Tacotron을 이용한 “책 읽어주는 주교수님”

COSE474

2015410008 윤현식
2015410112 최문영

목 차

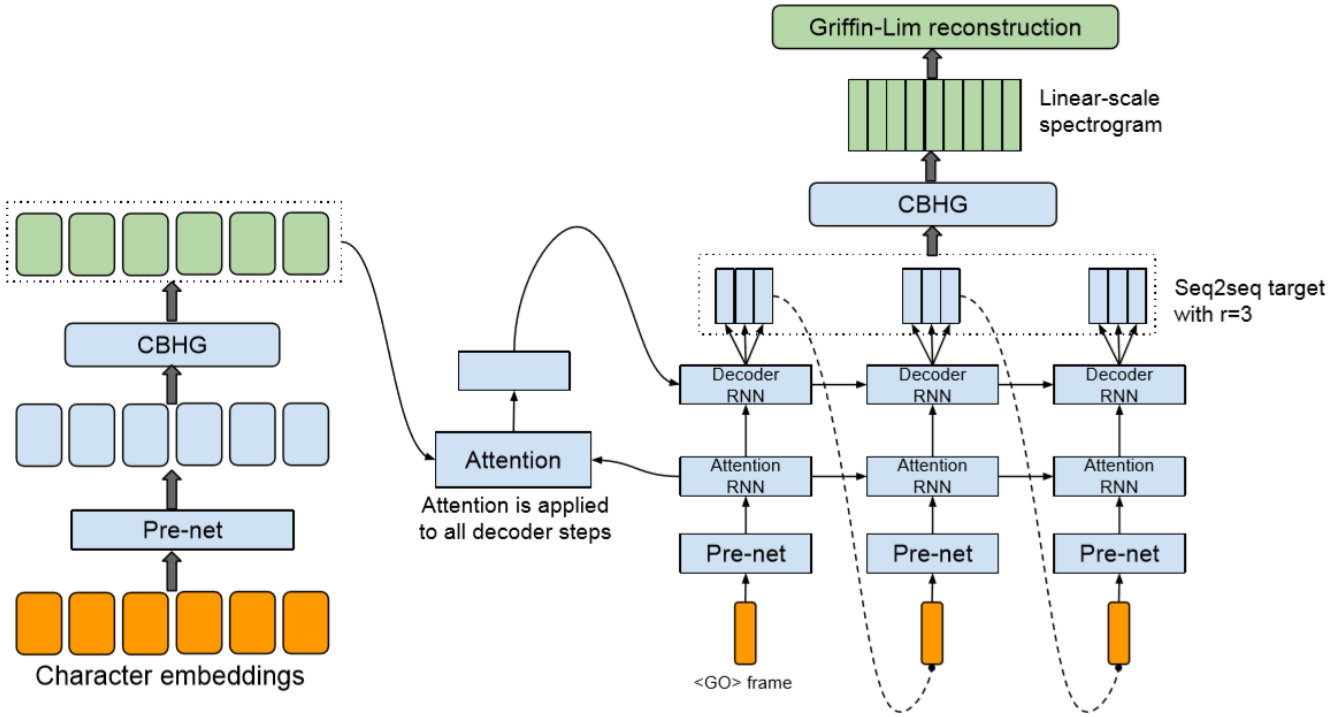
- 프로젝트 개요
- 프로젝트 상세
- 성능 및 시연
- 보완할 점/어려웠던 점

프로젝트 개요

모 델

Tacotron

- Tacotron: Towards end-to-end speech synthesis 모델을 구현하였다.
- Character level에서 작동하며, feature를 직접 입력하거나 이전에 존재하는 TTS system으로부터 뽑아 입력할 필요가 없다.



데이터셋

- 주재걸 교수님의 온라인 강의(KMOOC)에서 오디오 데이터셋과 자막을 얻었다.
- 오디오를 silence 단위로 잘라서 자막과 매칭했다.*

*https://github.com/carpedm20/multi-speaker-tacotron-tensorflow 참조

모델 인풋/아웃풋

- 모델 인풋
  - Train
    - Encoder : text sequence
      - e.g. “Let’s start the lecture.”
    - Decoder : 해당 text sequence의 80-band mel-scale spectrogram
  - Test
    - Encoder : 음성으로 합성할 text sequence
- 모델 아웃풋
  - input text sequence에 해당하는 waveform

