TAB-VCR: Tags and Attributes based VCR Baselines

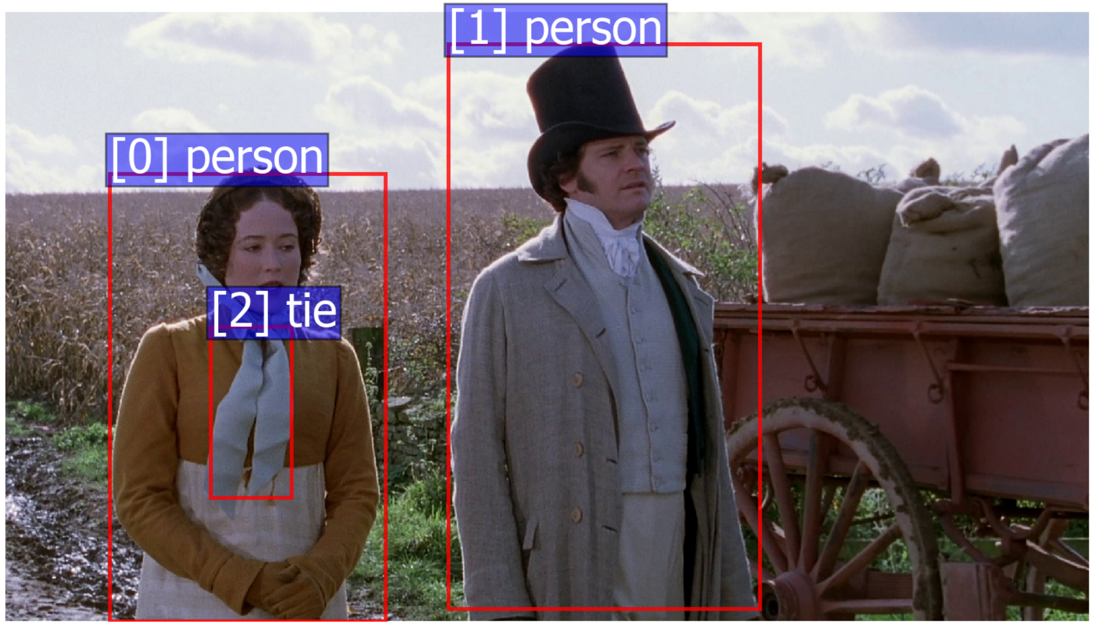
Jingxiang Lin, Unnat Jain, Alexander Schwing

University of Illinois at Urbana-Champaign

NeurIPS 2019

Visual Commonsense Reasoning (VCR)

Subtask 1
Question
Answering



Question

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

[Correct]

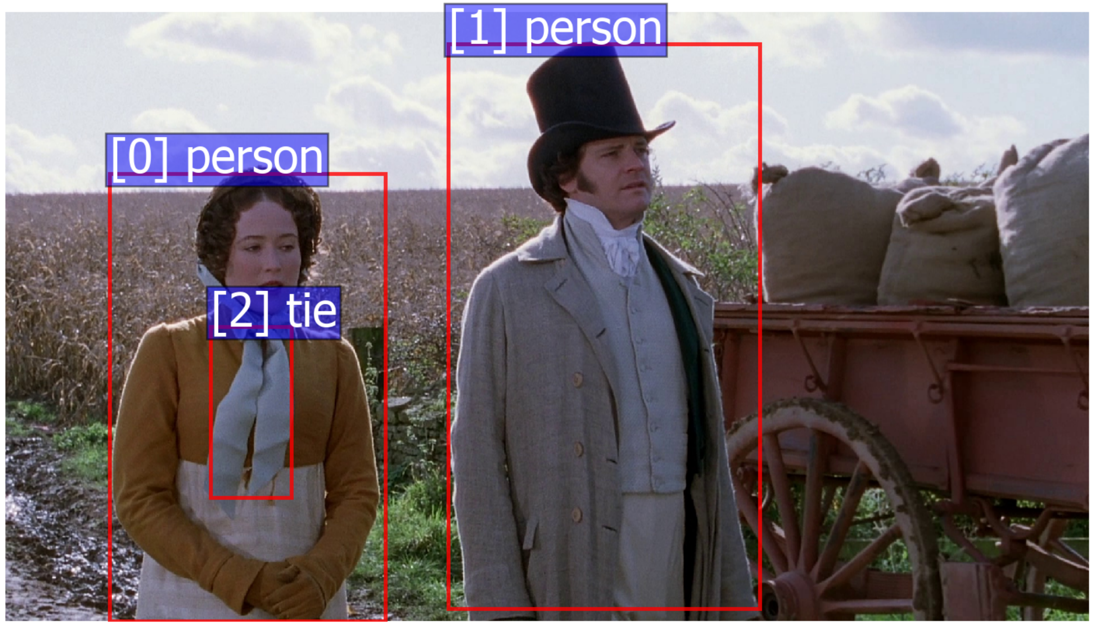
Answers

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.

