MobileNet을 이용한 한국어 입모양 인식 시스템

*이원종 **김주아 ***손서원 †김동호

서울과학기술대학교

*ekdma761@ **jooah04@ ***jessica7654@ †dongho.kim@seoultech.ac.kr

Korean Lip Reading System Using MobileNet

*Won-Jong Lee **Joo-Ah Kim ***Seo-Won Son †Dong Ho Kim

Seoul National University Of Science And Technology

요 약

Lip Reading(독순술(讀脣術)) 이란 입술의 움직임을 보고 상대방이 무슨 말을 하는지 알아내는 기술이다. 본 논문에서는 MBC, SBS 뉴스 클로징 영상에서 쓰이는 문장 10개를 데이터로 사용하고 CNN(Convolutional Neural Network) 아키텍처 중 모바일 기기에서 동작을 목표로 한 MobileNet을 모델로 이용하여 발화자의 입모양을 통해 문장 인식 연구를 진행한 결과를 제시한다. 본 연구는 MobileNet과 LSTM을 활용하여 한국어 입모양을 인식하는데 목적이 있다. 본 연구에서는 뉴스 클로징 영상을 프레임 단위로 잘라 실험 문장 10개를 수집하여 데이터셋(Dataset)을 만들고 발화한 입력 영상으로부터 입술 인식과 검출을 한 후, 전처리 과정을 수행한다. 이후 MobileNet과 LSTM을 이용하여 뉴스 클로징 문장을 발화하는 입모양을 학습 시킨 후 정확도를 알아보는 실험을 진행하였다.

1. 서론

현대 사회는 다수의 청인(聽人)들로 구성되어 있어 사회 시스템이 음성언어 위주로 돌아가고 있다. 청인 중심 사회에서 생존하기 위한 농인들은 상대의 입모양을 읽고 그 내용을 이해하는 독순술(Lip Reading)을 수년간 익혀야 한다. 또한 발음의 인식을 위해서 음성 정보의 분석이 필요하지만, 실제 환경에서 주변 소음으로 인한 잡음 요소가 음성 인식률을 떨어뜨리기 때문에 보다 향상된 인식 결과를 얻기 위해서 영상 정보를 동시에 활용하는 연구뿐만 아니라, 시각적 특성만을 사용해 발음을 인식하는 연구가 제안되고 있다 [1-4].

본 연구에서는 CNN 아키텍처 중 모바일 기기에서 동작을 목표로 한 MobileNet을 모델로 사용하여 한국어 입모양 인식 시스템의 정확도를 알아보는 실험을 진행하였다. 한국어 기반의 Lip-Reading 데이터셋이 존재하지 않았기에 데이터셋을 자체적으로 만들었다. 뚜렷한 입모양과 정확한 발음으로 발화하는 아나운서의 뉴스 클로징 문장 10개를 선정하여 한국어 입모양 인식 시스템의 학습 데이터셋으로 사용하였다.

2. 본론

2.1 데이터셋

한국어 입모양 인식을 위한 한국어 기반의 데이터셋을 자체적으로 만들어 사용하였다. 연구에 사용된 데이터셋은 MBC 뉴스테스크, SBS 8시 뉴스, SBS 나이트 라인의 클로징 문장인 '8시 뉴스 마칩니다.', '뉴스 마치겠습니다.', '오늘 뉴스 마치겠습니다.', '뉴스데스크 마칩니다.', '오늘도 함께해 주신 여러분 고맙습니다.', '시청해 주신 여러분 고맙습니다.', '고맙습니다.', '내일 뵙겠습니다.', '나이트라인 마칩니다.', '행복한 오늘 되십시오.'로 10개의 문장으로 구성하였다.

뚜렷한 입모양과 정확한 발음을 발화하는 데이터가 필요하여 아나운서의 뉴스 클로징 영상을 사용하였다. 뉴스 클로징 영상은 2~3문장으로 이루어져 있어 필요 문장을 데이터로 사용하기 위해 영상을 프레임 단위로 잘라 원하는 문장을 수집하였다. 데이터셋은 10개의 문장에 대해 40회씩 발화하는 영상 총 400개를 사용하였다.

