문자 수준의 글자 감지 성능 개선을 위한 연구 동향 : CRAFT 사례를 중심으로

김용균, 박정호*
고려대학교 전기전자공학과

A Review on the Character-level Text Detection Methodology Focused on CRAFT

Yong-Gyun Kim, James Jungjho Pak*
Department of Electrical and Electronic Engineering, Korea University

Abstract - 딥러닝 기술이 발전하면서 이미지로부터 글자를 감지하고 인식하는 방법이 상당 부분 단순화되어 관련 연구가 지속해서 진행되고 있다. 이미지에서 글자를 감지하는 방법(Text detection)으로는 기존의 물체 감지 신경망을 수정하여 이용하거나, 픽셀/문자/단어/문장 수준에서 구성요소를 학습하는 등의 방법을 사용하고 있다. 문자 수준의 글자 감지는 단어, 문장 수준의 학습을 통한 글자 감지보다 단어 내에서 구성될 수 있는 문자의 다양한 형태에 강인한 특성을 보여준다. 문자 수준에서 글자를 감지하는 초기 모델인 CRAFT는 예측 성능이 Feature map의 크기에 의존적이며, 실시간 처리에 어렵다는 특징이 있다. 본 초록에서는 CRAFT의 동작 원리를 소개하고, 성능을 개선하는 방향으로서, Graph Convolution Network를 적용하는 연구와 Generative Adversarial Network를 이용해 글자의 방향을 학습한 연구, Back-bone을 더 가벼운 모델로 대체하여 연산 시간을 줄인 연구들을 분석함으로써 글자 감지 모델의 성능 개선 방향을 제시하고자 한다.

1. 서 론

차량 번호판 인식을 위해 사용되던 OCR(Optical Character Recognition) 기술을 넘어, 주변 사물에서 글자를 감지하는 STR(Scene Text Recognition)에 이르기까지, 기존의 글자 감지 작업을 위해서는 CCA(Connected Component Analysis) 등의 Rule-based algorithm으로 영역을 추출했다 [1]. 이후 AlexNet[2]와 같은 CNN(Convolution Neural Network) 기반의 신경망을 통해 물체의 감지, 분류 작업이 활성화되면서 딥러닝 기술이 발전하였다. 이후 물체 감지 신경망이었던 SSD(Single Shot Detector)를 수정한 Textboxes[3]를 통해 이미지를 입력받아 글자를 감지하고, 감지한 글자에서 글자를 인식해 출력하는 End-to-End 신경망이 발표되면서 글자 감지, 인식의 다양한 연구가 진행되고 있다. 최근의 이미지에서 글자를 추출하는 연구는 글자의 감지, 인식, 감지와 인식을 한번에 처리하는 End-to-End 세 개의 분야로 나누어져 진행되고 있다. 글자의 감지는 이미지에서 글자를 감지하고 위치 좌표를 localization 하는 작업을 의미하고, 글자의 인식은 특정 영역의 글자가 어떠한 글자인지 분류하는 작업을 의미한다.

글자의 감지는 문장, 단어, 문자, 픽셀 수준의 다양한 연구가 진행되었다. 문자 수준에서 글자를 감지하는 Bottom-up 방식의 CRAFT[4] 모델은 단어, 문장 수준으로 학습한 다른 모델에 비해 다양한 크기와 방향으로 출력된 문자에 대해 강인하게 감지하는 경향이 있다. 한편, Feature map의 크기에 의존적이고 기존의 모델에 비해 처리 속도가 상대적으로 느리다는 지적이 있다. 본 초록에서는 문자 수준으로 글자를 감지하는 CRAFT 모델을 사례로 학습 원리와 성능을 개선하는 방법을 검토한다.

2. 본 론

2.1 문자 수준의 글자 감지

CRAFT 모델은 자연적인 이미지에 임의의 글자가 합성된 SynthText Dataset으로 사전 학습을 진행한다. SynthText Dataset은 문자 수준으로 위치 좌표가 주석되어 있어, 문자 수준으로 학습할 수 있다. CRAFT 모델은 사전 학습 단계에서 단어를 구성하는 각각의 문자가 위치하는 영역, 문자와 문자 사이의 영역을 학습한다. 단어의 문자 영역뿐만 아니라, 문자와 문자 사이의 영역도 학습함으로써, 단어를 구성하는 문자들의 문맥적 정보를 예측할 수 있게 된다. 합성 이미지를 사전 학습한 신경망을 ICDAR[5], TotalText 등의 실제 이미지 Dataset에서 전이 학습한다. 일반적으로 ICDAR을 비롯한 실제 이미지 Dataset에는 단어/문장 수준의 주석은 되어 있으나, 문자 수준의 주석은 되어 있지 않으므로, 사전 학습과 동일하게 문자 단위로 학습할 수 없다. 따라서 각 문자의 명확한 위치는 알 수 없지만, 단어의 길이 정보와 같은 간접적인 정보를 이용하여 단어의 글자 수 대비 예측한 문자의 개수를 측정하여 반영하는 약한 지도학습의 형태로 전이 학습을 진행한다.

