Paper Review: Representation Learning with Contrastive Predictive Coding

Beomsu Kim*

July 19, 2021

Paper Information.

- Aaron van den Oord, Yazeh Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748, 2018.

1 InfoNCE

In InfoNCE, there are two inputs.

- Data joint distribution $p_{XY}(x, y)$ with marginal distributions $p_X(x)$ and $p_Y(y)$,
- Parametrized model $f(x, y; \theta)$.

Our goal is to estimate the mutual information

$$I(X;Y) = D_{\mathrm{KL}}(p_{XY} \| p_X \otimes p_Y) = \mathbb{E}_{p_{XY}}\left[\log \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}\right].$$

To achieve this, we first need to estimate the density ratio. That is, we need to have

$$f(x, y; \theta) \propto \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}.$$

Once we have this ratio, we can approximate the mutual information (Section 2.2).

*Department of Mathematical Sciences, KAIST. Email beomsu.kim@kaist.ac.kr

1.1 Estimating the Density Ratio

Define the categorical random variable

$$I \in \{1, \dots, N\}, \qquad p_I(i) = \frac{1}{N}$$

and the joint distribution

$$p_{IXY^N}(i, x, y_{1:N}) = p_I(i) p_{XY}(x, y_i) \prod_{j \neq i} p_Y(y_j)$$

where we denote

$$y_{1:N} = (y_1, \dots, y_N).$$

Since

$$p_{XY^N|I}(x, y_{1:N} \mid i) = \frac{p_{IXY^N}(i, x, y_{1:N})}{p_I(i)} = p_{XY}(x, y_i) \prod_{j \neq i} p_Y(y_j),$$

sampling from p_{IXY^N} means

1. we draw $i \in \{1, \dots, N\}$ uniformly at random,
2. draw $(x, y_i) \sim p_{XY}(x, y)$,
3. draw $y_j \sim p_Y(y)$ for $j \neq i$.

This also means $(x, y_i) \sim p_{XY}$ and $(x, y_j) \sim p_X \otimes p_Y$ for $j \neq i$. InfoNCE solves

$$\min_\theta \ \mathcal{L}_{\texttt{InfoNCE}}(\theta) = \mathbb{E}_{p_{IXY^N}} \left[-\log \frac{f(x, y_i; \theta)}{\sum_{k=1}^N f(x, y_k; \theta)} \right].$$

To understand this objective, let us observe that

$$\begin{aligned}
p_{I|XY^N}(i \mid x, y_{1:N}) &= \frac{p_{IXY_{1:N}}(i, x, y_{1:N})}{\sum_{k=1}^N p_{IXY_{1:N}}(k, x, y_{1:N})} \\
&= \frac{p_{XY}(x, y_i) \prod_{j \neq i} p_Y(y_j)}{\sum_{k=1}^N p_{XY}(x, y_k) \prod_{j \neq k} p_Y(y_j)} \\
&= \frac{p_{XY}(x, y_i)/p_Y(y_i)}{\sum_{k=1}^N p_{XY}(x, y_k)/p_Y(y_k)} \\
&= \frac{p_{XY}(x, y_i)/p_X(x)p_Y(y_i)}{\sum_{k=1}^N p_{XY}(x, y_k)/p_X(x)p_Y(y_k)}.
\end{aligned}$$

Also, define the classifier

$$q_{I|XY^N}(i \mid x, y_{1:N}; \theta) = \frac{f(x, y_i; \theta)}{\sum_{k=1}^N f(x, y_k; \theta)}.$$

It follows that

$$
\begin{aligned}
\mathcal{L}_{\texttt{InfoNCE}}(\theta) &= \mathbb{E}_{p_{IXY^N}}\left[-\log \frac{f(x, y_i; \theta)}{\sum_{k=1}^{N} f(x, y_k; \theta)}\right] \\
&= \mathbb{E}_{p_{IXY^N}}\left[-\log q_{I|XY^N}(i \mid x, y_{1:N}; \theta)\right] \\
&= \mathbb{E}_{p_{XY^N}}\left[\mathbb{E}_{p_{I|XY^N}}\left[-\log q_{I|XY^N}(i \mid x, y_{1:N}; \theta)\right]\right] \\
&= \mathbb{E}_{p_{XY^N}}\left[H(p_{I|XY^N}, q_{I|XY_{1:N}})\right].
\end{aligned}
$$

So, if we successfully find a solution θ^* to InfoNCE, we will have

$$
\frac{f(x, y_i; \theta^*)}{\sum_{k=1}^{N} f(x, y_k; \theta^*)} = q_{I|XY^N}(i \mid x, y_{1:N}; \theta^*) = p_{I|XY^N}(i \mid x, y_{1:N}) = \frac{p_{XY}(x, y_i)/p_X(x)p_Y(y_i)}{\sum_{k=1}^{N} p_{XY}(x, y_k)/p_X(x)p_Y(y_k)}
$$

for all $(i, x, y_{1:N})$. This means

$$
f(x, y; \theta^*) \propto \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}
$$

for all (x, y). Thus,

$$
\mathcal{L}_{\texttt{InfoNCE}}(\theta^*) = \mathbb{E}_{p_{IXY_{1:N}}}\left[-\log \frac{p_{XY}(x, y_i)/p_X(x)p_Y(y_i)}{\sum_{k=1}^{N} p_{XY}(x, y_k)/p_X(x)p_Y(y_k)}\right].
$$

From here, we can lower bound the mutual information between X and Y.