Zellers et al. CVPR 2019
From Recognition to Cognition: Visual Commonsense Reasoning

Visual Commonsense Reasoning (VCR)

Subtask 1
Question Answering



[Correct] [Tags]

Question

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

Answers

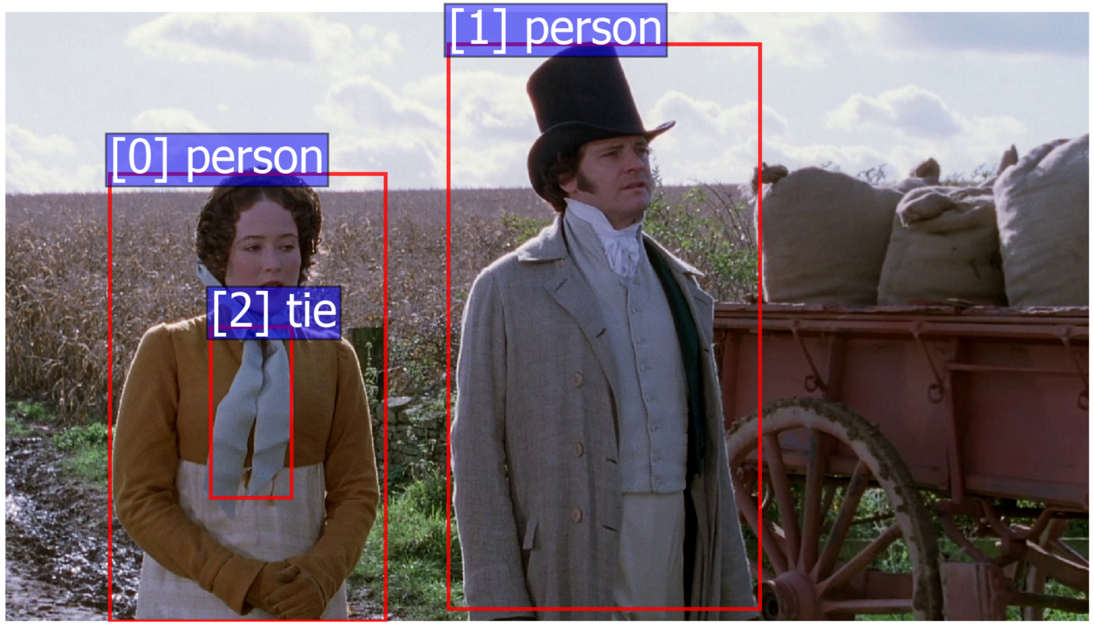
a) They traveled in a cart.
b) [0, 1] got [1] released from jail.
c) [0, 1] took the stairs to get up there.
d) They both got splashed.

Zellers et al. CVPR 2019
From Recognition to Cognition: Visual Commonsense Reasoning

Visual Commonsense Reasoning (VCR)

Subtask 2

Answer Justification



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.

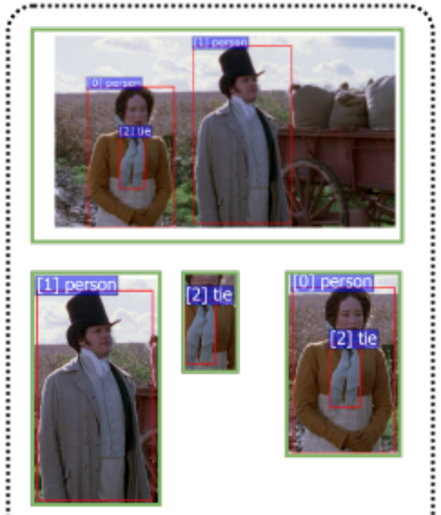
Zellers et al. CVPR 2019
From Recognition to Cognition: Visual Commonsense Reasoning

