

# Attention Correctness in Neural Image Captioning

Chenxi Liu<sup>1</sup> Junhua Mao<sup>2</sup> Fei Sha<sup>2,3</sup> Alan Yuille<sup>1,2</sup>

Johns Hopkins University<sup>1</sup>  
University of California, Los Angeles<sup>2</sup>  
University of Southern California<sup>3</sup>

## Abstract

Attention mechanisms have recently been introduced in deep learning for various tasks in natural language processing and computer vision. But despite their popularity, the “correctness” of the implicitly-learned attention maps has only been assessed qualitatively by visualization of several examples. In this paper we focus on evaluating and improving the correctness of attention in neural image captioning models. Specifically, we propose a quantitative evaluation metric for the consistency between the generated attention maps and human annotations, using recently released datasets with alignment between regions in images and entities in captions. We then propose novel models with different levels of explicit supervision for learning attention maps during training. The supervision can be strong when alignment between regions and caption entities are available, or weak when only object segments and categories are provided. We show on the popular Flickr30k and COCO datasets that introducing supervision of attention maps during training solidly improves both attention correctness and caption quality, showing the promise of making machine perception more human-like.

## Introduction

Recently, attention based deep models have been proved effective at handling a variety of AI problems such as machine translation (Bahdanau, Cho, and Bengio 2014), object detection (Mnih et al. 2014; Ba, Mnih, and Kavukcuoglu 2014), visual question answering (Xu and Saenko 2015; Chen et al. 2015), and image captioning (Xu et al. 2015). Inspired by human attention mechanisms, these deep models learn dynamic weightings of the input vectors, which allow for more flexibility and expressive power.

In this work we focus on attention models for image captioning. The state-of-the-art image captioning models (Kiros, Salakhutdinov, and Zemel 2014; Mao et al. 2015; Karpathy and Fei-Fei 2015; Donahue et al. 2015; Vinyals et al. 2015) adopt Convolutional Neural Networks (CNNs) to extract image features and Recurrent Neural Networks (RNNs) to decode these features into a sentence description. Within this encoder-decoder framework (Cho et al. 2014), the models proposed by (Xu et al. 2015) apply an attention mechanism, i.e. attending to different areas of the image when generating words one by one.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Figure 1: Image captioning models (Xu et al. 2015) can attend to different areas of the image when generating the words. However, these generated attention maps may not correspond to the region that the words or phrases describe in the image (e.g. “shovel”). We evaluate such phenomenon quantitatively by defining attention correctness, and alleviate this inconsistency by introducing explicit supervision. In addition, we show positive correlation between attention correctness and caption quality.

Although impressive visualization results of the attention maps for image captioning are shown in (Xu et al. 2015), the authors do not provide *quantitative evaluations* of the attention maps generated by their models. Since deep network attention can be viewed as a form of alignment from language space to image space, we argue that these attention maps in fact carry important information in understanding (and potentially improving) deep networks. Therefore in this paper, we study the following two questions:

- How often and to what extent are the attention maps consistent with human perception/annotation?
- Will more human-like attention maps result in better captioning performance?

Towards these goals, we propose a novel quantitative metric to evaluate the “correctness” of attention maps. We define “correctness” as the consistency between the attention maps generated by the model and the corresponding region that the words/phrases describe in the image. More specifically, we use the alignment annotations between image regions and noun phrase caption entities provided in the Flickr30k Entities dataset (Plummer et al. 2015) as our ground truth

maps. Using this metric, we show that the attention model of (Xu et al. 2015) performs better than the uniform attention baseline, but still has room for improvement in terms of attention consistency with human annotations.

Based on this observation, we propose a model with explicit supervision of the attention maps. The model can be used not only when detailed ground truth attention maps are given (e.g. the Flickr30k Entities dataset (Plummer et al. 2015)) but also when only the semantic labelings of image regions (which is a much cheaper type of annotations) are available (e.g. MS COCO dataset (Lin et al. 2014)). Our experiments show that in both scenarios, our models perform consistently and significantly better than the implicit attention counterpart in terms of both attention maps accuracy and the quality of the final generated captions. To the best of our knowledge, this is the first work that quantitatively measures the quality of visual attention in deep models and shows significant improvement by adding supervision to the attention module.

