TAB-VCR: Tags and Attributes based VCR Baselines

Jingxiang Lin, Unnat Jain, Alexander Schwing
University of Illinois at Urbana-Champaign

NeurIPS 2019
Subtask 1
Question Answering

Question: How did [0] and [1] get here?

Answers:

a) They traveled in a cart.
b) [0, 1] got [1] released from jail.
c) [0, 1] took the stairs to get up there.
d) They both got splashed.

[Correct]
How did \([0]\) and \([1]\) get here?

- a) They traveled in a cart.
- b) \([0, 1]\) got \([1]\) released from jail.
- c) \([0, 1]\) took the stairs to get up there.
- d) They both got splashed.
Visual Commonsense Reasoning (VCR)

Subtask 2
Answer
[0] person

Question & correct answer
How did [0, 1] get here?
They traveled in a cart.

Correct]
[Tags]

Rationales
a) Presumably they came here to get something from the store.
b) They are at a market and [0]'s clothes look like the locals in the background.
c) [1] is holding a bag which people often use to carry groceries.
d) The cart beside them is likely their mode of transportation.
Contributions

1. **Simple base network** with fewer than half the trainable parameters
2. **Leverage attribute information** about objects
3. Improve image-text grounding by **adding new tags**
1. Simple base network

[0, 1] got [1] released from jail.
1. Simple base network

[0, 1] got [1] released from jail.
1. Simple base network

[0, 1] got [1] released from jail.

BERT

Object detections

Faster RCNN

Faster RCNN

Casey

jail
1. Simple base network

[0, 1] got [1] released from jail.

BERT

Object detections

Faster RCNN

Faster RCNN

Casey

jail
1. Simple base network

[0, 1] got [1] released from jail.

BERT

Image + Sentence Embedding

Pooling

LSTM

Faster RCNN

Object detections

Casey

jail
1. Simple base network

[0, 1] got [1] released from jail.
2. Leverage attribute information
2. Leverage attribute information
Q: How did [0] and [1] get here?
   a) They traveled in a cart.
   b) [0, 1] got [1] released from jail.
   c) [0, 1] took the stairs to get up there.
   d) They both got splashed.

Q: How did [0, 1] get here
A: They traveled in a cart
   a) Presumably they came here to get something from the store.
   b) They are at a market and [0]'s clothes look like the locals in the background.
   c) [1] is holding a bag which people often use to carry groceries.
   d) The cart beside them is likely their mode of transportation.
3. Adding *new tags*

Q: How did [0] and [1] get here?

a) They traveled in a [cart].
b) [0, 1] got [1] released from jail.
c) [0, 1] took the stairs to get up there.
d) They both got splashed.

Q: How did [0, 1] get here

A: They traveled in a cart

a) Presumably they came here to get something from the store.
b) They are at a market and [0]'s clothes look like the locals in the background.
c) [1] is holding a bag which people often use to carry groceries.
d) The [cart] beside them is likely their mode of transportation.
## Results

<table>
<thead>
<tr>
<th>Vision Backbone</th>
<th>Subtask 1 Q → A</th>
<th>Subtask 2 QA → R</th>
<th>Trainable Params (Mn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2C (Zellers et al.)</td>
<td>Resnet 50</td>
<td>63.8</td>
<td>67.2</td>
</tr>
<tr>
<td>(1) Base (ours)</td>
<td>Resnet 101</td>
<td>67.50</td>
<td>69.75</td>
</tr>
<tr>
<td>(2) Base + attributes (ours)</td>
<td>Resnet 101</td>
<td>69.51</td>
<td>71.57</td>
</tr>
<tr>
<td>(3) Base + attributes + new tags (TAB-VCR)</td>
<td>Resnet 101</td>
<td><strong>69.89</strong></td>
<td><strong>72.15</strong></td>
</tr>
</tbody>
</table>