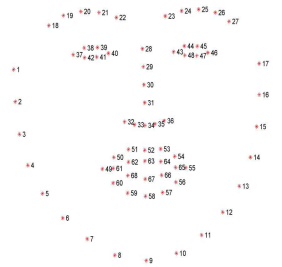
	문장
1	8시 뉴스 마칩니다.
2	뉴스 마치겠습니다.
3	오늘 뉴스 마치겠습니다.
4	뉴스데스크 마칩니다.
5	오늘도 함께해 주신 여러분 고맙습니다.
6	시청해 주신 여러분 고맙습니다.
7	고맙습니다.
8	내일 뵙겠습니다.
9	나이트라인 마칩니다.
10	행복한 오늘 되십시오.

표 1. 한국어 데이터셋으로 활용한 뉴스 클로징 문장

2.2 전처리

연구에 사용될 영상 데이터에서 입술 영역을 인식하고 검출하기 위해 Dlib을 사용하였다 [5]. Dlib은 얼굴에서 68개의 landmark를 찾아 **그림 1**과 같이 나타낸다. 데이터로 사용된 영상에는 아나운서의 얼굴과 수어 통역사의 얼굴이 함께 인식되는 문제가 발생하여 image resize를 사용해 우측 하단의 수어 통역사를 잘라내었다. 그다음, 아나운서의 얼굴을 인식한 후, 얼굴을 Crop 하여 **그림 2**로 나타내었다. Crop된 얼굴

에서 입술 영역인 49 ~ 68번째 landmark를 1:2 비율로 잘라내어 **그림 3**과 같이 사용하였다. 이후 모든 image를 MobileNet의 입력으로 사용하기 위해 numpy 배열로 변환했다.



[그림 1] Facial landmarks

[그림 2] Crop_face

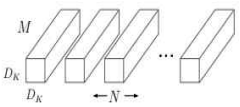
[그림 3] Crop_lip

2.3 MobileNet

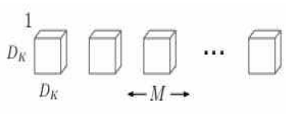
MobileNet은 핸드폰, 임베디드 시스템 등 저용량 메모리 환경에 딥러닝을 적용하기 위해 연산량을 감소시켜 경량화를 진행한 모델이다. 본 연구는 한국어 기반의 Lip-Reading을 모바일 기기에서 작동할 수 있도록 하기 위해 MobileNet 모델을 활용하여 학습하였다. 기존 CNN의 연산량은 **그림 4**와 같다. D_K는 입력값의 크기, M은 입력 channel의 수, N은 출력 channel의 수다. 출력값의 크기가 D_F일 때 연산량은 $D_K^2MND_F^2$이다. MobileNet의 Depthwise Separable Convolution은 3차원 계산을 2차원으로 계산하는 Depthwise Convolution 이후에 나머지 한 차원을 계산하는 Pointwise Convolution을 적용한다. Depthwise Convolution과 Pointwise Convolution은 각각 **그림 5, 그림 6**과 같다. Depthwise Convolution의 연산량은 $D_K^2MD_F^2$이고 Pointwise Convolution의 연산량은 MND_F^2이다. 따라서 Depthwise Separable Convolution의 연산량은 $D_K^2MD_F^2+MND_F^2$ 이고 기존 CNN와의 연산량 차이는 다음과 같다.

$$\frac{D_K^2MD_F^2+MND_F^2}{D_K^2MND_F^2}=\frac{1}{N}+\frac{1}{D_K^2}$$

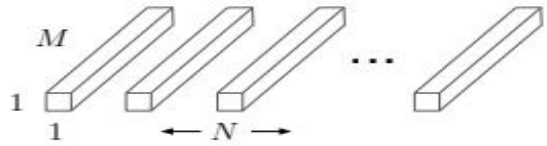
MobileNet은 3x3 Convolution을 사용하므로 D_K에 3을 대입해보면 약 8~9배 가량 연산량이 줄어듦을 확인할 수 있다 [6].



