Knowing When to Look:
Adaptive Attention via a Visual Sentinel for Image Captioning

CVPR'17

2019.11.26

Junsoo Lee

MOTIVATION

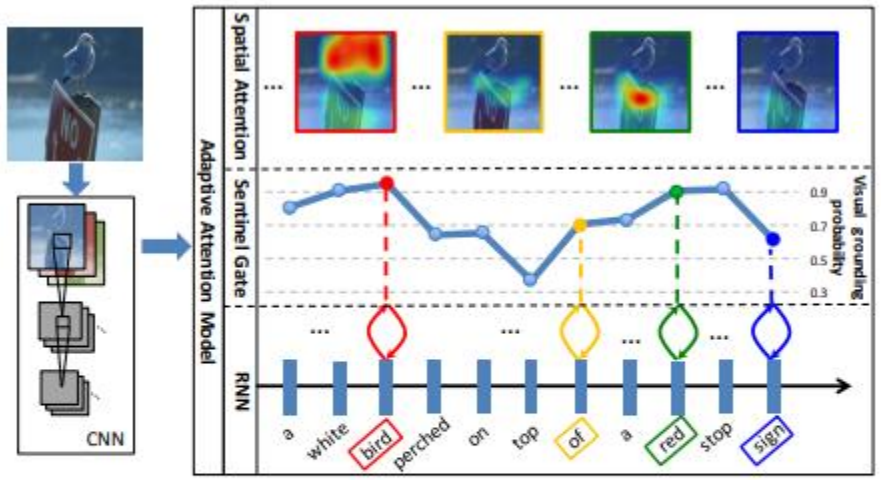

Figure 1: Our model learns an adaptive attention model that automatically determines when to look (**sentinel gate**) and where to look (**spatial attention**) for word generation, which are explained in section 2.2, 2.3 & 5.4.

- Attention-based neural encoder-decoder frameworks **force** "visual attention" to be active for every generated work.
- For example, "a" or "of" do not have corresponding canonical visual signals.
- Moreover, language correlations make the visual signal unnecessary, e.g., "sign" following "a red stop".

CONTRIBUTIONS

- In this work, authors introduce an adaptive attention encoder-decoder framework which can automatically decide **when to rely on visual signals** and **when to just rely on the language model**.
- Authors propose a new LSTM extension, which produces an additional **"visual sentinel" vector** instead of a single hidden state.
- The proposed model significantly outperforms other SOTA methods on MSCOCO and Flickr30k.
- Authors perform an extensive analysis of their adaptive attention model, including visual grounding and weakly supervised localization of generated attention maps.

ENCODER-DECODER FOR IMAGE CAPTIONING

$$\theta^* = argmax_\theta \sum_{(I,y)} \log p(y|I;\theta)$$

where θ are the parameters of the model, I is the image, and $y = \{y_1, \ldots, y_t\}$ is the corresponding caption. Using the chain rule, the log likelihood of the joint probability distribution can be decomposed into ordered conditionals:

$$logp(y) = \sum_{t=1}^{T} logp(y_t|y_1, \ldots, y_{t-1}, I)$$

With recurrent neural network (RNN), each conditional probability is modeled as:

$$logp(y_t|y_1, \ldots, y_{t-1}, I) = f(h_t, c_t)$$

where f is a nonlinear function that outputs the probability of y_t, c_t is the visual context vector at time t extracted from image I.

SPATIAL ATTENTION MODEL

For computing the context vector c_t which is defined as:

$$c_t = g(V, h_t)$$

where g is the attention function, $V = [v_1, \dots, v_k]$, $v_i \in \mathbb{R}^d$ is the spatial image features, each of which is a d dimensional representation corresponding to a part of the image.

$$z_t = w_h^T \tanh(W_v V + (W_g h_t)\mathrm{I}^T$$

$$\alpha_t = softmax(z_t)$$

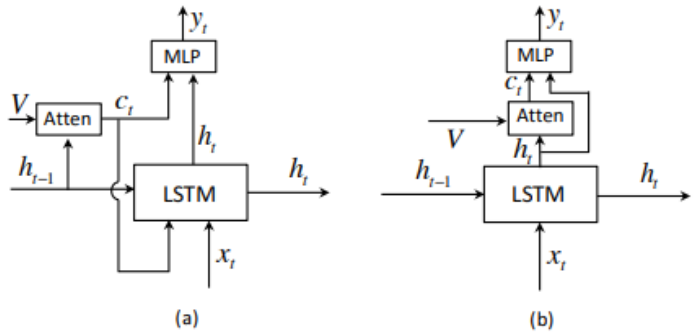

Figure 2: A illustration of soft attention model from [30] (a) and our proposed spatial attention model (b).

