Grounding of Textual Phrases in Images by Reconstruction

(ECCV, 2016)

Anna Rohrbach[1] Marcus Rohrbach[2,3] Ronghang Hu[2]
Trevor Darrell[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany

[2]UC Berkeley EECS, CA, United States

[3]ICSI, Berkeley, CA, United States

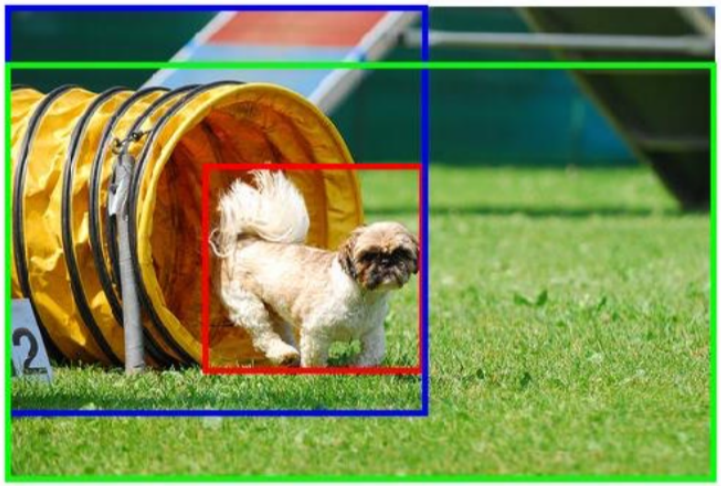
Presenters: Jiaxuan Chen, Meng Li
March 1, 2018

Outline

Outline

Motivation

- Language grounding (i.e.localizing) in visual data is a challenging problem.



A little brown and white dog emerges from a yellow collapsable toy tunnel onto the lawn.

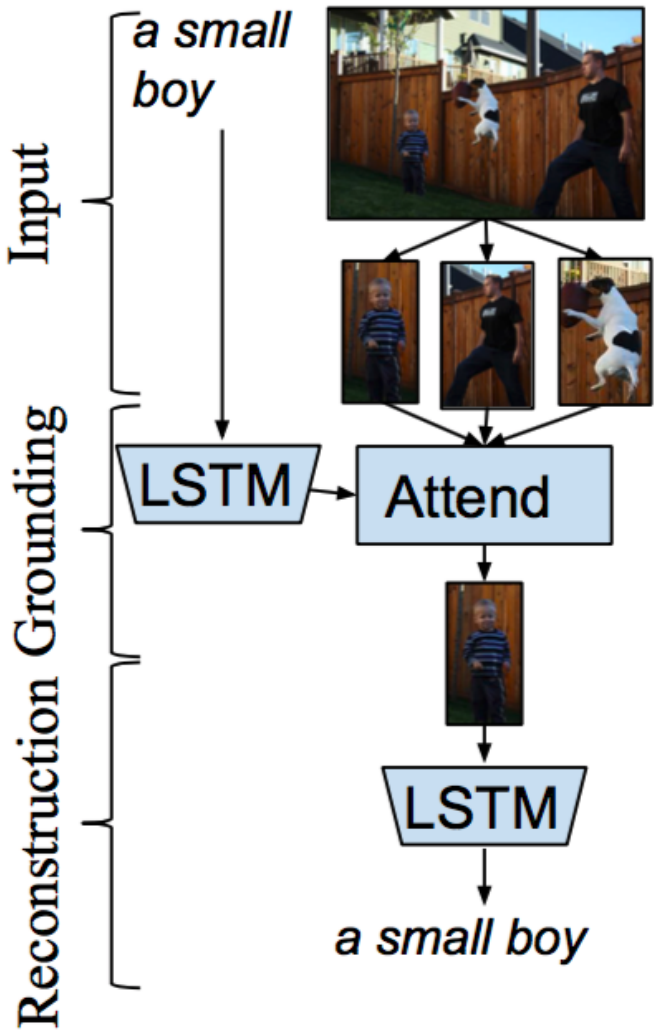
- Prior studies focus on constrained settings with a small number of nouns to ground.
- Few datasets provide the ground truth localization of phrases

Problem of this paper

- Problem:
  Given images paired with natural language phrases, grounding arbitrary textual phrases in images, with no, a few, or all bounding box annotations available.
- Relation to other works:
  This paper uses CNN, LSTM, soft-attention mechanism and a bi-directional mapping (Phrase $\Rightarrow$ Attended Image Region $\Rightarrow$ Reconstructed Phrase).

