Continuous Relaxation Training of Discrete Latent Variable Image Models

Casper Kaae Sønderby[*]
University of Copenhagen
casperkaae@gmail.com

Ben Poole[*]
Stanford University
poole@cs.stanford.edu

Andriy Mnih
DeepMind
amnih@google.com

Abstract

Despite recent improvements in training methodology, discrete latent variable models have failed to achieve the performance and popularity of their continuous counterparts. Here, we evaluate several approaches to training large-scale image models on CIFAR-10 using a probabilistic variant of the recently proposed Vector Quantized VAE architecture. We find that biased estimators such as continuous relaxations provide reliable methods for training these models while unbiased score-function-based estimators like VIMCO struggle in high-dimensional discrete spaces. Furthermore, we observe that the learned discrete codes lie on low-dimensional manifolds, indicating that discrete latent variables can learn to represent continuous latent quantities. Our findings show that continuous relaxation training of discrete latent variable models is a powerful method for learning representations that can flexibly capture both continuous and discrete aspects of natural data.

1 Introduction

Unsupervised learning of useful representations remains a key challenge in machine learning. Continuous latent variables models have made considerable progress, largely due to the ease of training with variational inference and the reparameterization trick (Kingma & Welling, 2013; Rezende et al., 2014). However, datasets are naturally modelled as discrete variables or mixtures of discrete and continuous variables where the reparameterization trick is not directly applicable. Furthermore, recent findings suggest that discrete categorical distributions can effectively capture the properties of continuous variables (see e.g. PixelRNN Oord et al. (2016)), and are more flexible at modelling multimodal distributions than simple continuous distributions such as Gaussians. Still, the usage of discrete variables has largely been limited to low-dimensional class labels in semi-supervised models (Kingma et al., 2014) or model classes tailored specifically to discrete variables (Bornschein et al., 2016; Rolfe, 2016; Bornschein et al., 2017). Here we study whether we can train large probabilistic discrete latent variable models and what representations they learn.

Recently, two types of gradient estimators have shown promising results for learning discrete latent variable models: 1) score-function gradient estimators such as VIMCO (Mnih & Rezende, 2016) and NVIL (Mnih & Gregor, 2014) and 2) continuous relaxation approaches that approximate discrete distributions with a continuous counterpart (Jang et al., 2016; Maddison et al., 2016). While these estimators work well in small models, it remains unclear whether they can be utilized for large generative models of images. For larger-scale models, the recently proposed Vector Quantized VAE (VQ-VAE) (Oord et al., 2017) model has shown success by combining a learned codebook used for deterministic nearest-neighbour vector quantization with a novel learning algorithm based on a straight-through gradient estimator and an additional regularizer. This model has achieved impressive generative performance on images, comparable to similar continuous latent variable models. However,

[*]Equal contribution. Work done during an internship at DeepMind.

Bayesian Deeplearning Workshop @ NIPS2017

it is not clear whether the gain in performance is primarily due to the specific parameterization of the nearest-neighbour vector quantization or to the specific algorithm used to learn the codebook vectors.

In this work we develop a discrete generative model using a similar codebook lookup as in the VQ-VAE model, but with probabilistic sampling from a categorical distribution instead of deterministic nearest neighbour selection. We evaluate several gradient estimators for training these models, and find that the probabilistic model trained with continuous relaxation using Gumbel-Softmax (Jang et al., 2016; Maddison et al., 2016) achieves a higher variational lower bound (ELBO) than the deterministic VQ-VAE on the CIFAR-10 dataset. Furthermore, we show that probabilistic discrete latent variable models can be trained with both large numbers of discrete latent variables and a large number of categories per latent variable. Finally, we analyze the learned discrete representations and find that they flexibly learn to represent both discrete and continuous structures.

2 Theory and Method

To compare deterministic and probabilistic models we interpret all models as optimizing the ELBO:

$$\log p(x) \geq \mathcal{L}(x) = \mathbb{E}_{z\sim q(z|x)}[\log p(x|z)] - \mathrm{KL}[q(z|x)|p(z)], \tag{1}$$

with a factorial prior $p(z) = \mathrm{Cat}(\theta)$ and observation model $p(x|z) = \mathrm{Cat}(f_\phi(m_z))$ with trainable parameters θ and ϕ respectively. The parameterization of the inference model $q(z|x)$ differs between the VQ-VAE and the probabilistic discrete models as described below.

