Adversarial Learning

Home Assignment I

Ido Calman & Matan Levy

Part I - TF Implementation of PGD Attack

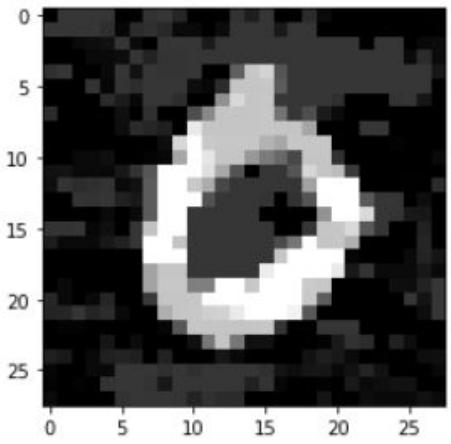
1. Below is a comparison between the FGSM and TGSM results (epsilon=0.3) on two platforms. Each result is a pair of mean perturbation distance and success rate.

Platform / Attack	FGSM	TGSM
NumPy	(0.29, 0.94)	(0.29, **0.56**)
TensorFlow	(0.29, **0.98**)	(0.29, 0.53)

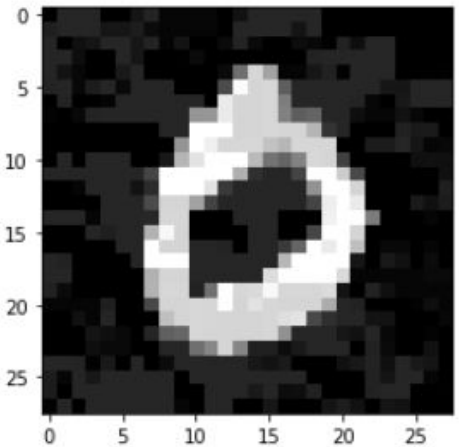
2. PGD Experiments:

Attack / Experiment	eps=4 iter_eps=0.05 iterations=30	eps=3 iter_eps=0.03 iterations=30	eps=3 iter_eps=0.03 iterations=60
Untargeted	(**0.97**, 0.14)	(0.8, 0.11)	(0.8, 0.11)
Targeted	(**0.8**, 0.14)	(0.42, 0.11)	(0.42, 0.11)

Untargeted example:



Targeted example:



Part II - Beyond PGD

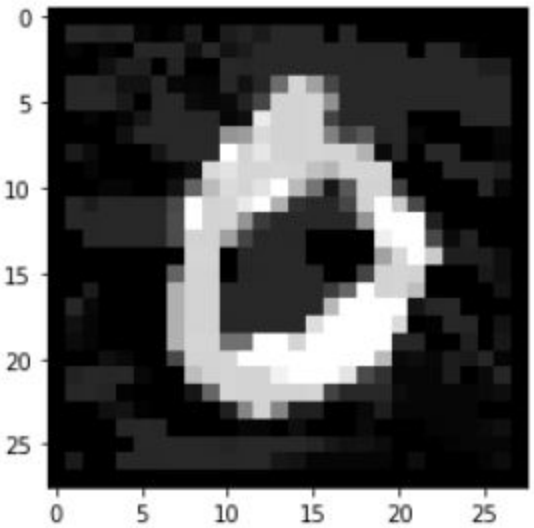
Frame perturbation restricted results:

Attack / Experiment	eps=4 iter_eps=0.05 iterations=30	eps=3 iter_eps=0.03 iterations=30	eps=3 iter_eps=0.03 iterations=60
Untargeted	(**0.97**, 0.14)	(0.83, 0.11)	(0.83, 0.11)
Targeted	(**0.83**, 0.14)	(0.47, 0.11)	(0.47, 0.11)

Non-lit pixels perturbation restricted results:

Attack / Experiment	eps=4 iter_eps=0.05 iterations=30	eps=3 iter_eps=0.03 iterations=30	eps=3 iter_eps=0.03 iterations=60
Untargeted	(**0.78**, 0.14)	(0.51, 0.1)	(0.5, 0.1)
Targeted	(**0.4**, 0.14)	(0.16, 0.1)	(0.16, 0.1)

Frame perturbation example:



Non-lit perturbation example: