Implementing Deepfake Detection with FACTOR

Eva Azazoglu, Zoey Katzive, Subash Matu, Huda Saeed
CSCI1470, Brown University, Spring 2025

Introduction

- Our project aims to detect deepfakes, particularly "zero-day" deepfakes which a model has not encountered before. Traditional methods fail against zero-day deepfakes since they rely on supervised learning, which is tied to known deepfake patterns
- We re-implement FACTOR in Tensorflow, which is a training-free approach that computes the similarity between a visual input and its associated identity claim
- Through capturing inconsistencies inherent to deepfakes, our project hopes to address the growing threat of misinformation and political manipulation that arises from increasingly convincing deepfake generation

Methodology

We used 2 datasets for 2 applications: 50 videos from FaceForensics++ for face swapping (FS) and 50 videos from PolyGlotFake for audio-visual (AV).

Preprocessing

- FS: Extract frames from videos; AV: Extract audio + regions of interest via off-the-shelf detector

Feature extraction

- FS: Extracted via FaceX-Zoo open-source face encoder
- AV: Audio + visual features extracted via Av-HUBERT encoder

Evaluation

- Cosine similarity is computed between the visual and its identity claim; this is called the truth score → then applied average precision and ROC-AUC as evaluation metrics

Ablation

- FS: See if varying false reference set sizes has an impact on ROC-AUC and AP scores
- AV: Changed λ in choosing the λ% percentile truth score per video

MLP extension for AV

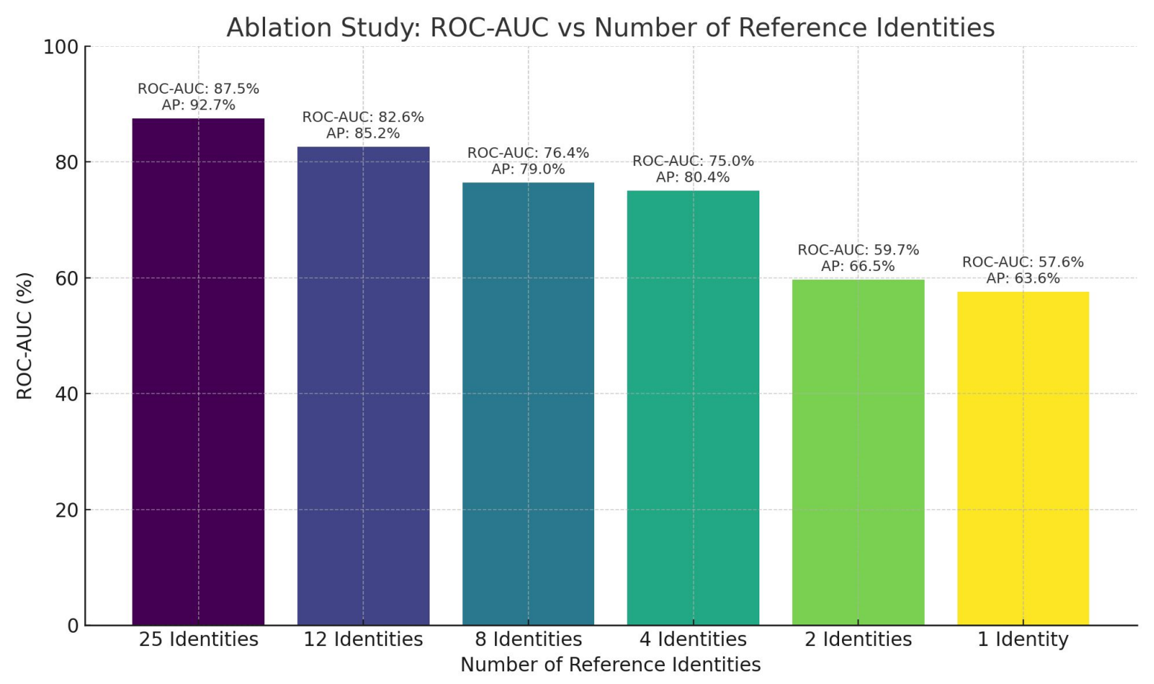
- Trained a 3-layer MLP using Torch (linear layer, ReLU, and softmax layer) which utilizes cosine similarity and audio/visual embeddings to calculate the deepfake probabilities

