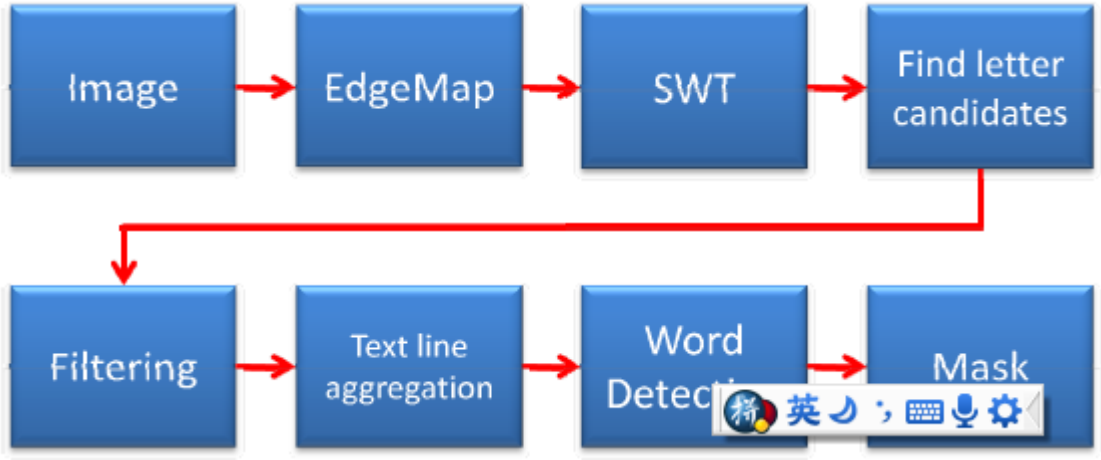
OCR and Text Spotting

Liangliang Cao

http://llcao.net/cu-deeplearning15/

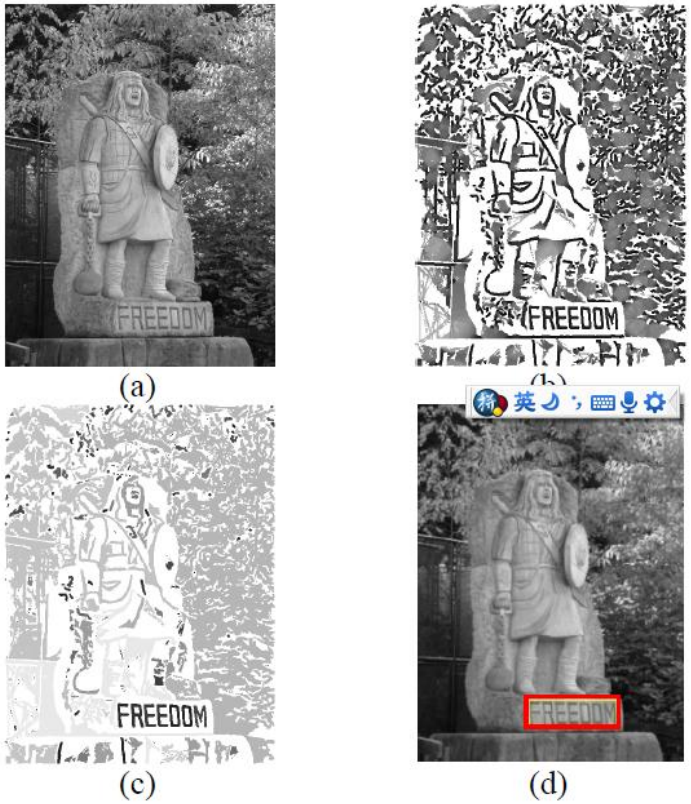
Traditional Methods
(No CNN or RNN)

