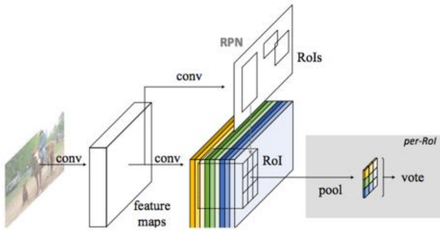
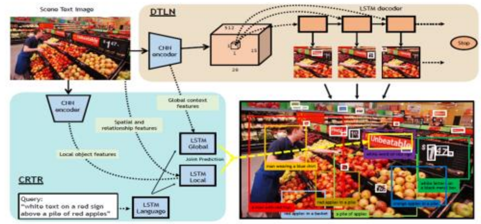


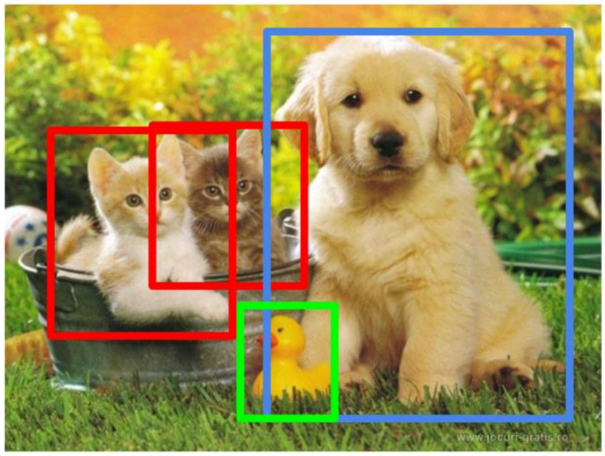


Text Detection

Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." In *Advances in neural information processing systems*, pp. 379-387. 2016.

Rong, Xuejian, Chucai Yi, and Yingli Tian. "Unambiguous Text Localization and Retrieval for Cluttered Scenes." In CVPR 2017.

Object Detection



DOG, (x, y, w, h)
CAT, (x, y, w, h)
CAT, (x, y, w, h)
DUCK (x, y, w, h)

= 16 numbers

Need variable sized outputs

Fei-Fei Li & Andrej Karpathy & Justin Johnson

Detection as Classification

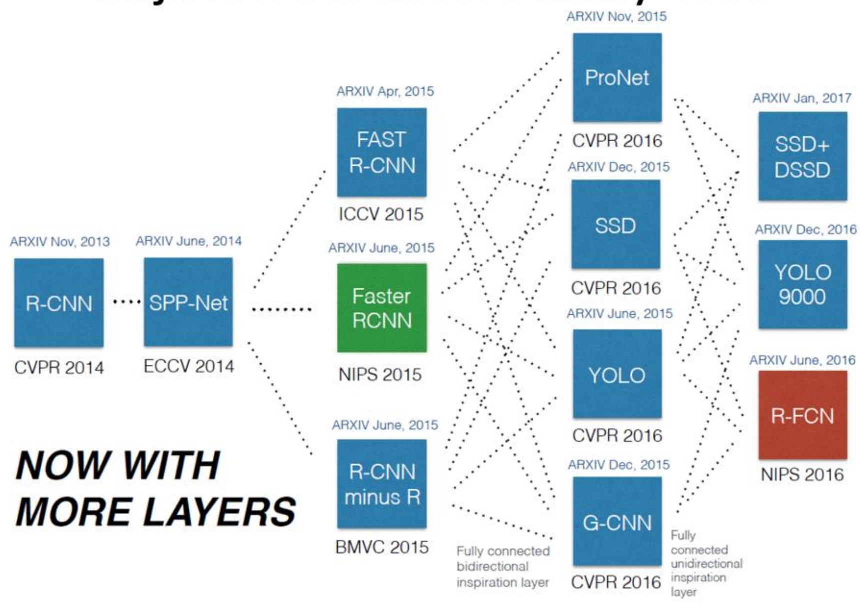
CAT? YES!

