Boosted Attention: Leveraging Human Attention for Image Captioning

Shi Chen[0000−0002−3749−4767] and Qi Zhao[0000−0003−3054−8934]

Department of Computer Science and Engineering,
University of Minnesota
{chen4595,qzhao}@umn.edu

Abstract. Visual attention has shown usefulness in image captioning, with the goal of enabling a caption model to selectively focus on regions of interest. Existing models typically rely on top-down language information and learn attention implicitly by optimizing the captioning objectives. While somewhat effective, the learned top-down attention can fail to focus on correct regions of interest without direct supervision of attention. Inspired by the human visual system which is driven by not only the task-specific top-down signals but also the visual stimuli, we in this work propose to use both types of attention for image captioning. In particular, we highlight the complementary nature of the two types of attention and develop a model (Boosted Attention) to integrate them for image captioning. We validate the proposed approach with state-of-the-art performance across various evaluation metrics.

Keywords: Image Captioning, Visual Attention, Human Attention

1 Introduction

Image captioning aims at generating fluent language descriptions on a given image. Inspired by the human visual system, in the past few years, visual attention has been incorporated in various image captioning models [21, 26, 32, 33]. Attention mechanisms encourage models to selectively focus on specific regions while generating captions instead of scanning through the whole image, avoiding information overflow as well as highlighting visual regions related to the task.

Following the success made in [32], visual attention in most conventional image captioning models is developed in a top-down fashion on a word basis. That is, visual attention is computed for each generated word based on visual information from the image and the partially generated natural language description. While such mechanism (*i.e.*, top-down attention) aims at connecting natural language and visual content, without prior knowledge on the visual content in terms of salient regions (*i.e.*, stimulus-based attention), the computed visual attention can fail to concentrate on objects of interest and attend to irrelevant regions. As shown in Figure 1, a model with only top-down attention focuses on non-salient regions in the background (Figure 1(c)) and does not capture salient objects in the image, *i.e.*, *bulldog* and *teddy bear* according to the human-generated caption.
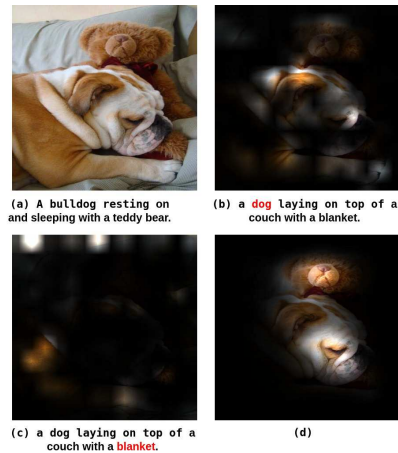
(a) A bulldog resting on and sleeping with a teddy bear.

(b) a dog laying on top of a couch with a blanket.

(c) a dog laying on top of a couch with a blanket.

(d)

Fig. 1. Top-down attention may fail to focus on objects of interest. (a): original image with human-generated caption, (b-c) two top-down attention maps and their corresponding model-generated captions, and (d) stimulus-based attention map for the image. Words related to the top-down attention maps are colored in red.

Human attention is driven by both task-specific top-down signals and task-independent visual stimuli. For visual tasks such as image captioning, humans would naturally deploy their gaze based on both top-down and stimulus-based information during the exploration. As a result, the objects being mentioned in the same image by different people are largely consistent and correlated with the objects highlighted by the stimulus-based attention [35]. Therefore, we propose that the visual stimuli can be a reasonable source for locating salient regions in image captioning, which can also complement top-down attention that relates to specific tasks. In Figure 1(d), we see that stimulus-based attention successfully attends to regions corresponding to objects of interest as mentioned in the human-generated caption.

In this work, we conduct qualitative analyses to understand the role of human stimulus-based attention in image captioning. We then present a Boosted Attention method that leverages stimulus-based attention for image captioning. More specifically, we combine the stimulus-based attention with top-down captioning attention to construct a novel attention mechanism that encourages models to attend to visual features based on task-specific top-down signals from natural language while at the same time focusing on salient regions highlighted by task-independent stimulus. Quantitative results on the Microsoft COCO [19] (MSCOCO) and Flickr30K [24] datasets show that incorporating stimulus-based attention is able to significantly improve the model performance across various evaluation metrics. We also visualize the results to qualitatively illustrate the complementary role of the two types of attention in image captioning. Our method is general and works with various image captioning models.

2   Related Works

Image Captioning. Generating natural language description based on a still image has gained increasing interest in the recent years. To generate captions, [4, 14, 17] first extract a set of attributes related to elements within an image and then generate language description based on the detected attributes. Several works [6, 9, 22] view image captioning as a ranking description problem and tackle the problem by conducting a query to retrieve descriptions lies close to an image on embedding space. With the successes of Deep Neural Networks (DNNs), a number of works [2, 5, 12, 20, 25, 31, 32] have developed neural network based methods to generate image captions. Typically, these methods use Convolutional Neural Networks (CNNs) as visual encoder to extract visual features and generate captions with Recurrent Neural Networks (RNNs) such as Long Short Term Memory (LSTM) [8].

