WaveNet :
A Generative Model for Raw Audio

한양대학교 산업공학과
IDSL 석사과정 고경준

HANYANG UNIVERSITY

Contents

1. PixelCNN
2. Introduction
3. WaveNet
4. Experiments

2. Introduction

Introduction

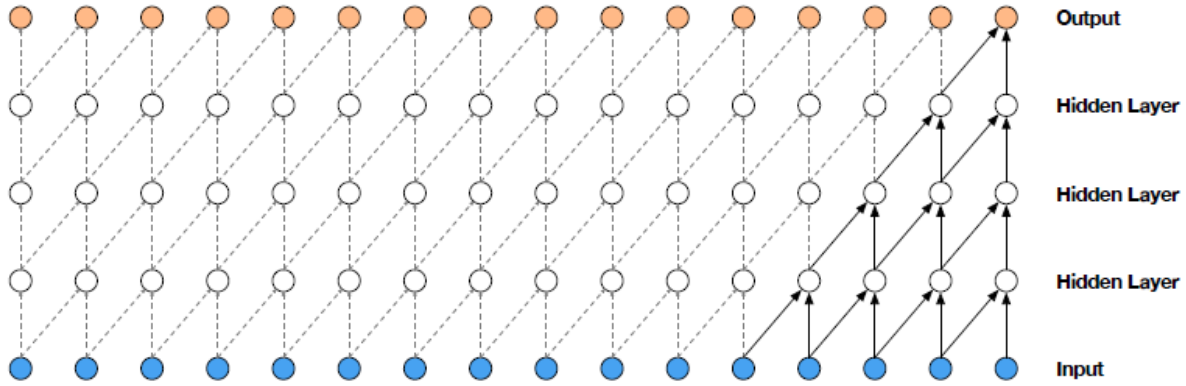
- Image, text 등 복잡한 distributions를 모델링하는 생성모형의 사례를 참고하여 raw audio waveforms를 생성해보려는 Idea에서 시작
- PixelCNN architecture를 base로 dilated causal convolution 모형을 제안

contribution

- Raw speech signal을 생성하는 모델을 text-to-speech(TTS) 분야에 처음 적용
- Long-range temporal dependency를 반영하기 위해 convolution에 dilation을 적용
- Speaker identity를 single model에 반영할 수 있는 방안 제시
- 같은 architecture로 audio generation에 flexible한 application이 가능

3. WaveNet

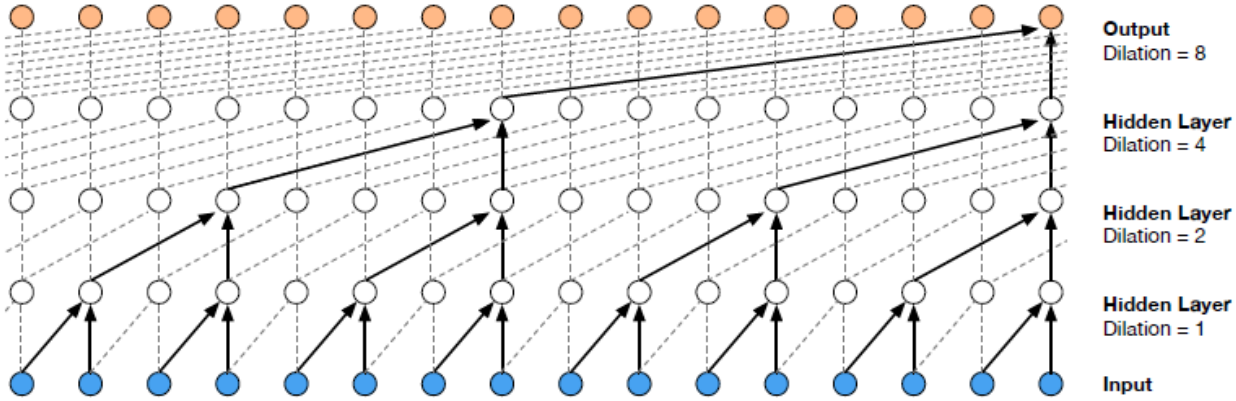
3.1 Dilated causal convolutions



- Main ingredient of the model은 causal convolution
- Image에서는 causal convolution은 mask tensor를 만들고 mask와 convolution kernel을 elementwise multiplication하는 masked convolution과 동일
- 이 과정은 audio data(1-D data)에서는 output of convolution의 shifting으로 적용 가능

