JingleBot:
Raw Audio Generation with WaveNet

Hanna Berger, Anna Boyd, Izzy Hargrave

UNIVERSITY of ROCHESTER

AUDIO & MUSIC ENGINEERING
AT THE UNIVERSITY OF ROCHESTER

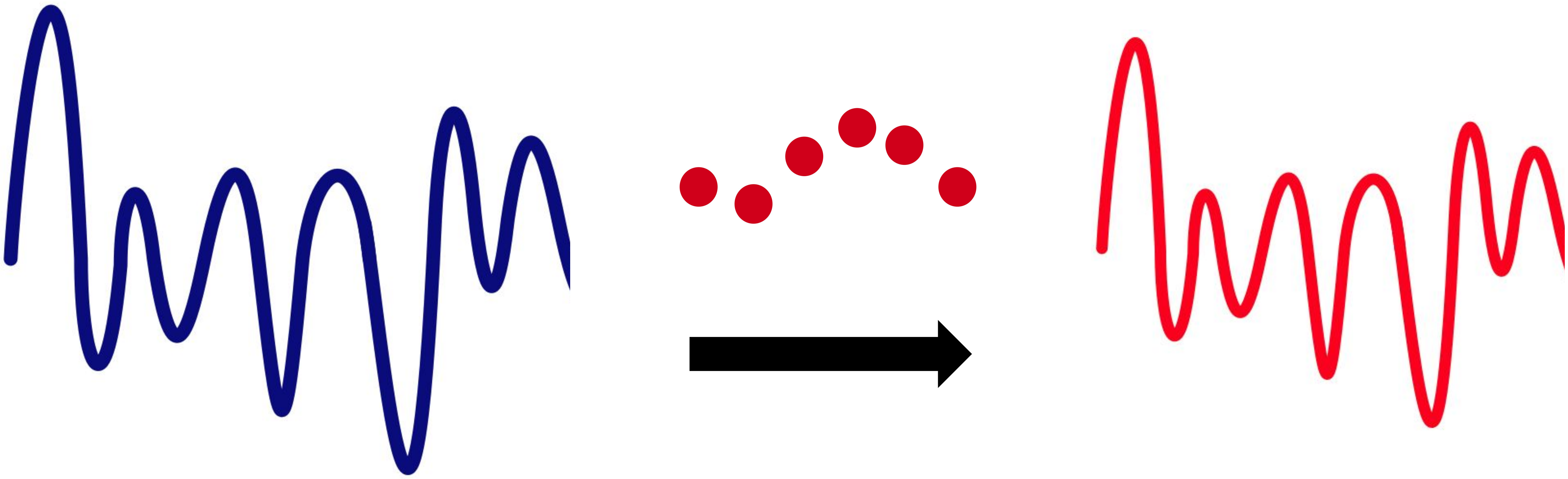
Conditioned Short-Form Raw Audio Generation

- Main goal is to build a model to generate jingles based off a predetermined set of moods that the user can select from
- Steps include:
  - Create WaveNet model based off of previous work
  - Train model on a large music dataset as a form of pre-training
  - Continue training model using a database of jingles
  - Generate jingles from the model

