סיכום תוצאות מטלה 1 – Defensive Distillation

מצורף בתקיה מחברת העבודה, מסודרת לפי סדר השאלות.

הרשת ששומשה לאורך כל האימון זהה לרשת שעבדנו איתה בכיתה למעט השכבה האחרונה אשר שונתה לפי הצורך (עם או בלי טמפרטורה, מספר מחלקות) הרשת מגיע ל98 אחוז דיוק על סט האימון.

- תקיפת הרשת הפשוטה:
  - FGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי – 0.77
    - מרחק פטרובציה אוקלידי ממוצע – 0.23
  - TGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי - 0.77
    - אחוז הגעה לסיווג המטרה – 0.34
    - מרחק פטרובציה אוקלידי ממוצע – 0.23
  - Untargeted PGD : epsilon - 4 , iterations – 30, epsilon step – 0.05
    - אחוז שינוי הסיווג המקורי – 1
    - מרחק פטרובציה אוקלידי ממוצע – 0.14
  - Targeted PGD : epsilon - 4 , iterations – 30, epsilon step – 0.05
    - אחוז שינוי הסיווג המקורי - 1
    - אחוז הגעה לסיווג המטרה – 1
    - מרחק פטרובציה אוקלידי ממוצע – 0.23
- תקיפת הרשת תחת הגנת Defensive Distillation: temperature – 30
  - FGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי – 0.09
    - מרחק פטרובציה אוקלידי ממוצע – 0.02
  - TGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי - 0.53
    - אחוז הגעה לסיווג המטרה – 0.28
    - מרחק פטרובציה אוקלידי ממוצע – 0.23
  - Untargeted PGD : epsilon - 4 , iterations – 30, epsilon step – 0.05
    - אחוז שינוי הסיווג המקורי – 0.09
    - מרחק פטרובציה אוקלידי ממוצע – 0.01
  - Targeted PGD : epsilon - 4 , iterations – 30, epsilon step – 0.05
    - אחוז שינוי הסיווג המקורי – 0.81
    - אחוז הגעה לסיווג המטרה – 0.76
    - מרחק פטרובציה אוקלידי ממוצע – 0.14

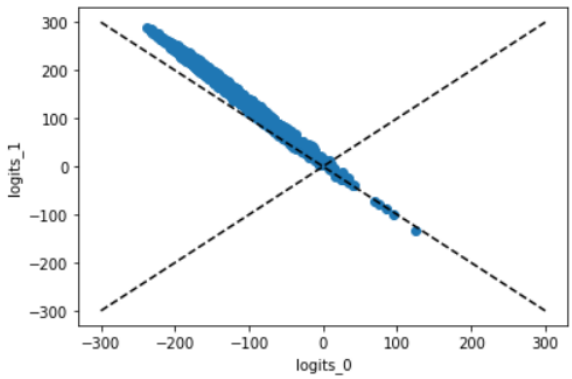
מעבר לבעיית סיווג בינארית (שתי מחלקות בלבד, נבחרו רנדומית).

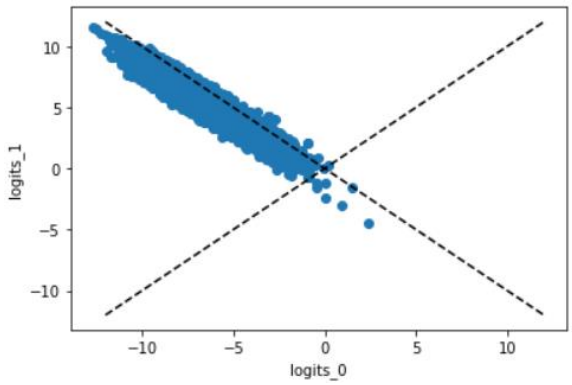
- תקיפת הרשת הפשוטה:
  - FGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי – 0.71
    - מרחק פטרובציה אוקלידי ממוצע – 0.24
  - TGSM : epsilon – 0.3
    - אחוז שינוי הסיווג המקורי - 0.7
    - אחוז הגעה לסיווג המטרה – 0.7
    - מרחק פטרובציה אוקלידי ממוצע – 0.24

