Gallil Maimon 208943951

**Introduction to Adversarial Machine Learning – Ex1**

**Regular classifier results – 10 classes**

loss: 0.1928 - categorical\_accuracy: 0.9425

val\_loss: 0.0952 - val\_categorical\_accuracy: 0.9704

Attack method results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attack\metric | parameters | Out of original | moved to target | Euclidean distance |
| FGSM | epsilon=0.3 | 0.98 | - | 0.29 |
| TGSM | epsilon=0.3 | 0.93 | 0.55 | 0.29 |
| PGD | iterations=20, epsilon=6.0, iter\_eps=0.07 | **0.99** | **0.99** | **0.22** |

**Student (Defensive distillation) classifier results – 10 classes**

Temp: 20

\* metrics are on teacher labels

loss: 0.3565 - categorical\_accuracy: 0.9252

val\_loss: 0.2338 - val\_categorical\_accuracy: 0.9767

Attack method results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attack\metric | parameters | Out of original | moved to target | Euclidean distance |
| FGSM | epsilon=0.3 | 0.42 | - | **0.1** |
| TGSM | epsilon=0.3 | 0.9 | 0.55 | 0.23 |
| PGD | iterations=20, epsilon=6.0, iter\_eps=0.07 | **0.94** | **0.89** | 0.21 |

**Regular classifier – 2 classes**

Classes: [2, 7], num\_train\_samples: 12223, num\_validation\_samples: 2060

loss: 0.0391 - categorical\_accuracy: 0.9871

val\_loss: 0.0489 - val\_categorical\_accuracy: 0.9830

Attack method results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attack\metric | parameters | Out of original | moved to target | Euclidean distance |
| FGSM | epsilon=0.3 | 0.74 | - | 0.29 |
| TGSM | epsilon=0.3 | 0.74 | 0.74 | 0.29 |
| PGD | iterations=20, epsilon=6.0, iter\_eps=0.07 | **0.92** | **0.92** | **0.21** |

**Student (Defensive distillation) classifier results – 2 classes**

Temp: 20

\* metrics are on teacher labels

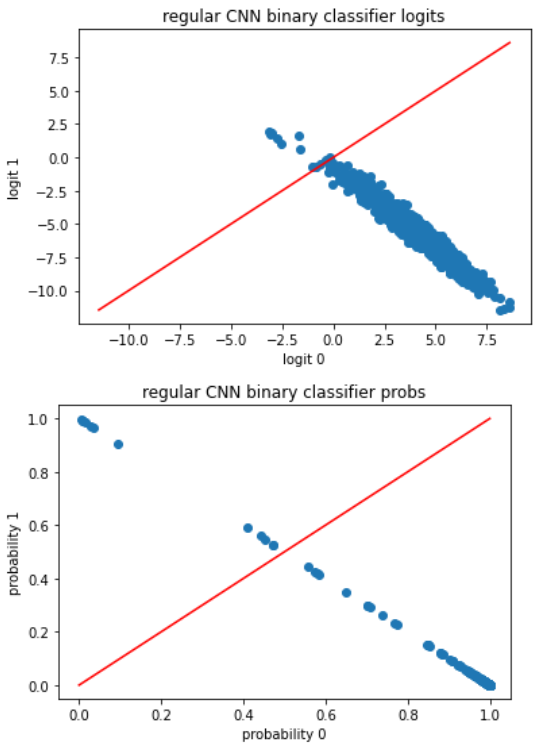
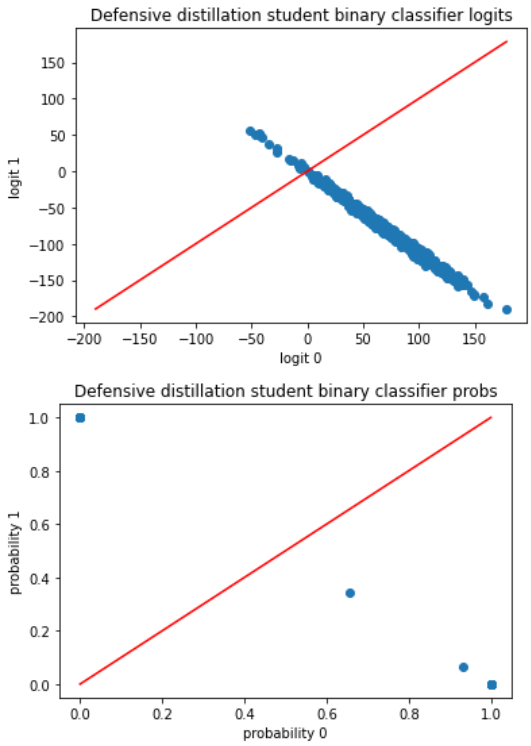
loss: 0.0604 - categorical\_accuracy: 0.9899

val\_loss: 0.0505 - val\_categorical\_accuracy: 0.9966

Attack method results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attack\metric | parameters | Out of original | moved to target | Euclidean distance |
| FGSM | epsilon=0.3 | 0.15 | - | **0.04** |
| FGSM | epsilon=10.0 | 0.15 | - | 1.44 |
| TGSM | epsilon=0.3 | 0.78 | 0.78 | 0.29 |
| PGD | iterations=20, epsilon=6.0, iter\_eps=0.07 | **0.83** | **0.83** | 0.19 |

**Logits and probability plots**

כפי שאפשר לראות, ערכי ה logits גדולים בסדר גודל (בערך פי 20) במודל שאומן באמצעות defensive distillation. בגישה זו שני המודלים התאמנו עם טמפרטורה ב softmax של 20 (מה שמסביר את פקטור גדילת ה logits). ניתן לראות שתופעה זו מתבטאת גם בהסתברויות יותר "נחרצות" כלומר יותר קרובות לאפס או אחד (כאשר עושים inference עם טמפרטורה של 1).

התופעה המתקבלת הגיונית משום שה loss קשור ל"הסתברות" אחרי ה softmax, וכן עבור טמפרטורה מסוימת T נדרשים logits גדולים פי T כדי להגיע לאותה הסתברות ובהתאמה לאותו loss.

זוהי גם ההתנהגות הרצויה עבור defensive distillation, שכן אם כל ההסתברויות הן זהות (1 או 0), אין גרדיאנט בנקודה הזו.