LIP SYNC MATTERS: A NOVEL MULTIMODAL FORGERY DETECTOR

Authors:
Sahibzada Adil Shahzad∗†, Ammarah Hashmi∗‡, Sarwar Khan∗†, Yan-Tsung Peng†, Yu Tsao§, Hsin-Min Wang

PROBLEM STATEMENT

Briefly elaborate on what you want to discuss.

Unimodal deep-fake detectors

Lack of DATASETS

Back to Agenda

04

Deepfake detectors work based on two sets of inputs: video input and audio input. Unimodal Deepfake detectors are primarily designed to detect one type of manipulation, either video or audio. The currently available multimodal deep fake detectors have 2 separate models for catching audio and video manipulations.
Recently, a new kind of Deepfakes has emerged on social networks and online, in which audio and video modalities are manipulated, making such content more challenging to detect due to their multimodal manipulations. The lack of well organised and labeled multimodal Deepfake datasets is also an issue for designing robust and general multimodal Deepfake detection systems.

Deepfakes are usually generated by swapping the source person's face with the target person's face, aiming to make the target person do what the source person does. These artificial intelligence-synthesized media content can be roughly grouped into three [12] categories, namely Lip-syncing, Head puppetry, and Face swapping.
Lip syncing is a video generated in a way that keeps the mouth motion consistent with a specific speech recording, so only the lip region is manipulated. While in the case of Head puppetry or Puppet-master, the target person is the puppet and the person whose action is followed is the master. The puppet-master video animates in a way that the puppet follows the expressions and head and eye motions of the master. Face swapping refers to replacing a source face with a target person's face without manipulating facial expressions.

PROPOSED SOLUTION

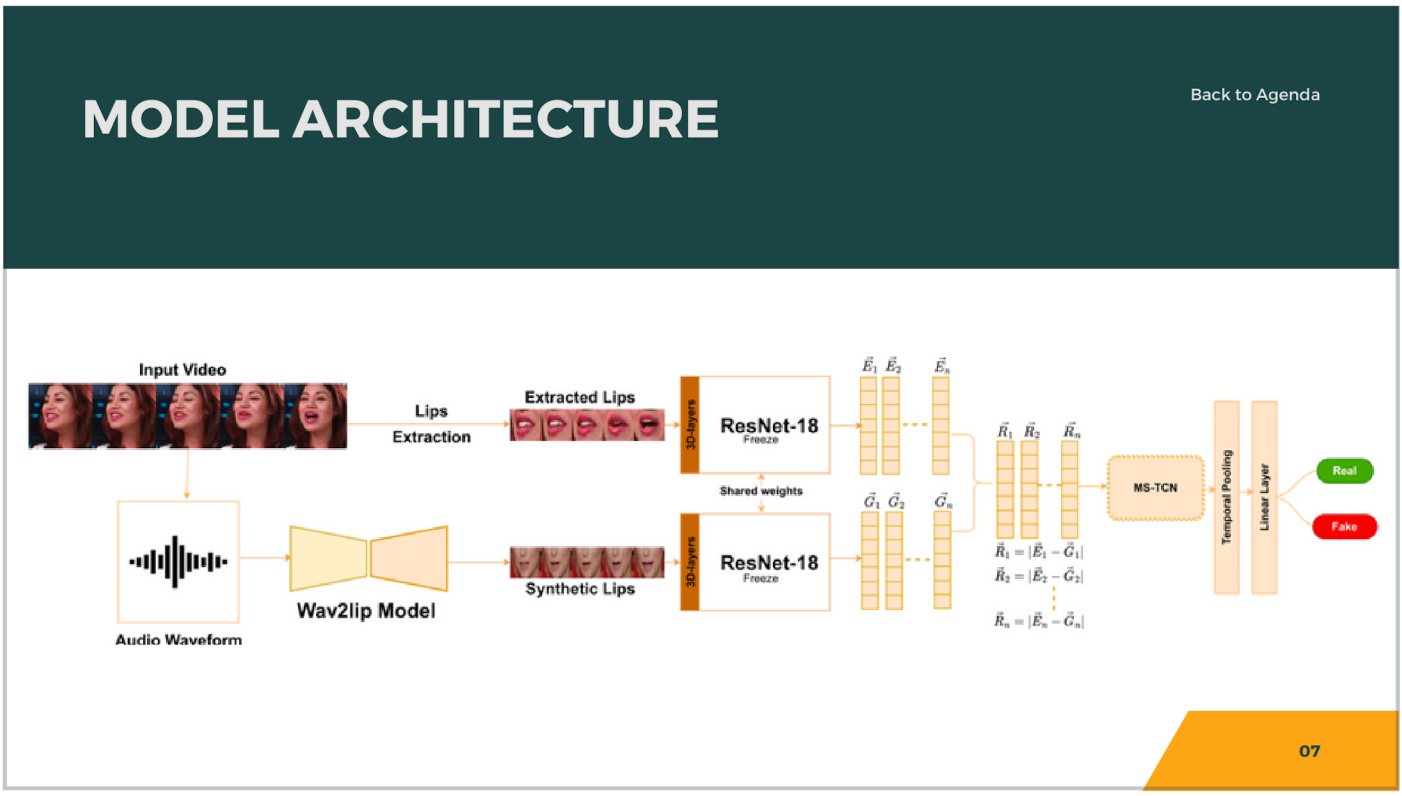
New DATASET-FakeAVceleb	**Game-changing ideology**
Siamese training based idea	**Outperforms state-of-the-art methods**

Back to Agenda 05

The proposed model assumes that the lip movements of the face will not be in sync with the actual lip movements for a particular audio.

