# Review 88: Secure Machine Learning in the Cloud Using One Way Scrambling by Deconvolution

**Paper: https://arxiv.org/abs/2111.03125v1**

פינת הסוקר:

המלצת קריאה מעדן ומייק: מומלץ לאנשים העוסקים בתחום ה-Cybersecurity או/ו לאנשים שמתעניינים בטכניקות לסיווג תמונות (בלי לראות אותן!!)

בהירות כתיבה: טובה

ידע מוקדם:

הצפנה

Deconvolution

יישומים פרקטיים אפשריים: מערכת להצפנת מידע

פרטי מאמר:

פורסם בתאריך: 4.11.2021, בארקיב.

הוצג בכנס: - KDD

תחומי מאמר:

פרטיות המידע

מודלים בענן

כלים מתמטיים, מושגים וסימונים:

אנקודר(להפקת ייצוג לטנטי של תמונה) ודקודר

מבוא:

הרצת מודלים גדולים של למידה עמוקה איננה משימה קלה. נדרש מאמץ רב ויכולות מחשוב כבירות (בד״כ GPU-ים או TPU-ים) סביב אימון המודל והפעלתו (Inference) שלו. חסם זה מנע משחקנים רבים את היכולת להשתמש במודלים ענקיים אלו . אך כיום זה משתנה עקב שימוש הולך וגדל בשירותי למידת מכונה בענן בו חברות פורסות מודלים גדולים ומאפשרים גישה לכלל הציבור דרך קריאות API שנעשות על גבי האינטרנט. גישה זו אמנם הנגישה מודלים רבים ואפשרה למספר רב של משתמשים להריץ אותם שלא היה ניתן קודם אך גם הביאה איתה מספר סוגיות רציניות. בעיה אחת מרכזית הינה פרטיות המידע: בשביל להעביר למודל מידע הנחוץ להרצתו יש צורך לשלוח אותו על גבי האינטרנט - - דבר שעלול לגרום לזליגת מידע רגיש. הבעיה השניה הינה היכולת של גורמים זרים להסיק מסקנות לגבי המידע של הארגון שלך בצורה עקיפה. ניתן לעשות זאת על ידי ניתוח הפלט של המודל בענן וזאת בשל העובדה שלכולם יש גישה לאותו מודל.

כדי להתגבר על הבעיות המוזכרות לארגונים היום קיימות שתי אופציות:

להצפין את המידע לפני שליחתו, פענוח בצד השני, הצפנת התוצאה ושליחתו חזרה לפענוח אצל הלקוח.

שימוש בהצפנה הומומורפית (Homomorphic Encryption - HE). הצפנה זו מאפשרת לצד המקבל את המידע המוצפן לא לפענח אותו, אלא לנתח אותו ולהגיע למסקנות דומות באותה צורה שבה היה מגיע אילולא המידע לא היה מוצפן מלכתחילה.

אך לפתרונות אלו יש כמובן בעיות. הבעיה עם הפתרון הראשון הינו שלאחר פענוח המידע בצד של ספק הענן, לאותו ספק תהיה גישה למידע הלא מוצפן. כלומר, יש פה מקרה של הפרת פרטיות המידע. הבעיה עם הפתרון השני הינו שהוא איטי שכן עיבוד המידע המוצפן קשה יותר.

כדי להתגבר על הבעיות של פתרונות אלה בבת אחת, כותבי המאמר מציעים שיטה הנקראת Encoding via Private-Key Deconvolution (EPKD).

השיטה:

השיטה המוצעת דומה ל-HE בכך שהיא מאפשרת למודל בענן לקבל את המידע מוצפן ולעבד אותו כאילו לא היה מוצפן אך עם מהירות דומה לעיבוד המידע הרגיל. כלומר החיסרון של מהירות העיבוד שמציג HE לא משחקת פה תפקיד. השיטה משתמשת בשלושה רכיבים:

Encoder(מקודד): נועד כדי ליצור ייצוג דחוס של המידע. בעבודה השתמשו החוקרים בכמה מודלים מסוג Resnet כגון ResNet5, ResNet101, ResNet50v2. כולם היו מאומנים מראש ולא כוילו בנוסף. השכבה האחרונה של ה-softmax הוסרה כדי להשתמש בוקטור הייצוג(Embedding) שהרשת יוצרת.

Generative Model(מודל גנרטיבי): המודל המשמש בתור שיטת ההצפנה. כאן משתמשים ברשת מסוג Deconvolution אשר המשקולות שלה מאותחלות באופן רנדומלי. המשקולות הרנדומליות משתמשות כמפתח ההצפנה. הרשת מקבל את וקטור הייצוג מהמקודד ויוצרת ממנו תמונה חדשה שלעין האנושית תראה כמו רעש מוחלט. התמונה החדשה נשלחת למודל בענן שמעבד אותה כמו תמונה רגילה ושולח את וקטור התחזיות שלו חזרה.

