Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

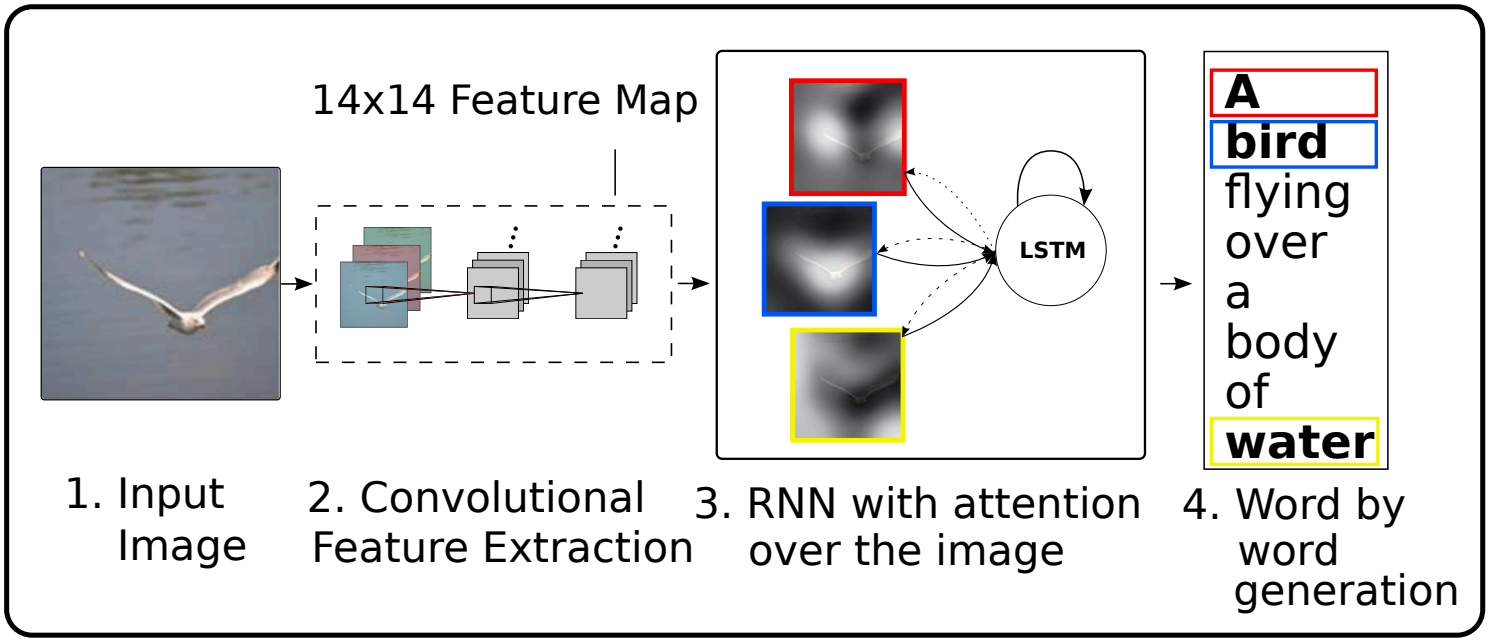
K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel & Y. Bengio

