צהריים טובים חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה(שוב לא הספקתי לסיים בלילה)
היום בחרתי לסקירה את המאמר שנקרא
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, an Comprehension
שיצא לפני כשנה.

תחומי מאמר: טרנספורמרים, denoising autoencoder, מודלים גנרטיביים

תמצית מאמר: המאמר מציע ארכיטקטורה מסוג denoising autoencoder לשחזור דאטה טקסטאולי מורעש. אחרי האימון אפשר להשתמש בו למספר משימות NLP ו NLU גנרטיביות ודיסקרימינטיביות שונות כגון תרגום, מענה אוטומטי על שאלות, תמצות אבסטרטיבי לאחרים

תקציר מאמר: המאמר מנסה לשלב את היתרונות של GPT (וכל צאצאי) גנרוט מוצלח של טקסט ע"י למידת מודל שפה אוטורגרסיבי (משתמשת רק בטוקנים שקודמים לו) בצורה מפורשת עם הצד החזק של BERT שהצליח ללמוד מודל שפה דו-כיווני בצורה לא מפורשת. הבעיה המשמעותית של BERT נמצאת באי יכולתו לגנרט טקסטים בצורה פשוטה ושקופה (נכון שיש עבודות שמציעות שיטות "להכריח" את BERT לגנרט טקסט אך בדרך כלל זה די מסורבל ואיכות הטקסטים המגונרטים תמיד יותר נחותה ממודלי SOTA. אז מה בעצם החידוש ש- BART מציע? BART מורכב מהמקודד ומהמפענח(encoder) ו- decoder) ומאומן כמו denoising autoencoder קלאסי (הכוונה כאן ל- pretraining כי בעיקרון כל משימה downstream דורשת כיול של הרשת). כלומר הקלט למקודד הוא טקסט מורעש שהמקודד ממפה אותו למרחב לטנטי כאשר המטרה של המפענח זה לשחזר את הטקסט המקורי.
אפשר להסתכל על BART כהכללה מסוימת של BERT ו- GPT כאשר המקודד שלו משלב את הארכיטקטורה הדו-כיוונית של BERT והגישה אוטורגרסיבית (קרי בונה את הפלט משמאל לימין -הכוונה כאן לשפות שכותבים בהן משמאל לימין :)). גישה זו מאפשרת להרעיש את הקלט במגוון דרכים שבהחלט תורם חיובית לעוצמה הייצוגים שהמודל בונה. אזכיר שלהבדיל מ BERT שמסווה חלק מהמילים ומנסה לשחזר אותם (יש גם את הזיהוי האם זוג המשפטים באים אחד אחרי השני ב BERT - אתייחס לזה בהמשך), BART משתמש במספר שיטות מעניינות להרעשת הטקסטים.

שיטת אימון: בנוסף להסוואת הטוקנים, הם ניסו לאמן את BART עם שיטות הרעשה הבאות:

1. <u>מחיקת טוקנים</u>: והמודל צריך להחליט באילו מיקומים יהיו הטוקנים החסרים
2. <u>הסוואה של קבוצות טוקנים רציפים (text infilling</u>): הם מגרילים את מספר הטוקנים הרצופים שהם מסווים מהתפלגות פואסון (לא מסבירים למה פואסון ולא התפלגות דיסקרטית אחרת) ומחליפים את כל הטוקנים האלו ב טוקן MASK . צריך לציין לכל מספר הטוקנים המוסווים רק טוקן MASK אחד מחליף אותם. המודל מאומן לחזות כמה טוקנים הוסוו. הם מציין שמספר הטוקנים המוסווים יכול להיות אפס כלומר אף טוקן לא מוסתר ו- MASK פשוט הוכנס אל תוך הטקסט.
3. <u>פרמוטציה של משפטים</u>: סדר המשפטים שונה בהתאם לפרמוטציה אקראית. המודל צריך לחזות את הסדר הנכון של המשפטים.
4. <u>סיבוב המסמך</u> :טוקן נבחר באקראי והטקסט סובב באופן כזה שהטוקן הנבחר הופך להיות הטוקן הראשון. המטרה של המודל לזהות את התחלת הטקסט

הם בחנו את הגישות הנז"ל והביצועים הכי טובים מתקבלים כאשר משלבים את text infilling יחד עם פרמוטציה של משפטים (2 ו- 3 ברשימה). מעניין שימוש ב 2 ו- 3 יחד לאימון מכליל את הגישה שב BERT המסווה טוקנים בודדים ומנסה לחזות האם זוג המשפטים באים אחד אחרי השני. במאמר נטען שזה גורם

למודל להתחשב יותר באורך המשפט ולקחת בחשבון תלויות ארוכות טווח (כלומר להתחשב ביותר טוקנים/משפטים בשביל לחזות את הטוקן הבא.

