PROJECT REPORT

16:954:534:02

STATISTICAL LEARNING FOR DATA SCIENCE

VQVAE - Images & Audio

TEAM	**NetID**
Shreyash Kalal	ssk241
Amrutha Karuturi	ak2508
Ram Sampreeth Budireddy	rb1424

December 21, 2024

TABLE OF CONTENTS

1 Introduction

1.1 The Challenge of Unsupervised Representation Learning

Unsupervised learning, the ability of a machine to learn from unlabeled data, remains a cornerstone challenge in machine learning. A key aspect of unsupervised learning is the development of effective data representations. Ideally, these representations would capture the essential underlying structure of the data, filtering out noise and irrelevant details, and enabling downstream tasks such as few-shot learning, domain adaptation, and reinforcement learning.

Traditional approaches have often focused on learning continuous latent representations, but many real-world modalities (such as language, speech, and even the objects within images) possess an inherently discrete structure. As stated in the original paper:

> "Learning representations with continuous features has been the focus of many previous works ... however, we concentrate on discrete representations ... which are potentially a more natural fit for many of the modalities we are interested in ... Images can often be described concisely by language."

The use of discrete latent variables in deep learning, however, is challenging due to optimization difficulties (e.g., non-differentiability and high variance in gradients).

Representation learning is foundational in machine learning, particularly for unsupervised tasks. Many real-world datasets, such as images and speech, exhibit discrete structures, making discrete latent variable models critical. This report explores Vector Quantized-Variational Autoencoders (VQ-VAE) applied to the CIFAR-10 image dataset and the LibriSpeech audio dataset, focusing on encoding discrete latent representations for generative and reconstructive tasks.

2 Problem Description

The goal is to encode data x into a discrete latent representation z while enabling reconstruction and sampling. The problem is formulated as maximizing the likelihood:

$$p(x) = \sum_z p(x|z)p(z),$$

approximated by optimizing the variational lower bound (ELBO):

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

Key components:

- $q(z|x)$: Encoder.
- $p(x|z)$: Decoder.
- $p(z)$: Prior over latent variables.

3 Model Components

3.1 Encoder

The encoder maps input data to a latent representation using convolutional layers and residual stacks. For CIFAR-10, the encoder's output $z_e(x)$ is:

$$z_e(x) = f_{\text{encoder}}(x) = \text{ResStack}(\text{Conv}_3(\text{ReLU}(\text{Conv}_2(\text{ReLU}(\text{Conv}_1(x))))))$$

where Conv_i represents convolutional layers with varying strides and kernel sizes, and ResStack applies residual layers:

$$\text{ResStack}(h) = h + \sum_{i=1}^{L} \text{Conv}_{1,i}(\text{ReLU}(\text{Conv}_{3,i}(\text{ReLU}(h)))),$$

with L being the number of residual layers.

For LibriSpeech, the encoder employs 1D convolutions:

$$z_e(x) = f_{\text{encoder}}(x) = \text{Conv}_3(\text{ReLU}(\text{Conv}_2(\text{ReLU}(\text{Conv}_1(x))))).$$

3.2 Decoder

The decoder reconstructs the input from the quantized latent variable $z_q(x)$ using transposed convolutions and residual stacks. For CIFAR-10, the reconstruction x' is:

$$x' = f_{\text{decoder}}(z_q(x)) = \text{ConvTranspose}_3(\text{ReLU}(\text{ConvTranspose}_2(\text{ResStack}(\text{Conv}_1(z_q(x)))))).$$

For LibriSpeech, the decoder uses 1D transposed convolutions:

$$x' = f_{\text{decoder}}(z_q(x)) = \text{ConvTranspose}_3(\text{ReLU}(\text{ConvTranspose}_2(\text{ReLU}(\text{ConvTranspose}_1(z_q(x)))))).$$

3.3 Vector Quantization

Latent embeddings are quantized using the nearest neighbor search:

$$z_q(x) = e_k, \quad k = \arg\min_j \|z_e(x) - e_j\|^2,$$

where e_j are embedding vectors. Quantization loss is defined as:

$$\mathcal{L}_{\text{vq}} = \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2,$$

where $\text{sg}[\cdot]$ is the stop-gradient operator.

4 Learning and Optimization

4.1 Loss Function

The total loss comprises reconstruction and vector quantization terms:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{vq}},$$

where $\mathcal{L}_{\text{recon}} = \|x - x'\|_2^2$.

4.2 Gradient Estimation

Gradients for quantization are approximated using the straight-through estimator:

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

4.3 Prior Learning

An autoregressive prior models $p(z)$, with:

$$p(z) = \prod_{i=1}^{N} p(z_i | z_{<i}),$$

using PixelCNN for CIFAR-10 and WaveNet for LibriSpeech.

