WaveNet: A Generative Model for Raw Audio

Aäron van den Oord et al., 2016

Rui Yoshinaga

July 21, 2021

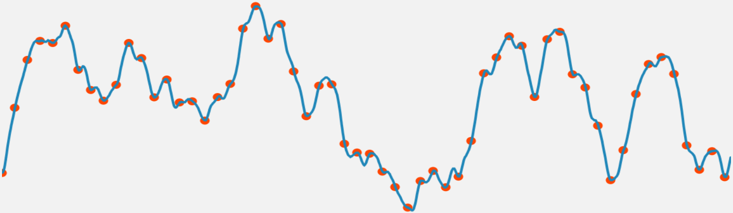
Overview

WaveNet: A Generative Model for Raw Audio

- Topic: Audio Generation
- Contributions:
  - Text-to-Speech (TTS) : Generating natural speech
  - Develop new architectures
  - Various audio generation by conditioning
  - Application to music generation, etc.

- Generated samples:
  - `https://deepmind.com/blog/article/wavenet-generative-model-raw-audio`

Composition



waveform $\boldsymbol{x} = \{x_1, \ldots, x_T\} \in [-1, 1]^T$

Modeling of $p(\boldsymbol{x})$

- The joint probability of waveform $\boldsymbol{x}$ is factorised:

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \quad (1)$$

- Sample x_t is conditioned on the previous samples $x_1, \ldots, x_{t-1}$

Composition

Architecture

- NN that takes inputs $x_{t-L}, \ldots, x_{t-1}$ and outputs x_t
- Modeling by CNN (Dilated Causal Convolution)
  - c.f. image generation [1, 2], w/o pooling layer

Training and Generation

- Training: maximize the log-likelihood
- Generation: autoregression

Softmax Distribution

Input/Output

- Input: previous samples $x_{t-L}, \dots, x_{t-1}$
- Output: the probability distribution of next sample x_t

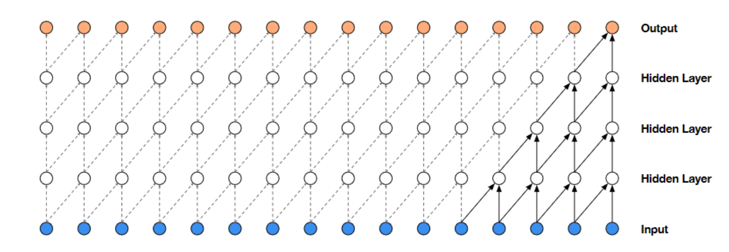
μ-law companding transofrmation

- raw audio: quantized at 16bits, amplitude: 2^{16} levels
- Apply μ-law companding transformation $\rightarrow$ 256 levels

$$f(x_t) = \mathrm{sign}(x_t) \frac{\ln(1 + \mu |x_t|)}{\ln(1 + \mu)}$$

where $-1 < x_t < 1$ and $\mu = 255$

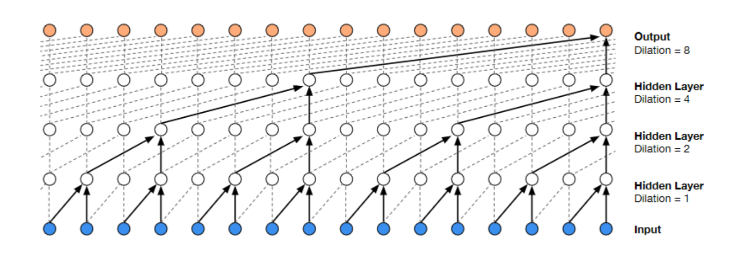
Causal Convolution



- Basic structure of WaveNet
- Independent future samples $x_{t+1}, x_{t+2}, \dots$
- No recurrent connections, training in parallel

- Increase in the size of receptive field
  - In the above figure, receptive field $= 5$
  - Increase layers or filter size

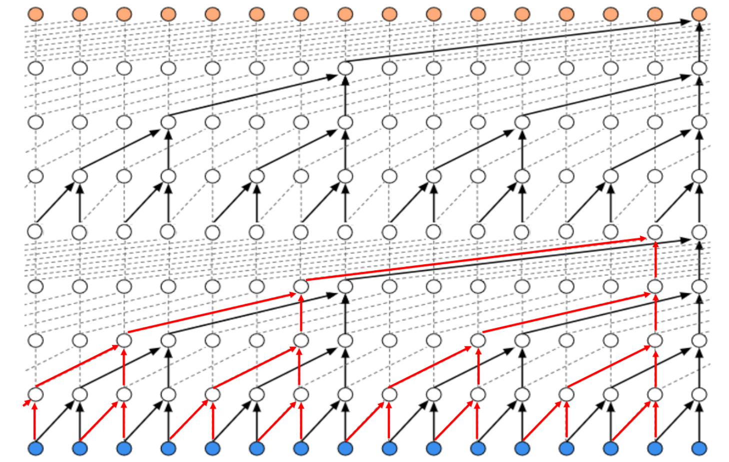
Dilated Convolution



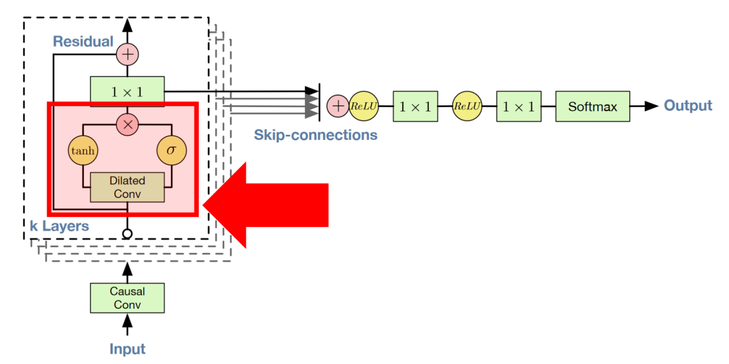
receptive field = 16

- Skip inputs at fixed intervals
- In WaveNet, dilation is doubled $(1, 2, 4, \ldots, 512, 1, 2, \ldots)$
  - Exponential increase in dilation $\rightarrow$ larger receptive field
  - Stacking blocks further increases the size of receptive field