프로젝트 상세

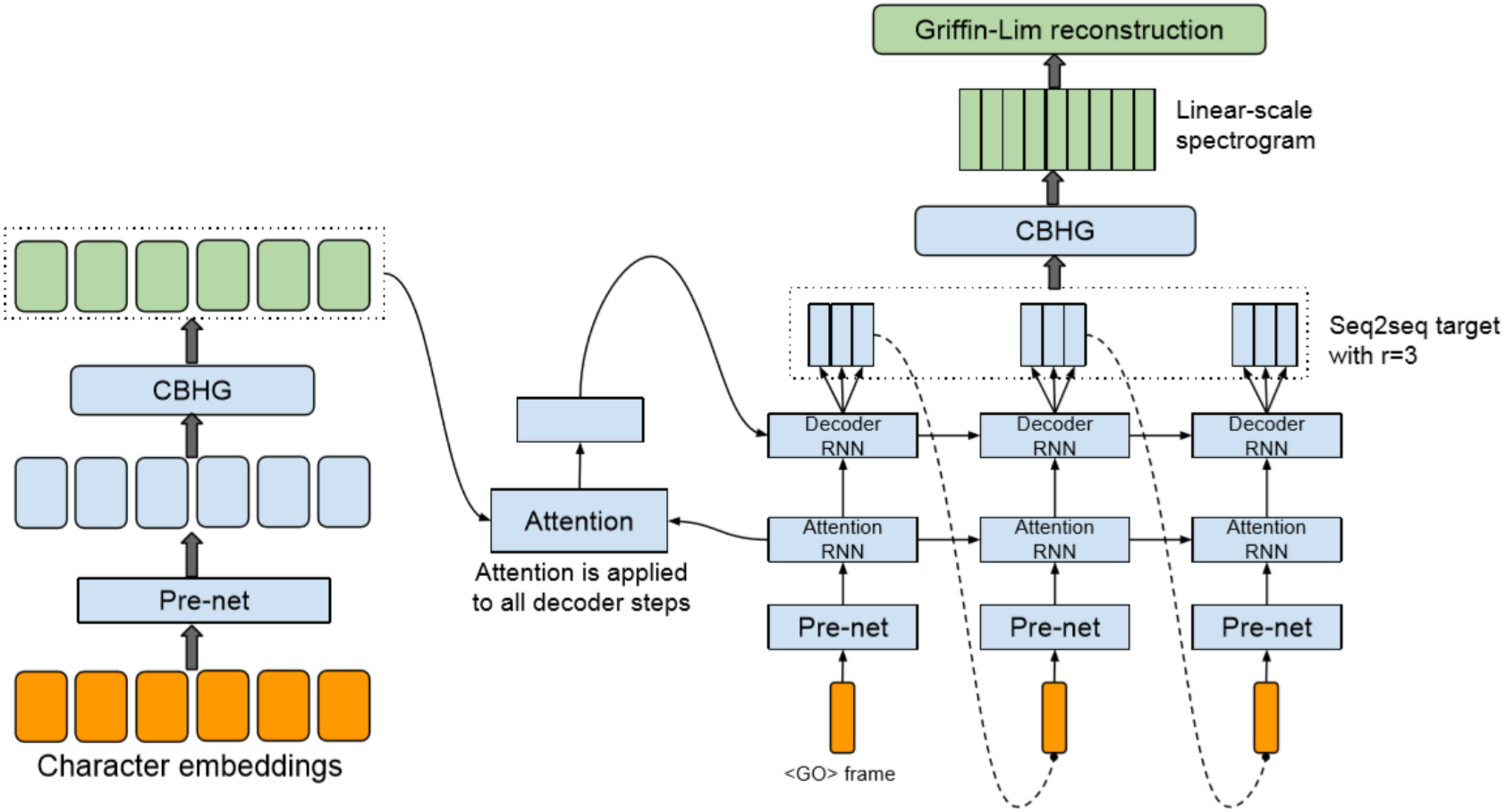
Tacotron

개요

- Tacotron은 generative text-to-speech 모델이다.
- Text sequence로부터 spectrogram을 합성하고, 이를 post processing algorithm (Griffin-Lim) 으로 phase estimation하여 waveform을 만든다.
- ⟨text, audio⟩ pair를 데이터셋으로 사용한다.

