Paper Review: Breaking the log-K Curse on Contrastive Learners With FlatNCE

Beomsu Kim*

July 20, 2021

Paper Information.

- Junya Chen et. al. Breaking the log-K Curse on Contrastive Learners With FlatNCE. arXiv preprint arXiv:2107.01152, 2021.

1 Introduction

- There are many unresolved issues with contrastive learning.
  - Contrastive learners need a very large number of negative samples to work well.
  - The bias, variance, and performance tradeoffs are in debate.
  - There is a lack of training diagnostic tools for contrastive learning.
- Our development starts with two simple intuitions.
  - The contrasts between positive and negative data should be as large as possible.
  - The objective should be properly normalized to yield minimal variance.

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

2 Contrastive Representation Learning with InfoNCE

- With $y_{1:K} = (y_1, \ldots, y_K)$, define

$$p_{XY^K}(x, y_{1:K}) = p_{XY}(x, y_1) \prod_{k \neq 1} p_Y(y_k).$$

  This means $(x, y_1) \sim p_{XY}$ and $(x, y_k) \sim p_X \otimes p_Y$ for $k \neq 1$.
- Let $g(x, y)$ be a parametrized function, such as a neural network.
- Given $(x, y_{1:K}) \sim p_{XY^N}$, define $g_k = g(x, y_k)$ for $k = 1, \ldots, K$. Also, define

$$I^K_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = \log \frac{\exp(g_1)}{\frac{1}{K} \sum_{k=1}^K \exp(g_k)},$$

  and

$$I^K_{\texttt{InfoNCE}}(X; Y \mid g) = \mathbb{E}_{p_{XY^K}} \left[I^K_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) \right]$$

  and

$$I^K_{\texttt{InfoNCE}}(X; Y) = \max_g \; I^K_{\texttt{InfoNCE}}(X; Y \mid g).$$

  Note that g_1 is the logit for the "positive pair" and g_k for $k \neq 1$ are the logits for the "negative pairs".
- We also define

$$I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = -\log \frac{\exp(g_1)}{\sum_{k=1}^K \exp(g_k)}$$

  such that

$$I^K_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = -I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) + \log K.$$

Proposition 2.1. *$I^K_{\texttt{InfoNCE}}(X; Y)$ is an asymptotically tight lower bound to mutual information, i.e.,*

$$I(X; Y) \geq I^K_{\texttt{InfoNCE}}(X; Y \mid g), \qquad \lim_{K \to \infty} I^K_{\texttt{InfoNCE}}(X; Y) \to I(X; Y).$$

Proof. See my paper review on InfoNCE. □

3 FlatNCE and Generalized Contrastive Representation Learning

- Define

$$I_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = \frac{\sum_{k \neq 1} \exp(g_k - g_1)}{\texttt{stop\_grad}[\sum_{k \neq 1} \exp(g_k - g_1)]}$$

  and

$$I^K_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1).$$

- We observe that

$$\sum_{k \neq 1} \exp(g_k - g_1) = \left(\frac{\exp(g_1)}{\sum_{k \neq 1} \exp(g_k)} \right)^{-1}$$

  and so (the gradient is w.r.t. the parameters of g)

$$\begin{aligned}
\nabla I_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) &= \frac{\nabla \sum_{k \neq 1} \exp(g_k - g_1)}{\texttt{stop_grad}[\sum_{k \neq 1} \exp(g_k - g_1)]} \\
&= \nabla \log \sum_{k \neq 1} \exp(g_k - g_1) \\
&= \nabla \log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1) \\
&= -\nabla I^K_{\texttt{FlatNCE}}(x, y_{1:N} \mid g).
\end{aligned}$$

  This shows that gradient descent on $I_{\texttt{FlatNCE}}$ is equivalent to gradient descent on $I^{\oplus,K}_{\texttt{FlatNCE}}$.

- We also observe that

$$\begin{aligned}
I^K_{\texttt{FlatNCE}}(x, y_{1:N} \mid g) &= -\log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1) \\
&= -\log \frac{\frac{1}{K} \sum_{k \neq 1} \exp(g_k)}{\exp(g_1)} \\
&= \log \frac{\exp(g_1)}{\frac{1}{K} \sum_{k \neq 1} \exp(g_k)}
\end{aligned}$$

  which is just $I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g)$ with the positive pair logit g_1 removed from the denominator sum.

