Enhancing MNIST Image Generation: VAEs with KL Divergence and GMM Techniques

This final project is a requisite for the 'Data Science Analysis' course, PH6130

1st Koushik Maji
Dept. Artificial Intelligence
Indian Institute of Technology, Hyderabad
AI23MTECH11004

2nd Rahul Verma
Dept. Artificial Intelligence
Indian Institute of Technology, Hyderabad
AI23MTECH11008

Abstract—**This project investigates the use of Variational Autoencoders (VAEs) for generating MNIST images through two approaches. The first approach employs a standard VAE with a loss function combining reconstruction error and Kullback-Leibler (KL) divergence regularization to promote smooth latent space mapping.**

The second approach introduces a novel method that omits the KL divergence term from the loss function and applies a Gaussian Mixture Model (GMM) to the latent space post-training. This strategy aims to enhance the diversity and quality of generated samples by modeling more complex latent distributions.

Comparing these approaches assesses their effectiveness in generating MNIST images in terms of quality, variety, and coherence. The findings provide insights into the trade-offs between regularization and flexibility, contributing to the advancement of robust generative models for digit recognition and similar applications.The code for this project is available in a Kaggle notebook.

I. Introduction

The MNIST (Modified National Institute of Standards and Technology) dataset, comprising a large database of handwritten digits, serves as a benchmark for evaluating machine learning models in the context of visual recognition tasks. The dataset's simplicity and ubiquity make it an ideal testbed for exploring generative modeling techniques such as **Variational Autoencoders (VAEs)**.

VAEs are a class of probabilistic autoencoders designed to learn low-dimensional representations of high-dimensional data and generate new data samples from the learned latent space. They excel in tasks such as data generation due to their capacity to map input data to a distribution, rather than a fixed vector. However, the conventional VAE approach presents certain challenges. Specifically, the model's performance is contingent upon the balance between the reconstruction term and the regularization term in the loss function, based on **Kullback-Leibler (KL) divergence**. This regularization helps enforce meaningful latent representations but can also limit the diversity and complexity of generated samples.

In this project, we investigate two approaches to address these challenges. The first approach involves a traditional VAE setup, where the model is trained using a loss function combining a reconstruction error and a KL divergence regularization term. This regularization constrains the latent space, promoting a smooth and continuous mapping of inputs to a learned distribution.

The second approach proposes a novel modification by omitting the KL divergence term from the loss function. Instead, a **Gaussian Mixture Model (GMM)** is applied to the latent space after training the VAE. This approach aims to overcome the limitations of the regularization term by modeling more complex and flexible latent distributions, potentially enhancing the quality and diversity of generated samples.

The primary objective of this project is to assess the effectiveness of these two approaches in generating MNIST images, particularly in terms of the quality, variety, and coherence of the generated data. The comparison will provide insights into the trade-offs between regularization and flexibility in the latent space, contributing to the development of more robust and efficient generative models for digit recognition and other related tasks.

II. MNIST Dataset

The MNIST dataset is a widely recognized benchmark in the fields of machine learning and computer vision. It comprises $70,000$ grayscale images of handwritten digits (0-9), each sized 28x28 pixels. The dataset is split into two subsets: $60,000$ images for training and $10,000$ images for testing. Each image is labeled with the corresponding digit, providing a comprehensive and well-balanced dataset for training and evaluating models.

A. Statistics

- **Size**: 70,000 images (60,000 training, 10,000 testing).
- **Image Dimensions**: 28x28 pixels, grayscale.
- **Classes**: 10 (digits 0-9).
- **Data Distribution**: Balanced distribution across all classes.

B. Reason for Choosing MNIST

- **Benchmarking**: MNIST serves as a standard reference for evaluating model performance, allowing direct comparisons across different methods.
- **Accessibility**: The dataset is publicly available and well-documented, making it easy to use for research purposes.
- **Simplicity**: MNIST's low complexity provides a controlled environment for testing new algorithms and fine-tuning models.
- **Historical Significance**: MNIST has played a pivotal role in advancing the field of machine learning and continues to serve as a proving ground for innovative techniques.

Fig. 1 shows few samples of MNIST images, showcasing one example from each digit class (0-9).

