המאמר היומי של מייק: 11.09.25

Fantastic Pretraining Optimizers and Where to Find Them

סקירה 502

במשך שנים, עולם אימון הקדם (pretraining) של מודלים גדולים נשלט על ידי AdamW. זהו סוס העבודה האמין והמובן היטב שעומד מאחורי כמעט כל מודל תשתית מרכזי. עם זאת, הממלכה נתונה במצור תמידי מצד טוענים לכתר. כמעט מדי חודש מופיע מאמר חדש המבשר על אופטימייזר "פנטסטי", Sophia, Muon, Mars, שטוען שהוא מהיר פי שניים, ומבטיח לקצץ את עלויות האימון האדירות בחצי.

הדבר מעלה שאלה של מיליוני $, שהדירה שינה מעיניהם של אנשי מקצוע בתחום: אם השיטות החדשות הללו כל כך מהפכניות, מדוע כמעט אף אחד לא משתמש בהן בריצות האימון היקרות שלו?. המאמר שאסקור היום עונה על שאלה זו. מסקנת המאמר היא כמו מקלחת צוננת עבור אלו הניזונים מהייפ, ושיעור מעמיק במתודולוגיה מדעית: שיפורי המהירות הפנטסטיים הם במידה רבה פנטזיה, שנולדה מהשוואות פגומות ולא הוגנות.

### שני החטאים של בחינת אופטימייזרים

המחברים מאבחנים את הבלבול בתחום על ידי זיהוי שני ליקויים מתודולוגיים בסיסיים, אשר ניפחו באופן שיטתי את ביצועי האופטימייזרים החדשים תוך שהם פוגעים באלוף המכהן, AdamW.

#### 1. החטא של "כוונון" לא הוגן

רוב המאמרים המציגים אופטימייזר חדש חוטאים בחטא הנקרא: העברה עצלה של היפר-פרמטרים. הם לוקחים סט סטנדרטי של היפר-פרמטרים עבור AdamW (קצב למידה, דעיכת משקולות וכו'), לעיתים קרובות ממתכון בן כמה שנים, ואז מכווננים בקפידה את השיטה המוצעת שלהם כדי שתגבור על בסיס ההשוואה (baseline) הסטטי והלא מכוונן הזה. זה כמו להריץ מכונית פורמולה 1 מכווננת היטב נגד מכונית סדאן מהמפעל ולהכריז על ניצחון. המחברים חושפים את הכשל הלוגי הזה בניסוי פשוט והרסני. הם לקחו בסיס השוואה נפוץ של AdamW מעבודות קודמות וכיווננו היפר-פרמטר אחד בלבד, קצב הלמידה. התוצאה?

האצה של פי 2 על פני בסיס ההשוואה עצמו, מה שמחק לחלוטין את היתרון לכאורה של האופטימייזרים "החדשניים" שאליהם הושווה. יתרה מכך, הם מראים שהיפר-פרמטרים אופטימליים אינם ניתנים להחלפה; דעיכת המשקולות האידיאלית עבור Lion, למשל, שונה באופן קיצוני מזו של AdamW, מה שהופך השוואות עם היפר-פרמטרים קבועים לבלתי סבירות מיסודן.

#### 2. החטא של הערכה קצרת-רואי

החטא השני הוא לשפוט מרתון לפי 100 המטרים הראשונים. מחקרים רבים מכריזים על ניצחון על סמך המהירות שבה עקומת הלוס של אופטימייזר צונחת בשלבים המוקדמים של האימון. המחברים מראים שזה מטעה באופן מסוכן. מכיוון שאופטימייזרים שונים מגיבים ללוחות זמנים של קצב למידה בדרכים מורכבות, דירוגי הביצועים שלהם יכולים ממש להתהפך במהלך ריצת האימון המלאה. אופטימייזר שנראה עדיף ב-20K צעדים עשוי להגיע לרוויה (plateau) ולהיעקף על ידי הירידה היציבה והאמינה של אחר עד סוף הריצה. ההשוואה המשמעותית היחידה היא הלוס הסופי לאחר שנוצל תקציב האימון המלא.

### היכן באמת מוצאים אופטימייזרים פנטסטיים

המאמר בונה תמונה חדשה וברורה יותר של נוף האופטימייזרים. התובנה העמוקה ביותר נובעת מהבחנה בסיסית, כמעט פיזיקלית, באופן פעולתם של אלגוריתמים אלה.