ICML 2015

Discussion by Yunchen Pu

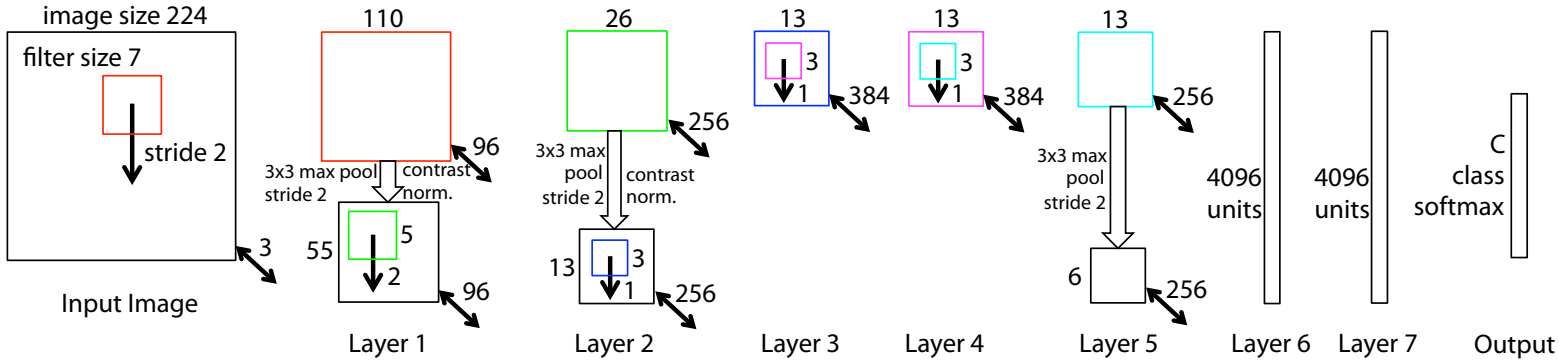
September 25, 2015

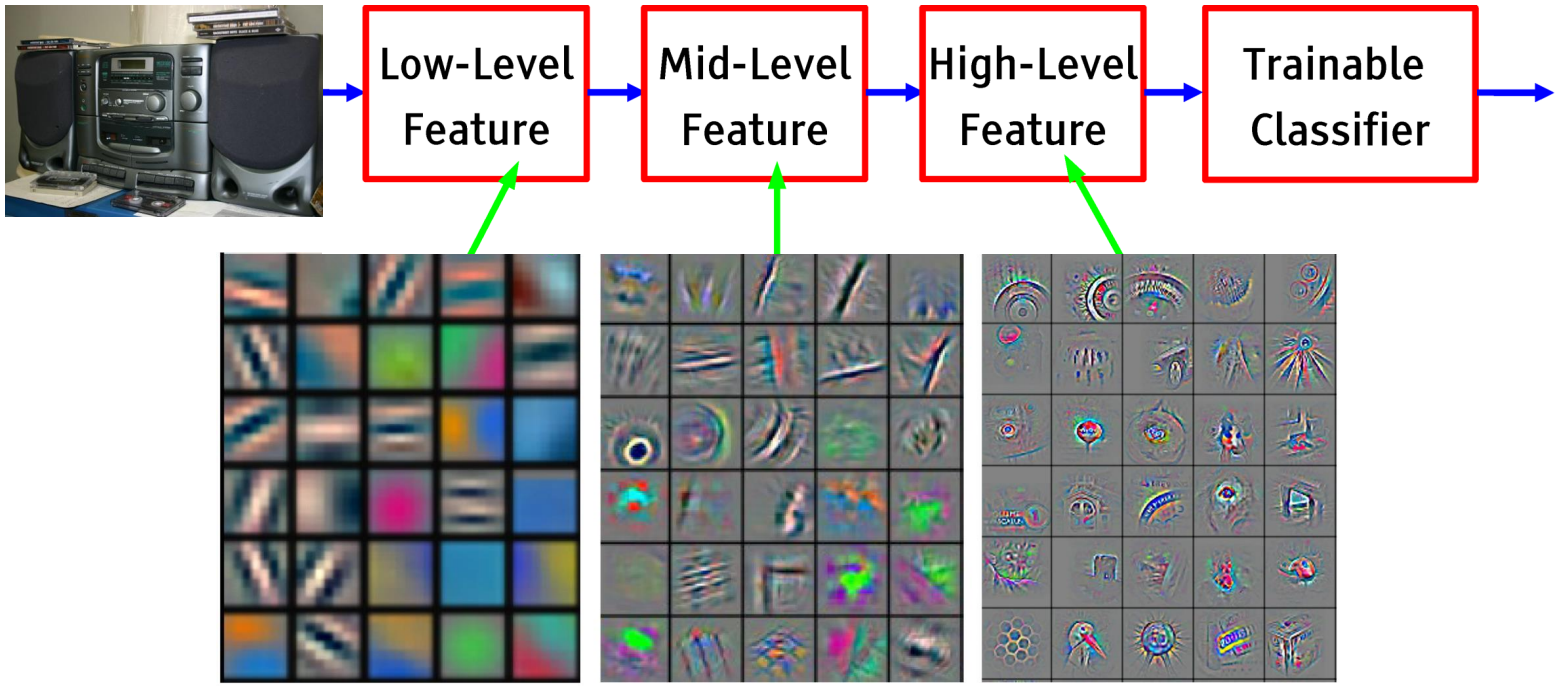
Model



- Use low-level features to represent the images.
- Use attention-based model to dynamically select needed features.

