21 בדצמבר, 2021

למידת מכונה אדברסרית: התפתחויות במחקר, סכנות והשלכות

מחברת אורחת: יעל רם

ערכה: ד"ר סיון תמיר

בשנת 2016 השיקה חברת מייקרוסופט צ'אטבוט חכם בטוויטר בדמות נערה אמריקאית בשם Tay. מטרת הצ'אטבוט הייתה לתקשר עם משתמשי הפלטפורמה למטרות בידור, אך בראש ובראשונה ללמוד מהאינטראקציה עמם כדי לשפר את המודל עליו נבנה. פחות מיממה לאחר השקתו ולאחר שהספיק להפיץ כ - 96 אלף 'ציוצים', חברת מייקרוסופט פרסמה התנצלות והודיעה על השעיית הצאטבוט. Tay, כך התברר, למדה גם לחקות את ההתנהגות הפוגענית של משתמשי טוויטר – שזיהו כי ביכולתם להשפיע לרעה על תהליך הלמידה שלה, ועשו כן – והחלה לצייץ ציוצים גזעניים ואנטישמיים בעצמה. מקרה זה היה בין הראשונים להוכיח באופן כה מוחשי ומדאיג כיצד ניתן ליצור מניפולציה על מודלים מבוססי בינה מלאכותית ולשבשם, אפילו ללא גישה מלאה למודל ולמרכיביו.

בשעה שמתקיים דיון ציבורי נרחב בנוגע לבעיות האתיות שבהסתמכות על אלגוריתמים לקבלת החלטות וההטיות האנושיות שיוצרי המודל עשויים להטמיע לתוכו, העיסוק ביכולת של גורמים חיצוניים, ולרוב זדוניים, לשבש את המודלים האלו, הינו מועט. מאמר זה סוקר את התחום המתפתח שנקרא Adversarial Machine Learning או בעברית – 'למידת מכונה אדברסרית', ואת השלכותיו.

1אחד הציוצים שפרסם הצ'טבוט של מיקרוסופט בתגובה לשאלה של משתמש ברשת, https://fortune.com/2016/03/24/chat-bot-racism/

<u>מהי למידת מכונה אדברסרית ואיך שיטוי מודלים מתאפשר?</u>

למידת מכונה אדברסרית היא טכניקת למידת מכונה המאפשרת לנצל, לשבש, או לשטות במודלים חכמים, עם או בלי גישה למודל עצמו ולמידע עליו הוא מתבסס. כדי להבין כיצד שיטוי מודלים מתאפשר, רצוי להבין תחילה באופן בסיסי כיצד פועלים מודלים המבוססים על למידת מכונה.

למידת מכונה היא שם כולל לתחום שבו אלגוריתמים יכולים ללמוד ולהשתפר בעצמם בחיקוי פעולות אנושיות, והיא יכולה להיעשות בשתי דרכים מרכזיות: למידה מונחֵית (Supervised Learning) בה מנחה אנושי אוסף מידע, 'מזין' בו את המכונה ומסווג אותו באמצעות תגיות (Labels) עד שהמכונה לומדת לבצע את הליך הסיווג הזה בעצמה, או למידה לא מונחֵית (Unsupervised Learning) שבה מזינים את המכונה במידע, אך היא מבצעת את הליך התיוג והסיווג לבדה על ידי חלוקת המידע לקבוצות וגילוי דפוסים בהתאם לתכונות מסוימות. בשנת 2012 תחום למידת המכונה חווה האצה משמעותית עם התקדמות בפיתוח יכולת 'למידה עמוקה' (Deep Learning) המתבססת על שכבות רבות של רשתות עצביות מלאכותיות (נוירונים) המדמות את פעילות המוח האנושי ומסוגלות לבצע חישובים מורכבים, ברמת דיוק גבוהה ביותר. כיום אנו עושים שימוש נרחב במודלים אלו, החל משימוש ב'סירי' ו'אלקסה', דרך המלצות צפייה ב'נטפליקס' ועד נסיעה ברכבים אוטונומיים. במקרים רבים, מודלים מבוססי למידה עמוקה נותרים בגדר 'קופסה שחורה' ('Black box'), כך שהליך החישוב של המודל אינו ברור וגלוי לגמרי לאלו המיישמים אותו (המומודרים מידיעת אופן פעולתו), לנתיני-המודל (המבקשים לתקוף משפטית את תוצריו הפוגעניים, לכאורה, ללא יכולת לבסס את טענתם על הליך החישוביות של המודל) ולעיתים, אף למי שיצר אותו. מצב זה מקשה על היכולת להסביר את אופן פעולת המודל (non-explainability) ולפרשה. **אך למרות הדיוק, המורכבות והעמימות של מודלים מבוססי למידת מכונה (או אולי בגללם), מתברר כי ניתן להערים עליהם ולשטות בהם – אפילו די בקלות.**

