CS179 Project: Image Generation

Kyle Yin Xu and Zackery He

Introduction

For this project, we studied class-conditional image generation on Fashion-MNIST. The goal was to generate and compare 28 by 28 pixel grayscale clothing images that match a requested class such as sneaker, bag, dress, or ankle boot. We compared two approaches:

- Conditional Variational Autoencoder (CVAE): learns a compact latent representation and decodes random latent vectors together with a class label.
- Class-Conditioned Denoising Diffusion Model (DDPM): starts from noise and repeatedly denoises it while conditioning on the requested class.

They minimize the following loss functions:

$$\mathcal{L}_{CVAE} = \mathrm{BCE}(x, \hat{x}) + D_{KL}(q(z|x, y) \,\|\, \mathcal{N}(0, I))$$

$$\mathcal{L}_{DDPM} = \mathbb{E}[\|\varepsilon - \varepsilon_\theta(x_t, t, y)\|_2^2]$$

Resources

We used the Fashion-MNIST dataset with 60,000 training images and 10,000 test images. The implementation used PyTorch and TorchVision for the main modeling and data-loading code, Matplotlib for plotting, SciPy for the feature-FID calculation, and Hugging Face's diffusers library for the DDPM scheduler and UNet implementation.

We wrote three notebooks notebooks ourselves, one for CVAE training, reconstruction, and generation, class-conditioned DDPM training and generation, and shared benchmark code for loading checkpoints (of trained models), generating samples, timing generation, computing conditional accuracy, computing feature-FID, and making comparison plots.

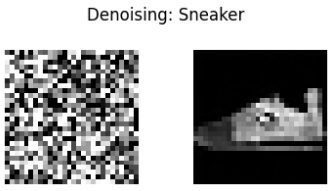

Figure 1: Example denoising steps from our diffusion model.

Methodology

We compared the models on three practical questions: how well they learned during training (training loss), how recognizable and class-consistent their generated images were (sample quality), and how long the generation took (sampling efficiency).
The comparison used the same dataset, class labels, and benchmark procedure for both models. After training each model for 10 epochs, we generated an equal number of samples from every Fashion-MNIST class, converted the outputs into the same pixel range, and evaluated them with the same classifier-based metrics and timing code. This made the comparison more about the modeling approach than about differences in preprocessing or evaluation setup. Each model was asked to generate 200 independent samples for each of the 10 classes. In order to measure our qualifications for a strong image generator, we used the following metrics:

1. **Conditional Accuracy:** For class consistency, we used conditional accuracy. After generating images for a requested class, we passed those images through a trained Fashion-MNIST CNN and checked whether the classifier predicted the intended label. For example, if the model was asked to generate sneakers, the sample counted as correct if the classifier also predicted "sneaker."
2. **Frechet Inception Distance (FID):** For overall sample quality, we used a Fashion-MNIST-specific version of FID, or Fréchet Inception Distance. Instead of comparing raw pixels, FID compares feature distributions from real and generated images. Because Fashion-MNIST is grayscale and much simpler than natural-image datasets, we used features from our Fashion-MNIST classifier rather than the usual Inception network. We ran real images from the test split through the convolutional layers to extract latent space embeddings for each of the real images, then we ran the same process for the CVAE and DDPM synthetic images and measured the Frechet distance between distributions from 2,000 generated images and the baseline dataset. Lower FID means the generated images are closer to the real test images in the classifier's feature space.
3. **Generation time:** For sampling efficiency, we measured the average seconds per image. The timing code recorded how long each model took to generate samples for every class, then divided by the number of images. This was especially important because the CVAE generates an image in one decoder pass, while the DDPM requires many denoising steps per image.

Data and Analysis

Conditional Accuracy

Across almost all classes, the diffusion-based model showed higher alignment with the classifier. The CVAE averaged an accuracy of 0.632 while the DDPM scored 0.831. While the DDPM was relatively consistent across most classes, the CVAAE appeared to excel at some classes (trousers, ankle boots, sneakers), but struggle greatly on others (T-shirts, sandals, shirts).

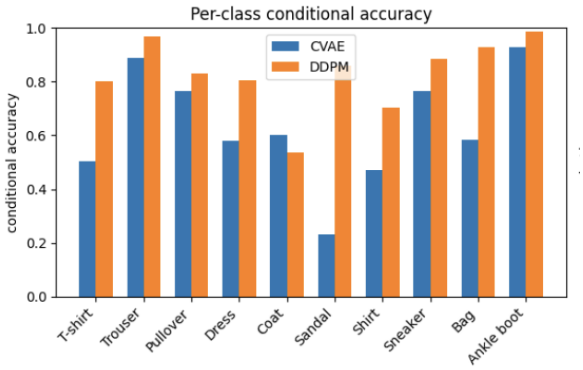

Figure 2: Conditional accuracy between CVAE and DDPM models for each class.

