TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

Vector Quantized Variational Autoencoders (VQ-VAEs)

Gaussian VAEs for Faces 2014

We can sample faces from the VAE by sampling noise z from $p_{\text{pri}}(z)$ and then sampling an image y from $p_{\text{dec}}(y|z)$.

[Alec Radford]

VQ-VAEs 2019

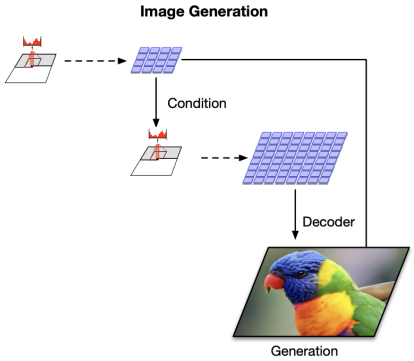
VQ-VAE-2, Razavi et al. June, 2019

VQ-VAEs 2019

Figure 1: Class-conditional 256x256 image samples from a two-level model trained on ImageNet.

VQ-VAE-2, Razavi et al. June, 2019

VQ-VAE-2 Model

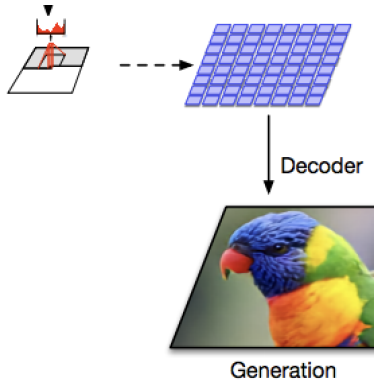


The probability of an image y is defined by the generator.

The generator is top-down and is similar to that of a progressive GAN.

VQ-VAE-2 Model

We describe the case of just one layer.



A separate unit describes progressive VAEs which are inspired by an analogy with progressive GANs. Progressive VAEs provide a speculative method of multi-layer VQ-VAE training.

The Encoder

Let s denote the image (we are using y for an image coordinate).

We have an encoder network, such as a CNN, which produces a layer $L_{\text{enc}}(s)[X, Y, I]$ with a vector at each pixel position.

Intuitively we cluster the vectors $L_{\text{enc}}(s)[x, y, I]$ using K-means clustering to produce cluster centers where $C[k, I]$ is the cluster center vector of cluster k.

The set of cluster centers $C[K, I]$ is a trained parameter of the encoder.

The Encoder

We then find the nearest cluster center to each vector $L_{\text{enc}}(s)[x, y, I]$.

$$k_{\text{enc}}(s)[x, y] = \underset{k}{\text{argmin}} \ ||L_{\text{enc}}(s)[x, y, I] - C[k, I]||$$

The "symbolic image" $k_{\text{enc}}(s)[X, Y]$ is the latent variable z.

The Decoder

To decode z we first construct a layer using the cluster center vectors at each x, y position.

$$\hat{L}_{\text{dec}}(z)[x, y, I] = C[k[x, y], I]$$

Finally we decode $\hat{L}_{\text{dec}}(z)[X, Y, I]$ to get an image $\hat{s}_{\text{dec}}(z)$.

Two-Phase Optimization

Phase 1: Train the encoder and the decoder to minimize reconstruction error.

$$\text{enc}^*, \text{dec}^* = \underset{\text{enc,dec}}{\text{argmin}} \; E_{s\sim\text{Pop},z\sim P_{\text{enc}}(z|s)} \left[-\ln P_{\text{dec}}(s|z)\right]$$

Phase 2: Train the prior holding the encoder and decoder fixed.

$$\text{pri}^* = \underset{\text{pri}}{\text{argmin}} \; E_{s\sim\text{Pop},z\sim P_{\text{enc}}(z|s)} \left[-\ln P_{\text{pri}}(z)\right]$$

Under the universality assumption this two phase method succeeds in modeling the popluation — for any encoder it suffices to optimize the decoder and the prior.

