Technical Report on Generative Models for Image-to-Image Translation and Generation

Wasif Mehboob
Department of Data Science
Fast University (NUCES), Islamabad

***Abstract*—This report presents our work on implementing a variety of deep generative models to perform image generation and translation tasks. We describe our experiments that involve:**

- **A Variational Autoencoder (VAE) and a simple Generative Adversarial Network (GAN) developed to generate synthetic signature images.**
- **A custom GAN tailored for the CIFAR-10 dataset (using only cats and dogs) that uses a similarity-based, Siamese-style discriminator.**
- **A Conditional GAN (cGAN) that conditions on input sketches to generate realistic face images.**
- **A CycleGAN that handles bidirectional image-to-image translation between real face images and sketches.**

The models are designed with different architectures, loss functions, and data augmentation techniques to overcome the challenges of limited datasets and to extract meaningful latent representations. We discuss our training strategies, present our observed loss curves, and provide qualitative results along with challenges we encountered.

***Index Terms*—Generative Models, Variational Autoencoder, GAN, Conditional GAN, CycleGAN, Image-to-Image Translation, Data Augmentation.**

I. Introduction

In recent years, deep learning has made tremendous strides in generating and translating images. In this project, we had a chance to explore various generative models by implementing and testing different architectures. Our work focuses on four main tasks:

1) **Question #1:** We built a Variational Autoencoder (VAE) and a simple GAN to generate synthetic signature images. Because our signature dataset was quite small, we used different data augmentation techniques such as scaling, rotation, and noise addition to enrich the data, ensuring that the models could learn a robust latent representation.
2) **Question #2:** We implemented a custom GAN on the CIFAR-10 dataset (filtered for cats and dogs). Unlike traditional GANs that simply classify images as real or fake, our discriminator is designed in a Siamese style so that it takes a generated image and a real image together and outputs a similarity score.
3) **Question #3:** We created a Conditional GAN (cGAN) that utilizes the Person Face Sketches dataset. The goal is to enable the generator to learn the mapping from an input sketch to a realistic face image, thereby creating a system where a user-provided sketch results in a convincing photo.
4) **Question #4:** We implemented a CycleGAN to perform bidirectional image-to-image translation using the same Person Face Sketches dataset. This model not only converts a real face image into a sketch but also translates a sketch back into a realistic face image.

In this report, we explain our methodology, describe our model architectures, discuss our results and challenges, and finally, summarize our findings.

II. Methodology

A. Dataset and Preprocessing

For Questions #1 and #2, our training data came from two sources:

- For the signature generation task, we used the signature dataset provided in Assignment #1. Each signature image is preprocessed to the shape `[1, 64, 64]` (grayscale) and augmented using transformations such as scaling, rotation, and noise addition.
- For the custom GAN on CIFAR-10, we use the CIFAR-10 dataset but only select the images of cats and dogs. These images are normalized to the range $[-1, 1]$.

For Questions #3 and #4, we use the Person Face Sketches dataset. The dataset is structured in three folders (`train`, `val`, and `test`), each with two subdirectories: `photos` (real face images) and `sketches` (corresponding sketches). All images are resized (e.g., 64×64) and normalized using standard transformations.

B. Model Architectures

1) Question #1: VAE and Simple GAN for Signatures:

VAE:

The Variational Autoencoder is built with a series of convolutional layers in the encoder that reduce the 64×64 signature image down to a 4×4 feature map. This is then flattened and passed through fully connected layers to yield the latent mean and log variance. In the decoder, the latent vector is projected back using a fully connected layer, reshaped, and fed through transposed convolutional layers to reconstruct the signature image. A combination of binary cross-entropy (BCE) and KL divergence losses guides the training.

Simple GAN:

Our simple GAN consists of:

- A **Generator** that starts from a 100-dimensional noise vector and uses fully connected and transposed convolutional layers to generate a fake signature image.

- A **Discriminator** that takes an image of shape `[1, 64, 64]` and uses convolutional layers to classify the image as real or fake (using a sigmoid output).

2) Question #2: Custom GAN for CIFAR-10: In this custom GAN, the generator is a DCGAN-style network that maps a 100-dimensional noise vector to a 32×32 RGB image. The discriminator is built with a Siamese architecture — it processes two inputs (a generated image and a real image) through a shared convolutional encoder and calculates the absolute difference between their feature representations before outputting a similarity score via fully connected layers. This similarity score is trained using binary cross-entropy loss.

3) Question #3: Conditional GAN (cGAN) for Person Face Sketches: For the conditional GAN, the generator leverages a two-branch architecture:

- One branch encodes the input sketch into a 100-dimensional feature vector.
- This feature vector is concatenated with a latent noise vector.
- The combined vector then passes through a series of layers to produce a realistic face image.

The discriminator for the cGAN is designed to accept a pair (sketch and image) concatenated along the channel dimension and produces a similarity score to determine if the generated image aligns with the given sketch.

4) Question #4: CycleGAN for Person Face Sketches: The CycleGAN consists of:

- Two generators (ResNet-based) that learn to translate between the two domains (photos and sketches). Each generator has an initial convolution block, two downsampling layers, several residual blocks (default 6, though this can be reduced), and two upsampling layers, with a final Tanh activation to ensure the output is in the range $[-1, 1]$.
- Two PatchGAN discriminators that take in images and provide a patch-wise decision on whether the image is real or fake.

