⚡️🚀המאמר היומי של מייק 01.07.24:⚡️🚀

Grokfast: Accelerated Grokking by Amplifying Slow Gradients

המאמר הזה משך את עיניי משתי סיבות. הסיבה הראשונה היא הופעת מילי Grokking בכותרת. מה זה בעצם Grokking בהקשר של אימון רשתות. אתם בטח יודעים אם אנו מאמנים את הרשת שלנו ליותר מדי זמן (כלומר אפוקים) אז באיזושהי נקודה היא מגיעה למצב של אוורפיט. כלומר הלוס על טריין סט ממשיך לרדת בזמן שהלוס על סט ולידציה מתחיל לעלות כלומר יכולת הכללה של המודל נפגעת.

אבל אם אנו נמשיך לאמן את הרשת שלנו עוד עוד באיזשהו שלב הלוס על סט ולידציה מתחיל לרדת לאט לאט כלומר יכולת הכללה של המודל משתפרת. כלומר אנו יוצאים מ״משטר האוורפיט״ אחרי שלב מסוים של אימון וזה נקרא grokking. התופעה הזו נחקרת רבות על ידי המדענים בתחום למידה עמוקה אבל אין הבנה מלאה למה זה קורה. השורשים של grokking הזו נמצאים כנראה בתופעה שנקראת double descent.

הסיבה השנייה שבחרתי לסקור את המאמר כי נוכחתה של התמרת פורייה שם אלא אחרי התעמקות קלה התברר שניתן היה להסתדר גם בלעדיו ולהסביר את המאמר בצורה פשוטה יותר בהרבה (מה שאני עושה בסקירה הזו).

גרוקינג זו תופעה מאוד נחמדה וכל אדם המאמן את המודלים שלו חפץ להגיע אליך אך הבעיה שצריך לאמן את הרשת למשך מאות אלפי ולפעמים יותר איפוקים וזה מאוד יקר. השאלה האם ניתן לזרז את התהליך הזה ולהגיע לגרוקינג מהר יותר.

וזה בדיוק מה זה המאמר רוצה לעשות. המאמר טוען שאם נחליק טיפה את עדכון המשקלים של הרשת (כלומר את הגרדיאנטים) אז ניתן להגיע לגרוקינג מהר יותר. נשמע לא מופרך בגדול (למשל PPO בלמידה עם חיזוקים גם מרככת את עדכון הגרדיאנט וגם שיטות אימון כמו ADAM ומומנטום של נסטרוב) - אבל כמובן ההוכחה לא נמצאת במאמר. וכאן המחברים דוחפים התמרת פורייה מהסיבה הפשוטה שהחלקה זו היא למעשה העברת גרדיאנים דרך מסנן low-pass אבל כאמור אפשר היה להסתדר בקלות בלעדיהם.

בסופה של דבר המאמר מציע למצע כמה גרדיאנטים, להחליק(להוסיף) באמצעות הממוצע הזה את הגרדיאנט הנוכחה ואז לעדכן את משקלי הרשת (עם adam למשל). כמובן שזה דורש לשמור כמה גרדיאנטים וזה מצריך הרבה זכרון והמחברים הציע החלקה מעריכית (exponential smoothing) במקום זה בלי כמעט לפגוע בתוצאות (התוצאה היא כמובן זירוז של הגעה לגרוקינג).

מאמר חמוד אבל ציפיתי ממנו קצת יותר..

https://arxiv.org/abs/2405.20233