Faster Convergence for Transformer Fine-tuning with Line Search Methods

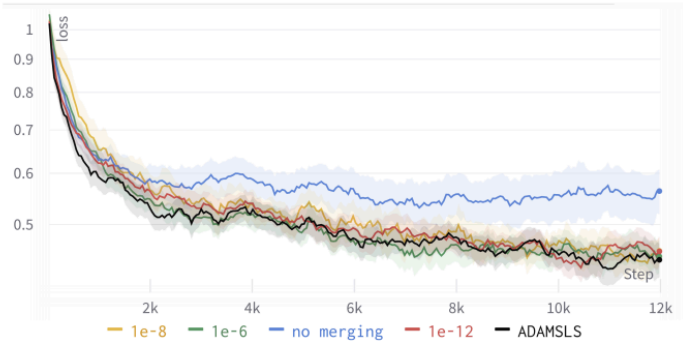

Fig. 2. Different merging thresholds on one epoch of the MNLI dataset. Standard error is indicated around each line.

מזמן לא כתבתי סקירה ונתקלתי במאמר חמוד הדן במשפחת גישות לאימון רשתות נוירונים שלא הייתי מודע לקיומה. אתם בטח יודעים שהיום מאמנים רשתות נוירונים ומודלי ML אחרים (=ממזערים את פונקציית הלוס שלהן). כמובן עם כל מיני שכלולים של גישת מורד הגרדיאנט הסטוכסטי (או Stochastic Gradient Descent או SGD בקצרה).

בגדול SGD היא שיטה איטרטיבית בכל איטרציה המשקלים של המודל מוזזים בכיוון של הגרדיאנט השלילי שהוא הכיוון (הלינארי) שבו פונקצית לוס קטנה "הכי הרבה". כאמור קיימים לא מעט שפצורים של SGD כמו ADAM, RMSProp ושיטות רבות נוספות המערבות צבירה גרדיאנט (מומנטום) שמטרתם היא להאיץ את קצב ההתכנסות של SGD ולהפוך אותו ליותר יציב. נזכיר שבכל השיטות האלו בכל איטרציה מעדכנים את המשקלים על סמך מיני-באץ' ולא דוגמא אחת כמו ב-SGD קלאסי.

המאמר שנסקור היום מציע גישה אחרת (מהירה יותר לטענת המחברים) למזעור של פונקציית לוס עבור מודלי מבוססי טרנספורמרים. קודם כל נציין כי ערך של פונקצית לוס לא בהכרח יורד (על מיני-באטץ' של איטרציה זו) אחרי עדכון של פרמטרי המודל באיטרציה של כל שיטה מבוססת מורד הגרדיאנט.

לפעמים הלוס על מיני-באטץ' עשוי לעלות אחרי העדכון גם אם אתם משתמשים בשיטות מתקדמות כמו ADAM או RMSProp. עבור SGD (כלומר MiniBatch GD) זה נובע בגדול מכך שקצב למידה (learning rate) גדול מדי. בשיטות עם מומנטום כמו ADAM המצב מורכב יותר (כי הכיוון שבו מזיזים את המשקולות הוא לא הגרדיאנט הממוצע של מיני-באטץ') אך הבעיה עדיין קיימת.

חשוב להבין שעלייה זמנית של פונקציית לוס עבור מיני-באצ'ים פה ושם היא "לא אסון" אם המגמה הכללית של ירידת לוס נשמרת במהלך האימון. אבל נשאלת השאלה האם שיטת אימון שתבטיח אי עלייה של פונקציית לוס בכל איטרציה תוביל לאימון יעיל ומהיר יותר בלי לפגוע באיכות המודל המתקבל בסוף האימון. זו השאלה שמחברים המאמר מנסים לענות עליה.

השיטה הנדונה במאמר מציעה דרך מאוד אינטואיטיבית לעקוף את הסוגיה הזו. כאמור אי ירידת של ערך פונקצית לוס ב-SGD נובעת מקצב למידה גדול מדי. המאמר מציע לבחור את קצב הלמידה כך שיבטיח ירידה של פונקצית לוס עבור כל המיני-באץ' בכל איטרציה של אימון.

בגדול בכל איטרצית אימון מתחילים בקצב למידה אקראי ומקטינים אותו (נגיד מחלקים ב 2) עד שהלוס אחרי העדכון יירד (על המיני-באץ'). ניתן לעשות את זה גם עם שיטות מתקדמות שנזכרו לעיל השלב האחרון הוא הזזה של משקלים בכיוון מסוים עם מקדם מסוים (קצב למידה). השיטה הזו נקראת Armijo Line Search או ALS.

המאמר מציע להפעיל את ALS על כל שכבה (בלוק של טרנספורמר) בנפרד. כלומר מחלקים משקלים של המודל ל-L קבוצות כאשר L הוא מספר השכבות ברשת. לאחר מכן מבצעים עדכון של המשקלים לכל שכבה בנפרד עם שיטת אופטימיזציה שבחרתם (SGD, ADAM etc) משולבת עם ALS. כלומר מורידים קצב למידה לכל שכבה בנפרד עד שהערך של פונקצית לוס אחרי העדכון יקטן כאשר שאר המשקלות מוקפאות.

לדעתי עם השיטה המוצעת האימון ייקח יותר זמן (כי מעדכנים כל שכבה בנפרד) אבל ניתן לעבוד עם באצ'ים גדולים יותר שזה תורם ליציבות תהליך האימון. השיטה מראה תוצאות לא רעות על פיין טיון של הטרנספורמרים.