Neural Baby Talk
By Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh (Georgia Tech and Facebook AI)

In Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018

14 March 2019
Presentors: Soojin Lee, Austin Rothwell, and Shane Sims
CPSC 532S
UBC Department of Computer Science
Motivation

- Aid for the visually impaired
- Personal assistants
- Human robot interaction

https://www.organicauthority.com/image/t_share/MTU5MzMwMTE3MjM5MTIxN1A0/dog-toy-cc.jpg
Current Methods

“A dog is sitting on a couch with a toy”

Source: Neural Baby Talk by Jiasen Lu et. al.
Current Issues


2. A dog is sitting on a couch with a toy.

3. What is covering the windows? blinds  Human Attention  SAN-2 (Yang et al.)  HieCoAtt-Q (Lu et al.)  Judd et al.
Proposed Method

- Use object recognition to bolster image captioning
- 11% increase in average precision on COCO dataset in the last year
- Encourages visual grounding (i.e., associates named concepts to pixels in the image)

Source: Neural Baby Talk by Jiasen Lu et. al.
Proposed Method

Source: Neural Baby Talk by Jiasen Lu et. al.
Related Work

Slot Filling
- Farhadi et al., 2010
- Kulkarni et al., 2011

Encoder-Decoder
- Kiros, Salakhudinov, & Zemel, 2014
- Mao et al., 2015
- Xu et al., 2015
- Vinyals et al., 2015
- You et al., 2016
- Rennie et al., 2017
- Anderson et al., 2018

Grounded Language
- Johnson, Karpathy, & Fei-Fei, 2016
- Kazemzadeh et al., 2014
- Plummer et al., 2015
- Hu et al., 2016
- Luo & Shakhnarovich, 2017
High Level Methodology

Result: caption entities are grounded in image regions

A puppy with a tie is sitting at table with a cake
Slotted Caption Template Generation

1. Use CNN for object detection:
   a. Detected objects and their bounding boxes become candidate grounding regions
2. Use CNN to generate feature map of input image (as in Assignment 3)
3. Use RNN to generate caption template
   a. Initialize hidden state with CNN image features
   b. Training time: each input is ground truth caption word
   c. Inference time: each input is sampled from previous output
   d. Slots for visual words are generated using a pointer network [give blackbox description of Ptr-Net]
      i. Pointer networks map (point) tokens in the output sequence to tokens in the input sequence
Slotted Caption Template Generation

In:
- cabinet
- dog
- tie
- chair
- table
- cake
- frame

Out: A <region-2> with a <region-3> is sitting at <region-4> with a <region-5>

1) Obtain candidate grounding regions: cabinet, dog, tie, chair, table, cake, frame
2) Use a (Ptr-Net) RNN to generate caption template:
   a) Given candidate regions, Ptr-Net “points” from a token in the caption to an associated image region.

- **Note:** Ptr-Net can be used whenever we want to map output elements back to input elements exactly.

Image source: Neural Baby Talk by Jiasen Lu et. al.
Pointer Networks
(Vinyals O, Fortunato M, Jaitly N. NIPS 2015)

2) Use a (Ptr-Net) RNN to generate caption template

sequence-to-sequence (Assignment 3)
2) Use a (Ptr-Net) RNN generate caption template
In: - A <region-2> with a <region-3> is sitting at <region-4> with a <region-5> 
- cabinet, dog, tie, chair, table, cake, frame  (coarse names from object detector)

Out: A puppy with a tie is sitting at table with a cake
Objective

$$L(\theta) = - \sum_{t=1}^{T} \log \left( p(y_t^* | \tilde{r}, y_{1:t-1}^*) p(\tilde{r} | y_{1:t-1}^*) 1(y_t^* = y_{\text{gt}}) \right) + p \left( b_t^*, s_t^* | r_t, y_{1:t-1}^* \right) \left( \frac{1}{m} \sum_{i=1}^{m} p(r_t^i | y_{1:t-1}^*) 1(y_t^* = y_{\text{vis}}) \right)$$

**Text word probability**

Probability of predicting the correct text word, given the sentinel features, and the previous ground truth words and probability of generating the sentinel features given ground truth caption.

