Is the Discrete VAE's Power Stuck in its Prior?

Anonymous Author(s)
Affiliation
Address
email

Abstract

We explore the factors that make probabilistic neural models with discrete latent representations effective at generating high-quality samples by jointly training a discrete model inspired by the Vector Quantized Variational AutoEncoder (VQ-VAE). While we hypothesize that joint training should ultimately improve the distribution fit by the model, we find that modifying the VQ-VAE to estimate a true variational posterior only slightly improves likelihood of held-out data, while having no distinguishable impact on the quality of images sampled from the model.

1 Introduction

While generative adversarial networks (GANs) excel at generating realistic-looking natural images [Karras et al., 2020, Park et al., 2019], probabilistic models including variational autoencoders (VAEs) [Kingma and Welling, 2014, Higgins et al., 2017] and normalizing flows [Dinh et al., 2015, Kingma and Dhariwal, 2018, Ho et al., 2019] generally lag behind as measured by the qualitative assessment of images sampled from the models. However, GANs do not offer many attractive features that probabilistic models naturally provide, such as efficient encoding of inputs into a latent space representation and likelihood estimation for held-out data. Indeed, while probabilistic models must posit some probability mass over every possible image, there is no constraint in the GAN architecture that ensures that all images correspond to a representation in latent space; due to mode collapse, the typical samples from a GAN's regularly under-represent the training data distribution.

Vector Quantized Variational AutoEncoders (VQ-VAEs) introduce a discrete latent space, and as a result they produce clearer image reconstructions than non-quantized VAEs [van den Oord et al., 2017]. While the samples from a VQ-VAE with a simple PixelCNN prior trained on ImageNet appear high quality, upon close inspection it becomes apparent that the model is only matching general textures (for example, most of the animals are missing heads). We study image sampling with human faces in this paper, because we believe the quality of samples is more immediately apparent to a human evaluator [Seyama and Nagayama, 2007], and it is commonly accepted that likelihood of held-out data is not a good metric for sample quality [Theis et al., 2016].

Recent extensions of the VQ-VAE method show that pairing a VQ-VAE with a hierarchical, autoregressive prior distribution can efficiently model diverse, high resolution images well, and even produce samples that compete with state of the art GANs as measured by Inception Score and Fréchet Inception Distance [Razavi et al., 2019]. However, this model is trained in two steps–the encoder-decoder are trained as if training a standard autoencoder, and then a hierarchical PixelSNAIL prior [Chen et al., 2018] is fit to model the discrete encodings. This approach seems to offload much of the hard problem to the prior distribution, while the encoder learns to predict a point estimate rather than a full variational posterior distribution. This also means that it is not possible to use the encoder distribution to accurately measure the probability of held-out data [Burda et al., 2016].

In this work, we investigate the extent to which the sample quality from a VQ-VAE depends on the model's discrete representation, and whether image quality can be improved without increasing the

Submitted to 1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020). Do not distribute.

power of a marginally trained prior distribution. Although PixelCNN models are highly flexible, the training procedure in the VQ-VAE only emits one-hot vectors that the prior needs to learn to extrapolate. The decoder is never fine-tuned with samples from the prior. We hypothesized that jointly training the encoder/decoder with the prior distribution would allow the model to 1) fit the prior to more closely match a fully-factored variational posterior and 2) train the encoder/decoder to emit images on the natural image manifold for a wider range of samples from the prior.

We jointly train a VAE, inspired by the VQ-VAE, by posing a fully-factored categorical distribution, rather than a deterministic quantization, at the bottleneck. Our model, dubbed the Categorical VQ-VAE, preserves the discrete latent space while allowing joint training of the model from end-to-end (encoder, decoder, and prior).

Surprisingly, we find that "smoothing" the latent space through sampling and jointly training the prior only slightly improves the likelihood of held-out images, and neither of these supposed improvements improve the quality of samples.

2 Background

Our goal is to learn a latent space where every point emitted by the prior maps to a realistic-looking image. The following sections describe closely-related models, as well as the building blocks of our model. We denote the encoder of the (Categorical) VQ-VAE $z_e(x)$ and the posterior distribution it parameterizes as $q(z = k|x)$. The probabilistic decoder of the (Categorical) VQ-VAE is denoted as $p(x|z)$ and the learned prior distribution over the discrete latent space is written as $p(z)$.

