

## A Appendix

This appendix consists of four sections (A.1–A.4). Section A.1 provides the theoretical properties of InfoLOOB and InfoNCE. It is shown how to derive that InfoNCE is a lower bound on mutual information. Further it is shown how to derive that InfoLOOB is an upper bound on mutual information. The proposed loss function and its gradients are discussed. Section A.2 provides details on the experiments. Section A.3 briefly reviews continuous modern Hopfield networks. Section A.4 discusses further related work.

### Contents of the appendix

A.1	InfoLOOB vs. InfoNCE	21
A.1.1	InfoNCE: Lower Bound on Mutual Information	21
A.1.2	InfoLOOB: Upper Bound on Mutual Information	24
A.1.3	InfoLOOB: Analysis of the Objective	29
A.1.4	InfoNCE and InfoLOOB: Gradients	37
A.1.5	InfoLOOB and InfoNCE: Probability Estimators	39
A.1.6	InfoLOOB and InfoNCE: Losses	40
A.2	Experiments	44
A.2.1	Ablation studies	44
A.2.2	Hyperparameters	45
A.2.3	Datasets	45
A.2.4	Zero-shot evaluation	46
A.2.5	Linear probing	47
A.2.6	Image-Text retrieval	49
A.2.7	Analysis of the image and text embeddings	49
A.2.8	Training time and memory consumption	50
A.3	Review of Modern Hopfield Networks	51
A.4	Further Related Work	53

### List of theorems

A1	Theorem (InfoNCE lower bound)	23
A2	Theorem (InfoLOOB upper bound)	27
A3	Theorem (Weighted Covariances)	43
A4	Theorem (Modern Hopfield Networks: Retrieval with One Update)	52
A5	Theorem (Modern Hopfield Networks: Exponential Storage Capacity)	53

### List of definitions

A1	Definition (Pattern Stored and Retrieved)	52
----	---	----

### List of figures

A1	Eigenvalues of the covariance matrix of image embeddings	50
A4	Visualization of zero-shot classification of three examples from each dataset	55

### List of tables

A1	Influence of loss functions and Hopfield retrieval (CC)	44
A2	Influence of loss functions and Hopfield retrieval (YFCC)	44
A3	Influence of learning rate scheduler	45
A4	Datasets used for downstream evaluation	46
A7	Linear probing for CLIP (reimplementation) and CLOOB trained on YFCC	48

## A.1 InfoLOOB vs. InfoNCE

### A.1.1 InfoNCE: Lower Bound on Mutual Information

We derive a lower bound on the mutual information between random variables  $X$  and  $Y$  distributed according to  $p(\mathbf{x}, \mathbf{y})$ . The mutual information  $I(X ; Y)$  between random variables  $X$  and  $Y$  is

$$I(X ; Y) = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} \right] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x} | \mathbf{y})}{p(\mathbf{x})} \right] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right]. \quad (\text{A1})$$

“InfoNCE” has been introduced in [van den Oord et al. \(2018\)](#) and is a *multi-sample bound*. In the setting introduced in [van den Oord et al. \(2018\)](#), we have an anchor sample  $\mathbf{y}$  given. For the anchor sample  $\mathbf{y}$  we draw a positive sample  $\mathbf{x}_1$  according to  $p(\mathbf{x}_1 | \mathbf{y})$ . Next, we draw a set  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  according to  $p(\tilde{X})$ , which are  $n - 1$  negative samples drawn iid according to  $p(\mathbf{x})$ . We have drawn a set  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  according to  $p(X | \mathbf{y})$ , which is one positive sample  $\mathbf{x}_1$  drawn by  $p(\mathbf{x}_1 | \mathbf{y})$  and  $N - 1$  negative samples  $\{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  drawn iid according to  $p(\mathbf{x})$ .

The InfoNCE with probabilities is

$$I_{\text{InfoNCE}}(X_1 ; Y) = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N} \sum_{i=1}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] \right], \quad (\text{A2})$$

where we inserted the factor  $\frac{1}{N}$  in contrast to the original version in [van den Oord et al. \(2018\)](#), where we followed [Poole et al. \(2019\)](#); [Tschannen et al. \(2019\)](#); [Cheng et al. \(2020\)](#); [Chen et al. \(2021\)](#).

The InfoNCE with score function  $f(\mathbf{x}, \mathbf{y})$  is

$$I_{\text{InfoNCE}}(X_1 ; Y) = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right]. \quad (\text{A3})$$

The InfoNCE with probabilities can be rewritten as:

$$\begin{aligned} I_{\text{InfoNCE}}(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N} \sum_{i=1}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] \right] \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{y}|\mathbf{x}_1)}{p(\mathbf{y})}}{\frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{y}|\mathbf{x}_i)}{p(\mathbf{y})}} \right) \right] \right] \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)}}{\frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{x}_i|\mathbf{y})}{p(\mathbf{x}_i)}} \right) \right] \right]. \end{aligned} \quad (\text{A4})$$

This is the InfoNCE with  $f(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x})$ .

**Set of pairs.** The InfoNCE can be written in a different setting [Poole et al. \(2019\)](#), which is used in most implementations. We sample  $N$  pairs independently from  $p(\mathbf{x}, \mathbf{y})$ , which gives  $Z = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . The InfoNCE is then

$$I_{\text{InfoNCE}}(X ; Y) = \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right) \right]. \quad (\text{A5})$$

Following [van den Oord et al. \(2018\)](#) we have

$$\begin{aligned}
I_{\text{InfoNCE}}(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{y}|\mathbf{x}_1)}{p(\mathbf{y})}}{\frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{y}|\mathbf{x}_i)}{p(\mathbf{y})}} \right) \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)}}{\frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{x}_i|\mathbf{y})}{p(\mathbf{x}_i)}} \right) \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y}) \prod_{l=2}^N p(\mathbf{x}_l)}{\sum_{i=1}^N p(\mathbf{x}_i | \mathbf{y}) \prod_{l \neq i} p(\mathbf{x}_l)} \right) \right] \right] + \ln(N) \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} [\ln p(i = 1 | X, \mathbf{y})] \right] + \ln(N),
\end{aligned} \tag{A6}$$

where  $p(i = 1 | X, \mathbf{y})$  is the probability that sample  $\mathbf{x}_1$  is the positive sample if we know there exists exactly one positive sample in  $X$ .

The InfoNCE is a lower bound on the mutual information. The following inequality is from [van den Oord et al. \(2018\)](#):

$$\begin{aligned}
I(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right) \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1|\mathbf{y})} \left[ - \ln \left( \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1 | \mathbf{y})} \right) \right] \right] \\
&\geq \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1|\mathbf{y})} \left[ - \ln \left( \frac{1}{N} + \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1 | \mathbf{y})} \right) \right] \right] \\
&\approx \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ - \ln \left( \frac{1}{N} + \frac{1}{N} \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1 | \mathbf{y})} \sum_{i=2}^N \frac{p(\mathbf{x}_i | \mathbf{y})}{p(\mathbf{x}_i)} \right) \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)}}{\frac{1}{N} \frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)} + \frac{1}{N} \sum_{i=2}^N \frac{p(\mathbf{x}_i|\mathbf{y})}{p(\mathbf{x}_i)}} \right) \right] \right] \\
&= I_{\text{InfoNCE}}(X_1 ; Y),
\end{aligned} \tag{A7}$$

where the " $\geq$ " is obtained by bounding  $\ln(1/N + a)$  by  $\ln(a)$ , which gives a bound that is not very tight, since  $a = \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1|\mathbf{y})}$  can become small. However for the " $\approx$ " [van den Oord et al. \(2018\)](#) have to assume

$$\frac{1}{N} \sum_{i=2}^N \frac{p(\mathbf{x}_i | \mathbf{y})}{p(\mathbf{x}_i)} = \frac{1}{N} \sum_{i=2}^N \frac{p(\mathbf{y} | \mathbf{x}_i)}{p(\mathbf{y})} \geq 1, \tag{A8}$$

which is unclear how to ensure.

For a proof of this bound see [Poole et al. \(2019\)](#).

We assumed that for the anchor sample  $\mathbf{y}$  a positive sample  $\mathbf{x}_1$  has been drawn according to  $p(\mathbf{x}_1 | \mathbf{y})$ . A set  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  of negative samples is drawn according to  $p(\mathbf{x})$ . Therefore, we have a set  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  that is drawn with one positive sample  $\mathbf{x}_1$  and  $N - 1$  negative samples  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$ . We have

$$p(\tilde{X}) = \prod_{i=2}^N p(\mathbf{x}_i), \tag{A9}$$

$$p(X | \mathbf{y}) = p(\mathbf{x}_1 | \mathbf{y}) \prod_{i=2}^N p(\mathbf{x}_i), \tag{A10}$$

$$p(X) = \prod_{i=1}^N p(\mathbf{x}_i). \tag{A11}$$

Next, we present a theorem that shows this bound, where we largely follow [Poole et al. \(2019\)](#) in the proof. In contrast to [Poole et al. \(2019\)](#), we do not use the NWJ bound [Nguyen et al. \(2010\)](#). The mutual information is

$$I(X_1; Y) = \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right) \right]. \quad (\text{A12})$$

**Theorem A1** (InfoNCE lower bound). *InfoNCE with score function  $f(\mathbf{x}, \mathbf{y})$  according to Eq. (A3) is a lower bound on the mutual information.*

$$I(X_1; Y) \geq \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] = I_{\text{InfoNCE}}(X_1; Y). \quad (\text{A13})$$

*InfoNCE with probabilities according to Eq. (A2) is a lower bound on the mutual information.*

$$I(X_1; Y) \geq \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N} \sum_{i=1}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] = I_{\text{InfoNCE}}(X_1; Y). \quad (\text{A14})$$

*The second bound Eq. (A14) is a special case of the first bound Eq. (A13).*

*Proof. Part (I):* Lower bound with score function  $f(\mathbf{x}, \mathbf{y})$ .

For each set  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$ , we define as data-dependent (depending on  $\tilde{X}$ ) score function  $g(\mathbf{x}_1, \mathbf{y}, \tilde{X})$  that is based on the score function  $f(\mathbf{x}, \mathbf{y})$ . Therefore we have for each  $\tilde{X}$  a different data-dependent score function  $g$  based on  $f$ . We will derive a bound on the InfoNCE, which is the expectation of a lower bound on the mutual information over the score functions. For score function  $g(\mathbf{x}_1, \mathbf{y}, \tilde{X})$ , we define a variational distribution  $q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})$  over  $\mathbf{x}_1$ :

$$q(\mathbf{x}_1 | \mathbf{y}, \tilde{X}) = \frac{p(\mathbf{x}_1) g(\mathbf{x}_1, \mathbf{y}, \tilde{X})}{Z(\mathbf{y}, \tilde{X})}, \quad (\text{A15})$$

$$Z(\mathbf{y}, \tilde{X}) = \mathbb{E}_{p(\mathbf{x}_1)} [g(\mathbf{x}_1, \mathbf{y}, \tilde{X})], \quad (\text{A16})$$

which ensures

$$\int q(\mathbf{x}_1 | \mathbf{y}, \tilde{X}) d\mathbf{x}_1 = 1. \quad (\text{A17})$$

We have

$$\frac{q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})}{p(\mathbf{x}_1)} = \frac{g(\mathbf{x}_1, \mathbf{y}, \tilde{X})}{Z(\mathbf{y}, \tilde{X})}. \quad (\text{A18})$$

For the function  $g$ , we set

$$g(\mathbf{x}_1, \mathbf{y}, \tilde{X}) = \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})}, \quad (\text{A19})$$

For the function  $f$  we use

$$f(\mathbf{x}_1, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y})), \quad (\text{A20})$$

where  $\text{sim}(\mathbf{x}, \mathbf{y})$  is typically the cosine similarity.

We next show that InfoNCE is a lower bound on the mutual information.

$$\begin{aligned}
I(X_1 ; Y) &= \mathbb{E}_{p(\tilde{X})} [I(X_1 ; Y)] = \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] \right] & (A21) \\
&= \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})} \frac{q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})}{p(\mathbf{x}_1)} \right) \right] \right] \\
&= \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})}{p(\mathbf{x}_1)} \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}, \tilde{X}))] \right] \\
&\geq \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y}, \tilde{X})}{p(\mathbf{x}_1)} \right] \right] = \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{g(\mathbf{x}_1, \mathbf{y}, \tilde{X})}{Z(\mathbf{y}, \tilde{X})} \right] \right] \\
&= \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln g(\mathbf{x}_1, \mathbf{y}, \tilde{X}) - \ln \left( \mathbb{E}_{p(\mathbf{x}_1)} [g(\mathbf{x}_1, \mathbf{y}, \tilde{X})] \right) \right] \right] \\
&= \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1 | \mathbf{y})} \left[ \ln g(\mathbf{x}_1, \mathbf{y}, \tilde{X}) \right] - \ln \left( \mathbb{E}_{p(\mathbf{x}_1)} [g(\mathbf{x}_1, \mathbf{y}, \tilde{X})] \right) \right] \right] \\
&= \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1 | \mathbf{y})} \left[ \ln g(\mathbf{x}_1, \mathbf{y}, \tilde{X}) \right] \right] - \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(\mathbf{x}_1)} [g(\mathbf{x}_1, \mathbf{y}, \tilde{X})] \right) \right] \right] \right] \\
&\geq \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln g(\mathbf{x}_1, \mathbf{y}, \tilde{X}) \right] - \mathbb{E}_{p(\tilde{X})} \left[ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1)} [g(\mathbf{x}_1, \mathbf{y}, \tilde{X})] - 1 \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X)} \left[ \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - 1 \right] \\
&= \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{p(X)} \left[ \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - 1 \right] \\
&= \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X)} \left[ \frac{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] - 1 \right] \\
&= \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right] \\
&= I_{\text{InfoNCE}}(X_1 ; Y) .
\end{aligned}$$

For the first " $\geq$ " we used that the Kullback-Leibler divergence is non-negative. For the second " $\geq$ " we used the inequality  $\ln a \leq a - 1$  for  $a > 0$ .

**Part (II):** Lower bound with probabilities.

If the score function  $f$  is

$$f(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x}) , \quad (A22)$$

then the bound is

$$\begin{aligned}
I(X_1 ; Y) &\geq \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] = \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N} \sum_{i=1}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] & (A23) \\
&= \mathbb{E}_{p(\mathbf{y})p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{y} | \mathbf{x}_1)}{p(\mathbf{y})}}{\frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{y} | \mathbf{x}_i)}{p(\mathbf{y})}} \right) \right] = I_{\text{InfoNCE}}(X_1 ; Y) .
\end{aligned}$$

This is the bound with probabilities in the theorem.  $\square$

### A.1.2 InfoLOOB: Upper Bound on Mutual Information

We derive an upper bound on the mutual information between random variables  $X$  and  $Y$  distributed according to  $p(\mathbf{x}, \mathbf{y})$ . The mutual information  $I(X ; Y)$  between random variables  $X$  and  $Y$  is

$$I(X ; Y) = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} \right] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x} | \mathbf{y})}{p(\mathbf{x})} \right] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ \ln \frac{p(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right] . \quad (A24)$$

In [Poole et al. \(2019\)](#) Eq. (13) introduces a variational upper bound on the mutual information, which has been called "Leave one out upper bound" (called "L1Out" in [Cheng et al. \(2020\)](#)). For simplicity, we call this bound "InfoLOOB", where LOOB is an acronym for "Leave One Out Bound". In contrast to InfoNCE, InfoLOOB is an upper bound on the mutual information. InfoLOOB is analog to InfoNCE except that the negative samples do not contain a positive sample. Fig. 1 and Fig. 2 in [Cheng et al. \(2020\)](#) both show that InfoLOOB is a better estimator for the mutual information than InfoNCE ([van den Oord et al., 2018](#)), MINE ([Belghazi et al., 2018](#)), and NWJ ([Nguyen et al., 2010](#)).

The InfoLOOB with score function  $f(\mathbf{x}, \mathbf{y})$  is defined as

$$I_{\text{InfoLOOB}}(X_1 ; Y) = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right]. \quad (\text{A25})$$

The InfoLOOB with probabilities is defined as

$$I_{\text{InfoLOOB}}(X_1 ; Y) = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N-1} \sum_{i=2}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] \right]. \quad (\text{A26})$$

This is the InfoLOOB Eq. (A25) with  $f(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x})$ .

