ⓟ Prototypical Contrastive learning

Prototypical Contrastive learning

Preliminaries

$X = \{x_1 \cdots x_n\}$ n images.

$f_\theta \rightarrow$ embedding function.

$X \rightarrow V = \{v_1, \cdots v_n\}$

$v_i = f_\theta(x_i)$

$$\mathcal{L}_{infonce} = \sum_{i=1}^{n} -\log \frac{\exp(v_i \cdot v_i'/\tau)}{\sum_{j=0}^{r} \exp(v_i \cdot v_j'/\tau)}$$

r negs. include the v_i'

$v_i' = f_{\theta'}(x_i)$

$\theta' \rightarrow$ moving avg of θ

PCL as EM:

$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \log p(x_i|\theta)$$ // maximize log-likelihood.

$$= \arg\max_\theta \sum_{i=1}^{n} \log \sum_{c_i \in C} p(x_i, c_i; \theta)$$

c_i: latent variable

law of total prob.

!! How to optimize this ??

$$\sum_{i=1}^{n} \log \sum_{c_i \in C} p(x_i, c_i|\theta) \geq \sum_{i=1}^{n} \sum_{c_i \in C} Q(c_i) \log \frac{p(x_i, c_i; \theta)}{Q(c_i)}$$ // ELBO

Focus

$$\sum_{c_i} Q(c_i) = 1$$

the equality holds if

$$Q(c_i) = p(c_i|x_i, \theta_j) = \frac{p(x_i, c_i|\theta)}{\sum_{c_i} p(x_i, c_i|\theta_t)}$$

E step:

Estimate $p(c_i|x_i;\theta)$,

k means on feature $v_i' = f_{\theta'}(x_i)$

↓

momentum encoder.

prototype c_i → centroid of the cluster.

compute $p(c_i|x_i,\theta) = \mathbb{1}(x_i \in c_i)$

⇓

sharp PDF.

M-step:

$$\sum_{i=1}^{n} \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i|\theta) = \sum_{i=1}^{n} \sum_{c_i \in c_i} p(c_i|x_i,\theta) \log p(x_i, c_i|\theta)$$

$$= \sum_{i=1}^{n} \sum_{c_i \in C} \mathbb{1}(x_i \in c_i) \log p(x_i, c_i|\theta)$$

$$p(x_i, c_i|\theta) = p(x_i|c_i,\theta)p(c_i|\theta) = \underbrace{\frac{1}{K}}_{\text{uniformity assumption}} p(x_i|c_i,\theta)$$

(P) Prototypical CL

Assuming isotopic gaussian.

$$p(x_i \mid c_i, \theta) = \exp\left(\frac{-(v_i - c_s)^2}{2\sigma_s^2}\right) \Big/ \sum_{j=1}^{K} \exp\left(\frac{-(v_i - c_j)^2}{2\sigma_j^2}\right)$$

$p(x_i \mid c_i, \theta)$ — catable form

By applying normalization of v & c we get.

$$p(x_i \mid c_i, \theta) = \exp\left(\frac{-(2 - 2\hat{v}_i \cdot \hat{c}_s)^2}{2\sigma_s^2}\right) \Big/ \sum_{j=1}^{K} \exp\left(\frac{-(2 - 2\hat{v}_i \cdot \hat{c}_j)^2}{2\sigma_j^2}\right)$$

$\hat{c}_j$ — vary

so maximizing log likelihood falls into.

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} -\log \frac{\exp\left[v_i \cdot c_s / \phi_s\right]}{\sum_{j=1}^{k} \exp\left[v_i \cdot c_j / \phi_j\right]} \quad ; \phi \propto \sigma^2$$

c_j — centroid of j cluster.

in practice the overall objective becomes.

$$\mathcal{L}_{\text{PROTONCE}} = \sum_{i=1}^{n} -\left[\log \frac{\exp(v_i \cdot \hat{v}_j^{+} / \tau)}{\sum_{j=0}^{r} \exp(v_i \cdot \hat{v}_j / \tau)} + \exp \frac{1}{m} \sum_{m=1}^{M} \log \frac{\exp\left(v_i \cdot c_s^m / \phi_s^m\right)}{\sum_{j=0}^{r} \exp\left(v_i \cdot c_j^m / \phi_j^m\right)}\right]$$

??

NCE

class prototype.

cluster m times !!

with different number of cluster ??

⇒ what if 1 is bad??

(iv)

Concentration estimation : ϕ (variance) (smaller means high concentration)

$\phi \leftarrow$ momentum features $\{v_z^{\wedge}\}_{z=1}^{Z}$ of same cluster c.

$$\phi = \frac{\sum_{z=1}^{Z} \| v_z^{\wedge} - c \|_2^2}{Z \log(Z + \alpha)}$$

should be smaller

α → smooth parameter.

scaling factor for e_s^m