⚡️🚀המאמר היומי של מייק 08.07.24: ⚡️🚀

Mixture of A Million Experts

המאמר של היום מציע לקחת את שיטת (Mixture of Experts(MoE לבניית ארכיטקטורות של מודלים עמוקים פופולרית במיוחד במודלי שפה. מאוד בגדול ב- MoE הרשת מורכבת מתת-רשתות (בד״כ מחלקים את שכבת ה-FFN של הטרנספורמר לכמה חלקים זרים). MoE מאומן להשתמש כל בפעם בחלק מתת-רשתות אלו (הנקראות מומחים) כאשר רשת gating רדודה יחסית באיזה מומחים צריך להשתמש כל פעם. כלומר יש לנו כן סוג של מימוש הגישה שנקראת ״lottery ticket hypothesis" דינמי כאשר כל פעם בוחרים להריץ רק חלק מהרשת.

כנראה שככל יש ברשת יותר מומחים בעלי אותה הארכיטקטורה וכל פעם בוחרים אותו מספר של המומחים הביצועים אמורים להשתפר אולם המחיר הוא המודל גדול יותר.המאמר מנסה לבדוק האם שווה להשתמש בהרבה מאוד במומחים רזים מאוד. המחרים מציעים לעבוד עם מיליון של מומחים של כל אחד מהם היא דל במיוחד. כמובן שכל פעם צריך לבחון את המומחים כל פעם ומכיוון שיש מיליון מומחים אז נדרש מאמץ חישובי לא קטן. המאמר מציע להשתמש בטכניקה הנקראת product key retrieval כדי להקטין את הסיבוכיות (בגדול זה חלוקה של וקטור המפתחות (keys) לשני חלקים, ביצוע חישוב לכל אחד בנפרד ושילובם).

וגיליתי משהו מעניין במאמר הזה - יש scaling law גם ל-MoEs. אולי אסקור אותו בקרוב…

https://arxiv.org/abs/2407.04153