Ruoye Wang
Jinchen Wu
Weijian Zhang

A Multimodal Method to Detect DeepFake Videos

Overall Project Goals

Motivation

- DeepFake threats
- Manual detection fails
- Learning-based methods in need

Goals

- Prior works: single modality
- Multimodal method
- Increase accuracy

Specific Aims

Specific Aims

Aims

- Multimodal deepfake detection
- Facial and vocal information
- Lip-speech synchronization

Goals

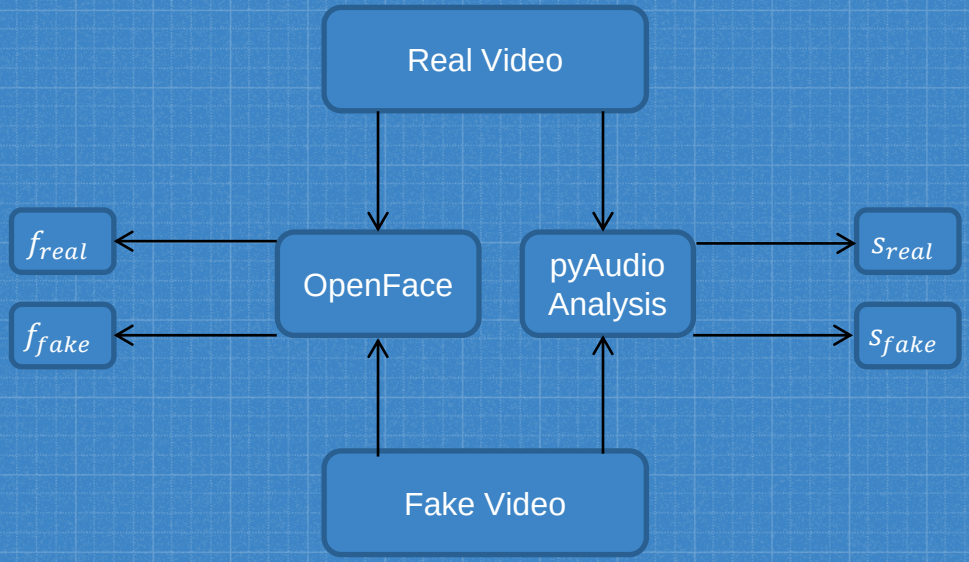
- Design a proper architecture
- Combine features
- Distinguish real/fake

Technical Approach

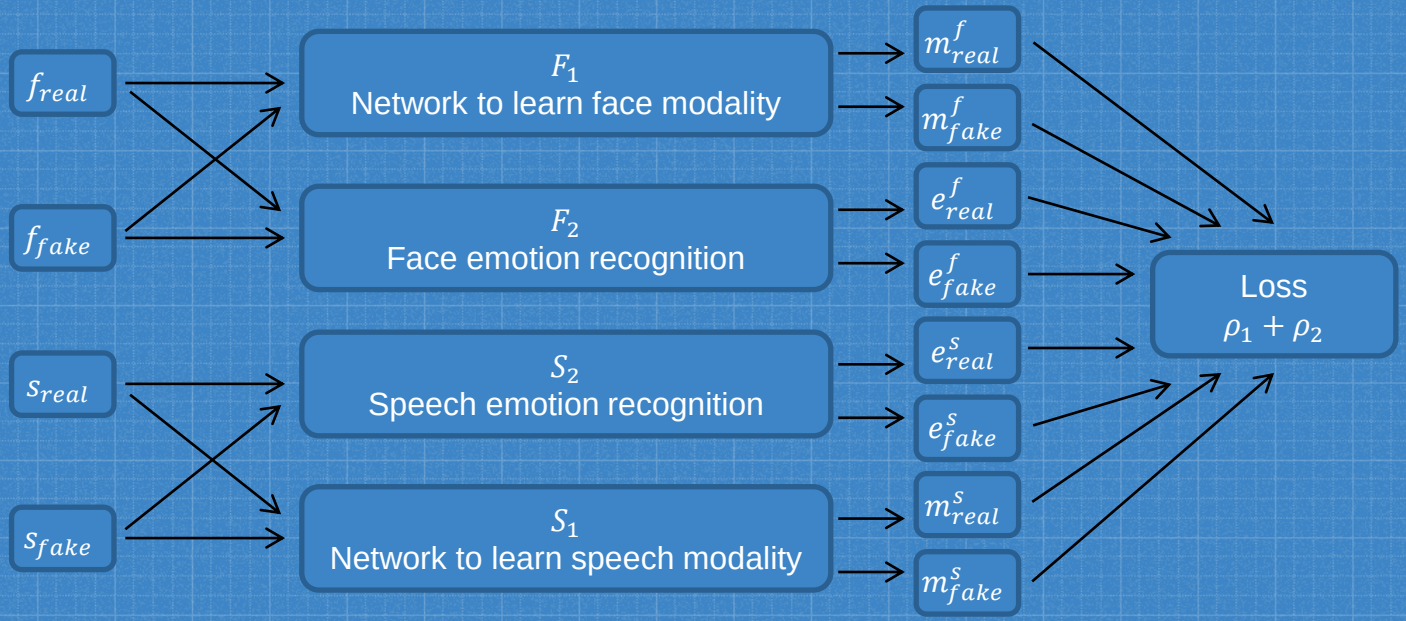
Basic ideas

- Strong correlation exists between different modalities of the same subject, e.g. visual and audio modalities
- Affective cues from different modalities would point to the same perceived emotions.

