Listen, Attend and Spell 2

Winter Vacation Capstone Study

TEAM Kai.Lib

발표자 : 김수환

2020.02.03 (MON)

LAS Model

- **Listen, Attend and Spell Model**

현재 우리가 사용하는 모델과 거의 같다고 해도 무방.

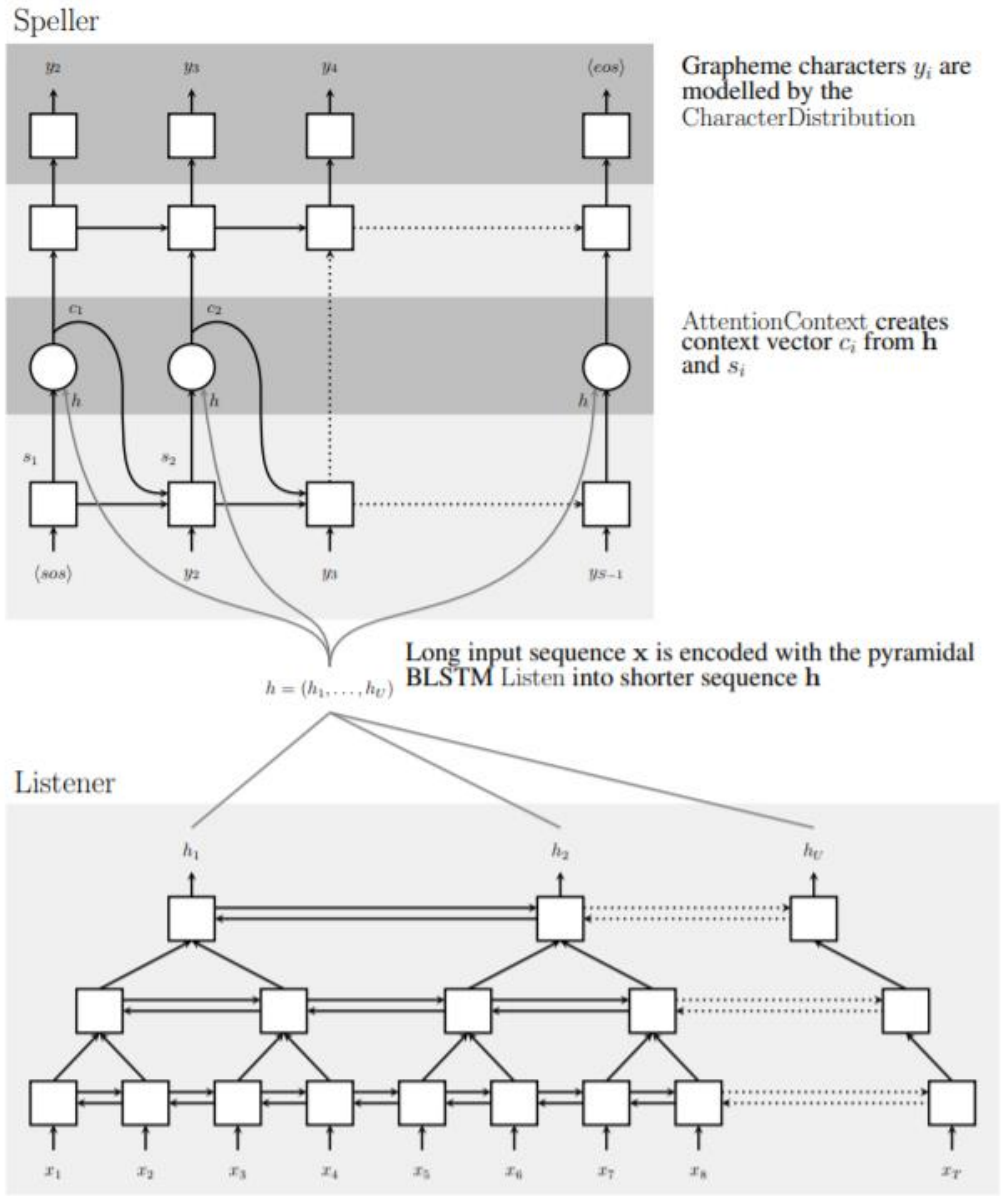
Seq2seq with Attention을 음성 인식에 적용한 모델.

character 단위로 학습.

