Implicit Feature Decoupling with Depthwise Quantization

Iordanis Fostiropoulos
University of Southern California
Los Angeles, CA
fostirop@usc.edu

Barry Boehm
University of Southern California
Los Angeles, CA
boehm@usc.edu

A. Appendix

A.1. Representation Capacity

Proofs in this section correspond to claims and results in the main text. Where applicable, a proposition will refer to the the equation in the main text for which the result is applied.

Proposition 1 (Depthwise Quantization Channel Capacity - Result for Equation 2 in the main text)**.** *The capacity $\mathcal{C}$ of Depthwise Quantization (DQ) channel for set of codebooks C is the entropy of the codebooks s.t. $\mathcal{C} = H(C)$*

Proof. Let the capacity of a channel $\mathcal{C} = I(x; z)$ [6] , where $I(::)$ is the mutual information. It is sufficient to show $\mathcal{C} = I(x; C) = H(C) - H(x|C)$ where $z = C = \{C_i : i \in N\}$ is the set of codebooks. Since the quantization channel is a noiseless discrete channel with deterministic quantization function, $P(x|C) = 1$ and thus $H(x|C) = 0$. □

$$\mathcal{C} = I(x; C) = H(C) \tag{1}$$

Proposition 2 (Representation Capacity - Result for Equation 2 in the main text)**.** *The channel capacity is bounded by the number of discrete latent factors S that can be represented by DQ. Let N be the cardinality of the set of codebooks C with K codes. Representation Capacity is defined as $C_R = -H(C) = \log S$*

Proof. Let $S = K^N$ be the sample space for the set of codebooks $C = C_i : i \in N$ with K codes. By definition

$$H(C) = -\sum_{C_i \in N} P(C_1, ..., C_n) \log P(C_1, ..., C_n) \tag{2}$$

where $P(C_i)P(C_j) > P(C_i)P(C_j|C_i)$.
$H(C)$ is maximized when C_i, C_j are independent variables and are uniformly distributed (uniform prior) s.t. $P(C_i) = \frac{1}{K}$. Thus:

$$\begin{aligned} H_{\max}(C) = &-\sum_{i \in K} [P(C_1) \times ... \times P(C_n) \\ &\log [P(C_1) \times ... \times P(C_n)] \\ &= \log K^N = \log S \end{aligned}$$ □

$$C_R = -H(C) = \log S \tag{3}$$

Proposition 3 (ELBO for Depthwise AutoEncoder - Result for Equation 6)**.** *The variational lower bound of DQ-AE is*

$$\mathbb{L} \geq max[\mathbb{E}_{q(z|x)} \log p(x|z) - C_R] \tag{4}$$

Proof. By definition [1]

$$\mathbb{L} \geq \mathbb{E}_{q(z|x)} \log p(x|z) - \beta D_{KL}(q(z|x)||p(z)) \tag{5}$$

Thus, it is sufficient to show that C_R is the bound of the divergence of the uniform prior $p(z)$ and inferred prior $q(z|x)$ s.t.

$$D_{KL}(q(z|x)||p(z)) = C_R - S \tag{6}$$

Let $p(z)$ be the uniform distribution and $q(z|x)$ the inferred prior. Therefore,

$$\begin{aligned} D_{KL}(q(z|x)||p(z)) &= \sum_{i \in N} q(z_i|x) \log \left(\frac{q(z_i|x)}{p(z_i)} \right) \\ &= \sum_{i \in N} q(z|x) \log \left(q(z|x) K^{-1} \right) \\ &= \sum_{i \in N} q(z|x) \log \left(q(z|x) \right) - N \log \left(K \right) \\ &< -H(q(z|x)) \end{aligned}$$ □

Since S is constant, it does not affect the optimization objective, the ELBO is

$$\mathbb{L} \geq \max[\mathbb{E}_{q(z|x)} \log p(x|z) - C_R] \tag{7}$$

A.2. Architecture

In this section we provide details on the Hierarchical DQ-AE architecture.

