הסבר כללי על טכניקת QLoRA ותאימותה לפרוייקט

הסבר זה הינו הסבר מקוצר על טכניקת QLoRA ובונה על גבי ההסבר על טכניקת LoRA בקובץ הקודם.

מה שטכניקה זו מוסיפה על גבי טכניקת LoRA זה את האות Q כלומר את ה-Quantized, המטרה של טכניקה זו היא לעשות את אימון המודלים מחדש אפילו יותר יעילה מאשר שיטת LoRA הרגילה.  
הבסיס של שיטה זו היא שיטת LoRA, בעיקר זה אומר להשתמש ב4 ביטים. מודלי שפה גדולים משתמשים ב32, שיטת LoRA הופכת את זה ל8 ושיטת QLoRA הופכת את זה ל4.  
מה שגורם לכך שאפשר לאמן את המודל על GPU אחד בלבד לדוגמא אפילו למודלי שפה עם מילארדים של פרמטרים.

השיטה QLoRA משתמשת במה שנקרא Paged Optimizers על מנת לטפל בקפיצות של השימוש בזיכרון שיכולות לקרות בזמן אימון, הoptimizers האלו משתמשים בgradient בחתיכות (chunks) קטנות יותר מה שעוזר למנוע בעיות זיכרון שעלולות לקרות בזמן האימונים ומאפשר אימון חלק יותר על GPU בודד אפילו.

הייתרונות העיקריים של QLoRA הינם:

* נכון לכיום זאת השיטה הכי יעילה לאימון מודלי שפה גדולים מבחינת זיכרון
* מהירה בצורה ניכרת מאימון מודלים רגיל, אך עלולה להיות קצת יותר איטית מLoRA
* באופן כללי היעילות של שיטה זו הינה בין הטובות שיש כרגע

ביחד עם כול היתרונות של שיטה זו, כן בא עם זה סיבוכיות קצת יותר מLoRA רגילה, שכן למצוא את רמת ה-quantization הנכונה זה מאוד קריטי שכן מידה לא נכונה עלולה לפגוע משמעותית בתוצאות של המודל. אחת הדרכים שניתן להתמודד עם זה זה להשתמש במה שנקרא NF4 (4-bit NormalFloat, טכניקת quantization שנחשבת מאוד "נאמנה").