Deepfake Video Detection using Vision Transformers

Shubh Khandelwal

CS22B1090

1 Introduction

This project implements a Vision Transformer-based model to detect fake videos in the Celeb-DF dataset. This report details the development and evaluation of this project for classifying deepfake videos using a Vision Transformer (ViT) architecture. The Celeb-DF v2 dataset is used for training and testing the model. Key evaluation metrics such as Accuracy, AUC, Precision, and Equal Error Rate (EER) are reported.

2 Requirements

```
torch
torchvision
scikit-learn
numpy
opencv-python
pandas
```

3 Dataset

- **Dataset:** Celeb-DF v2
- **Classes:** Real (0) and Fake (1)
- **Total Videos:** Approximately 5639
- **Format:** MP4

4 Preprocessing

- Created custom dataset and dataloader for loading videos from Celeb-DF dataset
- Converted video frames from BGR to RGB
- Applied resizing and normalization using ImageNet statistics
- Selected a fixed number of maximum frames per video (e.g., 32)
- Used padding (repeating the last frame) if a video had fewer frames

5 Vision Transformer (ViT) Architecture

- Patch embedding of video frames with embed dimension 768
- Implementation of multi head self attention
- MLP classification head
- Transformer encoder blocks with variable depth to increase or decrease complexity
- Use of patch embedding and transformer encoder blocks along with normalization layer and head

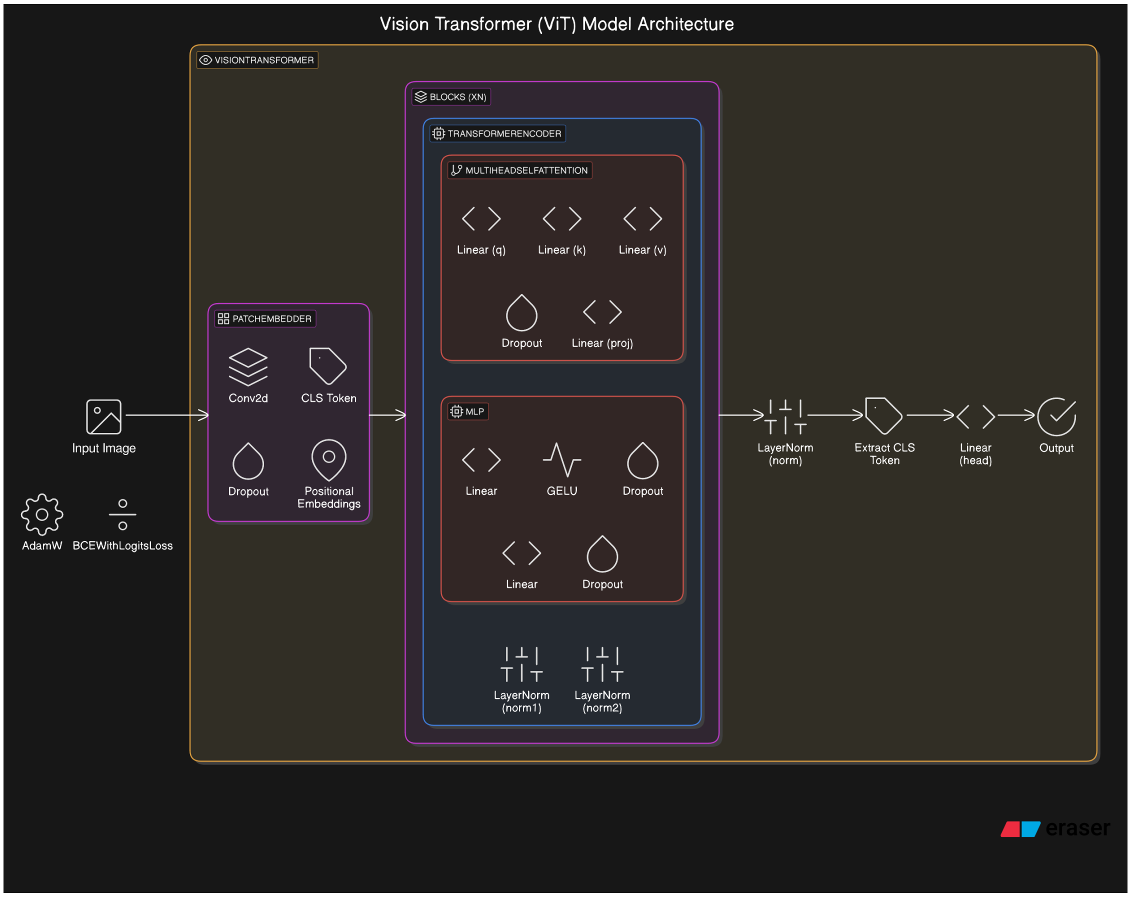

Figure 1: Block diagram of the Vision Transformer architecture

6 Training Details

- Optimizer: AdamW
- Learning Rate: 3e-4
- Weight Decay: 0.05
- Loss Function: BCEWithLogitsLoss
- Epochs: 10
- Batch Size: 2

7 Evaluation Metrics

Per-video evaluation using:

- Accuracy: 88.52%
- Area Under Curve (AUC): 49.87
- Precision: 88.52%
- Equal Error Rate (EER): 0.00

8 References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

2. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). *Deepfakes and beyond: A survey of face manipulation and fake detection.* Information Fusion, 64, 131-148.

3. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3207-3216).

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need.* Advances in neural information processing systems, 30.