# Reading Scene Text in Deep Convolutional Sequences

Pan He*, Weilin Huang*, Yu Qiao, Chen Change Loy, Xiaoou Tang

Shenzhen Key Lab of Comp. Vis and Pat. Rec.,
Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences

Multimedia Laboratory,
Department of Information Engineering,
The Chinese University of Hong Kong

# Text Recognition

Traditional OCR —— texts from printed documents

Largely black-and-white, nearly horizontal text line

In 1939 the Yorkshire Parish Register Society, of which the Parish Register Section of the Yorkshire Archaeological Society is the successor (the publications having been issued in numerical sequence without any break) published as its Volume No. 108 the entries in the Register of Wensley Parish Church from 1538 to 1700

# TEXT RECOGNITION

Scene text understanding —— texts from natural scene images

Larger diversity of text patterns

- low resolution
- low constrast
- blurring

Highly complicated background clusters

# Text Recognition

Numerous practical applications

License Plates Recognition

Street View House Number Recognition

Automated CAPTCHA Character Recognition

Text-based Image Retrieval


Keyword: baby



# Text Recognition

Retrieve text string from cropped word image




TRANSFORMERS HEPP Royal

# TEXT RECOGNITION

State-of-the-art —— character-level classification

DeepFeatures [ Jaderberg, Vedaldi, and Zisserman 2014. ECCV ]

- 2D character probability map and multiple visual cues

PhotoOCR [ Bissacco et al. 2013. ICCV ]

- static n-gram and word language model

# TEXT RECOGNITION

State-of-the-art —— word-level classification

Sub-regression [ Almazan et al. 2014. TPAMI ]

- subspace regression
- embedded text attributes

Dictionary-based [ Jaderberg et al. 2015. IJCV ]

- multi-class classification
- 4096 FC feature for 90K word classes

# TEXT RECOGNITION

State-of-the-art —— unconstrained text recognition

[ Jaderberg et al. 2015. ICLR ]

- character sequence model
- 4096 FC feature for up to 23 output classifiers of 37 classes

# TEXT RECOGNITION

Limitation

Character-level methods

- difficult character separation
- heuristic post-processing
- Ignore context info

Word-level methods

- ignore spatial info (sub-regression)
- constrained recognition (dictionary-based)
- length-limited recognition (23-multi-softmax)