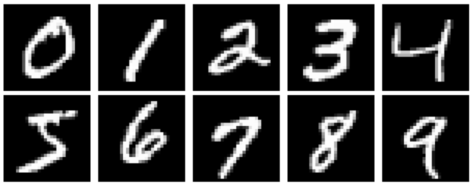

Fig. 1. MNIST Images from different classes

III. Methods

A. Vanilla VAE Architecture

Consider a generative model with a conditional distribution $p(x|z, w)$ over the D-dimensional data variable x governed by the output of a deep neural network $g(z, w)$. For example, $g(z, w)$ might represent the mean of a Gaussian conditional distribution. Also, consider a distribution over the M-dimensional latent variable z that is given by a zero-mean unit-variance Gaussian:

$$p(z) = N(z|0, I).$$

To derive the VAE approximation, first recall that, for an arbitrary probability distribution $q(z)$ over a space described by the latent variable z, the following relationship holds:

$$\ln p(x|w) = L(w) + KL(q(z) \parallel p(z|x, w)).$$

where L is the evidence lower bound, or ELBO, also known as the variational lower bound, given by

$$L(w) = \int \frac{p(x|z, w)p(z)}{q(z)} \ln q(z) dz.$$

The Kullback-Leibler divergence (KLD) regularization in VAEs plays a critical role in shaping the continuity of the latent space and the quality of generated samples. Minimizing the KLD between the approximate posterior distribution $q(z|x)$ and the prior distribution $p(z)$ (typically $N(z|0, I)$) ensures a smooth and continuous latent space:

$$\mathrm{KL}(q(z|x) \| p(z)) = \int q(z|x) \ln\left(\frac{q(z|x)}{p(z)}\right) dz.$$

The smooth latent space leads to gradual and realistic variations in generated samples, facilitating effective interpolation and high-quality, diverse outputs.

B. VAE with GMM

In this work, we propose a novel approach to improve the diversity of the latent space in Variational Autoencoders (VAEs). Our method involves modifying the loss function and fitting a Gaussian Mixture Model (GMM) to the latent space.

The standard loss function for VAEs consists of two terms: the reconstruction loss (BCE) and the Kullback-Leibler divergence (KLD). The KLD term encourages the latent variables to follow a unit Gaussian distribution. However, this can lead to a lack of diversity in the learned representations, as different classes may end up being represented close together in the latent space.

To address this issue, we propose a modified loss function:

$$L = L_{\text{BCE}} + \lambda L_{\text{Center}}$$

where L_{Center} is the center loss and λ is a hyper-parameter controlling the strength of the regularization. Center loss is a strategy for constructing widely-separated classes. It augments the standard supervised loss by adding a penalty term proportional to the distance of a class's examples from its center. It is defined as,

$$L_{\text{Center}} = \frac{1}{2} \sum_{i=1}^{m} \|\mathrm{x}_i - \mathrm{c}_{y_i}\|_2^2$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the deep feature of the i-th training sample (i.e., taken from the latent space), $\mathbf{c}_{y_i} \in \mathbb{R}^d$ denotes the center of class label y_i, d is the feature dimension. This encourages the model to learn widely-separated class representations. It tries to increase the inter-class distance of the embeddings and decrease the intra-class distance for embeddings of each class.

Under the assumption that each class in the latent space follows a Gaussian distribution, we propose to fit a GMM to the latent space. The GMM consists of 10 Gaussians, corresponding to the 10 classes of digits in the dataset.

After training the VAE with the modified loss function, we fit a GMM to the learned representations in the latent space. Each Gaussian in the GMM corresponds to one of the 10 classes. This allows us to generate new data samples that are likely to belong to a specific class by sampling from the corresponding Gaussian.

IV. Results

A. Outcomes from the Vanilla Variational Autoencoder

The model was trained employing a Vanilla Variational Autoencoder (VAE) with Kullback-Leibler (KL) divergence. The distribution of the latent space is visualized in a three-dimensional plot, as shown in 2. Each class is represented by

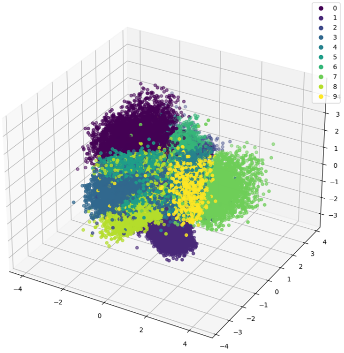

Fig. 2. Clusters obtained from vanilla VAE

a distinct color in this plot. It is noteworthy that some images do not appear to represent any specific digit. These images are sampled from a region in the latent space that represents the intersection between two classes. Consequently, these images may not hold meaningful representations.

The synthetic data generated from the Vanilla VAE is depicted in 3. This figure provides a comprehensive view of the synthetic data produced by the model and using 6-dimensional latent space. The implications of these results and their impact on the overall performance of the model are discussed in the following sections.

Please refer to Figures 2 and 3 for a more detailed understanding of these results.

