DUET: Detection Utilizing Enhancement for Text in Scanned or Captured Documents

Eun-Soo Jung*, HyeongGwan Son*, Kyusam Oh, Yongkeun Yun, Soonhwan Kwon, and Min Soo Kim

* Equal contributions

SAMSUNG SDS
Most of the previous studies on text detection focus on text in the wild\textsuperscript{[1-6]}

Text in the wild (or scene text)
- (Relatively) More labeled data\textsuperscript{[7-11]}

Text in documents
- Very few labeled data\textsuperscript{[12]}
  → insufficient to train deep neural model

Text detection models trained with scene text: *Limited* to cover features of document images

\textbf{Goal:} Deep neural text detector with improved accuracies for document images
**Training data**

- Data synthesizing
- *Overcome* the shortage of labeled document image data

**Learning strategy**

- Multi-task learning\[^{13-16}\]
- Weakly-supervised learning\[^{17-18}\]
- *Overcome* various types of noise in document images
- Fully convolutional encoder-decoder structure
  - Feature Pyramid Network (FPN)[19]
  - ResNeXt101[20]

- Multi-task learning
  - Branch for text detection → detection loss ($L_D$)
  - Branch for text enhancement
    - pixel-level binary classification (text/non-text) → enhancement loss ($L_E$)
  - Multi-task loss
    → $L = \lambda_1 L_D + (1 - \lambda_1) L_E$

($\lambda_1$: balancing parameter between $L_D$ and $L_E$)
Training

- **Phase 1**
  - Training *synthesized* data
  - Fully-supervised learning

- **Phase 2**
  - Training *synthetic* data and *real* data
  - Detection GTs: given
  - Enhancement GTs: ??
    \[ \rightarrow \text{weakly-supervised learning} \]
    - Binarized $\text{GT}_D (\text{GT}_E')$
      \[ \rightarrow \text{false positive loss (} L_{FP} \text{)} \]
Phase 1
- Training synthesized data
- Fully-supervised learning

Phase 2
- Training synthetic data and real data
- Detection GTs: given
- Enhancement GTs: ??
  → weakly-supervised learning
  - Binarized \( \text{GT}_D (\text{GT}_E') \)
  → false positive loss \((L_{FP})\)
  - Using interim trained detector
    → detection loss for enhanced output \((L_{D2})\)

Multi-task loss
→ \( L = \lambda_1 L_D + (1 - \lambda_1)L_E \)
\( L = \lambda_1 L_D + (1 - \lambda_1)(\lambda_2 L_{D2} + (1 - \lambda_2)L_{FP}) \)

\((\lambda_2: balancing parameter between L_{D2} and L_{FP})\)
Experiment and Results

- Benchmark database
  : Form Understanding in Noisy Scanned Documents (FUNSD) dataset\textsuperscript{[12]}
  
  (120 train, 29 validation, 50 test)

- Comparisons with previous studies

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract\textsuperscript{[21]}</td>
<td>45.4</td>
<td>68.0</td>
<td>54.4</td>
</tr>
<tr>
<td>Google Vision (API)\textsuperscript{[22]}</td>
<td>79.8</td>
<td>62.0</td>
<td>69.8</td>
</tr>
<tr>
<td>Faster R-CNN\textsuperscript{[23]}</td>
<td>70.4</td>
<td>84.8</td>
<td>76.9</td>
</tr>
<tr>
<td>EAST\textsuperscript{[4]}</td>
<td>51.6</td>
<td>84.0</td>
<td>63.9</td>
</tr>
<tr>
<td>CRAFT\textsuperscript{[5]}</td>
<td>91.2</td>
<td>84.2</td>
<td>87.6</td>
</tr>
<tr>
<td>CharNet\textsuperscript{[6]}</td>
<td>95.1</td>
<td>57.4</td>
<td>71.6</td>
</tr>
<tr>
<td>DUET (proposed)</td>
<td>93.1</td>
<td>92.2</td>
<td>92.6</td>
</tr>
</tbody>
</table>

- IoU threshold @ 0.5
- Results from \textsuperscript{[21]-23} and \textsuperscript{[4]} are provided by \textsuperscript{[12]}
- For \textsuperscript{[5]} and \textsuperscript{[6]}, the trained models and test codes from the original studies were used

Output examples

Detection

Enhancement
- Text detector for document images

- Enhance robustness for noisy documents
  - Auxiliary task: text enhancement

- Overcome data insufficiency
  - Synthesized training data
  - Weak-supervision to train enhancement of the real training data