**BERT**

(Bidirectional Encoder Representations from Transformers) BERT הוא מאמר שפורסם ע"י גוגל ומציג צורת אימון חדשנית ופורצת דרך בה האימון מתרחש על ידי הרצה דו-כיוונית של Transformer (מודל attention) למודלים של שפות. תוצאות המאמר מראות כי מודל שפה אשר מאומן באופן דו-כיווני יכול להיות בעל תחושה עמוקה יותר של הקשר וזרימה של השפה מאשר מודלים חד-כיוונים. במאמר, החוקרים מפרטים טכניקה חדשה בשם Masked LM (MLM) המאפשרת אימון דו-כיווני במודלים שבהם זה היה בלתי אפשרי בעבר.

BERT עושה שימוש ב-Transformer, מנגנון קשב הלומד יחסי הקשר בין מילים (או מילות משנה) בטקסט. בצורת הבסיסית (ונילה) שלו, Transformer כולל שני מנגנונים נפרדים - מקודד שקורא את קלט הטקסט ומפענח שמייצר חיזוי למשימה. המטרה של BERT היא ליצור מודל שפה, ולכן יש צורך רק במנגנון המקודד. בניגוד למודלים כיוונים, שקוראים את קלט הטקסט ברצף (משמאל לימין או מימין לשמאל), מקודד ה-Transformer קורא את כל רצף המילים בבת אחת. לכן זה נחשב דו-כיווני, אם כי יהיה נכון יותר לומר שהוא לא-כיווני. מאפיין זה מאפשר למודל ללמוד את ההקשר של מילה בהתבסס על כל סביבתה (משמאל ומימין למילה).

כאשר מאמנים מודלים של שפה, יש אתגר של הגדרת מטרת חיזוי. מודלים רבים מנבאים את המילה הבאה ברצף – המאמר טוען כי זו גישה כיוונית אשר מטבעה מגבילה את למידת ההקשר.

כדי להתגבר על אתגר זה, BERT משתמש בשתי אסטרטגיות אימון:

1. Masked LM (MLM)

בBERT מתבצע תהליך המקדים את הזנת רצפי המילים, במהלכו 15 אחוז מהמילים בכל רצף מולחפות ב token והמודל מנסה לחזות את הערך המקורי של המילים האלו בהתבסס על ההקשר שמספקות יתר המילים ברצף (שלא כוסו).

פונקציית הloss של BERT לוקחת בחשבון רק את החיזוי של ערכי המסיכה ולכן המודל מתכנס לאט יותר ממודלים כיווניים, אך זה משתנה כאשר המודל מפתח מודעות להקשר בין במילים בתהליך האימון.

1. Next Sentence Prediction (NSP)

במהלך הכשרת BERT, המודל מקבל זוגות משפטים כקלט ולומד לחזות אם המשפט השני בזוג הוא המשפט העוקב במסמך המקורי. במהלך האימון, רק 50% מהקלטים הם צמד משפטים עוקב.

כדי לעזור למודל להבחין בין שני המשפטים באימון, הקלט מעובד באופן הבא לפני הכניסה למודל:

1. CLS token מוכנס בתחילת המשפט הראשון ו SEP token מוכנס בסוף כל משפט.
2. לכל token נוסף משפט המציין משפט א' או משפט ב'.
3. לכל token מתווסף שיבוץ מקומי המציין את מיקומו ברצף.

כדי לחזות אם המשפט השני אכן קשור לראשון, מבצעים את השלבים הבאים:

1. כל רצף הקלט עובר דרך מודל Transformer.
2. פלט של CLS token (מסמן תחילת המשפט) הופך לווקטור בצורת 2×1
3. חישוב ההסתברות של IsNextSequence עם softmax.

בעת אימון מודל BERT, נשתמש ב- Masked LM וב-Next Sentence Prediction מאומנים יחד, במטרה למזער את פונקציית ההפסד המשולבת של שתי האסטרטגיות.

BERT יכול לשמש למגוון רחב של משימות שפה, תוך הוספת שכבה קטנה למודל הליבה:

* 1. משימות סיווג כגון ניתוח סנטימנט נעשות בדומה לסיווג משפט הבא.
  2. במשימות מענה שאלות, התוכנה מקבלת שאלה לגבי רצף טקסט ונדרשת לסמן את התשובה ברצף. ניתן לאמן מודל זה ע"י לימוד שני וקטורים נוספים – אחד לתחילת התשובה והשני לסופה.
  3. בזיהוי ישות לפי שמן (NER), התוכנה מקבלת רצף טקסט ונדרשת לסמן את סוגי הישויות השונים (אדם, ארגון, תאריך וכו') המופיעים בטקסט. באמצעות BERT, ניתן לאמן מודל מסוג זה.

באימון ה-Fine-tuning, רוב הפרמטרים ההיפר נשארים זהים לאימון BERT.

נקודות חשובות:

* נעדיף שגודל הדגם יהיה גדול ככל הניתן ( - BERT\_largeבעל 345 מיליון פרמטרים)
* עם מספיק נתוני אימון, יותר שלבי אימון נוכל להגיע לדיוק גבוה יותר.
* אימון דו-כיווני (MLM) מתעלה על אימון משמאל לימין, אפילו לאחר מספר קטן של שלבים לפני אימון .

לסיכום, BERT הוא ללא ספק פריצת דרך בשימוש בלמידת מכונה לעיבוד שפה טבעית. העובדה שהוא נגיש ומאפשר Fine-tuning מהיר ככל הנראה תאפשר מגוון רחב של יישומים מעשיים בעתיד.