**Caption refinement prob.**

Probability of predicting correct plurality and fine-grained name given image region features and previous ground truth words in caption.

**Averaged target region probability**

Probability of produced the anchored image grounding region given previous characters of ground truth caption.

Training: minimize this cross-entropy loss
Evaluation

Datasets
Flickr30k: 31,783 images, 5 captions per image, 275,555 annotated bounding boxes
COCO: 164,062 images, 5 captions per image

Object category to words
For COCO dataset. (e.g., mapping <person> to [“child”, “baker”, ...])

Caption pre-processing
Caption truncation (if > 16 words)
Building vocabulary (9,587 words for COCO, 6,864 words for Flickr30k)
Evaluation

1. Standard Image Captioning
BLEU: precision
METEOR: averaged precision and recall
CIDEr: averaged cosine similarity
SPICE: defined over scene graphs

Flickr30k dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU1</th>
<th>BLEU4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard-Attention [16]</td>
<td>66.9</td>
<td>19.9</td>
<td>18.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ATT-FCN [50]</td>
<td>64.7</td>
<td>23.0</td>
<td>18.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Adaptive [27]</td>
<td>67.7</td>
<td>25.1</td>
<td>20.4</td>
<td>53.1</td>
<td>14.5</td>
</tr>
<tr>
<td>NBT</td>
<td>69.0</td>
<td>27.1</td>
<td>21.7</td>
<td>57.5</td>
<td>15.6</td>
</tr>
<tr>
<td>NBT*oracle</td>
<td>72.0</td>
<td>28.5</td>
<td>23.1</td>
<td>64.8</td>
<td>19.6</td>
</tr>
</tbody>
</table>

COCO dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU1</th>
<th>BLEU4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive [27]</td>
<td>74.2</td>
<td>32.5</td>
<td>26.6</td>
<td>108.5</td>
<td>19.5</td>
</tr>
<tr>
<td>Att2in [39]</td>
<td>-</td>
<td>31.3</td>
<td>26.0</td>
<td>101.3</td>
<td>-</td>
</tr>
<tr>
<td>Up-Down [3]</td>
<td>74.5</td>
<td>33.4</td>
<td>26.1</td>
<td>105.4</td>
<td>19.2</td>
</tr>
<tr>
<td>Att2in* [39]</td>
<td>-</td>
<td>33.3</td>
<td>26.3</td>
<td>111.4</td>
<td>-</td>
</tr>
<tr>
<td>Up-Down† [3]</td>
<td>79.8</td>
<td>36.3</td>
<td>27.7</td>
<td>120.1</td>
<td>21.4</td>
</tr>
<tr>
<td>NBT</td>
<td>75.5</td>
<td>34.7</td>
<td>27.1</td>
<td>107.2</td>
<td>20.1</td>
</tr>
<tr>
<td>NBT*oracle</td>
<td>75.9</td>
<td>34.9</td>
<td>27.4</td>
<td>108.9</td>
<td>20.4</td>
</tr>
</tbody>
</table>

source: Neural Baby Talk by Jiasen Lu et al.
Evaluation

COCO

**Success**

- A dog is laying in the grass with a Frisbee.
- A bride and groom cutting a cake together.
- A little girl holding a cat in her hand.

**Failure**

- A woman sitting on a boat in the water.

Flickr30k

**Success**

- A cat is standing on a sign that says “UNK”.
- A young boy with blond-hair and a blue shirt is eating a chocolate.
- A band is performing on a stage.

**Failure**

- Two people are sitting on a boat in the water.

