המאמר היומי של מייק: 21.08.25

Scaling Laws for Forgetting When Fine-Tuning Large Language Models

כל מי שאי פעם ביצע פיינטיון למודל שפה חזק מכיר את הפשרה הכואבת. אתה מתאים את המודל למשימה חדשה, ובתוך כך הוא מפתח סוג של אמנזיה ושוכח את הידע הכללי שהיה כל כך יקר לרכוש. "שכחה קטסטרופלית" זו היא אתגר בסיסי. תרופה נפוצה היא לערבב כמות קטנה מנתוני האימון-המקדים המקוריים במהלך הכוונון העדין, אך זה תמיד הרגיש יותר כמו תרופת סבתא מאשר מדע.

מאמר שנסקור היום מרומם את הטריק הזה למדרגה של מדע מדויק וניתן לחיזוי. המחברים עושים הרבה יותר מאשר רק לציין ש"הזרקת נתונים עוזרת". הם מציגים מודל חיזוי מדויק המתאר את הריקוד המורכב בין גודל המודל, כמות דאטה של פיינטיון, ואחוז דאטה מהאימון המקדים המוזרק לתוכו. בעוד שהכותרת הבולטת היא שהזרקה של אחוז אחד בלבד יכולה לעצור את השכחה, החידוש האמיתי של המאמר טמון במסגרת המתמטית שבבסיסו, המסבירה את כל התהליך.

החדשנות המרכזית היא חוק סקיילינג חדש שנועד לחזות את הלוס הסופי על נתוני האימון-המקדים, מדד ישיר לכמה המודל שכח. במקום נוסחה פשוטה, חשבו על כך כמערכת יחסים בין כוחות מתחרים. המבנה של המודל אלגנטי. הוא מתחיל מקו בסיס ה-loss ההתחלתי של המודל על דאטה מהאימון-המקדים עוד לפני שפיינטיון החל. לאחר מכן, הוא מוסיף איבר שני המחשב את עוצמת השכחה שתתרחש. איבר השכחה הזה הוא שבר, עם גורמים המחמירים את השכחה בַּמּוֹנֶה וגורמים המונעים אותה בַּמְּכַנֶּה.

מה מחמיר את השכחה? בַּמּוֹנֶה, אנו מוצאים איבר המייצג את כמות נתוני הכוונון העדין הייחודיים. הדבר חושף תובנה מרתקת: ככל שמאמנים מודל על יותר נתונים חדשים, כך הוא שוכח יותר את הידע הישן שלו. הסיבה לכך היא שיותר צעדי אימון גורמים לפרמטרים של המודל לסטות רחוק יותר ממצבם המקורי והכללי.

מה נלחם בשכחה? בַּמְּכַנֶּה, אנו מוצאים את הגורמים הממתנים. הראשון הוא גודל המודל (ספירת הפרמטרים שלו). זה מאשש את האינטואיציה שלמודלים גדולים יותר יש יותר קיבולת ללמוד מידע חדש מבלי לדרוס ידע קיים.

מרכיב הקסם: הנה החלק המבריק ביותר במודל. הזרקת נתוני האימון-המקדים ממוּדֶלֶת כמכפיל רב-עוצמה על גודלו האפקטיבי של המודל. כאשר המודל רואה אפילו אחוז קטן של נתונים מקוריים, הוא מתנהג כאילו יש לו ספירת פרמטרים גדולה בהרבה לצורך זכירת האימון המקורי שלו. מקדם מיוחד, שהמאמר מכנה "יעילות יחסית של פרמטרים" (B), קובע עד כמה האפקט הזה חזק עבור תחום נתון. עבור דומיין שונה מאוד מנתוני האימון-המקדים (כמו מתמטיקה), מקדם יעילות זה הוא עצום, מה שמסמן שההזרקה היא קריטית. עבור דומיין דומה יותר (כמו ויקיפדיה), המקדם קטן בהרבה, מכיוון שהמודל פחות נוטה לשכוח מלכתחילה.

המודל הזה אינו תיאורטי בלבד; הוא מדויק להפליא. על פני 12 דומיינים שונים, הוא חוזה את ה-loss הסופי על נתוני האימון-המקדים עם שגיאה יחסית ממוצעת של 0.49% בלבד.

מודל רב-עוצמה זה לשכחה מניב מספר תובנות חדשניות ומעשיות נוספות.

#### 1. ביצועי פיינטיון אינם נפגעים

חשש טבעי הוא שערבוב נתונים ישנים יפגע בביצועי המודל במשימה החדשה. המחברים מראים שלא כך הדבר. ה-loss הסופי על נתוני הוולידציה של הפיינטיון כמעט ואינו מושפע מהזרקת כמות קטנה של נתונים מקוריים. למעשה, עבור מודלים קטנים יותר, ההזרקה פועלת כרגולריזטור (regularizer) בריא, המונע האוברפיט ולעיתים אף מוביל לביצועים טובים יותר על דומיין המטרה.

#### 2. אקסטרפולציה היא כוח-על

הערך האמיתי של חוק סקיילינג הוא היכולת לחזות את עתיד האימון (במידה מסוימת). המחברים מאשרים שהמודל שלהם מצוין לאקסטרפולציה. על ידי הרצת ניסויים זולים על מודלים קטנים יותר (למשל, מודל של 334 מיליון פרמטרים), הם הצליחו לחזות במדויק את השכחה וביצועי פיינטיון של מודלים גדולים ויקרים בהרבה (1.3 מיליארד פרמטרים ומעלה). הדבר מאפשר למעבדות לחזות את התוצאות של ריצה בת 7 שעות על 8 GPUs באמצעות ניסוי של 30 דקות על 4 GPUs, ובכך לחסוך כמויות אדירות של זמן ואנרגיה.

#### 3. לא צריך את כל ערימת השחת (וזה מאוד חשוב)

במונחים מעשיים, האם הטכניקה הזו דורשת הזרמת דאטה מדאטהסט מקורי בגודל פטה-בייטים? התשובה היא לא (לפי המאמר). ניסוי מעניין מראה כי דאטהסט קטן באופן מפתיע הנדגם בצורה מסוימת מנתוני האימון-המקדים מספיק כדי שההזרקה תהיה יעילה (כלומר שהמודל ילמד ממנו). זה הופך את השיטה לנגישה וקלה הרבה יותר ליישום ממה שניתן היה להניח.

https://arxiv.org/abs/2401.05605