Top-down Attention in Captioning. Top-down visual attention has been widely used on various image captioning models in order to allow models to selectively concentrate on objects of interest. Xu *et al.* [32] combine the memory vector of LSTM with visual features from CNN and feed the fused features to an attention network to compute the weights for features at different spatial locations. Yang *et al.* [33] propose a reviewer module that applies the visual attention mechanism for multiple times during generating the next word. In [21], an adaptive mechanism is proposed that assigns weights not only to visual features but also to a feature vector obtained based on the memory state of LSTM, since it is unnecessary to attend to the visual features for generating specific words such as 'the' and 'a'. Besides applying the attention mechanism on the spatial domain, Chen *et al.* [2] introduce channel-wise attention which is operated on different filters within a convolutional layer. Most of these models generate visual attention in a top-down fashion using the original visual features and top-down language information from the partially generated caption. Without direct supervision or prior knowledge with stimulus-based attention from the images, however, the computed top-down attention can fail to concentrate on the correct objects of interest and attend to irrelevant background.

Stimulus-based Attention in Captioning. To boost the performance of image captioning models, a few works attempt to use human stimulus-based attention. Sugano *et al.* [28] utilize ground truth human gaze to split top-down attention for gazed and non-gazed regions. Cornia *et al.* [3] integrate human attention in a captioning model similar as [28] but replace the human gaze with predicted saliency maps. In [29], Tavakoli *et al.* analyze the effects on stimulus-based attention in captioning by substituting the top-down attention with stimulus-based attention. While these models suggest that human attention can have positive effects on image captioning, they either incorporate only stimulus-based attention or use stimulus-based attention to separate the top-down attention at different locations, resulting in relatively marginal improvement over corresponding baselines.

In this work, we propose a Boosted Attention method that incorporates stimulus-based human attention with existing top-down visual attention. While also using human attention, our method differs from the aforementioned works in the following aspects: 1) Different from [29] which solely relies on stimulus-based attention, we emphasize that it is necessary to integrate stimulus-based attention with top-down attention. 2) Unlike [3, 28] which utilize stimulus-based attention to split top-down attention and extract features from regions either attended by both attention (gazed) or not attended by stimulus-based attention (non-gazed), our method extracts features from regions attended by either attention so both contribute directly with an equal role, naturally enabling the two types of attention to complement each other. Experimental results validate the complementary nature of them, which contributes to the significant boost in captioning performance. 3) Instead of using the spatial map for encoding stimulus-based attention like [3, 28, 29], we integrate the attention via attentional CNN features. Compared to spatial map, our features encode more abundant information and introduce channel-wise attention in addition to spatial attention.

3   The role of Stimulus-based Attention in Image Captioning

Though human-generated captions are relatively free-form, and with considerable inter-subject variance in descriptions, there exists a large degree of agreement in what people describe (*i.e.*, mentioned words in the captions) and what people look (*i.e.*, fixated objects with stimulus-based attention). In this section, we explore the role of stimulus-based attention in image captioning. Specifically, we show the correlations between stimulus-based attention and captioning attention by comparing them on the SALICON [11] dataset under different evaluation metrics. Note that to provide insights on how stimulus-based attention could contribute to the captioning task, the captioning attention we use here is derived from ground truth labels from MSCOCO and seen as ground truth attention for generating the captions.

Similar to [29], we generate captioning attention using visual object category to sentence's noun (VOS) mapping (please refer to the supplementary materials for details). The evaluation metrics used in the comparison include Coefficient Correlation (CC), Spearman's Rank Correlation (Spearman) and Similarity (SIM) [16]. Additionally, we also compute the probability of objects being described given that they are fixated by stimulus-based attention, *i.e.*, $P(d|f)$. To compute this probability, we first set up a small threshold (*i.e.*, 0.1) to filter out the false positive introduced during map re-scaling, then traverse all saliency fixations within the captioning attention map. For each fixation, if the attention value is above the predefined threshold, we consider at that fixation the corresponding object is mentioned in the captions.

