Image Generation and Translation Models in Deep Learning: GAN, VAE, CycleGAN, and Transformer for English-to-Urdu Translation

Usman Zafar i210608

No Institute Given

Abstract. This report presents the implementation and evaluation of multiple deep learning models, including Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), CycleGAN, and Transformer for machine translation tasks. The CIFAR-10 dataset is used for the image generation tasks, while a custom English-to-Urdu translation task is tackled using the Transformer model. Various techniques like feature matching and custom discriminators are explored to improve the quality and diversity of the generated images. Results are presented for both qualitative and quantitative evaluation.

1 Introduction

Deep learning models have revolutionized various fields in recent years, particularly in computer vision and natural language processing. In this report, we explore the implementation of four distinct models for different tasks:

- Image Generation using Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) on the CIFAR-10 dataset.
- CycleGAN implementation for person face sketch-to-image conversion.
- Machine Translation using Transformers for English-to-Urdu translation.

This work aims to demonstrate the power of deep learning models in generating realistic images, translating between languages, and transforming visual data.

2 Image Generation using GAN and VAE on CIFAR-10

2.1 Objective

The primary objective of this task was to use the CIFAR-10 dataset, focusing on the "cat" and "dog" classes, to train and compare two generative models: a custom GAN network and a Variational Autoencoder (VAE). The goal was to evaluate the visual quality, convergence behavior during training, and the model performance based on quantitative metrics such as reconstruction loss and similarity scores.

2.2 Dataset

The CIFAR-10 dataset consists of 60,000 32x32 color images across 10 classes, with 6,000 images per class. We used only the ”cat” and ”dog” classes for the task.

2.3 Implementation Details

Generative Adversarial Network (GAN) For the GAN, the generator and discriminator were designed to follow the standard architecture with modifications:

- The discriminator evaluates the similarity between real and generated images, aiming to output a dissimilarity score.
- A Siamese network-like architecture was used for the custom discriminator.
- Feature Matching and Mini-Batch Discrimination were implemented to reduce mode collapse and improve diversity in generated images.

Variational Autoencoder (VAE) The VAE was trained using the same dataset, and its performance was compared to that of the GAN. The VAE focused on minimizing the reconstruction loss, which was used as a performance metric.

2.4 Results and Evaluation

The models were evaluated based on:

- Visual Quality of Generated Images: GAN produced sharper images with better-defined features compared to VAE, which generated smoother but less detailed images.
- Convergence Behavior: The GAN showed slower convergence due to its adversarial nature, while the VAE converged more steadily.
- Quantitative Metrics: The GAN demonstrated higher similarity scores, whereas the VAE had lower reconstruction loss.

2.5 Comparison of GAN and VAE

3 CycleGAN Implementation for Person Face Sketches

3.1 Objective

The goal was to implement a CycleGAN model to perform image-to-image translation tasks. Specifically, the model converts face sketches to real face images and vice versa.

Metric	**GAN**	**VAE**
Generator Loss	170-180	N/A
Discriminator Loss	25-30	N/A
Reconstruction Loss	N/A	0.15
Visual Quality	High sharpness, clear features	Smooth but blurry images
Training Convergence	Slower, unstable at times	Faster, stable
Diversity of Output	High, due to feature matching	Limited, no diversity promoting

Table 1: Comparison between GAN and VAE models on CIFAR-10.

3.2 Dataset

The Person Face Sketches dataset from Kaggle was used. The dataset includes sketches and real face images, which were split into training and validation sets.

3.3 Implementation Details

The CycleGAN model was trained using an encoder-decoder architecture with two generators and two discriminators. Each generator learns to map an image from one domain (e.g., face sketches) to another (e.g., real face images). The discriminators evaluate the authenticity of the generated images, while the generators aim to fool the discriminators.

3.4 Results and Evaluation

The model was evaluated based on its ability to generate high-quality and realistic sketches and images. A Flask-based user interface was developed to demonstrate real-time image conversion.

3.5 CycleGAN Performance

Metric	**CycleGAN Performance**
Visual Quality of Generated Sketch	High similarity to real sketches
Visual Quality of Generated Face	Clear facial features, high realism
Training Convergence	Stable convergence with minimal overfitting
Model Size	10M parameters

Table 2: CycleGAN Performance on Person Face Sketches Dataset.

4 Quesstion-3 Machine Translation using Transformers for English-to-Urdu

Machine translation (MT) is a subfield of NLP that focuses on the automatic translation of text from one language to another. This project aims to build a translation model from English to Urdu using the Transformer architecture introduced by Vaswani et al. (2017). Urdu, being morphologically rich, presents unique challenges that are addressed through careful preprocessing and modeling.

4.1 Objectives

- Implement a Transformer-based model for English-to-Urdu translation.
- Train the model on a suitable parallel corpus.
- Evaluate the model using BLEU and ROUGE metrics.

4.2 Dataset

Two main datasets were considered:

- **Parallel Corpus for English-Urdu Language**: `https://www.kaggle.com/datasets/zainuddin123/parallel-corpus-for-english-urdu-language`, containing over 24,000 aligned sentence pairs.
- **UMC005 English-Urdu Parallel Corpus**: `https://ufal.mff.cuni.cz/umc/005-en-ur/`, containing texts from the Quran, Bible, Penn Treebank, and Emille corpus.