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

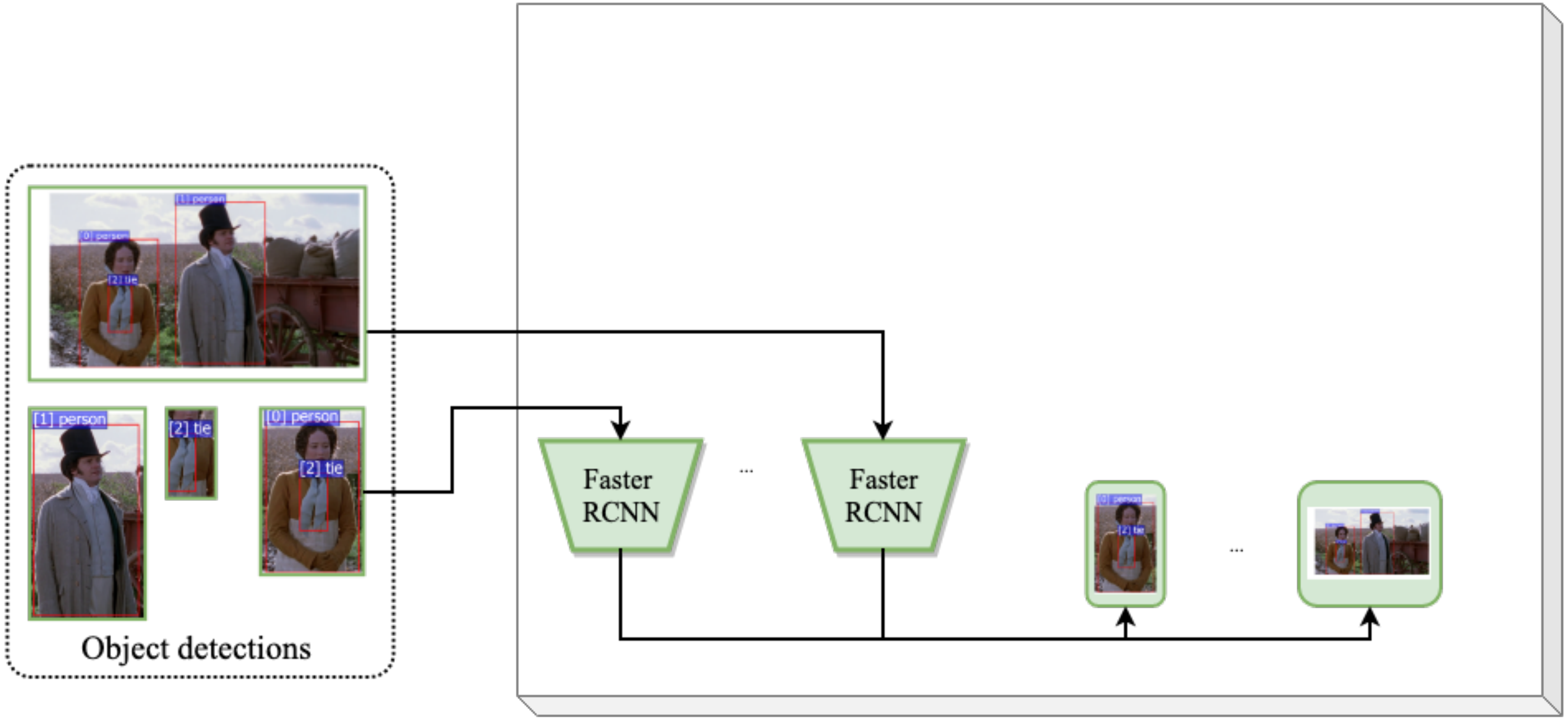
[0, 1] got [1] released from jail.