Algorithm 1 N-Hierarchical Depthwise Vector Quantizer

given encoder E, decoder $\mathcal{D}$, N $\times$ { quantizers Q, decoders D, up-samplers U } for each hierarchy, Reconstruction Loss function $\mathcal{L}$ and Optimizer $\mathcal{O}$ and training sample x

▷ Stack of N encoded representations bottom to top
$\mathbf{e}_{all} \leftarrow E(x)$
$e_{top} \leftarrow pop(\mathbf{e}_{all})$

▷ Quantize using DVQ
$q \leftarrow Q_{top}(e_{top})$
$d \leftarrow D_{top}(q)$
$u_{all} \leftarrow list()$
for e in $\mathbf{e}_{all}$ **do**
  $q,d,u \leftarrow$ DECODE(e, d)
  $u_{all} \leftarrow append(u)$
$\hat{x} \leftarrow \mathcal{D}$ (u_{all})
Update $\theta_{[E,Q,D,U]}$ based on $\mathcal{L}(x, \hat{x})$, using Optimizer $\mathcal{O}$

procedure DECODE(e_{cur}, d_{prev})
  Input Current level encoding e_{cur} and previous decoding d_{prev}
  Output Current Level quantization q, upsampling u and decoding d

  $q \leftarrow Q_{cur}$ (e_{cur}, d_{prev})
  $u \leftarrow U_{prev}(q)$
  $d \leftarrow D_{cur}(q) + d_{prev}$
  return q, d, u

Algorithm 1. As opposed to VQ-VAE [4] we use skip connections on the decoded quantized representations from top hierarchies to bottom and thus increase interaction between hierarchies to avoid prior collapse of top-level hierarchies. The decoder accepts quantized upsampled representations as opposed to independently decoding each hierarchy. Fig. 1 shows an overview of the architecture.

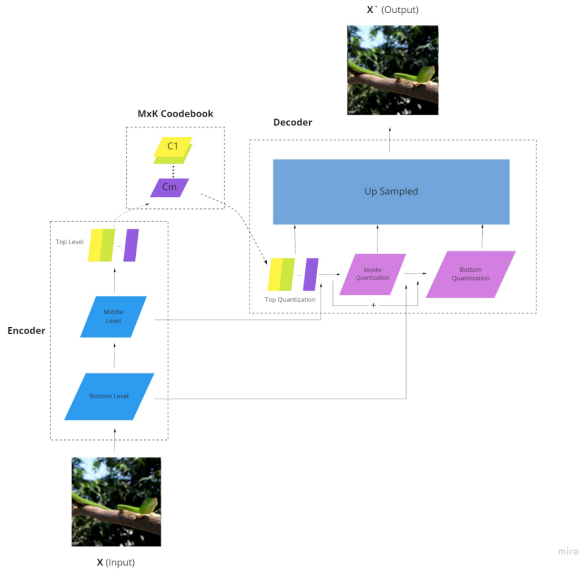

Figure 1. Architecture of N-Hierarchical Depthwise AutoEncoder. X is input to the model and is progressively encoded to finer grain representations. Each hidden representation in the decoder is decoded using previous hierarchy's decoded quantized representation as well as the encoded representation. The quantized representations are up-sampled and decoded jointly. Quantization of top use no prior decoding.

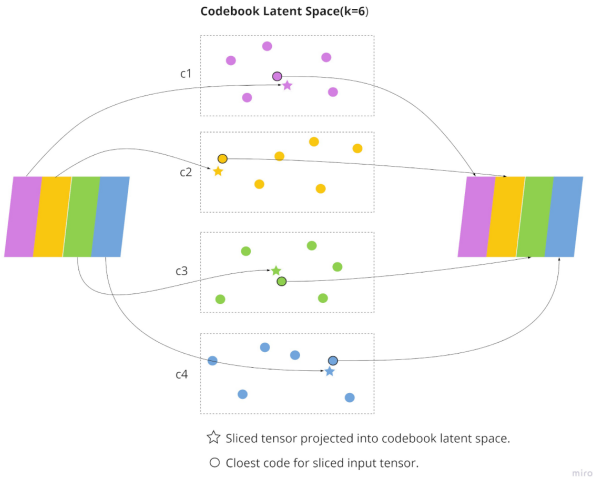

Figure 2. Each input latent representation is sent to the corresponding codebook. The closest code in the codebook latent space is the output of DQ.

