Wavenet

WaveNet: A Generative Model for Raw Audio(Google DeepMind, 2016)

WaveNet

보코더(Vocoder)의 한 종류

- 음성의 등장 확률을 학습하는 확률론적 모델
- 과거(t-1) 시점까지 음성데이터와 멜 스펙트로그램을 조건으로 현재(t) 시점의특정 음성 등장 확률을 추출

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

WaveNet 입출력

- 입력: 멜스펙트로그램 / 출력: 음성

- 일반적인 음성 데이터는 각 샘플을 16(bit) 정수 값으로 저장
  → $-2^{15} \sim 2^{15}-1$ 사이의 수로 표현
  → 즉, 65,536개의 숫자가 나올 확률($P(-2^{15}|x_1, \dots, x_{t-1}) \sim P(2^{15}-1|x_1, \dots, x_{t-1})$)을 계산

- 따라서 이를 256개의 숫자로 변환(총 256개의 확률)
  → μ-law Companding Transformation한 값이 WaveNet에서 입력으로 사용

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

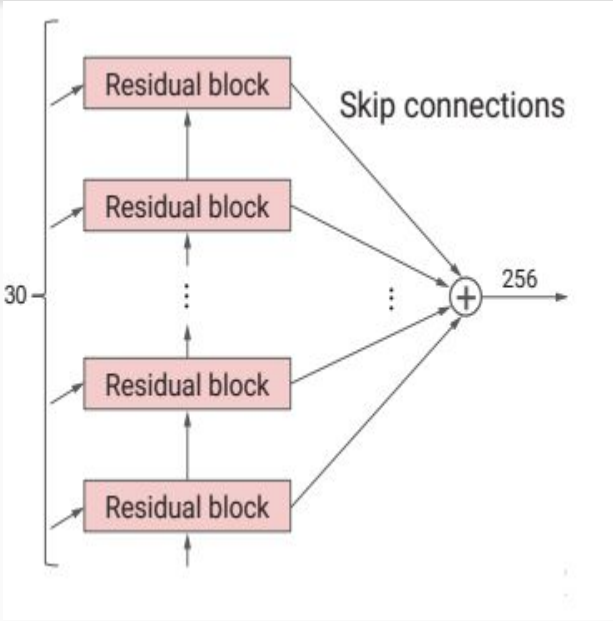
- 출력 값 역시 -128 ~ 127(256개) 범위의 정수로 최종적으로 이 정수를 이용해 음성 디지털 데이터로 변형

WaveNet 구조

30개의 Residual Block으로 구성

각 Residual Block의 output이
skip connection으로 합쳐져 최종 출력으로 활용

Residual Blcok
= Dilated Casual Convolution + Gated Activation Units



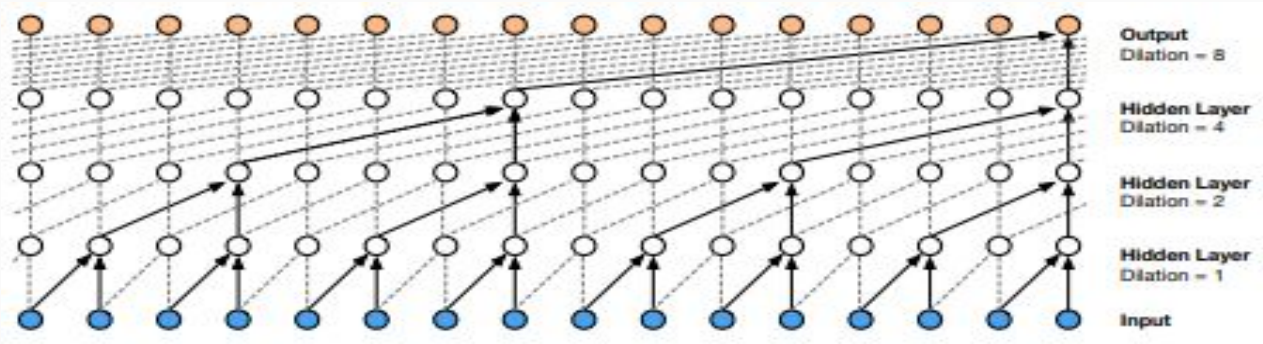
출처: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45882.pdf

WaveNet 구조

Dilated Casual Convolution: 과거 음성정보를 이용해 현재 시점의 정보를 생성

보통의 RNN/LSTM을 이용하게되면 비효율적(수용 범위가 넓으면 그 만큼 많은 연산)

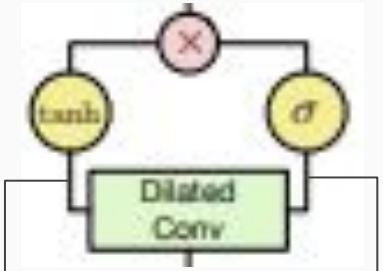
Casual Convolution: 시간 순서를 고려하여 과거 정보만을 접근하여 정보를 추출
Dilated Convolution: 추출간격을 조절하여 Layer를 적게 쌓아도 **더 넓은 수용범위**를 갖음



WaveNet 구조

Gated Activation Units: Dilated Convolution에서 생성된 정보를 다음 Layer에 얼마나 전달할지 결정

filter 경로: Dilated Conv에서 생성된 정보를 가공(tanh 부분)
gate 경로: filter 경로에서 가공된 정보를 다음 Layer에 얼마나 전달할지 결정(0~1) (시그모이드 부분)



Dilated Convolution된 값이 각각 convolution 연산을 거치고 각각 활성함

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

력 벡터 계산

filter index

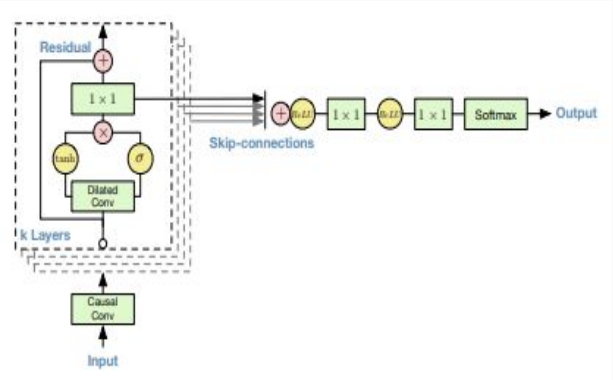
layer index

gate index

WaveNet 구조

Residual Block의 Output

→ 최종적으로 convolution과 gate를 통과하여 생성된 벡터는 Residual Connection으로 입력과 연결되어 최종 결과 생성



그레디언트 소멸 문제를 방지하여 많은 Layer를 쌓을 수 있는 구조

Conditional WaveNet 구조

Conditional WaveNet 특징

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

- Conditional Modeling P(x | h)을 통해 특징을 추가하고 특징에 맞는 음성 생성
- Condition으로 전역적 특징(화자), 지역적 특징(멜스펙트로그램) 부여 가능
- 전역적 특징: 출력 분포 전체에 영향을 미치는 요소($\mathbf{h}$)
- 지역적 특징: 특정한 시간에 영향을 미치는 요소(h_t)