한국어 음성인식 후처리를 위한 주의집중 기반의 멀티모달 모델*)

정영석[1,O], 오병두[2], 허탁성[2], 최정명[2], 김유섭[1]

한림대학교 소프트웨어융합대학[1]
한림대학교 융합소프트웨어학과[2]
{dnfkdi1995, iambd822, gjxkrtjd221, jeong5905}@gmail.com, yskim@hallym.ac.kr

Attention based multimodal model for Korean speech recognition post-editing

Yeong-Seok Jeong[1,O], Byoung-Doo Oh[2], Tak-Sung Heo[2], Jeong-Myeong Choi[2], Yu-Seop Kim[1]
Hallym University, College of Software[1]
Hallym University, Department of Convergence Software[2]

요 약

최근 음성인식 분야에서 신경망 기반의 종단간 모델이 제안되고 있다. 해당 모델들은 음성을 직접 입력받아 전사된 문장을 생성한다. 음성을 직접 입력받는 모델의 특성상 데이터의 품질이 모델의 성능에 많은 영향을 준다. 본 논문에서는 이러한 종단간 모델의 문제점을 해결하고자 음성인식 결과를 후처리하기 위한 멀티모달 기반 모델을 제안한다. 제안 모델은 음성과 전사된 문장을 입력 받는다. 입력된 각각의 데이터는 Encoder를 통해 자질을 추출하고 주의집중 메커니즘을 통해 Decoder로 추출된 정보를 전달한다. Decoder에서는 전달받은 주의집중 메커니즘의 결과를 바탕으로 후처리된 토큰을 생성한다. 본 논문에서는 후처리 모델의 성능을 평가하기 위해 word error rate를 사용했으며, 실험결과 Google cloud speech to text모델에 비해 word error rate가 8% 감소한 것을 확인했다.

주제어: 음성인식, 후처리, 주의집중 메커니즘, 멀티모달

1. 서론

음성인식이란, 발화자의 음성을 입력 받아 이를 컴퓨터가 인식해 텍스트 형태로 변환하는 것을 말한다[1]. 고전적 음성인식 모델은 Hidden Markov model(HMM)과 여러 종류의 모델(음성모델, 발음모델, 언어모델)로 구성되어 있었다 [2,3]. 최근에는 많은 연구에서 신경망 기반 모델이 제안되고 있다 [4]. 이러한 신경망 기반 모델은 기존의 HMM 기반의 모델에 비해 큰 성능 향상을 이루었다. 또한 하나의 모델을 통해 음성 형태의 입력을 텍스트로 전사하는 종단 간 모델을 구성하는 것이 가능 해졌다.

최근 제안되는 신경망 기반의 음성인식 모델은 하나의 모델을 통해 입력된 음성신호를 텍스트로 전사한다. 이러한 신경망 기반의 음성인식 모델은 기존 모델에 비해 성능 향상을 이루었다. 하지만 종단간 모델은 음성 형태의 데이터를 직접 입력 받기 때문에, 입력된 데이터의 품질이 결과에 많은 영향을 미친다 [5]. 이를 보완하기 위해 음성인식 모델의 결과값에 대한 후처리 과정이 필요하다.

본 논문에서는 음성인식 모델의 결과인 전사된 문장을 후처리하기 위한 모델을 제안한다. 제안 모델은 음성신호와 음성인식 모델의 결과인 문장을 입력 받는다. 각각의 입력을 구분하여 처리하기 위해 제안 모델은 2개의 Encoder를 통해 문장과 음성의 자질을 추출한다. Encoder는 Bidirectional Long Short-term Memory(Bi-LSTM)와 레이어 정규화(Layer Normalization)로 구성되어 있다. 각각의 Encoder를 통해 추출된 문장과 음성의 자질은 Decoder의 LSTM의 장기 의존성 문제(Long-term-Dependency)를 보완하기 위해 주의집중 메커니즘(Attention Mechanism)을 수행한다 [6,7]. Decoder에서는 Attention Mechanism을 통해 전달받은 값을 사용해 후처리 된 토큰을 생성한다.

