University of California
Los Angeles

Attention Correctness in Neural Image Captioning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Chenxi Liu

2016

Abstract of the Thesis

Attention Correctness in Neural Image Captioning

by

Chenxi Liu

Master of Science in Statistics

University of California, Los Angeles, 2016

Professor Alan Loddon Yuille, Chair

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 how well the attention maps align with human judgment, 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.

The thesis of Chenxi Liu is approved.

Yingnian Wu

Hongjing Lu

Alan Loddon Yuille, Committee Chair

University of California, Los Angeles

2016

To my parents

for their exceptional vision and unconditional support

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Prof. Alan Yuille for his support for this project. His precious comments and advice helped a lot in writing this thesis.

I would also like to thank Prof. Fei Sha in the computer science department, as this project was first shaped in his seminar class.

My sincere thanks to my lab-mate Junhua Mao. His guidance and expertise was crucial for this project.

I also want to thank the mentors in my earlier research career. They are Prof. Raquel Urtasun, Prof. Sanja Fidler, Prof. Greg Shakhnarovich, Prof. Michael Maire, and Prof. Jie Zhou.

CHAPTER 1

Introduction

Attention based deep models have been proved effective at handling problems such as machine translation [2], object detection [3, 4], visual question answering [5, 6], and image captioning [1]. In these tasks, the input consists of a number of vectors with the same dimension. Deep models with attention address these tasks by learning a dynamic combination of these vectors.

In this work we focus on attention models for image captioning. The state-of-the-art image captioning models [7, 8, 9, 10, 11] adopt Convolutional Neural Networks (CNNs) to extract image features and Recurrent Neural Networks (RNNs) to decode these features into a sentence description. These models can be interpreted within a sequence-to-sequence [12] or encoder-decoder [13] framework, so it is natural to apply attention mechanisms in these models [1].

Although impressive visualization results of the attention maps for image captioning are shown in [1], the authors do not provide quantitative evaluations of the attention maps generated by their models. This is a common issue for attention models, because defining and evaluating the attention maps can be hard for attention models for most tasks. However, an accurate quantitative metric is important and can provide further insight in understanding and improving attention models.

In this work, 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 those annotated by humans (i.e. the ground truth maps). We use the alignment annotation between image regions and noun phrase caption entities provided in the recently released Flickr30k Entities dataset [14] as our ground truth maps.

