Image Classification Using Discrete Representations

Barto Radman
Tilburg University / M CSAI
b.radman@tilburguniversity.edu

Abstract

This study investigates the performance of continuous and discrete latent spaces in variational autoencoders (VAEs) and vector quantized VAEs (VQ-VAEs) for image classification and reconstruction tasks. Using two image datasets, latent representations were evaluated through a probing classifier and reconstruction metrics, including mean squared error (MSE) and structural similarity index (SSIM). Results indicate similar performance between the methods on simple data, with continuous spaces outperforming discrete ones in reconstruction quality. Discrete latent spaces demonstrate superior classification performance on the more complex CIFAR100 dataset. Visualizations further reveal effective clustering in discrete spaces, particularly for simpler datasets. These findings underline the potential of discrete latent spaces for feature learning and classification tasks.

1 Introduction

Image classification is one of the most explored uses of deep neural networks. The ability to assign accurate labels without direct human supervision continues to serve as a building block for novel applications. Despite the plethora of research, the problem of image classification is at the core of computer vision systems. Being applied in domains such as earth remote sensing observation technology (Li et al., 2018), computer aided medical diagnostics (Zhang et al., 2019), classification of biological images and their structures (Affonso et al., 2017), and reinforcement learning world model where the agent learns directly from the vision input (Ha & Schmidhuber, 2018; Gelada et al., 2019) - all require labeled data.

Said applications rely on learning useful and efficient representations of large quantities of data, often in the form of latent spaces (Kingma & Welling, 2022). Latent spaces capture features most important for the task through reducing the dimension of the data, while positioning similar data points close to each other in the embedded space (Liu et al., 2019). A popular family of methods, called representation learning methods, utilize variational autoencoders (VAE) to map points from complex data into distributions of continuous latent spaces (Kingma & Welling, 2022). These representations usually take form of continuous valued vectors drawn from a Gaussian distribution (Gelada et al., 2019). The continuous nature of the representation space allows for simple interpolation between data close in proximity.

While continuous spaces assure that close points in the latent space have similar properties in real data, discrete data such as labeled images could be better represented by discrete latent spaces (Oord et al., 2018). In prior works, vector quantization (VQ) techniques are used in combination of VAEs to categorize the encoded distributions, and index them into discrete embedding tables. Thus, discretizing the representations could potentially aid in interpretability and compactness of the latent space (Oord et al., 2018).

VQ-VAEs, a representation learning model learning discrete VAEs latent spaces, showed success in visual and auditory modeling, the potential for using discrete latent spaces directly in image classification remains largely unexplored. This work aims to investigate how discrete latent spaces, as opposed to continuous representations or raw pixel data, can be effectively utilized for image classification tasks.

Thus, the central research question is: *How well do the discrete latent representations of a VQ-VAE correlate with the labels of annotated images?*

Additionally, this paper aims to interpret the reconstructed output of said model.

To address this, the paper focuses on extraction of latent representations, together with assessing the quality of image reconstructions and visualizing the latent space. Experiments on simple data

show parity between the methods, while discrete latent spaces outperform continuous ones on more complex data.

2 Related Work

Image classification methods rely on the use of convolutional neural networks (CNNs) in order to learn useful spatial image features. Generally, the methods work with raw pixel data which is then used to label new unseen data as belonging to one class or another. This is done by learning spatial hierarchical features through applying convolutional layers smaller than input in size. These representations are then used for downstream tasks such as classification (Gelada et al., 2019).

Pretrained classification models like ResNet (Targ et al., 2016) and VGG16 (Tammina, 2019) remain state-of-the-art with high accuracies on standard benchmark datasets such as MNIST and CIFAR100. Such networks are usually deep, containing many hidden layers, and trained on a large amounts of data (Tammina, 2019). Data mixtures for such models further augment the data in order to assure generalization of the models. Due to their size, these networks can be difficult to train and interpret - often being used for transfer learning on new task.

What is in common with these networks is that they represent the data in continuous latent space. One similar method utilized representation learning for classification of human motion (Butepage et al., 2017). In the paper, human dynamics were encoded in order to classify directly produced actions. Authors of the paper show that trained representation learning networks can be utilized for classification and prediction of human motion.

To the same effect, VQ-VAEs can be used to learn more suitable, interpretable features of image data - while having fewer connected layers and lower training time than the transfer learning methods. The property of discrete latent space is that the encoded input vector is embeded into a discrete space by finding the closest vector in a finite set of learnable codebook (Oord et al., 2018). Finite representations further compress the data and project it to a meaningful embedding space - where both the spatial and label information is preserved. While not directly using the decoder part of the network for its results, this work provides some reconstitution examples from the decoder output and evaluations of their structure.