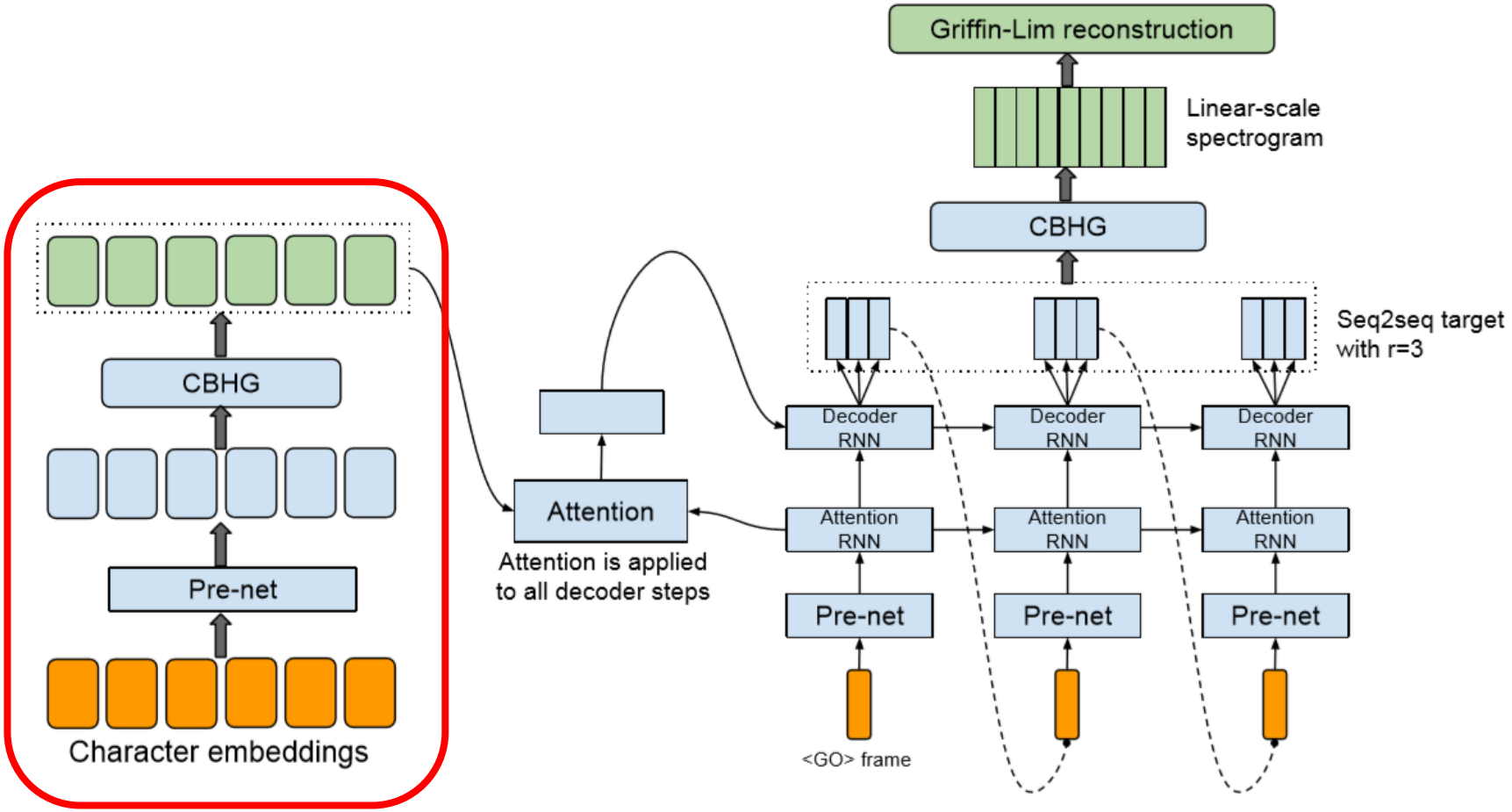
Tacotron

Model Architecture



Tacotron

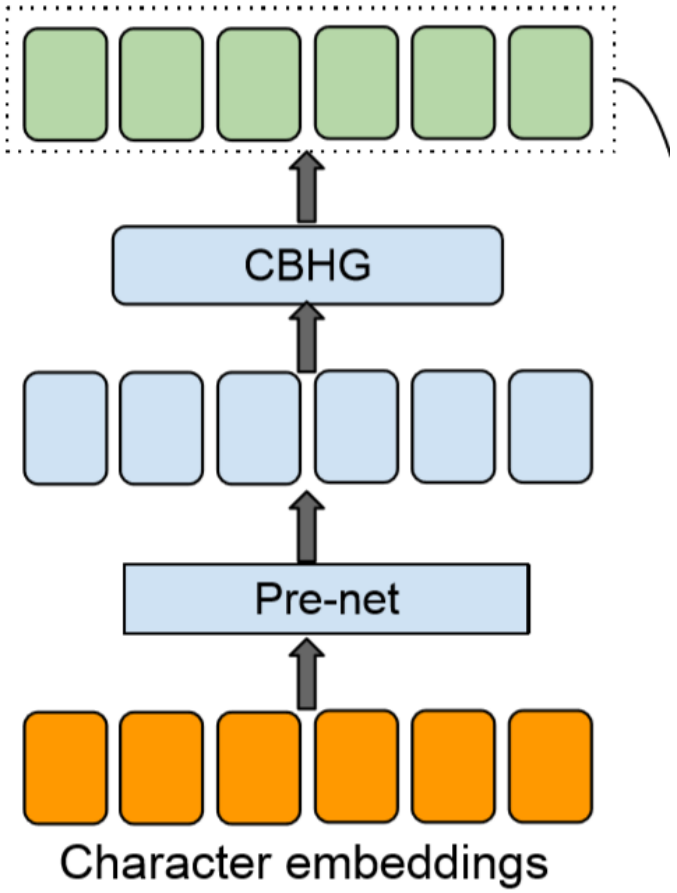
Encoder



Tacotron

Encoder

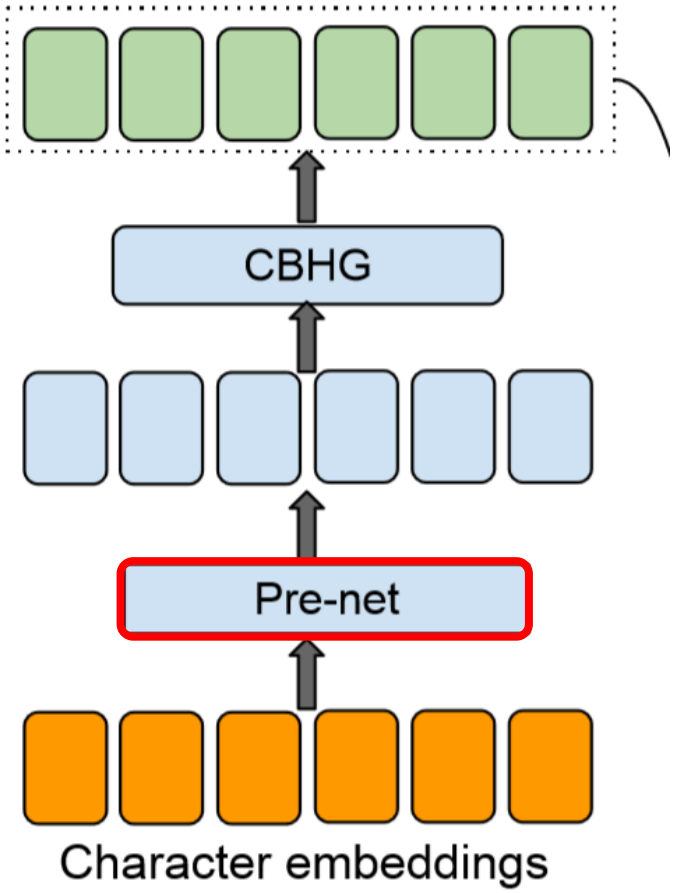
- Text의 representation을 추출하기 위한 모듈이다.
- Encoder의 인풋은 각 character들이 one-hot vector로 표현된 sequence이다.
- Encoder의 아웃풋은 마지막 GRU에서 나온 sequential feature이다.



Tacotron

Pre-net

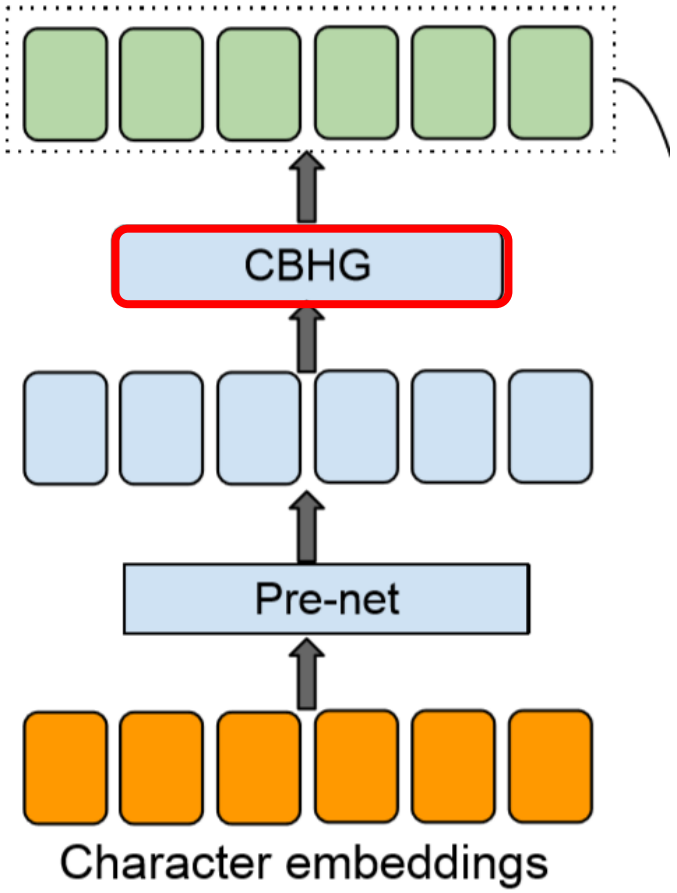
- FC-ReLU-Dropout-FC-ReLU-Dropout 으로 구성된다.
- Convergence와 일반화에 도움이 된다.



