A spelling correction model for end-to-end speech recognition

Jinxi Guo, Tara N.Sainath, Ron J. Weiss

2020 IEEE International Conference on
Acoustics, Speech, and Signal Processing

정영석

Contents

논문 선정 이유

▪ 논문 선정 이유

- End-to-End 음성인식 모델은 음성 신호를 직접 입력 받음

  → 입력 데이터의 품질에 많은 영향을 받음

- 음성인식 교정 분야에서 영어를 제외한 다른 언어에 대한 연구 부재

Review

• Abstract

▪ 기존 연구의 한계

- End-to-End 음성인식 모델

  → 데이터가 적어 드물게 나타나는 단어에 대한 처리가 어려움

- 사전학습된 RNN-LM과 cold · shallow fusion을 수행

  → 음성인식의 결과에서 생성된 단어를 고려하기 어려움

▪ 제안 방법론

- STT → TTS → STT 방법으로 데이터 생성

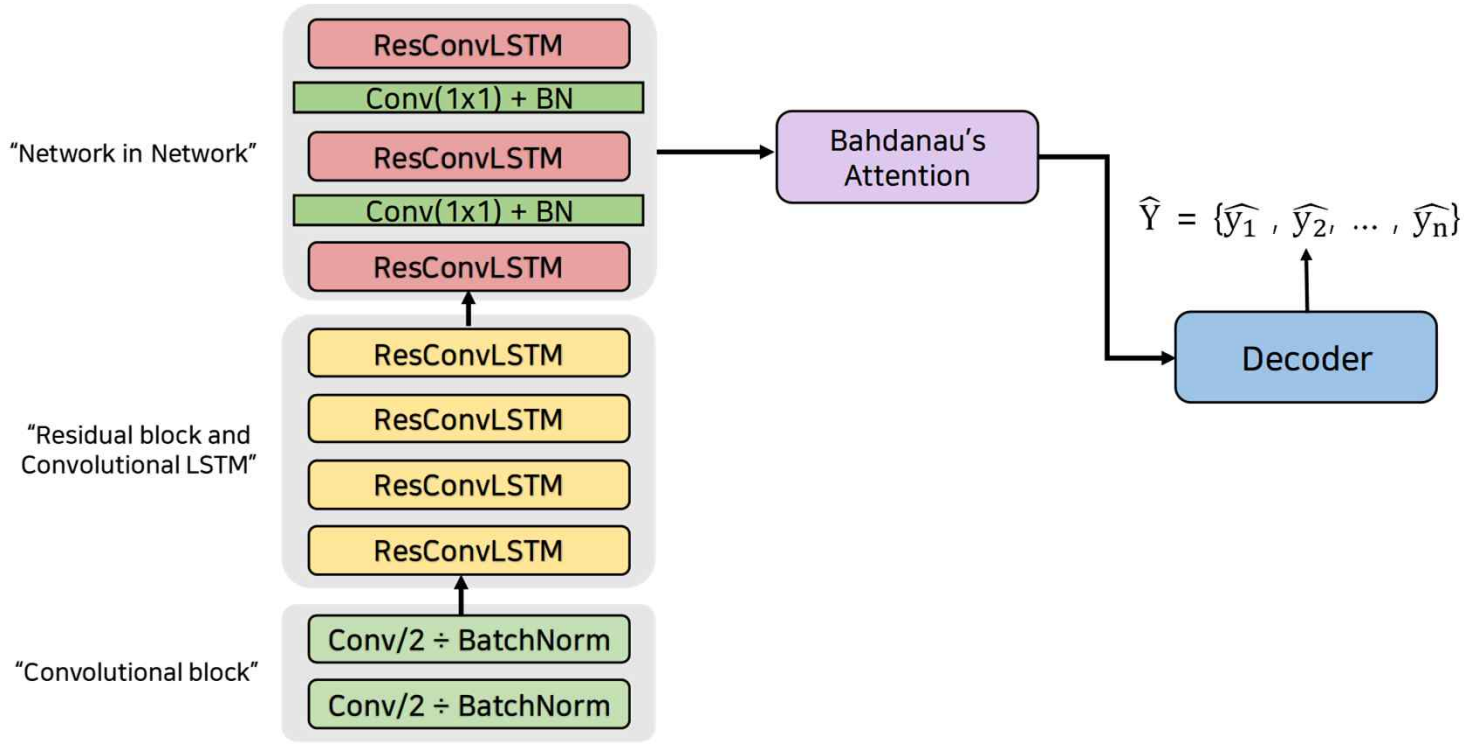
  (신경망 기반 기계번역 분야에서 사용되는 backtranslation 응용)

