ADVERSARIAL MACHINE LEARNING

DEFENSE: VIRTUAL ADVERSARIAL TRAINING

38

- Uses both labeled and unlabelled datapoints
- Loss function:
- Regularizer:
- Objective Function:

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

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

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

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

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

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

NO DEFINITIVE ANSWER

DEFENSE: VIRTUAL ADVERSARIAL TRAINING

- Uses both labeled and unlabelled datapoints
- Loss function:

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

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

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

- 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.