Tacotron

CBHG

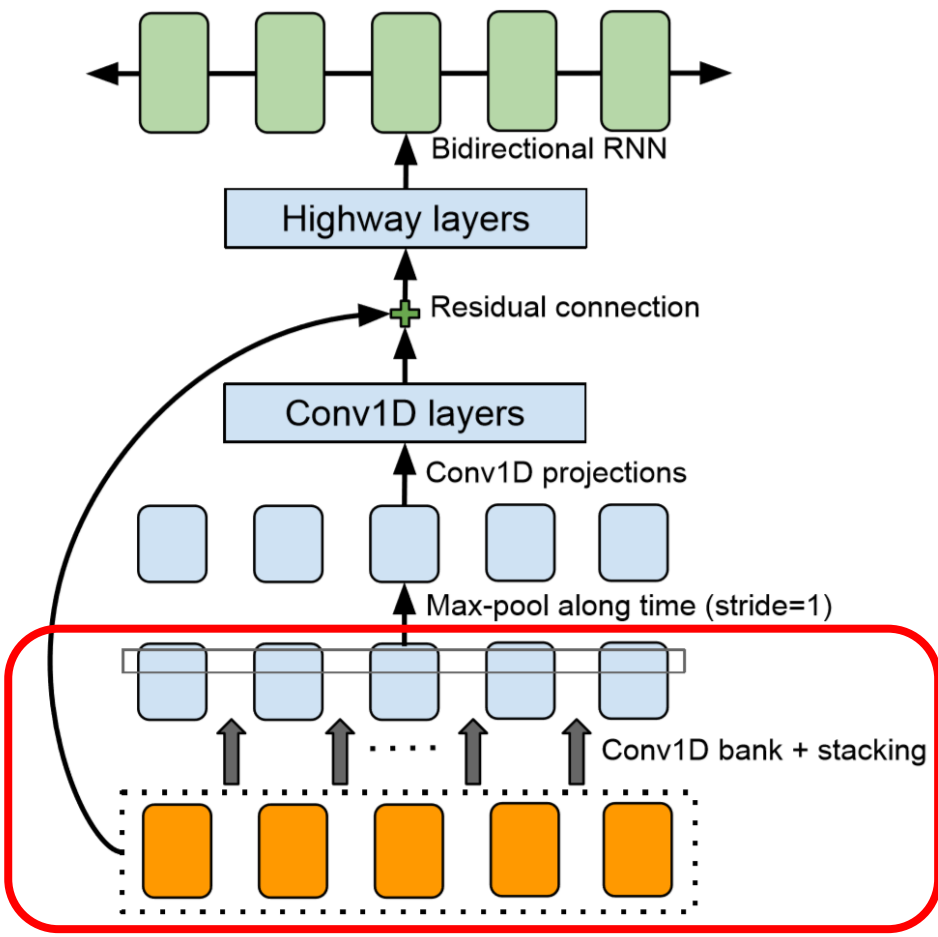
- Convolution Bank, Highway Net, GRU로 구성되어 CBHG라고 한다.
- Overfitting을 줄여준다.
- 일반적인 multi-layer RNN encoder보다 mispronunciation이 적다.



Tacotron

CBHG – Conv1D bank + stacking

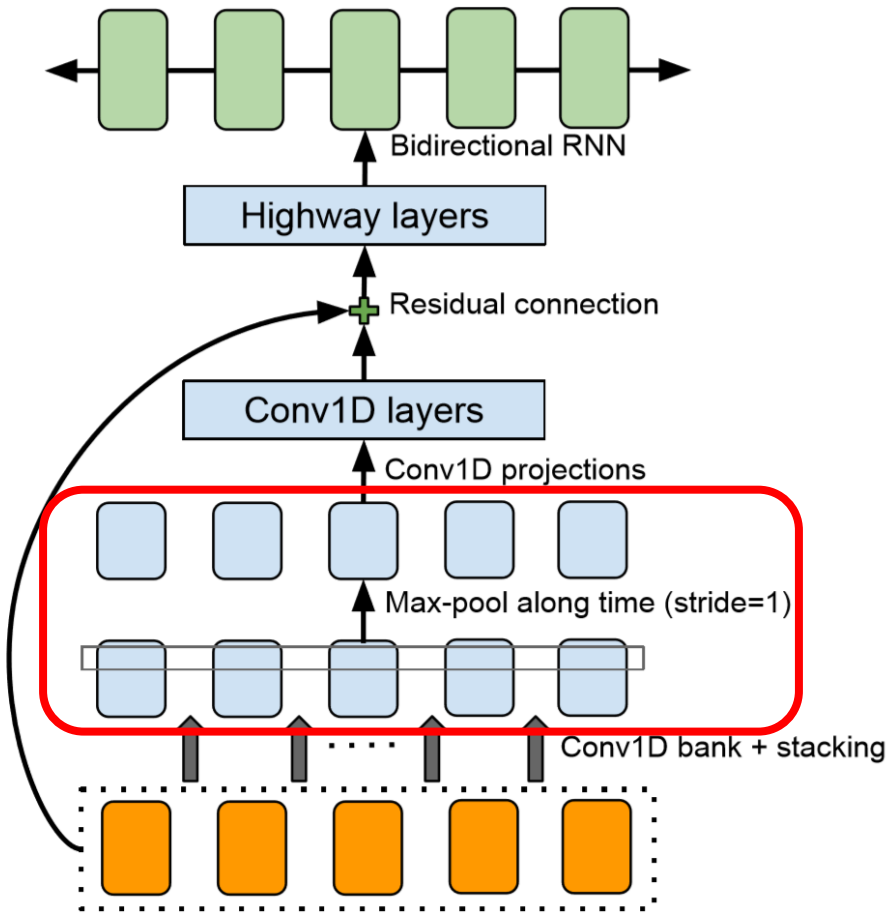
- 여러 사이즈(k)의 filter로 Conv1D Layer를 구성한다.
- k-gram 을 얻기 위함이다.
- 해당 Layer의 input들을 쌓는다.



