DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson, Andrej Karpathy, Li Fei-Fei (CVPR 2015)

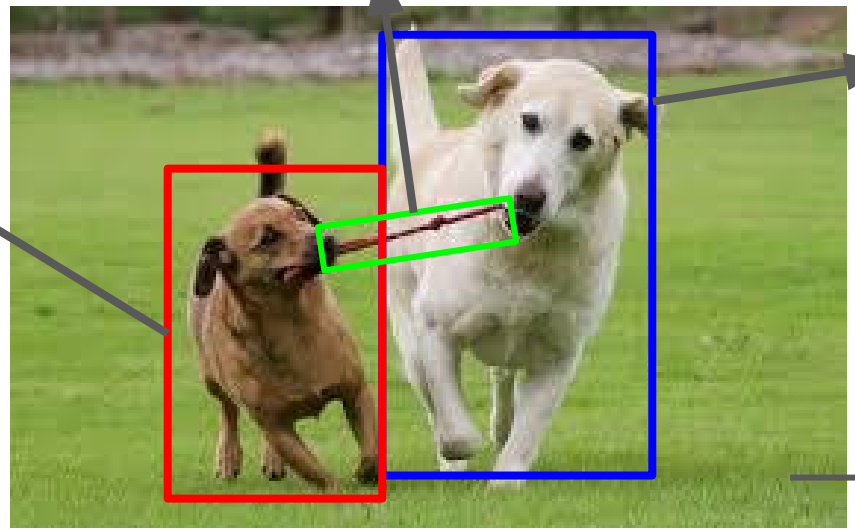
Gaurav (13274)
Palash Chauhan (13455)
Shubham Agarwal (13674)

CS698: Recent Advances in Computer Vision
State of Art Presentation

Contents

1. Introduction
2. Related Work
3. Model Architecture
4. Dataset
5. Evaluation and Results
6. Conclusion

Introduction



Two dogs running on grass are pulling on a rope

Whole Image

Image Regions

label density

Single Label

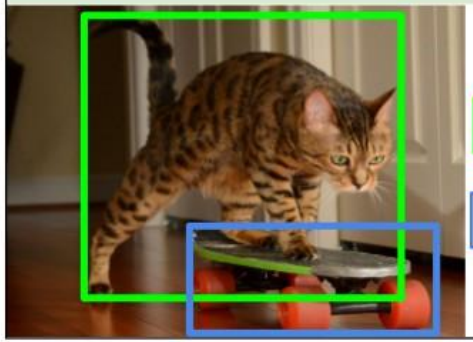
Sequence

label complexity

Classification

Cat

Detection



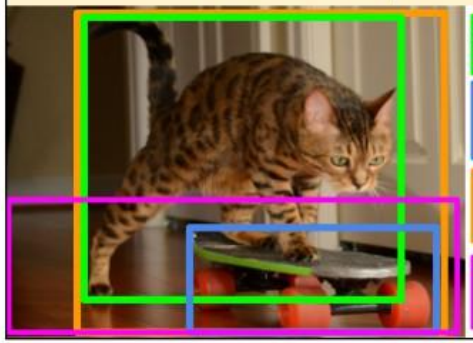
Cat

Skateboard

Captioning

A cat riding a skateboard

Dense Captioning



Orange spotted cat

Skateboard with red wheels

Cat riding a skateboard

Brown hardwood flooring

Related Work

Object Detection

- [O. Russakovsky et. al. 2015]
- [S. Ren et. al. 2015]
- [R. Girshick et. al. 2014]
- [A. Krizhevsky et. al. 2012]

