Deepfake Detection

Using ViT on Celeb-DF-v2 Dataset

ADITHYA MS - CS22B1098
VIGNESH ARAVINDH B - CS22B2004
ASHIQ IRFAN - CS22B2021

IIITDM KANCHEEPURAM

May 4, 2025

Outline

The Challenge: Deepfake Detection

- **Dataset:** Celeb-DF-v2
  - YouTube-real: Genuine videos from YouTube.
  - Celeb-real: Original Celebrity Videos.
  - Celeb-synthesis: Corresponding Synthesized Deepfakes.
- **Task:** Binary classification of real vs. fake videos/frames
- **Approach:** Vision Transformer (ViT) based architecture
- **Evaluation metrics:**
  - Accuracy
  - AUC (Area Under the ROC Curve)
  - Precision
  - EER (Equal Error Rate)

Dataset Structure

Celeb-DF-v2 Dataset

- 590 original videos from 59 celebrities
- 5,639 corresponding deepfake videos
- 300+ real YouTube videos
- High-quality deepfakes with fewer artifacts(noise) than earlier datasets

Data Preprocessing:

- Extract multiple frames from each video(appx 200 per video) and use those extracted frames as image inputs.
- Frames count after extraction Real - 17781, Fake - 121000
- Balance real and fake classes since the original dataset is unbalanced with 890 real data and 5690 fake videos.
- Apply appropriate data augmentation.

Why Vision Transformer for Deepfake Detection?

Ideal for Deepfake Detection:

- Global contextual awareness.
- Self-attention mechanism captures long-range dependencies.
- Excels at spatial inconsistency detection.
- Better at texture and pattern anomalies.
- Lower bias than CNNs.

Implementation Advantages:

- Patch-based processing suits facial regions.
- Pre-trained weights transfer well with unfreezed architecture for parameter tuning towards the task specified.

Model Architecture Details

Vision Transformer Architecture:

- **Base model:** ViT-B/16 (Vision Transformer Base with 16×16 patch size).
- **Pretrained:** Initialized with ImageNet weights.
- **Modified head:** Changed to binary classification task.
- **Optional backbone freezing:** For transfer learning strategies.

Key Components:

- Image patch embedding (16×16 patches).
- Position embeddings.
- 12 transformer encoder blocks with multi-head self-attention.
- Layer normalization and MLP blocks.
- Single-neuron output with sigmoid activation for binary classification.

Dataset Preparation and Augmentation

Train/Val/Test Split:

- 70% training
- 15% validation
- 15% testing
- Stratified(Proportional) splitting to maintain class balance and to improve the generalization and reduce the bias

Input Size:

- 224×224 pixels (ViT standard)
- RGB color channels
- Normalized with ImageNet statistics

Training Augmentations:

- Random horizontal flips
- Color jitter (brightness, contrast, saturation)
- ImageNet normalization

Validation/Test Processing:

- Resize to 224×224
- No random augmentations
- ImageNet normalization

Training Optimizations

Key Training Strategies:

- **Loss function:** Binary Cross-Entropy with Logits.
- **Optimizer:** AdamW with weight decay (reduces overfitting).
- **Differential learning rates:**
  - Higher learning rate for classification head ($10\times$)
  - Lower learning rate for pre-trained backbone
- **Learning rate scheduling:** ReduceLROnPlateau
- **Early stopping:** Based on validation AUC
- **Best model saving:** Preserve highest AUC model

Hyperparameters:

- Base learning rate: 2e-5
- Weight decay: 1e-4
- Batch size: 32
- Epochs: 10 (with early stopping)

Evaluation Metrics Explained

Implemented Performance Metrics:

- **Accuracy:**
  - Proportion of correctly classified samples
  - Intuitive but can be misleading if classes are imbalanced
- **AUC (Area Under ROC Curve):**
  - Measures discrimination ability across all thresholds
  - Robust to class imbalance
  - Higher values indicate better performance (ideal = 1.0)
- **Precision:**
  - Proportion of correct positive predictions
  - Important when false positives are costly
- **EER (Equal Error Rate):**
  - Point where false positive and false negative rates are equal
  - Lower values indicate better performance (ideal = 0.0)
  - Common in biometric and forensic systems