Based on the attention distribution, the context vector c_t can be obtained by:

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti}$$

where c_t and h_t are combined to predict next word y_{t+1}.

ADAPTIVE ATTENTION MODEL

The model learns to extract a new component that the model can fall back on when it chooses to not attend to the image.

For computing the "visual sentinel" vector s_t which is defined as:

$$g_t = \sigma(W_x x_t + W_h h_{t-1})$$

$$s_t = g_t \odot \tanh(m_t)$$

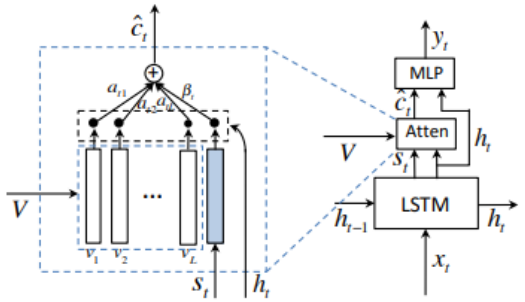

Figure 3: An illustration of the proposed model generating the t-th target word y_t given the image.

where W_x, W_h are projection matrices, x_t is the input to the LSTM at time step t, g_t is the gate applied on the memory cell m_t. $\odot$ represents the element-wise product and σ is the logistic sigmoid activation.

New adaptive context vector is defined as $\widehat{c_t}$, which is modeled as a mixture of the spatially attended image features and the visual sentinel vector. This trades off how much new information the network is considering from the image with what it already knows in the decoder memory (i.e., the visual sentinel).

$$\widehat{c_t} = \beta_t s_t + (1 - \beta_t) c_t$$

$$\widehat{\alpha_t} = softmax([z_t; w_h^T \tanh\left(W_s s_t + \left(W_g h_t\right)\right)]), \widehat{\alpha_t} \in \mathbb{R}^{k+1}$$

$$\beta_t = \widehat{\alpha_t}[k+1]$$

QUANTITATIVE RESULTS

	Flickr30k						MS-COCO					
Method	B-1	B-2	B-3	B-4	METEOR	CIDEr	B-1	B-2	B-3	B-4	METEOR	CIDEr
DeepVS [11]	0.573	0.369	0.240	0.157	0.153	0.247	0.625	0.450	0.321	0.230	0.195	0.660
Hard-Attention [30]	0.669	0.439	0.296	0.199	0.185	-	0.718	0.504	0.357	0.250	0.230	-
ATT-FCN† [34]	0.647	0.460	0.324	0.230	0.189	-	0.709	0.537	0.402	0.304	0.243	-
ERD [32]	-	-	-	-	-	-	-	-	-	0.298	0.240	0.895
MSM† [33]	-	-	-	-	-	-	0.730	0.565	0.429	0.325	0.251	0.986
Ours-Spatial	0.644	0.462	0.327	0.231	0.202	0.493	0.734	0.566	0.418	0.304	0.257	1.029
Ours-Adaptive	**0.677**	**0.494**	**0.354**	**0.251**	**0.204**	**0.531**	**0.742**	**0.580**	**0.439**	**0.332**	**0.266**	**1.085**

Table 1: Performance on Flickr30k and COCO test splits. † indicates ensemble models. **B-n** is BLEU score that uses up to n-grams. Higher is better in all columns. For future comparisons, our ROUGE-L/SPICE Flickr30k scores are 0.467/0.145 and the COCO scores are 0.549/0.194.

