# Review 4: Deep Double Descent: Where Bigger Models and More Data Hurts

**Paper: https://arxiv.org/abs/1912.02292v1**

בין המחברים של המאמר נמנה Ilya Sutskever שהוציא מאמרים בעלי השפעה רבה בתחום.

תחום המאמר: הייתי מגדיר אותו כחקר תכונות אימון של רשתות הנוירונים.

תקציר המאמר בשני משפטים: המאמר מצא שיש מצבים שבהם הוספה של טריין דאטה כמו כן הגדלה של כמות המשקלים ברשת גורמת לעלייה בשגיאה טסט. הקטע הוא שהתופעה הזו הינה זמנית, כלומר אם אנחנו ממשיכים להגדיל את גודל הטריין דאטה/להגדיל כמות המשקלים הרשת שגיאת טסט מתחילה לרדת

תקציר המאמר:

הטענה העיקרית של המאמר אפשר לסכם באופן הבא: יש בעצם 3 רג'ימים (מקטעים) בציר סיבוכיות המודל מבחינת השפעתו על שגיאת הטסט. תופעה זו התגלתה עי״ מ.בלקין ושותפיו ב 2018(Belkin\_DD1, Belkin\_DD2, Belkin\_youtube). בלקין הוכיח ריגורוזית את קיומה למגוון סוגים של רשתות הנוירונים . במקטע הראשון המודל שנמצא במצב של under parameterized וכל הגדלה של סיבוכיות המודל כגון (הגדלה של מספר פרמטרים של המודל, אורך האימון, כמות הרעש בלייבלים וכדומה) גורמת לירידה של שגיאת הטסט. יש את רג'ים הביניים הנקרא במאמר אינטרוול קריטי כל הגדלה של סיבוכיות המודל גורמת לעלייה בשגיאת הטסט. במקטע השלישי על ציר הסיבוכיות שגיאת הטסט שוב מתחילה לרדת ככל שסיבוכיות המודל עולה(המודל overparameterized במקטע הזה). במאמר התופעה הזו נקראת double-descent phenomenon.

המאמר מגדיר שלושה מצבים בהם תופעה זו יכולה להתרחש ומראים את זה על מגוון רחב של ארכיטקטורת, דאטה סטים, שיטת אימון ושיטות אוגמנטציה:

Model-wise double-descent

מודל יותר גדול מזיק לשגיאת טסט. הם מראים שיש כאשר מספר הפרמטרים במודל נמצא בתחום מסוים (אינטרוול קריטי) הגדלת מספר הפרמטרים גורם לעלייה בשגיאת טסט, שמתחלפת בירידה כאשר מספר הפרמטרים משיג רף מסוים (יוצא האינטרוול הקריטי). כמובן כאשר בודקים את הזה שאר הפרמטרים של המודל נותרים קבועים

אי מונוטוניות של השפעת גודל טריין סט.

יש אינטרוול של גדלי טריין סט שבו הגדלתו גורמת לעלייה בשגיאת הטסט וכאשר עוברים אותו שגיאת הטסט חוזרת לרדת שוב. במאמר מגדירים את מה שנקרא סיבוכיות מודל אפקטיבית כמספר דגימות מקסימלי כאשר המודל מצליח להשיג שגיאת אימון קרובה לאפס (המאמר מגדיר את הקרבה הזו בצורה מתמטית ריגורוזית). אציין ש- EMC כמובן תלוי במודל. אז האינטרוול הקריטי במקרה הזה נמצא סביב EMC

Epoch Number Double Descent

המשמעות כאן שקיים אינטרוול בציר האפוקס שקיימת בו תופעה דומה ואחר כך שגיאת טסט מתחילה לרדת שוב. דרך אגב אחת המשמעות של התופעה הזו כאן היא שאימון ממש ארוך עשוי למנוע אוורפיטינג.

בנוסף הם מראים תופעה דומה גם לרמת הרעש הלייבלים אולם הם מציינים שיש במודלים מסוימים ההתנהגות של שגיאת הטסט באינטרוול קריטי יכולה להיות שונה.

דאטה סטים: CIFAR-10, CIFAR-100, IWSLT‘14 de-en

ארכיטקטורות רשת: CNNs, ResNets, Transformers

נ.ב. מאוד אהבתי את המאמר הזה, המסקנות שלו נראות מאוד מעניינות. עם זאת אני חושב שיש צורך לבדוק את המסקנות האלו על ארכיטקטורות רשת מורכבות יותר כמו כן על מגוון יותר רחב של דאטה סטים. מומלץ!!