## Related Work

**Image Captioning Models** There has been growing interest in the field of image captioning, with lots of work demonstrating impressive results (Kiros, Salakhutdinov, and Zemel 2014; Xu et al. 2015; Mao et al. 2015; Vinyals et al. 2015; Donahue et al. 2015; Fang et al. 2015; Karpathy and Fei-Fei 2015; Chen and Zitnick 2014). However, it is uncertain to what extent the captioning models truly understand and recognize the objects in the image while generating the captions. (Xu et al. 2015) proposed an attention model and qualitatively showed that the model can attend to specific regions of the image by visualizing the attention maps of a few images. Our work takes a step further by quantitatively measuring the quality of the attention maps. The role of the attention maps also relates to referring expressions (Mao et al. 2016; Hu et al. 2015), where the goal is predicting the part of the image that is relevant to the expression.

**Deep Attention Models** In machine translation, (Bahdanau, Cho, and Bengio 2014) introduced an extra softmax layer in the RNN/LSTM structure that generates weights of the individual words of the sentence to be translated. The quality of the attention/alignment was qualitatively visualized in (Bahdanau, Cho, and Bengio 2014) and quantitatively evaluated in (Luong, Pham, and Manning 2015) using the alignment error rate. In image captioning, (Xu et al. 2015) used convolutional image features with spatial information as input, allowing attention on 2D space. (You et al. 2016) targeted attention on a set of concepts extracted from the image to generate image captions. In visual question answering, (Chen et al. 2015; Xu and Saenko 2015; Shih, Singh, and Hoiem 2016; Zhu et al. 2015) proposed several models which attend to image regions or questions when generating an answer. But none of these models quantitatively evaluates the quality of the attention maps or imposes supervision on the attention. Concurrently, (Das et al. 2016) analyzed the consistency between human and deep network attention in visual question answering. Our goal differs in that we are interested in how attention changes with the progression of the description.

**Image Description Datasets** For image captioning, Flickr8k (Hodosh, Young, and Hockenmaier 2013), Flickr30k (Young et al. 2014), and MS COCO (Lin et al. 2014) are the most commonly used benchmark datasets. (Plummer et al. 2015) developed the original caption annotations in Flickr30k by providing the region to phrase correspondences. Specifically, annotators were first asked to identify the noun phrases in the captions, and then mark the corresponding regions with bounding boxes. In this work we use this dataset as ground truth to evaluate the quality of the generated attention maps, as well as to train our strongly supervised attention model. Our model can also utilize the instance segmentation annotations in MS COCO to train our weakly supervised version.

## Deep Attention Models for Image Captioning

In this section, we first discuss the attention model that learns the attention weights implicitly (Xu et al. 2015), and then introduce our explicit supervised attention model.

### Implicit Attention Model

The implicit attention model (Xu et al. 2015) consists of three parts: the encoder which encodes the visual information (i.e. a visual feature extractor), the decoder which decodes the information into words, and the attention module which performs spatial attention.

The visual feature extractor produces  $L$  vectors that correspond to different spatial locations of the image:  $a = \{\mathbf{a}_1, \dots, \mathbf{a}_L\}$ ,  $\mathbf{a}_i \in \mathbb{R}^D$ . Given the visual features, the goal of the decoder is to generate a caption  $y$  of length  $C$ :  $y = \{y_1, \dots, y_C\}$ . We use  $\mathbf{y}_t \in \mathbb{R}^K$  to represent the one-hot encoding of  $y_t$ , where  $K$  is the dictionary size.