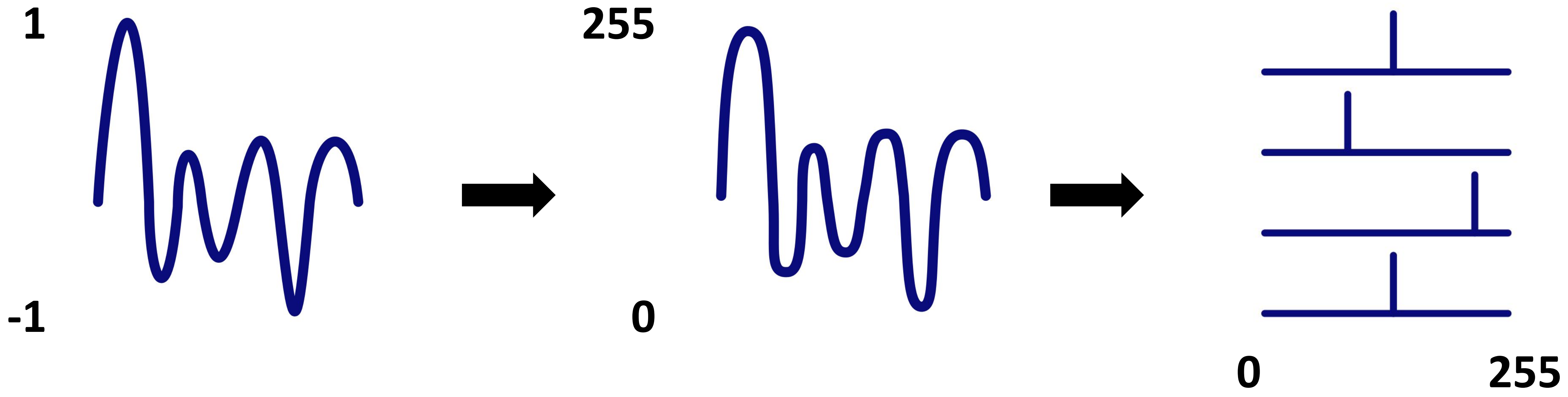
Background: WaveNet

- 2016 generative model for raw audio from Google Deepmind
- Useful for many audio tasks, including text-to-speech, speech-to-text, and music generation
- Can have a “memory” of a few seconds

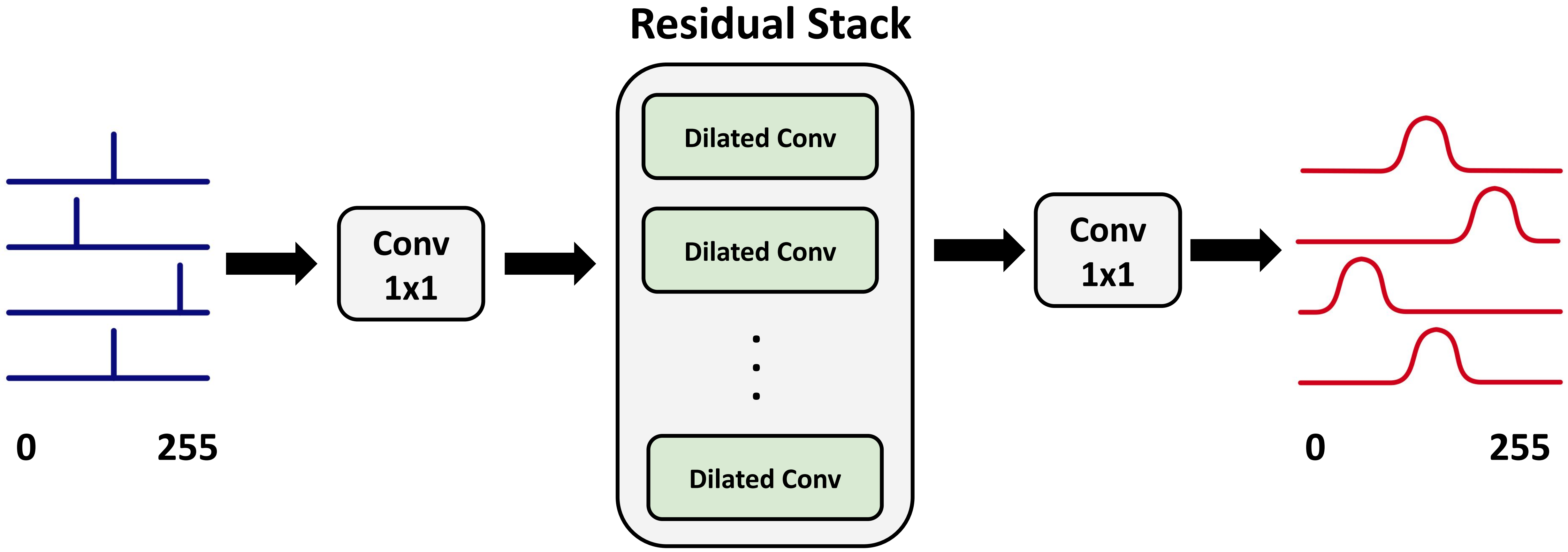
WaveNet: How does it work?



