Joint Audio-Visual Deepfake Detection

Yipin Zhou Ser-Nam Lim
Facebook AI
{yipinzhou, sernamlim}@fb.com

1. Converted audio deepfakes

To enable the training of joint detection framework on visual and audio deepfakes, we utilize existing video deepfake datasets with the audio channel available and convert these authentic audios into deepfakes by extracting Mel-spectrogram representations and reconstructing the speech from various vocoders [4, 6, 7, 9, 11, 5, 12]. By doing so, the content, speed and the speaker identity of the original speech will be maintained meanwhile synthesizing artifacts are inserted.

To prove the effectiveness of this authentic-to-synthetic conversion, we train an audio deepfake detector sharing the same architecture with the audio stream proposed in the paper using authentic and converted speech from **DFDC** dataset (English speech). We run this detector on in-the-wild (ITW) audio deepfakes from the Internet generated by unknown TTS systems. Specifically, we collect a dataset including 742 videos with synthetic speech from viral YouTube channels [3, 1] and an online demo [2]. We use speech from VoxCeleb [8] (original from YouTube) as real data to balance the dataset.

As comparison, we train the same detector using speech data from ASVSpoof2019 LA [10], which is a benchmark spoofing dataset including synthetic speech generated from 19 types TTS or VC systems. We compare the performance of models trained with converted data (**CVT model**) and ASV data (**ASV model**) on in-the-wild audio deepfake dataset in Table. 1 and Fig. 1. We can see that model trained with our converted fake audios could generalize well on challenging ITW audio deepfakes, except for 'Speaking of AI' YouTube videos, majority of which contain background music or sound effects. **ASV model** has been trained with clean data collected under experimental settings so that it achieves unsatisfying performance on noisy 'VoxCeleb' speech (authentic).

2. Cross-modality attention

To understand how the cross-modality (inter) attention helps the visual and audio deepfake joint detection, we visualize patches around mouth regions from video deepfakes, which achieves top and bottom 10% of attention weights in Fig. 2. On the other hand, we also visualize the audio Mel-spectrogram under the same attention setting in Fig. 3. To show the audio quality we show both real audios and the fake counterparts. The attention weights tend to fire on the sequence location where synthesizing artifacts appear in visual or audios, which might cause the mis-match between lip motions and the speech.

	Acc (%)	AUC
CVT model	81.96	89.55
ASV model	66.62	79.24

Table 1: Accuracy and AUC results on ITW audio deepfakes for models trained with our converted fake audios (CVT) and benchmark audio spoofing dataset (ASV).

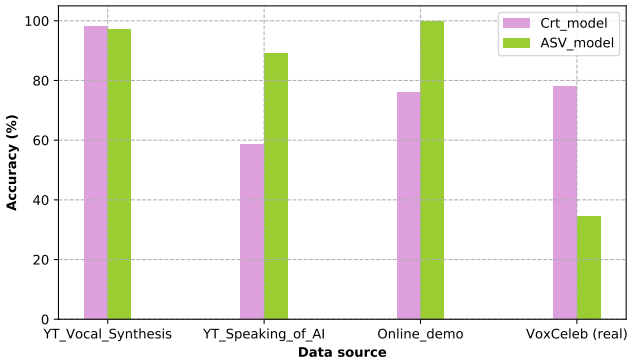


Figure 1: Accuracy of sub-categories of ITW audio deepfakes.

3. Inference

Our proposed model has 3 outputs from the sync-stream and video / audio streams indicating the probability of the whole sequence and each modality being modified. During training, the sync-stream enriches the video and audio representation by discriminating synchronization patterns

Figure 2: Visualization of patches in deepfake video sequences with lowest and highest attention weights at the locations.

