Hierarchical Residual Learning Based Vector Quantized Variational Autoencoder for Image Reconstruction and Generation

Mohammad Adiban[1,2], Kalin Stefanov[2], Marco Siniscalchi[1], Giampiero Salvi[1,3]
[1]Norwegian University of Science and Technology (NTNU), [2]Monash University, [3]KTH Royal Institute of Technology

Overview

- We proposed a hierarchical residual learning based vector quantized variational Autoencoder (namely HR-VQVAE) for the image reconstruction and generation tasks.
- HR-VQVAE maps the continuous latent representations to several layers of discrete representations through hierarchical codebooks.
- The vector selected within bottom layer determines the codebook that is activated in the top layer.
- Such a hierarchical searching procedure provides the advantage of local access to codebook indexes, which dramatically reduces search time.
- A novel objective function is proposed to provide contrastive learning by pushing each layer to extract information not learned by its preceding layers.
- The objective optimizes the output image from the combination of representations obtained from all layers.

Background: VQVAE, VQVAE-2

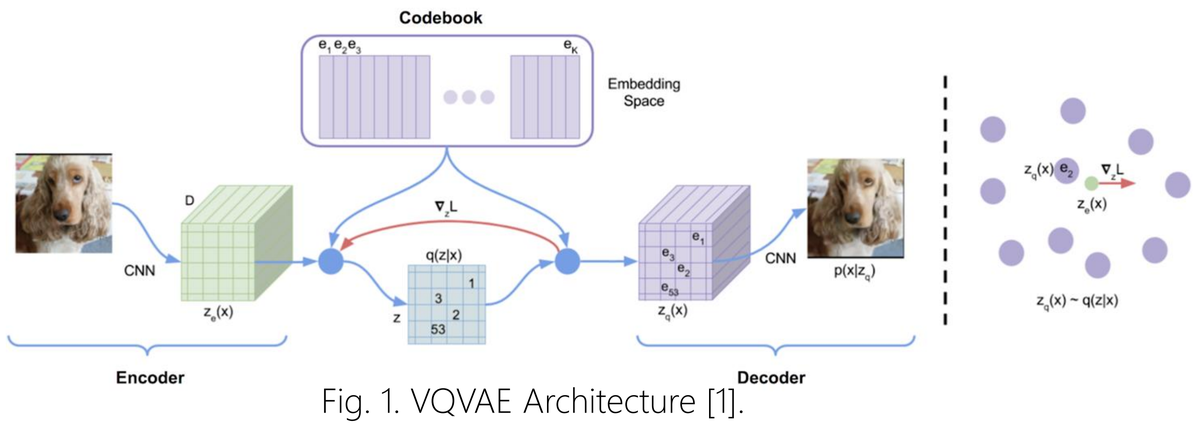


Fig. 1. VQVAE Architecture [1].

$$\text{Quantize}(\mathbf{z}_{hw}) = \mathbf{e}_k \text{ where } k = \arg\min_j ||\mathbf{z}_{hw} - \mathbf{e}_j||_2, \quad (1)$$

$$\mathcal{L}(\mathbf{x}, \mathcal{D}(\mathbf{e})) = \|\mathbf{x} - \mathcal{D}(\mathbf{e})\|_2^2 + \|\text{sg}[\mathbf{z}] - \mathbf{e}\|_2^2 + \beta\|\text{sg}[\mathbf{e}] - \mathbf{z}\|_2^2. \quad (2)$$

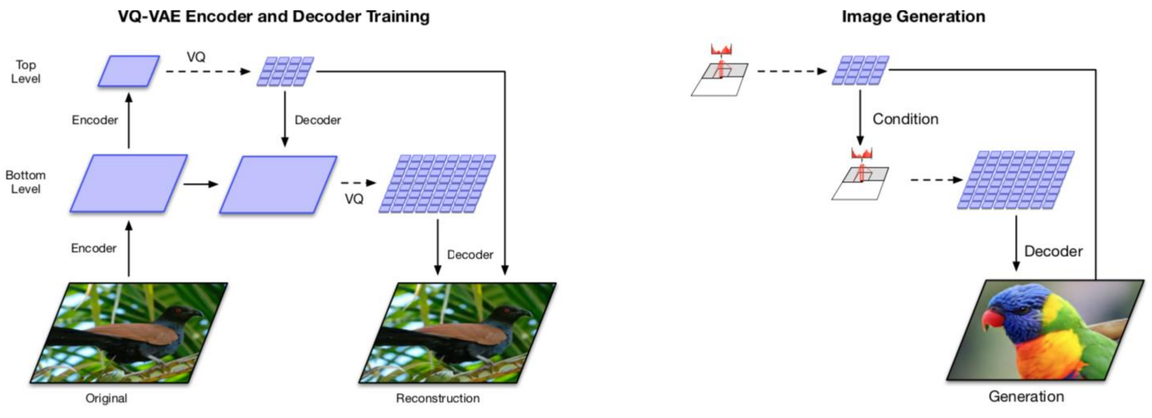


Fig. 2. VQVAE-2 Architecture [2].

