Modeling Context between Objects for Understanding Referring Expressions

Varun Nagaraja, Vlad Morariu, Larry Davis

ECCV 2016

Referring Expressions

Descriptions that identify a particular object instance

Man sitting on the left holding a game controller

Woman in the middle sitting on the bed

Man wearing a red jacket and blue jeans sitting on the right

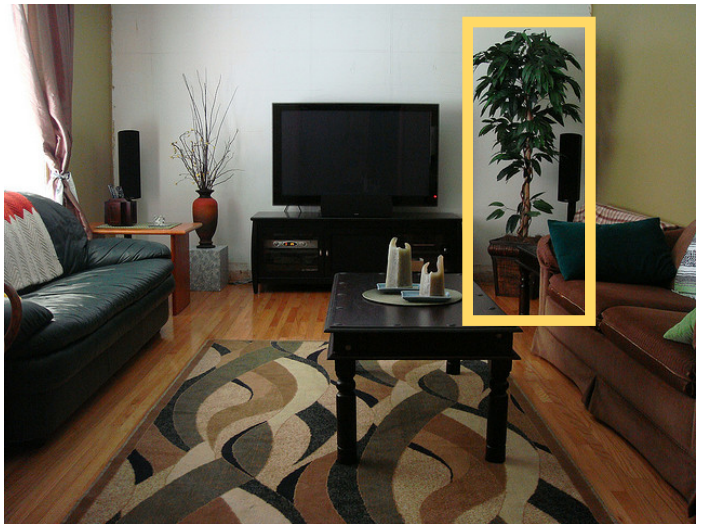
Referring expressions rely on attributes and context

Blonde fluffy dog

Giraffe bending down

Person riding a blue motorcycle

Tan colored sofa



Plant on the right side of the TV

Problem Formulation

Input

$\text{Sentence}(S)$

Girl wearing a red jacket

$\text{Image}(I)$

Output

$$R^* = \underset{R}{\text{argmax}}\ p(R|S, I)$$

Solution Framework

Hypothesize a set of region candidates

Generation and Comprehension of Unambiguous Object Descriptions
J. Mao et al., CVPR 2016

Solution Framework

Pick the region candidate with the highest probability of ***generating*** the query referring expression

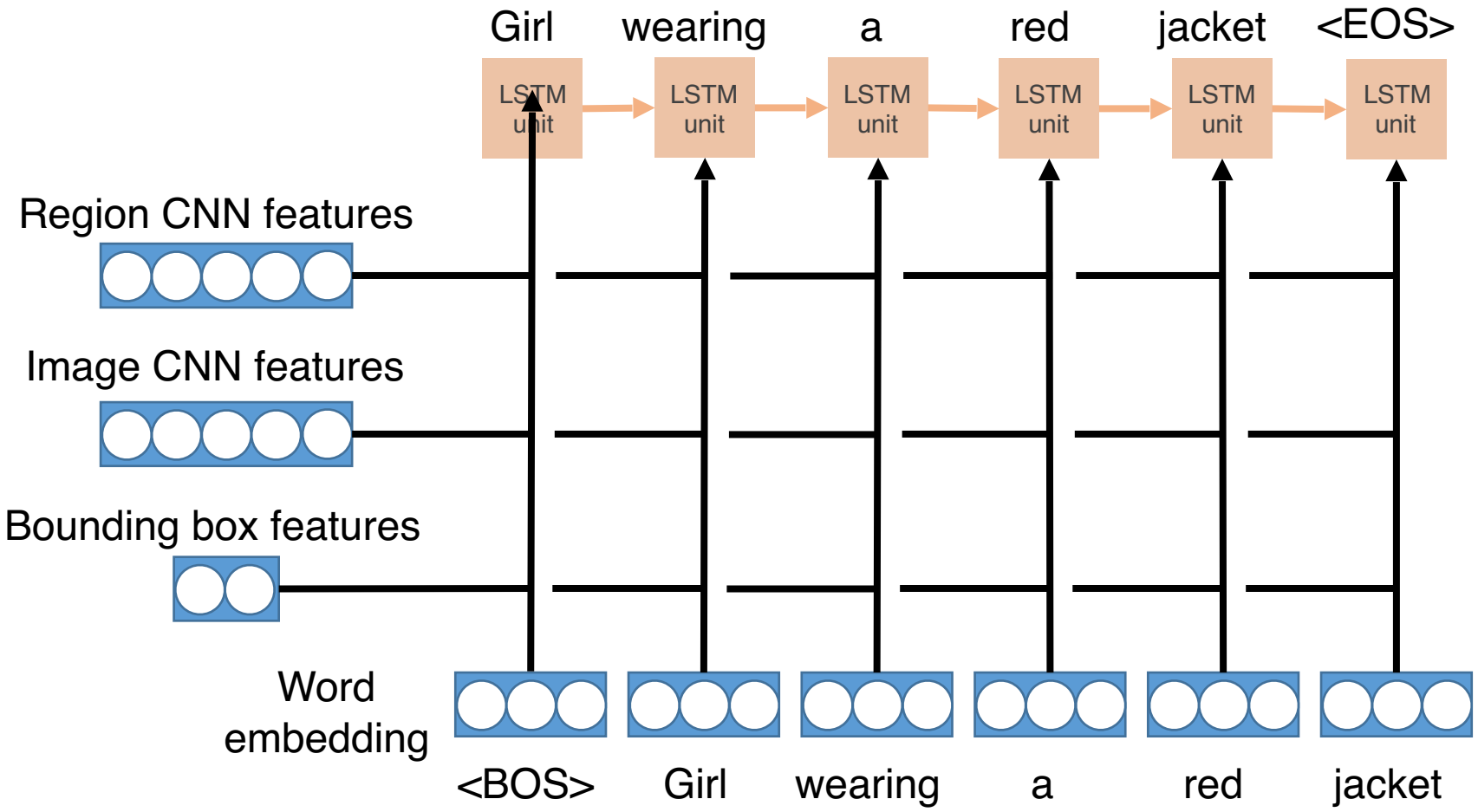
$$p(R|S,I) = \frac{p(S|R,I)p(R|I)}{\sum_{R' \in \mathcal{C}} p(S|R',I)p(R'|I)}$$

$$\Rightarrow \quad p(R|S,I) \propto p(S|R,I)$$

Generation and Comprehension of Unambiguous Object Descriptions
J. Mao et al., CVPR 2016

Baseline Method

Modeling referring expression probability using an LSTM



Generation and Comprehension of Unambiguous Object Descriptions
J. Mao et al., CVPR 2016

