Understanding Self supervised learning without negative

Two layer model example:

setup

online net weight $W \in \mathbb{R}^{n_2 \times n_1}$

predictor : $W_p \in \mathbb{R}^{n_1 \times n_2}$

target net : $W_a \in \mathbb{R}^{n_2 \times n_1}$

$x \in \mathbb{R}^{n_1}$ // input data

Two augmentation: $x : x_1, x_2 \sim P_{aug}(\cdot \mid x)$

$f_1 = W x_1 \in \mathbb{R}^{n_2}$ // online rep.

$f_{2a} = W_a x \in \mathbb{R}^{n_2}$ // target rep.

(I) BYOL objective

$$J(W, W_p) := \frac{1}{2} \mathbb{E}_{x_1, x_2} \left[\| W_p f_1 - \text{stopgrad}(f_{2a}) \|_2^2 \right]$$

(II) whereas $W_a \simeq \text{Exponential MA}(W)$

BYOL learning Dynamics:

learning rate ratio

weight decay

gradient

$$\dot{W}_p = \frac{\partial J}{\partial W_p} = \alpha_p \left(-W_p W (X + X') + W_a X\right) W^T - \eta W_p$$

$$\dot{W} = W_p^T \left(-W_p W (X + X') + W_a X\right) - \eta W$$

$$\dot{W}_a = \beta \left(-W_a + W\right)$$

[Everything is in here]

Expectation of outer product matrix.

$$\boxed{X := \mathbb{E}\left[\bar{x}_0 \bar{x}^T\right]} \quad \bar{x}(x) := \mathbb{E}_{x' \sim p_{aug}(\cdot|x)}\left[x'\right]$$

mean

Average augmented view of data point

Expected cov mat.

$$X' = \mathbb{E}\left[\underline{V_{x'|x}}\left[x'\right]\right] \quad \mathbb{E}(\text{cov})$$

covariance matrix of aug view x'

Requires Simplified Assumption: for analysis

→ (i) Proportional EMA

$$W_a(t) = \tau(t)\, W(t)$$

→ (ii) Isotopic Data augmentation:

Avg data covariance : $X = I$

data Avg avg cov : $X' = \sigma^2 I$

→ w_p is symmetric

→ Eigen decomposition.

F : correlation matrix of output of W

$F = W X W^T$

Findings ①

Eigen space of W_p aligns to F

[we can approximate W_p from F]

Theorm 3:

under some condition $F W_p - W_p F \to 0$

DirectPred Method:

Estimate $\hat{F}$ → Set W_p by the following

$$\hat{F} = \hat{U} \Lambda_F \hat{U}^T$$

$$\Lambda_F = \text{diag}[s_1, s_2 \ldots s_d]$$

then, define

$$p_j = \sqrt{s_j} + \epsilon \max_j s_j$$

$$\Rightarrow \quad W_p = \hat{U} \, \text{diag}[p_j] \, \hat{U}^T$$

To estimate correlation matrix

$$\hat{F} = \rho \hat{F} + (1-\rho) \mathbb{E}_B [f f^T]$$

Expectation over batch.