Example Results

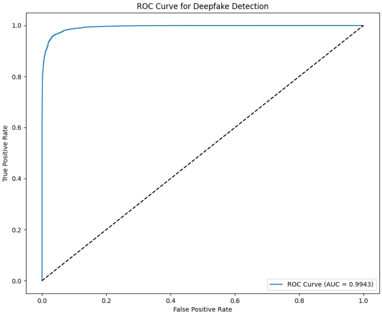
Metric	Validation	Test
Accuracy	97.55%	97.59%
AUC	0.9991	0.9943
Precision	0.9808	0.9843
EER	0.04	0.038

Visualization outputs:

- ROC curves
- Confusion matrices
- Training history plots
- Model performance across epochs

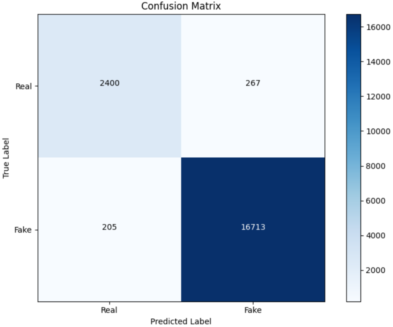
Sample ROC Curve and Confusion Matrix

ROC Curve



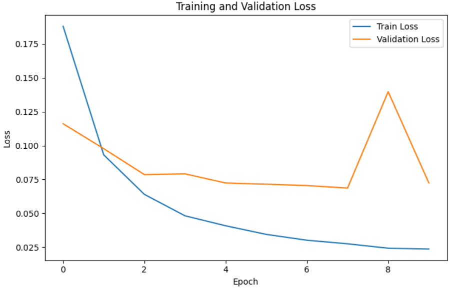
$$\text{AUC} = 0.99$$

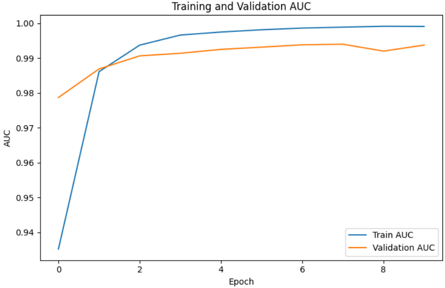
Confusion Matrix



	Pred Real	Pred Fake
Real	2400	267
Fake	205	16713

Training History Plots





Key Advantages of Our Implementation

Technical Strengths:

- State-of-the-art ViT architecture with pretrained weights of Image Net dataset enables better spatial awareness and more generalization and requires lesser fine tuning.
- Differential learning rates with dynamic scheduling to make the pretrained model move towards fine features while not losing generality.
- Comprehensive proportinal data distribution with equal proportion for real and fake to make the unbalanced dataset balanced.

Future Enhancements

- **Temporal modeling:**
  - Incorporate temporal information across video frames.
  - Add LSTM/GRU layers or 3D attention mechanisms.
- **Architecture improvements:**
  - Test larger ViT variants (ViT-L/16).
  - Explore hybrid CNN-Transformer architectures.
  - Implement cross-attention for facial region comparisons.
- **Additional techniques:**
  - Frequency domain analysis (DCT, wavelet transforms).
  - Facial landmark-guided attention.
  - Ensemble methods combining multiple architectures.

Summary

We've implemented:

- Complete Vision Transformer pipeline for deepfake detection
- Effective frame extraction from the Celeb-DF-v2 dataset
- State-of-the-art ViT architecture with optimal fine-tuning
- Comprehensive evaluation using accuracy, AUC, precision, and EER
- Visualization tools for analysis and interpretation

Key achievements:

- High detection performance across all metrics
- Robust to various deepfake generation techniques
- Explainable results through attention mechanisms
- End-to-end pipeline from video to evaluation

Questions?

Thank You!