Attention Correctness in Neural Image Captioning

Chenxi Liu

Joint work with Junhua Mao, Fei Sha, Alan Yuille

11/27/2016

Outline

- “Classic” Image Captioning Models
- Deep Attention in Image Captioning
- Evaluation of Visual Attention
- Supervision on Visual Attention
- Results and Discussion

“Classic” Image Captioning Models

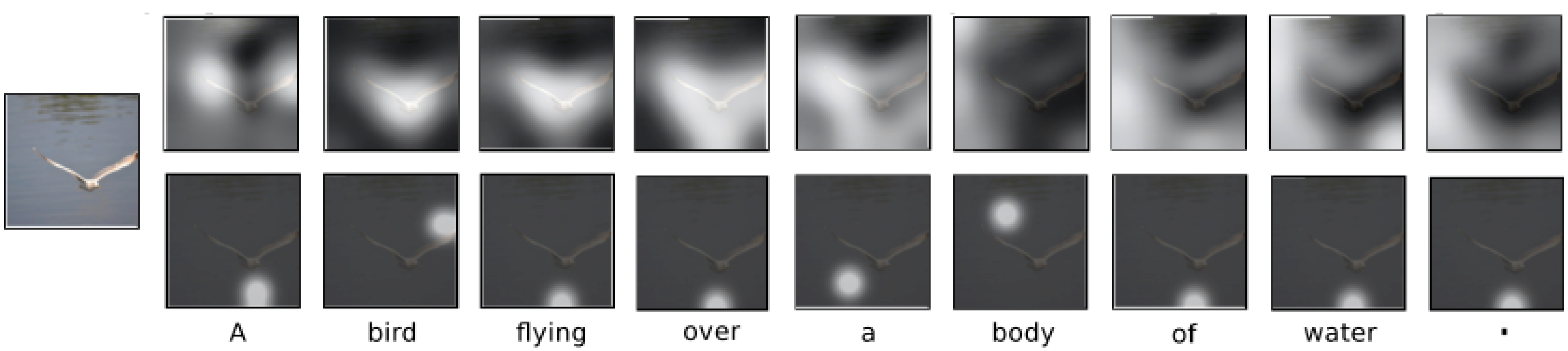
A little boy playing with a yellow shovel

$I \longrightarrow y_1, \ldots, y_T$

- $a = CNN(I)$ — CNN is usually pretrained on ImageNet
- $h_t = RNN(y_{t-1}, h_{t-1}, a)$ — RNN can be an LSTM
- $p(y_t | y_1, \ldots, y_{t-1}, I) = g(h_t)$ — g is usually a MLP

Deep Attention in Image Captioning

- Xu, Kelvin, et al. “Show, attend and tell: Neural image caption generation with visual attention.” *ICML 2015*
- Intuition
  - The image feature does not contain location information
  - Different words describe different regions of the image
  - Can this dynamic alignment be modeled and learned?



Deep Attention in Image Captioning

- $a_{1:L} = CNN(I)$ Now conv layer feature
- $h_t = RNN(y_{t-1}, h_{t-1}, z_t)$ Context vector is dynamic

- $z_t = \sum_{i=1}^{L} \alpha_{ti} a_i$ Weighted sum of per-location features
- $\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^{L} \exp(e_{ti})}$ Softmax: attention sums to 1
- $e_{ti} = \mathrm{f}(a_i, h_{t-1})$ f is usually a MLP

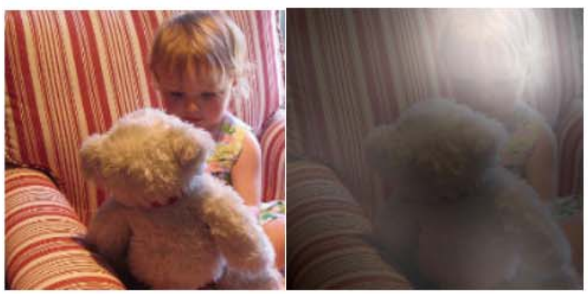
- Amazingly, the whole thing is differentiable