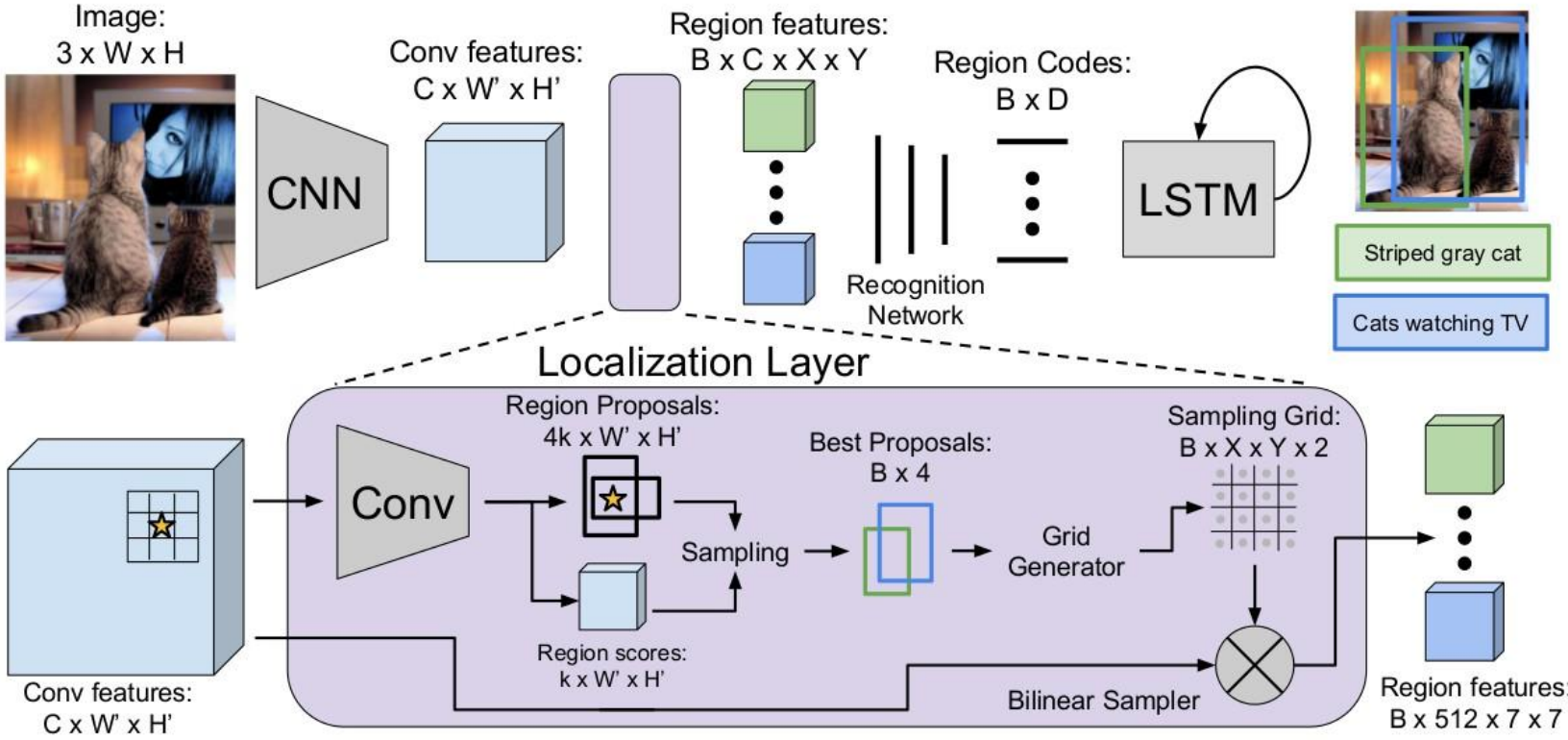
Image Captioning

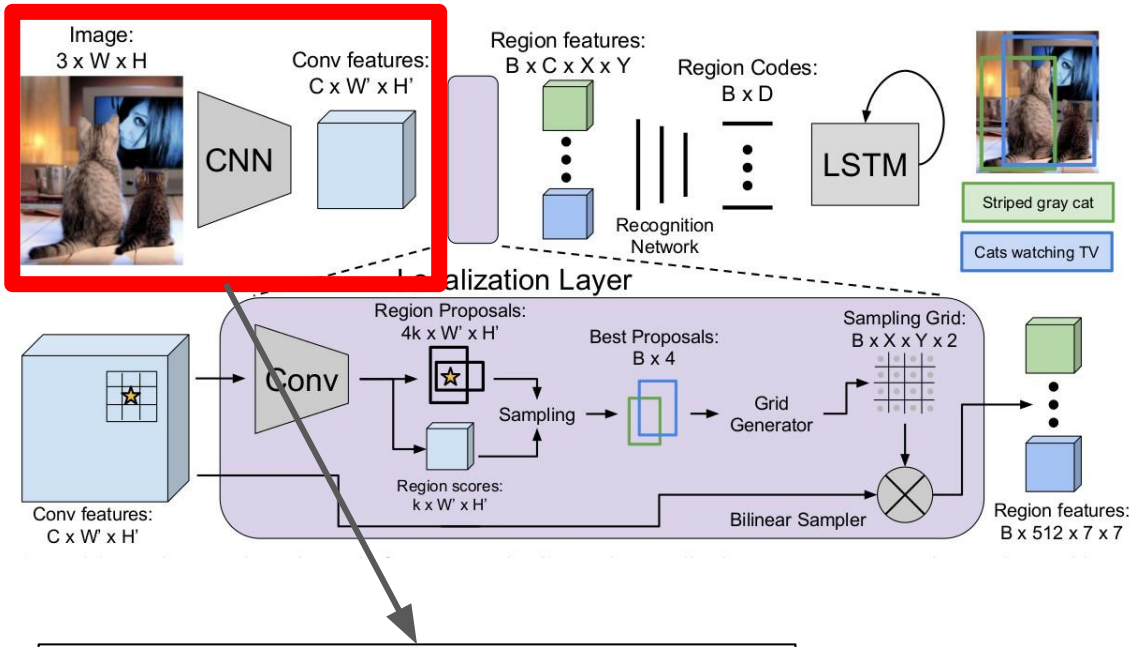
- [Xu et. al. 2015]
- [Vinyals et. al. 2014]
- [Vedantam et. al. 2014]
- [A. Graves et. al. 2013]
- [G. Kulkarni et. al. 2011]
- [A. Farhadi et. al. 2010]

