Detecting Lies in Video

Shashwat Sharma [1] Enzo Pinchon [1]

[1]Institut Polytechnique de Paris

Introduction - Is this person lying?

Misinformation dominates public discourse, showcased by politicians like Trump. Traditional lie detection methods rely on physiological signals like ECG or brain activity, making them invasive and impractical. AI represents the next step (non invasive, real time detection), and it's why we decided to work on it!

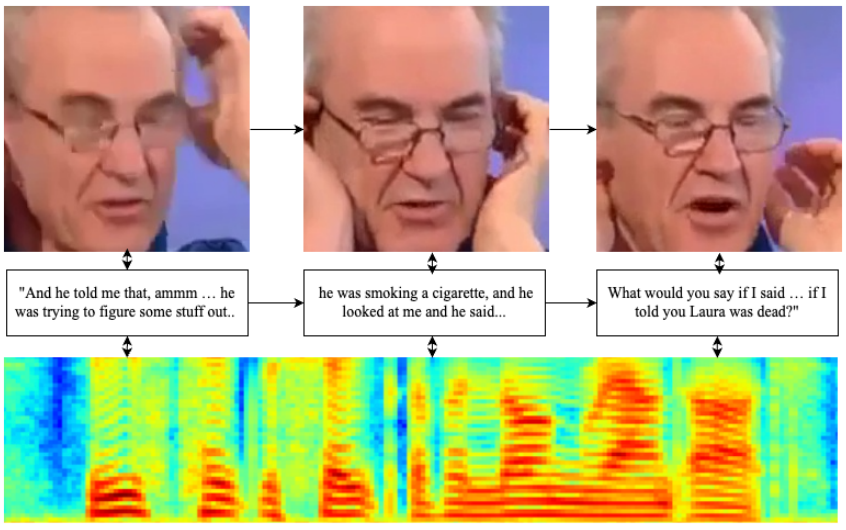

Figure 1. A Training Sample.

Above is part of a data sample that our model processes. Can you identify if the sample is a lie or a truth?

Bibliography

We use the DOLOS dataset which is the Largest gameshow-based deception dataset: 1,675 video clips featuring 213 subjects. Natural conversational deception vs. laboratory-induced scenarios. Rich multimodal data: video, audio with ground truth labels. All participants naturally motivated to deceive (unlike lab settings)

Why Gameshows for Deception Data?

Game shows create high-stakes deception in a controlled setting, perfect for data collection

- **Naturalistic deception:** Participants genuinely motivated to deceive.
- **Clear ground truth:** Definitive knowledge of deceptive vs. truthful statements.
- **Rich conversational context:** Realistic social interaction with stakes.

Baseline Methodology

We employ Parameter-Efficient Crossmodal Learning (PECL) [1] . It integrates Uniform Temporal Adapters (UT-Adapters) within transformer-based encoders for video and audio, ensuring efficient tuning . Additionally, Audio-Visual Fusion module facilitates cross-modal attention, capturing interactions between spoken content and visual cues.

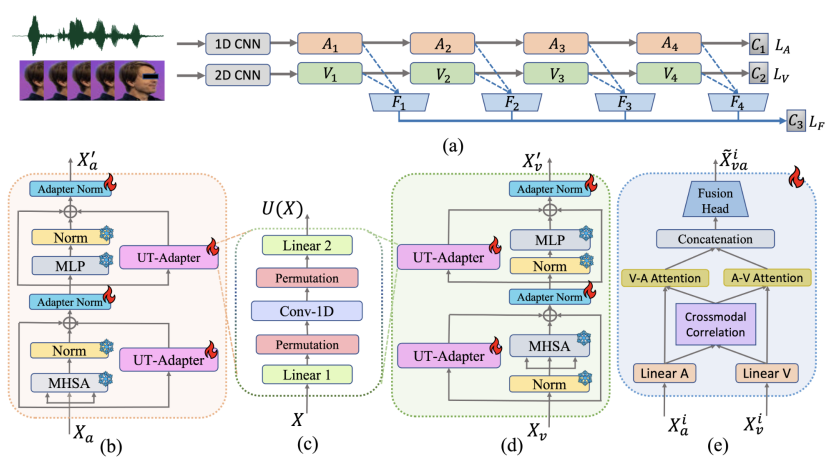

Figure 2. Architectural Overview. [1]

1. **Fine-Tuning:** , UT-Adapters enable lightweight adaptation without full model retraining.
2. **Multimodal Fusion** PAVF captures meaningful interactions between speech and visual expressions.
3. **Multi-task Training:** Jointly learns deception labels and linguistic features for improved performance.