In (Xu et al. 2015), an LSTM network (Hochreiter and Schmidhuber 1997) is used as the decoder:

$$\mathbf{i}_t = \sigma(W_i E \mathbf{y}_{t-1} + U_i \mathbf{h}_{t-1} + Z_i \mathbf{z}_t + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(W_f E \mathbf{y}_{t-1} + U_f \mathbf{h}_{t-1} + Z_f \mathbf{z}_t + \mathbf{b}_f) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh(W_c E \mathbf{y}_{t-1} + U_c \mathbf{h}_{t-1} + Z_c \mathbf{z}_t + \mathbf{b}_c) \quad (3)$$

$$\mathbf{o}_t = \sigma(W_o E \mathbf{y}_{t-1} + U_o \mathbf{h}_{t-1} + Z_o \mathbf{z}_t + \mathbf{b}_o) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t) \quad (5)$$

where  $\mathbf{i}_t, \mathbf{f}_t, \mathbf{c}_t, \mathbf{o}_t, \mathbf{h}_t$  are input gate, forget gate, memory, output gate, and hidden state of the LSTM respectively.  $W, U, Z, \mathbf{b}$  are weight matrices and biases.  $E \in \mathbb{R}^{m \times K}$  is an embedding matrix, and  $\sigma$  is the sigmoid function. The context vector  $\mathbf{z}_t = \sum_{i=1}^L \alpha_{ti} \mathbf{a}_i$  is a dynamic vector that represents the relevant part of image feature at time step  $t$ , where  $\alpha_{ti}$  is a scalar weighting of visual vector  $\mathbf{a}_i$  at time step  $t$ , defined as follows:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^L \exp(e_{tk})} \quad e_{ti} = f_{attn}(\mathbf{a}_i, \mathbf{h}_{t-1}) \quad (6)$$

$f_{attn}(\mathbf{a}_i, \mathbf{h}_{t-1})$  is a function that determines the amount of attention allocated to image feature  $\mathbf{a}_i$ , conditioned on the LSTM hidden state  $\mathbf{h}_{t-1}$ . In (Xu et al. 2015), this function is implemented as a multilayer perceptron. Note that by construction  $\sum_{i=1}^L \alpha_{ti} = 1$ .

The output word probability is determined by the image  $\mathbf{z}_t$ , the previous word  $y_{t-1}$ , and the hidden state  $\mathbf{h}_t$ :

$$p(y_t|a, y_{t-1}) \propto \exp(G_o(E\mathbf{y}_{t-1} + G_h\mathbf{h}_t + G_z\mathbf{z}_t)) \quad (7)$$

where  $G$  are learned parameters. The loss function, ignoring the regularization terms, is the negative log probability of the ground truth words  $w = \{w_1, \dots, w_C\}$ :

$$L_{t,cap} = -\log p(w_t|a, y_{t-1}) \quad (8)$$

### Supervised Attention Model

In this work we are interested in the attention map generated by the model  $\alpha_t = \{\alpha_{ti}\}_{i=1,\dots,L}$ . One limitation of the model in (Xu et al. 2015) is that even if we have some prior knowledge about the attention map, it will not be able to take advantage of this information to learn a better attention function  $f_{attn}(\mathbf{a}_i, \mathbf{h}_{t-1})$ . We tackle this problem by introducing explicit supervision.

Concretely, we first consider the case when the ground truth attention map  $\beta_t = \{\beta_{ti}\}_{i=1,\dots,L}$  is provided for the ground truth word  $w_t$ , with  $\sum_{i=1}^L \beta_{ti} = 1$ . Since  $\sum_{i=1}^L \beta_{ti} = \sum_{i=1}^L \alpha_{ti} = 1$ , they can be considered as two probability distributions of attention and it is natural to use the cross entropy loss. For the words that do not have an alignment with an image region (e.g. “a”, “is”), we simply set  $L_{t,attn}$  to be 0:

$$L_{t,attn} = \begin{cases} -\sum_{i=1}^L \beta_{ti} \log \alpha_{ti} & \text{if } \beta_t \text{ exists for } w_t \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The total loss is the weighted sum of the two loss terms:  $L = \sum_{t=1}^C L_{t,cap} + \lambda \sum_{t=1}^C L_{t,attn}$ .

We then discuss two ways of constructing the ground truth attention map  $\beta_t$ , depending on the types of annotations.