# TEXT RECOGNITION

## Motivation

Recent advancement on recurrent neural network community

- image caption
- speech recognition
- handwritten digit recognition

Scene text recognition is similar to speech recognition

- context information
- variable length

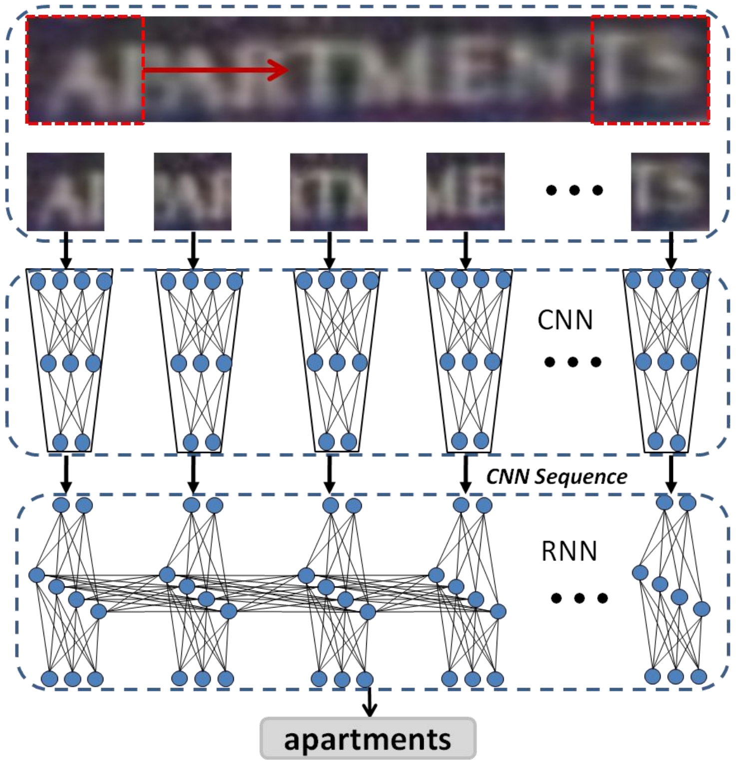
# OVERVIEW

Deep-Text Recurrent Networks(DTRN) for text recognition

- Sequence generation with Maxout CNN
- Sequence labelling with RNN

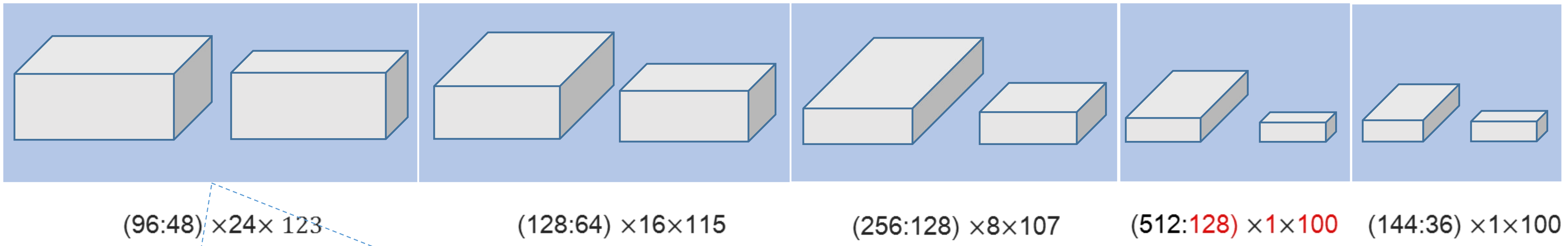
Experiments

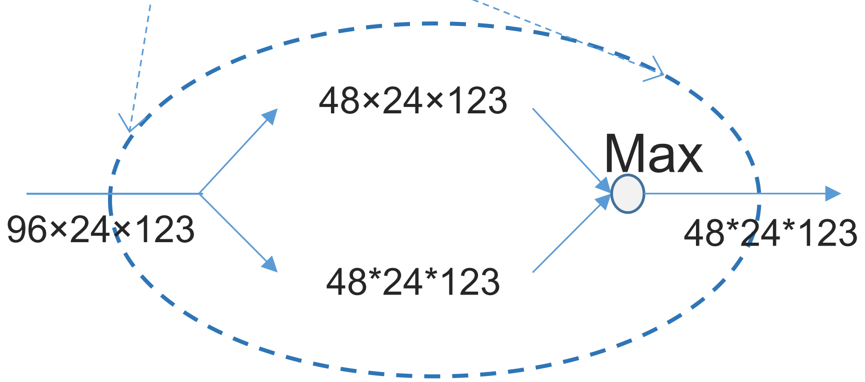
- DTRN vs DeepFeatures
- Comparisons with State-of-the-Art



# Sequence Generation Model

Suppose input is 32*131, 5 maxout convolutional layers, 128×100 (100 is T for RNN ) as the sequential feature