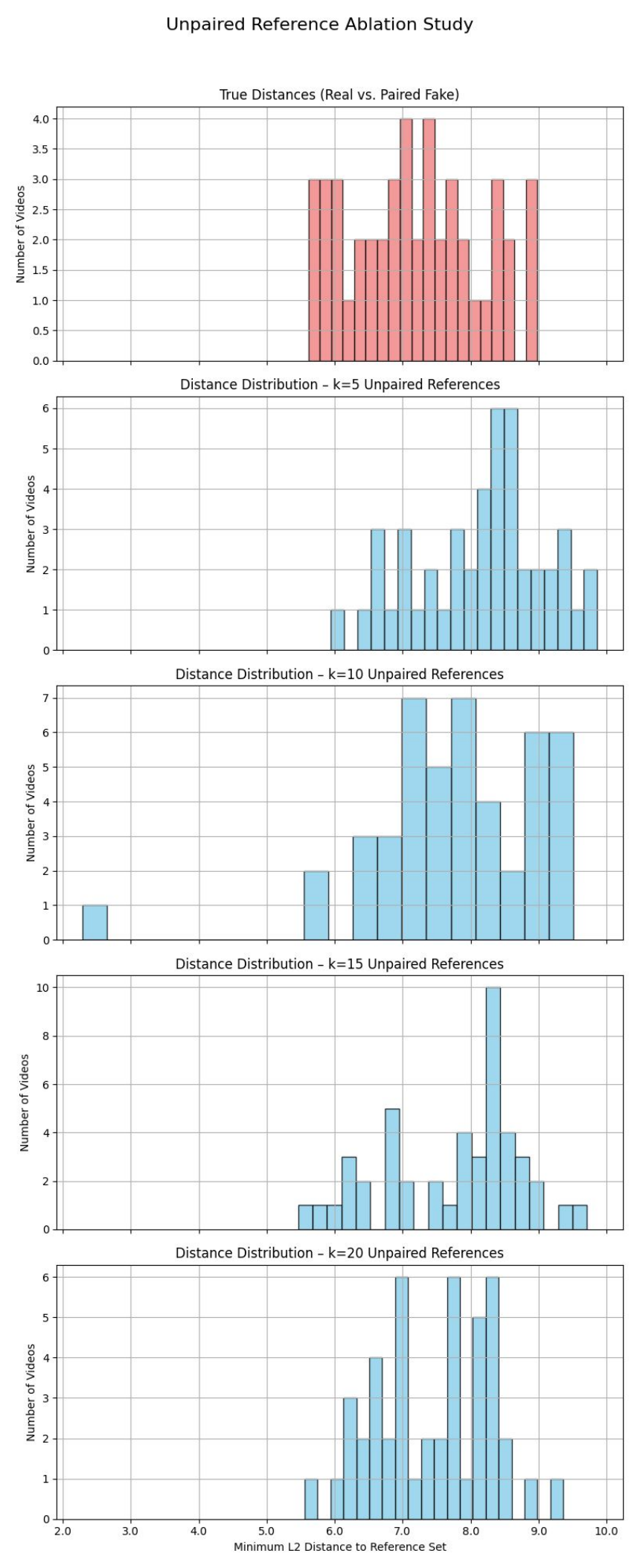
Results: Face Swapping

Ablation Study and Evaluation Results

- We ran the model on varying numbers of false reference identities and calculated ROC-AUC and AP scores to see if the model can still perform without pairs
- Effect of reference identity size: AP and ROC-AUC tend to rise as number of reference identities increases
- Performance still performs moderately well with lower number of reference identities

Left: real identity
Right: deepfake
Model correctly identified the deepfake





- We also compared fakes to true identity (baseline) and to unpaired references (k=5-20) and analyzed minimum L2 distance distributions instead of ROC-AUC/AP
- Effect of reference set size: distance to unpaired references increases with k
- FACTOR distinguishes fakes without pairs; performance improves with larger unpaired sets