Stroke Width Transform



Algorithm	Precision	Recall	*f*	Time (sec.)
Our system	**0.73**	**0.60**	**0.66**	**0.94**
Hinnerk Becker*	0.62	0.67	0.62	14.4
Alex Chen	0.60	0.60	0.58	0.35
Qiang Zhu	0.33	0.40	0.33	1.6
Jisoo Kim	0.22	0.28	0.22	2.2
Nobuo Ezaki	0.18	0.36	0.22	2.8
Ashida	0.55	0.46	0.50	8.7
HWDavid	0.44	0.46	0.45	0.3
Wolf	0.30	0.44	0.35	17.0
Todoran	0.19	0.18	0.18	0.3
Full	0.1	0.06	0.08	0.2

Performance on ICDAR dataset



Epshtein, Ofek, and Wexler, Detecting Text in Natural Scenes with Stroke Width Transform, *CVPR 2010.*
Dataset: research.microsoft.com/en-us/um/people/eyalofek/text_detection_database.zip

Google's PhotoOCR [ICCV 2013]

- Detectors
  - Viola-Jones
  - MRF
- Character region resized to 65 x 65 pixels
- HOG + 5 layer fully connected network
  - 422-960-480-480-480-4800-100
  - Trained by SGD with Adagrad and dropout
- Training set
  - 2.2M manually labeled characters
  - Augmented with 4M characters
    - Find 200K matches from 5M images
    - 200K x 200K => 40M labeled characters
    - Select 4M hard examples

Algorithm	Word Recognition Rate (%)
PhotoOCR	**82.83**
NESP	64.20
PicRead	57.99
Baseline (ABBYY)	45.30

ICDAR 2013 scene text

Algorithm	Word Recognition Rate (%)
PhotoOCR	**90.39**
Goel et al. [9]	77.28
Mishra et al. [15]	73.26
Novikova et al. [20]	72.9
Wang et al. [26]	70.0
Baseline (ABBYY) [9]	35.0

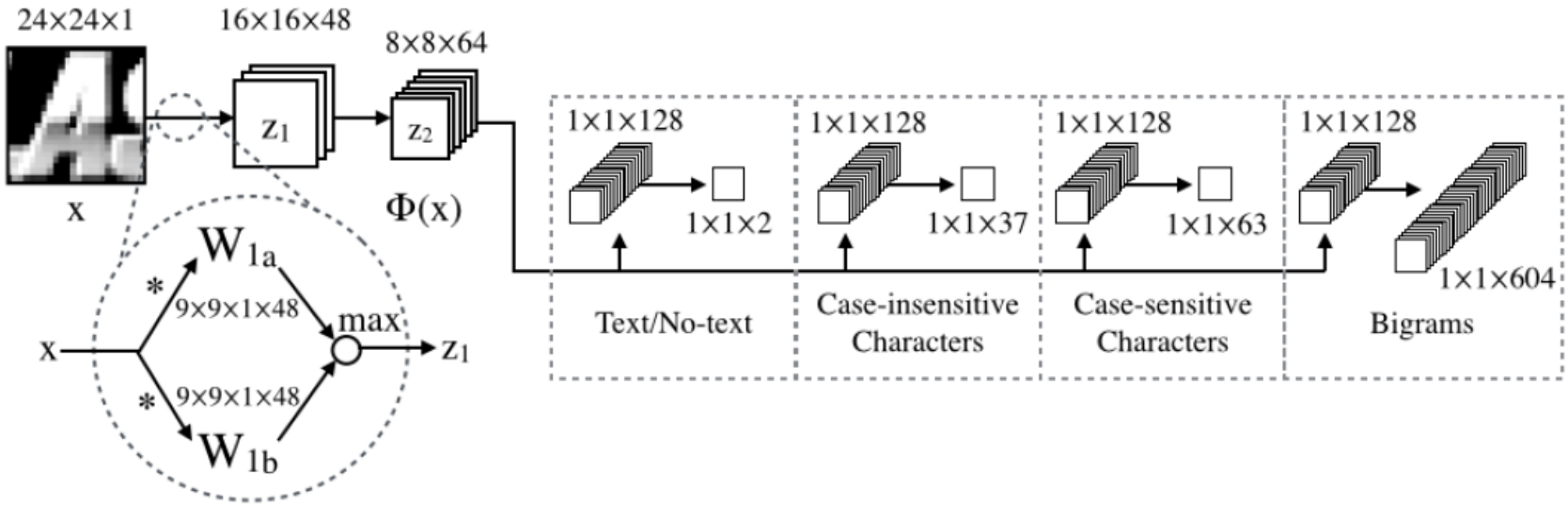
UCSD street view

The authors did not use CNN coz it is more computational expensive although more accurate.