Quantitative evaluations show that the objects described in the captions are likely to be fixated by stimulus-based attention with the probability $P(d|f) = 0.465$. According to [29], the probability of an object being mentioned given that
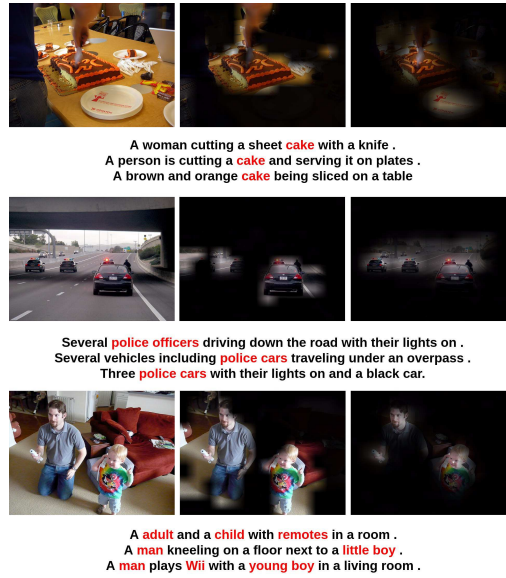
A woman cutting a sheet cake with a knife .
A person is cutting a cake and serving it on plates .
A brown and orange cake being sliced on a table

Several police officers driving down the road with their lights on .
Several vehicles including police cars traveling under an overpass .
Three police cars with their lights on and a black car.

A adult and a child with remotes in a room .
A man kneeling on a floor next to a little boy .
A man plays Wii with a young boy in a living room .

Fig. 2. Visualization for image captioning attention and stimulus-based attention. From left to right: original images, ground truth image captioning attention maps, stimulus-based attention maps. Captions are shown at the bottom of the images with objects of interest mentioned in multiple captions highlighted by red color.

it exists (*i.e.* $P(d|e)$) is around 0.2, thus stimulus-based attention increases the probability of selecting objects of interest by more than 2, providing reasonably good prior knowledge of the objects of interest for image captioning. However, note that since stimulus-based attention commonly attends to only parts of the salient objects instead of covering all or sometimes even majority of the pixels in the objects, the correlations between stimulus-based attention and captioning attention are not high, with $CC = 0.222$, $SIM = 0.353$ and $Spearman = 0.324$. Thus, even though stimulus-based attention is capable of partially capturing objects of interest for image captioning, solely relying on stimulus-based attention may not be sufficient for an image captioning model. Figure 2 shows examples of captioning attention and corresponding stimulus-based attention. We see that stimulus-based attention, while correctly locating objects of interest (*i.e.*, *cake*, *police car*, *man*, *remote* and *boy*), it typically covers part of the salient regions displayed in the captioning attention maps.

4   Boosted Attention Method

As mentioned in section 3, on the one hand, objects of interest in stimulus-based attention are reasonably consistent with objects of interest in image captioning, suggesting that stimulus-based attention can be used to provide prior knowledge

for image captioning. On the other hand, however, with certain level of discrepancy, in both location and coverage, stimulus-based attention alone could lead to loss of visual information and thus decreasing the quality of generated captions.

We therefore propose a Boosted Attention method for image captioning that incorporates stimulus-based attention into the conventional top-down attention framework of a captioning model. The stimulus-based attention is combined with top-down attention to construct a new attention mechanism called Boosted Attention, which encourages the model to focus on certain visual features based on top-down language signals while at the same time attending to the salient regions highlighted by the stimulus-based attention. In all of our experiments, the stimulus-based attention is obtained from a pre-trained saliency prediction network and details about the network can be found in section 5.
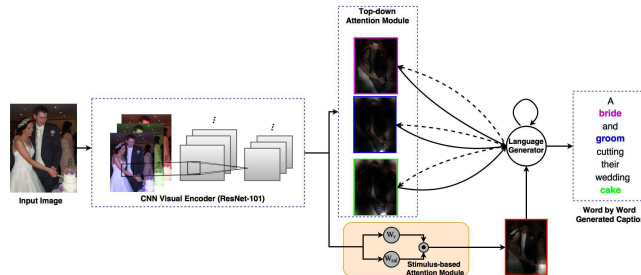

Fig. 3. An illustration of architecture design for proposed Boosted Attention method. Top-down attention maps and their corresponding words are highlighted in purple, blue, green color, while stimulus-based attention map is shown in the red frame.

Figure 3 illustrates the high-level architecture of our method. The model first takes a single raw image as input and encodes it with a CNN Visual Encoder to obtain the visual features. The encoded features are then passed through a Top-down Attention Module and our Stimulus-based Attention Module in parallel, computing the top-down attention and integrating stimulus-based attention. The proposed Stimulus-based Attention Module mainly consists of three parts, a convolutional layer W_{sal} pre-trained on saliency prediction for producing the stimulus-based attention features (attentional CNN features, section 4.1), a convolutional layer W_v that further encodes the visual features, and an integration module $\odot$ that combines stimulus-based attention and visual features. After processing with both the Top-down Attention Module and Stimulus-based Attention Module, visual features integrated with two attention are fed into the Language Generator to sequentially produce the caption.

Note that the proposed method is general and works with different top-down attention and language generation algorithms (*i.e.*, the Top-down Attention Module and the Language Generator in Figure 3). Details about the modules depend on a selected baseline model and the ones used in this work are described in section 5.