Max-margin Method

The baseline method can be improved by training the model to have lower probability for negative regions

Girl wearing a red jacket

Referred region

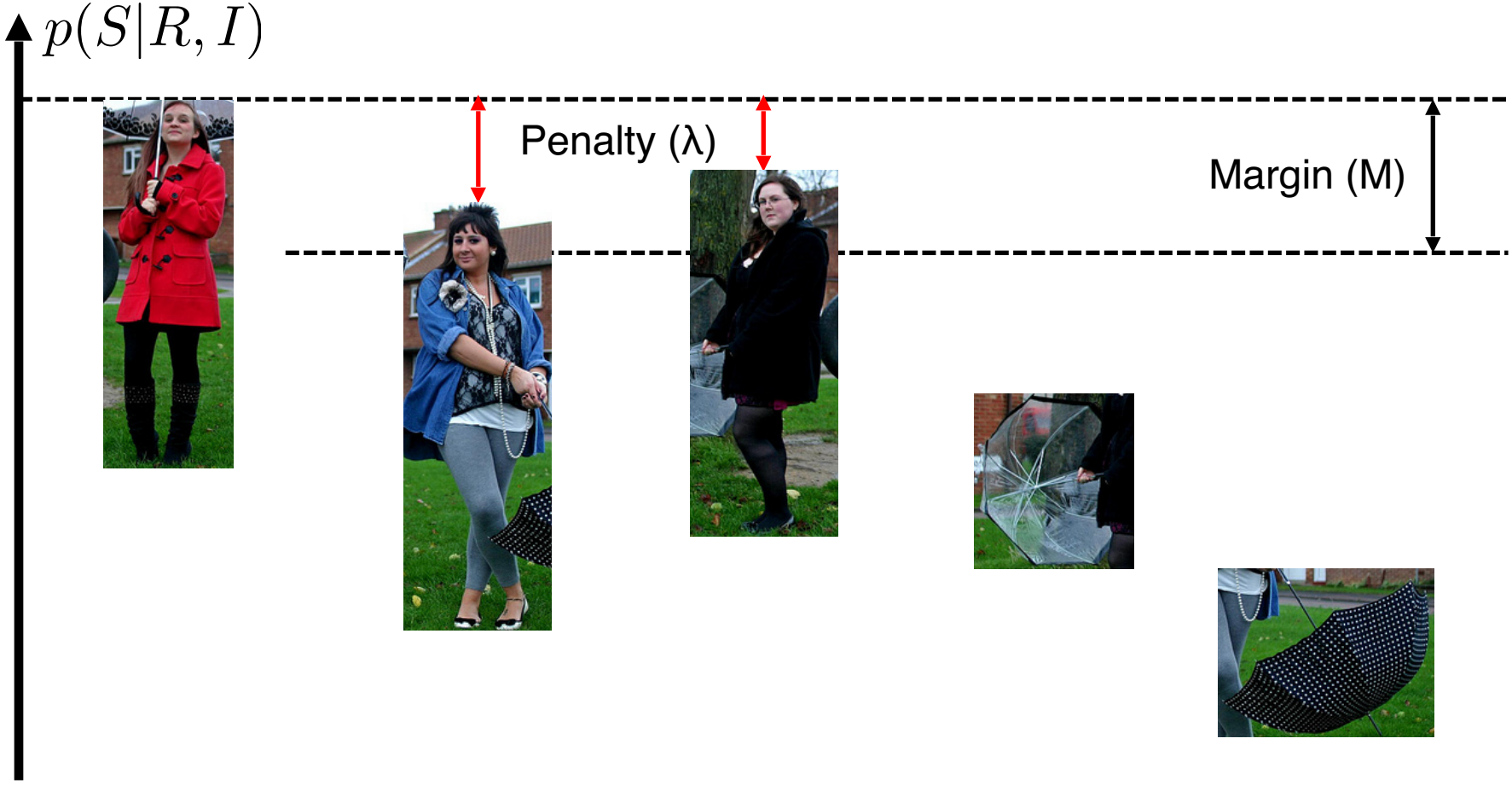
Negative regions

Generation and Comprehension of Unambiguous Object Descriptions
J. Mao et al., CVPR 2016