Proposed Method: HR-VQVAE

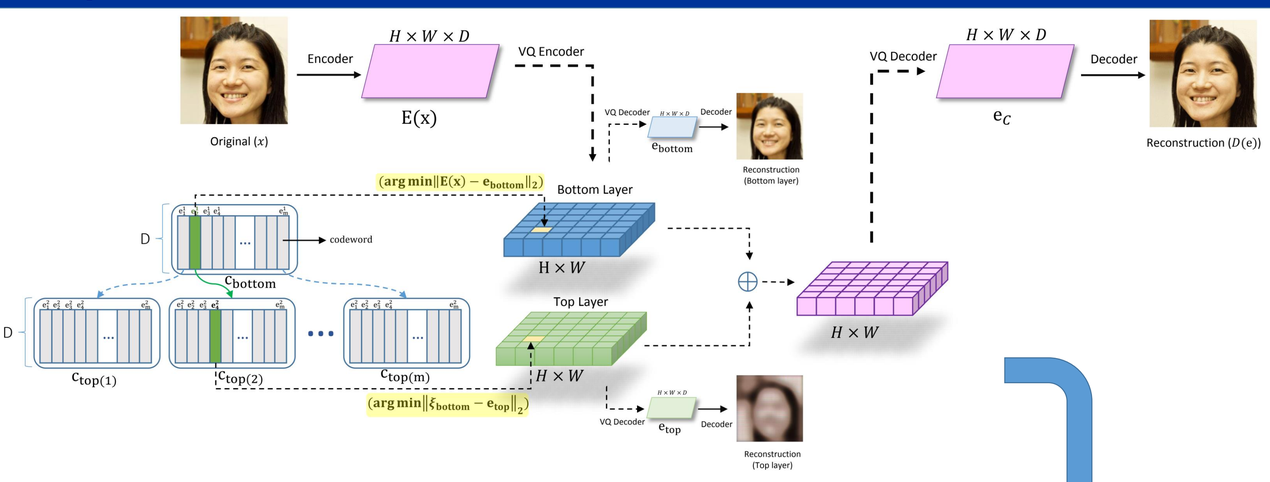
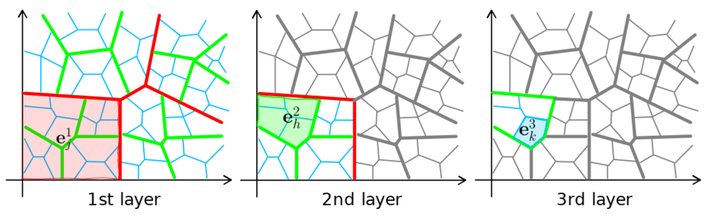
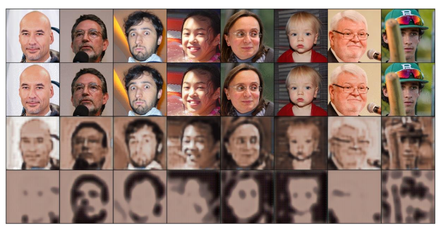


Fig.3. HR-VQVAE Architecture.





Objective functions

$$\text{Quantize}^i(\xi_{hw}^{i-1}) = \mathbf{e}_k^i \text{ where } k = \arg\min_j \|\xi_{hw}^{i-1} - \mathbf{e}_j^i\|_2, \quad (3)$$

$$\mathbf{e}_C = \sum_{i=1}^{n} \mathbf{e}^i, \quad (4)$$

$$\mathcal{L}(\mathbf{x}, \mathcal{D}(\mathbf{e}_C)) = \|\mathbf{x} - \mathcal{D}(\mathbf{e}_C)\|_2^2 + \|\text{sg}[\xi^0] - \mathbf{e}_C\|_2^2 + \beta_0\|\text{sg}[\mathbf{e}_C] - \xi^0\|_2^2 + \sum_{i=1}^{n} \mathcal{L}(\xi^{i-1}, \mathbf{e}^i), \quad (5)$$

$$\mathcal{L}(\xi^{i-1}, \mathbf{e}^i) = \|\text{sg}[\xi^{i-1}] - \mathbf{e}^i\|_2^2 + \beta_i\|\text{sg}[\mathbf{e}^i] - \xi^{i-1}\|_2^2, \quad (6)$$

