TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

Vector Quantized Variational Autoencoders (VQ-VAEs)

Gaussian VAEs

$$\text{VAE:} \quad \Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,z} \; \ln \frac{\hat{p}_\Psi(z|y)}{p_\Phi(z)} - \ln p_\Phi(y|z)$$

All models are Gaussian densities.

$$p_\Phi(z[i]) \; \propto \; \exp((z[i] - \mu_\Phi[i])^2 / 2\sigma_\Phi^2[i])$$

$$p_\Psi(z[i]|y) \; \propto \; \exp((z[i] - \hat{z}_\Psi(y)[i])^2 / 2\sigma_\Psi^2(y)[i])$$

$$p_\Phi(y[i]|z) \; \propto \; \exp((y[i] - \hat{y}_\Phi(z)[i])^2 / 2\sigma_\Phi^2(z)[i])$$

WLOG $p_\Phi(z) = \mathcal{N}(0, I)$

There is a simple reparameterization Φ' and Ψ' of Φ and Ψ such that Φ' and Ψ' give the same value of the ELBO but $p_{\Phi'}(z) = \mathcal{N}(0, I)$.

Gaussian VAEs for Faces 2014

We can sample faces from the VAE by sampling noise z from $p_\Phi(z)$ and then sampling an image y from $p_\Phi(y|z)$.

[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.

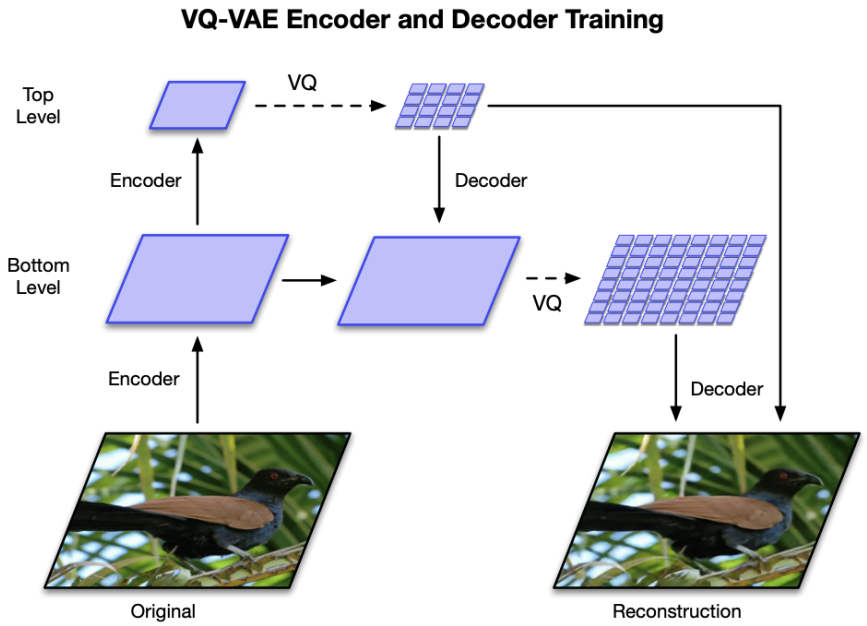
VQ-VAE-2, Razavi et al. June, 2019

Vector Quantized VAEs (VQ-VAE)

VQ-VAEs effectively perform k-means on vectors in the model so as to represent vectors by discrete cluster centers.

We use x and y for spatial image coordinates and use s (for signal) to denote images.

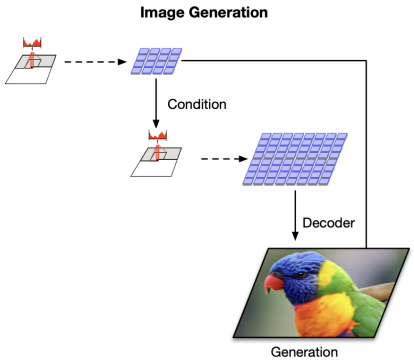
VQ-VAE Encoder-Decoder



In the one-layer case the latent variable (compressed image) is a "symbolic image" $z[X, Y]$ where $z[x, y]$ is a symbol (a cluster index).

Naively $z[X, Y]$ can be represented with $XY \log_2 K$ bits.

VQ-VAE Image Sampler



But they also train an autoregressive probability model (pixel CNN) giving a probability $P_\Phi(z[X, Y])$ for the symbolic image $z[X, Y]$.

This gives $-\log_2 P_\Phi(z[X, Y])$ bits per image (much lower).

To sample an image they sample $z[X, Y]$ from $P_\Phi(z[X, Y])$.

VQ-VAE Encode-Decode Training

We train a code book $C[K, I]$ where $C[k, I]$ is the center vector of cluster k.

$$
\begin{aligned}
L[X, Y, I] &= \mathrm{Enc}_\Phi(s) \\
z[x, y] &= \operatorname*{argmin}_k \; ||L[x, y, I] - C[k, I]|| \\
\hat{L}[x, y, I] &= C[z[x, y], I] \\
\hat{s} &= \mathrm{Dec}_\Phi(\hat{L}[X, Y, I])
\end{aligned}
$$

The "symbolic image" $z[X, Y]$ is the latent variable (compressed image) with naive bit length $XY \log_2 K$.

