לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה
היום בחרתי לסקירה את המאמר שנקרא:
Sharpness-Aware Minimization for Efficiently Improving Generalization
שהתפרס בארקיב רק לפני שבוע.

תחום מאמר: חקר שיטות אופטימיזציה לאימון של רשתות נוירונים

מאמר הוצג בכנס: טרם ידוע

תמצית מאמר: המאמר מציע ניסוח חדש לבעיית אופטימיזציה לאימון רשתות נוירונים. במקום מציאת וקטור משקלים שבו פונקציית לוס מקבלת ערך מינימלי (לסט דוגמאות נתון), המאמר מציע בעיית אופטימיזציה שהמטרה בה למצוא וקטור משקלים שהערך "המקסימלי של פונקצית לוס בסביבתו של וקטור זה הוא מינימלי". כלומר הם בפועל מגדירים וקטור משקלים אופטימלי בתור לא כזה שמביא למינימיזציה של פונקצית לוס אלא כזה שפונקציית לוס מקבלת ערכים "הכי נמוכים" בסביבתו (משטח לוס בסביבת נקודה פחות חדה). בנוסף הם מוכיחים באופן ריגורוזי שהפתרון בעיית אופטימיזציה שהם מציעים (הנקרא sharpness aware minimization -SAM) תורם באופן חיובי ליכולת הכללה של המודל.

רעיון בסיסי: כמו שאתם בטח יודעים הרוב המוחלט של רשתות נוירונים מודרניות הינן overparameterized בצורה מאוד משמעותית. במקרה כזה אופטימיזציה של משקלי רשת על סמך ערך של פונקצית לוס בנקודה בלבד(!!) עלול להוביל למודלים בעלי יכולת הכללה נמוכה. הסיבה המרכזית לכך זה מבנה גיאומטרי מאוד מורכב ולא קמור של משטח הלוס. הדוגמא הקלאסית לכך היא המקרה שבו המינימום שהתקבל הינו "חד" מאוד כלומר אפילו בסביבתו המאוד קרובה הערכים של פונקציית לוס הינם גבוהים משמעותית מערך המינימום. נקודה זו יכולה להיות תוצאה של דאטה רועשת ותוביל למודל עם יכולת הכללה נמוכה. המאמר מציע פתרון למצב זה עי" ניסוח בעיית אופטימיזציה שמתחשבת לא רק בערך של פונקצית לוס בנקודה אלא לוקחת בחשבון את ערכי הלוס בסביבתה. כלומר הניסוח המוצע (SAM) לוקח בחשבון גם את התכונות הגיאומטריות של משטח הלוס בסביבות הנקודה באופן מפורש.

תקציר מאמר: קודם כל בעיה זו ידועה כבר הרבה זמן ויש מספר רב של השיטות המסנות למזער את השפעתה השלילית על מודלים. הפתרונות שהוצעו אפשר לחלק לשתי משפחות עיקריות: הראשונה של בחירת האופטימייזר (Momentum, RmsProp, ADAM, עצירה מוקדמת וכדומה) והשנייה שינויים בתהליך האימון עצמו (באצצ'נורם, עומק סטוכסטי, אוגמנטציות שונות והרבה אחרים). כל השיטות האלו מנסות לפתור את אותה בעיית אופטימיזציה בדרכים שונות. לעומתם המאמר הנסקר מציע לשנות את בעיית אופטימיזציה עצמה (!!!) ומוכיחים שזה משפר את יכולת הכללה של המודל המתקבל כתוצאה מכך.

<u>פרטים טכניים:</u> פונקציית לוס המוצעת L מכילה שני איברים - הראשון זה הלוס המקסימלי בסביבה קטנה של הנקודה w (גודלה של סביבה זו הינו היפרפרמטר) והשני איבר רגולריזציה סטנדרטי עם נורמת Lp. (זה דומה אופטימיזציית proximal point). מעניין כי ניתן לרשום את הפונקציה הזו בצורה הבאה: סכום של ההפרש המקסימלי בין הנקודות בסביבת w ולבין ערך הלוס בנקודה w (הפרש זה נקרא במאמר חדות - sharpness) ואיבר רגולריזציה חדש שהוא סכום של נורמת Lp של w ערך הלוס בנקודה w.

