Improving LLM Final Representations with Inter-Layer Geometry

הרעיון המרכזי של המאמר

המאמר יוצא נגד ההנחה המקובלת ב - LLMs, שבה משתמשים רק ב - final layer representation לצורך prediction.

החוקרים טוענים שמידע חשוב עבור semantic understanding ו – reasoning מפוזר לאורך שכבות שונות במודל, ולכן שימוש רק בשכבה האחרונה גורם לאיבוד מידע.

מה המאמר מציע?

במקום לבחור רק final layer או רק ב - best layer, המאמר מציע לאפשר לכל שכבות ה - LLM "לתקשר" ביניהן, באמצעות Graph Neural Network (GNN).

כל Layer במודל הופך ל - Node בגרף, וה - GNN לומד כיצד לשלב מידע מכל השכבות ל - representation סופי טוב יותר.

הגישה הטובה ביותר במאמר

הגישה שהשיגה את הביצועים הטובים ביותר הייתה **Cayley-Encoder**.

במקום לחבר כל שכבה לכל שכבה אחרת (Fully Connected, שגם נבדקה), המודל משתמש ב - Cayley Graph - מבנה מתמטי sparse ויעיל, שמאפשר communication מהיר בין שכבות, בלי העלות החישובית הגדולה של Fully Connected Graph.

המאמר הראה ש - Cayley-Encoder השיג תוצאות טובות יותר כמעט בכל המשימות, עם פחות זיכרון, פחות פרמטרים, ופחות סיכון ל - overfitting.