DOG? NO

- **Problem**: Need to test many positions and scales
  - use a computationally demanding classifier (CNN)
- **Solution:** If your classifier is fast enough, just do it

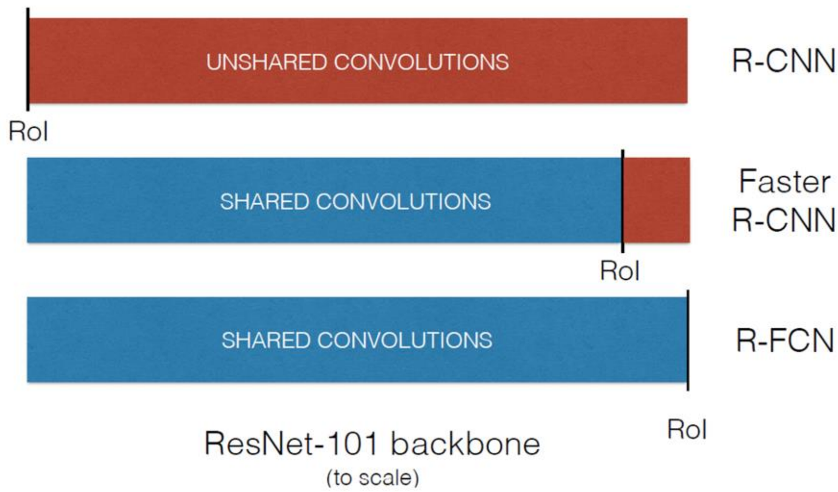
Fei-Fei Li & Andrej Karpathy & Justin Johnson

Object Detection Family Tree



Motivation

- Sharing is Caring



R-FCN: Object Detection via Region-based Fully Convolutional Networks

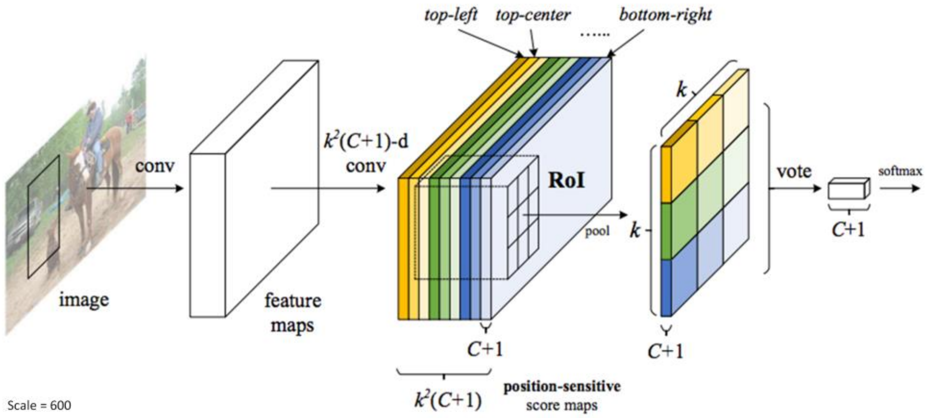
- In previous work, a RoI pooling layer has been inserted before the final convolutions to break the invariance at the cost of reduced sharing

Problem

- For image classification we want location **invariance**
- For object detection, we want location **variance**

- **Solution:** Position-Sensitive Score Maps

Position-Sensitive Score Maps

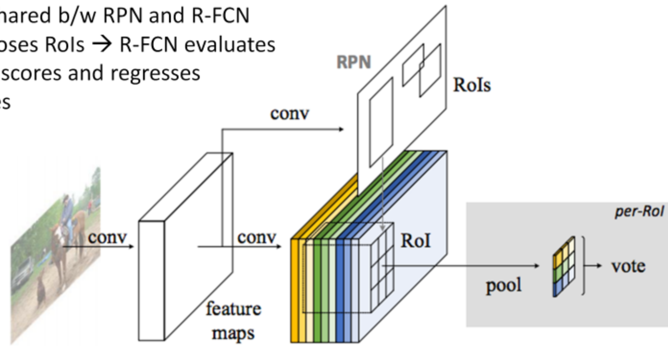


Channels take responsibility for relative spatial locations

C = class, k^2 = position-sensitive score map (kxk = 3x3 = {top-left, top-center, ..})

Efficient Sharing of Diagrams

Feature map shared b/w RPN and R-FCN
RPN part proposes RoIs → R-FCN evaluates
Category-wise scores and regresses bounding boxes



- Backbone: Res-101 on ImageNet
- *Minor modifications:*
  - Remove the GAP
  - Dimensionality reduction layer (1024)

ResNet-101's effective stride from 32 pixels to 16 pixels
Increasing the score map resolution