The InfoLOOB with probabilities can be written in different forms:

$$\begin{aligned} I_{\text{InfoLOOB}}(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N-1} \sum_{i=2}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] \right] \quad (\text{A27}) \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{y}|\mathbf{x}_1)}{p(\mathbf{y})}}{\frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{y}|\mathbf{x}_i)}{p(\mathbf{y})}} \right) \right] \right] = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)}}{\frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{x}_i|\mathbf{y})}{p(\mathbf{x}_i)}} \right) \right] \right]. \end{aligned}$$

**Set of pairs.** The InfoLOOB can we written in a different setting ([Poole et al., 2019](#)), which will be used in our implementations. We sample  $N$  pairs independently from  $p(\mathbf{x}, \mathbf{y})$ , which gives  $X = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . The InfoLOOB is then

$$I_{\text{InfoLOOB}}(X ; Y) = \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right) \right]. \quad (\text{A28})$$

We assume that an anchor sample  $\mathbf{y}$  is given. For the anchor sample  $\mathbf{y}$  we draw a positive sample  $\mathbf{x}_1$  according to  $p(\mathbf{x}_1 | \mathbf{y})$ . Next, we draw a set  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  of negative samples according to  $\tilde{p}(\mathbf{x} | \mathbf{y})$ . **For a given  $\mathbf{y}$ , the  $\mathbf{x}$  that have a large  $p(\mathbf{x} | \mathbf{y})$  are drawn with a lower probability  $\tilde{p}(\mathbf{x} | \mathbf{y})$  compared to random drawing via  $p(\mathbf{x})$ .** The negatives are indeed negatives. We have drawn first anchor sample  $\mathbf{y}$  and then  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , where  $\mathbf{x}_1$  is drawn according to  $p(\mathbf{x}_1 | \mathbf{y})$  and  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  are drawn iid according to  $\tilde{p}(\mathbf{x} | \mathbf{y})$ . We have

$$\tilde{p}(\tilde{X} | \mathbf{y}) = \prod_{i=2}^N \tilde{p}(\mathbf{x}_i | \mathbf{y}), \quad (\text{A29})$$

$$\tilde{p}(X | \mathbf{y}) = p(\mathbf{x}_1 | \mathbf{y}) \prod_{i=2}^N \tilde{p}(\mathbf{x}_i | \mathbf{y}), \quad (\text{A30})$$

$$\tilde{p}(\tilde{X} | \mathbf{y}) p(\mathbf{x}_1) = p(\mathbf{x}_1) \prod_{i=2}^N \tilde{p}(\mathbf{x}_i | \mathbf{y}). \quad (\text{A31})$$

We assume for score function  $f(\mathbf{x}, \mathbf{y})$

$$\forall_{\mathbf{y}} \forall_{\mathbf{x}} : 0 < f(\mathbf{x}, \mathbf{y}). \quad (\text{A32})$$

We ensure this by using for score function  $f$

$$f(\mathbf{x}, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}, \mathbf{y})), \quad (\text{A33})$$

where  $\text{sim}(\mathbf{x}, \mathbf{y})$  is typically the cosine similarity.

InfoLOOB with score function  $f(\mathbf{x}, \mathbf{y})$  and with undersampling via  $\tilde{p}(X | \mathbf{y})$  is (compare the definition of InfoLOOB Eq. (A25) without undersampling):

$$I_{\text{InfoLOOB}}(X ; Y) = E_{p(\mathbf{y})} \left[ E_{\tilde{p}(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right]. \quad (\text{A34})$$

The reference constant  $Z(\mathbf{y})$  gives the average score  $f(\mathbf{x}, \mathbf{y})$ , if the negatives for  $\mathbf{y}$  are selected with lower probability via  $\tilde{p}(\mathbf{x} | \mathbf{y})$  than with random drawing according to  $p(\mathbf{x})$ .

$$Z(\mathbf{y}) = E_{\tilde{p}(\mathbf{x}|\mathbf{y})} [f(\mathbf{x}, \mathbf{y})]. \quad (\text{A35})$$

We define the variational distribution

$$q(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}) f(\mathbf{x}, \mathbf{y})}{Z^*(\mathbf{y})}, \quad Z^*(\mathbf{y}) = E_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]. \quad (\text{A36})$$

With the variational distribution  $q(\mathbf{x} | \mathbf{y})$ , we express our main assumption. **The main assumption for the bound is:**

$$E_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x} | \mathbf{y}) \| q(\mathbf{x} | \mathbf{y}))] \leq E_{p(\mathbf{y})} [\ln Z^*(\mathbf{y}) - \ln Z(\mathbf{y})]. \quad (\text{A37})$$

This assumption can be written as

$$E_{p(\mathbf{y})} \left[ E_{p(\mathbf{x}|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}) Z(\mathbf{y})}{p(\mathbf{y}) f(\mathbf{x}, \mathbf{y})} \right) \right] \right] \leq 0. \quad (\text{A38})$$

This assumption ensures that the  $\mathbf{x}$  with large  $p(\mathbf{x} | \mathbf{y})$  are selected with lower probability via  $\tilde{p}(\mathbf{x} | \mathbf{y})$  than with random drawing according to  $p(\mathbf{x})$ . The negatives are ensured to be real negatives, that is,  $p(\mathbf{x} | \mathbf{y})$  is small and so is  $f(\mathbf{x}, \mathbf{y})$ . Consequently, we make sure that we draw  $\mathbf{x}$  with sufficient small  $f(\mathbf{x}, \mathbf{y})$ . The Kullback-Leibler gives the minimal required gap between drawing  $f(\mathbf{x}, \mathbf{y})$  via  $p(\mathbf{x})$  and drawing  $f(\mathbf{x}, \mathbf{y})$  via  $\tilde{p}(\mathbf{x} | \mathbf{y})$ .

**EXAMPLE.** With  $h(\mathbf{y}) > 0$ , we consider the setting

$$f(\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})}, \quad (\text{A39})$$

$$\tilde{p}(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}) p(\mathbf{y})}{h(\mathbf{y}) p(\mathbf{y} | \mathbf{x}) C(\mathbf{y})}, \quad C(\mathbf{y}) = E_{p(\mathbf{x})} \left[ \left( \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} \right)^{-1} \right]. \quad (\text{A40})$$

The main assumption becomes

$$E_{p(\mathbf{y})} \left[ E_{p(\mathbf{x}|\mathbf{y})} \left[ \ln \frac{Z(\mathbf{y})}{h(\mathbf{y})} \right] \right] \leq 0. \quad (\text{A41})$$

The main assumption holds since

$$\begin{aligned} Z(\mathbf{y}) &= E_{\tilde{p}(\mathbf{x}|\mathbf{y})} \left[ \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} \right] = \int \frac{p(\mathbf{x}) p(\mathbf{y})}{h(\mathbf{y}) p(\mathbf{y} | \mathbf{x}) C(\mathbf{y})} \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} d\mathbf{x} \quad (\text{A42}) \\ &= \int p(\mathbf{x}) C(\mathbf{y})^{-1} d\mathbf{x} = C(\mathbf{y})^{-1} = \left( E_{p(\mathbf{x})} \left[ \left( \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} \right)^{-1} \right] \right)^{-1} \\ &\leq \left( E_{p(\mathbf{x})} \left[ \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} \right] \right)^{-1} = E_{p(\mathbf{x})} \left[ \frac{p(\mathbf{y} | \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} \right] \\ &= \int \frac{p(\mathbf{y}, \mathbf{x}) h(\mathbf{y})}{p(\mathbf{y})} d\mathbf{x} = h(\mathbf{y}), \end{aligned}$$

where we used for the  $\leq$  Jensen's inequality with the function  $f(a) = 1/a$ , which is convex for  $a > 0$ .

For score function  $f(\mathbf{x}, \mathbf{y})$  and distribution  $\tilde{p}(\mathbf{x} | \mathbf{y})$  for sampling the negative samples, we have defined:

$$Z(\mathbf{y}) = \mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{y})} [f(\mathbf{x}, \mathbf{y})] , \quad (\text{A43})$$

$$Z^*(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})] , \quad (\text{A44})$$

$$q(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}) f(\mathbf{x}, \mathbf{y})}{Z^*(\mathbf{y})} . \quad (\text{A45})$$

Next theorem gives the upper bound of the InfoLOOB on the mutual information, which is

$$I(X_1 ; Y) = \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] . \quad (\text{A46})$$

**Theorem A2** (InfoLOOB upper bound). *If  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  are drawn iid according to  $\tilde{p}(\mathbf{x} | \mathbf{y})$  and if the main assumption holds:*

$$\mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x} | \mathbf{y}) \| q(\mathbf{x} | \mathbf{y}))] \leq \mathbb{E}_{p(\mathbf{y})} [\ln Z^*(\mathbf{y}) - \ln Z(\mathbf{y})] . \quad (\text{A47})$$

*Then InfoLOOB with score function  $f(\mathbf{x}, \mathbf{y})$  and undersampling positives by  $\tilde{p}(X | \mathbf{y})$  is an upper bound on the mutual information:*

$$I(X_1 ; Y) \leq \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] = I_{\text{InfoLOOB}}(X_1 ; Y) . \quad (\text{A48})$$

*If the negative samples  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  are drawn iid according to  $p(\mathbf{x})$ , then InfoLOOB with probabilities according to Eq. (A26) is an upper bound on the mutual information:*

$$I(X_1 ; Y) \leq \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N-1} \sum_{i=2}^N p(\mathbf{y} | \mathbf{X}_i)} \right) \right] \right] = I_{\text{InfoLOOB}}(X_1 ; Y) . \quad (\text{A49})$$

*The second bound Eq. (A49) is a special case of the first bound Eq. (A48).*



*Proof. Part (I):* Upper bound with score function  $f(\mathbf{x}, \mathbf{y})$ .

$$\begin{aligned}
I(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] & (A50) \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{q(\mathbf{x}_1 | \mathbf{y})} \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right) \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&\leq \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] + \mathbb{E}_{p(\mathbf{y})} [\ln \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] - \ln Z(\mathbf{y})] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} + \ln \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{Z(\mathbf{y})} \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]} \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{Z(\mathbf{y})} \right) \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{Z(\mathbf{y})} \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right]} \right) \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} [\ln f(\mathbf{x}_1, \mathbf{y})] - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \right) \right] \\
&\leq \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} [\ln f(\mathbf{x}_1, \mathbf{y})] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] \\
&= I_{\text{InfoLOOB}}(X_1 ; Y),
\end{aligned}$$

where the first " $\leq$ " uses assumption Eq. (A37), while Jensens's inequality was used for the second " $\leq$ " by exchanging the expectation and the "ln". We also used

$$\mathbb{E}_{\tilde{p}(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] = \frac{1}{N-1} \sum_{i=2}^N \mathbb{E}_{\tilde{p}(\mathbf{x}_i|\mathbf{y})} [f(\mathbf{x}_i, \mathbf{y})] = \frac{1}{N-1} \sum_{i=2}^N Z(\mathbf{y}) = Z(\mathbf{y}). \quad (A51)$$

**Part (II):** Upper bound with probabilities.

If the score function  $f$  is

$$f(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x}) \quad (A52)$$

and

$$\tilde{p}(\mathbf{x} | \mathbf{y}) = p(\mathbf{x}), \quad (A53)$$

then

$$\tilde{p}(X | \mathbf{y}) = p(X | \mathbf{y}), \quad (A54)$$

$$Z(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x})} [p(\mathbf{y} | \mathbf{x})] = p(\mathbf{y}), \quad (A55)$$

$$Z^*(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x})} [p(\mathbf{y} | \mathbf{x})] = p(\mathbf{y}), \quad (A56)$$

$$q(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}) p(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} = p(\mathbf{x} | \mathbf{y}), \quad (A57)$$

$$\text{KL}(p(\mathbf{x} | \mathbf{y}) \| q(\mathbf{x} | \mathbf{y})) = \text{KL}(p(\mathbf{x} | \mathbf{y}) \| p(\mathbf{x} | \mathbf{y})) = 0. \quad (A58)$$

Therefore, the main assumption holds, since

$$0 = \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x} | \mathbf{y}) \| q(\mathbf{x} | \mathbf{y}))] = \mathbb{E}_{p(\mathbf{y})} [\ln Z^*(\mathbf{y}) - \ln Z(\mathbf{y})] . \quad (\text{A59})$$

The bound becomes

$$\begin{aligned} I(X_1 ; Y) &\leq \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{y} | \mathbf{x}_1)}{\frac{1}{N-1} \sum_{i=2}^N p(\mathbf{y} | \mathbf{x}_i)} \right) \right] \right] \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{y}|\mathbf{x}_1)}{p(\mathbf{y})}}{\frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{y}|\mathbf{x}_i)}{p(\mathbf{y})}} \right) \right] \right] = I_{\text{InfoLOOB}}(X_1 ; Y) . \end{aligned} \quad (\text{A60})$$

An alternative proof is as follows:

$$\begin{aligned} I(X_1 ; Y) &= I(X_1 ; Y) - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{y})}{p(\mathbf{y})} \right) \right] \\ &= I(X_1 ; Y) - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{y} | \mathbf{x}_i)}{p(\mathbf{y})} \right] \right) \right] \\ &\leq I(X_1 ; Y) - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{y} | \mathbf{x}_i)}{p(\mathbf{y})} \right) \right] \right] \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1|\mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right) \right] \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{x}_i | \mathbf{y})}{p(\mathbf{x}_i)} \right) \right] \right] \\ &= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{\frac{p(\mathbf{x}_1|\mathbf{y})}{p(\mathbf{x}_1)}}{\frac{1}{N-1} \sum_{i=2}^N \frac{p(\mathbf{x}_i|\mathbf{y})}{p(\mathbf{x}_i)}} \right) \right] \right] \\ &= I_{\text{InfoLOOB}}(X_1 ; Y) . \end{aligned} \quad (\text{A61})$$

where we applied Jensen's inequality for the exchanging the expectation and the "ln" to obtain the " $\leq$ " inequality. □

Experiments that compare upper and lower bounds as mutual information estimates are provided in [Cheng et al. \(2020\)](#) and in [Poole et al. \(2019\)](#). In Fig. 2 in [Cheng et al. \(2020\)](#) it is shown that InfoLOOB is a good estimator of the mutual information.

### A.1.3 InfoLOOB: Analysis of the Objective

This subsection justifies the maximization of the InfoLOOB bound for contrastive learning. Maximizing the InfoLOOB bound is not intuitive as it was introduced as an upper bound on the mutual information in the previous subsection. Still maximizing the InfoLOOB bound leads to a good approximation of the mutual information, in particular for high mutual information.

InfoLOOB with a neural network as a scoring function is not an upper bound on the mutual information when not under-sampling. As we use InfoLOOB on training data for which we do not know the sampling procedure, we cannot assume under-sampling. Therefore, we elaborate more on the rationale behind the maximization of the InfoLOOB bound. (I) We show that InfoLOOB with neural networks as scoring function is bounded from above. Therefore, there exists a maximum and the optimization problem is well defined. (II) We show that InfoLOOB with neural networks as scoring function differs by two terms from the mutual information. The first term is the Kullback-Leibler divergence between the variational  $q(\mathbf{x} | \mathbf{y})$  and the posterior  $p(\mathbf{x} | \mathbf{y})$ . This divergence is minimal for  $q(\mathbf{x} | \mathbf{y}) = p(\mathbf{x} | \mathbf{y})$ , which implies  $f(\mathbf{y} | \mathbf{x}) = p(\mathbf{y} | \mathbf{x})$ . The second term is governed by the difference between the mean  $\mathbb{E}[f(\mathbf{x}, \mathbf{y})]$  and the empirical mean  $1/(N-1) \sum_i f(\mathbf{x}_i, \mathbf{y})$ . The Hoeffding bound can be applied to this difference. Therefore, the second term is negligible for large  $N$ . In contrast, the KL term is dominant and the relevant term, therefore maximizing InfoLOOB leads to  $f(\mathbf{y} | \mathbf{x}) \approx p(\mathbf{y} | \mathbf{x})$ .

We assume that an anchor sample  $\mathbf{y}$  is given. For the anchor sample  $\mathbf{y}$ , we draw a positive sample  $\mathbf{x}_1$  according to  $p(\mathbf{x}_1 | \mathbf{y})$ . We define the set  $\tilde{X} = \{\mathbf{x}_2, \dots, \mathbf{x}_N\}$  of negative samples, which are drawn iid according to  $p(\mathbf{x})$ . We define the set  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .

We have

$$p(\tilde{X}) = \prod_{i=2}^N p(\mathbf{x}_i), \quad (\text{A62})$$

$$p(X | \mathbf{y}) = p(\mathbf{x}_1 | \mathbf{y}) \prod_{i=2}^N p(\mathbf{x}_i) = p(\mathbf{x}_1 | \mathbf{y}) p(\tilde{X}), \quad (\text{A63})$$

$$p(X) = \prod_{i=1}^N p(\mathbf{x}_i) = p(\mathbf{x}_1) p(\tilde{X}). \quad (\text{A64})$$

We use the score function

$$f(\mathbf{x}, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}, \mathbf{y})), \quad (\text{A65})$$

where  $\text{sim}(\mathbf{x}, \mathbf{y})$  is typically the cosine similarity.

