CNN과 주의기반LSTM망을 리용한 조선어시각소인식방법

리광철, 리윤미

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《오늘 세계는 경제의 지식화에로 전환되고있으며 우리앞에는 나라의 경제를 지식의 힘으로 장성하는 경제로 일신시켜야 할 시대적과업이 나서고있습니다.》

숨은마르꼬브모형에 기초한 입놀림해득방법은 특징추출기에 의하여 얻어진 입놀림특징을 숨은마르꼬브모형이나 중첩신경망을 리용하여 인식하는 방법으로서 이 방법에서는 특징추출문제와 모형의 로바스트성제고문제와 같은 많은 문제들이 제기되고 그 정확성[1, 2]이 높지 못하다.

론문에서는 특징추출기로서의 CNN과 시계렬처리기로서의 주의에 기초한 LSTM을 리용하여 조선어시각소인식의 정확도를 개선하기 위한 한가지 방법을 제안하였다.

1. 조선어시각소인식을 위한 심층신경망

1) 전처리

입력으로서는 하나의 시각소에 해당한 입놀림화상렬이며 출력은 입력에 대한 시각소류형이다.

매 시각소는 동적인 특성과 정적인 특성을 다 가지는데 동적인 특성이라고 할 때에는 하나의 시각소가 일정한 시간지연을 가지고 매 시각에 대한 흐레임들사이의 시간적련관성을 가진다는것이며 정적인 특성을 가진다는것은 시각소마다 그 시각소를 규정하는 열쇠흐레임을 가진다는것이다.

일반적으로 시각소마다 시간길이 즉 흐레임렬의 길이는 서로 다르기때문에 론문에서는 전체 시간구간을 10개의 구간으로 나누고 매 부분구간에서 하나의 흐레임을 우연적으로 선택하였다. 그리하여 임의의 시각소에 대한 흐레임렬의 길이를 10으로 고정시키고 매 흐레임에서 입술령역만을 따내고 입술부분령역의 크기를 112×112크기로 정규화하였다.

2) 시각소인식을 위한 심층신경망의 구조

론문에서 제안한 시각소인식을 위한 심층신경망의 구조를 그림 1에 보여주었다.

그림 1에서 $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_n$ 은 n개의 입놀림화상렬로부터 CNN을 통하여 추출한 공간특징벡토르이다.

시각소인식을 위한 심층신경망의 입력은 입놀림화상렬의 매 흐레임에서 입술령역만을 따내여 얻어진 입놀림부분령역들의 렬이다.

망에서 CNN은 입력된 입령역화상렬에서 특징들을 추출하며 주의기구를 가진 LSTM은 시계렬정보와 주의무게들을 학습한다.

마지막으로 512차원특징이 전결합층을 거쳐 사영되며 softmax층을 통하여 최종적으

로 시각소류형이 결정된다. 여기서 CNN은 부호화기로 리용되고 LSTM은 복호화기로 리용된다.

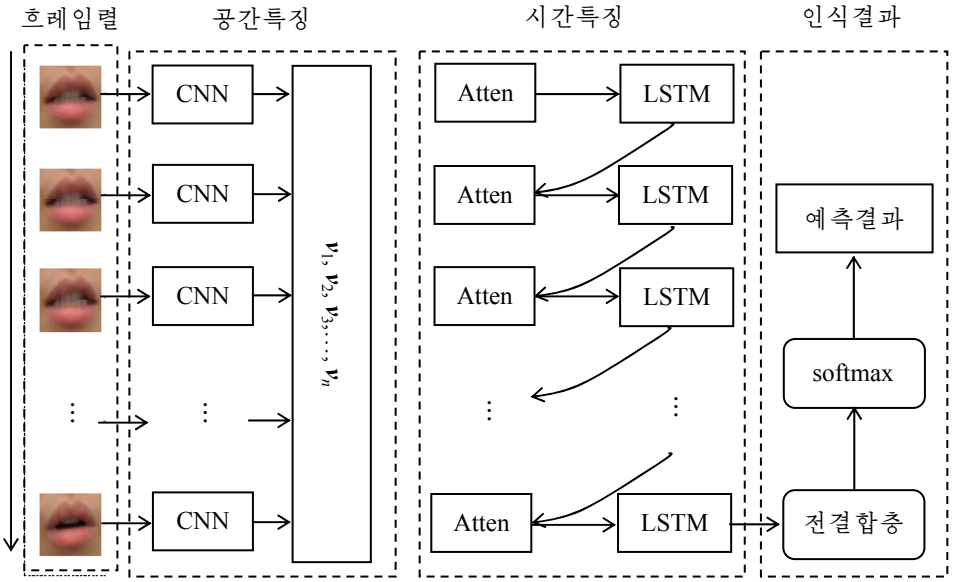

그림 1. 론문에서 제안한 시각소인식을 위한 심층신경망의 구조

복호화단계에서 단순한 LSTM망이 아니라 주의기구를 도입하여 주의무게값(α)들을 학습한다.

따라서 모형은 흐레임렬에서 흐레임들사이의 상관성을 학습하면서도 더 유효한 령역(흐레임)에 더 많은 주의를 집중하도록 한다.

주의기반LSTM망의 구조를 그림 2에 보여주었다.

