

## Deep CNN을 통한 손 제스처 인식 및 ASL-파벳 변환

이건, 윤리나, Ida Bagus Krishna Yoga Utama, 장영민

국민대학교

kmuleegun@naver.com, flsk0102@naver.com, krishna@kookmin.ac.kr, yjang@kookmin.ac.kr

### Hand gesture recognition and ASL-phabet conversion through deep convolutional neural network

Gun Lee, Ri Na Yoon, Ida Bagus Krishna Yoga Utama, Yeong Min Jang

Kookmin Univ

#### 요약

수화는 청각장애와 언어장애를 가진 사람들에게 없어서는 안 되는 중요한 대화수단이다. 본 논문에서는 영상인식 기술을 통해 알파벳을 인식하고 CNN 딥러닝을 통해 영상을 통해 구별되는 손 제스처를 문자로 변경해주는 알고리즘에 대해 연구하였다. 우선 CNN을 통해 손 제스처를 인식하여 학습시키고 이후에 새로운 손 제스처를 인식시키고 ASL-파벳을 예측하는 것이다. 제안된 CNN 아키텍처를 사용한 ASL-파벳 예측 정확도는 테스트 데이터에서 96.7%, 훈련 과정에서 98.3% 정확도를 보여주었다. 제안한 CNN 아키텍처는 다른 방법과 비교하여 가장 높은 정확도를 달성한다.

#### I. 서론

소리를 듣지 못하는 청각장애인들은 말을 제대로 하지 못하는 언어장애까지 수반하는 경우가 많다. 듣지 못하기 때문에 자신의 말을 제대로 교정할 수 없기 때문이다. 그들은 타인과의 대화를 위해선 수화를 사용한다. 하지만 이러한 수화는 많은 사람들에게 이야기 할 때는 불편할 수 밖에 없다. 언어를 시각으로 전달해야하기 때문이다. 또한 장애인들은 항상 도움을 구해야 하는 상황에 처하거나 수화를 알지 못하는 사람과의 대화에 처할 수 있다. 수화를 사용하는 사람들을 돕기 위해 수화를 영상인식을 통해 그들이 표현하는 언어, 수화를 자동적으로 문자로 변환해주는 아키텍처를 개발하였다. 여기서는 수화의 손짓을 감지하고 분류하는 방법을 제안한다. 데이터베이스는 미국 수화의 개방형 데이터베이스를 사용한다. 이 데이터베이스는 온라인에서 사용할 수 있으며 무료로 다운로드할 수 있다. 미국 수화 데이터베이스는 손 제스처를 분류하고 손 제스처를 알파벳으로 번역하기 위해 제안된 수정된 CNN 아키텍처를 훈련하는 데 활용된다.

#### II. 데이터 세트

청각 장애와 언어 장애를 가지고 있는 사람들이 사용하는 수화에도 다양한 종류가 존재한다. 무려 100개 이상의 다른 수화가 사용되며 인도 수화, 미국 수화, 이탈리아어 수화 및 많은 수화가 존재한다. 그 중 미국 수화(ASL)는 북미에서 4번째로 많이 사용될 정도로 많이 사용되는 언어이다. ASL은 미국뿐만 아니라 영어가 주요 의사소통으로 사용되는 30개 이상의 다른 국가에서도 사용된다. ASL은 American Manual Alphabet으로 널리 알려진 26개의 제스처 기호 집합으로, 영어 사전에 있는 다양한 단어를 표현하는데 사용할 수 있다. 우리가 사용하는 데이터 세트는 미국 수화용 MNIST(Modified National Institute of Standards and Technology database) 수화이다. 미국 수화는 원래 26개의 알파벳으로 구성되어 있다. 그러나 알파벳 J와 Z는 모션 기반 알파벳이기 때문에 이 두 알파벳은 이

데이터 세트에서 제외된다. 따라서 이 데이터 세트에서는 24개의 알파벳만 사용한다. 데이터 세트는 총 34,267개의 이미지로 구성되어 있으며, 학습 데이터용 이미지 27,455개와 테스트용 이미지 7,172개로 나뉜다. 데이터 세트의 각 이미지는 28 x 28 픽셀의 비슷한 크기를 가지며 모두 회색조 형식을 사용한다.

