Non-local Neurons

Generic Non-local operation $y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j)$

weighted sum

step 1

Normalization

output position

Instantiation: Design choice of the f & g

$g(x_j) = W_g x_j$ // [1×1 convolution or 1×1×1]

Gaussian: $f(x_i, x_j) = e^{x_i^T x_j}$

$C(x) = \sum_{\forall j} f(x_i, x_j)$

embedded gaussian: $f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}$ → $W_\theta x_i$, → $W_\phi(x_j)$

* Generalization of Non local

Dot Product: $f(x_i, x_j) = \theta(x_i)^T \phi(x_j)$

Concentration: $f(x_i, x_j) = \text{ReLU}(W_f^T [\theta(x_i), \phi(x_j)])$

* Wrapping Everything in Non local box

$z_i = W_z y_i + x_i$ // step 2

(P) Slowfast video Network.

Slowfast video Network

Slow Path	Fast path
(i) CNN (3D ResNet)	(i) Another CNN
(ii) Large temporal stride. (T frames) process	(ii) Temporal stride τ/α ; $\alpha > 1$ (αT frames) process
(iii) Process one out of τ frames.	(iii) No temporal downsampling
(iv) Has higher channel $\Rightarrow$ high computation	(iv) Low no of channel $\Rightarrow$ low computation. $\beta = 1/\alpha$ (weakened spatial ability)
(v) (T, s^2, C)	$(\alpha T, s^2, \beta C)$

Lateral Connection

3 ways (i) Time to channel $\{\alpha T, s^2, \beta C\} \rightarrow (T, s^2, \alpha\beta C)$

(ii) Time Strided sampling $(T, s^2, \beta C)$

(iii) Time strided conv. (filter with 5×1^2, $2\beta C$)

MutliSiam

Preliminary : BYOL:

$x \rightarrow$ image

$v \sim \mathcal{T}(x)$

$v' \sim \mathcal{T}'(x)$

Online Net : Backbone with GAPL, MLP projection, MLP predictor

Target Net : similar to Online But No MLP predictor.

BYOL → stop gradient prevents collapse. (on target Net)

$$\ell_{1D\_img} \triangleq -\cos(q, z') = \frac{\langle q, z' \rangle}{\|q\|_2 \cdot \|z'\|_2}$$

BYOL target parameters update

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

$$\tau \in [0, 1]$$

Positive Samples in multi-instance Data:

Clustering with DL

① clustering with DL

Reconstruction loss : $L = \underset{\text{AE}}{d}(x_i, f(x_i)) = \sum_i \| x_i - f(x_i) \|^2$

Auto encoder : Non-cluster loss.

cluster losses :

② K-means loss : (Yang et al. 2006)

$$L(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} s_{ik} \| z_i - \mu_k \|^2$$

N : Data point, K : No of cluster, μ_k : cluster center, z_i : embedded Data point, s_{ik} : boolean value for cluster assignment

(Any Soft clase option) ??

② cluster assignment hardening :

$$q_{ij} = \frac{\left(1 + \| z_i - \mu_j \|^2 / v\right)^{-\frac{v+1}{2}}}{\sum_{j'} \left(1 + \| z_i - \mu_{j'} \|^2 / v\right)^{-\frac{v+1}{2}}}$$

v : constant

// lower distance to maximum

q_{ij} : softmax value.

$$P_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'} \left(q_{ij'}^2 / \sum_i q_{ij'} \right)}$$

for all instances ~~in j class~~

// target Distribution P

Finally, loss function (for Network to minimize)

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

(iii) Balanced assignment loss:

(Dizaji et. al 2017)

$$L_{ba} = KL(G||U)$$

→ uniform Distribution

where, $g_k = P(y=k) = \frac{1}{N} \sum_i q_{ik}$

(iv) Locality preserving loss: (Huang 2014)

$$L_{Lp} = \sum_i \sum_{j \in N_k(i)} S(x_i, x_j) \|z_i - z_j\|^2$$

$S(x_i, x_j)$ ← similarity measure between x_i, x_j

$N_k(i)$ ← set of k nearest Neighbor of x_i

(v) Group Sparsity loss: (Ng... 2002)

$$L_{gs} = \sum_{i=1}^{N} \sum_{g=1}^{G} \lambda_g \|\phi^g(x_i)\|$$

weight of sparsity $\lambda_g = \lambda \sqrt{n_g}$

λ ← constant, n_g ← group size.

Combined losses: $L(\theta) = \alpha \underbrace{L_c(\theta)}_{\text{Clustering}} + (1-\alpha) \underbrace{L_n(\theta)}_{\text{Non clustering loss}}$

⑥ Theoretical understanding of CL

Theoretical Understanding of CL

framework:

setup:
- $\mathcal{X}$ → Dataset
- $D_{sim} \sim (x, x^+)$
- $D_{neg} \sim x_1^-, x_2^-, \cdots x_k^-$
- Encoder family f
- $\therefore f: \mathcal{X} \to \mathbb{R}^d$ such that, $\|f(\cdot)\| \leq R$; $R > 0$

latent class: (x, x^+) → similar pair

more setup:
- family of latent class C
- $c \in C$
- D_c over the $\mathcal{X}$
- $D_c(x)$: How relevant x to c.?
- ρ : how class occurs naturally.

Semantic Similarity:

$$D_{sim}(x, x^+) = \mathop{E}_{c \sim \rho} D_c(x)\, D_c(x^+)$$

$$D_{neg}(x^-) = \mathop{E}_{c \sim \rho} D_c(x^-)$$

Supervised Tasks: for a labeled pair (x, c) ; $c \in \{1 \cdots l_{c+1}\}$

$$D_{\mathcal{T}}(x, c) = D_c(x)\, D_{\mathcal{T}}(c)$$

⑧ Theoretical understanding of CL

framework:

setup:
- $\mathcal{X}$ → Dataset
- $D_{sim} \sim (x, x^+)$
- $D_{neg} \sim x_1^-, x_2^-, \cdots x_k^-$
- Encoder family f
- $\therefore f: \mathcal{X} \to \mathbb{R}^d$ such that, $\|f(\cdot)\| \leq R \; ; R > 0$

latent class: (x, x^+) → similar pair

more setup:
- family of latent class C
- $c \in C$
- D_c over the $\mathcal{X}$
- $D_c(x)$: How relevant x to c.??
- ρ : how class occurs naturally.

Semantic Similarity:

$$D_{sim}(x, x^+) = \underset{c \sim \rho}{E} D_c(x) D_c(x^+)$$

$$D_{neg}(x^-) = \underset{c \sim \rho}{E} D_c(x^-)$$

Supervised Tasks: for a labeled pair (x, c) ; $c \in \{1 \cdots c+1\}$

$$D_{\mathcal{T}}(x, c) = D_c(x) D_{\mathcal{T}}(c)$$

Evaluation metric for representation:

Task $\mathcal{T} = \{c_1 \cdots c_{k+1}\}$

function $g: \mathcal{X} \to \mathbb{R}^{k+1}$ // linear classifier.

point $(x, y) \in \mathcal{X} \times \mathcal{T}$

loss $\triangleq L(\{g(x)_y - g(x)_{y'}\}_{y \neq y'})$ [Different from true class should be high]

~~f~~(k dim vector of Differences in Coordinate)

By considering Standard hinge loss:

$$L(v) = \max\{0, 1 + \max_i\{-v_i\}\}$$

logistic loss $L(v) = \log_2\left(1 + \sum_i \exp(-v_i)\right)$; $v \in \mathbb{R}^k$

$$L_{sup}(\mathcal{T}, g) := \mathop{\mathbb{E}}_{(x,c) \sim D_c}\left[L\{g(x)_c - g(x)_{c'}\}_{c \neq c'}\right]$$

For linear Classifier: $g(x) = W f(x)$ finally (k+1)

Further, $L_{sup}(\mathcal{T}, f) = \inf_{W \in \mathbb{R}^{(k+1) \times d}} L_{sup}(\mathcal{T}, Wf)$

TP: $W \ni$ mean for each class representation.

Mean classifier: W^*

c-th row $\Rightarrow \mu_c : \mathop{\mathbb{E}}_{x \sim D_c}[f(x)]$

Now, $L^u_{sup}(\mathcal{T}, f) = L_{sup}(\mathcal{T}, W^* f)$

(P) Theor. understanding of CL

Arg Supervised loss:

$$L_{sup}(f) := \mathop{E}_{\{c_i\}_{i=1}^{k+1} \sim \rho^{k+1}} \left[L_{sup}\left(\{c_i\}_{i=1}^{k+1}, s\right) \mid c_i \neq c_j \right]$$

for mean class

$$L^{\mu}_{sup}(f) := \mathop{E}_{\{c_i\}_{i=1}^{k+1} \sim \rho^{k+1}} \left[L^{\mu}_{sup}\left(\{c_i\}_{i=1}^{k+1}, f\right) \mid c_i \neq c_j \right]$$

CL Algorithm:

unsupervised loss: Population loss:

Neg number

$$L_{un}(f) := E\left[L\left(\left\{ f(x)^T \left(f(x^+) - f(x_i^-) \right) \right\}_{i=1}^{k} \right) \right]$$

Empirical counterparts: $\left(x_j, x_j^+, x_{j1}^-, \ldots x_{jk}^- \right)_{j=1}^{n} \in D_{sim} \times D_{neg}^{k}$

$$\hat{L}_{un}(f) = \frac{1}{n} \sum_{j=1}^{n} L\left(\left\{ f(x_j)^T \left(f(x_j^+) - f(x_{ji}^-) \right) \right\}_{i=1}^{k} \right)$$

Now,

$$L_{un}(f) = \mathop{E}_{c^+, c^-} \mathop{E}_{x, x^+ \sim D_{c^+}^2} \left[L\left(\left\{ f(x)^T \left(f(x^+) - f(x_i^-) \right) \right\} \right) \right]$$

$$\sim \rho^{k+1} \qquad x_i \sim D_{c^-}$$

Results and theorem:

Th. 1. $L_{sup}(\hat{f}) \leq \alpha L_{un}(f) + \eta\, Gen_M + \delta \quad \forall f \in f_0$

upper bound?? generalization error.

$M \to \infty$, $Gen_M \to 0$

$\alpha, \eta \to 1$, $\delta \to 0$

[if c is large, $L_{un}(f)$ can be small]

$$L_{sup}(\hat{f}) \leq L^{\epsilon}_{un}(f) + \beta s(f) + \eta\, Gen_M \quad \forall f \in f_0$$

β: dependent.

$P \to$ uniform, $|\mathcal{A}| \to \infty$ then $\beta \to 0$, $\eta \to 1$

Ideal result should be: $\boxed{L_{sup}(\hat{f}) \leq \alpha L_{sup}(f) + \eta\, Gen_M}$ $\forall f \in f_0$

However not true?!