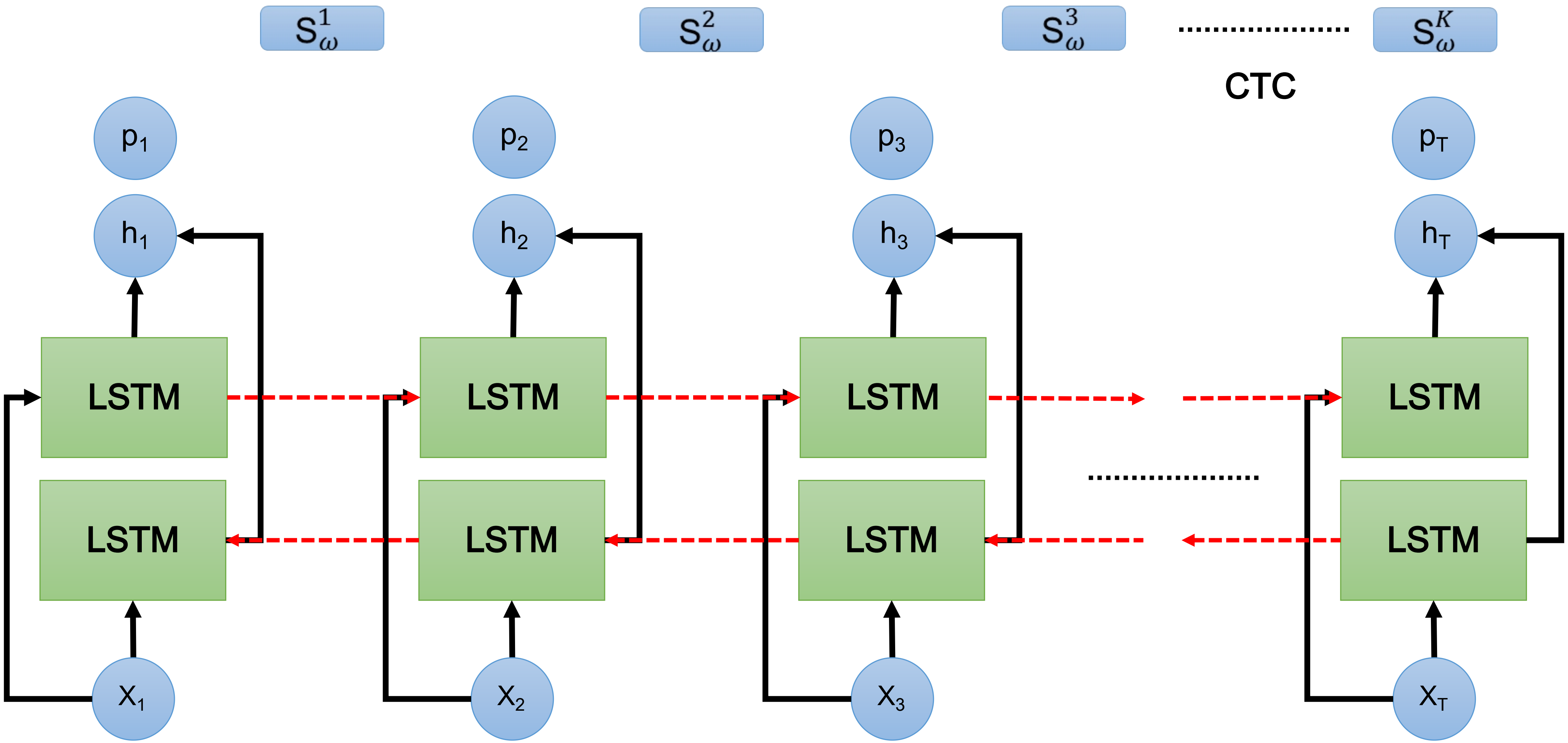



Maxout constructed in convolutional network

Pooling across channels

# Sequence Labelling Model



# Sequence Labelling Model

## Variables

$\mathrm{x} = \{x_1, x_2, x_3, \ldots, x_T\}$, $\quad$ $128 \times \mathrm{T}$ sequence from maxout convolutional activations

$\mathrm{h} = \{\mathrm{h}_1, \mathrm{h}_2, \mathrm{h}_3, \ldots, \mathrm{h}_T\}$, $\quad$ $256 \times \mathrm{T}$ sequence of the LSTM output

$p = \{p_1, p_2, p_3, \ldots, p_T\}$, $\quad$ $37 \times \mathrm{T}$ sequence of the estimation, $p_i = \mathrm{W}_{37\times 256}\mathrm{h}_\mathrm{i}$

$S_\omega \approx B\left(\operatorname*{argmax}_\pi P(\pi|p)\right)$, $\quad$ $S_\omega$ *is the target string with* $|S_\omega| = \mathrm{K}$ *and* $S_\omega = B(\pi)$

Projection $B$ removes the repeated labels and non $-$ character labels

for example, $B(-\mathrm{gg} - \mathrm{o} - \mathrm{oo} - \mathrm{d} -) = \mathrm{good}$

# Sequence Labelling Model

## Loss Function

$$L(I, S_\omega) = -\sum_{i=1}^{K} \log P(S_\omega^i | I)$$

$(I\ , S_\omega) \in \Omega$, sample pair (image sample I, corresponding target string $S_\omega$)

$$L(I, S_\omega) \approx -\sum_{t=1}^{T} \log P(\pi_t\,|I)$$

which can be optimized effectively with the Forward $-$ Backward algorithm proposed in [ Graves et al. 2006. ICML]

