Barlow Twin

Ⓕ Barlow twin

Batch of images x → y^A (Augmentation 1) ; y^B (Augmentation 2)

$z^A = f_\theta(y^A)$; $z^B = f_\theta(y^B)$ // mean centered.

$$\mathcal{L}_{BT} \triangleq \underbrace{\sum_i (1 - c_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} c_{ij}^2}_{\text{redundancy reduction term}}$$

cross correlation matrix.

[proxy entropy]

where, $c_{ij} = \dfrac{\sum_b z^A_{bi} z^B_{bj}}{\sqrt{\sum_b (z^A_{bi})^2 \sum_b (z^B_{bj})^2}}$

→ b indexes batch sample

→ i, j vector dimension of the network.

↳ summation over all the samples.

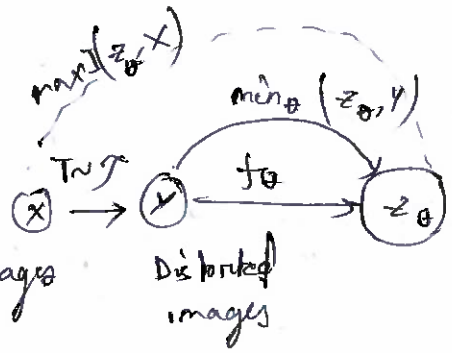
comparison with others:

$$\mathcal{L}_{\text{infoNCE}} \triangleq \underbrace{-\sum_b \frac{\langle z^A_b z^B_b \rangle_i}{\tau \|z^A_b\|_2 \|z^B_b\|_2}}_{\text{similarity term}} + \underbrace{\sum_b \log\left(\sum_{b' \neq b} \exp \frac{\langle z^A_b z^B_{b'} \rangle_i}{\tau \|z^A_b\| \|z^B_{b'}\|} \right)}_{\text{contrastive term.}}$$

non parametric estimation of Entropy of reps.

(F) Barlow twin (Appendix)

Information Bottleneck connection



$$IB_\theta \triangleq \underbrace{I(z_\theta, Y)}_{\text{close means invariant to } Y \text{ (Transformation)}} - \beta I(z_\theta, X) \qquad \text{As related to } X \text{ as possible.}$$

$$= \left[H(z_\theta) - H(z_\theta \mid Y)\right] \quad (H(z_\theta \mid Y) \to 0 \text{ (deterministic)})$$

$$- \beta \left[H(z_\theta) - H(z_\theta \mid X)\right]$$

(can be negative) the smaller the better.

β controlable

$$\approx H(z_\theta \mid X) + \underbrace{\frac{1-\beta}{\beta}}_{+ve} H(z_\theta) \quad \text{if } \beta = 1 \text{ // always positive}$$

then total positive (Entropy)

$z_\theta \sim$ gaussian constraints.

$$IB_\theta = E_X \log|C_{z_\theta|X}| + \frac{1-\beta}{\beta} \log|C_{z_\theta}|$$

($C_{z_\theta|X}$: covariance function; C_{z_θ}: covariance function)

connected to IB with some modification & constraints.

⑥ Demystifying CL

Demystifying Contrastive Learning

Contrastive loss.

$$\mathcal{L}(D, D^+) = -\sum_{(x, x^+) \in D^+} \frac{\exp\left(f(x)^T f(x^+)/\tau\right)}{\exp\left[f(x)^T f(x^+)/\tau\right] + \sum_{\substack{x^- \in D \\ x, x^- \notin D^+}} \exp\left(f(x)^T f(x^-)/\tau\right)}$$

Measuring Invariance:

transformation t

invariant function h iff. $h(x) = h(t(x))$

↳ label of image $t(x)$ / (x)

formal Notion iff $y(x) = y(t(x))$; where, $t : x \to x$

the $h^*(x) = h^*(t(x))$

⇓

invariant for $t(x)$ & label (y)

definition of firing unit

$h(x) \in \mathbb{R}^n$; fire if $s_i h_i(x) > t_i$; $s_i \in \{-1, 1\}$

$f_i(x) = \mathbb{1}(s_i h_i(x) > t_i)$; $f(x) \in \mathbb{F}^n$

Global firing rate, $G(i) = E[f_i(x)]$ // t_i dependency.

(i: ith neuron)

t_i chosen such that if $G(i) = 1/|y|$ ↳ no of class.

weaker ⇓

↳ number of firing unit

· one class → one section firing; equal parts

(P) Demystifying CL

Local trajectory: $T(x) = \{ t(x, \gamma) \mid \forall \gamma \}$ // set of transformed version of x image.

Local firing rate is defined as below

$$L_y(i) = \frac{1}{|X_y|} \sum_{x \in X_y} \frac{1}{|T(x)|} \sum_{x \in T(x)} f_i(x) \quad // \quad X_y = \{ x \mid x \in X, y(x) = y \}$$

- i: ith neuron
- $L_y(i)$: fraction of time i neuron fires.
- $\frac{1}{|X_y|}$: Avg
- $\sum_{x \in X_y}$: All image in X_y
- $\frac{1}{|T(x)|}$: Avg
- $\sum_{x \in T(x)} f_i(x)$: measuring local firing for x & their transformation.

Target conditioned invariance $I_y(i) = \frac{L_y(i)}{G(i)}$ // → find (top-k) neurons.

Representation Invariance Score: (RIS):

commonalities in top k neurons for each classes.

ⓟ 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