ערב טוב חברים ושנה טובה, היום אנחנו עם המהדורה הראשונה של הפינה DeepNightLearners בשנה העברית החדשה עם סקירה קצרה של מאמר בתחום של הלמידה העמוקה.

היום בחרתי לסקור מאמר הנקרא:
Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks

שיצא לפני קצת יותר משנה
תחום מאמר: חקר שיטות אופטימיזציה לאימון של רשתות נוירונים

תקציר מאמר ללא מושגים מתמטיים כבדים: המאמר טוען(ומוכיח) שעבור רשתות overparameterized שעצירה מוקדמת של gradient descent(GD) תורמת לרובסטיות של הרשת נגד לייבלים מורעשים. המאמר בעצם מוכיח שבאיטרציות הראשונות של GD מצליח ללמוד את הלייבלים הנכונים כאשר אם ממשיכים להריץ אותו הרשת לומדת גם את הלייבים המורעשים שזה כמובן פוגע בשגיאת טסט (עוצמת ההכללה יורדת).

תקציר מאמר: למרות שמסקנות המאמר די ברורות וקלות להבנה הוכחתם הריגורוזית כוללת שימוש בכלים מתמטיים לא פשוטים והגדרות מתמטיות לא טריוויאליות. עקב כך אתמקד בפוסט הזה בהסבר התנאים והטענות של המשפטים האלו.

ארכיטקטורה של הרשת ואתחול: כל התוצאות במאמר הוכחו לרשת דו-שכבתית (1 שכבה hidden) כאשר השכבה השנייה הינה קבועה ולא נלמדת (מאמנים רק את המשקלים בשכבה הראשונה). הפלט של השכבה השנייה הינו סקלר. אתחול של המשקלים הינו גאוסי (כמו ברוב המאמרים התיאורטיים ברשתות נוירונים).

הנחות על דאטה סט: הנחה נוספת במאמר היא שהנקודות בדאטה סט הלא מורעש המתאימים לקלאסים שונים הם מספיק רחוקות אחד מהשני מצד אחד ומאידך הנקודות ששייכים לאותם הלייבלים (קלאסים) מספיק קרובים (פרמטר אפסילון 0) לסנטרויד של הקלאס (בעצם הגדרה במאמר טיפה מורכבת יותר ומגדירה נקודות השייכות לכל קלאס כאיחוד של כמה קלאסטרים). הלייבלים מוגדרים כמספרים ממשים (!!) כאשר גם הם מספיק רחוקים אחד מהשני (פרמטר דלתא במאמר). דאטה סט בעל תכונות אלו נקרא באופן לא מפתיע clusterable dataset. בואו נחשוב מה ההיגיון הטמון בהגדרה הזו. הרי ברור שכל שהסנטרואידים של הנקודות השייכים לקלאסים שונים (בואו נקרא להם label set להבא) קרובים אחד לשניה, נהיה יותר קשה לאמן רשת נוירונים (או כל קלאסיפייר אחר) המבדיל ביניהם (לקלסטר אותם נכון). הדאטה סט המורעש מוגדר כ- clusterable dataset כאשר אחוז נתון של הנקודת השייכות לכל קלאס שונה ללייבל אחר.

פונקציית לוס: הרשת מאומנת לשערך את ערך הלייבל של הנקודה כאשר פונקציית לוס הינה הפרש בריבוע בין פלט של הרשת לבין הלייבל של הדוגמא קרי יש לנו כאן בעיית רגרסיה עם לוס ריבועי ולא בעיית סיווג.

מטריצת קוריאנס של רשת נוירונים :זה מושג מרכזי במאמר שבעזרתו מוכיחים אם כל הטענות העיקריות. כמו שכבר הזכרנו אם יש לנו שתי נקודות קרובות עם לייבילים שונים ולא מורעשים(!!) אז רשת "צריכה לעבוד קשה" בשביל להבחין ביניהם. אז המטרה של מטריצה זו היא לכמת את היכולת הזו עבור רשת נוירונים נתונה וסט מרכזי קלאסטרים נתון עי" condition number של המטריצה הזו. אזכיר ש condition number של מטריצה מוגדר כיחס בין ערך העצמי הגדול ביותר לבין הקטן ביותר. ככל ש- condition number של מטריצת קוריאנס של הרשת נמוך יותר אז קל יותר לרשת להבחין בין קלאסטרים שונים. האינטואיציה כאן היא די פשוטה: תניחו שיש שני מרכזי label sets שונים באותה נקודה. קל לראות שבמקרה זה המטריצה תהיה עם

שורות תלויות כלומר יהיה לה ע״ע 0. במקרה הזה condition number שלה יהיה אינסוף שמסתדר עם טענה המנוסחת למעלה.