A.3. Ablation Study

Results for the ablation study on the quantization process can be found in Tab. 1. Results for the ablation study on DQ-AE can be found in Tab. 2. We also perform additional experiments on MNIST where DQ ("Our") outperforms VQ with 1.92e-04 in l_2 reconstruction loss as compared to 3.41e-04, and similarly for CelebA with 9.57e-03 compared to 3.70e-02.

A.4. Training Configuration

For all experiments and for the quantizer we use $\beta = 0.25$ and dimensionality of each code $D = 64$, decay factor $\gamma = 0.99$ and $\epsilon = 1.00e-05$ unless otherwise noted. We use a different random seed for all experiments and for every trial. For the discretized logistic mixture loss ("mix") [5], we use 10 components and discretize on 8-bit (lossless). We use Adam with weight decay regularization [3] for optimization for all training settings. We use automatic mixed precision (*amp*)[1].We use a batch size of 128, learning rate 2.00e-04 and train for 400 epochs.

Ablation Study For DQ-AE we use 2 Encoder Block composed of 4 Resnet Block with Conv2D layer of 256 channel and 256 hidden unit and stride 2.

Likelihood estimation DQ-AE for the likelihood estimation task uses 2 hierarchies with $K_{\text{bot}} = 128$ and $K_{\text{top}} = 256$. For each hierarchical encoder, it uses 2 encoder block composed of 4 resnet block with Conv2D layer of 256 channel and 256 hidden unit.

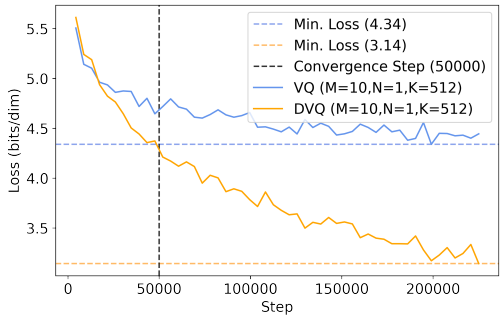

Figure 3. NLL Loss in bits/dim over time. Comparison between VQ and DVQ with an equivalent training set up. DVQ matches the best NLL reported for VQ by step 50,000 in contrast to step 200,000.

loss func.	M	K	DQ (nats/dim)	VQ (nats/dim)
ce	1	32	4.16e+00	4.16e+00
ce	1	128	4.01e+00	4.01e+00
ce	1	512	3.85e+00	3.85e+00
ce	3	32	**3.55e+00**	3.92e+00
ce	3	128	**3.31e+00**	3.80e+00
ce	3	512	**3.13e+00**	3.68e+00
ce	5	32	**3.25e+00**	3.84e+00
ce	5	128	**2.96e+00**	3.71e+00
ce	5	512	**2.75e+00**	3.59e+00
ce	10	32	**2.71e+00**	3.71e+00
ce	10	128	**2.37e+00**	3.51e+00
ce	10	512	**2.13e+00**	3.44e+00
loss func.	M	K	DQ (L_2)	VQ (L_2)
mse	1	32	1.22e-01	1.22e-01
mse	1	128	8.75e-02	8.75e-02
mse	1	512	6.78e-02	6.78e-02
mse	3	32	3.90e-02	7.70e-02
mse	3	128	**2.49e-02**	6.17e-02
mse	3	512	**1.67e-02**	4.78e-02
mse	5	32	**2.08e-02**	6.48e-02
mse	5	128	**1.15e-02**	5.52e-02
mse	5	512	**7.33e-03**	4.14e-02
mse	10	32	**6.84e-03**	5.54e-02
mse	10	128	**3.08e-03**	4.07e-02
mse	10	512	**1.68e-03**	3.25e-02

