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:

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

---

פינת הסוקר:

המלצת קריאה ממייק: חובה לאנשי NLP, במיוחד לחוקרים העוסקים במודלי שפה, מבוססי טרנספורמרים.

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות עם מודלי שפה, המבוססים על טרנספורמרים כמו BERT ו-GPT.

יישומים פרקטיים אפשריים: גנרוט טקסטים ברמה גבוהה יותר ובדרך פשוטה יותר מאלו של BERT.

פרטי מאמר:

לינק למאמר: זמין כאן

לינק לקוד: זמין כאן (בתוך פייטורץ')

פורסם בתאריך: 29.10.19, בארקיב

הוצג בכנס: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

---

תחומי מאמר:

- טרנספורמרים
- denoising autoencoder
- מודלים גנרטיביים

תמצית מאמר:

המאמר הנסקר מציע ארכיטקטורת רשת חדשה מסוג denoising autoencoder לשחזור דאטה טקסטואלי מורעש. לאחר האימון ניתן להשתמש במודל לביצוע של מספר משימות NLP ו-NLU גנרטיביות ודיסקרימינטיביות כגון תרגום, מענה אוטומטי על שאלות, תמצות אבסטרקטיבי וכמה סוגי משימות נוספים.

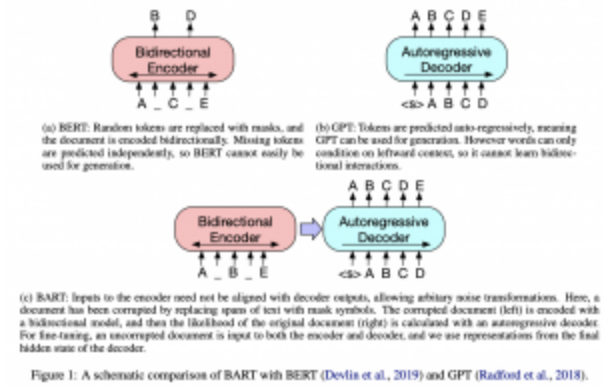

Figure 1: A schematic comparison of BART with BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

תקציר מאמר:

המאמר מנסה לשלב את התכונות החזקות של שני מודלי טרנספורמרים הכי פופולריים בתחום NLP היום:

- **GPT** (וצאצאיו): יכולת לגנרט טקסטים ברמה "גבוהה" (יחסית למתחרים) ע"י למידה של מודל שפה באופן אוטורגרסיבי (המודל משתמש רק בטוקנים שקודמים לטוקן הנחזה כדי לחזות את הטוקן הבא).
- **BERT**: למידת ייצוג לטנטי חזק (בעל מימד נמוך) של טקסט באמצעות למידת מודל שפה דו-כיוונית.

נזכיר כי החולשה המשמעותית של BERT נמצאת באי יכולתו לגנרט טקסטים בצורה פשוטה ושקופה (נכון שיש עבודות שמציעות שיטות המצליחות "להכריח" את BERT לגנרט טקסטים אך בדרך כלל זה די מסורבל ואיכות הטקסטים המגונרטים תמיד יותר נחותה ממודלי SOTA).

אז מה בעצם החידוש ש-BART מציע לנו? קודם כל 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 הלטנטי כאשר המטרה של המפענח זה לשחזר את הטקסט המקורי.

ניתן להסתכל על BART כהכלאה של BERT ו-GPT כאשר הוא משלב את הארכיטקטורה הדו-כיוונית של BERT והגישה אוטורגרסיבית של GPT (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 הלטנטיים של טקסט שהמודל בונה. אזכיר שלהבדיל מ-BERT שמסווה חלק מהמילים ומנסה לשחזר אותם (האימון של BERT מכיל גם את הזיהוי של סדר בין משפטים - אתייחס לזה בהמשך), BART משתמש במספר שיטות מעניינות להרעשת הטקסטים.

Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

שיטת אימון:

כאמור בנוסף להסוואת הטוקנים, המחברים הציעו 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 של טוקנים רציפים (text infilling):</u> כאן מגרילים **מספר** טוקנים רצופים שיוסוו, באמצעות שימוש בהתפלגות פואסון (לא מסבירים למה בחרו דווקא בפואסון ולא כל התפלגות דיסקרטית אחרת) ומחליפים את כל הטוקנים האלו בטוקן MASK. צריך לציין שלכל הטוקנים המוסווים יש טוקן MASK אחד בלבד שמחליף אותם. המודל מאומן לחזות כמה טוקנים הוסוו. המאמר מציין כי מספר הטוקנים המוסווים יכול להיות אפס כלומר אף טוקן לא מוסתר ו-MASK פשוט הוכנס אל תוך הטקסט.
3. <u>תמורה (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 מסובב באופן כזה שהטוקן הנבחר הופך להיות הטוקן הראשון. המטרה של המודל לזהות את התחלת הטקסט.

