ADVERSARIAL MACHINE LEARNING

DEFENSE: EXTENDED DEFENSIVE DISTILLATION

37

- Uncertainty Measure:

- Labelling Vector:

$$\sigma(x) = \frac{1}{N} \sum_{m \in 0..N-1} \left(\sum_{j \in 0...n-1} (z_j^m(x) - \bar(z_j)^2) \right)$$

$$k_j(x) = \begin{cases} 1 - \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = l \text{ (correct label)} \\ \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = n \text{ (outlier class)} \\ 0 & \end{cases})$$

Papernot, N., & McDaniel, P. (2017). Extending Defensive Distillation.

DEFENSE: VIRTUAL ADVERSARIAL TRAINING

- Uses both labeled and unlabelled datapoints
- Loss function:

$$\mathrm{LDS}(x_*, \theta) := D\left[p(y|x_*, \hat{\theta}), p(y|x_* + r_{\mathrm{vadv}}, \theta)\right]$$

$$r_{\mathrm{vadv}} := \arg\max_{r; \|r\|_2 \leq \epsilon} D\left[p(y|x_*, \hat{\theta}), p(y|x_* + r, \theta)\right],$$

$$\text{where } x_* \in \{D_l, D_{ul}\}$$

- Regularizer:

$$\mathcal{R}_{\mathrm{vadv}}(\mathcal{D}_l, \mathcal{D}_{ul}, \theta) := \frac{1}{N_l + N_{ul}} \sum_{x_* \in \mathcal{D}_l, \mathcal{D}_{ul}} \mathrm{LDS}(x_*, \theta).$$

- Objective Function:

$$\ell(\mathcal{D}_l, \theta) + \alpha \mathcal{R}_{\mathrm{vadv}}(\mathcal{D}_l, \mathcal{D}_{ul}, \theta),$$

-

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2017). Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning.

DEFENSE: EXTENDED DEFENSIVE DISTILLATION

- Uncertainty Measure:

$$\sigma(x) = \frac{1}{N} \sum_{m \in 0..N-1} \left(\sum_{j \in 0...n-1} (z_j^m(x) - \bar{(} z_j)^2) \right)$$

- Labelling Vector:

$$k_j(x) = \begin{cases} 1 - \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = l \text{ (correct label)} \\ \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = n \text{ (outlier class)}) \\ 0 \end{cases}$$

Papernot, N., & McDaniel, P. (2017). Extending Defensive Distillation.