Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions

Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy
INTRODUCTION

Early captioning approaches:
• Global image feature vector

Attention-based approaches:
• Weakly interpretable (through attention)
• Not controllable.
  • We can’t decide which regions get processed
  • No control over the generation process.

Show, control and tell
• Controllable via regions
  • A sequence (ordered)
  • A set (unordered)

CONTROLLABLE IMAGE CAPTIONING

A step back: what is the right abstraction level for generating sentences?

- So far: word level
- Ours: noun chunks

**Noun chunk**: a noun grouped together with its modifiers.

Easy to obtain via NLP tools (dependency tree), and...

Noun chunks can be mapped to image regions

CONTROLLABLE IMAGE CAPTIONING

- A sentence is a sequence of noun chunks, eventually associated with an image region
- Different captions $\rightarrow$ different chunks $\rightarrow$ different regions
- Captions differ in terms of:
  - The set of selected regions
  - The order in which they are described
  - Their mapping to noun chunks (linguistic abilities of the annotator) $\rightarrow$ Learned!

CONTROLLABLE IMAGE CAPTIONING

CONTROLLABLE IMAGE CAPTIONING

- Language model takes as input a sequence of regions
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it’s done with the generation of chunk, it moves to the next region in the sequence

Language model takes as input a sequence of regions

Switches between one region and the next one via a learned chunk-shifting gate

When it’s done with the generation of chunk, it moves to the next region in the sequence

• Language model takes as input a sequence of regions
• Switches between one region and the next one via a learned chunk-shifting gate
  • When it’s done with the generation of chunk, it moves to the next region in the sequence

CONTROLLABLE IMAGE CAPTIONING

- Language model takes as input a sequence of regions
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it’s done with the generation of chunk, it moves to the next region in the sequence

Train on GT words and shifting gate values (obtained via NLP)

$L(\theta) = - \sum_{t=1}^{T} \left( \log p(y_t^*| r_{1:t}^*, y_{1:t-1}^*) + g_t^* \log p(g_t = 1| r_{1:t}^*, y_{1:t-1}^*) + (1 - g_t^*) \log (1 - p(g_t = 1| r_{1:t}^*, y_{1:t-1}^*)) \right)$

CONTROLLABLE IMAGE CAPTIONING

- Train on GT words and shifting gate values (obtained via NLP)
- then, finetune using Reinforcement Learning
  - CIDEr wrt GT caption (caption quality)
  - Plus, use the alignment between the predicted and GT chunks as reward (Needleman-Wunsch algorithm)

CONTROLLABLE IMAGE CAPTIONING

• What if we have an unordered set as input?

• We can learn a network to do the sorting! → SINKHORN NETWORK
  
  • Approximates a derivable permutation matrix
  
  • Train on real data, then use the Hungarian to get the true permutation matrix.

CONTROLLABLE IMAGE CAPTIONING

- Controllability via a set of regions
  - 75.5% intersection-over-union with GT chunks!
  - Adds more diversity than methods tailored for diversity 😊

<table>
<thead>
<tr>
<th>Method</th>
<th>B-4</th>
<th>M</th>
<th>R</th>
<th>C</th>
<th>S</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC-2K† [36]</td>
<td>12.5</td>
<td>18.5</td>
<td>39.6</td>
<td>116.5</td>
<td>26.6</td>
<td>61.0</td>
</tr>
<tr>
<td>Up-Down† [3]</td>
<td>14.4</td>
<td>20.0</td>
<td>42.2</td>
<td>132.8</td>
<td>29.7</td>
<td>63.2</td>
</tr>
<tr>
<td>Neural Baby Talk† [24]</td>
<td>13.1</td>
<td>19.2</td>
<td>40.5</td>
<td>119.1</td>
<td>29.2</td>
<td>62.6</td>
</tr>
<tr>
<td>Controllable LSTM</td>
<td>12.9</td>
<td>19.3</td>
<td>41.3</td>
<td>123.4</td>
<td>28.7</td>
<td>64.2</td>
</tr>
<tr>
<td>Controllable Up-Down</td>
<td>18.1</td>
<td>23.6</td>
<td>48.4</td>
<td>170.5</td>
<td>40.4</td>
<td>71.6</td>
</tr>
<tr>
<td>Ours w/ single sentinel</td>
<td>17.4</td>
<td>23.6</td>
<td>48.4</td>
<td>168.4</td>
<td>43.7</td>
<td>75.4</td>
</tr>
<tr>
<td>Ours w/o visual sentinel</td>
<td>17.6</td>
<td>23.4</td>
<td>48.5</td>
<td>168.9</td>
<td>43.6</td>
<td>75.3</td>
</tr>
<tr>
<td>Ours</td>
<td>18.0</td>
<td>23.8</td>
<td>48.9</td>
<td>173.3</td>
<td>44.1</td>
<td>75.5</td>
</tr>
<tr>
<td>AG-CVAE [44]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.1</td>
</tr>
<tr>
<td>POS [8]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44.9</td>
</tr>
<tr>
<td>Ours</td>
<td>20</td>
<td>44.8</td>
<td>36.6</td>
<td>68.9</td>
<td>156.5</td>
<td>30.9</td>
</tr>
</tbody>
</table>

CONTROLLABLE IMAGE CAPTIONING

Results when Controlling with a sequence of regions

A man sitting at a desk with a computer and a man holding a camera.

A man sitting at a desk with a computer.

CONTROLLABLE IMAGE CAPTIONING

Results when Controlling with a sequence of regions

A giraffe standing in front of a zebra in a field.

A zebra standing next to a giraffe in a field.

CONTROLLABLE IMAGE CAPTIONING

Results when Controlling with a set of regions

A dog holding a frisbee in its mouth.  
A dog standing in the grass with a frisbee in its mouth.

A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

CONTROLLABLE IMAGE CAPTIONING

For training and evaluation, we collect COCO-Entities
→ more than 120,000 images

- COCO with noun chunks associated to regions
- Semi-automatically annotated

CONTROLLABLE IMAGE CAPTIONING

Get your hands dirty

Dataset, code, pre-trained models are available at

https://github.com/aimagelab/show-control-and-tell

CONTROLLABLE IMAGE CAPTIONING

Want to see more?

Rita Cucchiara’s keynote @ EPIC Workshop
Monday, 13:30

Poster #195
Wednesday, 15:20 – 18:00