**Strong Supervision with Alignment Annotation** In the simplest case, we have direct annotation that links the ground truth word  $w_t$  to a region  $R_t$  (in the form of bounding boxes or segmentation masks) in the image (e.g. Flickr30k Entities). We encourage the model to “attend to”  $R_t$  by constructing  $\hat{\beta}_t = \{\hat{\beta}_{ti}\}_{i=1,\dots,\hat{L}}$  where:

$$\hat{\beta}_{ti} = \begin{cases} 1 & i \in R_t \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Note that the resolution of the region  $R$  (e.g.  $224 \times 224$ ) and the attention map  $\alpha, \beta$  (e.g.  $14 \times 14$ ) may be different, so  $\hat{L}$  could be different from  $L$ . Therefore we need to resize  $\hat{\beta}_t$  to the same resolution as  $\alpha_t$  and normalize it to get  $\beta_t$ .

**Weak Supervision with Semantic Labeling** Ground truth alignment is expensive to collect and annotate. A much more general and cheaper annotation is to use bounding boxes or segmentation masks with object class labels (e.g. MS COCO). In this case, we are provided with a set of regions  $R_j$  in the image with associated object classes  $c_j$ ,  $j = 1, \dots, M$  where  $M$  is the number of object bounding boxes or segmentation masks in the image. Although not ideal, these annotations contain important information to

guide the attention of the model. For instance, for the caption “a boy is playing with a dog”, the model should attend to the region of a person when generating the word “boy”, and attend to the region of a dog when generating the word “dog”. This suggests that we can approximate image-to-language (region  $\rightarrow$  word) consistency by language-to-language (object class  $\rightarrow$  word) similarity.

Following this intuition, we set the likelihood that a word  $w_t$  and a region  $R_j$  are aligned by the similarity of  $w_t$  and  $c_j$  in the word embedding space:

$$\hat{\beta}_{ti} = \begin{cases} \text{sim}(\tilde{E}(w_t), \tilde{E}(c_j)) & i \in R_j \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where  $\tilde{E}(w_t)$  and  $\tilde{E}(c_j)$  denote the embeddings of the word  $w_t$  and  $c_j$  respectively.  $\tilde{E}$  can be the embedding  $E$  learned by the model or any off-the-shelf word embedding (e.g. pre-trained word2vec). We then resize and normalize  $\hat{\beta}_t$  in the same way as the strong supervision scenario.

### Attention Correctness: Evaluation Metric

At each time step in the implicit attention model, the LSTM not only predicts the next word  $y_t$  but also generates an attention map  $\alpha_t \in \mathbb{R}^L$  across all locations. However, the attention module is merely an intermediate step, while the error is only backpropagated from the word-likelihood loss in Equation 8. This opens the question of whether this implicitly-learned attention module is indeed effective.

Therefore in this section we introduce the concept of *attention correctness*, an evaluation metric that quantitatively analyzes the quality of the attention maps generated by the attention-based model.

#### Definition

For a word  $y_t$  with generated attention map  $\alpha_t$ , let  $R_t$  be the ground truth attention region, then we define the word attention correctness by

$$AC(y_t) = \sum_{i \in R_t} \hat{\alpha}_{ti} \quad (12)$$

which is a score between 0 and 1. Intuitively, this value captures the sum of the attention score that falls within human annotation (see Figure 2 for illustration).  $\hat{\alpha}_t = \{\hat{\alpha}_{ti}\}_{i=1,\dots,\hat{L}}$  is the resized and normalized  $\alpha_t$  in order to ensure size consistency.

In some cases a phrase  $\{y_t, \dots, y_{t+l}\}$  refers to the same entity, therefore the individual words share the same attention region  $R_t$ . We define the phrase attention correctness as the maximum of the individual scores<sup>1</sup>.

$$AC(\{y_t, \dots, y_{t+l}\}) = \max(AC(y_t), \dots, AC(y_{t+l})) \quad (13)$$

<sup>1</sup>In the experiments, we found that changing the definition from maximum to average does not affect our main conclusion.

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

Figure 2: Attention correctness is the sum of the weights within ground truth region (red bounding box), in this illustration  $0.12 + 0.20 + 0.10 + 0.12 = 0.54$ .