Table 1. We vary the number of codebook vectors K and codebooks M, while we keep the same $D = 64$. We evaluate our results on CIFAR10 using an identical training configuration between all models and multiple random initialization. Note that the DQ model do not fully converge, due to the limited number of computational resources. We train for 400 epochs and pick the best test loss for each architecture. The comparison between the models shows a statistical trend of improved likelihood estimation for $DQ - AE$. Figure 5 in the main text, shows the aggregate results of the likelihood estimation. The top, middle, and bottom line correspond to K having values 32,128,and 512, respectively. The effect of K is not as significant as the effect of M. For M=1 both VQ and DVQ are identical in terms of theoretical and experimental performance. As we increase M, we find that the loss significantly decreases. Moreover, K, is not the limiting factor to the channel capacity but M is. This can also be seen on the graph as the loss for all different K converges as we increase M.

[1] https://pytorch.org/docs/stable/amp.html

loss func.	M	K	DQ (nats/dim)	VQ (nats/dim)
ce	5	[128,128,128]	**2.96e+00**	3.73e+00
ce	5	[128,128]	**2.95e+00**	3.60e+00
ce	5	[128,256]	**2.96e+00**	3.69e+00
ce	5	[128,32]	**2.94e+00**	3.70e+00
ce	5	[256,128]	**2.84e+00**	3.63e+00
ce	5	[256,256]	**2.85e+00**	3.63e+00
ce	5	[32,128]	**3.21e+00**	3.72e+00
ce	5	[32,32,32]	**3.24e+00**	3.69e+00
ce	5	[64,64,64]	**3.08e+00**	3.79e+00
loss func.	M	K	DQ (nats/dim)	VQ (nats/dim)
mix	5	[128,128,128]	**2.55e+00**	3.04e+00
mix	5	[128,128]	**2.56e+00**	3.12e+00
mix	5	[128,256]	**2.52e+00**	3.11e+00
mix	5	[128,32]	**2.55e+00**	3.18e+00
mix	5	[256,128]	**2.49e+00**	3.15e+00
mix	5	[256,256]	**2.49e+00**	3.11e+00
mix	5	[32,128]	**2.79e+00**	3.26e+00
mix	5	[32,32,32]	**2.80e+00**	3.26e+00
mix	5	[64,64,64]	**2.65e+00**	3.06e+00
loss func.	M	K	DQ (L_2)	VQ (L_2)
mse	5	[128,128,128]	**1.15e-02**	5.53e-02
mse	5	[128,128]	**1.24e-02**	5.05e-02
mse	5	[128,256]	**1.02e-02**	5.52e-02
mse	5	[128,32]	**1.15e-02**	5.07e-02
mse	5	[256,128]	**9.31e-03**	4.52e-02
mse	5	[256,256]	**9.27e-03**	5.13e-02
mse	5	[32,128]	**1.94e-02**	5.82e-02
mse	5	[32,32,32]	**2.03e-02**	6.29e-02
mse	5	[64,64,64]	**1.50e-02**	5.81e-02

Table 2. Hierarchical Depthwise Quantizers for 2 and 3 hierarchies. DQ outperforms equivalent VQ. The "mix" objective function refers to 8-bit mixture of logistics [5] following the methodology by Child *et al*. [2]. The hierarchy capacity K is reported from top to bottom, i.e. $[K_{top}, K_{mid}, K_{bot}]$.

