# Review 37: Teaching with Commentaries

**Paper: https://arxiv.org/abs/2011.03037v2**

פינת הסוקר:

המלצת קריאה ממייק: מומלץ לאוהבי מטה-למידה ובעלי רקע בחדו"א2 מתקדם.

בהירות כתיבה: בינונית.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: רקע טוב בתחום מטה-למידה, חדו"א ברמה גבוהה.

יישומים פרקטיים אפשריים: ניתן להשתמש בגישה זו למשל לזיהוי דוגמאות המשפיעות ביותר על הביצועים או איתור פאטצ'ים בתמונות מהדאטהסט החשובים למשימה במהלך האימון של הרשת.

פרטי מאמר:

פורסם בתאריך: 5.11.20, בארקיב.

יוצג בכנס: ICLR 2021.

תחומי מאמר:

שיטות אימון של רשתות נוירונים.

שיטות מטה למידה (meta-learning) בתחום רשתות הנוירונים.

כלים מתמטיים, מושגים וסימונים:

משפט הפונקציה הסתומה.

חישוב נגזרת של פונקציה וקטורית דרך ההופכית של מטריצת הסיאן (hessian).

קירוב ניומן (neumann) לחישוב הופכית של אופרטור (מטריצה) לינארי.

רשת לומדת פנימית (inner student network).

רשת מלמדת (נקראת הרשת המפרשנת במאמר - commentary network).

אימון פנימי/חיצוני (inner/outer optimization).

מטה-אימון ,(meta-training).

תמצית מאמר:

כמו שאתם בטח יודעים, למרות הפעילות המחקרית האינטנסיבית בתחום הלמידה העמוקה עדיין קיימות לא מעט סוגיות פתוחות בנושאי אימון, רגולריזציה והבנה של מה שקורה בתוך רשתות נוירונים עמוקות. השאלות האלו נוגעות בסוגיות בסיסיות כמו: איך לאמן רשתות בצורה יותר מהירה, איך להקטין את כמות הדאטה הנדרשת לאימון, איך לשפר את יכולת הכללה רובסטיות של הרשתות.

אחת הגישות המעניינות שהוצעה לאחרונה שמנסה לתת מענה לשאלות אלו נקראת "לומדים ללמד״ (learning to teach) המציע לבנות מנגנון חיצוני (רשת בדרך כלל) בשביל לספק לרשת הלומדת תובנות (נקרא לזה גם מטה-מידע בהמשך) לגבי המשימה תוך כדי תהליך האימון. למשל מנגנון כזה יכול לבצע משקול של דוגמאות במטרה לעזור לרשת הלומדת ״לרכז את המאמץ״ בדוגמאות החשובות. דוגמא אחרת של גישה זו יכולה להיות רשת עזר ה״מייעצת״ איך לבנות דוגמאות (למשל ע״י ערבוב של דוגמאות מדאטה סט) הגורמים לרשת הלומדת לבנות ייצוג חזק של דאטה.

מאמר זה מציע מסגרת כללית לגישה זו(הנקראת למידה עם פרשנויות) ומציע תהליך אחיד להפקה של מטה-מידע (פרשנות) עי״ רשת חיצונית (מפרשנת) N\_com, תוך כדי ״הסקת מסקנות״ העולות בתהליך האימון של רשת N\_st (נקראת הרשת הלומדת) על סט האימון. אז בואו נבין איך כל זה עובד בעצם? נניח שאנו רוצים למצוא איזושהי טרנספורמציה (לדוגמא משקול/ערבוב) של דוגמאות בדאטה סט במטרה לשפר את הביצועים של הרשת הלומדת N\_st והמטרה של הרשת המפרשנת N\_com הינה למצוא את הטרנספורמציה הזו וזה למעשה מהווה הפלט שלה Out\_com. במקרה זה תהליך האימון מכיל את השלבים הבאים:

אופטימיזציה פנימית: עבור סט משקלים נתון W\_com של הרשת N\_com, מאמנים את N\_st (כמה איטרציות של GD על משקלי W\_st). במקרה הזה מפעילים טרנספורמציה Out\_com המופקת עי" (N\_com, W\_com) על הדאטה של סט האימון ומאמנים את N\_st עליו. הפלט של השלב הזה הוא המשקלים W\_st של N\_st.

אופטימיזציה חיצונית: מחשבים את הלוס של N\_st עם סט המשקלים W\_st מהשלב הקודם על סט ולידציה שעובר טרנספורמציה הניתנת עי" (N\_com, W\_com). כאן מאמנים את N\_st על משקלי W\_com כלומר מבצעים כמה איטרציות של GD אבל הפעם על למשקלי W\_com.

חוזרים על הצעדים אלו T פעמים כאשר T זה מספר האיטרציות של מטה אימון.

הסבר של רעיונות בסיסיים:

קודם כל נציין כי הגרדיאנט של משקלי N\_com משערך את השינוי בלוס של N\_st ביחס לשינוי במשקלים של N\_com. אבל צריך לזכור שבשביל לחשב את הלוס של האופטימלי של N\_st עבור משקלי N\_com נתונים, המשקלים של N\_st עוברים כמה איטרציות (אולי די הרבה) של GD במטרה למזער את הלוס שלה. לכן כדי לחשב את הגרדיאנט של הלוס של N\_st לפי משקלי W\_com צריך ״לגלגל את כל האיטרציות על משקלי N\_st״ ל- W\_com שזה יכול להיות די כבד חישובית כאשר N\_com הינה רשת גדולה.