3. WaveNet

3.1 Dilated causal convolutions



- Dilation을 추가하여 receptive field를 exponentially increase
  (the receptive field = #layers + filter length -1)
- Recurrent connection이 없으므로 training 시 layer 내의 계산이 parallel하게 진행
- Dilation이 1,2,4,…,512인 architecture를 하나의 block이라고 할 때, 여러 block을 쌓은 것이 WaveNet의 최종 architecture가 되고, block별로 out을 뽑아 최종 output을 산출

3. WaveNet

3.2 Softmax distribution

- Dilation을 추가하여 receptive field를 exponentially increase
  (the receptive field = #layers + filter length -1)
- Recurrent connection이 없으므로 training 시 layer 내의 계산이 parallel하게 진행
- Dilation이 1,2,4,…,512인 architecture를 하나의 block이라고 할 때, 여러 block을 쌓은 것이 WaveNet의 최종 architecture가 되고, block별로 out을 뽑아 최종 output을 산출

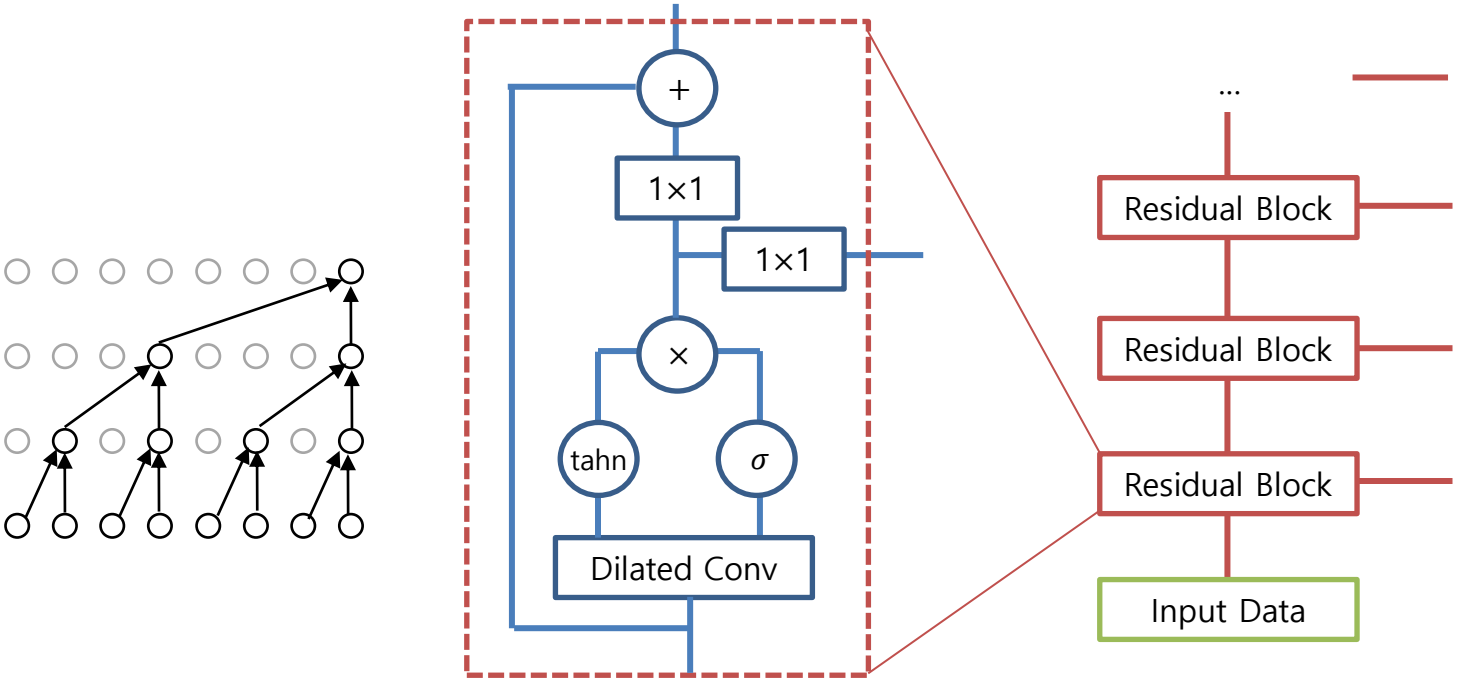
3. WaveNet

3.3 Gated activation units

- Convolution에서의 계산은 gated activation unit을 이용

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

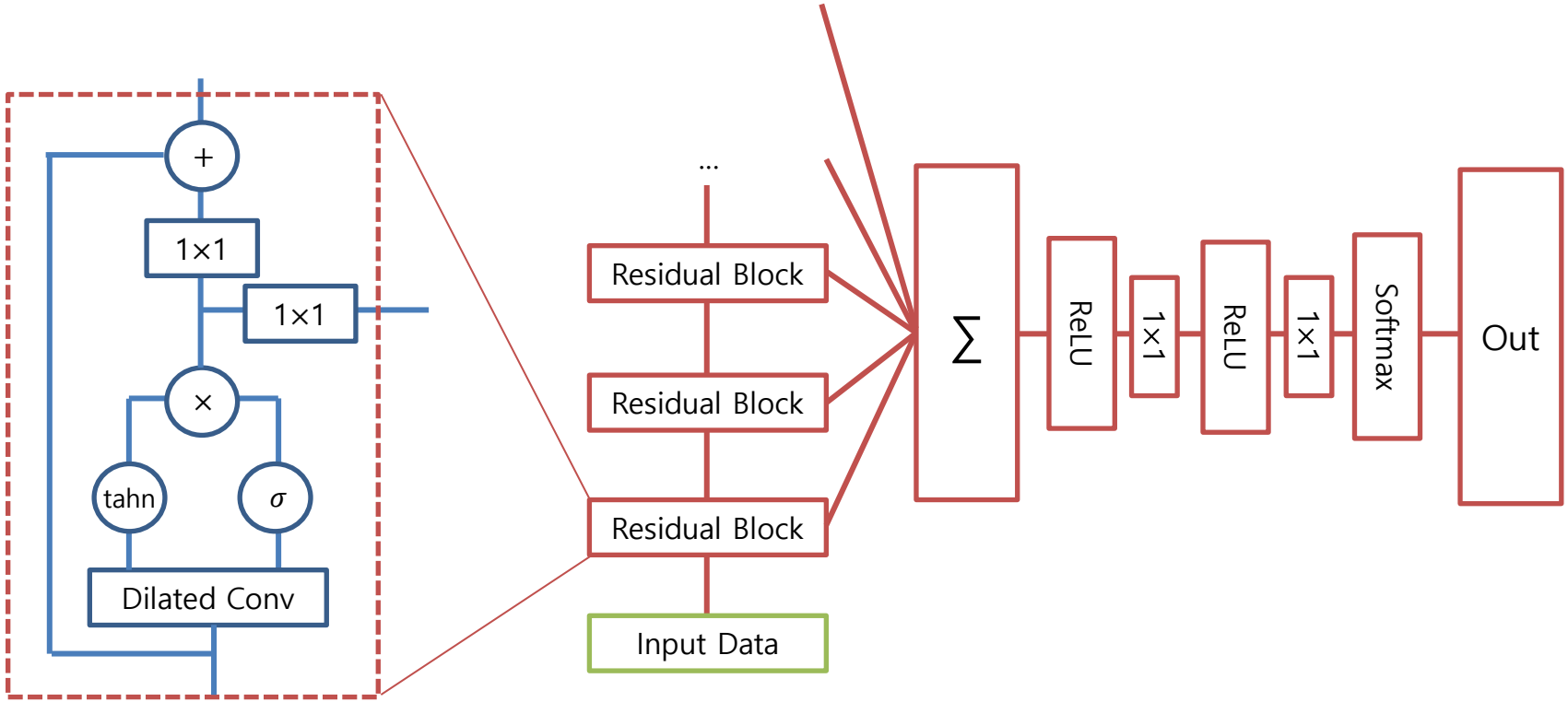
3.4 Residual and skip connection