[그림 4] 기존 CNN의 연산 [6]



[그림 5] Depthwise Convolution Filters [6]



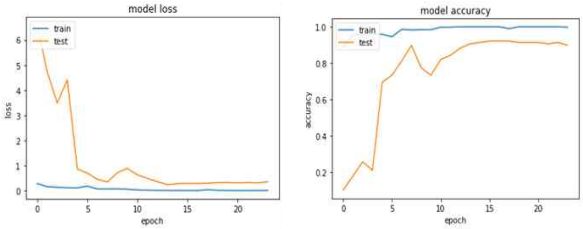
[그림 6] 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution [6]

2.4 MobileNet + LSTM

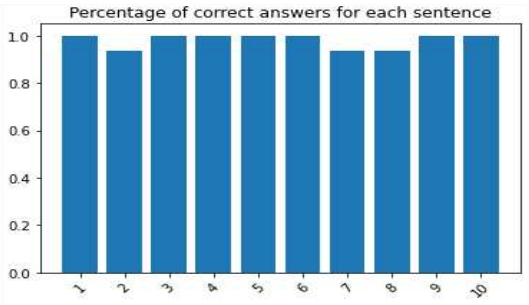
MobileNet의 마지막 Convolutional 레이어까지 진행 후, 추출된 결과를 두 번째 모델인 양방향 LSTM의 상단부분에 입력으로 넣어 학습을 진행하였다. LSTM(LongShort-Term Memory)이란 장기/단기 기억을 가능하게 설계한 신경망의 구조로, 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 가진 기존의 RNN(Recurrent Neural Network)을 보완한 것이다. MobileNet 계층 위에 이러한 LSTM 계층을 쌓음으로써, CNN 계층이 시간 정보를 추적하는 데 시간을 사용하지 않게 함으로써 더 나은 성능을 보여줄 것으로 기대하였다. 해당 모델에서 LSTM은 입술 모양의 변화를 포함하는 시퀀스를 패턴화하여, 이 시퀀스 차이를 사용하여 분류를 진행하게 된다. 활용한 LSTM 모델에는 크게 2가지 특징을 고려하였다 [7]. 첫 번째는 문장이 길어짐에 따라 한 벡터가 포함하고 있는 단어의 정보량의 증가하는 것을 고려하여 LSTM 레이어에 Bidirectional 함수를 사용하였다. Bidirectional LSTM 이란 기존의 LSTM 계층에 역방향으로 처리하는 LSTM 계층을 추가한 것이다. 이를 통해 프레임 단위의 입모양 이미지와 그 주변 이미지 정보를 균형 있게 담을 수 있다. 두 번째는 문장 내 단어의 양쪽의 정보를 활용하기 위해 Bidirectional LSTM 중 many-to-many 방식을 선택하였다. 추가적으로 영어 데이터셋을 활용한 LipNet에서 성능이 높게 나온 VGG16 + LSTM 방식을 바탕으로 MobileNet + LSTM 모델을 가지고 실험을 진행하였다 [8].

3. 연구 결과

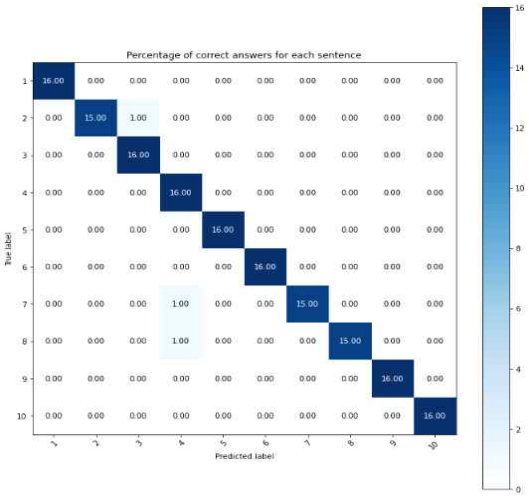
연구 결과 해당 모델은 98.12%의 학습 정확도를 확인할 수 있었다. 이와 같은 높은 수치가 나온 이유로는 MobileNet이 CNN의 사전 학습 모델 중 하나로 ImageNet에서 수만 개의 사진으로 사전 학습이 이미 진행되었기 때문이다. 실험 결과의 성능 수치를 높이기 위해 추가적으로 전처리 과정 내 밝기 조정, 데이터 증감을 진행하였고, 데이터셋의 분포를 균일하게 하기 위해 각 문장 별 반복 횟수를 동일하게 맞추었다. 또한 Batch size 및 Epoch 횟수 등 모델의 파라미터 값을 조절하여 실험을 진행하였다. Epoch 횟수의 증가를 통해 중도 탈락률이 증가하면 모델 성능이 나빠지는 것을 알 수 있었고 이를 통해 Epoch횟수의 크기가 모델의 성능과는 비례하기 않다는 것을 발견하였다. 그 결과, Batch size 32와 Epoch 25에서 중도 탈락 없이 좋은 성능을 보이는 것을 **그림 7**로 확인할 수 있었다. 정확도 **그림 8**과 **그림 9**을 보게 되면 지정한 문장이 전체적으로 고르게 다 학습이 되는 것을 확인할 수 있었다.