4.1   Attentional CNN Features

Instead of using the final output of the saliency prediction network (*i.e.*, the saliency map), we propose to make use of features from intermediate layers of the network which could encode richer information about stimulus-based attention. In this section, we formulate and provide intuitions behind using the attentional CNN features to encode stimulus-based attention.

Considering a fully-convolutional saliency prediction network, we denote it as the equation below (for simplicity we only take the last two layers into considerations):

$$S = softmax(W_m \ \delta(W_{sal}I)) \tag{1}$$

where I is the output of previous layers with ReLU activation, W_{sal} and W_m represent weight parameters in the layer that is used to produce attentional CNN features and output saliency map respectively, δ denotes the ReLU activation and S is the saliency map. The kernel size of both convolutional layers is 1, which enables the model to better capture cross-filter correlations as discussed in [10].

As shown in Equation 1, W_{sal} here constructs both channel-wise attention and spatial attention. Specifically, with the use of ReLU activation ensures non-negativity, to highlight salient regions in the saliency map W_{sal} needs to construct the correlations between filters and stimulus-based attention (*i.e.* suppressing filters that have negative correlations and emphasizing those have positive correlations). These correlations (channel-wise attention) are determined by the signs and magnitude of weights in W_m, *e.g.*, negative weights lead to decrease of activation in S and thus indicate negative correlations, larger weights emphasize more significant contributions. Furthermore, due to the use of spatial softmax activation, W_{sal} also considers the correlations between features and stimulus-based attention on spatial domain, resulting in the spatial attention.

Therefore, we in this work use W_{sal} to produce attentional CNN features for encoding stimulus-based attention, constructing not only spatial attention widely used in various captioning models but also channel-wise (filter-wise) attention [2] that recently found beneficial for image captioning. In Figure 4 we visualize attention maps computed with the CNN features, the results demonstrate that attentional CNN features utilized by our model are capable of highlighting various regions of interest.

4.2   Integrating Stimulus-based Attention

This section discusses our integration method for introducing stimulus-based attention. We first integrate stimulus-based attention with visual features using an asymmetric function as follows:

$$I^{'} = W_v I \circ log(W_{sal}I + \epsilon) \tag{2}$$

where I and $I^{'}$ are the visual features before and after integrating stimulus-based attention, W_v represents weights in an additional convolutional layer that further

encodes the visual features and W_{sal} is the same as in Equation 1, $\circ$ denotes hadamard product, and ϵ is a hyper-parameter. Note that $\odot$ in Figure 3 denotes the whole integration process of Equation 2.

The intuitions behind this integration method are three-fold: First, W_v further encodes visual features, allowing them to adapt to the cross-filter correlations with stimulus-based attention that are stored in W_{sal}. Second, by introducing logarithm, we aim at alleviating the effects of co-adaptation between W_v, W_{sal} and smoothing the contributions of stimulus-based attentional features. Third, with the hyper-parameter ϵ we form a residual mechanism, preserving the original information in visual features and thus preventing potential information loss caused by applying stimulus-based attention. This mechanism is crucial in the proposed integration method, because stimulus-based attention alone may fail to attend to all regions of interest and it is reasonable to allow the model to extract features attended by either one of the attention (stimulus-based or top-down). In our experiments, we define ϵ as a mathematical constant e to preserve the identity of the original visual features. Additional discussion on selecting the hyper-parameter is provided in the supplementary materials.

After obtaining the visual features attended by the stimulus-based attention (*i.e.* $I^{'}$), we apply top-down attention on them via hadamard product, enabling two attention to complement to each other. That is, when stimulus-based attention fails to attend to some regions of interest, top-down attention can attend to those regions via assigning larger weights, and vice versa. We further study the corporation between the two types of attention in section 5.3.

5   Experiments

Dataset and Evaluation. We evaluate our method on two popular datasets: 1) Microsoft COCO [19], where most images contain multiple objects in complex natural scenes with abundant context information. The dataset includes 82783, 40504, 40775 images for training, validation and online evaluation, each has 5 corresponding captions. We use the publicly available Karapthy's split [12] for both training and offline evaluation. 2) Flickr30K [24], where most images depict human performing various activities. It has a total of 31000 images from Flickr, each has 5 corresponding captions. Due to the lack of official split, in order to compare with other works we follow split from [12]. Four automatic metrics are used for evaluation, including BLEU [23], ROUGEL [18], METEOR [15] and CIDEr [30].

Saliency Prediction Network. In order to integrate stimulus-based attention, we construct a saliency prediction network with 2 convolutional layers (note that features from the last convolutional layer of a ResNet-101 are viewed as inputs). The first convolutional layer has 2048 filters while the second layer projects the CNN features to spatial saliency map using a single filter. The kernel size for both layers is set as 1 and the whole saliency network can be represented as Equation 1. We optimize the model on SALICON dataset with cross-entropy

loss and SGD optimizer using learning rate 2.5×10^{-4}. Batch size is set to 1. Weights from the first layer of saliency prediction network is utilized to initialize the stimulus-based attention module in the proposed method (*i.e.* W_{sal} in Equation 2).

