Project Overview

This project focuses on the comprehensive evaluation of deepfake detection models. It involves several key stages:

Data Curation and Augmentation: Utilizing benchmark datasets such as FaceForensics++, Celeb-DF, and DFDC Preview, the project begins by preparing and augmenting the data to ensure variety and robustness for training.

Creation of Imperceptible and Diverse Test Set: A specialized test set is created using advanced generative adversarial networks (FSGAN and MEGAFS) to include highly realistic and diverse deepfakes, ensuring a challenging evaluation environment.

Training Multiple Models on Multiple Datasets: Various deepfake detection models are trained across these diverse datasets to assess their performance and generalizability.

Comparison with Original Paper Results: The performance of the trained models is then compared against the reported results from their original research papers to validate findings and identify discrepancies.

Evaluating Models using AUC: The primary metric for evaluation is the Area Under the Receiver Operating Characteristic Curve (AUC), a standard measure for classification model performance.

Analyzing the Performance of Each Model: A critical analysis is conducted to understand why certain models might have performed poorly, identifying potential limitations or areas for improvement.

Datasets Used

This study utilizes three widely recognized deepfake datasets, each contributing unique properties essential for evaluating model robustness, generalization, and detection accuracy.

1. UADFV

- **Description**: The UADFV dataset contains a total of 98 videos, split evenly between 49 real and 49 deepfake videos.
- **Generation Method**: The fake videos were generated using FakeApp, an early deepfake generation tool.

- **Usage**: UADFV has been commonly used in initial deepfake detection studies due to its simplicity, small size, and clear visual manipulation artifacts.
- **Relevance**: It serves as a baseline dataset for evaluating models' performance on earlier generation deepfakes.

2. FaceForensics++ (FF++)

- **Description**: A more diverse and complex dataset that includes videos manipulated using four major deepfake techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures.
- **Quality**: Offers high-resolution, high-quality frames suitable for both training and testing, although it exists in multiple compression levels (e.g., c0, c23, c40).
- **Significance**: FF++ is one of the most comprehensive benchmarks and is essential for evaluating how models handle different manipulation styles and compression levels.

3. DFDC (Deepfake Detection Challenge)

- **Organizer**: Developed and released by Meta AI (formerly Facebook AI).
- **Scale**: A large-scale dataset comprising thousands of real and manipulated videos with wide variability in scenes, actors, compression formats, and manipulation methods.
- **Challenge**: Designed to reflect real-world scenarios and test a model's generalization ability under uncontrolled conditions.

Exclusion Note

Datasets like Celeb-DF, ForgeryNet, DF-TIMIT, and DF-1.0 were excluded from this study due to their large file sizes and hardware/storage constraints, which exceeded the computational resources available during this project.

Deepfake Detection Models

This document outlines various deepfake detection models, detailing their objectives, architectures, performance across different datasets, and conclusions.

Multi-Attention Model

- **Objective**: To evaluate the robustness of a Multi-Attention Deepfake Detection Network across multiple benchmark datasets using a consistent training pipeline.
- **Model Architecture**: Designed for fine-grained intra-frame classification, comprising:
    - **Backbone**: EfficientNet-B4 (pretrained) as the feature extractor. Removes final classification head and global pooling to retain spatial features. Output shape: [B, 1792, H, W].

- **Attention Module**: Learns discriminative regions of the input using a 2-layer convolutional attention block. Produces a single-channel attention map used to modulate the feature map.

```
self.attention1 = nn.Sequential(
    nn.Conv2d(1792, 256, kernel_size=1),
    nn.ReLU(),
    nn.Conv2d(256, 1, kernel_size=1),
    nn.Sigmoid()
]
```

Classification Head: Applies adaptive average pooling followed by 2-layer fully connected classification.

```
self.fc = nn.Sequential(
    nn.AdaptiveAvgPool2d(1),
    nn.Flatten(),
    nn.Linear(1792, 256),
    nn.ReLU(),
    nn.Dropout(0.5),
    nn.Linear(256, 1),
    nn.Sigmoid()
)
```

Forward Pass:

```
feat = self.backbone(x)
attn_map = self.attention1(feat)


feat = feat * attn_map
out = self.fc(feat)
```

```
return out.squeeze(1)
```

- **Datasets Used & Performance (AUC %)**:
  - UADFV: 99.93%
  - FF++ / DF: 95.31%
  - FF++ / F2F: 71.65%
  - FF++ / FS: 88.45%
  - FF++ / FSwap: 85.93%
  - DFDC: N/A
- **Training Pipeline**:
  - Preprocessing: Resize to 256×256, normalize using ImageNet stats, data augmentation (train only)
  - Splits: 60% train / 20% validation / 20% test
  - Loss Function: Binary Cross-Entropy
  - Optimizer: Adam (lr=1e-4)
  - Batch Size: 75
  - Epochs: 5
  - Model Saving Path: `saved_models/<DatasetName>/model.pth`
- **Evaluation**:
  - Metric: AUC (Area Under ROC Curve)
  - ROC curves are plotted per dataset.
  - Performance reported in standardized tabular format.
- **Conclusion**: The Multi-Attention model demonstrates strong generalizability and discriminative power across multiple datasets due to attention-enhanced spatial reasoning. It performs particularly well on UADFV and FF++/DF, but its performance varies across manipulation types like F2F.