Challenges

Deceptive cues can be subtle temporal patterns across facial expressions, body language, speech, and physiological signals.

- Deceptive behavior **evolve over time** and is context-dependent
- Need for **interpretable models** to explain detection decisions
- **Multiple modalities** need to be fused correctly

We tackle these challenges using an enhanced fused model which can chunk long video sequences and predict lies.

Audio Visual Adapter

The Audio-Visual Fusion module learns crossmodal attention between visual and audio features in an efficient lower-dimensional space. The key is the learnable crossmodal correlation matrix that captures sequence-level interactions:

$$P^i = X_a^i W_P^i X_v^{i\intercal}, \quad \widetilde{X}_v^i = \text{Softmax}(P^i)X_v^i + X_v^i, \quad \widetilde{X}_a^i = \text{Softmax}(P^{i\intercal})X_a^i + X_a^i$$

Extensions

We integrate Whisper for speech transcription, treating it as an additional modality. Using BERT, we analyze linguistic patterns in deception alongside audio and visual cues. To enhance multimodal learning, we introduce cross-attention for pattern linking and a weighted sum for final fusion.

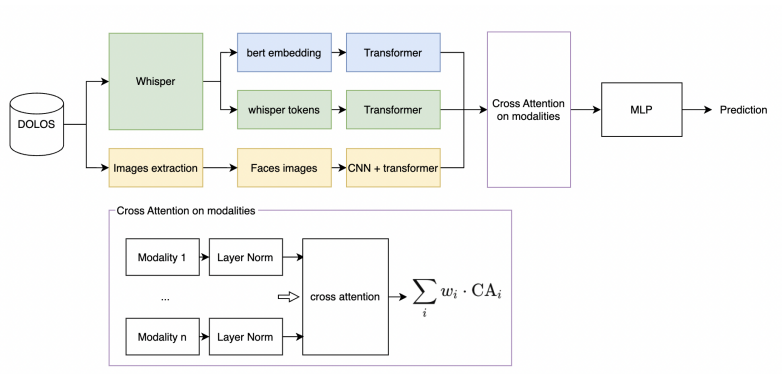

Figure 3. Proposed architecture

Custom Loss Functions

To effectively learn from multiple modalities, we introduce a loss function that balances modality-specific contributions while preserving cross-modal interactions. Our proposed loss combines a contrastive alignment term, modality consistency, and classification objectives.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cross-modal}} + \lambda_2 \mathcal{L}_{\text{modality-consistency}} + \lambda_3 \mathcal{L}_{\text{task}}$$

To align representations across different modalities, we employ a contrastive loss inspired by CLIP:

$$\mathcal{L}_{\text{cross-modal}} = -\sum_{i=1}^{N} \log \frac{\exp(\cos(\mathbf{z}_i^A, \mathbf{z}_i^B)/\tau)}{\sum_{j=1}^{N} \exp(\cos(\mathbf{z}_i^A, \mathbf{z}_j^B)/\tau)}$$

To encourage consistency between unimodal and fused representations, we introduce a modality consistency loss:

$$\mathcal{L}_{\text{modality-consistency}} = \sum_{m} \|\mathbf{z}_i^M - \mathbf{z}_i^{\text{fused}}\|_2^2$$

This loss function ensures that our multimodal model learns discriminative and well-aligned representations across audio, video, and text.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cross-modal}} + \lambda_2 \mathcal{L}_{\text{modality-consistency}} + \lambda_3 \mathcal{L}_{\text{task}}$$

References

[1] Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22135–22145, 2023.