Extract Features from CNN[1]

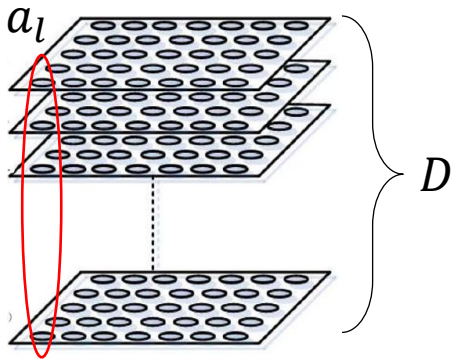




[1] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," ECCV, 2014.

Recurrent Neural Network

- Notation:
  - Caption sequence: $\boldsymbol{y} = \{y_1, \ldots, y_C\}$, $y_t \in \mathbb{R}^K$.
  - Image feature: $\boldsymbol{a} = \{a_1, \ldots, a_L\}$, $a_l \in \mathbb{R}^D$.
  - Context vector: $\hat{\boldsymbol{z}} = \{\hat{z}_1, \ldots, \hat{z}_C\}$, $\hat{z}_t \in \mathbb{R}^D$
  - Hidden unit: $\boldsymbol{h} = \{h_1, \ldots, h_C\}$, $h_t \in \mathbb{R}^n$.
  - Embedding matrix: $\mathbf{E} \in \mathbb{R}^{\mathbf{m} \times \mathbf{K}}$



- The hidden units h_t are computed by Long Short-Term Memory (LSTM).
- The output word probability:

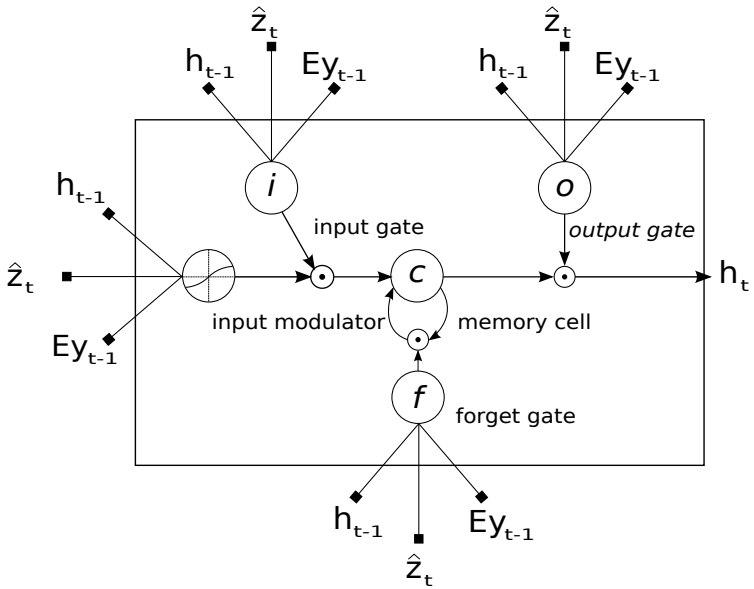
$$p(y_t|a, y_{j<t-1}) \propto \exp\{\mathbf{L}_o(\mathbf{E}\boldsymbol{y}_{t-1} + \mathbf{L}_h h_t + \mathbf{L}_z \hat{z}_t)\} \tag{1}$$

- Goal: maximize the log likelihood:

$$\log p(\boldsymbol{y}|a) = \sum_{t=1}^{C} \log p(y_t|a, y_{j<t-1}) \tag{2}$$

- $\mathbf{L}_o \in \mathbb{R}^{K \times m}$, $\mathbf{L}_h \in \mathbb{R}^{m \times n}$, $\mathbf{L}_z \in \mathbb{R}^{m \times D}$, and $\mathbf{E}$ are learned parameters.

Long Short-Term Memory



$$
\begin{aligned}
i_t &= \sigma(W_i Ey_{t-1} + U_i h_{t-1} + Z_i \hat{z}_t + b_t) \\
f_t &= \sigma(W_f Ey_{t-1} + U_f h_{t-1} + Z_f \hat{z}_t + b_f) \\
g_t &= \tanh(W_c Ey_{t-1} + U_c h_{t-1} + Z_c \hat{z}_t + b_c) \\
o_t &= \sigma(W_o Ey_{t-1} + U_o h_{t-1} + Z_o \hat{z}_t + b_o) \\
c_t &= f_t c_{t-1} + i_t g_t \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

where i_t, f_t, c_t, o_t and h_t are the input, forget, memory, output and hidden state of the LSTM repectively. $\mathbf{W}$, $\mathbf{U}$, $\mathbf{Z}$ and $\boldsymbol{b}$ are learned parameters.

