**פיין טיון של LLM ללא גרדיאנט: פריצת דרך אלגוריתמית, סקירה 521**

**הסקירה היומית של מייק: 08.10.25  
Evolution Strategies at Scale: LLM FINE-TUNING BEYOND REINFORCEMENT LEARNING**

אף שהרעיון הכללי של שימוש באסטרטגיות אבולוציוניות (ES) ברשתות נוירונים אינו חדש, ההנחה הרווחת הייתה שמגבלות זיכרון וחישוב הופכות אותו לבלתי-ישים עבור מודלים בעלי מיליארדי פרמטרים. החידוש המרכזי במאמר זה טמון באלגוריתם (אלגוריתם 2) מהונדס בקפידה, אשר מפרק שיטתית את המחסומים הללו באמצעות סדרת פתרונות הנדסיים מתוחכמים, המתמקדים בניהול זיכרון. זו אינה רק הדגמה של שימוש ב-ES, אלא בנייה מחדש של ארכיטקטורת התהליך כולו לסביבת מודלי ענק.

בואו ננתח את המכניקה האלגוריתמית שמאפשרת זאת. התהליך עוקף לחלוטין את הצורך בהתפשטות לאחור (backprop), מה שמפנה באופן מיידי את זיכרון ה-GPU המשמעותי שבדרך כלל שמור לגרפי חישוב וגרדיאנטים. עם זאת, האתגר עובר מניהול גרדיאנטים לניהול אחסון הפרמטרים של אוכלוסיית המודלים. יישום נאיבי היה דורש להחזיק בזיכרון N עותקים של מודל-ענק, דרישה שאינה מעשית. האלגוריתם של המחברים עוקף זאת באמצעות שלוש חדשנויות מרכזיות.

הראשונה היא רעיון ה"רעש הווירטואלי" באמצעות שחזור מ-seeds. אלגוריתם ES פועל על ידי החלת הפרעות (perturbations) אקראיות על וקטור הפרמטרים θ באמצעות רעש גאוסיאני, כדי ליצור N "צאצאים" להערכה. אחסון N וקטורי הרעש הללו, שכל אחד מהם בגודל המודל, הוא כשלעצמו בלתי אפשרי. הפתרון הוא לאחסן רק אוסף מצומצם של "seeds" אקראיים, אחד עבור כל חבר באוכלוסייה. במהלך שלבי הערכת הביצועים (דרך פונקציית תגמול) ועדכון המשקולות, כל תהליך מקבילי יכול לשחזר באופן מושלם את אותו וקטור רעש פסאודו-אקראי על ידי אתחול מחולל המספרים האקראיים עם ה-seed הספציפי. כך, הרעש הופך לישות ארעית המיוצרת "בדיוק בזמן" (just-in-time), נעשה בה שימוש, והיא נמחקת מבלי לתפוס זיכרון קבוע משמעותי.

השנייה היא מנגנון של פרטורבציות ושחזור "במקום" (in-place) ברמת השכבה**.** כדי להעריך חבר באוכלוסייה, האלגוריתם אינו יוצר מופע חדש של המודל. תחת זאת, הוא לוקח את מודל הבסיס היחיד והמשותף ומבצע בו שינוי כירורגי זמני. הוא עובר על שכבות המודל בזו אחר זו, מייצר עבור כל שכבה את מקטע הרעש המתאים (מה-seed) ומוסיף אותו ישירות למשקולותיה. לאחר שהתהליך הושלם, המודל מבצע "מעבר קדימה" (forward pass) כדי לייצר תגובה ולקבל ציון תגמול.

ההערכה עצמה היא שלב קריטי. המודל שעבר פרטורבציה מבצע "פענוח חמדני" (greedy decoding) דטרמיניסטי כדי לייצר תגובה. בחירה תכנונית זו חיונית, מכיוון שהיא מבטיחה שכל שונות בביצועים בין חברי האוכלוסייה נובעת אך ורק מהפרמטרים השונים שלהם, ולא מאקראיות בדגימת הטוקנים. התגובה שנוצרה מקבלת ציון מפונקציית תגמול (reward), המזקקת את הפלט המורכב של המודל לערך סקלרי יחיד המייצג את כשירותו (fitness). אופי הפונקציה תלוי במשימה: במשימת הסקה סמלית כמו Countdown, התגמול הוא אות דליל (sparse) הניתן רק אם הפתרון הסופי נכון. במשימות תמצות, התגמול מחושב על בסיס(אך לא רק) הפרש האורכים בין תגובת המודל לתשובה אידיאלית. לאחר קבלת התגמול, המודל משוחזר במדויק למצבו המקורי על ידי יצירה מחדש של אותו וקטור רעש וחיסורו מהמשקולות, שוב "במקום".

השלישית היא עדכון פרמטרים מצטבר ודיפרנציאלי**.** לאחר הערכת הביצועים של כל N חברי האוכלוסייה, יש לעדכן את פרמטרי המודל המרכזי. כלל העדכון של ES הוא סכום משוקלל של כל וקטורי הרעש, כשהמשקלים נקבעים על פי ציוני התגמול המנורמלים. יצירת וקטור העדכון הסופי, וקטור יחיד עם מיליארדי ממדים, תיצור צוואר בקבוק נוסף בזיכרון. האלגוריתם נמנע מכך על ידי פירוק העדכון: התהליך המרכזי חוזר על ה"seeds" פעם נוספת, ועבור כל אחד מהם, משחזר את וקטור הרעש, משקלל אותו, ומוסיף את תרומתו באופן מצטבר לפרמטרים של המודל הראשי, שכבה אחר שכבה. כך, וקטור העדכון המלא לעולם אינו מתממש בזיכרון כיחידה אחת (not materialized), אלא נבנה חלק אחר חלק ישירות בתוך טנסורי הפרמטרים הקיימים.

השיטה המוצעת משנה את פרדיגמת הכוונון העדין של LLMs. היא מחליפה את המנגנונים עתירי-החישוב של backpropagation ואת המורכבויות של למידת חיזוק, במערכת מבוססת "מעבר קדימה" בלבד, הניתנת למקבול מאסיבי, שהאתגר המרכזי שלה הוא "תזמור זיכרון" (memory orchestration). תרומתו של המאמר אינה רק בהצעת ES כחלופה, אלא באספקת יישום אלגוריתמי חדשני שהופך אותה למתחרה בת-קיימא בסדרי גודל שנחשבו בעבר מחוץ להישג יד.

https://arxiv.org/abs/2509.24372