# Review 67: Grokking: Generalization beyond Overfitting on small algorithmic datasets

**Paper: https://arxiv.org/abs/2210.01117v2**

מאמר די מסקרן של openai שיצא לאחרונה. אסקור אותו קצרות כמו שמקובל ב #shortdeepnightlearners

הטענה העיקרית של המאמר נשמעת לא מסובכת. הרי ידוע שאם נאמן רשת יותר מדי זמן, מלשב מסוים היא תיכנס למוד של overfitting כלומר שגיאת ההכללה שלה תתחיל לעלות. המחברים טוענים שאם נאמן אם נמשיך לאמן את הרשת גם לאחר שהיא נמצאת ב- overfitting מאיזשהו שלב שגיאת ההכללה שלה תתחיל לרדת כלומר הביצועים שלה על טסט (או ולידציה) יתחילו לעלות.

המחברים זיהו תופעות כאלו כאשר הם אימנו רשתות על דאטהסטים ״אלגוריתמיים״ קטנים. מה זה דאטהסט אלגוריתמי ואיזו משימה נותנים לרשת במקרה, אתם שואלים? אחת הדוגמאת הוא מטריצה המתארת פעולות במרחב התמורות (פרמוטציות). כלומר לוקחים תמורות בגודל מסוים (נגיד 120 תמורות עבור n=5) ובונים מטריצה בגודל 120x120 שמשבצת (i,j) מכילה תוצאה של הרכבת תמורה p\_i ו- p\_j. אז משחירים כמה מהמשבmות במטריצה זו ונותנים לרשת ללמוד אותם. זה משהו די מגניב שטרם ראיתי…

עכשיו כמה מילים על התופעה עצמה. התופעה הנצפית היא סוג של double descent, הנקרא epoch-wise (הטרמינולוגיה מ- https://www.kdnuggets.com/2020/04/double-descent-hypothesis-bigger-models-more-data-hurt-performance.html). אנו מכירים היטב double descent מהסוג Model-Wise שמשמעותו היא שהגדלה של מספר הפרמטרים לרשת גורמת לתופעה דומה משלב מסוים (שגיאת הכללה עולה ואחר-כך יורדת). מודה שבהתחלה התבלבלתי בעצמי וטענתי התופעה המתוארת במאמר היא לא double descent אך אחרי השיחה עם Liron Itzhaki (תודה רבה על תובנות מאוד מעניינות) הבנתי שיש לי טעות בזיהוי. דרך אגב גם Yannic Kilcher בסרטון שלו טען שזה לא double descent.

למה תופעה כזו מתרחשת. Misha Belkin הגדול חוקר את התופעה המעניינת הזו כבר כמה שנים אבל עדיין אין הסבר מתמטי לכך (תקנו אותי אם אני טועה כאן). לגבי הסברים בנפנופי ידיים יש לי שניים (אחד של לירון ואחד לי):

מרג'ינים: כאשר ממשיכים לאמן רשת (לסיווג נגיד) כשהיא נכנסה למוד overfitting והלוס ממשיך לרדת לאזור האפס, הרשת עשויה לגלות ״ מודל (המוגדר ע״י הפרמטרים של הרשת) המפריד בצורה רווחת יותר בין הקלאסים״ (עם מרג'ינים גדולים). מודל זה ״חסין״ לרעש בדאטה ומכלילים מספיק טוב לאחר שבהתחלת overfitting היא מצאה פתרונות עם מרג'ינים קטנים שמובילים לשגיאת הכללה גבוהה.

מינימום של פונקצית לוס: יש הנחה שפתרונות בעלי שגיאת הכללה נמוכה ״נמצאים״ במינימומים בעלי עקמומיות נמוכה (לא חדים) של פונקציית לוס. כלומר כאלו שפונקציית לוס סביבם לא משתנה הרבה וערכיה סביב נקודת מינימום זו היא נמוכים. נקודת מינימום כזו היא יותר טובה ממינימום חד (שאפילו בסביבתו הקטנה פונקציית לוס מקבלת ערכים גבוהים משמעותית מזה בנקודת מינימום). אז יש מצב שמאמנים רשת מספיק איפוקים אז יש סבירות גבוהה להגיע מתישהו לנקודה כזו (לא חדה) שכבר ״קשה״ לצאת ממנה כי פונקציית לוס מקבלת ערכים נמוכים סביבה.

כמובן של הסברים בנפנופי ידיים אבל לדעתי עשויים להיות מועילים לחשיבה על התופעה המעניינת הזו

סרטון של יניק: https://www.youtube.com/watch?v=dND-7llwrpw