The InfoLOOB with score function  $f(\mathbf{x}, \mathbf{y})$  is defined as

$$I_{\text{InfoLOOB}}(X_1; Y) = E_{p(\mathbf{y})} \left[ E_{p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right]. \quad (\text{A66})$$

We define the variational distribution

$$q(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}) f(\mathbf{x}, \mathbf{y})}{Z(\mathbf{y})}, \quad (\text{A67})$$

$$Z(\mathbf{y}) = E_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]. \quad (\text{A68})$$

The next inequality shows the relation between  $I(X_1 ; Y)$  and  $I_{\text{InfoLOOB}}(X_1 ; Y)$  for random variables  $X_1$  and  $Y$ .

$$\begin{aligned}
I(X_1 ; Y) &= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{p(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] & (\text{A69}) \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{q(\mathbf{x}_1 | \mathbf{y})} \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right) \right] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{q(\mathbf{x}_1 | \mathbf{y})}{p(\mathbf{x}_1)} \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \frac{f(\mathbf{x}_1, \mathbf{y})}{Z(\mathbf{y})} \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right]} \right) \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} [\ln f(\mathbf{x}_1, \mathbf{y})] - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \right) \right] \\
&+ \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y})} [\ln f(\mathbf{x}_1, \mathbf{y})] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] \\
&+ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \right) \right] \\
&+ \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{f(\mathbf{x}_1, \mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] + \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] \\
&- \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \right) \right] + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= I_{\text{InfoLOOB}}(X_1 ; Y) \\
&+ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \ln \left( \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \right) \right] \\
&+ \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= I_{\text{InfoLOOB}}(X_1 ; Y) \\
&+ \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(X|\mathbf{y})} \left[ \ln \left( \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right) \right] \right] - \mathbb{E}_{p(\mathbf{y})} [\ln (\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})])] \\
&+ \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= I_{\text{InfoLOOB}}(X_1 ; Y) \\
&- \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] \\
&+ \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))] \\
&= I_{\text{InfoLOOB}}(X_1 ; Y) - \text{DE} + \mathbb{E}_{p(\mathbf{y})} [\text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y}))],
\end{aligned}$$

where we used

$$\text{DE} = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] = \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \ln \left( \frac{Z(\mathbf{y})}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \right] \right] \quad (\text{A70})$$

(DE for difference of expectations) and

$$\begin{aligned} Z(\mathbf{y}) &= \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] = \mathbb{E}_{p(\bar{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right] \\ &= \mathbb{E}_{p(X|\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right]. \end{aligned} \quad (\text{A71})$$

Since both KL and DE are non-negative (for DE see below), to increase InfoLOOB we have either to decrease KL or to increase DE.

**Bounding DE.** Next we bound DE. We define

$$L = \mathbf{z}^T \mathbf{x} - \beta^{-1} \sum_{i=1}^N z_i \ln z_i. \quad (\text{A72})$$

The *log-sum-exp function* (lse) is

$$\text{lse}(\beta, \mathbf{a}) = \beta^{-1} \log \left( \sum_{i=1}^N \exp(\beta a_i) \right), \quad (\text{A73})$$

for  $\beta > 0$  and vector  $\mathbf{a} = (a_1, \dots, a_N)$ .

The lse is a convex function (Lemma 4 in (Gao & Pavel, 2017)). We obtain via Jensen's inequality and the convex lse:

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\bar{X})} \left[ \ln \left( \mathbb{E}_{p(\mathbf{x}_1)} \left[ \frac{\exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y}))}{\frac{1}{N-1} \sum_{i=2}^N \exp(\tau^{-1} \text{sim}(\mathbf{x}_i, \mathbf{y}))} \right] \right) \right] \right] \\ &\leq \mathbb{E}_{p(\mathbf{y})} \left[ \ln \mathbb{E}_{p(\mathbf{x}_1)} \left[ \exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y})) \right] - \tau^{-1} \mathbb{E}_{p(\mathbf{x}_1)} \left[ \text{sim}(\mathbf{x}_1, \mathbf{y}) \right] \right]. \end{aligned} \quad (\text{A74})$$

Again using Jensen's inequality and the concave ln, we get

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\bar{X})} \left[ \ln \left( \mathbb{E}_{p(\mathbf{x}_1)} \left[ \frac{\exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y}))}{\frac{1}{N-1} \sum_{i=2}^N \exp(\tau^{-1} \text{sim}(\mathbf{x}_i, \mathbf{y}))} \right] \right) \right] \right] \\ &\geq \mathbb{E}_{p(\mathbf{y})} \left[ \ln \mathbb{E}_{p(\mathbf{x}_1)} \left[ \exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y})) \right] - \ln \left( \frac{1}{N-1} \sum_{i=2}^N \mathbb{E}_{p(\mathbf{x}_i)} \left[ \exp(\tau^{-1} \text{sim}(\mathbf{x}_i, \mathbf{y})) \right] \right) \right] \\ &= 0. \end{aligned} \quad (\text{A75})$$

If we combine both previous inequalities, we obtain

$$0 \leq \text{DE} \leq \mathbb{E}_{p(\mathbf{y})} \left[ \ln \mathbb{E}_{p(\mathbf{x}_1)} \left[ \exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y})) \right] - \tau^{-1} \mathbb{E}_{p(\mathbf{x}_1)} \left[ \text{sim}(\mathbf{x}_1, \mathbf{y}) \right] \right]. \quad (\text{A76})$$

In particular, for bounded  $\text{sim}(\mathbf{x}_1, \mathbf{y})$ , we get

$$0 \leq \text{DE} \leq \tau^{-1} \left( \max_{\mathbf{y}, \mathbf{x}_1} \text{sim}(\mathbf{x}_1, \mathbf{y}) - \min_{\mathbf{y}, \mathbf{x}_1} \text{sim}(\mathbf{x}_1, \mathbf{y}) \right), \quad (\text{A77})$$

while Hoeffding's lemma gives

$$0 \leq \text{DE} \leq \frac{1}{8} \tau^{-2} \left( \max_{\mathbf{y}, \mathbf{x}_1} \text{sim}(\mathbf{x}_1, \mathbf{y}) - \min_{\mathbf{y}, \mathbf{x}_1} \text{sim}(\mathbf{x}_1, \mathbf{y}) \right)^2. \quad (\text{A78})$$

Thus, for bounded  $\text{sim}(\mathbf{x}_1, \mathbf{y})$ , DE is bounded, therefore also InfoLOOB.

Next, we show that DE is small. The Hoeffding bound (Proposition 2.5 in [Wainwright \(2019\)](#)) states that if the random variable  $S_f = f(X_1, \mathbf{y})$  with  $X_1 \sim p(\mathbf{x}_1)$  is  $\sigma_f^2$ -sub-Gaussian then with random variables  $S_f^i$  iid distributed like  $S_f$

$$p \left( \left| \mathbb{E}[S_f] - \frac{1}{N-1} \sum_{i=2}^N S_f^i \right| \geq \epsilon \right) \quad (\text{A79})$$

$$= p \left( \left| \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] - \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right| \geq \epsilon \right) \leq 2 \exp \left( - \frac{(N-1) \epsilon^2}{2 \sigma_f^2} \right). \quad (\text{A80})$$

If  $S_f \in [a, b]$  (e.g. if  $f(\mathbf{x}, \mathbf{y}) \in [a, b]$ ) then we can set  $\sigma_f = (b - a)/2$ .

For

$$\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] - \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \leq \epsilon \quad (\text{A81})$$

we have

$$\begin{aligned} \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) &\leq \ln \left( \frac{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) + \epsilon}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \\ &\leq \frac{\epsilon}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \leq \frac{\epsilon}{Z - \epsilon}, \end{aligned} \quad (\text{A82})$$

where we used  $\ln a \leq a - 1$  for  $0 < a$ . Analog for

$$\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) - \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] \leq \epsilon \quad (\text{A83})$$

we have

$$\begin{aligned} \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) &\geq \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] + \epsilon} \right) \\ &= - \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] + \epsilon}{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]} \right) \geq - \frac{\epsilon}{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]} = - \frac{\epsilon}{Z}, \end{aligned} \quad (\text{A84})$$

where we used  $-\ln a \geq 1 - a$  for  $0 < a$ .

In summary, if for all  $\mathbf{y}$

$$\left| \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] - \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right| \leq \epsilon, \quad (\text{A85})$$

then we have

$$- \frac{\epsilon}{Z} \leq \ln \left( \frac{\mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})]}{\frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y})} \right) \leq \frac{\epsilon}{Z - \epsilon}. \quad (\text{A86})$$

It follows that

$$- \epsilon \mathbb{E}_{p(\mathbf{y})} [Z(\mathbf{y})^{-1}] \leq \text{DE} \leq \epsilon \mathbb{E}_{p(\mathbf{y})} [(Z(\mathbf{y}) - \epsilon)^{-1}]. \quad (\text{A87})$$

Consequently, for large  $N$ , the Hoeffding bound Eq. (A79) holds with high probability, if  $\epsilon$  is chosen reasonably. Therefore, with high probability the term DE is small.

Next, we show that DE is governed by the variance of  $\text{sim}(\mathbf{x}, \mathbf{y})$  for unmatched pairs.

In Eq. (A76) the term DE is upper bounded:

$$\text{DE} \leq \mathbb{E}_{p(\mathbf{y})} [\ln \mathbb{E}_{p(\mathbf{x}_1)} [\exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y}))]] - \tau^{-1} \mathbb{E}_{p(\mathbf{x}_1)} [\text{sim}(\mathbf{x}_1, \mathbf{y})] \quad (\text{A88})$$

We express these equations via the random variable  $S = \text{sim}(X_1, \mathbf{y})$  with  $s = \text{sim}(\mathbf{x}_1, \mathbf{y})$ , which replaces the random variable  $X_1$  and its realization  $\mathbf{x}_1$ .

$$\begin{aligned} \mathbb{E}_{p(\mathbf{x}_1)} [g(\text{sim}(\mathbf{x}_1, \mathbf{y}))] &= \int p(\mathbf{x}_1) g(\text{sim}(\mathbf{x}_1, \mathbf{y})) d\mathbf{x}_1 \quad (\text{A89}) \\ &= \int p(\mathbf{x}_1) \int \delta(s - \text{sim}(\mathbf{x}_1, \mathbf{y})) g(s) ds d\mathbf{x}_1 = \int \int p(\mathbf{x}_1) \delta(s - \text{sim}(\mathbf{x}_1, \mathbf{y})) d\mathbf{x}_1 g(s) ds \\ &= \int p(s) g(s) ds = \mathbb{E}_{p(s)} [g(s)] , \end{aligned}$$

where we used the Dirac delta-distribution  $\delta$  and for  $s = \text{sim}(\mathbf{x}_1, \mathbf{y})$  we defined

$$p(s) = \int p(\mathbf{x}_1) \delta(s - \text{sim}(\mathbf{x}_1, \mathbf{y})) d\mathbf{x}_1 . \quad (\text{A90})$$

Eq. (A76) can be written as

$$\text{DE} \leq \mathbb{E}_{p(\mathbf{y})} [\ln \mathbb{E}_{p(s)} [\exp(\tau^{-1} s)]] - \tau^{-1} \mathbb{E}_{p(s)} [s] = \mathbb{E}_{p(\mathbf{y})} [\ln \mathbb{E}_{p(s)} [\exp(\tau^{-1} (s - \bar{s}))]] , \quad (\text{A91})$$

with  $\bar{s} = \mathbb{E}_{p(s)} [s]$ .

We assume that the random variable  $S$  with realization  $s = \text{sim}(\mathbf{x}_1, \mathbf{y})$  is sub-Gaussian, where  $\mathbf{y}$  is given and  $\mathbf{x}_1$  is drawn independently from  $\mathbf{y}$  according to  $p(\mathbf{x}_1)$ . Therefore, we assume that the similarity of a random matching is sub-Gaussian. This assumption is true for bounded similarities (like the cosine similarity) and for almost sure bounded similarities. The assumption is true if using vectors that are retrieved from a continuous modern Hopfield network since they are bounded by the largest stored vector. This assumption is true for a continuous similarity function if  $\mathbf{x}$ ,  $\mathbf{y}$ , and parameters are bounded, since the bounded  $\mathbf{x}$ ,  $\mathbf{y}$ , and parameters can be embedded in a compact set on which the similarity has a maximum. This assumption is true for learned similarities if the input is bounded.

For a random variable  $S$  that is  $\sigma^2$ -sub-Gaussian (Definition 2.2 in Wainwright (2019)) the constant  $\sigma^2$  is called a *proxy variance*. The minimal proxy variance  $\sigma_{\text{opt}}^2$  is called the *optimal proxy variance* with  $\text{Var}[S] \leq \sigma_{\text{opt}}^2$  (Arbel et al., 2019).  $S$  is *strictly* sub-Gaussian, if  $\sigma_{\text{opt}}^2 = \text{Var}[S]$ . Proposition 2.1 in Arbel et al. (2019) states

$$\sigma_{\text{opt}}^2 = \sup_{\lambda \in \mathbb{R}} \frac{2}{\lambda^2} \ln (\mathbb{E} [\exp(\lambda(S - \mu))]) , \quad (\text{A92})$$

with  $\mu = \mathbb{E}[S]$ . The supremum is attained for almost surely bounded random variables  $S$ . Eq. (4) in Arbel et al. (2019) states

$$\lim_{\lambda \rightarrow 0} \frac{2}{\lambda^2} \ln (\mathbb{E} [\exp(\lambda(S - \mu))]) = \text{Var}[S] . \quad (\text{A93})$$

Thus, for  $S$  being sub-Gaussian, we have

$$\text{DE} \leq \tau^{-2} \mathbb{E}_{p(\mathbf{y})} [\sigma_{\text{opt}}^2(S)] , \quad (\text{A94})$$

where  $\sigma_{\text{opt}}^2$  is the optimal proxy variance of  $S$ . For example, bounded random variables  $S \in [a, b]$  are sub-Gaussian with  $\sigma_{\text{opt}} \leq (b - a)/2$  (Exercise 2.4 in Wainwright (2019)).

KL is decreased by making the variation distribution  $q(\mathbf{x}_1 | \mathbf{y})$  more similar to the posterior  $p(\mathbf{x}_1 | \mathbf{y})$ . The value DE only depends on the marginal distributions  $p(\mathbf{y})$  and  $p(\mathbf{x})$ , since  $p(\tilde{X}) = \prod_{i=2}^N p(\mathbf{x}_i)$ . The value DE can be changed by adding an offset to  $f(\mathbf{x}, \mathbf{y})$ . However, scaling  $f(\mathbf{x}, \mathbf{y})$  by a factor does not change DE. Consequently, DE is difficult to change.

Therefore, increasing InfoLOOB is most effective by making  $q(\mathbf{x}_1 | \mathbf{y})$  more similar to the posterior  $p(\mathbf{x}_1 | \mathbf{y})$ .

**Gradient of InfoLOOB expressed by gradients of KL and DE.** Assume that the similarity is parametrized by  $\mathbf{w}$  giving  $\text{sim}(\mathbf{x}, \mathbf{y}; \mathbf{w})$ .

$$\begin{aligned} \text{KL}(p(\mathbf{x}_1 | \mathbf{y}) \| q(\mathbf{x}_1 | \mathbf{y})) &= \int p(\mathbf{x}_1 | \mathbf{y}) \ln \left( \frac{p(\mathbf{x}_1 | \mathbf{y})}{q(\mathbf{x}_1 | \mathbf{y})} \right) d\mathbf{x} \\ &= -\tau^{-1} \int p(\mathbf{x}_1 | \mathbf{y}) \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w}) d\mathbf{x}_1 + \ln Z + C, \end{aligned} \quad (\text{A95})$$

where  $C$  is independent of  $\mathbf{w}$ .

Next, we compute the derivative of KL with respect to parameters  $\mathbf{w}$ .

$$\begin{aligned} \frac{\partial \text{KL}}{\partial \mathbf{w}} & \quad (\text{A96}) \\ &= -\tau^{-1} \int p(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 + \frac{1}{Z} \int p(\mathbf{x}_1) \frac{\exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w}))}{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})} \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \\ &= -\tau^{-1} \int p(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 + \tau^{-1} \int p(\mathbf{x}_1) \frac{\exp(\tau^{-1} \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w}))}{Z} \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \\ &= -\tau^{-1} \int p(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 + \tau^{-1} \int q(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \\ &= \tau^{-1} \int (q(\mathbf{x}_1 | \mathbf{y}) - p(\mathbf{x}_1 | \mathbf{y})) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1. \end{aligned}$$

The derivative is the average difference between the posterior distribution  $p(\mathbf{x}_1 | \mathbf{y})$  and the variational distribution  $q(\mathbf{x}_1 | \mathbf{y})$  multiplied by the derivative of the similarity function. If both distribution match, then the derivative vanishes.