Dense Captioning

- [Karpathy et. al. 2014]

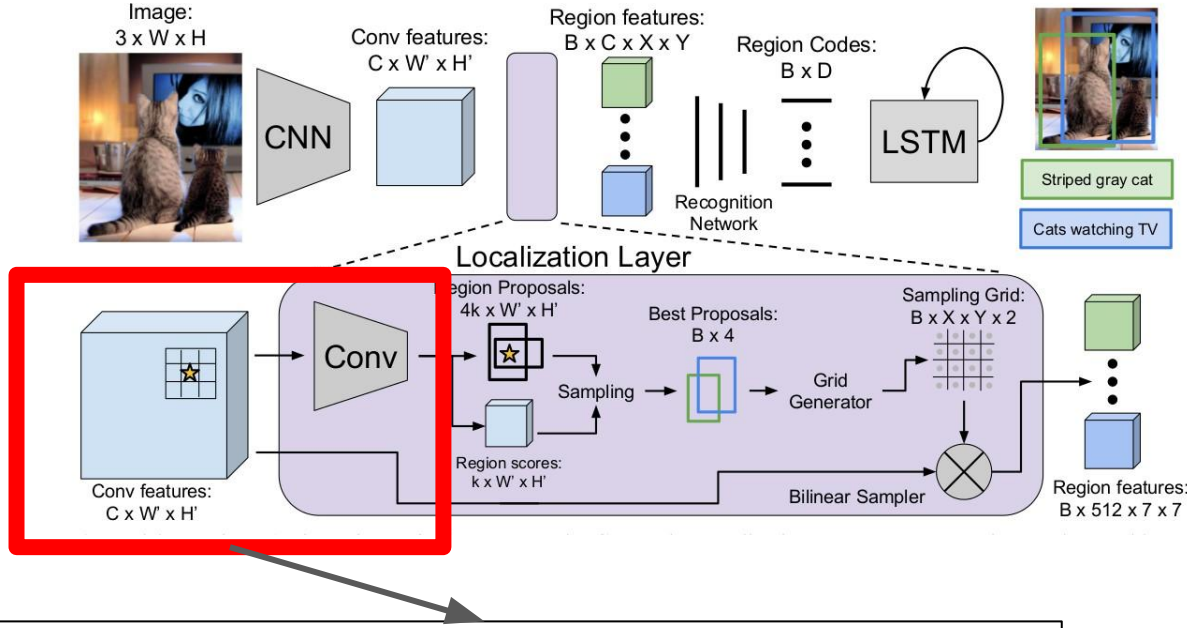
Model Architecture





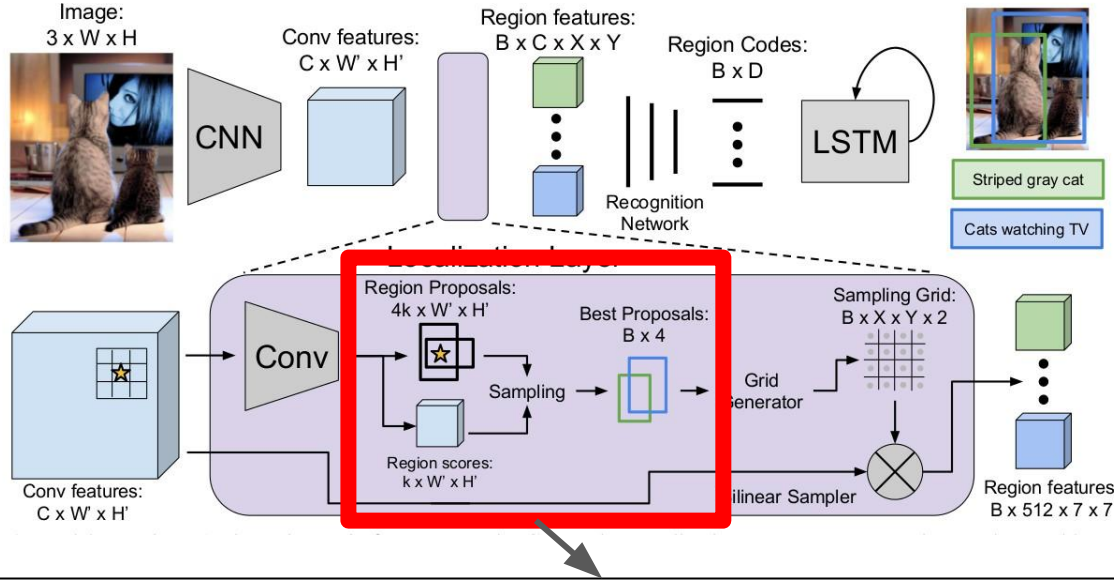
Convolutional Network

- VGG-16 architecture
- Final pooling layer removed
- Input: Image, 3 x W x H
- Output: Feature Tensor, C x W' x H'