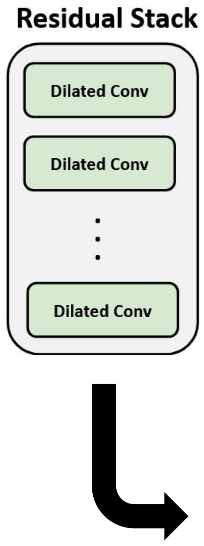
Data Pre-Processing

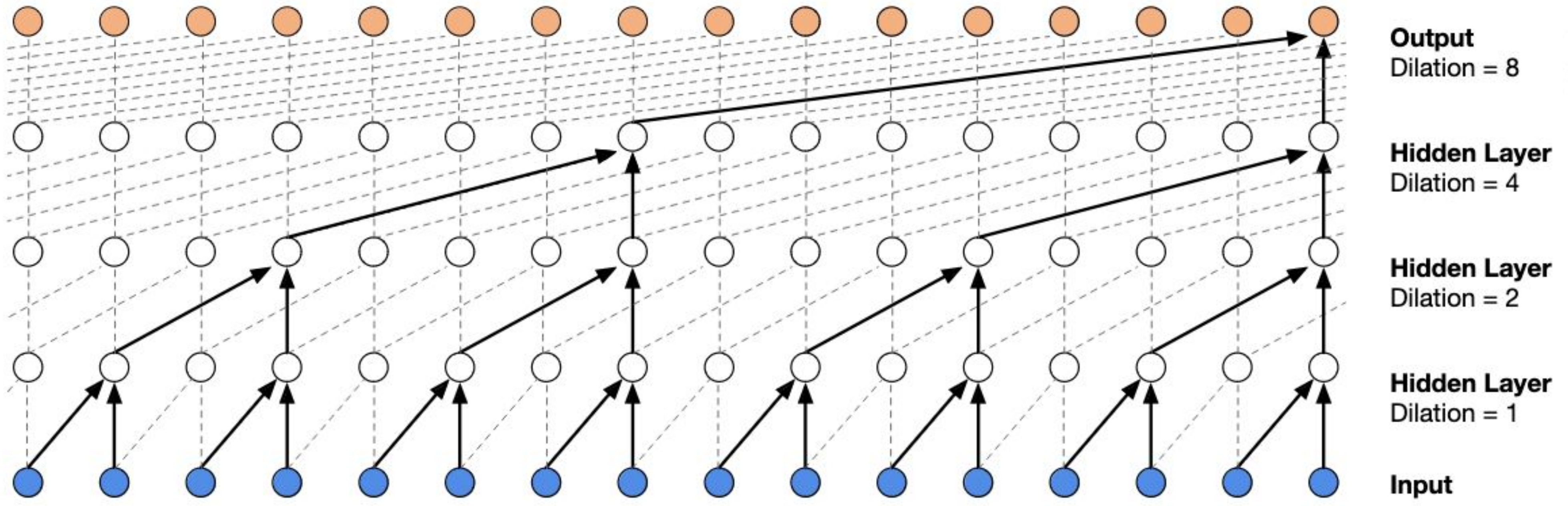


Model Structure



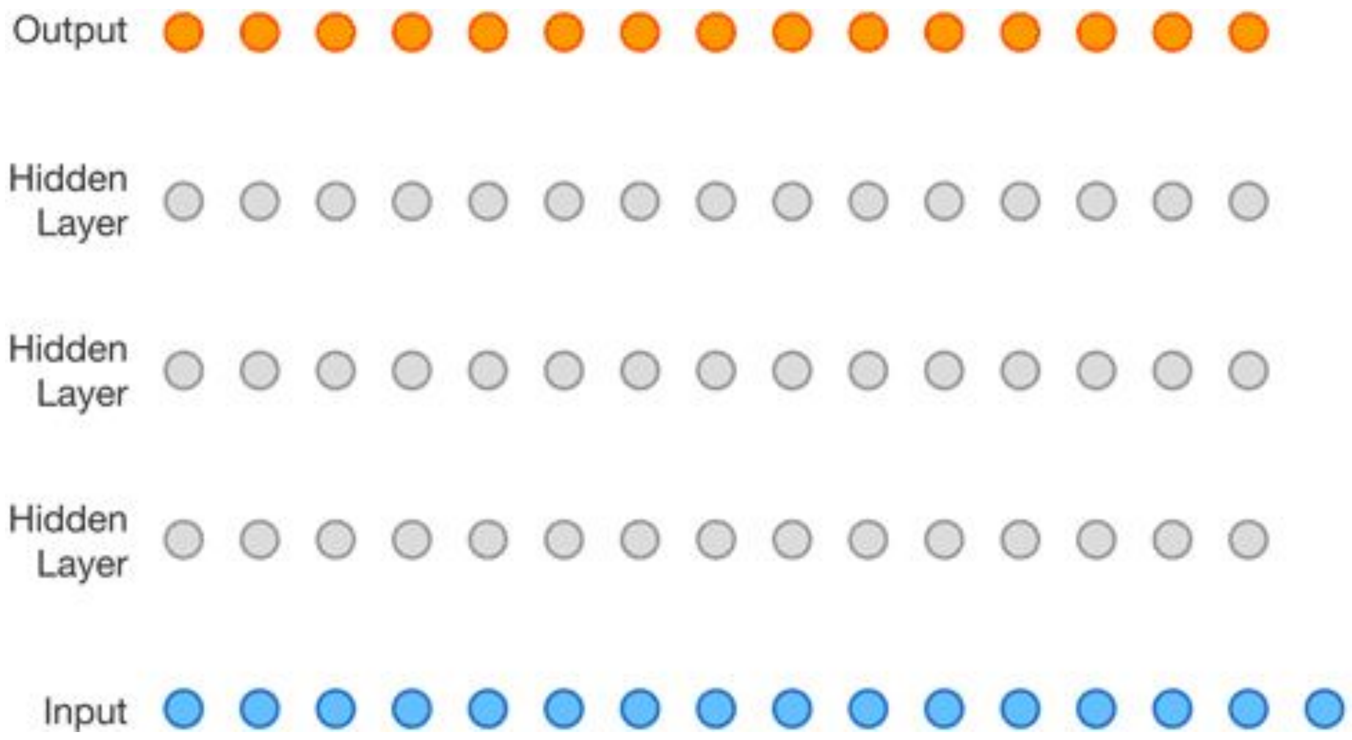
Model Structure





van den Oord et. al. (2016))

Generation



van den Oord et. al. (2016)

Datasets: FMA Dataset

- Used for preliminary training of the model to generate music
  - Not nearly enough jingles to create a dataset of that caliber
- Royalty-free music samples from Free Music Archive
- Multiple sizes (small, medium, large, etc.)
- We utilized the FMA Small dataset
  - 8000 audio samples, each 30-seconds long