Feature extraction



Training network



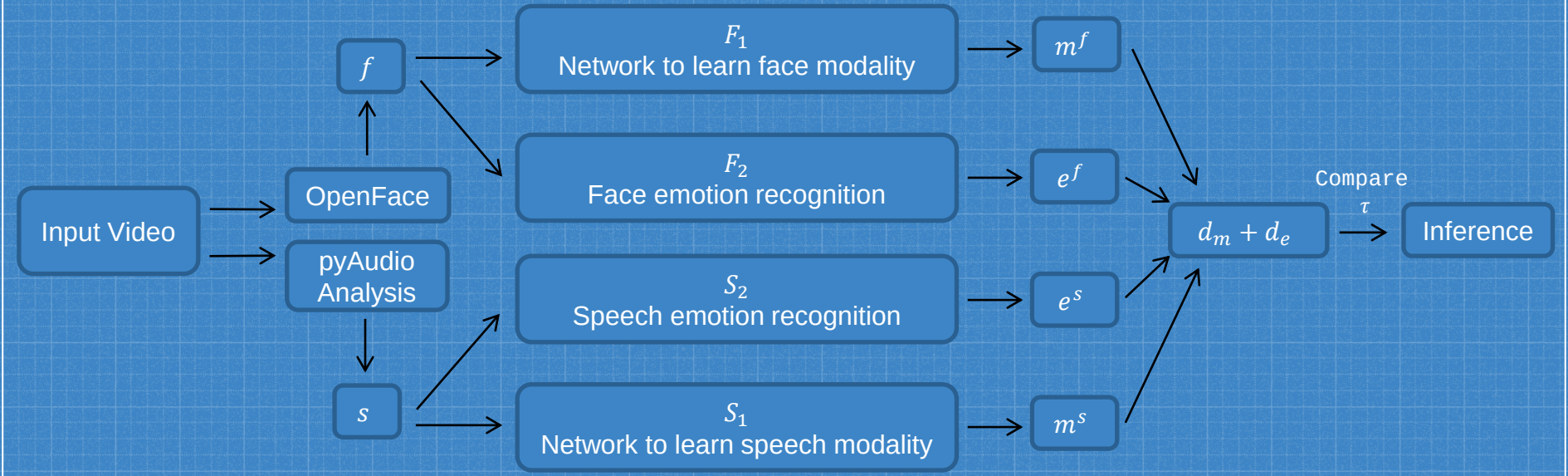
Similarity Loss 1: $\rho_1 = \max(L_1 + m_1, 0)$

$L_1 = d(m^s_{real}, m^f_{real}) - d(m^s_{real}, m^f_{fake})$

Similarity Loss 2: $\rho_2 = \max(L_2 + m_2, 0)$

$L_2 = d(e^s_{real}, e^s_{fake}) - d(e^s_{real}, e^f_{fake})$

Testing network



Distance 1: $d_m = d(m^f, m^s)$

Distance 2: $d_e = d(e^f, e^s)$

Datasets

DF-TIMIT

620 pairs of real and fake videos

DFDC

19,154 real videos and 99,992 fake videos

CMU-MOSEI

describes the perceived emotion space with 6 discrete emotions: happy, sad, angry, fearful, surprise, and disgust, and a “neural” emotion to denote the rest of the emotions

Related Work

Prior work

- Many prior works split the video into frames and study the connection between them:
- Li et al.[1] proposed to use DNN to detect artifacts observed during the face warping step of the generation algorithms
- Yang et al.[2] used the discontinuity of head pose in synthetic videos as a basis for recognition

Prior work

- In addition, different network structures have been employed by many previous works:
- The capsule network structure proposed by Nguyen et al.[3]
- XceptionNet proposed by Rossler et al.[4]

Prior work

- Previous researchers have also noticed that temporal coherence is not enforced effectively in the synthesis process of deepfakes:
- Sabir et al. [5] leveraged the use of spatio-temporal features of video streams to detect deepfakes
- Guera and Delp et al.[6] used a CNN with a Long Short Term Memory (LSTM) to detect deepfake videos