Deep Attention in Image Captioning

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

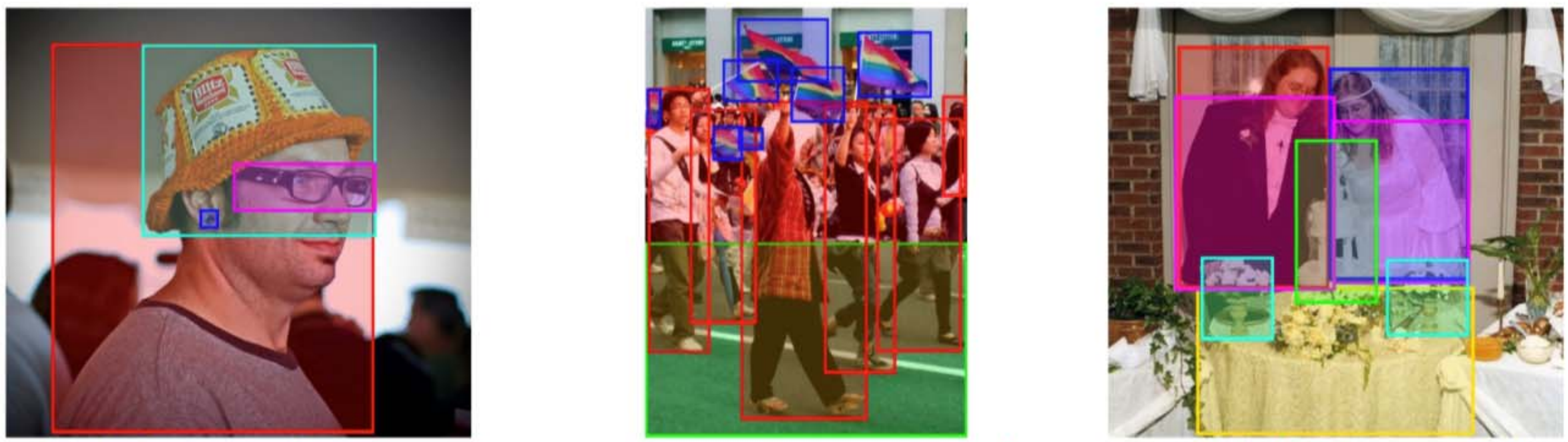
A giraffe standing in a forest with trees in the background.

So… what’s the problem?

- The attention maps carry important information in understanding (and potentially improving) deep networks
- Although impressive visualization results of the attention maps are shown, there are no quantitative evaluations
- In other words, the visualizations could be cherry-picked
- Therefore, we study the following two questions:
  - (Evaluation) How often and to what extent are the attention maps consistent with human perception/annotation?
  - (Supervision) Will more human-like attention maps result in better captioning performance?

But… where do we find GT attention?

- Plummer, Bryan A., et al. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” *ICCV* 2015.