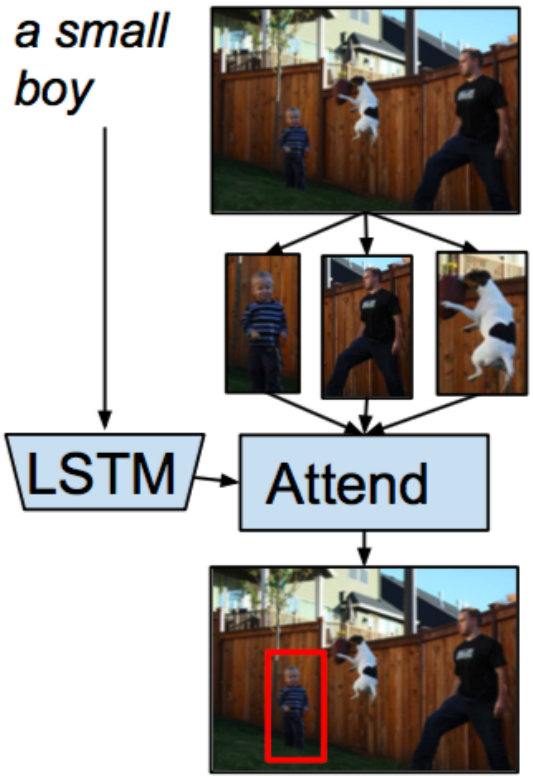
Outline

Training time



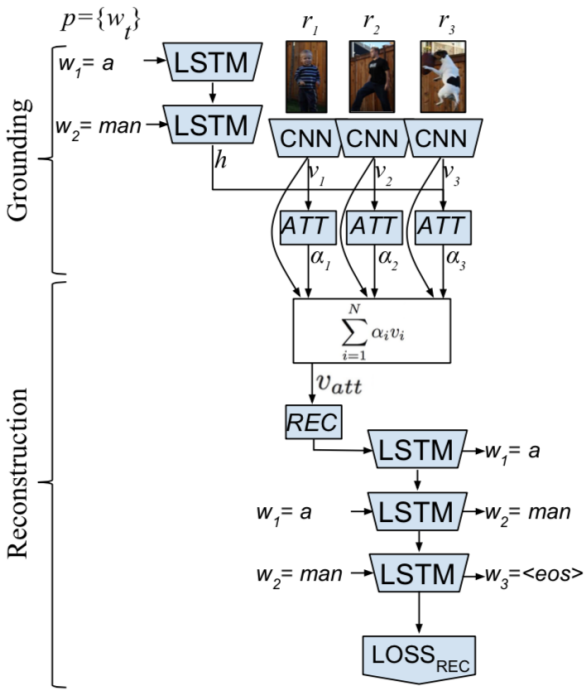
1. Given a phrase, attend to the most relevant region (or potentially also multiple regions) based on the phrase
2. Reconstruct the same phrase p from region(s) it attended to in the first phase