הערה: מטריצה הזו בעצם מהווה מטריצה קרנל אמפירית של הרשת (מטריצת קרנל של רשת נוירונים מודדת (בקירוב) את השפעת צעד אחד גרדיינט דסצנט על המשקלים ללוס של הרשת - תחת קירוב לינארי והיא מדויקת יותר ככל שגדלי השכבות גדולות יותר)

טענה עיקרית 1: בהינתן דאטה סט מורעש עם אחוז לייבלים מורעשים נמוך מספיק, וגודל השכבה הנלמדת מספיק גדול (,K * 4^condition number)O ^2 ,קיים קצב למידה שאחרי מספר צעדי GD (שהוא גם תלוי ב condition number) הרשת תלמד לזהות נכון את הלייבלים של כל (!!) הדוגמאות בדאטה סט. K מסמן את מספר הקלאסטרים (אזכיר של כל Label set מורכב מכמה קלאסטרים). מספר צעדי GD עד ההגעה לזיהוי מלא של כל הנקודות הלא מורעשות הוא מסדר K. בנוסף המרחק בין משקלי האתחול לבין המשקלים בכל(!!) האיטרציות לפני (עד ההגעה למצב שהרשת מזהה נכון את כל הדוגמאות עם הלייבלים הלא מורעשים) יהיה נמוך מספיק כלומר המשקלים של הרשת לא ישתנו ״יותר מדי״ במהלך האימון עד הזיהוי הנכון של הלייבלים הלא מורעשים שזה.

טענה עיקרית 2: עכשיו נשאלת השאלה מה קורה אם אנחנו לא עוצרים את האימון מוקדם וממשיכים לאמן את הרשת עם GD. המשפט העיקרי השני במאמר נותן מענה לשאלה הזו. המשפט הזה מוכיח שתחת אותם התנאים על ארכיטקטורת הרשת ועל הדאטה סט, בשביל לתת דיוק של 100% על דאטה סט עם לייבל מורעש אחד (לזהות נכון את כל הדוגמאות כולל זה עם הלייבל המורעש) המרחק שהמשקלים של הרשת צריכים לעבור (קרי המרחק בין המשקלים ההתחלתיים לבין אלו של הרשת המאומנת) צריך להיות לפחות 0δ/ε, כאשר המונה מהווה חסם על המרחק בין ערכי ההלייבלים השונם, המחנה מתאר את הרדיוס המקסימלי של הקלאסטר. ככל שהקלאסים יותר גדולים (המחנה עולה) אז המרחק מתקצר (הקלאסטרים יותר מרוחים וקרובים אחד לשני) וכאשר המרחק בין ערכי הלייבלים(המונה עולה) המרחק ה;ה מתארך (״מכריחים את הרשת לטעות גם כשיש לה ביטחון גבוה״). הכל תחת אתחוך גאוסי של המשקלים.

כלים מתמטיים המשמשים להוכחות: הם השתמשו ביקוביען של הרשת בשביל לנתח את מטריצת קוואריאנס של הרשת לדאטה סט נתון. הם הציגו את השארית המורעשת (הפרש בין פלט הרשת לבין הלייבל) כסכום של השארית הנקיה והרעש באותו לייבל מורעש. לאחר מכן הם הוכיחו שהשארית הנקיה ״מכוסה״ עי״ תת מרחב בעל מימד גבוה של מרחב המשקלים שמאפשר ללמוד אותם במהירות (גרדיאנטים חזקים בגדול) כאשר השאריות ״מכוסות״ עי״ תת-מרחב קטן שמקשה על האימון של הלייבלים המורעשים (גרדיאנטים חלשים). לדעתי זו מסקנה מאוד חזקה.

דאטה סטים: MNIST, CIFAR10

לינק למאמר: מאמר
לינק לקוד: לא פורסם