Baseline Model. To demonstrate the effectiveness of our method and the advantages of integrating stimulus-based attention, we apply the proposed method on our baseline model constructed based on Soft Attention [32] and several recent tips [2, 26] to enhance performance: we replace the VGG [27] based visual encoder with a more powerful ResNet-101 [7] based one. Instead of fine-tuning the encoder, we directly adopt visual features from the last convolutional layer of the visual encoder as input. When extracting the features, no cropping or re-scaling is applied to the original images, instead, an adaptive spatial average pooling layer is utilized to produce features with a fixed size of $2048 \times 14 \times 14$. Unlike [32] which trains the model solely on cross-entropy loss, we use the optimization method proposed in [26] which contains both supervised learning and reinforcement learning. The LSTM hidden size, word and attention dimensions are set as 512 in our baseline. The other settings remain the same as the original Soft Attention model.

Training. We train our models following the same settings from [26]: we use ADAM [13] optimizer for training all of the models and batch size is set as 50. Models are first trained on cross-entropy loss under supervised learning framework, with initial learning learning rate 5×10^{-4} and Scheduled Sampling [1] feedback probability being 0. During supervised learning, the learning rate is decayed by a factor of 0.8 every 3 epochs and feedback probability increased by 0.05 every 5 epochs. After 25 epochs of supervised learning, we further optimize the models under reinforcement learning framework on the CIDEr metric as [26]. The initial learning rate for reinforcement learning is set as 5×10^{-5} and also decayed by 0.8 every 3 epochs. In supervised learning we fix the weights for stimulus-based attention (W_{sal} in Equation 2) to establish correlations between filters within parallel layers (W_{sal} and W_v in Equation 2), while later on in reinforcement learning we fine-tune stimulus-based attention since the filter correlations have already been established.

5.1 Quantitative Results

In this section, we report quantitative results to demonstrate the effectiveness of the proposed method. We perform inter-model comparisons of the proposed method and 8 state-of-the-art models including Soft Attention [32], ATT [34], SCA-CNN [2], SCN-LSTM [5], RLE [25], AdaATT [21], Att2all [26] and PG-BCMR [20]. We also conduct intra-model comparisons on results with and without the proposed approach (*i.e.,* integrating the stimulus-based attention) and whether using pre-trained stimulus-attention for integration. During evaluation, beam search is utilized for generating the captions and the beam size is set as

3. Table 1 and Table 2 show the result comparison on Flickr30K and MSCOCO (Karpathy's test split [12] and online testing platform).

According to the comparative results, the proposed Boosted Attention method leads to significant performance increase across all evaluation metrics compared to the original baselines without stimulus-based attention. On Flickr30K, using our method results in 2.6%, 5.6%, 2.3% and 12% of relative improvements on BLEU-4, ROUGE-L, METEOR and CIDEr, while on MSCOCO the improvements are 5.7%, 2.0%, 2.7% and 5.6% for corresponding evaluation metrics. Moreover, boosted by the stimulus-based attention, our models are capable of achieving state-of-the-art performance on both datasets.

Model	Flickr30K				MSCOCO			
	B@4	MT	RG	CD	B@4	MT	RG	CD
Soft Attention [32]	0.191	0.185	-	-	0.243	0.239	-	-
ATT [34]	0.230	0.189	-	-	0.304	0.243	-	-
SCA-CNN [2]	0.223	0.195	0.449	0.447	0.311	0.250	0.531	0.952
SCN-LSTM [5]	0.265	0.218	-	-	0.330	0.257	-	1.012
RLE [25]	-	-	-	-	0.304	0.251	0.525	0.937
AdaATT [21]	0.251	0.204	0.467	0.531	0.332	0.266	0.549	1.085
Att2all [26]	-	-	-	-	0.342	0.267	0.557	1.140
ours-Baseline	0.267	0.197	0.471	0.523	0.335	0.258	0.551	1.062
ours-BAM*	0.270	0.204	0.477	0.571	0.350	0.262	0.559	1.111
ours-BAM	0.274	0.208	0.482	0.586	0.354	0.265	0.562	1.122
Improvement (%)	2.6%	5.6%	2.3%	12.0%	5.7%	2.7%	2.0%	5.6%

