WaveNet 논문 리뷰

윤태우

1. WaveNet 이전 모델의 한계

2016년 구글 딥마인드에서 오디오 생성 모델인 wavenet에 관한 논문을 발표했습니다.

기존에 많이 사용되던 concatenative TTS 방식은 녹음된 음성 데이터를 쪼개고, 합성하여 음성을 생성 했습니다.

이는 많은 음성 데이터를 필요로 했고, 화자나 톤을 바꾸는 등 변형을 할 때 마다 새로운 데이터가 필요했습니다.

wavenet은 이와 다르게 오디오 파형을 직접 모델링하여 훨씬 자연스러운 음성을 생성하고, 컨디션 모델링을 통해 다양한 음성을 생성할 수 있습니다.

2. WaveNet의 특징

WaveNet은 오디오 파형 데이터를 직접 사용하여 새로운 파형을 모델링 합니다. 파형 x = x_1,x_2,….,x_n은 조건부 확률을 이용하여 아래와 같이 나타냅니다.

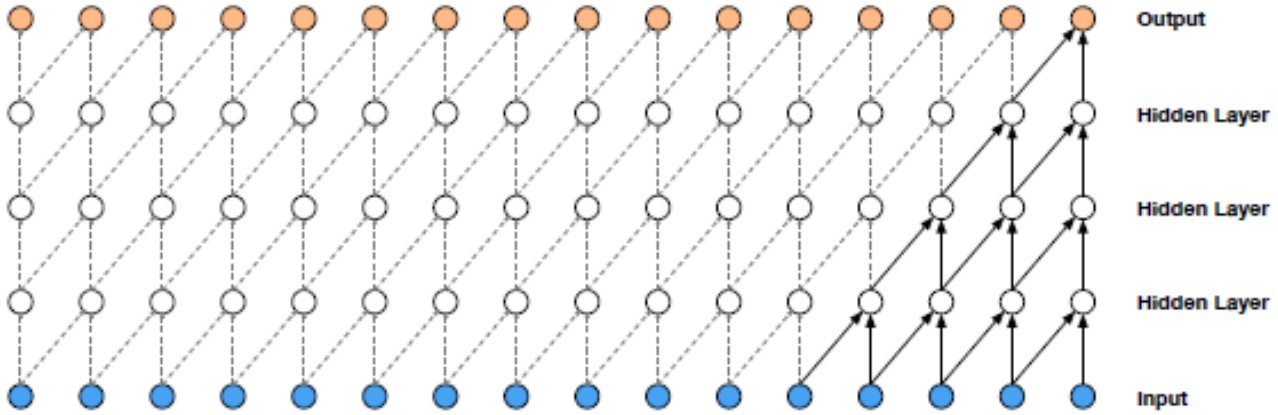
$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

식을 보면 각 샘플 x_i의 확률분포는 앞으로 올 데이터 x_t+1이 아닌 과거데이터, 즉 이전 샘플에 의해서 결정됩니다.

2. WaveNet의 특징

- Dilated Caual Convolutions

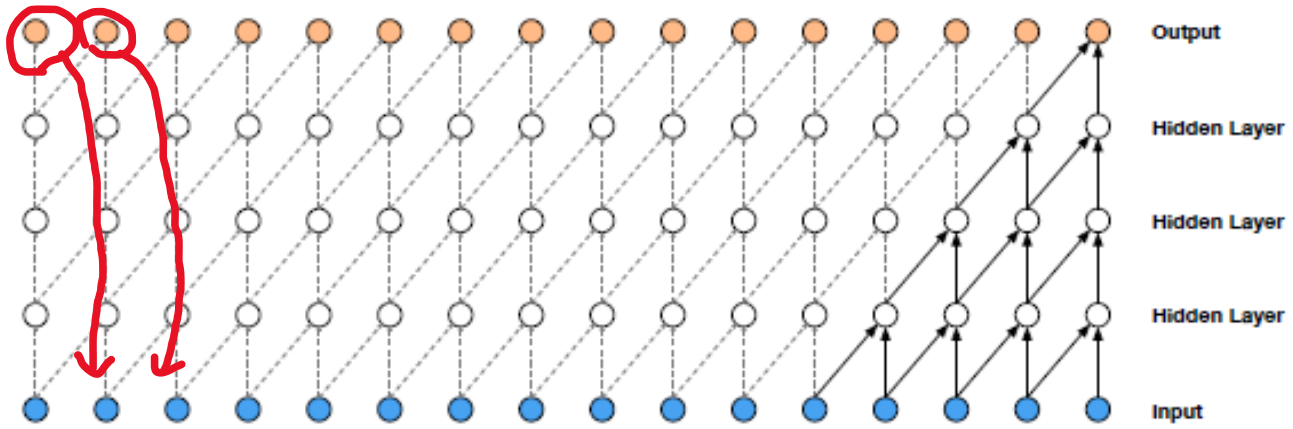
우선, 파형이 과거 샘플에 의해 결정되므로 과거 정보에만 접근 가능하도록 Caual Convolution layer를 여러겹 쌓았습니다.