Tacotron

CBHG – Max-pool along time

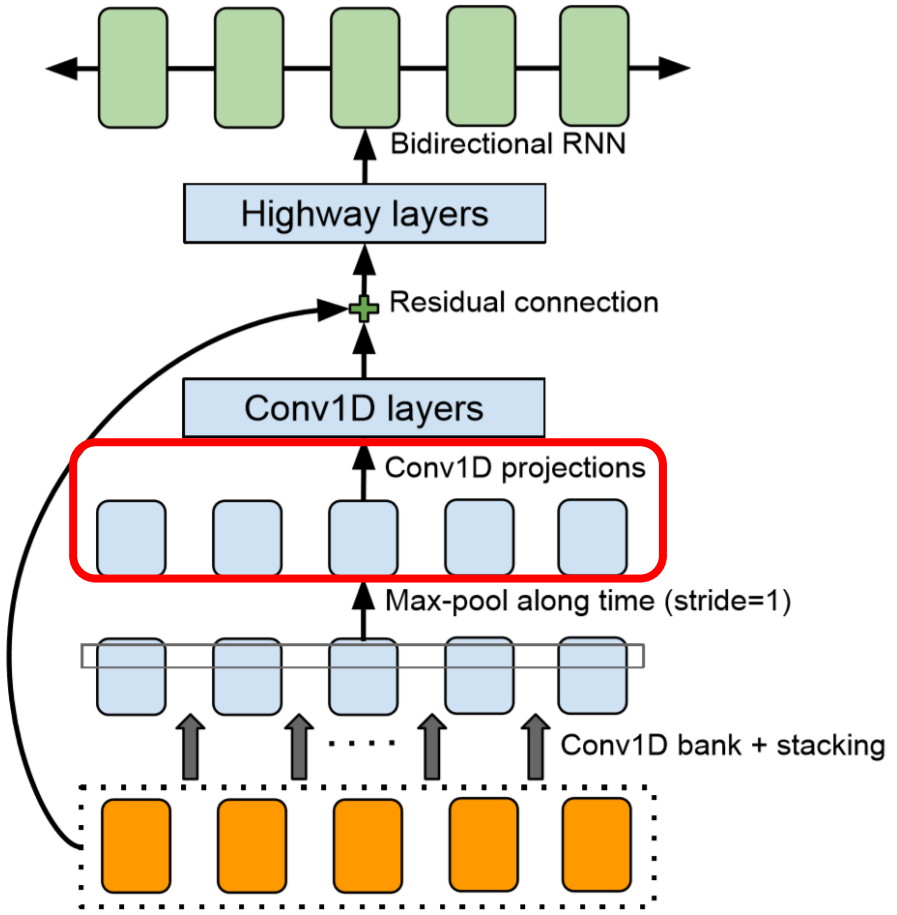
- Local invariance를 키우기 위해 max pooling을 적용한다.
- 시간축 상의 resolution을 유지하기 위해 stride는 1로 설정한다.



Tacotron

CBHG – Conv1D projections

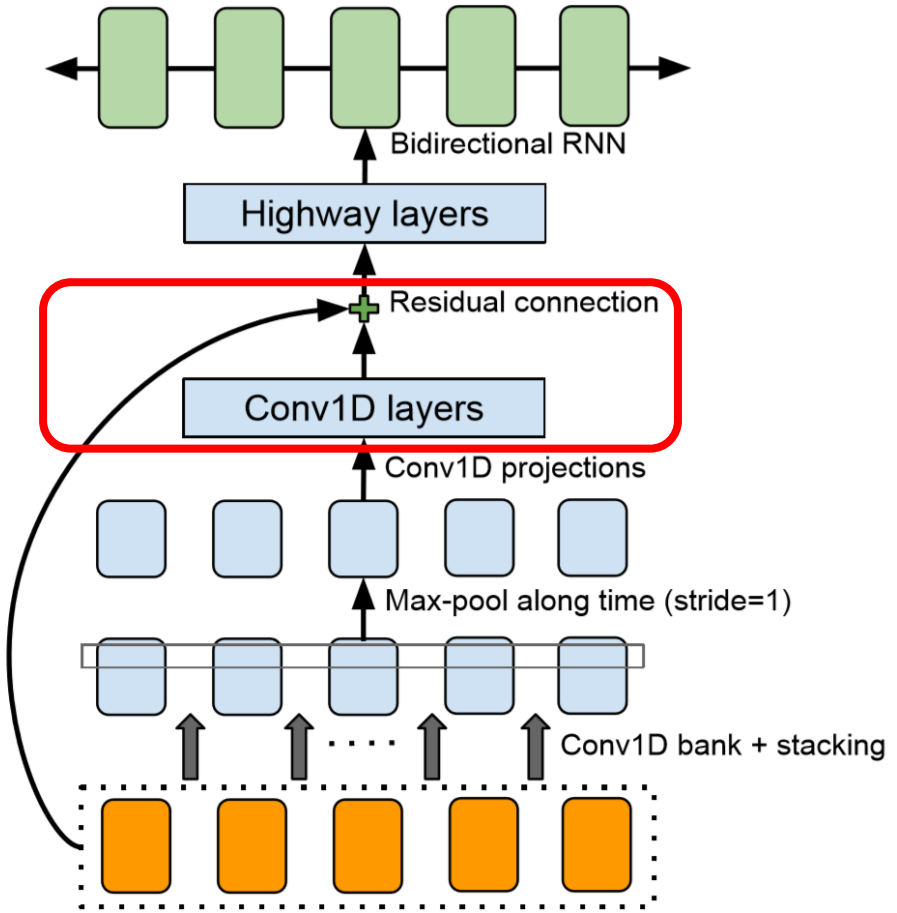
- Conv-ReLU-Conv 로 구성된다.
- High level feature들을 뽑기 위함이다.