Visualisation: Hit

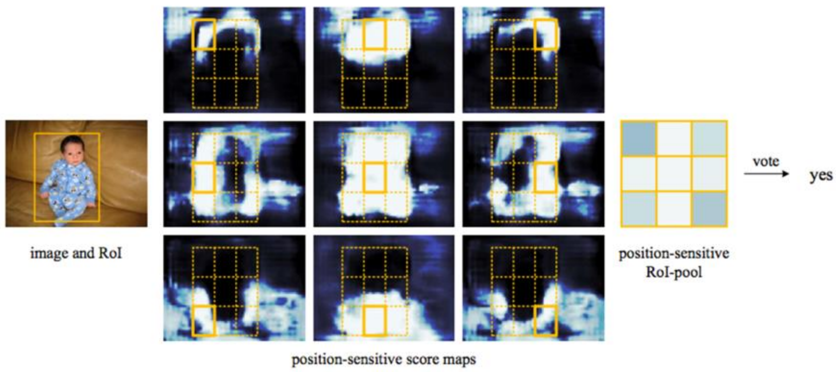


Figure 3: Visualization of R-FCN ($k \times k = 3 \times 3$) for the *person* category.

Visualisation: Miss

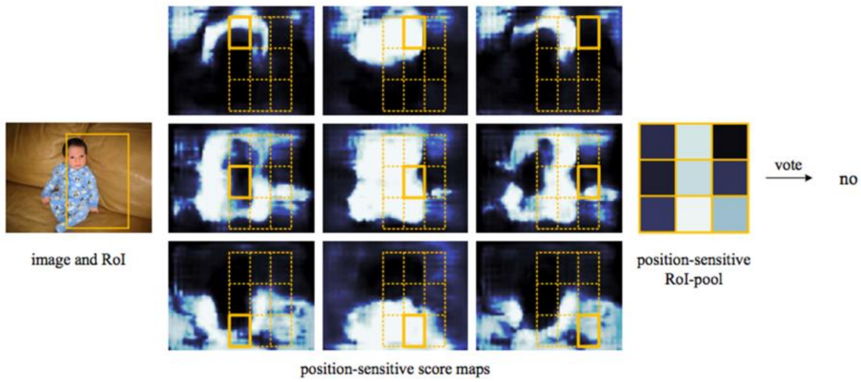


Figure 4: Visualization when an RoI does not correctly overlap the object.

Loss function = cross-entropy loss + box regression loss

Effect of Position Sensitivity on FC

- Without position sensitivity, Faster R-CNN takes a major performance hit when the RoI pooling is late in the network

Table 2: Comparisons among fully convolutional (or "almost" fully convolutional) strategies using **ResNet-101**. *All competitors in this table use the à trous trick.* Hard example mining is not conducted.

method	RoI output size ($k \times k$)	mAP on VOC 07 (%)
naïve Faster R-CNN	1×1 7×7	61.7 68.9
class-specific RPN	-	67.6
R-FCN (w/o position-sensitivity)	1×1	*fail*
R-FCN	3×3 7×7	75.5 **76.6**

- "naive" Faster R-CNN still has FC layer after RoI pooling
- 20x times faster than Faster R-CNN +++

Standard Benchmarks

- VOC 2007

Table 4: Comparisons on PASCAL VOC 2007 *test* set using **ResNet-101**. "Faster R-CNN +++" [9] uses iterative box regression, context, and multi-scale testing.

	training data	mAP (%)	test time (sec/img)
Faster R-CNN [9]	07+12	76.4	0.42
Faster R-CNN +++ [9]	07+12+COCO	**85.6**	3.36
R-FCN	07+12	79.5	0.17
R-FCN multi-sc train	07+12	80.5	0.17
R-FCN multi-sc train	07+12+COCO	**83.6**	0.17

Fine-tune R-FCN using a learning rate 0.001 for 20k mini-batches and 0.0001 for 10k mini-batches on VOC