A.5. Hierarchical Reconstruction

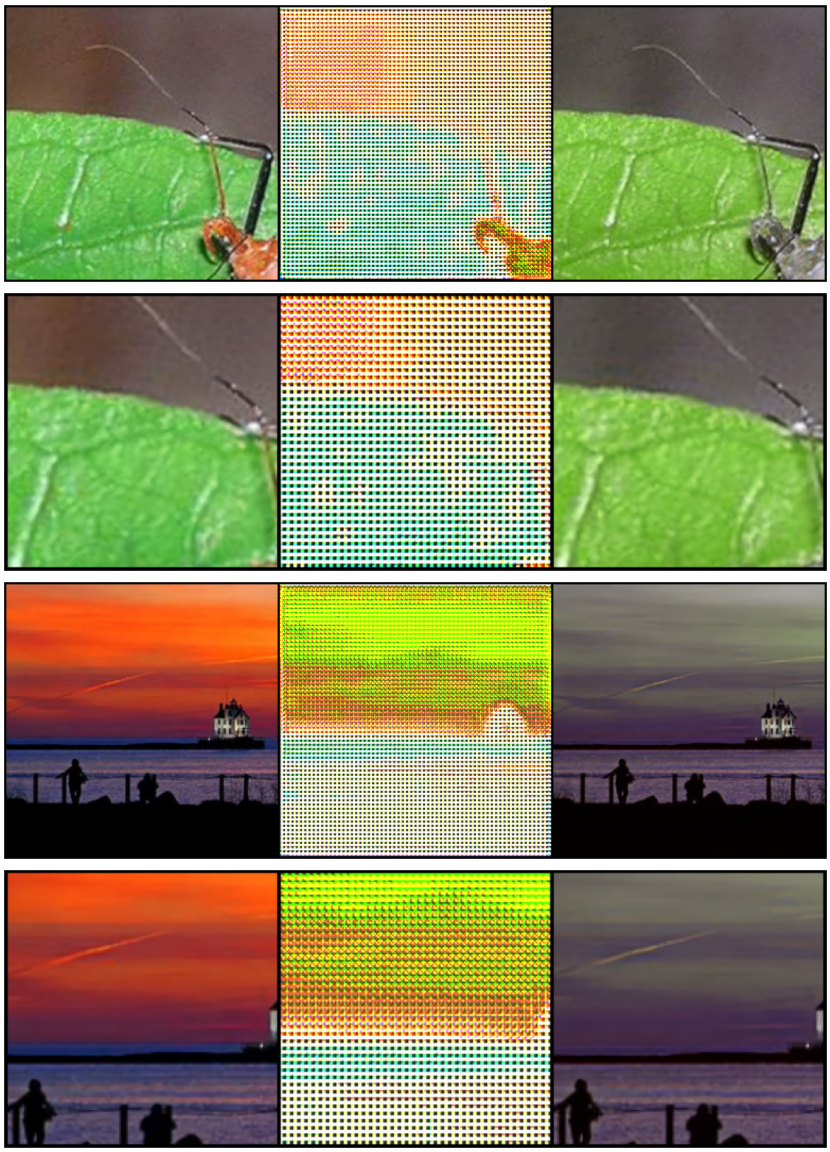

Figure 4. Image reconstructions from a model trained with L_2 for the reconstruction loss. Original image (left) is reconstructed using only **top** level codes (middle) and only **bottom** level codes (right). Top level hierarchy contains structural information, while bottom level hierarchy contains details.