- Define

$$I^{\oplus}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = \frac{\sum_{k=1}^{K} \exp(g_k - g_1)}{\texttt{stop\_grad}[\sum_{k=1}^{K} \exp(g_k - g_1)]}$$

and

$$I^{\oplus,K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\log \frac{1}{K} \sum_{k=1}^{K} \exp(g_k - g_1).$$

- Similar to $I_{\texttt{FlatNCE}}$, we have

$$\nabla I^{\oplus}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\nabla I^{\oplus,K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g)$$

and also

$$I^{\oplus,K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = \log \frac{\exp(g_1)}{\frac{1}{K}\sum_{k=1}^{K} \exp(g_k)} = I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g).$$

which shows that gradient ascent on $I^{\oplus}_{\texttt{FlatNCE}}$ is equivalent to gradient descent on $I^{K}_{\texttt{InfoNCE}}$.

Proposition 3.1. $\nabla I^{\oplus}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = \nabla I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g)$.

Proof. Since (see Section 2)

$$I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = -I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) + \log K,$$

we have (see the above bullet)

$$\nabla I^{\oplus}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\nabla I^{\oplus,K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\nabla I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = \nabla I_{\texttt{InfoNCE}}(x, y_{1:K} \mid g).$$

This concludes the proof. □

Proposition 3.2. *The gradient of $I_{\texttt{FlatNCE}}(x, y_{1:K} \mid g)$ is an importance-weighted estimator of the form*

$$\nabla I_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = \sum_{k \neq 1} w_k \nabla g_k - \nabla g_1, \qquad w_k = \frac{\exp(g_k)}{\sum_{k' \neq 1} \exp(g_{k'})}.$$

Proof. Observe that

$$\begin{aligned}
\nabla I_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) &= \frac{\nabla \sum_{k \neq 1} \exp(g_k - g_1)}{\sum_{k \neq 1} \exp(g_k - g_1)} \\
&= \sum_{k \neq 1} \frac{\exp(g_k - g_1)}{\sum_{k' \neq 1} \exp(g_{k'} - g_1)} (\nabla g_k - \nabla g_1) \\
&= \sum_{k \neq 1} \frac{\exp(g_k)}{\sum_{k' \neq 1} \exp(g_{k'})} (\nabla g_k - \nabla g_1) \\
&= \sum_{k \neq 1} w_k (\nabla g_k - \nabla g_1) \\
&= \sum_{k \neq 1} \nabla g_k - \nabla g_1
\end{aligned}$$

since $\sum_{k \neq 1} w_k = 1$. □

- In FlatNCE, larger weights will be assigned to the more challenging negative examples in the batch.
- The authors claim FlatNCE is also a formal MI lower bound using the below Lemma.

Lemma 3.3. *For arbitrary $u \in \mathbb{R}$, we have*

$$I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) \geq 1 - u - \frac{1}{K}\sum_{k=1}^{K} \exp(-u + g_k - g_1)$$

and the inequality holds when

$$u = \texttt{stop\_grad}\left[\log \frac{1}{K}\sum_{k=1}^{K} \exp(g_k - g_1)\right] = \texttt{stop\_grad}\left[-I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g)\right].$$

Proof. The Fenchel-Legendre dual for $f(t) = -\log t$ is $f^*(v) = -1 - \log(-v)$. That is,

$$f(t) = \sup_{v \in \mathbb{R}} \{vt - f^*(v)\}$$

and so

$$-\log t \geq vt + 1 + \log(-v)$$

for any $v \in \mathbb{R}$. Setting $v = -e^{-u}$, we get

$$-\log t \geq 1 - u - e^{-u} t$$

for any $u \in \mathbb{R}$. Since

$$I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = I^{\oplus,K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g) = -\log \frac{1}{K}\sum_{k=1}^{K} \exp(g_k - g_1),$$

setting

$$t = \frac{1}{K}\sum_{k=1}^{K} \exp(g_k - g_1)$$

proves the first part of the proposition. The second part can be checked by simple calculation. □

Corollary 3.4. $\mathbb{E}_{p_{XY^N}}\left[I^{K}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g)\right] \leq I(X;Y)$.

Proof. Plugging in the optimal value of u in the above Lemma, we have

$$I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g) = 1 + \texttt{stop\_grad}\left[I^{K}_{\texttt{InfoNCE}}(x, y_{1:K} \mid g)\right] - I^{\oplus}_{\texttt{FlatNCE}}(x, y_{1:K} \mid g).$$

Since the first two terms at the RHS are constant w.r.t. the parameters of g, the authors claim that the claimed inequality holds up to a constant. However, the above inequality is about $I^{\oplus}_{\texttt{FlatNCE}}$, not $I^{K}_{\texttt{FlatNCE}}$, so I don't think the proof is correct. □