Deep transfer cluster:

unlabeled dataset: $D^u = \{x_i^u ; i = 1, \dots m\}$

output class assignment: $y_i^u \in \{1, \dots K\}$

K: unknown

labeled dataset: $D^l = \{(x_i^l, y_i^l) ; i = 1 \dots v\}$

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

L: known.

objective: to learn what forms a good class ??

& transfer the knowledge for new class. dis.

Transfer clustering & Representation learning:

Representation $z = f_\theta(x) \in \mathbb{R}^d$

initiated by labeled data

fine-tuned by unlabeled data

} well something new.

Joint clustering & representation learning:

summarize DEC

Collection of vector prototypes:

$$U = \{ \mu_k, k = 1, \dots K \}$$

↳ cluster center.

Task: ① clustering ② Discriminative task.

$p(k|i)$; prob of ith sample to k cluster.

$$p(k|i) \propto \left(1 + \frac{\| z_i - \mu_k \|^2}{\alpha} \right)^{-\frac{\alpha+1}{2}} \quad \text{— ①}$$

$\propto$: prop sign

sampling chance: $p(i) = \frac{1}{N}$

Joint distribution

$$p(i, k) = p(k|i) \frac{1}{N}$$

instead of maximizing p

→ DEC matches p with some other dist. q

[suitably shaped]

Assuming: $q(i,k) = q(k|i)\frac{1}{N}$

$$E(q) = KL(q\|p) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q(k|i)\log\frac{q(k|i)}{p(k|i)} \quad (10)$$

[requires alternative optimization]

How to select q

$$q(k|i) \propto p(k|i)\,\underbrace{p(i|k)}_{??}$$

[equalization effect ;)
high if cluster is not too large.]

$$q(k|i) \propto \frac{p(k|i)^2}{\sum_{i=1}^{N} p(k|i)} \quad (11)$$

← power to 2 sharpening

not many should go to k.

[used with in combination of KL divergence.]

Bottleneck: initialization U.

$$Z^u = \{ z_i = f_\theta(x_i^u) ; \; i = 1 \cdots m \}$$

$$z_i \in \mathbb{R}^d$$

Dimension reduction layer: $\hat{z}_i = Az_i + b$

projection layer.

[PCA]

Temporal ensemble & consistency:

Accumulating the network prediction:

$$p^t(k|i) = \beta p^{t-1}(k|i) + (1-\beta) p^t(k|i) \quad \text{— (4)}$$

$$\tilde{p}^t(k|i) = \frac{1}{1-\beta^t} \cdot p^t(k|i) \quad \text{— (5)}$$

[$\hat{z}_i^{t+1}$ | i] & optimization

5 to 3 then again $(q(k|i))$ is ...

introducing consistency: Similar examples & together.

Eq 2 becomes ↑

$$E(q) = \frac{1}{N} \sum_{i=1}^{N} \sum_{i=1}^{K} q(k|i) \log \frac{q(k|i)}{p(k|i)}$$

$$+ \omega(t) \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \| p(k|i) - \hat{p}(k|i) \|^2$$

$\omega(t)$ → ramp up function.

consistency term.

$\hat{p}$ → prediction of f_θ or ensembled

Ensambling number of class:

L known classes into probe dataset

D_r^l of L_r class → class number prediction.

& D^l / D_r^l of $(L - L_r)$ class // training subset.

further, L_n into D^l_{na} of L^a_n classes

& D^l_{nv} of L^v_n class

$D^l_n \cup D^u$ → estimate no of classes.

D^l_{na} → forced to a subset.

D^l_{nv} → additional unlabeled data.

measure no of class C in $D^l_n \cup D^u$

& Quality on $D^l_n \cup D^u$

[constrained k-mean multiple time]

cluster Quality Index:

average cluster accuracy

$$\max \sum_{} \frac{\text{gt}}{\text{cluster assign.}}$$

$g \in Sym(L^k_n)$ s.t. $\sum_{i=1}^{n} \{ y_i = g(y_i) \}$

↓

group permutation of L^k_n elements.

cluster validity index

$$\sum_{x \in D^k} \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

$a(x)$: avg dist. between x & others data.

$b(x)$: smallest avg dist of x to all points in any other cluster.