The VQ-VAE is a deterministic autoencoder with a discrete latent space.The encoder produces a continuous vector representation, $v = f_{enc}(x) \in R^D$ which is then compared to each row in a codebook matrix $M \in \mathcal{R}^{C\times D}$ using Euclidean distance. The nearest-neighbour vector $m_i,\ i = \mathrm{argmin}_k(|m_k - v|_2), k = 1 \dots C$ is then used for reconstructing the data $\hat{x} = f_{dec}(m_i)$. The model learns to vector-quantize v using the codebook M, and the index i of the nearest neighbour can be interpreted as the discrete latent representation. Gradients are passed between the encoder and decoder using the straight through-estimator and the codebook matrix M is updated using a rule similar to k-means where the each m_k is moved towards the centroid of the v's assigned to it. The encoder can be interpreted as a deterministic inference model with $q(z|x) = \delta(i)$, thus the KL-divergence in Eq. 1 is the same fixed constant for all inputs.

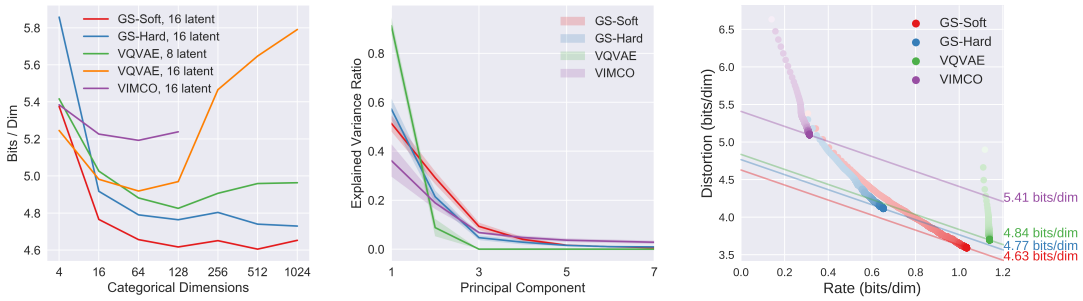
For the probabilistic models we let $q(z|x) = \mathrm{Cat}(f(x, M))$, where $f(x, M)$ calculates the logits for the categorical distribution using the distance between v and the rows of M. This model is a discrete VAE with a specific parameterization and *parameter sharing between the encoder and decoder*. If we interpret the codebook as a trainable memory and the encoder output v as memory query the model becomes reminiscent of generative models with memory (e.g. Bornschein et al. (2017)). We train the probabilistic models using either 1) VIMCO with 4 samples from $q(z|x)$ or 2) Gumbel-Softmax relaxation with a temperature of 0.5, a single sample from $q(z|x)$ and propagating either hard (discrete, denoted GS-Hard) or soft (continuous, denoted GS-Soft) samples forward during training . Note when evaluating we always use hard discrete samples. The encoders and decoders are fully convolutional and follow the settings in Oord et al. (2017): The encoder uses two strided convolutions (4x down-sampling) followed by two layers of two residual blocks. The decoder structure is similar but reversed, replacing the strided convolutions with transposed convolutions. We use 4, 8, or 16 latent variables per spatial position in the latent space and 256 feature maps in all other convolutions. All the experiments were performed on the CIFAR-10 dataset, using a minibatch size of 64 and the Adam optimizer with a learning rate of 5×10^{-4}.

3 Results and Conclusion

In Figure 1(a) we show bits/dim for the different learning algorithms as a function of the number of vectors C in the codebook. Across model classes the Gumbel-Softmax gradient estimator using soft samples (GS-Soft) achieve the best performance of 4.61 bits/dim using 512 categorical dimension. Importantly, we find that the probabilistic models (Fig 1(a), red and blue) are able to flexibly adapt the model capacity (KL term in ELBO) as the performance does not degrade as the categorical distributions C is increased beyond 128. For the VQ-VAE model (Fig 1(a), green and orange), the performance is maximized at 4.81 bits/dim for a specific value of C, since the deterministic inference model does not allow these models to adapt the latent capacity as C is increased. We note that

