המאמר היומי של מייק - 13.03.25  
SLIM: Let LLM Learn More and Forget Less with Soft LoRA and Identity Mixture

האם לפעמים קורה לכם שאתם מתחילים לקרוא את המאמר וככל שאתם מתקדמים ומתעמקים בו הוא מתחיל להיראות פחות ופחות טוב. לי זה לפעמים קורה עם אוכל אבל שם יותר קל לי להפסיק לאכול מאשר לקרוא מאמר. אז יאללה, אסקור אותו קצרות אך אל תצפו רבות…

המאמר מציע שיטה להלביש ערבוב של מומחים או MoE על LoRa. נזכיר ש-LoRa היא שיטת פיין טיון של רשתות נוירונים שבהם אנו לא מאמנים את כל משקולות המודל אלא רק מטריצות תוספות בעלת ראנק נמוך. MoE היא שיטה להורדה של העומס החישובי בטרנספורמרים כאשר אנו מחלקים את המטריצות בשכבת FFN של הטרנספורמרים לתת-מטריצות (מומחים) כאשר כל פעם לטוקן נתון אנו מפעילים רק חלק מהמומחים. שכבת ניתוב (routing layer) מחשבות את הציון של כל מומחים ובדרך כלל אנו בוחרים k מומחים בעלי ציון הגבוה ביותר (top-k).

אז המחברים משדכים LoRA עם MoE וזה בדיוק מה שמשך את עיניי. המאמר מציע להחליף LoRA רגיל עם כמה מומחי של LoRA שחלקם הינם מטריצות מראנק 0 או פשוט מטריצות אפסים. לטענת המאמר לא תמיד צריך להפעיל את LoRa. מומחי ה-LoRa נבחרים על ידי רשת ניתוב בדומה ל-MoE הסטנדרתי. עבור כל טוקן נבחרים K מומחים (בינם גם מומחי זהות) בעלי ציונים הגבוהים ביותר. שימו לב שבמאמר יש כמה שגיאות בנוסחאות המחשבים את התוצאה של המנגנון המוצע.

לאחר מכן המאמר מציע שיטה לשכלול הציונים של שכבת הניתוב בהתבסס על הסטטיסטיקות של הדאטהסט עליו בוצע הפיינטיון עם השיטה. סטטיסטיקה במקרה הזה מחושבת על המצבים החבויים של הרשת המחושבים על הדאטה של הפיין טיון (אופן החישוב המדויק לא מוגדר בצורה ברורה ולדעתי יש שגיאות בנוסחאות המגדירות אותו). המחברים מציעים לקלסטר את המצבים החבויים האלו לקלסטרים שמספרם כנראה שווה למספר הטוקנים בפרומפט (מוגדר כקבוע במאמר ועבור סדרות קצרות יותר משתמשים בטוקני ה-padding).

מרכזי הקלסטרים מתעדכנים במהלך הפיין טיון (כל פרומפט הקלט משויך לקלסטר הקרוב ביותר ואז מרכז הקלסטר מחושב מחדש). במהלך האינפרנס פרומפט הקלט משויך לקלסטר הקרוב ביותר (מרחק ריבוע) ואז ציוני המומחים המופקים על ידי שכבת הניתוב עבור מומחי הזהות מוזזים במקדם שעולה אם המרחק לקלסטר הקרוב עולה כאשר הציונים למומחי LoRA האחרים נותרים ללא שינוי. נציין שמרכזי הקלסטרים לא מתעדכנים במהלך האינפרנס.

לבסוף המאמר מציע דרך לשלב כמה MoE עם LoRa עבור כמה משימות פיין טיון שונות אבל אחרי שגיליתי טעיות גם בפרק הזה, ויתרתי….

https://arxiv.org/pdf/2410.07739