Test time

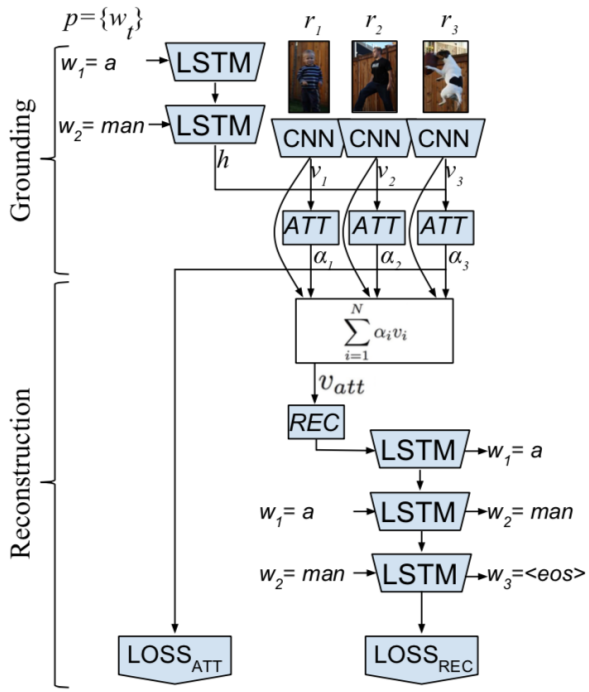


1. Given a phrase, attend to the most relevant region based on the phrase
2. Calculate the IOU (intersection over union) value between the selected box and the ground truth box.

