VECTOR QUANTIZED VARIATIONAL AUTOENCODERS (VQ-VAE) FOR IMAGE GENERATION ON CIFAR-10

Boqi Zhao, Hezhi Xie, Yinuo Tang,
University of California, Davis
{boqzhao, hezxie, fptang}@ucdavis.edu

1 MOTIVATION

Variational AutoEncoders (VAEs) represent a class of generative models extensively utilized in machine learning for capturing and generating complex data distributions. In our pursuit of high-quality image generation, we have explored the application of VAEs. However, both theoretical analysis and practical experimentation have demonstrated particular limitations inherent in VAEs. This prompts us to explore a variation of VAEs known as Vector Quantized Variational AutoEncoders (VQ-VAEs).

VAEs, while adept at generating a high diversity of results and ensuring training stability and efficiency, tend to generate blurred results, prompting the search for alternative approaches. The integration of the Vector Quantization (VQ) method with VAEs, known as VQ-VAEs, marks a significant advancement. VQ-VAEs, characterized by their discrete latent space, excel in capturing varied and unique modes within data distributions, yielding crisper image outputs. Embracing VQ-VAEs hinges on our priority to craft distinct images, recognizing the nuanced trade-offs involved—balancing a smooth latent space against sharper visual outputs.

2 RELATED WORK (METHODS)

2.1 IMAGE RECONSTRUCTION AND GENERATION

There are two main kinds of image generative models: likelihood-based models, which include VAEs, normalizing flows, and diffusion models; and likelihood-free models, the most prominent of which is Generative Adversarial Networks (GANs) (Methnani, 2023).

- Likelihood-Based Models: These models are trained by maximizing the likelihood of the observed data, which aims to find the best parameters that explain the samples (Methnani, 2023). The most popular subclass of likelihood-based generative models is VAE.
- Likelihood-Free Models: GANs follow a different approach, which consists of two neural networks: a generator and a discriminator. The discriminator aims to distinguish between real and generated images, and the generator learns how to generate realistic images that can deceive the discriminator. By iterating through this process, GANs are able to generate highly realistic images (de Souza et al., 2023).

VAE and its variant models have a broad application prospect. In recent years, they have played a very important role in data generation, especially in image generation (Wang et al., 2023). Therefore, we choose to implement VAE and its variant VQ-VAE to generate images and compare their performance in this work.

2.2 VAE

VAE was first proposed by Kingma et al. in 2013 (Kingma & Welling, 2013). It contains two parts: an encoder and a decoder, which are both neural networks. The encoder compresses the input images into a low-dimensional latent space, and the decoder generates new output images based on the latent space (Wang et al., 2023).

VAE is trained to minimize the loss comprising two components. Firstly, the reconstruction term quantifies the disparity between the decoded data and input data, driving the encoding-decoding

process toward optimal performance. Secondly, the regularization term evaluates the Kullback-Leibler divergence between the learned latent space distribution and a standard Gaussian (N(0,1)), ensuring regularization of the latent space to promote desirable properties for generating new data (Kingma & Welling, 2013).

Traditional VAE has many hypothetical preconditions and constraints, and the generated images are often fuzzy (Wang et al., 2023). Therefore, many improvements and variants have been proposed, such as Vector Quantized VAE (VQ-VAE) (Van Den Oord et al., 2017), Conditional VAE (CVAE) (Sohn et al., 2015), Very Deep VAE (VDVAE) (Child, 2020).

2.3 VECTOR QUANTIZATION (VQ)

Vector quantization, a signal compression technology that has gained prominence since the previous century, is commonly employed in the realms of audio, image, and video compression processing. In contrast to scalar quantization, which entails the rounding of individual numbers, vector quantization operates on entire sets of numbers known as input vectors, with corresponding quantization levels referred to as reproduction vectors(Cosman et al., 1993).

