המאמר היומי של מייק - 06.02.25  
SMALL LANGUAGE MODELS: SURVEY, MEASUREMENTS, AND INSIGHTS

תמצית:

המאמר זה חוקר את החשיבות הגוברת של מודלי שפה קטנים (SLMs) ומשווה את התפתחותם ל-LLMs. בעוד ש-LLMs דורשים משאבי מחשוב משמעותיים ובדרך כלל מופעלים בשרתים, SLMs מתוכננים לפעול במכשירים מוגבלי משאבים כמו מחשבים ניידים, טאבלטים, סמארטפונים ומכשירי IoT. המחקר מציע סקירה מקיפה של 59 מודלי SLM, מעריך אותם על בסיס התקדמויות ארכיטקטוניות, אלגוריתמי אימון ויעילות הסקה. באמצעות קידום אימוץ SLM, עבודה זו שואפת להפוך את הבינה המלאכותית לנגישה, זולה ויעילה יותר ליישום מעשי.

אני אתרגם את הטקסט לעברית:

ניתוח טכני

1. חידושים ארכיטקטוניים מאמר המחקר בוחן לעומק את האלמנטים הארכיטקטוניים המבדילים בין SLM ל-LLM, תוך הדגשת השינויים המשפרים את היעילות במכשירים עם משאבים מוגבלים.

מנגנוני self-attention: באופן מסורתי, attention מרובת-ראשים (MHA) הייתה המנגנון הדומיננטי ב-SLM. עם זאת, המאמר מזהה מעבר הדרגתי לטכניקת attention מבוססת-קבוצות (GQA). גרסה זו מפחיתה את המורכבות החישובית על ידי שיתוף ייצוגי שאילתות בין ראשים תוך שמירה על גיוון בייצוגי מפתח-ערך. הדוח מספק ראיות לכך שמודלי GQA, כמו Qwen2.5, עולים משמעותית על אלה עם MHA מבחינת ה-latencies ויעילות זיכרון, במיוחד בשלב ההיסק (אינפרנס).

רשתות feed-forward: האבולוציה הארכיטקטונית מראה העדפה ל-Gated FFNs (שזה שילוב של שני FFNs -אחת עם אקטיבציה א לינארית בסוף) על פני FFN סטנדרטיות. Gated FFN מפעיל באופן סלקטיבי חלקים מהרשת, מה שמוביל ליעילות פרמטרית טובה יותר. ממצא מעניין הוא הגיוון ביחסי הביניים ב-Gated FFNs, (כלומר המימד של השכבה הלינאריות הראשונה שם) הנעים בין פי 2 ל-8 מהממד החבוי, כאשר יחסים גדולים יותר משפרים בדרך כלל את הדיוק במשימות היסק מורכבות.

פונקציות אקטיבציה: נצפה מעבר משמעותי מ-GELU (שזה Gaussian Error Linear Unit) וגרסאותיה ל-SiLU (שזה Sigmoid Linear Unit). המאמר מציין כי SiLU יעילה מבחינה חישובית ומתאימה יותר למודלים מקוונטטים, שולטת במודלים שהושקו ב-2024.

נרמול שכבות: מודגש המעבר מ-LayerNorm ל-RMSNorm. RMSNorm מפחית את העומס החישובי על ידי ביטול הצורך בחישוב הממוצע במהלך הנורמליזציה, יתרון משמעותי במכשירי קצה.

גודל מילון: גדלי מילון של SLM גדלו משמעותית, ולעתים קרובות עולים על 50K טוקנים. המחברים מקשרים עלייה זו עם יכולות משופרות בהבנת שפה.

ניתוח דאטהסטים לאימון:

מגמות בשימוש בדאטהסטים: המחקר מתעד מעבר מדאטהסטים כלליים נפוצים כמו \*The Pile\* ו-\*RefinedWeb\* לדאטהסטים מאוגדים כמו \*FineWeb-Edu\* ו-\*DCLM\*. דאטהסטים החדשים הללו משלבים טכניקות סינון מבוססות-מודל המשפרות משמעותית את איכות דאטה.

איכות דאטה לעומת כמות: למרות ההסתמכות המוקדמת על נפח דאטה גדול, הדוח מוצא שדאטהסטים באיכות גבוהה מניבים ביצועי מודל טובים יותר, גם עם פחות טוקנים. לדוגמה, מודלים שאומנו על \*FineWeb-Edu\* משיגים דיוק תחרותי בהשוואה למודלים מסחריים (סגורי-קוד) מתקדמים.