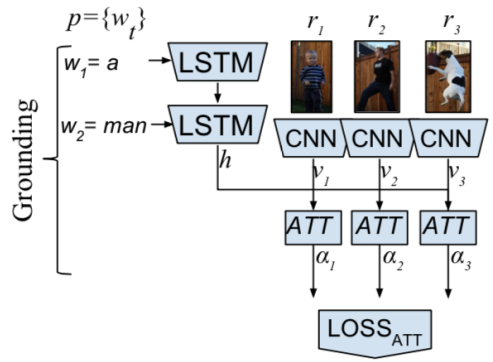
Model Overview



(a) Unsupervised

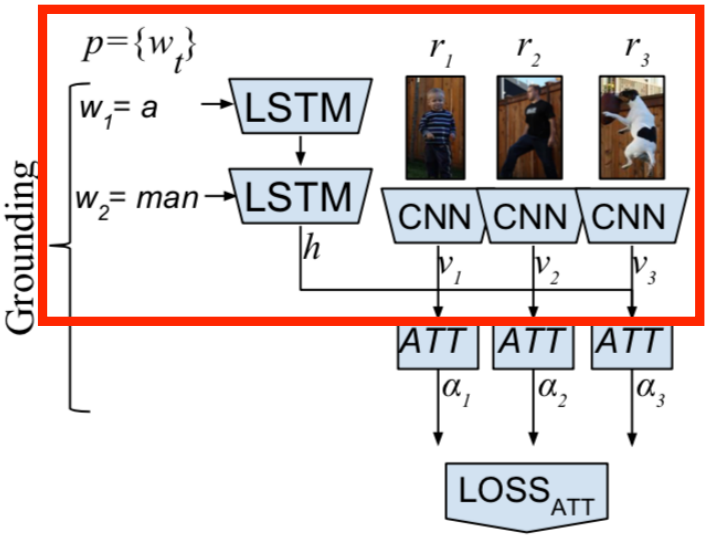


(b) Semi-supervised



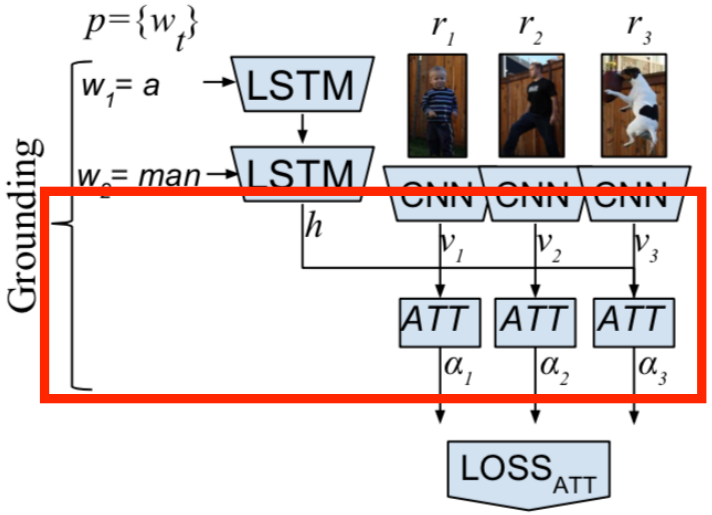
(c) Fully supervised

Learning to ground



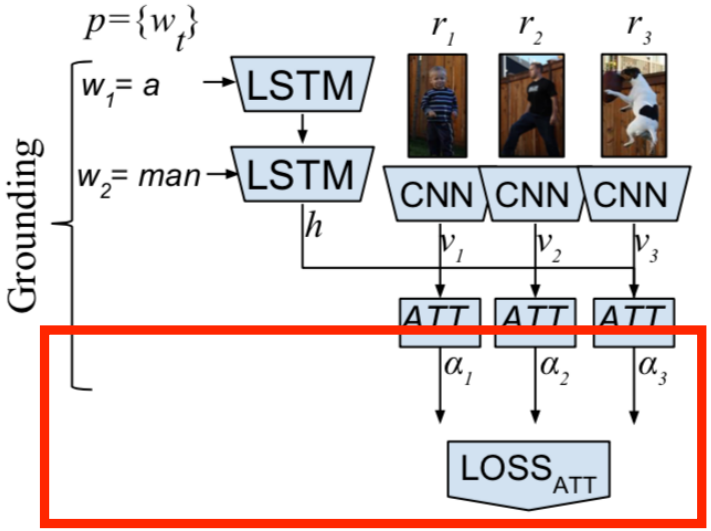
- Generate a set of image region proposals $\{r_i\}_{i=1,\ldots,N}$
- Encode phrase and proposal boxes $h = f_{LSTM}(p), v_i = f_{CNN}(r_i)$

Learning to ground



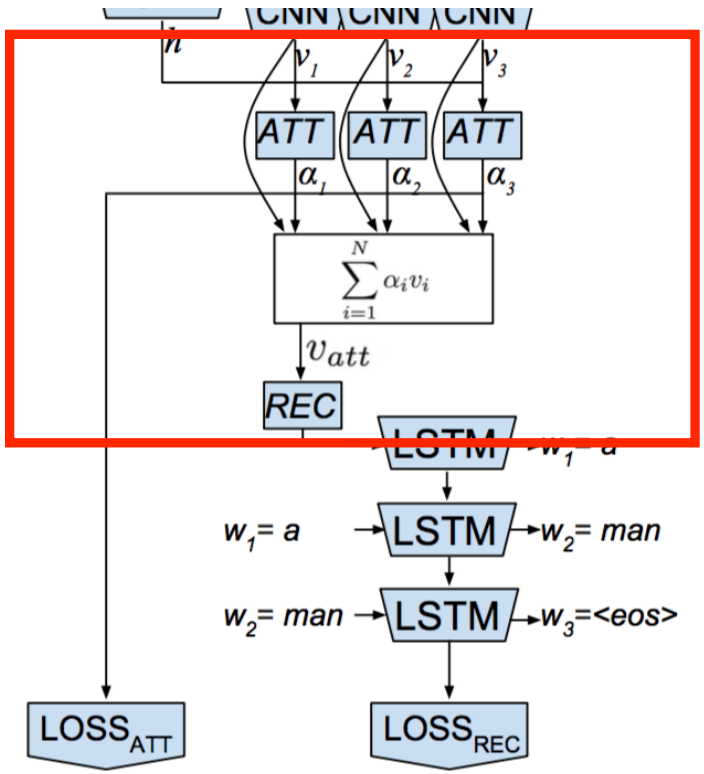
- Generate a set of image region proposals $\{r_i\}_{i=1,\ldots,N}$
- Encode phrase and proposal boxes $h = f_{LSTM}(p), v_i = f_{CNN}(r_i)$
- Apply Attention
  1. Attention weight: $\bar{\alpha}_i = f_{ATT}(p, r_i) = W_2 * \text{Relu}(W_h h + W_v v_i + b_1) + b_2$
  2. Normalized attention weight $\alpha_i = \text{softmax}(\bar{\alpha}_i)$
  3. E.g. $\alpha_1 = 0.2, \alpha_2 = 0.7, \alpha_3 = 0.1$

