⚡️🚀המאמר היומי של מייק -15.10.24: ⚡️🚀  
EFFICIENT DICTIONARY LEARNING WITH SWITCH SPARSE AUTOENCODERS

היום סוקרים מאמר קליל המשלב שני רעיונות די נחמדים שמשמים LLMs (במיוחד לאחרונה) והאמת השילוב שלהם נראה די טבעי. הרעיון הראשון הינו Mixture of Experts או MoE בקצרה.

MoE היא שיטה המאפשרת לנו להקל על האינפרנס על ידי שימוש רק בחלק ממשקלי המודל. בד״כ מטריצות משקלים ברשת feed-forward (יש שם 2 שכבות בסך הכל) בבלוק הטרנספורמרים (אחרי attention) מחוקלים לכמה קבוצות שכל אחת מהן נקראת מומחה או expert. באינפרנס המודל משתמש רק בחלק (לפעמים רק אחד) מהמומחים ובכך הוא מוריד את מחירו של האינפרנס. כלומר אותו המודל מופעל בצורה קצת שונה בהתאם לקלט (בנוסף ל-attention),

הקונספט השני הוא Sparse AutoEncoders או SAE בקצרה שהפך להיות די פופולרי אחרי החוקרים של אנטרופיק הציעו להשתמש בו למטרת חקר interpretability של מודלי שפה. לפני הבלוג הזה הסברה הרווחת (סוג של) היתה שבמודל שפה יש נוירונים שנדלקים חזק (מקבלים ערך גבוה) על קונספטים מסוימים כאשר כל נוירון כזה הינו מונו-סמנטי כלומר יש קונספט אחד בלבד שהוא ״אחראי״ עליו.

לעומת זאת החוקרים של אנטרופיק הציע להתבונן בכל נוירון כפולי-סמנטי כלומר ״אחראי״ על מספר קונספטים לא קשורים. לפי משנתם ניתן לגלות את הקונספטים האלו באמצעות SAE שבונה autoencoder דליל (הרוב אפסים) במימד גבוה הרבה יותר מגודל השכבה שבה נמצאים הנוירונים הפוליסמנטיים אלו. SAE כאן מורכב משתי שכבות בלבד, אחת לאנקודר ואחת לדקודר.

כאן כל רכיב שהוא לא אפס בווקטור אחרי שכבת ה-encoder של SAE הוא אחראי על קונספט מסוים כלומר מהווה נוירון מונוסמנטי. כך יוצא שכל נוירון בשכבה המקורית הוא שילוב לינארים של הנוירונים המונוסמנטיים אלו. SAE מאומן בצורה די סטנדרטית עם איבר רגולריזציה שאוכף את דלילות הייצוג אחרי האנקודר.

אז המאמר מציע לשלב את שני הקונספטים האלו כך שכל נוירון הוא צירוף לינארי אחר של הנוירונים המונוסמנטיים בשכבת ה-encoder. זה מאפשר גמישות נוספת ביחס לרעיון המקורי ובטח מאפשר לגלות קונספטים שונים המוסתרים בתוך ה-LLMs שלנו.

מאמר קליל - ממליץ להעיף מבט

https://arxiv.org/abs/2410.08201