# Review 143: [Short] Explaining grokking through circuit efficiency, 11.09.2023 https://arxiv.org/abs/2309.02390

**Paper: https://arxiv.org/abs/2309.02390v1**

יש תופעה מעניינת, הנקראת גרוקינג (grokking) המתרחשת (לפעמים) במהלך אימון של רשתות נוירונים. למעשה גרוקינג מחלק אימון של רשת נוירונים לשני שלבים עיקריים: הראשון הוא שלב השינון (memorization) כאשר הרשת משננת את הדוגמאות(overfit) ושלב ההכללה (למידה אמיתית).  
היום ב-#shorthebrewpapereviews אנחנו סוקרים מאמר המנסה להסביר למה התופעה הזו מתרחשת. למה בכלל רשת נוירונים ״משננת״ את הדוגמאות בתחילת האימון מרגע מסוים עוברת למשטר ההכללה, כלומר למידה אמיתית.   
  
קודם כל המחברים שמו לב שבסוף משטר השינון הדיוק של המודל הוא כמעט מושלם והלוס מאד נמוך ולמרות זאת מרגע מסוים הרשת מתחילה ללמוד להכליל משום מה למרות זאת. המחברים טוענים שהסיבה לכך היא הרשת מצליחה להוריד את הלוס עוד יותר (במצב של דיוק כמעט מושלם על הטריין סט) על ידי כך שהיא מתחילה להוציא חיזויים בטוחים יותר (עבור הקטגוריות הנכונות).   
  
אני חושב שראיתי מאמרים(לדעתי בנושא double descent) שטוענים שהירידה בלוס כזו במצב של ערכי לוס מאוד נמוכים הופכת את הפונקציה שהרשת ממדלת ליותר פשוטה (מאפס הרבה פרמטרים ומגדיל את האחרים). הם גם מתארים תופעה מעניינת נוספת: כאשר במשטר השינון מקפיאים את לוג'יטים (שמהם מחשבים את ההסתברות הקטגוריות) הנורמה של פרמטרי הרשת עולה עם הגדלת הדאטהסט.   
  
לעומת זאת במשטר ההכללה נורמה של פרמטרי הרשת לא תלויה בגודל הדאטהסט. בנוסף הם שמו לב שלפעמים ניתן לצפות תופעה שהם קראו לה semi-grokking כאשר המעבר למצב שינון מתרחש אבל הרשת מצליחה להגיע למצב ההכללה אבל לא מצליחה להגיע להכללה טובה ונתקעת איפשהו באמצע (semi-overfit). בקיצור מאמר מאוד מעניין – מומלץ בחום!