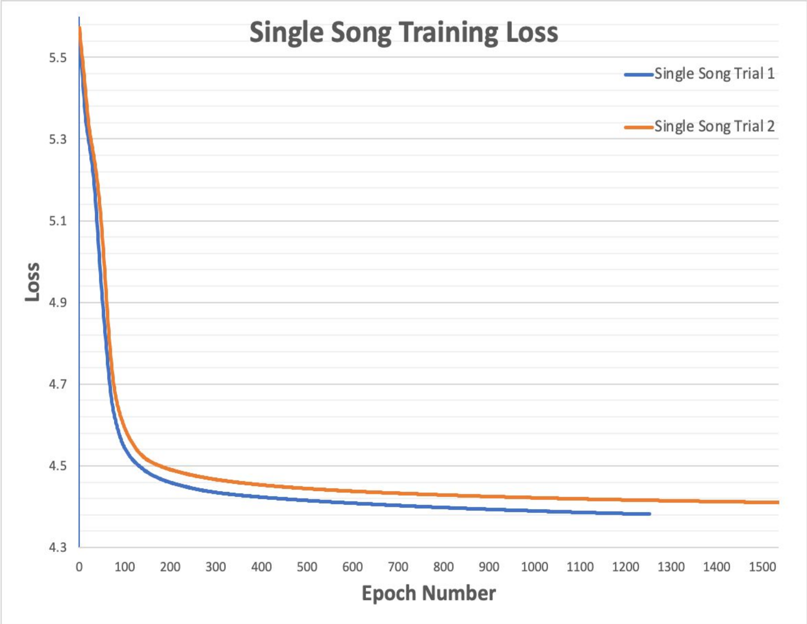
Datasets: Jingle Dataset

- Created our own jingle dataset since none preexisting
- Collected 420+ from various royalty free websites [8,9,10,11,12]
- Manually labeled and categorized each jingle into 5 different moods
- Moods were chosen based off of what was most commonly heard
  - 58 Melancholy (mel)
  - 53 Mysterious (mys)
  - 75 Playful (plf)
  - 90 Relaxing (rlx)
  - 144 Upbeat (upb)

Experimentation: One Song Experiment

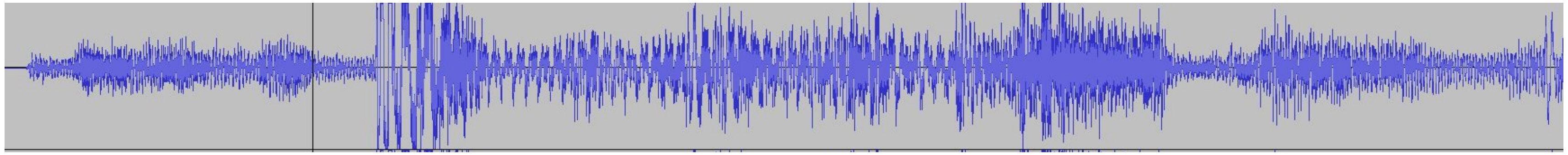
- Proof of concept experiment
- If only trained on a single song the model should learn to “predict” that song and recreate it perfectly
- Trial and error to find the right model hyperparameters
- Limited memory for computation
    - Multiple Stacks vs Larger Receptive Field

Hyperparameters	Trial #1	Trial #2
Batch Size:	1	1
Stack Size:	4	1
Layer Size:	8	24
Learning Rate:	0.003	0.003
Total Epochs Trained:	1252	1536

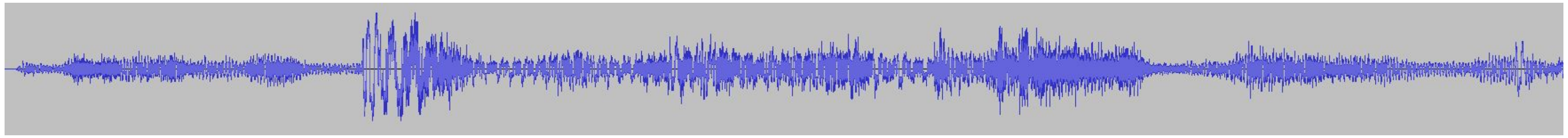


One Song Experiment: Results (Entire Song)

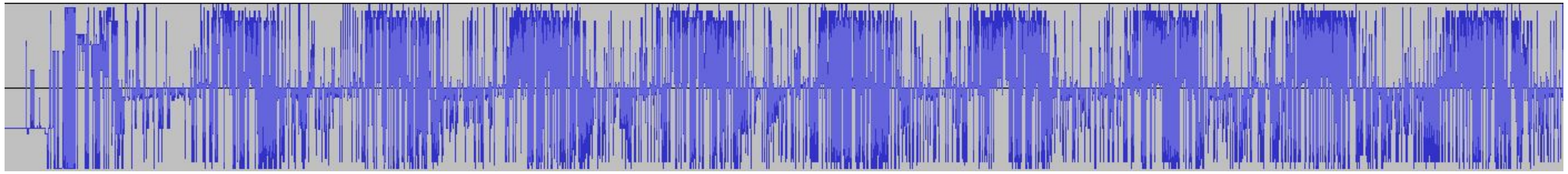
- Original Audio:



- Forward Pass through Model (predicted by the model, just one sample at a time)

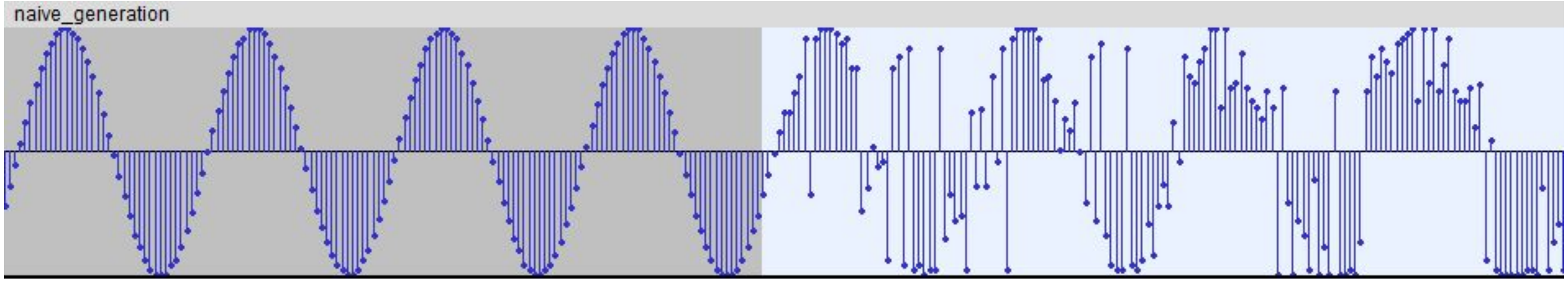


- Output of Generation Algorithm (predicted on previously predicted samples)



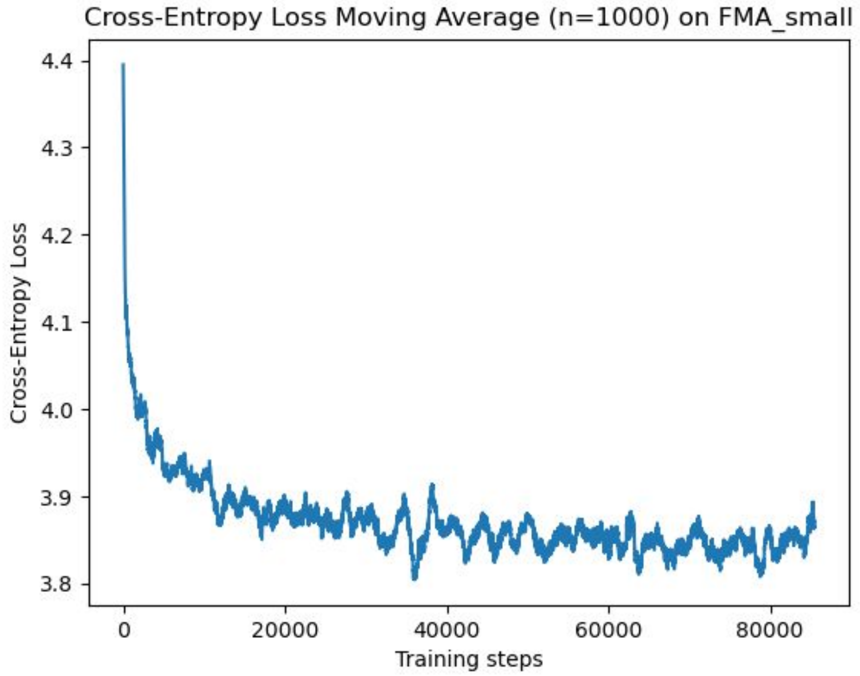
What if we simplified it?