Max-margin Method

Penalize negative regions if they are within a margin

Girl wearing a red jacket



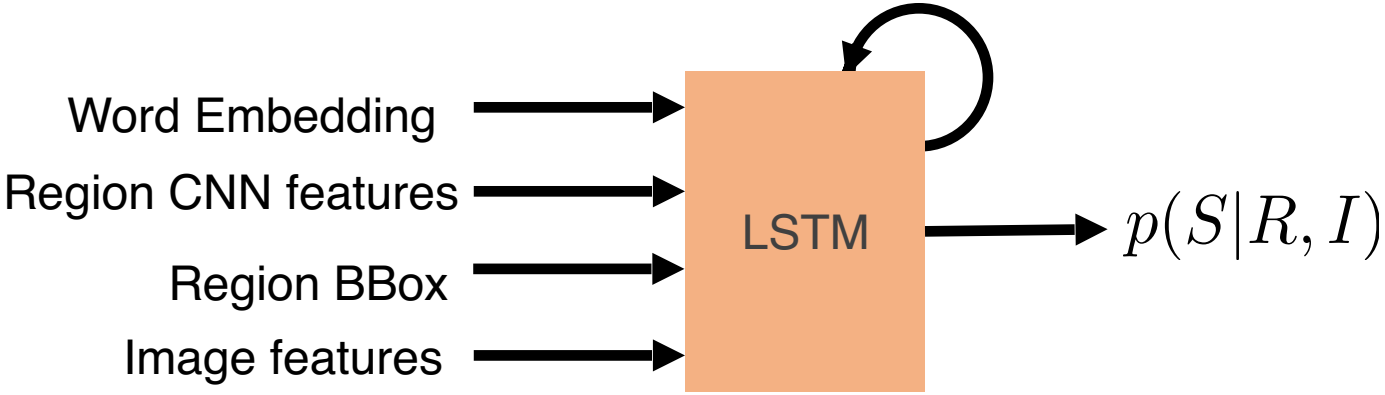
Modeling Context

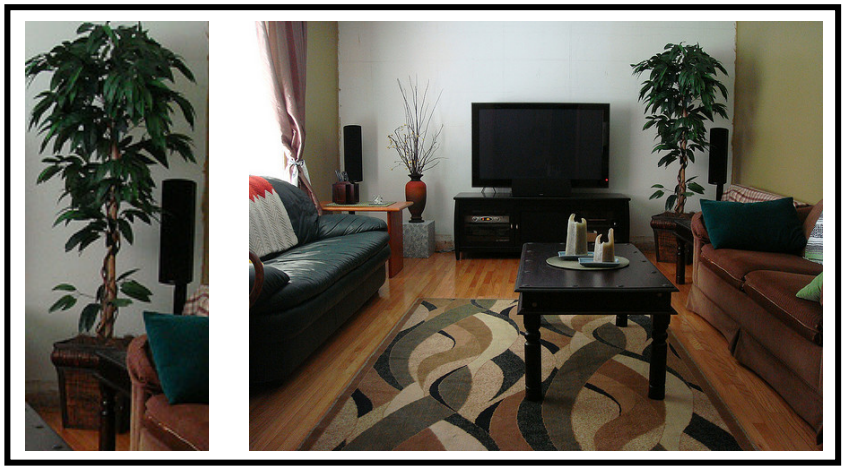
Baseline and Max-margin methods do not model locations of other objects

