InfoNCE loss and Comparison for DML and SSL

Shengqi Fang

Department of Statistics and Data Science, SUSTech

July 20, 2023

Different losses

Contrastive loss (Bromley et al. [1993] (Chopra et al. [2005])

- $\mathcal{L}_{\text{cont}}\left(\boldsymbol{Z}\right) = \sum_{(i,j)\in\mathbb{P}} \|\boldsymbol{z}_j - \boldsymbol{z}_i\|_2 + \sum_{(i,j)\notin\mathbb{P}} \operatorname{ReLU}\left(m - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2\right)^2, m > 0,$

Triplet loss ((Weinberger and Saul [2009] (Chechik et al. [2010])

- $\mathcal{L}_{\text{triplet}}\left(\boldsymbol{Z}\right) = \sum_{(i,j)\in\mathbb{P}} \sum_{\{(k,l)\notin\mathbb{P}, k=i\}} \operatorname{ReLU}\left(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2 - \|\boldsymbol{z}_i - \boldsymbol{z}_k\| + m\right), m > 0,$

Neighbourhood Component Analysis (Goldberger et al. [2004])

- $\mathcal{L}_{\text{NCA}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \frac{e^{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2}}{\sum_{(k,l)\in[N]^2} e^{-\|\boldsymbol{z}_k - \boldsymbol{z}_l\|_2^2}},$

Different losses

(N+1)-tuple loss(Sohn [2016])

- $\mathcal{L}_{\text{tuple}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \log\left(\frac{e^{\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle}}{\sum_{(k,l)\in\mathbb{P}} e^{\langle \boldsymbol{z}_i, \boldsymbol{z}_l\rangle}}\right) + \beta\|\boldsymbol{Z}\|_F^2,$

infoNCE loss(Oord et al. [2018, CPC])

- $\mathcal{L}_{\text{infoNCE}} = -\sum_{(i,j)\in\mathbb{P}} \log\left(\frac{e^{\text{CoSim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau}}{\sum_{k=1}^{N} e^{\text{CoSim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau}}\right)$

Contrastive Predictive Coding(CPC)

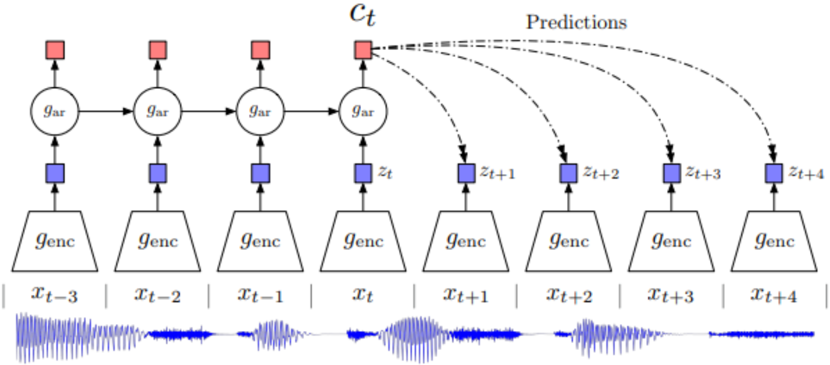
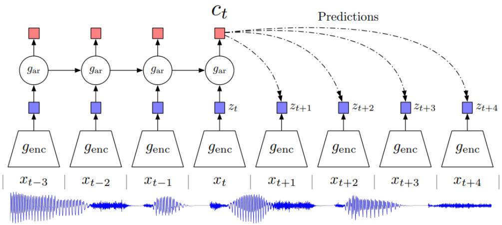


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

InfoNCE loss



- $I(x; c) = \sum_{x,c} p(x, c) \log \frac{p(x|c)}{p(x)}$
- $f_k\left(x_{t+k}, c_t\right) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$
- $f_k\left(x_{t+k}, c_t\right) = \exp\left(z_{t+k}^T W_k c_t\right)$

InfoNCE loss: $\mathcal{L}_\mathrm{N} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$

$I\left(x_{t+k}, c_t\right) \geq \log(N) - \mathcal{L}_\mathrm{N}$

InfoNCE loss

$$
\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\mathbb{E}_X \log \left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}} \right] \\
&= \mathbb{E}_X \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)} \right] \\
&\approx \mathbb{E}_X \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \mathbb{E}_{x_j} \frac{p(x_j|c_t)}{p(x_j)} \right] \\
&= \mathbb{E}_X \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \right] \\
&\geq \mathbb{E}_X \log \left[\frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N \right] \\
&= -I(x_{t+k}, c_t) + \log(N),
\end{aligned}
$$

- $I\left(x_{t+k}, c_t\right) \geq \log(N) - \mathcal{L}_{\mathrm{N}}$

InfoNCE loss

The infoNCE Offsprings

- He et al. [2020a, **MoCo**] introduces momentum encoder as an alternative to the memory bank regularization of eq. (5) and introduces a queue to store many negative samples from previous batches; [Chen et al., 2020d, **MoCoV2**] adds a projector, [Chen et al., 2021b, **MoCoV3**] adds ViTs
- Chen et al. [2020b, **SimCLR**] removes the momentum encoder and the i^{th} term from the denominator coining it **NT-Xent** (Normalized Temperature-scaled cross entropy)

$$\mathcal{L}_{\text{NT-Xent}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \frac{e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_j)}}{\sum_{k=1}^{N} \mathbf{1}_{\{k\neq i\}} e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_k)}},$$

- Yeh et al. [2021, **DCL**] additionally removes the positive pair in the denominator

$$\mathcal{L}_{\text{DCL}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \frac{e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_j)}}{\sum_{k=1}^{N} \mathbf{1}_{\{k\neq i \wedge (i,k)\neq\mathbb{P}\}} e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_k)}},$$

- Dwibedi et al. [2021, **NNCLR**] uses nearest neighbors from a queue $\mathbb{Q}$

$$\mathcal{L}_{\text{NNCLR}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \frac{e^{\text{CoSim}(\text{NN}(\boldsymbol{z}_i,\mathbb{Q}),\boldsymbol{z}_j)}}{\sum_{(k,l)\in\mathbb{P}}^{N} e^{\text{CoSim}(\text{NN}(\boldsymbol{z}_i,\mathbb{Q}),\boldsymbol{z}_l)}},$$

- Mitrovic et al. [2020, **RELIC**] adds a regularization term to enforce invariance

$$\mathcal{L}_{\text{RELIC}}(\boldsymbol{Z}) = -\sum_{(i,j)\in\mathbb{P}} \frac{e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_j)}}{\sum_{k=1}^{N} \mathbf{1}_{\{k\neq i\}} e^{\text{CoSim}(\boldsymbol{z}_i,\boldsymbol{z}_k)}} + KL(p(\boldsymbol{z}_i), p(\boldsymbol{z}_j)),$$

- Li et al. [2020, **PCL**] uses prototypes

Figure 3: Extensions of the infoNCE loss.

1. Different losses
2. InfoNCE loss
3. Deep Metric Learning V.S. Contrastive SSL

DML V.S. Contrastive SSL

DML

- positive/negative pairs come from labels or fixed transforms e.g. two halves of an image
- Hard-Negative Sampling for each mini-batch
- encoder DN
- small dataset (N<200k)
- zero-shot k-NN validation

Contrastive SSL

- positive pairs come from designed DAs that are continuously sampled, negative pairs are all nonpositive pairs regardless of class membership
- random sampling
- encoder DN + projector MLP
- large dataset
- -zero-shot k-NN validation
  -zero/few-shot/fine-tuning linear probing