딥러닝을 이용한 한글 OCR 정확도 향상에 대한 연구

강가현 · 고지현 · 권용준 · 권나영 · 고석주

경북대학교 컴퓨터학부

A Study on Improvement of Korean OCR Accuracy Using Deep Learning

Ga-Hyeon Kang · Ji-Hyun Ko · Yong-Jun Kwon · Na-Young Kwon · Seok-Ju Koh

Kyungpook National University

E-mail : rkgus6545@naver.com · iwlgusi1@naver.com · rtyuipl@naver.com · kny4262@naver.com · sjkoh@knu.ac.kr

요 약

다음은 본 논문에서는 딥러닝을 통한 한글 OCR 정확도 향상을 제안한다. OCR은 인쇄되거나 손으로 쓴 문자를 광학적 방법으로 감지 인식하여 디지털로 인코딩하는 프로그램이다. 현재 가장 많이 쓰이는 tesseract OCR의 경우, 영문 인식의 정확도가 높다. 하지만 한글은 복잡한 구조에 비해 학습 데이터가 적어 정확도가 떨어진다. 따라서 이 연구에서는 이미지 프로세싱을 통해 원하는 이미지에서 글자 영역을 추출하고, 이를 학습 데이터로 활용한 딥러닝으로 한글 OCR의 정확도를 향상시키는 방법을 제안한다. 기존 영문과 숫자 및 몇 가지 언어에만 국한되어 발전해왔던 OCR을 다양한 언어에도 응용할 수 있을 것으로 기대된다.

ABSTRACT

In this paper, we propose the improvement of Hangul OCR accuracy through deep learning. OCR is a program that senses printed and handwritten characters in an optical way and encodes them digitally. In the case of the most commonly used Tesseract OCR, the accuracy of English recognition is high. However, Hangul has lower accuracy because it has less learning data for a complex structure. Therefore, in this study, we propose a method to improve the accuracy of Hangul OCR by extracting the character region from the desired image through image processing and using deep learning using it as learning data. It is expected that OCR, which has been developed only by existing alphanumeric and several languages, can be applied to various languages.

키워드

OCR, 딥러닝, 이미지 프로세싱, 한글 인식률, 정확도 향상

Ⅰ. 서 론

컴퓨터의 등장 이전에는 문서 기록을 위한 수단으로 주로 종이가 사용되었다. 이후 컴퓨터가 대중화되면서 종이 문서가 스캔을 통하여 전자문서로 많이 변환되었다. 하지만 이렇게 변환된 전자문서는 사람이 읽기만 가능하고 가공하지 못한다. 가공을 위해서는 변환된 전자문서를 사용자가 편집 가능한 텍스트 문서로 바꾸어야 한다. 이를 사람의 수작업으로 수행한다면 노동의 낭비와 비효율, 그리고 처리 가능한 양의 한계에 부딪히게 된다. 이러한 문제는 이미지 파일에 존재하는 글자를 편집 가능한 텍스트 형태로 바꾸어 주는 Optical Character Recognition(OCR) 기술을 통해 많은 부분 해결되었다. 이 OCR 기술은 카드 인식, 우편물 분류 등 다양한 분야에서 활용되고 있다.[1][2]

OCR의 활용 가능성이 대두되면서 해당 분야의 성능 향상 연구 및 프로그램 개발이 활발히 이루어지고 있다. 현재 ABBYY FineReader, 아르미, Readdiris Corporate등 다양한 OCR 프로그램들이 존재한다. 그 중 Tesseract OCR은 구글사의 OCR 프로그램으로 현재 가장 많이 사용되고 있다. Tesseract는 다양한 언어의 텍스트 변환을 지원하고 있다. 하지만 모든 언어에서 동일한 변환 일치율을 보이지는 않는다. 라틴문자와 같

이 한 음운이 한 글자인 언어는 문자인식이 쉽지만, 그에 비해 한글은 음운 하나가 자음과 모음이 합쳐져 음운의 개수에 비해 글자의 경우의 수가 매우 많다. 따라서 한글의 이런 복잡한 구조 때문에 라틴계열의 언어보다 많은 학습 데이터가 필요해 문자인식의 정확도가 떨어진다.

따라서 본 논문에서는 딥러닝으로 한글 OCR의 정확도를 향상시키는 방법을 제안한다. 제안된 방법에서는 한글의 초성, 중성, 종성의 모든 경우의 수를 조합하여 글자의 이미지를 생성한다. 그리고 그 이미지를 Convolutional Neural Networks(CNN)을 이용해 딥러닝으로 학습시킨다.[3] 충분한 학습 후, 테스트할 이미지를 이미지 프로세싱을 통해 글자 영역을 추출한다. 추출한 글자 영역을 테스트 케이스로 OCR을 진행한다.

