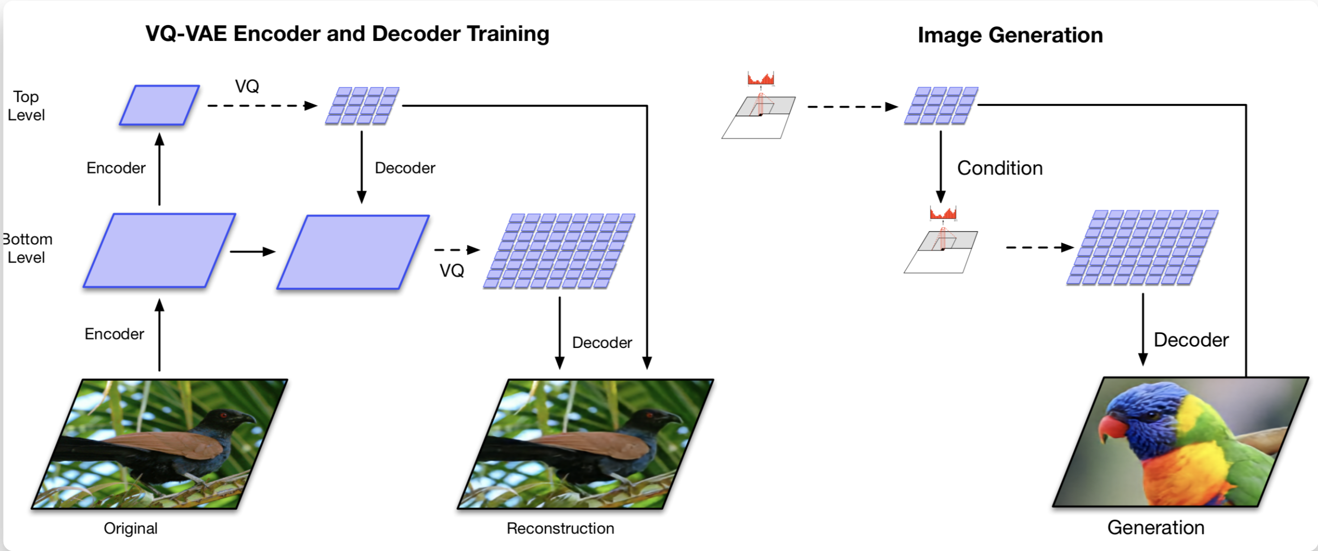
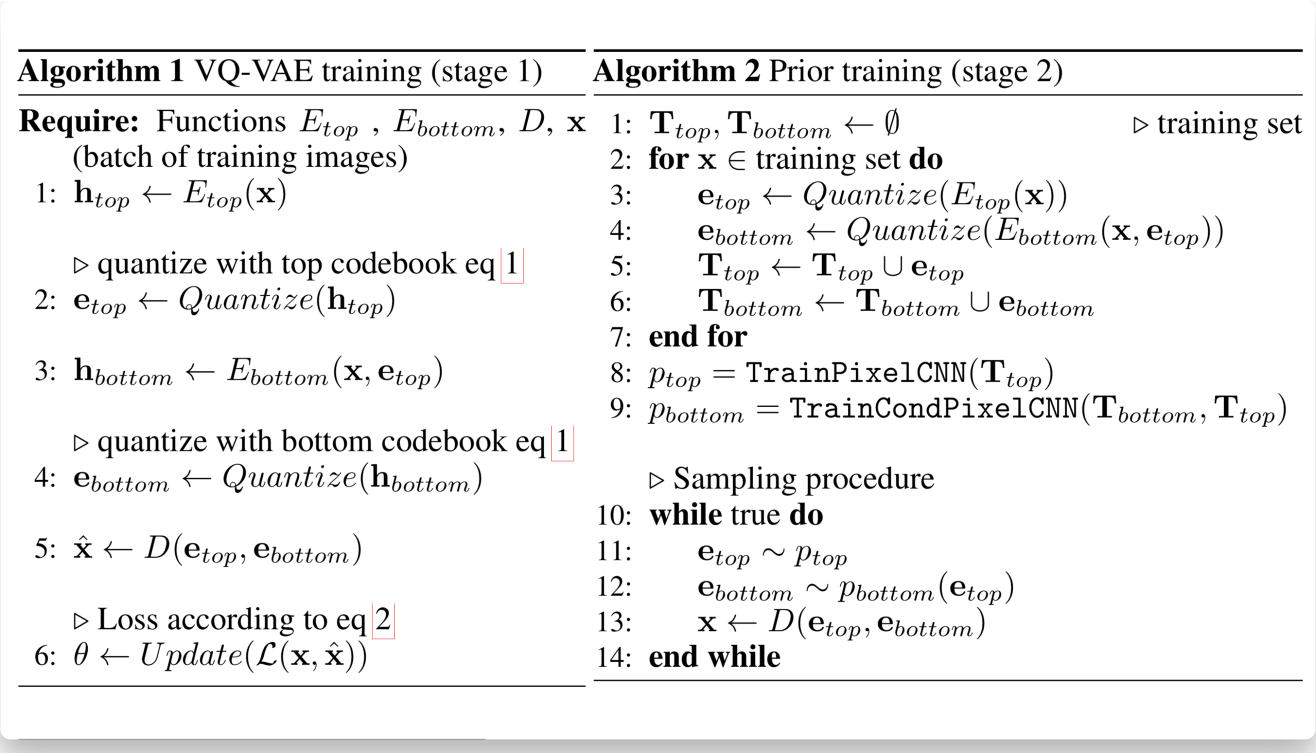
VQ-VAE-2