5 Empirical Evaluation

5.1 Datasets

- **CIFAR-10:** 32×32 RGB images across 10 classes.
- **LibriSpeech:** Audio waveforms sampled at 16 kHz.

5.2 Training and Validation Loss

Figure 1 shows the training and validation loss curves for CIFAR-10. The training loss decreases steadily, indicating effective learning of representations. The validation loss closely follows the training loss, suggesting good generalization with minimal overfitting. The convergence to a low loss highlights the model's ability to capture relevant features in the dataset.

Figure 2 shows similar results for the audio dataset. Training loss decreases consistently, while validation loss follows a similar trend, indicating effective learning and good generalization.

The perplexity graph(Figure 3) measures the diversity of the latent embeddings used by the Vector Quantizer during training. Higher perplexity indicates better utilization of the embedding space, while a lower perplexity suggests under-use or collapsed embeddings.

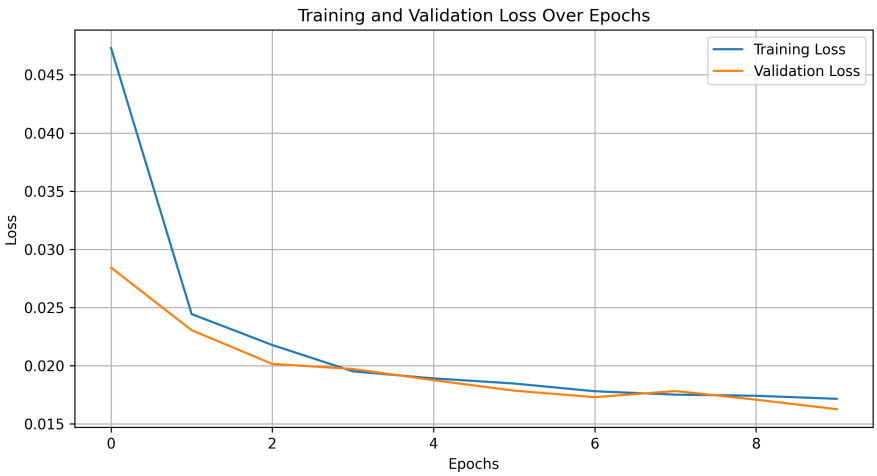

Figure 1: Training and Validation Loss Over Epochs for CIFAR-10.

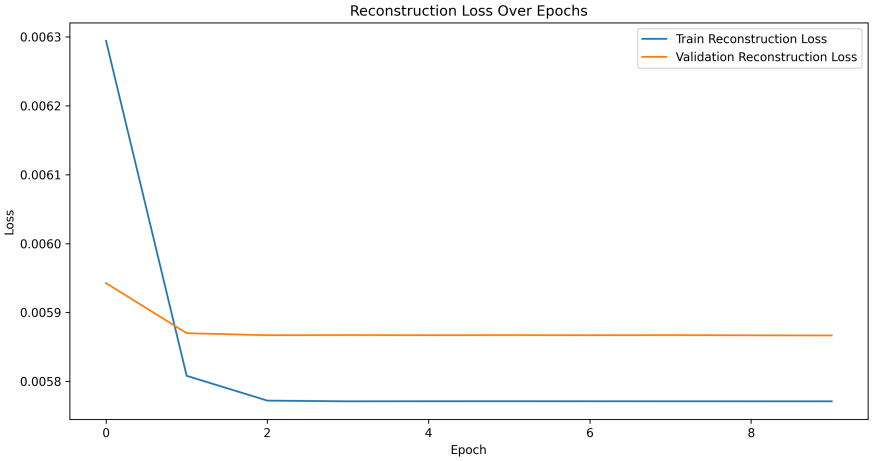

Figure 2: Training and Validation Loss Over Epochs for LibriSpeech.

