Momenta - Audio Deepfake Detection Take-Home Assessment

Devam Tanwani

APPROACH 1:

Analysis of the Wav2Vec 2.0 + SVM Approach

The wav2vec 2.0 + SVM approach is designed to strike a balance between performance and computational efficiency. It leverages a pre-trained self-supervised learning (SSL) model, wav2vec 2.0 to extract the speech features and uses the SVM classifier to perform the final decision-making. This configuration falls under the “Green AI” category, as it minimizes the carbon footprint and resource consumption while maintaining detection accuracy.

Key Technical Innovations

- **Utilization of Pre-Trained SSL Model:**
  Instead of training an entire deep neural network from scratch or fine-tuning a massive model, this approach uses the wav2vec 2.0, which has already been pre-trained on large-scale speech data.

- **Traditional Machine Learning Model:**
  The main classifier is an SVM employing an RBF (Radial Basis Function) kernel. SVMs are known for their robustness in high-dimensional spaces and ability to efficiently separate classes with clear margins. In this context, the SVM is used to distinguish between genuine human speech and deepfake speech by taking advantage of the features provided by wav2vec 2.0.

- **Green AI Considerations:**
  By avoiding the need for extensive GPU-based fine-tuning and reducing the overall number of trainable parameters (808 parameters only), this method dramatically cuts down energy consumption and carbon emissions. This makes it extremely useful for deployment in environments where resources are limited or when real-time, low-latency performance is required on CPU-only systems.

Performance Metrics

- Equal Error Rate (EER): 0.90%

- F1-score: 0.95

Why This Approach Is Promising for Our Use Case

- Using wav2vec 2.0, this approach uses embeddings from the second transformer layer. Studies indicate that these intermediate representations are often more distinguishable for classification tasks than those from the final layer.
- This decision not only reduces the computational overhead by avoiding the expensive deeper layers but also enhances the discriminative power for detecting subtle artifacts in AI speech.
- This low computational overhead allows for rapid inference, which is essential for analyzing ongoing conversations and delivering near real-time feedback on potential forgeries.
- This approach can also be scaled without incurring significant costs or requiring high performance GPUs.
- The emphasis on Green AI not only reduces environmental impact but also lowers operational costs. This is especially important when monitoring large volumes of audio data in real-world applications.

Potential Limitations or Challenges

- **Dependence on Pre-trained Representations:**
  The approach heavily relies on the quality of the wav2vec 2.0 embeddings. If the input audio deviates significantly from the conditions seen during pre-training (e.g., differing languages, extreme noise conditions), the extracted features might not be as effective.

- **Limited Adaptability:**
  Since the model does not involve fine-tuning the SSL component for the specific deepfake detection task, its ability to adapt to new or evolving deepfake techniques may be limited. Future iterations might need to explore hybrid approaches that allow some degree of fine-tuning without compromising efficiency.

- **SVM Scalability in High-Dimensional Spaces:**
  While SVMs are effective for binary classification tasks, their performance may degrade if the dimensionality of the features increases or if the volume of data becomes very large.

APPROACH 2:

Using Spectrogram-based Representations and Transfer Learning

This approach targets the deepfake audio detection task by transforming raw audio into diverse spectrogram representations and leveraging multiple deep learning strategies. The system is

designed to capture subtle spectral artifacts in audio and fuse insights from different models on the ASVspoof 2019 dataset. It not only enhances robustness but also mitigates overfitting.

Key Technical Innovations

- **Diverse Spectrogram Generation:**
  The system extracts six different spectrograms by applying transformations such as Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT) that can also be used to validate the classification.
  Auditory filters (e.g., Mel, Gammatone, linear filters) are applied. This combination enriches the feature set by highlighting different aspects of the frequency content and temporal dynamics in the audio signal.

- Study used three main types of deep learning strategies. I believe the best one for our cause is Transfer Learning. Fine-tuning pre-trained computer vision models (e.g., ResNet-18, MobileNet-V3, EfficientNet-B0, DenseNet-121) to adapt to the spectrogram domain, thereby leveraging the rich, pre-learned representations from ImageNet.