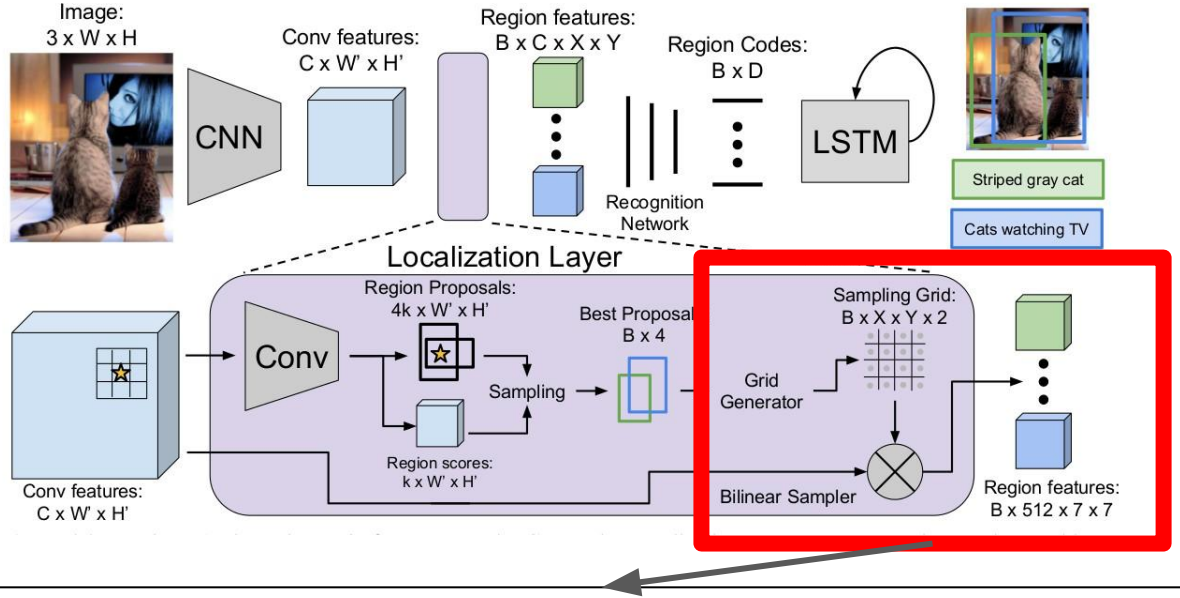
Convolutional Anchors, Box Regression

- Approach based on Faster R-CNN
- Predict region proposals by regressing offsets from a set of translation-invariant anchors
- *k* anchor boxes at every point
- Regress from anchors to the region proposals



