TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Vector Quantized Variational Autoencoders (VQ-VAEs)

Gaussian VAEs 2013

Sample $z \sim \mathcal{N}(0, I)$ and compute $y_\Phi(z)$

[Alec Radford]

VQ-VAEs 2019

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

VQ-VAEs 2019

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 $\hat{L}[X, Y, I]$.

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

This is a two-level rate-distortion auto-encoder where the rate can be no larger than $XY \ln K$.

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

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

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

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

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

The latent variables are discrete so some approximation to SGD must be used.

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.

Direct 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.

Image Compression

END