Learning to ground



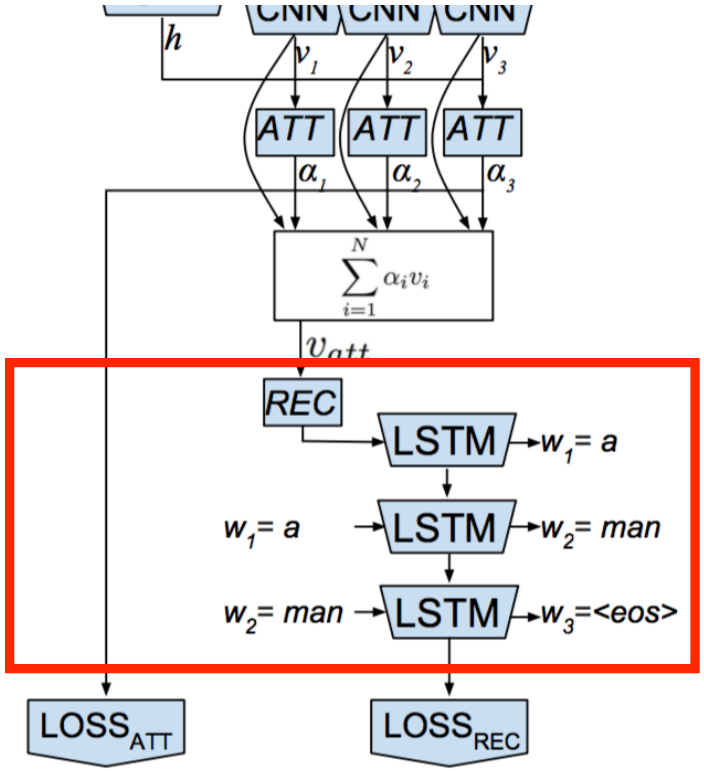
- Generate a set of image region proposals $\{r_i\}_{i=1,\ldots,N}$
- Encode phrase and proposal boxes $h = f_{LSTM}(p), v_i = f_{CNN}(r_i)$
- Apply Attention
  1. Attention weight: $\bar{\alpha}_i = f_{ATT}(p, r_i) = W_2 * \text{Relu}(W_h h + W_v v_i + b_1) + b_2$
  2. Normalized attention weight $\alpha_i = \text{softmax}(\bar{\alpha}_i)$
- Supervised training loss $L_{ATT} = -\frac{1}{B}\sum_{b=1}^{B} \log(\alpha_{\hat{j}})$, where $r_{\hat{j}}$ is the correct proposal box

Learning to reconstruct



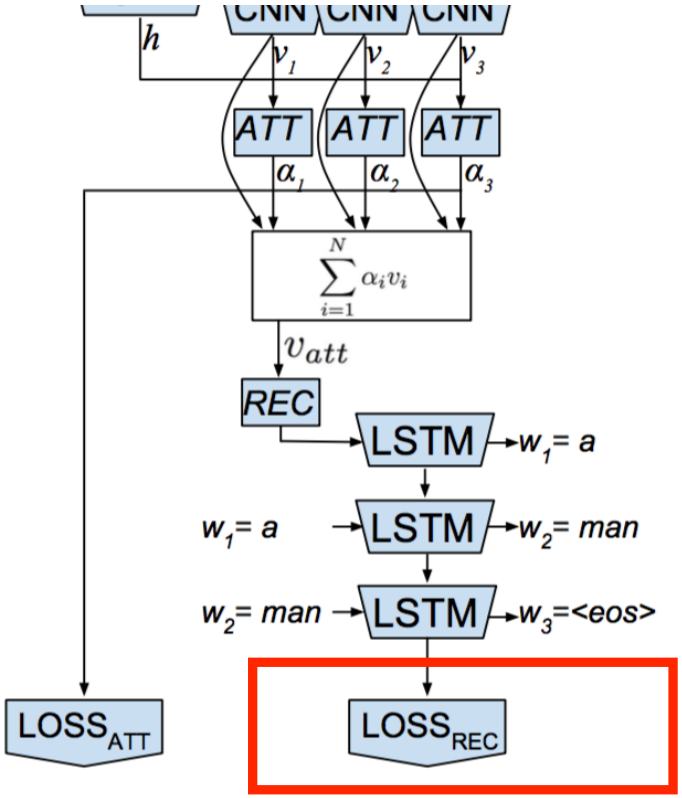
- Encode visual features
  1. $v_{ATT} = \sum_{i=1}^{N} \alpha_i v_i$
     $= 0.2v_1 + 0.7v_2 + 0.1v_3$
  2. $v'_{ATT} = f_{REC}(v_{ATT}) =$
     $\text{Relu}(W_a v_{ATT} + b_a)$

