בס"ד

תוכנית העבודה לפרוייקט

סטיב גוטפרוינד וישי אשר

טכנולוגיות וכן ספריות שיהיו בשימוש במהלך הפרוייקט

נתכנת בשפת python , כאשר עיקר העבודה שלנו הוא מול מודלים של רשתות עמוקות . על כן נשתמש בספרייה tensorflow שהיא הספריה המרכזית איתה עובדים בפיתוח רשתות עמוקות , כמובן גם נשתמש בnumpy לצורך חישובים (כגון פרמוטציה רנדומית). בנוסף נתעסק בהצפנות על כן נשתמש בספרייה pycrypto , ואולי עוד ספריות בדומה לה. בנוסף נייבא, ונשנה בהתאם לצורכנו ,את אלגוריתמי התקיפה (תקיפת Adversarial examples) CW , FGSM (וייתכנו עוד) שממומשים ב- python .

אבני דרך

1) תכנון הפרויקט ושחזור המודל [זמן משוערך : חודש]
בחלק זה ,**שכבר עשינו**, המטרות הן :
   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 .
2) הצפנת המודל[1] [זמן משוערך : חודש וחצי - חודשיים]
עוד לפני שאנחנו נפנה להתקפות על המודל , בשלב זה נצטרך 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 נממש ונבדוק 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.
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 .
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

[1] מינוח לא מדוייק (הכוונה להצפנת הדוגמאות בדרכן לאימון במודל)

במקרה בו אבן דרך 2 ואבן דרך 3 נגמרו בכך שאין לנו רמת דיוק מספקת :
<u>החלת המפתח במקום אחר במודל : [זמן משוערך : חודשיים]</u>
לשם ההצפנה , במאמר ובעקבותיו אנחנו באבני דרך הקודמים , ניגע בהצפנות על הDATA עצמו כלומר ההצפנה היא על התמונה לפני שנכנסת לאימון ברשת . אמנם ניתן לבחון חלופות אחרות לשימוש במפתח הסודי , למשל קביעת היפר-פרמטרים כגון :

א. Learning rate.
ב. Dropout level.

ננסה לחשוב על כאלו אפשרויות בין אם בקביעת ההיפר-פרמטרים השונים ובין אם קביעת כמות השכבות , או מודיפיקציה כלשהי על השכבות והמשקולות עצמן .

כמובן אחרי שלב זה ניתן לבצע את התקיפות המתוארות באפשרות דלעיל .

5) <u>בחינת שיטות התקפה נוספות : [זמן משוערך : חודש]</u>
אבן דרך זו תלויה בכמות הזמן שיישאר לנו בשלב זה , בהתחשב בשלב הסיכום . (אבן דרך 6)
בשלב זה נבחן שיטות התקפה נוספות מלבד אלו המתוארת ב4 , כמו IFGSM , ואחרות.

6) <u>סיכום הפרויקט ומצגת : [זמן משוערך : חודש]</u>
בשלב זה נבצע עיבוד של כל מה שנראה במהלך אבני הדרך הקודמים , כלומר נכין מצגת ופוסטר בהתאם לדרישות על בסיס מה שהצלחנו להשיג :
בין אם הצלחנו לאשש את הכתוב במאמר , כלומר ההצפנה אכן צלחה מול ההתקפות בהורדת אחוזי דיוק בכמות זניחה בעוד הבטיחות עלתה משמעותית .
ובין אם לא , כלומר גם אם מצאנו שיטה אחרת להצפנה , או אפילו שיטת התקפה ש"שוברת" את המודלים המדוברים במאמר .
זאת על סמך כל ההצפנות וההתקפות דלעיל .
נכין טבלאות דיוקים על סמך הצפנות שונות והתקפות שונות על מנת להמחיש ויזואלית את מה שנשיג במהלך כל הפרויקט .