לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה
היום בחרתי לסקירה את המאמר שנקרא:
Language Through a Prism: A Spectral Approach for Multiscale Language Representation
שיצא לפני קצת פחות מחודש.

הוגש לכנס: 2020 NeurIPS

תחומי מאמר:
- חקר תכונות מודלי NLP עמוקות

כלים מתמטיים, מושגים וסימונים:
- אנליזה ספקטרלית לגילוי של קשרים במגוון סקאלות בייצוג של טקסט (אמבדינג)
- התמרת קוסינוס דיסקרטיתDCT)), התמרת קוסינוס דיסקרטית ההופכית (iDCT)
- מסנן מעביר נמוכים (,(LPFמסנן מעביר גבוהים(,(HPFמסנן מעביר פס ((BPF

תמצית מאמר: שפות טבעיות מציגות תכונות מבניות שונות בכמה סקאלות שונות החל מהרמה של מילה עד רמת הפיסקה והמסמך. בהקשר זה נשאלת השאלה האם המודלים, המבוססים על רשתות הנוירונים בתחום ה NLP, תופסים את התכונות ההיררכיות אלו? האם ניתן "לשפר את הרשת אם מאלצים אותה" לחקות את התכונות הללו? איך תכונות אלו משתנות בין מודלים למשימות שונות? המאמר הנסקר מנסה לתת מענה על השאלות האלו.

למעשה המאמר מציע שיטה לבחון תכונות וביצועי מודל NLP בסקאלה נתונה עי" הורדותן של כל הסקאלות האחרות המודל. למשל בשביל לבדוק את ביצועי המודל בסקאלת קצרת טווח (רמת מילה) למשימה ספציפית, הם מאלצים את המודל "לא להשתמש" בסקאלות ארוכות טווח (משפטים, פסקאות וכדומה). זה נעשה עי" שימוש בטכניקות ספקטרליות מתחום עיבוד אותות המאפשרות לסנן(בתחום התדר) רק את התכונות בסקאלה הנדרשת. כאן סקאלות קצרות טווח (רמת מילה) מיוצגות עי" תדרים גבוהים כאשר סקאלות ארוכות טווח מיוצגות ע"י תדרים גבוהים יותר (נפרט על כך בהמשך).

השיטה מסתמכת על הפעלה של מסננים ספקטרליים על אקטיבציות של נוירונים לאורך הטקסט(זה מימד ה"זמן" שלנו !!). כלומר אם נרצה לבדוק עד כמה סקאלה קצרה (מילה או שתיים, תדרים גבוהים) משפיעה על ביצועי מודל, מוסיפים למודל שכבה המפלטרת החוצה את כל הסקאלות הארוכות (תדרים יותר נמוכים). אם ביצועי מודל לא משתנים בצורה משמעותית כתוצאה מהסינון הזה, המסקנה היא ש"תלויות(סקאלות) ברמת מילה" חשובות חשובות יותר לביצוע מוצלח של המשימה מאשר תלויות ארוכות טווח. כלומר במשימה זו "למודל מספיק להתמקד בתלויות קצרות טווח בטקסט" בשביל להשיג ביצועים טובים.

בנוסף טכניקה זו מאפשרת לבודד את התכונות(מידע) הקשורות לסקאלה ולהפריד אותן מהתכונות הסמנטיות בייצוגים של טוקנים. בשביל להגיע להפרדה זו מוסיפים למודל שכבה המעבירה חלקים שונים של וקטורי ייצוג של הטוקנים (אמבדינגס) דרך מסננים ספקטרליים שונים.

הערה: המאמר טוען שבעיקרון ניתן להוסיף שכבה כזו (שנקראת Prism) לא רק כהשכבה האחרונה של הרשת, אך בפועל בכל הניסויים שהם עשו, הם הוסיפו את Prism אחרי שכבת האמבדינגס של BERT. בעקבות זה אתייחס בהמשך רק לסינון הספקטרלי של שכבת ייצוג הטוקנים(אמבדינגס).

כמו שכבר אמרנו, המיקום של וקטורי הייצוג בטקסט משחק תפקיד של מימד ה"זמן" . בסוף מאמנים את הרשת עם שכבת Prism למשימות שונות. אז משווים את הביצועים של רשת עם Prism עם הרשת המקורית במשימה הזו בשביל לבדוק האם הפרדה זו תורמת לביצועים.