본 논문은 6장으로 구성되어 있다. 2장에서는 본 논문과 관련된 연구를 설명한다. 3장에서는 제안 모델의 구조 및 신경망을 설명한다. 4장에서는 본 논문에서 사용한 데이터와 실험을 위한 전처리 과정을 작성한다. 5장에서는 실험에 따른 결과를 분석한다. 6장에서는 본 연구를 요약하며 향후 연구에 대해 작성한다.

2. 관련 연구

본 논문에서는 음성인식의 결과인 전사된 문장과 음성신호를 사용해 전사된 문장을 후처리한다. 본 장에서는 제안모델과 관련된 Attention Mechanism 기반의 텍스트

*) 이 연구는 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행되었음.(NRF-2019S1A5A2A03052093)

교정 모델과 멀티모달 기반의 음성인식 모델에 대한 연구를 설명한다.

2.1. Attention based text correction model

순환신경망 기반의 시퀀스-투-시퀀스(Sequence-to-Sequence) 모델은 문장 교정, 기계번역 등 다양한 분야에서 사용되고 있다. 하지만 순환신경망 기반의 Sequence-to-Sequence 모델은 Long-term-Dependency 문제가 있다 [7].

[8] 연구에서는 해당 문제를 보완하기 위해 [6]연구에서 제안된 Attention Mechanism을 사용해 중국어 문장을 교정했다. 해당 연구에서는 입력된 문장을 음절 단위로 토큰화 하여 Encoder에 입력했다. Encoder에서는 음절단위로 토큰화 된 문장에서 자질을 추출했다. Encoder에서 추출된 자질은 Decoder의 은닉 값과 Attention Mechanism을 수행했다. 이를 통해 Encoder에 입력된 자질의 전체 의미를 나타내는 context vector를 생성했다. Decoder에서는 context vector를 통해 전달받은 값을 사용해 교정된 텍스트를 생성했다.

2.2. Multimodal based recognition model

일반적인 종단간 음성인식 모델의 경우 음성신호만 고려해 텍스트로 전사한다 [1,2,4]. 최근에는 모델의 성능 향상을 위해 사람의 입 모양 등을 모델에 삽입하는 멀티모달 기반의 음성인식 모델에 대한 연구가 증가하고 있다 [9,10].

[10]의 연구에서는 음성 인식의 성능 향상을 위해 음성(A)과 영상(V)을 고려한 멀티모달 기반 모델을 제안했다. [10]의 모델은 합성곱 신경망(CNN) 기반 모델을 통해 영상에서 자질을 추출했다. 또한 음성과 영상의 자질을 함께 고려하기 위해 시각적 적응 훈련(visual adaptive training)을 수행했다. 이 결합 결과를 Bi-LSTM에 입력 후 분석하여 각 시점의 토큰을 생성했다. [10]의 음성 인식 과정은 아래와 같다.

$$V' = VAT(CNN(V), W_{vat}) \quad (1)$$

$$O = BiLSTM(V' + A) \quad (2)$$

식 (1)에서 CNN 기반 모델을 통해 자질을 추출한 V는 차원축소를 위해 가중치 행렬 W_{vat}와 내적을 수행한다. 식 (2)에서 내적을 수행한 결과 V'는 음성신호와의 결합을 위해 A와 더해진다. 해당 결과는 Bi-LSTM을 통해 결과 텍스트를 예측한다.

3. Attention 기반 멀티모달 후처리 모델

본 연구에서는 음성인식 후처리를 수행하기 위해 Encoder-Decoder 구조의 멀티모달 모델을 제안한다. 이때, Encoder에서 Decoder로 자질을 전달하기 위해 Attention Mechanism을 사용한다. 그림 1은 본 논문에서 제안하는 한국어 음성인식 후처리를 위한 주의집중 기반 멀티모달 후처리 모델의 구조이다.

