Learning to Read by Spelling
Towards **Unsupervised** Text Recognition

Ankush Gupta      Andrea Vedaldi      Andrew Zisserman

Visual Geometry Group (VGG)
University of Oxford

ICVGIP 2018, Hyderabad
Text Recognition

Imaged Text  ➔  ASCII Text

tion of regular fits of the gout, one or more joints

Assumes \textbf{localisation} is given
• Word / line level bounding boxes
Let’s solve this
Text Recognition: Sequence Learning 101

Lots of paired data

ConvNet

Sequence Model (e.g. RNNs)

Paired Image / Annotations

tion of regular fits of the gout, one or more joints
part of the brain; the tunica arachnoides was rated and of a pale yellow colour; and with the
Text Recognition: **Paired Data?**

**Manual labour**
- Expensive
- Boring...

**Synthetic Data**
- New engine for each domain
- Complex pipelines
- Domain gap

Jaderberg et al., NIPS DLW 2014

Gupta et al., CVPR16, BMVC18
Can we do without *paired* data?
Language is **highly structured**

- There are \(~8\) billion random strings of length 7 (26 characters) but only 15K are valid English strings.

- The frequency of characters and words, and their co-occurrence (n-grams etc.) further constrain the output.

**We leverage this structure for supervision.**
Method
Text Recognition

2 sub-problems

1. **Segment** text image into characters, and cluster to **consistent class**

2. **Assign** each cluster to correct “character” label

→ Solve for a $|\mathcal{A}| \times |\mathcal{A}|$ permutation matrix

where, $\mathcal{A}$ is the alphabet, e.g.: 26 English letters \{a, b, c, ..., z\}
Unpaired Text Recognition
Learn from unaligned text corpora, and text-images

28:
English letters
\{a, b, c, ..., z\}
+ space
+ pad

| \mathcal{A} |

Fully-Conv Recognition Net

images with \leq L characters

Softmax each position

Discriminator

Adversarial Loss

“fake” text

“real” text

Valid Language Strings
e.g. from: WMT, NewsGroup etc.

English letters \{a, b, c, ..., z\} + space + pad

Softmax each position

Discriminator

Adversarial Loss

“fake” text

“real” text

Valid Language Strings
e.g. from: WMT, NewsGroup etc.

Images of regular fits of the gout, one or more joints
part of the brain; the tunica arachnoides was
rated and of a pale yellow colour; and with the
Unpaired Text Recognition

**read** (riːd) *verb* • Look at and comprehend the meaning of (written or printed matter) by interpreting the characters or symbols of which it is composed.

**spell** (spɛl) *verb* • Write or name the letters that form (a word) in correct sequence.

— *Oxford Dictionary of English*
**Pitfall: Uncorrelated Predictions**

The “recognizer” can generate valid text without “recognizing”

→ Fool the discriminator without solving the task
  
  *e.g.* use “text-image” as noise → learn generator for valid English strings

```
part of the brain; the tunica arachnoides was
```

```
“this is a valid English string which looks real to the discriminator”
```
Uncorrelated Predictions: **Solution**

Global discriminator checks validity of the entire sentence.

Local recognizer restricted receptive field (~3 characters).

No Reconstruction unlike CycleGAN.

Because text → image is highly ambiguous.
Experiments
Synthetic Text Images

- Fixed-width font
- WMT Newscrawl text source (EMNLP datasets)
- Control over nuisance factors → used for analysis
- 100K training sequences, 1K test sequences
Real Text Images

- Google Books scans
- **Non-fixed width font**
- **Varying word spacing** due to alignment
- “See-through” from back
- Different case (small / capital)
- Italics / noise / fading etc.
- 3K training lines, 300 test lines (no overlapping pages)

- Use Google OCR output as unaligned “text source”
Training Strategy
Sample images and valid strings independently

Adversarial Loss

Discriminator

fully-conv recognition net

softmax predictions

one-hot strings

Valid Language Strings

Text:

tion of regular fits of the gout, one or more joints part of the brain; the tunica arachnoides was rated and of a pale yellow colour; and with the
Results
Synthetic Text Images

- ~99% character accuracy
- ~95% word accuracy
- Trained on 24-length sequences → test on 3, 5, 7, 9, 11, 24, 32, 48 → generalization to other lengths
Real Text Images

Why?

