# Review 100: Fastformer: Additive attention is Can Be All you need

**Paper: https://arxiv.org/abs/2108.09084v6**

אתם אולי שמתם לב שיש לי חולשה למאמרים ששמם ניתן באמצעות הטמפלייט "X is not ("") is you need". הפעם המאמר שאני הולך לסקור (קצרות או #shortdeepnightlearners") לא מקיים את הטפלייט הזה במדויק אבל עדיין נמצא בסביבת אפסילון ממנו לאפסילון די קטן.

המאמר נקרא:

Fastformer: Additive attention is Can Be All you need

מטרת המאמר היא להתמודד עם הסיבוכיות הריבועית של הטרנספורמר (מבחינת אורך הקלט). המחברים עשו את זה בדרך פשוטה להפליא:

לוקטור נלמד w (אחד!!) מחשבים את מקדמי ה-attention עם וקטורי query. כלומר לכל ייצוג טוקן x\_i מחשבים את וקטור ה-query שלו בצורה הסטנדרטית (מכפלה במטריצה W\_q). לאחר מכן מחשבים את מקדמי ה-attention (סקלריים!!) הצורה הרגילה עם סופטמקס.

מחשבים את וקטור הקונטקסט הגלובלי c\_q (אחד!!) שזה הסכום הממושקל של וקטורי query עם מקדמי ה-attention שחושבו בסעיף הקודם.

מכפילים וקטורי ה-key k\_j בווקטור c\_q איבר-איבר (element wise) ומקבלים וקטורי K\_cj

וקטורי K\_cj מוכפלים איבר איבר בוקטורי v\_j value, מכפילים את התוצאה במטריצת משקלים נלמדת ומקבלים וקטורי r\_j

הפתעה: לווקטורי r\_j מחברים (למה - כי כנראה זה עבד ) את ווקטורי ה- query שמקבלים את הייצוג הסופי של FastFormer.

כמו ששמתם לב אין פה רכיב attention אמיתי - בשום שלב הייצוג של טוקן לא תלוי באופן מפורש בקרבתו עם כל טוקו אחר אלא רק עם ייצוג גלובלי של כל סדרת הקלט. ייצוג גלובלי של הפלט גם לא תלוי באופן מפורש בקרבה בין זוגות הטוקנים (יש attention עם וקטור נלמד אחד ומה שנמדד זו קרבה של ייצוגי הטוקנים אליו).

המאמר מראה שיפור בביצועים בכמה משימות כמו

זיהוי רייטינג של סרטים

זיהוי נושא

Text summarization

כמעט בכל המשימות Fastformer הצליח להכות את הטרנספורמר המקורי (בקצת) וכמה וריאנטים של הטרנספורמר כמו LinFormer ו- LongFormer.

שימו לב שתוצאות אלו הושגו ללא שימוש במנגנון attention באופן מפורש שזה קצת מפתיע. הייתי מאוד רוצה לראות איך Fastformer עובד על סוגי משימות נוספים ועל דאטהסטים אחרים.

מאמר: https://arxiv.org/abs/2108.09084