- Block 내에 residual connection이 존재하고, 1X1 Conv를 거쳐 다음 layer, out으로 전달

3. WaveNet

3.4 Residual and skip connection

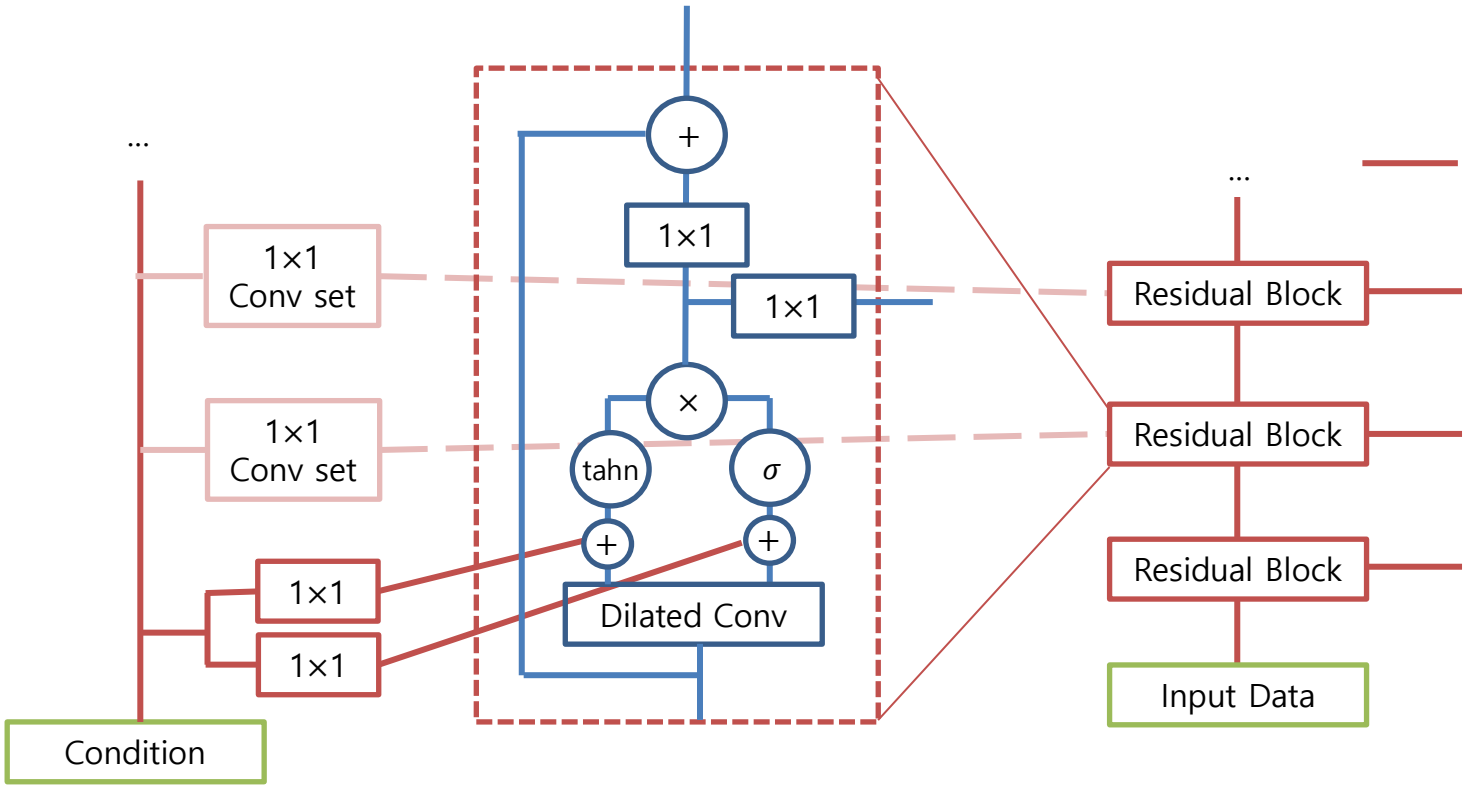


- Residual Block의 두 1X1 Conv는 각각 다음 Block과 out으로 전달
- Out으로 전달되는 1X1 Conv 값들을 합쳐 ReLU와 Conv를 거친 뒤 Softmax로 최종out 산출

3. WaveNet

3.5 Conditional WaveNets

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



- Additional input을 y라고 할 때, 위 식과 같이 계산되어서 Single machine에 적용 가능

4. Experiments

4.1 Multi-speaker speech generation

- English multi-speaker corpus (CSTR voice toolkit), conditioned on the speaker
- The experiment was conducted in two versions
  (one with added conditions and one without)
- Condition applied by feeding speaker ID int the form of one-hot vector
- It generate non-existent but human-like words (like language or image generation)
- Adding speaker ID vector resulted in better performance than solely trained model
- It also can mimic acoustics and the breathing etc.. with other characteristics

- There is no detail table or visualizations about this experiments in paper

4. Experiments

4.2 Text-to-Speech

- Google's North American English and Mandarin Chinese TTS system
- Using condition with logarithmic fundamental frequency value ($log\ F_0$)
- In the MOS(mean opinion score), subjects evaluate the output to five-point Likert scale score (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Exceleent)

	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

4. Experiments

4.3 Music

- MagnaTagATune datasets are used
- It's hard to evaluate quantitatively but good
- Model can generate much musical samples using larger receptive field
- Conditions like genre or instruments is work well in this model

4.4 Speech recognition

- This experiments used TIMIT dataset
- This model can also used to speech recognition
- It is much cheaper than RNN based model
  (maybe because there is no recurrent connection)

감사합니다.