Regenerating Audio Data from Silent Video through Deep Neural Networks

Adam Haile Konrad Rozpadek Samir Mahmud Alexander Neuwirth

Department of Electrical Engineering and Computer Science
Milwaukee School of Engineering
Milwaukee, WI, 53110
{hailea, rozpadekk, mahmuds, neuwirtha} @msoe.edu

Midwest Instruction and Computing Symposium, 2023

Abstract

Abundant internet video data with high-quality paired audio presents an opportunity to train a model capable of synthesizing new audio for silent video clips. Our work targets a first pass solution to this problem based on deep convolutional neural networks by encoding individual video frames and generating a relevant audio clip. The model is trained using the Youtube-8M dataset. We train a video classification model for frame context and use it to condition a custom WaveNet model trained for video audio generation. Finally, we propose a full audio generation pipeline using these techniques, and discuss the capabilities, limitations and further directions for this approach.

Introduction

The field of audio generation was established relatively recently. The relatively recent development of this research community likely stems from the inherent challenges of working with audio and the wide variety of signals that audio can represent - from speech to music to everyday ambient noises. Some of the more recent breakthroughs come mostly through attempts to generate musical audio. Developing an algorithm capable of writing music has been much more deeply explored and is generally much easier to model as it can be easily represented as a discrete set of musical tones and rhythms (e.g., sheet music or the digital MIDI format) rather than an audio waveform. To enable generation of arbitrary sounds, not just music, our work builds on the smaller body of work around arbitrary waveform generation: producing audio through spectrograms of the output waveform.

Video has become an ever more present form of media in the past few decades; Although not all video has the same form of presentation. Some videos might have a focus on action, change, and vision. On the other hand, some videos have a focus on verbal discussions, text, and other ways of expressing words. The central problem that we addressed has to do with the former, the focus on visuals. Not all videos have audio, or if they do have audio, it might not be very relevant to what is happening on screen. By addressing this problem through the generation of novel, relevant audio it is possible to enhance a video to better convey the ideas expressed to the viewer. Furthermore, it may also enable those with visual impairments to better sense context about the video, increasing accessibility.

Combining these two modalities poses a particularly interesting challenge. Some of the most recent research has come from Vladimir Iashin and others [1], in their research of utilizing transformers to generate a spectrogram from video. For the purpose of our research, we utilize a CNN encoder conditioning a WaveNet style generation approach in order to gain a better initial understanding of how audio can be generated with artificial intelligence, but we do have plans to develop an end-to-end transformer-based encoder/decoder in the future.

Dataset

The dataset that was chosen to train the model was the Youtube-8m dataset. The Youtube-8M dataset consists of video identifiers of YouTube videos that have been assigned labels to fit various common categories and ideas that are expressed in the labeled videos. These videos have been filtered to require at least 1000 views, a length between 120-500 seconds, and a clear display of one of the defined labels in the Youtube-8M label vocabulary. These requirements ensured that the videos that could be used to train the model had a significant amount of accurate audio for the idea that was being expressed.

The Youtube-8M dataset consists of over 6.1 million videos that have been labeled. This resulted in most labels having a large list of videos that can be used to train. In addition to that, Youtube-8M has a vocabulary of 3862 labels which allows the model to be generalized across a wide variety of topics and styles [2].

Video Input Preprocessing

All input video frames are preprocessed before being used as data in the categorizing pipeline. The frames are resized into a 224 by 224 square in order to make sure all input data from any video resolution is normalized. In addition to that, the frames are converted to grayscale to simplify the input. Videos used to train the audio synthesizer pipeline are fetched and have all but their audio tracks removed. The audio is then all converted into the same .wav audio format to be used to train the model.

Processed Image Classification

Each processed frame of video input is categorized via a convolutional neural network, this allows an effective and fast system for scanning our data. We use a 2D convolution to learn filters for our data with a collection of 3x3 kernels, following a flatten in order to reach a densely connected output layer. This is where the convolutional neural network classifies the output into a category to condition the audio generator. As a per-frame operation, this system decides the target audio type for the audio generation process, selecting the dataset the audio generation model pulls from.

WaveNet

For human observers, being able to look at a silent video and make inferences on the sound is often intuitive. By examining the different characteristics of features in a video, one can make grounded guesses as to what sounds they would make from countless examples in learned experience. WaveNet follows a very similar process. WaveNet is a type of generative model, designed by DeepMind, which is able to generate raw audio data. They apply this architecture to realistic text-to-speech generation. However, this is not where the capabilities of WaveNet end. WaveNet has the ability to generate audio on any type of data it is trained on by predicting new samples of audio based upon prior samples. It does this by taking a context window of the previous audio samples at each time step and passing it through a series of dilated convolutional layers. Each layer of convolution has a dilation factor which controls the size of the receptive field. Dilated convolutions were chosen because they have been successfully applied previously in signal processing and image segmentation and provide significant memory size optimizations when applied to very large inputs, such as long audio waveforms [3].