Tacotron

CBHG – Residual connection

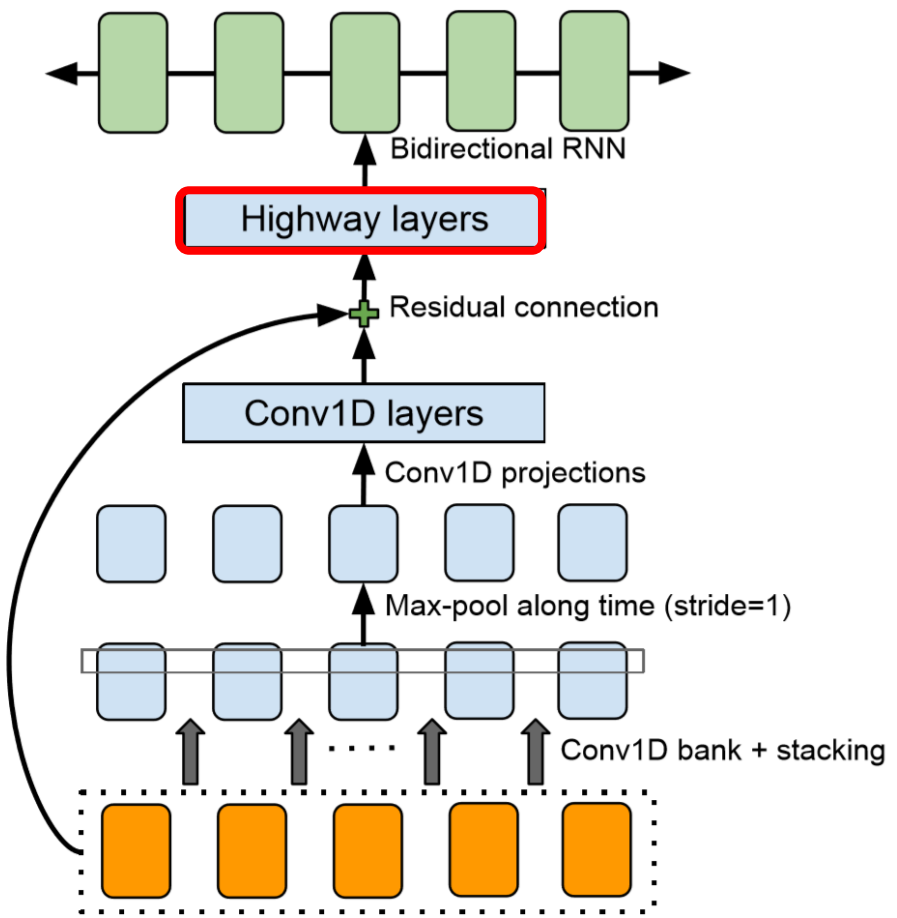
- Convergence 와 일반화를 돕기 위해 적용한다.
- Conv1DBank를 거치기 전의 인풋과 Conv1D projection의 아웃풋을 더한다.



Tacotron

CBHG – Highway layers

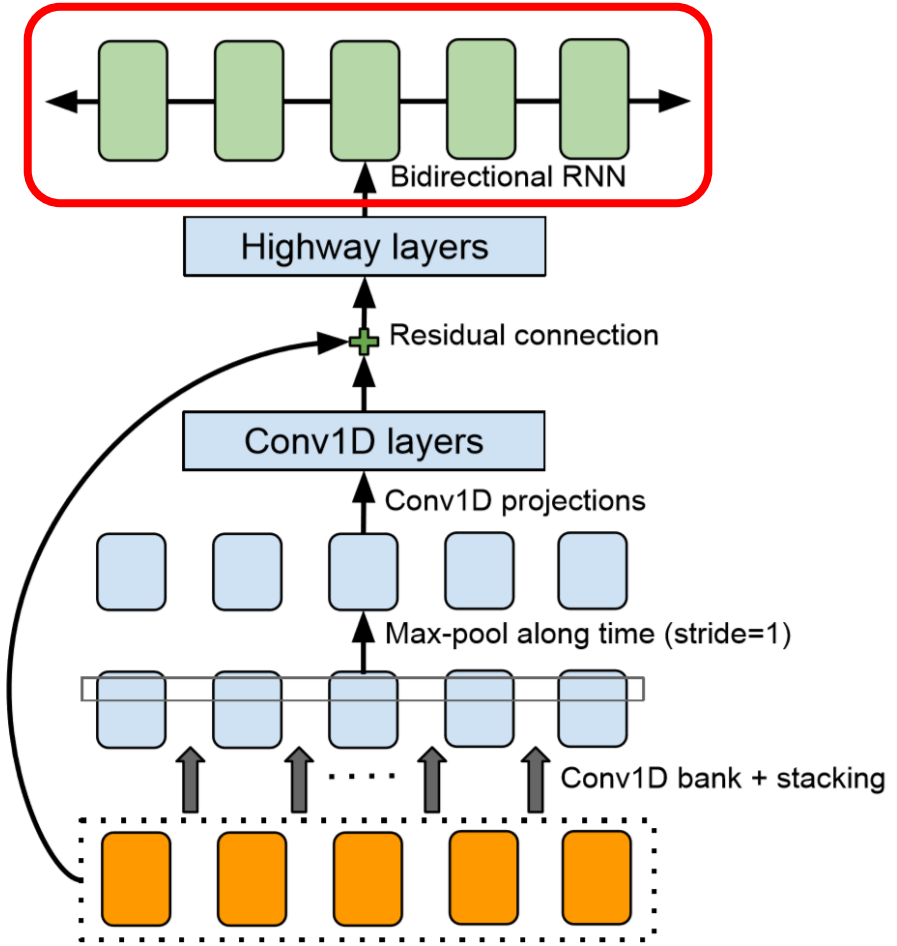
- 원래의 인풋과 한 층의 FC layer 를 거친 결과를 weighted sum 하는 구조이다.
- $y = H(x,WH) \cdot T(x,WT) + x \cdot (1-T(x,WT))$