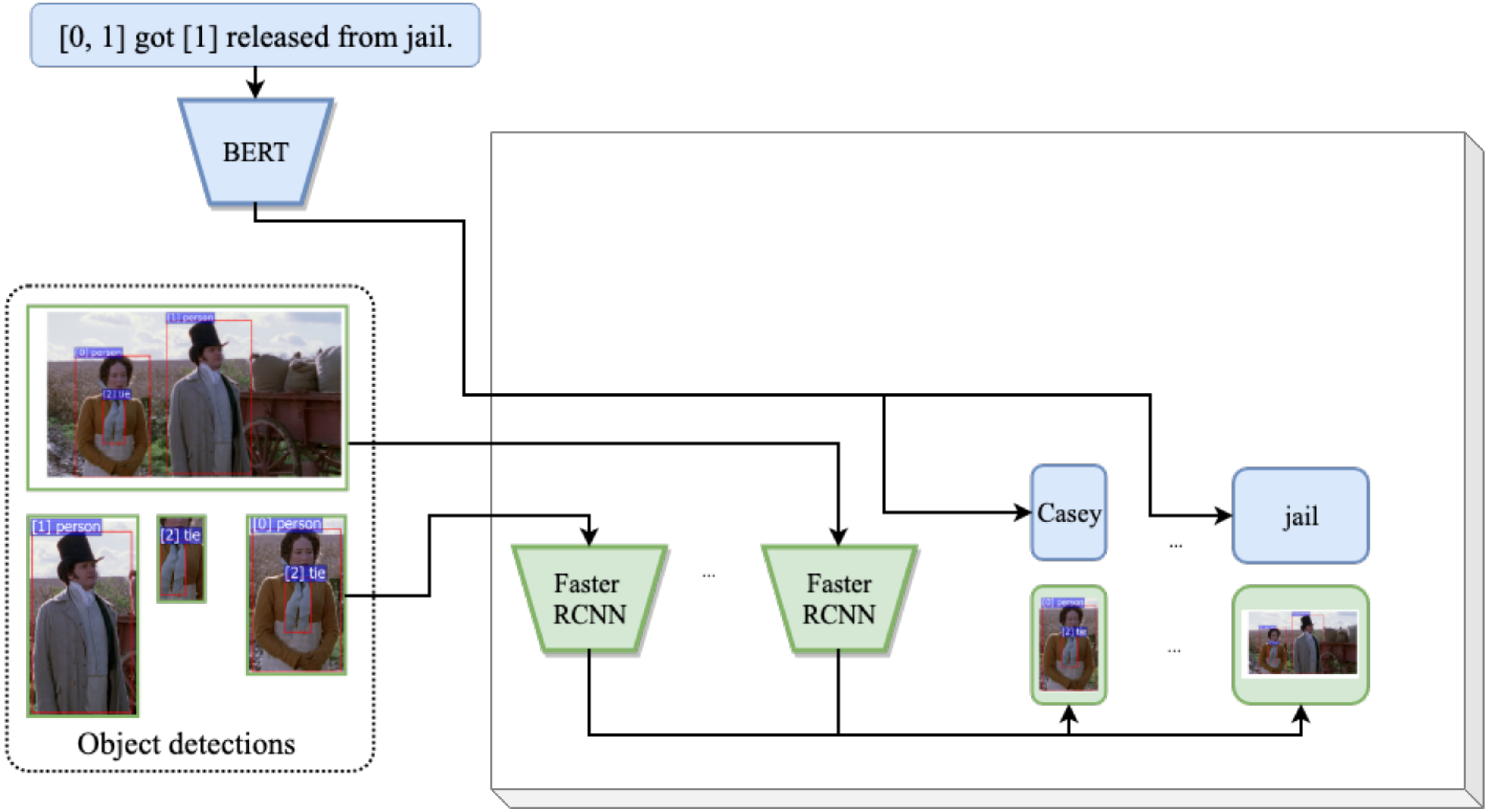
Object detections

1. Simple base network

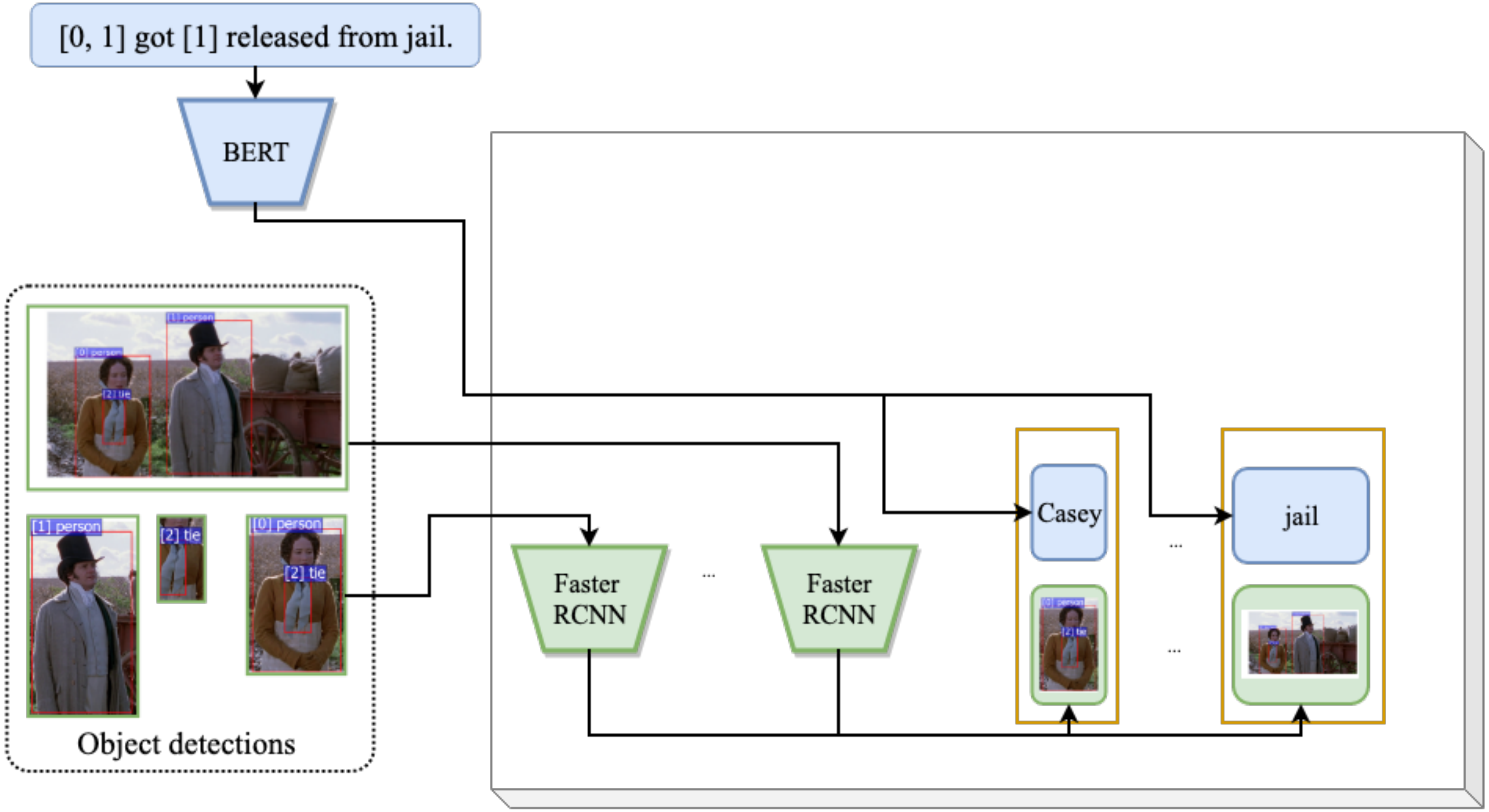
[0, 1] got [1] released from jail.