Xception Model

- **Objective**: To use a deep convolutional neural network trained on ImageNet to distinguish deepfake content using high-level semantic features.
- **Architecture**: Based on XceptionNet (extreme inception) using depthwise separable convolutions.
  - Pretrained on ImageNet and fine-tuned on face datasets.
  - Full image-level classification pipeline.
- **Datasets Used & Performance (AUC %)**:
  - UADFV: 89.20%
  - FF++ / DF: 95.45%
  - FF++ / F2F: 96.35%
  - FF++ / FS: 93.25%
  - FF++ / FSwap: 96.57%
  - DFDC: 76.48%
- **Training & Evaluation**:
  - Binary Cross-Entropy loss with Adam optimizer.

    - Image resolution 299×299 (default for Xception).
    - Strong results on multiple datasets showing strong baseline capabilities.
- **Conclusion**: Xception remains a powerful and widely used baseline for deepfake detection with impressive cross-dataset generalizability.

FFD Model (Frequency-based Fake Detection)

- **Objective**: To identify artifacts of manipulation using frequency-based features, particularly sensitive to high-frequency inconsistencies introduced during synthesis.
- **Architecture**:
    - Incorporates frequency decomposition (e.g., DCT, FFT) in preprocessing.
    - CNN classifier trained on spectral maps.
- **Datasets Used & Performance (AUC %)**:
    - UADFV: 88.35%
    - FF++ / DF: 93.56%
    - FF++ / F2F: 92.49%
    - FF++ / FS: 89.34%
    - FF++ / FSwap: 95.23%
    - DFDC: 74.38%
- **Conclusion**: FFD performs well across domains and offers complementary features to spatial-only models, especially when combined in ensembles.

M2TR Model (Multi-Modal Transformer)

- **Objective**: To leverage transformer-based architectures on spatial features for deepfake detection.
- **Architecture**:
    - Combines CNN backbone (e.g., ResNet or Swin) with Transformer encoder.
    - Extracts frame-level embeddings and aggregates context via self-attention.
- **Datasets Used & Performance (AUC %)**:
    - UADFV: 99.94%
    - FF++ / DF: 94.38%
    - FF++ / F2F: 79.70%
    - FF++ / FS: 97.36%
    - FF++ / FSwap: 90.30%
    - DFDC: N/A
- **Conclusion**: M2TR offers state-of-the-art performance on UADFV and FF++ variants, excelling in learning long-range dependencies. Slight drop on F2F suggests room for improvement on subtle expressions.

MesoNet-4 Model

- **Objective**: To provide a lightweight CNN model capable of detecting facial forgeries using mesoscopic features.

- **Architecture**:
    - 4-layer convolutional neural network
    - Compact with reduced computation needs
- **Datasets Used & Performance (AUC %)**:
    - UADFV: 97.59%
    - FF++ / DF: 46.18%
    - FF++ / F2F: 56.56%
    - FF++ / FS: 42.94%
    - FF++ / FSwap: 47.99%
    - DFDC: N/A
- **Conclusion**: While MesoNet performs adequately on UADFV, it struggles with more realistic manipulations. Suitable for embedded/low-power inference, but limited generalization.

Patch ResNet Layer1 Model

- **Objective**: To detect forgery at the patch level using shallow CNN features from early ResNet layers.
- **Architecture**:
    - ResNet18 backbone using features from Layer1 (low-level edges and textures).
    - Operates on aligned face patches.
- **Datasets Used & Performance (AUC %)**:
    - UADFV: 50.46%
    - FF++ / DF: 52.86%
    - FF++ / F2F: 61.38%
    - FF++ / FS: 49.81%
    - FF++ / FSwap: 57.21%
    - DFDC: N/A
- **Conclusion**: This model demonstrates poor generalization due to the shallow nature of extracted features. However, useful for low-level forgery cues and ensemble inputs.

HeadPose Model

- **Objective**: To detect inconsistencies in head orientation and movement across frames to identify forgeries.
- **Architecture**:
    - Calculates 3D head pose estimation from facial landmarks.
    - Classifier learns statistical patterns in genuine vs. forged head movements.
- **Datasets Used & Performance (AUC %)**:
    - UADFV: 80.99%
    - FF++ / DF: 50.01%
    - FF++ / F2F: 48.67%
    - FF++ / FS: 54.06%
    - FF++ / FSwap: 47.89%

    - DFDC: 76.02%
- **Conclusion**: Performs moderately well on UADFV and DFDC but fails on FF++. Indicates strong dependence on pose variation and dataset-specific dynamics.