Next, we compute the derivative of DE with respect to parameters  $\mathbf{w}$ .

$$\begin{aligned}
& \frac{\partial \text{DE}}{\partial \mathbf{w}} \tag{A97} \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \frac{\partial \ln Z}{\partial \mathbf{w}} \right] - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{\frac{1}{N-1} \sum_{i=2}^N \tau^{-1} \exp(\tau^{-1} \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})) \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}}}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \right] \right] \\
&= \mathbb{E}_{p(\mathbf{y})} \left[ \tau^{-1} \int q(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \right] \\
&\quad - \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{\frac{1}{N-1} \sum_{i=2}^N \tau^{-1} \exp(\tau^{-1} \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})) \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}}}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \int q(\mathbf{x}_1 | \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \right] \\
&\quad - \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \int \frac{p(\mathbf{x}_1) f(\mathbf{x}_1, \mathbf{y})}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]} \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} d\mathbf{x}_1 \right] \\
&\quad - \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\mathbf{x}_1)} \left[ \frac{f(\mathbf{x}_1, \mathbf{y})}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]} \frac{\partial \text{sim}(\mathbf{x}_1, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&\quad - \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \frac{1}{N-1} \sum_{i=2}^N \mathbb{E}_{p(\mathbf{x}_i)} \left[ \frac{f(\mathbf{x}_i, \mathbf{y})}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&\quad - \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&\quad - \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \frac{f(\mathbf{x}_i, \mathbf{y})}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \left( \frac{1}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y})]} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \right) f(\mathbf{x}_i, \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right] \\
&= \tau^{-1} \mathbb{E}_{p(\mathbf{y})} \left[ \mathbb{E}_{p(\tilde{X})} \left[ \frac{1}{N-1} \sum_{i=2}^N \left( \frac{1}{Z} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \right) f(\mathbf{x}_i, \mathbf{y}) \frac{\partial \text{sim}(\mathbf{x}_i, \mathbf{y}; \mathbf{w})}{\partial \mathbf{w}} \right] \right].
\end{aligned}$$

The derivative is the average of  $\frac{1}{Z} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})}$  multiplied by the score function and the derivative of the similarity function. The average is over  $\mathbf{y}$  and  $\tilde{X}$ , therefore the whole derivative becomes even smaller. Consequently, for small  $b - a$  and large  $N$ , the derivative of DE is small.

Note that for

$$\left| \mathbb{E}_{p(\mathbf{x}_1)} [f(\mathbf{x}_1, \mathbf{y})] - \frac{1}{N-1} \sum_{i=2}^N f(\mathbf{x}_i, \mathbf{y}) \right| \leq \epsilon \tag{A98}$$

we have

$$\frac{1}{Z} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \leq \frac{1}{Z} - \frac{1}{Z + \epsilon} = \frac{\epsilon}{Z(Z + \epsilon)}, \quad (\text{A99})$$

$$\frac{1}{Z} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \geq \frac{1}{Z} - \frac{1}{Z - \epsilon} = -\frac{\epsilon}{Z(Z - \epsilon)}, \quad (\text{A100})$$

therefore

$$\left| \frac{1}{Z} - \frac{1}{\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})} \right| \leq \frac{\epsilon}{Z(Z - \epsilon)}. \quad (\text{A101})$$

If the expectation  $Z$  is well approximated by the average  $\frac{1}{N-1} \sum_{j=2}^N f(\mathbf{x}_j, \mathbf{y})$ , then both DE and its gradient are small.

Derivative of InfoLOOB via KL and DE:

$$\frac{\partial I_{\text{InfoLOOB}}(X_1; Y)}{\partial \mathbf{w}} = \frac{\partial \text{DE}}{\partial \mathbf{w}} - \frac{\partial \text{KL}}{\partial \mathbf{w}}. \quad (\text{A102})$$

In this gradient, the KL term is dominating, therefore  $f(\mathbf{x}, \mathbf{y})$  is pushed to approximate the conditional probability  $p(\mathbf{y} | \mathbf{x})$ . Modern Hopfield networks lead to larger values of  $p(\mathbf{y} | \mathbf{x})$  as the mutual information becomes larger, therefore modern Hopfield networks help to push  $f(\mathbf{x}, \mathbf{y})$  to large values. Furthermore, modern Hopfield networks increase  $Z$ , which is in the denominator of the bound on DE and its derivative.

#### A.1.4 InfoNCE and InfoLOOB: Gradients

We consider the InfoNCE and the InfoLOOB loss function. For computing the loss function, we sample  $N$  pairs independently from  $p(\mathbf{x}, \mathbf{y})$ , which gives the training set  $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . InfoNCE and InfoLOOB only differ in using the positive example in the negatives. More precisely, for the sample  $\mathbf{y}_1$ , InfoNCE uses for the matrix of negative samples  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ , while InfoLOOB uses  $\tilde{\mathbf{X}} = (\mathbf{x}_2, \dots, \mathbf{x}_N)$ .

**InfoNCE.** The InfoNCE loss is

$$L_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right) = \frac{1}{N} \sum_{i=1}^N L_{\text{InfoNCE}}(\mathbf{y}_i), \quad (\text{A103})$$

where we used

$$L_{\text{InfoNCE}}(\mathbf{y}_i) = -\ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right). \quad (\text{A104})$$

For the score function  $f(\mathbf{x}, \mathbf{y})$ , we use

$$f(\mathbf{x}, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}, \mathbf{y})), \quad (\text{A105})$$

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T \mathbf{x} \quad (\text{A106})$$

with  $\tau$  as the temperature.

The loss function for this score function is

$$L_{\text{InfoNCE}}(\mathbf{y}) = -\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \text{lse}(\tau^{-1}, \mathbf{X}^T \mathbf{y}), \quad (\text{A107})$$

where lse is the *log-sum-exp function* (lse):

$$\text{lse}(\beta, \mathbf{a}) = \beta^{-1} \log \left( \sum_{i=1}^N \exp(\beta a_i) \right), \quad (\text{A108})$$

for  $\beta > 0$  and vector  $\mathbf{a} = (a_1, \dots, a_N)$ .

The gradient with respect to  $\mathbf{y}$  is

$$\frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{y}} = -\tau^{-1} \mathbf{x}_1 + \tau^{-1} \mathbf{X} \text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y}), \quad (\text{A109})$$

which is the positive example  $\mathbf{x}_1$  that fits to the anchor example  $\mathbf{y}$  minus the Hopfield network update with state pattern  $\mathbf{y}$  and stored patterns  $\mathbf{X}$  and then this difference multiplied by  $\tau^{-1}$ .

This gradient can be simplified, since the positive example  $\mathbf{x}_1$  is also in the negative examples. Using  $\mathbf{p} = (p_1, \dots, p_N)^T = \text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y})$ , we obtain

$$\begin{aligned} \frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{y}} &= -\tau^{-1} (1 - p_1) \left( \mathbf{x}_1 - \frac{1}{1 - p_1} \mathbf{X} (\text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y}) - (p_1, 0, \dots, 0)^T) \right) \\ &= -\tau^{-1} (1 - p_1) \left( \mathbf{x}_1 - \tilde{\mathbf{X}} \text{softmax}(\tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y}) \right) = (1 - p_1) \frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{y}}. \end{aligned} \quad (\text{A110})$$

where

$$\begin{aligned} &\frac{1}{1 - p_1} \mathbf{X} (\text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y}) - (p_1, 0, \dots, 0)^T) \\ &= \frac{1}{1 - p_1} \mathbf{X} ((p_1, p_2, \dots, p_N)^T - (p_1, 0, \dots, 0)^T) \\ &= \frac{1}{1 - p_1} \mathbf{X} (0, p_2, \dots, p_N)^T = \frac{1}{1 - p_1} \sum_{i=2}^N p_i \mathbf{x}_i \end{aligned} \quad (\text{A111})$$

is the softmax average over the negatives  $\mathbf{x}_i$  for  $2 \leq i \leq N$  without  $\mathbf{x}_1$ . It can be easily seen that  $\frac{1}{1 - p_1} \sum_{i=2}^N p_i = \frac{1 - p_1}{1 - p_1} = 1$ . For the derivative of the InfoLOOB see below.

The gradient with respect to  $\mathbf{x}_1$  is

$$\frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{x}_1} = -\tau^{-1} \mathbf{y} + \tau^{-1} \frac{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y})}{\sum_{i=1}^N \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y})} \mathbf{y} \quad (\text{A112})$$

$$= -\tau^{-1} (1 - p_1) \mathbf{y}. \quad (\text{A113})$$

Consequently, the learning rate is scaled by  $(1 - p_1)$ .

The sum of gradients with respect to  $\mathbf{x}_1$  and  $\mathbf{x}_i$  is

$$\begin{aligned} \frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{x}_1} + \sum_{i=1}^N \frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{x}_i} &= -\tau^{-1} \mathbf{y} + \tau^{-1} \mathbf{y} \mathbf{1}^T \text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y}) \\ &= -\tau^{-1} \mathbf{y} + \tau^{-1} \mathbf{y} = 0, \end{aligned} \quad (\text{A114})$$

where  $\mathbf{1}$  is the vector with ones. However, the derivatives with respect to the weights are not zero since the  $\mathbf{x}_i$  are differently computed.

**InfoLOOB.** The InfoLOOB loss is

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right) = \frac{1}{N} \sum_{i=1}^N L_{\text{InfoLOOB}}(\mathbf{y}_i), \quad (\text{A115})$$

where we used

$$L_{\text{InfoLOOB}}(\mathbf{y}_i) = -\ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right). \quad (\text{A116})$$

For the score function  $f(\mathbf{x}, \mathbf{y})$ , we use

$$f(\mathbf{x}, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}, \mathbf{y})), \quad (\text{A117})$$

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T \mathbf{x} \quad (\text{A118})$$

with  $\tau$  as the temperature.

The loss function for this score function is

$$L_{\text{InfoLOOB}}(\mathbf{y}) = -\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \text{lse} \left( \tau^{-1}, \tilde{\mathbf{X}}^T \mathbf{y} \right), \quad (\text{A119})$$

where  $\text{lse}$  is the log-sum-exponential function.

The gradient with respect to  $\mathbf{y}$  is

$$\frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{y}} = -\tau^{-1} \mathbf{x}_1 + \tau^{-1} \tilde{\mathbf{X}} \text{softmax} \left( \tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y} \right), \quad (\text{A120})$$

which is the positive example  $\mathbf{x}_1$  that fits to the anchor example  $\mathbf{y}$  minus the Hopfield network update with state pattern  $\mathbf{y}$  and stored patterns  $\tilde{\mathbf{X}}$  and then this difference multiplied by  $\tau^{-1}$ .

The gradient with respect to  $\mathbf{x}_1$  is

$$\frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{x}_1} = -\tau^{-1} \mathbf{y}. \quad (\text{A121})$$

The sum of gradients with respect to  $\mathbf{x}_1$  and  $\mathbf{x}_i$  is

$$\begin{aligned} \frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{x}_1} + \sum_i \frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{x}_i} &= -\tau^{-1} \mathbf{y} + \tau^{-1} \mathbf{y} \mathbf{1}^T \text{softmax} \left( \tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y} \right) \\ &= -\tau^{-1} \mathbf{y} + \tau^{-1} \mathbf{y} = 0, \end{aligned} \quad (\text{A122})$$

where  $\mathbf{1}$  is the vector with ones. However, the derivatives with respect to the weights are not zero since the  $\mathbf{x}_i$  are differently computed.

**Gradients with respect to  $\tau^{-1}$ .** The gradient of the InfoNCE loss Eq. (A103) using the similarity Eq. (A105) with respect to  $\tau^{-1}$  is

$$\frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \tau^{-1}} = -\mathbf{y}^T \mathbf{x}_1 + \mathbf{y}^T \mathbf{X} \text{softmax} \left( \tau^{-1} \mathbf{X}^T \mathbf{y} \right) \quad (\text{A123})$$

$$= -\mathbf{y}^T \left( \mathbf{x}_1 - \mathbf{X} \text{softmax} \left( \tau^{-1} \mathbf{X}^T \mathbf{y} \right) \right), \quad (\text{A124})$$

which is the similarity of the anchor  $\mathbf{y}$  with the difference of the positive example  $\mathbf{x}_1$  and the Hopfield network update with state pattern  $\mathbf{y}$  and stored patterns  $\mathbf{X}$ . The gradient of the InfoLOOB loss Eq. (A115) using the similarity Eq. (A117) with respect to  $\tau^{-1}$  is

$$\frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \tau^{-1}} = -\mathbf{y}^T \mathbf{x}_1 + \mathbf{y}^T \tilde{\mathbf{X}} \text{softmax} \left( \tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y} \right) \quad (\text{A125})$$

$$= -\mathbf{y}^T \left( \mathbf{x}_1 - \tilde{\mathbf{X}} \text{softmax} \left( \tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y} \right) \right), \quad (\text{A126})$$

with the difference compared to Eq. (A123) that the Hopfield network update is done with stored patterns  $\tilde{\mathbf{X}}$  instead of  $\mathbf{X}$ . Without the positive example  $\mathbf{x}_1$  in the stored patterns  $\tilde{\mathbf{X}}$ , the term  $\mathbf{x}_1 - \tilde{\mathbf{X}} \text{softmax} \left( \tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y} \right)$  in Eq. (A125) will not decrease like the term  $\mathbf{x}_1 - \mathbf{X} \text{softmax} \left( \tau^{-1} \mathbf{X}^T \mathbf{y} \right)$  in Eq. (A123) but grow even larger with better separation of the positive and negative examples.

### A.1.5 InfoLOOB and InfoNCE: Probability Estimators

In [McAllester & Stratos \(2018, 2020\)](#) it was shown that estimators of the mutual information by lower bounds have problems as they come with serious statistical limitations. Statistically more justified for representing the mutual information is a difference of entropies, which are estimated by minimizing the cross-entropy loss. Both InfoNCE and InfoLOOB losses can be viewed as cross-entropy losses.

We sample  $N$  pairs independently from  $p(\mathbf{x}, \mathbf{y})$ , which gives  $Z = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . We set  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  and  $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ , so that,  $Z = X \times Y$ . The score function  $f(\mathbf{x}, \mathbf{y})$  is an estimator for  $p(\mathbf{x}, \mathbf{y})$ . Then we obtain

estimators  $\hat{q}$  for the conditional probabilities.  $\hat{q}(\mathbf{y}_i | \mathbf{x}_i, Y \setminus \{\mathbf{y}_i\})$  is an estimator for  $p(\mathbf{y}_i | \mathbf{x}_i)$  and  $\hat{q}(\mathbf{x}_i | \mathbf{y}_i, X \setminus \{\mathbf{x}_i\})$  an estimator for  $p(\mathbf{x}_i | \mathbf{y}_i)$ . Each estimator  $\hat{q}$  uses beyond  $(\mathbf{x}_i, \mathbf{y}_i)$  additional samples to estimate the normalizing constant. For InfoNCE these estimators are

$$\hat{q}^1(\mathbf{y}_i | \mathbf{x}_i, Y \setminus \{\mathbf{y}_i\}) = \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_i, \mathbf{y}_j)} \approx \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\mathbb{E}_{p(\mathbf{y})} [f(\mathbf{x}_i, \mathbf{y})]}, \quad (\text{A127})$$

$$\hat{q}^2(\mathbf{x}_i | \mathbf{y}_i, X \setminus \{\mathbf{x}_i\}) = \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i)} \approx \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y}_i)]}. \quad (\text{A128})$$

The cross-entropy losses for the InfoNCE estimators are

$$L_{\text{InfoNCE}}^1 = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_i, \mathbf{y}_j)} \right), \quad (\text{A129})$$

$$L_{\text{InfoNCE}}^2 = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right). \quad (\text{A130})$$

For InfoLOOB these estimators are

$$\hat{q}^1(\mathbf{y}_i | \mathbf{x}_i, Y \setminus \{\mathbf{y}_i\}) = \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_i, \mathbf{y}_j)} \approx \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\mathbb{E}_{p(\mathbf{y})} [f(\mathbf{x}_i, \mathbf{y})]}, \quad (\text{A131})$$

$$\hat{q}^2(\mathbf{x}_i | \mathbf{y}_i, X \setminus \{\mathbf{x}_i\}) = \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i)} \approx \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y}_i)]}. \quad (\text{A132})$$

The cross-entropy losses for the InfoLOOB estimators are

$$L_{\text{InfoLOOB}}^1 = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_i, \mathbf{y}_j)} \right), \quad (\text{A133})$$

$$L_{\text{InfoLOOB}}^2 = -\frac{1}{N} \sum_{i=1}^N \ln \left( \frac{f(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i)} \right). \quad (\text{A134})$$

The InfoLOOB estimator uses for normalization

$$\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y}_i)] \approx \frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_j, \mathbf{y}_i), \quad (\text{A135})$$

$$\mathbb{E}_{p(\mathbf{y})} [f(\mathbf{x}_i, \mathbf{y})] \approx \frac{1}{N-1} \sum_{j=1, j \neq i}^N f(\mathbf{x}_i, \mathbf{y}_j), \quad (\text{A136})$$

in contrast to InfoNCE, which uses

$$\mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x}, \mathbf{y}_i)] \approx \frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_j, \mathbf{y}_i), \quad (\text{A137})$$

$$\mathbb{E}_{p(\mathbf{y})} [f(\mathbf{x}_i, \mathbf{y})] \approx \frac{1}{N} \sum_{j=1}^N f(\mathbf{x}_i, \mathbf{y}_j). \quad (\text{A138})$$

If InfoNCE estimates the normalizing constant separately, then it would be biased.  $(\mathbf{x}_i, \mathbf{y}_i)$  is drawn according to  $p(\mathbf{x}_i, \mathbf{y}_i)$  instead of  $p(\mathbf{x}_i)p(\mathbf{y}_i)$ . In contrast, if InfoLOOB estimated the normalizing constant separately, then it would be unbiased.