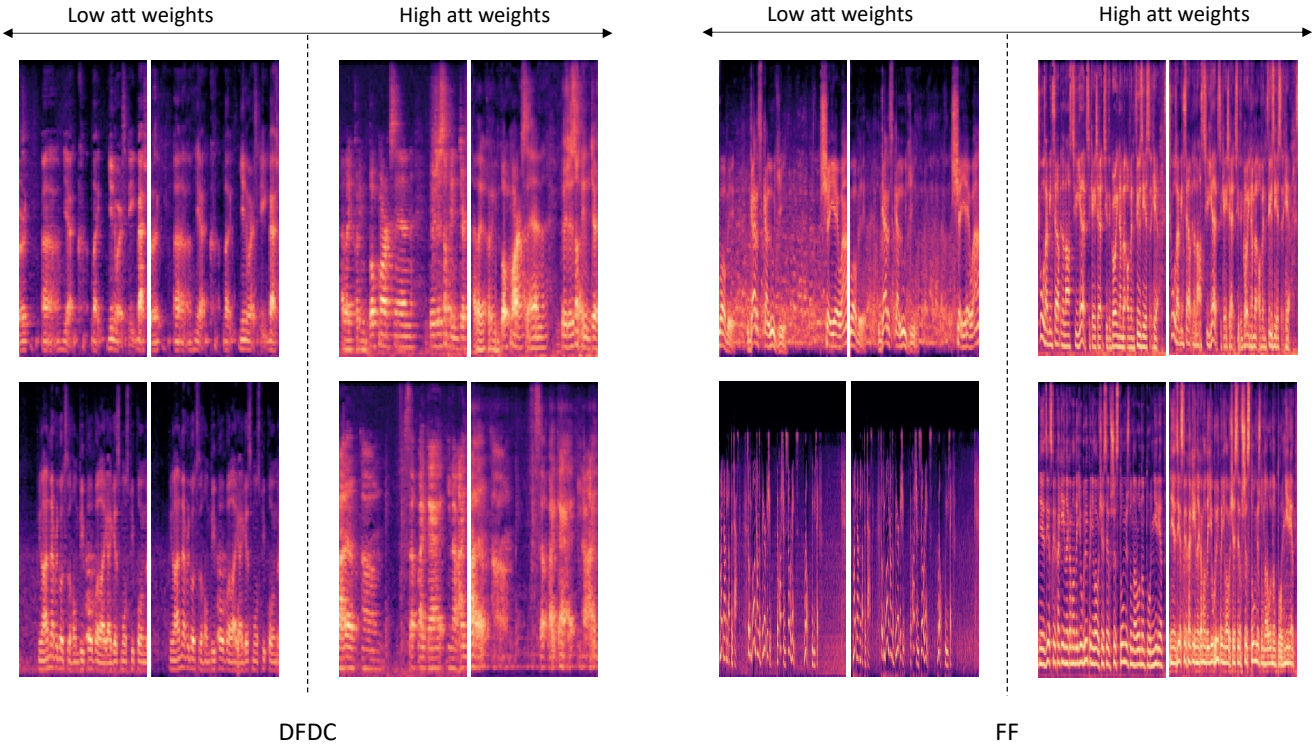


Figure 3: Visualization of audio Mel-spectrograms with lowest and highest attention weights at the locations. For each pair of results, left represents the real audio and right is the fake audio.

between authentic and fake pairs. For doing inference in practice, to understand whether the output of sync-stream can still be an indicator (**sync-stream**) or directly referring the predictions from each modality (**two-stream**) is a better option to determine deepfakes, we compare the performance of the proposed model making inference with these two ways in Fig. 4. The results demonstrate that the sync-stream prediction itself can be a more robust indicator to determine whether the whole sequence has been manipulated during inference.

4. Robustness to noise

We observe that jointly training audio and visual stacks is able to make the model more robust to the noise inserted into the visual modality, which could commonly hap-

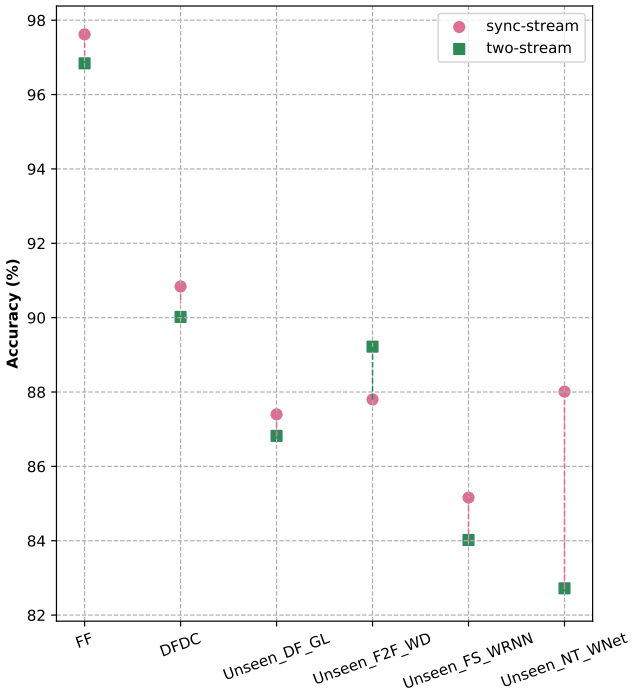


Figure 4: Inference accuracy of utilizing the output of sync-stream (**sync-stream**) compared with referring the probability of each modality (**two-stream**) to determine whether the input sequence is a deepfake. Results from FF, DFDC datasets as well as unseen category evaluation on FF are shown.

pen for deepfake videos due to the compression artifacts. Specifically, We randomly insert guassian noise, JPEG and video compression (with 0.5 probability) to the FF testing videos. For independently trained models, the accuracy of video stream goes down to 93.12% from 98.41% (5.29% drop). Our proposed model (**{2+1}-stream**) got 94.22% with original 98.81% (4.59% drop), which indicates the potential that our method (utilizing extra sync signal) could be more robust to noise.