Stacked Dilated Convolution



Gated Activation Unit

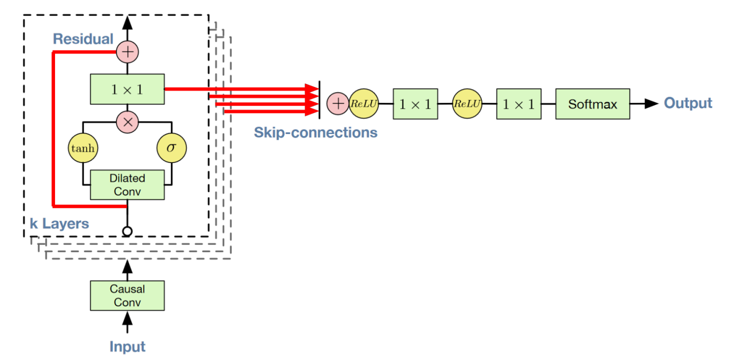


- Gated activation unit [2]

$$z = \tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x) \qquad (2)$$

  - $W_{*,k}$: learnable convolution filter in k-th layer
  - $*$: convolution operator
- Worked significantly better than ReLU

Residual and Skip Connection



- Residual [3] and Skip Connections
  - Speed up convergence
  - Enable training of much deeper model

Conditional WaveNets

- Modeling conditional distribution $p(\boldsymbol{x}|\boldsymbol{h})$ by additional input $\boldsymbol{h}$
- Eq.(1) now becomes

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \boldsymbol{h}) \tag{3}$$

- Generate audio with the required characteristics by conditioning
  - e.g.1 speaker ID (who speaks)
  - e.g.2 text for TTS (what to speaks)
- Conditioning in 2 different ways:
  - {Global, Local} conditioning

Global Conditioning

- Conditioning on single $\boldsymbol{h}$ that influences output across all timesteps
  e.g. TTS: speaker embedding (who speaks)

- The activation function from Eq.(2) now becomes:

$$\boldsymbol{z} = \tanh(W_{f,k} * \boldsymbol{x} + V_{f,k}^{T}\boldsymbol{h}) \odot \sigma(W_{g,k} * \boldsymbol{x} + V_{g,k}^{T}\boldsymbol{h})$$

  - $W_{*,k}$: learnable convolution filter in k-th layer
  - $*$: convolution operator
  - $V_{*,k}$: learnable linear projection in k-th layer
  - $V_{*,k}^{T}\boldsymbol{h}$: broadcast over the time dimention

Local Conditioning

- Condition on the second time series $\boldsymbol{h} = \{h_t\}$
  - $\boldsymbol{h}$ has a lower sampling frequency than audio $\boldsymbol{x}$

  e.g. TTS: linguistic features

- Map $\boldsymbol{h}$ to a new time series $\boldsymbol{y} = f(\boldsymbol{h})$
  - f : transposed convolutional network (learned upsampling)
  - (also possible to use $V_{f,k} * \boldsymbol{h}$ repeatedly across time)

- The activation function from Eq.(2) now becomes:

$$\boldsymbol{z} = \tanh(W_{f,k} * \boldsymbol{x} + V_{f,k} * \boldsymbol{y}) \odot \sigma(W_{g,k} * \boldsymbol{x} + V_{g,k} * \boldsymbol{y})$$

  - $W_{*,k}, V_{*,k}$: learnable convolution filter in k-th layer
  - $*$: convolution operator

Experiments (Music)

Music Datasets

1. MagnaTagATune dataset
   - each 29-second clip is annotated with tags
2. YouTube piano dataset

Receptive field

- several seconds: long-range consistency $\times$
- enlarging the receptive field $\rightarrow$ musical samples

Generating results

- Harmonic and aesthetically pleasing even no conditioning
- Conditioning on tags
  - Train with one-hot vector corresponds to the tags
  - Control generation by conditioning on the one-hot vector
- `https://www.deepmind.com/blog/wavenet-generative-model-raw-audio/`

Experiments

Multi-Speaker Speech Generation

- Conditioned on the speaker: feed the speaker ID as one-hot vector
- Generate non-existent but human language-like words realistically
- Mimicked the breathing and mouth movements of the speakers

Text-To-Speech

- Conditioned on the linguistic features and the value of $\log f_0$
- Achieved 5-scale MOSs in naturalness above 4.0
- Significantly better than the baseline

Speech Recognition

- Partially change architecture and loss
- Obtained the best score as model trained directly on raw audio

Conclusion

- WaveNet: A generative model for raw audio
  - Directly generate audio as waveform (raw audio domain)
- Dilated Causal Convolutions
  - Enables large receptive field with exponentially increasing dilation
- Can be applied in various applications about audio signals
  - TTS, Music generation, Speech recognition

References

[1] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu.
Pixel Recurrent Neural Networks, 2016.

[2] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu.
Conditional Image Generation with PixelCNN Decoders, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep Residual Learning for Image Recognition, 2015.