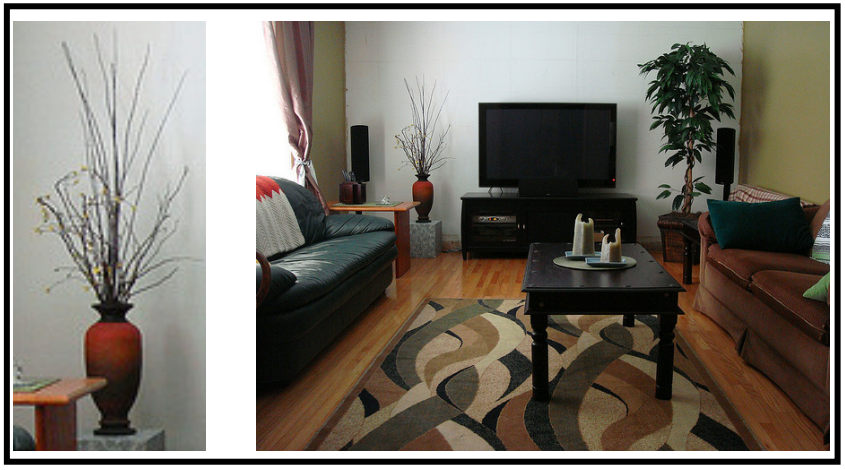
The plant on the right side of the TV

Modeling Context

Baseline and Max-margin architecture

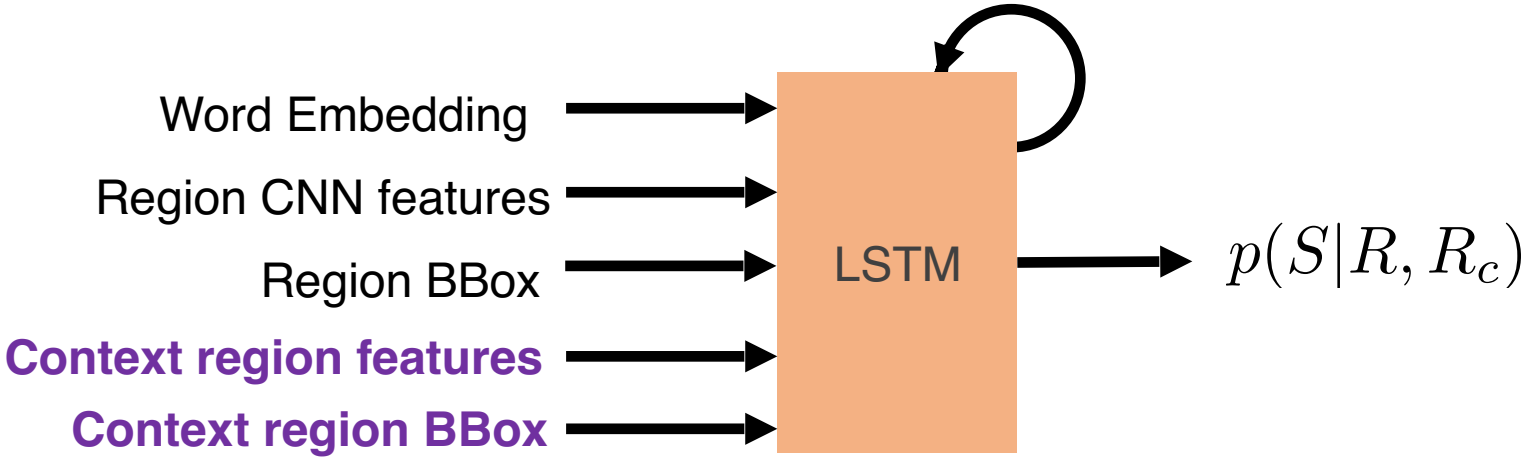




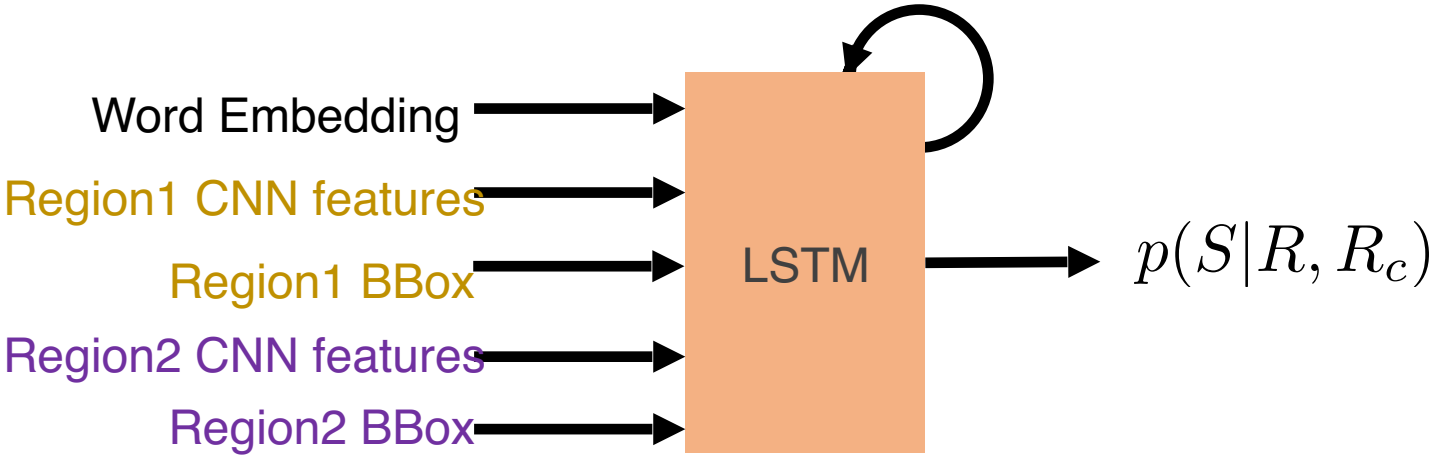


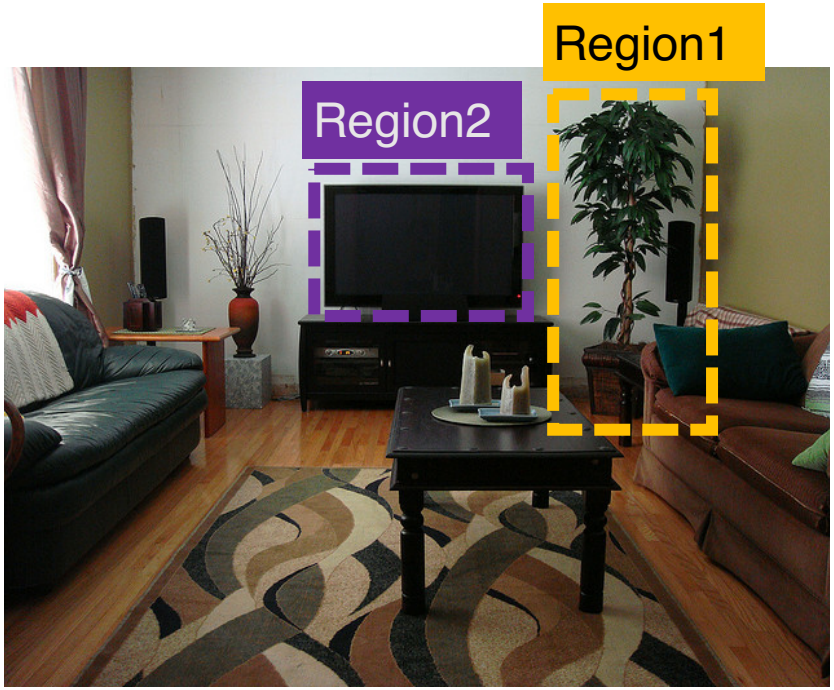
Modeling Context

Context model architecture

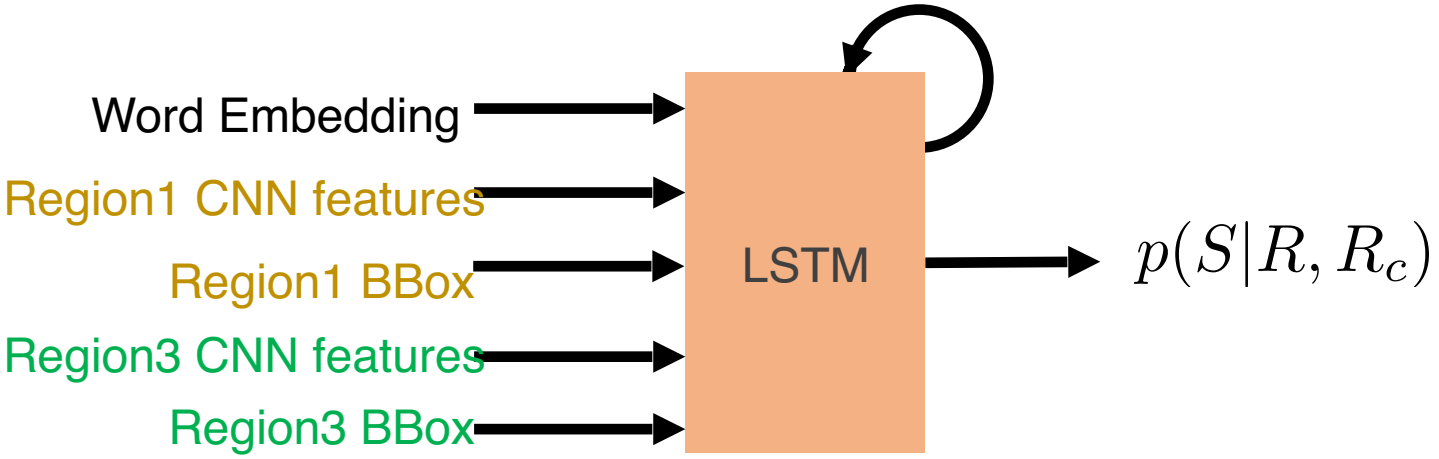


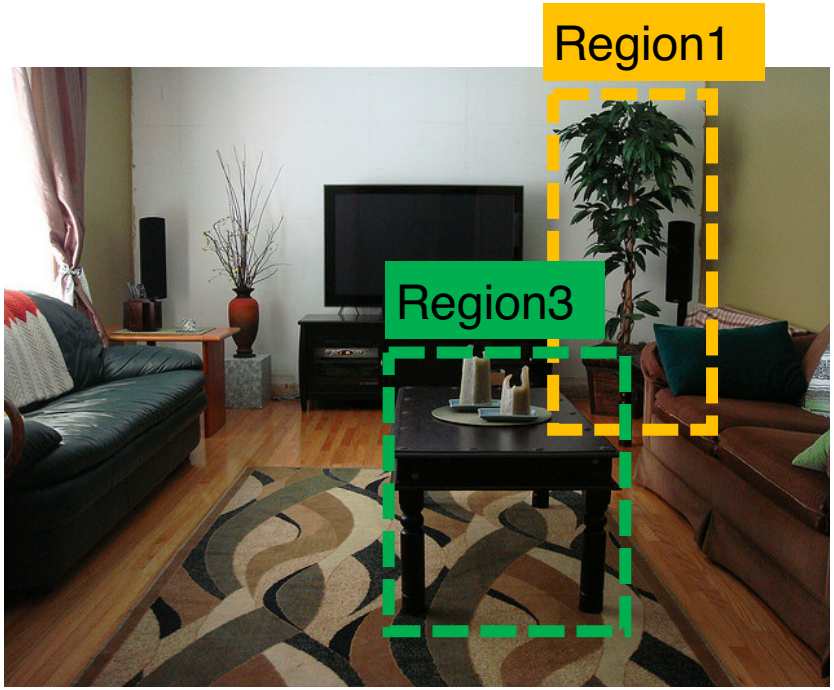
Modeling Context



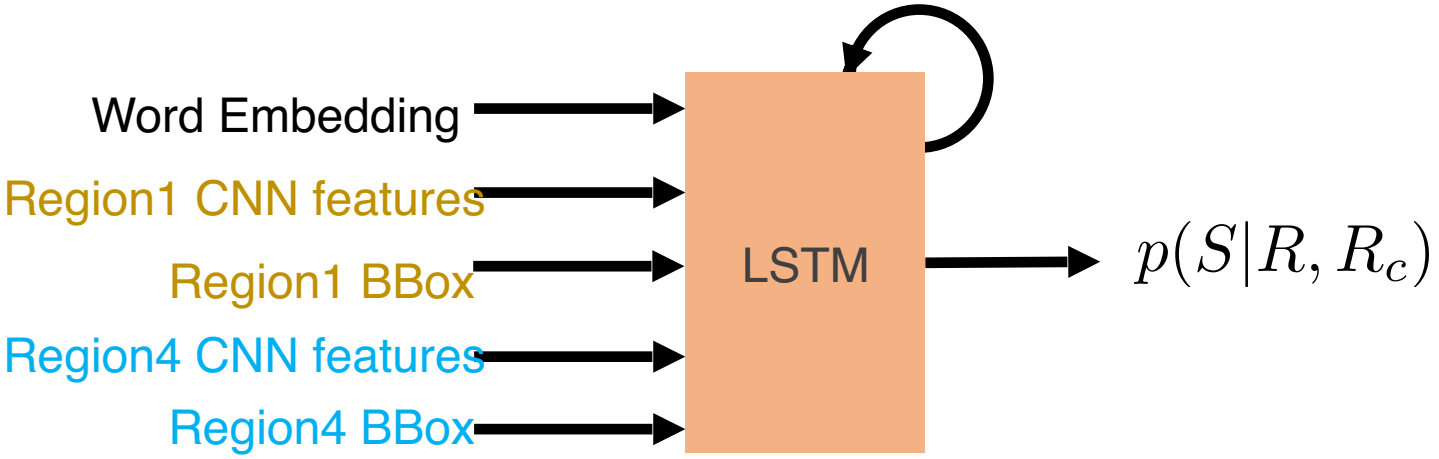


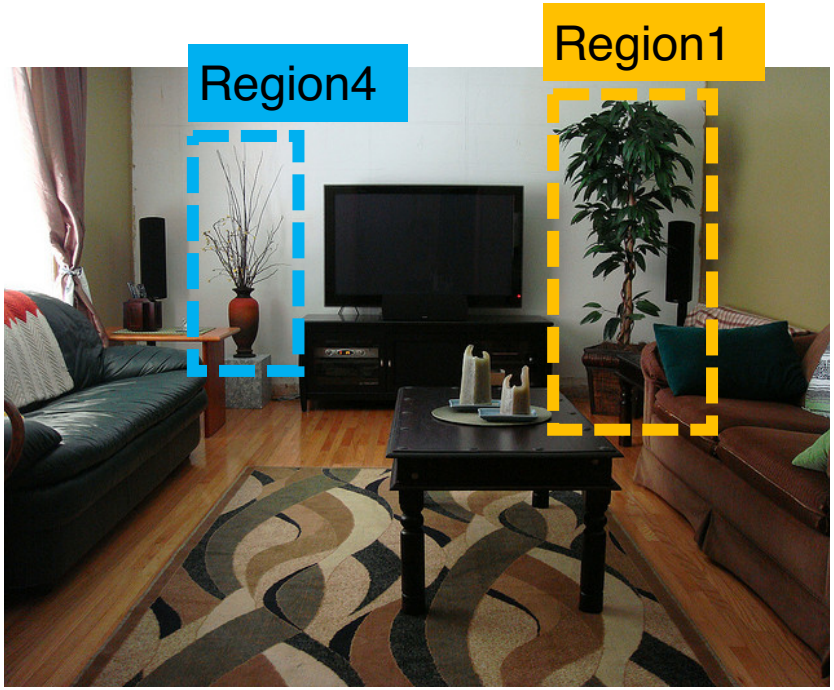
Modeling Context





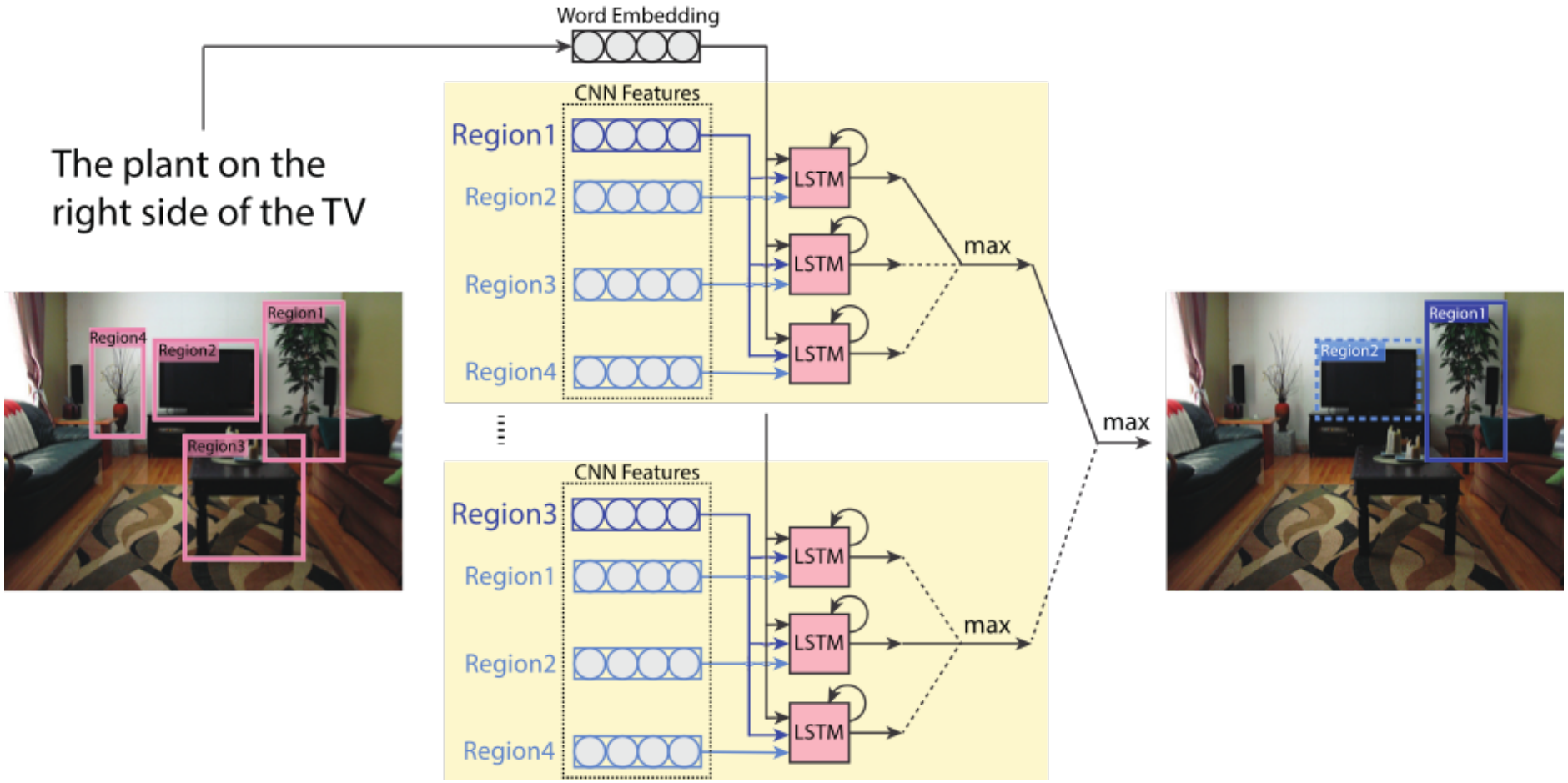