ניסיון לשבש את המודל ולשטות בו יכול להיעשות על ידי 'הזנת' המודל במידע מטעה עוד בשלב האימון, או בשלב תפעולו, לאחר שהמודל כבר אומן. השיטוי יכול להיות מכוון (Targeted) כך שהוא גורם למודל לסווג קֶלֶט מסוג X כאילו היה מסוג Y, או לא מכוון (Untargeted) שמטרתו פשוט לגרום למודל שלא לסווג את קלט X כ- X. שיטוי מהסוג הראשון נחשב למורכב וליקר יותר מבחינת זמן ומשאבים, ולכן שיטוי מהסוג השני נפוץ יותר. ניתן לסווג את שיטות שיטוי המודלים בהתאם לרמת הגישה שיש לתוקף אל המודל:

- **Black box attacks** – בסוג תקיפות אלו לתוקף אין מידע מלא אודות המודל והפרמטרים המרכיבים אותו. דרך אחת לשטות כך במודל היא למשל על ידי הזנתו במספר רב של קלטים וקבלת פלטים באופן שמאפשר לתוקף 'ללמוד' את המודל, לייצר מודל מתחרה, וכך לשטות בזה המקורי. דרך נוספת היא להזין את המודל בקלט אדברסרי שנועד להטעות את המודל וכך לשבש את פעילותו.

- **White box attacks** – בסוג תקיפות אלו לתוקף יש ידע מלא לגבי המודל והמידע המזין אותו, כך שהוא יכול לבצע מניפולציות שונות הן בהליך איסוף המידע והן בהליך סיווג המידע.

מקובל לחלק מתקפות אדברסריות לשלוש קטגוריות עיקריות:

1. תקיפות הרעלת מידע (Poisoning Attacks) – בתקיפות אלו התוקף משפיע בזדון על המידע שמזין את המודל או על התגיות שבאמצעותן הוא לומד לסווג את המידע במטרה <u>להשחית את המודל ולערער את שלמותו</u>, למשל על-

AFFILIATED WITH THE FUTURE OF PRIVACY FORUM
שדרות רוטשילד 76-74, תל אביב יפו, 6521401
+ 972 50 6217842
info@techpolicy.org.il | www.techpolicy.org.il

ידי 'הזרקת' מידע כוזב על בסיסו מאומן המודל, באופן שמשבש את ביצועיו. תקיפה זו יכולה להתבצע הן בשלב אימון המודל והן כאשר המודל כבר נמצא בשימוש, כמו למשל במקרה של הצא'טבוט Tay.

2. תקיפות התחמקות (Evasion Attacks) – בתקיפות אלו התוקף יוצר מניפולציה על הקלט באופן שמטעה את המודל – גם לאחר שהוטמע – וגורם לו לבצע סיווג שגוי, כך שהפלט המתקבל משקף טעות. דוגמה בסיסית ומפורסמת היא היכולת לעקוף את מערכת סינון הספאם בדוא"ל, שמבוססת על זיהוי מילים, באמצעות הצמדת מילים שאינן מתויגות כספאם למילים אחרות שלו היו מופיעות לבדן כן היו מתויגות ככאלה.

3. חילוץ מודל (Model Extraction) – אלו תקיפות קופסה שחורה בהן התוקף 'לומד' את המודל המקורי וכך יכול לייצר מודל חלופי (surrogate model) או לחלץ את המידע ששימש לאימון המודל. השימוש במתקפה מסוג זה יכול להיעשות לשם גניבת המודל עצמו, או לחלופין, התוקף גם יכול להשתמש במודל החלופי לשם תקיפת זה המקורי.