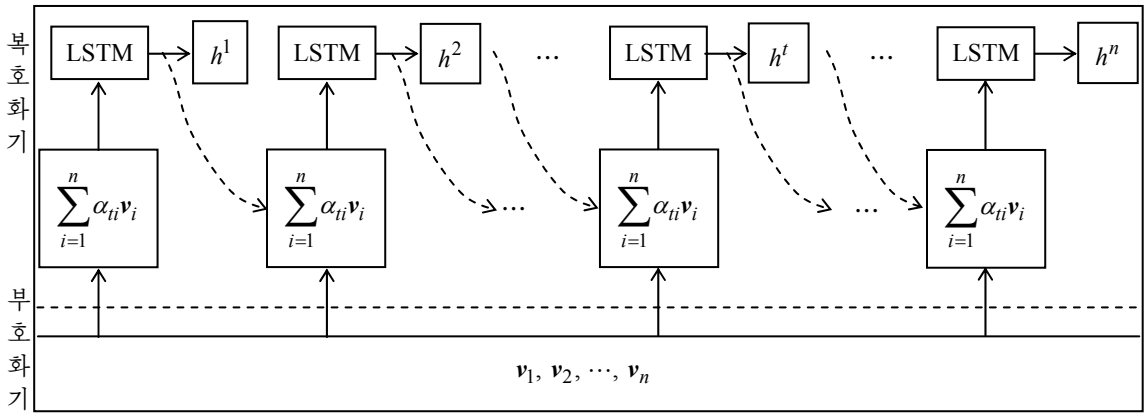

그림 2. 주의기반LSTM망의 구조

LSTM망의 입력은 다음과 같다.

$$\varphi(V) = \sum_{i=1}^{n} \alpha_{ti} v_i \tag{1}$$

여기서 $\boldsymbol{v}_i$는 i번째 흐레임의 특징벡토르로서 CNN에 의하여 추출된 공간특징이다.

t시각에서의 주의무게 α_{ti}는 전시각에서의 LSTM망의 출력과 현재시각의 특징벡토르로부터 식 (2), (3)에 의하여 결정된다.

$$e_{ti} = \tanh(\boldsymbol{W} \times h_{t-1} + \boldsymbol{U} \times \boldsymbol{v}_i + b) \tag{2}$$

여기서 h_{t-1}은 $t-1$시각에서의 LSTM망의 출력, $\boldsymbol{v}_i$는 현재시각의 특징벡토르, $\boldsymbol{W}$, $\boldsymbol{U}$, b는 각각 학습할 무게행렬과 편위파라메터들을 나타낸다.

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{n} \exp(e_{tk})} \tag{3}$$

t시각에 주의기반LSTM망의 출력은 다음과 같다.

$$h_t = f_{rnn}(h_{t-1},\ \varphi(\boldsymbol{V})) \tag{4}$$

여기서 f_{rnn}은 LSTM망을 의미하며 h_{t-1}은 $t-1$시각에서의 LSTM망의 출력이고 $\varphi(\boldsymbol{V})$는 주의무게들을 증가시킨 후에 t시각의 입력이다.

주의기구의 추가는 계산량을 증가시키지만 그것은 선택적으로 동화상에서 효과적인 정보에만 주목하고 비유효한 정보의 간섭을 줄임으로써 망의 성능을 상당히 개선시킨다.

2. 조선어시각소인식실험

조선어시각소인식을 위한 CNN-LSTM망에서 부호화기로 리용되는 CNN은 VGG19모형[3]을 리용한다.

VGG19의 입력은 크기가 112×112인 RGB입술령역화상이며 출력은 4 096차원의 벡토르이다.

따라서 CNN에 의하여 10개의 흐레임에 대한 10개의 4 096차원의 벡토르가 얻어지고 매 시각에 따르는 주의무게에 의하여 결합된 4 096차원의 벡토르 $\varphi(\boldsymbol{V})$가 주의에 기초한 LSTM망에 입력된다.

주의기반LSTM망에서 LSTM층의 수는 1개, 세포수는 512개로 하였다.

전결합층의 세포수는 조선어시각소류형의 개수와 똑같이 12개로 하였다.

CNN과 LSTM을 결합한 모형 CNN-LSTM, CNN과 주의기반LSTM을 결합한 모형 CNN-ATTEN-LSTM에 의한 조선어시각소인식결과를 표에 보여주었다.

표. 두가지 모형에 의한 조선어시각소인식결과

분 류	CNN-LSTM/%	CNN-ATTEN-LSTM/%
학습자료	70	76
비학습자료	58	64

표에서 보는것처럼 부호화기로서 같은 CNN망을 리용하였다고 해도 복호화기로서 주의기반LSTM을 리용할 때가 인식정확도가 더 높다.

맺 는 말

입놀림화상에서 공간특징을 추출하기 위한 CNN과 복호화기로서 주의기반LSTM망을 리용하여 조선어시각소를 인식하기 위한 한가지 방법을 제안하고 실험을 통하여 그 효과성을 검증하였다.

참 고 문 헌

[1] N. Puviarasan, S. Palanivel; Expert Syst. Appl., 38, 4477, 2011.

[2] Yiting Li et al.; IEEE ICIS2016, 26, 2016.

[3] K. Simonyan, A. Zisserman; ICLR2015, 1, 2015.

주체109(2020)년 5월 5일 원고접수

Korean Viseme Recognition Method by CNN and Attention-based LSTM

Ri Kwang Chol, *Ri Yun Mi*

In this paper, we propose a method for improving the Korean viseme recognition accuracy using CNN as encoder, and attention-based LSTM as decoder.

Keywords: lip reading, viseme, CNN, LSTM