Reported Performance Metrics

- EER - 0.03
- Area under the Curve (AuC) - 0.994

Advantages for Real-World Deployment

- **Robust Feature Extraction:** The use of multiple spectrogram types allows the model to capture a broad spectrum of auditory features. This diversity helps in recognizing deep fake artifacts that may be missed if only a single type of spectrogram were used.
- **Leverages Learned Visual Features**: These models already understand textures, patterns, and spatial features — which directly translates to time-frequency patterns in spectrograms.
- **Interpretable:**CNNs offer interpretable feature maps, saliency visualizations (e.g., Grad-CAM), useful for **explaining why** a sample is considered a deepfake.
- **Potential for Real-Time Applications:**
  With 2-second segmentation, the system is designed for fast inference, which is crucial for real-time detection scenarios in voice-activated systems, smart devices, and security applications.

Potential Limitations and Challenges

- **Computational Complexity:**
  The ensemble framework, while robust, can be computationally intensive. Running multiple models simultaneously might require more powerful hardware, which could be a

bottleneck in resource-constrained environments.

- **Domain Gap Between Images and Audio:** Pretrained models like ResNet or EfficientNet are trained on **natural image datasets** (ImageNet). Spectrograms, though image-like, represent **time-frequency audio patterns** — not objects, textures, or scenes. Without proper fine-tuning, the pretrained filters may misinterpret audio-specific patterns.

- **Generalization to Unseen Deepfake Models -** Spectrogram-based vision models might **struggle to generalize to novel synthetic voices** or unseen generation techniques if they were trained on a narrow set of deepfake generators.

APPROACH 3:

Graph-Based Audio Deepfake Detection Using Integrated Spectro-Temporal Graph Attention Networks (AASIST)

AASIST is a method which uses graph-based deep learning to capture complex time and frequency patterns in audio signals. The approach incorporates a graph attention mechanism within a neural network to effectively model the relationships among these components.

Key Technical Innovation

- **Integrated Spectro-Temporal Graph Attention:**
  AASIST constructs a graph where nodes represent time-frequency components extracted from the raw audio. By applying graph attention mechanisms, the model learns to emphasize the most discriminative relationships between these components. This allows it to capture both spectral nuances and temporal dynamics.

- **End-to-End Raw Audio Processing:**
  Unlike approaches that rely solely on handcrafted spectrogram features or pre-trained image models, AASIST can operate directly on raw audio inputs. Its design integrates spectral and temporal cues seamlessly, enabling the model to detect subtle artifacts introduced by AI-generated speech.

Reported Performance Metrics

- EER - 0.83%

Why It's Promising for Our Use Case

- **Real-Time or Near Real-Time Detection:**
  With careful optimization, the graph-based architecture can be streamlined for efficient inference. Its ability to capture audio relationships makes it a strong candidate for real-time applications where subtle differences between real and synthetic speech need to be quickly detected.

- **Feature Capture:**
  By using a graph attention mechanism, AASIST goes beyond traditional convolutional approaches. It learns to weigh the importance of different time-frequency components dynamically, which is particularly advantageous when analyzing real conversations that may contain overlapping speech or background noise.

- **Adaptability to Challenging Environments:**
  The integrated nature of the model helps generalize across different types of deepfake generation techniques.

Potential Limitations or Challenges

- **Computational Complexity:**
  Graph neural network architectures tend to be more complex compared to standard CNNs or transfer learning approaches. This could translate into higher computational overhead during both training and inference, which might require more powerful hardware or optimization for deployment on CPU-only systems.

- **Model Engineering and Integration:**
  Integrating graph-based models into existing systems may require additional engineering effort as well as a trade-off between model complexity and real-time performance.

- **Scalability:**
  Scaling the system to work reliably in diverse, real-world scenarios might necessitate further fine-tuning or adaptation, especially in environments with highly variable audio quality.