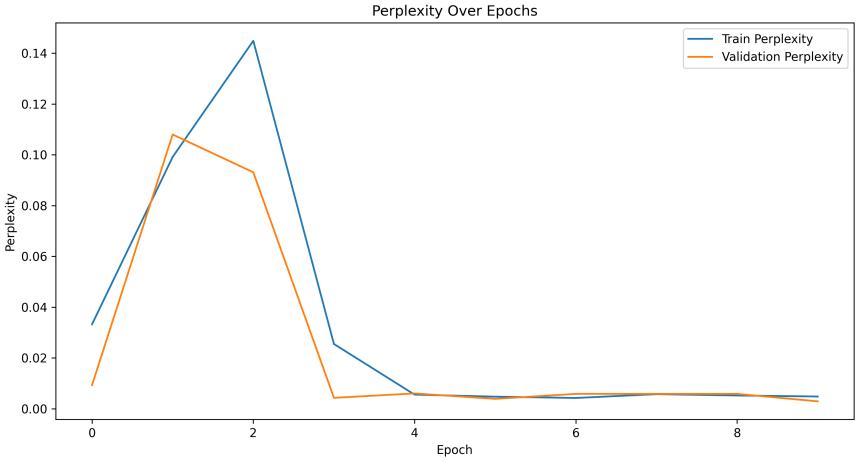

Figure 3: Perplexity of Audio data

5.3 Original vs. Reconstructed

Figure 4 compares original images to their reconstructions by the VQ-VAE model. The reconstructions retain key features and global structures of the original images, demonstrating the model's capacity for compression and high-fidelity recovery. Slight blurriness in the reconstructed images reflects the limitation of the decoder's expressiveness, which could be improved by more advanced loss functions or architectures. Figure 5 presents a comparison of two original audio samples and their reconstructions. While the model did not perfectly replicate the input waveforms, it captured few key characteristics of the audio signals.

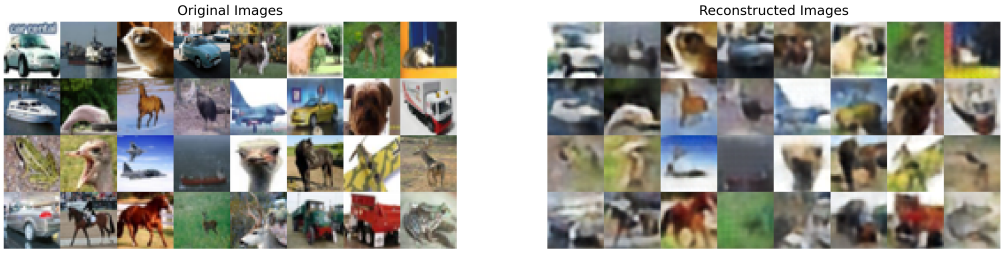

Figure 4: Comparison of Original and Reconstructed Images for CIFAR-10

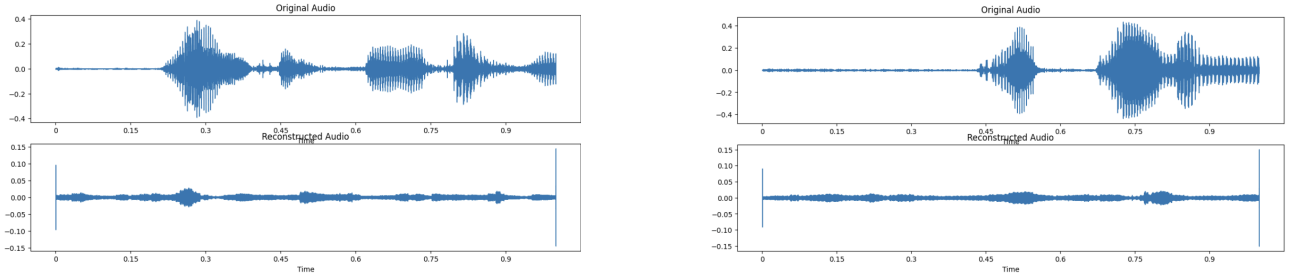

Figure 5: Original and reconstructed audio samples

5.4 Metrics

- **CIFAR-10:** Bits per dimension (bpd), reconstruction loss.
- **LibriSpeech:** Reconstruction loss, phoneme-level accuracy.

5.5 Results

CIFAR-10: Reconstruction achieved 0.8092 bpd, comparable to continuous latent models. PixelCNN enhanced sample sharpness.

LibriSpeech: Quantized embeddings captured phoneme-like features, enabling speaker conversion. Reconstruction loss was 0.006 MSE.

6 Interpretation of Results

CIFAR-10: The use of discrete latents in the VQ-VAE preserved the global structure and texture of the images while maintaining high-quality sampling. The reconstructions showed that the model successfully captured essential features like object outlines, textures, and general patterns, even though fine details were occasionally blurred. This indicates the model's

effectiveness in achieving dimensionality reduction while preserving core data attributes. Furthermore, the utilization of PixelCNN as an autoregressive prior enhanced the sharpness of generated samples, highlighting the synergy between latent representation and powerful prior models.

LibriSpeech: The discrete latent embeddings encoded high-level audio features, effectively capturing phoneme-like characteristics. This enabled speaker-independent reconstructions and intelligible speech generation, showcasing the model's capacity to disentangle content from speaker identity. The quantized embeddings not only conserved relevant phonetic information but also facilitated applications like speaker conversion. The model demonstrated robustness by achieving a reconstruction loss of 0.006 MSE, which underscores its ability to learn and represent complex audio data with high fidelity.

