Audio-Video Deepfake Detection

Rapcea Catalin, Voinescu David-Ioan

Introduction

The rise of sophisticated generative models has made deepfake detection a pressing challenge. These manipulated videos often maintain high visual and audio fidelity, making them difficult to detect through traditional unimodal approaches. Our work proposes a multimodal detection strategy that leverages the alignment between audio and visual streams.

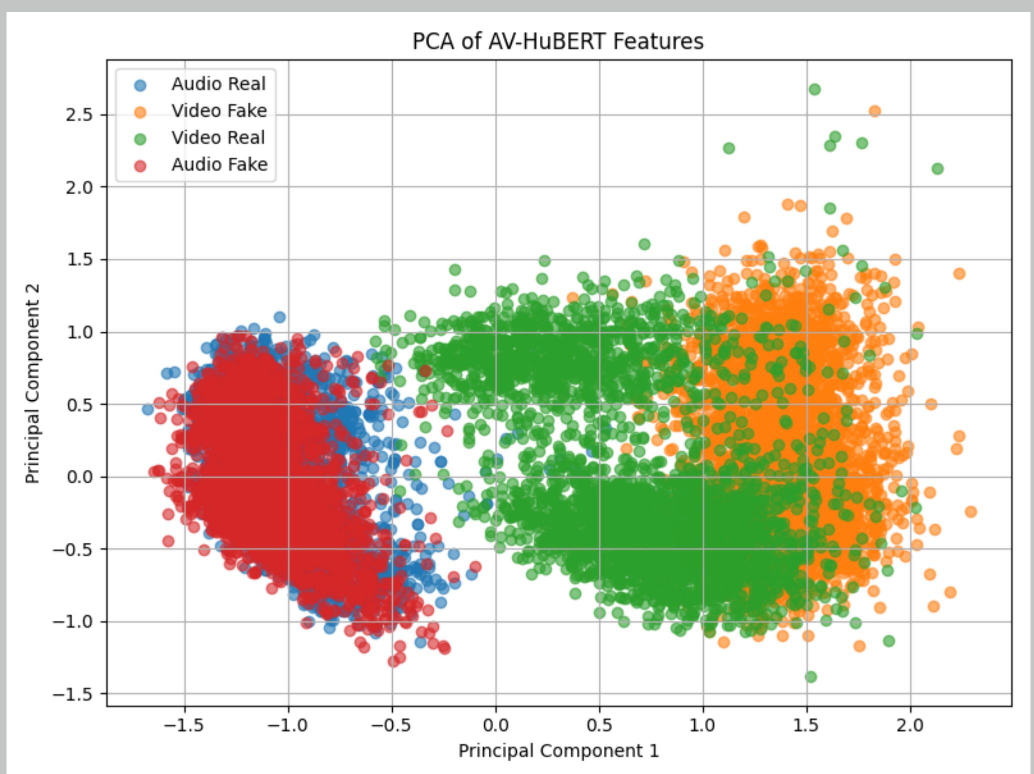
- **Approach**: Use AV-HuBERT, a self-supervised audio-visual model, to extract rich features from both modalities
- **Hypothesis**: Real videos exhibit natural audio-visual synchronization, while deepfakes introduce subtle desynchronization
- **Goal**: Detect deepfakes by identifying temporal and cross-modal inconsistencies using similarity metrics and simple classifiers
- **Dataset**: AVLips v1.0 – 7,000 videos (46.8% real, 53.2% fake) generated with Wav2Lip, TalkLip, and SadTalker

Feature Extraction

From each video, we extract synchronized audio and visual data:

- **Audio**: Extracted at 16kHz from the .mp4 video
- **Video**: Mouth region cropped (96×96) using dlib facial landmarks
- **AV-HuBERT Features**: Extracted separately from audio and video streams using a pre-trained AV-HuBERT model:

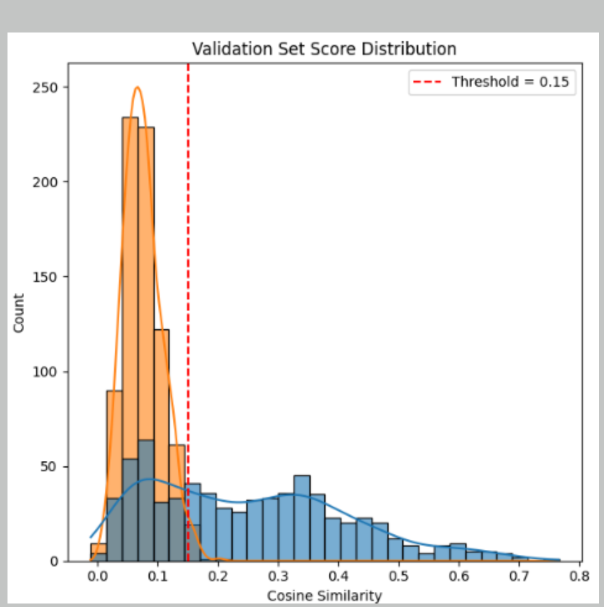
$$f_v = \text{AV-HuBERT}_{\text{video}}(x_v), \quad f_a = \text{AV-HuBERT}_{\text{audio}}(x_a) \tag{1}$$