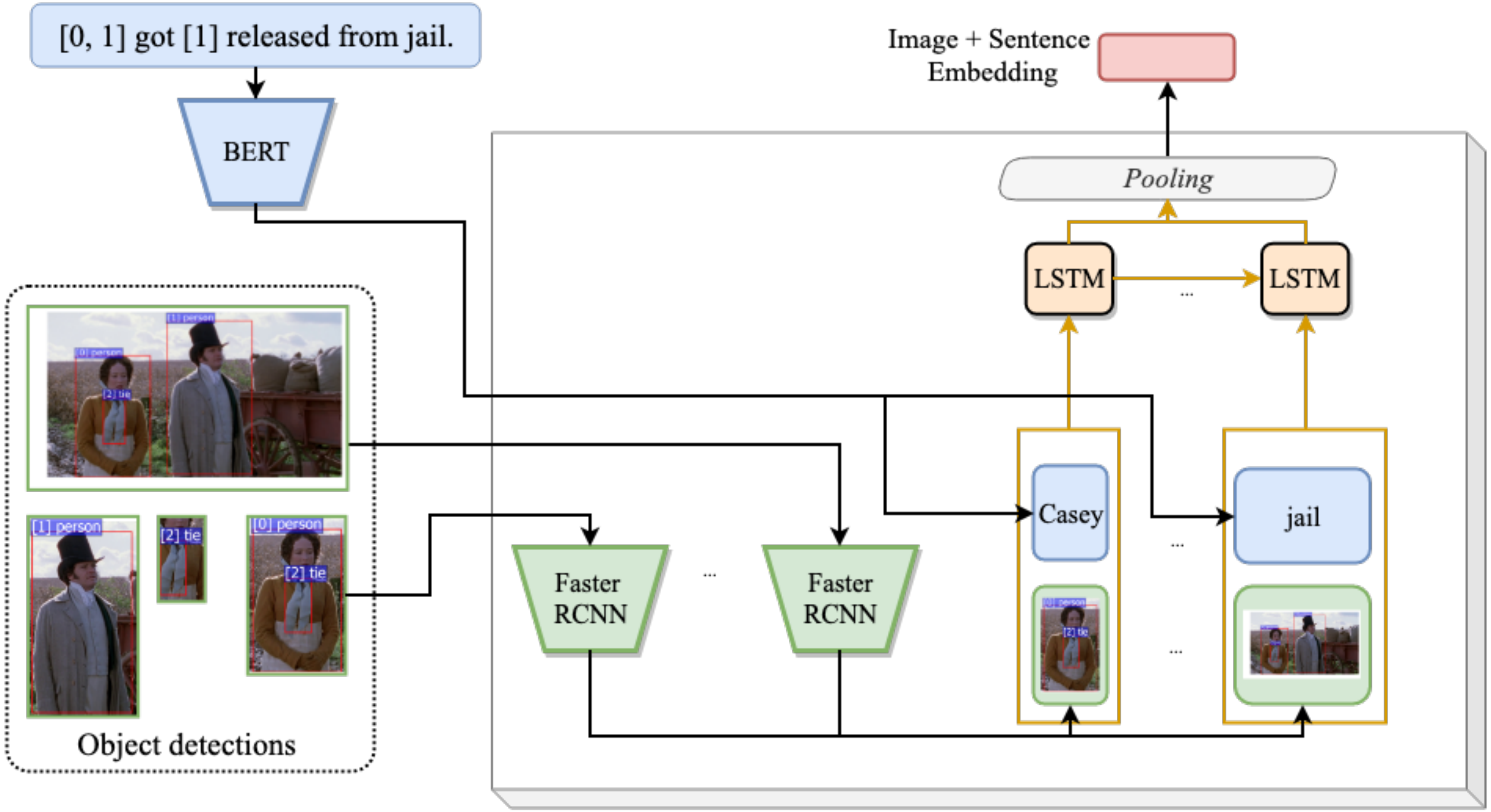
1. Simple base network



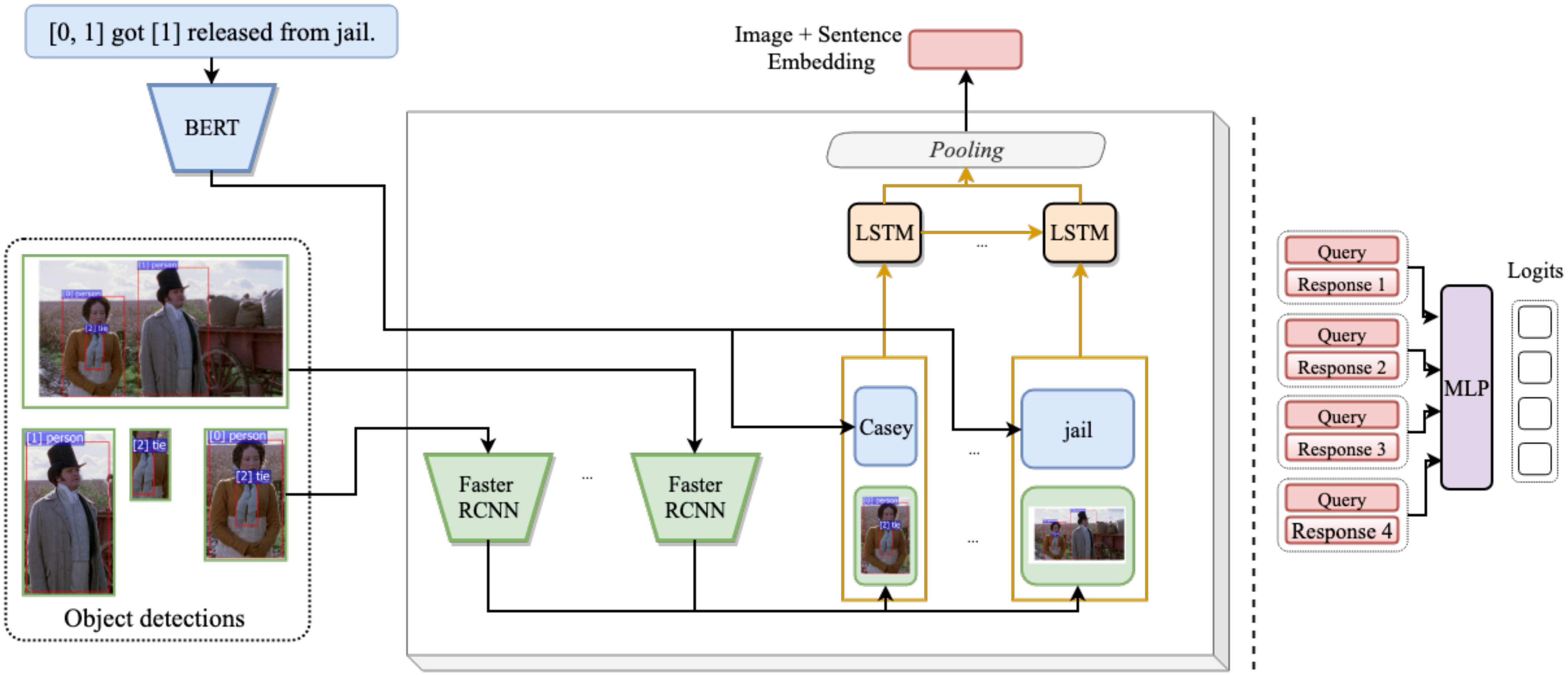
1. Simple base network



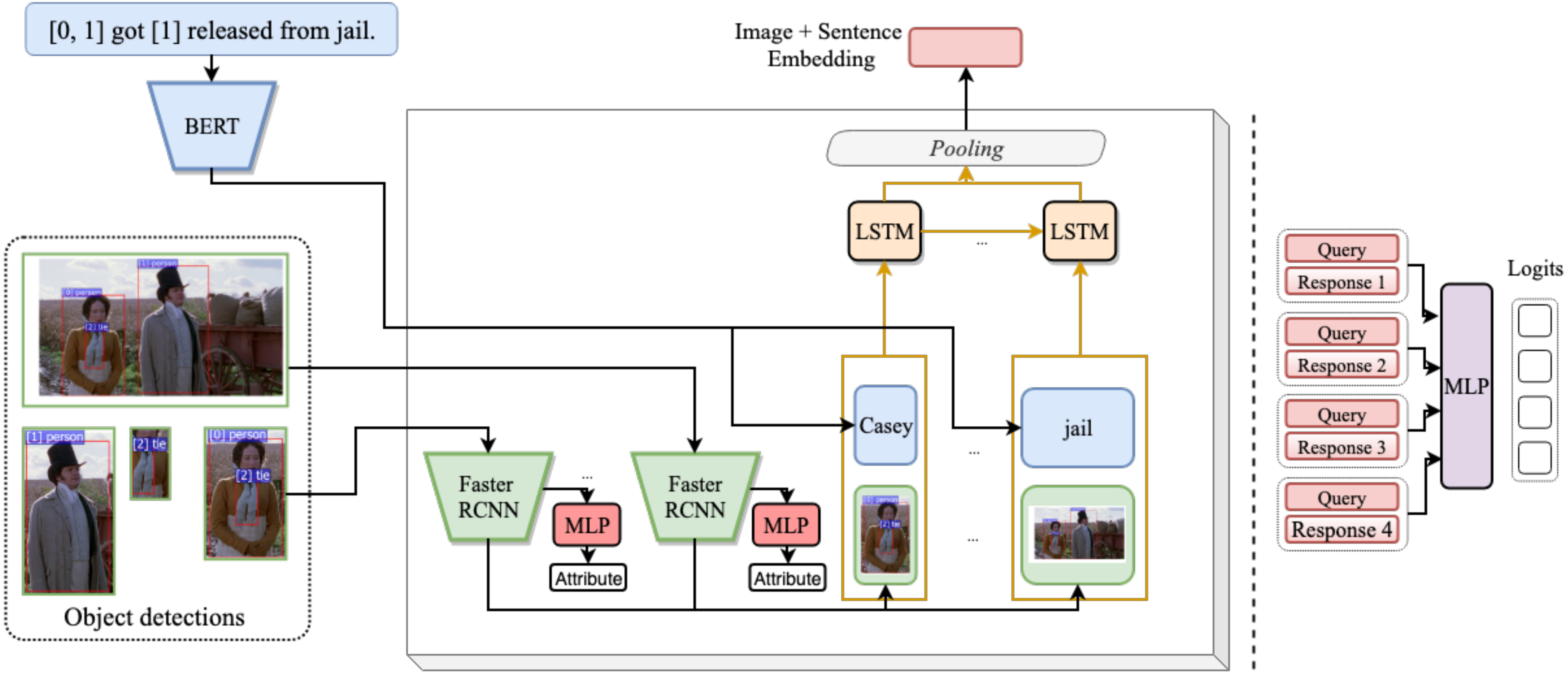
1. Simple base network



1. Simple base network



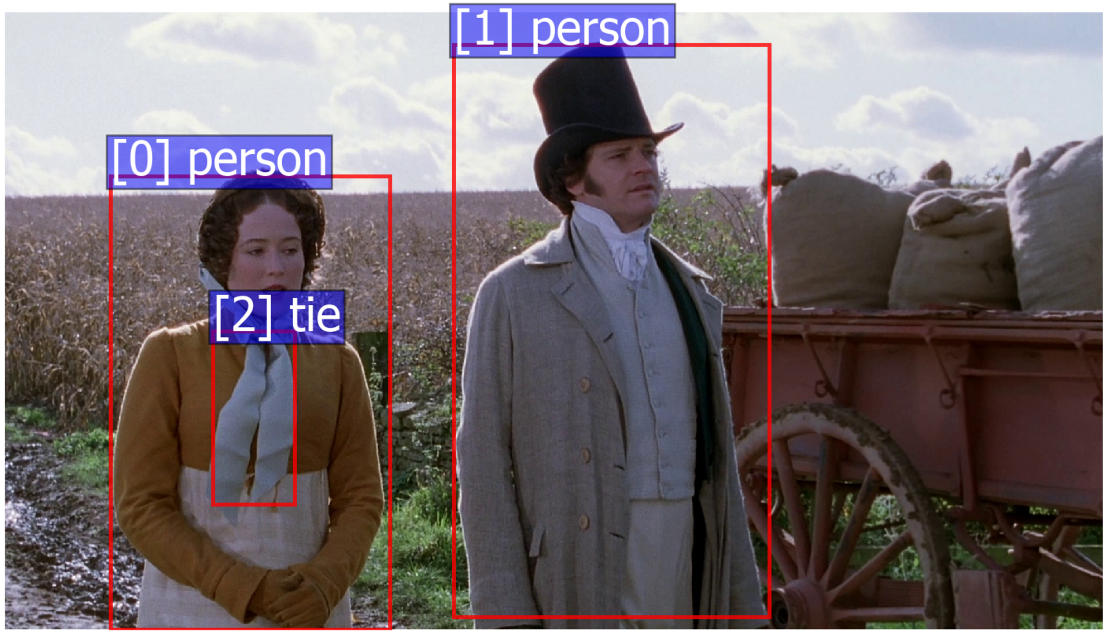
