המאמר היומי של מייק - 09.03.25  
THE SUPER WEIGHT IN LARGE LANGUAGE MODELS

זה די לא ייאמן, אבל מודלים שפה גדולים עם מיליארדי או אפילו עשרות מיליארדי פרמטרים עלולים לסבול ירידה כואבת בביצועים אם מורידים מהם אפילו משקל בודד. ממצא מפתיע זה חל לפחות על חלק מהמודלים העוצמתיים האלה.

מאמר זה מתעמק במאפיין ספציפי ובלתי צפוי של מודלים שפה גדולים: קיומם של "משקלים על (SWs)". המחברים מתקדמים מעבר לתצפית ידועה על כך ש-LLMs מכילים משקלים חריגים המשפיעים באופן ניכר על הביצועים, ומציגים ראיות לכך שמשקל בודד יכול להיות קריטי באופן לא פרופורציונלי לתפקוד המודל.

כאמור הממצא המרכזי הוא שהורדת SW בודד יכול לגרום לירידה קשה בביצועי LLM. השפעה דרסטית זו מתבטאת כעלייה חדה בפרפלקסיטי וירידה בדיוק zero-shot לרמות כמעט אקראיות. מה שראוי לציון במיוחד הוא העובדה שהסרת SW לבין ההשפעה הקטנה יחסית של הורדה של משקלים חריגים אחרים, אפילו בעלי גודל גדול יותר.

המאמר נותן דוגמה מעניינת להשפעה של הסרת משקל על-כבד עבור הפרומפט: "קיץ חם. חורף הוא…"(באנגלית). הטוקן הבא הנכון צריך להיות "קר" ועם המודל המקורי עם SW, הוא חוזה נכון את הטוקן הבא "קר" בהסתברות גבוהה של 81.4%. כאשר SW מוסר, החיזוי המוביל של המודל הוא stopword "ה"(the) בהסתברות נמוכה ולא בטוחה של 9.0%. זה מצביע על כך ש-SW חיוני למודל כדי לבצע חיזוי נכון ובטוח של מילים משמעותיות. המאמר לא רק מתעד את התופעה הזו; הוא גם בוחן את המנגנונים הבסיסיים הקשורים אליה. המחברים מקשרים SW ליצירת "אקטיבציות SW", שהן אקטיבציה גדולות וחריגות המתפשטות דרך המודל כמעט ללא קשר לקלט.

יתר על כן, המחקר בוחן את ההשלכות של SW עבור קוונטיזציה של מודלי שפה. נוכחותם של חריגים, כולל SW ואקטיבציות חריגות הנגרמות מהם, מציבה אתגר משמעותי לקוונטיזציה יעילה, שכן חריגים אלה יכולים לעוות את תהליך הקוונטיזציה ולהוביל לאובדן מידע משמעותי. המחברים מדגימים ששימור חריגי SW (גם משקלים וגם אקטיבציות) יכול לשפר את יעילות הקוונטיזציה מסוג "עיגול לערך הקרוב ביותר", אפילו לאפשר שימוש בגדלים גדולים יותר של בלוקים בקוונטיזציה (עבורם מחושבים קבועי קוונטיזציה). זה מושג על ידי השארת SW מחוץ לתהליך הקוונטיזציה ושחזור ערכיהם לאחר מכן, תוך צמצום ההשפעות השליליות של ערכים קיצוניים אלה על קוונטיזציה של פרמטרים אחרים. על ידי התמודדות עם האתגרים שמציבים חריגי על-כבד, הגישה המוצעת מאפשרת יישום של שיטות קוונטיזציה פשוטות ויעילות יותר, ומקלה על פריסת מודלים בסביבות עם משאבים מוגבלים.

עבודה זו יוצרת טיעון חזק ש-SW אינם רק אנומליות מבודדות אלא רכיבים אינטגרליים הממלאים תפקיד חיוני בעיצוב ההתנהגות והיעילות של LLMs, עם השלכות משמעותיות לדחיסה ולאינפרנס של מודלים. תרומת המאמר אינה טמונה רק בזיהוי SW אלא גם באפיון תפקידם הפונקציונלי בתוך LLMs. המחברים מנתחים כיצד משקלים משפיעים על פלט המודל, ומקשרים אותם ל״התפשטות״ של אקטיבציות חריגות.

https://arxiv.org/abs/2411.07191