QUANTITATIVE RESULTS

Method	B-1		B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC [27]	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946
MS Captivator [8]	0.715	0.907	0.543	0.819	0.407	0.710	0.308	0.601	0.248	0.339	0.526	0.680	0.931	0.937
m-RNN [18]	0.716	0.890	0.545	0.798	0.404	0.687	0.299	0.575	0.242	0.325	0.521	0.666	0.917	0.935
LRCN [7]	0.718	0.895	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.528	0.678	0.921	0.934
Hard-Attention [30]	0.705	0.881	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.516	0.654	0.865	0.893
ATT-FCN [34]	0.731	0.900	0.565	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958
ERD [32]	0.720	0.900	0.550	0.812	0.414	0.705	0.313	0.597	0.256	0.347	0.533	0.686	0.965	0.969
MSM [33]	0.739	0.919	0.575	0.842	0.436	0.740	0.330	0.632	0.256	0.350	0.542	0.700	0.984	1.003
Ours-Adaptive	**0.748**	**0.920**	**0.584**	**0.845**	**0.444**	**0.744**	**0.336**	**0.637**	**0.264**	**0.359**	**0.550**	**0.705**	**1.042**	**1.059**

Table 2: Leaderboard of the published state-of-the-art image captioning models on the online COCO testing server. Our submission is a ensemble of 5 models trained with different initialization.

EXPERIMENT CONSIDERATIONS

- Flickr30k contains 31,783 images, each of which is paired with 5 crowd-sourced captions.
- MSCOCO contains 82,783, 40,504 and 40,775 images for training, validation and test respectively.
- The authors truncate captions longer than 18 words for COCO and 22 for Flickr30k.
- Building a vocabulary of words that occur at least 5 and 3 times in the training set, resulting in 9567 and 7649 words for COCO and Flickr30k respectively.

QUALITATIVE RESULTS

Figure 4: Visualization of generated captions and image attention maps on the COCO dataset. Different colors show a correspondence between attended regions and underlined words. First 2 columns are success cases, last columns are failure examples. Best viewed in color.

QUALITATIVE RESULTS

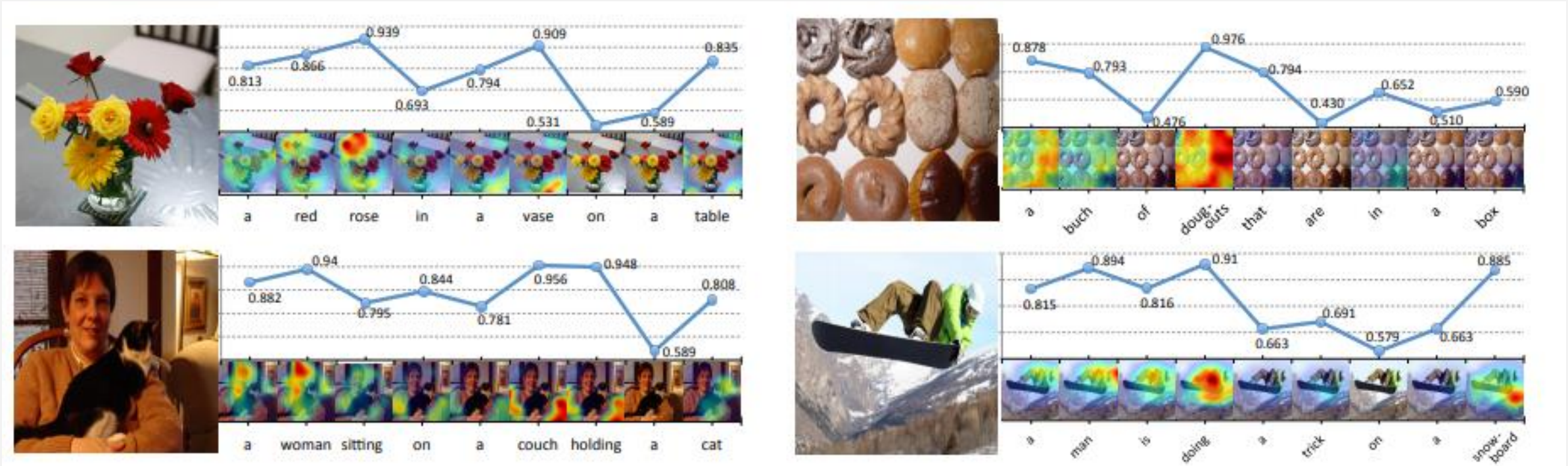

Figure 5: Visualization of generated captions, visual grounding probabilities of each generated word, and corresponding spatial attention maps produced by our model.

The authors define $1 - \beta$ as its visual grounding probability.

