Name: Amogh Agrawal

Part 1: Research & Selection

I have reviewed the provided repository and identified three models that are well-suited for our purpose.

1. RawNet2

- **Key Technical Innovation:**

  RawNet2 processes raw audio waveforms directly using a combination of convolutional and recurrent neural networks. This eliminates the need for handcrafted features and enables the model to learn subtle patterns indicative of deepfake audio.

- **Reported Performance Metrics:**

  Achieves state-of-the-art accuracy in deepfake detection without handcrafted features. However, its performance can degrade under certain manipulations like volume control and noise injection.

- **Why You Find This Approach Promising for Our Specific Needs:**

  Since RawNet2 does not rely on pre-extracted features, it can adapt to different types of audio forgeries effectively. Its real-time or near real-time processing capability makes it suitable for practical applications where fast detection is required.

- **Potential Limitations or Challenges:**

  Vulnerable to specific audio manipulations such as volume control. Additionally, training and inference require significant computational resources, making deployment on resource-constrained devices challenging.

2. AASIST

- **Key Technical Innovation:**

  AASIST employs a graph neural network (GNN) to analyze spectro-temporal features of audio. This architecture enhances its ability to capture complex relationships between different parts of the signal, improving deepfake detection.

- **Reported Performance Metrics:**

  Demonstrates strong performance on datasets like SONAR and often outperforms other models in specific scenarios. More robust than RawNet2 against certain manipulations.

- **Why You Find This Approach Promising for Our Specific Needs:**

  AASIST's use of GNN allows it to identify subtle structural changes in audio, making it more resistant to specific deepfake techniques. Its real-time processing capability aligns well with applications requiring quick decisions.

- **Potential Limitations or Challenges:**

  Although more robust than RawNet2, AASIST still struggles with certain types of audio manipulations, such as fading. Computational demands remain high, similar to other deep-learning-based models.

3. CNN-Based Model with EfficientNet for ASVspoof Dataset

- **Key Technical Innovation:**

  This approach leverages a pre-trained EfficientNet model as a feature extractor for analyzing spectrograms—visual representations of sound frequencies over time—on the ASVspoof dataset. Transfer learning allows us to utilize the rich feature representations learned from large-scale datasets, adapting them for deepfake detection.

- **Reported Performance Metrics:**

  Pre-trained models, such as EfficientNet, demonstrate high accuracy in deepfake detection when fine-tuned on spectrogram-based representations of audio data. This approach enhances model generalization and robustness, even with limited data.

- **Why This Approach is Promising for Our Specific Needs:**
    1. Pre-trained Feature Extraction: EfficientNet has learned powerful image representations from large-scale datasets, making it well-suited for analyzing spectrograms without requiring extensive retraining.
    2. Reduced Training Time: Since most of the model's parameters are already optimized, fine-tuning only requires training a few additional layers, significantly reducing computational costs.
    3. Scalability and Deployment Efficiency: Unlike heavier architectures like RawNet2 or AASIST, EfficientNet can be deployed efficiently on mid-range GPUs or even CPUs, making real-world deployment more feasible.
    4. Improved Generalization: Transfer learning allows the model to generalize better across different types of spoofed audio samples, even when data availability is limited.

- **Potential Limitations or Challenges:**
    1. Dependence on Pre-trained Weights: The model's effectiveness depends on how well the pre-trained weights generalize to spectrogram-based inputs, which are different from standard image inputs.
    2. Feature Alignment: Spectrogram-based features may require additional preprocessing steps to ensure compatibility with EfficientNet's convolutional layers.
    3. Data Augmentation Needs: The model may still require augmentation techniques to improve robustness against unseen spoofing techniques.

Part 2: Implementation

Selected Model: “CNN-Based Model with EfficientNet for ASVspoof Dataset”

GitHub Repository Link:

https://github.com/AmoghAgrawal1249/Audio-Deepfake-Detection-using-Pretrained-Model

Advantages of Using a Pre-Trained Model (EfficientNet) with ASVspoof Dataset:

1. Highly Optimized Representations – EfficientNet has been trained on millions of images, making it an excellent feature extractor when fine-tuned for deepfake detection on spectrogram data.
2. Computational Efficiency – Unlike deeper models like RawNet2, EfficientNet achieves a strong balance between accuracy and efficiency, allowing for fast inference on consumer-grade hardware.
3. Simplified Training Process – Transfer learning requires fewer epochs and less data compared to training a CNN from scratch, reducing the risk of overfitting.
4. Pre-Trained Models Reduce Data Dependency – Since EfficientNet already captures high-level feature representations, it requires fewer labeled samples for effective fine-tuning.
5. Scalability for Deployment – The model can run on embedded devices or cloud-based systems efficiently due to its optimized architecture.
6. Robustness to Overfitting – Pre-trained models generalize better than models trained from scratch, especially on limited datasets.