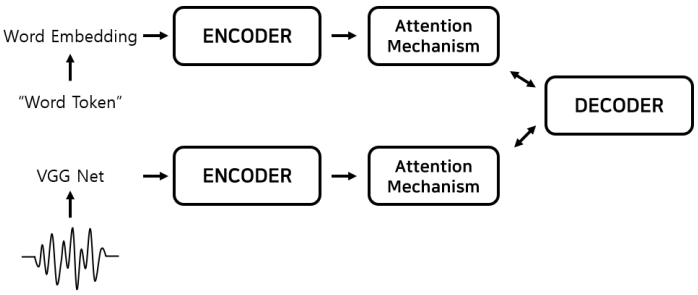

그림 1. Attention 기반 멀티모달 후처리 모델 구조

3.1. VGG Net

음성인식에서 VGG Net은 spectrogram 형태의 음성신호에서 자질을 추출하기 위해 사용된다 [10,11]. VGG Net은 CNN과 Max Pooling으로 구성된 Network 구조이다 [12]. 본 연구에서는 입력된 음성신호(A)의 자질을 VGG Net의 CNN을 통해 추출하며, 차원축소를 수행하기 위해 Max Pooling을 사용한다. VGG Net을 통한 자질(A')은 차원 축소를 수행하기 위해 1차원으로 변환한다. 그리고 투영 층(Projection Layer)을 통해 차원을 축소한다. 이를 표현한 수식은 아래와 같다.

$$A' = VGGNet(A) \quad (3)$$

$$A'_{flat} = Flat(A') \quad (4)$$

$$P = W_{proj}A'_{flat} + b_{proj} \quad (5)$$

식 (6)은 spectrogram 형태의 입력 A에 대해 VGG Net을 통해 자질을 추출하는 과정이다. 추출된 자질은 A'로 표현된다. 식 (7)은 1차원으로 변환한 결과는 A'_{flat}으로 표현된다. 식 (8)에서 A'_{flat}은 W_{proj}을 통해 투영 연산을 수행한다. 이때, $W_{proj} \in R^{a \times d}$이며 a는 1차원 형태로 변환한 A'_{flat}의 차원수 d는 투영 연산을 통해 축소된 차원수를 의미한다.

3.2. Encoder

본 논문에서 제안하는 모델은 음성과 문장 각각을 입력 받는다. 입력된 각각의 데이터의 자질을 구분하기 위해 2개의 Encoder를 사용했다. 각각의 Encoder는 임베딩[1] 된 단어 벡터(E)와 음성신호에서 자질을 추출한 VGG Net의 결과(P)를 입력 받는다.

각각의 Encoder는 모두 Bi-LSTM과 Layer Normalization 층으로 구성되어 있다. LSTM은 순환 신경

[1]https://keras.io/search.html?q=embedding

망 기반 구조로 장기의존성 문제를 보완한다 [13]. 본 논문에서는 Encoder에 입력된 값의 시간적 양방향성을 고려하여 자질을 추출하기 위해 Bi-LSTM을 적용했다 [15]. 또한 빠른 학습과 내부 공변량 변화(internal covariate shift) 문제를 줄이기 위해 Bi-LSTM의 결과값에 Layer Normalization를 수행했다 [16]. 아래는 각각의 입력을 처리하기 위한 수식표현이다.

$$H_{voice} = Bidirection\left(LSTM\left(P, h_p\right)\right) \quad (6)$$

$$H_{word} = Bidirection\left(LSTM\left(E, h_E\right)\right) \quad (7)$$

Encoder에서 추출된 음성신호와 단어 벡터 H_{voice} , H_{word}는 Attention Mechanism을 통해 Decoder로 매 예측 시점마다 전달된다.

