글자 수 정보를 이용한 이미지 내 글자 영역 검출 방법

김영우, *김원준
건국대학교
kyw6541@konkuk.ac.kr, *wonjkim@konkuk.ac.kr

Scene Text Detection with Length of Text

Yeong Woo Kim, *Wonjun Kim
Konkuk University

요 약

딥러닝의 발전과 함께 합성곱 신경망 기반의 이미지 내 글자 영역 검출(Scene Text Detection) 방법들이 제안됐다. 그러나 이러한 방법들은 대부분 데이터셋이 제공하는 단어의 위치 정보만을 이용할 뿐 글자 영역이 갖는 고유한 정보인 글자 수는 활용하지 않는다. 따라서 본 논문에서는 글자 수 정보를 학습하여 효과적으로 이미지 내의 글자 영역을 검출하는 모듈을 제안한다. 제안하는 방법은 간단한 합성곱 신경망으로 구성된 이미지 내 글자 영역 검출 모델에 글자 수를 예측하는 모듈을 추가하여 학습을 진행하였다. 글자 영역 검출 성능 평가에 널리 사용되는 ICDAR 2015 데이터셋을 통해 기존 방법 대비 성능이 향상됨을 보였고, 글자 수 정보가 글자 영역을 감지하는 데 유효한 정보임을 확인했다.

1. 서론

최근 자율주행차와 같이 특정 장면 내의 객체들을 분석할 필요가 있는 산업들의 수요가 높아짐에 따라 이미지 내의 객체, 차선, 글자 영역 등을 탐지하고 인식하는 기술이 활발하게 연구되고 있다. 특히, 딥러닝의 빠른 발전과 함께 이미지의 특성을 더욱 정교하게 추출할 수 있는 합성곱 신경망 기반의 뼈대 신경망들과 고도화된 객체 탐지 알고리즘들이 등장하면서 딥러닝 기반의 이미지 내 글자 영역 검출 방법들의 성능이 꾸준히 높아지고 있다.

기존의 이미지 내 글자 영역 검출 방법들은 데이터셋에서 제공하는 이미지 내 단어들의 위치 정보만을 사용하며, 글자 수와 같이 글자 영역이 갖는 고유한 정보는 활용하지 않는다. 이미지 내의 글자 영역은 비슷한 모양의 글자들이 연속적으로 모여 하나의 단어를 이루기 때문에, 글자 수 정보는 글자 영역을 검출하는 데 있어서 중요한 정보로 사용된다. 신경망이 글자 영역의 글자 수를 예측하도록 학습하여 기존의 위치 정보만을 고려했을 때보다 더 효과적으로 글자 영역을 검출할 수 있다.

따라서 본 논문에서는 기존의 방법들에서 사용되지 않았던 글자 영역의 글자 수 정보를 신경망의 학습 단계에 활용한 새로운 이미지 내 글자 영역 검출 모듈을 제안한다. 제안하는 글자 수 예측 모듈은 감지된 글자 영역에 픽셀 단위로 글자 수 값을 예측하여 출력한다. 글자 수 예측 모듈은 학습 단계에서 글자 영역의 글자 수를 학습하고 추론 단계에서는 사용되지 않는다.

이미지 내 글자 영역 검출에 사용되는 ICDAR 2015 데이터셋을 통해 제안하는 모듈을 추가한 신경망을 학습하고 평가하였다. 실험 결과를 통해 글자 수 예측 모듈을 추가한 방법에서 F1-score 가 약 1% 상승했으며, 글자 수 정보가 글자 영역을 감지하는데 효과적으로 사용될 수 있음을 확인했다.

