המאמר היומי של מייק: 15.06.25  
Random Teachers are Good Teachers

מאמר עתיק אך מאוד מעניין לדעתי…

מאמר זה מציג ממצא מעניין ונוגד אינטואיציה באופן עמוק, המאתגר הנחות יסוד בתחומים של זיקוק ידע (knowledge distillation) ולמידה בפיקוח עצמי (self-supervised learning או SSL). המחברים מדגימים כי מודל "סטודנט" יכול ללמוד ייצוגים איכותיים על ידי זיקוק ידע מרשת "מורה" שהמשקולות שלה(המורה) אקראיות לחלוטין ואינן מאומנות. העבודה מפרקת את ״מערכת היחסים המורה-סטודנט״ הסטנדרטי כדי לבודד ולחקור דינמיקת למידה עם זיקוק ידע, וחושפת כי התהליך הדומה לרגולריזציה לא מפורשת (implicit regularization), שאינה תלויה בכך שהמורה מחזיק ב"ידע" ממשי כלשהו.

כאמור המטרה העיקרית של המאמר לחקור את דינמיקת זיקוק ידע. המאמר בגדול בודק שני משטרי זיקוק ידע: עם דאטה מתויג ובלי דאטה מתויג (ללא תוויות).

ליבת תרומתו של המאמר טמונה במערך הניסוי הפשוט והאלגנטי שלו. המחברים יוצרים תרחיש שנועד להסיר גורמים מפריעים (confounding factors) שונים שבדרך כלל מיוחסת להם ההצלחה של שיטות זיקוק ו-SSL.

היעדר "ידע אפל" (Dark Knowledge): רשת המורה מאותחלת באמצעות בצורה אקראית ולאחר מכן "מוקפאת". היא לעולם אינה נחשפת לדאטה האימון או לתוויות, כלומר היא אינה מכילה שום מידע נלמד כלשהו על המשימה או על התפלגות הדאטה. מטרת הסטודנט היא פשוט למזער את מרחק KL בין התפלגות הפלט שלו לבין הפלט הסטטי והאקראי של המורה (אבל לפעמים מוסיפים לוס של הסטודנט על הדאטה)

היעדר אוגמנטציית דאטה (Data Augmentation): בניגוד לשיטות רווחות ב-SSL, עבודה זו מסירה במכוון את כל אוגמנטציות מהדאטה. הדבר מבטיח שהאינווריאנטיות הנלמדת אינה נובעת מהטיות מובנות (inductive biases) מפורשות שמוכנסות על ידי טכניקות כמו חיתוך (cropping), היפוך (flipping) או שינויי צבע (color jittering).

היעדר תוויות (Labels): כל תהליך הזיקוק מתבצע ללא פיקוח (unsupervised) וללא תוויות. התוויות האמיתיות של הקטגוריות משמשות רק בסוף התהליך כדי להעריך את איכות הייצוגים הנלמדים באמצעות בדיקה לינארית (linear probing) כלומר אימון מסווג לינארי על גבי הייצוגים הקפואים ממקודד הסטודנט.

מסגרת מינימליסטית זו מבטיחה שכל אפקט למידה שנצפה ניתן לייחס אך ורק לאינטראקציה בין ארכיטקטורת המודל, התפלגות הדאטה הטבעית ודינמיקת האופטימיזציה מבוססת-הגרדיאנט של מערך המורה-סטודנט.

תוצאותיו של ניסוי הינן מאוד מפתיעות. רשת הסטודנט משיגה באופן עקבי ומשמעותי ביצועים טובים יותר מהמורה האקראי שלה במונחים של דיוק בבדיקה לינארית, וזאת על פני דאטהסטים רבים כמו (CIFAR-100, STL10, TinyImageNet) וארכיטקטורות שונות כגון (ResNet, VGG). כלומר הדאטה יותר חשוב מהמורה.

מציאה נוספת במאמר היא "תופעת הלוקאליות" (locality phenomenon): הקרבה ההתחלתית של משקולות הסטודנט לאלו של המורה היא קריטית ללמידה מוצלחת. המחברים חוקרים זאת על ידי אתחול משקולות הסטודנט כצירוף קמור של משקולות המורה ומשקולות אקראיות, הנשלט על ידי פרמטר לוקאליות α. כאשר α קרוב לאפס (כלומר, הסטודנט מתחיל כמעט זהה למורה), הלמידה היא המהירה ביותר והביצועים הסופיים הם הגבוהים ביותר (כאן הסטודנט הוא באותו הארכיטקטורה של המורה - זה לא תרחיש פרקטי אך מעניין לחקירה).

