בס"ד

דו"ח תקופתי ראשון לפרוייקט

סטיב גוטפרוינד וישי אשר

נעבור על אבני הדרך שתיכננו עד התאריך הנוכחי (אמצע ינואר) , ונראה את ההתקדמות באבני הדרך השונים שהצבנו לעצמנו (זאת **בהתאמה** למסמך "תוכנית העבודה לפרוייקט") :

אבני דרך

1) תכנון הפרויקט ושחזור המודל

בחלק זה המטרות היו :

א. שחזור המודלים של DNN מהמאמר על הdatasets הבאים :
FASHION-MNIST,CIFAR-10,MNIST

ב. אימון המודלים .

ג. בדיקת הinference של המודלים השונים בהשוואה למוצג במאמר .

תיאור העבודה שהושלמה

א. שחזרנו את שלושת המודלים המתוארים במאמר. (הכוונה למודלים שונים שתוארו במאמר שניתן להריץ אותם על הdatasets הנ"ל)

ב. הרצנו ואימנו את המודלים השונים על כל הdatasets לעיל , פרט לCIFAR10 שהתברר במהלך העבודה על הפרוייקט שהוא אינו נכלל במאמר . אמנם בכל זאת הרצנו את אחד המודלים על CIFAR10 . (ומכיוון שהמודלים לא היו מותאמים לCIFAR10 אזי לא המשכנו בכיוון הנ"ל כי הם לא עסקו בזה , אמנם ייתכן ובהמשך ננסה לבנות מודל שיתאים לו)

ג. השווינו את הinference שקיבלנו על ההרצות השונות למוצג במאמר ואכן קיבלנו דיוק גבוה.

להלן טבלת הדיוק שלנו (משמאל) ושל המאמר (מימין) כאשר בשניהם מדובר באחוזי השגיאה .[1]

זוהי העמודה הרלוונטית (שלב ההתקפה הוא רק באבן דרך 4)

dataset \ model	unencrypted-model		
	CW_1	CW_2	FGSM
mnist	1.05%	0.77%	0.94%
fashion_mnist	8.90%	8.03%	9.6%
cifar10		25.75%	

Attack	Classical classifier	
	original	attacked
	MNIST	
CW ℓ_2	1.00	100.00
CW ℓ_0	1.00	100.00
CW ℓ_∞	1.00	99.99
FGSM	1.00	92.10
	Fashion MNIST	
CW ℓ_2	7.50	100.00
CW ℓ_0	7.50	100.00
CW ℓ_∞	7.50	99.90
FGSM	8.60	60.60

[1] גרף המתאר את תוצאות הדיוק מצורף בסוף הדו"ח

הסבר :
בטבלה שלנו (משמאל) זוהי טבלת אחוזי שגיאה לכל dataset לפי המודלים השונים – CW_1,CW_2,FGSM (המודלים הללו נקראו כך בהתאמה להתקפה שבהמשך ננסה לתקוף אותם)
בעוד שבטבלה שבמאמר היא טבלת אחוזי שגיאה לפי סוג **ההתקפה** (כאשר התוצאות בעמודה הרלוונטית הן על המודלים **שלא** מוצפנים , וכן ללא כל התקפה) , והמודלים השונים הם בהתאמה לסוג ההתקפה גם כן . (FGSM למשל מתאים למודל FGSM וכן כל הCW שם מתאים למודלים CW_1,CW_2)

פירוט שעות

ישבנו יחד כדי לתכנן את העבודה, להתקין את החבילות הנדרשות ולמדנו על tensorflow.
התחלקנו בבניית המודלים והאימון שלהם על הdatasets השונים (כלומר לפחות 6 אימונים אם אין באגים)

סה"כ: בערך 45 שעות עבודה כל אחד

2) הצפנת המודל[2] [זמן משוערך : חודש וחצי - חודשיים]
בחלק זה המטרות היו:
לבדוק שיטות שונות של הצפנה ולבדוק אם הדיוק של המודלים בtestים נשמר (עד כדי סטייה זניחה של 3% - 1.5%) .
זאת בשימוש בספרייה pycrypto ודומיה .
השיטות אותן היינו צריכים לבדוק את אחוזי הדיוק שלהן :
א. פרמטוציה רנדומית של הפיקסלים (כלומר הBytes) .
ב. AES in ECB mode of operation.
ג. AES in CBC mode of operation.
ד. AES in CTR mode of operation.

