한글 필기체 인식을 위한 딥러닝 기반의 모델 구현

김재정, 박민호, 정상중, *정도운
동서대학교 인공지능응용학과
e-mail : qkqhd74@gmail.com, pinkc47@naver.com, sjjung@dongseo.ac.kr,
dujeong@dongseo.ac.kr

CNN-based Hangul Handwriting Recognition System

Jae-Jung Kim, Min-Ho Park, Sang-Joong Jung, *Do-Un Jeong
Department of Applied Artificial Intelligence
Dongseo University

Abstract

In this paper, we implemented a Hangul handwriting recognition system using CNN. The 'PHD08' data set was used as training data, and the implemented model consists of a total of 10 covloution layers. As a result of the performance evaluation of the implemented model, a high recognition rate of 99.68% was confirmed, and the result of the recognition experiment using 100 actual handwritten images also showed high accuracy. In future research, we intend to continue research on database configuration and model learning speed and recognition rate improvement to improve Hangul recognition.

I. 서론

최근 인공지능과 빅데이터 관련 산업의 발전으로 인해 아날로그 문서들의 디지털화에 대한 사회적 요구가 증가하고 있다. 이에, 다양한 필체를 이미지 스캐너 등의 장비를 통해 컴퓨터가 편집할 수 있는 문자코드로 변환하는 광학 문자 인식(OCR) 기술이 주목받고 있다. 하지만 광학 문자 인식 시스템은 인쇄체의 경우 높은 정확성을 나타내지만, 필기체에 대해서는 명확하게 인식하지 못하고 있다. 특히, 한글 필기체의 경우 다른 언어와 다르게 특정 글자의 유사성, 자음과 모음의 조합으로 인한 형태적 특성 등으로 문제가 있어 인식에 어려움이 있다. 이를 해결하고자 OpenCV, Pytesseract 등을 사용한 선행연구가 진행되었지만, 글자 주변이 깨끗하지 못하거나, 문서의 상태가 좋지 못하는 경우 낮은 정확성을 나타내고 있다[1]. 따라서 잡음을 고려하여 필기체를 인식할 수 있는 시스템이 필요하다. 본 논문에서는 정확성이 높은 딥러닝 기반의 한글 필기체 인식 모델을 구현하였다.

II. 본론

2.1 PHD08 데이터 세트

기존의 한글 데이터베이스는 주로 표준 글꼴을 이용하여 제작되어 손글씨와 같은 분야에 사용하는데 한계가 있다. 하지만 'PHD08' 데이터베이스는 완성형 글자 2,350자를 대상으로 9종류에 글꼴을 포함하며 이진 임계치와 회전 각도를 변형시킨 샘플을 보유하고 있다.

이렇게 총 5,139,450개의 샘플을 가진 한글 데이터베이스는 인쇄 글자가 아닌 손글씨를 인식하는 것에 사용하기 용이하다.

III. 구현

구현된 딥러닝 기반의 한글 필기체 인식 모델은

'PHD08' 데이터 세트를 이용하여 한글 분류 학습을 수행하였다. 학습에 사용된 데이터는 전처리 과정을 통해 28×28형태의 동일한 사이즈로 재구성하였다. 학습 모델은 Pooling layer와 relu 활성화 함수를 사용한 Convolution layer로 구성되어있고 출력단에서는 softmax 활성화 함수를 이용하여 한글 글자 분류를 진행하였다. 이후, Adam을 이용한 최적화 기법을 적용하였으며 Dropout과 Early stop을 적용하여 과적합을 방지하였다. 알고리즘 내부에서 사용된 학습 데이터의 전처리 일례를 그림 1에 나타내었으며, 구현된 한글 필기체 분류 모델의 구성을 표 1에 나타내었다.