The intuition is that the phrase may contain some less interesting words whose attention map is ambiguous, and the attention maps of these words can be ignored by the max operation. For example, when evaluating the phrase “a group of people”, we are more interested in the attention correctness for “people” rather than “of”.

We discuss next how to find ground truth attention regions during testing, in order to apply this evaluation metric.

## Ground Truth Attention Region During Testing

In order to compute attention correctness, we need the correspondence between regions in the image and phrases in the caption. However, in the testing stage, the generated caption is often different from the ground truth captions. This makes evaluation difficult, because we only have corresponding image regions for the phrases in the ground truth caption, but not *any* phrase. To this end, we propose two strategies.

**Ground Truth Caption** One option is to enforce the model to output the ground truth sentence by resetting the input to the ground truth word at each time step. This procedure to some extent allows us to “decorrelate” the attention module from the captioning component, and diagnose if the learned attention module is meaningful. Since the generated caption exactly matches the ground truth, we compute attention correctness for all noun phrases in the test set.

**Generated Caption** Another option is to align the entities in the generated caption to those in the ground truth caption. For each image, we first extract the noun phrases of the generated caption using a POS tagger (e.g. Stanford Parser (Manning et al. 2014)), and see if there exists a word-by-word match in the set of noun phrases in the ground truth captions. For example, if the generated caption is “A dog jumping over a hurdle” and one of the ground truth captions is “A cat jumping over a hurdle”, we match the noun phrase “a hurdle” appearing in both sentences. We then calculate the attention correctness for the matched phrases only.

## Experiments

### Implementation Details

**Implicit/Supervised Attention Models** All implementation details strictly follow (Xu et al. 2015). We resize the image such that the shorter side has 256 pixels, and then center crop the  $224 \times 224$  image, before extracting the conv5\_4 feature of the 19 layer version of VGG net (Simonyan and Zisserman 2014) pretrained on ImageNet (Deng et al. 2009). The model is trained using stochastic gradient descent with the Adam algorithm (Kingma and Ba 2014). Dropout (Srivastava et al. 2014) is used as regularization. We use the hyperparameters provided in the publicly available code<sup>2</sup>. We set the number of LSTM units to 1300 for Flickr30k and 1800 for COCO.

**Ground Truth Attention for Strong Supervision Model** We experiment with our strong supervision model on the Flickr30k dataset (Young et al. 2014). The Flickr30k Entities dataset (Plummer et al. 2015) is used for generating the ground truth attention maps. For each entity (noun phrase) in

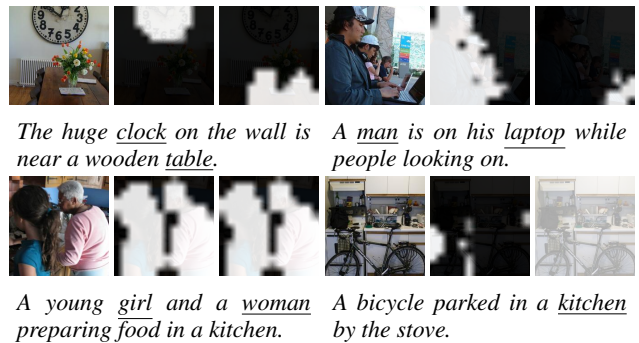


Figure 3: Ground truth attention maps generated for COCO. The first two examples show successful cases. The third example is a failed case where the proposed method aligns both “girl” and “woman” to the “person” category. The fourth example shows the necessity of using the scene category list. If we do not distinguish between object and scene (middle), the algorithm proposes to align the word “kitchen” with objects like “spoon” and “oven”. We propose to use uniform attention (right) in these cases.

the caption, the Flickr30k Entities dataset provides the corresponding bounding box of the entity in the image. Therefore ideally, the model should “attend to” the marked region when predicting the associated words. We evaluate on noun phrases only, because for other types of words (e.g. determiner, preposition) the attention might be ambiguous and meaningless.