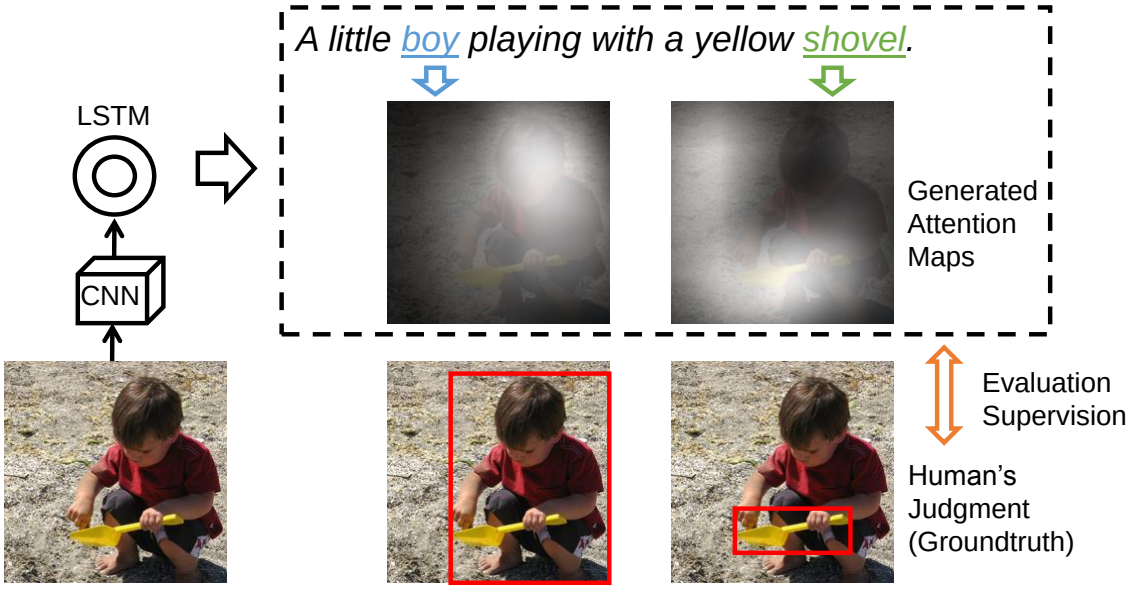


Figure 1.1: The illustration of the motivation of our work. We propose a quantitative evaluation metric for the quality of attention maps and find there is room for improvement of the implicitly learned attention maps of [1]. We then propose a novel model that is able to utilize either the strong labels of region-to-phrase correspondence, or weak labels of segmentation with object category, to improve both the quality of the attention maps and the generated captions.

Using this metric, we show that the attention model of [1] performs better than the uniform attention baseline, but still has big room for improvement in terms of attention consistency with humans.

Based on this observation, we propose a simple but effective model with explicit supervision of the attention maps. The model can be used not only in situations where detailed ground truth attention maps are given (e.g. the Flickr30k Entities dataset [14]) but also when only the object categories of image regions (which is a much cheaper type of annotations compared to [14]) are available (e.g. MS COCO dataset [15]). Our experiments show that 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 in

both scenarios. To the best of our knowledge, this is the first work that quantitatively measures the quality of attention in deep models and shows significant improvement by adding supervision to the attention module.

CHAPTER 2

Related Work

Image Captioning Models There has been growing interest in the field of image captioning, with lots of work demonstrating impressive results [7, 1, 8, 11, 10, 16, 9, 17]. However, it is not clear whether the captioning models truly understand and recognize the objects in the image while generating the captions. [1] 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. We build on their work and take a step further by quantitatively measuring the quality of the attention maps, which offers insight into understanding and improving current image captioning models.

Deep Attention Models Attention mechanism is an important property of human visual systems [18, 19]. Since deep neural networks are inspired by the structure of neurons in human brains, exploring the use of attention in these artificial models seems natural and promising.

In machine translation, [2] introduced an extra softmax layer in the RNN/LSTM structure that generates weights of the individual words of the sentence to be translated. This allowed the individual representations of the words to be preserved. In image captioning, [1] replaced the individual words in machine translation model by convolutional image features, allowing the model to attend to different areas of the image when generating words one by one. This model is discussed in details in section 3.1. [20] proposed to target attention on a set of concepts extracted from the image to generate image captions. In visual question answering, [6, 5, 21] 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.

Image Description Datasets For image captioning, Flickr8k [22], Flickr30k [23], and MS COCO [15] are the most commonly used benchmark datasets. Each image in these datasets has 5 accompanying captions. The original annotations of these datasets do not have alignment between the image regions and the entities (e.g. noun phrases) in the captions. Plummer et al. [14] developed the original caption annotations in Flickr30k by providing the region to phrase correspondences. Specifically, they align the noun phrases in the captions to image regions using human annotators. In this work, we use this dataset for constructing ground truth attention maps to evaluate the quality of the generated attention maps, as well as to train our strongly supervised attention model.

CHAPTER 3

Deep Attention Models for Image Captioning

In this section, we first introduce Xu et al. [1]'s attention model that learns the attention weights implicitly and then introduce our explicit supervised attention model.

3.1 Implicit Attention Models

Xu et al. [1] was the first attempt to introduce attention models to image captioning. The model 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: $a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\},\ \mathbf{a}_i \in \mathbb{R}^D$. Given the visual features, the goal of the decoder is to generate a caption y of length C: $y = \{y_1, \ldots, y_C\}$. 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 [1], an LSTM network [24] is used as the decoder:

$$\mathbf{i}_t = \sigma(W_i E\mathbf{y}_{t-1} + U_i\mathbf{h}_{t-1} + Z_i\mathbf{z}_t + \mathbf{b}_i) \tag{3.1}$$

$$\mathbf{f}_t = \sigma(W_f E\mathbf{y}_{t-1} + U_f\mathbf{h}_{t-1} + Z_f\mathbf{z}_t + \mathbf{b}_f) \tag{3.2}$$

