Areas of Attention in Image Captioning

Marco Pedersoli*, Thomas Lucas
Cordelia Schmid, Jakob Verbeek

INRIA Grenoble Rhône-Alpes, France
* Now at École de Technologie Supérieure Montreal, Canada
Image captioning

- Given an image, generate a natural language description

Figure taken from [Kiros et al., 2015]

- a black and white photo of a window.
- a young boy standing on a parking lot next to cars.
- a wooden table and chairs arranged in a room.
- a car is parked in the middle of nowhere.
- a ferry boat on a marina with a group of people.
- a little boy with a bunch of friends on the street.
Encoder-decoder models for captioning

- State of the art based on encoder-decoder approach
  [Kiros et al., 2014]
  - Inspired from encoder-decoder models in machine translation, see e.g. [Sutskever et al., 2014]

- Encoder transforms input to a internal representation

- Decoder maps internal representation to output

Figure taken from [Vinyals et al., 2015]
Limitations

- Only discriminative training
  - Pure-text corpus to better learn language?
  - Image-only data to learn image parser?

- Limited to a fixed vocabulary
  - How to generalize better from few examples?
  - Character-level prediction?

- Single image parse into a vector representation
  - Global image representation, how to get compositionality?
  - How to offload visual content from memory state?

Figure taken from [Kiros et al., 2015]
Leveraging locality and compositionality with attention

- **Sequentially** attend to different parts of the input
- Associate local image evidence with words in caption
- Also used in speech recognition and machine translation

Which areas to consider?

Which mechanism to exploit these areas?
Baseline: “vanilla” captioning system

Figure taken from [Vinyals et al., 2015]

A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.
Encoder

- CNN with VGG-16 architecture
  [Simonyan and Zisserman, 2015]
  - 16 layers with trainable weights, 138M parameters
  - Penultimate layer of ImageNet pre-trained model

Figure taken from [Noh et al., 2015]
Decoder

- GRU-based RNN [Chung et al., 2014]
  - State initialized with CNN code
  - Previous word used as input: “output feedback”

\[
\begin{align*}
  z_t &= \sigma (W_z \cdot [h_{t-1}, x_t]) \\
  r_t &= \sigma (W_r \cdot [h_{t-1}, x_t]) \\
  \tilde{h}_t &= \tanh (W \cdot [r_t \ast h_{t-1}, x_t]) \\
  h_t &= (1 - z_t) \ast h_{t-1} + z_t \ast \tilde{h}_t
\end{align*}
\]

Figures taken from [Karpathy and Fei-Fei, 2015] and http://colah.github.io
Baseline model: word prediction

- Baseline RNN is based on state-word interactions

\[
p(w_t|h_t) \propto \exp \left( w_t^T W \theta_{wh} h_t \right)
\]  

- \( w_t \): 1-hot coding of word at time \( t \)
- \( W \): contains word-embedding vectors in rows
- \( \theta_{wh} \): parameter matrix to score word-state combination
- Think: “a logistic discriminant word-classifier given state”

- Train: maximum-likelihood using ground-truth inputs for state evolution (“teacher forced”)

- Test: Generate approximate maximum-likelihood sentences with beam-search
Our “Areas of Attention” model

- Based on scoring state-word-region combinations
  - Which region-word pair “stands out” given the current state?

\[
p(w_t, r_t | h_t) \propto \exp s(w_t, r_t, h_t),
\]

\[
s(w_t, r_t, h_t) = w_t^T W \theta_{wh} h_t + w_t^T W \theta_{wr} R^T r_t + r_t^T R \theta_{rh} h_t + w_t^T W \theta_w + r_t^T R \theta_r,
\]

- \(w_t\): 1-hot coding of word at time \(t\)
- \(W\): contains word-embedding vectors in rows
- \(r_t\): 1-hot coding of region at time \(t\)
- \(R\): contains region feature vectors in rows
- \(\theta_{wh}, \theta_{wr}, \theta_{rh}\): region-word-state interaction matrices
- \(\theta_w, \theta_r\): region and word bias vectors
Our “Areas of Attention” model

- Predict words using \( p(w_t|h_t) = \sum_{r_t} p(w_t, r_t|h_t) \)
- Use appearance of attended regions for state update

\[
\nu_t = \sum_{r_t} p(r_t|h_t)r_t^\top R, \quad (4)
\]

\[
h_{t+1} = \text{GRU}(h_t, [w_t^\top W \nu_t^\top]^\top). \quad (5)
\]
And how about the regions?

- Our AoA model is agnostic to type of image region, experimentally we compare three different region types

  - **Activation grid**: take positions of conv5 layer as regions, descriptor is “column” of activations across feature channels
  
  - **Object proposals**: using EdgeBox object proposals [Zitnick and Dollár, 2014], average conv5 features over box
  
  - **Spatial transformer**: predict region from each conv4 position, compute conv5 features over warped $3 \times 3$ area
Spatial transformer regions

- Localization network regresses affine transformations for all feature map positions.
- Transformations are applied to the anchor boxes that are used to locally re-sample the feature map, before convolution.
- Reverts to “Activation grid” for identity transformation.
Microsoft Common Objects in Context (MSCOCO)

• 80k train, 40 development images, 5 sentences per image

1. A woman kneeling down next to a dog on a snow covered slope.
2. A boy and his dog are playing in the snow.
3. A snowboarder in a blue jacket and a black and brown dog.
4. Snowboarder sitting next to a dog in the snow.
5. A snowboarder sits in snow beside a dog.
Evaluation of model components