If we look at individual samples, we can observe the differences in how the CVAE and DDPM models function. Since the CVAE learns latent-space vectors shared across each class, it has a tendency to average out the concept of each class, which can explain the apparent blurriness in each of the generated images. Noticeably, the CVAE also tends to draw blurred squares (possibly of different sizes to correspond with how much of the image each of the classes typically tends to take up). Lots of the misclassifications shown may be due to this overall "unsureness" the model possesses.

Meanwhile, the DDPM samples show a completely different kind of error. The generated images show solid lines and shapes reminiscent of the original dataset, but we can see many hallucinations. For example, the model may hallucinate another set of sleeves, add random strips of fabric, or attempt to draw an object of another class. In contrast with the CVAE, the failures from the diffusion are very clear in what they portray, but show fundamental issues that go against what the classes actually represent. While they are more rare than the blurs from the CVAE, they show that the DDPM can be wrong and confidently so.

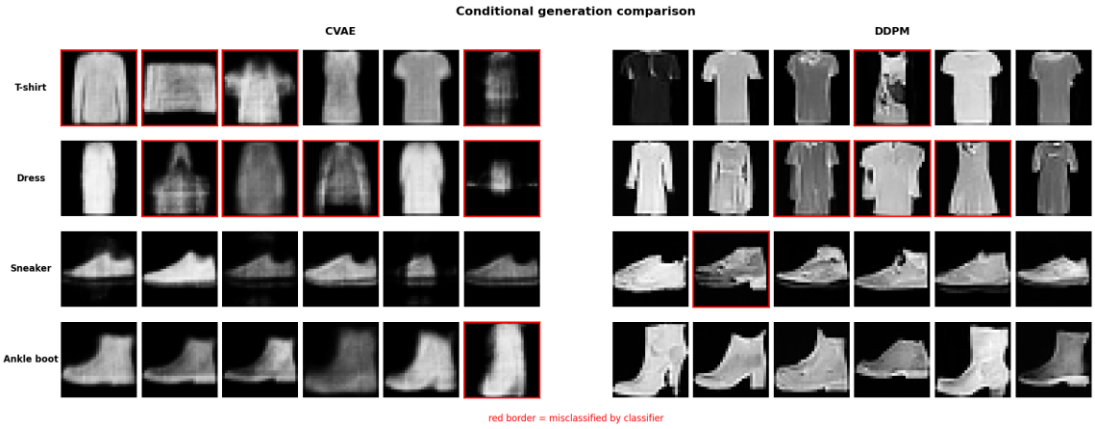

Figure 3: Examples of generated images from both models (highlighted red for misclassified).

Frechet Inception Distance

The FID results from our experiments reinforce the conclusions made by our conditional accuracy metric. The CVAE's average distance from the real dataset was 124.41 while the DDPM averaged 40.78.

While these numbers are calculated based on embeddings without real units, their scale shows that the diffusion model's images tended to be far more aligned with the classifier's idea of the concepts in comparison to the CVAE.

Generation Time

While the DDPM outperforms the CVAE in output performance, our last metric is meant to measure its efficiency in doing so. Generating 2,000 samples for the diffusion model took around 12 minutes in total, while the CVAE was near instantaneous. On average, the diffusion model took 0.3676ms per image, while the CVAE took 0.0000ms (our metrics rounded off at four decimals).

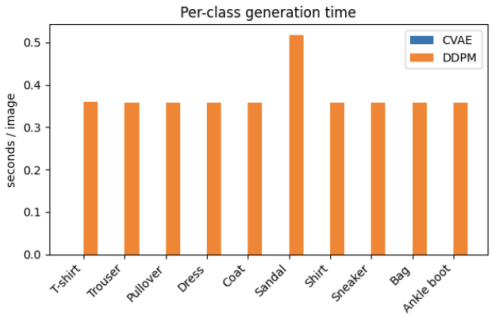

Figure 4: Generation times for each class between the two models.

Conclusion

The experiment showed a clear quality-speed tradeoff across the specific metrics we measured. For training convergence, both models improved over 10 epochs: the CVAE loss decreased from 285.55 to 237.99, while the DDPM noise-prediction MSE decreased from 0.0672 to 0.0370. For conditional fidelity, the DDPM performed better, reaching 0.831 conditional accuracy compared with 0.632 for the CVAE. For sample quality, the DDPM also had the better feature-FID score, 40.78 versus 124.41 for the CVAE, meaning its generated images were closer to real Fashion-MNIST images in the classifier feature space.

The main advantage of the CVAE was sampling efficiency. Its generation time printed as 0.0000 seconds per image, effectively under 0.0001 seconds per image in this benchmark, while the DDPM took 0.3736 seconds per image because it used 200 denoising steps. Overall, the DDPM was the stronger generator when judging by conditional accuracy and feature-FID, while the CVAE was the better choice for speed.

To summarize, the DDPM produced better and more class-accurate Fashion-MNIST images, but the CVAE generated images much faster. The better choice depends on whether image quality or sampling speed matters more for a given use case.