**חשיבה מחדש על המעבר קדימה: מבט קומפוזיציוני על מודלי שפה גדולים, סקירה 519**

**סקירת המאמר היומית של מייק: 04.10.25  
Skip a Layer or Loop it? Test-Time Depth Adaptation of Pretrained LLMs**

הנחת היסוד מאחורי LLMs היא קשיחות ארכיטקטונית. וקטור קלט מתחיל בשכבה הראשונה ומתקדם באופן סדרתי עד לשכבה *N*, תוך שהוא עובר סדרה קבועה של טרנספורמציות, ללא קשר לשאלה אם השאילתה טריוויאלית או מורכבת עד מאוד. מאמר זה קורא תיגר על פרדיגמה סטטית זו, וממסגר אותה מחדש לא כבעיה של אימון המודל, אלא כבעיה של הסקה בזמן-טסט (test-time inference). החידוש המרכזי הוא קפיצה רעיונית: המאמר מציע שנפסיק להתייחס למודל טרנספורמר מאומן מראש כמבנה מונוליטי, ובמקום זאת נראה בו ספרייה קומפוזיציונית של מודולים.

המחברים מציגים את "שרשרת-השכבות" (CoLa - Chain-of-Layers), שבה עבור כל קלט נתון, נבנה "מסלול" אופטימלי באופן דינמי מהשכבות הקיימות והקפואות של מודל הבסיס. לא מדובר רק במציאת נקודת יציאה מוקדמת. החדשנות טמונה בגמישות העצומה של המרחב הקומפוזיציוני שנפתח. עבודה זו מסנתזת באלגנטיות שני קווי מחקר נפרדים: שיטות חיסור (גיזום שכבות, יציאות מוקדמות) שמטרתן יעילות על ידי צמצום חישובים, ושיטות חיבור (טרנספורמרים לולאתיים, עומק אדפטיבי) השואפות לעיבוד עמוק יותר על ידי חזרה על חישובים.

מרחב החיפוש של CoLa מאפשר לשתי הפעולות, דילוג וחזרה לולאתית, לחול על כל מקטע רציף של שכבות. הדבר יוצר מגוון עשיר הרבה יותר של תת-ארכיטקטורות אפשריות בהשוואה לעבודות קודמות. כעת מודל יכול לבנות מסלול רדוד עבור שאילתה פשוטה על ידי דילוג על חלקים נרחבים מהשכבות האמצעיות שלו ("חשיבה מהירה"), בעוד שעבור משימת חשיבה מורכבת, הוא יכול לזהות שכבה שימושית במיוחד, למשל, כזו המתמחה בהפשטה של יחסים, ולחזור עליה בלולאה מספר פעמים לצורך עידון איטרטיבי ("חשיבה איטית"). החידוש האמיתי הוא היכולת לשלב בין אסטרטגיות אלו, וליצור מסלולים היברידיים שעשויים, לדוגמה, לדלג על שכבות מוקדמות, לעשות שימוש חוזר אינטנסיבי בכמה שכבות אמצעיות, ואז לדלג ישירות לשכבות הפלט הסופיות.

כמובן, מרחב החיפוש של תמורות שכבות כאלה הוא עצום מבחינה קומבינטורית, מה שהופך גישת bruteforce או גישה חמדנית ללא מעשית. כדי לנווט במרחב זה, המחברים ממסגרים את גילוי המסלול כבעיית חיפוש ומשתמשים בחיפוש מונטה קרלו על עץ (MCTS). זהו החידוש המרכזי השני של המאמר. MCTS מספק שיטה עקרונית לאיזון בין חקירה (exploration) של רצפי שכבות חדשניים, ואולי לא אינטואיטיביים, לבין ניצול (exploitation) של מסלולים שכבר הראו פוטנציאל.

מבחינה טכנית, MCTS מנוסח כמשחק חיפוש שבו כל "מצב" הוא רצף שכבות ספציפי, שמאותחל כמסלול המעבר הסטנדרטי של המודל. ה"פעולות" אינן תמורות שרירותיות אלא עריכות מובנות המיושמות על מקטעי שכבות רציפים, כגון דילוג על קבוצה של *k* שכבות או חזרה על מקטע *r* פעמים. לאחר כל simulated rollout של העץ, שבה מסלול מלא מופעל על קלט, מחושב "תגמול" (reward) המבוסס על נכונות חיזוי. החידוש בהנחיית החיפוש טמון בפונקציית המטרה רבת-החלקים שלו. ציון ה-UCB מאזן בין 3 לחצים מתחרים: רכיב ניצול המעדיף מסלולים עם היסטוריה של גמול גבוה, רכיב חקירה המעודד דגימה של נתיבים ארכיטקטוניים שטרם נבחנו, ועונש מכריע על אורך המסלול, אשר מטה במפורש את החיפוש לכיוון פתרונות חסכוניים מבחינה חישובית. לכן, החיפוש אינו שואף רק לנכונות; הוא שואף לאיזון אופטימלי, ובסופו של דבר מניב קבוצה של מסלולים יעילים-פרטו המציגים פשרה בין דיוק לעלות הסקה.

בסופו של דבר, תרומת המאמר אינה מודל חדש, אלא דרך חשיבה חדשה על הסקה. הוא מניח שהמעבר קדימה הסטנדרטי הוא רק אחד מאינספור מסלולים חישוביים אפשריים, וככל הנראה תת-אופטימלי עבור רוב הקלטים. ייתכן שכוחו האמיתי של מודל מאומן מראש אינו טמון בעומקו הקבוע, אלא במרחב הקומבינטורי הסמוי של ארכיטקטורות שהוא מכיל באופן מובלע. עבודה זו מספקת תכנית אב משכנעת ומבוססת-עקרונות חישוביים לאופן שבו ניתן להתחיל לחקור מרחב זה.

https://arxiv.org/pdf/2507.07996