Modeling Context





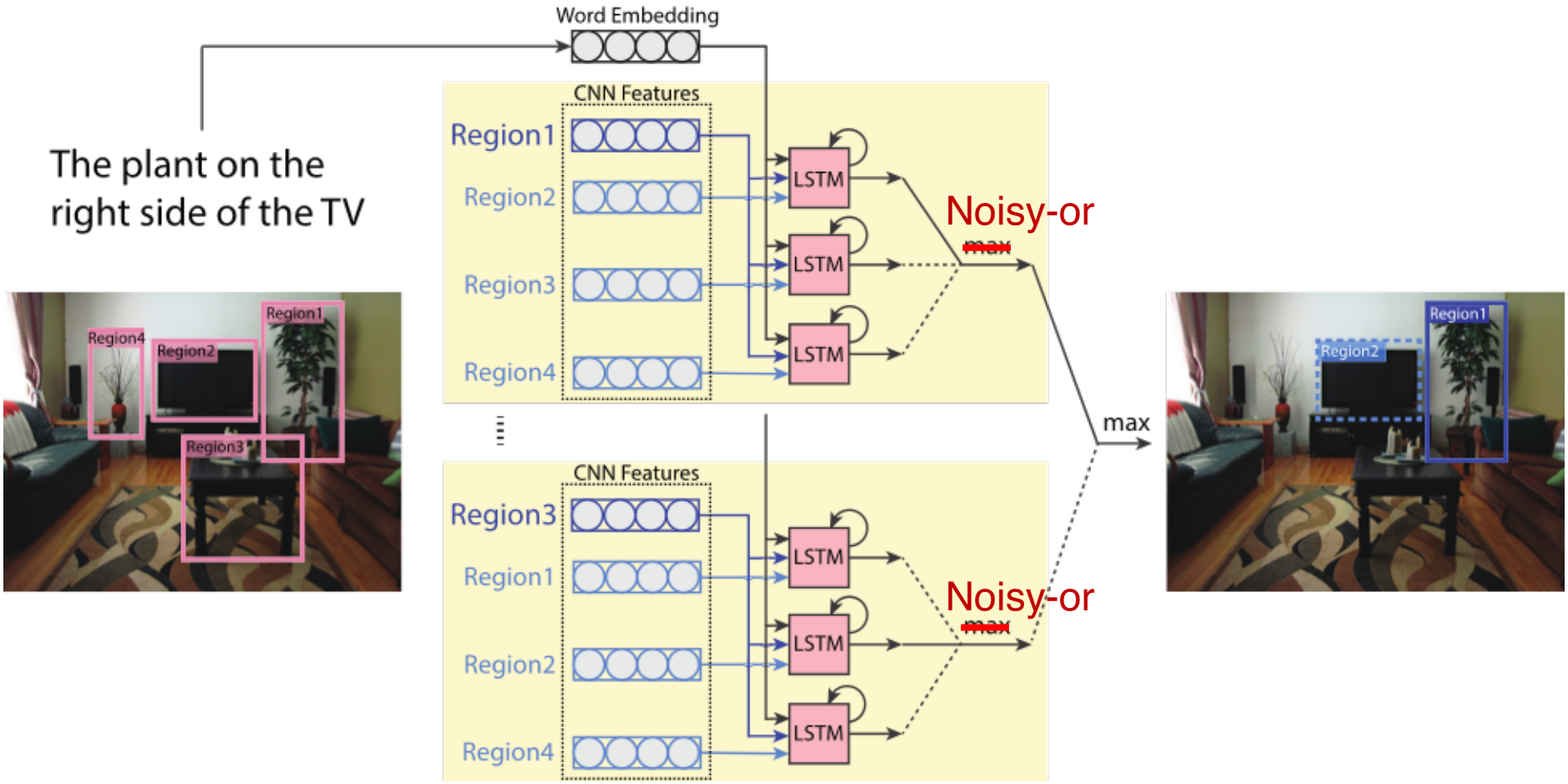
Modeling Context

Pooling context from multiple pairs of regions



Modeling Context

We can also use noisy-or pooling which is more robust



Training the Context Model

The challenge is that there are no annotations available for context objects

The plant on the right side of the TV

Multiple Instance Learning

So we use a MIL based technique and use the annotation of the referred object as weak supervision

The plant on the right side of the TV

Multiple Instance Learning

A **positive bag** of pairs where at least one of the pairs will be the positive sample.

Multiple Instance Learning

A **positive bag** of pairs where at least one of the pairs will be the positive sample.

Multiple Instance Learning

A **negative bag** where all the pairs are negative samples.

Multiple Instance Learning

A **negative bag** where all the pairs are negative samples.

Multiple Instance Learning

A **negative bag** where all the pairs are negative samples.

Implementation Details

Implemented in Caffe

Region and Image features
- VGG16 fc8 layer - fine-tuned.