5. Additional results and visualization

We further show the sub-category performance of visual and audio deepfakes manipulated by different methods in Fig. 5 and Fig. 6. For **DFDC**, manipulation methods are not released, so we only report accuracy for real and fake categories. We could observe that jointly trained deepfake framework (**{2+1}-stream**) outperforms or being competitive with independently trained framework on majority of visual and audio categories as well as averaged results.

We show more visualization results indicating which regions the network focus on to make decisions in Fig. 7. The same with paper, in (a) and (b), we show results from independently trained video deepfake detector (top) and from the video stack of the **{2+1}-stream+att** network (bottom) for **FF** and **DFDC** datasets respectively. By jointly training with the sync-stream, majority of the attention falls on the mouth regions indicating that the correspondence between lip motions and the audio channel could be automatically learned. In (c), we also visualize where the {2+1}-stream framework with shuffled audiovisual pairs pays attention to for making predictions and observe that such correspondence no longer exists for training with unpaired data.

References

[1] Speaking-of-ai-youtube-channel. youtube.com/channel/UCID5qusrF32kSj-oSGq3rJg.

[2] Tts-online-demo. vo.codes.

[3] Vocal-synthesis-youtube-channel. youtube.com/channel/UCRt-fquxnij9wDnFJnpPS2Q.

[4] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. In *ICASSP*, 1983.

[5] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *ICML*, 2018.

[6] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *NeurIPS*, 2019.

[7] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. In *IEICE Transactions on Information and Systems*, 2016.

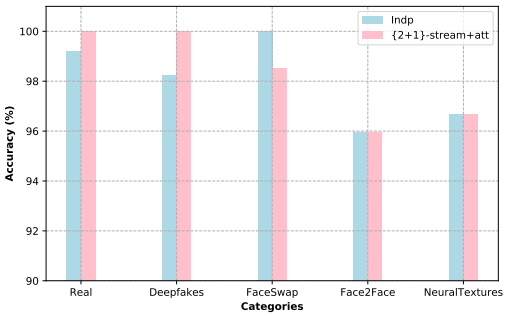
[8] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech and Language*, 2020.

[9] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP*, 2019.

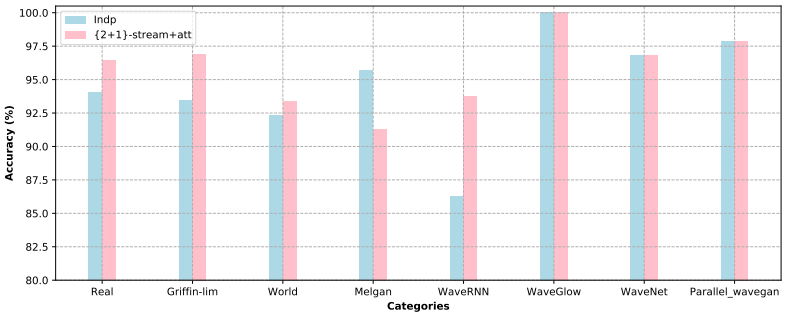
[10] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *INTERSPEECH*, 2019.

[11] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *CoRR*, 2016.

[12] R. Yamamoto, E. Song, and J. Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*, 2020.