The Internal Inference Network (INN) - רשת feed forward קטנה המקבל כקלט את הייצוג של התמונה המקורית יחד עם וקטור התחזיות שהמודל בענן החזיר על התמונה המוצפנת. הרשת מוציאה כפלט את התיוג האמיתי של תמונת המקור.

חשוב לשים לב שהמודל בענן אינו מקבל את התמונה המקורית אלא את התמונה אחרי תהליך הגנרוט. דבר נוסף לשים לב אליו הינו שהמודל המגנרט אינו מאומן אלא מאותחל בסט משקולות חדש ורנדומלי בכל פעם. למה כך? ובכן בתחום ההצפנה יש את מושג מפתח ההצפנה שבעזרתו מצפינים את המידע. אם מישהו יידע את המפתח הוא יוכל להשתמש בו כדי לשחזר את המידע המוצפן. לכן, חשוב שיהיה קשה להשיג את המפתח. במקרה של המודל המגנרט, מפתח ההצפנה הינו משקולת המודל. אם מישהו ישיג את המשקולות של המודל הוא יוכל להשתמש בהן כדי להעתיק את המודל ולהשתמש בו לתהליך ההפוך - מעבר ממרחב האמבדינג חזרה לתמונה המקורית. כדי למנוע זאת, מאתחלים רנדומלית את המשקולות בכל פעם כדי להקטין את הסיכוי שמישהו ישיג בנקודת זמן מסוימת בדיוק את המשקולות שיצרו תמונה מסוימת. התמונה המוצפנת שנוצרת מהמודל הגנרטיבי תראה כמו רעש מוחלט לעין האנושית כפי שניתן לראות בתמונה 1:

תהליך האימון:

הרכיב היחיד שצריך אימון הינו מודל ה-INN. כותבי המאמר טוענים שבשביל אימון המודל יש צורך בדאטהסט קטן בלבד של תמונות מתויגות או לחילופין דאטהסט ציבורי שזמין באינטרנט. תמונות אלו נשלחות למודל בענן שמחזיר את וקטור ההסתברויות (של הקטגוריות) עבור משימת סיווג. הקידוד של התמונה יחד עם וקטור ההסתברויות (כלומר שני וקטורים) משמשים כדגימה אחת לאימון מודל ה-INN. תהליך האימון מוצג בתמונה 1.

איך אבל מודל ה-INN יודע לקבל מידע מוקטור ההסתבוריות של מודל אותו הוא לא מכיר? אובחן מדובר בשילוב של שתי שיטות משני תחומים.

הראשונה שייכת לתחום ה-Adversarial Machine Learning ונקראת Membership Inference. בשיטה זו קיים משתמש הרוצה לנסות לגנוב מידע על אופן אימון המודל. במקרה זה מדובר ניצול של ידיעה של האם דגימה מסוימת היא חלק מסט האימון של המודל או לא. על ידי ניתוח וקטור ההסתברויות של המודל, ניתן לראות דפוסים בהם המודל מתנהג אחרת עבור דגימות מסט האימון וכאלו שלא. כדי לבצע את הניתוח מאמנים מודל שקובע עבור כל קלט (וקטור הסתברויות ותמונה) האם התמונה היא חלק מסט האימון. עבודה זו לוקחת השראה מתחום זה. הסיבה לכך הינה שגם בשיטה המוצעת כאן, הקלט הינו וקטור הסתברויות והדגימה שאליה שייך וקטור זה. השוני הוא הפלט, שכן בשיטה המקורית משתמשים במודל כדי לחזות האם דגימה שייכת לסט האימון או לא וכאן מנסים לחזות את הסיווג האמיתי של הדגימה מוקטור ההסתברויות. בעבודה זו, ניתן לראות שיש חפיפה שכן מדובר באותו קלט אך פלט שונה.

השיטה השניה היא מתחום ה-Distillation. השיטה מדברת על אימון מודל קטן המחקה מודל גדול יותר (ensemble של מודלים במקרה זה). המודל המאומן יהיה בסדרי גודל קטן יותר מהמודל אותו הוא מנסה ללמוד ובנוסף יהיה בעל ביצועים דומים. מודל ה-INN מנסה באמצעות אותה שיטה ללמוד את מודל הענן. הדגש של כותבי המאמר כפי שמוזכר הינו שמודל ה-INN הינו מודל קטן (ההנחה היא שקטן בסדרי גודל מהמודל בענן) וכולל רק שכבה חבויה אחת בלבד.