First, we extract its lip sequence from the video. Second, we convert the audio modality to the visual modality by synthesising the corresponding lip sequence from the audio stream using the Wav2lip student model [7].

dataset
group of researchers recently released a multimodal dataset called FakeAVCeleb [5]. This dataset is generated from the VoxCeleb2 [6] dataset by selecting videos of 500 celebrities. Each real video is a clean video with only one person's frontal face where the face is clearly visible. The FakeAVCeleb dataset is fairly balanced in terms of gender, race, geography, and visual and audio manipulations. Additionally, it covers many Deepfake generation techniques; thus, deep learning models trained with this balanced and diverse dataset can generalise well.



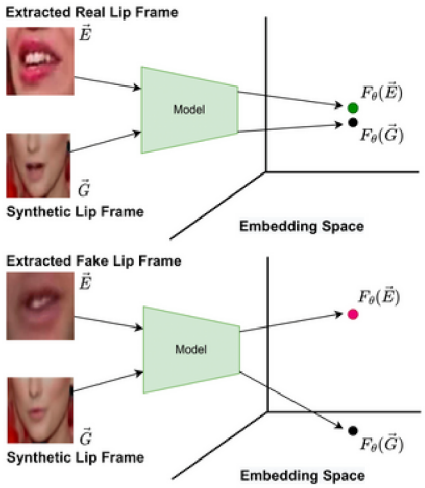
The lip sequence extracted from an input video is transformed into a vector representation sequence, {$\vec{E}$i}Ni=1, by a pre-trained ResNet-18 model. A pre- trained Wav2lip model is used to convert the audio track in the input video into a synthetic lip sequence, which is then transformed into a vector representation sequence, {$\vec{G}$i}Ni=1, by the same pre-trained ResNet-18 model. By sequentially subtracting each pair of vectors ($\vec{E}$ i and $\vec{G}$ i) corresponding to the same time and taking the absolute value, a vector representation sequence, {$\vec{R}$i}Ni=1, of the input video can be obtained. Finally, a MS-TCN model and a temporal pooling layer are used to extract a single vector representation of the input video, and a linear layer is used for the final prediction.

SPECIFICS OF THE ARCHITECTURE

Wav2lip model

MS-TCN

ResNet-18 model



Back to Agenda

08

WAV2lip

The Lip-sync generator [3] considered as the parent model is highly inaccurate on noisy speech input. The Wav2lip model is trained on a single identity to generate an accurate lip sequence for clean/noisy input speech. The student Wav2lip model is trained by a teacher Lip-sync expert [3], a pre-trained lip generating model that synthesizes lip movement on a static face by feeding clean speech. The student Wav2lip model was trained to mimic the Lip-sync expert model by feeding noisy speech with a static face image as input.

RESNET
ResNet-18 pre-trained on a lipreading task.
Training ResNet-18:

The authors of the paper have trained the ResNet-18 model using a large dataset.
The dataset consists of lip sequences extracted from videos and synthesized audio streams.
During training, the model is presented with pairs of lip sequences and corresponding audio strea ms.
The model learns to extract features from the lip sequences and audio streams, capturing their relationship and synchronization.
The training process involves adjusting the weights and biases of the network based on the prediction errors and a chosen optimization algorithm (e.g., stochastic gradient descent).
Fine-Tuning:

After the initial training of ResNet-18, the model is fine-tuned using the lip sequences extracted from videos and the synthesized audio streams.
Fine-tuning involves further training the model on the specific task of lip-sync forgery detection.
This step allows the model to adapt and specialize its learned features to better detect inconsistencies in lip-sync between the visual and audio modalities.

MS-TCN
The MS-TCN layer, also known as the Multi-Scale Temporal Convolutional Network layer, is used in the proposed multimodal forgery detector to capture temporal dependencies and patterns in the input data. In simple words, it helps the model understand how things change over time in the lip movements and audio streams.
Temporal Convolutional Network (TCN):

Think of a TCN as a tool that can recognize patterns in sequences of data.
It looks at neighboring data points together to understand how they relate to each other.
For example, in a video, it can look at neighboring frames to figure out how things are changing from frame to frame.
Multi-Scale:

"Multi-scale" means considering different levels or scales of time.
It's like looking at different time ranges, from short-term to long-term, to understand what's happeni ng.
For example, you might want to look at changes happening over a few seconds or changes happening over several minutes.
How MS-TCN Works:

The MS-TCN combines the power of TCN and the ability to analyze different time scales.
It can capture patterns and relationships between neighboring data points at various time ranges.
By doing this, it can understand both short-term changes and long-term trends in the video.
Importance of MS-TCN:

The MS-TCN is useful when we want to understand how things evolve over time.
In the context of the multimodal forgery detector, it helps us analyze the changes in lip movements and audio streams.
By using the MS-TCN, the model can identify subtle and complex variations in lip-sync over time, which is important for detecting forgery.
In simpler terms, the MS-TCN is like a tool that helps us see how things are changing in a video. It can look at neighboring frames and different time ranges to understand the patterns and trends in the video. This is crucial for detecting lip-sync forgery because it allows the model to spot inconsistencies and changes that indicate manipulation.