Bounding box features
- scaled <xmin, ymin, xmax, ymax, area>

Word embedding size – 1024
LSTM hidden dimension – 1024

Region candidates – MCG technique
Region filtering process
- Obtain scores from Fast-RCNN and select regions above a threshold

Results

Datasets

Google RefExp	# refexp	# images
Train partition	85408	23199
Val partition	9602	2600
Test partition	Not released yet	

UNC RefExp	# refexp	# images
Train partition	120624	17000
Val partition	10834	1500
Person centric → TestA partition	5657	750
Object centric → TestB partition	5095	750

Google RefExp Results

A detection is considered true positive if the IOU score is greater than 0.5

All results are from noisy-or pooling

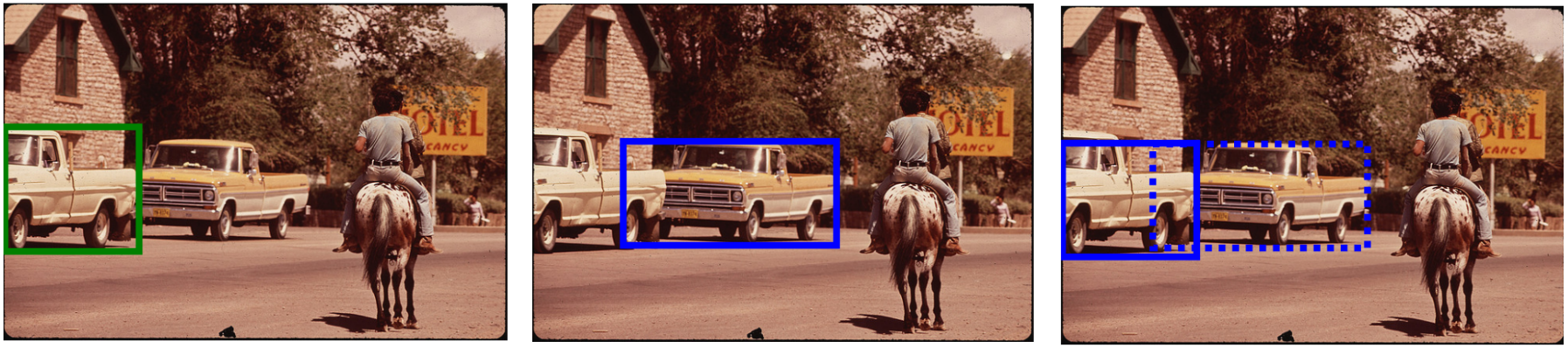
Google RefExp Validation Partition

Method \ Proposals	GT	MCG
Max Likelihood [Mao et al]	57.5	42.4
Max margin [Mao et al]	65.7	47.8
Ours, Neg. Bag margin	**68.4**	49.5
Ours, Pos. & Neg. Bag margin	**68.4**	**50.0**

Google RefExp Results

Groundtruth | Image context only | Noisy-or pooling

The chair closest to the lady



A white truck in front of a yellow truck

UNC RefExp Results

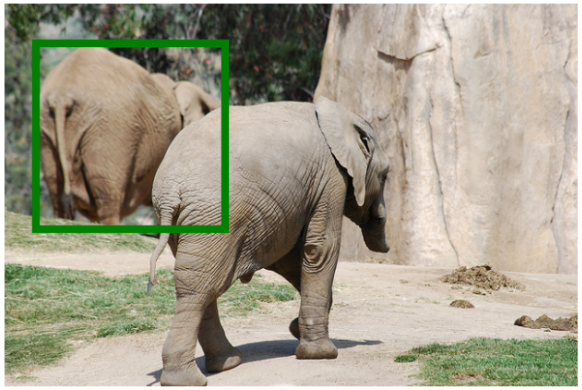
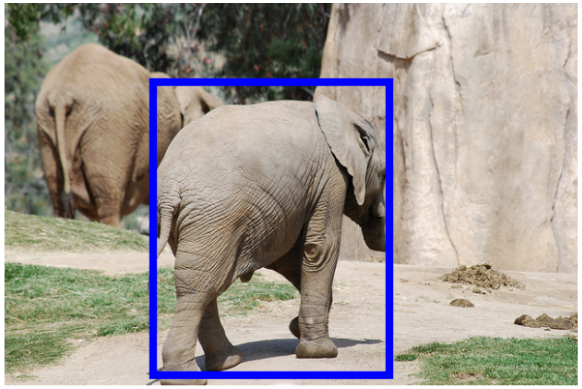
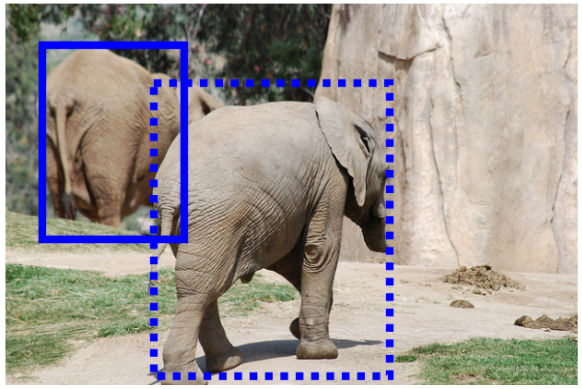
TestB Partition (Object centric)

Method \ Proposals	GT	MCG
Max Likelihood [Mao et al]	70.6	50.0
Max margin [Mao et al]	76.3	55.1
Ours, Neg. Bag margin	**78.0**	**56.4**
Ours, Pos. & Neg. Bag margin	76.1	56.3

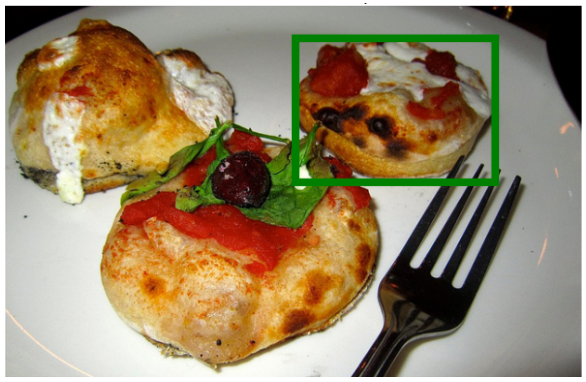
UNC RefExp Results

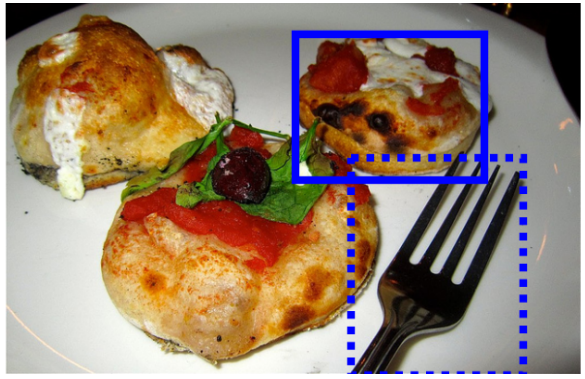
TestB Partition (Object centric)