In training, besides the adversarial loss (using MSE), we also employ cycle consistency loss (L1 loss) and identity loss (L1 loss) to ensure that the translated images can be converted back to their original domain with minimal distortion. The model weights are saved after each epoch, allowing training to resume in case of interruptions.

C. Training Strategy and Hyperparameters

For each model, we set up our training as follows:

- **Optimization:** Adam optimizers are used with learning rates typically around 0.0002 and betas set to (0.5, 0.999). In some cases, if training is unstable, a slightly lower learning rate (e.g., 0.0001) is recommended.
- **Checkpointing:** Model weights are saved after every epoch to prevent loss of progress.
- **Data Augmentation:** The signature dataset is augmented using scaling, rotation, and noise addition to expand the dataset and help models learn richer representations.
- **CycleGAN Specifics:** The training of CycleGAN is done in an end-to-end manner with alternated updates between the generators and discriminators.

III. Results

A. Training Performance

- **Q1 (VAE and Simple GAN):** The VAE showed a steady decrease in reconstruction loss, and the GAN exhibited typical adversarial dynamics where the discriminator's loss approached zero in early epochs while the generator loss was high, gradually decreasing with further training.
- **Q2 (Custom GAN for CIFAR-10):** The Siamese discriminator provided a similarity score between real and generated images, and over training, the generator improved in generating images closer to real ones, as observed from the evolving loss values.
- **Q3 (Conditional GAN for Face Sketches):** The model learned to condition on the sketch input, gradually generating more realistic face images. Visual results showed that with more epochs the generated images better matched the input sketches.
- **Q4 (CycleGAN):** The CycleGAN trained on the Person Face Sketches dataset took alot time for epochs but demonstrated successful bidirectional image translation for larger number of epochs. Real faces were consistently translated into sketches and vice versa as training advanced over many epochs.

B. Output Examples

Although space is limited here, sample outputs include:

- Reconstructed signature images from the VAE that closely resemble the original.
- Fake signatures from the simple GAN that gradually improve over training.

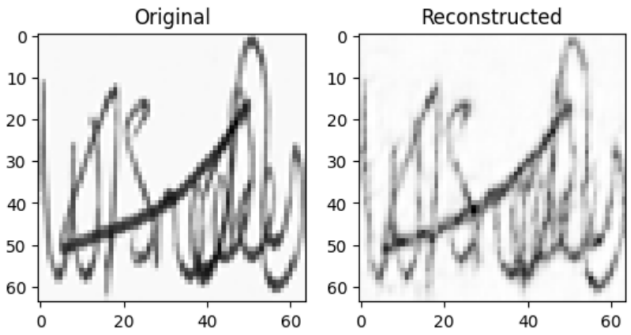

Fig. 1

- CIFAR-10 images (cats and dogs) generated by the custom GAN with high similarity scores.
- Generated face images conditioned on sketches (from the cGAN) and successful translations in the CycleGAN (both sketch to face and face to sketch).

IV. Discussion

Our experiments confirmed that:

- Data augmentation dramatically improves performance in tasks with limited data, as seen in the signature generation tasks.
- The Siamese-style discriminator in our custom GAN for CIFAR-10 helps the generator learn by emphasizing similarity, though early training shows low discriminator loss (a common occurrence in short training regimes).
- The conditional GAN successfully utilizes sketch information to generate realistic face images, yet balancing losses (adversarial, identity, and cycle) requires careful tuning.
- The CycleGAN model, despite its complexity, effectively handles bidirectional translation between faces and sketches. Training can be time-intensive, and adjustments (e.g., reducing the number of residual blocks or adjusting the learning rate) help speed up convergence.

We also noted challenges with training time and GPU memory usage, which were mitigated by checkpointing and cautious hyperparameter tuning.

V. Conclusion

In this report, we explored multiple generative modeling techniques:

- A VAE and a simple GAN for generating signature images, demonstrating that even limited and augmented data can yield meaningful latent spaces.
- A custom GAN for CIFAR-10 using a Siamese discriminator that directly compares generated images with real images.
- A Conditional GAN (cGAN) that learns to translate sketches into realistic face images.
- A CycleGAN that achieves bidirectional image-to-image translation between real face images and sketches.

These experiments highlight that, with proper data preparation, careful architectural design, and optimization techniques, deep generative models can effectively generate and translate images in complex settings.

VI. Prompts

The assignment instructions were as follows:

1) **Question #1:** Implement a VAE and a simple GAN to generate fake signatures. Augment the dataset to ensure sufficient diversity, and test the models on a reserved test set.
2) **Question #2:** Implement a custom GAN on the CIFAR-10 dataset (restricted to cats and dogs) with a discriminator that outputs a similarity score between a generated and real image.
3) **Question #3:** Implement a Conditional GAN (cGAN) using the Person Face Sketches dataset, conditioning on sketches to generate realistic face images.
4) **Question #4:** Implement a CycleGAN using the Person Face Sketches dataset to perform bidirectional image-to-image translation (face-to-sketch and sketch-to-face). Save model weights after every epoch to resume training if needed.

VII. References

References

[1] Kingma, D.P. and Welling, M., 2013. Auto-Encoding Variational Bayes. *ICLR*.

[2] Goodfellow, I., et al., 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

[3] Mirza, M. and Osindero, S., 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*.

[4] Zhu, J.Y., et al., 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *CVPR*.

[5] Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*.