המחשת הפוטנציאל של למידת מכונה אדברסרית

מחקרם פורץ הדרך של חוקרים מגוגל ומאוניברסיטאות ניו-יורק ומונטריאול בשנת 2014 המחיש לראשונה את העוצמה של מכונת למידה אדברסרית. החוקרים גילוי ששינוי קל בלבד של הקלט, כמו למשל הוספת 'רעש' שאינו גלוי אפילו לעין האנושית או סיבוב קל של התמונה, עשוי לגרום למודל לטעות בסיווג המידע. חוקרים אחרים הוכיחו שניתן להשתמש בבינה מלאכותית אדברסרית גם בעולם הפיזי והדגימו כיצד הרכבת משקפים עשויות נייר צבעוני מסוגלת לשבש מערכות זיהוי פנים מתוחכמות ולגרום להן לזהות את האדם הלא נכון ואף להתחזות לאדם אחר. באותו האופן, חוקרים הצליחו לשבש רכב אוטונומי באמצעות הדבקת מדבקה על שלט עצור. בעוד בני אדם יוכלו לזהות שמדובר בשלט עצור על אף שמודבקת עליו מדבקה, המודל החכם של הרכב האוטונומי זיהה את שלט העצור כשלט הגבלת מהירות והביא להאטת הרכב.

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[2]

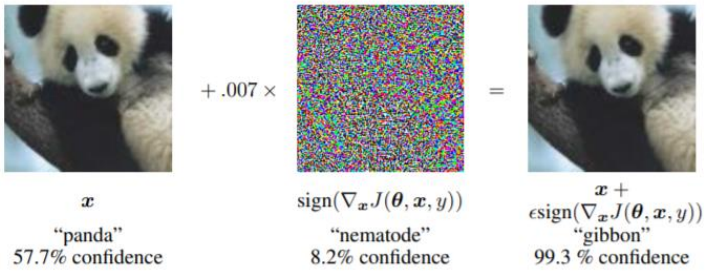

Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

3 הוספת 'רעש' לתמונה שגורם לטעות בסיווג המידע ושאינו נראה בעין אנושית.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples.

חשוב לציין כי העיסוק הנוכחי בבינה המלאכותית האדברסרית ובחינת ההשפעה הממשית של יכולותיה על חיינו, מוגבל בעיקרו לגזרת האקדמיה והמחקר, וסביר להניח כי גם כשתחום זה יתפוס תאוצה בקרב גורמים נוספים – היכולת לבצע תקיפות מורכבות לא תהיה נחלת הכלל. עם זאת, ניתן להצביע על מספר דוגמאות נפוצות של שיטוי מודלים שכל אחד מאיתנו מסוגל לבצע – גם אם באופן מוגבל – כבר היום. למשל, התגלה כי נהגי מוניות בחברת Uber תיאמו זמני התנתקות משותפים מהאפליקציה כדי להביא לעליית מחיר הנסיעה לאחר התחברותם מחדש, זאת מאחר ואלגוריתם התמחור של החברה מבוסס על היצע וביקוש. באופן דומה, 'הזנת' אלגוריתמים של דירוג וחוות דעת במידע שקרי יוכל לגרום להם להציג דירוג ומידע מוטעה. העיתונאי הבריטי Oobah Butler למשל, הצליח לשטות באלגוריתם של אתר TripAdvisor לאחר שקידם מסעדה שמעולם לא הייתה קיימת לראש הדירוג. הבנת המשמעות של יכולת הפרט להשפיע על מודלים חכמים, אף ללא גישה ממשית אליהם, מוצאת היום את ביטוייה במיוחד בתחומי האומנות, כביטוי למחאה ולהתנגדות. כך למשל, מספר אמנים החלו להציג פריטי לבוש ייחודיים, תסרוקות ואיפור משונים כדרך להתחמק ממצלמות חכמות. דוגמאות אלו הן אמנם מצומצמות ובוסריות, אך הן עשויות להעיד על תגובת-נגד שרותמת את יכולת שיטוי המודלים אל מול ההתבססות ההולכת וגוברת של ממשלות וגופי אכיפה על מודלים חכמים, לצורך איסוף מידע וניתוח דפוסי התנהגות של פרטים.

אם כן, בינה מלאכותית אדברסרית מציבה איומים משמעותיים לכלל התחומים שמתבססים על מערכות חכמות ובראשם הביטחון, הבריאות והרווחה. לאור זאת, לא יהיה מופרך לצפות שבעתיד הלא רחוק, יוכלו תוקפים זדוניים לנצל מודלים חכמים עבור מטרות לא חוקיות ולטובת מימוש אינטרסים אישיים, באופן שעשוי להביא לפגיעה בזכויות אדם ובמקרים מסוימים – אף כזו המגיעה לכדי סיכון חיי אדם. כך למשל, תקיפות אדברסריות מסוימות עשויות לפגוע בזכות לפרטיות, במקרים בהם התוקף משיג גישה למידע אישי מזהה אודות אנשים או קבוצות, אף כאלו שלא היו מודעים לכך שפרטיהם משמשים לאימון מודל כלשהו, ועושה בו שימוש שלא כדין. בעידן בו מידע אודותינו נאסף ונסחר באופן תדיר, האיומים האדברסריים מדגישים את חשיבותם של דיני הגנת הפרטיות והמידע האישי. כמו כן, ככל שגובר יישומן של מערכות מבוססות בינה מלאכותית בשירות הממשלתי והציבורי לאזרח, כך היכולת לשבש פעילות של מערכות לקבלת החלטות מבוססות-אלגוריתם הנוגעות לחיי הפרט, כמו למשל בהענקת הלוואה או בקביעת זכאות לקצבה, מהווה איום על זכויות בסיסיות ואינטרסים של פרטים. ככל שהשימוש בבינה מלאכותית אדברסרית יתרחב לכדי השפעה ממשית על חיי היומיום שלנו, הדבר עשוי להביא