Groundtruth | Image context only | Noisy-or pooling

Elephant towards the back





Food on the far back on the plate

UNC RefExp Results

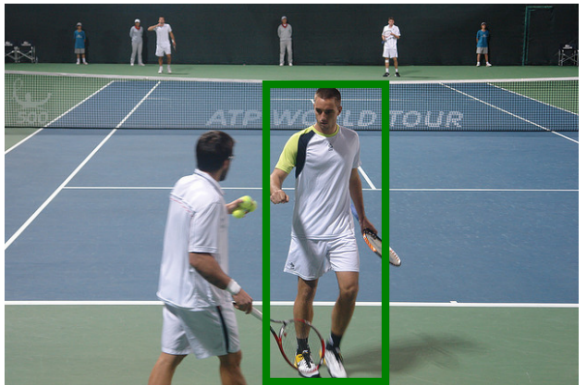
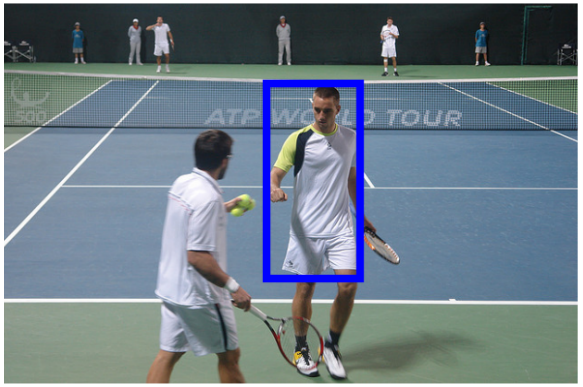
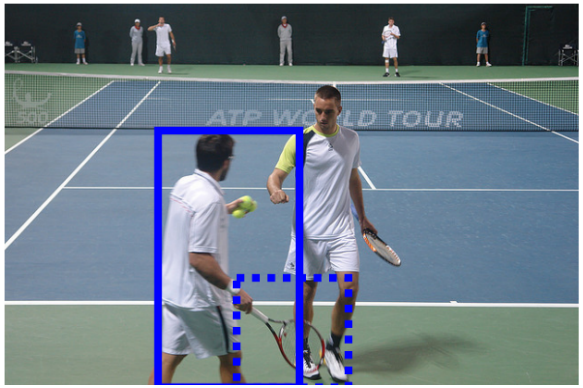
TestA Partition (Person centric)

Method \ Proposals	GT	MCG
Max Likelihood [Mao et al]	65.9	53.2
Max margin [Mao et al]	74.9	58.4
Ours, Neg. Bag margin	**75.6**	58.6
Ours, Pos. & Neg. Bag margin	75.0	**58.7**

UNC RefExp Results

TestA Partition (Person centric)

Groundtruth | Image context only | Noisy-or pooling

Guy on the tennis course



Blue on left

A few remarks about limitations

- Success depends on region proposal algorithm including candidates for the correct referred and context objects
  - Much more demanding than just requiring a candidate for the referred object
  - Ameliorated somewhat by having the entire image as a candidate context object
- Straightforward extension to include additional context objects (language can be deeply nested) intractable
- (Methodological) – would like to evaluate performance restricted to “relevant” referring expressions, but difficult to specify correct criteria for selection

Spatial Context Heatmap

We investigate the effect of context object on spatial locations of the referred object

A woman sitting on a bench

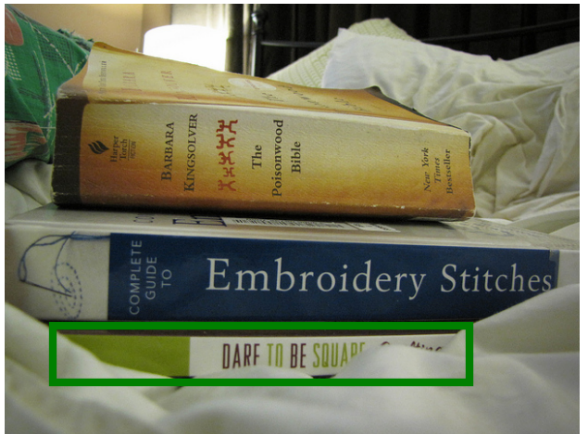
Spatial Context Heatmap

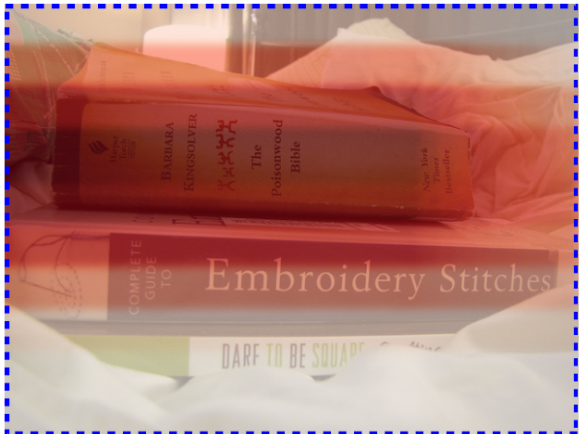
Groundtruth

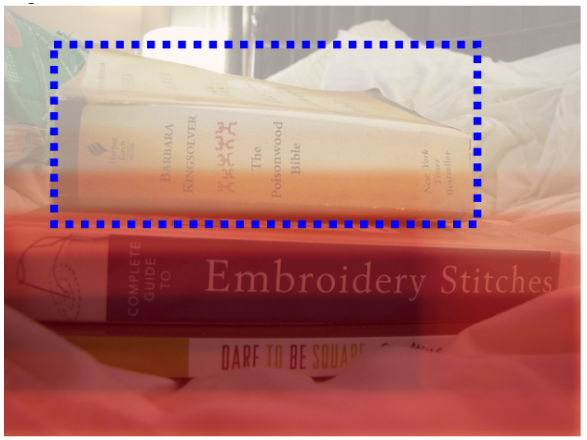
Image as context

Object as context

A woman sitting on a bench







A green and white book underneath two other books

Conclusion

Proposed a technique that models the probability of a referring expression as a function of a region and a context region.

MIL based objective functions can be used for training LSTMs to handle the lack of annotations for context objects.

Our models are capable of identifying the referred region along with the supporting context region.