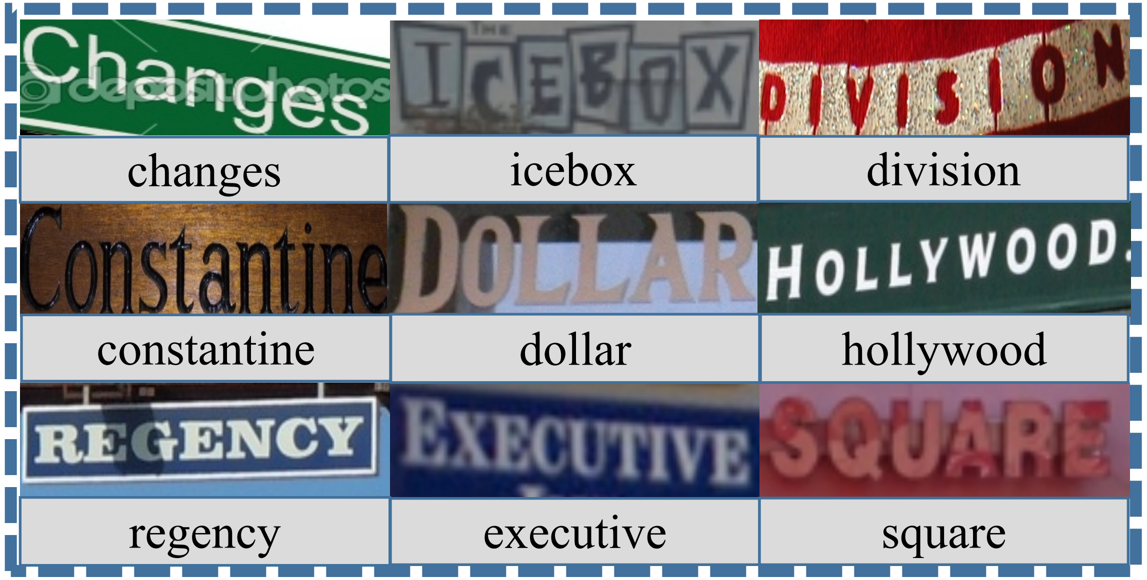
# EXPERIMENTS

# DATASETS

SVT 647 word images

ICDAR2003 860 word images

IIIT5K 3000 word images



# Training

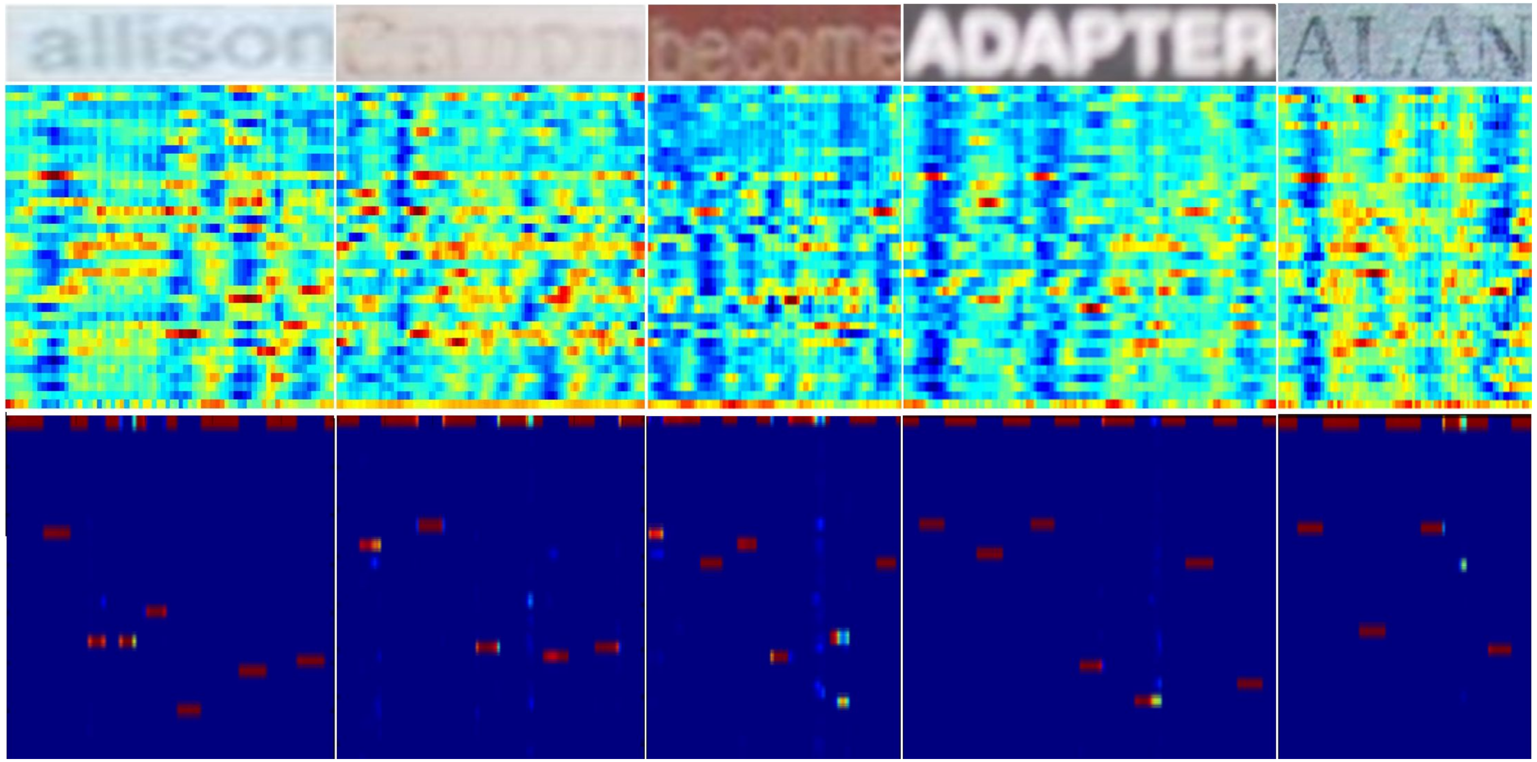
Using $1.8 \times 10^5$ character images for training sequence generation model