2. WaveNet의 특징

- Dilated Caual Convolutions

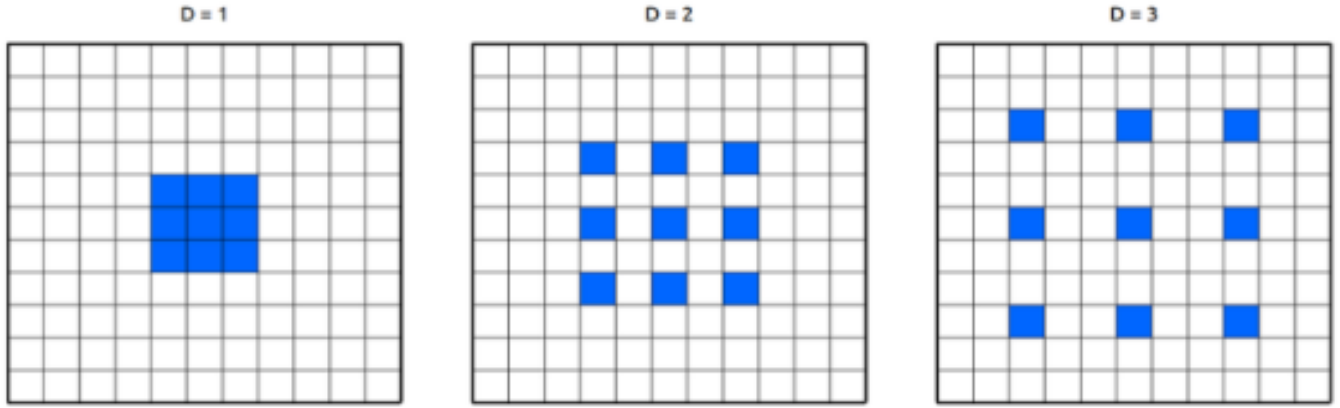
Caual Convolution layer 에서는 음성을 생성할 때 예측을 한 스텝 씩 진행하고, 예측한 결과값이 다음 입력값으로 함께 들어가게 됩니다.



2. WaveNet의 특징

- Dilated Caual Convolutions

Caual Convolution만 사용하기에는 receptive field가 넓어지기 위해 많은 층을 필요로 합니다. 이 문제를 해결하기 위해 Dilated Convolution을 사용합니다. Dilated Convolution은 일정 스텝을 건너 뛰며 filter를 적용합니다.



Receptive field : 3x3 = 9

Receptive field : 5x5 = 25

Receptive field : 7x7 = 49

2. WaveNet의 특징

- μ-law companding

  일반적으로 오디오는 16-bit 정수값으로 저장하여 사용하기 때문에 2^16 = 65536개의 확률을 다루어야 합니다. 때문에 이 수를 줄이기 위해 μ-law companding을 적용하여 256개의 값 중 하나로 양자화 했습니다.

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

* Companding : 제한된 동작범위의 채널이 가진 단점을 완화시키는 방식

2. WaveNet의 특징

- Gated Activation Unit

PixelCNN에서 사용된 Gated Activation Unit을 사용합니다.

매 층마다 입력값이 주어지면 filter와 gate에 대한 convolution을 각각 구한 뒤 elememt-wise 곱을 구합니다.