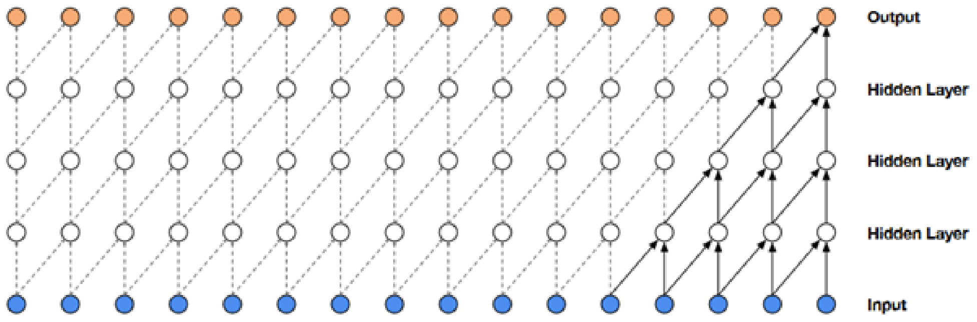

Figure 1: Visualization of a dilated convolutional later [1]

For the scope of our work, we utilized a preexisting implementation of WaveNet for TensorFlow, built by Kotaro Onishi [4], with additional reference to another implementation that was built on a different framework version, built by Igor Babuschkin and others [5].

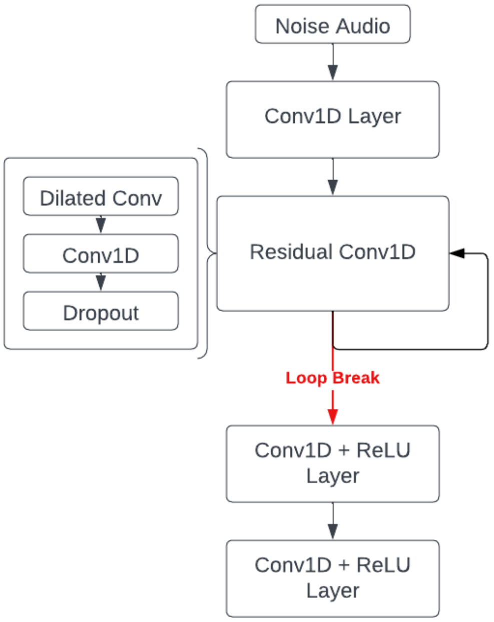

Figure 2: Flowchart of custom WaveNet implementation

Beyond simply extending the style and content found in its training data, WaveNet also enables fine-grained control of its output by operating on two types of conditioning: global and local conditioning. Within the context of our research, apply global conditioning to WaveNet. Global conditioning causes WaveNet to generate with reference to a single, fixed-length vector which possesses all of the style/category data of the audio sample. Local conditioning, which we do not apply in our first-pass implementation, allows for multiple inputs to be added into the sequence which, in our context, would allow for us to add the information of new frames to each time step of the audio. We do not apply global conditioning to this model due to the toolset limitations. Currently, the best public TensorFlow implementation of WaveNet does not include capabilities to generate audio for local conditioning. Implementing such a feature from scratch would require a much deeper analysis into WaveNet's architecture. It is, however, a logical future step for this project, as it would allow for a much closer generation alignment to the source video, rather than focusing the generation with only data from the first frame.

Pipeline

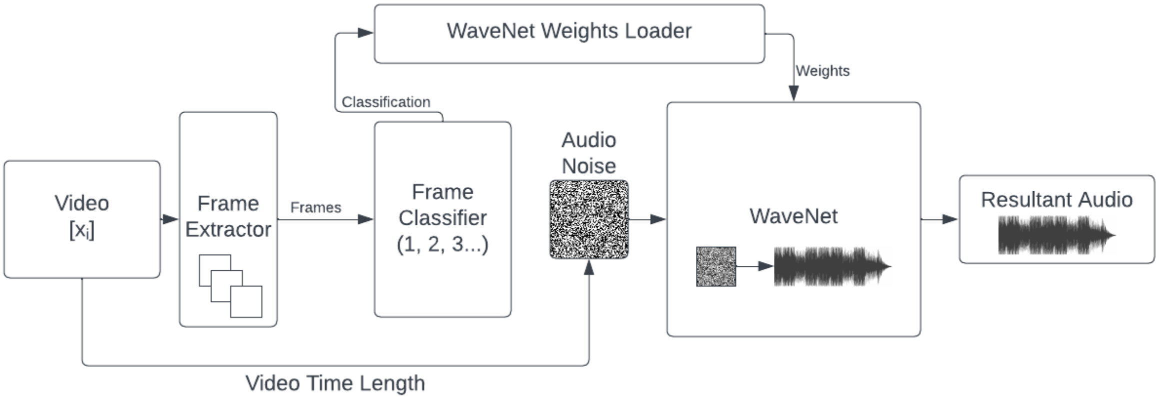

Figure 3: Visualization of implemented pipeline

The pipeline implements two key stages:
1. Categorizing a frame into one of the selected categories that WaveNet was trained on, and
2. Generating audio representative of the video frame’s category.