Learning to reconstruct



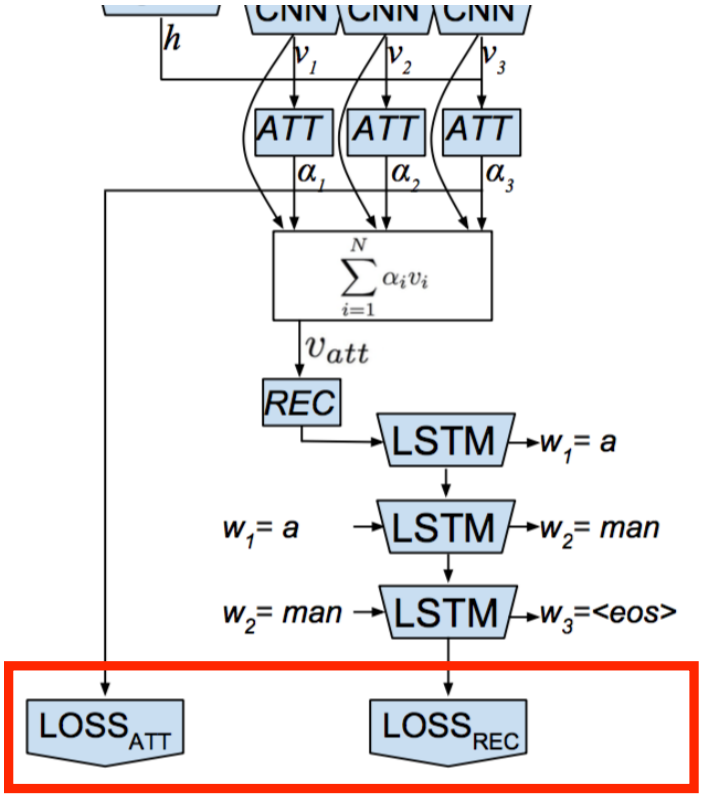
- Encode visual features
  1. $v_{ATT} = \sum_{i=1}^{N} \alpha_i v_i$
     $= 0.2v_1 + 0.7v_2 + 0.1v_3$
  2. $v'_{ATT} = f_{REC}(v_{ATT}) =$
     $\text{Relu}(W_a v_{ATT} + b_a)$
- Generated distribution over p
  $P(p|v'_{ATT}) = f_{LSTM}(v'_{ATT})$

Learning to reconstruct



- Encode visual features
  1. $v_{ATT} = \sum_{i=1}^{N} \alpha_i v_i$
     $= 0.2v_1 + 0.7v_2 + 0.1v_3$
  2. $v'_{ATT} = f_{REC}(v_{ATT}) =$
     $\text{Relu}(W_a v_{ATT} + b_a)$
- Generated distribution over p
  $P(p|v'_{ATT}) = f_{LSTM}(v'_{ATT})$
- Unsupervised training loss
  $L_{REC} = -\frac{1}{B} \sum_{b=1}^{B} \log(P(\hat{p}|v'_{ATT}))$,
  where $\hat{p}$ is the ground-truth phrase

