בס"ד

דו"ח תקופתי 2 לפרוייקט

סטיב גוטפרוינד וישי אשר

סיכום דו"ח תקופתי 1

- בנינו מודלים של רשתות עמוקות בהתאם למוצג במאמר .
- הרצנו את המודלים על הdatasets הבאים : MNIST FASHION-MNIST CIFAR-10 והדיוק של המודלים היה גבוה כמצופה . חוץ מאשר CIFAR-10 שהדיוק אצלו לא היה מספק (**וגם לא הוזכר במאמר**) על כן לא המשכנו לעבוד איתו .
- ביצענו שיטות הצפנה שונות על תמונות אשר נכנסות למודלים השונים כגון :
  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.
- בשיטה הראשונה (פרמוטציה) המודל אכן הצליח ללמוד בצורה משמעותית כמו שהראינו , אמנם בשאר השיטות אחוזי הדיוק היו שקולים להגרלה . (כלומר לא הייתה מטרה אמיתית להמשיך ולתקוף את המודלים הללו שכן הם בכלל לא לומדים) .

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

המשך - אבני דרך (אבני דרך 2-1 כוסו בדוח הקודם)

3) הצפנות יותר טובות בהתאמה למודלים / לDatasets . [זמן משוערך : חודש-חודשיים]
חלק זה נעשה במידה ולא נגיע לרמת דיוק מספקת באבן דרך 2 .
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 .

תיאור העבודה שהושלמה

בחלק זה כמובא בסוף הדוח הקודם ניסינו להצפין בצורת בלוקים בהצפנות השונות – אבל עדיין לא קיבלנו תוצאות יותר טובות בהצפנת AES בכל הmodes of operations [למעט ECB שבשיטה הזו התוצאות היו בערך 50% שגיאה עבור FASHION-MNIST ו 20% שגיאה על MNIST . אמנם זה מעניין אבל שוב לא דיוק מספיק טוב כדי שיהיה מעניין לתקוף אותו . (זהו עניין שמעניין לחקור בפני עצמו מעבר לפרוייקט)]

ניתן לראות את האחוזים המדוברים בטבלה הבאה המתארת את אחוזי השגיאה (האחוזים בעמודה השמאלית יותר הם עבור MNIST ואילו בעמודה הימנית הן עבור FASHION-MNIST , וכן מה שכתוב "irrelevant" הכוונה היא להתקפות שכן באחוזים שגיאה כאלו – ההתקפות לא רלוונטיות כמבואר לעיל) :

בס"ד

ECB	CW l_2	encrypt v1	16.58			55.66	irrelevant	irrelevant
		encrypt v2	18.11			41.97		
	FGSM	encrypt v1	20.88			59.23		
		encrypt v2	19.95			46.25		
CBC	CW l_2	encrypt v1	64.07	irrelevant	irrelevant	72.12	irrelevant	irrelevant
		encrypt v2	69.12			64.47		
	FGSM	encrypt v1	88.65			90		
		encrypt v2	88.65			90		
CTR	CW l_2	encrypt v1	88.65	irrelevant	irrelevant	90	irrelevant	irrelevant
		encrypt v2	88.65			90		
	FGSM	encrypt v1	88.65			90		
		encrypt v2	88.65			90		

פירוט שעות

סה"כ: בערך 20 שעות עבודה כל אחד (שכן היינו חייבים לעבור להתקפות עצמן) .

4) לשלב זה יש שתי אפשרויות :
במקרה בו אבן דרך 2 (או 3 במידת הצורך) נגמרה בכך שהגענו לרמת דיוק מספקת (ירידה בדיוק בכמות אחוזים זניחה מהדיוק המקורי) , אזי :
התקפות על המודל המוצפן : [זמן משוערך : חודש וחצי - חודשיים]
נבנה ונריץ התקפות Adversarial Examples על המודלים המוצפנים שיצאו ברמת דיוק מספקת.
התקפות כגון :
- CW
- FGSM

תיאור העבודה שהושלמה

ראשית כל , נקדים ונאמר שלאחר עיון נוסף במאמר התברר כי המודלים CW1 , CW2 הן המודלים שאותם תקפו במאמר המקורי של התקיפה CW. (של Carlini & Wagner) . אמנם , במאמר עצמו כתוב כי המודל CW2 הוא מודל שנועד לנתונים של CIFAR-10 ומכיוון שבמאמר עצמו לא תקפו (או אימנו) מודל שלומד את CIFAR-10 . אז התמקדנו במודל CW1 בעיקר (והוא המודל היחיד שיותקף בהתקפת CW) , וכמובן במודל FGSM שמתאים להתקפת הFGSM . (במאמר הציגו את שני המודלים)

כעת , מכיוון שרק עבור שיטת ההצפנה של פרמוטציה קיבלנו אחוזי דיוק מספקים תקפנו בעזרת שתי ההתקפות את המודלים המתאימים על שני הDATASETS : MNIST and FASHION-MNIST.