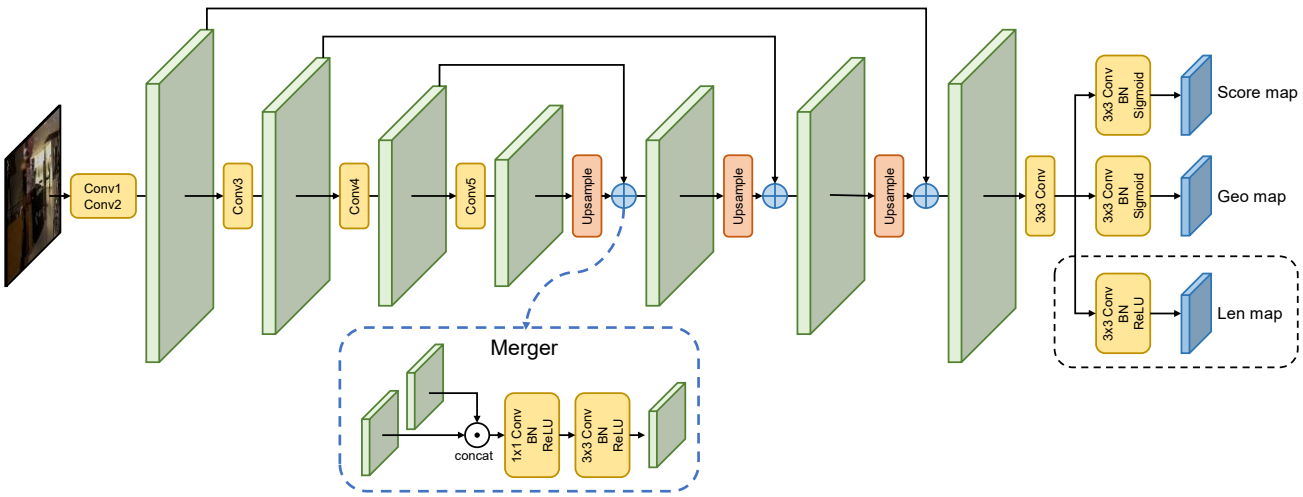

그림 1. 제안하는 글자 수 예측 모듈을 추가한 신경망의 전체 구조

2. 제안하는 방법

본 논문은 글자 수 정보를 학습 단계에서 활용할 수 있도록 기존의 이미지 내 글자 영역 검출 방법에 글자 수를 예측하는 모듈을 추가하였다. 제안하는 글자 수 예측 모듈은 글자 영역의 글자 수를 픽셀 단위로 출력하며 예측된 글자 수를 데이터셋의 단어로부터 얻어진 실제 글자 수와 비교하여 신경망을 학습한다. 제안하는 글자 수 예측 모듈을 추가한 전체 신경망의 구조는 그림 1 에 나타나 있다.

제안하는 모듈을 적용하기 위해 합성곱 신경망 기반의 이미지 내 글자 영역 검출 모델인 EAST [1]에 글자 수 예측 모듈을 추가했다. 실험에 사용된 EAST [1]의 전체 구조는 세 단계로 나눌 수 있다. 먼저 Feature Extractor Stem 단계에서는 뼈대 신경망을 통해 입력된 이미지의 특징을 나타내는 저해상도 Feature 로 인코딩한다. 기본적인 EAST [1]는 뼈대 신경망으로 PVANet [2]과 VGG16 [3]을 사용했으나, 제안하는 신경망에서는 ResNet-50 [4]을 사용했다. 다음으로 Feature-merging Branch 단계에서는 인코딩되었던 저해상도 Feature 를 다시 입력 이미지의 1/4 크기가 되도록 디코딩한다. 이때, 다양한 크기의 글자 영역을 고려하기 위하여 인코딩 부분과 디코딩 부분의 Feature 들을 병합한다 (그림 1 의 Merger). 마지막으로 Output Layer 단계에서는 글자 영역의 경계 상자를 추론하는데 사용되는 Score Map 과 Geo Map 을 출력한다. Score Map 은 현재 픽셀이 글자 영역에 속해 있을 확률을 나타내며, Geo Map 은 예측되는 경계 상자의 네 변과 현재 픽셀까지의 거리 그리고 예측된 경계 상자가 기울어진 각도를 나타낸다. 제안하는 글자 수 예측 모듈은 Output Layer 에 추가된다. 모듈은 3x3 크기의 합성곱 계층과 배치 정규화, 그리고 ReLU 활성화 함수로 구성된다. 글자 수 예측 모듈은 입력 이미지의 1/4 크기를 갖는 Len Map 을 출력한다. Len Map 의 각 픽셀값은 그 위치에 대응되는 글자 영역의 예측된 글자 수이다. 만약 해당 픽셀이 글자 영역에 속하지 않았다고 예측될 경우 0 을 출력한다. 예측된 글자 수는 데이터셋의 단어로부터 계산된 글자 수와 비교하여 L1 손실함수를 적용한다.