Model	Accuracy	MSE	SSIM
vae_mnsit	**80.05**	**0.002**	**0.981**
vq_mnist	69.60	0.018	0.906
vae_cifar	24.34	**0.001**	**0.950**
vq_cifar	**40.13**	0.064	0.598

Table 1: Classification accuracies for base VAE and VQ-VAE trained on MNIST and CIFAR100 datasets. MSE and SSIM measures for reconstruction similarity (Wang et al., 2004).

Representation learning models, ranging from simple VAEs to VQ-VAEs and their other variations, asses model reconstruction and generation qualities based on mean square errors between the original and reconstructed images, along with structural similarity index, Fréchet inception distance (FID) and inception score (Wang et al., 2004; Liu et al., 2019).

3 Methods

Experiments were conducted on two standard image classification datasets. MNIST is a dataset containing 28x28 grayscale handwritten digit images with ten unique classes (Deng, 2012). CIFAR100 is a more complex dataset of 32x32x3 RGB images with 100 classes and 20 superclasses (Krizhevsky, 2009). Datasets were normalized and split into train (50000) and test (10000) sets. For the purposes of this study, MNIST dataset was used as a base, however with the simple nature of the data, validity of the models was further shown on more difficult to model data.

As a basis for comparison of differences between continuous and discrete representations, standard VAE and VQ-VAE model architectures were used (Oord et al., 2018; Kingma & Welling, 2022). Both of the models shared the same encoder and decoder structure to ensure fair comparison. VQ-VAE model incorporated a vector quantization layer after the encoder block with the number of embedding vectors set to K = 512, each with D = 64 dimensions and commitment cost (beta) of 0.25, directly used to transform the encoded latent space.

Before the data entered the quantization layer, image label embeddings of the labels were added to the data in order to include textual information with the image spatial encoding. Encoder blocks consisted of three strided and padded convolutional layers with leaky ReLu activations and kernel size of four, decoder shared the same amount of layers. Models trained on CIFAR100 further utilized batch

normalization layers to increase the performance for more complex images. Models and their architectures were kept as simple as possible to aid in interpretability of results. Training was conducted with Adam optimizer, learning rate 0.001 and batch size of 64 for 10 epochs each.

To directly address the research question, assesing the correlation between the discrete latent representations of a VQ-VAE and annotated image labels, the models were evaluated by probing with a simple feed forward neural network classifier directly on the continuous and discrete latent representations. Accuracy was used as a measure of how well the representations capture label information. Other metrics were not necessary to use due to the balanced and structured nature of the dataset. The same evaluation and network structure was performed for both models, allowing for a comparison between the model performance on discrete and continuous representations.

Lastly, two reconstruction metrics were computed to assess the quality of the latent representations in preserving image structure (Wang et al., 2004). Reconstruction mean square error and structural similarity index were used due to their ease of interpretation. Some image reconstructions were generated for visual inspection of the models output, along with two visualizations of the discretized latent space for both of the datasets. For the purposes of more clear latent space visualization, CIFAR100 test latents were separated into twenty coarse superclasses which group the hundred previous classes based on using methods found on (https://github.com/ryanchankh/cifar100coarse).

Figure 1: Example reconstructions from the CIFAR100 dataset, VAE model. (Top) Original images (Bottom) Reconstructed images. Continuous representations interpolate between more complex data easier than discrete representations.

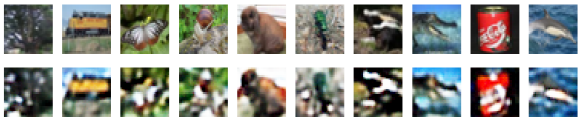

Figure 2: Example reconstructions from VQ-VAE model. Reconstructions suffer from spatial artifacts possibly due to discretization of the latent space before decoding.

4 Results and Discussion

To address the primary research question, the proposed implementation of a simple feed forward neural network was evaluated between variational autoencoder and its discrete variation. As shown in Table 1., original VAE implementation with continuous latent representations outperforms the discrete representations when evaluated on the MNIST dataset, with the VAE model achieving 80 percent accuracy. However, when evaluating on more complex data of colored images, we observe discrete representations outperforming, with 40 percent accuracy for hundred possible classes.

While for the purposes of this study the models were kept simple, the results indicate parity between the information contained in models latent spaces on simple data, however, discretizing the output of the encoder displays higher performance when probing more composite data.

Thus we can conclude that quantizing the vectors, that is categorizing the the posterior and prior distributions and indexing the samples into an embedding table, preserves label and spatial information of the images. Features of the high dimensional CIFAR100 dataset were captured more effectively in categorical spaces, as a product of categorical space being more discriminative for categorical labeled image data.

Study also directly compared the quality of the image reconstructions through calculating mean square error and structural similarity index between the original images and their reconstructions (Table 1). Again when comparing on the baseline MNIST dataset, both of the models achieve low reconstruction errors and high similarities, with the VAE slightly outperforming VQ-VAE. Original VAE model further outperforms vector quantized model on colored images, likely due to the simplicity of the networks and the power of simple encoder block. Figures 1 and 2 present randomly drawn reconstructions of both models for visual inspection.