Context Vector

- The context vector: $\hat{z}_t = \phi(\{a_i\}, \{\alpha_{ti}\})$.
- The positive weight: $\alpha_{ti} = \exp(e_{ti}) / \sum_k \exp(e_{tk})$.
- Attention model[2]:

$$e_{ti} = f_{att}(a_i, h_{t-1}) = \begin{cases} a_i^\top W_a h_{t-1} \\ V_a^\top \tanh(W_a[a; h_{t-1}]) \end{cases} \tag{3}$$

- Deterministic "Soft" Attention: $\hat{z}_t = \sum_{i=1}^{L} \alpha_{ti} a_i$
- Stochastic "Hard" Attention: $\boldsymbol{s}_t$ is a one-hot vector:

$$p(s_{ti} = 1 | s_{j<t}, \boldsymbol{a}) = \alpha_{ti} \qquad \hat{z}_t = \sum_{i=1}^{L} s_{ti} a_i \tag{4}$$

[2]Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. . "Effective Approaches to Attention-based Neural Machine Translation." EMNLP 2015.

Stochastic "Hard" Attention

- The variational lower bound:

$$\log p(\boldsymbol{y}|a) = \log \sum_s p(s|\boldsymbol{a}) p(\boldsymbol{y}|s, \boldsymbol{a}) \geq \sum_s p(s|\boldsymbol{a}) \log p(\boldsymbol{y}|s, \boldsymbol{a}) = L_s \tag{5}$$

- The gredient w.r.t. model parameter W is:

$$\frac{\partial L_s}{\partial W} = \sum_s p(s|\boldsymbol{a}) \left[\frac{\partial \log p(\boldsymbol{y}|s, \boldsymbol{a})}{\partial W} + \log p(\boldsymbol{y}|s, \boldsymbol{a}) \frac{\partial \log p(s|\boldsymbol{a})}{\partial W} \right] \tag{6}$$

$$\approxeq \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\partial \log p(\boldsymbol{y}|\tilde{s}^n, \boldsymbol{a})}{\partial W} + \log p(\boldsymbol{y}|\tilde{s}^n, \boldsymbol{a}) \frac{\partial \log p(\tilde{s}^n|\boldsymbol{a})}{\partial W} \right] \tag{7}$$

Where $\tilde{s}^n$ are samples from $\mathit{Multinoulli}(\{\alpha_i\})$

Experiments

- Three benchmark datasets: Flickr8k(8000 images), Flickr30K(30000 images) and MS COCO(82783 images).
- Optimization method: RMSProp for Flickr8k dataset; Adam for Flickr30k and MS COCO dataset.
- Use the feature map of the fourth convolutional layer before max-pooling from an ImageNet pretrained VGGnet (19 layers in total).
- Speed up: During training, set the caption in each mini-batch to be the same length.

Experiments: Quantitative Analysis[3]

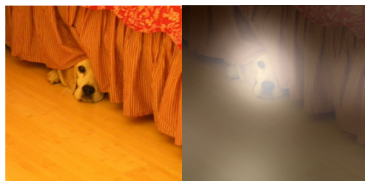
Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods, † indicates a different split, (—) indicates an unknown metric, ∘ indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, a indicates using AlexNet

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

[3] BLEU-n is the geometric average of the n-gram precision

Experiments: Qualitative Analysis

A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.

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

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

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

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

Figure : Examples of attending to the correct object

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Figure : Examples of attending to the wrong object