החלוקה המרכזית היא בין אופטימייזרים מבוססי-סקלר ומבוססי-מטריצה.

אופטימייזרים מבוססי-סקלר כמו AdamW, Nesterov AdamW ו-Lion הם פשוטים מבחינה רעיונית. הם מתייחסים לכל פרמטר במטריצות המשקולות העצומות של המודל כאל סוכן עצמאי. העדכון עבור כל משקולל הוא פעולה סקלרית, המחושבת על סמך היסטוריית הגרדיאנטים שלו (המומנטום מסדר ראשון ושני). זה כמו קהל עצום של מטיילים המנווטים על הר, כאשר כל אדם מסתכל רק על עד כמה תלול הקרקע שמתחת לרגליו כדי להחליט על צעדו הבא.

אופטימייזרים מבוססי-מטריצה כמו Muon, Soap ו-Kron הם "החיות הפנטסטיות" מהכותרת. אלגוריתמים אלו מתוחכמים יותר. הם אינם מתייחסים למשקולות כאל סקלרים עצמאיים; הם מבינים את המבנה המטריציוני המובנה של שכבות הרשת העצבית. במקום להחיל תיקון סקלרי פשוט, הם מיישמים preconditioner מבוסס-מטריצה, הם מכפילים את כל מטריצת הגרדיאנטים במטריצה אחרת.

במונחים פיזיקליים, זהו ההבדל בין לדעת מהו השיפוע תחת רגליך לבין הבנת העקמומיות של העמק כולו. ה-preconditioner המטריציוני מעצב מחדש את הגיאומטריה של בעיית האופטימיזציה עבור שכבה שלמה בבת אחת, ומוצא נתיב יעיל יותר באופן גלובלי לעבר המינימום. זו הסיבה שכאשר הם מכווננים כראוי ומושווים בהגינות, הם אכן מהירים יותר.

### המציאות המפכחת של הסקייל

אז, האם השיטות מבוססות-המטריצה הן אותו פתרון קסם המבטיח האצה של פי 2? הממצא הסופי והמכריע של המאמר הוא "לא" מהדהד. בעוד שאופטימייזרים מבוססי-מטריצה אכן מציגים ביצועים טובים יותר באופן עקבי מבני דודיהם מבוססי-הסקלר, שיפור המהירות צנוע בהרבה ממה שנטען. בנוסף יתרון זה דועך עם גודל המודל. עבור מודלים קטנים יותר (בסביבות 130 מיליון פרמטרים), אופטימייזרים כמו Muon ו-Soap מספקים שיפור מהירות מכובד של 1.3x-1.4x על פני AdamW מכוונן היטב. אך כאשר מטפסים למודלים של 1.2 מיליארד פרמטרים, היתרון הזה מתכווץ לכדי 1.1x זעום. בקנה המידה של מודלי חזית, מגמה זו מצביעה על כך שהיתרונות עשויים להפוך כמעט לזניחים.

הבחירה באופטימייזר הטוב ביותר תלויה גם במשטר האימון. בסביבות מוגבלות-דאטה (בערך כמו הכמות ה"אופטימלית של צ'ינצ'ילה"), Muon מנצח באופן עקבי. עם זאת, כאשר נכנסים למשטר של אימון-יתר עם יחס דאטה- גודל למודל גבוה (פי 8 מצ'ינצ'ילה או יותר), Soap ו-Kron תופסים את ההובלה, מה שמרמז שהטיפול שלהם במידע מסדר שני מועיל יותר באופקי אימון ארוכים יותר.

בסופו של דבר, המאמר אינו מספק לנו אלגוריתם קסם חדש. הוא מעניק לנו משהו יקר ערך בהרבה: מתודולוגיה לגילוי האמת. הוא מלמד אותנו שהתקדמות באופטימיזציה אינה עוסקת בהצהרות נוצצות על האצה של פי 2, אלא בכוונון קפדני, הערכה בקנה מידה, והבנת ההבדלים המבניים העמוקים בין אלגוריתמים. האופטימייזרים הפנטסטיים אמיתיים, אך הם אינם חיות מיתולוגיות; הם כלים מדויקים המציעים יתרונות צנועים, מותנים ותלויי-סקייל במשימה העצומה של בניית AI.

https://arxiv.org/abs/2509.02046

חרדה מלאכותית, השלכות אמיתיות