CNN Based Methods

VGG ECCV 2014

- Dataset:
  - FlickrType (6.7 images, 15K words, 71K characters)
  - Cropped 22K characters
  - 37 class (26 letters, 10 digits, + non-text)
- 4 layer CNN (with dropout and maxout)



Max Jaderberg, Andrea Vedaldi, Andrew Zisserman, Deep Features for Text Spotting, ECCV 2014

VGG ECCV 2014 (2)

Pipelines to make use of CNN features

1. Text detection:
   1. CNN + sliding window => text saliency map
   2. Recognize lines of text (run length smoothing)
   3. Split lines into words (Otsu thresholding)
2. Word rec
   1. From character hypothesis to word hypothesis (dynamic programming)

Character recognition

Method	Character Classifier (%)		Case-sensitive Character Classifier(%)	Text/No-text Classifier (%)		Bigram Classifier (%)
	IC03	SVT	IC03	IC03	SVT	IC03
Wang & Wu [47]	-	-	83.9	97.8*	-	-
Alsharif [6]	89.8	-	86.0	-	-	-
Proposed	**91.0**	**80.3**	**86.8**	**98.2**	**97.1**	**72.5**

Max Jaderberg, Andrea Vedaldi, Andrew Zisserman, Deep Features for Text Spotting, ECCV 2014

Google Street View Digit Recognition

- Results
  - 96% accuracy on SVHN
  - 97.84% per digit recognition
  - 90+% accuracy on Google street view number
  - 99.8% on reCAPTCHA (text recognition)

Goodfellow, Bulatov, Ibarz, Arnoud, Shet, Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks, ICLR 2014

Google Street View Digit Recognition

- Network for public house number dataset:
  - Input 64x64 image, with random crop of 54x54 image
  - 8 CNN + 1 locally connected + 2 fully connected
- Network for public house number dataset: (128 x 128 im)
  - 5 CNN + 1 locally connected + 2 fully connected
- Network for CAPTCHA
  - 9 CNN + 1 locally connected + 1 fully connected
- A new output layer for sequential estimation

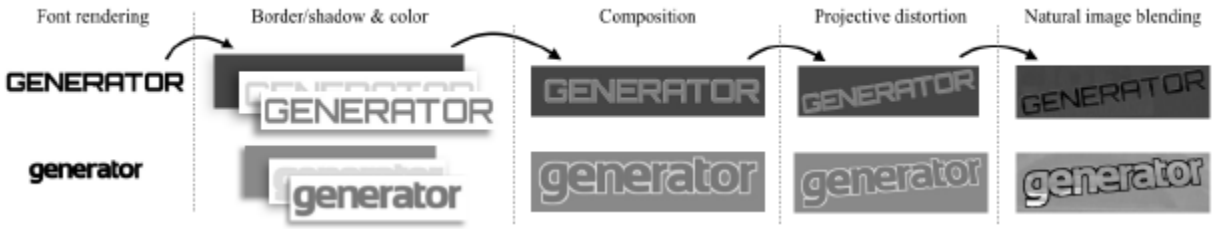
$$P(\mathbf{S} = \mathbf{s}|X) = P(L = n \mid X)\Pi_{i=1}^{n} P(S_i = s_i \mid X)$$

