⚡️🚀המאמר היומי של מייק 04.09.24: ⚡️🚀

Flexora: Flexible Low Rank Adaptation for Large Language Models

המאמר הזה נסקר קודם כל בגלל שהוא למעשה מימוש של רעיון שחשבתי עליו והוא גם רשום לי בבקלוג (שהוא באורך די אינסופי). הרעיון הוא למעשה שיטה לבחירה (לפעמים קוראים לזה אופטימיזציה) של ההייפרפרמטרים של LoRA (סוג של).

כמו שאתם בטח זוכרים LoRA היא משפחה (די גדולה שממשיכה לגדול) של שיטות מהמשפחה (גדולה עוד יותר) של שיטות חסכוניות פיינטיון של מודלי שפה ענקיים (או PEFT - Parameter Efficient Fine-Tuning). C ב-LoRA אנו מאמנים תוספת של משקלים לכל שכבה במקום לאמן את כל המשקלים במודל. כל תוספת כזו היא מטריצה בעלת רנק נמוך כלומר אפקטיבית מכילה מעט פרמטרים מאשר מטריצת המשקלים של השכבה.

פרקטית כל תוספת היא מכפלה של שתי מטריצות בעלות רנק נמוך (מלבניות) וככל הרנק נמוך יותר יש לנו פחות פרמטרים לאפטם במהלך פיינטיון. הבחירה של הרנק של מטריצות התוספות הנדרשת למקסום ביצועים איננה בעיה פשוטה ויש מספר מאמרים שדנים בנושא הזה (בד״כ עד רנק מסוים הביצועים משתפרים ומנקודה מסוימת מתחיל אוורפיט).

המאמר (וגם אני) חשבו על דרך אחרת של אופטימיזציה של LoRA. המחברים שואלים שאלה פשוטה - למה בנוסף לאימון של מטריצות התוספות לא נאמן את ה-importance שלה בכל שכבה. ה-importance במקרה הזה היא המקדם המכפיל את מטריצת התוספות לפני הוספתה מטריצת המשקלות המקורית במודל (שנותרת קבועה במהלך פיינטיון). האלגוריתם המוצע עושה כמה איטרציות של משקלי ה-importance לעדכון אחד של משקלות התוספות.

האמת שהרעיון שלי הכיל עוד שלב של pruning. כלומר אחרי מספר של איטרציות אימון מתחילים לאפס ומפסיקים לאמן מטריצות התופסות עם importances נמוכים מאיזה סף. כנראה שאצטרך לבדוק את זה לבדי :)

https://arxiv.org/abs/2408.10774