Progress Report: MIT 6.8300
Visually Indicated Sound Discriminator

Georgios Efstathiadis
925942659

Yassine El Janati Elidrissi
936329232

Abstract

Audio-visual synchronization is one of the most critical factors in creating an immersive and engaging experience for viewers and users in various applications such as movie production or video editing. One of the key challenges in achieving audio-visual synchronization is accurately synchronizing sound effects with the corresponding visuals. However, in many cases, it can be challenging to produce realistic and coherent sounds that match the visuals accurately. To address this challenge, we propose a novel approach, VAMM, based on a discriminator model using neural networks to detect fake sounds that are not coherent with the video. The model leverages the visual context provided by the video to distinguish between genuine and fake sounds.

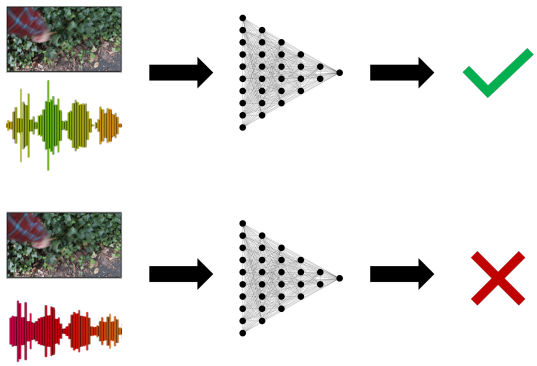

Figure 1. Description of how the discriminator operates. The model detects sound tracks that are not coherent with the video.

1. Introduction

Audio-visual synchronization is essential for creating a compelling and immersive experience in various applications such as movie production, gaming, and virtual reality. In these applications, sound effects must be synchronized with the corresponding visuals to provide a coherent and realistic experience for the user. However, producing realistic and coherent sounds that match the visuals can be challenging. This is where our proposed discriminator model based on neural networks can be useful.

In recent years, there has been growing interest in using neural networks to improve audio-visual synchronization. Discriminator models have been particularly useful in detecting fake sounds in videos, where the audio does not match the visual content. In this paper, we propose a novel approach for detecting fake sounds by utilizing the visual context provided by the video.

Our proposed model is a discriminator neural network that is trained to distinguish between genuine and fake sounds based on the visual context provided by the video. We use a large dataset of drumsticks hitting different objects to train and evaluate the proposed model. The dataset includes various objects such as tables, chairs, doors, and walls, each producing a unique sound when hit by a drumstick.

The proposed model can be combined with other neural networks to improve audio-visual synchronization further. For example, a generator model could be used to generate new sounds that match the visual context of the video, while the discriminator model could be used to detect fake sounds that do not match the visual content. The proposed model can help to create a more realistic and immersive audio environment and improve the overall quality of the audio-visual synchronization.

2. Related Work

There has been previous work in determining if a video sequence matches an audio sequence, with the work done by Arandjelovic and Zisserman [1] being the most similar to our own. They use separate sub-networks, convolutional neural networks, for sound and video to get audio and video embeddings and then combine them in a fusion neural network to get a probability of matching. In a subsequent work [2] Arandjelovic and Zisserman use the audio-visual correspondence as an objective to determine which area of an image produces the sound.

In [5], Morgado et al. propose a self-supervised neu-

ral network to contrast images to multiple sound samples and the opposite. Marcheret et al. [4] tried various network structures for synchrony detection among video and sound, including concatenating video and sound data before parsing through a neural network or concatenating video and sound features provided from other neural networks.

An application of audio-visual synchronization that has been researched extensively is lip-syncing. In [3], Chung and Zisserman present SyncNet which consists of two CNNs that model 5-frame videos and sound separately and apply a contrastive loss to their output to determine if they are matching. Prajwat et al. [7] use a variation of SyncNet as their lip-sync discriminator in their GAN network to generate lip-sync from video and audio.

3. Methods

3.1. Dataset

To investigate visually indicated sounds, we use an existing dataset of videos compiled by the authors of a related paper [6] where a drumstick is used to interact with various objects in different scenes. The use of a drumstick minimizes occlusion of the scene and enables better observation of the objects' reactions, and provides a consistent way of generating sounds across different objects.

In total we have 977 videos and their corresponding audios. The videos have a frame rate of 30 frames per second and the audios have a sample rate of 96,000 samples per second with 2 channels (radio). Each frame in the videos has 3 channels, RGB, a height of 256 and a width of 456. We split the dataset in two subsets - one where the sound tracks match the videos, and one where the sound tracks are split and shuffled/shifted. We also split the dataset into three datasets one for training, one for validation and one for testing with a 80-10-10 split.

When loading the data we normalize the videos to have pixel values between 0 and 1 and we average the two channels of audio. We also select a random 2 second sample of each video and audio to input to the model.