* Different colours show a correspondence between the visual words and grounding regions.
* Grey regions are the proposals not selected in the captions.

source: Neural Baby Talk by Jiasen Lu et. al.
Evaluation

2. Robust Image Captioning
To evaluate image captioning for novel scene compositions

Robust-COCO split
- Distribution of co-occurring objects in train data is different from test data
- Calculate the co-occurrence statistics for 80 object categories
- Sufficient examples from each category in train set
- Novel compositions (pairs) of categories in test set

Accuracy
- Whether or not a generated caption includes the new object combination
- 100% accuracy for at least one mention of the novel category pair
Evaluation

Robust-COCO split

: worse (2~3 points drop) performance for all models

COCO dataset with Robust split

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att2in [39]</td>
<td>31.5</td>
<td>24.6</td>
<td>90.6</td>
<td>17.7</td>
<td>39.0</td>
</tr>
<tr>
<td>Up-Down [3]</td>
<td>31.6</td>
<td>25.0</td>
<td>92.0</td>
<td>18.1</td>
<td>39.7</td>
</tr>
<tr>
<td>NBT</td>
<td>31.7</td>
<td>25.2</td>
<td>94.1</td>
<td>18.3</td>
<td>42.4</td>
</tr>
<tr>
<td>NBT_oracle</td>
<td>31.9</td>
<td>25.5</td>
<td>95.5</td>
<td>18.7</td>
<td>45.7</td>
</tr>
</tbody>
</table>

Success

A cat laying on the floor next to a remote control.
A man sitting on a bench next to a bird.
A dog is standing on a skateboard in the grass.

Failure

A bird sitting on a branch in a tree.

Image source: Neural Baby Talk by Jiasen Lu et. al.
Evaluation

3. Novel Object Captioning
Excludes all the image-sentence pairs that contain at least one of the eight objects in COCO ("bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra")

Test set is split into in-domain and out-of-domain subsets

F1 score
: checks if the excluded object is correctly mentioned in the generated caption
## Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>bottle</th>
<th>bus</th>
<th>couch</th>
<th>microwave</th>
<th>pizza</th>
<th>racket</th>
<th>suitcase</th>
<th>zebra</th>
<th>Avg</th>
<th>SPICE</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>In-Domain Test Data</th>
<th>SPICE</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCC [4]</td>
<td>4.6</td>
<td>29.8</td>
<td>45.9</td>
<td>28.1</td>
<td>64.6</td>
<td>52.2</td>
<td>13.2</td>
<td>79.9</td>
<td>39.8</td>
<td>13.4</td>
<td>21.0</td>
<td>59.1</td>
<td>15.9</td>
<td>23.0</td>
<td>77.2</td>
<td></td>
</tr>
<tr>
<td>NOC [43]</td>
<td>17.8</td>
<td>68.8</td>
<td>25.6</td>
<td>24.7</td>
<td>69.3</td>
<td>68.1</td>
<td>39.9</td>
<td>89.0</td>
<td>49.1</td>
<td>-</td>
<td>21.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C-LSTM [49]</td>
<td>29.7</td>
<td>74.4</td>
<td>38.8</td>
<td>27.8</td>
<td>68.2</td>
<td>70.3</td>
<td>44.8</td>
<td>91.4</td>
<td>55.7</td>
<td>-</td>
<td>23.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Base+T4 [2]</td>
<td>16.3</td>
<td>67.8</td>
<td>48.2</td>
<td>29.7</td>
<td>77.2</td>
<td>57.1</td>
<td>49.9</td>
<td>85.7</td>
<td>54.0</td>
<td>15.9</td>
<td>23.3</td>
<td>77.9</td>
<td>18.0</td>
<td>24.5</td>
<td>86.3</td>
<td></td>
</tr>
<tr>
<td>NBT*+G</td>
<td>7.1</td>
<td>73.7</td>
<td>34.4</td>
<td>61.9</td>
<td>59.9</td>
<td>20.2</td>
<td>42.3</td>
<td>88.5</td>
<td>48.5</td>
<td>15.7</td>
<td>22.8</td>
<td>77.0</td>
<td>17.5</td>
<td>24.3</td>
<td>87.4</td>
<td></td>
</tr>
<tr>
<td>NBT\†+G</td>
<td>14.0</td>
<td>74.8</td>
<td>42.8</td>
<td>63.7</td>
<td>74.4</td>
<td>19.0</td>
<td>44.5</td>
<td>92.0</td>
<td>53.2</td>
<td>16.6</td>
<td>23.9</td>
<td>84.0</td>
<td>18.4</td>
<td>25.3</td>
<td>94.0</td>
<td></td>
</tr>
<tr>
<td>NBT\†+T1</td>
<td>36.2</td>
<td>77.7</td>
<td>43.9</td>
<td>65.8</td>
<td>70.3</td>
<td>19.8</td>
<td>51.2</td>
<td>93.7</td>
<td>57.3</td>
<td>16.7</td>
<td>23.9</td>
<td>85.7</td>
<td>18.4</td>
<td>25.5</td>
<td>95.2</td>
<td></td>
</tr>
<tr>
<td>NBT\†+T2</td>
<td><strong>38.3</strong></td>
<td><strong>80.0</strong></td>
<td><strong>54.0</strong></td>
<td><strong>70.3</strong></td>
<td><strong>81.1</strong></td>
<td><strong>74.8</strong></td>
<td><strong>67.8</strong></td>
<td><strong>96.6</strong></td>
<td><strong>70.3</strong></td>
<td><strong>17.4</strong></td>
<td><strong>24.1</strong></td>
<td><strong>86.0</strong></td>
<td>18.0</td>
<td>25.0</td>
<td>92.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Evaluation of captions generated using the proposed method. G means greedy decoding, and T1—2 means using constrained beam search [3] with 1—2 top detected concepts. * is the result using VGG-16 [41] and † is the result using ResNet-101.
Evaluation