Detecting Lies in Video

Shashwat Sharma [1] Enzo Pinchon [1]

[1]Institut Polytechnique de Paris

Chunk based processing

To extend PECL for long-form video deception detection, we introduce a Hierarchical Temporal Processing framework. First, Whisper segments videos into 2-5 second clips, ensuring alignment with speech timestamps. Each chunk is processed independently using PECL's audio-visual encoders (Wav2Vec2 + ViT), fused via PAVF, and classified at the chunk level. To model long-range dependencies, we employ a sequence-level processing stage, utilizing either (i) a Transformer with cross-attention for capturing global context, or (ii) a sliding-window Transformer with a memory bank for adaptive temporal focus. This enables deception classification across entire videos while providing attention-based interpretability to highlight deceptive segments.

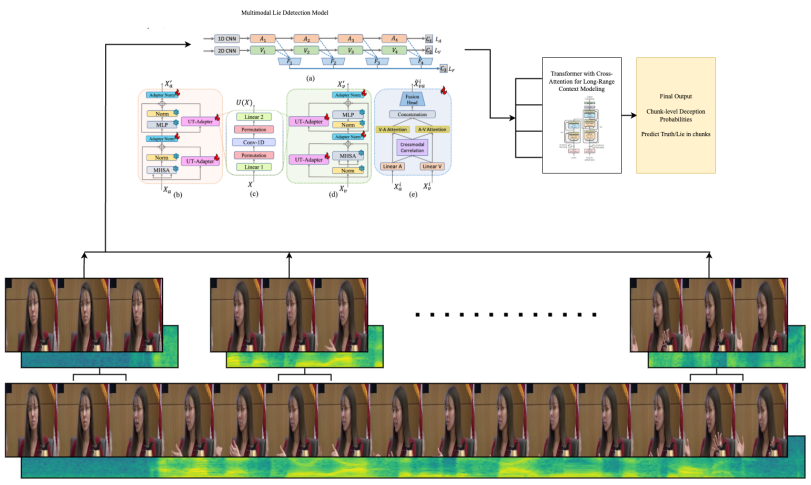

Figure 1. Chunk Processing for long videos.

Video Processing

Transformer-based models like TimeSformer [2] and Video Swin Transformer [3] apply self-attention to learn long-range dependencies. Hybrid approaches, such as ViViT[1] and X-CLIP [4], integrate transformers with CNN backbones for efficient video understanding.

Why chunk videos to smaller clips?

Not only does it make the task for the model much simpler (binary classification vs multi-label), there's also:

- **Long-range dependencies:** Capturing temporal coherence in extended videos is challenging.
- **Model limitations:** RNNs suffer from vanishing gradients, while transformers face quadratic complexity.
- **Multimodal fusion:** Aligning video, audio, and text over long durations increases complexity.
- **Efficient processing:** Chunking, memory mechanisms, and hierarchical models help mitigate these issues.

Baseline Results

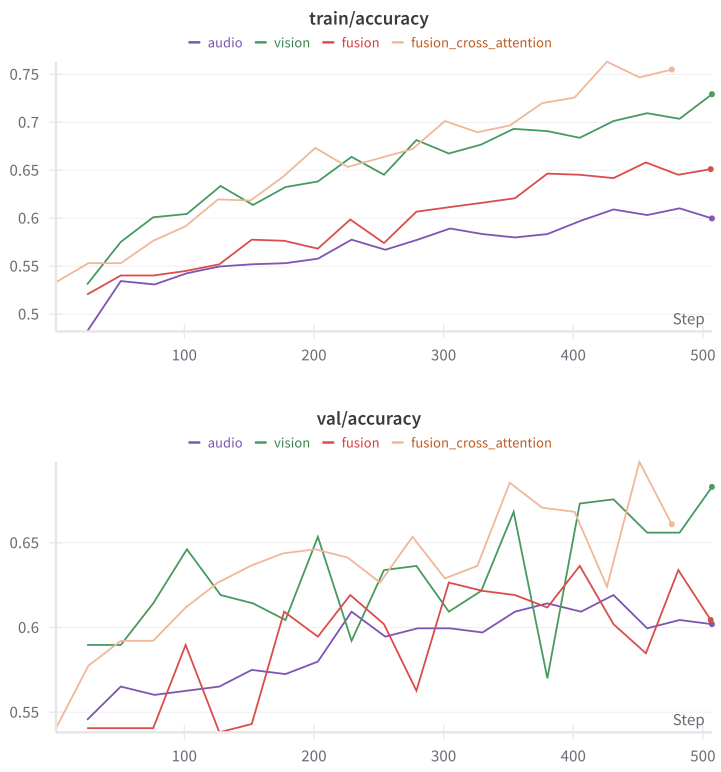

Figure 2. Train and Validation Accuracy.