2.1 Related Models

Sønderby et al. [2017] develop the most closely-related model to ours in the literature. Rather than jointly training a flexible prior, however, this work assumes that the prior distribution is uniform. The authors investigate VQ-VAE performance as the latent space becomes higher dimensional. The work does not explore the quality of images sampled from the model. The modifications explained in Section 3.3 that enable sampling from the Quantized VQ-VAE posterior are analogous to the GS-Hard model described in the paper.

Kaiser et al. [2018] demonstrates that VQ-VAE model fit degrades as the length of quantization vectors grows, a phenomenon they call "index collapse." Rather than introducing sampling to avoid the problem of index collapse, the authors develop a model to quantize sub-vectors of the encoder's output using multiple codebooks.

Several recent works have tried to improve the quality of samples by increasing the expressiveness of the model. In particular, Rezende and Mohamed [2015] replaces a fully-factored posterior distribution in the encoder with a normalizing flow, and Gulrajani et al. [2017] uses a powerful autoregressive model as the decoder. Both approaches are complementary to our model's joint training of an expressive autoregressive prior.

2.2 PixelCNN++

Salimans et al. [2017] introduced the PixelCNN++, inspired by the PixelCNN architecture from van der Oord et al. [2016]. Both the PixelCNN and PixelCNN++ are autoregressive convolutional neural networks that output a distribution over the input space where each location is conditioned exclusively on the previous spatial values of the input. This spatial conditioning is performed through masked convolutions or shifted convolutions that prevent the model from seeing any spatially future values throughout the network.

2.3 VQ-VAE

van den Oord et al. [2017] introduced the VQ-VAE, a generative model that uses set of discrete latent representations rather than continuous ones. It is comprised of an encoder $z_e(x)$ that parameterizes a deterministic posterior categorical distribution $q(z = k|x)$, given by

$$q(z = k|x) = \begin{cases} 1 & \text{for } k = \arg\min_j ||z_e(x) - e_j||_2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $z_e(x)$ is the encoder's approximation of a latent embedding and e_j is the jth latent embedding. The encoder's approximation of latent embeddings are then discretized to the closest embedding:

$$z_q(x) = e_k, \text{ where } k = \arg\min_j ||z_e(x) - e_j||_2 \tag{2}$$

These discretized representations are then passed to the decoder network which parameterizes a distribution $p(x|z)$ over the input data. Because Equation 2 is not differentiable, a straight-through gradient estimation is used.

After the training of the VQ-VAE encoder and decoder is completed, an autoregressive prior (such as a PixelCNN), is trained over the latent embeddings allowing one to sample from the from the latent space and then decode these samples back into input space.

3 Method

In this section we introduce the modifications of the original VQ-VAE. By reformulating the posterior distribution, incorporating the Gumbel-Softmax distribution and temperature annealing, the Categorical VQ-VAE offers a discretized Variational AutoEncoder architecture that allows training of the encoder, decoder, and prior simultaneously. We view the training of the prior through the lens of "empirical Bayes" and fit the model jointly by allowing gradients to flow between the encoder, decoder, and prior.

3.1 Categorical VQ-VAE posterior reformulation

As the deterministic posterior distribution $q(z = k|x)$ shown in Equation 1 does not readily admit sampling, we reformulate the VQ-VAE encoder to model a nondeterministic categorical posterior distribution directly. Thus the encoder of the Categorical VQ-VAE outputs logits that parameterize the probability of using using each latent embedding at each point in the bottleneck.

3.1.1 Gumbel softmax

Jang et al. [2017] introduced a reparameterization technique, similar to Gaussian reparameterization, that allows one to approximate sampling from a categorical distribution in a differentiable manner. This is done by adding i.i.d. samples, g_i, from Gumbel$(0, 1)$ to the logits of a categorical distribution and taking the softmax of the resulting values:

$$y_i = \frac{\exp((\text{logit}_i + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\text{logit}_j + g_j)/\tau)}$$

As the softmax temperature τ approaches 0, these samples approach one-hot vectors representing the maximum index of the Gumbel samples added to the logits, thus offering a differentiable sampling method.

This sampling technique is used to draw samples from the posterior distribution of the Categorical VQ-VAE and allows us to forgo using a straight-through gradient approximation in the discretization step: discretization is accomplished by matrix-multiplying the categorical samples into an embedding matrix. In the beginning of training this results in a weighted mixture of embeddings being passed to the decoder because the softmax temperature is low and thus the samples may not be one-hots, but as the temperature is annealed the embeddings become more pure. In our experiments we found that annealing τ to a value of 0.1 over the course of training brought samples sufficiently close to one-hots for our purposes.

