1. Overview

**“Unsupervised” text recognition using no paired supervisory data**

**Contributions**

- Unsupervised training of text recognition networks from scratch, from unaligned collections of text-images and text-strings harvested from corpora.
- Excellent accuracy on synthetic and real scans of books.

2. Supervised Methods

- Sequence recognition models trained on millions of (image, text-string) paired training examples.

3. Unsupervised Training

Text recognition can be broken into two sub-problems.

We exploit the structure in language, and solve the two jointly with end-to-end training.

**Model**

- Adversarial training to align the predicted strings with lexically valid strings sampled from target corpora.
- Local recognizer, global discriminator to avoid fooling the discriminator with valid strings without recognizing.

<table>
<thead>
<tr>
<th>Synthetic Data</th>
<th>Manual Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>New engine for each domain</td>
<td>Expensive</td>
</tr>
<tr>
<td>Complex pipelines</td>
<td>Not scalable</td>
</tr>
<tr>
<td>Domain gap</td>
<td></td>
</tr>
</tbody>
</table>

4. Experiments

**Synthetic Data**

- Fixed-width font
- Non-fixed width font
- Variable word spacing
- See-through from back
- Control over nuisance factors
- WMT NewsGroup text source
- 100K training, 1K testing

**Google Books Scans**

- Different case (small / capital)
- Italics / noise / fading etc.
- 3K training, 300 test lines
- Residual RNN lends pliability to features to counter varying character widths and spacing.

**Generalisation to different lengths**

- Fully-convolutional recognizer generalizes to lengths different (= 3, 5, 7, 9, 11, 13, 32, 48) than training length (= 24).

**Effect of length on convergence**

- Short lengths (3, 5): no convergence
- Longer sequence \(\rightarrow\) faster convergence

**Which character is learnt first?**

- Symbols are learnt roughly in the order of their frequency. Spearman’s rank correlation coefficient: \(p = 0.80, p\)-value < 1e–5.

**Text corpus ablation**

- Completely unrelated lexicon (#3): small adverse effect
- Related lexicon (#2): no such effect

Financial support was provided by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems Grant EP/L013885/1, EPSRC Programme Grant Sedabtyle (EP/M013774/1), ERC 677195 - IDIU, and the Clarendon Fund scholarship.

---

Learning to Read by Spelling Towards Unsupervised Text Recognition

Ankush Gupta
Andrea Vedaldi
Andrew Zisserman

Visual Geometry Group (VGG)
University of Oxford