Qualitative Results

Figure 3 shows the activation map of the first CNN layer, highlighting the regions of interest detected by the model.

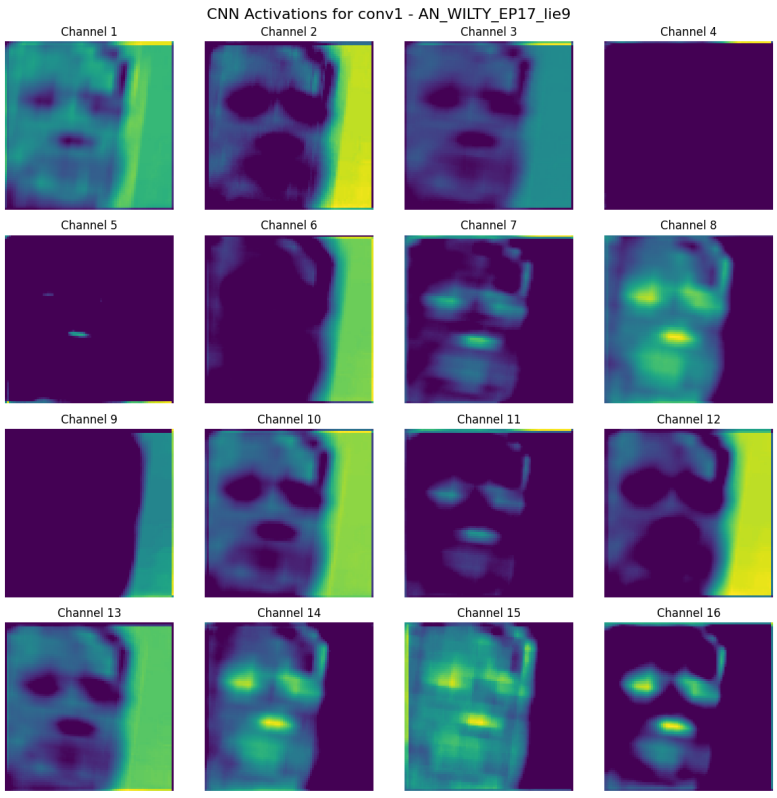

Figure 3. CNN layer-1 activation map.

Multi-task Learning

Multi-task Learning enhances model performance by simultaneously learning multiple tasks, allowing shared representations to improve generalization. In this framework, we predict K + 1 binary labels, where K labels correspond to MUMIN features and last is for lie/truth prediction. The fused multimodal embedding is used for classification. The multi-task loss function is formulated as follows:

$$L_M = -\sum_{k=1}^{K+1} \left(Y_k \log(S_k) + (1 - Y_k) \log(1 - S_k)\right)$$

Interesting Cases

Interestingly, when we trained XGBoost using only these MUMIN features, we achieved 65% accuracy, indicating that these features alone can capture enough information to get results as good as the base fusion model. This suggests that the MUMIN features may contain significant latent information, and if a deep learning model can predict these features more effectively, it could serve as an inherent feature selector.

Future work

3D Vision Transformer for Spatiotemporal Analysis We plan to explore a **3D Vision Transformer (ViT)** to better capture temporal dependencies in deception detection. This approach would tokenize video frames into spatiotemporal patches, leveraging self-attention mechanisms to model motion cues and expression dynamics.

Enhancing the Classifier with Downstream Tasks The DOLOS dataset contains auxiliary labels (e.g., facial expressions). We aim to integrate **multi-task learning (MTL)** by predicting deception alongside these labels, using an **MLP classifier** trained on learned embeddings to refine predictions and improve model robustness.

Framework We dedicated significant time to refactoring the existing code, addressing compatibility issues with libraries, deprecated dependencies, and correcting architectural mistakes, such as missing activation functions. We aim to further enhance the framework and integrate a graphical interface for **real-time video processing**.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 18–24 Jul 2021.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[4] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022.