$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t\tanh(W_c E\mathbf{y}_{t-1} + U_c\mathbf{h}_{t-1} + Z_c\mathbf{z}_t + \mathbf{b}_c) \tag{3.3}$$

$$\mathbf{o}_t = \sigma(W_o E\mathbf{y}_{t-1} + U_o\mathbf{h}_{t-1} + Z_o\mathbf{z}_t + \mathbf{b}_o) \tag{3.4}$$

$$\mathbf{h}_t = \mathbf{o}_t\tanh(\mathbf{c}_t) \tag{3.5}$$

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 σ is the sigmoid function. The context vector $\mathbf{z}_t \in \mathbb{R}^D$ is a dynamic vector that represents the relevant part of image feature at time step t. In Xu et al. [1]'s deterministic "soft" attention model,

$$\mathbf{z}_t = \sum_{i=1}^{L} \alpha_{ti} \mathbf{a}_i \tag{3.6}$$

where α_{ti} is a scalar weighting of visual vector $\mathbf{a}_i$ at time step t, defined as follows:

$$e_{ti} = f_{attn}(\mathbf{a_i}, \mathbf{h}_{t-1}) \qquad \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{3.7}$$

where $f_{attn}(\mathbf{a_i}, \mathbf{h}_{t-1})$ 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 [1], 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 $\mathbf{h}_t$:

$$p(y_t | a, y_{t-1}) \propto \exp(G_o(E\mathbf{y}_{t-1} + G_h \mathbf{h}_t + G_z \mathbf{z}_t)) \tag{3.8}$$

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

$$L_{t,cap} = -\log p(w_t | a, y_{t-1}) \tag{3.9}$$

3.2 Supervised Attention Models

Deep network attention can be viewed as a form of alignment from language space to image space. However, it is not clear how good this alignment is. Moreover, even if the ground truth of this alignment is provided in a dataset, the model in [1] will not be able to take advantage of this information to learn better attention function $f_{attn}(\mathbf{a_i}, \mathbf{h}_{t-1})$. In this work, we seek to enforce attention correctness by introducing explicit supervision.

Concretely, we first consider the case when the ground truth attention map $\boldsymbol{\beta}_t = \{\beta_{ti}\}_{i=1,\ldots,L}$ is provided for 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 introduce the cross entropy loss on the attention map. For the words that do not have an alignment with an image region (e.g. "a", "is"), we simply set $L_{t,attn}$ as 0:

$$L_{t,attn} = \begin{cases} -\sum_{i=1}^{L} \beta_{ti} \log \alpha_{ti} & \text{if } w_t \text{ aligns with a ground truth region} \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

The total loss is the weighted sum of the two loss terms:

$$L = \sum_{t=1}^{C} L_{t,cap} + \lambda \sum_{t=1}^{C} L_{t,attn} \tag{3.11}$$

We then discuss two ways of constructing the ground truth attention map $\boldsymbol{\beta}_t$, depending on the types of annotations available.

3.2.1 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 segment masks) in the image. We encourage the model to "attend to" R_t by first constructing $\hat{\boldsymbol{\beta}}_t = \{\hat{\beta}_{t\hat{i}}\}_{\hat{i}=1,..,\hat{L}}$ by:

$$\hat{\beta}_{t\hat{i}} = \begin{cases} 1 & \hat{i} \in R_t \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

Note that the resolution of the region R and the attention map $\boldsymbol{\alpha}, \boldsymbol{\beta}$ might be different, so $\hat{L}$ could be different from L. Then we resize $\hat{\boldsymbol{\beta}}_t$ to the same resolution as $\boldsymbol{\alpha}_t$ and normalize it to get $\boldsymbol{\beta}_t$.

3.2.2 Weak Supervision with Object Category Annotation

Ground truth alignment is expensive to collect and annotate. A much more general and cheaper annotation is to use bounding boxes or segments with object category. In this case, we are provided with a set of regions R_j in the image with associated object classes c_j,

$j = 1, \ldots, M$ where M is the number of object bounding boxes or segments in the image. Although not ideal, these annotations contain important information to guide the attention of the model. For instance, for a caption of "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". We can use this information to automatically find semantically related words in the sentences and regions with the object category labels in the image.

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:

$$\hat{\beta}_{t\hat{i}} = \begin{cases} \text{sim}(\tilde{E}(w_t), \tilde{E}(c_j)) & \hat{i} \in R_j \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where $\tilde{E}(w_t)$ and $\tilde{E}(c_j)$ denotes 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. We then resize and normalize $\hat{\boldsymbol{\beta}}_t$ in the same way as the strong supervision scenario.

