From Recognition to Cognition: Visual Commonsense Reasoning

Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi.
CVPR 2019

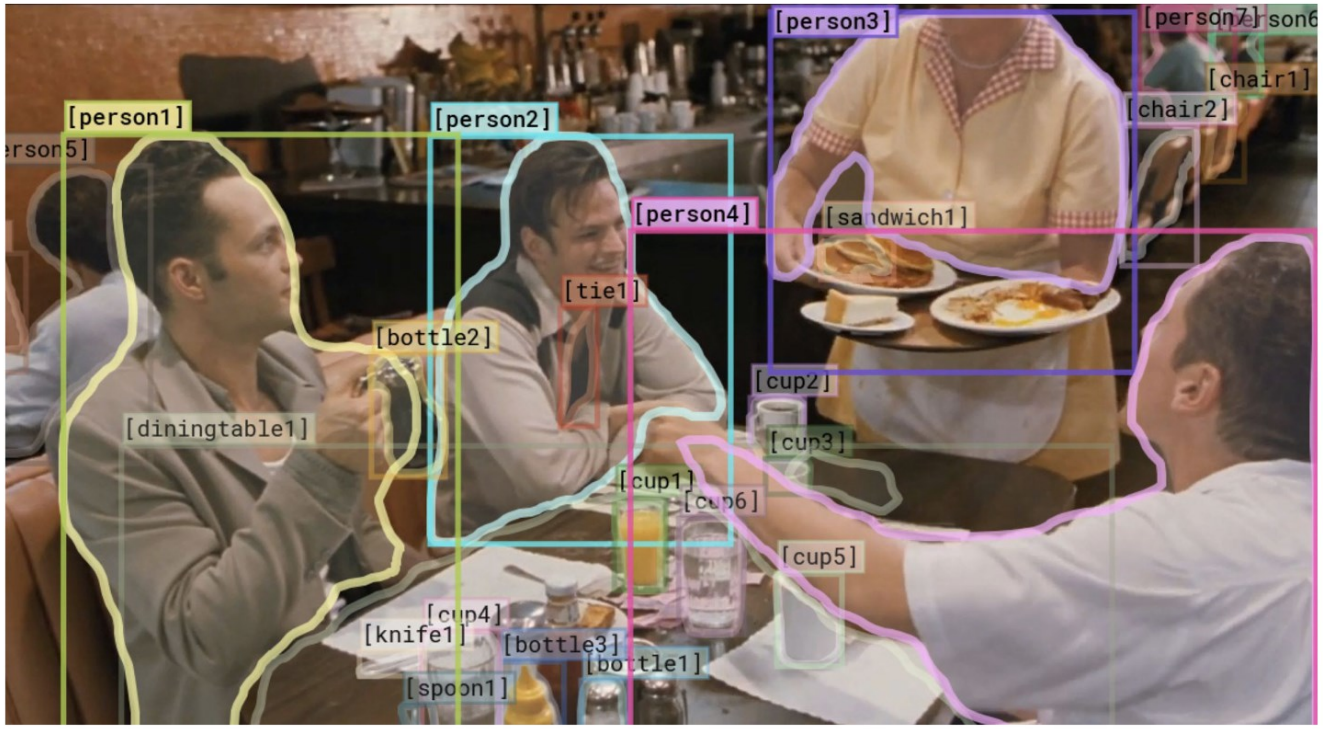
David Semedo df.semedo@fct.unl.pt

NOVA Search Reading group

Credits. Heavily based on original authors' presentation: https://www.cs.utexas.edu/~mooney/gnlp/slides/from_recognition_to_cognition.pdf