Image Autoencoding

The VQ-VAE for images uses

$$\text{enc}^*, \text{dec}^* = \underset{\text{enc,dec}}{\text{argmin}} \; E_{s \sim \text{Pop}} \left[||s - \hat{s}_{\text{dec}}(z_{\text{enc}}(s))||^2 \right]$$

Handling Discrete Latents

$$\text{enc}^*, \text{dec}^* = \underset{\text{enc,dec}}{\text{argmin}} \; E_{s \sim \text{Pop}} \left[||s - \hat{s}_{\text{dec}}(z_{\text{enc}}(s))||^2 \right]$$

Since $z = k_{\text{enc}}(s)[x, y]$ is discrete we have

$$k_{\text{enc}}(s)[x, y].\text{grad} = 0.$$

Handling Discrete Latents

$$\text{enc}^*, \text{dec}^* = \underset{\text{enc,dec}}{\text{argmin}} \; E_{s\sim\text{Pop}}\left[||s - \hat{s}_{\text{dec}}(z_{\text{enc}}(s))||^2\right]$$

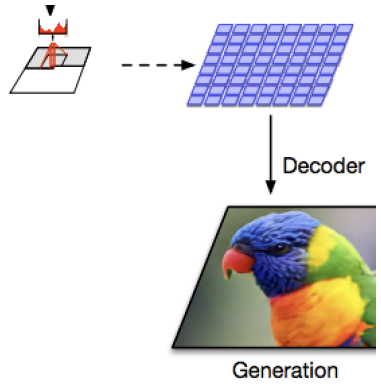
VQ-VAE uses "straight-through" gradients and "k-means" gradients.

$$L_{\text{enc}}(s)[x, y, I].\text{grad} = \hat{L}_{\text{dec}}(z)[x, y, I].\text{grad} + \beta(L_{\text{enc}}(s)[x, y, I] - C[k_{\text{enc}}(s)[x, y], I])$$

$$C[k, I].\text{grad} = \sum_{k_{\text{enc}}(s)[x,y]=k} \gamma(C[k, I] - L_{\text{enc}}(s)[x, y, I])$$

One Layer VQ-VAE Training Phase 2

Finally we hold the encoder fixed and train the prior $P_{\text{pri}}(z)$ to be an auto-regressive model of the symbolic image $k_{\text{enc}}(s)[X, Y]$.



Quantitative Evaluation

The VQ-VAE2 paper reports a classification accuracy score (CAS) for class-conditional image generation.

We generate image-class pairs from the generative model trained on the ImageNet training data.

We then train an image classifier from the generated pairs and measure its accuracy on the ImageNet test set.

	Top-1 Accuracy	Top-5 Accuracy
BigGAN deep	42.65	65.92
VQ-VAE	54.83	77.59
VQ-VAE after reconstructing	58.74	80.98
Real data	73.09	91.47

Image Compression

Figure 3: Reconstructions from a hierarchical VQ-VAE with three latent maps (top, middle, bottom). The rightmost image is the original. Each latent map adds extra detail to the reconstruction. These latent maps are approximately 3072x, 768x, 192x times smaller than the original image (respectively).

Rate-Distortion Evaluation.

Rate-distortion metrics for image compression to discrete representations support unambiguous rate-distortion evaluation.

Rate-distortion metrics also allow one to explore the rate-distortion trade-off.

	Train NLL	Validation NLL	Train MSE	Validation MSE
Top prior	3.40	3.41	-	-
Bottom prior	3.45	3.45	-	-
VQ Decoder	-	-	0.0047	0.0050

Table 1: Train and validation negative log-likelihood (NLL) for top and bottom prior measured by encoding train and validation set resp., as well as Mean Squared Error for train and validation set. The small difference in both NLL and MSE suggests that neither the prior network nor the VQ-VAE overfit.

DALL·E: A Text-Conditional Image dVAE

DALL·E is a text-conditional VQ-VAE model of images.

The Vector quantization is done independent of the text. However, the model of the probability distribution of the symbolic image $z[x, y]$ is conditioned on text.

Ramesh et al. 2021

DALL·E

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

END