- Spelling Correction model, RNN-LM, Recognition model과 shallow fusion

Review

- ## Baseline recognition model

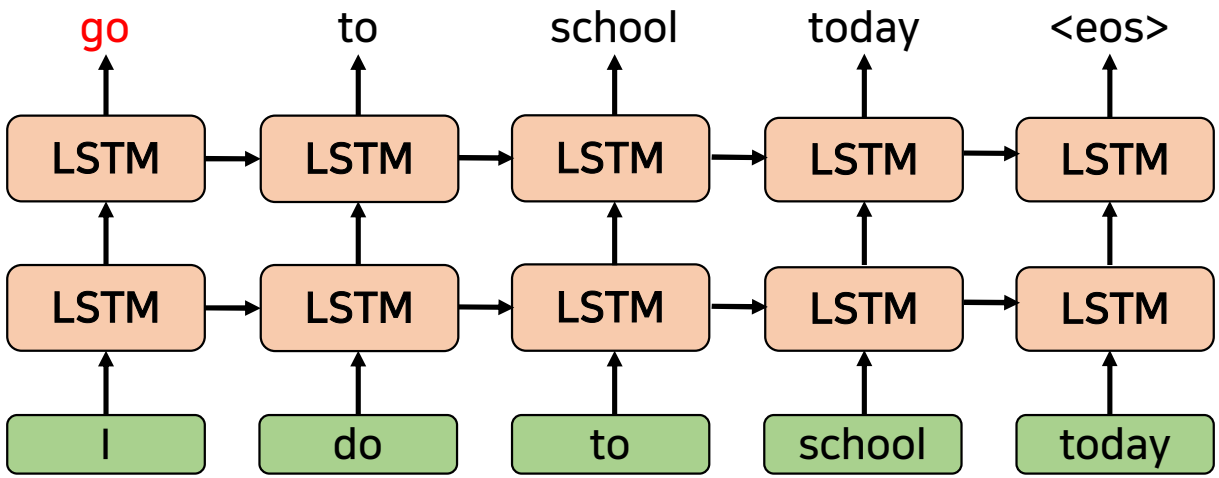
(Very deep convolutional networks for end-to-end speech recognition)



Review

▪ Methodology(External LM)

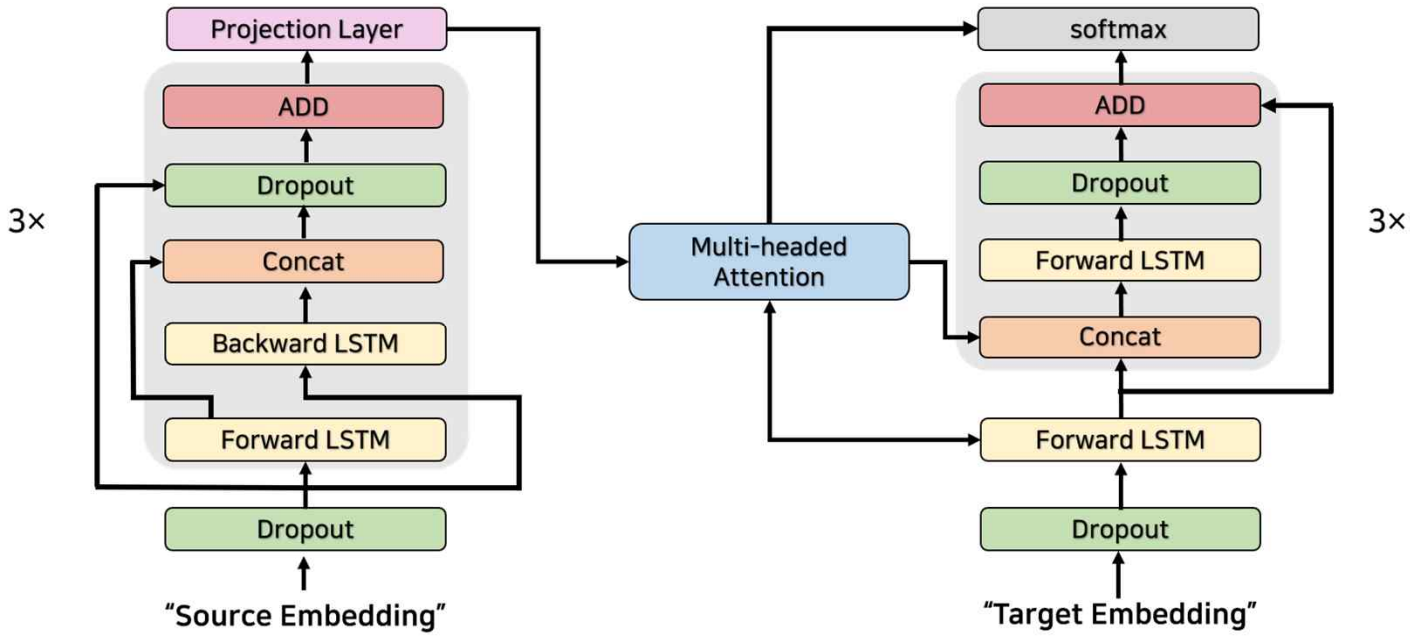
- 학습된 RNN-LM과 shallow fusion수행
- 외부 Language model을 통해 가장 높은 확률의 토큰 sequence 생성