Continuous latent space interpolation remains state of the art for reconstruction of images from their latents. Discretization causes the reconstructed images to contain artifacts, likely due to constraining of the latent space used to reconstruct the images.

Figure 3 shows t-SNE visualizations of VQ latent spaces for both of the datasets. As seen, the discrete latent spaces cluster well around ten

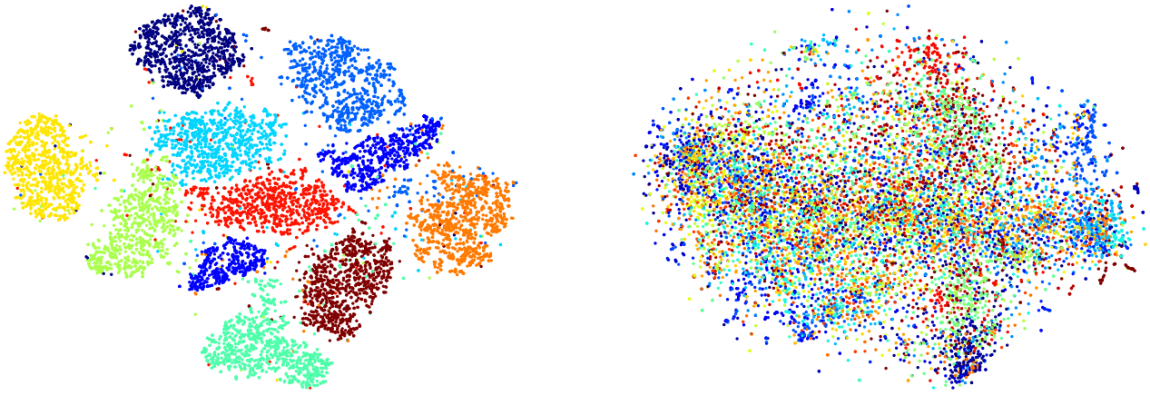

Figure 3: Visualization of VQ latent space for MNIST and CIFAR100 datasets. (Left) Model clusters 10 MNIST classes (Right) but only displays limited cluster success with CIFAR100 superclasses.

MINST classes, while the large amount of CIFAR100 classes show only moderate clustering, even when lowering the amount of classes to twenty superclasses.

This paper has shown that discrete latent spaces contain useful information for classification and can be used for efficient feature learning. It is left to be seen whether utilizing discrete latent spaces for more complicated and diverse categorical distributions can be used in downstream tasks apart from classification.

5 References

Affonso, C., Rossi, A. L. D., Vieira, F. H. A., & De Carvalho, A. C. P. D. L. F. (2017). Deep learning for biological image classification. Expert Systems with Applications, 85, 114–122. https://doi.org/10.1016/j.eswa.2017.05.039

Butepage, J., Black, M. J., Kragic, D., & Kjellstrom, H. (2017). Deep Representation Learning for Human Motion Prediction and Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1591–1599. https://doi.org/10.1109/CVPR.2017.173

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142. (n.d.).

Gelada, C., Kumar, S., Buckman, J., Nachum, O., & Bellemare, M. G. (2019). DeepMDP: Learning Continuous Latent Space Models for Representation Learning (arXiv:1906.02736). arXiv. https://doi.org/10.48550/arXiv.1906.02736

Ha, D., & Schmidhuber, J. (2018). World Models. https://doi.org/10.5281/zenodo.1207631

Https://github.com/ryanchankh/cifar100coarse. (n.d.). [Computer software].

Kingma, D. P., & Welling, M. (2022). Auto-Encoding Variational Bayes (arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images.

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. WIREs Data Mining and Knowledge Discovery, 8(6), e1264. https://doi.org/10.1002/widm.1264

Liu, Y., Jun, E., Li, Q., & Heer, J. (2019). Latent Space Cartography: Visual Analysis of Vector Space Embeddings. Computer Graphics Forum, 38(3), 67–78. https://doi.org/10.1111/cgf.13672

Oord, A. van den, Vinyals, O., & Kavukcuoglu, K. (2018). Neural Discrete Representation Learning (arXiv:1711.00937). arXiv. https://doi.org/10.48550/arXiv.1711.00937

Tammina, S. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications (IJSRP), 9(10), p9420. https://doi.org/10.29322/IJSRP.9.10.2019.p9420

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in Resnet: Generalizing Residual Architectures (arXiv:1603.08029). arXiv. https://doi.org/10.48550/arXiv.1603.08029

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Zhang, J., Xie, Y., Wu, Q., & Xia, Y. (2019). Medical image classification using synergic deep learning. Medical Image Analysis, 54, 10–19. https://doi.org/10.1016/j.media.2019.02.010