הסבר של רעיונות בסיסיים:
בואו ננסה להבין איך בעצם עובדת שכבת Prism:

- <u>חלוקה לסקאלות (תדרים):</u> מחלקים את הרכיבים של וקטורי הייצוג לכמה תת-קבוצות. כלומר אם יש לנו אמבדינגס באורך 360 ואנו רוצים לבחון 3 סקאלות שונות, הרכיבים 1,..., 120(קבוצת אינדקסים S_1) יהיו "אחראים" על הסקאלה ראשונה עם התדרים הגבוהים ביותר (ברמת מילה עד שתי מילים נגיד), הרכיבים 121,..., 240 (קבוצה 2S) ייצגו את הסקאלה השניה עם התדרים הבינוניים (ברמת "המשפט") , ו- 120 הרכיבים האחרונים S_3 "ישויכו" לסקאלה 3 של התדרים הנמוכים ביותר (ברמת "פסקה/המסמך")

<u>הסבר בעניין התדרים:</u> השאלה המתבקשת כאן למה "סקאלה של מילה" מייצגת דווקא תדרים גבוהים בזמן שה"סקאלה של מסמך" מייצגת דווקא את התדרים הנמוכים ביותר? התשובה לכך נובעת מהעובדה ש"התדר של סקאלה בטקסט" הינו ביחס הפוך ל"מחזור" של אותה סקאלה. למעשה "המחזור" של "מילה" הינו נמוך ביותר בזמן של מחזור של "סקאלת הפיסקה" הינו גבוה הרבה יותר. הסיבה לכך שהטקסט מורכב מהרבה מילים, פחות משפטים ועוד פחות פסקאות

- <u>בנייה של וקטורי דגימות T לכל נוירון באמבדינג</u>: לכל אינדקס i בווקטורי הייצוג על פני כל הטוקנים בטקסט, בונים וקטור דגימות T_i. למשל עבור רכיב מסוים בווקטור הייצוג (נגיד במיקום 213) ובונים וקטור דגימות T_213 המורכב מכל הרכיבים מס' 213 על פני כל הייצוגים של הטוקנים בטקסט.
- <u>העברה של וקטורי T_i דרך DCT:</u>מפעילים את התמרת קוסינוס דיסקרטיות DCT (יפורט בהמשך) על כל וקטור D_i ובונים להם את הייצוגים הספקטרליים שלהם(בתחום התדר). הייצוג הספקטרלי של וקטור דגימות T_i יסומן ב F_i. נציין כי כל וקטורי דגימות עוברים אותה התמרת כלומר אם יש לנו 360 וקטורי T_i, אנו צריכים לבצע 360 DCTים (לכל אחד בנפרד). חשוב לזכור שהמימד של כל וקטור T_i שווה למספר הטוקנים בטקסט(!!)
- <u>סינון ספקטרלי של וקטורי F_i:</u> לכל וקטור F_i בוחרים את המסנן הספקטרלי שלו לפי האינדקס i. וקטוריF_i עם אינדקסים מקבוצה 1S (ברמת מילה) יועברו דרך מסנן מעביר גבוהים HPF, האינדקסים מקבוצה 3S יועברו דרך מסנן מעביר נמוכים ואינדקסים מקבוצה 2S יועברו דרך מסנן מעביר פס BPF (ההסבר על איך עובדים המסננים נמצא בפרק הבא)

- <u>העברה של וקטורי F_i המסוננים דרך התמרת קוסינוס ההופכית iDCT:</u> למעשה iDCT מעבירה את הספקטרום המסונן של הייצוגים בחזרה לתחום "זמן" (נזכיר שאצלנו מימד הזמן זה האינדקסים של האמבדינגס לאורך הטקסט). נסמן את התוצאה של פעולה זו כ T_fi. שעבור כל i הווקטור T_fi בנוי מכל הרכיבים במיקום i של וקטורי הייצוג המסוננים.
-
- אימון רגיל של רשת (BERT) עם שכבת prism