CHAPTER 4

Attention Correctness: Evaluation Metric

At each time step in the base model, the LSTM not only predicts the next word y_t but also generates an attention map $\boldsymbol{\alpha_t} \in \mathbb{R}^L$ across all locations. However, the attention module serves only as an intermediate step, while the backpropagated error only comes from the captioning loss. This opens the question 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.

4.1 Definition

For a word y_t with generated attention map $\boldsymbol{\alpha}_t$, let R_t be the ground truth attention region, then we define the word attention correctness by

$$AC(y_t) = \sum_{\hat{i} \in R_t} \hat{\alpha}_{t\hat{i}} \tag{4.1}$$

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 4.1 for illustration). $\hat{\boldsymbol{\alpha}}_t = \{\hat{\alpha}_{t\hat{i}}\}_{\hat{i}=1,\ldots,\hat{L}}$ is the resized and normalized $\boldsymbol{\alpha}_t$ in order to ensure size consistency.

In some cases a phrase $\{y_t, \ldots, y_{t+l}\}$ refers to the same entity, therefore the individual words share the same attention region R_t. We define the phrase attention correctness as the

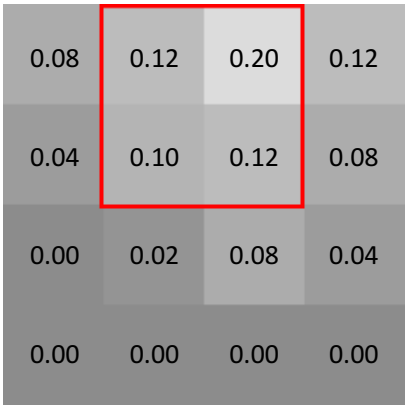


Figure 4.1: Attention correctness is the sum of the weights within ground truth region (red bounding box), in this illustration 0.12 + 0.20 + 0.10 + 0.12 = 0.54.

maximum of the individual scores[1].

$$AC(\{y_t, \ldots, y_{t+l}\}) = \max(AC(y_t), \ldots, AC(y_{t+l})) \tag{4.2}$$

The intuition is that the phrase may contain some words whose attention map is ambiguous. For example, when evaluating the phrase "a group of people", we are more interested in the attention correctness for "people" rather than "of".

4.2 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 ground truth captions. This makes evaluation difficult, because we only have R_t for the phrases in the ground truth caption, but not *any* possible phrase. To this end, we propose two strategies.

Ground Truth Caption One option is to enforce the trained model to output the ground truth sentence by resetting y_t at each time step. This procedure to some extent allows us to "decorrelate" the attention module from the captioning component, and diagnose if the

[1]We found that changing the definition from maximum to average does not affect our main conclusion.

learned attention module is meaningful. Since the generated caption exactly matches the ground truth, we can compute attention correctness for each noun phrase 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 POS tagger (e.g. Stanford Parser [25]), 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.