Learning to reconstruct



- Encode visual features
  1. $v_{ATT} = \sum_{i=1}^{N} \alpha_i v_i$
     $= 0.2v_1 + 0.7v_2 + 0.1v_3$
  2. $v'_{ATT} = f_{REC}(v_{ATT}) =$
     $\text{Relu}(W_a v_{ATT} + b_a)$
- Generated distribution over p
  $P(p|v'_{ATT}) = f_{LSTM}(v'_{ATT})$
- Unsupervised training loss
  $L_{REC} = -\frac{1}{B} \sum_{b=1}^{B} \log(P(\hat{p}|v'_{ATT}))$,
  where $\hat{p}$ is the ground-truth phrase
- Semi-supervised training loss
  $L = \lambda L_{ATT} + L_{REC}$

Outline

Experimental Setup I

Datasets: Flickr 30k Entities and ReferItGame

- Flickr 30k Entities: over 275K bounding boxes from 31K images with natural language phrases
- ReferItGame: over 99K regions from 20K images with natural language expressions

Generate 100 bounding box proposals for each image using

- Selective Search for Flickr 30k Entities
- Edge Boxes for ReferItGame

For semi- and fully supervised models, the ground-truth attention is obtained by selecting the proposal box that overlaps most with the ground-truth box.

Experimental Setup II

Flickr 30k Entities

- VGG-CLS: VGG16 network trained on ImageNet
- VGG-DET: VGG16 network fine-tuned for object detection on PASCAL, trained using Fast R-CNN

ReferItGame

- VGG+SPAT: VGG-CLS features and additional spatial features provided by Hu et al (2016).

Test time evaluation: the ratio of phrases for which the attended box overlaps with the ground-truth box by more than 0.5 IOU.

Outline

Experiments on Flickr 30k Entities dataset

Approach	Accuracy		
	Other	VGG-CLS	VGG-DET
Unsupervised training			
Deep Fragments [6]	21.78	-	-
GroundeR	-	24.66	28.94
Supervised training			
CCA [35]	-	27.42	-
SCRC [18]	-	27.80	-
DSPE [45]	-	-	43.89
GroundeR	-	41.56	47.81
Semi-supervised training			
GroundeR 3.12% annot.	-	33.02	42.32
GroundeR 6.25% annot.	-	37.10	44.02
GroundeR 12.5% annot.	-	38.67	44.96
GroundeR 25.0% annot.	-	39.31	45.32
GroundeR 50.0% annot.	-	40.72	46.65
GroundeR 100.0% annot.	-	42.43	48.38
Proposal upperbound	77.90	77.90	77.90

Table 1: Phrase localization performance on Flickr 30k Entities with different levels of bounding box supervision, accuracy in %.

Experiments on ReferItGame dataset

Approach	Accuracy		
	Other	VGG	VGG+SPAT
Unsupervised training			
LRCN [9] (reported in [18])	8.59	-	-
CAFFE-7K [15] (reported in [18])	10.38	-	-
GroundeR	-	10.69	10.70
Supervised training			
SCRC [18]	-	-	17.93
GroundeR	-	23.44	26.93
Semi-supervised training			
GroundeR 3.12% annot.	-	13.70	15.03
GroundeR 6.25% annot.	-	16.19	19.53
GroundeR 12.5% annot.	-	19.02	21.65
GroundeR 25.0% annot.	-	21.43	24.55
GroundeR 50.0% annot.	-	22.67	25.51
GroundeR 100.0% annot.	-	24.18	28.51
Proposal upperbound	59.38	59.38	59.38

Table 3: Phrase localization performance on ReferItGame with different levels of bounding box supervision, accuracy in %.