II. 본 론

2.1 학습 데이터 생성

현재 인식률이 높은 라틴계열의 언어들은 한 음운이 한 글자로 되어있어 구조가 단순하고 글자의 경우의 수가 적다. 그에 비하여 한글은 한 음운이 자음과 모음을 이용해 초, 중, 종성을 조합한 구조로 되어있어 복잡하고 글자의 경우의 수가 많다. 그로 인해 딥러닝하는 OCR 프로그램에 많은 이미지를 학습시키더라도 모든 글자를 학습시키지 못하는 경우가 많다. 따라서 본 장에서는 조합 가능한 모든 한글 글자 이미지를 생성하여 학습 데이터로 이용하려고 한다.

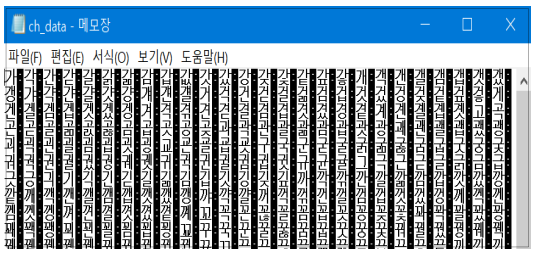

그림 1. 모든 한글 글자를 조합한 텍스트 파일

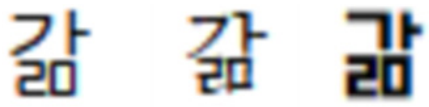

그림 2. 다양한 폰트를 사용한 학습 데이터의 예

먼저, 표기 가능한 모든 경우의 한글 글자를 조합하여 이를 그림 1처럼 텍스트 파일에 저장하도록 한다. 생성된 텍스트 파일에서 한 글자씩 읽어 들인 다음, 지정한 폰트를 이용해 모든 글자를 그림 2처럼 학습 이미지로 만든다. 이 과정을 여러 가지 폰트를 사용해서 다양한 학습 이미지를 생성한다. 그리고 이 이미지들을 딥러닝을 위한 학습 데이터로 사용한다.

2.2 OCR 프로그램 동작과정

이 절에서는 OCR 프로그램이 이미지에서 글자를 인식하고 텍스트로 출력하는 과정을 설명한다. 이때 사용하는 이미지는 임계처리 과정을 거친 것을 사용한다. 이는 이미지에서 검정 글자를 돋보이게 만들어 OCR의 품질을 높이기 위함이다. OCR 프로그램의 글자 인식 과정은 크게 Detection, Prediction, Semantic Analysis, Reconstruction 4가지 단계로 나뉜다.

1) Detection

이미지에서 글자로 추정되는 영역은 여러 줄로 이루어진 한 문단으로 인식된다. 이 과정에서는 인식된 문단에서 한 줄씩 사각형 모양으로 묶는다. 그리고 각 사각형 모서리의 위치 정보들을 담은 객체들을 얻는다. 이를 다음 과정으로 전달한다.

2) Prediction

전달받은 객체에서 각 줄에 있는 음운들이 하나씩 들어있는 사각형을 만들고, 그 위치를 담은 객체를 얻는다. 학습시켜놓은 모델을 통해 객체마다 예상되는 초성, 중성, 종성들을 찾는다. 그리고 초성, 중성, 종성들을 한 글자로 조합해 음운별 예상 후보를 배열에 저장하고, 다음 단계로 전달한다. 이때 음운별 예상 후보는 한 가지 이상일 수 있다.

3) Semantic Analysis

전달받은 음운별 예상 후보 중 이미지의 글자와 일치하는 정도가 가장 높은 글자를 반환한다. 이 반환 과정은 문단 내의 모든 글자에 대해서 수행한다.

4) Reconstruction

위 과정을 통해 반환받은 글자들을 합쳐서 한 줄로 만든다. 그리고 그 줄들을 이어서 문단으로 재구성한 결과물을 출력한다.

III. 실험 결과

tesseract 4.0.0-beta.1 버전과 tensorflow 1.7.0 버전을 사용하여 실험을 진행하였다. 흰 바탕에 검은 글씨로 된 무작위의 이미지를 선택하여 tesseract와 본 논문에서 설명한 방법을 이용하여 딥러닝한 OCR 프로그램에 실행하여 텍스트 파일을 얻었다.

