המאמר היומי של מייק: 25.08.25

Pulling Back the Curtain: Unsupervised Adversarial Detection via Contrastive Auxiliary Networks

מודלים של למידה עמוקה הם המנועים של הבינה המלאכותית המודרנית, אך יש להם פגיעות קריטית: התקפות אדברסריאליות (adversarial attacks). התקפות אלו מנצלות חולשה זו על ידי החדרת הפרעות זעירות, כמעט בלתי נראות, לקלטים, הגורמות למודלים לבצע תחזיות שגויות לחלוטין, עם השלכות שעלולות להיות קטסטרופליות. במשך שנים, קהילת הבינה המלאכותית לכודה במרוץ חימוש, כאשר הגנות רבות הן יקרות חישובית ולעיתים קרובות מוגבלות בהיקפן.

כאן נכנס לתמונה המאמר המסוקר המחברים מציעים הגנה חדשה וחכמה, בשם U-CAN. במקום לנסות להפוך את המודל עצמו לחסין יותר, U-CAN פועלת כמערכת חיסונית נפרדת, המזהה ומסמנת קלטים אדברסריאליים לפני שהם יכולים לגרום נזק. השיטה המוצעת עושה זאת בצורה לא מונחית (unsupervised), כלומר אינה צריכה לראות דוגמאות להתקפות במהלך האימון שלה. זהו חשוב מאוד ליישומים רבים שאין לנו דאטה אדברסריאלי.

כיצד U-CAN מבצעת את הקסם שלה? הרעיון המרכזי הוא לבחון את הייצוגים הפנימיים של המודל. חשבו על מודל למידה עמוקה כסדרה של שכבות, שכל אחת מהן יוצרת ייצוג מופשט יותר של הקלט. U-CAN מצמידה "רשתות עזר" קטנות לשכבות ביניים אלו, כמו גששים זעירים המאפשרים "לראות" מה המודל "חושב" בשלבים שונים.

התובנה המרכזית היא שבעוד שקלטים אדברסריאליים נראים לנו נורמליים, הם יוצרים כאוס בתוך השכבות הפנימיות של המודל. U-CAN מאומנת לזהות את חוסר העקביות הפנימי הזה. במהלך שלב הזיהוי, הפלטים מרשתות העזר השונות, שכל אחת מהן צופה בשכבה אחרת, מושווים זה לזה. עבור קלט תקין, פלטים אלו יהיו דומים מאוד, ויספרו סיפור עקבי בזמן שהנתונים זורמים במודל. לעומת זאת, קלט אדברסריאלי יוצר אותות פנימיים סותרים, הגורמים לפלטים של רשתות העזר להיות שונים זה מזה. U-CAN מחשבת את המרחק בין הפלטים הללו; ציון אי-דמיון גבוה משמש כדגל אדום, המזהה את הקלט כמועמד פוטנציאלי ל-adversarial. זוהי דרך אלגנטית לזהות התקפות על ידי האזנה לסתירות פנימיות.

כעת, בואו נהיה מעט יותר "מתמטיים", אך ללא משוואות. סוד הקסם מאחורי הצלחתה של U-CAN טמון בכמה מרכיבי מפתח ופילוסופיית אימון ייחודית.

רשתות עזר (Auxiliary Networks): רשתות קטנות כלומר כאלו שאינן משנות את הפרמטרים של המודל הראשי או משפיעות על ביצועיו במשימתו המקורית (כמו LoRA כזה).

שכבות היטל (Projection Layers): שכבות אלו מקבלות את הייצוגים מרובי הממדים מהמודל הראשי ומטילות אותם למרחב בעל ממדים נמוכים יותר, מה שמקל על זיהוי חריגות.

שכבות לינאריות מבוססות ArcFace: טכניקה זו, שמקורה בזיהוי פנים, מותאמת ליצירת ייצוגים שיכולים להבדיל ביעילות בין דפוסים תקינים לאדברסריאליים.

תהליך האימון עצמו הוא מה שהופך את U-CAN לפרקטית כל כך. בהינתן מודל יעד מאומן מראש, M, כל רשת עזר מאומנת לעדן את מפות התכונות הפנימיות, או אמבדינגס, תוך שמירה על המודל הראשי M קפוא. זוהי בחירה ארכיטקטונית חשובה ביותר. במהלך האימון, המטרה היא למקסם את הדמיון התוך-קבוצתי, כלומר, לגרום לדפוסים הפנימיים של כל הקלטים ה"נורמליים" מאותה קטגוריה (למשל, כל תמונות החתולים) להיראות דומים ככל האפשר זה לזה ובמקביל לאכוף מרווח שמפריד בין דגימות מקטגוריות שונות.

לגישת ה"מודל הקפוא" יש השלכות עמוקות. משמעותה היא שאין צורך לאמן מחדש את המודל המקורי, שלעיתים קרובות הוא עצום בגודלו. זה הופך את U-CAN ליעילה להפליא וקלה להטמעה על מערכות בינה מלאכותית קיימות, ללא עלויות חישוביות אדירות. יתרה מכך, מכיוון שמשקולות המודל הראשי אינן משתנות, ביצועיו במשימתו המקורית אינם נפגעים להבדיל משיטות הגנה רבות שיש להן תופעת לוואי זו. על ידי מיקוד רשתות העזר במשימה היחידה של יצירת קלאסטר הדוק וצפוף עבור נתונים "נורמליים", U-CAN למעשה לומדת חתימה נאמנה למקור של מה שנחשב לגיטימי. כל קלט שייצר ייצוג פנימי הנופל מחוץ לגבול זה, מחוץ ל"מרווח" , יסומן מיד כחריג.

ה-U-CAN מציעה מספר יתרונות על פני שיטות הגנה קיימות:

היא לא מונחית: היא יכולה לזהות סוגים חדשים של התקפות שמעולם לא ראתה, יתרון עצום על פני שיטות כמו אימון אדברסריאלי.

היא לא פולשנית: U-CAN אינה משנה את הפרמטרים של המודל הראשי, ולכן אינה פוגעת בביצועיו במשימתו המקורית.

היא ניתנת להרחבה (Scalable): ניתן להוסיף בקלות את רשתות העזר הקלות למודלים קיימים ללא תוספת חישובית משמעותית.

https://arxiv.org/pdf/2502.09110

המוח במיקור חוץ: הפרדיגמה החדשה שבה סוכני AI לומדים בלי לגעת ב-LLM.