בשלב זה אמנם , נתקלנו בבעיות מימוש :

בס"ד

ההתקפות CW וFGSM על המודלים הרגילים ללא הפעלת פרמוטציה עברו בהצלחה מרובה .[1] (100% הצלחה בהתקפה של CW עבור MNIST and FASHION-MNIST , ועבור התקפת FGSM הצלחה של 90% מול MNIST , והצלחה של 80% מול FASHION-MNIST)[2]

לגבי בעיות מימוש ובעיות לוגיות שנתקלנו בהן במהלך מימוש ההתקפה על מודלים שמעבירים את הקלטים שלהם בפרמוטציה נפרט בדוח הסופי .[3] כחלק מתהליך זה שלחנו בעצת המנחה שלנו לפרוייקט מייל לאחד מכותבי המאמר ,ולאחר שלא קיבלנו תגובה ממשית לבעיה שלנו שלחנו מייל גם לCARLINI אחד מכותבי התקפת CW, שהוא ייעץ לנו בתחום .

כתוצאה מהתכתבויות אלו וניסיון חוזר ונשנה למימוש ההתקפה ולהגיע לאחוזים דומים לשל המאמר , הרצנו שני גירסאות של תקיפה :

א) תקיפה בה לתוקף יש גישת אורקל לפונקציית predict שמבקשת פרדיקציה על תמונה , והפונקציה מעבירה את הפרמוטציה הסודית על התמונה לפני שנשלחת לסיווג במודל . (בטבלה בעמוד הבא התקפה מסוג זה מובאת בשם scenario 1)

ב) תקיפה בה לתוקף יש גישה רק לפונקציית predict ללא שום הפעלת הפרמוטציה . (בטבלה בעמוד הבא התקפה מסוג זה מובאת בשם scenario 2)

בהתקפה מסוג א – קיבלנו שהתוקף מצליח לתקוף כמו מודל שאינו מעביר בפרמוטציה סודית .

בהתקפה מסוג ב :

(1) כמבואר לעיל התוקף מקבל גישה לפונקציה הזו, ומייצר adversarial images והבדיקה של אחוזי הצלחת ההתקפה היא מול מודל שמעביר פרמוטציה סודית על הקלטים שלו . (ממדל את חוסר הידיעה של התוקף על הפרמוטציה הסודית)

(2) מקבלים אחוזי הצלחת התקפה נמוכים . כלומר הצלחה בהגנה נגד ההתקפה . (לפחות בגירסה הזו של התוקף)

בעמוד הבא מצורפת טבלת אחוזי השגיאה (=הצלחת ההתקפה בעמודות המתאימות להתקפה) על המודלים שאינם עושים שום דבר לקלט (המודלים הרגילים) , לעומת המודלים שעושים פרמוטציה . (חלק מן ההתקפות עדיין בהרצה והן מתוכננות לסיום עד חודש הבא)

- מה שרשום CW : l_0 l_2 l_∞ הכוונה היא לנורמות השונות שיש בהתקפת CW .

---

[1] הכוונה באחוזי הצלחה של התקפה היא בעצם בכמה אחוזים מתוך התמונות המודל **שגה** .

[2] נציין שקיבלנו אחוזים דומים לאלו של המאמר , אולם בFGSM האחוזים טיפה שונים אבל הפרמטרים של ההתקפה לא היו רשומים .

[3] על חלקים אלו הוקדש מירב הזמן בחלק זה . (וכמובן על הרבה הרצות)

Classification error (%) on the first 1000 test samples

		mnist			fashion_mnist		
		original	attacked		original	attacked	
			scenario 1	scenario 2		scenario 1	scenario 2
UNENCRYPTED	CW l_2	0.97	100		8.66	100	
	CW l_0		100			100	
	CW l_∞		100			100	
	FGSM	1.5	82.94		10.62	94.25	
PERMUTATED	CW l_2	3.63	100	4.5	12.4	100	12.7
	CW l_0		100	7.3		100	
	CW l_∞		100			100	
	FGSM	3.02	89.14		12.04	91.82	

פירוט שעות

סה"כ: בערך 180 שעות עבודה כל אחד .

תוספת :

PADDING

כחלק מהפרוייקט נתקלנו בכך שמודל שמעביר פרמוטציה רנדומית על הקלטים שלו עדיין מצליח ללמוד בצורה טובה .

רצינו לבדוק בעצת המנחה שלנו , מה יקרה אם התמונות שהמודל ילמד יהיו יותר גדולות (כלומר למשל בMNIST התמונות הן מגודל 28*28 , נרצה לבדוק מה יקרה אם נגדיל את התמונה לגודל יותר גדול)

ההגדלה היא ע"י ריפוד באפסים (כלומר בפיקסלים לבנים) מסביב לתמונות המקוריות . מה שרצינו לבדוק בצורה זו היא האם המודל עדיין יצליח ללמוד בצורה טובה על אף שהפרמוטציה הרנדומית עלולה לפזר את הפיקסלים באופן כזה שלא תהיה תבנית שיוכל המודל ללמוד על התמונות לאחר שעברו פרמוטציה .

כפי שניתן לראות על מה שהספקנו להריץ – עדיין המודל (CW1 שהוא המודל היותר חזק)

מצליח ללמוד בצורה מאוד טובה . (וזה מפתיע – כי היינו מצפים שככל שהמידע גדול יותר הפיזור של הפיקסלים יהיה עם פחות תבניות)

accuracies of permutation on different image sizes
(padding done with 0's around the original)

	image size	error rate	min/epoch
mnist	28x28	3.63	4
	40x40	2.65	5
	60x60	2.69	12
	100x100	2.3	14
fashion_mnist	28x28	12.4	
	40x40	12.07	13
	60x60		
	100x100		

פירוט שעות

סה"כ: בערך 40 שעות עבודה כל אחד .