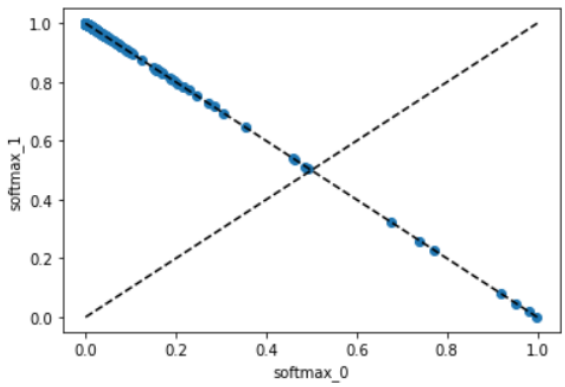
- Untargeted PGD : 4 - epsilon , 30 – iterations ,0.05 – epsilon step
  - אחוז שינוי הסיווג המקורי – 0.98
  - מרחק פטרובציה אוקלידי ממוצע – 0.14
- Targeted PGD : 4 - epsilon , 30 – iterations ,0.05 – epsilon step
  - אחוז שינוי הסיווג המקורי – 0.98
  - אחוז הגעה לסיווג המטרה – 0.98
  - מרחק פטרובציה אוקלידי ממוצע – 0.14

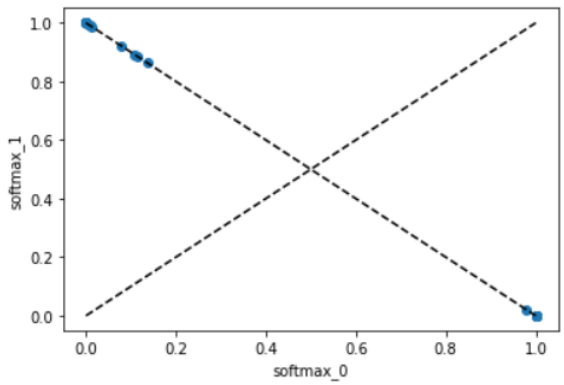
- תקיפת הרשת תחת הגנת Defensive Distillation: 30 – temperature
  - FGSM : 0.3 – epsilon
    - אחוז שינוי הסיווג המקורי – 0.04
    - מרחק פטרובציה אוקלידי ממוצע – 0.01
  - TGSM : 0.3 – epsilon
    - אחוז שינוי הסיווג המקורי - 0.55
    - אחוז הגעה לסיווג המטרה – 0.55
    - מרחק פטרובציה אוקלידי ממוצע – 0.24
  - Untargeted PGD : 4 - epsilon , 30 – iterations ,0.05 – epsilon step
    - אחוז שינוי הסיווג המקורי – 0.04
    - מרחק פטרובציה אוקלידי ממוצע – 0.01
  - Targeted PGD : 4 - epsilon , 30 – iterations ,0.05 – epsilon step
    - אחוז שינוי הסיווג המקורי – 0.7
    - אחוז הגעה לסיווג המטרה – 0.7
    - מרחק פטרובציה אוקלידי ממוצע – 0.13

ניתוח מעמיק לערכי הlogits, softmax עם\בלי (מימין עם, משמאל בלי) ההגנה:









כפי שניתן לראות בתרשימים, ערכי הlogits ברשת המוגנת גדלו בערך פי 30 לאומת הרשת המקורית. שינוי זה נכפה על הרשת בשל ההגנה על ידי אופן האימון המיוחד יחד עם שימוש בטמפרטורה גבוהה שמושכת את ערכי ההחלטה וגורמת למודל ליהיות יותר בטוח בהחלטות שלו. כאשר התהליך מתכנס ערכי הsoftmax והlogits גבוהים ואנו מקבלים מצב שהגרדיאנט "לכיוון" מטרת האמת כבר מתאפס ממש ולכן גם ערך הsign שלו מתאפס ומכאן שאנו לא מצליחים כלל לייצר פרטורבציה יעילה במקרה של untargeted. לאומת זאת במקרה של targeted attack , כן קיים גרדיאנט ולכן ההתקפה כן מצליחה לעבוד.
אינטואיטיבית המודל כבר מאוד בטוח בהחלטה שלו ולכן הוא לא מסוגל לאפטם את עצמו עוד לכיוון מטרת האמת == גרדיאנט 0 , לאומת זאת אם נשנה את המטרה המודל כן יכול לאפטם את עצמו לכיוון החדש == יש גרדיאנט.
נשים לב מהתוצאות המוקדמות יותר כי אומנן הuntargeted attack נחסם קליל, אך ניתן לראות שההגנה הקשתה מעט על הtargeted attack אך זה לא מספיק כדי לעצור את ההתקפה לגמרי.