2. Leverage attribute information

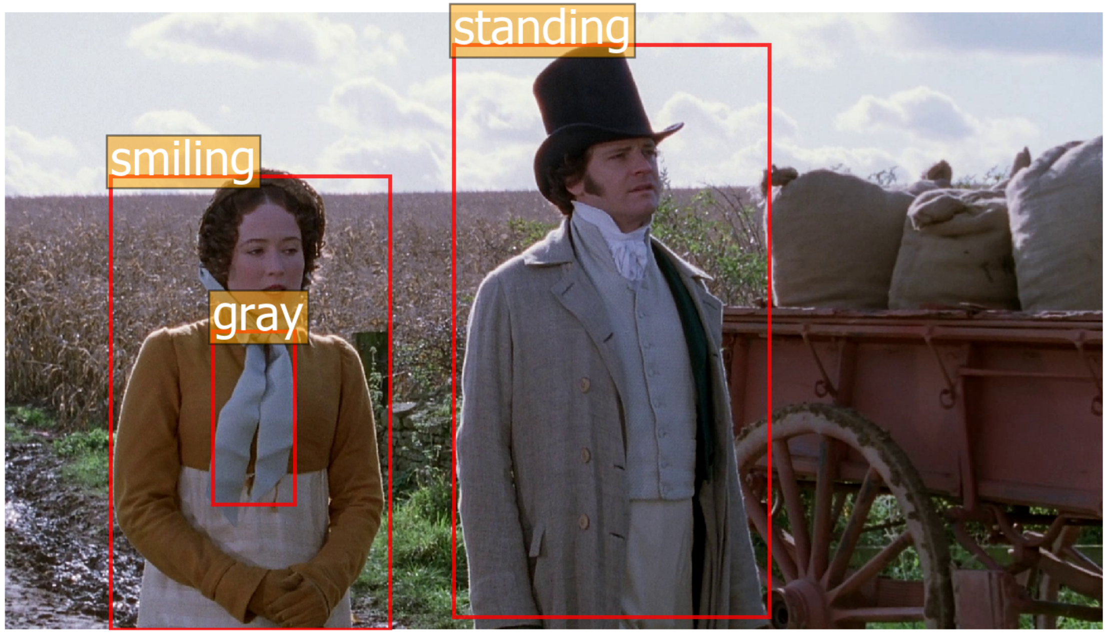


Anderson et al. CVPR 2018

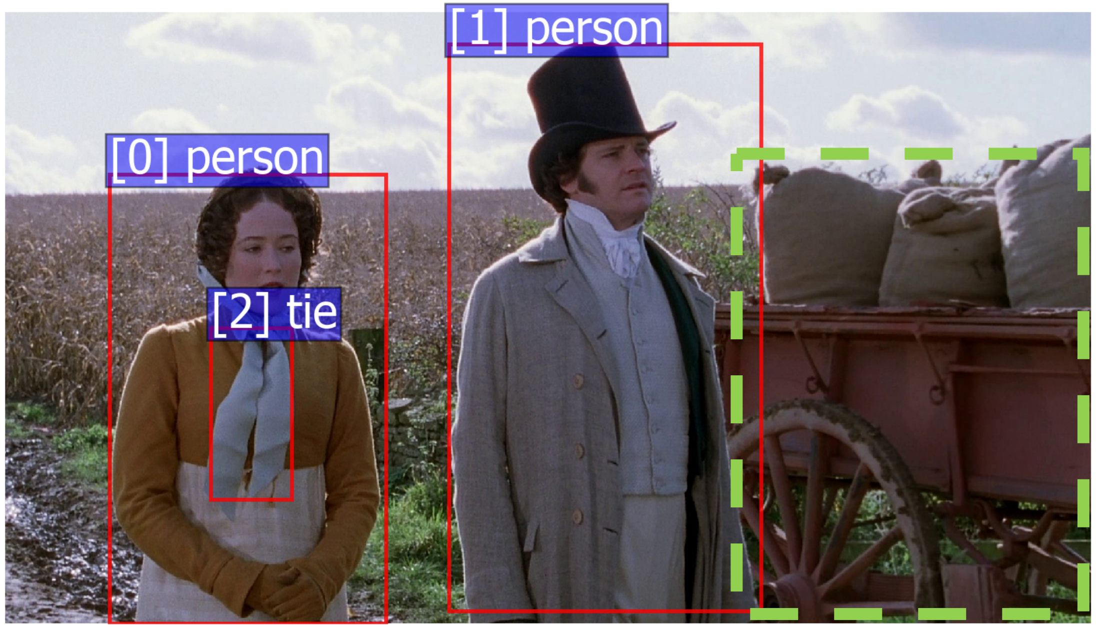
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

2. Leverage attribute information