※ 차이점 ※

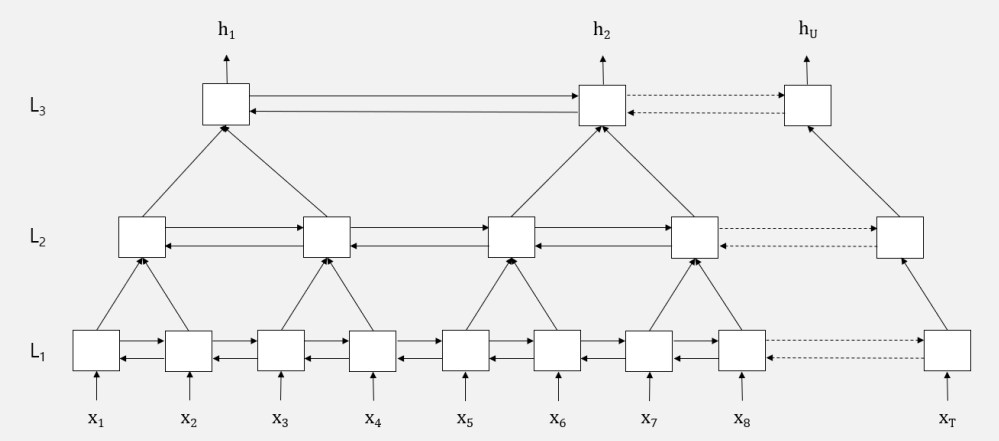
1. LAS는 인코더에서 pBLSTM(Pyrimidal LSTM)를 사용
2. 우리는 인코더 LSTM 레이어에 넣기 전 Convolution Layer 선행

(Deep Speech Style)

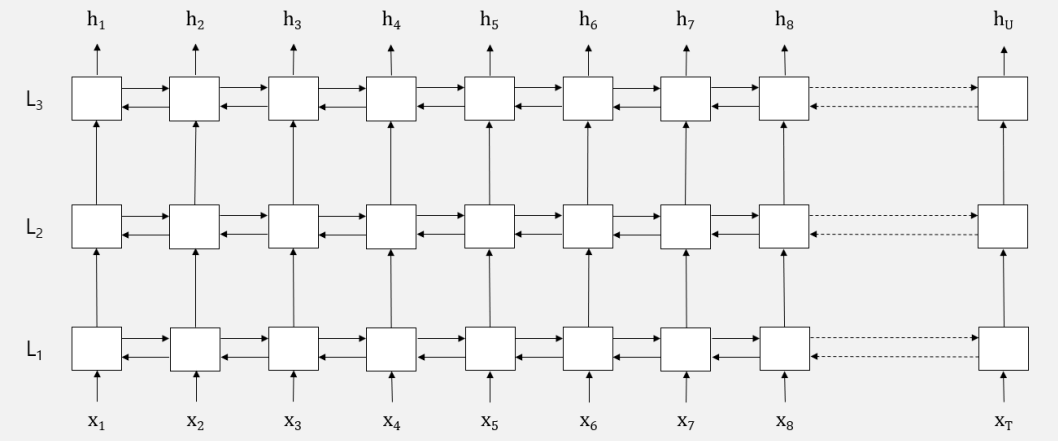


LAS Model

- **Pyrimidal LSTM**



pBLSTM



BLSTM

이전 레이어 2i,2i+1을 concatenate하여 다음 레이어의 i번째 RNN 셀의 입력으로 넣는 구조.
=> 더 압축하여 표현함으로써 Sequence Length를 줄일 수 있다.
이는 인코더 & 디코더 & 어텐션 메커니즘 모두에서 연산량 감소를 가능하게 한다.

