המאמר היומי של מייק - 20.03.25  
softmax is not enough (for sharp out-of-distribution)

המאמר הזה מציעה שיטה לשיפור ביצועי ההכללה עבור מודלי טרנספורמרים מזווית די לא צפויה. המחברים מציעים שיטה להתמודדות עם מה שנקרא דיספרסיה (או פיזור בעברית) של מקדמים ה-attention בטרנספורמרים. זה מתבטא למשל באי יכולת (לפי המאמר) של הטרנספורמרים למקד את מקדמי ה-attention במספר טוקנים קטן (יחסית לאורך הסדרה). זה חשוב למשל בשאלות כמו מציאת מקסימומים של סדרת מספרים נתונה או שאלות בסגנון ״מחט בערימת השחת״ (needle in a haystack) כאשר המודל מתבקש מקטע קצר לא קשור בטקסט מסוים (יחסית ארוך).

המחברים טוענים שאחת הסיבות לבעיות אלו היא פיזור מקדמי ה-attention במנגנון הטרנספורמרים. מקדמים אלו מחושבים עם פונקצית סופטמקס ה״מנרמלת״ את המכפלות הפנימיות של וקטורי K ו-Q עבור כל טוקני הסדרה. לפי המאמר הבעיה קשורה לכך שעבור קונטקסטים ארוכים לסופטמקס במיוחד בטרנספורמרים העמוקים יש ״נטיה למרוח את פלט הסופטמקס״.

אחת הדרכים להתמודד עם התופעה הזו היא להוריד את הטמפרטורה אבל זה עלול להעלות סיכוי לשגיאה במקרים בהם הלוגיט (משקל attention לא מנורמל) של הטוקן הנכון יותר קטן מהלוגית המקסימלי. כדי להתמודד עם התופעה המבחרים הציעו גרסה חדשה של סופטמקס בה הטמפרטורה תלויה באנטרופיה של הטוקנים.

הם אימנו מודל עבור מקרים שבהם הלוגיט של הטוקן הנכון אינו מקסימלי כאשר המטרה היתה למקסם את הסתברות הדגימה של הטוקן הנכון (אחרי מנגנון ה-attention ו-FFN). מטרת המודל היתה לחשב ערך אופטימלי של טמפרטורה כפונקצייה של אנטרופיית של משקלי attention לא מנורמלים. הנוסחה של הטמפרטורה יצאה הופכית (1 חלקי) של פולימום מחזקה 4. אציין כי הטמפרטורה מחושבת בזמן האינפרנס כתלות באינטרופיית הטוקנים לפי המודל הזה.

המחברים הראו אמפירית כי עם הטמרטורה האדפטיבית מקטינה פיזור משקלי ה-attention. למרות שהטמפרטורה האדפטיבית האופטימלית יורדת עם עלייה באנטרופיית הלוגיטים היא גורמת לפחות שגיאות של המודל יחסית למקרה שהיא נקבעת באופן קשיח.

https://arxiv.org/abs/2410.01104