CHAPTER 5

Experiments

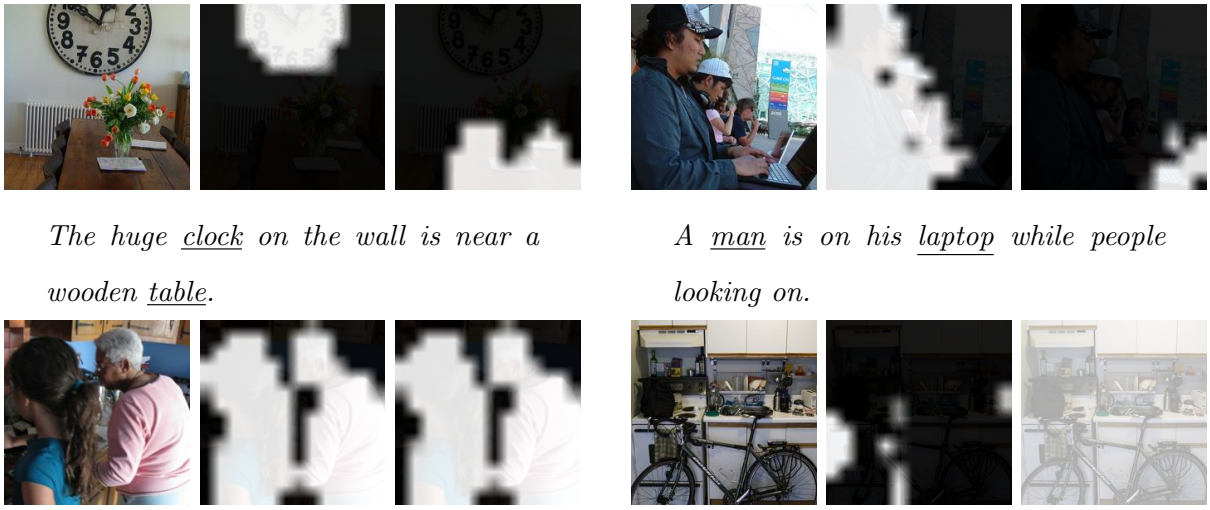
5.1 Implementation Details

Implicit/Supervised Attention Models All implementation details strictly follow [1]. We resize the image such that the shorter side has 256 pixels, and then center crop the 224×224 image. We then extract the conv5_4 feature of the 19 layer version of VGG net [26] pre-trained on ImageNet [27]. The model was trained using stochastic gradient descent with the Adam algorithm [28]. Dropout [29] was used as regularization. We use the hyperparameters provided in the publicly available code[1]. Specifically, 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 model in section 3.2.1 on Flickr30k dataset [23]. We use the Flickr30k Entities dataset [14] for generating ground truth attention maps. For each entity (noun phrase) in 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 [15] contains instance segmentation masks of 80 classes in addition to the captions, which makes it suitable for our model in section 3.2.2. We only construct $\boldsymbol{\beta}_t$ for the nouns in the captions, which we extract using the Stanford Parser [25]. The similarity function is chosen

[1] https://github.com/kelvinxu/arctic-captions



The huge clock on the wall is near a wooden table.

A man is on his laptop while people looking on.

A young girl and a woman preparing food in a kitchen.

A bicycle parked in a kitchen by the stove.

Figure 5.1: Ground truth attention maps generated for COCO. The first two examples (top row) show successful cases. The third example (bottom left) is a failed case where the proposed method aligns both "girl" and "woman" to the "person" category. The fourth example (bottom right) 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". Uniform attention (right) makes more sense.

to be the cosine distance between word vectors [30] pretrained on GoogleNews[2], and we set an empirical threshold of 1/3.

The $\boldsymbol{\beta}_t$ 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 [15], which provides a scene

[2]https://code.google.com/archive/p/word2vec/

Table 5.1: Attention correctness and baseline. Both the implicit and the (strongly) supervised models outperform the baseline. The supervised model performs better than the implicit model in both settings.

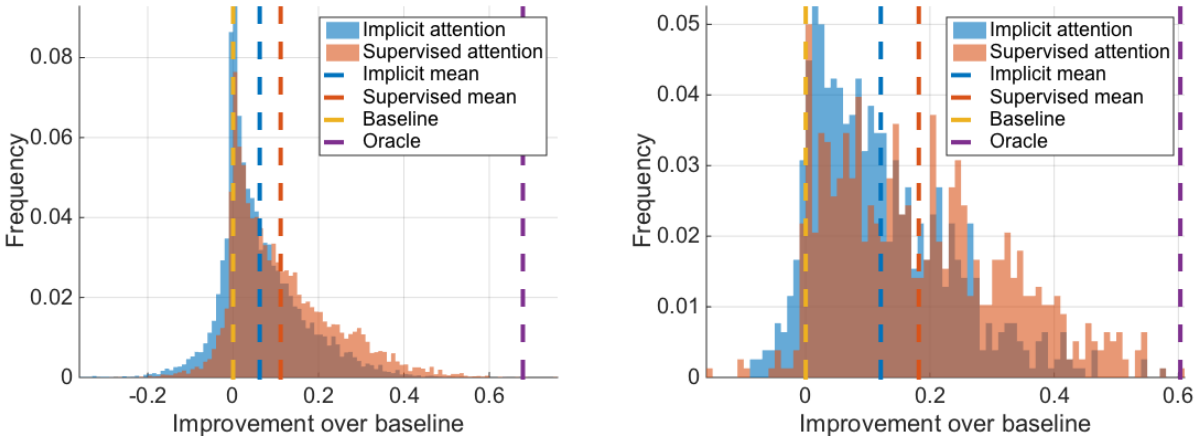
Caption	Model	Baseline	Correctness	# NP
Ground Truth	Implicit	0.3214	0.3836	14566
	Supervised	0.3214	**0.4329**	14566
Generated	Implicit	0.3995	0.5202	883
	Supervised	0.3968	**0.5787**	901

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 $\boldsymbol{\beta}_t$ to be uniform, i.e. equal attention across image. This greatly improves the quality of $\boldsymbol{\beta}_t$ (see illustration in Figure 5.1).

