Conrad Niewienda
Marcos Sanson

October 8, 2024

Lab 1: Implementing a Multi-View VAE for MNIST-FMNIST

Instructions: Jointly implement a VAE to model MNIST and FMNIST datasets. Both can be easily downloaded from the Pytorch vision package (https://pytorch.org/vision/main/datasets.html).
Every data point X is a pair of MNIST-FMNIST images corresponding to the same category. You can find the FMNIST labels in this link.

- It's highly recommended to first write down the generative model, the inference family, and the overall loss.
- You can base your implementation on some repository that you find out there. But ensure you understand the details; what is the point otherwise?
- When your model is roughly running, show a TSNE embedding of data points and how you implement the cross-domain generation. Namely, generate FMNIST from MNIST and vice versa.
- Summarize everything in a report with a link to your code (use Google Drive or Github)
- You can work individually or in groups of up to three people.

Conrad Niewienda
Marcos Sanson

October 8, 2024

Objective

The objective of this lab project was to set up a multi-view Variational Autoencoder (VAE) that utilizes paired data from the MNIST and Fashion-MNIST (FMNIST) datasets to generate corresponding samples across these two datasets. Specifically, we aimed to enable the VAE to generate FMNIST images (e.g., clothing items) from MNIST digit samples and vice versa, using the multi-view structure to enhance cross-dataset generation accuracy.

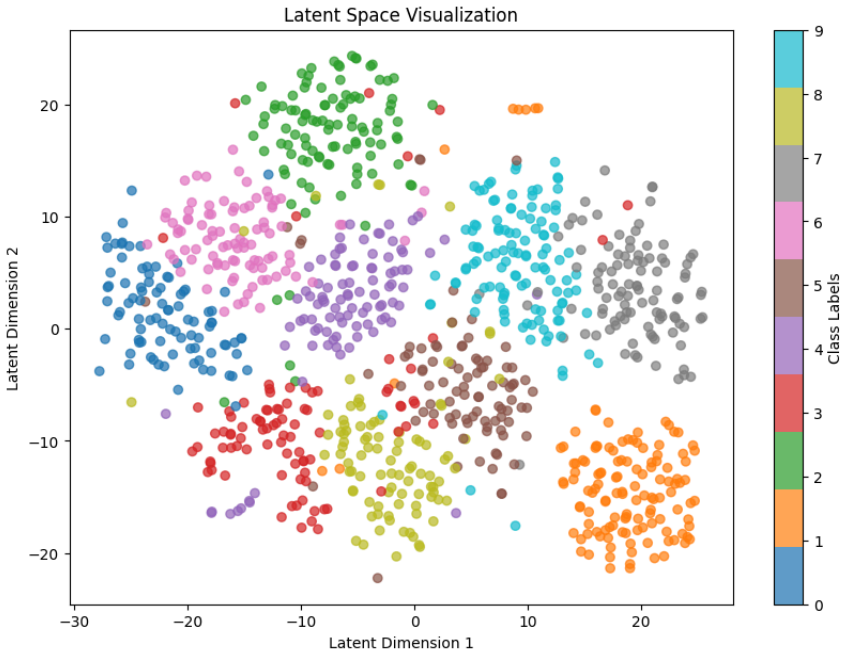
Methodology

We implemented a multi-view VAE, where two encoders and decoders were used to process the MNIST and FMNIST datasets, respectively. The latent space was shared across both views, which enabled the model to learn a common representation that could translate between the two datasets. Training was performed on paired samples where each MNIST image corresponded to an FMNIST image with the same label (e.g., MNIST “1” digit paired with FMNIST “Trouser”). This setup allowed the model to capture the relationship between visually distinct, but semantically similar, classes in both datasets.

The training process involved minimizing the reconstruction loss and the Kullback-Leibler (KL) divergence to ensure that the learned latent space approximates a Gaussian distribution. We experimented with various hyperparameters to optimize the performance of the VAE on cross-dataset generation.

Key Learnings

1. **Gaussian Structure in Latent Space**: Another key insight from our project was that as the VAE training improved, the latent space distribution approached a Gaussian distribution. This pattern, consistent with the VAE's design, indicates that a well-trained VAE will exhibit a smooth, approximately normal latent space, which is important for stable and high-quality cross-dataset generation.



2. **Learning Metadata with Paired Labels**: Using paired data with matching labels across MNIST and FMNIST enabled the VAE to capture metadata, effectively learning that an MNIST “1” corresponds to a FMNIST “Trouser”, an MNIST “0” to an FMNIST “T-shirt/top”, and so on. This pairing allowed the VAE to associate abstract characteristics between datasets, resulting in more accurate representations and effective cross-dataset mappings.
3. **Complexity of Data and Generation Difficulty**: Our experiments revealed that generating MNIST digits from reconstructed FMNIST images was less accurate and less readable compared to generating FMNIST images from MNIST digit samples. This highlights the challenges associated with cross-dataset generation in a multi-view VAE setup.
    a. Several factors may contribute to this, including overfeeding of the FMNIST dataset. Additionally, encoding might access the prior more frequently, leading to a more separated MNIST dataset, while FMNIST clothing data may exhibit greater overlap.
    b. Potential solutions include stopping training earlier or adjusting the loss function to better balance the Evidence Lower Bound (ELBO) reconstruction and Kullback-Leibler divergence (KLD). Without these adjustments, the data could be pushed farther apart in the latent space. More experiments, testing, and fine-tuning with loss comparisons could provide further insights.

Conclusion

This project highlighted the potential of multi-view VAEs to handle complex, cross-domain generation tasks. By aligning the latent space with paired MNIST and FMNIST data, we achieved a model that can generalize across distinct yet related datasets, while also underscoring the limitations imposed by data complexity on generative tasks. The insights on latent space structure and metadata learning also provide valuable guidelines for future work on multi-view generative models.

Link to code:

Google Colab:
https://colab.research.google.com/drive/1tgiVLylsvObbWxy2Zib1KcMVw2KZ3wm1?usp=sharing

GitHub:
https://github.com/Marcos-Sanson/UC3M-ML/blob/main/LAB01.ipynb