3. 실험 결과

제안하는 방법은 이미지 내 글자 영역 검출에 널리 사용되는 ICDAR 2015 데이터셋을 이용해 학습 및 평가했다. ICDAR 2015 데이터셋은 구글 글라스를 이용해 백화점 실내와 길거리 등을 촬영한 1,500 장의 이미지로 구성되어 있으며 각 이미지는 주로 간판과 표지판 등의 작고 다양한 형태의 글자 영역을 포함한다. 학습을 위한 배치 크기는 24 로 설정했으며 총 500 번을 반복하여 학습하였다.

Method	Precision	Recall	F1-score
EAST [1]	80.611 %	80.067 %	80.338 %
제안 방식	82.719 %	79.971 %	81.322 %

표 1. ICDAR 2015 데이터셋에서의 정량적 성능 평가

학습된 모델의 성능은 Precision(정밀도), Recall (재현율), F1-score 를 기준으로 평가했다. Precision 은 신경망으로부터 글자 영역으로 분류된 픽셀이 실제로 글자 영역인 비율을 나타내며, Recall 은 실제 이미지 내 글자 영역 중 신경망이 글자 영역이라고 예측한 픽셀의 비율이다. F1-score 는 Precision 과 Recall 의 조화평균으로 얻어진다.

(a) (b)

그림 2. (a) 글자 영역 검출 결과, (b) 글자 수 예측 결과

표 1 은 제안하는 방법을 ICDAR 2015 를 통해 정량적으로 평가한 결과이다. 글자 수 예측 모듈을 추가한 방법이 기존의 방법대비 약 1% 정도의 F1-score 가 향상된 것을 확인하였다.

그림 2 는 제안하는 방법을 이용한 이미지 내 글자 영역 검출 결과이다. 제시된 이미지 내의 예측된 글자 영역의 위치를 (a)에 경계상자로 표현했으며, (b)에는 제안하는 글자 수 예측 모듈의 출력을 시각화했다. (a)에 나타나 있는 각 글자 영역의 예측된 글자 수를 (b)에서 스펙트럼의 형태로 표현하였다. 두 번째 행의 예측 결과를 보면 'SALE'이라고 적혀 있는 글자 영역은 다른 글자 영역들과 비교해서 경계 상자가 크지만, 그 글자 수는 적게 예측되었다. 이를 통해 제안하는 글자 수 예측 모듈이 단순히 글자 영역 경계상자의 크기와 형태에 의존하지 않고 정상적으로 글자 영역의 글자 수를 인식하고 있음을 확인했다.

4. 결론

본 논문은 글자 영역의 고유한 정보인 글자 수를 이미지 내 글자 영역 검출에 사용하는 방법을 제안한다. 이는 기존의 합성곱 신경망 기반의 이미지 내 글자 영역 검출 모델에 글자 수를 예측하는 모듈을 추가하여 학습하는 방법으로 간단하게 적용할 수 있다. 합성곱 신경망 기반의 글자 영역 검출 모델인 EAST [1]에 제안하는 모듈을 적용하였고 효과적으로 이미지 내의 글자 영역을 감지하였다.

감사의 글

본 연구는 2022 년도 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 결과로 수행되었음. (No.2018-0-00213)

참고문헌

[1] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5551-5560.

[2] K. -H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "PVANET: Deep but lightweight neural networks for real-time object detection," *arXiv:1608.08021*. 2016.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.