ממצא זה מרמז על גיאומטריה מעניינת של משטח הלוס. הפרמטריזציה של המורה,θ\_T, מהווה מינימום לוקאלי טריוויאלי שבו לוס הזיקוק הוא אפס. עם זאת, תהליך האופטימיזציה לא נשאר שם. במקום זאת, הוא מוצא מינימום לוקאלי סמוך ולא טריוויאלי, θ\_S, המתאים לאזור עם דיוק גבוה בהרבה (עבור דאטהסט אימון, כלומר, ייצוגים טובים יותר). הדמיות של הנוף חושפות כי המורה יושב לעתים קרובות בתוך "עמק א-סימטרי" חד. נראה כי מודל הסטודנט נמלט מהפתרון הטריוויאלי על ידי תנועה לעבר הצד ה"שטוח" יותר של עמק זה, אזור שהגיאומטריה שלו ידועה כבעלת מתאם להכללה (generalization) טובה יותר.

אולי הממצא העמוק ביותר הוא שצ'קפוינט של הסטודנט, שפותחה כולה ללא תוויות (רק זיקוק ידע), מציגה ייצוגים מבניים שבעבר סברו כי הן מופיעות רק בשלבים המוקדמים של אימון מפוקח. כלומר רשת התלמיד מתקרבת למורה (גם אקראי) כאשר יש ״בתוכה כרטיס זוכה״ - תת-רשת קטנה היודעת לעשות את אותו הדבר.

הופעת (Lottery Ticket Hypothesis): המחברים מצאו כי כבר צ'קפוינט של 1 של הסטודנט מכיל "כרטיס לוטו זוכה" תת-רשת דלילה שניתן לאמן מחדש ממשקולותיה ההתחלתיות כדי להשיג דיוק גבוה במשימה מונחית. לרשת המאותחלת באופן אקראי אין תכונה זו; היא מופיעה ברשתות מונחות רק לאחר מספר אפוקי אימון. הדבר מרמז כי זיקוק ממורה אקראי מנחה את הרשת לתצורה פרמטרית שכבר מובנית ללמידה יעילה.

קישוריות מצבים לינארית (Linear Mode Connectivity): בד״כ כאשר משתמשים בצ'קפוינט מוקדם של הסטודנט כאתחול למספר ריצות אימון מונחות (כל אחת עם מיני-באצ'ים שונים), הפתרונות המתקבלים בד״כ הם "מקושרים לינארית". משמעות הדבר היא שניתן לבצע אינטרפולציה לינארית במרחב המשקולות בין כל שניים מהפתרונות הללו מבלי לקבל לוס גבוה בדרך. יציבות זו מצביעה על כך שהסטודנט כבר התכנס ל"אגן רחב ושטוח" במשטח הלוס ״הפוקח״, ובכך עוקף למעשה את השלב הכאוטי הראשוני של אופטימיזציה מפוקחת.

### מסקנה

המאמר טוען שהצלחתן של מסגרות מורה-סטודנט אינה מיוחסת אך ורק להעברת "ידע אפל" ממורה מאומן. במקום זאת, המאמר חושף כי הרגולריזציה הלא מפורשת הנוצרת מדינמיקת הלמידה היא מנוע רב-עוצמה ללמידת ייצוגים חזקים בפני עצמה. על ידי הדגמה שרשת לפתח ייצוגים מבנים מתוחכמים (כמו "כרטיסי לוטו זוכה") מאות אקראי לחלוטין, המחברים מאלצים הערכה מחדש של המנגנונים הבסיסיים מאחורי זיקוק-עצמי ולמידה -SSL. העבודה מספקת מצע ניסויים לעבודה עתידית שמטרתה להסיר את המסתורין מעל "השלב המוקדם" של אימון רשתות נוירונים והגיאומטריה המורכבת של משטח הלוס שלהן.

https://arxiv.org/abs/2302.12091