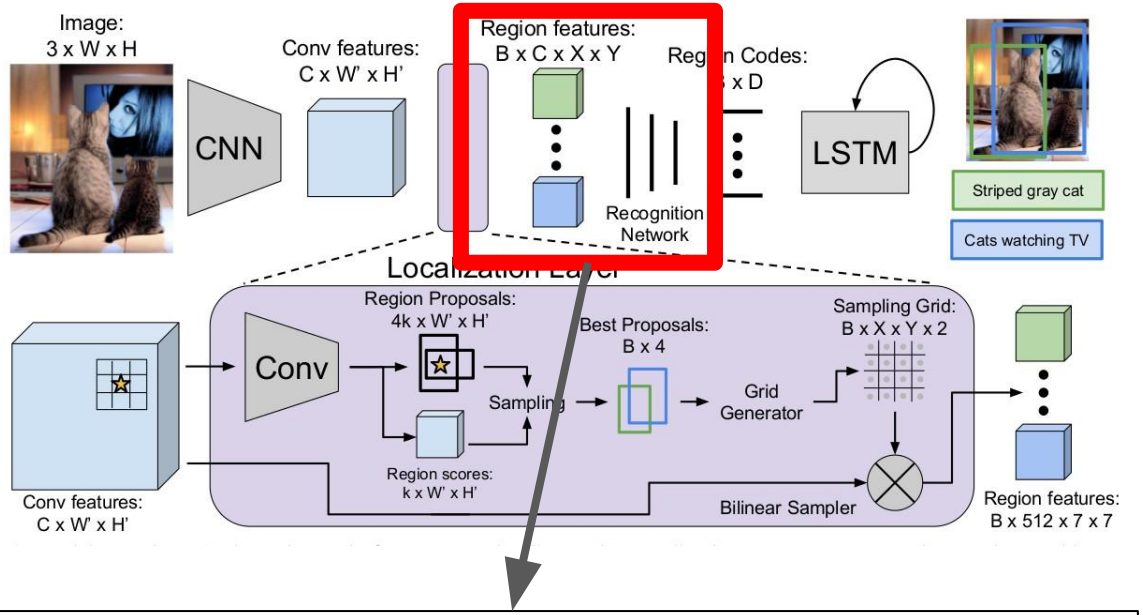
Box Sampling

- 5*k* x W' x H' : Coordinates and scores of regional proposals
- W = 720, H = 540, *k* = 12 → 17280 proposals → expensive
- During training, sample a minibatch of B=256 regions with atmost B/2 positives and rest negatives.
- Positive, if IoU > 0.7 with some ground-truth region
- Negative, if IoU < 0.3 with all ground-truth regions
- Also, the predicted region of maximal IoU with each ground-truth region is positive



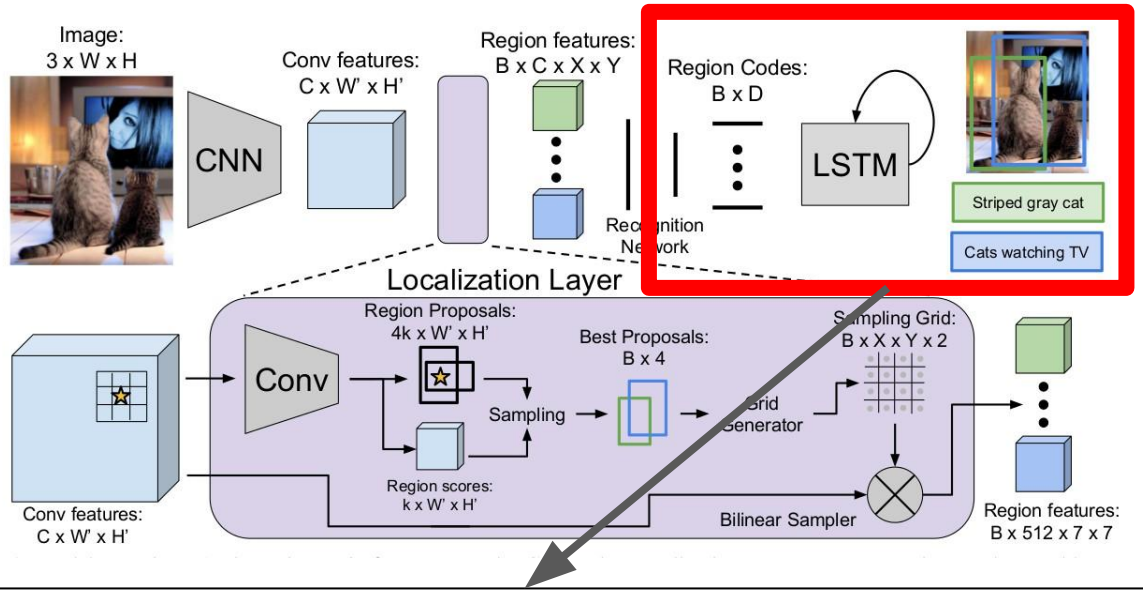
Bilinear Interpolation

- Must extract a fixed-size feaure representation for each variably sized region proposal
- Using RoI pooling layer as in Fast R-CNN, gradients can be propagated backward from the output features to the input features, but not to the input proposal coordinates.
- Replaced with bilinear interpolation
- Sampling grid is a linear function of the proposal coordinates, so gradients can be propagated backward into predicted region proposal coordinates



