🚀המאמר היומי של מייק -11.10.24: ⚡️🚀  
SELECTIVE ATTENTION IMPROVES TRANSFORMER

היום נסקור מאמר המציג רעיון לשיפור הליבה של הטרנספורמים, כלומר מנגנון ה-attention. להבדיל מהמאמר של סקרתי(Selective Transformer) הרעיון כאן די ברור לי מתמטית ולא ולא זיהיתי בו נוסחאות מתמטיות ״מפתיעות״. המאמר של היום מציע שיטה לשיפור ביצועים של הטרנספורמרים ועל הדרך מצליח להקטין את גודל הזכרון הנדרש עבורו.

המחברים טוענים (ובצדק) שלפעמים יש טוקנים שלא כדאי לטרוח ולחשב מקדמי attention עבור זוגות מסוימים של הטוקנים. בנוסף ניתן לדעת את זה על ידי הסתכלות על טוקנים ביניהם ואלו באים לפניהם (ההקשר). ֿ

המחברים נותנים את הדוגמא הבאה הממחישה את התופעה הזו. נניח שהטוקנים א, ב, ג הוזנו לטרנספורמר. בשכבה כלשהו עם מ attention סטנדרטי, טוקן ב מחליט ״כמה הוא מעוניין לקחת״ מטוקן א (מקדם attention), וטוקן ג יכול להחליט כמה לקרוא מטוקן א, אבל טוקן ב אינו יכול להשפיע על כמה טוקן ג ״לוקח״ מטוקן א. אם טוקן ב קבע שטוקן א אינו רלוונטי או אפילו מטעה לטוקנים עתידיים כמו ג, אין שום דבר שהוא יכול לעשות בשכבה הנתונה כדי לתקן זאת. השיטה המוצעת על ידי המחברים באה לתקן (להקל) את הבעיה הזו.

הרעיון המוצע הוא מאוד אינטואיטיבי ואלגנטי. המחברים מציע להחסיר מווקטור ה-attention (לפני חישוב הסופטמקס) של כל טוקן מטריצת מיסוך נלמדת F. איבר ij במטריצת F (עבור זוג טוקנים i- ו- j) מבטא עד כמה אנו רוצים להקטין את ה-attention בין טוקנים אלו. ערך גבוה של F\_ij מסמן לנו שהמודל ״מאמין״ שצריך ״להתעלם מהקשר בין טוקן i ל-j כלומר (אם i > j; מטריצה F הינה מטריצה קוזלית כלומר F\_ij =0 אם i<j) אפשר לא לדלג על חישוב מקדם ה-attention ביניהם.

אבל מה זה מטריצת F ואיך היא נבנית? גם בצורה מאוד אינטואיטיבית F) הינה שסכום של מטריצות מיסוך רכה S עבור כל הטוקנים בין j ל-i. כלומר טוקן j אינו משפיע על מקדמי מיסוך עבור ה-attention לטוקנים שקודמים ל-j. המחברים לא מסבירים למה הם בחרו לעשות את זה ככה (למישהו יש רעיון?). כמובן מטריצה S הינה אי שלילית (עושים ReLu).

השיטה המוצעת יכולה כאמור לעזור בהאצת האינפרנס על ידי הורדה של טוקנים עם מקדמי F הגדולים ביותר מחישוב ה-attention (לטוקן i נתון). למעשה זה סוג של pruning שהוא תחום מחקר די פעיל ברשתות הנוירונים. המחברים מציעים להגדיר ״תקציב attention" לכל שכבה (בלוק של טרנספורמר) ובאופן הדרגתי להעיף מספר קבוע של טוקנים מחישוב ה-attention (נעשה באיטרציות). כל פעם מורידים טוקנים עם ערכי F הגבוהים ביותר ובוחרים שכבה שעבורה הורדה כזו משפיעה באופן המועט ביותר על ה-perplexity (כלומר log-likelihood).

בנוסף כבר במהלך האימון של מטריצות S אנו יכולים לגרום למודל ״לבטל״ יותר נוירונים על ידי הוספה של איבר לפונקציית הלוס הרגילה שלה(log-likelihood), הקונס את המודל על S בעלת ערכים נמוכים מדי.

יש לי תחושה שהמאמר הזה הוא התחלה של משהו מעניין…

https://arxiv.org/pdf/2410.02703