TRIE: End-to-End Text Reading and Information Extraction for Document Understanding

Peng Zhang¹
Yunlu Xu¹
Zhuanzhan Cheng²¹
Shiliang Pu¹
Jing Lu¹
Liang Qiao¹
Yi Niu¹
Fei Wu²

1. Hikvision Research Institute, Hangzhou, China
2. Zhejiang University, Hangzhou, China
Background

◆ VRD (Visually Rich Document)

- Text
  - plain text

- Visual
  - layout
  - tabular
  - font size

◆ VRD Understanding

VRD Understanding Algorithm

Receipt.

Entities to extract.

- Name: OJC MARKETING SDN BHD
- Date: 15/01/2019
- Total: 193.00

Receipt.
**Problem of current framework**

- **Limitation 1:** Limited visual features in IE.
  - Keep: \(x_0, y_0, \ldots, x_3, y_3, '15/01/2019'\)
  - \(x_0, y_0, \ldots, x_3, y_3, '193.00'\)
  - Lost: Font, Color, Layout etc.

- **Limitation 2:** Ignoring relations between OCR & IE.

**Motivation**

- **Advantage 1:** Multimodal fusion in IE.
  - Keep: \(x_0, y_0, \ldots, x_3, y_3, '15/01/2019'\)
  - \(x_0, y_0, \ldots, x_3, y_3, '193.00'\)
  - Font, Color, Layout etc.

- **Advantage 2:** Bridging OCR & IE,
  - Forward: OCR boost IE
  - Backward: IE boost OCR
Method

◆ Overall Architecture.

Forward Process
Backward Process

Text Reading
Information Extraction

Multimodal context block

Context Feature

Visual Context

Textual Context

Hidden Layers

Textual 
features

Visual features

Position features

Shared
Convolutions

Text Detection
ROI Align
Encoder
Decoder

Textual Context

2018

abc

Martin

Name: Martin
E-mail: abc123***@gmail.com
Edu-period: 2018.09-2019.06

Phone: 132****6116
E-mail: abc123***@gmail.com

2018.09–2019.06
Computer Science and Technology
Fu*** University
GPA: 3.88 / 4.0

2014.09–2018.06
Computer Science and Technology
Bachelor of Science
Fu*** University
GPA: 3.83 / 4.0

Email: ***@gmail.com
Experiment

◆ Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training</th>
<th>Testing</th>
<th>Entities</th>
<th>Layout</th>
<th>Text Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxi Invoices</td>
<td>4000</td>
<td>1000</td>
<td>9</td>
<td>Fixed</td>
<td>Struct</td>
</tr>
<tr>
<td>SROIE</td>
<td>626</td>
<td>347</td>
<td>4</td>
<td>Variable</td>
<td>Struct</td>
</tr>
<tr>
<td>Resumes</td>
<td>1978</td>
<td>497</td>
<td>6</td>
<td>Variable</td>
<td>Semi-struct</td>
</tr>
</tbody>
</table>

Dataset Statics

Dataset Examples

(a) Taxi Invoices

(b) Receipt

(c) Resume

◆ Performance Summary

<table>
<thead>
<tr>
<th>Entities</th>
<th>Chargrid(TR)</th>
<th>NER(TR)</th>
<th>GCN(TR)</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>89.4</td>
<td>94.5</td>
<td>97.0</td>
<td>98.2</td>
</tr>
<tr>
<td>Number</td>
<td>85.3</td>
<td>92.4</td>
<td>93.7</td>
<td>95.4</td>
</tr>
<tr>
<td>Date</td>
<td>89.8</td>
<td>82.5</td>
<td>93.0</td>
<td>94.9</td>
</tr>
<tr>
<td>Pick-up time</td>
<td>82.9</td>
<td>60.0</td>
<td><strong>86.3</strong></td>
<td>84.6</td>
</tr>
<tr>
<td>Drop-off time</td>
<td>87.4</td>
<td>81.1</td>
<td>91.0</td>
<td><strong>93.6</strong></td>
</tr>
<tr>
<td>Price</td>
<td>93.0</td>
<td>94.5</td>
<td>93.6</td>
<td>94.9</td>
</tr>
<tr>
<td>Distance</td>
<td>92.7</td>
<td>93.6</td>
<td>91.4</td>
<td><strong>94.4</strong></td>
</tr>
<tr>
<td>Waiting</td>
<td>89.2</td>
<td>85.4</td>
<td>91.0</td>
<td>92.4</td>
</tr>
<tr>
<td>Amount</td>
<td>80.2</td>
<td>86.3</td>
<td>88.7</td>
<td><strong>90.9</strong></td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td>87.77</td>
<td>85.59</td>
<td>91.74</td>
<td><strong>93.26</strong></td>
</tr>
</tbody>
</table>

Taxi Invoices Dataset

ICDAR2019 SROIE Dataset

<table>
<thead>
<tr>
<th>Setting</th>
<th>Model</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting 1: Prediction of bboxes and transcript of texts</td>
<td>Chargrid(TR)</td>
<td>78.24</td>
</tr>
<tr>
<td></td>
<td>NER(TR)</td>
<td>69.09</td>
</tr>
<tr>
<td></td>
<td>GCN(TR)</td>
<td>76.51</td>
</tr>
<tr>
<td></td>
<td>Our model</td>
<td><strong>82.06</strong></td>
</tr>
<tr>
<td></td>
<td>LayoutLM[54]</td>
<td>95.24</td>
</tr>
<tr>
<td></td>
<td>PICK[58]</td>
<td>96.12</td>
</tr>
<tr>
<td></td>
<td>Our model</td>
<td><strong>96.18</strong></td>
</tr>
</tbody>
</table>
Experiment

◆ Performance Summary

<table>
<thead>
<tr>
<th>Entities</th>
<th>Chargrid/TR</th>
<th>NER(TR)</th>
<th>GCN(TR)</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>43.4</td>
<td>42.7</td>
<td>42.9</td>
<td>45.7</td>
</tr>
<tr>
<td>Phone</td>
<td>87.0</td>
<td>86.6</td>
<td>83.3</td>
<td>88.0</td>
</tr>
<tr>
<td>E-mail</td>
<td>70.9</td>
<td>69.6</td>
<td>68.0</td>
<td>74.9</td>
</tr>
<tr>
<td>Edu-period</td>
<td>77.1</td>
<td>68.7</td>
<td>62.2</td>
<td>81.4</td>
</tr>
<tr>
<td>University</td>
<td>74.7</td>
<td>86.0</td>
<td>82.3</td>
<td>87.4</td>
</tr>
<tr>
<td>Major</td>
<td>72.1</td>
<td>80.4</td>
<td>78.7</td>
<td>80.8</td>
</tr>
<tr>
<td>Avg</td>
<td>70.87</td>
<td>72.33</td>
<td>69.57</td>
<td>76.3</td>
</tr>
<tr>
<td>Speed(fps)</td>
<td>1.13</td>
<td>1.69</td>
<td>1.62</td>
<td>1.76</td>
</tr>
</tbody>
</table>

Resumes Dataset

◆ Discussion

- Effects of multimodal features on IE.
- Effects of E2E framework on text reading.
- Effects of layers and heads in textual context block.
See Far, Go Further