### A.1.6 InfoLOOB and InfoNCE: Losses

We have  $N$  pairs drawn iid from  $p(\mathbf{x}, \mathbf{y})$ , where we assume that a pair  $(\mathbf{x}_i, \mathbf{y}_i)$  is already an embedding of the original drawn pair. These build up the embedding training set  $Z =$

$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$  that allows to construct the matrices  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  of  $N$  embedding samples  $\mathbf{x}_i$  and  $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$  of  $N$  embedding samples  $\mathbf{y}_i$ . We also have  $M$  stored patterns  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_M)$  and  $K$  stored patterns  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_K)$ .

The state vectors  $\mathbf{x}_i$  and  $\mathbf{y}_i$  are the queries for the Hopfield networks, which retrieve some vectors from  $\mathbf{U}$  or  $\mathbf{V}$ . We normalize vectors  $\|\mathbf{x}_i\| = \|\mathbf{y}_i\| = \|\mathbf{u}_i\| = \|\mathbf{v}_i\| = 1$ . The following vectors are retrieved from modern Hopfield networks (Ramsauer et al., 2021):

$$\mathbf{U}_{\mathbf{x}_i} = \mathbf{U} \operatorname{softmax}(\beta \mathbf{U}^T \mathbf{x}_i), \quad (\text{A139}) \quad \mathbf{V}_{\mathbf{x}_i} = \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T \mathbf{x}_i), \quad (\text{A141})$$

$$\mathbf{U}_{\mathbf{y}_i} = \mathbf{U} \operatorname{softmax}(\beta \mathbf{U}^T \mathbf{y}_i), \quad (\text{A140}) \quad \mathbf{V}_{\mathbf{y}_i} = \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T \mathbf{y}_i). \quad (\text{A142})$$

where  $\mathbf{U}_{\mathbf{x}_i}$  denotes an image-retrieved image embedding,  $\mathbf{U}_{\mathbf{y}_i}$  a text-retrieved image embedding,  $\mathbf{V}_{\mathbf{x}_i}$  an image-retrieved text embedding and  $\mathbf{V}_{\mathbf{y}_i}$  a text-retrieved text embedding. The hyperparameter  $\beta$  corresponds to the inverse temperature:  $\beta = 0$  retrieves the average of the stored pattern, while large  $\beta$  retrieve the stored pattern that is most similar to the state pattern (query).

We consider the loss functions

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^N \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}, \quad (\text{A143})$$

$$\mathcal{L}_{\text{InfoLOOB}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}, \quad (\text{A144})$$

$$\mathcal{L}_{\text{InfoLOOB}}^{\text{H-UUVU}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_j})} - \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_j}^T \mathbf{V}_{\mathbf{y}_i})}, \quad (\text{A145})$$

$$\mathcal{L}_{\text{InfoLOOB}}^{\text{H-UUVV}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_j})} - \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\tau^{-1} \mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{V}_{\mathbf{x}_j}^T \mathbf{V}_{\mathbf{y}_i})}, \quad (\text{A146})$$

where for InfoLOOB the sum  $\sum_{j \neq i}$  in the denominator contains only negative examples  $j$ . We do not consider the loss function  $\mathcal{L}_{\text{InfoLOOB}}^{\text{H-UUVU}}$  because of the high variance in the dot product  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$ , as elaborated in the following.

Let us consider the dot product between the anchor retrieval with the positive pattern retrieval for the loss functions with Hopfield. In the first term of the loss function Eq. (A145),  $\mathbf{U}_{\mathbf{x}_i}$  is the anchor with  $\mathbf{V}_{\mathbf{y}_i}$  as the positive sample and  $\mathbf{V}_{\mathbf{y}_i}$  with  $\mathbf{U}_{\mathbf{x}_i}$  as the positive sample for the second term, since the anchor also appears in each term of the denominator. Equivalently the same is valid for Eq. (A146), but with positive samples  $\mathbf{V}_{\mathbf{x}_i}$  and  $\mathbf{U}_{\mathbf{y}_i}$  respectively. These dot products can be written as

$$\mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i} = \operatorname{softmax}(\beta \mathbf{U}^T \mathbf{x}_i)^T \mathbf{U}^T \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T \mathbf{y}_i), \quad (\text{A147})$$

$$\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i} = \operatorname{softmax}(\beta \mathbf{U}^T \mathbf{x}_i)^T \mathbf{U}^T \mathbf{U} \operatorname{softmax}(\beta \mathbf{U}^T \mathbf{y}_i), \quad (\text{A148})$$

$$\mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i} = \operatorname{softmax}(\beta \mathbf{V}^T \mathbf{x}_i)^T \mathbf{V}^T \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T \mathbf{y}_i). \quad (\text{A149})$$

**High variance of  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$ .** To compute the dot product  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$ ,  $M + K$  stored patterns are required ( $M$  of the  $\mathbf{u}_j$  and  $K$  of the  $\mathbf{v}_j$ ). In contrast, the dot products  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i}$  and  $\mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$  require only  $M$  or respectively  $K$  stored patterns. Therefore,  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$  has higher variance than both  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i}$  and  $\mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$ .

**Covariance structure extracted by  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i}$  and  $\mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i}$ .** The Jacobian  $\mathbf{J}$  of the softmax  $\mathbf{p} = \operatorname{softmax}(\beta \mathbf{a})$  is

$$\mathbf{J}(\beta \mathbf{a}) = \frac{\partial \operatorname{softmax}(\beta \mathbf{a})}{\partial \mathbf{a}} = \beta (\operatorname{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^T), \quad (\text{A150})$$

which is a symmetric, positive semi-definite matrix with one eigenvalue of zero for eigenvector  $\mathbf{1}$ .  $\mathbf{J}(\beta\mathbf{a})$  is diagonally dominant since  $|p_i(1-p_i)| - \sum_{j \neq i} |p_i p_j| = p_i - \sum_j p_i p_j = p_i - p_i = 0$ .

Next we give upper bounds on the norm of  $\mathbf{J}$ .

**Lemma A1.** *For a softmax  $\mathbf{p} = \text{softmax}(\beta\mathbf{x})$  with  $m = \max_i p_i(1-p_i)$ , the spectral norm of the Jacobian  $\mathbf{J}$  of the softmax is bounded:*

$$\|\mathbf{J}\|_2 \leq 2 m \beta, \quad (\text{A151})$$

$$\|\mathbf{J}\|_1 \leq 2 m \beta, \quad (\text{A152})$$

$$\|\mathbf{J}\|_\infty \leq 2 m \beta. \quad (\text{A153})$$

In particular everywhere holds

$$\|\mathbf{J}\|_2 \leq \frac{1}{2} \beta. \quad (\text{A154})$$

If  $p_{\max} = \max_i p_i \geq 1 - \epsilon \geq 0.5$ , then for the spectral norm of the Jacobian holds

$$\|\mathbf{J}\|_2 \leq 2 \epsilon \beta - 2 \epsilon^2 \beta < 2 \epsilon \beta. \quad (\text{A155})$$

*Proof.* We consider the maximum absolute column sum norm

$$\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}| \quad (\text{A156})$$

and the maximum absolute row sum norm

$$\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|. \quad (\text{A157})$$

We have for  $\mathbf{A} = \mathbf{J} = \beta (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)$

$$\begin{aligned} \sum_j |a_{ij}| &= \beta \left( p_i(1-p_i) + \sum_{j, j \neq i} p_i p_j \right) = \beta p_i (1 - 2p_i + \sum_j p_j) \\ &= 2 \beta p_i (1 - p_i) \leq 2 m \beta, \end{aligned} \quad (\text{A158})$$

$$\begin{aligned} \sum_i |a_{ij}| &= \beta \left( p_j(1-p_j) + \sum_{i, i \neq j} p_j p_i \right) = \beta p_j (1 - 2p_j + \sum_i p_i) \\ &= 2 \beta p_j (1 - p_j) \leq 2 m \beta. \end{aligned} \quad (\text{A159})$$

Therefore, we have

$$\|\mathbf{J}\|_1 \leq 2 m \beta, \quad (\text{A160})$$

$$\|\mathbf{J}\|_\infty \leq 2 m \beta, \quad (\text{A161})$$

$$\|\mathbf{J}\|_2 \leq \sqrt{\|\mathbf{J}\|_1 \|\mathbf{J}\|_\infty} \leq 2 m \beta. \quad (\text{A162})$$

The last inequality is a direct consequence of Hölder's inequality.

For  $0 \leq p_i \leq 1$ , we have  $p_i(1-p_i) \leq 0.25$ . Therefore,  $m \leq 0.25$  for all values of  $p_i$ .

If  $p_{\max} \geq 1 - \epsilon \geq 0.5$  ( $\epsilon \leq 0.5$ ), then  $1 - p_{\max} \leq \epsilon$  and for  $p_i \neq p_{\max}$   $p_i \leq \epsilon$ . The derivative  $\partial x(1-x)/\partial x = 1 - 2x > 0$  for  $x < 0.5$ , therefore  $x(1-x)$  increases with  $x$  for  $x < 0.5$ . Using  $x = 1 - p_{\max}$  and for  $p_i \neq p_{\max}$   $x = p_i$ , we obtain  $p_i(1-p_i) \leq \epsilon(1-\epsilon)$  for all  $i$ . Consequently, we have  $m \leq \epsilon(1-\epsilon)$ .  $\square$

For the softmax  $\mathbf{p} = \text{softmax}(\beta\mathbf{a})$  with Jacobian  $\partial \mathbf{J} / \partial \mathbf{a} = \mathbf{J}(\beta\mathbf{a}) = \beta (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)$  and for arbitrary  $N$ -dimensional vectors  $\mathbf{b}$  and  $\mathbf{c}$ , we have

$$\mathbf{b}^T \mathbf{J}(\beta\mathbf{a}) \mathbf{c} = \beta \mathbf{b}^T (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T) \mathbf{c} = \beta \left( \sum_i p_i b_i c_i - \left( \sum_i p_i b_i \right) \left( \sum_i p_i c_i \right) \right). \quad (\text{A163})$$

Therefore,  $\mathbf{b}^T \mathbf{J}(\beta \mathbf{a}) \mathbf{c}$  is  $\beta$  times the covariance between  $\mathbf{b}$  and  $\mathbf{c}$  if component  $i$  is drawn with probability  $p_i$  of the multinomial distribution  $\mathbf{p}$ . In our case the component  $i$  is sample  $i$ .

Using the mean  $\hat{\mathbf{u}} = 1/M \sum_{i=1}^M \mathbf{u}_i$ , the empirical covariance of data  $\mathbf{U}$  is

$$\text{Cov}(\mathbf{U}) = 1/M \mathbf{U} \mathbf{U}^T - \hat{\mathbf{u}} \hat{\mathbf{u}}^T, \quad (\text{A164})$$

$$[\text{Cov}(\mathbf{U})]_{kl} = \sum_{i=1}^M 1/M u_{ik} u_{il} - \left( \sum_{i=1}^M 1/M u_{ik} \right) \left( \sum_{i=1}^M 1/M u_{il} \right). \quad (\text{A165})$$

The weighted covariance (samples  $\mathbf{u}_i$  are drawn according to  $p_i$ )

$$\text{Cov}(\mathbf{U}) = \mathbf{U} \mathbf{J}(\beta \mathbf{a}) \mathbf{U}^T, \quad (\text{A166})$$

$$[\text{Cov}(\mathbf{U})]_{kl} = \beta \left( \sum_{i=1}^M p_i u_{ik} u_{il} - \left( \sum_{i=1}^M p_i u_{ik} \right) \left( \sum_{i=1}^M p_i u_{il} \right) \right), \quad (\text{A167})$$

which replaces  $1/M$  from equal sampling by the  $p_i$ , that is,  $\mathbf{u}_i$  is sampled with probability  $p_i$ .

The next theorem states how to express the dot product  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i}$  by weighted covariances of the data  $\mathbf{U}$ .

**Theorem A3** (Weighted Covariances). *Using the weighted covariances*

$$\text{Cov}(\mathbf{U}, \mathbf{y}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T, \quad \text{Cov}(\mathbf{U}, \mathbf{x}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T, \quad (\text{A168})$$

$$\mathbf{J}^m(\beta \mathbf{a}) = \int_0^1 \mathbf{J}(\lambda \beta \mathbf{a}) d\lambda, \quad (\text{A169})$$

where the mean Jacobian  $\mathbf{J}^m$  is symmetric, diagonally dominant, and positive semi-definite with spectral norm bounded by  $\|\mathbf{J}^m\|_2 \leq 0.5\beta$ .

The dot product  $\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i}$  can be expressed by the weighted covariances

$$\mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i} = (\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{x}_i) \mathbf{x}_i)^T (\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{y}_i) \mathbf{y}_i), \quad (\text{A170})$$

where the mean is  $\bar{\mathbf{u}} = 1/M \mathbf{U} \mathbf{1}$ .

*Proof.* We apply the mean value theorem to the softmax with the symmetric, diagonally dominant, positive semi-definite Jacobian matrix  $\mathbf{J}^m = \int_0^1 \mathbf{J}(\lambda \mathbf{a} + (1-\lambda)\mathbf{a}') d\lambda$ :

$$\text{softmax}(\mathbf{a}) - \text{softmax}(\mathbf{a}') = \mathbf{J}^m (\mathbf{a} - \mathbf{a}'). \quad (\text{A171})$$

We set  $\mathbf{a}' = \mathbf{0}$  and use  $\beta \mathbf{a}$  instead of  $\mathbf{a}$ , which gives:

$$\text{softmax}(\beta \mathbf{a}) = 1/M \mathbf{1} + \mathbf{J}^m(\beta \mathbf{a}) \mathbf{a}, \quad \mathbf{J}^m(\beta \mathbf{a}) = \int_0^1 \mathbf{J}(\lambda \beta \mathbf{a}) d\lambda, \quad (\text{A172})$$

which is exact. We obtain

$$\text{softmax}(\beta \mathbf{U}^T \mathbf{x}_i) = 1/M \mathbf{1} + \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T \mathbf{x}_i, \quad (\text{A173})$$

$$\text{softmax}(\beta \mathbf{U}^T \mathbf{y}_i) = 1/M \mathbf{1} + \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T \mathbf{y}_i. \quad (\text{A174})$$

The spectral norm of  $\mathbf{J}^m$  is bounded by  $\|\mathbf{J}^m\|_2 \leq 0.5\beta$ , since this bound holds for every  $\mathbf{J}(\lambda \beta \mathbf{a})$  in  $\mathbf{J}^m(\beta \mathbf{a}) = \int_0^1 \mathbf{J}(\lambda \beta \mathbf{a}) d\lambda$  according to Lemma A1.

The dot product between the anchor retrieval and the positive sample is:

$$\begin{aligned} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i} &= \text{softmax}(\beta \mathbf{U}^T \mathbf{x}_i)^T \mathbf{U}^T \mathbf{U} \text{softmax}(\beta \mathbf{U}^T \mathbf{y}_i) \\ &= (1/M \mathbf{1} + \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T \mathbf{x}_i)^T \mathbf{U}^T \mathbf{U} (1/M \mathbf{1} + \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T \mathbf{y}_i) \\ &= (1/M \mathbf{U} \mathbf{1} + \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T \mathbf{x}_i)^T (1/M \mathbf{U} \mathbf{1} + \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T \mathbf{y}_i) \\ &= (\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{x}_i) \mathbf{x}_i)^T (\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{y}_i) \mathbf{y}_i), \end{aligned} \quad (\text{A175})$$

where we used the mean  $\bar{\mathbf{u}} = 1/M \mathbf{U} \mathbf{1}$  and the weighted covariances

$$\text{Cov}(\mathbf{U}, \mathbf{y}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T, \quad \text{Cov}(\mathbf{U}, \mathbf{x}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T. \quad (\text{A176})$$

□



The Jacobian  $J^m$  is symmetric, diagonally dominant, and positive semi-definite. The weighted covariance  $\text{Cov}(\mathbf{U}, \cdot)$  is the covariance if the stored pattern  $\mathbf{u}_i$  is drawn according to an averaged  $p_i$  given by  $J^m(\cdot)$ . Analog for weighted covariance  $\text{Cov}(\mathbf{V}, \cdot)$ . When maximizing the dot product  $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ , the normalized vectors  $\mathbf{x}_i$  and  $\mathbf{y}_i$  are encouraged to agree on drawing the patterns  $\mathbf{u}_i$  with the same probability  $p_i$  to generate similar weighted covariance matrices  $\text{Cov}(\mathbf{U}, \cdot)$ . If subsets of  $\mathbf{U}$  have a strong covariance structure, then it can be exploited to produce large weighted covariances and, in turn, large dot products of  $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ . Furthermore, for a large dot product  $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ ,  $\mathbf{x}_i$  and  $\mathbf{y}_i$  have to be similar to one another to extract the same direction from the covariance matrices. All considerations are analog for  $\mathbf{V}_{x_i}^T \mathbf{V}_{y_i}$ .