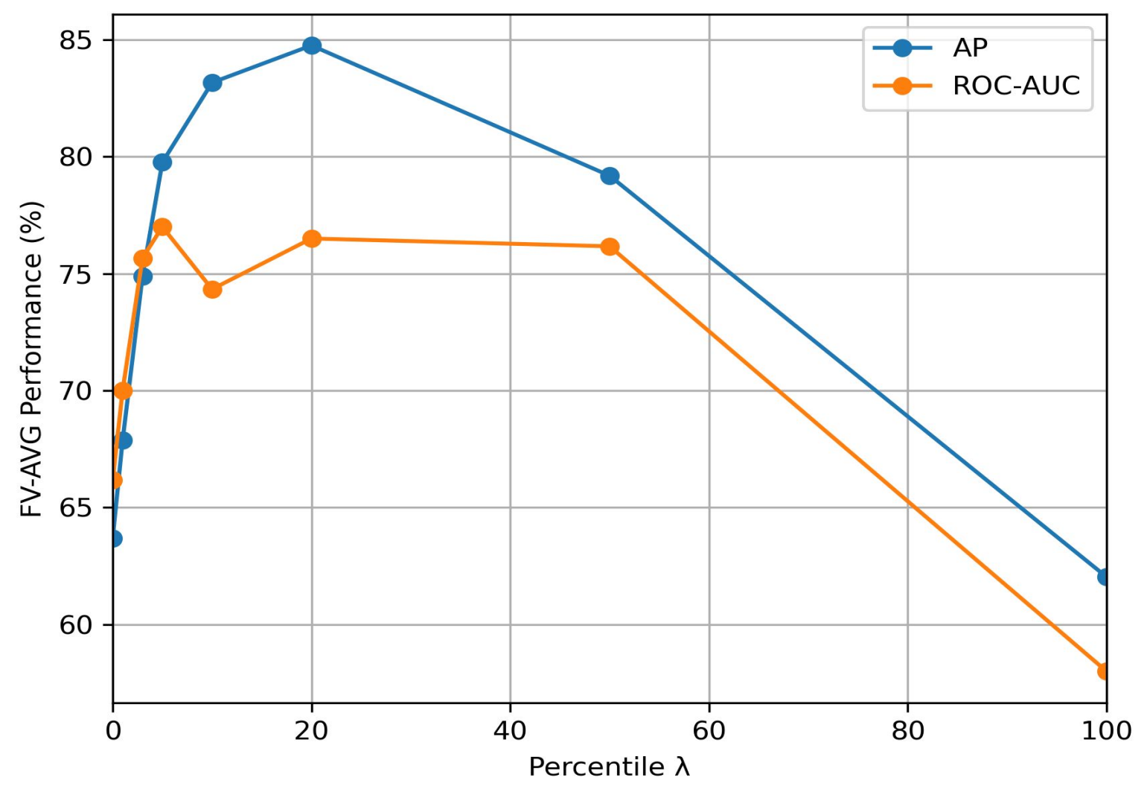
Discussion

- Lessons learned
  - It is not a trivial effort to convert between 2 frameworks (e.g. PyTorch and TensorFlow)
  - Using off-the-shelf encoders as feature extractors can lead to very powerful models
- Our evaluations were conducted on 50 videos, which limits the generalizability of our results and led to some variability between different runs
- Nevertheless, we observed similar results across runs; the trends resembled those in the paper, which supports FACTOR's high level of performance
- With more computational power, it would be interesting to build on the AV MLP and train it on more data

Results: Audio-Visual

Ablation Study and Evaluation Results

- λ percentile effect: AP and ROC-AUC rise with λ up to a peak, then drop off sharply
- Optimal λ≈20 (AP = 84.75, ROC-AUC = 76.50)
- Broad ablation metrics: max AP = 85%, max ROC-AUC ≈ 76%
- Mean ROC-AUC (32 frames) = 78.50
- Mean AP (32 frames) = 84.43



MLP Extension Results

- Training: Cross-entropy loss decreases from 0.67 to 0.6 after 10 epochs
- Testing: all real videos had fake probabilities < 0.5, while all fake videos had fake probabilities > 0.5 → 100% classification accuracy

Citations

Hou, Y., Fu, H., Chen, C., Li, Z., Zhang, H., & Zhao, J. (2024, December). Polyglotfake: A novel multilingual and multimodal deepfake dataset. In *International Conference on Pattern Recognition* (pp. 180-193). Cham: Springer Nature Switzerland.

Reiss, T., Cavia, B., & Hoshen, Y. (2023). Detecting deepfakes without seeing any. arXiv. https://arxiv.org/abs/2311.01458

Rößler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).