n can be 0-5 or more, and 10 values per digits

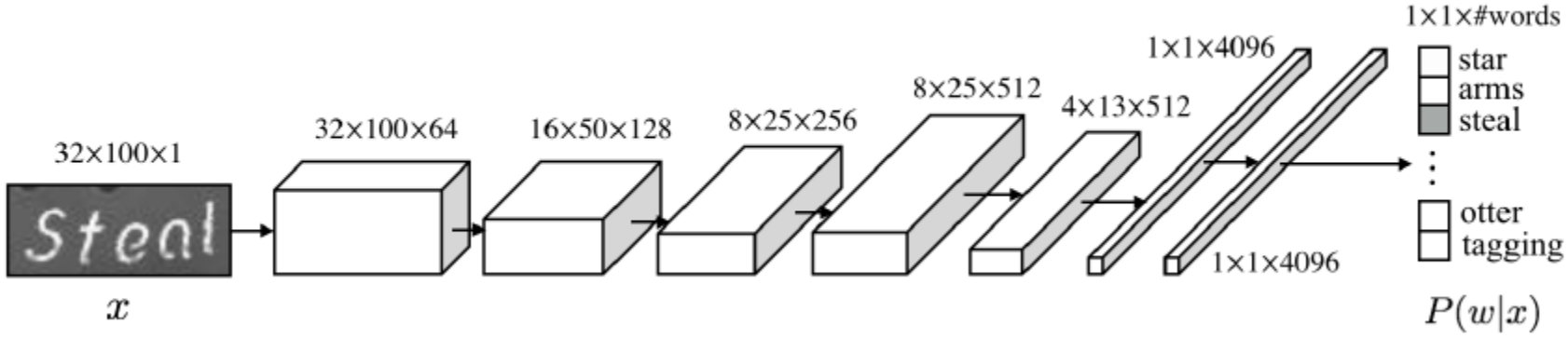
Goodfellow, Bulatov, Ibarz, Arnoud, Shet, Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks, ICLR 2014

VGG New CNN

- Generate Training Data by Synthesizing (32x100 image)



- 5 CNN + 2 fully connected (words as output directly)



M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman,
Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition, NIPS Deep Learning Workshop, 2014
Reading Text in the Wild with Convolutional Neural Networks , arXiv, 1412.1842, 2014

VGG New CNN

Stage	# proposals	Time	Time/proposal
(a) Edge Boxes	$> 10^7$	2.2s	$<$ 0.002ms
(b) ACF detector	$> 10^7$	2.1s	$<$ 0.002ms
(c) RF filter	10^4	1.8s	0.18ms
(d) CNN regression	10^3	1.2s	1.2ms
(e) CNN recognition	10^3	2.2s	2.2ms

Model	IC03-50	IC03-Full	IC03-50k	SVT-50	SVT	IC13	IIIT5k-50	IIIT5k-1k
Baseline ABBYY [26]	56.0	55.0	-	35.0	-	-	24.3	-
Wang [26]	76.0	62.0	-	57.0	-	-	-	-
Mishra [17]	81.8	67.8	-	73.2	-	-	-	-
Novikova [21]	82.8	-	-	72.9	-	-	64.1	57.5
Wang & Wu [27]	90.0	84.0	-	70.0	-	-	-	-
Goel [7]	89.7	-	-	77.3	-	-	-	-
PhotoOCR [3]	-	-	-	90.4	78.0	87.6	-	-
Alsharif [2]	93.1	88.6	85.1	74.3	-	-	-	-
Almazan [1]	-	-	-	89.2	-	-	91.2	82.1
Yao [29]	88.5	80.3	-	75.9	-	-	80.2	69.3
Jaderberg [11]	96.2	91.5	-	86.1	-	-	-	-
Gordo [9]	-	-	-	90.7	-	-	93.3	86.6
DICT-IC03-Full	**99.2**	**98.1**	-	-	-	-	-	-
DICT-SVT-Full	-	-	-	**96.1**	**87.0**	-	-	-
DICT+2-90k	**98.7**	**98.6**	**93.3**	**95.4**	**80.7**	**90.8**	**97.1**	**92.7**
CHAR+2	**96.7**	**94.0**	**89.5**	**92.6**	68.0	79.5	**95.5**	85.4
NGRAM+2-SVM	**96.5**	**94.0**	-	-	-	-	-	-

M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman,
Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition, NIPS Deep Learning Workshop, 2014
Reading Text in the Wild with Convolutional Neural Networks , arXiv, 1412.1842, 2014