Experiments

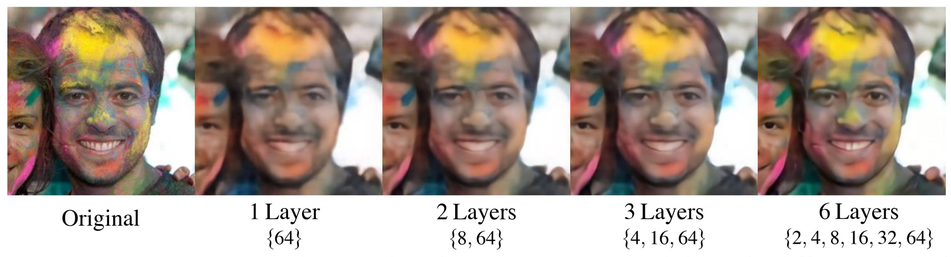


Fig. 4. Reconstructions obtained with HR-VQVAE models with different depths.

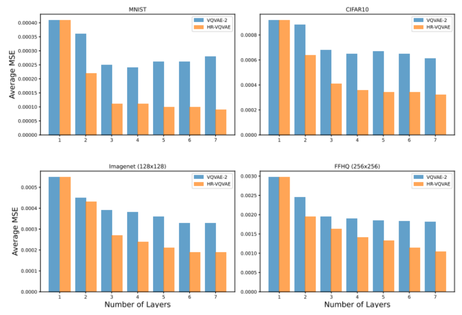


Fig. 5. The effect of model depth.

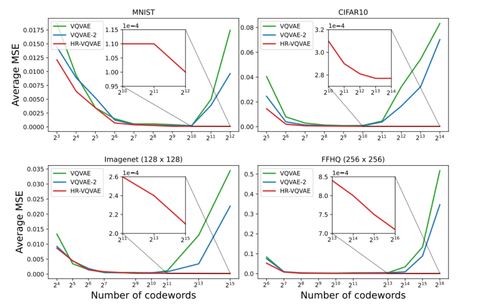


Fig. 6. Average MSE vs number of codewords.

Fig. 7. Reconstructions obtained with HR-VQVAE, VQVAE-2 and VQVAE.

Table 1. FID/MSE reconstruction results using HR-VQVAE, VQVAE-2 and VQVAE

Model	FID ↓ / MSE ↓			
	FFHQ	ImageNet	CIFAR10	MNIST
VQVAE [19]	2.86/0.00298	3.66/0.00055	21.65/0.00092	7.9/0.00041
VQVAE-2 [18]	1.92/0.00195	2.94/0.00039	**18.03**/0.00068	6.7/0.00025
HR-VQVAE	**1.26/0.00163**	**2.28/0.00027**	18.11/**0.00041**	**6.1/0.00011**

Table 2: Time for reconstructing 10,000 samples using HR-VQVAE, VQVAE-2 and VQVAE.

Model	Seconds			
	FFHQ	Imagenet	CIFAR10	MNIST
VQVAE [19]	5.0977652	4.6152677	2.7087896	0.062474
VQVAE-2 [18]	9.3443758	8.8135872	4.4492340	0.090778
HR-VQVAE	**0.8398101**	**0.6714823**	**0.4667842**	**0.010830**

Fig. 8. Random samples on FFHQ generated by VQVAE-2, VQ-GAN and HR-VQVAE, respectively.

Table 3: Generation results using HR-VQVAE, VQVAE-2 and VQVAE.

Model	Generation evaluation (FID ↓)			
	FFHQ	ImageNet	CIFAR10	MNIST
VQVAE [19]	24.93	44.76	78.90	16.69
VQVAE-2 [18]	19.66	39.51	74.43	11.81
HR-VQVAE	**17.45**	**35.29**	**71.38**	**11.75**

Conclusion

- We proposed a novel multi-layer variational Autoencoder method for image modeling that we call HR-VQVAE.
- HR-VQVAE learns hierarchical residual discrete representations in an iterative and hierarchical fashion.
- The objective of HR-VQVAE is designed to encourage different layers to encode different aspects of an image.
- Through experimental evidence, we show how HR-VQVAE can reconstruct images with a higher level of details than state-of-the-art models with similar complexity.
- We also show that we can increase the size of the codebooks without incurring the codebook collapse problem that is observed in methods such as VQVAE and VQVAE-2.

References

[1] *Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In Advances in Neural Information Processing Systems, pages 6306–6315, 2017.*
[2] *Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In Advances in neural information processing systems, pages 14866–14876, 2019.*