Recognition Network

- Fully-connected neural network
- Input (from bilinear interpolation): tensor, B x C x X x Y
- Produces a code for every region
- Hence, compactly encodes visual appearance
- D = 4096

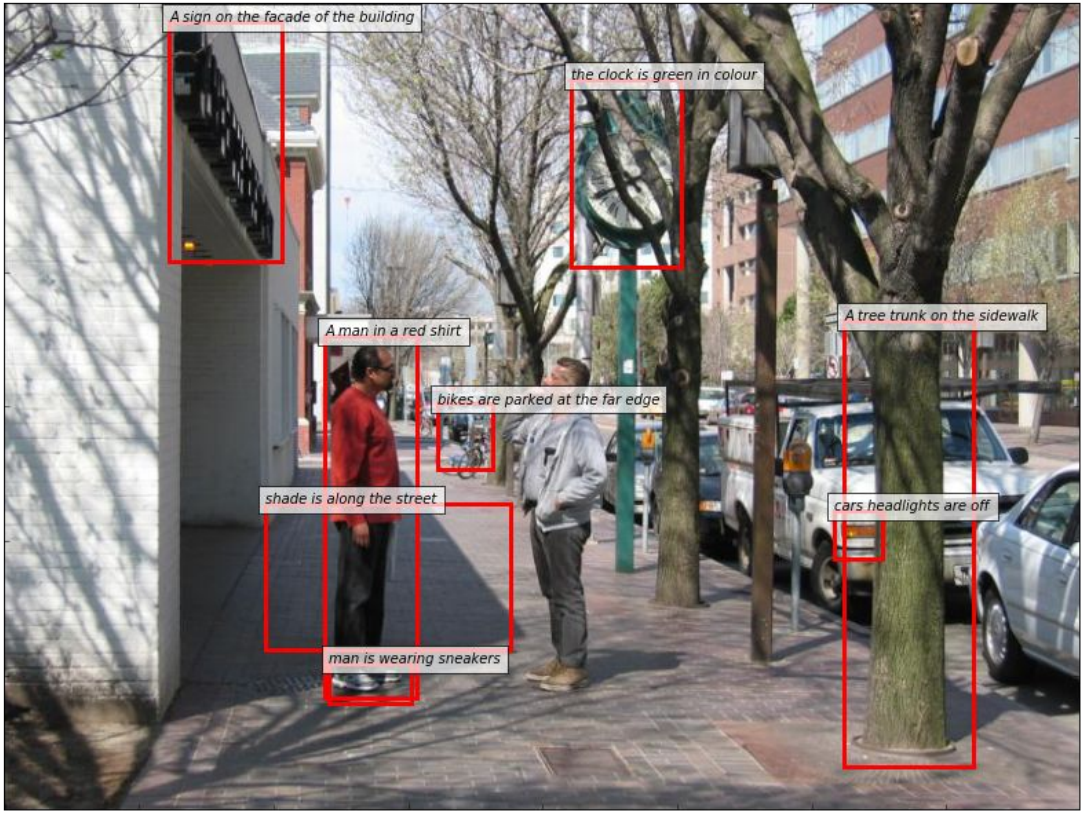


RNN Language Model

- use the region codes to condition an RNN language model
- Training: Feed word vectors appended with a START token and the region code. Obtain output vectors of length |V| + 1. |V| = token vocabulary
- Loss function on output vectors → average cross entropy
- Testing: Feed the region code in the beginning. Sample most likely next token and feed it to RNN in the next time step

Visual Gnome Dataset

- 94,313 images
- 4,100,413 snippets of text (43.5 per image)