כמה טוקנים צריך לאימון: המחברים מציינים מגמה מפתיעה: SLM רבים מאומנים על מספר טוקנים הגדול בהרבה ממה שחוק צ'ינצ'ילה מציע. לדוגמה Qwen2.0 - 500M מאומן על 12 טריליון טוקנים, בעוד ש-Qwen2.0 - 1.5B מאומן על 7 טריליון בלבד. אסטרטגיית "אימון-היתר" המכוונת הזו מוצגת כאופטימיזציה לסביבות מוגבלות-משאבים, המאפשרת למודלים להכליל טוב יותר כאשר הם מופעלים במכשירים עם כוח חישוב מוגבל.

3. חידושים באלגוריתם האימון

שיטת Maximal Update Parameterization - µP: בשימוש במודלים כמו Cerebras-GPT, שיטת µP מבטיחה אימון יציב על ידי בקרה על אתחול, קצבי למידה בכל שכבה, ועוצמות ראקטיבציה. טכניקה זו מאפשרת להעביר היפר-פרמטרים שאופטמו עבור מודלים גדולים ישירות למודלים קטנים יותר, מה שמייעל את תהליך האימון.

זיקוק ידע(דיסטילציה): LaMini-GPT ו-Gemma-2 מנצלים טכניקה זו להעברת ידע ממודלי מורה גדולים למודלי תלמיד קטנים יותר, מה שמוביל לביצועים משופרים ללא צורך באימון נרחב.

אסטרטגיית אימון מקדים דו-שלבית: אומצה על ידי MiniCPM, אסטרטגיה זו כוללת שלב ראשוני עם דאטה באיכות נמוכה ולאחר מכן פיין-טיון עדין עם דאטה באיכות גבוהה וספציפיים למשימה. השיטה מוכיחה את עצמה כיעילה באיזון בין יעילות חישובית לביצועי המודל.

4. הערכת ביצועים

היסק מבוסס שכל ישר: מודלים כמו Phi-3-mini משיגים ביצועים מתקדמים, (המתחרים ב-LLaMA 3.1 7B). תוצאות הבנצ'מרק מגלות שהשיפורים באיכות מערך הדאטה ואסטרטגיות האימון אפשרו ל-SLM לצמצם את הפער עם מודלים גדולים יותר.

פתרון בעיות ריזונינג: Phi-3-mini ו-SLM אחרים בעלי ביצועים גבוהים מציגים שיפור של 13.5% בביצועים משנת 2022 עד 2024, עולים על קצב השיפור של מודלי LLaMA. זה מדגים את הבשלות הגוברת של SLM בטיפול במשימות היסק מורכבות.

מתמטיקה: ביצועי ה-SLM נשארים תת-אופטימליים במתמטיקה, כאשר המודלים מתקשים לטפל במשימות הדורשות חשיבה לוגית. המחברים מייחסים פער זה למחסור בדאטהסטים באיכות גבוהה המתמקדים בלוגיקה.

למידת in-context: הניסויים מגלים ש-SLM מפיקים תועלת משמעותית מלמידה בהקשר, במיוחד עבור משימות כמו אתגר ARC, שם נצפים שיפורים בדיוק של עד 4.8%. עם זאת, חלק מהמודלים, כמו LaMini-LM, מציגים הידרדרות בביצועים עקב אוברפיט.

.5 ניתוח יעילות בזמן ריצה

לייטנסי וזיכרון נדרש: המחקר מוצא שהלייטנסי של היסק מושפעת הן מגודל המודל והן מהארכיטקטורה. לדוגמה, Qwen1.5 רץ 31.9% מהר יותר מ-Qwen2.0 על המעבד Jetson Orin, למרות שיש לו 25.4% יותר פרמטרים. זה מיוחס להבדלים במנגנוני attention ואסטרטגיות שיתוף פרמטרים (parameter sharing). צריכת זיכרון היא לינארית בד״כ ביחס לגודל המודל אך מושפע גם מגורמים כמו גודל אוצר המילים ומנגנוני attention מודלים כמו Bloom-1B1, שיש להם אוצרות מילים גדולים יותר, מציגים שימוש גבוה באופן לא פרופורציונלי בזיכרון.

קווינטוט: טכניקות קווינטוט, במיוחד 4 ביט, מוכיחות את עצמן כיעילות בהפחתת השהיה ושימוש בזיכרון. שיטת Q4 KM מפחיתה לייטנסי בממוצע ב-50% במהלך היסק, עולה על שיטות כימות 3-ביט ו-6-ביט, הסובלות מחוסר יעילות חומרתית.

סיכום:  
הניתוח הטכני המוצג במאמר זה מספק הבנה מקיפה של השיקולים הארכיטקטוניים, האימון, וזמן הריצה החיוניים לפיתוח ופריסה של SLM. על ידי התמודדות עם אתגרי היעילות ומגבלות המשאבים, המאמר מציע תובנות חשובות לקידום מחקר ה-SLM ויישומים מעשיים.

https://arxiv.org/abs/2409.15790