3.3. Attention Mechanism

본 연구에서는 Decoder의 예측 과정에서 생기는 Long-term-dependency 문제를 보완하기 위해 Attention Mechanism을 사용한다 [6]. Attention Mechanism은 Encoder의 입력 데이터에 대한 토큰별 추출된 자질 ($H_{enc} \in \{H_{voice}, H_{word}\}$)과 Decoder의 $t-1$ 번째 은닉 상태 ($\bar{h}^{t-1}$)의 관계 연산을 통해 t번째 시점의 참조 값을 계산한다. Attention Mechanism은 아래와 같이 표현된다.

$$score\left(h_{enc}^{t}, \bar{h}^{t-1}\right) = v_a^T tan\, h\left(W_1 h_{enc}^{t} + W_2 \bar{h}^{t-1}\right) \quad (8)$$

$$a_t = \frac{\exp\left(score\left(h_{enc}^{t}, \bar{h}^{t-1}\right)\right)}{\sum_{s=1}^{S} \exp\left(score(h_{enc}^{s}, \bar{h}^{t-1}\right)} \quad (9)$$

$$c_t = \sum_s a_{ts} \bar{h}_s \quad (10)$$

식 (8)에서 score 함수의 결과는 H_{enc}와 $\bar{h}^{t-1}$의 관계된 정도를 나타낸다. (9) 식의 a_t는 계산된 h_{enc}에 대한 각 토큰간 관계의 정도를 0과 1사이의 값으로 표현하기 위한 softmax 수식이다. 식 (10)의 c_t는 Decoder로 전달되는 문맥벡터(context vector)이다. 이는 t 시점의 관점으로 계산한 Encoder의 함축된 정보를 의미를 나타낸다. 본 연구에서는 음성신호에서 추출된 자질 H_{voice}와 문장을 통해 생성된 자질 H_{word} 에 대해 각각 구분하여 Attention Mechanism을 사용한다. 이를 통해 각각을 표현하는 context vector(c_t^{voice}, c_t^{text})를 생성한다.

3.4. Decoder

본 논문에서 Decoder는 Attention Mechanism의 결과인 문맥 벡터(c_t)와 입력된 단어 토큰을 사용해 문장을 생성한다. 먼저, Decoder는 문장 생성을 위해 이전 시점의 word token(w_{t-1})을 입력 받는다. 해당 토큰을 밀집표현(dense representation)으로 변환하기 위해 단어 임베딩을 수행한다. 임베딩 된 토큰은 Attention Mechanism의 결과(c_t^{voice}, c_t^{test})와 결합된다. 결합된 토큰($Concat_t$) 은 생성된 문장의 토큰별 어순을 고려하기 위해 순환신경망 중 하나인 LSTM을 사용해 자질을 추출한다. LSTM에 의해 추출된 자질(L_t)은 다시 한번 Encoder의 정보를 참조하기 위해 주의 집중 메커니즘의 각 결과와 더한다. 그리고 해당 결과를 완전 연결 층(fully connected layer)에 입력한다. 완전 연결 층을 통해 Decoder는 해당 시점의 후처리 된 토큰을 생성한다. Decoder의 토큰 예측과정은 아래와 같이 표현된다.

$$e_t = Embedding(w_{t-1}) \quad (11)$$

$$Concat_t = Concatenate\left(e_t, c_t^{voice}, c_t^{text}\right) \quad (12)$$

$$L_t = LSTM(Concat_t) \quad (13)$$

$$O_t = FC\left(\left(L_t + c_t^{voice} + c_t^{text}\right), W_{fc}\right) \quad (14)$$

식 (11)은 Decoder에 입력된 토큰의 임베딩 과정을 표현한다. 식 (12)는 Attention Mechanism의 결과와 임베딩된 결과를 결합하는 과정을 표현한다. 식 (13)은 LSTM을 통해 $Concat_t$에서 벡터를 추출하는 과정을 의미한다. 식 (14)는 완전 연결 층을 통해 최종적인 토큰을 예측하는 과정을 의미한다. 해당 식에서 W_{fc}는 가중치 행렬을 의미한다. 그림 2는 본 논문에서 제안한 모델의 Decoder 구조이다.

