①

Unsupervised Data Augmentation (UDA)

Ⓟ unsupervised Data Augmentation (UDA)

UDA : Target model $p_{\theta}(y \mid x)$

$P_L(x)$ // Labeled data dist

$P_U(x)$ // Unlabeled data dist.

Perfect model f^*

Supervised Augmentation : $\hat{x} \sim q(\hat{x} \mid x)$

UDA : input x ; $P_\theta(y \mid x)$ and $P_\theta(y \mid x, \epsilon)$ → minimize Divergence → $\mathcal{D}\left(P_\theta(y \mid x) \,\|\, P_\theta(y \mid x, \epsilon)\right)$

noise

quality??

$\hat{x} = q(x, \epsilon)$

Total Objective

$$\min_\theta \mathcal{J}(\theta) = \begin{cases} \overbrace{\mathbb{E}_{x_1 \sim P_L(x)}\left[-\log P_\theta(f^*(x) \mid x_1)\right]}^{\text{supervised}} \\ + \lambda \underbrace{\mathbb{E}_{x \sim P_U(x)} \mathbb{E}_{\hat{x} \sim q(\hat{x} \mid x_c)}\left[CE\left(P_{\tilde{\theta}}(y \mid x_c) \,\|\, P_\theta(y \mid \hat{x})\right)\right]}_{\text{unsupervised}} \end{cases}$$

fixed copy (no grad) — $\tilde{\theta}$; updated network — θ

Sharpening Prediction : indicator

$$\frac{1}{|B|} \sum_{x \in B} \overbrace{I\left(\max_{y'} P_{\tilde{\theta}}(y' \mid x) > \beta\right)}^{\text{indicator}} CE\left(P_{\tilde{\theta}}^{sharp}(y \mid x) \,\|\, P_\theta(y \mid \hat{x})\right)$$

$$P_{\tilde{\theta}}^{sharp}(y \mid x) = \frac{\exp(z_y / \tau)}{\sum_{y'} \exp(z_{y'} / \tau)}$$

logit label for y

Theory:

In-domain: $P_U(\hat{x}) > 0$ for $\hat{x} \sim q(\hat{x}|x)$, $x \sim P_U(x)$

Label preserving: $f^*(x) = f^*(\hat{x})$ for $q(\hat{x}|x)$; $x \sim P_U(x)$

Reversible: if $q(\hat{x}|x) > 0$; then $q(x|\hat{x}) > 0$

Theorem: under UDA, $Pr(A)$: Algo. can't infer the label of new test example | m labeled example from P_L

$$Pr(A) = \sum_i P_i \underbrace{(1 - P_i)^m}_{\text{geometric (succeed after m try)}} \quad // \text{ Prob bound}$$

$$= \sum_{x \in C_i} P_L(x) \quad // \text{ labeled component.}$$

$P_i \Rightarrow$ observed example fall in i-th component

Further, if $m = O(k/\epsilon) \Rightarrow \underbrace{Pr(A) = O(\epsilon)}_{\text{Error Rate..}}$

component number (k)