[논문 리뷰] Tacotron 2 : NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

diddu · 약 1시간 전

논문

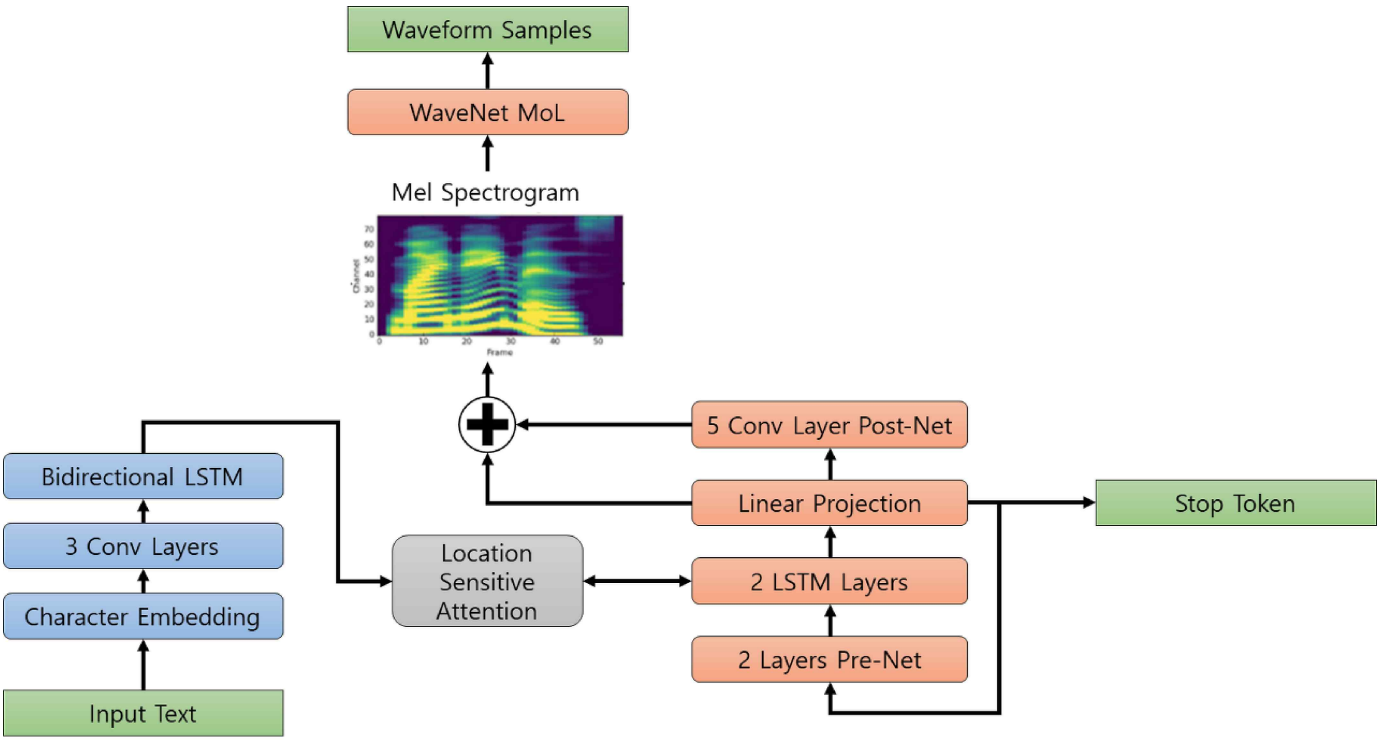
💡 Tacotron 2 : NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS_논문 리뷰

🔭 cf) Tacotron2 란?

2018년 Google에서 발표한 Tacotron2는 텍스트를 음성으로 변환하는 TTS(Text-to-Speech) 모델이다. Seq2Seq(Sequence-to-Sequence) 구조를 기반으로 하며, 주요 구성 요소로는 Encoder, Decoder, Attention이 있다.

- Encoder : 문자를 일련의 hidden 벡터(feature)로 변환
- Attention : 시간 순서에 따라 Encoder에서 생성된 hidden 벡터의 정보를 추출하여 Decoder에 전달
- Decoder : 이 정보를 사용하여 mel-spectrogram을 생성 (이렇게 생성된 mel-spectrogram은 수정된 WaveNet 모델인 vocoder에 의해 시간 영역 waveform으로 합성)

📌 Architecture



`Spectrogram Prediction Network` : 문자 시퀀스를 Mel spectrograms으로 변환
문자 임베딩을 통해 입력된 문자들이 3개의 컨볼루션 레이어를 거친 후 양방향 LSTM(512개의 뉴런)을 통해 인코딩된 피처로 변환된다.

`Modified WaveNet` : Mel spectrograms를 음성으로 전환
Tacotron에서 생성한 mel-spectrogram과 WaveNet Vocoder, 그리고 Griffin-Lim 알고리즘이 결합하여 Tacotron 2가 완성

1. ***Encoder***

- Character Embedding, 3 Convolution Layer, Bidirectional LSTM으로 구성
- 입력된 one-hot vector는 Embedding matrix를 통해 512차원의 embedding vector로 변환
- 이후 3개의 conv-layer와 bi-LSTM layer를 거쳐 최종적으로 encoded feature가 생성

2. ***Attention***

- Encoder와 Decoder의 LSTM에서 생성된 feature를 이용하여 alignment하는 과정 포함
- 이 모델은 Location Sensitive Attention을 사용 (Additive attention mechanism에 attention alignment 정보가 추가된 형태_

3. ***Decoder***

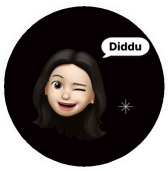
- Pre-Net, Decoder LSTM, 그리고 Post-Net이라는 세 부분으로 구성
    - Pre-Net : 중요 정보를 거르는 역할
    - Decoder LSTM : 특정 시점에 해당하는 정보를 생성
    - Post-Net : 생성된 mel-vector를 보정하여 최종적인 mel-spectrogram의 품질을 높임

📌 특징

`End-to-end architecture`
모델의 각 부분들이 서로 상호 작용하면서 최적화되기 때문에, 전체 시스템의 성능 향상에 도움이 된다. 반면 Tacotron 1은 여러 개별 컴포넌트들로 구성되어 있어 각 부분마다 따로 최적화를 해야 한다.

`Griffin-Lim Algorithm`
Tacotron 1에서 처음 등장한 Griffin-Lim 알고리즘이 Tacotron 2에서도 사용됐다.
(spectrogram으로 변환할 때 STFT에 의해 버려진 phase 정보를 예측하는 역할)

`Improved attention mechanism`
위치 정보를 고려하는 Location Sensitive Attention 메커니즘이 도입됐다. 때문에 긴 문장과 복잡한 문장 패턴도 잘 처리할 수 있게 되었다.



Diddu

다음 포스트
[논문 리뷰] VATT: Transformers for Multimodal Self-Supervised Learning fro...

이전 포스트
[논문 리뷰] PaLM-E: An Embodied Multimodal Language Model

0개의 댓글

댓글을 작성하세요

댓글 작성