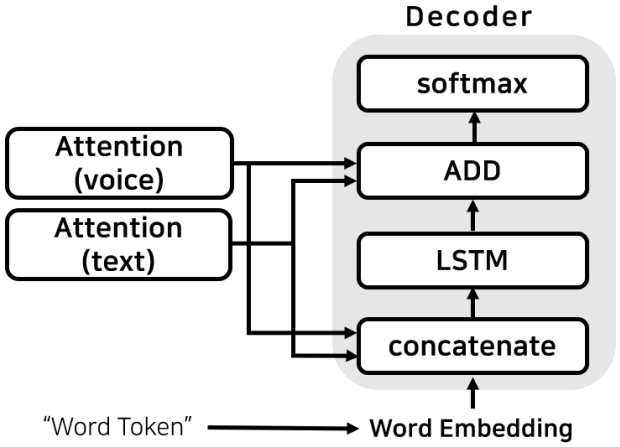

그림 2. 제안 모델의 Decoder 구조

4. 데이터

본 논문은 Naver ClovaCall Dataset[2]을 사용해 실험을 진행했다. 해당 데이터셋은 전화 기반의 한국어 음성 데이터셋으로 예약, 문의와 같은 간단한 문장으로 구성되어 있다. 총 61,287개의 음성-문장 쌍으로 구성되어 있다.

본 논문의 제안 모델은 음성인식의 모델의 결과인 전

[2] https://github.com/clovaai/ClovaCall

사된 문장과 음성 데이터를 모두 입력받는 멀티모달 모델이다. 따라서 전사된 문장과 음성 데이터 각각을 구분하여 전처리를 수행했다.

표 1. STT결과 및 정답 문장 비교

STT 결과	정답
고통받는 시간이 언제인가요	***보통 닫는*** 시간이 언제인가요
투움바파스타 하나랑 ***독립*** 하나 주세요	투움바 파스타 하나랑 ***폭립***하나 주세요
혹시 ***상가 번으로*** 가능할까요	***네*** 혹시 ***창가 자리로*** 가능할까요

첫 번째, Encoder에 입력되는 문장은 Google cloud Speech to Text[3](STT)를 통해 전사된 문장을 사용했다. 전사된 문장 가운데 3음절 이하의 문장은 음성인식 모델이 제대로 작동되지 않은 것으로 간주해 제거했다. 이후, 생성된 각 문장을 한국어 의미의 최소 단위인 형태소 단위로 토큰화 하였다. 표 1은 STT의 결과와 정답 문장을 비교한 표이다.

두 번째, Encoder에 입력되는 음성 데이터는 음성 없이 노이즈만 있는 부분을 제거했다. 다음으로 음성 데이터를 Mel Spectrogram으로 변환했다. 변환된 Mel Spectrogram은 46ns의 길이로 구분하여 잘랐다. 이때 음절의 음운을 고려하기 위해 23ns씩 중복하였다.

5. 실험 및 분석

5.1. 하이퍼 파라미터

표 2. 하이퍼 파라미터 설정

Hyper-Parameter	Setting
Embedding dimension	300
Projection Layer dimension	256
Bi-LSTM dimension	128
Layer Normalization epsilon	10^{-3}
LSTM dimension (Decoder)	256
LSTM dropout rate	0.2
optimizer	Adam
Learning rate	10^{-3}

표 2는 본 연구를 진행할 때 적용한 하이퍼 파라미터이다. 본 실험의 진행에서 각 단어를 표현하기 위한 벡터는 균일 분포로 초기화 했다. 그리고 학습 과정에서 각 단어의 벡터 값을 갱신했다. VGG Net의 구조는 [13] 연구의 13 weight layers와 동일하다.