5.2 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 uniform baseline, which attends equally across the whole image. Therefore the baseline score is simply the percentage of the bounding box size over the size of the whole image. The results are summarized in Table 5.1.

Ground Truth Caption Result In this setting, both 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 14×14 to 224×224 and perform normalization before we compute the



(a) Ground truth caption result

(b) Generated caption result

Figure 5.2: Histograms of attention correctness for implicit model and supervised model. The more to the right the better.

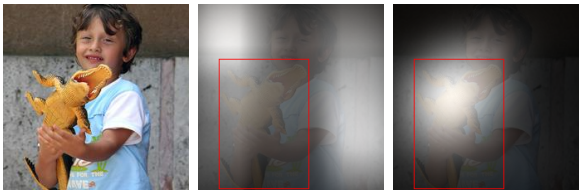
attention correctness for this noun phrase.

Both models are evaluated in Figure 5.2a. 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 [1], and 11.14% for the supervised version. Visually, the distribution of the supervised model is further to the right towards the oracle (where attention correctness is 1 for every match). This indicates that although the implicit model has captured some aspects of attention, the model learned with strong supervision has a better attention module.

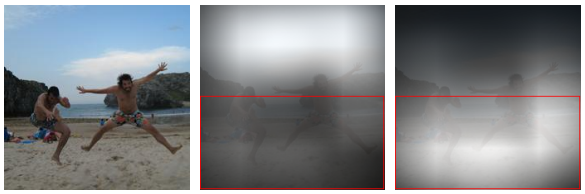
In Figure 5.3 we show some examples where the supervised model successfully recovered the spatial location of the underlined entity, while the implicit model attends to the wrong region.

Generated Caption Result In this experiment, our algorithm is able to align 883 noun phrases for the implicit model and 901 for the supervised version. Since the word-by-word match strategy is rather conservative, these alignments are correct and reliable, as verified by

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.

Figure 5.3: 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.

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 5.2b. In general the conclusion is the same: the supervised attention model produces attention maps that are more consistent with human judgment. In terms of numbers, 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 Figure 5.4 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.

Table 5.2: 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 [1]	28.8	19.1	18.49
	Implicit [1]*	29.2	20.1	19.10
	Strong Supervised	**30.2**	**21.0**	**19.21**
COCO	Implicit [1]	34.4	24.3	23.90
	Implicit [1]*	36.4	26.9	24.46
	Weak Supervised	**37.2**	**27.6**	**24.78**

5.3 Evaluation of Captioning Performance

In the previous subsection we showed 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. The intuition is that a meaningful dynamic weighting of the input vectors will allow later components to decode information more easily. In this subsection we give experimental support by showing that the supervised attention model also provides better captioning performance.

We report BLEU [31] and METEOR [32] scores to allow comparison with [1]. In Table 5.2 we show both the scores reported in [1] and our implementation. Note that our implementation of [1] 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.

