Sraavya Pradeep
ECE 180DA
Feb 9, 2023

The Future of Sound: Audio LM

What is Audio LM?

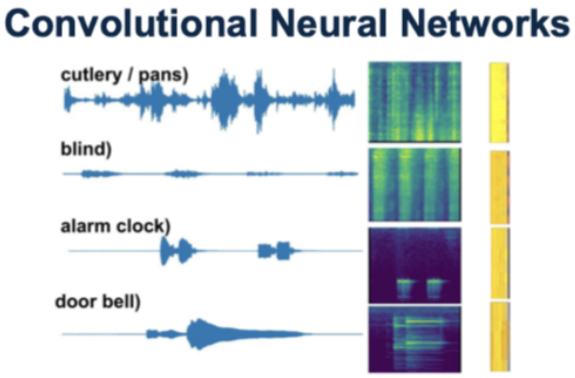
Audio LM is a type of machine learning model that is trained to predict and generate audio data based on the patterns found in a large dataset of audio recordings. It works by analyzing waveforms and using statistical techniques to identify different elements in the sound signal. It can be trained on a wide range of audio data, including speech, music, and environmental sounds, with practical application to speech recognition models, music transcription, and sound classification. Audio LMs have the potential to better equip existing technologies such as virtual assistants and music recommendation systems as it caters to the needs of growing auditory data with efficient audio processing.

What is it built on?

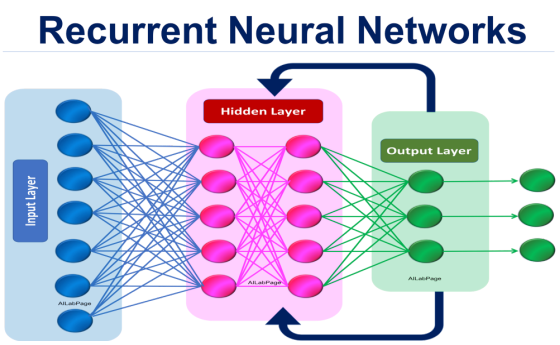
The Audio LM machine learning model is built on a combination of software and hardware technologies. It is typically designed with programming languages such as Python, MATLAB, and C++. Python provides designers with access to powerful libraries such as TensorFlow, PyTorch, and Keras, which have a variety of pre-built algorithms and neural network architectures that can be customized and built upon.

There are three important concepts used:

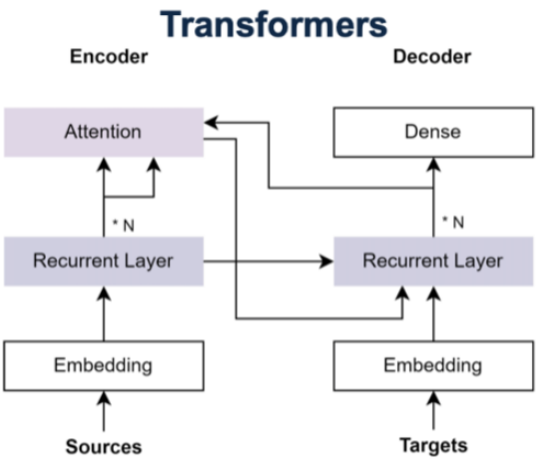
Convolutional Neural Networks (CNN) for Audio Classification: A deep learning algorithm that breaks data down into smaller features using filters that highlight specific aspects of the data; these aspects are called features of the data. The CNN then trains to adjust the parameters of the filters so that it can accurately recognize these features to make classification predictions on the data.



Recurrent Neural Networks (RNN) for Sequence Modeling: A type of neural network used for processing sequential data, in our case we focus on the time series data. RNNs work differently from feedforward neural networks, which process inputs in a single time forward pass. Instead, RNNs use the output of each time as the input for the next, allowing it to capture dependencies at different time steps. Combining this with CNN's, we now have a way to capture features in the data and process its temporal dependencies and sequences.



Transformers: A neural network designed to process sequences of tokens (usually words, in our case sound frequencies generated by a spectrogram), and evaluate them using contextual dependencies. It finds the significance of a token with respect to its neighbors using self-attenuation layers, and implements feedforward layers to change these into fixed-size representations that are used later on. When combined with RNNs, transformers are used to generate contextualized frequency embeddings in the audio track, which is then fed into an RNN for more processing.



Furthermore, it uses **digital signal processing (DSP)** libraries, such as Librosa, Kaldi, and Pymir which provide tools for analyzing sound by filtering, analyzing, transforming, and compressing digital signals. This can help us focus on specific features in the data by removing unwanted noise and better classify the type of waveform we are analyzing. We can also use DSP concepts for **data augmentation**, a technique used to increase the training dataset, by taking existing data and finding new variations through pitch changes, adding noise/reverb and doing time shifts.

We also use **GPU Processing** paired with **Cloud Computing** for efficient, scalable, and parallelizable processing of large data sets. The data is generally stored on cloud platforms such as Amazon Web Services (AWS) and Google Cloud Platform (GCP) which allows for the distribution of computations across many servers, which creates parallel processing. Focusing on GPU processing over CPU makes this even faster, as the GPU can create parallel processes on a given machine, has larger memory bandwidth, and can more efficiently perform calculations required for training audio LMs.

Applications of this Technology

Music Transcription: This is the process of converting an audio recording into sheet music or a MIDI file (a type of digital file that contains musical information on pitch, duration, notes, etc). This allows us to better preserve musical performances and provide a technical analysis on compositions. In the future, this can also allow us to computerize music creation by helping a computer predict the end of a composition, or even create a song from scratch. For example, Google has recently created a new AI system called MusicLM that can create music from any genre when given a simple text description.
Speech Recognition: Converting spoken language to text can be made easier with Audio LM. It is frequently used in consumer-based technologies like virtual assistants, transcribers, or closed-captions for movies and TV shows. We could better understand different accents, languages, and make communication easier for people with different linguistic backgrounds.

This would also make media more accessible, removing language as a barrier for comprehension.

Setbacks

Though Audio LM has paved the way for extraordinary advancement in musical technology, there are still many obstacles to overcome. For starters, the model requires an expansive dataset to accurately find patterns in the music. It can be hard to find large, diversified datasets with high-quality recordings, especially when the model is focusing on a less popular genre, or a specific artist's work.

Another issue is the cost of training this model. There is a significant amount of computational resources that are needed to implement audio Lms. This includes computer power, production environments, and most importantly, a lot of memory power. It can be difficult for smaller companies to get the resources required to get substantial results from Audio LM.

Work Cited

Pius, S. (2023, February 13). *MusicLM and audiolm google's text to music and Audio Tool*. Artificial Intelligence +. Retrieved March 24, 2023, from https://www.aiplusinfo.com/blog/musiclm-and-audiolm-googles-text-to-music-and-audio-tool/

Schmid, F., Koutini, K., & Widmer, G. (2023, February 28). *Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation*. arXiv.org. Retrieved March 24, 2023, from https://arxiv.org/abs/2211.04772

Thursday, O. 06, & Learning, A. M. A. M. (n.d.). *AudioLM: A language modeling approach to audio generation*. – Google AI Blog. Retrieved March 24, 2023, from https://ai.googleblog.com/2022/10/audiolm-language-modeling-approach-to.html