## A.2 Experiments

### A.2.1 Ablation studies

As detailed in the main part of the paper, CLOOB has two new main components compared to CLIP: (1) the modern Hopfield networks and (2) the InfoLOOB objective instead of the InfoNCE objective. To assess effects of the new major components of CLOOB, we performed ablation studies on the CC and YFCC datasets. The results are reported in Table A1 for models pre-trained on CC for 31 and 128 epochs and in Table A2 for models pre-trained on YFCC for 28 epochs.

Table A1: Influence of loss functions and Hopfield retrieval for models pre-trained on CC for 31 epochs (left) and 128 epochs (right, indicated by \*). Both InfoLOOB and InfoNCE with Hopfield decrease the performance compared to InfoNCE in most of the tasks. InfoLOOB with Hopfield has a strong synergistic effect and therefore considerably improves the performance in 5 out of 8 datasets (epoch 31) and 7 out of 8 datasets (epoch 128) compared to all other models.

Dataset	InfoNCE	InfoLOOB	InfoNCE Hopfield	InfoLOOB Hopfield	InfoNCE*	InfoLOOB*	InfoNCE Hopfield*	InfoLOOB Hopfield*
Birdsnap	<b>2.58</b>	2.37	1.67	2.53	2.15	1.89	2.15	<b>3.39</b>
Country211	0.53	0.63	0.54	<b>0.76</b>	0.62	0.62	0.66	<b>0.79</b>
Flowers102	13.16	13.03	11.53	<b>14.24</b>	11.79	11.57	10.86	<b>14.24</b>
GTSRB	4.47	4.39	5.76	<b>5.86</b>	<b>9.25</b>	6.93	6.24	8.67
UCF101	<b>23.68</b>	19.14	20.56	22.29	21.33	20.56	21.40	<b>24.05</b>
Stanford Cars	<b>1.38</b>	1.33	1.24	1.37	1.26	1.19	1.24	<b>1.62</b>
ImageNet	21.74	22.13	19.04	<b>24.21</b>	22.80	22.69	20.29	<b>25.59</b>
ImageNetV2	21.45	21.65	18.97	<b>23.80</b>	22.44	22.13	20.22	<b>25.50</b>

Table A2: Influence of loss functions and Hopfield retrieval for models pre-trained on YFCC for 28 epochs. Both InfoLOOB and InfoNCE with Hopfield decrease the performance compared to InfoNCE in most of the tasks. InfoLOOB with Hopfield has a strong synergistic effect and therefore considerably improves the performance in 6 out of 8 datasets compared to all other models.

Dataset	InfoNCE	InfoLOOB	InfoNCE Hopfield	InfoLOOB Hopfield
Birdsnap	22.1	19.6	19.1	<b>28.9</b>
Country211	7.8	7.5	6.4	<b>7.9</b>
Flowers102	48.2	50.4	43.5	<b>55.1</b>
GTSRB	<b>8.9</b>	3.7	5.9	8.1
UCF101	<b>26.7</b>	24.0	25.4	25.3
Stanford Cars	3.1	2.4	2.8	<b>4.1</b>
ImageNet	34.0	32.2	30.4	<b>35.7</b>
ImageNetV2	32.8	30.9	29.4	<b>34.6</b>

For the ablation studies above we use a fixed inverse temperature parameter  $\tau^{-1}$  of 30 for all compared models. The value of  $\tau^{-1}$  was determined via hyperparameter search (see Section A.2.2).

In contrast to CLIP, we use a learning rate scheduler with restarts (Loshchilov & Hutter, 2017) to be more flexible regarding the number of total training epochs and enable training up to a plateau. To investigate the influence of the learning rate scheduler, we performed experiments with and without

restarts. Table A3 shows the zero-shot performance for the different downstream tasks for CLIP and CLOOB respectively. For both CLIP and CLOOB, the performance at the majority of the tasks either increases or remains roughly the same with restarts.

Table A3: Influence of learning rate scheduler. For most of the tasks the performance either increases or remains roughly the same with restarts for both CLIP and CLOOB.

Dataset	CLIP		CLOOB	
	w/o restarts	w/ restarts	w/o restarts	w/ restarts
Birdsnap	<b>2.10</b>	1.94	<b>2.64</b>	2.53
Country211	<b>0.71</b>	0.62	0.63	<b>0.76</b>
Flowers102	11.00	<b>13.04</b>	11.50	<b>14.24</b>
GTSRB	6.16	<b>7.28</b>	5.05	<b>5.86</b>
UCF101	19.05	<b>21.00</b>	21.97	<b>22.29</b>
Stanford Cars	<b>1.29</b>	0.90	1.22	<b>1.37</b>
ImageNet	20.19	<b>20.31</b>	23.29	<b>24.21</b>
ImageNet V2	20.53	<b>20.63</b>	22.97	<b>23.80</b>

### A.2.2 Hyperparameters

The hyperparameter search was done on a validation split of CC with about 15,000 samples. For the hyperparameter  $\tau^{-1}$  several values were considered (14.3, 30, 50, 70), where 30 leads to the best results for both YFCC and CC. Analogously to CLIP, we use the Adam optimizer (Kingma et al., 2014) with decoupled weight decay regularization (Loshchilov & Hutter, 2019). The weight decay is only applied to weights that are not gains or biases. As proposed in OpenCLIP (Ilharco et al., 2021) weight decay was set to 0.1. Different choices of weight decay (0.2 or 0.05), did not lead to a relevant performance change. We use the same learning rate of  $1 \times 10^{-3}$  for CC and  $5 \times 10^{-4}$  for YFCC as used in OpenCLIP. For the hyperparameter  $\beta$  we considered values in the range of 5 to 20. A value of 8 resulted in the best performance for CC and 14.3 for YFCC. The batch size for CC was reduced to 512 due to computational restraints which did not result in performance losses. The batch size for YFCC was kept at 1024 as reported by OpenCLIP since a reduction resulted in a significant drop in performance. The learning rate scheduler for all experiments is cosine annealing with warmup and hard restarts (Loshchilov & Hutter, 2017) with a cycle length of 7 epochs. For models trained on YFCC the warmup was set to 10000 steps and for models trained on CC to 20000 steps.

### A.2.3 Datasets

For pre-training we considered two datasets, Conceptual Captions (CC) (Sharma et al., 2018) and YFCC100M (Thomee et al., 2016). The CC dataset consists of 2.9 million images and corresponding high-quality captions. Images and their corresponding notations for CC have been gathered via an automated process from the web and therefore represent a wide variety of styles. Raw descriptions of images are collected from the *alt-text* HTML attribute. Both images and texts were filtered such that only image-text pairs above a certain quality threshold are part of this dataset. The dataset we refer to as YFCC is a subset of the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset. It was created by filtering for images which contain natural language descriptions and/or titles in English resulting in 15 million image-caption pairs. The textual descriptions contain less useful information than CC because they are not filtered by quality. Occasionally they also contain metadata like camera settings or web addresses.

We evaluate and compare our method on several downstream classification tasks. We evaluate on the same set of datasets as CLIP reported for a model trained on YFCC. This set contains Birdsnap (Berg et al., 2014), Country211 (Radford et al., 2021), Flowers102 (Nilsback & Zisserman, 2008), GTSRB (Stallkamp et al., 2011), UCF101 (Soomro et al., 2012), Stanford Cars (Krause et al., 2013) and ImageNet (Deng et al., 2009). We also include ImageNet V2 in our analysis (Recht et al., 2019). Additionally we added zero-shot results for Caltech101 (Fei-Fei et al., 2004), CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009), DTD (Cimpoi et al., 2014), Eurosat (Helber et al., 2018, 2019), FER2013 (Goodfellow et al., 2013), FGVC-Aircraft (Maji et al., 2013), Food101 (Bossard et al.,

Table A4: Datasets used for downstream evaluation. In the case of several train or test sets per dataset we report the total number of samples. It should be noted that at the time of this work some images from the Birdsnap dataset were not accessible anymore.

Dataset	Classes	Train size	Test size	Evaluation metric
Birdsnap	500	38,411	1,855	accuracy
Country211	211	42,200	21,100	accuracy
Flowers102	102	2,040	6,149	class-weighted accuracy
GTSRB	43	26,640	12,630	accuracy
ImageNet	1,000	1,281,167	50,000	accuracy
ImageNet V2	1,000	1,281,167	30,000	accuracy
Stanford Cars	196	8,144	8,041	accuracy
UCF101	101	28,747	11,213	accuracy
Caltech101	102	3,120	6,024	class-weighted accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
DTD	47	3,807	1,833	accuracy
Eurosat	10	10,000	5,000	accuracy
FER2013	7	28,709	7,178	accuracy
FGVC-Aircraft	100	10,000	3,333	class-weighted accuracy
Food101	101	75,750	25,250	accuracy
Pets	37	3,696	3,694	class-weighted accuracy
RESISC45	45	6,300	25,200	accuracy
STL10	10	1,000	8,000	accuracy
SUN397	397	72,763	35,991	accuracy

2014), Pets (Parkhi et al., 2012), RESISC45 (Cheng et al., 2017), STL10 (Coates et al., 2011) and SUN397 (Xiao et al., 2010).

Table A4 shows an overview of training and test set sizes, number of classes and the applied evaluation metric. In the case of several test sets per dataset the metric is calculated for every set individually and the average performance is reported. The set size in Table A4 corresponds to the total number of samples across all test and training sets of a dataset respectively.

**Birdsnap** contains images of North American bird species, however our dataset is smaller than reported in CLIP as some samples are no longer available. The **Country211** dataset was published in CLIP and is a small subset of the YFCC100m dataset. It consists of photos that can be assigned to 211 countries via GPS coordinates. For each country 200 photos are sampled for the training set and 100 for testing. For the **Flowers102** images of 102 flower categories commonly occurring in the United Kingdom were collected. Several classes are very similar and there is a large variation in scale, pose and lighting. The German Traffic Sign Recognition Benchmark (**GTSRB**) was a challenge held at the IJCNN 2011. The dataset contains images of german traffic signs from more than 40 classes. Note that two versions of this dataset exist, one used for the challenge and an official dataset released after the competition. For CLIP the linear probing classifiers were trained using the competition training set but tested on the official test set. **Stanford Cars** contains images of 196 car models at the level of make, model and year (e.g. Tesla Model S Sedan 2012). **UCF101** (Soomro et al., 2012) is a video dataset with short clips for action recognition consisting of three training sets and three test sets. We follow the procedure reported in CLIP and extract the middle frame of every video to assemble the dataset. The **ImageNet** Large Scale Visual Recognition Challenge was held from 2012 through 2017 and is one of the most widely used benchmarks for object detection and localization. Several years later **ImageNet V2** assembled three new test sets with images from the same 1,000 classes to test for generalization of models optimized for the original ImageNet benchmark. Every test set comprises 10,000 samples.

#### A.2.4 Zero-shot evaluation

Class names for all downstream tasks were adopted from CLIP, that is, among other changes special characters like hyphens or apostrophes were removed. Furthermore, some class names of the datasets

Table A5: Zero-shot results for models trained on CC with ResNet-50 vision encoders for two different checkpoints over 20 datasets. Results are given as mean accuracy over 5 runs. Statistically significant results are shown in bold. CLIP and CLOOB were trained for 31 epochs while CLIP\* and CLOOB\* were trained for 128 epochs.

Dataset	CLIP RN-50	CLOOB RN-50	CLIP* RN-50	CLOOB* RN-50
Birdsnap	2.26 ± 0.20	<b>3.06 ± 0.30</b>	2.8 ± 0.16	<b>3.24 ± 0.31</b>
Country211	0.67 ± 0.11	0.67 ± 0.05	0.7 ± 0.04	0.73 ± 0.05
Flowers102	12.56 ± 0.38	13.45 ± 1.19	13.32 ± 0.43	14.36 ± 1.17
GTSRB	7.66 ± 1.07	6.38 ± 2.11	8.96 ± 1.70	7.03 ± 1.22
UCF101	20.98 ± 1.55	22.26 ± 0.72	21.63 ± 0.65	<b>23.03 ± 0.85</b>
Stanford Cars	0.91 ± 0.10	<b>1.23 ± 0.10</b>	0.99 ± 0.16	<b>1.41 ± 0.32</b>
ImageNet	20.33 ± 0.28	<b>23.97 ± 0.15</b>	21.3 ± 0.42	<b>25.67 ± 0.22</b>
ImageNet V2	20.24 ± 0.50	<b>23.59 ± 0.15</b>	21.24 ± 0.22	<b>25.49 ± 0.11</b>
Caltech101	45.59 ± 0.44	<b>48.73 ± 0.94</b>	46.39 ± 1.58	<b>50.62 ± 0.84</b>
CIFAR10	<b>50.18 ± 1.52</b>	40.95 ± 2.24	<b>53.75 ± 1.49</b>	43.48 ± 2.84
CIFAR100	20.82 ± 1.45	21.59 ± 0.87	23.45 ± 1.99	24.41 ± 1.27
DTD	14.7 ± 1.32	<b>17.96 ± 2.04</b>	16.29 ± 1.30	16.51 ± 0.98
Eurosat	14.86 ± 5.98	21.47 ± 4.66	16.84 ± 2.28	19.56 ± 6.19
FER2013	<b>24.67 ± 1.34</b>	18.50 ± 1.74	22.70 ± 3.99	23.52 ± 2.73
FGVC-Aircraft	1.40 ± 0.27	1.31 ± 0.13	1.53 ± 0.19	1.30 ± 0.37
Food101	13.08 ± 0.36	<b>16.20 ± 0.38</b>	14.88 ± 0.51	<b>16.57 ± 0.39</b>
Pets	12.13 ± 1.87	12.93 ± 1.00	12.68 ± 0.86	13.45 ± 0.68
RESISC45	25.85 ± 2.01	<b>28.01 ± 1.02</b>	25.97 ± 1.56	<b>30.54 ± 1.21</b>
STL10	<b>82.97 ± 1.82</b>	79.11 ± 1.69	<b>84.02 ± 0.71</b>	82.28 ± 1.22
SUN397	38.96 ± 0.35	<b>42.29 ± 0.54</b>	39.86 ± 0.55	<b>44.15 ± 0.27</b>

were slightly changed (e.g. “kite” to “kite (bird of prey)” in ImageNet). For zero-shot evaluation, we use the same prompt as published in CLIP. Depending on the dataset the number of prompts can vary from one prompt (e.g. “a photo of a {label}, a type of bird.” for Birdsnap) up to 80 prompts for ImageNet covering various settings (e.g. “a cropped photo of a {label}.”, “a origami {label}.”). In case of several prompts an average embedding over all prompt embeddings is calculated. Figure A4 shows the zero-shot results for all evaluation tasks with the ResNet-50x4 model reported in Table 3.

In addition to the results of the main paper, the zeroshot performance of the models was tested on additional datasets. For details about the additional datasets we refer the reader to Section A.1 of Radford et al. (2021). Table A5 shows the results for models trained on CC. Table A6 shows the results for models trained on YFCC.

### A.2.5 Linear probing

We tried to follow the evaluation procedure in Radford et al. (2021) as closely as possible. We note one difference with respect to the implementation: Instead of scikit-learn’s logistic regression using the L-BFGS solver, we use cuML’s logistic regression classifier with L-BFGS algorithm to utilize GPUs for efficiency. All hyperparameters are the same as described in Radford et al. (2021), the maximum number of iterations was set to 1000, and the L2 regularization strength  $\lambda$  was determined by using a parametric binary search.

We tried to reproduce the CLIP results with the correspondingly published models, however, failed to produce the exact numbers. This could be due to several factors:

- The train and validation split. Same as in Radford et al. (2021), we use the provided validation set to perform the hyperparameter search. When there is none provided, we use a random half of the training dataset for validation.
- In case of a tie in the validation score, we use the maximal  $\lambda$  for the strongest regularization. We note though that we came closer to reproducing the results published in CLIP when using the mean  $\lambda$  over all ties when these exist.

Table A6: Zero-shot results for the CLIP reimplementation and CLOOB using different ResNet architectures trained on YFCC over 20 datasets.