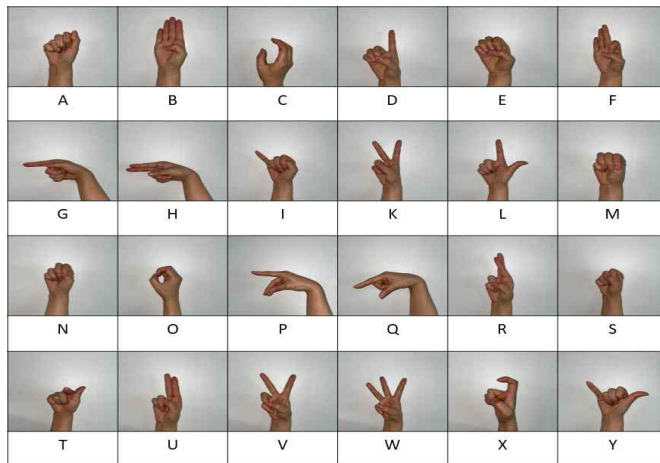


그림 1. 미국식 수화 (ASL)

#### III. 시스템 아키텍처

Convolutional neural network(CNN 또는 ConvNet)은 수동으로 특징을 추출할 필요 없이 데이터로부터 직접 학습하는 딥 러닝을 위한 신경망 아키텍처이다. CNN은 영상에서 객체, 얼굴, 장면 인식을 위한 패턴을 찾을 때 특히 유용하다. 자율주행 차량 및 얼굴 인식 응용 분야와 같이 객체 인식과 컴퓨터 비전이 필요한 응용 분야에서는 CNN을 많이 사용한다. 이 CNN을 사용하여 영상에서 손 제스처를 인식하여 표현하고자 하는 문자를 가져올 수 있다. 이 연구에서 우리는 6개의 컨볼루션 레이어, 6개의 배

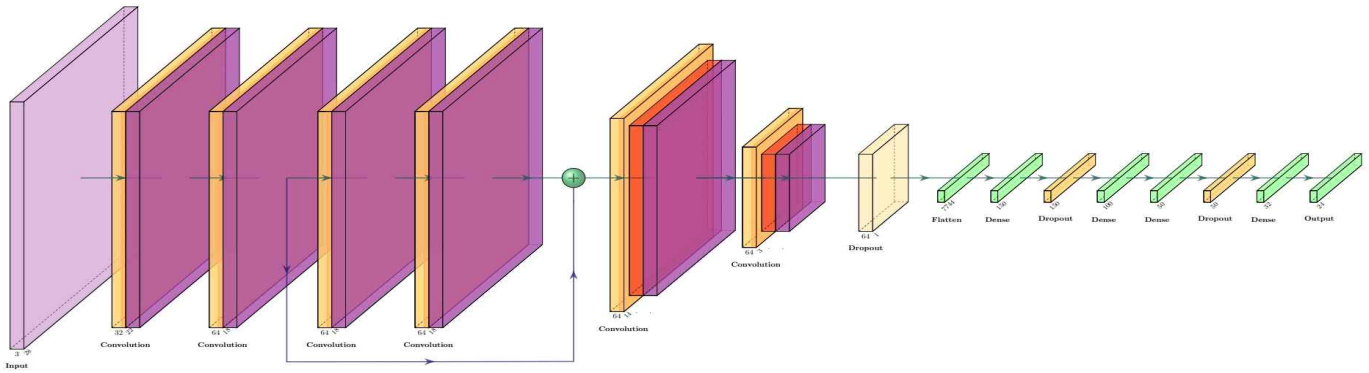


그림2. 제안된 CNN 아키텍처

치 정규화 레이어, 2개의 풀링 레이어, 3개의 드롭아웃 레이어, 5개의 고밀도 레이어로 구성된 CNN 아키텍처를 제안한다. 우리는 또한 두 번째 컨볼루션 레이어와 네 번째 컨볼루션 레이어 사이에 잔여 연결을 구현한다. 각 컨볼루션 레이어에서 서로 다른 크기의 커널을 적용하고 ReLU 활성화 함수를 사용하여 컨볼루션 레이어를 활성화한다. 또한 모든 컨볼루션에 대해 배치 정규화를 적용한다. 여러 드롭아웃 레이어가 CNN 아키텍처 전체에 구현되어 모델을 정규화하여 네트워크의 과적합을 줄이는 데 도움이 된다. 컨볼루션 레이어는 이미지에 포함된 특별한 정보를 추출하는 것을 목적으로 한다. 그런 다음 이미지를 분류하기 위해 고밀도 레이어를 사용한다. 조밀한 레이어는 각 뉴런이 ReLU에 의해 활성화되는 서로 다른 수의 뉴런으로 구성된다. 데이터 세트의 각 클래스를 나타내는 24개의 뉴런으로 구성된 출력 레이어이기 때문에 Softmax는 마지막 밀집 레이어만 활성화한다. Softmax 활성화 함수는 출력 레이어의 각 뉴런에 대한 확률 값을 생성한다.

