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) $I(\mathbf{X}, \mathbf{C})$ 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.

Finally,

\begin{equation}
I(\mathbf{X}, \tilde{\mathbf{C}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{e}|\mathbf{c}_i} \left[ \log \frac{\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_i) \frac{1}{2} \log \frac{1-p}{p})}{\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j) \frac{1}{2} \log \frac{1-p}{p})} \right] \tag{4}
\end{equation}

\begin{equation}
\begin{aligned}
&= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{e}|\mathbf{c}_i} \left[ \log \frac{\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_i) \frac{1}{2} \log \frac{1-p}{p})}{\exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j) \frac{1}{2} \log \frac{1-p}{p})} \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{e}|\mathbf{c}_i} \left[ \left( \log \sum_{i=1}^{N} \exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_i) \frac{1}{2} \log \frac{1-p}{p}) \right) - \log \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} \left( D(1-2p) \log \frac{1-p}{p} \right) \\
&\leq \frac{1}{N} \sum_{i=1}^{N} \left( \text{Const} - \mathbb{E}_{\mathbf{e}|\mathbf{c}_i} \left[ \log \sum_{i=1}^{N} \exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_i) \frac{1}{2} \log \frac{1-p}{p}) \right] \right) \\
&= \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 \left( \frac{1-p}{p} \right) d_H + \text{Const.} \tag{5}
\end{aligned}
\end{equation}

\begin{equation}
\begin{aligned}
&\leq \frac{1}{N} \sum_{i=1}^{N} \left( \text{Const} - \mathbb{E}_{\mathbf{e}|\mathbf{c}_i} \left[ \log \sum_{j=1}^{N} \exp((\tilde{\mathbf{c}} \cdot \mathbf{c}_j) \frac{1}{2} \log \frac{1-p}{p}) \right] \right) \\
&= \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 \left( \frac{1-p}{p} \right) d_H + \text{Const.} \tag{6}
\end{aligned}
\end{equation}

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