Tacotron

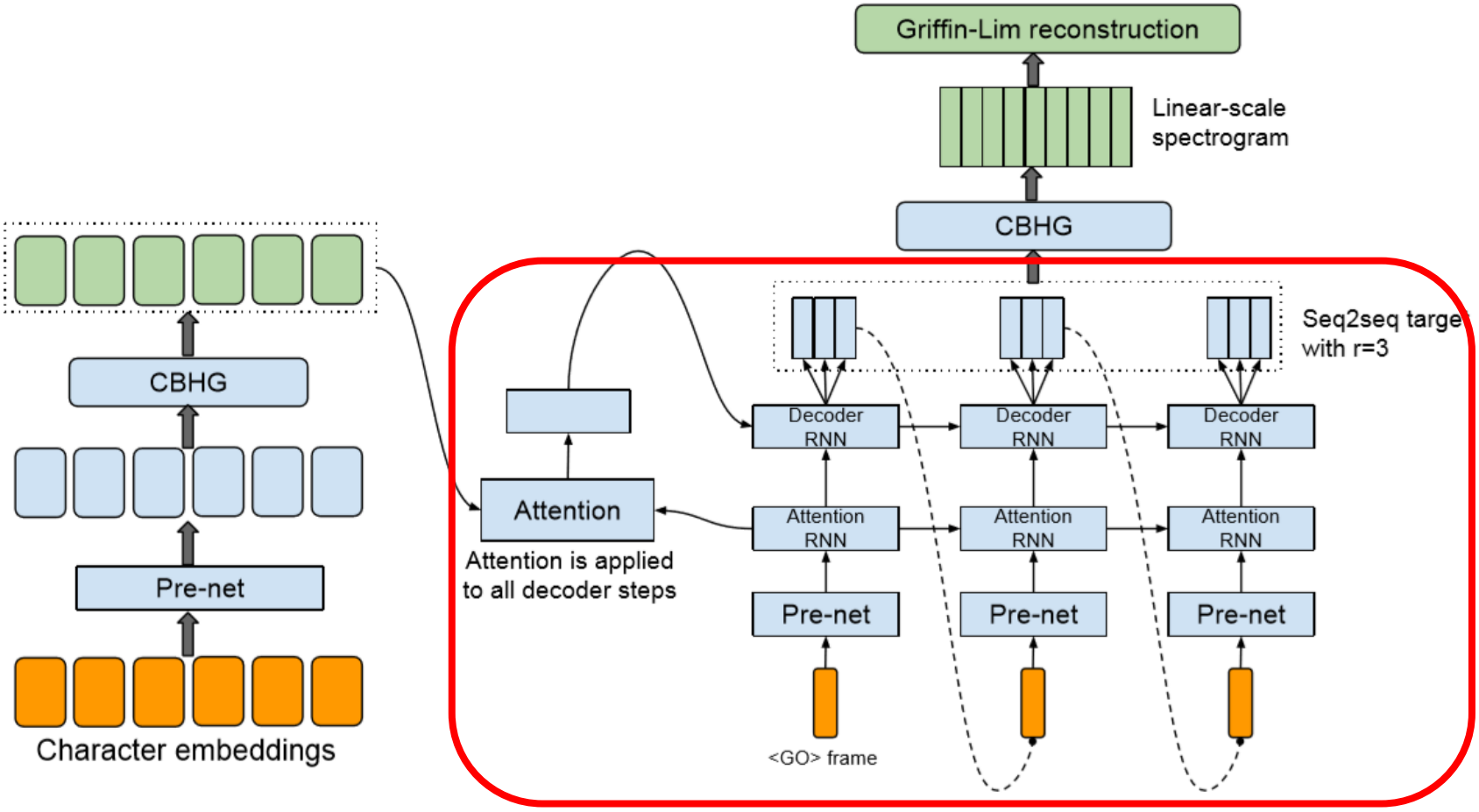
CBHG – Bidirectional RNN

- GRU를 사용한다.
- Forward, backward 방향 각각에서 sequential feature를 뽑아내기 위해 적용한다.



