Attention Correctness in Neural Image Captioning

Chenxi Liu[1] Junhua Mao[2] Fei Sha[2, 3] Alan Yuille[1, 2]
[1]Johns Hopkins University [2]University of California, Los Angeles [3]University of Southern California

Motivation

- Attention models are effective at handling many AI tasks
- We focus on attention models in image captioning
- Although impressive visualization of the attention maps are shown, no *quantitative* evaluations are available

We study the following two questions:

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

Contributions:

- Evaluation metric "attention correctness"
- Quantitative analysis of attention quality
- Propose a supervised attention model
- Close gap between machine and human perception

Implicit Attention Model

Proposed in (Xu et al. 2015)

- Visual features for different spatial locations $a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}$
- LSTM network

$$\mathbf{i}_t = \sigma(W_i E\mathbf{y}_{t-1} + U_i \mathbf{h}_{t-1} + Z_i \mathbf{z}_t + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(W_f E\mathbf{y}_{t-1} + U_f \mathbf{h}_{t-1} + Z_f \mathbf{z}_t + \mathbf{b}_f)$$
$$\mathbf{c}_t = \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh(W_c E\mathbf{y}_{t-1} + U_c \mathbf{h}_{t-1} + Z_c \mathbf{z}_t + \mathbf{b}_c)$$
$$\mathbf{o}_t = \sigma(W_o E\mathbf{y}_{t-1} + U_o \mathbf{h}_{t-1} + Z_o \mathbf{z}_t + \mathbf{b}_o)$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t)$$

- Context vector $\mathbf{z}_t = \sum_{i=1}^{L} \alpha_{ti} \mathbf{a}_i$

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

- Word generation and loss function

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

Supervised Attention Model

- Minimize the distance between generated attention map $\boldsymbol{\alpha}_t$ and ground truth attention map $\boldsymbol{\beta}_t$
- Use cross entropy loss since $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ both sum to 1

$$L_{t,attn} = \begin{cases} -\sum_{i=1}^{L} \beta_{ti} \log \alpha_{ti} & \text{if } \boldsymbol{\beta}_t \text{ exists for } w_t \\ 0 & \text{otherwise} \end{cases}$$

- Total loss

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

- What remains is how to construct $\boldsymbol{\beta}_t$

Strong supervision:

- Write 1 inside the ground truth region, and 0 elsewhere

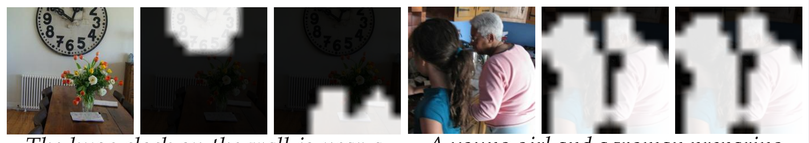
$$\hat{\beta}_{t\hat{i}} = \begin{cases} 1 & \hat{i} \in R_t \\ 0 & \text{otherwise} \end{cases}$$

Weak supervision:

- Alignment is expensive to collect to annotate
- Segmentation masks with object class labels are more common
- Approximate image-to-language consistency by language-to-language similarity

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

- In practice we use cosine distance of word2vec



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

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

Attention Correctness

Word level:

- Sum of attention score that falls within the ground truth region
- A score between 0 and 1

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

0.08	0.12	0.20	0.12
0.04	0.10	0.12	0.08
0.00	0.02	0.08	0.04
0.00	0.00	0.00	0.00

Phrase level:

- Maximum of individual word attention correctness

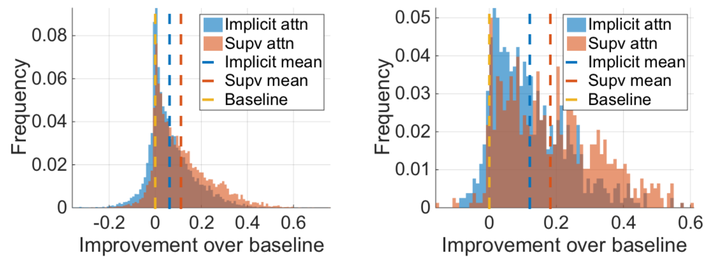
$$AC(\{y_t, \ldots, y_{t+l}\}) = \max(AC(y_t), \ldots, AC(y_{t+l}))$$

Results

Evaluation of Attention Correctness:

- On Flickr30k test set

Caption	Model	Baseline	Correctness
Ground Truth	Implicit	0.3214	0.3836
	Supervised	0.3214	**0.4329**
Generated	Implicit	0.3995	0.5202
	Supervised	0.3968	**0.5787**



- Improvement is greatest for small objects

Evaluation of Captioning Performance:

- Caption quality consistently increases with supervision
- No matter strong or weak

Dataset	Model	BLEU-3	BLEU-4	METEOR
Flickr30k	Implicit	28.8	19.1	18.49
	Implicit*	29.2	20.1	19.10
	Strong Sup	**30.2**	**21.0**	**19.21**
COCO	Implicit	34.4	24.3	23.90
	Implicit*	36.4	26.9	24.46
	Weak Sup	**37.2**	**27.6**	**24.78**

- Higher attention correctness results in better captions

Correctness	BLEU-3	BLEU-4	METEOR
High	38.0	28.1	23.01
Middle	36.5	26.1	21.94
Low	35.8	25.4	21.14

A man in a red jacket and blue pants is snowboarding.

A man in a red jumpsuit and a black hat is snowboarding.

Conclusion:

- Attention models attend to meaningful regions if compared against uniform baseline, but still have room for improvement
- There exists a positive correlation between attention correctness and captioning quality
- Closing the gap between machine attention and human perception is necessary and important