QUALITATIVE RESULTS

Learning "where" to attend

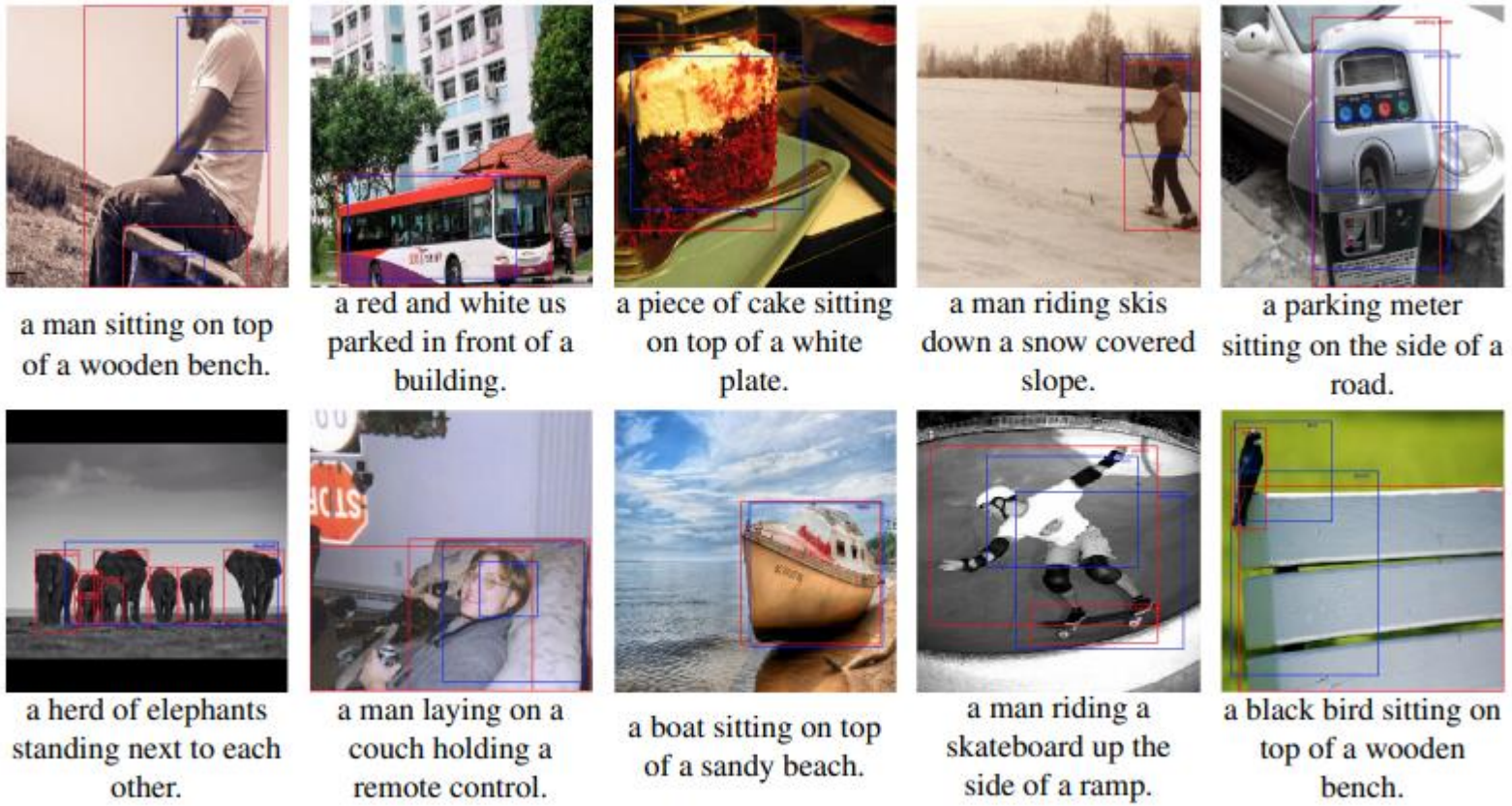

Figure 11: Visualization of generated captions and weakly supervised localization result. Red bounding box is the ground truth annotation, blue bounding box is the predicted location using spatial attention map.

Given the word w_t and attention map α_t, the authors first segment the regions of the image with attention values larger than threshold, where it is a per-class threshold estimated using the COCO validation split.