Cosine Similarity

Cosine similarity is used as a **baseline** method to evaluate temporal alignment between audio and video features:

- Compute frame-level cosine similarity between audio and video features.
- Aggregate frame scores using the median to obtain a single score per video.
- Classify videos as real or fake by applying a threshold to the similarity score.
- Evaluate performance using AUC across various thresholds to select the optimal one.



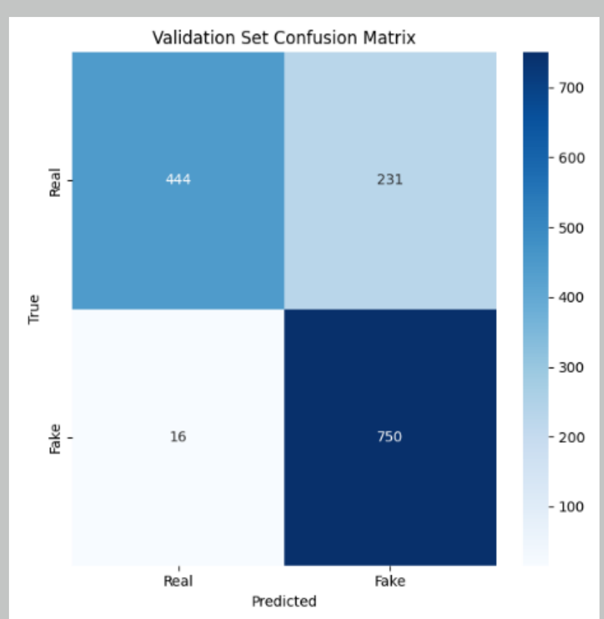


Table 1: Cosine Similarity Results by AV-HuBERT Layer

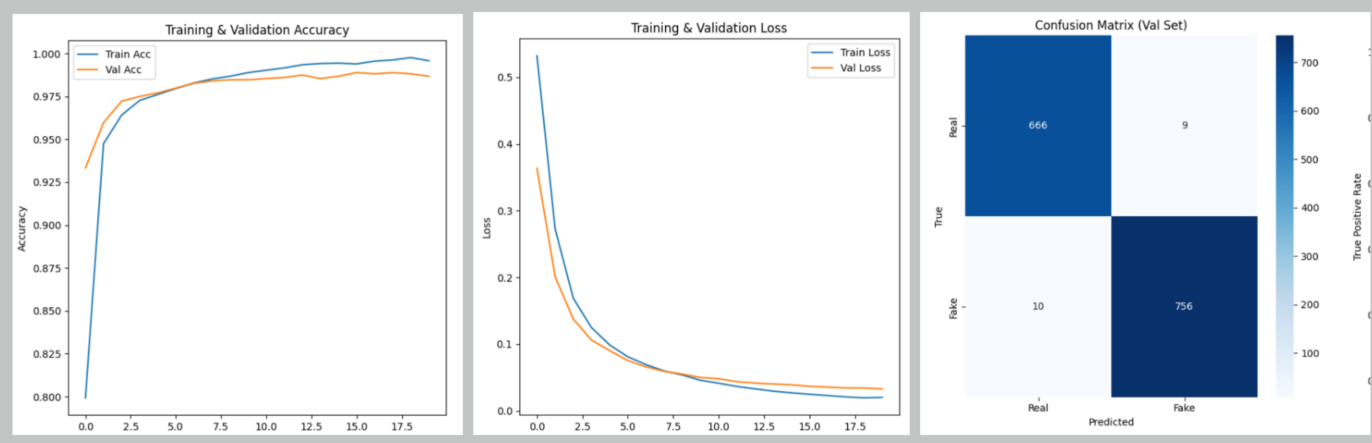
Layer	Threshold	AUC	Accuracy	F1-Score
Layer 6	0.0500	0.7664	0.7738	0.8060
Layer 9	0.0500	0.7477	0.7495	0.7669
Final Layer	**0.1500**	**0.8183**	**0.8286**	**0.8589**

Key Observations:

- **Layer Progression**: Performance improves with deeper layers(Final > 6 > 9)
- **Threshold Sensitivity**: Final layer requires higher threshold (0.15 vs 0.05)
- **F1 Advantage**: Higher F1 than accuracy suggests good precision-recall balance
- **Anomaly**: Layer 9 underperforms Layer 6 - warrants investigation

