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

by Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, ICML 2015
Introduction

- “Scene understanding”
- Purpose of attention?
  - allows for salient features to dynamically come to the forefront as needed.
  - “hard” attention & “soft attention
Model - Encoder

• Model takes a single raw image and generates a caption $y$ encoded as a sequence of 1-of-$K$ encoded words.

• Caption: $y = y_1, \ldots, y_C, y_i \in K$ dimensional

• Image: $a = a_1, \ldots, a_L, a_i \in K$ dimensional

$K$: vocab size, $C$: caption length $D$: dim. of representation corresponding to a part of the image
• The features are extracted from a lower conv layer unlike previous works which used a FC layer
Model - Decoder

- Use a LSTM that produces a caption by generating one word at every time step ($y_t$) conditioned on a context vector ($\hat{z}_t$), the previous hidden state ($h_{t-1}$) and the previously generated words ($y_{t-1}$).

- $i_t = \sigma(W_i E y_{t-1} + U_i h_{t-1} + Z_i \hat{z}_t + b_i)$,
- $f_t = \sigma(W_f E y_{t-1} + U_f h_{t-1} + Z_f \hat{z}_t + b_f)$,
- $c_t = f_c c_{t-1} + i_t \tanh(W_c E y_{t-1} + U_c h_{t-1} + Z_c \hat{z}_t + b_c)$,
- $o_t = \sigma(W_o E y_{t-1} + U_o h_{t-1} + Z_o \hat{z}_t + b_o)$,
- $h_t = o_t \tanh(c_t)$. 
Model – Decoder: Context vector, $\hat{z}_t$

- Dynamic representation of the relevant part of the image input at time, $t$

$$\hat{z}_t = \phi(\{a_i\}, \{\alpha_i\})$$

- (Stochastic attention) : the probability that location $i$ is the right place to focus for producing the next word

- (Deterministic attention) : the relative importance to give to location $i$ in blending the $ai$’s together
Stochastic “Hard” Attention

• The location variable $s_t$ as where the model decides to focus attention when generating the $t$ th word. $s_{t,i}$ is an indicator one-hot variable which is set to 1 if the $i$-th location (out of $L$) is the one used to extract visual features.

$$p(s_{t,i} = 1|s_{j<t}, a) = \alpha_{t,i}$$

$$\hat{z}_t = \sum_i s_{t,i}a_i$$
Deterministic “Soft” Attention

- Take the expectation of the context vector \( \hat{z}_t \) directly and formulate a deterministic attention model by computing a soft attention weighted annotation vector \( \phi \)

\[
E_p(s_t|a)[\hat{z}_t] = \sum_{i=1}^{L} \alpha_{t,i} a_i
\]

- This is the same as the original attention mechanism
- Loss for soft-attention

\[
L_d = -\log(p(y|a)) + \lambda \sum_i \left(1 - \sum_t \alpha_{ti}\right)^2
\]
Training

• Both variants of attention model were trained with SGD using adaptive learning rate
• To create $a_i$, they used VGG pretrained on ImageNet without finetuning
Experiments

• Data

<table>
<thead>
<tr>
<th>Flickr8k</th>
<th>Flickr30k</th>
<th>MS COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,000 images</td>
<td>30,000 images</td>
<td>82,738 images</td>
</tr>
<tr>
<td>5 reference sentences / image</td>
<td></td>
<td>More than 5 / image</td>
</tr>
</tbody>
</table>

• Metric: BLEU (Bilingual Evaluation Understudy)
  • Metric used to evaluate Machine Translation
  • We know this from earlier discussions
Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods. † indicates a different split, (—) indicates an unknown metric, o indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, α indicates using AlexNet

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr8k</td>
<td>Google NIC (Vinyals et al., 2014)†Σ</td>
<td>63</td>
<td>41</td>
<td>27</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear (Kiros et al., 2014a)ο</td>
<td>65.6</td>
<td>42.4</td>
<td>27.7</td>
<td>17.7</td>
<td>17.31</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>67</td>
<td>44.8</td>
<td>29.9</td>
<td>19.5</td>
<td>18.93</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>67</td>
<td>45.7</td>
<td>31.4</td>
<td>21.3</td>
<td>20.30</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>Google NIC†οΣ</td>
<td>66.3</td>
<td>42.3</td>
<td>27.7</td>
<td>18.3</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear</td>
<td>60.0</td>
<td>38</td>
<td>25.4</td>
<td>17.1</td>
<td>16.88</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>66.7</td>
<td>43.4</td>
<td>28.8</td>
<td>19.1</td>
<td>18.49</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>66.9</td>
<td>43.9</td>
<td>29.6</td>
<td>19.9</td>
<td>18.46</td>
</tr>
<tr>
<td>COCO</td>
<td>CMU/MS Research (Chen &amp; Zitnick, 2014)α</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>20.41</td>
</tr>
<tr>
<td></td>
<td>MS Research (Fang et al., 2014)†α</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>20.71</td>
</tr>
<tr>
<td></td>
<td>BRNN (Karpathy &amp; Li, 2014)ο</td>
<td>64.2</td>
<td>45.1</td>
<td>30.4</td>
<td>20.3</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Google NIC†οΣ</td>
<td>66.6</td>
<td>46.1</td>
<td>32.9</td>
<td>24.6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinearο</td>
<td>70.8</td>
<td>48.9</td>
<td>34.4</td>
<td>24.3</td>
<td>20.03</td>
</tr>
<tr>
<td></td>
<td>Soft-Attentionο</td>
<td>70.7</td>
<td>49.2</td>
<td>34.4</td>
<td>24.3</td>
<td>23.90</td>
</tr>
<tr>
<td></td>
<td>Hard-Attentionο</td>
<td>71.8</td>
<td>50.4</td>
<td>35.7</td>
<td>25.0</td>
<td>23.04</td>
</tr>
</tbody>
</table>
Results

• Achieve state-of-the-art results on MS COCO dataset