3.2. Model architecture

We are going to try different model architectures and techniques to model the probability of a video and a audio sample matching. In our current implementation we are using FCNNs, so we can use video and audio samples of varying time intervals. We are following a similar architecture as in [1], where we use two sub-networks for video and audio with a fully connected layer in the end with an output of 512 nodes each. We then combine the video and audio embeddings by concatenating them and we apply another fully connected layer with an output of 1 node with a softmax activation which will be the probability of matching.

Both sub-models leverage transfer learning to benefit from pre-trained models, specifically ResNet-18, for better feature extraction. The video sub-model consists of a pre-trained ResNet-18 model followed by a Gated Recurrent Unit (GRU) layer. The ResNet-18 model extracts features from individual video frames, while the GRU layer captures temporal information across frames. The video frames are first passed through the ResNet-18 model, where the fully connected layer has been replaced with an identity operation to obtain the feature vectors. These feature vectors are then fed into the GRU layer, resulting in a single output vector with 512 nodes for the entire video sequence. The audio sub-model also uses a pre-trained ResNet-18 model, but the input is first preprocessed with a MelSpectrogram layer followed by an AmplitudeToDB layer from the torchaudio library. The MelSpectrogram layer computes the spectrogram and maps it to the Mel scale, while the AmplitudeToDB layer converts the amplitude values to decibels. The output of the AmplitudeToDB layer is a single-channel spectrogram, which is then replicated to create a 3-channel input, as required by the ResNet-18 model. The audio features are then extracted by passing the 3-channel input through the ResNet-18 model which output a 512 node vector.

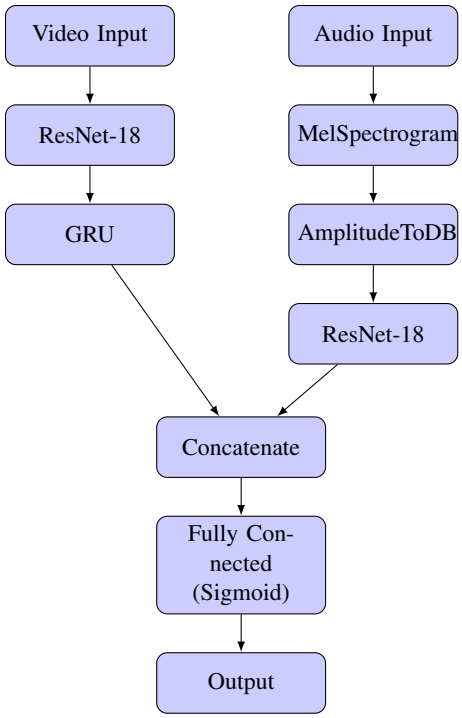

Figure 2. Diagram of the proposed audio-visual synchronization model.

Finally, the output vectors from the video and audio sub-models are concatenated and passed through a fully connected layer with a sigmoid activation function to produce a single output value. This output value indicates the prob-

ability of the input video and audio streams matching.

4. Conclusion

In this project, we aim to present a novel approach for detecting fake sounds in audio-visual synchronization using a discriminator model based on neural networks. Our proposed model, *Video Audio Matching Model* or *VAMM*, utilizes the visual context provided by the video to distinguish between genuine and fake sounds.

We hope our method demonstrates promise in detecting fake sounds, improving the overall quality of audio-visual synchronization. However, there are some limitations and potential areas for improvement. One limitation is the reliance on a dataset containing only drumstick-generated sounds, which may not generalize well to other types of sounds or real-world scenarios. Future work could involve expanding the dataset to include a wider variety of objects and sounds to enhance the model's generalizability.

Additionally, the current model architecture could be further optimized by experimenting with different neural network configurations, such as incorporating attention mechanisms or exploring alternative pre-trained models. Furthermore, our model could be combined with other neural networks, such as a generator model, to not only detect fake sounds but also generate new sounds that match the visual context of the video.

In conclusion, our proposed discriminator model represents a significant step towards improving audio-visual synchronization in various applications, such as movie production, gaming, and virtual reality. Future work will focus on addressing the limitations and exploring potential enhancements to further refine the model and its applicability across a broad range of scenarios.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. 10 2017. 1, 2

[2] Relja Arandjelovic and Andrew Zisserman. *Objects that Sound*. Springer Science+Business Media, 9 2018. 1

[3] Joon Son Chung and Andrew Zisserman. *Out of Time: Automated Lip Sync in the Wild*. Springer Science+Business Media, 11 2016. 2

[4] Etienne Marcheret, Gerasimos Potamianos, Josef Vopicka, and Vaibhava Goel. Detecting audio-visual synchrony using deep neural networks. 9 2015. 2

[5] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement. 6 2021. 1

[6] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds, 2016. 2

[7] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. Cornell University, 8 2020. 2