WaveNet - Generative Music Production

Simon Scapan

DHBW - Mannheim

June 19, 2021

WaveNet in General

- Developed by DeepMind in London
- Generate raw speech signals with subjective naturalness never before reported in the field of Text-to-Speech (TTS) (Oord et al., 2016)
- Performance improvement by over 50% (van den Oord and Dieleman, 2016)
- Advantage : one model for different purposes

WaveNet in General

- Architecture based on dilated causal convolutions
- WaveNets provide a generic and flexible framework for many applications relying on audio generation :
  - Text-to-Speech
  - Music generation
  - Speech enhancement
  - Voice conversion
  - Source separation

Source : (Oord et al., 2016)

Model Explanation - Theoreticaly

- Generative model operating on raw audio waveform
- Joint probability of a waveform is factorised as product of conditional probabilities
- Each audio sample is therefore conditioned on the samples at all previous timesteps
- Conditional probability distribution is modelled by stack of convolutional layers
- No pooling layers in network
- Output of the model has same time dimensionality as input

Source : (Oord et al., 2016)

Model Explanation - Visual

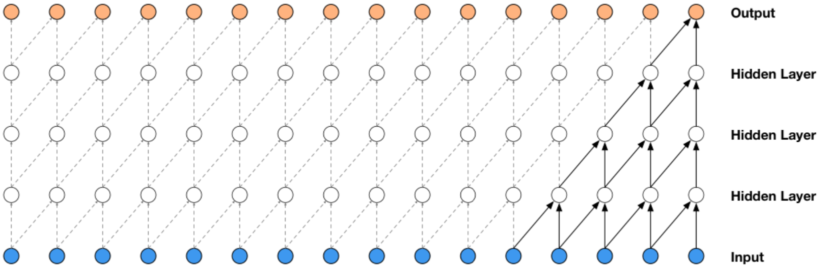

Figure : Visualization of a stack of causal convolutional layers
Source : (Oord et al., 2016)

Model Explanation - Causal Convolutions

- Main ingredient of WaveNet are causal convolutions
- Based on that, the model cannot violate the ordering in which the data is modeled
- Predictions emitted by model at timestep t cannot depend on any of the future timesteps
- At training, conditional predictions for all timesteps can be made in parallel (all timesteps of ground truth x are known)

Source : (Oord et al., 2016)

Model Explanation - Causal Convolutions

- At generation of outputs with model, predictions are sequential : after each sample is predicted, it is fed back into network to predict next sample
- Models with causal convolutions do not have recurrent connections, they are typically faster to train than RNNs
- Problem of causal convolutions is : they require many layers, or large filters to increase the receptive field

Source : (Oord et al., 2016)

Model Explanation - Dilated Convolutions

- A dilated convolution is a convolution where the filter is applied over an area larger than its length by skipping input values with a certain step
- It is equivalent to a convolution with a larger filter derived from the original filter by dilating it with zeros, but significantly more efficient
- Similar to pooling or strided convolutions, but here the output has the same size as the input
- Stacked dilated convolutions enable networks to have very large receptive fields with just a few layers

Source : (Oord et al., 2016)

Model Explanation - Visual

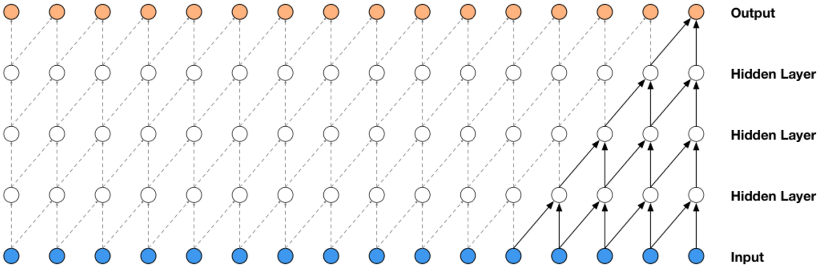

Figure : Visualization of a stack of dilated causal convolutional layers
Source: (Oord et al., 2016)

Music Production with WaveNet

- "WaveNets can be used to model any audio signal"
- Unlike the TTS experiments networks were not conditioned on an input sequence telling it what to play (such as a musical score)
- Instead : simply let it generate whatever it wanted to
- Fact that directly generating timestep per timestep with deep neural networks works at all for 16kHz audio is really surprising

Source : (van den Oord and Dieleman, 2016)

Implementation Example

Let's have a look at the Results Chen written down in following article :

Generating Ambient Music from WaveNet

Rachel Chen Dec 13, 2017 · 20 min read

Stefan Bordovsky, Rachel Chen, Kyle Grier, Danny Sutanto

Source : Medium (Chen, 2017)

Implementation Example - Prerequisites

- Model trained on Tensorflow implementation of WaveNet
- 150 000 steps at a default of 0.001 learning rate
- Amazon Web Services' p2.xLarge EC2 instance to train the WaveNet model with a GPU
- 118 500 steps trained in approximately 3.5 days (then AWS costs get to high) :
  - with each step taking roughly 2.5 seconds
  - their laptops took approximately 1 minute just to train one step

Source : Medium (Chen, 2017)

Implementation Example - Some Results

- Based on Happy Music from YouTube the model results are :
  - 9950 steps
  - 10800 steps
  - 14450 steps
  - 25650 steps

Source : Medium (Chen, 2017)

Implementation Example - Challenges

- Very much iterations are needed in order to achieve approximately good results
- Model requires at least 20 000 steps to generate something somewhat recognizable
- And around 80 000 steps for something somewhat coherent
- Learning on local machines takes very long for only semi good results

Source : Medium (Chen, 2017)

Implementation Example - Chances

- Scientist at DeepMind implemented a model playing Piano :
  - WaveNet Piano example (van den Oord and Dieleman, 2016)
- Advantage is, that they input exactly one instrument
- WaveNet achieves good results on simple inputs
- Complex inputs require a lot of learning steps

Wrap up

- WaveNet is basically a good model for generating music
- Good results can be achieved quickly with individual instruments
- If whole songs are used as input, the model has to make significantly more learning steps
- This extensive learning is very computationally, time-consuming and costly

Bibliography

Chen, R. (2017). Generating ambient noise from wavenet.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.

van den Oord, A. and Dieleman, S. (2016). Wavenet: A generative model for raw audio.