The architecture of vector quantization technology is similar to that of an autoencoder, typically comprising an encoder, a decoder, and a codebook. The encoder is responsible for assigning a sequence of input vectors to the code vector within the codebook that exhibits the closest proximity. Simultaneously, the encoder adheres to mapping rules during its operations, typically formulated based on distance measures, such as the Euclidean distance.

A codebook, constituting representative vectors generated through clustering algorithms, is established before the vector quantization process begins. Each grouping of vectors in the codebook is denoted as a code vector, efficiently representing each input vector.

The decoder takes the assigned code vector in the codebook. Subsequently, the decoder utilizes this designated code vector to substitute the input vector for reconstruction.

2.4 VQ-VAE

VQ-VAE, as a variant of VAE, its main improvement lies in replacing continuous variables in the latent space with discrete variables. Such discrete variables enhance model performance in complex inference and prediction (Van Den Oord et al., 2017). Specifically, VQ-VAE designs $q(z|x)$, the smoothior and $p(x|z)$, prior distributions as categorical, and the samples drawn from these distributions are indexed into embedded tables. These embeddings are then used as input to the decoder network (Roy et al., 2018). Given a latent embedding space (the codebook) $e \in \mathbb{R}^{K \times D}$, where K is the number of categories, and D is the dimensionality of each code vector e_i, an input x is passed to the encoder to produce $z_e(x)$. By looking up the codebook, $z_e(x)$ matches with the code vector e_k. Then $z_q(x)$ is assigned by e_k, and piped to the decoder to finish the reconstruction (Van Den Oord et al., 2017). In this case, the posterior categorical distribution $q(z = k|x)$ is defined as one-hot, **which means it is deterministic so KL divergence will be constant** ($\log K$).

There are still some tricks worth mentioning during training. We think they are based on the discreteness of the embedding spaces and the design of the Encoder-Decoder network. Primarily, the absence of a gradient between $z_e(x)$ and $z_q(x)$ arises due to the non-differentiability of $q(z = k|x)$. Nevertheless, the authors circumvented this challenge by directly propagating gradients from the decoder input $z_q(x)$ to the encoder output $z_e(x)$, leveraging the symmetrical structure inherent in the design of the encoder and decoder (Van Den Oord et al., 2017).

$$L = \log p(x|z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2$$

In addition, the authors introduce a commitment loss, denoted as the third term in the equation above, to ensure that each encoder output firmly adheres to an embedding e_i. This measure prevents embeddings from exhibiting arbitrary growth. Moreover, the authors incorporate a loss term grounded in Vector Quantization to adjust the embedding vectors e_i to closely align with $z_e(x)$. In this scenario, the influence of the codebook is on the rise.

3 Experiments and Results

3.1 Dataset

Training deep neural networks on ImageNet can be computationally demanding due to its large size and high-resolution images. Therefore, we decided to switch our dataset to CIFAR-10. CIFAR-10 consists of 60,000 32x32 color images, divided into 10 classes, with 6,000 images per class. The ten classes in CIFAR-10 include common objects and animals, including airplanes, birds, etc.

3.2 Experiments

Upon reviewing existing papers on VQ-VAEs, the reconstruction performance evaluation task serves as a primary experimental method. Therefore, we opted to follow a similar approach for our study.

3.2.1 Model Design

We initially trained a conventional VAE model on CIFAR-10 to establish a baseline performance. The model comprises an encoder with four convolutional layers and a decoder with four deconvolutional layers. The latent space is 128 dimensions. Utilizing a learning rate of 5e-4, we assessed the reconstruction performance after 40 epochs across various batch sizes. We trained our VQ-VAE model on CIFAR-10 as well. The network architecture setup is similar to that in Neural Discrete Representation Learning (Van Den Oord et al., 2017). To compare with VAE, we used the same optimizer, maintaining the same learning rate and number of epochs. For training hyperparameters, we set the `num_hiddens=128`, `num_residual_hiddens=32`, `embedding_dim=64`, and `num_embeddings=512`. When we implemented the loss function, instead of using the simplest dictionary learning algorithm to optimize the embeddings in latent space, we chose to implement the Exponential Moving Averages algorithm.