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

⦿ : element-wise 곱

W_f : filter convolution

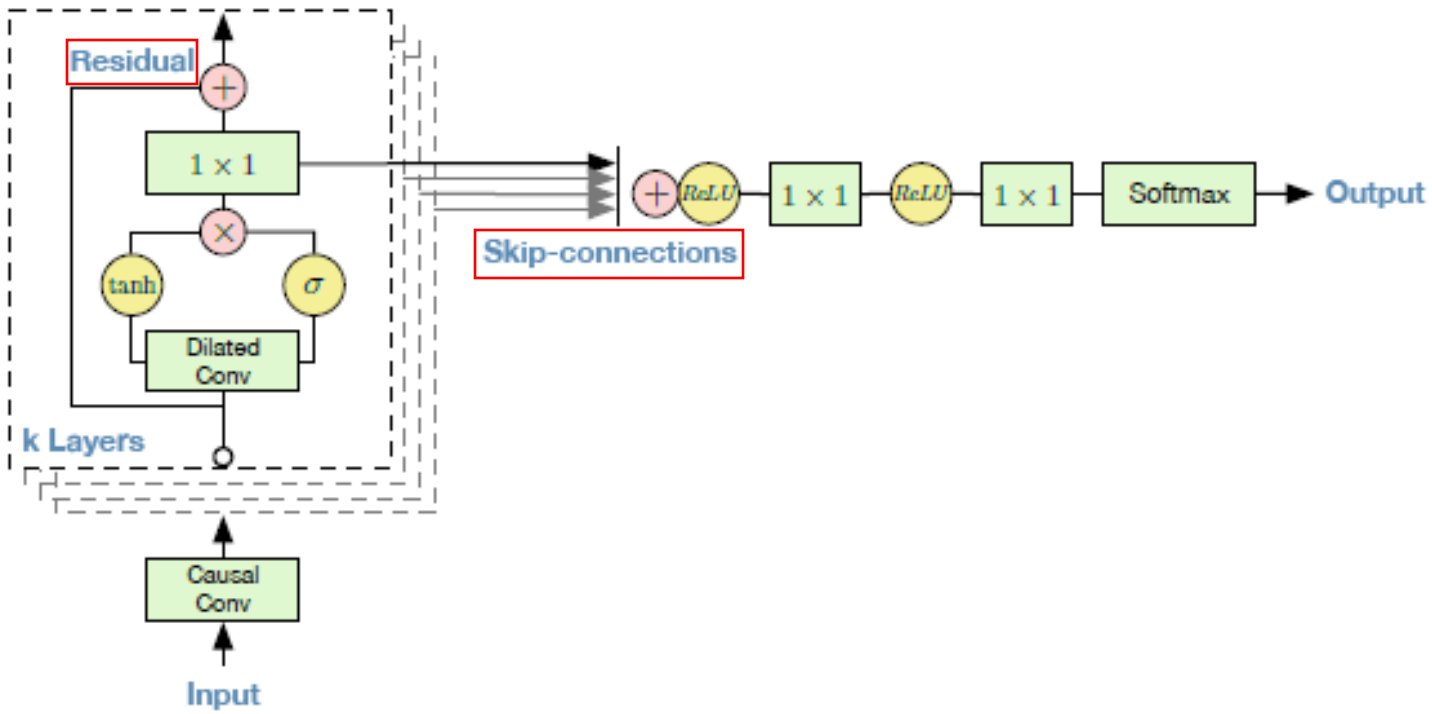
W_g : gate convolution

k : layer number

2. WaveNet의 특징

- Residual & Skip connections

학습 시간 단축을 위하여 잔차 연결과 skip-connection을 사용합니다.



2. WaveNet의 특징

- Conditional WaveNet

  wavenet에서는 특정 조건 h가 추가되었을 때, 조건부 확률을 다음과 같이 구할 수 있습니다.

$$p(x \mid h) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}, h)$$

  예를 들어, 여러 화자의 음성 데이터를 입력 받고 오디오를 생성 할 때, 각 화자의 정보들을 조건 h로 설정하여 생성 할 수 있습니다.

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

  ← 조건 h가 추가되었을 때 activation function