- Using activation grid as attention areas

<table>
<thead>
<tr>
<th>Method</th>
<th>B1</th>
<th>B4</th>
<th>Meteor</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: $\theta_{wh}$</td>
<td>66.3</td>
<td>26.4</td>
<td>22.2</td>
<td>78.9</td>
</tr>
<tr>
<td>Ours: $\theta_{wh}, \theta_{wr}$</td>
<td>68.0</td>
<td>28.0</td>
<td>22.9</td>
<td>83.6</td>
</tr>
<tr>
<td>Ours: $\theta_{wh}, \theta_{wr}, \theta_{rh}$</td>
<td>68.2</td>
<td>28.4</td>
<td>23.3</td>
<td>85.5</td>
</tr>
<tr>
<td>Ours: conditional feedback</td>
<td>68.3</td>
<td>28.7</td>
<td>23.7</td>
<td>86.8</td>
</tr>
<tr>
<td>Ours: full model</td>
<td>69.1</td>
<td>28.8</td>
<td>23.7</td>
<td>87.4</td>
</tr>
</tbody>
</table>

- Local word-region interaction improves
- Local region-state interaction improves
- Word-conditioning visual feedback, i.e. using $p(r_t | w_t, h_t)$ instead of $p(r_t | h_t)$, degrades w.r.t. full model
Evaluation of attention areas

- Object proposals: top regions by “objectness”
- Grids + transformers: regular sampling
Effect of CNN fine-tuning

- RNN training only: fixed pre-trained CNN
- CNN-RNN fine-tuning: second stage trains all

<table>
<thead>
<tr>
<th>Method</th>
<th>B1</th>
<th>B4</th>
<th>Meteor</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RNN training only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>66.3</td>
<td>26.4</td>
<td>22.2</td>
<td>78.9</td>
</tr>
<tr>
<td>Spatial transformers</td>
<td><strong>70.2</strong></td>
<td><strong>30.2</strong></td>
<td><strong>24.2</strong></td>
<td><strong>91.1</strong></td>
</tr>
<tr>
<td><strong>CNN-RNN fine-tuning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>68.6</td>
<td>28.7</td>
<td>23.5</td>
<td>87.1</td>
</tr>
<tr>
<td>Spatial transformers</td>
<td><strong>70.8</strong></td>
<td><strong>30.7</strong></td>
<td><strong>24.5</strong></td>
<td><strong>93.8</strong></td>
</tr>
</tbody>
</table>
Comparison of attention areas
Comparison of attention areas

Grids
A large jetliner flying through a cloudy sky.

Props.
A plane flying in the sky with a sky background.

Transformers
A plane is flying low over a field.
Comparison of attention areas

**Grids**

A couple of elephants standing next to each other.

**Proposals**

A large elephant standing in a field of grass.

**Transformers**

A couple of elephants standing in a field.
Comparison to the state of the art

- Competitive with state-of-the-art methods
- More data (80k+30k) improves performance
- Ensemble of training with different seeds expected to improve

<table>
<thead>
<tr>
<th>Method</th>
<th>B1</th>
<th>B4</th>
<th>Meteor</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinyals et al. [Vinyals et al., 2015]</td>
<td>-</td>
<td>27.7</td>
<td>23.7</td>
<td>85.5</td>
</tr>
<tr>
<td>Xu et al. [Xu et al., 2015], soft</td>
<td>70.9</td>
<td>24.3</td>
<td>23.9</td>
<td>-</td>
</tr>
<tr>
<td>Xu et al. [Xu et al., 2015], hard</td>
<td><strong>71.8</strong></td>
<td>25.0</td>
<td>23.0</td>
<td>-</td>
</tr>
<tr>
<td>Yang et al. [Yang et al., 2016]</td>
<td>-</td>
<td>29.0</td>
<td>23.7</td>
<td>88.6</td>
</tr>
<tr>
<td>Jin et al. [Jin et al., 2015]</td>
<td>69.7</td>
<td>28.2</td>
<td>23.5</td>
<td>83.8</td>
</tr>
<tr>
<td>Donahue et al. [Donahue et al., 2015]</td>
<td>71.1</td>
<td>30.0</td>
<td>24.2</td>
<td>89.6</td>
</tr>
<tr>
<td>Ranzato et al. [Ranzato et al., 2016]</td>
<td>-</td>
<td>29.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bengio et al. [Bengio et al., 2015]</td>
<td>-</td>
<td>30.6</td>
<td>24.3</td>
<td>92.1</td>
</tr>
<tr>
<td>Areas of Attention (ours)</td>
<td>70.8</td>
<td><strong>30.7</strong></td>
<td><strong>24.5</strong></td>
<td><strong>93.8</strong></td>
</tr>
<tr>
<td>AoA, data augmentation</td>
<td>72.1</td>
<td>31.1</td>
<td>25.0</td>
<td>95.6</td>
</tr>
</tbody>
</table>
More examples

a woman is sitting on a elephant with a baby elephant

a double decker bus is parked on the street

a woman is walking down the street with a bag
More examples

- A vase filled with purple flowers on a table
- A large jetliner flying over a city street
- A person on a snowboard in the snow
More examples

a man on a motorcycle is on the street

a man on a skateboard is going down a hill

a baseball player pitching a ball on a field
Areas of Attention in Image Captioning

Marco Pedersoli*, Thomas Lucas
Cordelia Schmid, Jakob Verbeek

INRIA Grenoble Rhône-Alpes, France
* Now at École de Technologie Supérieure Montreal, Canada
References I