The Problem

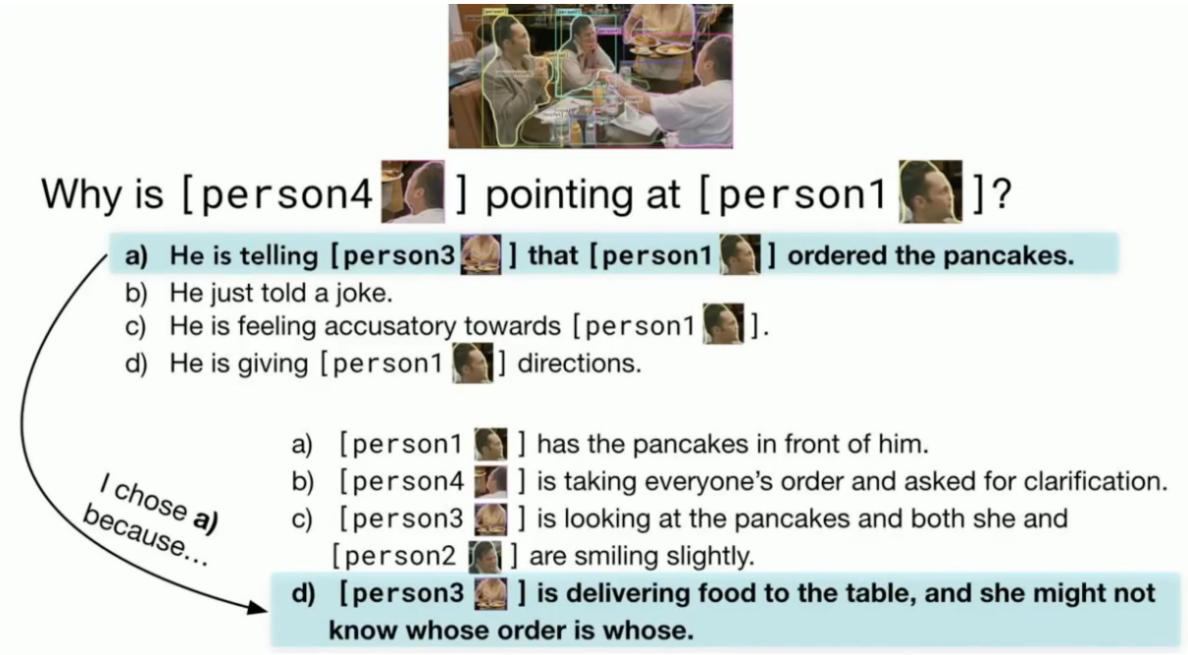
- Infer the entire situation: what is happening and why is it happening



1. Three people dining and already ordered food
2. **Person 3** is serving and is not with the group
3. **Person 1** ordered pancakes and bacon
4. How? **Person 4** is pointing to **Person 1** while looking at the server (**Person 3**)