References

[1] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018). https://arxiv.org/abs/1811.00656

[2] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 8261–8265. https://arxiv.org/abs/1811.00661

[3] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2307–2311. https://arxiv.org/abs/1810.11215

[4] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision. 1–11. https://arxiv.org/abs/1901.08971

[5] EkraamSabir,JiaxinCheng,AyushJaiswal,WaelAbdAlmageed,IacopoMasi,and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3 (2019), 1

[6] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–6. https://ieeexplore.ieee.org/document/8639163

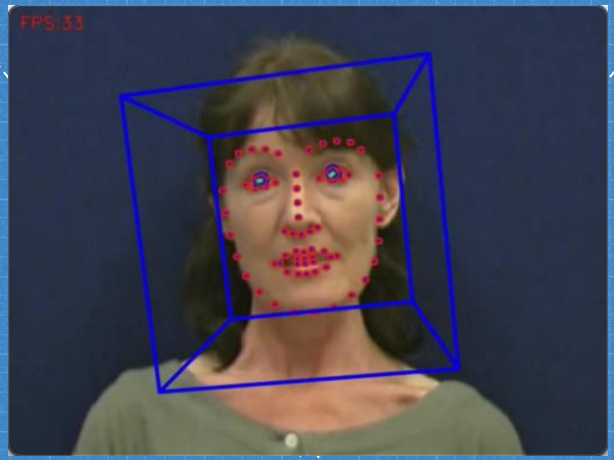
Current Status

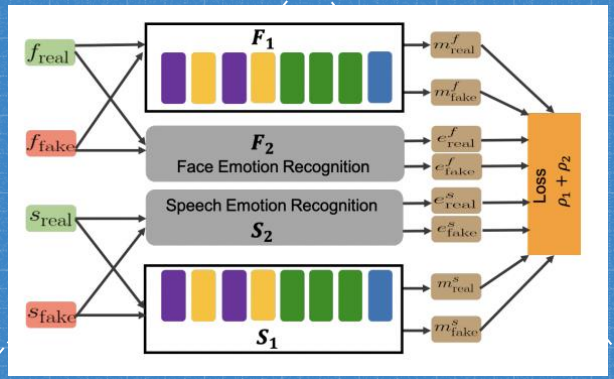
Data preprocessing

- OpenFace
- pyAudioAnalysis

Network building

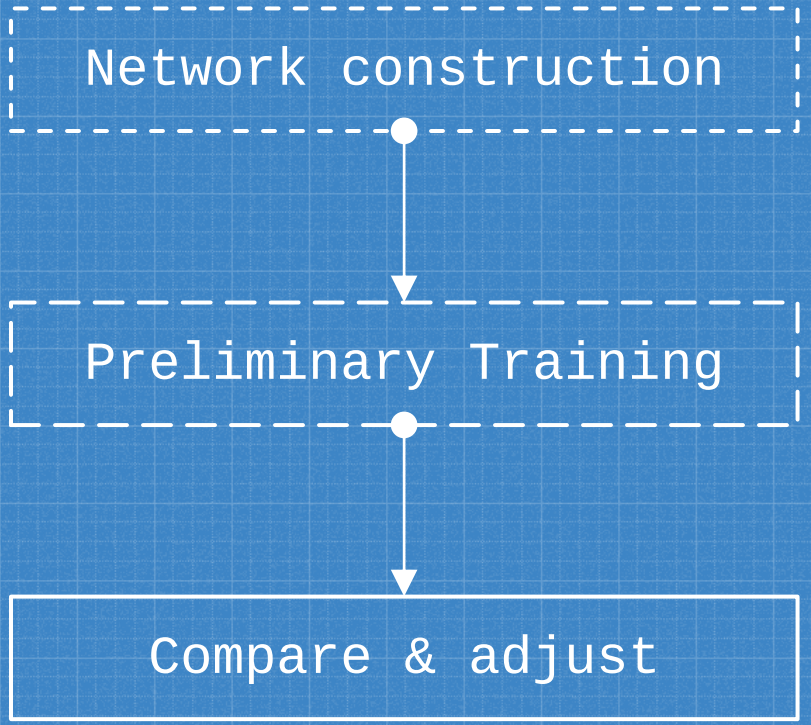
- *Emotions Don't Lie*
- Restore using PyTorch





Next steps

Experiment:
new features



References

[1] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018). https://arxiv.org/abs/1811.00656

[2] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 8261–8265. https://arxiv.org/abs/1811.00661

[3] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2307–2311. https://arxiv.org/abs/1810.11215

[4] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision. 1–11. https://arxiv.org/abs/1901.08971

[5] EkraamSabir,JiaxinCheng,AyushJaiswal,WaelAbdAlmageed,IacopoMasi,and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3 (2019), 1

[6] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–6. https://ieeexplore.ieee.org/document/8639163