WordFences: Text Localization and Recognition

ICIP 2017

Andrei Polzounov (Universitat Politecnica de Catalunya, Barcelona, Spain),

Artsiom Ablavatski (A*STAR Institute for Infocomm Research, Singapore),

Dr Sergio Escalera (Universitat de Barcelona, Barcelona, Spain),

Dr Shijian Lu (A*STAR Institute for Infocomm Research, Singapore),

Dr Jianfei Cai (Nanyang Technological University, Singapore)
Sponsors

- Institute for Infocomm Research (I²R), at Singapore’s Agency for Science, Technology and Research (A*STAR)
- CERCA Program, Government of Catalonia
Problem Description

- Text detection and recognition in natural scene imagery.
- Good test case problem for AI research and for uses in industry: mapping business from StreetView, translating menus or billboards, etc.
Motivation

- OCR can be used on scanned text.
- Natural images have a ton of variety in fonts, scales, kerning and features.
Proposed Solution

• 2-stage deep learning network
  • Find locations/ROIs (text localization) with CNN.
  • Detect characters (end-to-end text recognition) with RNN.

• 1st stage is the more difficult one. It is related to object recognition and semantic segmentation problems in Computer Vision.
Contributions

- Two main contributions:
  - Use a word separator – a „WordFence“ for delimiting individual words.
  - Using a novel weighted pixelwise softmax function for training the semantic segmentation.
Sample Detections
Related Work (CV)

- Maximally Stable Extremal Regions by Huang et al. (2014) – works by first using an MSER transform and then training a CNN.
- Stroke Width Transform by Epshtein et al. (2010) – an edge detection method, that relies on the fact that a given text font should have similar width/thickness for each stroke in a character.
- Edge Boxes by Zitnick and Dollár (2014) – simple object score based on the number of edges within a given sliding window. Sparse and fast to evaluate, but results could be better.
Related Deep Learning


Related Text Detection


- Jaderberg et al. (2014) – used CNN to generate high number of proposals and filter with Random Forest (HoG).

- He et al. (2016) – cascaded CNN with false positive rejection to detect textlines (no split).
Model Overview

ResNet 101

Image → conv → dilated conv → Bilinear Interpolation → background, WordFence, text → Word extraction via connected component analysis
Word Localization as Semantic Segmentation

• Semantic segmentation is a well known problem.

• Able to handle different scales using wide fields of view and multi-scale inference.

• Ground truth word separators created by dilation.
ResNet of Exponential Receptive Fields

- ResNets help to train huge networks without a vanishing gradient.
- Receptive fields can be enlarged using convolutional dilations in the ConvNets.
- Deep network + exponential receptive fields = effective multi-scale detection of different sized text.
Weighted Pixelwise Softmax Loss

• Background pixels are the majority. More emphasis for text and WordFence pixels is needed.

Algorithm 1 Pixelwise Weighted Softmax Loss

Require: Predicates after fusion $Pr$, ground truth labels $L$

1: $probs \leftarrow \text{Softmax}(Pr)$ $\triangleright$ pixel probabilities
2: $m \leftarrow \text{NumberOfUniqueLabels}(L)$
3: $n_1, n_2, \ldots, n_m \leftarrow \text{CountsOfUniqueLabels}(L)$ $\triangleright$ get counts of each label on a ground truth image
4: $loss \leftarrow -\sum \frac{1}{n_{gt}} \log(probs_{gt})$ $\triangleright$ weighted loss calculation
5: $\text{Backpropagate}(loss, \frac{1}{n_1}, \frac{1}{n_2}, \ldots, \frac{1}{n_m})$ $\triangleright$ loss backpropagation with normalization factors

• Algorithm allows us to rebalance per image weights on the fly.
WordFence vs No-WordFence

- WordFences act as penalization for merged words.
Text Datasets

- COCO-Text – subset of the popular MS-COCO dataset for object recognition (21 classes)
- SVT – Google Street View data
- SynthText – synthetic text mixed with scene images from Gupta et al.
## Localization Results

<table>
<thead>
<tr>
<th>Model</th>
<th>PASCAL VOC IoU = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICDAR11</td>
</tr>
<tr>
<td>Tian et al. [32]</td>
<td>0.89</td>
</tr>
<tr>
<td>Gupta et al. [5]</td>
<td>0.78</td>
</tr>
<tr>
<td>Jaderberg et al. [11]*</td>
<td>0.89</td>
</tr>
<tr>
<td>Gupta et al. [5]*</td>
<td>0.94</td>
</tr>
<tr>
<td>WDN (ours)</td>
<td>0.64</td>
</tr>
</tbody>
</table>

- Methods marked with * use multi-stage false-positive detectors.
The recognition stage uses proposals given from the detection network.

Recognition stage is based on the CRNN network by Shi et al. (2016).

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>IC11</th>
<th>IC13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neumann et al.</td>
<td>2013</td>
<td>0.45</td>
<td>-</td>
</tr>
<tr>
<td>Jaderberg et al.</td>
<td>2015</td>
<td>0.69</td>
<td>0.76</td>
</tr>
<tr>
<td>Gupta et al.</td>
<td>2015</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>WDN</td>
<td>2016</td>
<td>0.84</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Tiredness kills

A short break could save your life

<table>
<thead>
<tr>
<th>P=0.875, R=0.778, F=0.824</th>
<th>P=0.625, R=1.000, F=0.769</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;short break&quot; not split correctly</td>
<td>Some false positives</td>
</tr>
</tbody>
</table>
ICDAR 2013

P=1.000, R=0.889, F=0.941 | P=0.833, R=0.714, F=0.769
Some overlaps are <0.5 IoU

<table>
<thead>
<tr>
<th>P=1.000, R=0.400, F=0.571</th>
<th>P=0.300, R=0.750, F=0.429</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenging case - word is within another word</td>
<td>Too much word splitting, due to tricky font and spacing</td>
</tr>
</tbody>
</table>
Problems Encountered

• Going from segmentations to bounding boxes.
• Noisy detections and many false positives:
  • Many small detected regions.
  • Camera artifacts such as glare.
• Need for balancing precision vs recall.
Conclusion

- Text recognition as semantic segmentation
- WordFences as penalization
- SOTA recall on detection, which provides high quality samples to the recognition stage (which in itself is able to throw away false positives).
- SOTA F-scores on recognition
Future Work

• WDN relies on visual information to split words.

• Humans also use word semantics and memory.
  • “Raeding wrods with jubmled letetrs“.

• A smarter system would be able to read text directly from images without a midpoint CV representation.
  • Learn directly from neural net to a dictionary word output?