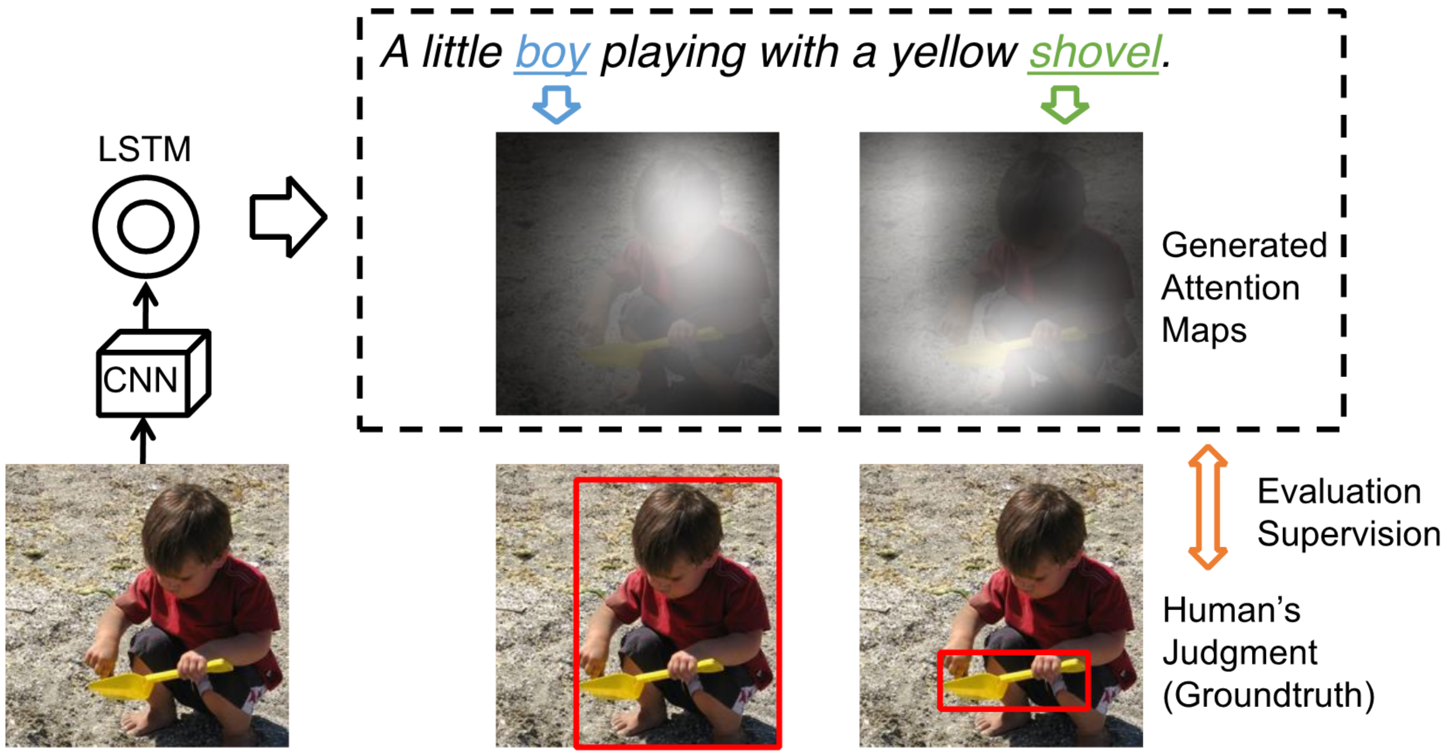


A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Summary



A little boy playing with a yellow shovel.
LSTM
CNN
Generated Attention Maps
Evaluation Supervision
Human's Judgment (Groundtruth)

Evaluation of Visual Attention

- In answer to Q1
- We define attention correctness as a metric that scores he consistency between an attention map and the ground truth region
- Attention Correctness of a word:
- $AC(y_t) = \sum_{i \in R_t} \alpha_{ti}$
- Attention Correctness of a phrase:
- $AC(\{y_t, \dots, y_{t+l}\}) = \max(AC(y_t), \dots, AC(y_{t+l}))$

0.08	0.12	0.20	0.12
0.04	0.10	0.12	0.08
0.00	0.02	0.08	0.04
0.00	0.00	0.00	0.00

Supervision on Visual Attention

- In answer to Q2
- We encourage the generated attention to resemble GT attention by introducing explicit supervision
- $L_{attn} = \begin{cases} -\sum_{i=1}^{L} \beta_{ti} \log \alpha_{ti} \\ 0 \end{cases}$
- $L = L_{orig} + \lambda L_{attn}$
- The question remains is how to construct β_{ti}