Future Improvements:

- Experiment with Different Pre-trained Architectures: Testing models like ResNet or Vision Transformers for further improvements in accuracy.
- Integrate Attention Mechanisms: Enhancing the model’s ability to focus on critical frequency patterns in the spectrograms.
- Domain-Specific Pretraining: Fine-tuning EfficientNet on a broader range of deepfake audio datasets to improve domain adaptation.
- Hybrid Approaches: Combining EfficientNet with LSTM or Transformer-based models to capture both spatial and temporal dependencies in the audio data.

Part 3: Audio Deepfake Detection Implementation Report

1. Implementation Process

Challenges Encountered

- **Dataset Handling Issues:**
  - The dataset was large, making it difficult to load and process efficiently.
  - File sorting inconsistencies caused repeated selection of specific patterns in file indexing.
- **Model Input Shape Mismatch:**
  - The pre-trained EfficientNetB0 model expected a 3-channel input, but the spectrograms were grayscale (single channel).
  - This was resolved by duplicating the grayscale channel to create a 3-channel input.
- **Training Instability:**
  - Initial training results showed very low accuracy and precision.
  - Experimentation with different learning rates and dropout layers helped stabilize training.
- **Evaluation Metrics Misalignment:**
  - Initially, results showed high recall but extremely low accuracy and precision.
  - Investigated label parsing, confirming that bonafide/spoof labels were correctly extracted.

Solutions Implemented

- **Efficient File Selection:** Changed random sampling to fixed indexing for dataset consistency.
- **Data Preprocessing Fixes:** Corrected label mapping and ensured input shape compatibility with the model.
- **Model Architecture Tweaks:** Added dropout layers to reduce overfitting.
- **Performance Tracking:** Monitored metrics like accuracy, precision, recall, and F1-score.

Assumptions Made

- The dataset is representative of real-world deepfake detection scenarios.
- The bonafide/spoof classification is binary with well-separated features.
- EfficientNetB0 is a suitable backbone for feature extraction.

2. Model Analysis

Why This Model Was Selected

- **Transfer Learning Advantage:** EfficientNetB0 provides pre-trained feature extraction, reducing training time.
- **Compact Yet Powerful:** EfficientNetB0 is lightweight compared to deeper models like ResNet.
- **Proven Performance:** It has shown strong results in image-based classification tasks, making it suitable for spectrogram-based analysis.

High-Level Technical Explanation

1. **Feature Extraction:**
   - Converts audio into spectrograms, making it suitable for image-based deep learning models.
   - EfficientNetB0 processes these images to extract deep feature representations.
2. **Classification Layer:**
   - Global Average Pooling reduces dimensionality.
   - Fully connected layers learn decision boundaries for bonafide vs. spoof classification.
   - The final layer uses a sigmoid activation to output a probability score.

Performance Results(for the first 300 rows of the dataset)

- **Accuracy:** 71.00%
- **Precision:** 34.69%
- **Recall:** 23.61%
- **F1 Score:** 28.10%

Observed Strengths and Weaknesses

Strengths:

- The model is lightweight and computationally efficient.
- Moderate accuracy achieved by fine-tuning the prediction threshold during evaluation.

Weaknesses:

- Poor precision indicates many false positives.
- May struggle with real-world deepfake variations due to dataset bias.

Future Improvements

- **Enhance Data Augmentation:** Introduce pitch shifting, time stretching, and noise injection.
- **Improve Model Complexity:** Fine-tune deeper EfficientNet variants (B2, B3) for richer feature extraction.
- **Refine Decision Boundaries:** Adjust class weights to address class imbalances.

- **Use Multi-Modal Learning:** Combine spectrograms with raw waveform analysis for better accuracy.

3. Reflection Questions

1. Significant Challenges in Implementation

- Handling large datasets efficiently while ensuring reproducibility in file selection.
- Tuning model hyperparameters to improve precision without sacrificing recall.
- Ensuring correct label assignments when loading the dataset.

2. Real-World Performance vs. Research Datasets

- The model may perform worse in real-world conditions due to unseen attack types.
- Controlled datasets may lack variability seen in practical deepfake scenarios.
- Environmental noise and different recording conditions could impact accuracy.

3. Additional Data or Resources for Improvement

- A larger and more diverse dataset covering various deepfake attack techniques.
- Higher-quality audio recordings to preserve subtle speech patterns.
- Computational resources for deeper model fine-tuning.

4. Deployment Considerations for Production

- **Latency Optimization:** Convert the model to TensorFlow Lite for real-time detection.
- **Continuous Learning:** Periodically retrain on new deepfake samples to adapt to evolving threats.
- **Integration with Real-World Systems:** Deploy in media forensics, telecom security, and content authentication workflows.
- **Scalability:** Use cloud-based deployment for handling large volumes of audio data.