Documentation & Analysis

Implementation Process

Challenges Encountered:

1. **Low Accuracy and Undefined Precision:**
    - The model exhibited poor performance, with accuracy around 50%.
2. **IndexError During Training:**
    - An IndexError: list index out of range occurred during the training loop, disrupting the process.

Solutions Implemented:

1. **Addressing Low Accuracy:**
    - Introduced data augmentation techniques to increase dataset variability and enhance model generalization.
    - Adjusted the learning rate and experimented with different optimizers to improve convergence.
2. **Resolving IndexError:**
    - Ensured that dataset splitting was performed correctly, verifying that the sum of training and validation sizes matched the total dataset length.
    - Checked for any off-by-one errors in indexing and ensured that the DataLoader was properly configured.

Assumptions Made:

- Assumed that the dataset labels were accurate and representative of real and AI-generated speech.
- Presumed that the Mel-Spectrogram features effectively captured the necessary audio characteristics for deepfake detection.

Analysis

Model Selection Rationale:

The ResNet architecture was chosen for its proven ability to capture complex patterns in image data, making it suitable for analyzing Mel-Spectrograms of audio signals. Its depth allows for learning hierarchical features, which is beneficial in distinguishing subtle differences between genuine and AI-generated speech.

Model Functionality:

ResNet introduces residual learning through skip connections, allowing gradients to flow directly through the network, mitigating the vanishing gradient problem. This enables the training of very deep networks, capturing intricate patterns in the input data.

Performance Results:

- **Accuracy:** Approximately 50%
- **Confusion Matrix:**
  - True Negatives (TN): 4737
  - False Positives (FP): 0
  - False Negatives (FN): 4743
  - True Positives (TP): 0

Observed Strengths and Weaknesses:

- **Strengths:**
  - ResNet's architecture is adept at capturing complex patterns, which is essential for analyzing Mel-Spectrograms.
- **Weaknesses:**
  - The model failed to predict any positive cases, leading to undefined precision and recall.
  - High computational requirements may hinder real-time deployment without optimization.

Suggestions for Future Improvements:

- **Data Augmentation:** Implement techniques such as time stretching, pitch shifting, and adding background noise to enhance dataset diversity.

- **Feature Engineering:** Explore additional audio features like MFCCs or Chroma features to provide the model with more information.
- **Model Optimization:** Consider using a more lightweight architecture or pruning techniques to reduce computational load for real-time applications.
- **Threshold Adjustment:** Analyze the decision threshold to balance sensitivity and specificity better.

Reflection

1. **Significant Challenges:**
   - Achieving meaningful predictions was difficult, as the model tended to predict only one class.
   - Handling the large dataset required substantial computational resources.
2. **Real-World Performance:**
   - The model's current performance suggests it would struggle in real-world scenarios. Enhancements in data preprocessing and model training are necessary for practical applications.
3. **Additional Data or Resources:**
   - Access to a more extensive and diverse dataset, including various AI-generated speech examples, would likely improve model robustness.
   - Utilizing more powerful computational resources would facilitate experimenting with more complex models and training techniques.
4. **Deployment Considerations:**
   - Optimize the model to reduce latency and computational demands.
   - Implement continuous learning to adapt to new deepfake techniques.
   - Ensure the system can handle real-time audio streams and integrate seamlessly with existing communication platforms.