어텐션 중심을 이용한 글자 단위 영역 검출

김지인*, 정창성*
*고려대학교 전기전자공학과
e-mail : skygazer227@korea.ac.kr

Character-level Region Detection Using Attention Center

Jiin Kim*, Chang-Sung Jeong*
*Dept. of Electrical Engineering, Korea University

요 약

최근 딥러닝으로 진행되는 광학 문자 인식 분야는 대부분 단어 단위로 인식하는 것으로 글자 단위의 영역을 검출하는 데에는 적합하지 못하다. 본 연구는 각 글자의 영역을 검출하기 위해 기존의 딥러닝을 이용한 광학 문자 인식 절차인 단어 분리 과정과 단어 인식 과정을 유지하면서 어텐션 중심을 이용하여 각 글자의 영역을 보다 정확하게 검출하는 것을 목표로 한다. 제안하는 모델은 CRAFT 와 Attention Network 를 사용한 OCR 과정을 확장한 모델로 각 단어 문자열 결과물에 각 글자의 영역을 추가로 나타내게 되며 각 글자와 라벨 간의 IOU 평균은 0.671 로 나타났다.

1. 서론

최근 이미지상의 문자를 인식하는 광학 문자 인식(Optical character recognition; OCR) 분야의 관심이 증가하고 있다. 초기의 OCR 분야는 각 글자의 형태를 검출하여 단어를 조합하는 방식으로 진행되었으나 딥러닝으로 전환이 이루어지면서 최근의 연구는 대부분 글자 단위가 아닌 단어 단위로 진행이 되고 있다.

이러한 단어 단위의 문자인식은 반대로 글자 단위 인식에는 취약한 단점이 있다. 대부분의 경우에는 단어 단위로 인식해도 큰 문제가 없지만 특수문자가 끼어 있으면서 붙어있는 두 단어를 한 단어로 인식하는 등 예외사항이 발생할 경우 인식률이 떨어지는 현상이 생기게 된다.

따라서 이러한 문제를 해결하기 위해 기존의 딥러닝을 이용한 방법을 유지하되 추가적인 과정을 통해 각 글자의 영역을 검출하는 방법을 제안하며 이 방법을 통해 궁극적으로 위에 나타난 예외사항 등을 해결하고자 한다.

2. 관련 연구

최근의 광학 문자 인식분야의 연구는 대부분 딥러닝으로 진행되고 있다. 이 분야는 일반적으로 다시 두 분야로 나뉘는데 하나는 단어 분리(word segmentation) 분야로써 원본 이미지로부터 각 단어의 영역을 검출하는 분야이고 다른 하나는 단어 인식(word recognition) 분야로써 단어 분리 과정을 통해 얻은 단어 영역 안의 이미지로부터 해당 단어 문자열을 찾는 과정이다.

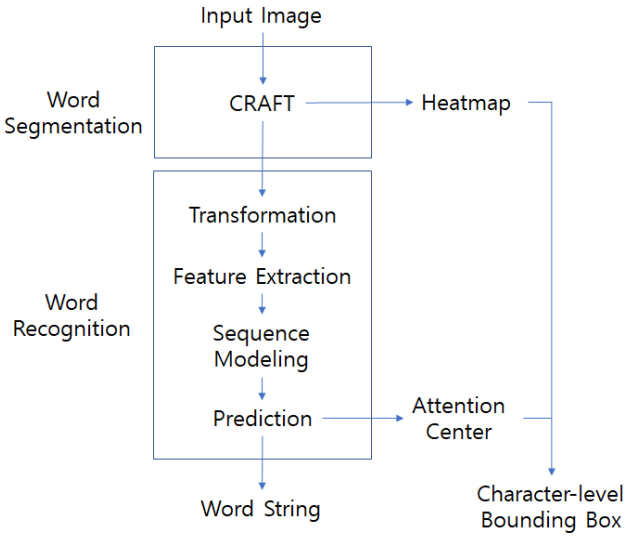
최근에는 두 과정 모두 딥러닝을 적용하는 방법이 주로 사용되며 단어 분리 과정에는 SSD[1], R-CNN[2], CRAFT[3] 등이 제안되었고 단어 인식 과정에는 R2AM[4], RARE[5], 등이 제안되었다.

단어 분리 과정의 방법들 중 비교적 최근에 나온 CRAFT 의 경우 이전에 존재하던 단어 영역을 직접적으로 검출하는 방식과는 다르게 각 단어에 존재하는 글자들에 대한 영역 정보와 단어 내부의 글자들을 연결하는 링크에 대한 정보 각각 히트맵 형태로 학습하는 방식을 사용하며 검출된 글자와 링크 정보로부터 단어의 바운딩 박스를 형성하게 된다. 단 이 방법의 경우 한글의 자모들 각각에 히트맵을 형성하는 경우가 존재하여 링크가 포함되는 단어의 검출에는 영향이 없으나 각 글자에 대한 바운딩 박스를 검출하는 데에는 적합하지 않다.