המחברים 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 של BERT המסווה, כאמור, טוקנים בודדים ומנסה לחזות סדר של זוג נתון של משפטים. במאמר נטען שזה גורם למודל להתחשב יותר באורך המשפט ולקחת בחשבון תלויות ארוכות טווח (כלומר להתחשב ביותר טוקנים/משפטים לחיזוי הטוקן הבא).

(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

Figure 3: Fine tuning BART for classification and translation.

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 באיזו פונקציית לוס השתמשו במאמר, כנראה קרוס אנטרופי רגיל על הטוקנים המשוחזרים.

שיטות כיול (fine-tuning) של BART באמצעות משימות שונות:

- **משימות סיווג של סדרת טוקנים:** אותו קלט מוזן למקודד ולמפענח והשכבה האחרונה של המפענח משמשת כמסווג מולטי-קלאס לינארי. זה קצת דומה לשימוש בטוקן CLS של BERT אבל כאן מוסיפים טוקן בסוף הטקסט ולא בהתחלה כדי (לטענתם) שהמפנעח יוכל לנצל את הפלט של כל השכבות הקודמות שלו (hidden) עבור.
- **משימות סיווג טוקן:** המסמך המלא הוכנס למקודד ולפענח והפלט של השכבה האחרונה של המפענח משמשת לסיווג של הטוקן.
- **משימות גנרוט דאטה טקסטואלי:** כמו שאתם בטח זוכרים המפענח של BART הינו אוטורגרסיבי וניתן לכייל אותו בצורה פשוטה בשביל משימות גנרטיביות כמו גנרוט תשובה על שאלה או יצירה של תמצות אבסטרקטיבי. המקודד פשוט מקבל את הקלט והמפענח מגנרט את הפלט בצורה אוטורגרסיבית.
- **משימת תרגום אוטומטי:** כאן המאמר עשה משהו מעניין. המחברים החליפו את שכבת אמבדינג של טוקנים ב-BART במקודד נוסף שהקלט שלו הוא השפה שמתרגמים ממנה (הם ניסו 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, שכבת self-attention הראשונה של המקודד של BART ו-positional encoding. בשלב השני מאמנים את כל הפרמטרים של BART במשך מספר קטן של איטרציות.

Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

Table 1: Comparison of pre-training objectives. All models are of comparable size and are trained for 1M step on a combination of books and Wikipedia data. Entries in the bottom two blocks are trained on identical dat using the same code-base, and fine-tuned with the same procedures. Entries in the second block are inspired by pre-training objectives proposed in previous work, but have been simplified to focus on evaluation objectives (se §4.1). Performance varies considerably across tasks, but the BART models with text infilling demonstrate the mos consistently strong performance.

-

הישגי מאמר:

המחברים בחרו בדרך השואה מעניינת (לא סטנדרטית). קודם כל הם השווה את הביצועים של BART עם BERT ביחס למשימות רבות על מספר דאטהסטים (שזה דווקא שגרתי לגמרי). אז הם אימנו את BART באמצעות כמה גישות שהוצעו בעבר לאימון מודלי שפה מבוססי טרנספורמרים (UniLM, XLNet, MASS וכמה אחרות) והוכיחו כי שיטת האימון המוצעת ל-BART מצליחה להשיג ביצועים יותר טובים מכולן בכל המשימות שנבדקו פרט למשימה אחת (רשימת המשימות מפורטת בסעיף הבא). הייתי רוצה לראות השוואה של המודל המוצע מול ארכיטקטורות נוספת (למרות שדי השתכנעתי ששיטת האימון המוצעת היא טובה אבל חסרה לי השוואה של BART מול מודלים אחרים).

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	**89.0**/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

Table 2: Results for large models on SQuAD and GLUE tasks. BART performs comparably to RoBERTa and XLNet, suggesting that BART's uni-directional decoder layers do not reduce performance on discriminative tasks.

משימות להשוואה:

המחברים ניסו את המודל שלהם על מגוון רחב של משימות כגון: SQuAD, MNLI, ELI5, XSum, CNN/DM , ConvAI2.

נ.ב.

מאמר כתוב היטב עם רעיון פשוט שמצליח להוסיף יכולת גנרוט יעילה ל-BERT. גם שיטת אימון המוצעת נראית לי מעניינת ויעילה. המאמר כתוב מעולה, קל מאוד לקרוא אותו. בקיצור מומלץ בחום.

#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.