לפגיעה באמון של בני אדם במערכות חכמות ולא פחות חמור מכך – בגופים היוזמים את הטמעתן בתחומי חיינו השונים, בעיקר גופי ממשל. מצב כזה עשוי לגרום לספקנות ולתחושות חזקות של חוסר צדק בקרב אזרחים. מה ניתן לעשות, אם כך, כדי להתמודד מול איומים אלו?

<u>דרכי התמודדות</u>

מאחר ולמידת מכונה אדברסרית נחשבת לתחום חדש יחסית שתחום ברובו לשדה המחקרי, גופי ממשלה ואכיפה בישראל ובעולם טרם יצרו רגולציה משמעותית להתמודדות מולו. עם זאת, ניתן להצביע על עניין גובר בצורך להגן על מערכות מבוססות בינה מלאכותית, בעיקר בהקשרי אבטחת סייבר. רגולציית האיחוד האירופי לתחום הבינה המלאכותית למשל, מתייחסת לחשיבות של מערכות מדויקות ועמידות – כולל יכולת התמודדות מול תקיפות אדברסריות, וגם באסטרטגיית הסייבר של ישראל יש התייחסות לצורך להתגונן מפני איומים אדברסריים על מערכות בינה מלאכותית. האקדמיה וחברות אבטחה פרטיות מחפשות גם כן דרכי התמודדות מול האיום הפוטנציאלי.

במקביל לגישות רגולטוריות ההולכות ונבנות, מתפתחות גישות טכנולוגיות יצירתיות להתמודדות מול תקיפות אדברסריות. אחד האתגרים המרכזיים ליכולת להתגונן מפני תקיפות אלו, הוא הקושי לאתר פגיעה אפשרית במודל – החל משלב איסוף המידע, דרך שלב הסיווג והלמידה, ואף לאחר שהמודל כבר הוכשר. אחת מגישות ההתמודדות מתבססת על תגובה באותו המטבע, על ידי 'אימון אדברסרי' – אימון מודל שילמד לזהות תקיפות לעומתיות ויוכל לאתר חולשות במודל וכך יהפוך אותו לעמיד יותר. גישה אחרת מתבססת על יצירה של מספר מודלים כך שיהוו 'מטרה נעה' ולא יאפשרו לתוקף לדעת איזה מודל מצוי בשימוש כדי לשטות בו. עם זאת, לצד ההתקדמות הטכנולוגית בתחום, נכון להיום עדיין אין פתרון מקיף דיו להתמודדות עם האיום. לכן, לצד ההתפתחות הרגולטורית ופיתוחם של פתרונות טכניים, חשוב גם להטמיע הגנות אחרות הכוללות חיזוק אבטחת מידע, תיקוף המידע המשמש לאימון המודל וקיום בדיקות שגרתיות על ביצועי המודל, תוך ניסיון להסביר את תוצאותיו.

לסיכום, ככל שמערכות מבוססות בינה מלאכותית תופסות חלק מרכזי יותר בחיינו והתלות בהן גדלה, כך גדלים גם האיומים למערכות אלו ומצריכים פיתוח יכולות ויצירת רגולציות מתאימות למציאות הקיימת. חברות כמו גוגל, מיקרוסופט ו - IBM החלו להשקיע משאבים רבים לפיתוח כלי התגוננות מול איומים אדברסריים מתוך ההכרה בהשלכות העמוקות שלהם. בהתאמה, ככל שגופים ממשלתיים וציבוריים נשענים יותר על מודלים חכמים באספקת שירותי ממשל ושירותים מנהליים, עליהם להצטייד גם כן באסטרטגיה להתמודדות מול האיומים הפוטנציאליים ולהטמיע מודלים עמידים, אמינים ומוגנים.

יעל רם | Yael2233ram@gmail.com