<u>ההיבט התיאורטי</u>: במאמר מוכיחים עבור טריין סט נתון שהלוס של SAM בכל נקודה w מהווה חסם עליון על הלוס על ה- population (שממנה הטריין סט הדגם) בהסתברות גבוהה.(המשפט טיפה יותר כללי ועובד על משפחה יותר רחבת של פונקציות רגולריזציה). כמובן הכל תחת תנאים טכניים על התפלגות שממנה הדאטה סט נדגם.

בעצם המשפט הזה אומר שפתרון בעיית SAM מוביל למודל בעל יכולת הכללה טובה. ההוכחה די לא טריויאלית ומערבת חסמי PAC בייסיאניים (מוכללים)

<u>פתרון בעיית SAM</u>: קודם כל משתמשים בקירוב טיילור מסדר ראשון בשביל למצוא את הנקודה בסביבת של w שעבורה בלוס הוא מקסימלי. אחר כך הבעיה בנידון מתורגמת לבעית נורמה דואלית שיש לה פתרון מפורש. אחרי שמוצאים את הערך המקסימלי בסביבת w בעצם נותרת לנו בעיית אופטימיזציה רגילה שפותרים אותה דרך gradient descent קלאסי. מכיוון שהפתרון של בעיית המקסימום מכיל נגזרת לפי w, הביטוי של גרדיאנט מכיל את הסיאן (hessian) של פונקצית הלוס שזה מאוד לא כיף כאשר w הוא ממימד של מאות מיליונים אבל הם טוענים מה שמופיע בביטוי זה מכפלה שלו בוקטור מסיום ואפשר לחשב זאת ללא חישוב של ההסיאן). בסופו של דבר האלגוריתמים שלהם הינה הכללה מסוימת של GD שניתן להריץ אותה עם כלי גזירה אוטומטיים כמו TF או PyTorch

הישגי מאמר: הם הראו שהאלגוריתם שלהם מציג ביצועים טובים יותר על שיטות אופטימיזציה שונות ומגוונות (שבהן סוגים שונים אוגמנטציה, אופטימייזרים שונים ועוד) על מגוון מאוד רחב של דאטה סטים ארכיטקטורות רשת שונות. בכל השוואה הם פשוט החליפו את האופטימיזציה מקורית ב SAM והשוו את הביצועים על הטסט סט.

<u>לייבלים רועשים:</u>SAM הציג שיפור ניכר כאשר מופעל באימון על דאטה סטים עם לייבלים רועשים. בעצם זה לא מפתיע כי החוזק העיקרי של האלגוריתם הוא מניעת התכנסות למינימום "חד" ונוכחות לייבלים רוכשים בכמות ניכר עלול להוביל בקלות למינימומים כאלו באלגוריתמים אופטימיזציה קלאסיים.

<u>מבנה ההסיאן בסביבת נקודת אופטימום</u>: בשביל לאשש את ההנחות לגבי יכולות של SAM במניעת המינימומים חדים הם בדקו את הערכים העצמיים (המקסימלי וגם היחס בין המקסימלי לכמה הבאים) של ההסיאן בנקודות אופטימום שנמצאו ע"י SAM מול אלגוריתמים אחרים. הרי ידוע שככל שהמינימום יותר חד, יש להיסאין ערכים עצמיים גבוהים יותר ויחס בין ע"ע המקסימלי להבאים גבוה יותר גם כן. אז הם הראו ש- SAM מוריד את שני הצדדים האלו בצורה מאוד משמעותית.

דאטה סטים: CIFAR10, CIFAR100, Flowers, Stanford_cars, Birdsnap, Food101, Oxford_IIIT_Pets, FGVC_Aircraft, Fashion-MNIST וכמה אחרים

ארכיטקטורות רשת שנבדקו: Wide-ResNet-28-10, Shake-Shake , EffNet, TBMSL-Net, Gpipe וכמה אחרים

לינק למאמר:<u>מאמר</u>

לינק לקוד: לא מצאתי

נ.ב. לדעתי הצנועה זה מאמר מאוד חשוב המציע שיטה מאוד מעניינת לשיפור יכולת הכללה של הרשתות. יש לה פוטנציאל רציני להיכנס לארגז כלים סטנדרטי לאימון רשתות. התרשמתי גם המשוואות הרבות והמגוונות מול שיטות אחרות שנעשו במאמר.