⚡️🚀המאמר היומי של מייק 14.09.24: ⚡️🚀  
Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning

חוקי סקיילינג זה נושא מאוד מעניין אך לצערי אני מתקשה למצוא מאמרים באמת שווים סקירה (שמכילים מעבר לניסויים אינסופיים עם הייפרפרמטרים שונים). הפעם התמזל מזלי ונתקלתי במאמר הלא חדש הזה שהוא נראה די שווה.

המאמר מציע סוג חדש חוקי סקיילינג בנוגע ל- Data Pruning (צמצום דאטה או DP). המחברים מספקים ראיות תיאורטיות (זו הסיבה שאני סוקר אותו) ואמפיריות לכך שצמצום פיסות דאטה מיותרות או פחות אינפורמטיביות יכול לשבור את חוקי הסקליינג המסורתיים, ולהשיג הפחתה מהירה יותר בשגיאה תוך שימוש בפחות משאבים.

רקע: חוקי הסקיילינג של רשתות נוירונים מתארים כיצד השגיאה(טסט) יורדת עם הגדלת גודל המודל, כמות הדאטה או כמות הקומפיוט, בהתאם לחוק חזקה (Power Law). עם זאת, סקלינג זה אינו יעיל, שכן שיפור בביצועים דורש כמות דאטה/משאבים אקספוננציאלית. המחברים שואלים האם ניתן להשיג סקלינג טוב יותר מחוק חזקה על ידי בחירה מושכלת של דאטה.

התמצית: המחברים מפתחים מסגרת תיאורטית המבוססת הלקוחה ממכניקה סטטיסטית, תוך שימוש במודל בסגנון זיקוק מידע (מודלי סטודנט-מורה). מודל זה מתאים לבחינה תיאורטית של data pruning (זריקת נתונים) בגלל פשטותו המתמטית, תוך שמירה על תכונות הכללה (generalization) שנשמרות במודלים מורכבים יותר.

המסגרת המתמטית המוצעת מורכבת מ"מורה" שמייצר דאטה, ומודל "סטודנט" שמנסה ללמוד אותו. הרעיון המרכזי הוא ״להעיף דוגמאות על על בסיס הקושי שלהן״. קושי של דוגמא נמדד על פי המארג'ין(המרחק של דוגמא מגבול ההחלטה). בגדול הם הראו כי יש לשמור דוגמאות קלות (עם מארג'ינים גדולים) עבור דאטהסטים קטנים, בעוד שדוגמאות קשות יותר (עם מארג'ינים קטנים) הן אינפורמטיביות יותר דאטהסטים גדולים. המחברים מראים כי שגיאת ההכללה E\_g, תלויה ביחס בין מספר דוגמאות כולל לפרמטרי המודל (alpha) ובחלק מהדאטה f שהוסר. המסקנה המרכזית היא שחיתוך אופטימלי שובר את חוק החזקה בסקיילינג, ומוביל לסקיילינג מעריכי של הפחתת השגיאת הכללה.

אז אלו דוגמאות להשאיר: כאמור עבור דאטהסטים קטנים, עדיף לשמור דוגמאות קלות כדי להימנע אוברפיט, בעוד שעבור דאטהסטים גדולים, משתלם להשאיר דוגמאות קשות כדי ללמוד גבולות החלטה עדינים יותר. יש טענה במאמר שברגע ששומרים את הדוגמאות הקשות ביותר, מתאפשר סקלינג מעריכי של הפחתת שגיאת ההכללה E\_g, עבור דאטהסטים גדולים. המחברים מצאו כי הדעיכה המעריכית מחזיקה עד לנקודת שבירה קריטית, שבה הדוגמרו הנותרים כבר אינם מספקים מספיק מידע. מעבר לנקודה זו, דעיכת השגיאה מאטה ועוברת לחוק חזקה.

רווח מידע (Information gain או IG): המחברים טוענים כי בלמידה עם רשתות המידע השולי שמספקת כל דוגמא נוספת פוחת עם מספר הדוגמאות, מה שמוביל ליחס חוק חזקה בין גודל הדאטהסט להפחתת שגיאת הכללה. אולם, עם אסטרטגיית בחירה חכמה, המצב משתנה. חיתוך מסיר נתונים מיותרים או בלתי אינפורמטיביים, ומאפשר לכל דוגמה שנותרה לספק מידע ייחודי יותר על המשימה. מתמטית, תכולת המידע של דאטהסט (לסטודנט) פרופורציונלית למספר דוגמאות שנותרו, אך ניתן להאט את קצב הירידה עם בחירה מושכלת של הדוגמאות. כלומר רווח המידע לדוגמא נשאר משמעותי גם כשהדטאהסט נחתך, מה שמאפשר דעיכה מעריכית של השגיאה.

חוסר איזון בין קטגוריות: המאמר דן בכך שבחירת דוגמאות ללא התחשבות בהתפלגות קטגוריות עלול להוביל לחוסר איזון בינן יגרום לירידה בביצועי המודל. המחברים מציעים טכניקת איזון קטגוריות שמבטיחה שכל אלו יישארו מיוצגות היטב בדאטהסט החתוך.

https://arxiv.org/abs/2206.14486