Overlap with a ground-truth box of at least 0.5

Effect of Depth

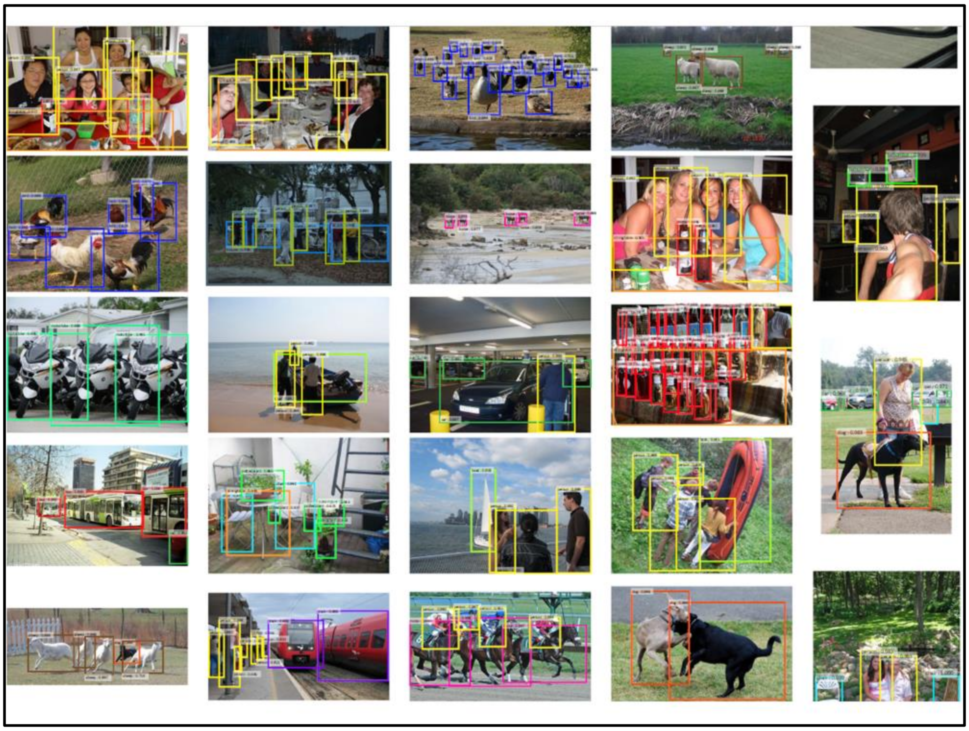
	training data	test data	ResNet-50	ResNet-101	ResNet-152
R-FCN	07+12	07	77.0	79.5	79.6
R-FCN multi-sc train	07+12	07	78.7	80.5	80.4

Saturates at ResNet-101

Effect of Proposal Type

	training data	test data	RPN [18]	SS [27]	EB [28]
R-FCN	07+12	07	**79.5**	77.2	77.8

Works pretty well with any proposal method



Conclusion

- Curated examples of R-FCN results on the PASCAL VOC 2007 test set (83.6% mAP). The network is ResNet-101, and the training data is 07+12+COCO. A score threshold of 0.6 is used for displaying. The running time per image is 170ms on one Nvidia K40 GPU.

- System naturally adopts the state-of-the-art image classification backbones, such as ResNets, that are by design fully convolutional.

Unambiguous Text Localization and Retrieval for Cluttered Scenes

- Text instance as one category of self-described objects provides valuable information for understanding and describing cluttered scenes.

Problem

- How to find the text instance that you desire?

Cluttered background
How quickly

Cluttered Scene

- Text Localization

Text Recognition → Content Matching

Using NLP

Relationship:
Text and Surrounding Object

Retrieve Interested Text

Framework

- Recurrent Text Localization
- Text-Context Relationship Modeling
- Unambiguous Text Retrieval

  – VGG-16 (ImageNet) to encode a scene I into a feature map in a MxN grid of 512-d feature descriptors.

Proposed Architecture

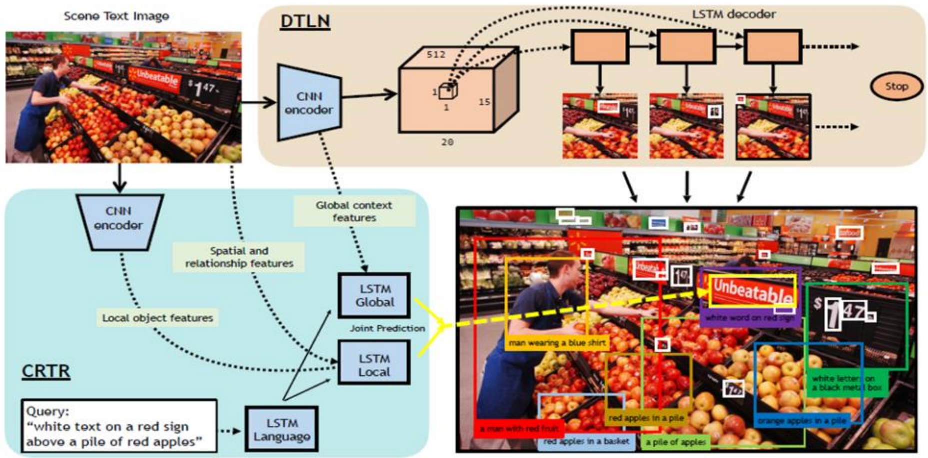


Figure 2: The architecture of the proposed Dense Text Localization Network (DTLN) and Context Reasoning Text Retrieval (CRTR) Models. For an input image, the DTLN model directly decodes the CNN features into a variable length set of text instance candidates. The CRTR model pools the information from three different LSTM models, and jointly scores and ranks the candidate text regions which are generated by DTLN.