Categorization

There are 3 steps that go into categorization. The first step is the extraction and preprocessing of frames from a video. These frames are then fed into the frame classifier neural network that outputs a resulting category. Frames from various videos in selected categories are used to train this classifier. This category is then matched with the associated data that should be fed into WaveNet as a global conditioning parameter.

Audio Generation

First, the audio extracted from our training videos is used to train WaveNet along with the relevance weights for the video that was being represented. During the generation of new audio, the output from the categorization step is set as the global conditioning parameter. The length of the video is also fed as a parameter in order to generate an audio track of the same length that consists of random noise. This noise is fed as the input into WaveNet, and the resulting conditioned output is the generated audio output of the pipeline.

Limitation

This pipeline limits how specific the generated audio can be to a video. The implementation is limited to one category, from one frame, to be fed into WaveNet and is unable to synthesize multiple ideas into the generated audio. A more expanded implementation could apply local conditioning in order to have multiple frames feed multiple categories or semantic information unique to specific subjects in the frames as additional conditioning into WaveNet. Furthermore, an expanded implementation could generate multiple segments of audio tracks, then stitch them in order to better reflect rapid changes in visuals.

Challenges

During the process of building this pipeline, we encountered a variety of challenges, many of which we overcame, such as the complex modality-specific issues of working with both video and audio, and also several which we have plans to address in future work. Among these issues is the enhancement from global conditioning to local conditioning.

As described earlier when outlining the differences between global and local conditioning, the WaveNet implementation we have been working with lacks local conditioning capabilities. Local conditioning would prove the capability to factor new information into the generation of different parts of the waveform. A workaround to this is to create short, global conditioning, audio samples. These could then be stitched together to create the full-length sample. This loses out on the major factor of what makes WaveNet powerful though and it is no longer able to use these past audio segments as context for future generation. This is why local conditioning is important, to create a much more detailed generation specific to each frame of the video. Local conditioning should be able to generate much faster than this proposed workaround, as it would be able to continually generate without need to restart generation and be more consistent too.

Translating the meaning of video to audio in general is challenging due to the fact that there is no direct way to express the ideas of a video in an easily transferable way. We attempted this by attaching specific categories that express what is happening in a frame, but this does not express more subtle expressions such as an object moving faster, or an image becoming brighter. These ideas are expressed over multiple frames and cannot be simply determined by examining a single frame.

Future Directions

Much possible improvement remains, but this work demonstrates a full pipeline for extracting category themes from a video and generating relevant audio. We demonstrate that it can be implemented with off-the-shelf tools and components. For a more complete design, we plan to create an end-to-end encoder and decoder to encode an input video and generate the audio corresponding to the frames of the video. This end-to-end encoder/decoder model could be much faster and would condense down much of the bulk of the current model into a format which would be more easily used by others. This encoder would look something similar to what was developed by Vladimir Iashin and others in their transformer design [1].

Prior to the end-to-end encoder and decoder, the implementation of local conditioning into our WaveNet model could provide significant improvements. Local conditioning is necessary to utilize the full capacity of WaveNet.

Conclusion

We propose a deep learning pipeline for audio generation conditioned on paired video frames. Throughout this project, we've explored the capabilities and limitations of tuned, pre-existing components. Through the development of a frame classification architecture, and a custom trained WaveNet conditioned to generate audio in each classified category, we have been able to form the beginnings of future audio generation based entirely on off-the-shelf convolutional neural networks. This structure can be further developed and serves as a baseline for our future development of an end-to-end transformer structure. This type of structure should allow for a greatly improved speed of audio generation, and improved accuracy in the results of generated audio as well. Further research into WaveNet is necessary to properly tune and condition it for optimal generation quality, and to generate based on local conditioning rather than global conditioning. As we continue to further our research into this field, we plan to greatly improve our architecture, and create a full-scale model which can be utilized for accessibility, restoration, and further data recognition. This work represents a concrete step forward toward our eventual goal to have a single model take the entirety of any length video and generate the full audio to match.

References

[1] “Papers with Code - Taming Visually Guided Sound Generation,” *paperswithcode.com*. https://paperswithcode.com/paper/taming-visually-guided-sound-generation (accessed Mar. 13, 2023).

[2] “YouTube-8M: A large and diverse labeled video dataset for Video Understanding Research,” *Google*. [Online]. Available: https://research.google.com/youtube8m/index.html. [Accessed: 14-Mar-2023].

[3] A. Van Den Oord *et al.*, “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO.” Available: https://arxiv.org/pdf/1609.03499.pdf

[4] K. Onishi, “WaveNet Tensorflow v2,” *GitHub*, Sep. 15, 2022. https://github.com/kokeshing/WaveNet-tf2/.

[5] I. Babuschkin, “A TensorFlow implementation of DeepMind’s WaveNet paper,” *GitHub*, Jul. 13, 2022. https://github.com/ibab/tensorflow-wavenet