3.2.2 Evaluation of Reconstructed Images

The reconstructed images generated by VAE and VQ-VAE are shown in Figure 1. Notably, the reconstructed images from VAE display a considerable level of fuzziness. This lack of clarity might result from the constraints imposed by the Gaussian assumptions in the encoder/decoder, potentially limiting the VAE's ability to generate more realistic images (Dai & Wipf, 2019). In contrast, the reconstructed images generated by VQ-VAE exhibit significantly enhanced realism, making it challenging to distinguish them from the original images. This illustrates the robust capability of VQ-VAE in reconstructing high-fidelity images.

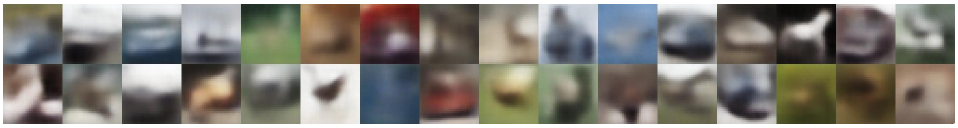
Structural Similarity Index Measure (SSIM) gauges the likeness between two images, providing a valuable metric for assessing image quality against an original, uncompressed, or undistorted reference image. We computed SSIM scores for both VAE and VQ-VAE with different batch sizes. As depicted in Figure 2, the SSIM scores for VAE are just around 0.78, whereas all batch sizes of VQ-VAE surpass 0.96, demonstrating a substantial improvement of image quality.

Furthermore, our examination encompassed an analysis of VQ-VAE's training losses across varying batch sizes, as shown in Figure 3. We noted a consistent downward trend in training loss across these different batch sizes. The initial five epochs displayed a rapid decline, succeeded by a more gradual decrease in subsequent epochs. With a batch size of 256, the model's training loss stabilized at approximately 0.107, accompanied by a corresponding reconstruction loss of approximately 0.055. In contrast, with a batch size of 16, the final convergence of the training loss hovered around 0.076, while the reconstruction loss settled at approximately 0.051.

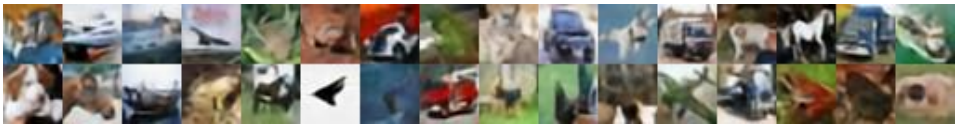
4 Conclusion

In this project, our exploration focused on addressing the challenges faced by VAEs in generating high-quality images, leading us to investigate the VQ-VAE variant for enhanced image reconstruction capabilities. Our findings substantiate that VQ-VAE exhibits remarkable power in reproducing high-quality images, blurring the boundaries between the generated and original images. This highlights the ability of VQ-VAE's discrete latent space to adeptly encapsulate crucial data features.

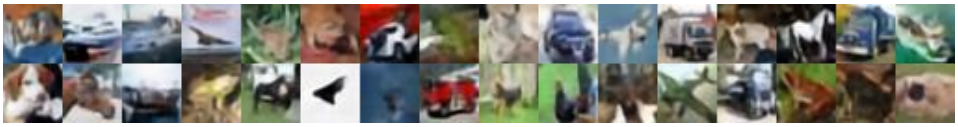
(a) Original Images



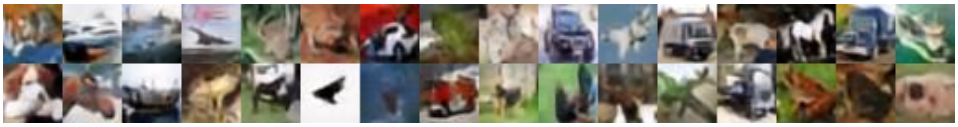
(b) Reconstructed images: VAE (batch size =16)



(c) Reconstructed images: VQ-VAE (batch size =16)