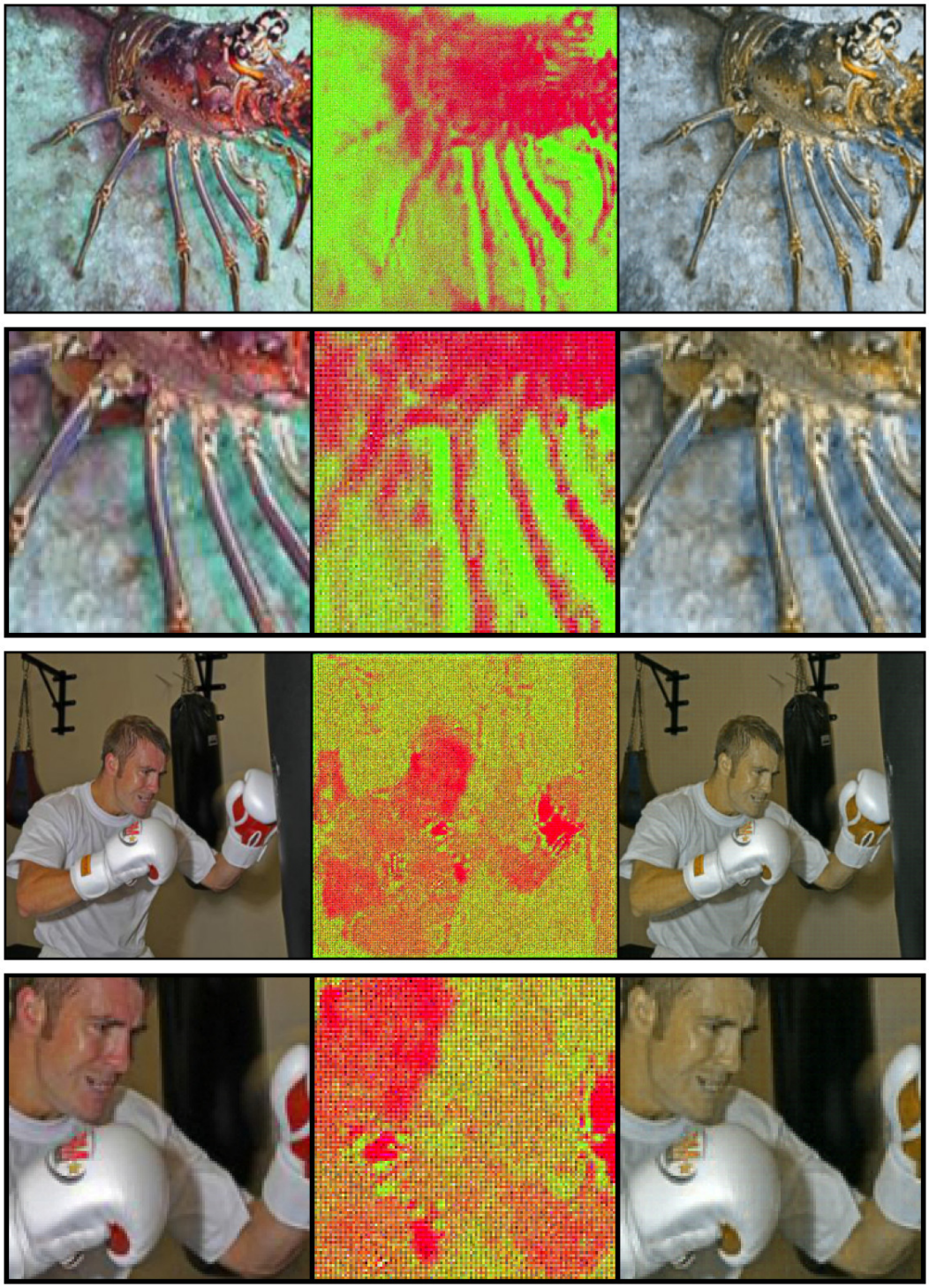

Figure 5. Image reconstructions from a model trained with discretized mixture of logistic loss (dmol) [5] for the reconstruction loss. Original image (left) is reconstructed using only **top** level codes (middle) and only **bottom** level codes (right). Top level hierarchy contains structural information, while bottom level hierarchy contains details.