Tacotron

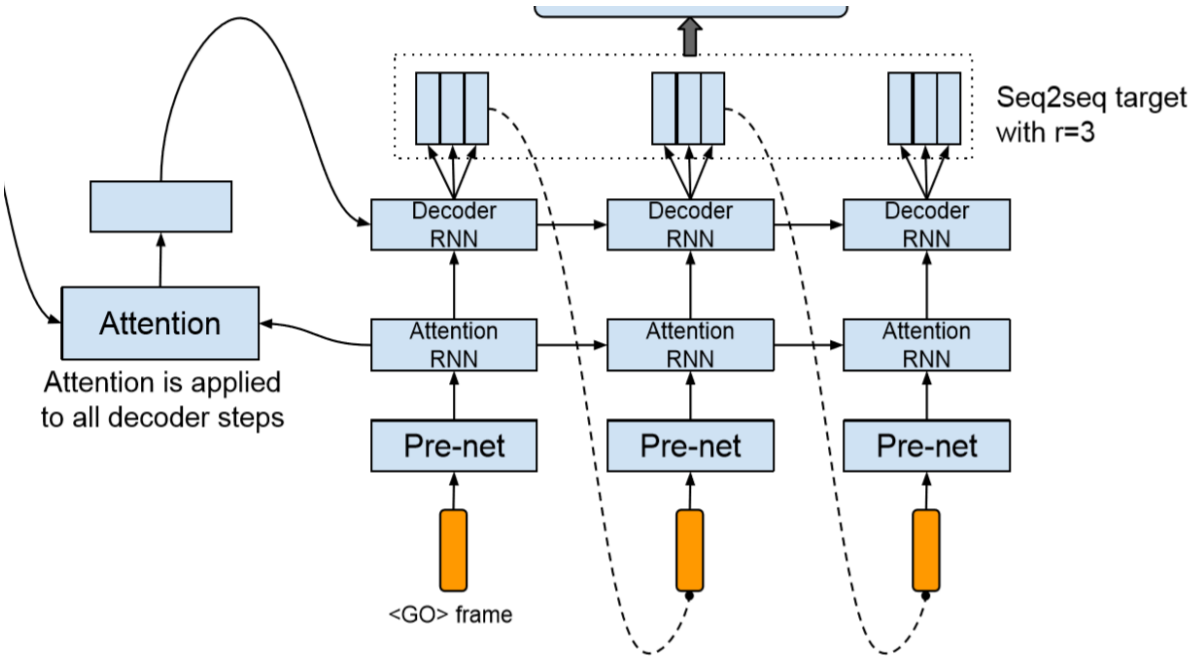
Decoder



Tacotron

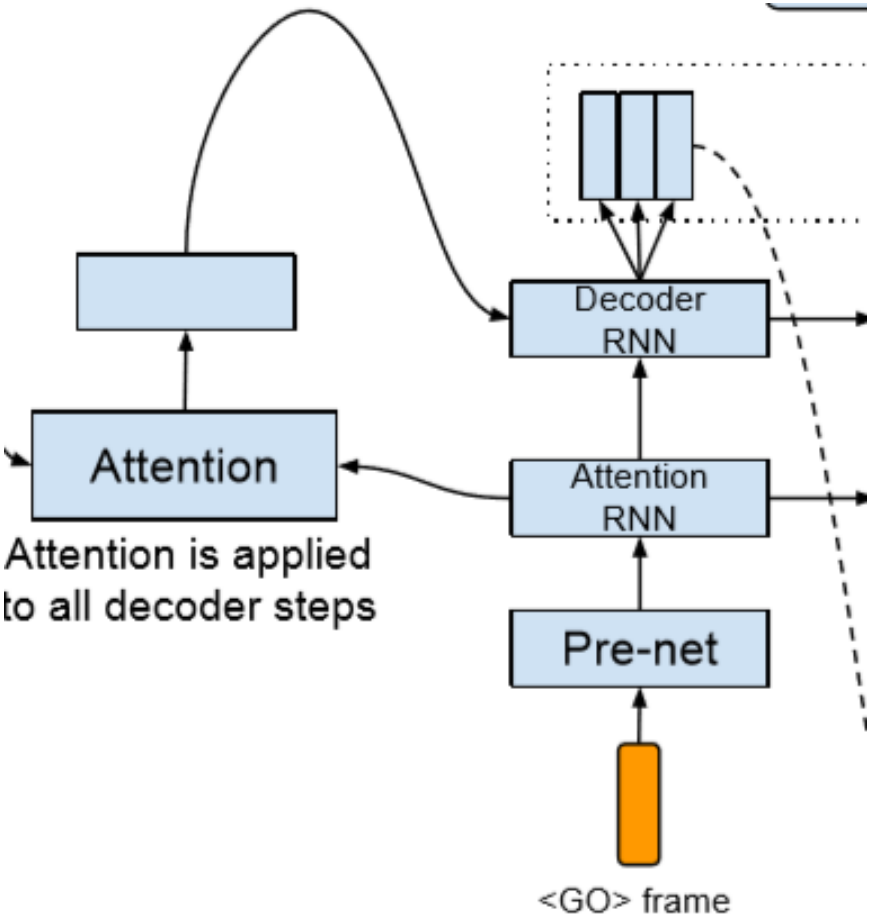
Decoder

- Train time
  - Spectrogram을 input으로 받는다.
- Test time
  - 전 단계에서 예측한 target spectrogram의 전부, 혹은 마지막 frame만을 다음 단계의 input으로 받는다.



Tacotron

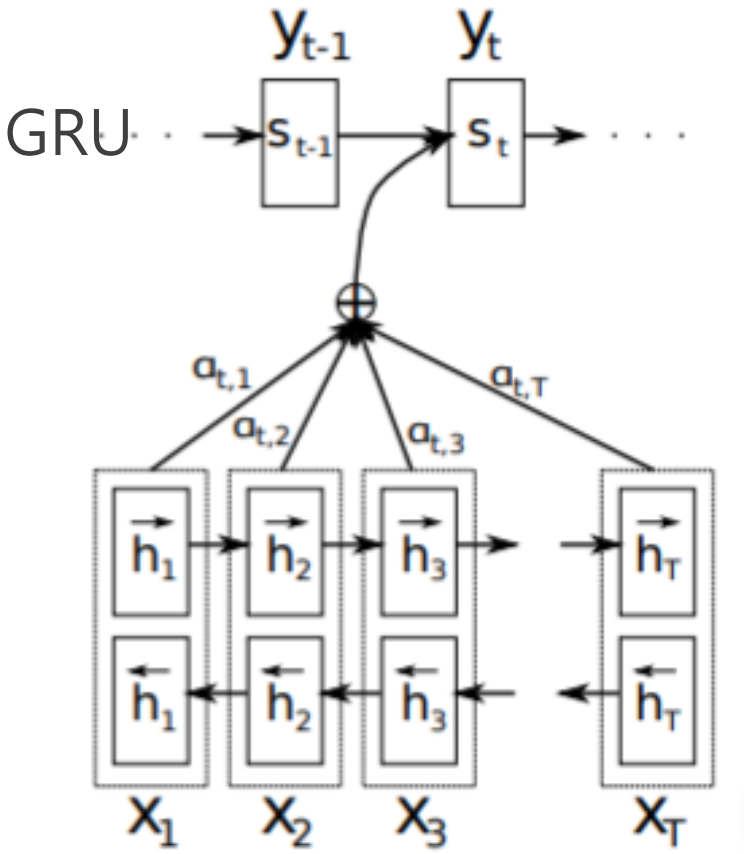
Decoder



- ⟨GO⟩ frame
  - all zero frame
- Prenet
  - Encoder단계의 prenet과 동일, dropout
- Attention RNN
  - Content based attention (Bahdanau Attention)
  - GRU와 연결
- Decoder RNN
  - AttentionRNN output과 context vector concatenation+projection 하여 사용
  - Stack of GRUs + residual connection
- Next Frame
  - Multiple spectrogram frames

Tacotron

Attention



- GRU

$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i),$$
$$s_i = f(s_{i-1},y_{i-1},c_i).$$

- Context vector

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

- Attention weights

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

$$e_{ij} = a(s_{i-1}, h_j)$$

성능 및 시연

추후 추가 예정

보완할 점 / 어려웠던 점

추후 추가 예정