Review

▪ Methodology(Spelling Correction)

- Recognition Model에서 생성된 오류를 참조하기 위한 모델 학습



Review

▪ Methodology(Spelling Correction)

- Spell Correction model의 특징

  ✓ Residual Connection

  ✓ Layer Normalization Per-gate

  ✓ Multi-head additive attention

  ✓ LAS model의 결과 n-best lists를 모두 학습 데이터로 이용해 LAS 모델의 오류 분포를 충분히 고려

Review

Inference

- 최적의 결과 생성을 위한 수식

$$A^* = \underset{A}{\operatorname{argmax}} \lambda_{LAS} * p_i + \lambda_{SC} * \mathrm{q}_{ij} + \lambda_{LM} * r_{ij}$$

- A^* : 최적의 결과
- λ_{LAS} , λ_{SC} , λ_{LM} : LAS, SC, LM에서 생성된 결과를 고려하기 위한 각각의 가중치
- p_i : LAS 통해 생성된 단위 token의 확률 분포
- q_{ij} : Spelling Correction을 통해 생성된 단위 token의 확률 분포
- r_{ij} : Language Model을 통해 생성된 단위 token의 확률 분포

Review

- **실험 데이터 및 결과**
  - LibriSpeech 960 hours["Clean"]
  - STT-TTS-STT (backtranslation) 에서 multi-style training을 통해 데이터 증폭
  - STT에서 beam size를 8로 지정

System	Dev-clean	Test-clean
LAS	5.80	6.03
LAS→LM(8)	4.56	4.72
LAS-TTS	5.68	5.85
LAS-TTS→LM(8)	4.45	4.52
LAS→SC(1)	5.04	5.08
LAS→SC(8) →LM(64)	4.20	4.33
LAS→SC-MTR(1)	4.87	4.91
LAS→SC-MTF(8) →LM(64)	4.12	4.2

Review

- 구현
  - 데이터
    1. BBC 뉴스 데이터
    2. Kaggle 뉴스 기사 데이터
  - Hyper-Parameter
    1. Optimizer : Adam
    2. Learning rate : warm up
    3. Dropout : 0.2
    4. Embedding : Glove(256)

논문의 한계

▪ 논문의 한계점

- LSTM기반의 모델을 겹겹이 쌓은 구조
  - ➢ 학습 시간 및 처리시간↑
- 복잡한 모델 구조로 실제로 상용화 하거나 모델을 학습시키는데 제한적
  - ➢ 학습데이터↑
- 다양한 음성인식 모델의 오류를 학습하지 못하는 Spelling Correction Model
  - ➢ 음성인식 모델의 오류만 고려

향후 연구계획

▪ 향후 연구계획

- Transformer 기반의 spell correction 모델 생성
- 한국어와 같은 low resource 언어에서도 학습이 가능한 비교적 간단한 형태의 모델
- 음성신호의 특징이 고려되어 다양한 음성인식 모델에서 적용이 가능한 모델 생성

THANK YOU