$3 \times 10^3$ word images for optimizing the sequence labelling model

# COMPARISON

DTRN vs DeepFeatures

We can get clearer 2D character probability map due to our recurrence property



# COMPARISON

State-of-the-art

| Method | Cropped Word Recognition Accuracy(%) | | | | | |
|---|---|---|---|---|---|---|
| | IC03-50 | IC03-FULL | SVT-50 | IIIT5k-50 | IIIT5k-1K | |
| Wang et al. 2011 | 76.0 | 62.0 | 57.0 | 64.1 | 57.5 | Other |
| Mishra et al. 2012 | 81.8 | 67.8 | 73.2 | - | - | |
| Novikova et al. 2012 | 82.8 | - | 72.9 | - | - | |
| TSM+CRF(Shi et al. 2013) | 87.4 | 79.3 | 73.5 | - | - | |
| Lee et al. 2014 | 88.0 | 76.0 | 80.0 | - | - | Mid-level representation |
| Strokelets(Yao et al. 2014) | 88.5 | 80.3 | 75.9 | 80.2 | 69.3 | |
| Wang et al. 2012 | 90.0 | 84.0 | 70.0 | - | - | Deep neural network |
| Alsharif and Pineau 2013 | 93.1 | 88.6 | 74.3 | - | - | |
| Su and Lu 2014 | 92.0 | 82.0 | 83.0 | - | - | |
| DeepFeatures | 96.2 | 91.5 | 86.1 | - | - | |
| Goel et al. 2013 | 89.7 | - | 77.3 | - | - | Whole image representation |
| Almazán et al. 2014 | - | - | 87.0 | 88.6 | 75.6 | |
| DTRN | **97.0** | **93.8** | **93.5** | **94.0** | **91.5** | Proposed method |
| PhotoOCR | - | - | 90.4 | - | - | Training on additional large datasets |
| Jaderberg2015a | 97.8 | 97.0 | 93.2 | 95.5 | 89.6 | |
| Jaderberg2015b | 98.7 | 98.6 | 95.4 | 97.1 | 92.7 | |

# Results


(Left) Correct recognitions, (Right) Incorrect samples

# SUMMARY

Cast scene text recognition as sequence labelling problem

Leverage word context information to recognize highly ambiguous images

Process unknown words and arbitrary strings

# Thank You

mybestsonny@gmail.com