### Ground Truth Attention for Weak Supervision Model

The MS COCO dataset (Lin et al. 2014) contains instance segmentation masks of 80 classes in addition to the captions, which makes it suitable for our model with weak supervision. We only construct  $\beta_t$  for the nouns in the captions, which are extracted using the Stanford Parser (Manning et al. 2014). The similarity function in Equation 11 is chosen to be the cosine distance between word vectors (Mikolov et al. 2013) pretrained on GoogleNews<sup>3</sup>, and we set an empirical threshold of 1/3 (i.e. only keep those with cosine distance greater than the threshold).

The  $\beta_t$  generated in this way still contains obvious errors, primarily because word2vec cannot distinguish well between objects and scenes. For example, the similarity between the word “kitchen” and the object class “spoon” is above threshold. But when generating a scene word like “kitchen”, the model should be attending to the whole image instead of focusing on a small object like “spoon”.

To address this problem, we refer to the supplement of (Lin et al. 2014), which provides a scene category list containing key words of scenes used when collecting the dataset. Whenever some word in this scene category list appears in the caption, we set  $\beta_t$  to be uniform, i.e. equal attention across image. This greatly improves the quality of  $\beta_t$  in some cases (see illustration in Figure 3).

**Comparison of Metric Designs** To show the legitimacy of our attention correctness metric, we compute the spearsman

<sup>2</sup><https://github.com/kelvinxu/arctic-captions>

<sup>3</sup><https://code.google.com/archive/p/word2vec/>

Table 1: Attention correctness and baseline on Flickr30k test set. Both the implicit and the (strongly) supervised models outperform the baseline. The supervised model performs better than the implicit model in both settings.

Caption	Model	Baseline	Correctness
Ground Truth	Implicit	0.3214	0.3836
	Supervised	0.3214	<b>0.4329</b>
Generated	Implicit	0.3995	0.5202
	Supervised	0.3968	<b>0.5787</b>

Table 2: Attention correctness and baseline on the Flickr30k test set (generated caption, same matches for implicit and supervised) with respect to bounding box size. The improvement is greatest for small objects.

BBox Size	Model	Baseline	Correctness
Small	Implicit	0.1196	0.2484
	Supervised	0.1196	<b>0.3682</b>
Medium	Implicit	0.3731	0.5371
	Supervised	0.3731	<b>0.6117</b>
Large	Implicit	0.7358	0.8117
	Supervised	0.7358	<b>0.8255</b>

correlation of our design and three other metrics: negative L1 distance, negative L2 distance, and KL divergence between  $\hat{\beta}_t$  and  $\hat{\alpha}_t$ . On the Flickr30k test set with implicit attention and ground truth caption, the spearsman correlations between any two are all above 0.96 (see supplementary material), suggesting that all these measurements are similar. Therefore our metric statistically correlates well with other metrics, while being the most intuitive.

### Evaluation of Attention Correctness

In this subsection, we quantitatively evaluate the attention correctness of both the implicit and the supervised attention model. All experiments are conducted on the 1000 test images of Flickr30k. We compare the result with a uniform baseline, which attends equally across the whole image. Therefore the baseline score is simply the size of the bounding box over the size of the whole image. The results are summarized in Table 1.

**Ground Truth Caption Result** In this setting, both the implicit and supervised models are forced to produce exactly the same captions, resulting in 14566 noun phrase matches. We discard those with no attention region or full image attention (as the match score will be 1 regardless of the attention map). For each of the remaining matches, we resize the original attention map from  $14 \times 14$  to  $224 \times 224$  and perform normalization before we compute the attention correctness for this noun phrase.

Both models are evaluated in Figure 4a. The horizontal axis is the improvement over baseline, therefore a better attention module should result in a distribution further to the right. On average, both models perform better than the baseline. Specifically, the average gain over uniform attention baseline is 6.22% for the implicit attention model (Xu et al.

