A Novel Audio-Video Multimodal Deep Learning Model for Improved Deepfake Detection To Combat Disinformation

Huang, Louis (School: Herbert Henry Dow High School)
Huang, Emma (School: Herbert Henry Dow High School)

Deepfakes are a rapidly improving tool for disinformation. In 2025, around 8 million deepfakes were shared online, with numbers doubling every 6 months. Research has found that people struggle to detect deepfakes, but deep learning has more potential. Existing models lack generalizability due to using only a singular modality (video or audio). Using both video and audio data to detect deepfakes has been little explored. This project finds a solution to the deepfake crisis: creating and comprehensively testing an audio-video multimodal deep learning model for deepfake detection. A multimodal audio-video deepfake dataset containing 20k+ video files was used. Embeddings were extracted using deep learning models (Vision Transformers for video embeddings and Wav2Vec2 for audio embeddings) with three variants: general, deepfake, and emotion. These embeddings were then sent as features to 13 different downstream machine learning models. Over 3,300 models were created using this process, iterating over 71 million files. t-SNE embedding visualizations showed that deepfake embeddings successfully distinguished between real and deepfake samples. Deepfake embeddings had superior performance. The best model overall had an accuracy of 99.49% and an AUC score of 99.71%, significantly outperforming current state-of-the-art deepfake detection methods. A cost analysis model found that this multimodal approach is extremely cost-effective, with a one-time cost of just $7,200 to process the 95 million videos posted on Instagram daily. Moreover, the parts of the face that are most important to deepfake detection were determined. In conclusion, audio-video multimodal deep learning is an effective technique to detect deepfakes and stop disinformation.

Awards Won: