🚀המאמר היומי של מייק -10.10.24: ⚡️🚀  
DIFFERENTIAL TRANSFORMER

המאמר הזה עשה הרבה גלים ביומיים האחרונים וזו הסיבה שבחרתי אותו לסקירה היומית שלי. המאמר החזיר אותי 3-4 שנים אחורה לתקופה שבה על בסיס ימי יצאו מאמרים המציעים שכלולים שונים לליבה של הטרנספורמרים כה אהובים עלינו. כמובן אני מתכוון למנגנון ה-attention שמאפשר לנו לכמת קשרים בין הטוקנים השונים בטקסט.

המחברים הציעו להחליף את חישוב הסופטמקס הרגיל שיש לנו בטרנספורמרים בהפרש משוקלל (רק הסופטמקס השני משוקלל) של הסופטמקסים. כל סופטמקס מחושב עם מטריצת Q ו-K משלה כאשר המשקול λ של הסופטמקס השני מחושב באופן הבא: λ = exp(λ\_q1 · λ\_k1 ) − exp(λ\_q2 · λ\_k2 ) + λ\_init כאשר

λ\_q1 , λ\_k1 , λ\_q2 , λ\_k2 ∈ R^d הינם נלמדים ו- ((λ\_init = 0.8 − 0.6 × exp(−0.3 · (l − 1, כאשר l זה מספר השכבה (של בלוק הטרנספורמר). אם הנוסחה עבור λ איכשהו מובנת ודי סטנדרטית הנוסחה עבור λ\_init נותרת בגדר תעלומה (אלא אם כן זה ניסוי ותהיה רגרסיה של הערכים שהתקבלו עם פונקציה מצורה מסוימת).

המאמר טוען לשיפור תוצאת אבל הבדיקות נעשו בעיקר למודלים עם 3B פרמטרים. יש גם טענות לקנסול של רעש כלשהו שאני לא בטוח שאני מבין. בקיצר אני קצת סקפטי, מודה….

https://arxiv.org/abs/2410.05258