפוט המורעש (אדוורסרי) ואותו החיזוי המסומנת כ- I(T',Y). בנוסף אימון עם פונקצית מטרה כזה ממקסם את המידע ההדדי בין ייצוג האינפוט וחיזוי הלייבלים, המתורגם לביצועי מודל במשימת downstream.

למה זה טוב, אתם שואלים? שימו לב שבסופו של דבר המטרה של האימון האדוורסרי הינה הפיכת הרשת לרובסטית לדוגמאות אדוורסריות כלומר החיזוי של הרשת לא אמור להשתנות כאשר הופכים דוגמא רגילה לדוגמא אדוורסרית. עקב העובדה ש I(T',Y) ו- I(T, Y) מהווים מדד לביצועי הרשת עם האינפוט האדוורסרי והרגיל בהתאמה, מזעור ההפרש ביניהם מתורגם לביצועים טובים יותר של המודל בתרחיש אדוורסרי.

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

<u>עקרון צוואר בקבוק למידע ברשתות נוירונים</u>:

עיקרון זה (שהומצא עי״ פרופ' תשבי ב 2015) מגדיר את מטרת הלמידה העמוקה (כלומר אימון של רשת נוירונים) כטרייד אוף בין דחיסת מידע עי״ הרשת (בניית ייצוג מקומפרס של אינפוט) לבין יכולת החיזוי שלה. עקרון זה מתורגם למיקסום המידע ההדדי בין ייצוג האינפוט (משפט במקרה שבנדון) T ובין החיזוי של Y, המסומן I(T, Y) ובאותו זמן למינימיזציה של המידע הדדי בין האינפוט X לייצוג עצמו, המסומן כ- I(X, T). שימו לב ש- I(T, Y) מהווה אינדיקציה לביצועי הרשת על הטריין סט. לעומת זאת I(X, T) אפשר לפרש כאיבר רגולריזציה למזעור אוברפיטינג.

אז המאמר מציע לאמן מודל שפה על משימת downstream על פונקציית מטרה הנוסעת עקרון צוואר הבקבוק של מידע. כמו שכבר ראינו קודם הפונקציה מכילה את המידע ההדדי בין ייצוג האינפוט T לבין האינפוט עצמו X שבמקרה שלנו הינו משפט. הייצוג T המכיל את האמבדינגס של כל הטוקנים מהמשפט X של מימד הייצוג של כל טוקן הוא 768 (עבור BERT Base). המימד הגבוה של T אינו מאפשר לחשב/לשערך את I(X, T) כמו שהוא. המאמר מציע (ומוכיח ריגורוזית) שניתן לחסום I(X, T) מלמטה עי״ הסכום של I(X, T_i) מוכפל במספר הטוקנים במשפט שהופך את בעיית אופטימיזציה זו לקלה יותר בהרבה מהבחינה החישובית.

<u>דוגמא אדוורסרית בעולם NLP:</u> בשביל להמשיך את ניתוח המאמר בואו נבין מה זה דוגמא אדוורסרית בדומיין של NLP. נזכיר שדוגמא אדוורסרית נוצרת עי" הוספת רעש קטן לדוגמא רגילה כדי לעוות את חיזויה ע"' הרשת. בדומיין טקסטואלי משפט אדוורסרי נוצר עי״ שינוי קטן של המשפט המקורי (כך המרחקים בין האמבדינגס של המילים משתנים בצורה מינורית בלבד) שלא משנה את חיזויו (עי״ אדם) אך ״מבלבל את הרשת״ שכן משנה את החיזוי שלה לגבי המשפט המורעש.