[3] https://cloud.google.com/speech-to-text?hl=ko

5.2. 결과 분석

본 연구에서 제안한 멀티모달 기반 모델과 기존 텍스트 기반 모델과 성능을 비교하기 위해 음성인식 분야에서 성능평가에 사용되는 word error rate(WER)를 사용했다. WER은 아래 수식과 같이 표현된다.

$$WER = \frac{S + D + I}{N} \times 100\ (\%) \qquad (15)$$

식 (15)에서 S 는 교체할 단어, D 는 삭제할 단어, I 는 추가할 단어, N 은 전체 단어의 수를 의미한다. WER 은 오류가 포함된 문장에서 정답 문장을 생성하기 위한 연산횟수를 수치화 한 성능평가 방법이다.

본 실험에서는 멀티모달 기반 모델과 텍스트 기반 모델의 성능을 비교한다. 이를 위해 기계번역에서 활용되는 LSTM Encoder-Decoder, Transformer, Attention 기반의 sequence-to-sequence모델과 성능을 비교했다. 표 3은 성능을 비교한 결과를 나타낸 표이다.

표 3. 제안 모델과 관련연구 모델의 WER비교

Model	WER (%)
STT	15.9%
LSTM Encoder-Decoder [17]	36.17%
Transformer [18]	21.76%
Text based Attention Mechanism [6]	8.05%
LSTM Encoder-Decoder + VAT	18.75%
Multimodal based Transformer	20.08
Proposed Model	**7.77%**

먼저 LSTM 기반의 Encoder-Decoder 모델은 Encoder의 LSTM이 추출한 자질을 사용해 Decoder에서 전체 문장을 생성한다. 이러한 모델의 경우 Long-term-Dependency가 있어 전체 문장을 생성하는데 어려움을 겪는다 [7].

Transformer 기반 모델은 학습에 많은 데이터를 필요로 한다 [19]. 본 연구에서는 비교적 적은 데이터를 사용해 실험을 진행했다. 따라서 모델이 요구하는 충분한 학습데이터를 제공하지 못해 STT의 결과에서 후처리를 제대로 수행하지 못하는 것을 확인할 수 있다.

다음으로 멀티모달 모델의 효율성을 검증하기 위한 구조를 표 3을 통해 비교와 분석을 한다. 표 3의 LSTM Encoder-Decoder + VAT는 [17] 연구에서 제안된 모델을 기반으로 하여 VGG Net을 통해 자질을 추출한 음성신호와 텍스트를 입력 받는다. 음성신호와 텍스트에서 추출된 자질은 VAT 모듈을 통해 각 자질을 통합해 Decoder에 전달한다. Decoder는 전달받은 통합된 자질을 이용해 후처리된 문장을 생성한다. 음성신호와 텍스트 자질을 함께 입력 받는 해당 모델은 텍스트에 기반한 [17] 연구의

제안 모델보다 성능이 개선되었다.

Transformer 모델에서 멀티모달 모델의 성능을 비교했다. 표 3의 Multimodal based Transformer의 구조는 [20]의 모델 구조를 사용해 실험을 진행했다. 제안 모델과 동일하게 VGG Net을 사용해 음성 정보에서 자질을 추출해 모델에 입력했다. 실험 결과 음성과 텍스트 정보를 모두 고려한 멀티모달 기반 모델의 성능이 텍스트 기반 모델인 [18] 연구의 모델에 비해 성능이 개선된 것을 확인할 수 있었다.

제안 모델은 입력된 문장과 음성정보를 바탕으로 수행한 Attention Mechanism을 사용해 후처리된 결과를 생성한다. 따라서 제안 모델은 다양한 자질을 고려해 결과를 생성할 수 있어 텍스트 기반 모델인 [6]에 비해 성능이 향상되었다. 또한 텍스트 기반 모델 [6,17,18]은 음성인식 모델이 전사한 문장만 고려한 모델은 음성인식 모델 자체의 성능에 의존적이다. 하지만 제안 모델은 음성의 자질도 고려해 후처리를 진행한다. 따라서 음성인식 모델의 결과에 비교적 의존성이 떨어지는 장점도 존재한다.