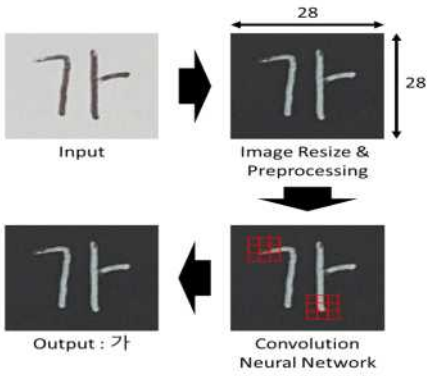

그림 1. 학습 데이터의 전처리 일례

표 1. 구현된 한글 필기체 분류 모델의 구성

Layer	Output size
Conv2D(Input, relu)	28 × 28 × 32
Max-Pooling	13 × 13 × 32
Conv2D(relu)	11 × 11 × 64
Max-Pooling	5 × 5 × 64
Conv2D(relu)	3 × 3 × 64
Dense	64
Dropout	0.2
Dense	32
Dropout	0.2
Dense(softmax)	2350

구현된 한글 필기체 분류 모델의 성능평가를 위해 학습데이터는 전체 데이터에서 80%를 무작위로 선별하여 학습용 데이터로 사용하였고 나머지 20%는 검증용 데이터로 구분하여 학습을 수행하였다. 학습 결과 구현된 모델은 검증용 데이터에 대하여 99.68%의 정확성과 0.0149의 손실값을 확인하였다. 이후 한글 필기체 인식 시스템의 정확도를 평가하기 위해 실제 손글씨 100개를 촬영하여 실험을 진행하였으며, 실제 손글씨를 인식한 결과 역시 정확히 분류되는 것을 확인하였다. 학습 결과의 정확성(test_acc)과 손실값(test_loss) 결과를 표 2와 그림 1에 나타내었다.

표 2. 구현된 CNN 모델의 성능 평가

test_acc	test_loss
0.9968	0.0149

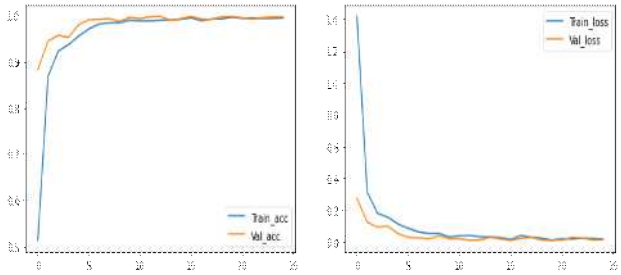

그림 2. 한글 필기체 인식 모델의 실험 결과

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 정확성이 높은 딥러닝 기반의 한글 필기체 인식 모델을 구현하였다. 구현된 인공지능 모델은 CNN을 기반으로 'PHD08' 데이터 세트를 이용하여 학습데이터로 사용하였으며 Pooling layer와 relu 활성화 함수를 사용한 Convolution layer로 구성하고 softmax 활성화 함수를 이용하여 한글 글자 분류를 진행하였다. Adam을 이용한 최적화 기법을 적용하였고 과적합 방지를 위해 Dropout과 early stop을 적용하였다. 따라서 총 10개의 레이어로 구성된 CNN 모델을 구현하였다. 구현된 모델의 정확성을 평가한 결과 99.68%의 높은 인식률을 확인하였다. 최종 구현된 시스템의 성능평가를 위해 실제 손글씨 이미지 100장을 촬영하여 인식 실험을 진행하였으며, 실제 손글씨를 인식한 결과 역시 정확히 분류되는 것을 확인하였다. 향후 연구에서는 한글 인식 개선을 위한 데이터베이스 구성과 모델의 학습 속도 및 인식률 향상을 위한 연구를 지속하고자 한다.

감사의 글

본 연구는 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구(No.2018R1D1A1B07045337) 사업 및 중소벤처기업부의 산학연 collabo R&D 사업(S3247582)에 의해 지원된 연구 결과물임을 밝힙니다.

참고문헌

[1] Y. S. Dho, S. M. Nam, S. J. Yoon, J. H. Lim, B. Y. Park. Development of letter recognizing and speech system using raspberry pi. JKIIT. 2018