Table 1. Performance comparison with the state-of-the-art on Flickr30K and MSCOCO (test split in [12]). Baseline is our augmented baseline model without stimulus-based attention, BAM indicates the proposed Boosted Attention model and BAM* denotes the model without using pre-trained stimulus-based attention but with the same architecture as BAM. Reported scores are BLEU-4 (B@4), METEOR (MT), ROUGE-L (RG) and CIDEr (CD). The relative improvement by using the proposed method over its baseline is shown in percentage.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGEL	METEOR	CIDEr
ATT[†] [34]	0.731	0.565	0.424	0.316	0.535	0.250	0.953
SCA-CNN [2]	0.712	0.542	0.404	0.302	0.524	0.244	0.912
SCN-LSTM[†] [5]	0.740	0.575	0.436	0.331	0.543	0.257	1.003
PG-BCMR [20]	0.754	0.591	0.445	0.332	0.550	0.257	1.013
AdaATT[†] [21]	0.748	0.584	0.444	0.336	0.550	0.264	1.042
Att2all[†] [26]	0.781	0.619	0.470	0.352	0.563	0.270	1.147
ours-BAM[†]	0.794	0.622	0.470	0.349	0.560	0.264	1.083

Table 2. Online results (C5) on the MSCOCO evaluation platform, † indicates ensemble of models. Our result is obtained from an ensemble of 4 models trained under different random seeds.

To further study the contributions of stimulus-based attention, we conduct experiments using a model with the same architecture as the proposed model but not initialized on pre-trained weights for stimulus-based attention . In this case, the stimulus-based attention W_{sal} is trained end-to-end and not fixed during supervised learning. As shown in Table 1, models with pre-trained stimulus-based attention (BAM) are able to consistently outperform models without stimulus-based attention (BAM*), indicating that stimulus-based attention plays an essential role on boosting the performance and the improvement of our method is not merely due to advantages of modifications on architecture.

5.2   Qualitative Results

In addition to quantitative evaluations, in this section we further demonstrate the effectiveness of proposed method via comparing qualitative results computed by models with and without using our method. Figure 4 shows the captions generated based the two models, together with the corresponding stimulus-based attention maps computed by models using the Boosted Attention method. Stimulus-based attention maps are generated by normalizing the average activation within the CNN features at different spatial locations.
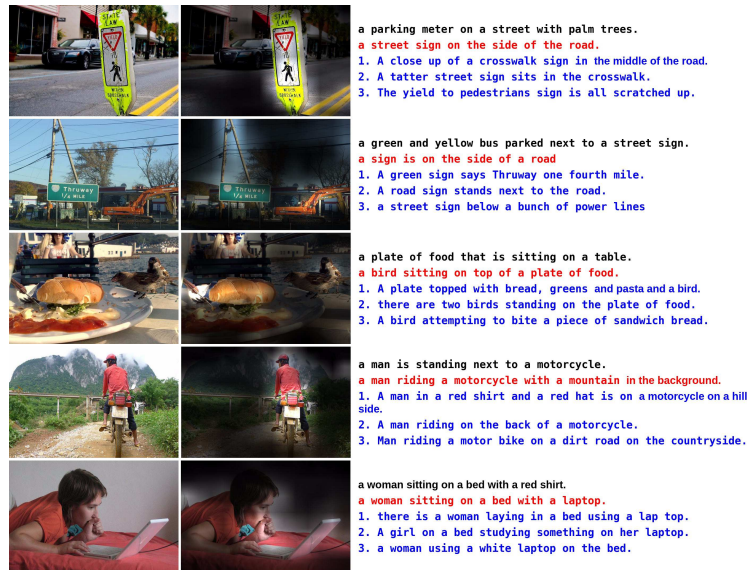

Fig. 4. Qualitative results for models with and without using the Boosted Attention method. From left to right: original images, stimulus-based attention map, and captions corresponding to the images. Captions generated by models with and without using Boosted Attention method are colored in red and black respectively, while the ground-truth human generated captions are colored in blue.

According to the results, introducing stimulus-based attention helps the model efficiently locate the objects of interest within the visual scenarios and generate better captions. For example, in the top two images, the model using Boosted Attention successfully focuses on the *street signs* similar as humans do (as shown in the attention maps as well as the captions in red) while the model without incorporating stimulus-based attention gets lost in the background objects such as the *palm trees* and *bus* (see the captions in black). Furthermore, the results also indicate that the model with the proposed Boosted Attention method is capable of capturing multiple salient objects within images. For example, for the bottom three images, by incorporating stimulus-based attention, the model is able to concentrate on objects including the *bird*, *mountain* and *laptop* (see the attention maps and captions in red). These objects are missing in the captions generated by the model without using Boosted Attention (captions in black) but mentioned in multiple human generated captions (captions in blue).

5.3   Attention Corporation in Image Captioning

To explore how the two types of attention, *i.e.*, stimulus-based attention and top-down model attention, corporate with each other during the caption generation process, we first evaluate the correlations between the attention maps from the two types of attention. The stimulus-based attention map is extracted using the same method described in section 5.2. Since top-down attention maps are generated for each corresponding word within a caption, we compute the average correlations between stimulus-based attention map and top-down attention maps for different words.