Dataset	RN-50		RN-101		RN-50x4	
	CLIP	CLOOB	CLIP	CLOOB	CLIP	CLOOB
Birdsnap	21.8	<b>28.9</b>	22.6	<b>30.3</b>	20.8	<b>32.0</b>
Country211	6.9	<b>7.9</b>	7.8	<b>8.5</b>	8.1	<b>9.3</b>
Flowers102	48.0	<b>55.1</b>	48.0	<b>55.3</b>	50.1	<b>54.3</b>
GTSRB	7.9	<b>8.1</b>	7.4	<b>11.6</b>	9.4	<b>11.8</b>
UCF101	<b>27.2</b>	25.3	28.6	<b>28.8</b>	31.0	<b>31.9</b>
Stanford Cars	3.7	<b>4.1</b>	3.8	<b>5.5</b>	3.5	<b>6.1</b>
ImageNet	34.6	<b>35.7</b>	35.3	<b>37.1</b>	37.7	<b>39.0</b>
ImageNet V2	33.4	<b>34.6</b>	34.1	<b>35.6</b>	35.9	<b>37.3</b>
Caltech101	<b>55.8</b>	53.5	<b>57.7</b>	56.4	57.8	<b>58.7</b>
CIFAR10	<b>44.3</b>	42.4	<b>53.9</b>	51.4	<b>57.0</b>	47.4
CIFAR100	<b>21.9</b>	18.8	22.8	<b>23.1</b>	<b>23.1</b>	21.9
DTD	19.6	<b>20.3</b>	<b>22.5</b>	18.1	<b>22.4</b>	21.3
Eurosat	25.0	<b>25.9</b>	<b>24.9</b>	23.0	22.1	<b>28.5</b>
FER2013	<b>14.4</b>	11.0	<b>29.5</b>	17.2	<b>31.4</b>	16.3
FGVC-Aircraft	3.5	<b>5.6</b>	3.0	<b>5.8</b>	4.5	<b>6.4</b>
Food101	47.8	<b>50.5</b>	47.8	<b>54.4</b>	50.1	<b>57.9</b>
Pets	28.9	<b>30.4</b>	28.5	<b>31.4</b>	32.1	<b>32.6</b>
RESISC45	<b>23.2</b>	22.1	22.4	<b>22.9</b>	22.1	<b>26.6</b>
STL10	<b>85.8</b>	81.9	<b>88.2</b>	81.7	<b>88.9</b>	83.2
SUN397	46.2	<b>47.3</b>	47.7	<b>47.9</b>	47.6	<b>47.8</b>

- For the Birdsnap dataset, the resources that we have got online at the time of this writing could be different from the resources that CLIP’s authors obtained at the time.

Linear probing evaluation of YFCC pre-trained models is shown in Table A7. Comparing our reimplementation of CLIP and CLOOB with different ResNet encoders, we observe mixed results. The reason for this effect might be attributed to the observed task-dependence of multimodal models (Devillers et al., 2021). Another potential reason is that the benefit of the restrictions to more reliable patterns that occur in both modalities does not directly translate to an evaluation of just the encoding part of one modality. Again, as expected in self-supervised training, increasing the capacity of the CLOOB models benefits accuracy.

Table A7: Linear probing results for the reimplementation of CLIP and CLOOB using different ResNet architectures trained on YFCC for 28 epochs. The performance of CLOOB scales with increased encoder size.

Dataset	RN-50		RN-101		RN-50x4	
	CLIP	CLOOB	CLIP	CLOOB	CLIP	CLOOB
Birdsnap	50.9	<b>56.2</b>	51.6	<b>58.1</b>	57.6	<b>62.2</b>
Country211	19.5	<b>20.6</b>	20.8	<b>21.8</b>	22.5	<b>24.2</b>
Flowers102	94.8	<b>96.1</b>	94.5	<b>96.1</b>	95.1	<b>96.2</b>
GTSRB	<b>82.5</b>	78.9	<b>80.3</b>	77.9	<b>84.6</b>	80.6
UCF101	<b>75.2</b>	72.3	<b>76.0</b>	72.8	<b>77.3</b>	75.3
Stanford Cars	36.2	<b>37.7</b>	34.9	<b>39.0</b>	38.5	<b>44.3</b>
ImageNet	<b>66.9</b>	65.7	<b>67.9</b>	67.0	<b>70.0</b>	69.7
ImageNet V2	<b>60.2</b>	58.7	<b>61.0</b>	60.3	<b>62.8</b>	62.2

Table A8: Results for image-to-text and text-to-image retrieval on the CC validation set containing 13,330 samples. Results are given as mean accuracy over 5 runs. Statistically significant results are shown in bold. CLIP and CLOOB were trained for 31 epochs while CLIP\* and CLOOB\* were trained for 128 epochs.

Task	CLIP RN-50	CLOOB RN-50	CLIP* RN-50	CLOOB* RN-50
image-to-text R@1	0.297 ± 0.001	<b>0.319 ± 0.002</b>	0.316 ± 0.002	<b>0.342 ± 0.002</b>
image-to-text R@5	0.540 ± 0.003	<b>0.557 ± 0.001</b>	0.563 ± 0.003	<b>0.586 ± 0.002</b>
image-to-text R@10	0.638 ± 0.003	<b>0.651 ± 0.002</b>	0.660 ± 0.002	<b>0.678 ± 0.001</b>
text-to-image R@1	0.300 ± 0.003	<b>0.324 ± 0.001</b>	0.316 ± 0.002	<b>0.348 ± 0.001</b>
text-to-image R@5	0.542 ± 0.003	<b>0.566 ± 0.001</b>	0.565 ± 0.002	<b>0.593 ± 0.002</b>
text-to-image R@10	0.638 ± 0.002	<b>0.655 ± 0.001</b>	0.661 ± 0.001	<b>0.679 ± 0.001</b>

### A.2.6 Image-Text retrieval

In addition to zero-shot and linear probing, we tested the models trained on CC in image-to-text retrieval and text-to-image retrieval. The task is to find the matching image to a given text (image-to-text) or, respectively, finding the matching text to a given image (text-to-image). The dataset used for this task is the validation set of CC, which contains 13,330 image-text pairs. We report the results in Table A8. CLOOB significantly outperforms CLIP in both image-to-text and text-to-image retrieval.

### A.2.7 Analysis of the image and text embeddings

Following our ablation studies, we use models trained on CC to disentangle the effects of InfoLOOB and modern Hopfield networks. To track the behaviour during learning, we calculate the embeddings of the validation set of CC (consisting of 13,330 image-caption pairs) for all epochs before a restart of the learning rate scheduler.

We apply the extended uniformity test  $A_n$  of Ajne (Ajne, 1968; Prentice, 1978) to the respective embeddings of the image and text encoders. Let  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  be the embeddings of one modality (text or image), consisting of  $n$  samples of dimension  $d$ . The samples are normalized:  $\mathbf{x}_i = 1$ . As specified in Eq. (A177),  $A_n$  calculates the difference between a uniform distribution, where all samples on the hypersphere are orthogonal to each other, and the actual distribution of the embedding  $\mathbf{X}$ . Consequently, an embedding with low uniformity results in a high Ajne  $A_n$  test statistic:

$$A_n = \frac{n}{4} - \frac{1}{\pi n} \sum_{i=1}^n \sum_{j>i}^n \cos^{-1}(\mathbf{x}_j^T \mathbf{x}_i). \quad (\text{A177})$$

To understand the influence of modern Hopfield networks, we analyzed the covariance structure of the image and caption embeddings. Similar to Jing et al. (2022), we calculated the sorted eigenvalues of the covariance matrices of image embeddings. Figure A1 shows the sorted eigenvalues of InfoNCE and InfoLOOB with and without Hopfield retrievals during training. Both models without modern Hopfield networks struggle to increase the number of effective eigenvalues which contribute to the variance of the embeddings. InfoNCE with modern Hopfield networks starts with a small number of effective eigenvalues. We attribute this to the saturating effect of InfoNCE, which impedes the modern Hopfield network to extract more covariance. During training the effective eigenvalues steadily increase at a consistent rate. InfoLOOB with Hopfield starts with a high number of effective eigenvalues and strongly improves them during training.

Additionally, we looked at the distribution of similarity scores (dot product) of matched pairs and unmatched pairs over the validation set of CC of embeddings from models trained for 128 epochs. In Figure A2, we contrast the distributions of similarity scores of matched pairs with the distributions of the similarity score for the 1,000 unmatched pairs that have the highest similarity score with the anchor. InfoNCE with Hopfield results in a moderate increase of the similarity scores of matched pairs, as well as in an increase of the similarity score of unmatched pairs. The latter is an undesired side effect of Hopfield networks as unmatched pairs get also more similar to one another. Compared to InfoNCE, InfoLOOB does not saturate. Therefore, it considerably increases the similarity between



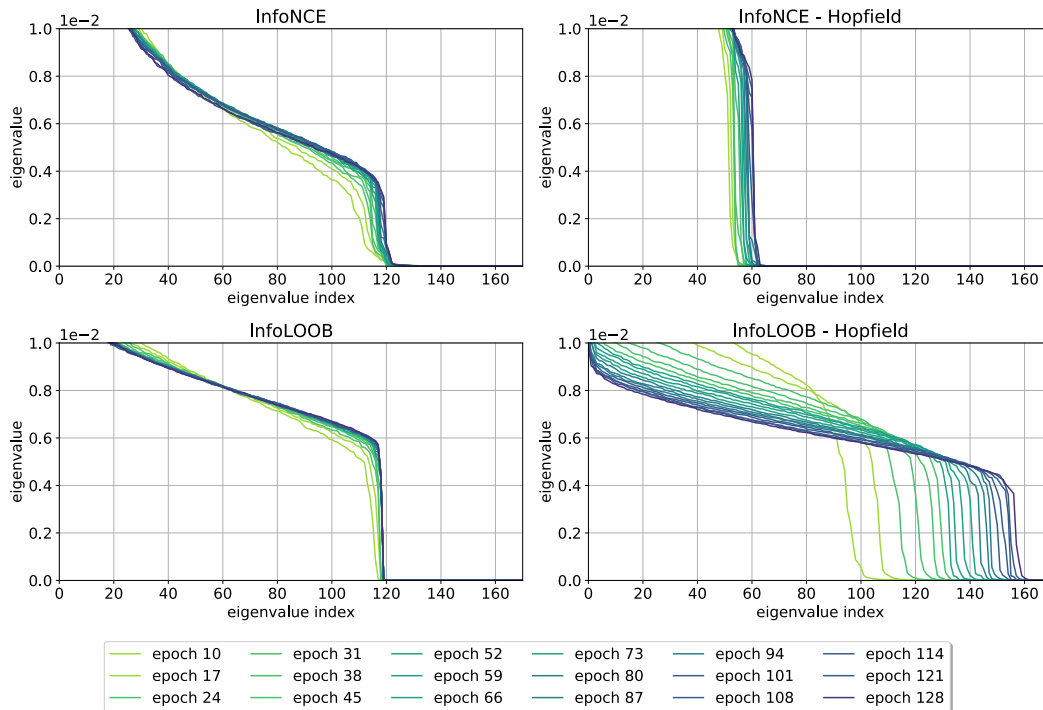


Figure A1: Sorted eigenvalues of InfoNCE and InfoLOOB with and without Hopfield retrievals during training.

matched pairs and also reduces the average similarity of the top-1000 unmatched pairs. InfoLOOB attributes a cosine similarity of one to many pairs, which is not plausible for multi-modal pairs. Clearly, this is an overfitting problem of InfoLOOB. Noteworthy, an observed increase in alignment between epoch 31 and epoch 128 does not benefit the downstream performance. The combination of Hopfield and InfoLOOB increases the similarity score of matched pairs and simultaneously reduces the average similarity of unmatched pairs compared to InfoNCE without Hopfield (the CLIP setting).

Figure A3 illustrates the overfitting problem of InfoLOOB. The distributions of unmatched pairs takes into account the ten unmatched pairs per anchor with the highest similarity score. In particular in the case of InfoLOOB without Hopfield, high similarity scores of the matched pairs correlate with high similarity scores of the top-10 unmatched pairs. In contrast, InfoLOOB with Hopfield does not suffer from this overfitting problem.

### A.2.8 Training time and memory consumption

To elaborate on the time and memory consumption of CLOOB, Table A9 compares the experiments done in Section 5.1. The memory consumption of CLOOB is the same as CLIP since only embeddings from the mini-batch are used both in the objective and in the Hopfield memories. The time consumption is approximately 5% higher, which is because of the retrieval of the modern Hopfield networks. The added complexity of modern Hopfield networks is  $\mathcal{O}(N)$  per sample, where  $N$  denotes the batch size.

Table A9: Memory and time consumption of CLIP and CLOOB when trained for 31 epochs on CC.

Model	Batch size (per GPU)	Memory (per GPU)	GPU hours	ImageNet zero-shot
CLIP	128	13.7GB	141	20.3
CLOOB	128	13.7GB	148	24.0

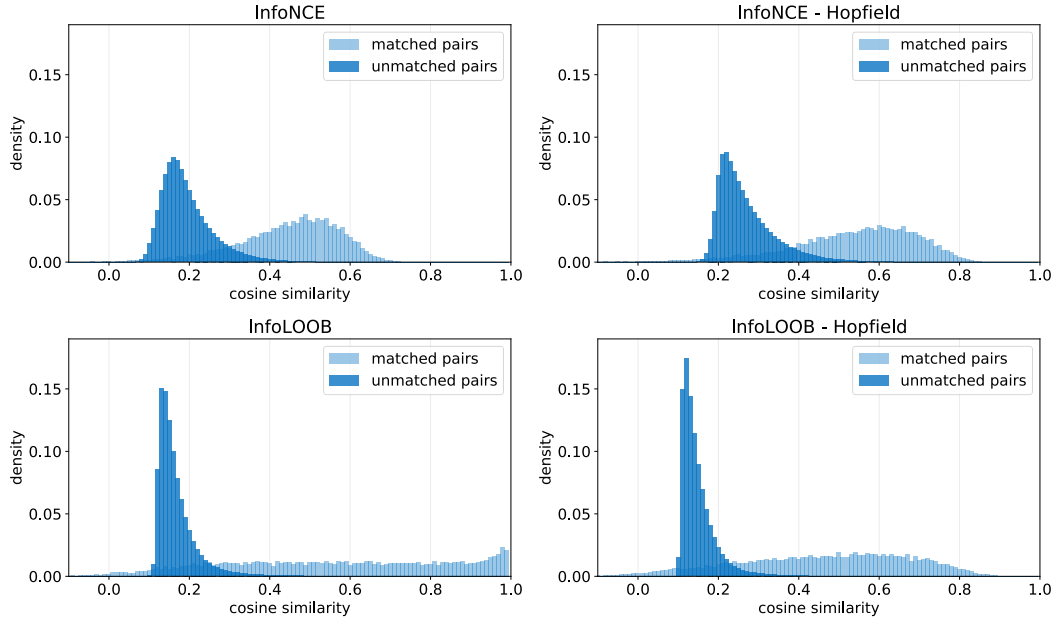


Figure A2: Distribution of the cosine similarity of matched pairs and the cosine similarity of the 1,000 unmatched pairs that have the highest similarity score with the anchor.

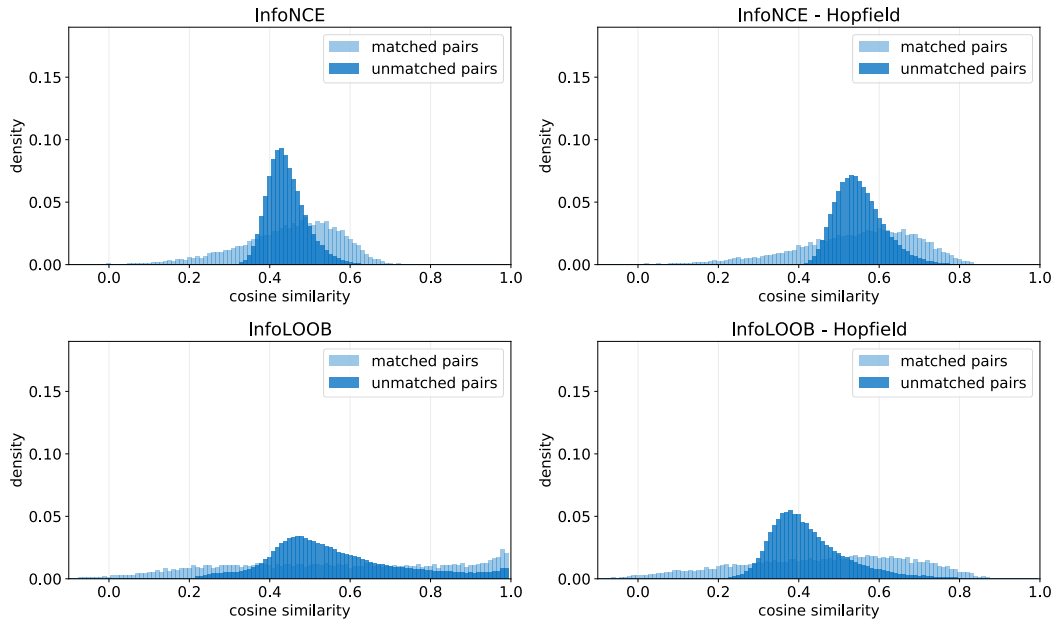


Figure A3: Distribution of the cosine similarity of matched pairs and the cosine similarity of the 10 unmatched pairs that have the highest similarity score with the anchor.