1.2 Estimating the Mutual Information

Since

$$p_{XY^N|I}(x, y_{1:N} \mid i) = \frac{p_{IXY^N}(i, x, y_{1:N})}{p_I(i)} = \frac{p_I(i) p_{XY}(x, y_i) \prod_{j \neq i} p_Y(y_j)}{p_I(i)} = p_{XY}(x, y_i) \prod_{j \neq i} p_Y(y_j),$$

conditioned on $i \sim p_I(i)$, we have (using the notation $y_{k \neq i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N)$)

$$(x, y_i) \sim p_{XY}, \qquad y_{k \neq i} \sim \prod_{k \neq i} p_Y(y_k).$$

It follows that

$$\begin{aligned}
\mathcal{L}_{\texttt{InfoNCE}}(\theta^*) &= \mathbb{E}_{p_{IXY^N}} \left[-\log \frac{p_{XY}(x, y_i)/p_X(x)p_Y(y_i)}{\sum_{k=1}^N p_{XY}(x, y_k)/p_X(x)p_Y(y_k)} \right] \\
&= \mathbb{E}_{p_{IXY^N}} \left[\log \left(1 + \frac{p_X(x)p_Y(y_i)}{p_{XY}(x, y_i)} \sum_{k \neq i} \frac{p_{X|Y}(x \mid y_k)}{p_X(x)} \right) \right] \\
&= \mathbb{E}_{p_I} \left[\mathbb{E}_{p_{XY^N|I}} \left[\log \left(1 + \frac{p_X(x)p_Y(y_i)}{p_{XY}(x, y_i)} \sum_{k \neq i} \frac{p_{X|Y}(x \mid y_k)}{p_X(x)} \right) \right] \right] \\
&= \mathbb{E}_{i \sim p_I} \left[\mathbb{E}_{(x, y_i) \sim p_{XY}} \left[\mathbb{E}_{y_{k \neq i} \sim \prod_{k \neq i} p_Y} \left[\log \left(1 + \frac{p_X(x)p_Y(y_i)}{p_{XY}(x, y_i)} \sum_{k \neq i} \frac{p_{X|Y}(x \mid y_k)}{p_X(x)} \right) \right] \right] \right] \\
&= \mathbb{E}_{(x, y_1) \sim p_{XY}} \left[\mathbb{E}_{y_{k \neq 1} \sim \prod_{k \neq 1} p_Y} \left[\log \left(1 + \frac{p_X(x)p_Y(y_1)}{p_{XY}(x, y_1)} \sum_{k \neq 1} \frac{p_{X|Y}(x \mid y_k)}{p_X(x)} \right) \right] \right] \\
&\geq \mathbb{E}_{(x, y_1) \sim p_{XY}} \left[\log \left(1 + \frac{p_X(x)p_Y(y_1)}{p_{XY}(x, y_1)} \sum_{k \neq 1} \mathbb{E}_{y_k \sim p_Y} \left[\frac{p_{X|Y}(x \mid y_k)}{p_X(x)} \right] \right) \right] \\
&= \mathbb{E}_{(x, y_1) \sim p_{XY}} \left[\log \left(1 + \frac{p_X(x)p_Y(y_1)}{p_{XY}(x, y_1)} (N-1) \right) \right] \\
&\geq \mathbb{E}_{(x, y_1) \sim p_{XY}} \left[\log \frac{p_X(x)p_Y(y_1)}{p_{XY}(x, y_1)} (N-1) \right] \\
&= \log(N-1) - \mathbb{E}_{(x, y_1) \sim p_{XY}} \left[\log \frac{p_{XY}(x, y_1)}{p_X(x)p_Y(y_1)} \right] \\
&= \log(N-1) - I(X, Y)
\end{aligned}$$

where we have used Jensen's inequality at the first inequality. This proves

$$I(X, Y) \geq \log(N-1) - \mathcal{L}_{\texttt{InfoNCE}}(\theta^*)$$

and this bound becomes tight as $N \to \infty$. However, $\mathcal{L}_{\texttt{InfoNCE}}(\theta) \geq 0$ for all θ. So, the estimate of mutual information by InfoNCE cannot exceed $\log(N-1)$.

Reference material.

- *Noise Contrastive Estimation* by Karl Stratos.