- DTLN
  - Trained on SynthText in the
  - Wild dataset (800k images)
  - Resized to 480x640
  - CRTR → fine-tune DenseCap [36]
    - On COCO-TextRef

New Dataset

COCO-TextRef Dataset

- scene text and language annotations
- **5,000** images for training/tuning
- **1,638** images for testing
- **31,870** expressions
- **11,342** distinct objects
- **17,355** text instances

Experiments

- Text Localization

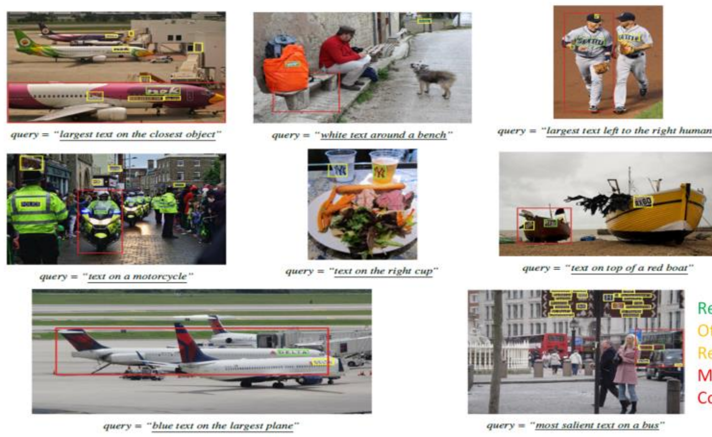
Detected Most Text Instances

Figure 3: Example results of scene text localization. The green bounding boxes contain correct detections; Red bounding boxes contain false positives; Red dashed box (e.g., the one at the bottom-right image) contains the false negative.

Experiments

- Text Retrieval

Quality Results



Retrieved Text Region
Other Detected Text Regions
Mostly Related Context Concept

Figure 4: Text region retrieval results of the proposed Context Reasoning Text Retrieval (CRTR) model on the COCO-TextRef dataset. At first, red boxes are employed to denote context concepts. Then green boxes are added to identify the successfully retrieved text regions associated with the context concepts. The remaining text regions are marked by yellow boxes.

Experiments

- Comparison

Table 1: Performance comparison between our proposed framework with previous scene text localization approaches on ICDAR 2013 [30] and SVT datasets [31] in terms of the measures of PASCAL Eval [32] and DetEval [33]. Precision (*P*) and Recall (*R*) at maximum F-measure (*F*) and the average computation time (*T*) are reported. Bold number indicates the best performance for each measure metric. Average time spent on these scene text localization approaches (the last column) demonstrates that the proposed DTLN achieves state-of-the-art F-measure while running in comparable speed as competing approaches.

	PASCAL Eval						DetEval						Time
	IC13			SVT			IC13			SVT			Avg.
	F	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*T/s*
TH-TextLoc [30]	-	-	-	-	-	-	0.67	0.70	0.65	-	-	-	-
Text Spotter [8]	-	-	-	-	-	-	0.74	0.88	0.65	-	-	-	0.3
Yin et al. [9]	-	-	-	-	-	-	0.76	0.88	0.66	-	-	-	0.43
Lu et al. [34]	-	-	-	-	-	-	0.78	0.89	0.70	-	-	-	-
Jaderberg [12]	0.76	0.87	0.68	0.54	0.63	0.47	0.77	0.89	0.68	0.25	0.28	0.23	7.3
Zhang et al. [35]	-	-	-	-	-	-	0.80	0.88	0.74	-	-	-	60.0
FCN [13]	-	-	-	-	-	-	0.83	0.88	0.78	-	-	-	2.1
FCRNall+filts [15]	0.84	**0.94**	0.76	0.63	0.65	0.60	0.83	**0.94**	0.77	0.27	**0.29**	0.26	1.27
Tian et al. [17]	**0.88**	0.93	**0.83**	**0.66**	**0.68**	**0.65**	-	-	-	-	-	-	**0.14**
DTLN	0.85	0.92	0.79	0.64	0.65	0.63	**0.85**	0.92	**0.78**	**0.28**	**0.29**	**0.27**	0.35

Conclusion

- To utilize text instances for understanding natural scenes, we have proposed a framework that combines image-based text localization with language-based context description for text instances.
- Future work will focus on combining the models of scene text localization and scene text retrieval to produce an end-to-end system.