המאמר היומי של מייק: 19.08.25

Training Transformers with Enforced Lipschitz Bounds

בעולם הלמידה העמוקה, אנו מודדים את ביצועי המודל על כמה בנמארקים פופולריים ושמחים כאשר המודל מפגין ביצועים גבוהים עליהם. עם זאת, מתחת לפני השטח של הישגים מרשימים אלה מסתתרת בעיה עיקשת שלעיתים קרובות מתעלמים ממנה: חוסר יציבות. כל מי שאימן מודל טרנספורמר גדול נתקל בוודאי בתסכול של גרדיאנטים מתפוצצים או נעלמים, בצורך בתכנון עדין של קצבי למידה, ובערך ה-"NaN" המסתורי בפונקציית הלוס שיכול לשבש ריצת אימון שלמה. סוגיות אלה מצביעות על חוסר שליטה יסודי בהתנהגות המודל.

המאמר שנסקור היום מציע פתרון נחמד לבעיה זו. במקום להסתמך על אוסף של טריקים אמפיריים, המחברים מציגים מתודולוגיית אימון חדשנית שאוכפת תכונה מתמטית הידועה בשם תנאי ליפשיץ. גישה זו לא רק  מרסנת את חוסר היציבות של הטרנספורמר, אלא גם מובילה לשיפור ביכולת ההכללה ורובסטיות של המודל. בואו נצלול לעומק החידושים המרכזיים של עבודה מרתקת זו.

בבסיסו, תנאי ליפשיץ הוא מדד ל"חלקוּת" או ל"רגישות" של פונקציה. פונקציה עם קבוע ליפשיץ קטן אינה יכולה להשתנות מהר מדי; שינויים קטנים בקלט יובילו רק לשינויים קטנים בפלט. על ידי אכיפת חסם ליפשיץ על רשת נוירונים, אנו למעשה מציבים "מגבלת מהירות" על מידת השינוי בפלט המודל כתגובה להפרעות בקלט שלו.

זהו רעיון רב עוצמה. בהקשר של טרנספורמרים, משמעות הדבר היא שאנו יכולים לשלוט ברגישות של כל רכיב במודל, ממנגנון ה-attention ועד לשכבות FFN לשליטה מדויקת ועדינה זו יש השלכות עמוקות על יציבות האימון וביצועי המודל. כדי לאכוף את תנאי ליפשיץ, המחברים מציעים סדרה של שינויים חדשניים בארכיטקטורת הטרנספורמר הסטנדרטית. לא מדובר בתיקונים קלים, אלא בתכנון מחדש עקרוני של רכיבי הליבה של המודל:

שכבות עם נרמול ספקטרלי (Spectrally Normalized Layers): המחברים מיישמים נרמול ספקטרלי על מטריצות המשקולות הן במנגנון ה-attention והן ב-FFN. טכניקה זו נבחרה בשל דיוקה המתמטי: הנורמה הספקטרלית של מטריצת משקולות שווה בדיוק לקבוע ליפשיץ של אותה שכבה לינארית. הדבר מאפשר שליטה ישירה והדוקה ברגישות המודל בכל שלב.

בלוקים של רשת FFN שהם 1-ליפשיץ: חידוש מרכזי הוא האופן שבו המאמר מטפל באי-הלינאריות של ה-FFN. המחברים מראים כיצד לבנות את כל בלוק ה-FFN כך שיהיה 1-ליפשיץ באמצעות פונקציות אקטיבציה סטנדרטיות כמו ReLU או GeLU. הדבר מושג על ידי שילוב של מטריצות משקולות מנורמלות ספקטרלית עם טיפול בפונקציית האקטיבציה, מה שמבטיח שהטרנספורמציה השלמה בתוך הבלוק עומדת באילוץ ליפשיץ המחמיר.

חיבורי שארית (Residual Connections): המחברים מספקים גם ניתוח עמוק של חיבורי השארית, שהם יסודיים בארכיטקטורת הטרנספורמר. הם מדגימים כיצד לשנות את קנה המידה (scaling) של נתיבי השארית כראוי כדי להבטיח שהוספתם אינה מפרה את תכונת הליפשיץ של המודל כולו. הרכבה זהירה זו של רכיבים חסומים באופן מוכח היא שמאפשרת לרסן את ארכיטקטורת הטרנספורמר כולה.

חידושים ארכיטקטוניים אלה, יחדיו, יוצרים סוג חדש של טרנספורמר שהוא, מעצם תכנונו, יציב וממושמע יותר מקודמיו. היתרונות של הטרנספורמר מרוסן-ליפשיץ ניכרים מיד במהלך האימון. המחברים מדגימים שהמודל שלהם יציב באופן יוצא דופן, אפילו ללא צורך בנרמול שכבה (Layer Normalization), רכיב שנחשב לעיתים קרובות חיוני עבור טרנספורמרים סטנדרטיים.

יציבות זו מאפשרת תהליך אימון פשוט וחזק יותר. המחברים מראים שניתן לאמן את המודל שלהם עם קצבי למידה גדולים יותר ושהוא פחות רגיש לבחירת היפר-פרמטרים. הדבר לא רק הופך את תהליך האימון ליעיל יותר, אלא גם פותח דלת לאפשרויות חדשות להגדלת מודלי טרנספורמר. היתרונות של אכיפת חסמי ליפשיץ חורגים מעבר ליציבות האימון בלבד. המחברים מדגימים גם שהמודל שלהם מציג יכולת הכללה ועמידות משופרות:

יכולת הכללה טובה יותר: אילוץ ליפשיץ פועל כצורה של רגולריזציה לא מפורשת רבת עוצמה, המונעת מהמודל לבצע אוברפיט לדאטה האימון. הדבר מוביל לביצועים טובים יותר על נתונים שלא נראו בעבר.

עמידות מוגברת למתקפות אדברסריאליות (Adversarial Attacks): על ידי הגבלת רגישות המודל להפרעות קטנות בקלט, אילוץ ליפשיץ הופך את המודל לעמיד יותר באופן אינהרנטי למתקפות אדברסריאליות. המחברים מראים שהמודל שלהם חסין למתקפות אלה באופן משמעותי יותר מטרנספורמרים סטנדרטיים.

https://arxiv.org/abs/2507.13338

כלל האחוז האחד לריפוי אמנזיה של בינה מלאכותית: צלילת עומק