ADVERSARIAL MACHINE LEARNING

DEFENCE STRATEGIES

30

- Varies based on goal of adversary
- For multi class targeted or non-targeted attack defence means making classifier robust against adversarial perturbations (e.g. Self Driving Cars)
- For anomaly detection etc scenario defence means detecting adversarial examples

DEFENCE STRATEGIES

- Adversarial Training
- Distillation
- Regularisation (Dropout, Wright Decay etc., Label Smoothing)
- Ensemble
- Virtual Adversarial Training

DEFENCE STRATEGIES

- Varies based on goal of adversary
- For multi class targeted or non-targeted attack defence means making classifier robust against adversarial perturbations (e.g. Self Driving Cars)
- For anomaly detection etc scenario defence means detecting adversarial examples