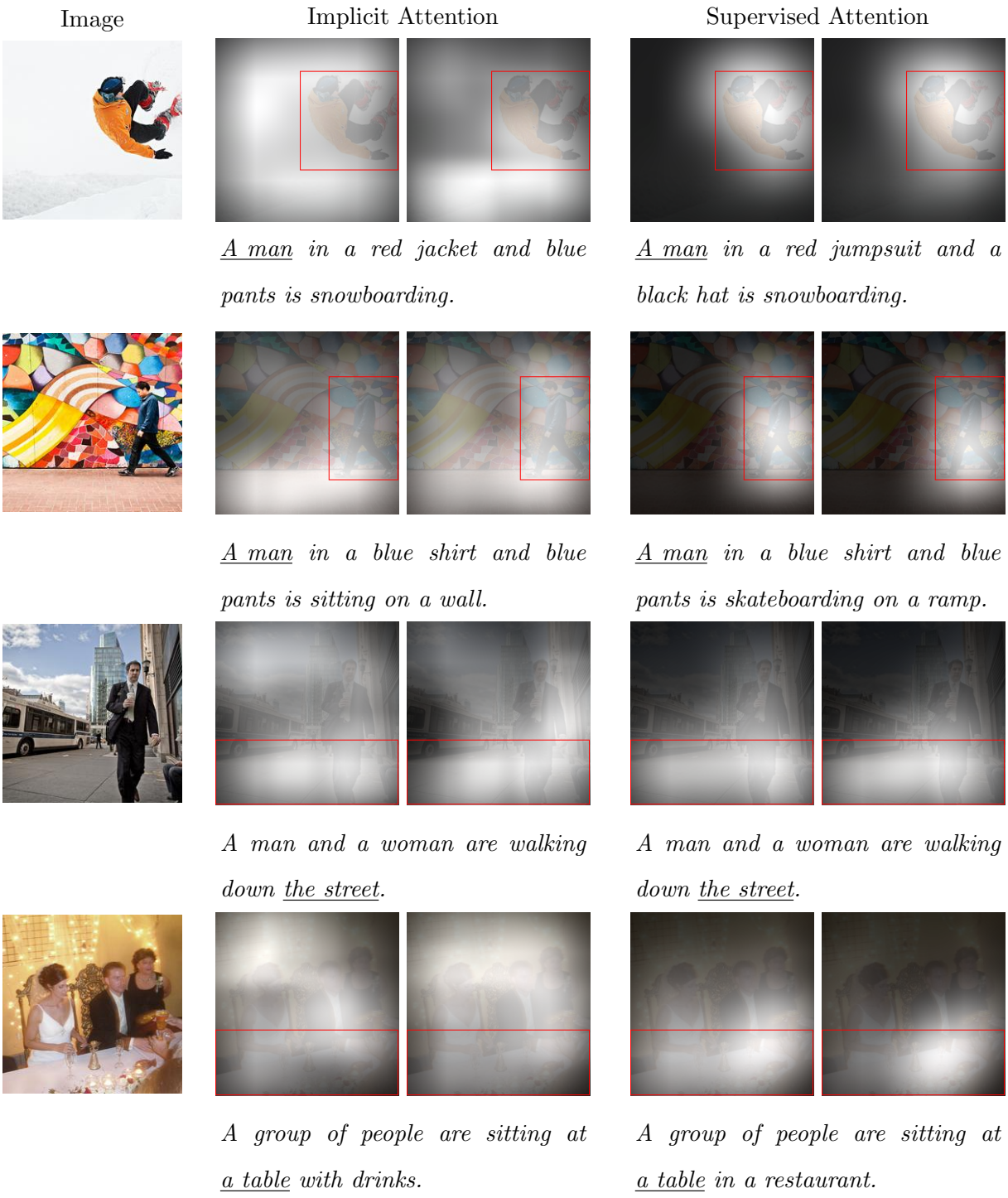


Figure 5.4: Attention correctness using generated captions. The red box marks correct attention region (from Flickr30k Entities). We show two attention maps for the two words in a phrase. In general the attention maps generated by our supervised model have higher quality.

CHAPTER 6

Discussion

In this work we make a first attempt to give a quantitative answer to the question: to what extent are attention maps consistent with human perceptions? We first define attention correctness 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 evaluation results suggest that although the implicit models outperform the baseline, they still have big 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. Even when attention ground truth 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 other tasks, such as visual question answering.

References

[1] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, pp. 2204–2212, 2014.

[4] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[5] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," *arXiv preprint arXiv:1511.05234*, 2015.

[6] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *ICLR*, 2015.

[9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[14] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," *arXiv preprint arXiv:1405.0312*, 2014.

[16] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482, 2015.

[17] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[18] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[19] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[20] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," *arXiv preprint arXiv:1603.03925*, 2016.

[21] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," *arXiv preprint arXiv:1511.03416*, 2015.

[22] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, pp. 853–899, 2013.

[23] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit.," in *ACL (System Demonstrations)*, pp. 55–60, 2014.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[32] S. Banerjee and A. Lavie, "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*, vol. 29, pp. 65–72, 2005.