We compute the correlations on the 5000 images from Karpathy's test split [12]. Two evaluation metrics commonly used for estimating correlations between spatial maps, *i.e.* Coefficient Correlation (CC) and Spearman's Rank Correlation (Spearman), are utilized for analysis. According to the experimental results, CC and Spearman scores are negative ($CC = -0.256, Spearman = -0.369$), indicating that stimulus-based attention tends to focus on regions different from top-down attention thus the two can potentially complement each other.

Next, we show qualitative results to demonstrate that two attention corporate in a complementary manner. Figure 5 compares top-down attention and its corresponding stimulus-based attention, three typical scenarios for the corporation between attention are summarized as follows:

Scenario I: Stimulus-based attention has successfully captured all of the objects of interest corresponding to generated caption. In this case, top-down attention tends to play a minor role on discriminating the salient regions related to the task. As shown in the first two images, since stimulus-based attention has already concentrated on the objects of interest mentioned in the captions (*i.e.*, *horse* and *church* in the first image, *man* and *giraffe* in the second image), when generating the words corresponding to the objects, top-down attention either does not have a clear focused region (the 1st image) or attends to similar regions
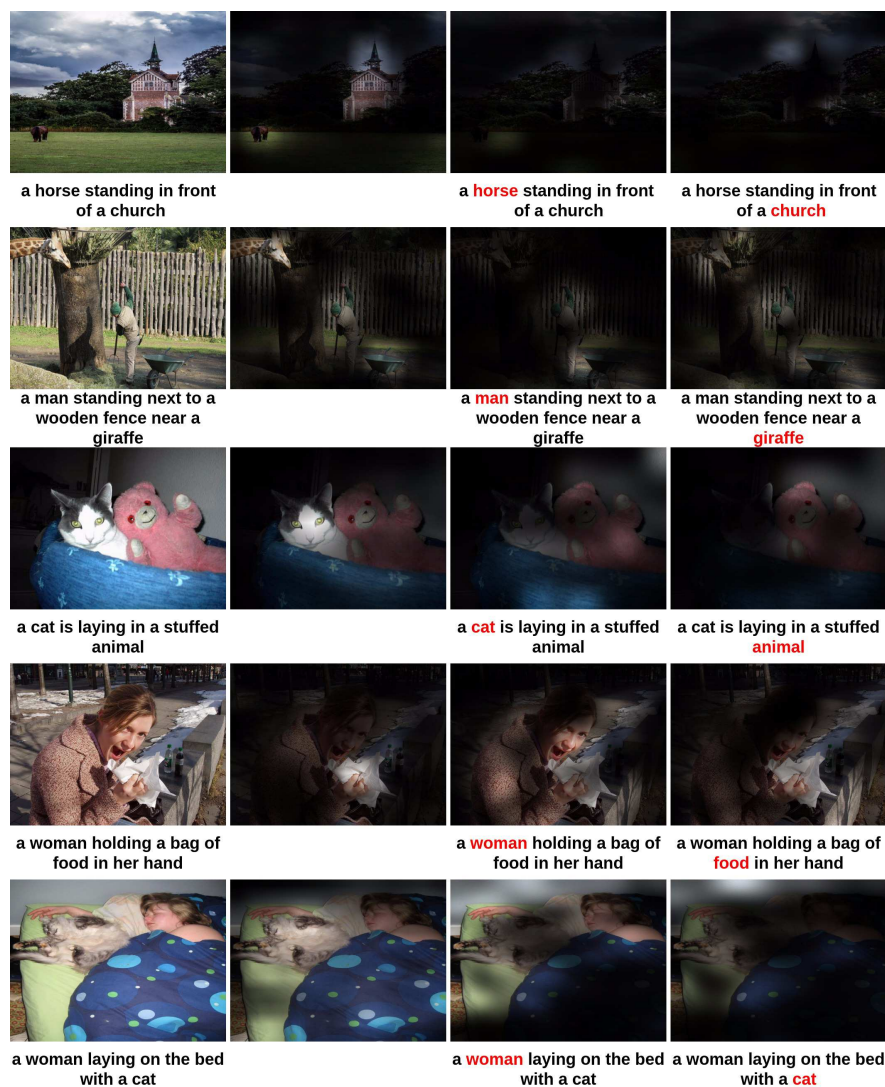
Fig. 5. Qualitative results illustrating that the two types of attention complement each other in various situations. From left to right: original images with generated captions, stimulus-based attention maps, top-down model attention maps for different words within the captions. The word associated with a specific top-down attention map is highlighted in red color.

as stimulus-based attention (the 2nd image).