Supervision on Visual Attention

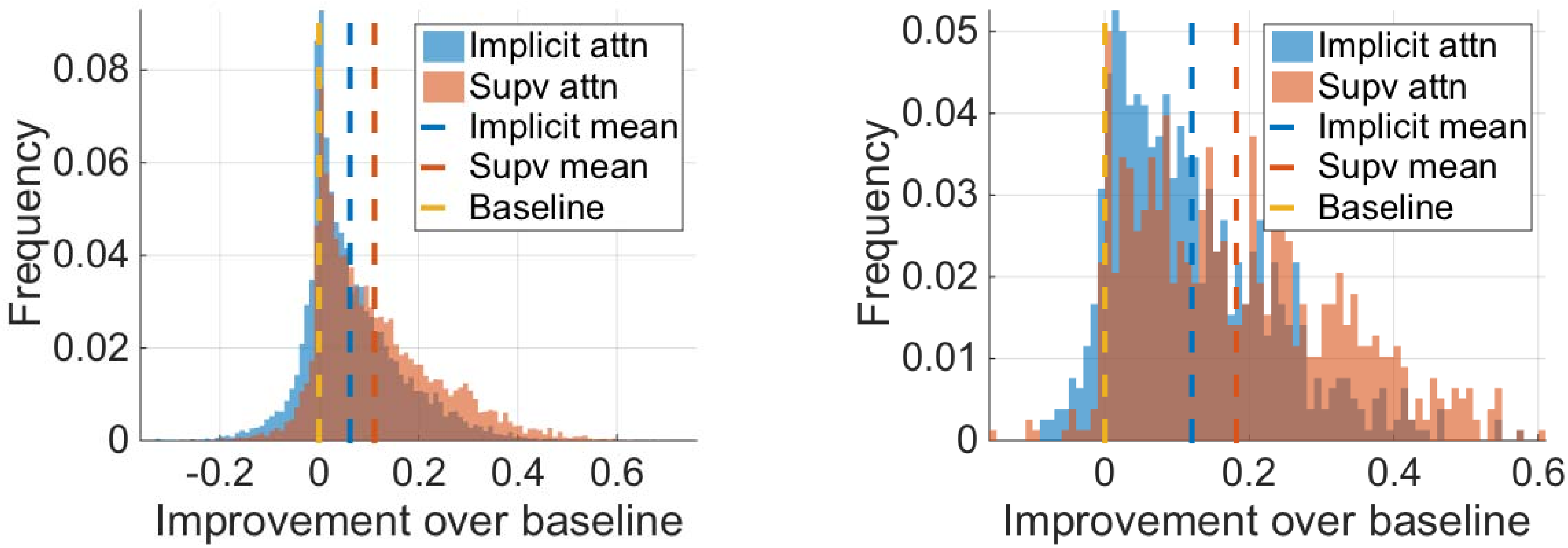
- Strong Supervision with Alignment Annotation
  - In Flickr30k Entities, the corresponding region of a phrase is given
  - So we construct β_{ti} from the corresponding region
- Weak Supervision with Semantic Labeling
  - In MS COCO, the corresponding region of a phrase is not annotated
  - We can "guess" β_{ti} from the instance segmentation masks with 80 semantic classes. For example, for the caption "A boy is playing with a dog", the model should probably attend to the region of "person" class when generating the word "boy"
  - This is not ideal of course

Results of Attention Correctness

Caption	Model	Baseline	Correctness
Ground Truth	Implicit	0.3214	0.3836
	Supervised	0.3214	**0.4329**
Generated	Implicit	0.3995	0.5202
	Supervised	0.3968	**0.5787**

- Baseline: attending equally everywhere (not learning any meaningful attention)
- The implicit attention model outperforms the baseline by 12%, so the model is indeed learning some meaningful attention
- The supervised attention model outperforms the baseline by 18%, i.e. our model is better at localizing the corresponding region

Results of Attention Correctness



Results of Caption Quality

- The fact that our model has better attention correctness is not too much of a surprise
- We may be more interested in whether supervised attention model also has better captioning performance
- The intuition is that a meaningful dynamic weighting of the input vectors will allow later components to decode information more easily

Results of Caption Quality

Table 3: Comparison of image captioning performance. * indicates our implementation. Caption quality consistently increases with supervision, whether it is strong or weak.

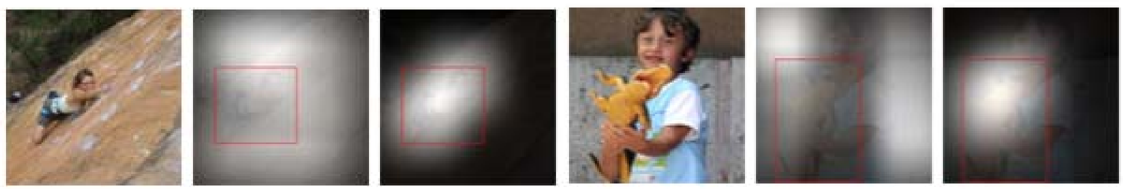
Dataset	Model	BLEU-3	BLEU-4	METEOR
Flickr30k	Implicit	28.8	19.1	18.49
	Implicit*	29.2	20.1	19.10
	Strong Sup	**30.2**	**21.0**	**19.21**
COCO	Implicit	34.4	24.3	23.90
	Implicit*	36.4	26.9	24.46
	Weak Sup	**37.2**	**27.6**	**24.78**

