העתקה, נכון? מודל לבדיקת הגנת זכויות יוצרים במודלים של למידה עמוקה

Chen et al., 2022

1. מבוא. מודלים של למידה עמוקה הם אלגוריתמים שמיועדים לפתור בעיות מורכבות מהעולם האמיתי כמו למשל זיהוי תמונות, דיבור או עיבוד שפה טבעית, והם כיום אחד מהמודלים העיקריים עליהם מתבססת בינה מלאכותית. העתקה של מודלים אלה יכולה לגרום לנזקים כלכליים משמעותיים. בשל כך, פותחו בשנים האחרונות מספר טכניקות להגנה על מודלים אלה. המחקר הנוכחי מציע טכניקה חדשה שנקראת DeepJudge, מתאר את יתרונותיה בהשוואה לטכניקות הגנה קיימות, ובוחן את יעילותה אל מול סוגים שונים של העתקה.

2. רקע. רשת נוירונים עמוקה היא מודל המבוסס על שכבות קלט, שכבות נסתרות, ושכבת הפלט. אחת הטכניקות למניעת העתקה של מודלים אלה נקראת "סימן-מים" (Watermarking), הכוללת הטמעה ואימות. בשלב ההטמעה מוכנסת חתימה כלשהי, לדוגמא רצף תווים, אל מרחב הפרמטרים של המודל, והיוצר יכול להבחין אם מודל חשוד יוצר חתימה זו. טכניקה נוספת נקראת "טביעת-אצבע" (Fingerprinting), הכוללת חילוץ ואימות. בשלב החילוץ, אחד הפיצ'רים שיוצר המודל הוא "טביעת-אצבע" ייחודית, כאשר ניתן לזהות מודל מועתק אם הוא יוצר טביעת-אצבע זו. יתרונה של "טביעת-אצבע" הוא שהיא אינה פולשנית, כלומר היא אינה מוטמעת בשלב האימון של המודל.

3. מודל האיום על זכויות היוצרים של רשתות נוירונים עמוקות. קיימים שלושה סגנונות העתקה אפשריים: א. כוונון עדין, ב. קיצוץ, ג. חילוץ. כוונון עדין הוא מצב שבו האויב מכיר את כל המודל של הקורבן, והגניבה היא אימון של המודל על מאגר נתונים אחר. קיצוץ הוא מצב שבו ההעתקה היא תחילה "קיצוץ" של חלק מהמודל, ולאחר מכן כוונון עדין שלו. חילוץ הוא מצב שבו לאויב אין ידע מלא לגבי המודל, אך הניסיון הוא לגניבת הפונקציונאליות, למשל את הפרדיקציות של המודל.

4. בחינת הגנה על זכויות יוצרים של רשתות נוירונים עמוקות. בחלק זה מוצגת טכניקת DeepJudge. לטכניקה קיימים שלושה מרכיבים: 1. מטריצות בחינה, 2. ייצור של מקרה המבחן, 3. שיפוט סופי. מטריצות הבחינה הן מערך המשתנים התיאורטי שבוחן הבדלים בין המודל החשוד למודל הקורבן. ייצור מקרה המבחן הוא חישוב משתנים אלה בפועל בהשוואה בין שני המודלים. השיפוט הסופי כולל את ההחלטה האם המודל החשוד מועתק או לא. לטכניקת DeepJudge יש יתרון על פני "סימן-מים" בכך שהיא אינה פולשנית, ויתרון על פני "טביעת-אצבע" בכך שהיא מסוגלת לזהות את כל סוגי ההעתקות: כוונון עדין, קיצוץ וחילוץ.

5. ניסויים. התבצעה הערכה של ביצועי DeepJudge בתרחישי העתקה בסגנונות השונים. בסך הכל, הערכת DeepJudge התבצעה ביחס ל-11 שיטות העתקה שונות, ובקרב למעלה מ-300 מודלים עמוקים. DeepJudge נמצאה אפקטיבית בזיהוי העתקות מסוג כוונון עדין וקיצוץ. זיהוי העתקה מסוג חילוץ היה מאתגר יותר, אך השתפר ככל שההעתקה הייתה ברמה גבוהה יותר.

6. חוזק נגד תוקפים מסתגלים. בחלק זה נבחן החוזק של DeepJudge כאשר האויב יודע מהם הפרמטרים של מטריצות הבחינה ומקרה המבחן. נמצא כי DeepJudge אפקטיבית בזיהוי העתקה גם במצבים אלה, אך במקרים מסוימים היה צורך בהפעלה של הטכניקה מספר פעמים.

7. מסקנה. המאמר מציג את DeepJudge, טכניקה חדשה להגנה על זכויות יוצרים של מודלים של למידה עמוקה. בהשוואה לטכניקה קיימת של "סימן-מים", DeepJudge אינה מצריכה הטמעה בשלב האימון של המודל. בהשוואה לטכניקה של "טביעת-אצבע", DeepJudge מאפשרת זיהוי של העתקות מסוגים מגוונים יותר, ועמידה יותר להעתקות מסתגלות. בוצעו ניסויים שהראו את יעילותה של הטכניקה. יישומה יכול לסייע בהתמודדות עם הפרת זכויות יוצרים של מודלים של למידה עמוקה.

מקור

Chen, J., Wang, J., Peng, T., Sun, Y., Cheng, P., Ji, S., ... & Song, D. (2022). Copy, right? A testing framework for copyright protection of deep learning models. In: *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 824-841). San Francisco: IEEE Security and Privacy.