Scenario II: Stimulus-based attention concentrates on only part of an object but not covering the entire object (*e.g.*, the 3rd image), or it covers some but not all objects of interest (*e.g.*, the 4th image). Under these situations, top-down attention will focus on the missing regions to enhance the objects of interest and complement stimulus-based attention. In the 3rd image, stimulus-based attention highlights the *cat* but only the bottom part of the *stuffed animal*, therefore in order to collect enough visual information when generating the word *'animal'*, top-down attention is placed on the upper part of the *stuffed animal*. Furthermore, in the 4th image we can see that since stimulus-based attention does not quite focus on the *woman*, during generating the word *'woman'* top-attention significantly emphasizes the face of the *woman* and reveals the lost visual information.

Scenario III: Stimulus-based attention fails to distinguish salient objects with irrelevant background. In this case, top-down attention will play a major role in extracting regions corresponding to the objects of interest. As shown in the 5th image, due to the complexity of the visual scenario, stimulus-based attention confuses the objects of interest (*i.e. woman* and *cat* according to the caption) with background objects such as bed and blanket. As a result, the model relies on top-down attention to filter out the irrelevant information and concentrate on regions related to the word being generated.

6   Conclusion

In this work, we propose a Boosted Attention method that leverages human stimulus-based attention to improve the performance of image captioning models. Stimulus-based attention provides prior knowledge on salient regions within the visual scenarios and plays a complementary role to the top-down attention computed by the image captioning models. Experimental results on the MSCOCO and Flickr30K datasets show that the proposed method leads to significant improvements in captioning performance across various evaluation metrics and achieves state-of-the-art results. The proposed method is also general and compatible with various image captioning models using top-down visual attention.

Acknowledgements

References

1. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. pp. 1171–1179. NIPS'15, MIT Press, Cambridge, MA, USA (2015), http://dl.acm.org/citation.cfm?id=2969239.2969370

2. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6298–6306 (2017). https://doi.org/10.1109/CVPR.2017.667

3. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Visual saliency for image captioning in new multimedia services. In: 2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW). pp. 309–314 (2017). https://doi.org/10.1109/ICMEW.2017.8026277

4. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: Proceedings of the 11th European Conference on Computer Vision: Part IV. pp. 15–29. ECCV'10, Springer-Verlag, Berlin, Heidelberg (2010), http://dl.acm.org/citation.cfm?id=1888089.1888092

5. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1141–1150 (2017). https://doi.org/10.1109/CVPR.2017.127

6. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: ECCV (2014)

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)

8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735

9. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Int. Res. **47**(1), 853–899 (2013), http://dl.acm.org/citation.cfm?id=2566972.2566993

10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR **abs/1709.01507** (2017), http://arxiv.org/abs/1709.01507

11. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1072–1080 (2015). https://doi.org/10.1109/CVPR.2015.7298710

12. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015). https://doi.org/10.1109/CVPR.2015.7298932

13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980

14. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th CVPR (2011)

15. Lavie, A., Agarwal, A.: Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: Proceedings of the

Second Workshop on Statistical Machine Translation. pp. 228–231. StatMT '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1626355.1626389

16. Li, J., Xia, C., Song, Y., Fang, S., Chen, X.: A data-driven metric for comprehensive evaluation of saliency models. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 190–198 (2015). https://doi.org/10.1109/ICCV.2015.30

17. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 220–228. CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2018936.2018962

18. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004) (2004)

19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). Zürich (2014), oral

20. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 873–881 (2017). https://doi.org/10.1109/ICCV.2017.100

21. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

22. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Neural Information Processing Systems (NIPS) (2011)

23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1073083.1073135, https://doi.org/10.3115/1073083.1073135

24. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2641–2649 (2015). https://doi.org/10.1109/ICCV.2015.303

25. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1151–1159 (2017). https://doi.org/10.1109/CVPR.2017.128

26. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014), http://arxiv.org/abs/1409.1556

28. Sugano, Y., Bulling, A.: Seeing with humans: Gaze-assisted neural image captioning. CoRR **abs/1608.05203** (2016), http://arxiv.org/abs/1608.05203

29. Tavakoliy, H.R., Shetty, R., Borji, A., Laaksonen, J.: Paying attention to descriptions generated by image captioning models. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2506–2515 (2017). https://doi.org/10.1109/ICCV.2017.272

30. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 4566–4575. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7299087, https://doi.org/10.1109/CVPR.2015.7299087
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164. IEEE Computer Society (2015)
32. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Blei, D., Bach, F. (eds.) Proceedings of the 32nd International Conference on Machine Learning (ICML-15). pp. 2048–2057. JMLR Workshop and Conference Proceedings (2015), http://jmlr.org/proceedings/papers/v37/xuc15.pdf
33. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R.: Review networks for caption generation. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 2361–2369. Curran Associates, Inc. (2016), http://papers.nips.cc/paper/6167-review-networks-for-caption-generation.pdf
34. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4651–4659 (2016). https://doi.org/10.1109/CVPR.2016.503
35. Yun, K., Peng, Y., Samaras, D., Zelinsky, G.J., Berg, T.L.: Studying relationships between human gaze, description, and computer vision. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Computer Society Conference on. IEEE (2013)