DEFENCE STRATEGIES

31

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

DEFENSE: ADVERSARIAL TRAINING

- Modified loss function:

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon sign(\nabla_x J(\theta, x, y)))$$

- Augment the adversarial examples into training dataset.

Algorithm 1 Adversarial training of network N.
Size of the training minibatch is m. Number of adversarial images in the minibatch is k.

1: Randomly initialize network N
2: **repeat**
3: Read minibatch $B = \{X^1, \ldots, X^m\}$ from training set
4: Generate k adversarial examples $\{X^1_{adv}, \ldots, X^k_{adv}\}$ from corresponding clean examples $\{X^1, \ldots, X^k\}$ using current state of the network N
5: Make new minibatch $B' = \{X^1_{adv}, \ldots, X^k_{adv}, X^{k+1}, \ldots, X^m\}$
6: Do one training step of network N using minibatch B'
7: **until** training converged

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale

DEFENCE STRATEGIES

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