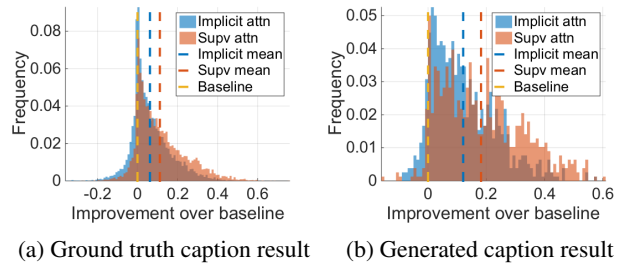


Figure 4: Histograms of attention correctness for the implicit model and the supervised model on the Flickr30k test set. The more to the right the better.

2015), and 11.14% for the supervised version. Visually, the distribution of the supervised model is further to the right. This indicates that although the implicit model has captured some aspects of attention, the model learned with strong supervision has a better attention module.

In Figure 5 we show some examples where the supervised model correctly recovers the spatial location of the underlined entity, while the implicit model attends to the wrong region.

**Generated Caption Result** In this experiment, word-by-word match is able to align 909 noun phrases for the implicit model and 901 for the supervised version. Since this strategy is rather conservative, these alignments are correct and reliable, as verified by a manual check. Similarly, we discard those with no attention region or full image attention, and perform resize and normalization before we compute the correctness score.

The results are shown in Figure 4b. In general the conclusion is the same: the supervised attention model produces attention maps that are more consistent with human judgment. The average improvement over the uniform baseline is 12.07% for the implicit model and 18.19% for the supervised model, which is a 50% relative gain.

In order to diagnose the relationship between object size and attention correctness, we further split the test set equally with small, medium, and large ground truth bounding box, and report the baseline and attention correctness individually. We can see from Table 2 that the improvement of our supervised model over the implicit model is greatest for small objects, and pinpointing small objects is stronger evidence of image understanding than large objects.

In Figure 6 we provide some qualitative results. These examples show that for the same entity, the supervised model produces more human-like attention than the implicit model. More visualization are in the supplementary material.

### Evaluation of Captioning Performance

We have shown that supervised attention models achieve higher attention correctness than implicit attention models. Although this is meaningful in tasks such as region grounding, in many tasks attention only serves as an intermediate step. We may be more interested in whether supervised attention model also has better captioning performance, which is the end goal. The intuition is that a meaningful dynamic

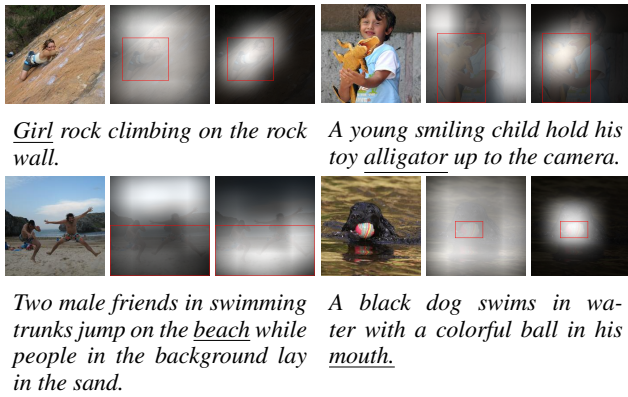


Figure 5: Attention correctness using ground truth captions. From left to right: original image, implicit attention, supervised attention. The red box marks correct attention region (from Flickr30k Entities). In general the attention maps generated by our supervised model have higher quality.

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	<b>30.2</b>	<b>21.0</b>	<b>19.21</b>
COCO	Implicit	34.4	24.3	23.90
	Implicit*	36.4	26.9	24.46
	Weak Sup	<b>37.2</b>	<b>27.6</b>	<b>24.78</b>

weighting of the input vectors will allow later components to decode information more easily. In this subsection we give experimental support.

We report BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005) scores to allow comparison with (Xu et al. 2015). In Table 3 we show both the scores reported in (Xu et al. 2015) and our implementation. Note that our implementation of (Xu et al. 2015) gives slightly improved result over what they reported. We observe that BLEU and METEOR scores consistently increase after we introduce supervised attention for both Flickr30k and COCO. Specifically in terms of BLEU-4, we observe a significant increase of 0.9 and 0.7 percent respectively.