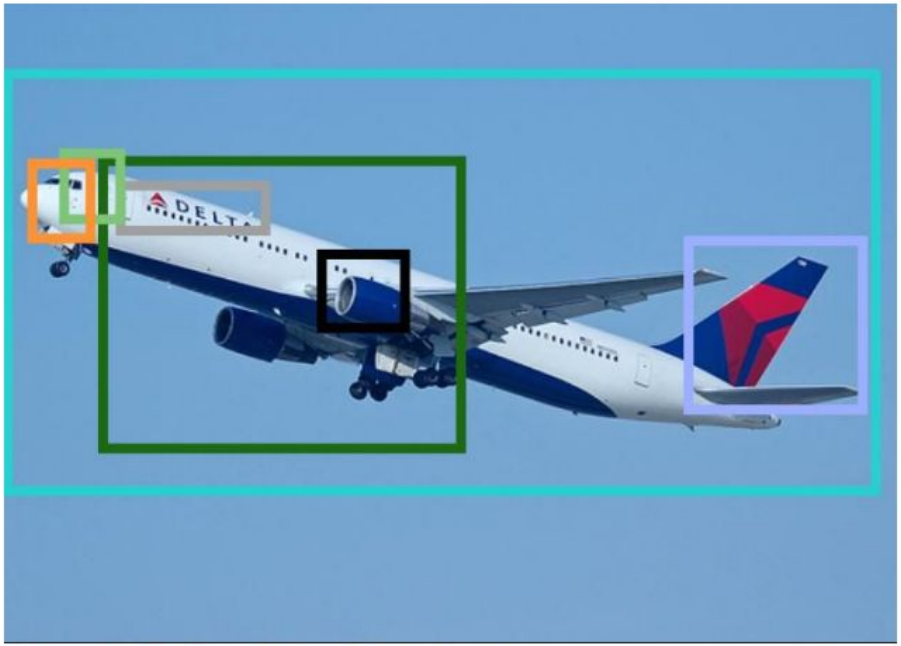
Evaluation Metric

- Mean Average Precision
  - Localization
    - Intersection over union
  - Language accuracy
    - METEOR score

Baseline Models

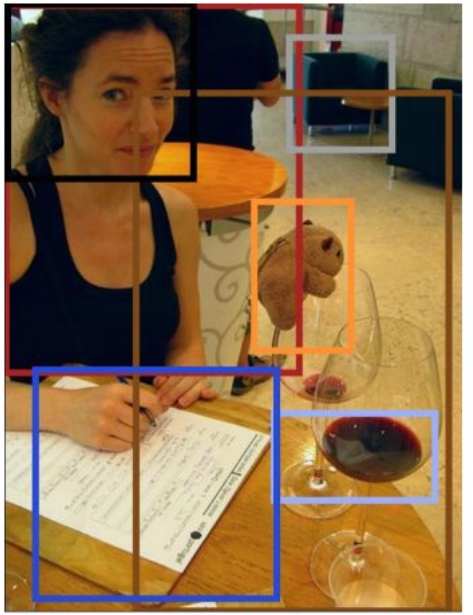
- Baseline Models
  - Region RNN model
  - Full Image RNN model
  - FCLN on EB
- Region Proposals during testing
  - Ground truth boxes(GT)
  - Edge Boxes(EB)
  - Region Proposal Network(RPN)

Results - Qualitative



plane is flying. tail of the plane. red and white plane. plane is white. engine on the plane. windows on the plane. nose of the plane.

A large jetliner flying through a blue sky.



woman wearing a black shirt. teddy bear is brown. chair is black. glass of wine. table is brown. woman with brown hair. paper on the table.

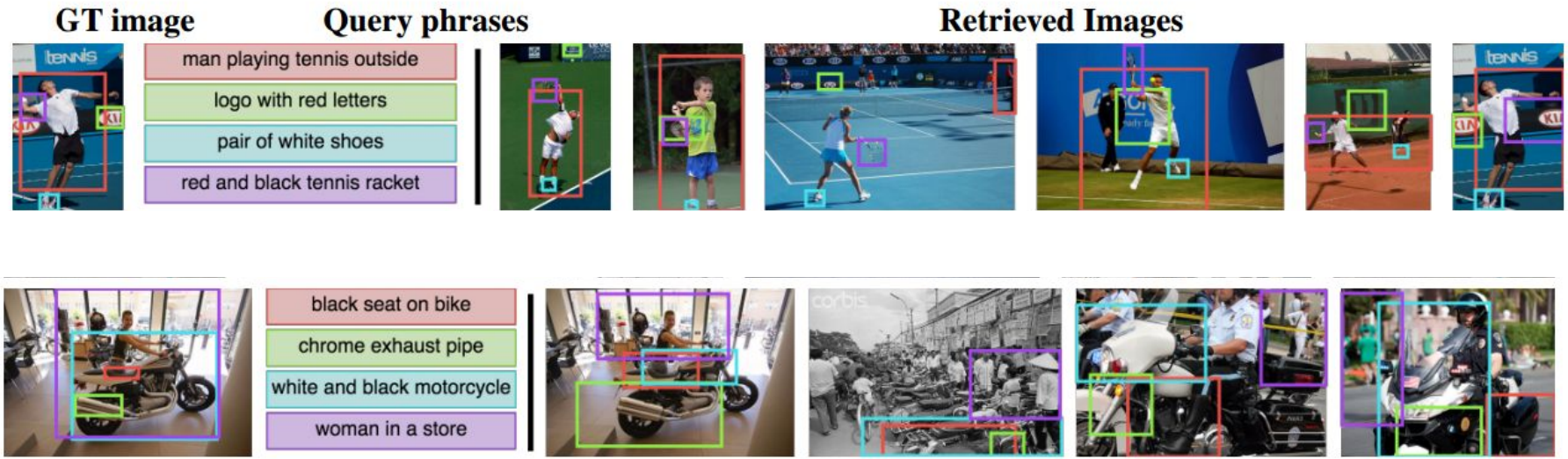
A man and a woman sitting at a table with a cake.

