המאמר היומי של מייק - 13.02.25  
One Initialization to Rule them All: Fine-tuning via Explained Variance Adaptation

היום נסקור קצרות מאמר המציע שיטת LoRa לשיפור של טכניקת טיוב (fine-tuning) של LLMS. כמו שאתם בטח זוכרים LoRA מוסיפה למשקלי המודל (בשכבות מסוימות) מטריצה נלמדת בעלת ראנק משמעותית נמוך יותר מהמימד של מטריצת המשקולות. משקולות המודל נשארות קבועות (לא מאומנות) במהלך הטיוב.

המחברים כאומר מציע גישה לשכלול של LoRA המכיל שלב מקדים שנקרא במאמר אתחול Date-Driven. מטרת אתחול זה היא ״להתאים את הראנק של מטריצות של LoRA לכל שכבה של המודל״. הרי אם אנו מאמנים תוספת משקלים מסוימת (במהלך אנו יכולים לפזר אותם בצורה ״אופטימלית״ בין שכבות המודל. האופטימליות כאן נמדדת באמצעות השונות של האקטיבציות של השכבה (כלומר הפלט של שכבת FFN) עבור הדאטה שאנו מאמנים עליו.

הרי אם שונות האקטיבציות על דאטה האימון היא נמוכה זה אומר שערכי השכבה פחות או יותר קבועים ולא כדאי לבזבז עליה את המשקלים של LoRA. כלומר אפשר להשתמש ב-LoRA בעלת ראנק נמוך מאוד (אם בכלל) לשכבה זו. אבל איך ניתן למדוד את השונות הזו באמצעות ערכים סינגולריים של מטריצות האקטיציות המחושבים באמצעות פירוק SVD של מטריצה זו. מימדים מטריצת האקטיבציה כאן היא המימד החבוי של המודל וגודל הבאץ'.

אז מחשבים את הערכים הסינגולריים של מטריצת האקטיבציות על דאטהסט האימון עד שהוקטורים הסינגולריים (הימניים מתיצבים). וקטורים אלו מתעדכנים במהלך הרצות הבאצ'ים(המודל מתאמן) ויצירתם(של הוקטורים) נעצרת כאשר הם מתייצבים ומפסיקים להשתנות באופן מהותי (המאמר מודד את הדמיון באמצעות מרחק קוסיין - אם הוא גבוה מדי עבור שכבה מסוימת מפסיקים את עדכון הוקטורים עבור שכבה זו(אימון זה המתבצע לפני LoRA).

לאחר שהוקטורים הסינגולריים התכנסו עבור כל השכבות, לוקחים את הערכים העצמיים ומחשבים את אחוז השונות המוסבר על ידי כל שכבה (מחושב על ידי סכום הריבועים של הערכים הסינגולריים שלהם) ביחס לשונות המוסברת על ידי כל המודל (שהיא סכום הריבועים של הערכים הסינגולריים עבור אקטיבציות של כל שכבות המודל).

בשלב הבא מקצים את הראנקים של מטריצות LoRA לשכבות שפונקציות של השונות המוסברת על ידי. כלומר ככל השונות המוסברת של שכבה עולה, מקצים יותר ראנקים שלי LoRa. בשלב האחרון מאמנים LoRa עם הקצאה ״אוםטימלית״ של ראנקי מטריצות LoRA בהתבסס על דאטה האימון. רעיון די מעניין שמראה תוצאות לא רעות.

https://arxiv.org/abs/2410.07170