כמו שכבר הזכרתי המאמר מציע להוסיף לפונקציית המטרה המקורית L_ib איבר רגולריזציה נוסף הממקסם את סכום של המידעים ההדדיים של ייצוג משפט Z והייצוגים של הטוקנים הנקראים local anchored(LA)

<u>טוקני LA</u>: ייצוגים של טוקנים כאלו מקיימים שני התנאים הבאים: הם צריכים להיות רובסטיים בתרחישים אדוורסריים ובנוסף הם צריכים להכיל מידע מועיל למשימת downstream. המאמר מציע לאתר טוקנים בעלי תכונות אלו עי״ חיפוש של הטוקנים שלא מקיימים את הדרישות האלו. כדי לאתר את הטוקנים הלא רובסטיים הם מבצעים התקפות אדוורסריות כדי לזהות טוקנים <u>ששינוי קטן בייצוגם</u> מביא לעלייה משמעותית בלוס של מהמשימה downstream כלומר טוקנים המהווים 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. הבעיה שטוקנים אלו כלל לא מועילים למשימה. אז מה שהמאמר מציע זה לבחור טוקנים ששינוי מתון בהם גורם לשינוי מתון בלוס של משימת downstream ולנצל אותם כ״עוגני אמבדינג של המשפט״. כלומר אנחנו רוצים לנצל כמה שיותר את המידע מ- local anchored tokens בשביל לבנות ייצוג משפט עמיד לדוגמאות אדוורסריות. זו הסיבה שמוסיפים סכום של המידעים ההדדיים בין ייצוג המשפט Z וטוקנים אלו.

<u>איך משערכים את פונקצית המטרה בפועל</u>: אז הכל טוב ויפה אבל נשאלת השאלה איך אנחנו נאמן רשת כאשר פונקצית מטרה שלה כוללת כל מיני מידעים הדדיים בין וקטורים שונים. הרי ידוע ששיערוך של מידע הדדי הינו untractable ובדרך כלל משתמשים בחסמים בשביל לבנות פונקצית מטרה שהיא יותר נוחה לאימון. במאמר משתמשים ב- InfoNCE שעבר פונקציית מרחק נתונה g בין הייצוגים (זה יכול להיות מרחק קוסיין או לפעמים רשת MLP עם שתיים-שלוש שכבות) בונים פונקצית מטרה ש״מקרבת״ את הייצוגים שאנו רוצים למקסם את המידע ההדדי ביניהם (כמו ייצוג המשפט והטוקנים local anchored ממנו) ״ולהרחיק״ את הייצוגים של טוקנים ומשפטים הנבחרים בצורה רנדומלית. אז בונים מיני באטץ' המורכב ״מזוג ייצוגים אמיתי״ ובשאר הזוגות המשפט והטוקנים נבחרים רנדומלית. פונקצית reward במקרה הזה מורכבת מאקספוננט של המרחק בין הייצוגים של ״הזוג האמיתי״ במונה כאשר המכנה מכיל את סכום אקספוננטים של המרחקים בין כל הזוגות. Oord et al הוכיח ב 2018 שמיקסום

של פונקציה מטרה מצורה זו מגדילה את המידע הדדי בין הייצוגים של הזוגות האמיתיים כלומר משיגה את המטרה שלנו כאן.

הישגי מאמר: הם מראים את עליונותה של שיטת אימון (כיול) InfoBert בהתמודדת נגד דוגמאות אדוורסריות עבור BERT ו- ROBERTA עבור כמה דאטה סטים אדוורסריים בעלי דרגות קושי שונות. כמו שכבר הזכרתי גם אימון של InfoBERT על דאטה סט ללא דוגמאות אדוורסריות וגם עם דוגמאות אדוורסריות.

לינק למאמר: **https://arxiv.org/pdf/2010.02329.pdf**

לינק לקוד: אין

נ.ב. מאמר עם רעיון מאוד מעניין המתבסס על עקרון צוואר בקבוק של מידע עבור רשתות נוירונים בשביל לשפר את עמידות הרשת בתרחישים אדוורסריים. אהבתי שהמאמר מנסח את הקשר הזה בצורה ריגורוזית ומוכיח אותו (במשפט 3.2). בקיצור המלצת קריאה חמה!

#DeepNightLearners