Capstone Proposal: Deep Reinforcement Learning of Physically Simulated Percussive Instruments

cm5409a@american.edu

March 2021

1 Introduction

1.1 Motivation - The drummer shortage

In my experience, musicians trying to start a band are always short a drummer. Since the drum kit is a very cumbersome and loud instrument to have, most of us don't own one, especially if we live in an apartment building. With a high demand and lack of supply, good drummers are hard to find and even harder to record. It seems fair to assume that more excellent music could be made if all musicians had access to skilled drummers.

1.2 Existing Static Solutions

Logic Pro and GarageBand's virtual drummer are examples of virtual instruments which attempt to provide realistic drum accompaniments. A user can place the plugin onto a track which automatically generates MIDI drum patterns. However, it lacks some dimensionality and customizability - it's mainly controlled by a two-dimensional slider with a range between "simple" and "complex" on the x-axis, and "soft" and "loud" on the y axis. You can choose from a variety of "drummers" with different styles and different kits. They even have names, like "Kyle" (Rock), "Rose" (Modern R&B), and "Benny" (Modern Motown).

With Logic Pro's Drummer, users can furthermore adjust several one dimensional sliders like "Hi-Hat," "Kick & Snare," "Swing," and "Fills." What these actually do is not very clear on the surface. For instance, dragging the fills slider higher might increase the frequency at which fills are played, but also may change the style of the fills, or even where they are placed in the rhythm. These controls are here to allow a user without prior experience of playing a drum kit to dial in a style that suits the arrangement. The virtual drummer itself is static, and does not change dynamically in response to the rest of the song.

Steven Slate Drums 5.5 (SSD 5) is a more advanced and customizable virtual instrument, and assumes more knowledge of the drum kit from the user (it's also a great way to learn.) In SSD 5, you can tweak every part of the kit, from the dynamics to the envelope, the bleed into the large array of virtual mics that all attempt to simulate a real drum kit in a studio. Although SSD 5 comes with a large amount of included MIDI grooves, it cannot respond dynamically to the arrangement, like a real drummer would. Therefore, you have to manually tweak the MIDI notes to match your rhythm.

1.3 Existing Dynamic Solutions

Right now is a very exciting time in machine learning research, especially when it comes to music. One of the biggest projects aimed at providing computers with a greater understanding of musical composition is Magenta, a Tensorflow research blog which has provided lots of research, data, software, and even a hardware synth, all relating to the field of generative music. Large publicly available datasets have hundreds of hours of musicians playing instruments, including drums, with appropriately labeled data for training machine learning models. They also spotlight researchers in the community publishing papers on the topics of music generation, like GrooVAE, which is especially relevant to this project.

Magenta Studio is their free software suite that can dynamically generate new beats or melodies based on the existing arrangement in the DAW. It uses neural networks trained on these large datasets to generate relevant musical inferences.

2 Neural Networks

A neural network refers to the web of connections between neurons inside a biological brain. Neurons are interconnected in a highly complex and dynamic structure that performs computations through action potentials - electrical "spikes" of activity that can suppress or activate other neurons.

2.1 Computational Models

Neural networks implemented in computation are modeled after these biological networks, and work on similar (but not identical) principles. These artificial neural networks take in some sort of input data and push it through a series of layers with mathematical operations (often non-linear) that transform the data into a useful output. To obtain a useful output, the model needs to be trained, which in supervised learning, involves showing the model a large amount of examples of what it should output given an input, and minimizing loss via the backpropogation algorithm.

Models that identify patterns or objects in data are known as “classifiers.” Our brain does this constantly - with sound we can identify audible events and draw conclusions as to what caused them. For instance, a car alarm can go off outside your house, outside of our view, and you don’t need to actually see the car to know it’s a car. The same can apply for music - we can identify musical patterns, tones, textures, grooves, and even specific instruments, because our brain has been trained throughout life to maintain strong connections between audible stimuli and a set of memories and associations with the sound.

Computational models which aim to generate new information are referred to as “generative” models. We also do this all the time in our brains. Although our thoughts are confined to our own heads, we can take external actions on the world by moving our muscles. These voluntary movements are processed in the Limbic system and the Basal ganglia.

In music, we can combine our auditory classification abilities with our sense of creativity, to generate a series of muscle movements that allow us to sing, dance, and play instruments. We are simultaneously harnessing several different types of networks to generate new musical information in response to our surroundings.

2.2 Generative Audio Networks

There’s a few areas of active research which have gained momentum in the last few years due to the proliferation of better datasets and faster computing hardware. Some research has focused on musical scores and the information associated with it. MusicVAE takes existing melodies and morphs them into alternate styles, by condensing down high level musical ideas into a set of latent variables within the network. Projects like MidiMe take this one step further, creating a network that can learn an individual’s musical style, and present new musical suggestions. For these, the neural network can take an input of MIDI notes, and output another series of complementary MIDI notes. This can help people compose music with automatically generated musical suggestions. Often these research projects express their intent to revolutionize creativity by providing creators with new tools.

Others have focused on generating raw audio, synthesizing new or contextually useful sounds with neural networks. A few notable examples include WaveNet, WaveGan, and GANSynth. These projects are exciting because they provide entirely new ways to synthesize sounds.

3 Human Music

The way we make music is vastly different from the existing computational implementations. We use our muscles to generate movements, which in turn, impart a force on a physical object which generates an acoustic resonance or percussive impact. Even with singing, the same applies - with the physical object being our larynx, and the acoustic resonance occurring in our trachea.

