⚡️🚀 המאמר היומי של מייק 10.06.24:⚡️🚀

Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks

אחד התופעות המרתקות בלמידה עמוקה היא גרוקינג - שהיא מעבר ״פתאומי״ של רשתות עמוקות למצב של הכללה מהמצב של overfitting למשל אחרי אימון מאוד ארוך. הרי ידוע שאם עבור דאטהסט נתון ורשת עמוקה בעלת יכולת ייצוג גבוהה מספיק (representativeness) אחרי שלב מסוים באימון אנו נגיע ל-overfitting כלומר למצב שבו ביצועי המודל יילכו וישתפרו עבור סט האימון אולם הביצועים על סט הולידציה יספגו ירידה בביצועים.

מה שמגניב ומפתיע בגרוקינג שעבור אימון ארוך מספיק מגיע המצב שביצועי המודל על סט הוולידציה מתחילים לעלות יחד עם אלה על סט האימון כלומר המודל מגיע לשלב של הכללה אמיתית. מעניין שתופעה דומה מתרחשת בתנאים מסוימים אם אנו מגדילים את קיבולת המודל (מס' הפרמטרים) כאשר גודל הדאטהסט ומשך האימון נותרים קבועים) וגם כאשר אנו מגדילים את גודל הדאטהסט תוך שמירה של משך האימון קיבולת המודל קבועים.

למעשה תופעות אלו שייכות למשפחת double descent (יש גם multiple descent) שנחקרה רבות על ידי חוקר דגול מישה בלקין. התופעה עצמה נתגלתה לפני יותר מ- 30 שנה (מי שרוצה להתעמק בנושא תעקבו אחרי https://www.linkedin.com/in/charlesmartin14/ - הוא אחד המומחים הגדולים).

אוקיי, אז מה עשה המאמר הנסקר? הוא חקר תופעת גרוקינג כאשר מתרחשת אם מגדילים את מספר המשימות (כל משימה היא סוג של רגרסיה לינארית בשדה המודולו(שארית)) שעבורן אנו מאמנים את המודל (כמובן לקחו מודל שפה). מתברר כי יש כמה משטרים (מודים) של יכולת הכללה של המודל כאשר משחקים עם היחס של מספר הדוגמאות פר משימה ועם מספר המשימה. בגדול מאוד אם נותנים מספיק משימות גדול מספיק ומספר דוגמאות פר משימה גדול מספיק אז מגיעים להכללה אמיתית כאשר המודל אכן לומד את המשימה במלואה).

https://arxiv.org/abs/2406.02550

קריאה מהנה!