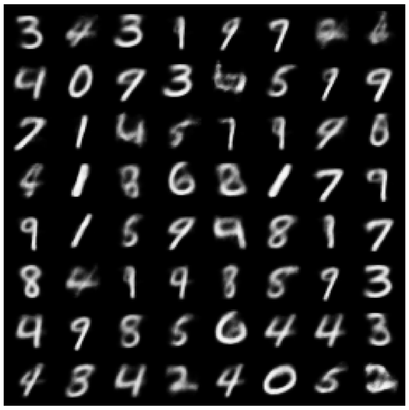

Fig. 3. Digits generated using vanilla VAE

B. Outcomes from the Variational Autoencoder with GMM and Center Loss

In this section, we present the results obtained from the Variational Autoencoder (VAE) model that was trained using a Gaussian Mixture Model (GMM) and center loss (CL). The model's performance was evaluated based on the quality of the learned representations in the latent space and the synthetic data generated.

The distribution of the latent space is visualized in a three-dimensional plot 4. It is evident from the plot that the latent space is well-separated, indicating that the center loss has played a significant role in enhancing the diversity of the learned representations. This is a marked improvement over the results obtained from the Vanilla VAE model, where the latent space showed overlapping regions between different classes.

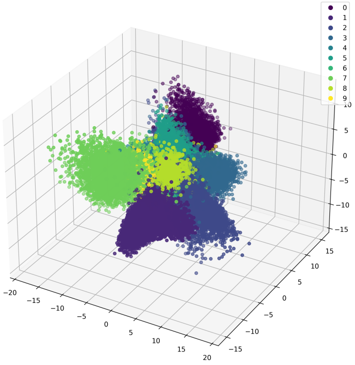

Fig. 4. Clusters obtained from VAE with GMM

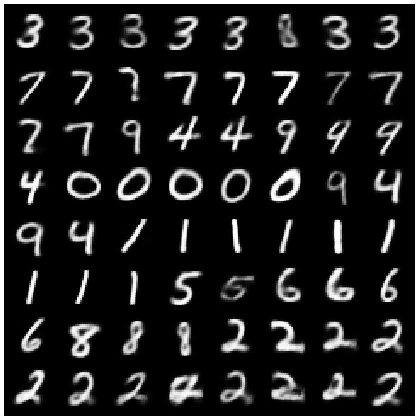

Fig. 5. Digits generated using VAE with GMM

The synthetic data generated from the VAE model with GMM and CL with a 6-dimensional latent space is depicted in 5. We have strategically reduced the variance of each Gaussian by a factor of 1.5 in the GMM to generate samples that are closer to the class mean. This approach ensures that the generated synthetic data is more likely to belong to a specific class, thereby improving the interpretability of the learned representations. A noticeable reduction in the number of redundant images is observed, which signifies the effectiveness of our proposed method in generating diverse and meaningful representations.

The mean of each class in the latent space is printed out in 6. We have strategically reduced the variance of each Gaussian in the GMM to generate samples that are closer to the class mean. This approach ensures that the generated synthetic data

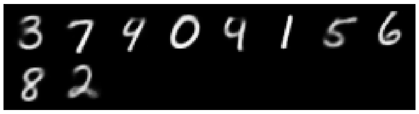

Fig. 6. Means of each Gaussian generated using VAE with GMM

is more likely to belong to a specific class, thereby improving the interpretability of the learned representations.

V. Conclusion

Our proposed method provides a novel way to improve the diversity of the learned representations in VAEs. By modifying the loss function and fitting a GMM to the latent space, we encourage the model to learn more scattered and class-specific representations. This approach could be beneficial in various applications where diverse and interpretable representations are desired.

The results demonstrate the effectiveness of our proposed method in improving the diversity and interpretability of the learned representations in VAEs. By fitting a GMM to the latent space and modifying the loss function with center loss, the model is encouraged to learn more scattered and class-specific representations. This approach holds promise for various applications where diverse and interpretable representations are desired.

References

[1] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” arXiv preprint arXiv:1312.6114v11 [stat.ML], 2013. [Online]. Available: https://doi.org/10.48550/arXiv.1312.6114

[2] Q. T. Ngo and S. Yoon, “Weighted-center Loss for Facial Expressions Recognition,” in 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 54-56. doi: 10.1109/ICTC49870.2020.9289472.

[3] L. Yoon, “PyTorch MNIST VAE,” GitHub repository, 2018. [Online]. Available: https://github.com/lyeoni/pytorch-mnist-VAE/blob/master/pytorch-mnist-VAE.ipynb