VQ-VAE

A Preprint

1 Connection to VAE

Standard Variational Autoencoders (VAEs) consist of:

- an encoder $q(z|x)$ (typically Gaussian),
- a prior $p(z)$ (usually standard Gaussian),
- a decoder $p(x|z)$.

They are trained by maximizing the ELBO:

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)\|p(z)).$$

VQ-VAE modifies this by introducing discrete latent variables and a vector quantization bottleneck instead of a continuous latent space.

2 Discrete Latent Variables

Instead of sampling $q(z|x)$, VQ-VAE selects the nearest neighbor:

$$q(z = k|x) = \begin{cases} 1 & \text{if } k = \arg\min_j \|z_e(x) - e_j\|^2, \\ 0 & \text{otherwise,} \end{cases}$$

where $z_e(x)$ is the encoder output and $\{e_j\}_{j=1}^K$ is a learned embedding dictionary in $\mathbb{R}^D$.

Quantized latent:

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

3 Loss Function

The full training loss combines three terms:

$$\mathcal{L} = \underbrace{\log p(x|z_q(x))}_{\text{reconstruction}} + \underbrace{\|\mathrm{sg}[z_e(x)] - e\|_2^2}_{\text{codebook update}} + \underbrace{\beta\|z_e(x) - \mathrm{sg}[e]\|_2^2}_{\text{commitment loss}}.$$

- $\mathrm{sg}[\cdot]$ is the stop-gradient operator: identity in the forward pass, zero in the backward pass.
- The decoder is optimized using the first term.
- The embeddings are updated using the second term.
- The encoder is optimized using the first and third terms.

4 Gradients and Training

Since quantization is non-differentiable, gradients are passed via a **straight-through estimator**:

$$\frac{\partial \mathcal{L}}{\partial z_e(x)} \approx \frac{\partial \mathcal{L}}{\partial z_q(x)}.$$

5 Prior and Generation

During training, the prior is fixed and uniform:

$$p(z) = \frac{1}{K}, \quad \Rightarrow \quad \mathrm{KL}(q(z|x)\|p(z)) = \log K = \text{constant}.$$

After training, we fit an **autoregressive prior** over z:

- PixelCNN for images,
- WaveNet for audio.

To generate, we sample from $p(z)$ autoregressively and decode via $p(x|z)$.

6 Log-likelihood Approximation

Since the decoder is trained using $z_q(x)$, we approximate:

$$\log p(x) \approx \log p(x|z_q(x)) + \log p(z_q(x)).$$

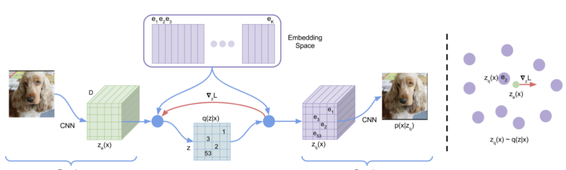
From Jensen's inequality:

$$\log p(x) \geq \log p(x|z_q(x)) + \log p(z_q(x)).$$

7 Scaling to Multiple Latents

VQ-VAE uses N discrete latent variables (e.g., 32×32 grid for ImageNet), and the loss becomes:

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{(i)}.$$



Algorithm 1 VQ-VAE Forward and Training Pass

Input: data sample x, codebook $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$
$\mathbf{z}_e \leftarrow \text{Encoder}(x)$
$k \leftarrow \arg\min_j \|\mathbf{z}_e - \mathbf{e}_j\|^2$
$\mathbf{z}_q \leftarrow \mathbf{e}_k$
$\hat{x} \leftarrow \text{Decoder}(\mathbf{z}_q)$
Compute losses:
    $\mathcal{L}_{\text{rec}} \leftarrow \|x - \hat{x}\|^2$
    $\mathcal{L}_{\text{cb}} \leftarrow \|\text{sg}[\mathbf{z}_e] - \mathbf{e}_k\|^2$
    $\mathcal{L}_{\text{com}} \leftarrow \|\mathbf{z}_e - \text{sg}[\mathbf{e}_k]\|^2$
$\mathcal{L} \leftarrow \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cb}} + \beta\mathcal{L}_{\text{com}}$
Update:
    Update encoder using $\nabla\mathcal{L}_{\text{rec}} + \nabla\mathcal{L}_{\text{com}}$
    Update decoder using $\nabla\mathcal{L}_{\text{rec}}$
    Update codebook using $\nabla\mathcal{L}_{\text{cb}}$