(d) Reconstructed images: VQ-VAE (batch size =64)



(e) Reconstructed images: VQ-VAE (batch size =256)

Figure 1: Reconstructed images using VAE and VQ-VAE

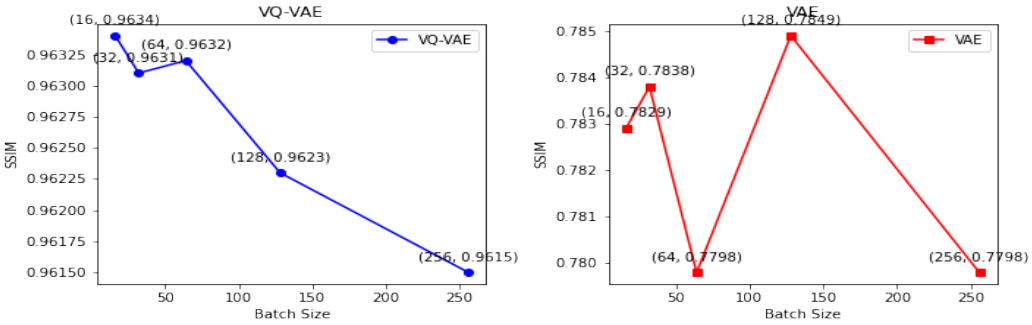


Figure 2: SSIM of VQ-VAE (left) and VAE (right)

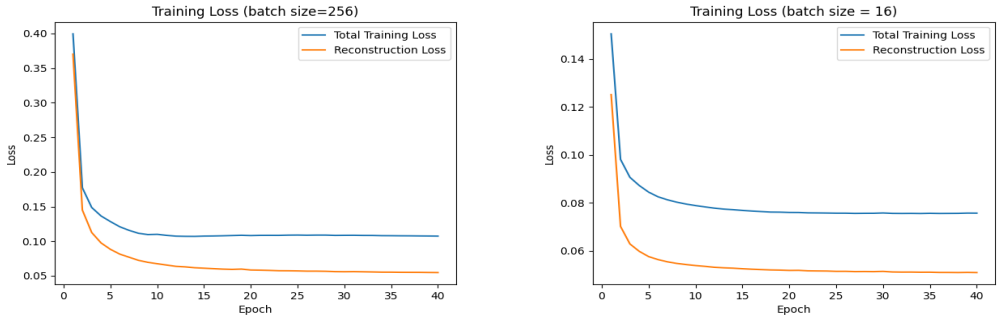


Figure 3: Training loss of VQ-VAE

REFERENCES

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

P.C. Cosman, K.L. Oehler, E.A. Riskin, and R.M. Gray. Using vector quantization for image processing. *Proceedings of the IEEE*, 81(9):1326–1341, 1993. doi: 10.1109/5.237540.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Vinicius Luis Trevisan de Souza, Bruno Augusto Dorta Marques, Harlen Costa Batagelo, and João Paulo Gois. A review on generative adversarial networks for image generation. *Computers & Graphics*, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Shams Methnani. Deep generative modeling: An overview of recent advances in likelihood-based models and an application to 3d point cloud generation. 2023.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and experiments on vector quantized autoencoders, 2018.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Chuang Wang, Jianwen Song, and Lijun Wang. P-2.9: A review of image generation methods based on deep learning. *SID Symposium Digest of Technical Papers*, 54(S1):507–512, 2023. doi: https://doi.org/10.1002/sdtp.16343. URL `https://sid.onlinelibrary.wiley.com/doi/abs/10.1002/sdtp.16343`.

APPENDIX

Our group's distribution of tasks is as follows:

Boqi Zhao: Developed the VQ-VAE model, performed analyses of its results, and explained VQ and VQ-VAE in the Methods Section.

Hezhi Xie: Developed the baseline VAE model, performed analyses of its results, and explained Image Reconstruction and VAE in the Methods Section.

Yinuo Tang: Contributed to the motivation and dataset section and assisted in results analysis.