LAS Model

- **Exposure Bias Problem**

Teacher Forcing은 Seq2seq 학습 기반에서 Default였다. (Default Teacher Forcing Ratio : 1.0)
하지만 Exposure Bias Problem이 존재한다. 실제 추론 시에는 정답이 아닌 잘못 예측한 인풋이 들어올 수도 있다.
본 논문에서는 이러한 문제점을 완화하기 위해 Teacher Forcing Ratio를 0.9로 낮춰서 실험을 진행함.

다만 노출 편향 문제가 생각만큼 큰 영향을 미치지는 않는다는 연구 결과가 나와 있다고 한다.
(T. He, J. Zhang, Z. Zhou, and J. Glass, Quantifying Exposure Bias for Neural Language Generation (2019), arXiv.)

=> 과연 그럴지는 실제로 테스트해봐야 된다고 생각 (발표자 의견)

LAS Model

▪ Decoding and Rescoring

Decoding

본 논문에서는 Beam-Search를 사용했다고 함. (Beam width = 32)

기존 방식으로는, 모든 Beam이 <eos>를 만나고 나면 가장 높은 확률을 갖는 Beam을 선택한 후,

Dictionary (사전) 을 통해 언어 교정을 하는 방식을 많이 사용했음.

하지만, 본 논문에서는 실험시, 이 Dictionary 방식이 별로 필요가 없다고 주장함. (거의 항상 완벽한 단어를 내놓았기 때문)

Rescoring

본 논문에서는 Dictionary 방식보다는, Recoring 방식이 더 낫다고 주장함.

32개의 Beam 후보를 뽑아놓은 후, 여기에 Language Model을 적용하여 가장 높은 점수를 받는 텍스트를 채택하는 방법

(여기서는 자신들의 방대한 텍스트데이터로 자연스러운 정도를 점수매겼다고 함)

$$s(\mathbf{y}|\mathbf{x}) = \frac{\log P(\mathbf{y}|\mathbf{x})}{|\mathbf{y}|_c} + \lambda \log P_{\mathrm{LM}}(\mathbf{y})$$

Experiment

- **LAS Model's Experiment**

Data

Category	hours
train	2000h
valid	10h
test	16h

Feature

Parameter	Use
feature_size	40
feature_extraction	log mel
frame_length	10ms

Hyperparameters

Hyperparameter	Use
encoder_layer_size	3
decoder_layer_size	2
hidden_size	256
batch_size	32
teacher_forcing	1.0 & 0.9

Experiment

Performance

No Language Model

teacher forcing (%)	clean WER (%)	noisy WER (%)
100	16.2	19.0
90	14.1	16.5

Apply Language Model

teacher forcing (%)	clean WER (%)	noisy WER (%)
100	12.6	14.7
90	10.3	12.0

Experiment

- **Performance**

Model	clean WER (%)	noisy WER(%)
CLDNN-HMM (SOTA)	8.0	8.9
LAS + Sampling + LM Rescoring	10.3	12.0

LAS 모델은 당시 SOTA (State-Of-The-Art) 모델인 CLDNN-HMM과 2.3%의 WER 정도만의 차이를 기록
본 논문에서는 이 차이를 Convolutional filter의 유무일 것이라고 추정함

Proposal

- **해당 내용들을 바탕으로 우리 모델에 추가하고 싶은 점 제안**

1. pBLSTM 적용
2. Teacher Forcing Ratio 0.99 → 0.90으로 하향 조정
3. Training Data 증가 (현재 800시간을 사용했지만, 너무 많은 데이터를 테스트 데이터로 사용했다고 생각)

   (980시간 정도로 학습하고 20시간으로 테스트 및 15시간 정도를 홀드아웃 검증 데이터로 사용하는 방법 제안)
4. Language Model 도입 (Rescoring 기법 사용)

   (공부가 필요한 부분)