⚡️🚀המאמר היומי של מייק 15.08.24: ⚡️🚀

Your Classifier Can Be Secretly a Likelihood-Based OOD Detect

משנים טיפה את הכיוון היום וסוקרים מאמר לא על LLM. המאמר דן בזיהוי של דאטה שלא מתפלג לפי התפלגות הדאטה במהלך אימון המודל. למשל אימנתם מודל לזהות חתולים, כלבים וסוסים ופתאום מפעילים את המודל שלכם על תמונה של טנק. אם לא נקטתם אמצעים נגד זיהוי דאטה מחוץ להתפלגות האימון (או OOD) אתם עלולים לזהות את הטנק הזה בתור אחת הקטגוריות שאימנתם את המודל עליהם כלומר בתור כלב, חתול או סוס.

כמובן שהמצב הזה מאוד בעייתי ועקב כך הוא נחקר רבות במהלך השנים האחרונות. המאמר שנסקור קצרות היום מציע שיטה מאוד אלגנטית וטבעית להתמודד עם הסוגיה הזו. המאמר מציע לאמן מודל לזהות קטגוריות היעד (שמופיעות בסט האימון) אלא גם לכפות התפלגות מסוימת על הייצוג שלהם המופק על ידי המודל (כלומר של הפלט של השכבה האחרונה של הרשת).

הפרמטרים של ההתפלגות הזו נקבעים מראש (הממוצע ופרמטר ששולט בכמה ההתפלגות מרוכזת סביב הממוצע - סוג של מטריצת קווריאנס). ואם עבור דוגמא נתונה וקטור הייצוג יוצא רחוק מספיק מכל וקטורי הממוצע של כל הקטגוריות (כאשר מקדם הפיזור נלקח בחשבון) אז הדוגמא הזו מזוהה בתור OOD.

בתור התפלגות היעד המחברים לקחו התפלגות von Mises-Fisher על ספרה במימד של וקטור הייצוג p (כלומר הספרה היא במימד p-1). המחברים טוענים שזה עובד טוב יותר מאשר התפלגות גאוסית.

https://arxiv.org/abs/2408.04851