- Trained the model on a sine wave to scale it down and look into generation algorithm further
  - Gave it 512 samples and had it predict the rest
  - Proof that it can generate something harmonic, we just were not able to train a complex enough model or train it for long enough



Experimentation: Model Training on FMA_small

- Training environment:
  - 2x Nvidia 1080ti GPUs (20gb GPU RAM total)
  - Learning rate=0.001, Adam optimizer
  - 90,000 training steps (and counting…)
  - Stack_size=4, Layer_size=8
  - Maximum receptive field=256 samples (~2ms)
  - No help from Google on any of this
  - Couldn’t fit the whole model on two cards!



Computation (and Lack Thereof)

- The biggest bottleneck, by far, was computation
- Simply did not have access to the compute required to generate realistic musical fragments
- Forced us to simplify the scope of the project to modeling one song – or even just a sine wave
- We couldn't fit a WaveNet model large enough to generate realistic music fragments on 20gb of GPU RAM

```
OutOfMemoryError: CUDA out of memory. Tried to allocate 2.00 MiB (GPU 0; 11.00 GiB total capacity; 9.09 GiB already allocated; 0 bytes free; 10.23 GiB reserved in total by PyTorch)
```

The story of this project

What’s next? Future Work

- Experimenting with the generation model for better results
  - Training on sine wave converged with 0.00012 cross-entropy loss. Why isn’t our generated audio better? Something’s fishy…
- Adding a third GPU to our training computer to increase receptive field and stack size
- Training on FMA-Full (1TB of audio)
- Adjusting hyperparameters
  - Learning schedule
  - Length of input audio
- Fine-tuning the model on our custom dataset
- Global conditioning of the model
- AWS/Azure for more compute? Different model?

References:

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," arXiv preprint arXiv:1609.03499, 2016.

[2] K. Chen, W. Zhang, S. Dubnov, G. Xia and W. Li, "The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation," 2019 International Workshop on Multilayer Music Representation and Processing (MMRP), Milan, Italy, 2019, pp. 77-84, doi: 10.1109/MMRP.2019.00022.

[3] R. Manzelli, et al., "An end to end model for automatic music generation: Combining deep raw and symbolic audio networks," in Proceedings of the Musical Metacreation Workshop at the 9th International Conference on Computational Creativity, Salamanca, Spain, 2018.

[4] S. Luo, "Bach Genre Music Generation with WaveNet—A Steerable CNN-based Method with Different Temperature Parameters," in Proceedings of the 4th International Conference on Intelligent Science and Technology, 2022, pp. 40-46.

[5] "BandNet: A Neural Network-based, Multi-Instrument Beatles-Style MIDI Music Composition Machine," arXiv preprint arXiv:1812.07126, 2018

[6] T. Le Paine *et al.*, “Fast Wavenet Generation Algorithm,” Nov. 2016.

[7] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: a Dataset for Music Analysis,” Sep. 2017. Available: https://arxiv.org/pdf/1612.01840.pdf

[8] Hibou Music Library, “Jingle and Jingles,” www.hibou-music.com. https://www.hibou-music.com/jingle-jingles.html (accessed Nov. 09, 2023)

[9] Pixabay, “Royalty Free Music Downloads,” Pixabay. https://pixabay.com/music/ (accessed Nov. 10, 2023).

[10] Freepik, “Videvo: Royalty Free Music Download Background Stock Audio,” Royalty Free Music Download Background Stock Audio, 2023. https://www.videvo.net/royalty-free-music/ (accessed Nov. 10, 2023).

[11] H. Chahidi, “Royalty Free Music | Jingles,” Music Screen, 2023. https://www.musicscreen.org/royalty-free-jingle-music.php (accessed Nov. 09, 2023).

[12] Tribe of Noise, “Free Music Archive,” Free Music Archive, 2023. https://freemusicarchive.org (accessed Nov. 29, 2023).