VQ-VAE-2 (Ali Razavi, et al. 2019) is a two-level hierarchical VQ-VAE combined with self-attention autoregressive model.

01. Stage 1 is to **train a hierarchical VQ-VAE**: The design of hierarchical latent variables intends to separate local patterns (i.e., texture) from global information (i.e., object shapes). The training of the larger bottom level codebook is conditioned on the smaller top level code too, so that it does not have to learn everything from scratch.
02. Stage 2 is to **learn a prior over the latent discrete codebook** so that we sample from it and generate images. In this way, the decoder can receive input vectors sampled from a similar distribution as the one in training. A powerful autoregressive model enhanced with multi-headed self-attention layers is used to capture the prior distribution (like PixelSNAIL; Chen et al 2017).

Considering that VQ-VAE-2 depends on discrete latent variables configured in a simple hierarchical setting, the quality of its generated images are pretty amazing.





Algorithm 1 VQ-VAE training (stage 1)

```
Require: Functions $E_{top}$ , $E_{bottom}$, $D$, $\mathbf{x}$ (batch of training images)
$\mathbf{h}_{top} \leftarrow E_{top}(\mathbf{x})$
    ▷ quantize with top codebook eq 1
$\mathbf{e}_{top} \leftarrow Quantize(\mathbf{h}_{top})$
$\mathbf{h}_{bottom} \leftarrow E_{bottom}(\mathbf{x}, \mathbf{e}_{top})$
    ▷ quantize with bottom codebook eq 1
$\mathbf{e}_{bottom} \leftarrow Quantize(\mathbf{h}_{bottom})$
$\hat{\mathbf{x}} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$
    ▷ Loss according to eq 2
$\theta \leftarrow Update(\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}))$
```

Algorithm 2 Prior training (stage 2)

```
$\mathbf{T}_{top}, \mathbf{T}_{bottom} \leftarrow \emptyset$                ▷ training set
for $\mathbf{x} \in$ training set do
    $\mathbf{e}_{top} \leftarrow Quantize(E_{top}(\mathbf{x}))$
    $\mathbf{e}_{bottom} \leftarrow Quantize(E_{bottom}(\mathbf{x}, \mathbf{e}_{top}))$
    $\mathbf{T}_{top} \leftarrow \mathbf{T}_{top} \cup \mathbf{e}_{top}$
    $\mathbf{T}_{bottom} \leftarrow \mathbf{T}_{bottom} \cup \mathbf{e}_{bottom}$
end for
$p_{top} = \texttt{TrainPixelCNN}(\mathbf{T}_{top})$
$p_{bottom} = \texttt{TrainCondPixelCNN}(\mathbf{T}_{bottom}, \mathbf{T}_{top})$

    ▷ Sampling procedure
while true do
    $\mathbf{e}_{top} \sim p_{top}$
    $\mathbf{e}_{bottom} \sim p_{bottom}(\mathbf{e}_{top})$
    $\mathbf{x} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$
end while
```

VQ-VAE2 has 3 main contributions:

01. Extend the VQ-VAE model to ImageNet
02. Propose a hierarchical structure for VQ-VAE
03. Use pixel-cnn to do generation in the structured latent space, which takes the discrete quantization result $q(z|x)$as inputs and the reconstruction target.

主要变化就是把 VQ-VAE 的 encoder 和 decoder 进行了**分层，bottom 层对 local feature 进行建模，top 层对 global feature 进行建模；为了让 top 层能更有效地提取 global 信息，在网络中加入了 self attention**. 初次之外，在 prior 上进行采样的时候，考虑到 autoregressive 的 sample 会累积误差，即如果 x_1 出现了一点误差，那么由于 $x_2|x_1$的误差会更大，以此类推。因此加入了 rejection sampling, 即生成某类的图片后，通过一个 classifier 判断一下生成的图片像不像这个类的 sample, 如果不像就弃掉。文章效果很惊艳，但是理论上没有特别大的改进。