גם אם נחליט להשתמש רק באיטרציה אחת בתהליך האופטימיזציה הפנימית (זה מה שעשו במאמר) עדיין של לנו בעיה עם חישוב הגרדיאנט של הלוס לפי W\_com. הבעיה הזו נובעת מהעובדה שגרדיאנט זה שווה למכפלה של הגרדיאנט של הלוס לפי W\_st (שזה ניתן לחשב אותו בצורה הסטנדרטית של גזירת הלוס של רשתות) והנגזרת של וקטור משקלים W\_st לפי לוקטור משקלים W\_com . נזכיר ש W\_st תלוי ב- W\_com בצורה לא מפורשת כי בשלב האופטימיזציה הפנימית W\_st מחושב על הדאטה סט אחרי הפעלת עליו טרנספורמציה המוגדרת עי״ W\_com. נגזרת זו היא בעצם מטריצה (W\_com ו- W\_st הם וקטורים) שמימדיה עלולים להיות די גבוהים. נניח ש N\_st ו- N\_com הם רשתות לא גדולות בגודל של מיליון משקלים אז הנגזרת הזו תהיה מטריצה בגודל מיליון על מיליון ותידרש כמות זיכרון עצומה בשביל לאחסן אותה.

לכן המאמר מציע להשתמש במשפט הפונקציה הסתומה עבור הנגזרת של הלוס (של השלב החיצוני L\_out) לפי W\_st. משפט זה מאפשר לתאר את הנגזרת הבעייתית עי״ מכפלה של הופכית הסיאן של לוס של השלב הראשון L\_in לפי W\_st והמטריצה של הנגזרות המעורבות לפי W\_com ו- W\_in של L\_in. למי שלא זוכר הסיאן זה מטריצה המורכבת מהנגזרות שניות של L\_in לפי הזוג של רכיבים של W\_com ו- W\_st. צריך לזכור שהפירוק לעיל מתקיים בסביבת נקודה שבה הנגזרת של L\_in לפי מW\_st מתאפסת. העובדה שלא ניתן למצוא אותה במדויק וזה יכול להשפיע בצורה שלילית על התהליך המטה-למידה.

גם אחרי הפירוק הזה יש לנו בעיה - והיא טמונה בחישוב של הופכית של ההסיאן של L\_in לפי W\_st. אפילו עבור W\_st בגודל יחסית לא גדול חישוב ההופכית (ולפעמים ההסיאן עצמו) יכול להיות מאוד כבד וידרוש כמות עצומה של הזיכרון והזמן. בשביל להקל על ההיבט החישובי משתמשים בקירוב נוימן עבור ההופכית מוכפלת בגרדיאנט של L\_in לפי W\_st (מאמר של לוריין) תוך שימוש בצורת עדכון של GD .דרך אגב נוסחת נוימן מאתרת הופכית של אופרטור לינארי בתור טור אינסופי של החזקות שלה (מוזזות במינוס מטריצה היחידה). במאמר משתמשים רק באיבר אחד של קירוב זה.

הישגי מאמר:

במאמר מראים 3 דרכים להשתמש בגישה זו לשיפור תהליך האימון של רשתות נוירונים:

חישוב של משקול על דוגמאות מסט האימון עי״ W\_com (דוגמאות עם משקל גבוה משפיעים יותר על הלוס). מעניין שהם גם בדקו את הביצועים של השיטה שלהם בתרחיש למידת few-shot שזו משימת מטה-למידה קלאסית. המטרה בלמידת few-shot היא לאמן רשת חיצונית (מטה) על מספר משימות (שנלמדות בפועל עי״ הרשת הפנימית) במטרה ללמד אותה להפיק תכונות משותפות של כל המשימות (שימו לב שזה מקרה פרטי של הפרדיגמה הכללית שהוצעה במאמר). כאשר מגיע משימה חדשה הרשת החיצינות מסוגלת לכייל את את עצמה עם כמות קטנה של דאטה במשימה הזו. MAML זו אחת דרכים לפתור בעיה זו והיא מאמנת רשת חיצונית לאתחול המשקלים של הרשת הפנימית שתאפשר לה להגיע לביצועים טובים על משימה חדשה במספר איטרציות GD קטן. אז הם מראים שהשילוב של MAML עם משקול דוגמאות הנבנה עי״ W\_com גורם לשיפור ביצועים משמעותי. המאמר מראה שיפור בביצועים עבור הדאטה סטים MiniImageNet ו- CUB200-2011.

בנייה של מקדמים ערבוב אופטימליים עבור הדוגמאות (בדומה ל mixup). כאן דוגמא מעורבבת נבנית כסכום קמור של שתי דוגמאות: x\_mix = ax\_1 + (1 - a)x\_2 כאשר מטרת W\_com הינה לחשב את מקדמי a האופטימליים לביצועי משימת סיווג(מאמר mixup המקורי מגריל אותם מהתפלגות בטה). מעניין שכאן ״למידה עם פרשנויות״ מנצחת את mixup ב- CIFAR10 ומציגה ביצועים קצת פחות טובים ממנה על CIFAR10 (דאטה סט יותר קטן).

חישוב מסכות על תמונות לשיפור הפיצ'רים המופקים עי״ רשת. כאן W\_com בעצם מחפשת אזורים ״חשובים״ בתמונה שכדאי לרשת הלומדת להתרכז עליהם. המאמר מראה באופן ויזואלי שהמסכות שהוא מפיק אכן מתמקדות באזורים החשובים של תמונות ומראים באופן כמותי את עדיפותן על פני שיטות אחרות לבניית מסכות.

נ.ב.

מאמר מציע מסגרת כללית לשיפור של תהליך למידה של רשתות נוירונים שניתן להשתמש בה למגוון רחב של משימות של הלמידה העמוקה. הגישה שלהם גם עוזרת להפיק תובנות חדשות תוך כדי תהליך האימון של רשתות. אני מני שעוד נשמע על שימושים רבים של גישה זו…