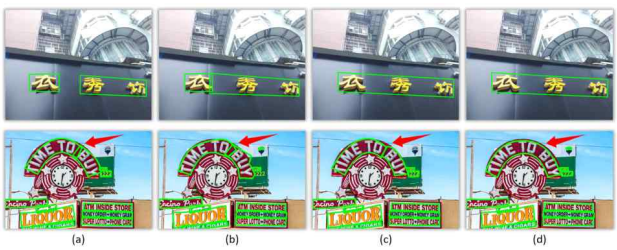
2.2 문자 수준의 글자 감지 모델 성능 개선

CRAFT는 단어 내 문자의 문맥적 정보를 학습하기 위해 문자와 문자 사이의 영역을 학습하는데, 글자 사이의 간격이 너무 큰 단어의 경우에는 영역의 학습 및 예측이 불가능하다. 또한, 모델의 FPS가 상대적으로 저조해 실시간 처리에 어려움이 존재한다. 해당 문제점들에 대해 성능을 개선한 연구를 소개함으로써, 문자 수준의 글자 감지 모델의 개선 방향을 제시한다.

2.2.1 문자 간 관계 검출성능 개선 방법

CRAFT는 Unet[6] 구조로 이루어진 CNN 기반의 신경망이다. 문자와 문자 사이의 영역을 학습할 때, Convolution feature map에 글자의 특징값이 기록되는데, 해당 픽셀에 인접한 지역적인 정보를 취합하므로, 문자 간의 거리가 Feature map보다 크면 인접한 다른 문자를 찾을 수 없어 학습 및 추론이 불가능하다.

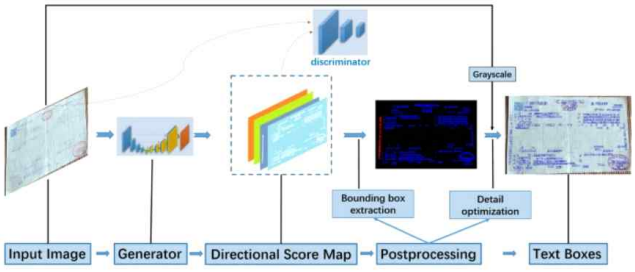
ReLaText는 2차원 Convolution이 아닌 GCN(Graph Convolution Network)를 이용해 단어의 문맥적 정보를 얻는 것을 제안한다. GCN을 사용하게 되면 문자의 간격이 Feature map보다 클 때뿐만 아니라 더 좁게 위치할 때도 분별할 수 있다.[7] 그림 1은 기존의 연구를 3종류로 나눠 (a), (b), (c)에 도시하고 GCN을 도입한 결과를 (d)에 나타낸 결과이다. CRAFT의 경우 (a)에 속하며 Feature map이 같은 단어 내의 다른 문자를 감지하지 못해 분리되었으나, GCN을 도입했을 때 (d)와 같이 문자 간의 관계를 감지하고 온전히 단어 전체를 감지한 결과를 얻었음을 볼 수 있었다.



<그림 1> GCN 도입을 통한 문자 간 관계 검출성능 개선 (a,b,c) 기존 연구 결과 (d)GCN 결과

2.2.2 문자의 방향 검출을 통한 성능 개선 방법

일상과 문서 속 글자 모두 인쇄 방향에 따라 다른 의미를 가질 수 있다. 글자를 감지하고, 인식 작업을 통해 이미지를 텍스트로 변환해야 할 때, 인쇄 방향의 정보의 유무는 후처리에 많은 영향을 줄 수 있다. DetectGAN[8]는 단어의 위치 좌표뿐만 아니라, 인쇄 방향의 정보를 얻는 방법을 제안한다. CRAFT의 경우, 이미지를 입력하고 Encoding, Decoding 과정을 거쳐 글자의 위치 정보를 얻어낸다. DetectGAN은 이미지를 입력받아서, 4방향의 이미지를 생성한다. 각기 다른 방향의 4장의 이미지에서 글자를 감지함으로써, 기존의 연구 방법보다 더 많은 정보를 통해 좋은 성능을 이끌어 낼 수 있다.



