TextBoxes: A Fast Text Detector with a Single Deep Neural Network

by jjayy

논문 링크

https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewFile/14202/14295

Abstract

textboxes는 기존의 scene text detection을 end-to-end로 제시한다.
scene text를 single network를 통해 검출하며 NMS를 제외하고는 다른 후처리를 하지 않는 것이 특징이다.
기존의 다른 모델들에 비해 text localization의 정확도가 높고 빠르며, text detection만 수행하는 textboxes에 다양한 text recognition 모델들을 연계할 시 SOTA 모델들 사이에서 두각을 나타냈다.

1. Introduction

기존의 scene text detection은 '글자/단어 후보 생성, 후보 필터링, grouping'으로 이어지는 단계가 필요했기 때문에 모델 학습 중의 tuning이 매우 어려웠고 속도가 매우 느리다는 단점이 존재한다.
일상 상황에서 많이 접하는 상황에서는 높은 효용성을 가지고 있지만, 전통적인 OCR과 비슷한 task를 가지고 있음에도 실제 상황에서의 노이즈 및 왜곡이 많아 발전이 거의 없었다.
때문에 본 논문의 저자는 text detection을 빠르게 수행하기 위해 single network를 사용하여 object detection에서 큰 성능 향상을 보인 SSD의 구조를 사용하여 textboxes 모델을 구성했다.

2. Related Works

scene text detection은 character based와 word based, 그리고 text-line based로 나뉘고, textboxes는 그 중에서도 SSD 모델을 활용한 word based 모델이다. word based 모델은 R-CNN과 YOLO를 base로 한 모델들을 참고했다.
word based 모델로 detection을 진행했기 때문에 word recognizer로 성능이 뛰어난 CRNN을 뒷단에 연결하여 SOTA에서 end-to-end 모델로 높은 성능을 기록했다.

3. Textboxes

구조는 SSD와 동일, detection layer를 text detection에 맞게 textbox layer로 변형했다.

Architecture

cov1_1부터 conv4_3까지는 VGG-16의 구조를 사용하고 있으며, 이후 conv6와 pool11부터는 오직 convolution과 pooling으로만 이루어진 fully-convolutional-network이다. 이를 Base Network로 지칭한다.
SSD의 구조처럼 Textboxes의 Base Network에서 중간중간 map point에서 feature map을 가져온다.
이렇게 가져온 feature map마다 새로운 convolution을 수행해서 Textbox layer를 구성하게 된다. 즉 textboxes 모델의 output은 textbox layer.

Textbox Layers

SSD의 detection layer와 texbox layer의 만드는 차이는 default box에 있다.

기존의 SSD는 feature map에 따른 다양한 resolution을 바탕으로 다양한 aspect ratio의 box를 생성한다.
object와 다르게 높이에 비해 가로가 긴 특징을 가지고 있기 때문에 SSD와 다르게 가로의 길이가 긴 aspect ratio만 존재한다. 때문에 SSD와는 다르게 세로 간격에 공백이 생기게 되어 이를 채워주기 위해 vertical offset을 적용했다.
vertical offset은 기존에 설정된 높이의 1/2만 아래로 내리는 방식으로 설정한다.

이렇게 default box를 생성하기 때문에 각 feature map에서 뽑은 각각의 layer는 72개의 벡터를 가지게 된다.
기존에 설정된 12개의 aspect ratio를 바탕으로 predicted box를 생성하게 되는데, 이 box는 위치값(dx,dy)과 box의 너비높이(dw, dh) 총 4개의 offset으로 이루어져 있으며, 해당 box에 text가 있을 확률 및 없을 확률인 confidence(c1, c2) 두 가지를 추가하여 총 6개의 값으로 구성되어 있다. 때문에 12 x 6 = 72의 벡터 값을 가지게 된다.

Loss function

loss = location loss + confidence loss
location loss에는 smooth_L1 loss, confidence loss에는 softmax loss를 사용한다.

textbox layer의 모든 output을 사용하지 않고 default box을 Hard Negative Mining 과정을 통해 positive와 negative로 나누어 사용한다.
Jaccard Overlap 0.5를 기준으로 이상은 positive, 0.5 미만은 negative로 지정하며, 이 경우 negative가 현저히 많기 때문에 3배수만 남기고 나머지는 날린다.

location loss에서는 positive box만 사용하고 confidence loss는 positive box에서 c1, negatvie box에서는 c2를 선택적으로 사용하여 계산한다.

total loss = (location loss + confidence loss)/N (N은 positive box 개수)

Non Maximum Supression(NMS)

textboxes에 있는 유일한 후처리 방법으로 수많은 box들 중 진짜 positive box를 가려내는 방법이다.
가장 confidence가 높은 box와 IOU가 일정 이상인 box가 같은 text를 검출했다고 판단하여 지우는 방식이다.