OPTICAL CHARACTER RECOGNITION WITH CRNN MODEL

MyungJun Lee
chancecatch@icloud.com

ABSTRACT

Optical Character Recognition (OCR)은 이미지 속 텍스트 정보를 검출하는 데 사용되는 기술로, Text Detection과 Text Recognition의 두 단계로 구성되어 있다. OCR의 역사는 오래되었으며, 딥러닝이 도입되기 전까지 Tesseract OCR (See Figure 1) 같은 OCR 엔진이 다양한 텍스트 관련 작업에 널리 쓰였다. 그러나 이러한 전통적인 OCR 엔진은 많은 단계를 필요로 하고 복잡하며, 손글씨나 낮은 품질의 이미지에서는 정확도가 떨어진다는 단점이 있다.

본 연구에서는 딥러닝을 OCR 과정에 적용하였다. 텍스트 검출에는 세그멘테이션(Segmentation) 기반의 CRAFT를 활용한 keras-ocr을 사용하였으며, 텍스트 인식에는 컨볼루션 레이어(Convolution Layer)와 RNN을 결합하고 CTC(Connectionist Temporal Classification)로 학습된 CRNN(Convolutional Recurrent Neural Network) 모델을 사용하였다. 이 연구를 통해 CRNN 모델이 다른 모델들에 비해 텍스트를 성공적으로 인식함을 확인할 수 있었다. 그러나 CRNN 모델에도 여러가지 한계가 있으며, 최근 등장한 Transformer 기반의 인식 모듈이 이러한 한계를 극복할 수 있는 가능성을 제시하기를 희망한다.

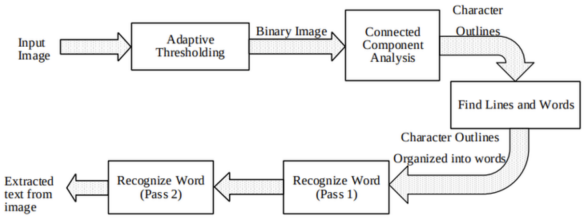

Figure 1: Architechture of Tesseract OCR

1 Introduction

최근의 Deep Neural Network연구는 대부분 객체 탐지나 분류에 초점을 맞추고 있다. 본 논문에서는 컴퓨터 비전의 오래된 과제 중 하나인 이미지 기반의 Unsegmented Sequence text Recognition에 대해서 다루려고 한다. 이미지 기반의 텍스트 인식을 위해서 몇 가지 시도가 선행되었다. 글자 단위로 텍스트를 탐지하는 방식이 있었으나 이 방법은 글자를 탐지한 후 맥락에 맞게 다시 단어로 묶어주어야 하는 단계가 요구되었고, 다른 방법으로는 텍스트를 이미지 분류 문제로 인식하고, 각 단어에 클래스 레이블을 할당하는 방식도 있었으나 대규모 훈련 모델을 필요로 하며, 다양한 유형의 Sequence 인식에는 한계가 있었다. 본 논문에서는 CNN과 RNN의 조합을 통해서 새로운 신경망을 설계하고 이미지 내의 Sequence를 인식함으로써, 최근의 연구에 기여하고자 한다. 이 새로운 신경망을 CRNN(Convolutional Recurrent Neural Network)이라고 제시한다.

2 Method

본 논문에서는 텍스트 검출에는 CRAFT를 활용한 keras-ocr을 사용하였다. Text Detection 부터 받은 이미지를 Input으로 CNN 모델을 통해 특징을 추출하고, Map-To-Sequence로 Sequence형태의 feature로 변환한 후 다양한 길이의 Input을 처리할 수 있는 RNN으로 넣는다. 여기서 RNN은 텍스트가 지니고 있는 특성상 장기 정보 유지 및 앞과

뒤의 텍스트 정보를 활용하기 위해 Bidirectional LSTM으로 설계하였고, 단계마다 나오는 결과는 Transcription layer(Fully connected layer)를 통해서 단계마다 어떤 character의 확률이 높은지 예측하도록 했다.(See Figure 2)

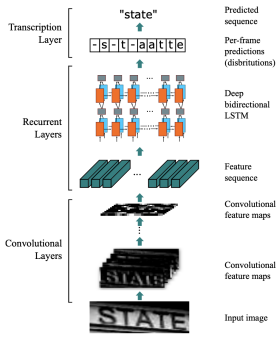

Figure 2: The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature se- quence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

2.1 structure of RCNN

본 논문에서 RCNN Recognition model(See Figure 3)의 정량적인 평가를 위해서 MJSynth 데이터셋을 활용하였다. 입력 이미지는 텍스트 검출 및 전처리를 통해서 높이가 32가 되도록 조정하였다. 합성곱 레이어 구조는 VGG-VeryDeep 아키텍처를 기반으로 구성하였다. 3번째와 4번째 최대 풀링 레이어에서는 1×2 크기의 직사각형 풀링 윈도우를 사용하여, 긴 특징 시퀀스를 형성하게 함으로써, 긴 특징을 지닌 문자를 인식할 수 있도록 하였다. Recurrent Layer(Bidirectional-LSTM)는 은닉 노드를 256개로 설정하였다. 끝으로 Transcription layer(Fully connected layer)를 연결하였다.

Type	Configurations
Transcription	-
Bidirectional-LSTM	#hidden units:256
Bidirectional-LSTM	#hidden units:256
Map-to-Sequence	-
Convolution	#maps:512, k:2 × 2, s:1, p:0
MaxPooling	Window:1 × 2, s:2
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
MaxPooling	Window:1 × 2, s:2
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:128, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:64, k:3 × 3, s:1, p:1
Input	$W \times 32$ gray-scale image

Figure 3: Network configuration summary. The first row is the top layer. ‘k’, ‘s’ and ‘p’ stand for kernel size, stride and padding size respectively

2.2 CTC(Connectionist Temporal Classification)

CTC는 순환 신경망에서 Unsegmented data를 Labelling하는 방법으로, 기존의 방법들은 입력 데이터와 출력 레이블이 서로 다른 길이의 시퀀스를 가질 때는 정렬(Alignment)이 필요했다. CTC는 이러한 정렬 없이도 연속적인 시퀀스 데이터를 효과적으로 처리할 수 있는 방법이다. 본 연구에서 CRNN 모델에 CTC를 적용함으로써, 입력 이미지의 다양한 길이의 텍스트 시퀀스를 효과적으로 인식하고, 이를 연속적인 문자 레이블로 변환할 수 있었다.

3 Conclusion

본 연구를 통해 CRNN 모델이 이미지 내의 텍스트를 효과적으로 인식하는 데 매우 유용하다는 것을 확인할 수 있었으며, CTC와 결합된 CRNN 모델은 Unsegmentation 텍스트 시퀀스를 처리하는 데 강력한 성능이 있는 것을 확인할 수 있었다. 그러나 CRNN 모델은 여전히 한계가 있으며, 특정 부분에서는 개선이 요구된다. 최근 등장한 Transformer 기반의 인식 모듈과 같은 새로운 기술들이 이러한 한계를 극복하고 OCR 기술을 더욱 발전시킬 수

있는 가능성을 제시하고 있으므로, 향후 더욱 발전된 모델들이 개발되어 더 다양한 분야에서 활용될 수 있기를 기대한다.

Acknowledgments

This was was supported in part by Aiffel