המאמר היומי של מייק - 28.02.25  
The FFT Strikes Back: An Efficient Alternative to Self-Attention

מי עוקב אחרי מספיק זמן בוודאי יודע שיש לי חולשה למאמרים שמופיע בהם התמרת פוריה או כל התמרה אחרת (כמו התמרת קוסיין DCT). הסיבה לחיבה זו היא 5 שנים שביליתי בתור חוקר, מהנדס אלגוריתמים ומרצה בתחום עיבוד אותות (מערכות תקשורת אלחוטיות). המאמר הזה מתיימר להציע מנגנון שהוא מחליף את מנגנון attention בשכבה המבצעת טרנספורמציה של ייצוגי טוקנים בתחום התדר (כלומר התמרת פוריה). הטענה במאמר שהיא בעלת expressiveness (מסוגלת למדל את אותן הפונקציות) קרובה לזו של הטרנספורמרים. הטענות מוכחות בצורה חצי תיאורטית (הוכחה מלאה לא הוצגה במאמר).

אבל מה היתרון של הגישה המוצעות המבוססת על טרנספורמציה לא לינאריות במרחב התדר? כמובן סיבוכיות יותר נמוכה מהסדר (O(NlogN המשמעותית יותר נמוכה מ-(O(N^2 הסיבוכיות התיאורטית של מנגנון ה-attention. כאן N מסמן את אורך הסדרה. ידוע שניתן לעשות התמרת פורייה עם הסיבוכיות (O(NlogN ולמרות שהמאמר מכניס אי לינאריות בטרנספורמציות מעל מרחב התדר עדיין הסיבוכיות של המנגנון המוצע נותרת (O(NlogN.

אז איך כל הסיפור הזה עובד? קודם כל מעבירים את כל הטוקנים דרך התמרת פוריה כאשר כל מימד ייצוג הטוקנים עובר FFT בנפרד. כלומר אם יש לנו 10K טוקנים שכל אחד מהם מיוצג על ידי וקטור באורך 1024 יש לנו 1024 התמרות פוריה כל אחת באורך של 10K. לאחר מכן מחשבים את הממוצע של כל הייצוגים בתחום התדר ומעבירים את התוצאה דרך MLP (כלומר כמה שכבות fully-connected) כאשר הפלט שלו הינו בגודל המקורי של הסדרה (בדוגמא זה יהיה 10Kx1024). לאחר מכן מחברים את התוצאה למטריצה W\_base שכולה מורכבת מאחדות.

בשלב הבא מכפילים (איבר-איבר) את התוצאה עם התוצאה הראשונית של התמרת פוריה. כלומר מה שיש לנו כאן הוא משקול מחדש של התמרת פורייה של ייצוגי הטוקנים כאשר המשקלות מחושבות בצורה לא לינארית. לבסוף עושים לתוצאה של של ReLu למספרים מרוכבים(modReLU) ומעבירים חזרה למרחב המקורי עם ה-IFFT.

והתוצאות כמובן לא רעות בכלל…

https://arxiv.org/abs/2502.18394