최근의 단어 인식 과정은 일반적으로 변형(Transformation), 특징점 추출(Feature extraction), 시퀀스 모델링(Sequence modeling), 예측(Prediction)의 4 단계를 순차적으로 적용하여 진행된다. 이 중 본문에서 사용한 예측 모델인 Attention Network 의 경우 내부적으로 위치에 대한 중간 결과물인 비중값을 사용하여 시퀀스의 각 벡터의 문자를 찾게 되는데 이 값을 단독으로 사용하여 이미지의 크기와 매핑하면 해당 벡터의 어텐션 중심을 구할 수 있다. [6]

3. 제안하는 모델

한국어 이미지에서 각 글자의 바운딩 박스를 의도대로 찾지 못하는 문제를 해결하기 위해 Attention Network 에서 얻어지는 중심점 정보와 CRAFT 모델에서 얻어지는 히트맵 정보를 활용하여 기존 Attention Network 혹은 CRAFT 에서 얻을 수 있는 글자단위 바운딩 박스를 개선하는 모델을 제안한다.



(그림 1) 제안하는 모델 모식도

먼저 CRAFT 모델을 활용한 단어 분리(word segmentation) 과정에서 발생하는 히트맵 정보를 메모리에 저장해두는 과정이 필요하다. 단어 분리 과정과 단어 인식(word recognition) 과정은 서로 독립적인 과정으로써 단어 분리 과정의 결과물인 각 단어의 바운딩 박스를 단어 인식 과정의 입력으로 사용하기 때문에 별다른 저장과정이 없다면 중간 결과물인 히트맵은 버려지게 된다.

이후 일반적으로 단어 인식에 사용되는 딥러닝 모델을 활용하되 최종단계인 예측 단계를 Attention Network 를 사용하여 최종 결과물을 도출한다. 이 때 네트워크의 중간 결과물인 위치에 대한 비중값을 저장하여 각 글자의 위치에 대한 정보를 얻을 수 있도록 한다. 여기서 얻은 비중값을 통하여 현재 단어 바운딩 박스 내에서 찾은 각 글자의 중심점을 찾고 이를 원본 이미지에 대응하는 위치로 매핑한다.

위의 과정을 거쳐 각 글자의 중심점을 찾은 후 최종적으로 각 글자의 바운딩 박스를 찾는 과정이 필요하다. 처음 단어 분리 과정에서 저장한 히트맵에서 현재 단어 바운딩 박스에 해당하는 영역을 추출한 후 히트맵에서 나타난 경계 중 각 글자의 중심점들의 중간에 해당하는 지점을 기준으로 가장 가까운 경계들로 글자를 분리하여 각 글자의 바운딩 박스를 구할 수 있다.

4. 실험 과정

기존에 존재하는 데이터셋은 모두 단어 단위의 문자 인식을 위한 데이터셋으로 각 글자의 바운딩 박스를 검증하기 위한 데이터셋은 존재하지 않기 때문에 SynthText 를 활용하여 각 글자의 바운딩 박스를 포함한 데이터셋을 생성하여 실험에 사용하였다.

실험에 활용한 데이터셋은 SynthText 를 통해 생성한 한글 500 장, 영어 500 장으로 구성된 데이터셋이며 각 이미지는 각 언어의 5-10 개의 단어로 구성된다.

실험에 활용된 Word Segmentation 모델과 Word Recognition 모델은 모두 ICDAR2013-2017 의 Training Set 을 사용하여 훈련되었다.

본 실험은 해당 모델에서 추론한 글자 단위 영역과 데이터셋의 라벨과의 IOU(Intersection over Union)을 성능 지표로 삼았다.

5. 실험 결과

<표 1>

데이터셋	한국어	영어	평균
IOU	0.715	0.627	0.671

위의 표는 SynthText 로 생성된 라벨에 포함된 글자 단위 영역과 모델에서 검출한 글자 단위 영역의 IOU 결과를 나타낸 것이다.

6. 결론

제안된 모델은 존의 CRAFT 만으로는 불가능한 여러 글자의 바운딩 박스를 찾는데 활용할 수 있으며 특히 영어보다 한국어에 특화된 것을 확인할 수 있다.

하지만 제안된 모델은 사용된 CRAFT 및 Attention Network 의 학습 정도에 매우 의존적인 경향을 보인다. 단어 단위 트레이닝 셋을 통해 학습된 해당 모델들은 모두 중심점이 실제 중심점에 비해 대체로 오른쪽으로 어긋나는 Attention Drift 현상[6]이 나타나게 되어 IOU 가 낮아지는 결과를 보이게 된다.

따라서 향후 추가적인 연구를 통해 이러한 Attention Drift 현상을 억제하고 더 정밀하게 각 글자의 영역을 검출할 수 있도록 개선할 예정이다.

참고문헌

[1] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: “single shot multibox detector.” In ECCV, 2016.

[2] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” CoRR, vol. abs/1311.2524, 2013.

[3] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection,” in CVPR, 2019.

[4] Chen-Yu Lee and Simon Osindero. “Recursive recurrent nets with attention modeling for ocr in the wild.” In CVPR, pages 2231–2239, 2016.

[5] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. “Robust scene text recognition with automatic rectification.” In CVPR, pages 4168–4176, 2016.

[6] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. “Focusing attention: Towards accurate text recognition in natural images.” In ICCV, pages 5086–5094, 2017.