Training the Code Book

We preserve information about the image s by minimizing the distortion between $L[X, Y, I]$ and its reconstruction $\hat{L}[X, Y, I]$.

$$\Phi^* = \underset{\Phi}{\text{argmin}} \; E_s \; \beta||L[X, Y, I] - \hat{L}[X, Y, I]||^2 + ||s - \hat{s}||^2$$

Parameter-Specific Learning Rates

$$||L[X,Y,I] - \hat{L}[X,Y,I]||^2 = \sum_{x,y} ||L[x,y,I] - C[z[x,y],I]||^2$$

For the gradient of this they use

$$\begin{aligned} \text{for } x,y \;\; L[x,y,I].\text{grad } \texttt{+=} \; 2\beta(L[x,y,I] - C[z[x,y],I]) \\ \text{for } x,y \;\; C[z[x,y],I].\text{grad } \texttt{+=} \; 2(C[z[x,y],I] - L[x,y,I]) \end{aligned}$$

This gives a parameter-specific learning rate for $C[K,I]$.

Parameter-specific learning rates do not change the stationary points (the points where the gradients are zero).

The Relationship to K-means

for x, y $C[z[x, y], I]$.grad += $2(C[z[x, y], I] - L[x, y, I])$

At a stationary point we get that $C[k, I]$ is the mean of the set of vectors $L[x, y, I]$ with $z[x, y] = k$ (as in K-means).

Straight Through Gradients

$$z[x, y] = \underset{k}{\text{argmin}} \; ||L[x, y, I] - C[k, I]||$$

$$\hat{L}[x, y, I] = C[z[x, y], I]$$

$z[x, y]$.grad $= 0$ and back-propagation fails. This is true for any discrete value in a network.

They use "straight-through" gradients.

$$\text{for } x, y \;\; L[x, y, I].\text{grad } \texttt{+=} \; \hat{L}[x, y, I].\text{grad}$$

This assumes low distortion between $L[X, Y, I]$ and $\hat{L}[X, Y, I]$.

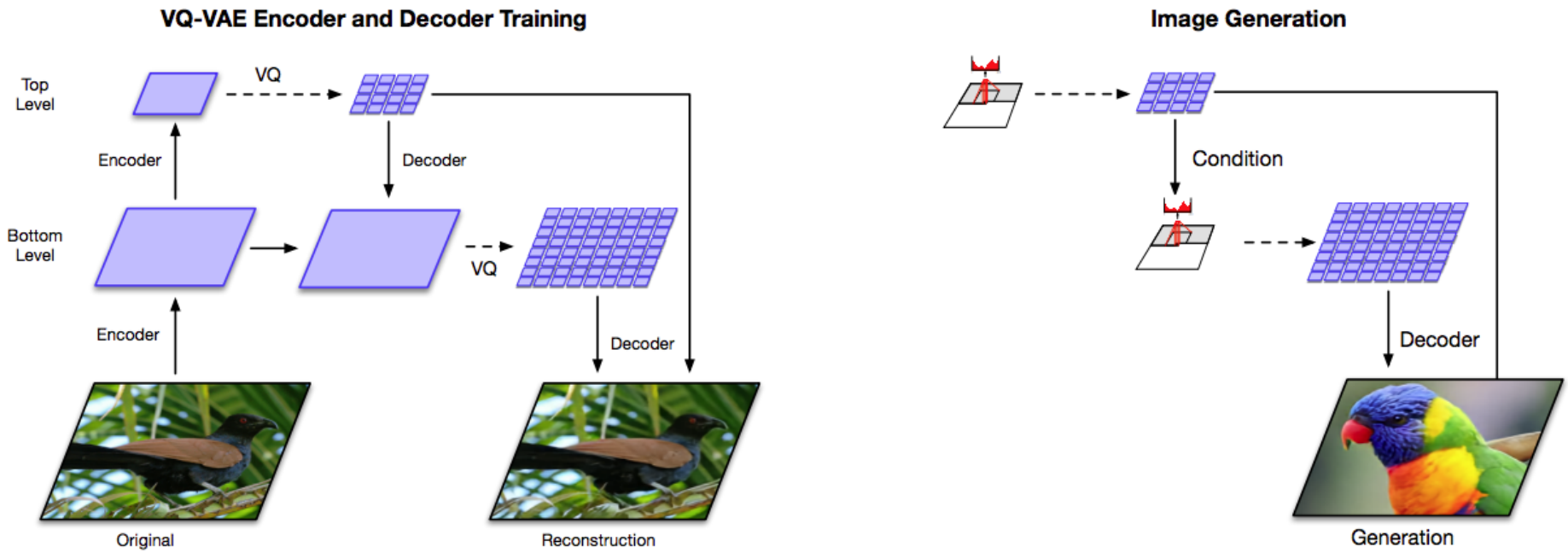
A Suggested Modification

The parameter β is paying two roles

- It controls the relative weight of the two distortion losses.
- It controls the learning rate adjustment for the codebook.

Shouldn't we have separate parameters for these two roles?

Multi-Layer Vector Quantized VAEs



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

h_{top} $h_{\text{top}}, h_{\text{middle}}$ $h_{\text{top}}, h_{\text{middle}}, h_{\text{bottom}}$ Original

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