To show the positive correlation between attention correctness and caption quality, we further split the Flickr30k test set (excluding those with zero alignment) equally into three sets with high, middle, and low attention correctness. The BLEU-4 scores are 28.1, 26.1, 25.4, and METEOR are 23.01, 21.94, 21.14 respectively (see Table 4). This indicates that higher attention correctness means better captioning performance.

## Discussion

In this work we make a first attempt to give a quantitative answer to the question: to what extent are attention maps

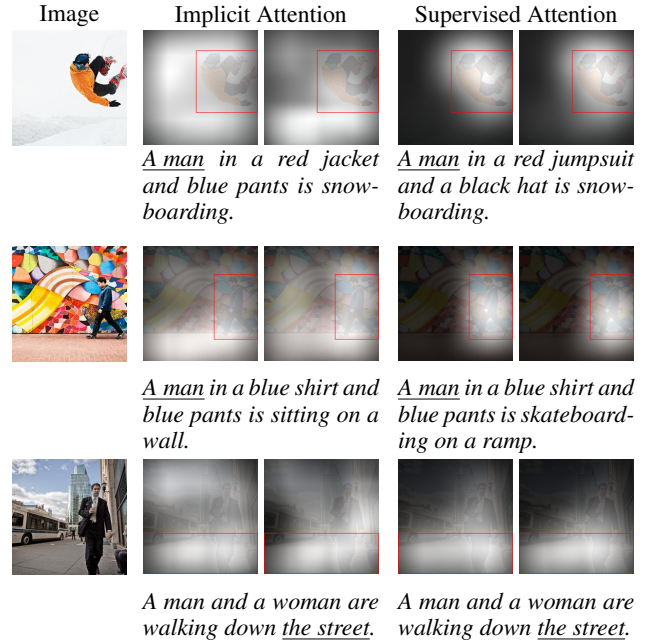


Figure 6: Attention correctness using generated captions. The red box marks correct attention region (from Flickr30k Entities). We show two attention maps for the two words in a phrase. In general the attention maps generated by our supervised model have higher quality.

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

consistent with human perceptions? We first define attention correctness in terms of consistency with human annotation at both the word level and phrase level. In the context of image captioning, we evaluated the state-of-the-art models with implicitly trained attention modules. The quantitative results suggest that although the implicit models outperform the uniform attention baseline, they still have room for improvement.

We then show that by introducing supervision of attention map, we can improve both the image captioning performance and attention map quality. In fact, we observe a positive correlation between attention correctness and captioning quality. Even when the ground truth attention is unavailable, we are still able to utilize the segmentation masks with object category as a weak supervision to the attention maps, and significantly boost captioning performance.

We believe closing the gap between machine attention and human perception is necessary, and expect to see similar efforts in related fields.

## Acknowledgments

We gratefully acknowledge support from NSF STC award CCF-1231216 and the Army Research Office ARO 62250-CS. FS is partially supported by NSF IIS-1065243, 1451412, 1513966, and CCF-1139148. We also thank Tianze Shi for helpful suggestions in the early stage of this work.

## References

- Ba, J.; Mnih, V.; and Kavukcuoglu, K. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.
- Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, 65–72.
- Chen, X., and Zitnick, C. L. 2014. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*.
- Chen, K.; Wang, J.; Chen, L.-C.; Gao, H.; Xu, W.; and Nevatia, R. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.
- Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Das, A.; Agrawal, H.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? *arXiv preprint arXiv:1606.03556*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.
- Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2625–2634.
- Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *CVPR*, 1473–1482.
- Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 853–899.
- Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2015. Natural language object retrieval. *arXiv preprint arXiv:1511.04164*.
- Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.
- Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2014. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*.
- Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.
- Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.
- Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*.
- Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, 2204–2212.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.
- Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2641–2649.
- Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Learning to localize little landmarks. In *Computer Vision and Pattern Recognition*.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.
- Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.
- Xu, H., and Saenko, K. 2015. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv preprint arXiv:1511.05234*.
- Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.
- You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. *arXiv preprint arXiv:1603.03925*.
- Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.
- Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2015. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*.