Varying spacing / non-fixed width font challenging for fully-conv. recognizer

45% word accuracy!
Real Text Images

Why?

Varying spacing / non-fixed width challenging for fully-conv. recognizer

→ Let features travel using a “skip-RNN” in the last layer
Real Text Images

85% character accuracy (now vs. 45% before)!

96.2% character accuracy.
Real Text Example

Ground Truth

The different forms in which this disease appears have rendered it necessary to divide it into regular and irregular gout. In the former, the attacks of which are known by the denomination of regular fits of the gout, one or more joints of the extremities become inflamed, painful, and tender and frequently in an exquisite degree. A symptomatic fever proportioned to the degree of pain and inflammation, with evening exacerbations accompany the other complaints which distress the patient for uncertain periods sometimes for several weeks when the fit goes off, the joints which have been the seat of the disease are always found to have become rigid and inflexible in proportion to the degree in which the disease has existed in them; frequently remaining enlarged and incapable of free motion for a considerable time on the other hand the patient at the same time experiences so perfect an exemption from disease as generally to lead to the opinion that the fit has occasioned the most salutary changes in the system.

In the irregular gout, the affection of the joints is much less confined than in the former. Sometimes it leaves the joints at first attacked and fixes on some distant parts; and sometimes after harassing the patient by making a circuit in

Image

Prediction

the different forms in which this disease appears have rendered it necessary to divide it into regular and irregular gout. In the former, the attacks of which are known by the denomination of regular fits of the gout, one or more joints of the extremities become inflamed, painful, and tender and frequently in an exquisite degree. A symptomatic fever proportioned to the degree of pain and inflammation, with evening exacerbations accompany the other complaints which distress the patient for uncertain periods sometimes for several weeks when the fit goes off, the joints which have been the seat of the disease are always found to have become rigid and inflexible in proportion to the degree in which the disease has existed in them; frequently remaining enlarged and incapable of free motion for a considerable time on the other hand the patient at the same time experiences so perfect an exemption from disease as generally to lead to the opinion that the fit has occasioned the most salutary changes in the system.

In the irregular gout, the affection of the joints is much less confined than in the former. Sometimes it leaves the joints at first attacked and fixes on some distant parts; and sometimes after harassing the patient by making a circuit in

in the irregular gout the affection of the joints is much less confined than in the former. Sometimes it leaves the joints at first attacked and fixes on some distant parts; and sometimes after harassing the patient by making a circuit in
Analysis
Effect of Sequence Length on Training Convergence

Training with sequences of different lengths:

- Short lengths 3-5: no convergence
- Longer sequence → faster convergence 13 > 11 > 9 > 7
• Symbols are learnt roughly in the order of their frequency.
(Spearman’s rank correlation coefficient $\rho = 0.80$, p-value $< 1e^{-5}$).

Which symbol is learnt first?

Symbols are learnt roughly in the order of their frequency.
(Spearman’s rank correlation coefficient $\rho = 0.80$, p-value $< 1e^{-5}$).
• Completely unrelated lexicon (#3): small adverse effect
• Related lexicon (#2): no such effect
Extensions

• Not text-image specific
  → apply to **any input** domain, as long as **output is still language**

• Examples:
  • Speech
  • Sign language / gestures
  • Lip reading
Any Questions?

Learning to Read by Spelling
Towards **Unsupervised** Text Recognition

Ankush Gupta      Andrea Vedaldi      Andrew Zisserman

Visual Geometry Group (VGG)
University of Oxford

ICVGIP 2018, Hyderabad
Which symbol is learnt first?
Learning Dynamics

brought to view by dissection, it was discovered

1. First {space} / segmentation into words is learnt
2. Next, common words like {to, it}
3. Last, less frequent characters like {v, w}.

The last transcription also corresponds to the ground-truth (punctuations are not modelled). The colour bar on the right indicates the accuracy (darker means higher accuracy).