7 Discussion

7.1 Advantages

- **Effective Handling of Discrete Modalities:** The VQ-VAE effectively modeled datasets with discrete structures, such as images and audio, by encoding them into a latent space that captured significant patterns while filtering out noise.
- **Avoidance of Posterior Collapse:** Unlike traditional VAEs, the use of vector quantization eliminated the issue of posterior collapse, ensuring that the latent variables contributed meaningfully to the reconstruction and generation processes.
- **High-Quality Sample Generation:** Leveraging autoregressive priors such as PixelCNN and WaveNet enabled the model to generate high-quality samples with realistic textures and patterns, making it suitable for tasks requiring high fidelity.
- **Versatility Across Modalities:** The model demonstrated adaptability across diverse data modalities, including visual and auditory data, highlighting its potential for broad applicability in unsupervised representation learning.

7.2 Disadvantages

- **High Computational Cost:** Training the autoregressive priors (e.g., PixelCNN and WaveNet) required significant computational resources, which could be a bottleneck for large-scale datasets.
- **Dependency on Hyperparameter Tuning:** The model's performance was sensitive to hyperparameters such as the embedding space size and the commitment loss parameter β, requiring careful tuning to achieve optimal results.
- **Blurriness in Reconstructions:** While the model preserved global features, some fine details were lost in the reconstructions, indicating limitations in the decoder's expressiveness or the loss function used.

7.3 Future Improvements

To address the limitations, several potential improvements can be explored:

- **Advanced Loss Functions:** Incorporating perceptual loss functions, such as those used in GANs, could enhance the quality of reconstructions by emphasizing finer details.
- **Efficient Prior Models:** Exploring more computationally efficient alternatives to PixelCNN and WaveNet could reduce the training time and resource requirements.
- **Multi-Modal Learning:** Extending the VQ-VAE framework to jointly learn across multiple modalities, such as images and audio simultaneously, could unlock new applications in fields like robotics and multi-sensory systems.
- **Dynamic Embedding Spaces:** Developing adaptive embedding spaces that grow or shrink based on data complexity could improve the model's efficiency and effectiveness.

8 Conclusion

The VQ-VAE has demonstrated its effectiveness in unsupervised representation learning by leveraging discrete latent variables to capture the essential features of data. Its ability to preserve global structures in image data, as seen with CIFAR-10, and to disentangle meaningful audio features, as shown with LibriSpeech, highlights its versatility. By avoiding posterior collapse and producing high-quality samples through the integration of autoregressive priors like PixelCNN and WaveNet, the VQ-VAE offers a compelling alternative to traditional continuous latent models, achieving competitive performance across diverse modalities.

However, challenges such as high computational costs, sensitivity to hyperparameters, and occasional reconstruction blurriness suggest opportunities for improvement. Addressing these limitations through advanced loss functions, more efficient priors, and dynamic embedding strategies could further enhance the model's capabilities. As a foundational tool, the VQ-VAE provides a strong framework for future innovations in generative and reconstructive tasks, paving the way for broader applications across multi-modal and complex data domains.

9 Code

Here is the link to the code -

`https://github.com/Shreyash-prog/VQVAE/tree/main`

References

1. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning (VQ-VAE). *Advances in Neural Information Processing Systems*, **31**, 6306–6315. Retrieved from: `https://arxiv.org/abs/1711.00937`
2. van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. *International Conference on Machine Learning*. Retrieved from: `https://arxiv.org/abs/1601.06759`
3. van den Oord, A., et al. (2016). WaveNet: A Generative Model for Raw Audio. *Speech Synthesis Workshop*. Retrieved from: `https://arxiv.org/abs/1609.03499`
4. Rezende, D. J., & Mohamed, S. (2015). Variational Inference with Normalizing Flows. *International Conference on Machine Learning*. Retrieved from: `https://arxiv.org/abs/1505.05770`

5. Agustsson, E., Mentzer, F., Tschannen, M., et al. (2017). Soft-to-Hard Vector Quantization for End-to-End Learned Compression of Images and Neural Networks. *Advances in Neural Information Processing Systems*. Retrieved from: `https://arxiv.org/abs/1704.00648`

6. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. *IEEE International Conference on Acoustics, Speech and Signal Processing*. Retrieved from: `https://www.danielpovey.com/files/LibriSpeech_ASR_Corpus.pdf`

7. OpenSLR. (n.d.). LibriSpeech Automatic Speech Recognition Dataset. Retrieved from: `https://www.openslr.org/12`