[Appendix]
Unsupervised Representation Learning via Neural Activation Coding

Yookoon Park [1] **Sangho Lee** [2] **Gunhee Kim** [2] **David M. Blei** [1]

1. Experimental Details for VAE Pretraining

We use diagonal Gaussians for both the variational posterior and the generative distribution of VAEs. For the encoder, we attach a linear output layer on ResNet-18 to predict the mean and the variances of a Gaussian distribution. The decoder takes a similar architecture with transposed convolution layers. We jointly train the encoder and the decoder on CIFAR-10 for 100 epochs with a batch size of 128. We use the Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and no weight decay. The global learning rate is set to 5e-4. When finetuning, we apply a smaller learning rate of 5e-5 to the pretrained ResNet encoder, while keeping the learning rate high for the linear output layer.

2. MI and Average Hamming Distance

We argue that maximizing the mutual information (MI) $I(\mathbf{X}, \widetilde{\mathbf{C}})$ over a noisy communication channel learns the codewords that have high Hamming distance to each other. We here show a relationship between the mutual information and the average Hamming distance between the codewords. Specifically, the mutual information *lower-bounds* the average Hamming distance. Recall that the Hamming distance between two codewords $\mathbf{c}_i, \mathbf{c}_j \in \{-1, 1\}^D$ is

$$d_H(\mathbf{c}_i, \mathbf{c}_j) = D - \frac{\mathbf{c}_i \cdot \mathbf{c}_j}{2}. \tag{1}$$

The average Hamming distance is defined as

$$\overline{d_H} = \frac{1}{N(N-1)} \sum_{j \neq i} d_H(\mathbf{c}_i, \mathbf{c}_j). \tag{2}$$

Let $\tilde{\mathbf{c}}$ be the noisy message transmitted through a binary symmetric channel with flip probability p. Then

$$\mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[\tilde{\mathbf{c}} \cdot \mathbf{c}_j] = (1-2p)(\mathbf{c}_i \cdot \mathbf{c}_j). \tag{3}$$

Finally,

$$\begin{aligned}
& I(\mathbf{X}, \widetilde{\mathbf{C}}) && (4)\\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}\left[\log \frac{\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_i)\frac{1}{2}\log\frac{1-p}{p})}{\frac{1}{N}\sum_{j=1}^{N}\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j)\frac{1}{2}\log\frac{1-p}{p})}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\Bigg(\mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[(\tilde{\mathbf{c}} \cdot \mathbf{c}_i)\frac{1}{2}\log\frac{1-p}{p}] && (5)\\
&\qquad - \mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[\log\frac{1}{N}\sum_{j=1}^{N}\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j)\frac{1}{2}\log\frac{1-p}{p})]\Bigg) \\
&= \frac{1}{N}\sum_{i=1}^{N}\Bigg(\frac{D(1-2p)}{2}\log\frac{1-p}{p} && (6)\\
&\qquad - \mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[\log\frac{1}{N}\sum_{j=1}^{N}\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j)\frac{1}{2}\log\frac{1-p}{p})]\Bigg) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left(\text{Const} - \mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[\frac{1}{N}\sum_{j=1}^{N}(\tilde{\mathbf{c}} \cdot \mathbf{c}_j)\frac{1}{2}\log\frac{1-p}{p}]\right) \\
&\qquad \text{(Jensen's inequality)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\text{Const} + \sum_{j \neq i}(1-2p)(\mathbf{c}_i \cdot \mathbf{c}_j))\frac{1}{2}\log\frac{1-p}{p}\right) \\
&= \frac{N-1}{N}(1-2p)\log(\frac{1-p}{p})\overline{d_H} + \text{Const}. && (7)
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{N}\sum_{i=1}^{N}\left(\text{Const} - \mathbb{E}_{\tilde{\mathbf{c}}|\mathbf{c}_i}[\frac{1}{N}\sum_{j=1}^{N}(\tilde{\mathbf{c}} \cdot \mathbf{c}_j)\frac{1}{2}\log\frac{1-p}{p}]\right) \\
&\qquad \text{(Jensen's inequality)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\text{Const} + \sum_{j \neq i}(1-2p)(\mathbf{c}_i \cdot \mathbf{c}_j))\frac{1}{2}\log\frac{1-p}{p}\right) \\
&= \frac{N-1}{N}(1-2p)\log(\frac{1-p}{p})\overline{d_H} + \text{Const}. && (8)
\end{aligned}$$

Therefore, maximizing the mutual information objective increases the average Hamming distance between the codewords.

[1]Computer Science Department, Columbia University, New York, USA [2]Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea. Correspondence to: David M. Blei <david.blei@columbia.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

References

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.