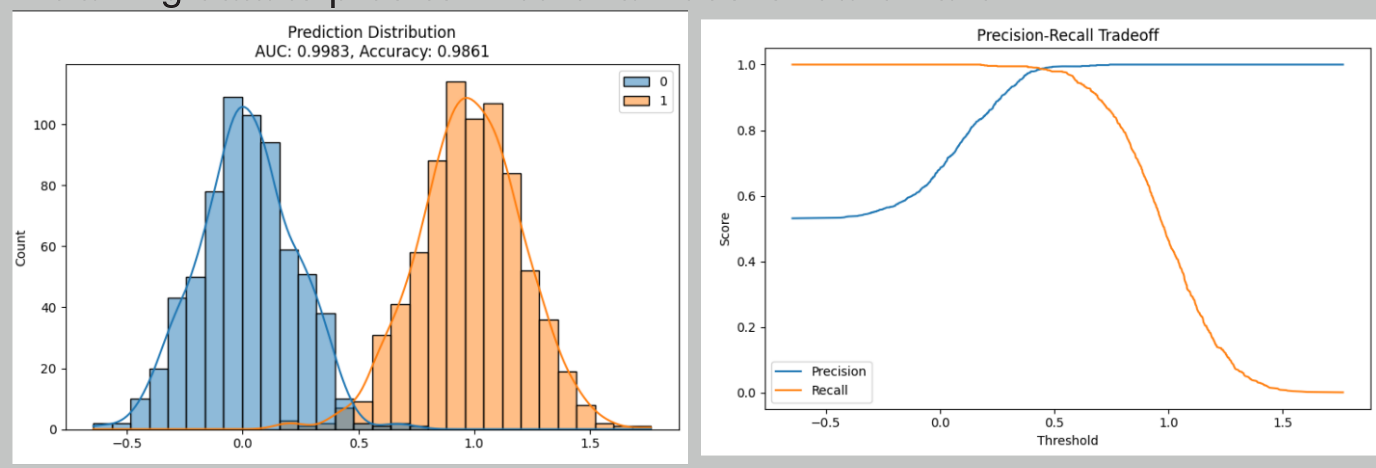
Fully Connected Linear Model

- **Model Architecture**: A feedforward neural network with:
  - Input layer combining audio and video features
  - Hidden layer with 256 units (ReLU activation)
  - Dropout layer (p=0.3) for regularization
  - Output layer with sigmoid activation
- **Training**: Uses Adam optimizer (lr=0.001) and BCELoss
- **Metrics**: Tracks accuracy, F1-score, and AUC-ROC



Linear Regression Model

- **Data Preparation**: Audio and video features for real and fake videos are loaded, flattened, and split into training, validation, and test sets.
- **Feature Scaling**: Features are standardized using `StandardScaler` to ensure consistent input for the model.
- **Model Training**: A linear regression model is trained on the scaled training data to predict whether a video is real or fake.



Models Results Comparison

Table 2: Linear Regression Performance by AV-HuBERT Layer (Test Set)

Method	Layer	AUC	Accuracy	F1	Precision	Recall
Linear Regression	6	0.9934	0.9598	0.9617	0.9700	0.9500
	9	**0.9944**	**0.9702**	**0.9718**	**0.9800**	**0.9700**
	Final	0.9901	0.9604	0.9630	0.9600	0.9700

- **Best Performance**: Layer 9 achieves highest scores across all metrics
- **Decision Threshold**: 0.5 used for all classifications
- **Key Trend**: Middle layers (6-9) outperform the final layer

Table 3: Feed Forward Model Performance by AV-HuBERT Layer (Test Set)

Method	Layer	AUC	Accuracy	F1	Precision	Recall
Linear Model	6	0.9998	0.9931	0.9935	0.9940	0.9930
	9	0.9997	**0.9938**	**0.9941**	0.9945	**0.9937**
	Final	0.9994	0.9903	0.9908	**0.9950**	0.9866

- **Performance Gap**: Linear models outperform cosine similarity by 17-28% AUC across layers
- **Layer Trends**:
  - Cosine: Improves monotonically with depth (Final layer best)
  - Linear: Middle layers (6-9) perform slightly better than final
- **Practical Choice**: Layer 9 offers best tradeoff for both methods

Conclusions

- **Effective Detection Framework**:
  - Demonstrated that AV-HuBERT features effectively capture audio-visual inconsistencies in deepfakes
  - Achieved 99.4% AUC using simple linear models, significantly outperforming cosine similarity baselines (81.8% AUC)
  - Middle layers (6-9) showed optimal performance, suggesting they capture the most discriminative features
- **Key Insights**:
  - Multimodal approaches are crucial - unimodal methods miss critical cross-modal artifacts
  - Feature quality matters more than model complexity (simple linear models outperformed complex architectures)
  - Temporal synchronization provides strong signals for detection