A.6. Perceptual Evaluation of Image Reconstructions

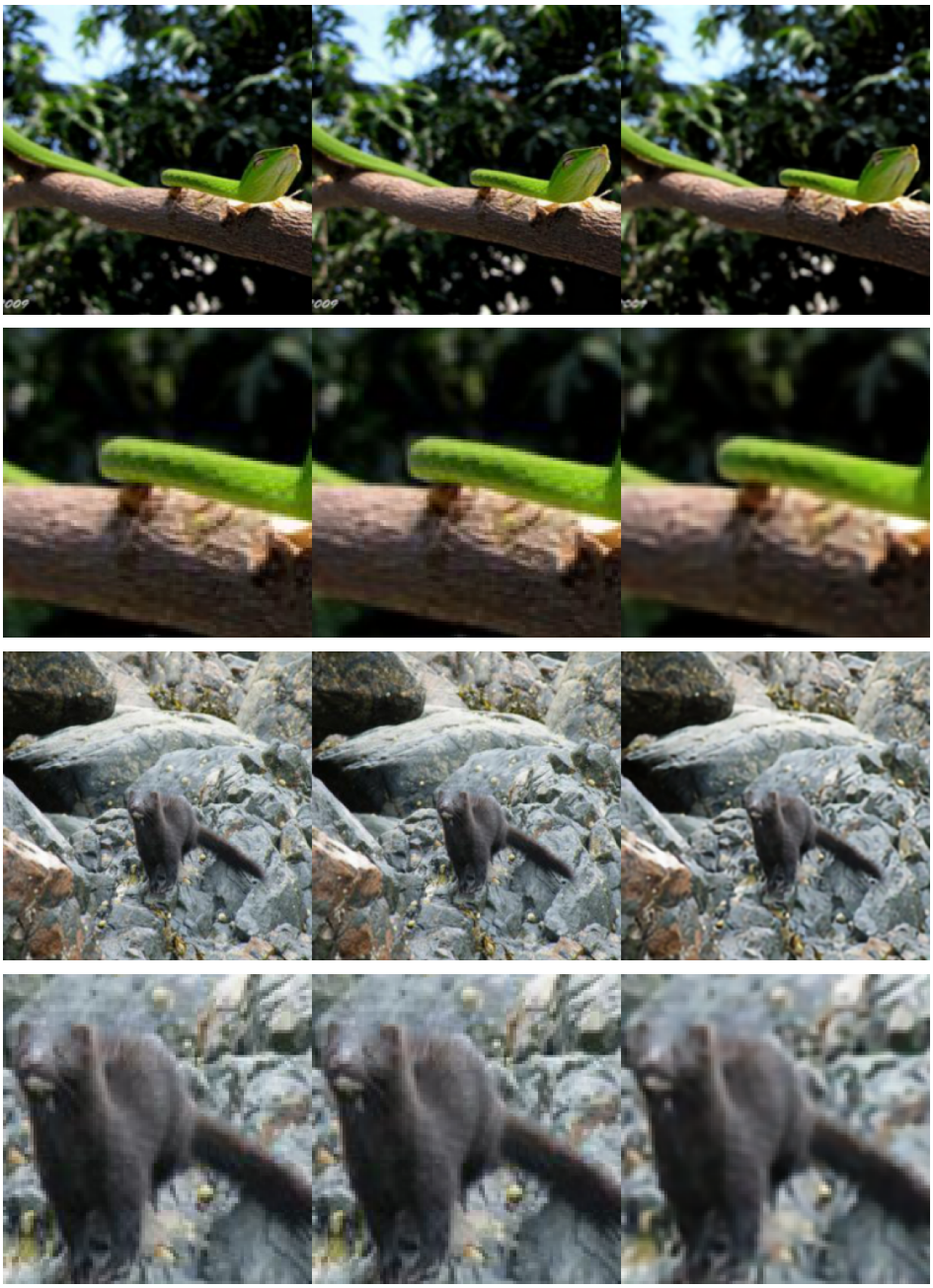

Figure 6. The original image (left) is fed through and reconstructed by a model trained with DQ (middle) and VQ (right). The model is trained using identical settings. Perceptual quality of DQ outperforms VQ.

Figure 7. The original image (left) is fed through and reconstructed by a model trained with DQ (middle) and VQ (right). The model is trained using identical settings. Perceptual quality of DQ outperforms VQ.

References

[1] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018. 1

[2] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. 4

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[4] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019. 2

[5] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017. 3, 4, 6

[6] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 1