Success
A zebra that is standing in the dirt.
A little girl wearing a helmet and holding a tennis racket.
A woman standing in front of a red bus.

Failure
A plate of food with a bottle and a cup of beer.

Image source: Neural Baby Talk by Jiasen Lu et. al.
Conclusion

A novel image captioning framework
: natural language + grounded in detected objects

Two-stage approach
: 1) generate hybrid template
: 2) fills the slots with categories recognized by object detector

NBT outperforms the state-of-art models on standard, robust, and novel object captioning
Limitations

1. Caption template generation and slot filling parts are trained end-to-end. Why not add the CNN to the model for fine tuning during training?

   a. Authors note a strength of NBT is that you can swap object detectors to suit problem.

   b. Would have been interesting to see difference in evaluation metrics.

2. Not clear how useful fined grained category name assignment is.

   a. In general, authors could have compared performance with and without this sub-model.

Table source: Neural Baby Talk by Jiasen Lu et. al.

Image source: CPSC 532S lecture slides
Possible Extensions

1. Compare NBT performance with and without end-to-end training of the CNN
2. Perform object detection in model that maximizes accuracy-specificity tradeoff.
   a. Critical for real world tasks that authors use as their motivation (ex: helping visually impaired)
      i. Likely harm COCO evaluation metrics -> points to the need for a new metric for this real world task
   b. Accomplish this by:
      i. Pretrain CNN pre-trained on ImageNet
      ii. Classify according to semantic hierarchy (eliminating part of slot filling model)
         1. Already organized by WordNet hierarchy
         2. See my course project for doing this with modern CNN architectures
Thank you.

Questions?
Back-up Slides
Compared Models

Hard-attention
: Attention-based image caption
(“soft” or “hard” attention)

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. “soft” (top row) vs “hard” (bottom row) attention. (Note that both models generated the same captions in this example.)
Compared Models

ATT-FCN
: Attention-based image caption (semantically important regions)

You et al. (2016). Image Captioning with Semantic Attention
Compared Models

Adaptive

Attention-based model with a visual sentinel ("when" to look at + "which" region)

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.