### A.3 Review of Modern Hopfield Networks

We briefly review continuous modern Hopfield networks that are used for deep learning architectures. They are continuous and differentiable, therefore they work with gradient descent in deep architectures. They retrieve with one update only, therefore they can be activated like other deep learning layers. They have exponential storage capacity, therefore they can tackle large problems. Hopfield networks are energy-based, binary associative memories, which popularized artificial neural networks in the 1980s (Hopfield, 1982, 1984). Associative memory networks have been designed to store and



retrieve samples. Their storage capacity can be considerably increased by polynomial terms in the energy function (Chen et al., 1986; Psaltis & Cheol, 1986; Baldi & Venkatesh, 1987; Gardner, 1987; Abbott & Arian, 1987; Horn & Usher, 1988; Caputo & Niemann, 2002; Krotov & Hopfield, 2016). In contrast to these binary memory networks, we use continuous associative memory networks with very high storage capacity. These modern Hopfield networks for deep learning architectures have an energy function with continuous states and can retrieve samples with only one update (Ramsauer et al., 2021). Modern Hopfield Networks have been successfully applied to immune repertoire classification (Widrich et al., 2020) and chemical reaction prediction (Seidl et al., 2021).

We assume a set of patterns  $\{\mathbf{u}_1, \dots, \mathbf{u}_N\} \subset \mathbb{R}^d$  that are stacked as columns to the matrix  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_N)$  and a state pattern (query)  $\boldsymbol{\xi} \in \mathbb{R}^d$  that represents the current state. The largest norm of a stored pattern is  $M = \max_i \|\mathbf{u}_i\|$ . Continuous modern Hopfield networks with state  $\boldsymbol{\xi}$  have the energy

$$E = -\beta^{-1} \log \left( \sum_{i=1}^N \exp(\beta \mathbf{u}_i^T \boldsymbol{\xi}) \right) + \beta^{-1} \log N + \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} + \frac{1}{2} M^2. \quad (\text{A178})$$

For energy  $E$  and state  $\boldsymbol{\xi}$ , the update rule

$$\boldsymbol{\xi}^{\text{new}} = f(\boldsymbol{\xi}; \mathbf{U}, \beta) = \mathbf{U} \mathbf{p} = \mathbf{U} \text{softmax}(\beta \mathbf{U}^T \boldsymbol{\xi}) \quad (\text{A179})$$

has been proven to converge globally to stationary points of the energy  $E$ , which are almost always local minima (Ramsauer et al., 2021). The update rule Eq. (A179) is also the formula of the well-known transformer attention mechanism (Ramsauer et al., 2021), therefore Hopfield retrieval and transformer attention coincide.

The *separation*  $\Delta_i$  of a pattern  $\mathbf{u}_i$  is defined as its minimal dot product difference to any of the other patterns:  $\Delta_i = \min_{j, j \neq i} (\mathbf{u}_i^T \mathbf{u}_i - \mathbf{u}_i^T \mathbf{u}_j)$ . A pattern is *well-separated* from the data if  $\Delta_i \geq \frac{2}{\beta N} + \frac{1}{\beta} \log(2(N-1)N\beta M^2)$ . If the patterns  $\mathbf{u}_i$  are well separated, the iterate Eq. (A179) converges to a fixed point close to a stored pattern. If some patterns are similar to one another and, therefore, not well separated, the update rule Eq. (A179) converges to a fixed point close to the mean of the similar patterns. This fixed point is a *metastable state* of the energy function and averages over similar patterns.

The next theorem states that the update rule Eq. (A179) typically converges after one update if the patterns are well separated. Furthermore, it states that the retrieval error is exponentially small in the separation  $\Delta_i$ .

**Theorem A4** (Modern Hopfield Networks: Retrieval with One Update). *With query  $\boldsymbol{\xi}$ , after one update the distance of the new point  $f(\boldsymbol{\xi})$  to the fixed point  $\mathbf{u}_i^*$  is exponentially small in the separation  $\Delta_i$ . The precise bounds using the Jacobian  $\mathbf{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$  and its value  $\mathbf{J}^m$  in the mean value theorem are:*

$$\|f(\boldsymbol{\xi}) - \mathbf{u}_i^*\| \leq \|\mathbf{J}^m\|_2 \|\boldsymbol{\xi} - \mathbf{u}_i^*\|, \quad (\text{A180})$$

$$\|\mathbf{J}^m\|_2 \leq 2\beta N M^2 (N-1) \exp(-\beta(\Delta_i - 2 \max\{\|\boldsymbol{\xi} - \mathbf{u}_i\|, \|\mathbf{u}_i^* - \mathbf{u}_i\|\} M)). \quad (\text{A181})$$

For given  $\epsilon$  and sufficient large  $\Delta_i$ , we have  $\|f(\boldsymbol{\xi}) - \mathbf{u}_i^*\| < \epsilon$ , that is, retrieval with one update. The retrieval error  $\|f(\boldsymbol{\xi}) - \mathbf{u}_i\|$  of pattern  $\mathbf{u}_i$  is bounded by

$$\|f(\boldsymbol{\xi}) - \mathbf{u}_i\| \leq 2(N-1) \exp(-\beta(\Delta_i - 2 \max\{\|\boldsymbol{\xi} - \mathbf{u}_i\|, \|\mathbf{u}_i^* - \mathbf{u}_i\|\} M)) M. \quad (\text{A182})$$

For a proof see (Ramsauer et al., 2021).

The main requirement of modern Hopfield networks to be suited for contrastive learning is that they can store and retrieve enough embeddings if the batch size is large. We want to store a potentially large set of embeddings. We first define what we mean by storing and retrieving patterns from a modern Hopfield network.

**Definition A1** (Pattern Stored and Retrieved). *We assume that around every pattern  $\mathbf{u}_i$  a sphere  $S_i$  is given. We say  $\mathbf{u}_i$  is stored if there is a single fixed point  $\mathbf{u}_i^* \in S_i$  to which all points  $\boldsymbol{\xi} \in S_i$  converge, and  $S_i \cap S_j = \emptyset$  for  $i \neq j$ . We say  $\mathbf{u}_i$  is retrieved for a given  $\epsilon$  if iteration (update rule) Eq. (A179) gives a point  $\tilde{\mathbf{x}}_i$  that is at least  $\epsilon$ -close to the single fixed point  $\mathbf{u}_i^* \in S_i$ . The retrieval error is  $\|\tilde{\mathbf{x}}_i - \mathbf{u}_i\|$ .*

As with classical Hopfield networks, we consider patterns on the sphere, i.e. patterns with a fixed norm. For randomly chosen patterns, the number of patterns that can be stored is exponential in the dimension  $d$  of the space of the patterns ( $\mathbf{u}_i \in \mathbb{R}^d$ ).

**Theorem A5** (Modern Hopfield Networks: Exponential Storage Capacity). *We assume a failure probability  $0 < p \leq 1$  and randomly chosen patterns on the sphere with radius  $M := K\sqrt{d-1}$ . We define  $a := \frac{2}{d-1}(1 + \ln(2\beta K^2 p(d-1)))$ ,  $b := \frac{2K^2\beta}{5}$ , and  $c := \frac{b}{W_0(\exp(a+\ln(b)))}$ , where  $W_0$  is the upper branch of the Lambert  $W$  function (Olver et al., 2010, (4.13)), and ensure  $c \geq \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}$ . Then with probability  $1 - p$ , the number of random patterns that can be stored is*

$$N \geq \sqrt{p} c^{\frac{d-1}{4}}. \quad (\text{A183})$$

Therefore it is proven for  $c \geq 3.1546$  with  $\beta = 1$ ,  $K = 3$ ,  $d = 20$  and  $p = 0.001$  ( $a + \ln(b) > 1.27$ ) and proven for  $c \geq 1.3718$  with  $\beta = 1$ ,  $K = 1$ ,  $d = 75$ , and  $p = 0.001$  ( $a + \ln(b) < -0.94$ ).

For a proof see (Ramsauer et al., 2021).

This theorem justifies to use continuous modern Hopfield networks for using retrieved embeddings instead of the original embeddings for large batch sizes. Even for hundreds of thousands of embeddings, the continuous modern Hopfield network is able to retrieve the embeddings if the dimension of the embeddings is large enough.

#### A.4 Further Related Work

With the advent of large corpora of unlabeled data in vision and language, self-supervised learning via contrastive learning has become highly successful. Some contrastive learning objectives, such as those of BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021), do not require negative samples. However, the most popular objective for contrastive learning is InfoNCE (van den Oord et al., 2018), in which for an anchor sample, a positive sample is contrasted with negative samples.

The idea to use objectives with negative samples is well known in deep learning (Gutmann & Hyvärinen, 2010; Chen et al., 2017; Mikolov et al., 2013). For contrastive learning, the most successful objective is InfoNCE, which has been introduced as Contrastive Predictive Coding (CPC) (van den Oord et al., 2018). InfoNCE has been applied to transfer learning (Hénaff et al., 2019), to natural language response suggestion (Henderson et al., 2017), to learning sentence representations from unlabelled data (Logeswaran & Lee, 2018), and to unsupervised feature learning by maximizing distinctions between instances (Wu et al., 2018). InfoNCE has been used for learning visual representations in Pretext-Invariant Representation Learning (PIRL) (Misra & vanDerMaaten, 2020), in Momentum Contrast (MoCo) (He et al., 2020), and in SimCLR (Chen et al., 2020). SimCLR became well known as it was highly effective for transfer learning. Zero-shot transfer learning (Lampert et al., 2009) is one of the most ambitious goals in vision, since it would improve various real-world downstream applications. Current models in natural language processing and vision perform very well on standard benchmarks, but they fail at new data, new applications, deployments in the wild, and stress tests (D’Amour et al., 2020; Recht et al., 2019; Taori et al., 2020; Lapuschkin et al., 2019; Geirhos et al., 2020). A model with high zero-shot transfer learning performance will not fail on such data, therefore will be trusted by practitioners.

Multiple works have proposed improvements to InfoNCE. Joint Contrastive Learning (JCL) studies the effect of sampling multiple positives for each anchor. (Cai et al., 2020). Sampling negatives around each positive leads to higher bias but lower variance than InfoNCE (Wu et al., 2021). InfoNCE has been generalized to C-InfoNCE and WeaC-InfoNCE, which are conditional contrastive learning approaches to remove undesirable information in self-supervised representations (Tsai et al., 2021). ProtoNCE is a generalized version of the InfoNCE, which pushes representations to be closer to their assigned prototypes (Li et al., 2021). ProtoNCE combines contrastive learning with clustering. SimCSE employs InfoNCE for contrastive learning to learn sentence embeddings (Gao et al., 2021). InfoNCE has been extended to video representation learning (Han et al., 2020).

CLOOB uses InfoLOOB, which is an upper bound on the mutual information. An alternative upper bound on the mutual information would be Contrastive Log-ratio Upper Bound (CLUB), which was used for minimizing the mutual information (Cheng et al., 2020). So far CLUB was only used for minimizing the mutual information, except for the analysis in (Wang & Liu, 2021). In our

experiments, maximizing CLUB failed as confirmed in (Wang & Liu, 2021). The reason is that the embedding distribution is not uniform as required for successful contrastive learning (Wang & Isola, 2020; Wang & Liu, 2021).

Many follow up works have been based on the CLIP model. The CLIP model is used in Vision-and-Language tasks (Shen et al., 2021). The CLIP model guided generative models via an additional training objective (Bau et al., 2021; Galatolo et al., 2021; Frans et al., 2021) and improved clustering of latent representations (Pakhomov et al., 2021). It is used in studies of out of distribution performance (Devillers et al., 2021; Milbich et al., 2021; Miller et al., 2021), of fine-tuning robustness (Wortsman et al., 2021), of zero-shot prompts (Zhou et al., 2021) and of adversarial attacks to uncurated datasets (Carlini & Terzis, 2021). It stirred discussions about more holistic evaluation schemes in computer vision (Agarwal et al., 2021). Multiple methods utilize the CLIP model in a straightforward way to perform text-to-video retrieval (Fang et al., 2021; Luo et al., 2021; Narasimhan et al., 2021).

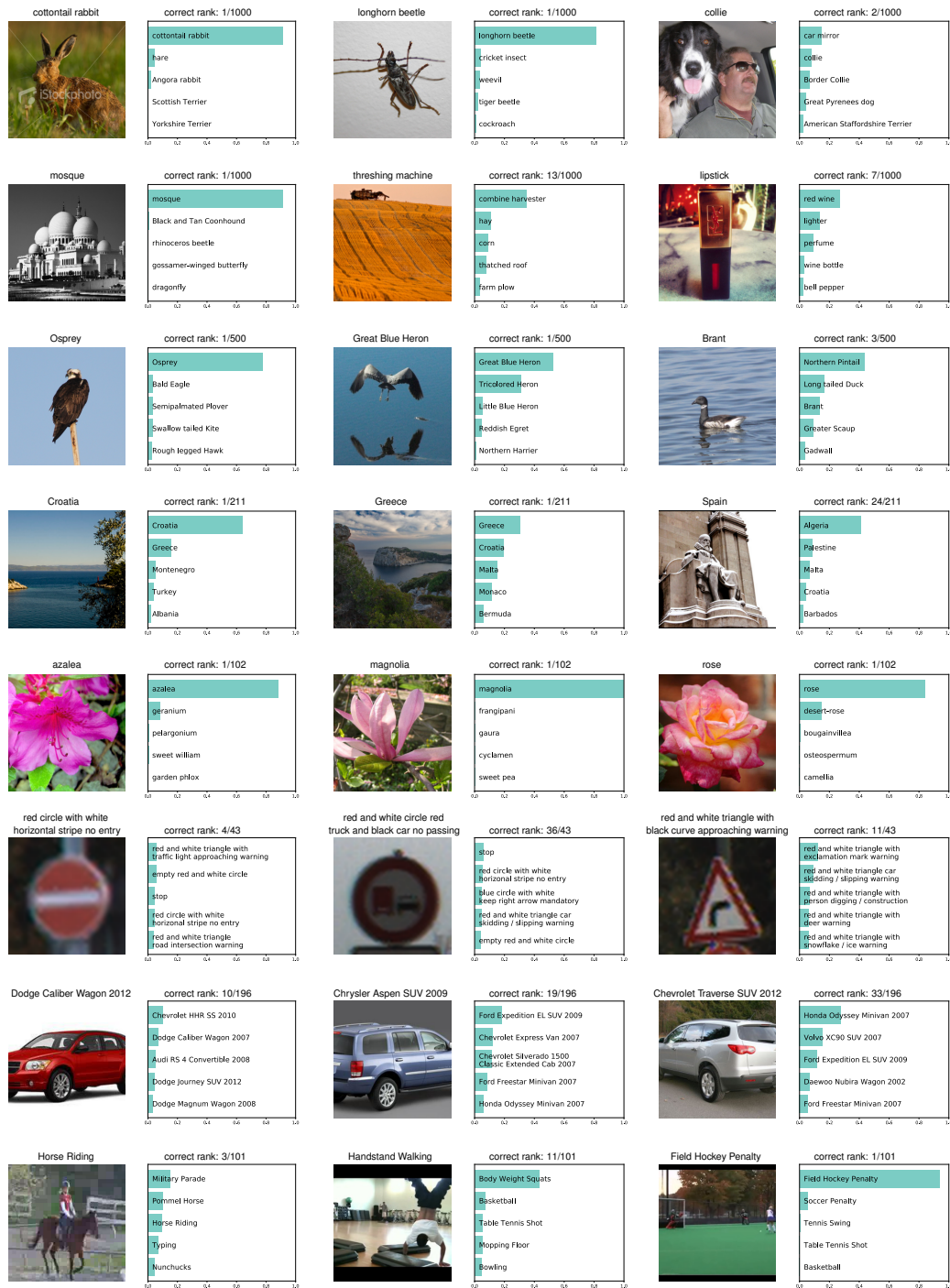


Figure A4: Visualization of zero-shot classification of three examples from each dataset. The following datasets are used (top to bottom): ImageNet, ImageNet V2, Birdsnap, Country211, Flowers102, GTSRB, Stanford Cars and UCF101. The ground truth label is displayed above the picture. The bar plots show the softmax values of the top 5 classes.