Oord et al. (2017) report a somewhat better VQ-VAE result of 4.67 bits/dim. Models trained with VIMCO and GS-Soft perform similarly for small categorical dimensions, but VIMCO struggles to take advantage of the additional capacity of larger dimensions. As a point of reference, Gregor et al. (2016) report 4.54 bits/dim for a comparable convolutional VAE model with continuous latent variables, though we note that better performance can be achieved with the current state-of-the-art autoregressive models (Oord et al., 2016; Salimans et al., 2017) and hierarchical models (Kingma et al., 2016). Visualizing the rate-distortion trade-off ($KL(q(z|x)\|p(z))$ vs. $\log p(x|z)$) reveals very different learning dynamics for the probabilistic and deterministic models (Figure 1(c)). The VQ-VAE model operates at a fixed capacity throughout training and learns to reduce distortion only — the small changes in KL are due to the prior adapting to the empirical distribution of the code usage. In contrast, the probabilistic models slowly introduce capacity into the model to decrease distortion, yielding models that operate at both lower distortion and lower rate for the models trained with GS-Soft.

Finally, we analyzed the learned discrete representations in the codebook matrix M. Figure 1(b) shows the fraction of explained variance as a function of the number of principal components of M for models with 64 categorical dimensions and 16 latent variables per spatial position. For all models, the learned codebook matrices are low-dimensional, spanning fewer than 5 dimensions. To further understand the structure of the low-dimensional codes we visualized their projection onto the first 3 principal components (Figure 2). We find that the codes typically tile a 1- or 2-D continuous manifold, indicating that the discrete latent variables are learning to represent a low-dimensional continuous signal. In conclusion, our results show that when trained with the Gumbel-Softmax/Concrete relaxation discrete latent variable models can perform on par with continuous latent variable models, and capture both discrete and continuous aspects of natural data.



(a) (b) (c)

Figure 1: Comparison of Gumbel-Softmax (GS), VIMCO and VQ-VAE models. a) ELBO in bits/dim as a function of the dimension C of the codebook M. b) Explained variance for the principal components of M c) Rate-distortion curves and learning progression.

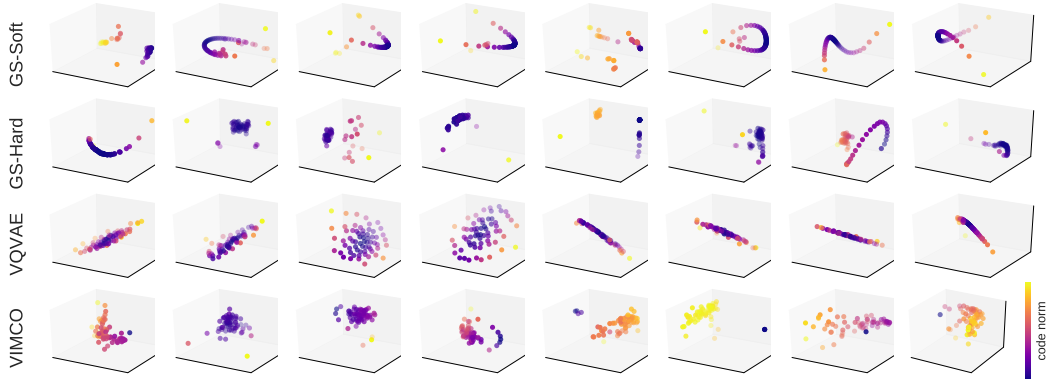


Figure 2: Visualization of codebooks. Columns are different categorical latent variables, and rows are different training methods. Points within each plot are the location of a learned code (row of M) projected onto the first 3 prinicpal components of M. Colors show the norm of each code (blue small, yellow large).

References

Jorg Bornschein, Samira Shabanian, Asja Fischer, and Yoshua Bengio. Bidirectional helmholtz machines. In *International Conference on Machine Learning*, pp. 2511–2519, 2016.

Jörg Bornschein, Andriy Mnih, Daniel Zoran, and Danilo J Rezende. Variational memory addressing in generative models. *Advances In Neural Information Processing Systems*, 2017.

Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. In *Advances In Neural Information Processing Systems*, pp. 3549–3557, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *International Conference on Machine Learning*, 2014.

Andriy Mnih and Danilo Rezende. Variational inference for monte carlo objectives. *International Conference on Machine Learning*, 2016.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Aaron van den Oord, Koray Kavukcuoglu, and Oriol Vinyals. Neural discrete representation learning. *Advances In Neural Information Processing Systems*, 2017.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

Jason Tyler Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.