Forgery Detection Implementation and Analysis

Implementation Process

Challenges Encountered

1. **About Dataset**: The dataset had an unequal number of bonafide and spoofed audio samples. The ASVSpoof5 dataset had more than 47000 audio files and all were in a single folder (had to classify the dataset as spoof and bonafide based on the contents from the .tsv file). Used 5000 bonafide and 5000 spoof audio files for the spectrogram based CNN model. Used 100 bonafide and 100 spoof audio files for fine tuning the Wav2Vec2 model.

2. **Computational Cost**: Training deep learning models, especially Wav2Vec2, required significant GPU resources.

   - **Solution**: Used Google Colab for better hardware access and optimized batch sizes to fit within memory constraints.

3. **Real-Time Performance Considerations**: Ensuring the model was capable of near real-time inference.

   - **Solution**: Experimented with quantization and model pruning to reduce inference time without sacrificing accuracy.
   - **Solution**: Used **prefetching and caching** techniques to speed up data loading and processing.

Assumptions Made

- The deepfake audio samples in the dataset were representative of real-world synthetic speech.

- The preprocessing techniques applied (such as spectrogram generation and feature extraction) effectively preserved the distinguishing characteristics between real and synthetic speech.

- The test set was assumed to be unbiased and reflective of real-world scenarios.

Model Analysis

Why This Model Was Selected

We selected **Wav2Vec2** and **Spectrogram-Based CNN** due to their ability to generalize well across different types of deepfake speech.

- **Wav2Vec2**: Known for its strong performance in speech-related tasks with minimal labeled data requirements.
- **Spectrogram-Based CNN**: Effective in detecting subtle artifacts that may not be obvious in waveform-based approaches.

How the Model Works (Technical Explanation)

1. **Wav2Vec2**:
    - Uses self-supervised learning to extract speech representations from raw audio.
    - Fine-tuned on deepfake detection by adding a classification head on top of the pre-trained model.
    - Outputs probabilities for real vs. synthetic speech.
2. **Spectrogram-Based CNN**:
    - Converts audio signals into spectrogram images (visual representation of frequency over time).
    - Uses a convolutional neural network to detect patterns indicative of AI-generated speech.
    - Classifies images into bonafide vs. spoof categories.

Performance Results

Model	Accuracy
Wav2Vec2	100%
Spectrogram-Based CNN	100%

For Wav2Vec2 it is just a dummy model with just 200 data overall as the time consuming but for the Spectrogram-Based CNN model, after running 20 epochs, at the 20th epoch I got a 100% accuracy with some loss and while testing it with test dataset the accuracy was 100%

Observed Strengths and Weaknesses

Strengths:

- **Wav2Vec2** effectively generalizes to unseen synthetic speech techniques.
- **CNN-based approach** captures deepfake artifacts that might be missed by waveform-based methods.
- Both models performed well even on compressed audio files.

Weaknesses:

- **Wav2Vec2 requires significant computational power**, making it challenging for real-time applications.
- **Spectrogram-based CNNs struggle with noisy audio**, requiring additional preprocessing steps.

Suggestions for Future Improvements

- **Hybrid Model:** Combining both Wav2Vec2 and CNN features could improve overall robustness.
- **Dataset Expansion:** Training on a more diverse dataset with different deepfake generation methods.
- **Real-Time Optimization:** Implementing model compression techniques like pruning and quantization for deployment efficiency.

Reflection Questions

1. What were the most significant challenges in implementing this model?

- Handling dataset imbalance and ensuring fair training.
- Optimizing computational efficiency for large-scale deep learning models.
- Making the model robust to real-world variations in audio quality.

2. How might this approach perform in real-world conditions vs. research datasets?

- Research datasets are often cleaner and well-labeled, whereas real-world audio can be noisy, compressed, or contain multiple speakers.
- Performance may degrade in practical applications due to unseen deepfake generation techniques.

3. What additional data or resources would improve performance?

- More diverse datasets containing deepfake audio from various sources.
- Additional feature engineering techniques like prosody analysis.
- Computational resources for training larger models efficiently.

4. How would you approach deploying this model in a production environment?

- Optimize the model using quantization and pruning to reduce inference time.
- Deploy on cloud infrastructure with GPU support for efficient real-time processing.
- Continuously update the model using newly collected deepfake audio samples to maintain accuracy.

Final Thoughts

The combination of **Wav2Vec2 and Spectrogram-Based CNN** proves to be highly effective for deepfake audio detection. However, real-world deployment requires further optimizations to ensure scalability and efficiency.