תיאור העבודה שהושלמה

א. הרצנו את המודלים השונים כאשר לפני כניסת כל תמונה לאימון היא עוברת פרמוטציה רנדומית של הפיקסלים (באופן שטוח , כלומר תמונה בגודל 28*28 משוטחת לגודל 784 , עוברת בפרמוטציה , ואז מוחזרת לגודל 28*28) .
זוהי השיטה המוצגת במאמר , ואכן בשיטה זו קיבלנו תוצאות דיוק גבוהות[3] כמו של המאמר:

[2] מינוח לא מדוייק (הכוונה להצפנת הדוגמאות בדרכן לאימון במודל)
[3] גרף המתאר את תוצאות הדיוק מצורף בסוף הדו"ח

בס"ד

זוהי העמודה הרלוונטית (שלב ההתקפה הוא רק באבן דרך 4)

אחוזי השגיאה במאמר

Attack	Classical classifier		Classifier on permuted data	
	original	attacked	original	attacked
MNIST				
CW ℓ_2	1.00	100.00	3.00	8.64
CW ℓ_0	1.00	100.00	3.00	14.53
CW ℓ_∞	1.00	99.99	3.00	12.24
FGSM	1.00	92.10	1.40	18.00
Fashion MNIST				
CW ℓ_2	7.50	100.00	11.50	12.12
CW ℓ_0	7.50	100.00	11.50	13.48
CW ℓ_∞	7.50	99.90	11.50	12.55
FGSM	8.60	60.60	11.20	27.50

אחוזי השגיאה אצלנו

	permutated-model		
dataset \ model	CW_1	CW_2	FGSM
mnist	3.37%	3.50%	3.17%
fashion_mnist	12.22%	12.30%	12.33%

ב. לאחר מכן , ניסינו את שיטות ההצפנה (**שאינן מופיעות במאמר**) המתוארות בסעיפים ב – ד , אבל כמו שניתן לראות בתוצאות דלקמן (התוצאות הן אחוזי שגיאה) שאחוזי הדיוק הם נעים סביב 10% , ומכיוון שבdatasets עליהם אנו עובדים יש 10 מחלקות לסיווג , המשמעות היא שהמודלים כלל אינם אפקטיבים ואינם לומדים כלל שכן המודלים צודקים בסיווג (לפי הtest) רק ב10% מהמקרים , לכן זה כמו להגריל את המחלקה .

	AES-ECB-model			AES-CBC-model			AES-CTR-model		
dataset \ model	CW_1	CW_2	FGSM	CW_1	CW_2	FGSM	CW_1	CW_2	FGSM
mnist	88.65%	90.18%	90.42%	88.65%	88.65%	88.65%	88.65%	89.90%	89.72%
fashion_mnist	58.60%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%

- ניתן לראות במודל שמוצפן בהצפנת AES במוד הפעלה של ECB על הdataset fashion_mnist, שהשגיאה אמנם גדולה אך קטנה משמעותית מהשאר . נשים לב לעובדה זו כאשר ננסה להצפין בדרכים אחרות באבן דרך 3 .

את ההצפנות דלעיל הצפנו בשיטה של השטחת התמונה (שיטה שתוארה מקודם) . נחקור באבן דרך 3 (המתוארת לקמן) שיטות אחרות .

פירוט שעות

ישבנו יחד כדי לבחון ולהבין את ההצפנות השונות, ולהבין איזה ספריות קוד צריך להתקין (לקח לא מעט זמן), ולאחר מכן התפצלנו שוב כדי לממש את ההצפנות ושוב לאמן את כולם על הdatasets השונים (כלומר לפחות 24 אימונים, אם אין באגים). ולבסוף שוב נפגשנו בכדי לרכז את התוצאות ולבדוק ציפיות מול המאמר ודיון על מה נעשה בהמשך.

סה"כ: בערך 85 שעות עבודה כל אחד .

3) <u>הצפנות יותר טובות בהתאמה למודלים / לDatasets . [זמן משוערך : חודש-חודשיים]</u>
ננסה להצפין את הdata בלי "לשבש" את התמונה , כלומר לנסות להשאיר יחסים בין פיקסלים בתמונה על מנת שהמודל יהיה מדויק , כלומר לנסות לשמור על תתי-מבנה של התמונה שמלמדים על סיווג התמונה לאחר ההצפנה .

- ממה שראינו , מה שננסה בשלב הבא יהיה לנסות להצפין את התמונות שיטה של הצפנת בלוקים (מתוך המטריצה בגודל 28*28) , למשל בגדלים של 4*4 .
- אכן קיבלנו אחוזי דיוק כמו שרצינו באבן דרך 2 , סעיף א (כלומר על ה permutated models) לכן לא נעסוק באבן דרך זו הרבה זמן לפני שנעבור להתקפות עצמן (המתאורות באבן דרך 4) .

בס"ד

גרפים המתארים את הדיוק :