(a)

(b)

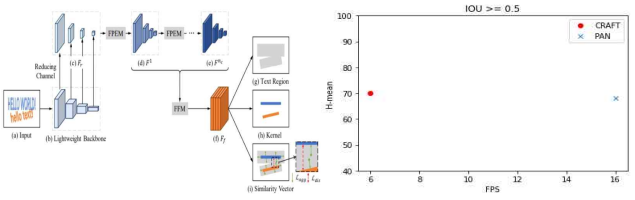
<그림 2> DetectGAN 도입을 통한 문자 감지 성능 개선
(a) Overview (b) DetectGAN 결과

2.2.2 FPS 개선을 위한 방법

PAN[9]은 FPN[10]의 신경망을 수정한 모델이다. VGG16 기반으로 만들어진 CRAFT의 Unet 신경망에 비해 PAN은 ResNet 기반으로 만들어져 더 빨리 연산 결과를 얻을 수 있다. PAN은 속도와 성능의 균형을 맞추기 위해 Decoder의 여러층에서 예측값을 얻을 수 있는 FPN을 도입하였다. 그 결과, 그림 3과 같이 PAN은 CRAFT와 동일한 데이터 셋에서 유사한 예측 성능을 보이면서도 실시간 처리가 가능함을 보여준다.

Method	ICDAR 13			ICDAR 15			COCO-Text		
	P	R	H-mean	P	R	H-mean	P	R	H-mean
CRAFT	72.77%	**77.62%**	75.12%	82.2%	**77.85%**	**79.97%**	56.73%	**55.99%**	**56.36%**
PAN	**83.83%**	69.13%	**75.77%**	**85.95%**	73.66%	79.33%	**59.07%**	43.64%	50.21%

(a)



(b) (c)

<그림 3> CRAFT와 PAN의 성능 비교
(a) 데이터셋 검출성능 (b) PAN overview (c) CRAFT와 PAN의 FPS 비교

3. 결 론

CRAFT 모델을 구성하는 Unet의 Convolution 신경망은 문자 사이의 문맥 정보를 저장하는 데 있어 물리적인 한계가 있다. Graph Convolutional Network를 도입하여 feature 사이의 관계를 정의하면, Feature map의 크기와 관계없이 문자 간의 관계 검출이 가능해지므로 글자 감지 성능 개선이 가능할 것으로 생각된다. 또한 DetectGAN에 소개된 방식과 같이 Generative Adversarial Network을 이용하여 이미지 내 글자의 인쇄 방향 정보를 얻을 수 있다면 다양한 다른 감지 모델과의 결합을 시도해 성능 개선을 기대해 볼 수 있을 것으로 생각된다. PAN의 경우에는 CRAFT와 예측성능이 비슷하면서 실시간 처리가 가능하다는 점에서 괄목할 만하다. CRAFT의 Unet은 PAN의 ResNet에 비해 연산 과정에서 느릴 뿐만 아니라, 파라미터의 개수도 수 배가량 많아 공간을 많이 차지하는 문제도 존재한다. 다만 PAN은 문자 수준에서의 글자 감지가 아닌 단어/문장 수준의 감지기이므로, CRAFT와 같은 문자 수준의 글자 감지 모델과 함께 앙상블 기법으로 활용하면 글자 감지 작업에서 좋은 성능을 기대해 볼 수 있을 것으로 생각된다.

[참 고 문 헌]

[1] Huang, W., Lin, Z., Yang, J., & Wang, J. (2013). Text localization in natural images using stroke feature transform and text covariance descriptors. In Proceedings of the IEEE international conference on computer vision (pp. 1241 - 1248).

[2] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

[3] Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. In AAAI (pp. 4161 - 4167).

[4] Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019b). Character region awareness for text detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 9365 - 9374).

[5] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras. Icdar 2013 robust reading competition. In ICDAR, pages 1484 - 1493. IEEE, 2013

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234 - 241. Springer, 2015

[7] Chixiang Ma, Lei Sun, Zhuoyao Zhong, Qiang Huo, ReLaText: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks, Pattern Recognition, Volume 111, 2021, ISSN 0031-3203

[8] Zhao, J., Wang, Y., Xiao, B. et al. DetectGAN: GAN-based text detector for camera-captured document images. IJDAR 23, 267 - 277 (2020).

[9] Wang, Wenhai, et al. “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network.” Proceedings of the IEEE International Conference on Computer Vision. 2019.

[10] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, ´ Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2017.