המאמר היומי של מייק: 16.05.25  
GROKKING AT THE EDGE OF NUMERICAL STABILITY

לא יכולתי לפספס את המאמר הזה - לא היה שום סיכוי. הרי מילה grokking מופיעה בשם המאמר וזה משהו שאני לא מפספס בגלל שזו אחת התופעות הכי מרתקות ובלתי מוסברות כרגע (כמו in-context learning) בלמידה עמוקה. אבל מה זה בעצם גרוקינג?

גרוקינג זו תופעה במהלך אימון של מודלים שונים כאשר אחרי הגעה ל״ביצועים אופטימליים״ על סט ולידציה. אם נמשיך לאמן לאחר מכן בהתחלה נראה ירידה בביצועים על סט הולידציה מלווה בעלייה של הביצועים על הטריין סט שזה כלומר אוברפיט. אם נמשיך לאמן עוד ועוד אז במצבים מסוימים (למשל במצב over-parameterized כאשר הקיבולת של המודל גדולה בהרבה מ״מה לדאטהסט שלנו צריך״) הלוס על סט הולידציה יתחיל לרדת שוב. כלומר האוברפיט נגמר והמודל נכנס למשטר הכללה, כלומד למידה אמיתית של הבעיה - וזה בדיוק גרוקינג.

גרוקינג כאשר לתופעה אחרות המתרחשות באימון של רשתות נוירונים: double descent ו-lottery ticket hypothesis. ניתן לאפיין תופעות אלו באמצעות כלים מפיזיקה סטטיסטית (עשו זאת עוד בתחילת שנות ה-90). גרוקינג לא קורה אוטומטית במהלך כל אימון מאוד ארוך, לפעמים צריך להשתמש ברגולריזציה כדי שזה יקרה. המאמר חוקר את הסיבות שאי הופעה של גרוקינג דרך ניתוח של שינוי משקולות המודל במהלך האימון - כלומר גרדיאנטים.

המחברים טוענים שאי הופעה של גרוקינג קשורה לקריסת הגרדיאנט במודל כלומר המודל מפסיק לעדכן את משקלותיו ועקב כך הגרוקינג לא קורה. המודל פשוט לא לומד. זה קורה בגלל שגיאות נומריות של פעולות floating point (או fp בקצרה). עבר פונקציית סופטמקס רגילה העדכונים הם כה קטנים שהמודל פשוט לא רואה אותם. כלומר הם מעבר לדיוק של FP אחרי הנקודה העשרונית. המחברים טוענים שטריקים ידועים שמונעים overflow ו-underflow כמו logsumexp (שזה חלוקה בערך המקסימלי שיש באקספוננט והוצאותו משם אחר הלוג) - יש לנו log בלוס אחרי הסופטמקס הרי.

אז הדבר הראשון שהמחברים מציעים הוא השכלול של סופטמקס הנקרא StableMax המקל על קריסת הגרדיאנט. פשוט לוקחים פונקציה שעולה בקצב נמוך יותר מהאקספוננט (שיש לנו בסופטמקס). אך שיפור זה בלבד לא מספיק והמחברים מציעים שכלול של שיטת עידכון הגרדיאנט במהלף האימון. המחברים שמו לב שבלא מעט מקרים שגרוקינג לא קורה כי הלוס על הטריין יורד בעיקר באמצעות הכנסת ״טמפרטורה גבוהה״ לסופטמקס במהלך אימון. כלומר הרשת ״בוחרת״ לעדכן את משקליו על ידי הכפלתם בקבוע מסוים כל איטרציה של GD. כך הלוגיט של הקטגוריה מקבל ערכים מאוד גבוהים וחיוביים והאחרים מקבלים ערכים שליליים מאוד נמוכים.

כאמור המחברים טוענים שסיבה לתופעה זו שינוי של משקלות המודל בכיוון של משקולות המודל כלומר מכפילים אותן בקבוע מסיום. אז המחברים מציעים לעדכן את המשקולות במהלך האימון (GD) בכיוון של הגרדיאנט מוטל על היפר-מישור האורתוגונלי לוקטור המשלקולות הנוכחי. כלומר הם מונעים מהמודל לנפח את משקלותיו בצורה שתוארה קודם (אין שינוי בכיוון משקולות המודל). המחברים ראו שככה ניתן להגיע לגרוקינג מהר יותר מאשר אימון עם רגולריזציה.

מאמר בהחלט שווה קריאה,

https://arxiv.org/abs/2501.04697