⚡️🚀המאמר היומי של מייק 03.07.24:⚡️🚀

The Remarkable Robustness of LLMs: Stages of Inference?

מאמר מעניין החוקר איזה שכבות ניתן לזרוק ממודל השפה ועדיין לשמור על ביצועים נאותים. אתם אולי מכירים lottery ticket hypothesis הטוען כי ברשתות עתירות פרמטרים (overparameterized) בד:כ ניתן למצוא קטנה הרבה יותר עם ביצועים מאוד קרובים אך הבעיה שאנו לא יודעים לאתר אותה.

המאמר כאמור בחן איזה שכבות הן סוג של מיותרות במודלי שפה והגיע לתופעות מעניינות לגבי תהליך האינפרנס שלהם. הם זיהו 4 שלבים עיקריים

דה-טוקניזציה או רכישה התחלתית של קשרים קונטקסטואליים: טרנספורמציה ראשונית של ייצוג ה-raw (מהמילון) של הטוקנים לייצוג תלוי הקשר (חישובי attention כבדים לכל אורך הקונטקסט).

הנדסת פיצ'רים התחלתיים מהייצוגים תלוי הקשר מהשלב הקודם ו״הכנת קרקע״ לחיזוי של הטוקנים הבאים. עדיין לא ניתן לחזות את הטוקנים האלו מהפיצ'רים בשלב הזה אבל המודל מתחיל ״להבין הקשרים מרחבים ועתיים בטקסט (היה מחקר מעניין הזה)

בניית קבוצות נוירונים (אנסמבל) לחיזוי הטוקן הבא. בשלב הזה הרשת מתחילה להתכנס ולבנות קבוצות ״prediction neurons" שישולבו יחד למטרת חיזוי הטוקן הבא.

חידוד של prediction neurons: הרשת ״בוחרת״ את הנוירונים החשובים ביותר לחיזוי הטוקן הבא על ידי הדעכה של חלק מה-prediction neurons מהשלב הקודם.

והכי חשוב שהשכבות מעורבות בשלב 1 ובשלב 4 הם הכי חשובות לביצוע המודל כאשר חלק מהשכבות של שלב 2 ו-3 ניתן להסיר ללא פגיעה משמעותית בביצועים.

הרבה טענות מעניינות במאמר הזה (חלקם הגדול זה סיכום של העבודות הקודמות בנושא הזה).

https://arxiv.org/abs/2406.19384