We operate in a physical system with external constraints, and it defines how we turn sounds into music. Maybe it's possible the music we generate in our heads is somewhat related to things like GANSynth, WaveNet, or MusicVAE (although these are incredibly simple comparatively,) but we must always use physical means to actually create and share it.

This is why learning a new instrument is hard - it's a highly complex process in which our multi-modal experience of hearing, physical touch, sight, emotion, memory of past experiences, etc., are all continuously combined into discrete and deliberate physical actions which result in an acoustic phenomenon that other people can agree is "musical."

The unifying theme of *human* music is a set of physical actions that create pleasant, rhythmic sounds. Computers have it easy, since they can just generate a waveform representation of audio. But they don't understand the physical expression, which is one of the most important components. Can we teach a computational model how to express such things?

4 Deep Reinforcement Learning of Physically Simulated Percussive Instruments

4.1 Goals

The goal of this project is to train a neural network to generate rhythmically useful actions in a simulated environment, in a way analogous to how people perform music. Furthermore, such a system will ideally generate these rhythmic responses given a musical input, meaning the network could work out a rhythmic solution that is complementary to the input audio stimuli.

The virtual "agent" will be a simulated drummer, with physical constraints modeled after the way our arms can move in 3D space while holding a drumstick, which is able to create percussive impacts at precise rhythmic timings. This brings up a lot of questions that won't be covered in this proposal. This will be a more basic overview of the steps I plan to take to implement such a system.

4.2 Training

The virtual drummer agent and its environment will be created in the Unity Engine. Unity is a game engine that is also an ideal solution for conducting machine learning research, and has been used extensively by researchers through the ML Agents toolkit. Training of physically simulated agents is nothing new. Check out this video of simulated agents learning to walk on two feet, or this one of agents learning to play hide and seek.

The agent will be trained using reinforcement learning, a machine learning technique in which a system of sparse rewards are applied at appropriate times within a simulation, in order to modify an agent's policy (or operational behavior) with the goal of improving the agent's ability to accomplish a task. In this case, the aim is to improve the agent's ability to move their virtual hands, attached to drum sticks, in 3D space, so that they may hit a drum's surface in a rhythmic pattern. Furthermore, they should make these actions in reaction to an audio stimuli, generating a groove that attempts to match the musical structure. The aim is to train an AI agent that has the ability to make rhythmic inference from a real musician's playing, and convert it into simulated physical actions which imitates a real drummer's response.

At its worst, it could sound like a two-year-old smashing pots and pans together. In its ideal form, it could actually hit pads on the simulated instrument in a rhythmic fashion that corresponds to external musical stimuli. To be clear, this is an extremely far out prospect that will likely be closer to the two-year-old than to John Bonham.

4.3 Virtual Reality Component

To train the neural network, there needs to be proper training data for the agent to correct against. This is why I've decided to use virtual reality to provide a simulated environment for a drummer to generate training data with. Within this environment, there will be a simulated drum kit, in which the player (me) will play along to music. The pads of the virtual drums will be placed relative to the real-world space of an actual electronic drum kit (Roland TD-11,) in order to record precise timing of impacts using MIDI data. Although it's not ideal for playing, I will be holding VR controllers in my hands along with the sticks, so that I can capture exactly where my hands are in world space.

4.4 Performance Data Capture

During a drum performance, data will be sampled at a given frequency and stored for training the model. This data will include:

- Velocity vectors [x,y,z] of the hand controllers gripping the sticks
- Position vectors [x,y,z] of the hand controllers
- MIDI velocity and timing from the TD-11 pads
- Waveform audio of musical stimuli presented to the player for drumming
- Detected virtual collisions in the simulation between the simulated drumstick and virtual drum head

4.5 Agent Simulation

With this performance data, I will then construct a simulation for the virtual agent to move its hands. It has the option at any given moment to increase the velocity of this hand towards any three dimensional vector. The virtual agent's main duty is to make a decision about how to apply force, given a 3D vector input to the neural network of where its hands currently are, and their velocity. If I make it this far, I will also condition this input on the audio stimuli, so that it may imitate actions based on an inferred groove. This may require an entirely separate neural network, which will analyze the audio signal for features, and return a latent vector for conditioning the generative network.

5 Conclusion

Because this project is ambitious, I do not know exactly how much of this proposal I will get to. The easy parts are already in the works, which includes creating the virtual environment and drum kit, as well as the data sampling programs. Other parts have already taken some form in previous projects of mine, and this is somewhat a culmination of other things I've done in the past. I have explored reinforcement learning before with Unity's ML Agents, but not nearly to this extent. I have also done work with generative neural networks including GANs and VAEs, as well as work with audio classification networks. This project also touches on my experience with VR application development, and XR virtual instrument design, which I hope to continue getting better at, as well as publish work for. Signal processing and timbral analysis also may come in to play, but these are part of the stretch goals of conditioning the drummer agent. The nature of this project stretches my knowledge, and I hope to learn a lot.

Although I aim to teach the computer how to play drums, I too am still learning. This project combines my newfound passion for playing drums with machine learning, which is why I chose to approach this. In fact, I think the fact that I am actively learning is helpful, because I'm noticing the immediate difficulties to overcome and errors to correct in my playing.

Eventually, work like this could be useful for many things, from providing a virtual band member to your drummer-less band, to gaining a greater understanding of how the human brain learns to play an instrument. It could also have applications for robotics - robots have a huge potential to be musical and rhythmic. If you don't believe they can, watch this.