Results - Quantitative

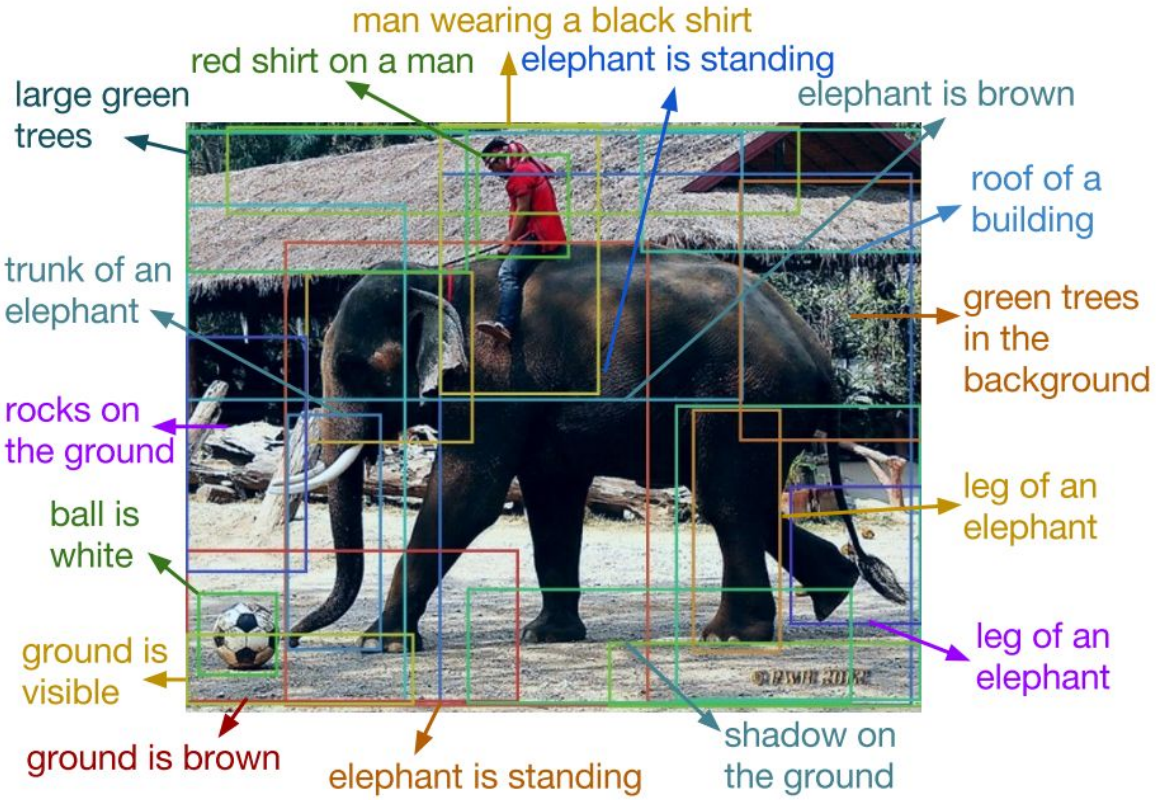
Region source	**Language** (METEOR) EB	RPN	GT	**Dense captioning** (AP) EB	RPN	GT	**Test runtime** (ms) Proposals	CNN+Recog	RNN	Total
Full image RNN [21]	0.173	0.197	0.209	2.42	4.27	*14.11*	210ms	2950ms	**10ms**	3170ms
Region RNN [21]	0.221	0.244	0.272	1.07	4.26	*21.90*	210ms	2950ms	**10ms**	3170ms
FCLN on EB [13]	**0.264**	**0.296**	0.293	4.88	3.21	*26.84*	210ms	**140ms**	**10ms**	360ms
Our model (FCLN)	**0.264**	0.273	**0.305**	**5.24**	**5.39**	***27.03***	**90ms**	**140ms**	**10ms**	**240ms**

Further Applications

- Image retrieval using regions and captions



Conclusion


man wearing a black shirt
red shirt on a man
elephant is standing
large green trees
elephant is brown
roof of a building
trunk of an elephant
green trees in the background
rocks on the ground
leg of an elephant
ball is white
ground is visible
leg of an elephant
ground is brown
elephant is standing
shadow on the ground

Questions ?