ארכיטקטורה: היא די דומה לארכיטקטורה של BERT עם שני הבדלים משמעותיים: כל השכבות של המפענח מבצעות חישוב של cross- attention עם השכבה האחרונה של המקודד לעומת BERT שמבצע את זה רק בשכבה האחרונה של המפענח. ההבדל השני הוא העדר שכבות feed-forward לפני השכבה האחרונה. חוץ מזה יש הבדלים קלים נוספים כמו שימוש בפונקציית אקטיבציה מסוג GELU במקום RELU

פונקציית לוס: לא מצאתי אזכור במאמר, כנראה קרוס אנטרופי רגיל על הטוקנים המשוחזרים.

שיטות כיול fine-tuning)) למשימות שונות:

1. <u>משימות סיווג סיקוונס</u> **:**אותו קלט מוכנס למקודד ולמפענח והשכבה האחרונה של המפענח משמשת כמסווג מולטי-קלאס לינארי. זה קצת דומה ל CLS של BERT אבל כאן מוסיפים טוקן בסוף הטקסט כדי (לטענתם) שהמפנעח יוכל לנצל את הפלט של כל השכבות הקודמות שלו (hidden).

2. <u>משימות סיווג טוקן</u> **:**המסמך המלא הוכנס למקודד ולפענח והפלט של השכבה האחרונה של המפענח משמשת לסיוג הטוקן

3. <u>משימות גנרוט דאטה טקסטואלי:</u> כמו שאתם בטח זוכרים המפענח של BART הינו אוטורגרסיבי ואפשר לכייל אותו בצורה פשוטה בשביל משימות גנרטיביות כמו גנרוט תשובה על שאלה או תמצות אבסטרקטיבי. המקודד פשוט מקבל את הקלט והמפענח מגנרט את הפלט בצורה אוטורגרסיבית.
4. <u>משימת תרגום אוטומטי</u> כאן הם עשו משהו מעניין. הם הציעו להחליף את שכבת אמבדינג של BART במקודד נוסף הקלט אליו הינו השפה שמתרגמים ממנה (הם ניסו את להשתמש במודל הזה רק לתרגום לאנגלית). מקודד זה אומן מאפס למפות מילים מהשפה המתורגמת ל "אנגלית מקולקלת/מורעשת" ואז המפענח "מנקה" אותה והופך אותה לאנגלית תקנית. המודל אומן בשני שלבים שהלוס בשניהם הוא קרוס אנטרופי על הפלט של BART. בשלב הראשון מאמנים את המקודד החדש, השכבה הראשונה המקודד של BART(כולל positional embedding). בשלב השני מאמנים את הכל הפרמטרים מספר קטן של איטרציות.

הישגי מאמר: המחברים בחרו בדרך השואה מעניינת (לא סטנדרטית). קודם כל הם השווה את הביצועים של BART עם BERT ביחס למשימות רבות על דיאטה סטים שונים(שזה דווקא שגרתי לגמרי). בנוסף הם ניסו לאמן את BART במספר דרכים לאמן מודלים מהסוג הזה שהוצעו במאמרים אחרים (UniLM, XLNet, MASS וכדומה) והוכיחו ששיטתם יותר טובה מהם בכל המשימות שהם בדקו פרט למשימה אחת (רשימת המשימות פורטה בסעיף הבא). הייתי רוצה לראות השוואה של המודל שלהם מול ארכיטקטורות אחרות(השתכנעתי ששיטת האימון שלהם טובה אבל חסרה בהשוואה מול מודלים אחרים).

דאטה סטים ומשימות להשוואה: המחברים הניסו את המודל שלהם על מגוון רחב של משימות כגון: SQuAD, MNLI, ELI5, XSum, ConvAI2 **,** CNN/DM

נ.ב. מאמר עם רעיון פשוט שמצליח להוסיף יכולת גנרוט יעילה ל BERT. גם שיטת אימון שהם מציעים נראית לי מעניינת ויעילה. המאמר כתוב מעולה, קל מאוד לקרוא אותו. בקיצור מומלץ.

לינק למאמר: **paper**
לינק לקוד: **Code**(בתוך פייטורץ)

#DeepNightLearners