Results of Caption Quality

Table 4: Captioning scores on the Flickr30k test set for different attention correctness levels in the generated caption, implicit attention experiment. Higher attention correctness results in better captioning performance.

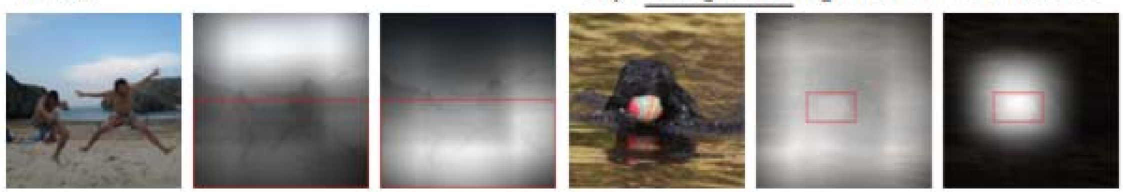
Correctness	**BLEU-3**	**BLEU-4**	**METEOR**
High	38.0	28.1	23.01
Middle	36.5	26.1	21.94
Low	35.8	25.4	21.14

Qualitative Results



Girl rock climbing on the rock wall.

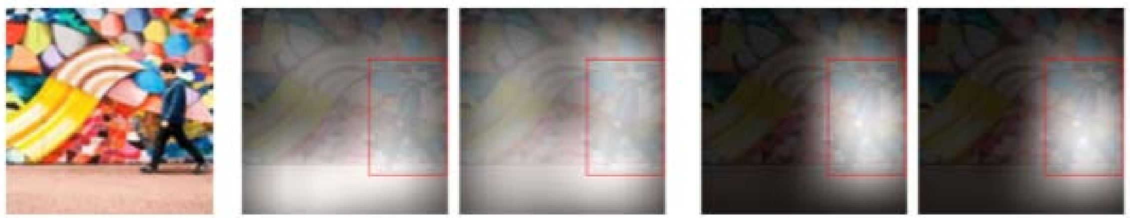
A young smiling child hold his toy alligator up to the camera.



Two male friends in swimming trunks jump on the beach while people in the background lay in the sand.

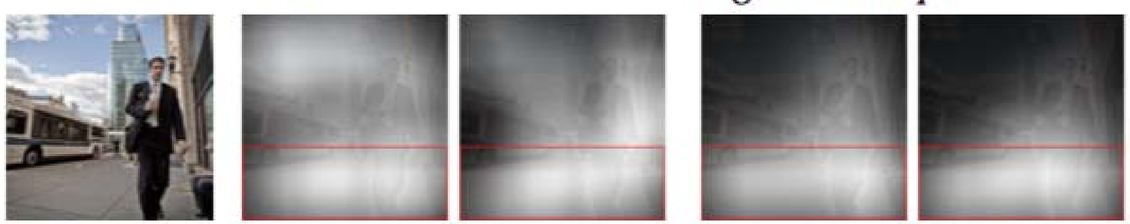
A black dog swims in water with a colorful ball in his mouth.

Qualitative Results



A man in a blue shirt and blue pants is sitting on a wall.

A man in a blue shirt and blue pants is skateboarding on a ramp.



A man and a woman are walking down the street.

A man and a woman are walking down the street.

Discussion

- Visual attention allows us to peek into the deep learning black box, and shows us how machines interpret the image
- However, its interpretation is not entirely consistent with human perception, which is arguably a more “reasonable” and “low energy” interpretation. A similar conclusion was also reached recently in visual question answering
- Attention is essentially a (normalized) similarity function that bears resemblance to semantic segmentation. In the future I plan to draw more connection between attention and semantic segmentation

Thank you!