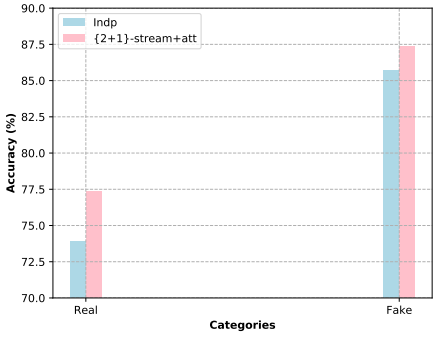


(a) Video stream

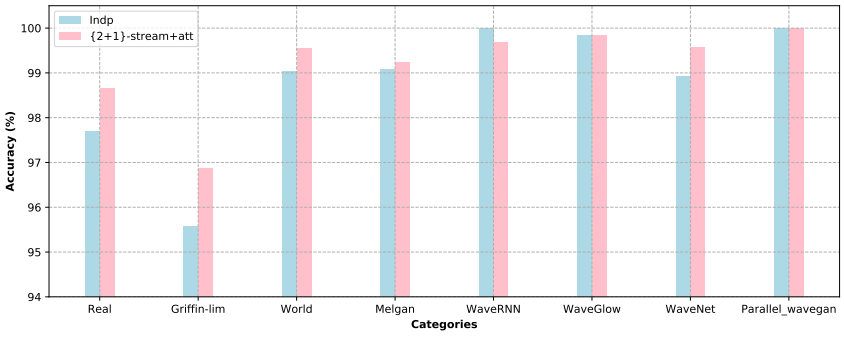


(b) Audio stream

Figure 5: Sub-category accuracy (%) of video and audio streams for FF dataset. The first category represents real data and the rests are deepfake categories.



(a) Video stream



(b) Audio stream

Figure 6: Sub-category accuracy (%) of video and audio streams for DFDC dataset

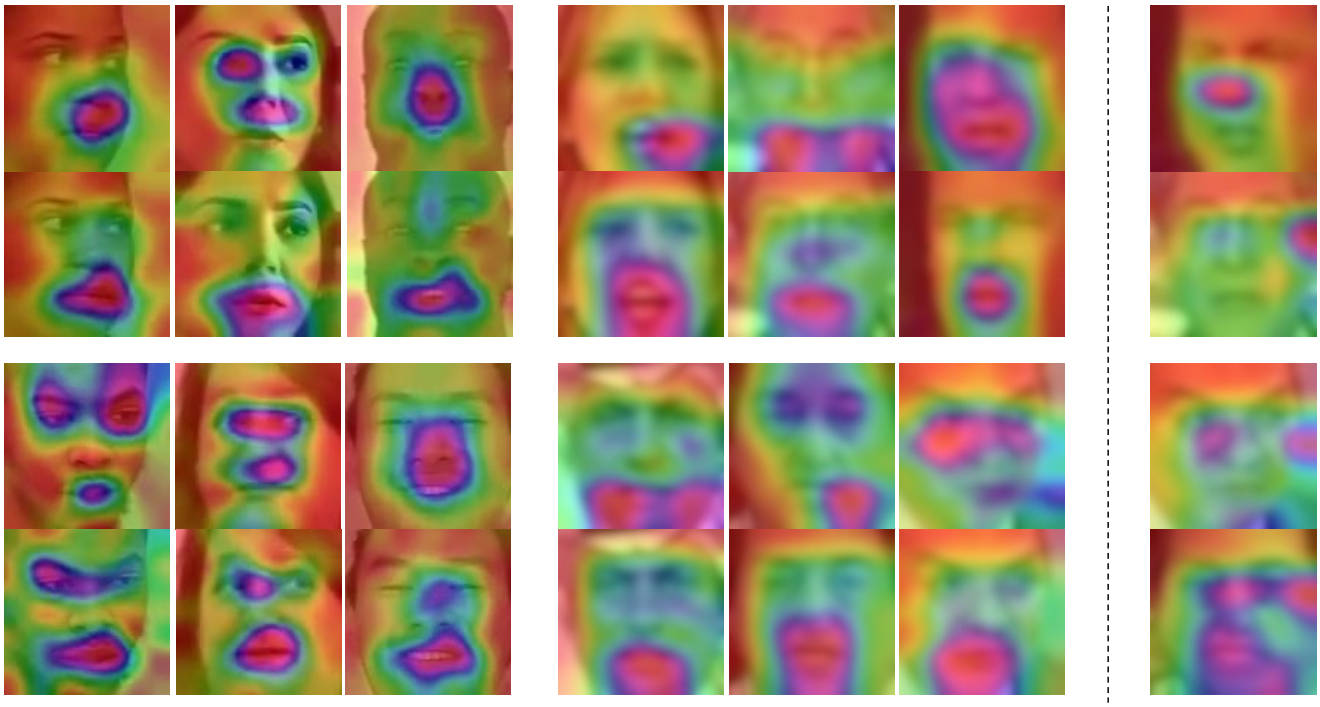


Figure 7: Visualization on where the network focus on while making predictions. (a) Frames from DFDC. (b) Frames from FF (blurred for identity protection). For each set of results, the top row is from independently trained network and the bottom row is from joint detection framwork with sync-stream. (c) Visualization from the network with shuffled audios.