This softmax temperature annealing can also be performed without the addition of Gumbel samples to attempt replicating something similar to the original VQ-VAE deterministic posterior by only taking the highest probability embedding.

3.2 PixelCNN++

In all experiments the prior is a slightly modified and downscaled PixelCNN++ that forgoes the logistic mixture at output. We use 3 residual layers with 64 channels instead of the original 5 residual layers with 192 channels.

3.3 VQ-VAE posterior reformulation

To evaluate the similarities and differences of the Categorical VQ-VAE and the Quantized VQ-VAE we implemented an optional reformulation of the Quantized VQ-VAE posterior that admits sampling. Jang et al. [2017] discusses a straight-through Gumbel-Softmax estimator in which samples from the categorical distribution are always one-hot, so rather than taking the $\arg\min$ distance embedding, we use the distances as logits to parameterize the posterior distribution. Similar to the Categorical VQ-VAE discretization, these one-hot samples are matrix-multiplied into an embedding matrix, and because they are one-hots there is no weighted mixture of embeddings.

3.4 Marginal and joint training

In the typical VQ-VAE training regime, constructing a model that can generate samples from the latent space is done through marginal training: first the VQ-VAE is trained on the task of encoding and reconstructing a dataset, and then a prior is trained afterward on the discrete codes.

In the case of a Categorical VQ-VAE the encoder directly parameterizes a distribution over latent samples, so both marginal and joint training methods are possible. In the joint training case the objective function we minimize is the negation of ELBO:

$$-\text{ELBO} = -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}) + D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

where $q(z|x)$ is the distribution parameterized by the encoder, $p(x|z)$ is the decoder, and $p(z)$ is the prior. Thus the Categorical VQ-VAE and the prior distribution over latent samples may be trained simultaneously in a fully differentiable manner.

During the marginal training regime we split the ELBO objective apart and first train the the VQ-VAE variants to minimize

$$-\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}) \tag{3}$$

This is calculated by centering normal distributions $N(x, \sigma^2)$ on the input pixel values and we allow each model to optimize the parameter σ, following Lin et al. [2019]. Afterward, while holding the VQ-VAE variants weights fixed, we train the prior to minimize the KL-Divergence:

$$D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \tag{4}$$

As both the log-likelihood and KL-Divergence objectives are expectations over the latent space, we approximate them with a Monte-Carlo estimate.

4 Experiments

Here we present a qualitative and quantitative analysis of our findings when the Categorical VQ-VAE is tasked with modeling images of human faces from the Large-scale CelebFaces Attributes (CelebA) dataset [Liu et al., 2015]. In all models, we take one sample from the encoder distribution. All models have been trained using the Adam optimizer for 500 epochs with a batch size of 128 and a learning rate of 0.0003. All training experiments have been performed using the CelebA dataset down-scaled spatially to 64x64.

We use 4 residual layers with 128 feature maps in both the encoder and decoder of all VQ-VAE architectures. The latent space is comprised of 256 embeddings, and the PixelCNN++ parameters are described in Section 3.2.

4.1 Model options

As both the Categorical VQ-VAE and the Quantized VQ-VAE allow for optional Gumbel-Softmax sampling, there are are 6 straightforward model and training combinations available. However, despite Categorical VQ-VAE with no Gumbel-Softmax sampling being a training option, we find in our experiments that it is not suitable as a discretized Variational AutoEncoder. Even with annealing the softmax temperature to a small value, after the softmax the weights remain somewhat dense so the embeddings passed to the decoder are mixtures. Thus these configurations will be left out of our experiments and the remaining 4 configurations will be discussed.

Figure 1: Samples from the following models: 1st row: Quantized VQ-VAE, no posterior sampling. 2nd row: Quantized VQ-VAE, with posterior sampling. 3rd row: Categorical VQ-VAE, marginal training. 4th row: Categorical VQ-VAE, joint training. Based on evaluating a large number of such samples, the researchers concluded that neither sample quality nor diversity significantly change between models.

Table 1: Final model performances on held-out test images. Under the training column, "M" stands for marginal training and "J" represents joint training.