כל תמונה רגישה מועברת דרך המקודד ומוצפנת על ידי המודל הגנרטיבי. בנוסף כדי להקשות עוד יותר על פענוח המידע על ידי גורם זר, נוספו לתמונה הרגישה 9 תמונות הנבחרות רנדומליות כל פעם אשר הוצפנו גם הם. התמונות נועדו כדי להוסיף רעש, כך לא ניתן יהיה לדעת מה מהמידע הוא רגיש ומה הוא לא. בכל שליחה של מידע התמונה הרגישה והתמונות הלא רגישות עורבבו כך שמיקום התמונה הרגישה במערך שנשלח היה נבחר באופן רנדומלי כדי להקשות על זיהוי שלה. מודל ה-INN נמדד על היכולת לסווג נכון כל תמונה בין אם היא מוצפנת ובין אם לא. המדדים אשר כותבי המאמר השתמשו בהם הינם: Top1, Top5 שבקיצור אומרים האם הסיווג האמיתי הוא בעל ההסתברות הגבוה ביותר או בין החמש הכי גבוהות.

את התוצאות ניתן לראות בקצרה בגרף המוצג:

בדיקת אבטחה:

כיוון שקיים תהליך הצפנה בכל השיטה חשוב לבדוק את איכות ההצפנה והאם ניתנת לפיצוח על ידי גורם זר. החוקרים בדקו קודם את המקרה הטריוויאלי שבו התוקף ינסה את כל הקומבינציות של פיקסל בתמונה המוצפנת כדי לשחזר את התמונה המקורית. הם ציינו שאפילו במקרה הפשוט של תמונה 2x2 ומפתח הצפנה (משקולות המודל הגנרטיבי) בגודל 2x2 עם טווח ערכים של מרחב האפשרויות הינו גדול מדי בשביל לפצח את ההצפנה ולשחזר את התמונה המקורית. לכן הם התמקדו בהתקפת סייבר יותר מתוחכמת הנקראת (Distinguishing attacks (DA. בהתקפה זו מטרת התוקף היא לא לגנוב מידע כל שהוא אלא לבצע אנליזה למידע ולראות האם הוא יכול למצוא קשרים בין חלקים שונים, למשל בין המידע המוצפן למידע שהצפינו או בין שני פריטי מידע שהוצפנו. מציאת קשר שכזה יוכל לשרת את התוקף לאחר מכן בשביל לפצח חלקים מההצפנה. במקרה שלנו נרצה להימנע מהמקרה בו התוקף יוכל ללמוד על הקשר של שתי תמונות מוצפנות. אם אכן קיים כזה התוקף יוכל להסיק שמקורן באותו סיווג (למשל שתי תמונות של כלב).

התקפה נוספת אותה בחנו הכותבים הינה Model Extraction. בהתקפה זו התוקך מנסה לשחזר את משקולות המודל (משמע לגנוב את המודל). המחברים ניסו את השיטות המתקדמות כיום בתחום תחת ההנחה שלתוקף יש גישה למודל בצורה של שליחת קלט וקבלת פלט. תחת שיטות אלו צריך מספר קטן של תמונות והגרסה המוצפנות שלהן. את זה ניתן על ידי שליחת תמונות למודל וקבלת הפלט שלו (ההצפנה במקרה זה). התוצאות האמפיריות מראות שבשביל שההתקפה תעבוד על התוקף להחזיק ב-10% דוגמאות מגודל המודל. כלומר אם גודל המודל הינו 10 מיליון פרמטרים, התוקף יצטרך בערך מיליון דוגמאות של קלט ופלט של המודל. הטענה היא שהמודל חסין לסוג זה של התקפה בשל העובדה בשל העובדה שמפתח ההצפנה משתנה לעיתים תכופות תוך כדי שימוש המודל. כלומר אם התוקף ירצה לייצר מיליון דגימות הוא לא יוכל לעשות זאת תחת ההנחה שמפתח הצפנה אחד (משקולות המודל) הוא זה שיצר את התמונות המוצפנות ולכן ההתקפה נשברת.

סיכום:

בעבודה זו הראנו שניתן להשתמש במודלים שונים כדי להצפין את המידע בצורה יעילה, כדי לשלוח מידע בצורה מאובטחת על גבי האינטרנט למודל גדול הנמצא בענן. אותה שיטה לא סובלת מבעיות של איטיות בהצפנת HE או בבעיות אבטחה שונות. השיטה משתמשת בשני רשתות ניורונים אשר מאותחלות באופן רנדומלי וללא אימון מקודדים תמונות. הקידוד משמשת כהצפנה של התמונה אשר נשלחת מוצפנת לענן. המודל בענן שולח לנו חזרה את וקטור ההסתברויות של התמונה המוצפנת. נשתמש ברשת ניורונים נוספת, קטנה משמעותית מהמודל בענן כדי ללמוד לשחזר את הסיווג האמיתי של תמונה בהינתן וקטור ההסתברויות והקידוד של התמונה. שיטה זו מראה את השימוש של רשתות ניורונים לא רק כמשחזרי מידע טובה אלא גם כמצפיני מידע טובים ופותחת דלת לשימוש השיטה על סוגי דאטה נוספים כגון טקסט או דאטה טבלאי.

שיתוף פעולה: הסקירה נכתבה בשיתוף פעולה עם עדן יבין.