המאמר היומי של מייק: 30.04.25  
THE COMPLEXITY DYNAMICS OF GROKKING

זה אחד המאמרים החזקים והכי עמוקים שקראתי לאחרונה. ולא, הוא לא אימן מודל שהשיג ציונים הכי גבוהים בכל הבנצ'מרקים, לא הציע ארכיטקטורה או שיטת אימון חדשה. מה שהמחברים ניסו לעשות זה להסביר את תופעת הנקראת גרוקינג (grokking) דרך הפריזמה של דחיסה. דחיסה של דאטה וגם דחיסה של המודלים. הנושא קצת מורכב ואנסה להסביר אותו לאט לאט בצורה פשוטה.

תופעת גרוקינג מתרחשת במהלך אימון של רשת נוירונים כאשר אנו ממשיכים לאמן את הרשת אחרי שהגענו למינימום של לוס ולידציה. כמובן שבהתחלה אנו נכנסים למוד של אוברפיט ולוס הולידציה שלנו עולה ועולה. אבל בנקודת מסוימת קורה משהו מוזר - פתאום לוס הוולידציה מתחיל לרדת וזה מצביע על כך המודל עובר ממוד של זיכרון (memorization) למוד של הכללה. במילים פשוטת המודל אשכרה ״פיצח את הבעיה״.

תופעה זו מתרחשת במודלי overparameterized כאשר מספר המשקלים במודל גבוה הרבה יותר מזה ש״צריך בשביל ללמוד הדאטהסט״ (ניתן להסביר זאת בצורה מדויקת יותר אך זה מערב מתמטיקה לא טריויאלית שלא נחוצה להבנת סקירה זו). גרוקינג קשור לתופעה הנקראת double descent וגם lottery ticket hypothesis. דרך אגב אם ממשיכים לאמן את המודל אז לוס הולידציה ממשיך לרדת ולא עוצר (כלומר מתכנס לאפס).

אוקי, אבל איך כל הסיפור הזה קשור לדחיסה? בשביל כך אנו צריכים להסביר שני מונחים מאוד חשובים: הראשון הוא עקרון שנקרא minimum description length או MDL. עקרון זה טוען אם אנו רוצים לדחוס את הדאטהסט שלנו בצורה הטובה ביותר באמצעות מודל אנו צריכים למעזר סכום של אנטרופיית הדאטה אחרי שהועבר דרך המודל פלוס הקומפלקסיטי (complexity) של המודל עצמו. עקרון זה מתבסס על משפט הקידוד של שנון הטוען שככל האנטרופיה של הדאטה קטנה יותר ניתן לדחוס את הדאטה בצורה יעילה יותר (כלומר לדחוס אותו יותר).

אוקיי, למדוד אנטרופיה של הדאטהסט אחרי שהוא הועבר דרך המודל אנו פחות או יותר יודעים. עבור משימת סיווג זה יכול להיות פשוט לוס cross-entropy. עבור לשערך את הקומפלקסיטי של המודל אנו צריכים לעבוד יותר קשה. קודם כל צריך להגדיר מה זה הקומפלקסיטי של קולמוגורוב או KC. למעשה KC עבור דאטה d נתון מוגדר בתור אורך תוכנית מחשב (=קוד) הקצר ביותר שיכול לפלוט d. למשל עבור שורה של אחדות אנו צריכים קוד מאוד קצר (KC נמוך) וכדי להדפיס שורה של 0 ו- 1 רנדומליים צריך קוד ארוך בערך באורך השורה (KC גבוה). כמובן שלא ניתן לחשב KC במדויק.

עוד מושג חשוב שצריך לדעת להבנת המאמר הוא פונקצית rate distortion r שבהינתן קלט x ואפסילון חיובי מגדירה מהו ייצוג מספר הביטים המינימלי (או KC) של קלט y עם שהוא רחוק מ-x באפסילון לכל היותר. כמובן ״רחוק״ תלויה בפונקצית מרחק ובמאמר תפקיד של x ו-y הם המשחקים מודל מאומן ״רגיל״ M ומודל coarse-grained או CS. מודל CS הוא מודל מאומן שעבר סוג מסוים של ״פישוט״ של M למשל קווינטוט, pruning או החלפת מטריצות משקולות בייצוגן על ידי מטריצות בעלות ראנק נמוך. גם מודל M שאומן עם רגולריזציה יכול להיחשב CS יחסית למודל שאומן ללא רגולריזציה. פונקציית מרחק שהמחברים השתמשו בה עבור חישוב של r הוא הפרש בין הלוסים של מודל רגיל M למודל CS.

אוקיי, אחרי שהבנו את המושגים הנחוצים בואו נחזור לגרוקינג. הטענה העיקרית של המאמר היא שככל שאנו מתקדמים באימון המודל שמתקבל נהיה דחיס יותר כלומר קיים מודל CS עם הפרש ביצועים זניח (אפסילון) מהמודל המאומן (במהלך גרוקינג). כל זה קורה בזמן של ה-description length של הדאטה באמצעות המודל רק יורד כלומר המודל אכן לומד את הדאטה באמצעות מודל דחיס (סוג של פשוט יותר). למה זה קורה בעצם? המודל מצליח להגיע ללוס קרוס-אנטרופי נמוך באמצעות מודל דחיס (בעל rate distortion נמוך לפי ההגדרה בפסקה הקודמת).

מקווה שהצלחתי להסביר את המאמר הזה בצורה ברורה יחסית.

https://arxiv.org/abs/2412.09810