VQ-VAE Variant	Sample	Training	Bits/dim	KL-Divergence	σ
Quantized	No	M	5.295	864	0.0355
Quantized	Yes	M	5.632	**539**	0.0444
Categorical	Yes	M	5.393	702	0.0394
Categorical	Yes	J	**5.276**	801	0.0355

4.2 Quantitative evaluation

The results of all training combinations and model configurations are laid out in Table 1. In terms of log likelihood, all models performed similarly. However, the jointly trained Categorical VQ-VAE marginally outperformed all other model variants, with the original Quantized VQ-VAE architecture following quite closely behind. With respect to the prior, the Quantized VQ-VAE model with posterior sampling had the lowest KL-Divergence followed by the marginally trained Categorical VQ-VAE. In terms of ELBO the jointly trained Categorical VQ-VAE architecture out-performed the marginally trained Quantized VQ-VAE both with and without posterior sampling due to its slightly higher log likelihood.

In training the prior, encoder, and decoder jointly one may logically expect the prior performance to increase, perhaps even significantly. However, we find that is not the case as the joint model out-performed only the Quantized VQ-VAE with no posterior sampling. To investigate why this is the case, we took the converged marginally trained Categorical VQ-VAE and jointly trained it to see if the performance would improve. Jointly training the marginal Categorical model brought the KL-Divergence down from 702 to 674 and lowered the bits/dim from 5.393 to 5.353. Despite this, the marginally then jointly trained model still fell short of the categorically top performing models in all metrics.

4.3 Samples

Samples from all trained models are shown in Figure 4.3. It is clear that these samples are not of the quality of state of the art GANs, however, the VQ-VAE architecture in Razavi et al. [2019] that generates samples competing with GANs is multi-scale and hierarchical whereas ours is not. We believe this to be the reason for the comparatively poor samples.

When trained on CIFAR-10 and MNIST we find that these architectures consistently produce samples of high quality so we believe that the problem also gets harder with high resolution natural images.

5 Qualitative evaluation

Our experiments demonstrate that modifying the VQ-VAE model to be structured more like a traditional VAE—by introducing a more expressive posterior distribution, adding sampling, and jointly training—have practically no impact on the model fit or the quality of images sampled from the model. This finding indicates that the PixelCNN prior is flexible enough that it can learn practically any distribution fit by a discrete autoencoder. This finding is very similar to "posterior collapse," experienced in VAEs with expressive autoregressive decoders, where the decoder learns to ignore the posterior distribution entirely when generating samples Lucas et al. [2019].

Although joint training does not add significant flexibility to the model, learning an accurate posterior distribution is critical for a wide variety of inference tasks. Our finding that learning a full variational posterior does not outperform a simple, quantized point estimate raises the larger problem: do existing discrete neural models truly amortize posterior model inference for discrete latent variable models?

References

Y. Burda, R. B. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel. PixelSNAIL: An improved autoregressive generative model. volume 80 of *Proceedings of Machine Learning Research*, pages 864–872, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018.

L. Dinh, D. Krueger, and Y. Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taïga, F. Visin, D. Vázquez, and A. C. Courville. Pixelvae: A latent variable model for natural images. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. volume 97 of *Proceedings of Machine Learning Research*, pages 2722–2730, Long Beach, California, USA, 09–15 Jun 2019.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

L. Kaiser, S. Bengio, A. Roy, A. Vaswani, N. Parmar, J. Uszkoreit, and N. Shazeer. Fast decoding in sequence models using discrete latent variables. volume 80 of *Proceedings of Machine Learning Research*, pages 2390–2399, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, pages 10215–10224. 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

S. Lin, S. J. Roberts, N. Trigoni, and R. Clark. Balancing reconstruction quality and regularisation in ELBO for vaes. *CoRR*, abs/1909.03765, 2019.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Understanding posterior collapse in generative latent variable models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*, 2019.

T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems 32*, pages 14866–14876. 2019.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015.

T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

J. Seyama and R. S. Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, 16(4):337–351, 2007.

C. Sønderby, B. Poole, and A. Mnih. Continuous relaxation training of discrete latent variable image models. In *Bayesian Deeplearning Workshop @ NIPS2017*, 2017.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

A. van den Oord, O. Vinyals, and k. kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30*, pages 6306–6315. 2017.

A. van der Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756, New York, New York, USA, 20–22 Jun 2016.