Qualitative results on Flickr 30k Entities

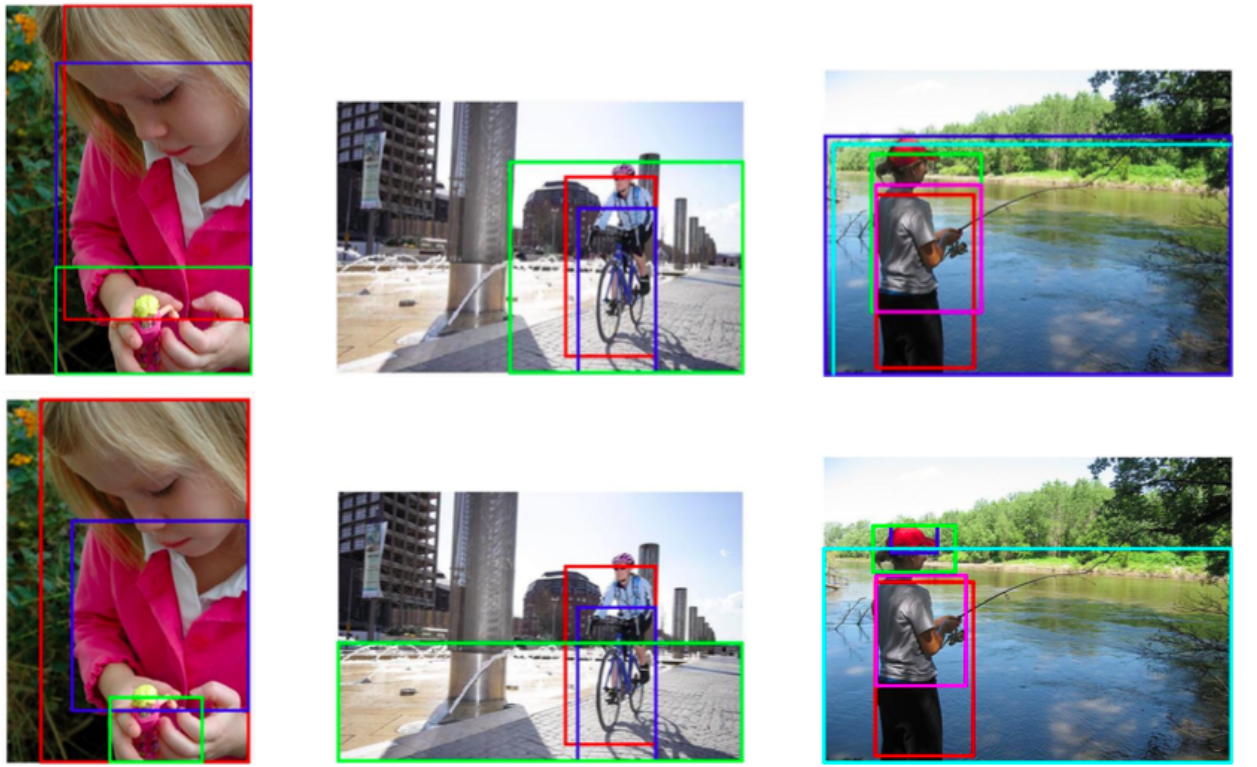
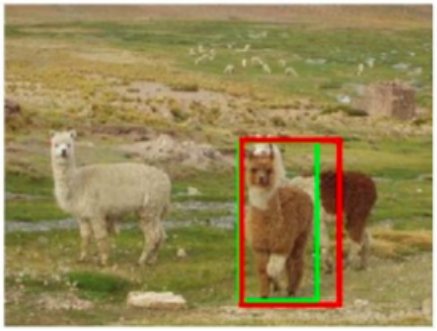


Fig. 3: Qualitative results on the test set of Flickr 30k Entities. Top : GroundeR (VGG-DET) unsupervised, bottom: GroundeR (VGG-DET) supervised.

Qualitative results on ReferItGame

two people on right

picture of a bird flying above sand



dat alpaca up in front, total coffeelate swag

palm tree coming out of the top of the building

guy with blue shirt and yellow shorts

hut to the nearest left of the person on the right

Fig. 4: Qualitative results on the test set of ReferItGame: GroundeR (VGG+SPAT) supervised. Green: ground-truth box, red: predicted box.

Outline

Comments

Contribution:

- Applicability to un-, semi-, and supervised training regimes

Limitations:

- Ignore the object relationships
- Select a single proposal box that overlaps the most with ground-truth box but not the union

Outline

Potential Extensions

- Fine tune VGG
- Apply attention on input phrases
- Different language model

Thanks for your attention!