[그림 7] Loss and Accuracy Plots for MobileNet + LSTM



[그림 8] Percentage of correct answers for each sentence



[그림 9] Confusion Martix for MobileNet + LSTM

4. 결론

본 논문에서는 자체 제작한 한국어 데이터셋으로 MobileNet과 LSTM을 사용하여 한국어 입모양 인식 시스템 구현을 위한 학습 정확도를 알아보는 연구를 진행하였다. 그중 가장 좋은 성능을 보이는 Batch size와 Epoch 구하였다. 최적화된 값은 Batch size 32와 Epoch 25 일 때, 학습 정확도 98.12%가 나타났다. 다만, 한정된 Data를 사용하여 해당 문장만을 학습한 결과이기 때문에 그 이외의 단어와 문장에 대한 결과를 내지 못한 아쉬움이 있다.

현재 존재하는 다양한 입모양 인식 기술이 있다. 본 연구는 영어 데이터셋이 아닌 한국어 데이터셋을 활용했다는 점에서 큰 의미가 있다. 또한 기존의 VGG-16과 달리 연산의 효율성에 집중하여 보다 경량화된 네트워크인 MobileNet을 사용함으로써 모바일 기기에서의 동작 가능성에 대해 살펴볼 수 있었다.

마지막으로 이 연구를 통해 청각장애인의 구화 보조 수단으로서의 활용과 소음 환경에서의 음성 인식 성능 향상에 도움이 될 것으로 생각하고, 다양하고 풍부한 데이터셋을 활용한 추가적인 연구가 진행된다면 더 많은 단어와 문장을 인식하는 시스템이 구축될 것으로 기대한다.

5. Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [IITP-2021-0-01816 , 메타버스 자율트윈 핵심기술 연구]

6. 참고문헌

[1] Y. Xianoyi, Lipreading Recognition of English Vowels Using Convolutional Neural Network and Recurrent Neural Network, Master's Thesis of Chonbuk National University, 2017.

[2] Y.K. Kim, J.G. Lim, and M.H. Kim, "Lip Reading Method Using CNN for Utterance Period Detection," Journal of Digital Convergence, Vol. 14, No. 8, pp. 233-243, 2016.

[3] D.Y. Lim, S.G. Kim, and K.T. Chong, "Development of a Real-time Lip Recognition for Improving English Pronunciation Using Deep Learning," J ournal of Institute of Control, Robotics and Systems, Vol. 24, No. 4, pp. 327-333, 2018.

[4] C.G. Lee, E.S. Lee, S.T. Jung, and S.S. Lee,"Design and Implementation of a Real-Time Lipreading System Using PCA & HMM,"J ournal of Korea Multimedia Society, Vol. 7, No. 11, pp. 1597-1609, 2004.

[5] D. E. King, "Dlib-ml: A machine learning toolkit," Journal of Machine Learning Research, vol. 10, pp. 1755-1758, 2009.

[6] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017.

[7] Garg Amit, Jonathan Noyola, and Sameep Bagadia. .Technicalreport,StanfordUniversity,CS231nprojectreport,2016.

[8] Gutierrez, Abiel, and Zoe-Alanah Robert. "Lip Reading Word Classification".

[9] Parth Khetarpal, Shayan Sadar, and Riaz Moradian. ""‚InternationalJournalofComputerApplications,2017.