הישגי מאמר:
<u>בחינת "חשיבות" של סקאלות לביצועי מודל עבור משימות שונות:</u> בשביל לבדוק את רמת ההשפעה של "סקאלה מסוימת" על הביצועים המחברים סיננו את כל הסקאלות האחרות. נניח שאנו רוצים לבחון את ההשפעה של סקאלת "המילים" (תדרים גבוהים) על ביצועי מודל במשימה מסוימת. אז מפעילים מסנן שמסנן את כל התדרים האחרים (הנמוכים והבינוניים) עי" העברה של ייצוגי הטוקנים לאורך הטקסט דרך HPF בצורה המפורטת בסעיף הקודם. צריך לציין שהמאמר חילק את כל הסקאלות (תדרים) ל 5 תחומים שווים באורך:

1. מילה - תדרים גבוהים
2. פסוקית (clause) - תדרים גבוהים-בינוניים
3. משפט - תדרים בינוניים
4. פיסקה - תדרים נמוכים בינוניים
5. מסמך - תדרים נמוכים

מהבדיקות שהם עשו עולה שלמשימת זיהוי נושא, התדרים הנמוכים הם הכי חשובים שזה דווקא די הגיוני כי צריך להבין את הטקסט כולו פחות או יותר בשביל לזהות את הנושא שלו. מה שמפתיע בתוצאות שלהם זה השיפור המשמעותי בביצועים של המודל מול המודל המקורי אחרי סינון של התדרים הגבוהים. במשימת סיווג אופי תגובה בדו-שיח, התדרים החשובים הם הבינוניים אבל לא בפער גדול על התדרים האחרים. במשימת זיהוי חלקי דיבור התדרים הגבוהים יצאו הכי משמעותיים שזה די מובן בהתחשב לאופי המשימה. הרי בשביל להבין לאיזה חלק דיבור לשייך מילה צריך לקחת בחשבון מילה או שתיים סמוכות.

מעניין שלמשימת זיהוי מילה ממוסכת שעליה אומן BERT (בנוסף לזיהוי סדר המשפטים) התדרים הכי חשובים הם הגבוהים ביותר כלומר בשביל לנחש מילה "תחת מסכה" מספיק לדעת מילה או שתיים מסביב אליה. בעיני זו תגלית מאוד מסקרנת(!!)

ביצועי מודל עם שכבת Prism: כאן הם הוסיפו שכבת prism ל BERT ובדקו את ביצועיה שלוש המשימות מהפיסקה הקודמת. הם הצליחו לשפר את הביצועים בצורה משמעותית לשתי משימות מתוך שלוש, כאשר עבור משימת זיהוי חלקי דיבור הם קיבלו תוצאות נמוכות טיפה מ BERT המקורי. הם השתמשו הדור סט 103-WikiText

הסבר על מושגים חשובים במאמר:

התמרת קוסינוס דיסקרטית DCT והופכית שלה :IDCT למעשה זה מקרה פרטי של התמרת פוריה הסטנדרטית. היא פועלת על סדרה של מספרים ממשיים ומעבירה אותה לסדרה ממשית מאותו אורך בתחום התדר. אינטואיטיבית, התמרה זו מחפשת דמיון בין הסדרה לפונקציות קוסינוס מתדרים שונים.

דאטה סטים:

- משימת זיהוי אופי תגובה בדו-שיח:(Dialog speech act classification) השתמשו ב Switchboard Dialog Speech Acts corpus
- משימת זיהוי נושא:20 Newsgroups dataset
- משימת זיהוי חלקי דיבור Penn Treebank :

לינק למאמר: https://arxiv.org/abs/2011.04823

לינק לקוד: לא הוגש

נ.ב. מאמר עם תוצאות מאוד מסקרנות המשתמש בטכניקות ספקטרליות לבחינה של תבניות (אורכי תלויות) עבור מודלי NLP עמוקים במשימות שונות. הבעיה שהם בדקו את התוצאות שלהם על מעט משימות ורק על דאטה סט אחד בלבד לכל משימה שקצת מקשה עליי להשתכנע שהתופעות שהם גילו מתרחשים במשימות NLP אחרות במגוון דאטה סטים. אני מצפה שהמשך של המחקר המאוד מעניין הזה...

#deepnightlearners