4.3 Methodology

4.4 Data Preprocessing

Data was cleaned and tokenized using NLTK and custom scripts. Subword tokenization using Byte Pair Encoding (BPE) was applied to handle out-of-vocabulary words effectively.

4.5 Model Architecture

The model follows the Transformer architecture. TensorFlow was used with mixed precision training enabled. Pretrained multilingual models such as mBART were also fine-tuned.

4.6 Training

The model was trained using a custom learning rate scheduler and gradient clipping. Batch size and epochs were tuned to achieve optimal results. Training was performed on GPUs to leverage hardware acceleration.

4.7 Evaluation

Evaluation was carried out using BLEU and ROUGE metrics. The BLEU score provided a quantitative measure of translation quality. Example translations were also inspected for qualitative assessment.

4.8 Results

The model achieved competitive BLEU scores on the validation set. Example translations showed syntactic and semantic alignment with the ground truth.

4.9 Conclusion and Future Work

This implementation demonstrates the feasibility of using Transformers for English-to-Urdu translation. Future improvements may include:

- Extensive hyperparameter tuning.
- Further fine-tuning of pretrained models.
- Synthetic data generation using LLMs to augment training sets.

5 Question-4 CNN vs Vision Transformer on CIFAR DATASET

The CIFAR-10 dataset, comprising 60,000 32x32 color images distributed across ten distinct classes, is a critical benchmark for assessing image classification models. With the evolution of deep learning technologies, Vision Transformers (ViT) have emerged as a groundbreaking approach, adapting the transformer architecture—originally devised for natural language processing—to the realm of image recognition. This study endeavors to implement a Vision Transformer for CIFAR-10 image classification, evaluate its effectiveness, and compare its performance against traditional and hybrid architectures. Our investigation will encompass data preprocessing, model development, hyperparameter optimization, and performance evaluation through metrics such as accuracy, precision, recall, and F1-score.

5.1 Models

5.2 Vision Transformer (ViT)

The Vision Transformer (ViT) implemented for this study adapts the transformer architecture, traditionally used for natural language processing, to the image classification task on the CIFAR-10 dataset. The images are divided into fixed-size patches, which are then linearly embedded. Positional encodings are added to these patch embeddings to retain positional information. The transformer uses self-attention mechanisms to process these embeddings, facilitating the model to focus on relevant parts of the image. This architecture was implemented using PyTorch, with specific attention to layer configurations and attention heads to optimize performance.

5.3 Hybrid Architecture

The hybrid architecture combines convolutional neural networks (CNN) and multi-layer perceptrons (MLP) to leverage the strengths of both. In this model, a CNN is first used to extract spatial hierarchies from image patches, capturing essential local features without the complexity of global attention mechanisms. These features are then processed through an MLP, allowing the model to learn higher-level abstractions. This approach aims to blend the efficiency of CNNs in handling images with the sequential data processing capabilities of MLPs, creating a robust framework for image classification.

5.4 Comparison of Models

This section provides a comparative analysis of the three models implemented: Vision Transformer (ViT), ResNet, and the Hybrid architecture. Each model's performance is evaluated based on the confusion matrices generated during testing on the CIFAR-10 dataset.

5.5 Confusion Matrices

The confusion matrices for each model are instrumental in visualizing their performance, highlighting correct and incorrect classifications across the different classes.

5.6 Performance Discussion

CNN The hybrid architecture combines convolutional neural networks (CNN) and multi-layer perceptrons (MLP) to leverage the strengths of both. In this model, the CNN first extracts spatial hierarchies from image patches, capturing essential local features, and then the MLP processes these features, allowing the model to learn higher-level abstractions. The Hybrid Model appears to perform the best among the three, as indicated by its confusion matrix, which shows higher numbers of correct classifications across most classes.

Vision Transformer (ViT) The Vision Transformer shows a promising approach but has underperformed relative to the other models. This underperformance is likely due to fewer epochs of training and limited computational resources, which restricted its ability to fully learn from the dataset. The ViT's performance is still competitive, but it is evident that further optimization and training are required to realize its full potential.

5.7 Model Analysis

From the results, it is clear that the Hybrid Model's combination of CNN for feature extraction followed by an MLP for classification effectively captures both local and global image features. The pretrained ResNet Model benefits from transfer learning but seems to need more fine-tuning on the CIFAR-10 dataset. Meanwhile, the Vision Transformer, which divides images into patches and processes them through self-attention, showed potential but requires more extensive training and resources to outperform the other models.

5.8 Conclusion

In this study, we implemented and compared three distinct models for image classification on the CIFAR-10 dataset: the Vision Transformer (ViT), ResNet, and a Hybrid architecture combining CNN and MLP. The Hybrid Model demonstrated superior performance, effectively utilizing both local and global image features, followed by the ResNet model, which benefited significantly from transfer learning. The ViT, although innovative, lagged behind due to insufficient training resources and fewer epochs, indicating a need for more extensive training and optimization. These findings underscore the potential of hybrid models in achieving high accuracy in image classification tasks and highlight the necessity for adequate resources when training computationally intensive models like ViT. Future studies should explore the scalability of these models and their applicability to other complex image datasets.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I. (2017). Attention is All You Need. *Neural Information Processing Systems (NeurIPS)*.
2. Radford, A., Metz, L., Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434*.