3. Adding *new tags*



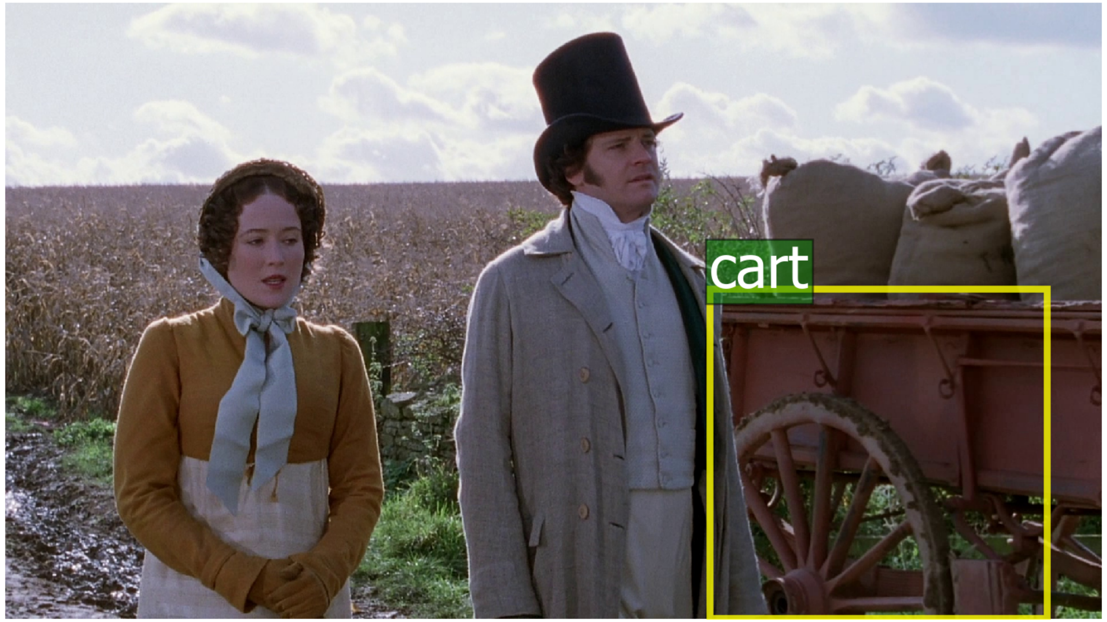
Q: How did [0] and [1] get here?

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.

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

	Vision Backbone	Subtask 1 Q → A	Subtask 2 QA → R	Trainable Params (Mn)
R2C (Zellers et al.)	Resnet 50	63.8	67.2	26.8
(1) Base (ours)	Resnet 101	67.50	69.75	4.9
(2) Base + *attributes* (ours)	Resnet 101	69.51	71.57	4.7
(3) Base + attributes + *new tags* (**TAB-VCR**)	Resnet 101	**69.89**	**72.15**	**4.7**