6. 결론

본 연구에서는 기존 문장 대 문장 모델에서 사용되는 주의 집중 메커니즘을 사용해 모델을 구성했다. 또한 모델에 음성정보를 고려하기 위해 음성과 텍스트를 동시에 입력 받는 멀티모달 모델을 제안하였다. 실험결과 텍스트만 입력 받는 기존 모델에 비해 성능이 향상된 것을 확인할 수 있었다.

본 연구는 비교적 적은 데이터를 사용해 실험을 진행했다. 향후 연구에서는 대용량 데이터를 사용해 기존 모델에 비해 복잡한 구조의 모델을 구성함으로써 성능을 개선할 것이다.

참고문헌

[1] Laurent Besacier, Etienne Barnard, Alexey Karpov, Tanja Schultz, Automatic speech recognition for under resourced lanugages: A survey, Speech Communication, pp. 85-100, 2014

[2] Yoshua Bengio, Renato De Mori, Giovanni Flammia, Ralf Kompe, Global optimization of a neural network hidden markov model hybrid, IJCNN-91-Seattle IEEE, 789-794, 1991

[3] Ali Yazgan, Murat Saraclar, Hybrid language models for out of vocabulary word detection in largy vocabulary conversational speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 745-748, 2004

[4] Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, ICASSP, pp.4960-4964, 2016

[5] 권세도, 정홍, 지능로봇에 적합한 잡음 환경에서의 원거리 음성인식 전처리 시스템. 대한전자공학회 학술대회, pp. 671-672, 2006.

[6] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio. End-to-End Attention-based Large Vocabulary Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4960-4964, 2016

[7] Bengio, Yoshua, Patrice Simard, Paolo Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks, pp.157-166, 1994

[8] Jianyong Duan, Yang Yuan, Hao Wang, Xiaopeng Wei, Zheng Tan, Research on Chinese Text Error Correction Based on Sequence Model, International Conference on Asian Language Processing (IALP), 2019

[9] Di Hu, Xuelong Li, Xiaoqiang Lu, Temporal Multimodal Learning in Audiovisual Speech Recognition. CVPR, pp.3574-3582, 2016

[10] Shurti Palaskar, Ramon Sanabria, Florian Metze, End-to-end Multimodal Speech Recognition. ICASSP, pp. 5774-5778, 2018

[11] Takaaki Hori, Shinji Watanabe, Yu Zhang, William Chan. Adbances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM. INTERSPEECH, pp.265-271, 2017

[12] William Hartmann, Roger Hsiao, Tim Ng, Jeff Ma, Francis Keith, Man-Hung Siu, Improved Single System Conversational Telephone Speech Recognition with VGG Bottleneck Features, INTERSPEECH, pp.112-116, 2017

[13] Simonyan, K, Zisserman, A. Very deep convolutional networks for large-scale image recognition, ICLR, 2015

[14] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-term Memory. Neural Computation.pp.1735-1780, 1997

[15] Xuezhe Ma, Eduard Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. ACL, pp.1064-1074, 2016

[16] Lei Jimmy Ba, Ryan Kiros Geoffrey E.Hinton, Layer Normalization, CoRR, abs/1607.06450, https://arxiv.org/abs/1607.06450, 2016

[17] llya Sutskever, Oriol Vinyals, Quoc V.Le. Sequence to Sequence Learning with Neural Networks. NIPS Proceedings of the 27th International Conference on Neural Information Processing System.pp.3104-3112, 2014

[18] 15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, Neural Information Processing Systems (NIPS), pp.6000-6010, 2017

[19] Martin Popel, Ondrej Bojar. Training Tips for the Transformer Model, Prague Bulletin of Mathematical Linguistics, pp.43-47, 2018
[20] Jaehun shin, Jong-hyeok Lee. Multi-encoder Transformer Network for Automatic Post-Editing, *ACL*, pp.840-845, 2019