#### IV. 결과 및 토론

##### IV-I. 데이터 전처리

데이터 세트의 이미지 크기는 모두 28 x 28픽셀이며 이미 회색조 형식이다. CNN에 이미지를 삽입하기 전에 이미지의 각 픽셀 값을 다시 조정한다. 원래 픽셀 값의 범위는 0-255이다. 여기에서 0-1 범위의 값을 다시 조정한다. 모든 이미지를 동일한 방식으로 처리하고 출력 범위가 0-1인 Softmax 기능을 사용하기 때문에 이 범위 값을 선택한다.

##### IV-II. 교육 및 테스트

훈련 과정을 위해 훈련 데이터를 80%-20%로 나눈다. 80% 데이터는 훈련에 사용되고 20% 데이터는 훈련 프로세스의 모든 시대에 대한 네트워크 성능 검증에 사용된다. 이 연구에서는 100개의 epoch 훈련을 실행한다. 훈련 과정에서 우리는 항상 검증 과정에서 나온 정확도 값을 모니터링한다. 훈련에서는 항상 최고의 정확도를 달성하는 최고의 모델을 저장한다. 훈련 과정이 끝난 후, 우리는 이전에 모델에서 볼 수 없었던 이미지 데이터 세트를 사용하여 테스트 과정을 수행한다. 훈련 과정에서 저장한 최상의 모델을 로드하고 해당 모델을 사용하여 테스트 데이터 세트를 기반으로 예측한다. 교육 및 테스트 과정에서 Crosse entropy loss 함수를 사용하여 손실을 측정한다. 손실을 측정하는 것은 모델 무게를 수정하여 모델 성능을 향상시키는 데 중요하다.

##### IV-III. 결과 및 분석

훈련 과정에서 우리는 0.005의 손실 값과 98.3%의 정확도를 달성했다. 한편 테스트 과정에서 손실 값 0.09와 정확도 96.7%를 달성했다. 훈련과 테스트 과정의 정확도가 크게 다르지 않기 때문에 만족할만한 결과를 얻

은 것 같다. 이는 모델이 과적합 또는 과소적합 되지 않고 모델이 이전에 해당 데이터에 대해 학습한 적이 없음에도 불구하고 올바르게 분류할 수 있도록 모델이 매우 잘 정규화 되었음을 의미한다. 이 결과는 네트워크의 잔여 연결 때문이기도 하다. 비선형 활성화 함수가 네트워크를 따라 그래디언트가 사라지거나 폭발할 수 있기 때문에 잔차 연결은 그래디언트가 직접 흐르도록 하고 비선형 활성화 함수의 영향을 받지 않는다. 우리는 본 논문의 결과를 다른 논문의 결과와 비교한다. 표 1로 다른 방법과의 비교한 결과를 제시한다.

표 1. 다른 방법과의 정확도 비교 결과

| 방법                    | 정확도 (%) |
|-----------------------|---------|
| LeNet [1]             | 82.19   |
| CapsNet [1]           | 88.93   |
| CapsNet augmented [1] | 95.08%  |
| CNN [2]               | 95.3%   |
| Ours                  | 96.7%   |

#### V. 결론

본 논문에서는 MNIST 수화 데이터 세트를 사용하여 훈련되는 CNN 아키텍처를 사용하여 손 제스처를 감지하고 분류하는 작업을 발표했다. 제안된 방법의 성능은 훈련 과정에서 98.3%의 정확도와 테스트 과정에서 96.7%의 정확도를 생성할 수 있음을 약속한다. 본 결과를 통해 deep CNN을 사용한 손 제스처 인지 방식을 더 보완하여 수화를 사용하는 사람들에게 게 보다 더 편한 삶을 제공할 수 있기를 바란다.

#### 참고 문헌

- [1] Bilgin, M., & Mutludogan, K. (2019). American sign language character recognition with capsule networks. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE.
- [2] Chakraborty, D., Garg, D., Ghosh, A., & Chan, J. H. (2018). Trigger detection system for American sign language using deep convolutional neural networks. Proceedings of the 10th International Conference on Advances in Information Technology - IAIT 2018. New York, New York, USA: ACM Press.
- [3] MathWorks "https://kr.mathworks.com/discovery/convolutional-neural-network-matlab.html"