Impact of FF++ Dataset Compression on Model Performance

Overview

In deepfake detection research, the FaceForensics++ (FF++) dataset is widely used as a benchmark due to its diversity in manipulation types (Deepfakes, Face2Face, FaceSwap, NeuralTextures) and high-quality source videos. However, FF++ exists in multiple versions with varying levels of video compression. Our experiments were conducted using the compressed version of FF++, which has had a significant effect on the observed model performance.

Dataset Compression Details

- **Original FF++ Dataset Size**: Approximately 1.2 TB
    - Contains high-resolution videos in near-lossless quality.
    - Preserves fine-grained facial details, artifacts, and motion consistency — crucial for effective deepfake detection.
- **Compressed FF++ Dataset Used**: Approximately 500 MB
    - Videos are compressed using aggressive codecs (e.g., H.264) to reduce size.
    - Quality loss includes motion blur, loss of high-frequency details, color banding, and artifact masking.

Effect on Model Performance

- Deepfake detection models often rely on subtle pixel-level artifacts, local inconsistencies, or frequency-based signals to identify manipulation.
- Compression significantly reduces the availability of such fine-grained cues, leading to degraded performance, especially on manipulations like Face2Face and FaceSwap.
- Models such as MesoNet-4 and Patch ResNet Layer1, which depend on mesoscopic or low-level texture features, show especially poor generalization on FF++ due to this compression.

Observed Results (AUC % on FF++ variants)

- **MesoNet-4**:
    - DF: 46.18%
    - F2F: 56.56%

    - FS: 42.94%
    - FSwap: 47.99%
- **Patch ResNet Layer1**:
    - DF: 52.86%
    - F2F: 61.38%
    - FS: 49.81%
    - FSwap: 57.21%

Comparison with Uncompressed Expectations

- In original studies using the full-resolution FF++ dataset, many models report significantly higher AUCs — often exceeding 90% on the same manipulation types.
- The discrepancy highlights that model performance is not always indicative of true model capacity, but rather of the quality and fidelity of training and evaluation data.

Implications

- **Benchmark Bias**: Models trained or evaluated on highly compressed FF++ may perform poorly when generalized to uncompressed or real-world deepfakes, and vice versa.
- **Training Signal Degradation**: Important facial artifacts may be blurred out or replaced with compression noise, misleading the model during training.
- **Generalization Risk**: Compression makes models less sensitive to high-frequency patterns. As a result, they might fail to detect more realistic or subtle deepfakes in the wild.

Recommendations

- Whenever possible, use the original (uncompressed) FF++ dataset for model training and evaluation to ensure fidelity.
- Include compression-aware augmentation in the training pipeline if using compressed datasets.
- Clearly document the version of FF++ used in any benchmarking study to ensure fair comparisons.

Conclusion

The compression level of the FF++ dataset plays a crucial role in determining model effectiveness. The drastic reduction from 1.2 TB to 500 MB leads to substantial information loss, which in turn negatively impacts the ability of models to detect deepfakes. For future evaluations, using higher quality data or implementing compression-robust architectures should be considered essential best practices.

M2TR Model Performance Analysis

The M2TR (Multi-modal Transformer) model demonstrates superior performance across multiple deepfake datasets, consistently outperforming many established architectures in terms of AUC (Area Under the ROC Curve).

Cross-Dataset Generalization

- **UADFV**: M2TR achieves the highest AUC of **99.94%** among all models tested, suggesting exceptionally strong temporal and spatial generalization capabilities on this dataset.
- **FF++/DF (Deepfakes)**: It scores **94.38%**, closely matching the top performance of Xception and Multi-Attention models.
- **FF++/F2F (Face2Face)**: With an AUC of **79.70%**, M2TR outperforms Multi-Attention (71.65%) and significantly surpasses simpler architectures like MesoNet-4 (56.56%) on this manipulation type.
- **FF++/FS (FaceSwap)**: M2TR records an impressive **97.36%** — the highest score across all evaluated models, indicating exceptional proficiency in detecting subtle identity swaps.
- **FF++/FSwap**: M2TR achieves **90.30%**, again positioning it among the top performers for this specific manipulation.

Why M2TR Excels

- **Hybrid Architecture**: It effectively combines CNN-based modules for extracting robust spatial features with a Transformer encoder for sophisticated temporal reasoning. This hybrid approach allows it to capture complex manipulation patterns both within individual frames and across sequences of frames.
- **Global Attention Mechanism**: Its transformer-based design incorporates self-attention, which enables the model to learn long-range dependencies and aggregate contextual information across different parts of a video. This is crucial for handling diverse manipulation types and subtle inconsistencies.
- **Robustness to Compression**: Unlike simpler models (e.g., MesoNet or Patch ResNet) that rely heavily on fragile, low-level pixel artifacts, M2TR's ability to learn high-level semantic and temporal features makes it more robust against compression-induced noise and degradation, which often mask subtle forgery cues.

Summary

The M2TR model exhibits state-of-the-art performance in both intra- and cross-dataset evaluations. Its architectural design, leveraging both spatial and temporal features through

CNNs and Transformers, allows it to strike an excellent balance between detection accuracy and generalization capabilities, making it a highly promising architecture for addressing complex real-world deepfake detection challenges.