Experiments: Qualitative Analysis

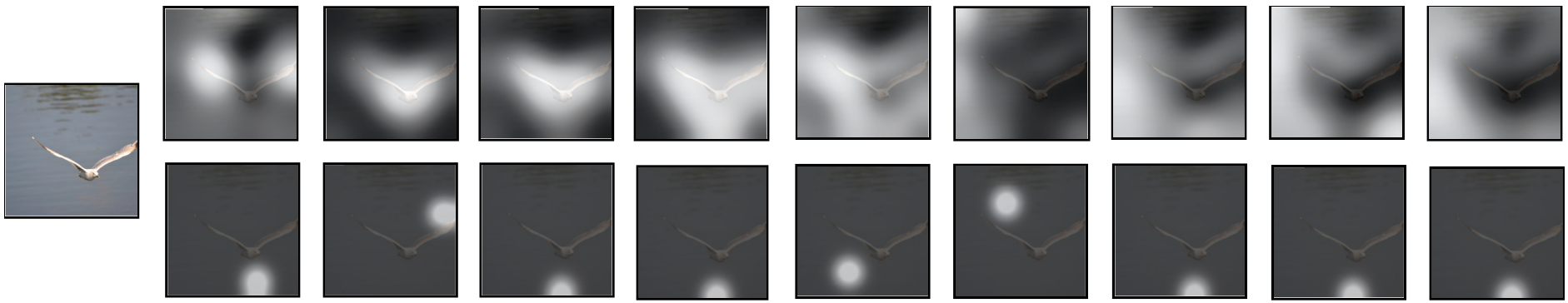
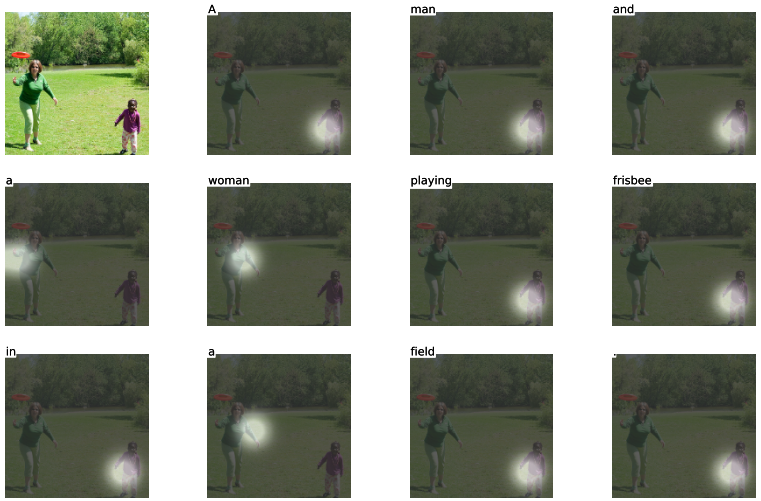


Figure : Visualizations from our "hard" (bottom) and "soft" (top) attention model

Experiments: Qualitative Analysis

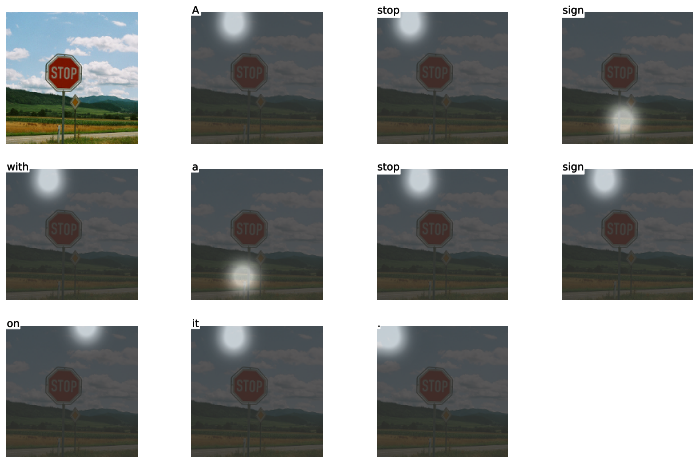


(a) A man and a woman playing frisbee in a field.

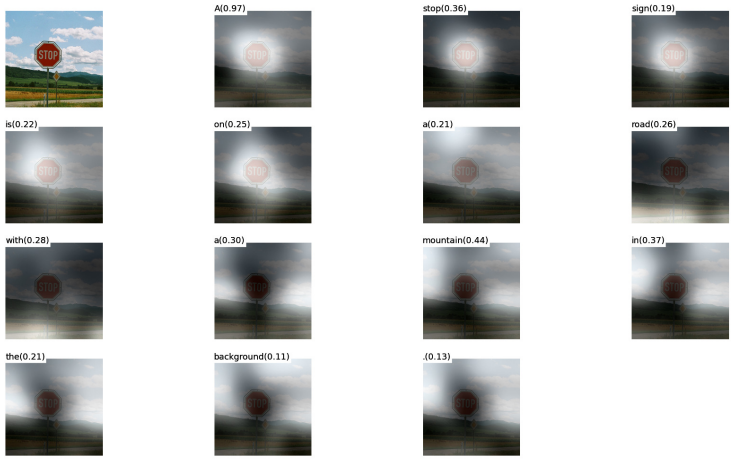
(b) A woman is throwing a frisbee in a park.

Figure : Visualizations from our "hard" (a) and "soft" (b) attention model

Experiments: Qualitative Analysis



(a) A stop sign with a stop sign on it.



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

Figure : Visualizations from our "hard" (a) and "soft" (b) attention model