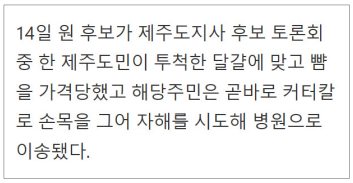

그림 3. OCR 테스트 이미지

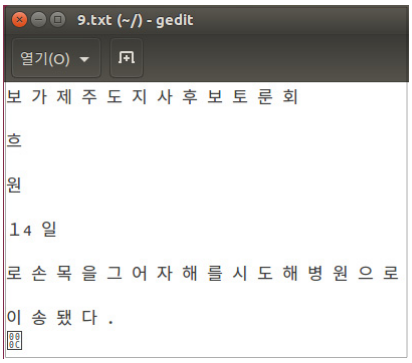

그림 4. tesseract 테스트 결과

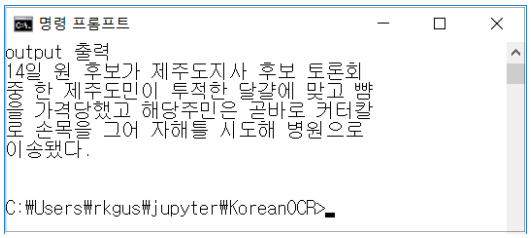

그림 5. 딥러닝 프로그램 테스트 결과

그림 3은 테스트용 이미지이고 그림 4는 tesseract를 사용, 그림 5는 본 논문의 딥러닝 프로그램을 사용하여 테스트한 결과이다. 그림 4에서 보는 것처럼 tesseract는 한글을 인식하여도 순서가 엉망인 경우가 많지만, 그림 5처럼 딥러닝을 사용했을 때는 tesseract 보다 성능이 향상된 것을 확인할 수 있다.

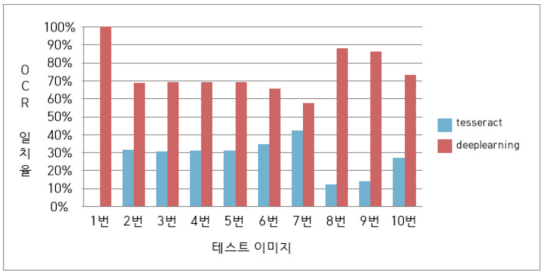

그림 6. 프로그램에 따른 OCR 결과 일치율

앞서 설명한 방식으로 10개의 이미지를 실험하였다. 이미지의 글자와 특수문자 '.', ','를 포함한 모든 글자의 개수를 C라 하였다. 테스트 이미지와 비교하여 각 글자의 형태와 순서가 일치할 시 정확히 출력한 것으로 판단하였다. 이를 기준으로 tesseract로 인식하였을 때 정확히 출력된 글자의 개수를 L_t, 딥러닝 프로그램으로 인식하였을 때 정확히 출력된 글자의 개수를 L_d라 하였다.

이때, 각각의 일치율을 $\frac{L_t}{C}\times 100(\%), \frac{L_t}{C}\times 100(\%)$로 계산하였다. 그림 6은 tesseract와 딥러닝 프로그램에서 추출한 텍스트들이 실제와 얼마나 일치하는지를 실험한 결과를 나타낸 그래프이다. tesseract와 비교해 딥러닝 프로그램의 일치율이 평균 약 58.8%만큼 높은 것을 확인하였다.

Ⅳ. 결 론

본 논문에서는 한글 OCR 프로그램 성능 향상을 위해 조합 가능한 모든 한글 글자를 이미지화하여 이를 학습 데이터로 이용하는 방법을 제안한다. 이후 이를 이용해 OCR 프로그램을 딥러닝 시킨다. 본 OCR 프로그램은 Detection, Prediction, Semantic Analysis, Reconstruction의 과정을 통해 동작한다. 제안한 방법으로 딥러닝 시킨 프로그램이 현재 많이 사용되고 있는 OCR 프로그램인 tesseract에 비해 한글 인식에서 향상된 일치율을 보임을 증명하였다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학사업의 연구결과로 수행되었음(2015-0-00912)

참고문헌

[1] 왕진원, "광학문자판독(OCR)기술을 이용한 우편물 구분원리와 판독율 향상 방안", 우정정보 1993권 2호

[2] 정민철, "신용카드 번호의 광학적 인식", 반도체디스플레이기술학회지(한국반도체디스플레이기술학회) 13권 1호

[3] 임수창, 김승현, 김연호, 김도연, "소 부류 객체 분류를 위한 CNN기반 학습망 설계", 한국정보통신학회 논문지