Residual training

Labelled data $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N^l}$

$y_i^l \in \{1, \dots e^l\}$

output / Basic feature $b(x_i^l) = Net(x_i^l, \theta_{net}, \theta_6)$

previous layers.

final layer

labeled data feature

labeled data : stage 1

Objective loss:

Softmax w.r.t. y_i:

$$\mathcal{L}_{ce} = \frac{1}{N^l} \sum_{i=1}^{N^l} -\log p(y^i \mid x^i)$$

unlabeled data : stage 2

Residual loss

Dataset $\mathcal{D}^u = \{x_i^u\}_{i=1}^{N^u} \in C^u$

Objective 1 : Disentangling Residual feature for Clustering:

Residual feature:

$$r(x_i^u) = \text{Net}(x_i^u; \theta_{net}, \theta_m)$$

// parallel to basic feature.

MTL Setting

Total features:

Additive feature (ResNet)

$$c(x_i^u) = b(x_i^u) + r(x_i^u)$$

Obtain initial clusters from $\{b(x_i^u)\}_{i=1}^{N^u}$

$$U = \{\mu_k\}_{k=1}^{K}$$

↳ C^u dimensional.

feature Reduction

$$\tilde{c}(x_i^u) = fc(c(x_i^u); \theta_{fc}) \in \mathbb{R}^{C^u}$$

Soft assignment between each $\tilde{c}(x_j^u)$ and cluster center U.

student-t similarity

$$\left(1 + \|\tilde{c}(x_i^u) - \mu_k\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}$$

$$q_k(x_i^u) = \frac{\left(1 + \|\tilde{c}(x_i^u) - \mu_k\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_j \left(1 + \|\tilde{c}(x_i^u) - \mu_j\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}$$

Auxiliary target Distribution:

$$t_k(x_i^u) = \frac{q_k^2(x_i^u)/f_k}{\sum_j q_j^2(x_i^u)/f_j}$$

$$f_k = \sum_i q_k(x_i^u)$$ / normalization.

Overall cluster loss [Retraining]

??

$$\mathcal{L}_{cluster} = \frac{1}{N^u} \sum_{i=1}^{N^u} \sum_{k=1}^{K} t_k(x_i^u) \log \frac{t_k(x_i^u)}{q_k(x_i^u)}$$

just increase $q_k(x_i^u)$

Sharpening & matching

Objective 2: Preserving Basic without f_θ.

fine tune Output

$$\mathcal{L}_{distill} = \frac{1}{N^u} \sum_{i=1}^{N^u} \sum_{y=1}^{C^t} - \hat{p}(y|x_i^u) \log p(y|x_i^u)$$

$\hat{p}(y|x_i^u)$: frozen layer output

Objective 3: Pairwise labeling for residual feature:

$$\mathcal{L}_{pair} = -\frac{1}{(N^u)^2} \sum_{i,j}^{N^u} \left[A_{ij} \log s_{ij}^u + (1 - A_{ij}) \log (1 - s_{ij}^u) \right]$$

s_{ij}^u: $p(x_i^u) \odot p(x_j^u)$

A_{ij}: $\{1, 0\}$ if x_j is within top k-nn of sample x_i

full loss objective:

$$\mathcal{L}_{total} = \mathcal{L}_{cluster} + \beta \mathcal{L}_{distill} + \mathcal{L}_{pair}$$

Evolution Matrix

ACC defined by

predicted cluster

$$\max_{perm \in P} \frac{1}{N} \sum_{i=1}^{N} \delta(perm(\hat{y}_i) = y_i)$$