Motivation and Importance

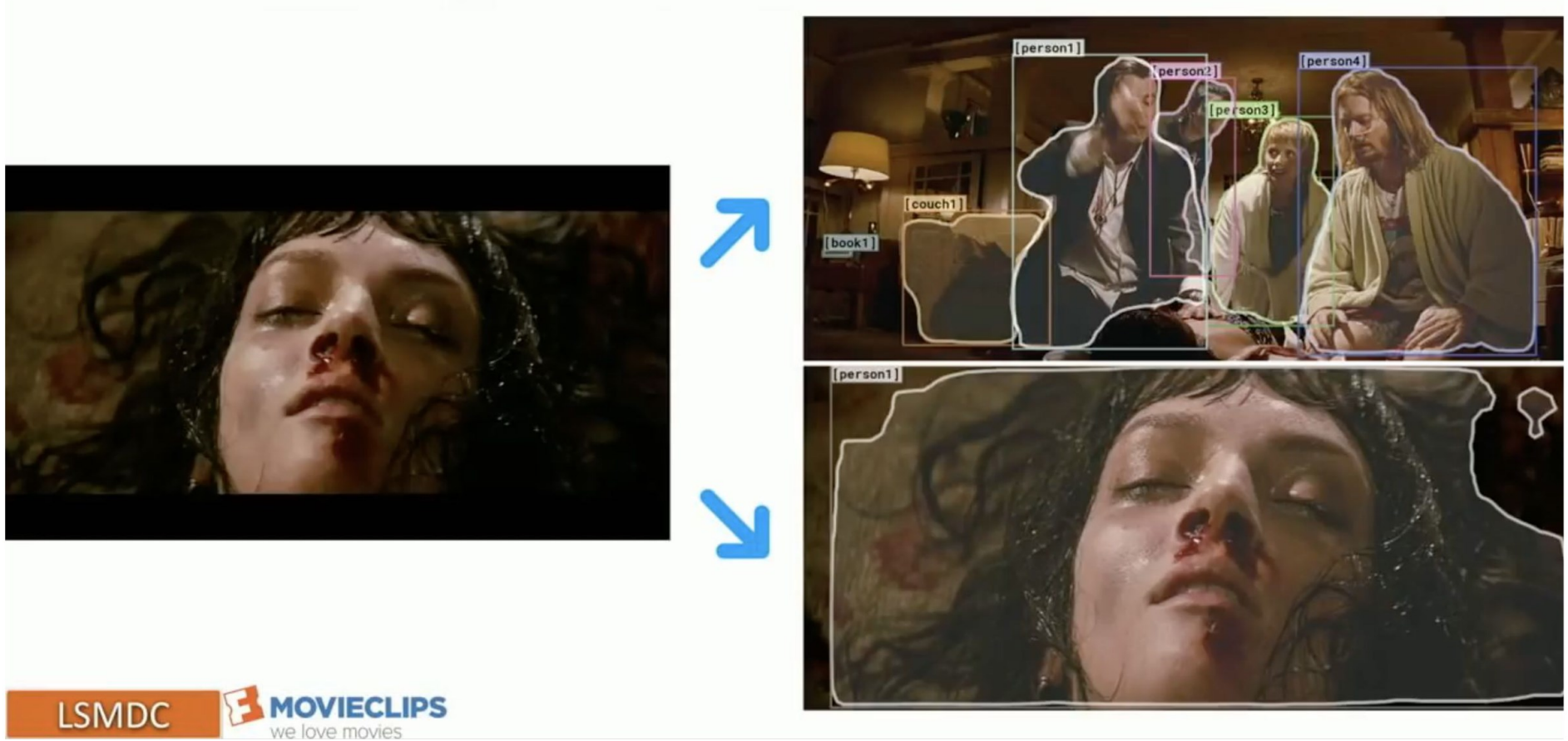
- ➔ Recognition [find objects] and cognition [infer interaction]
- ➔ Good vision systems
- ➔ **Good cognition systems** at scale.



- Image Captioning : High level understanding, Difficult Evaluation
- VQA : No rationale, Easy Evaluation.
- Multiple choice setting.
- Justification has to include details about the scene and background knowledge about how the world works.

Creating a dataset for VCR

Collecting commonsense inferences



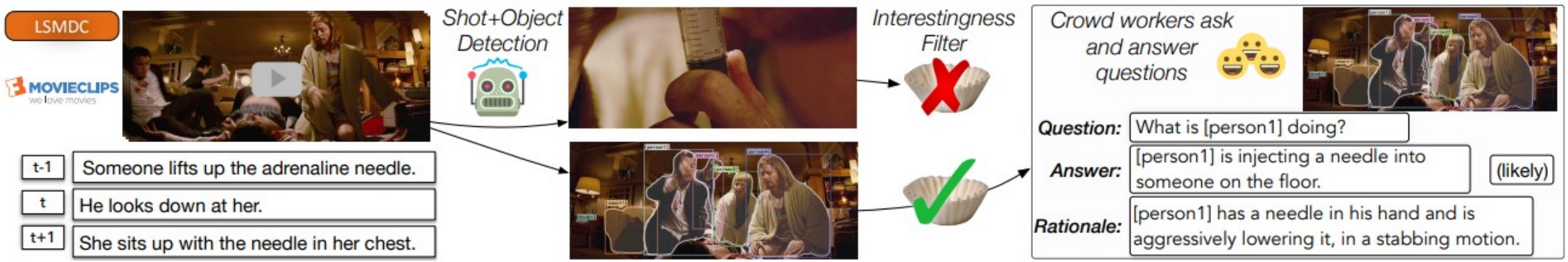
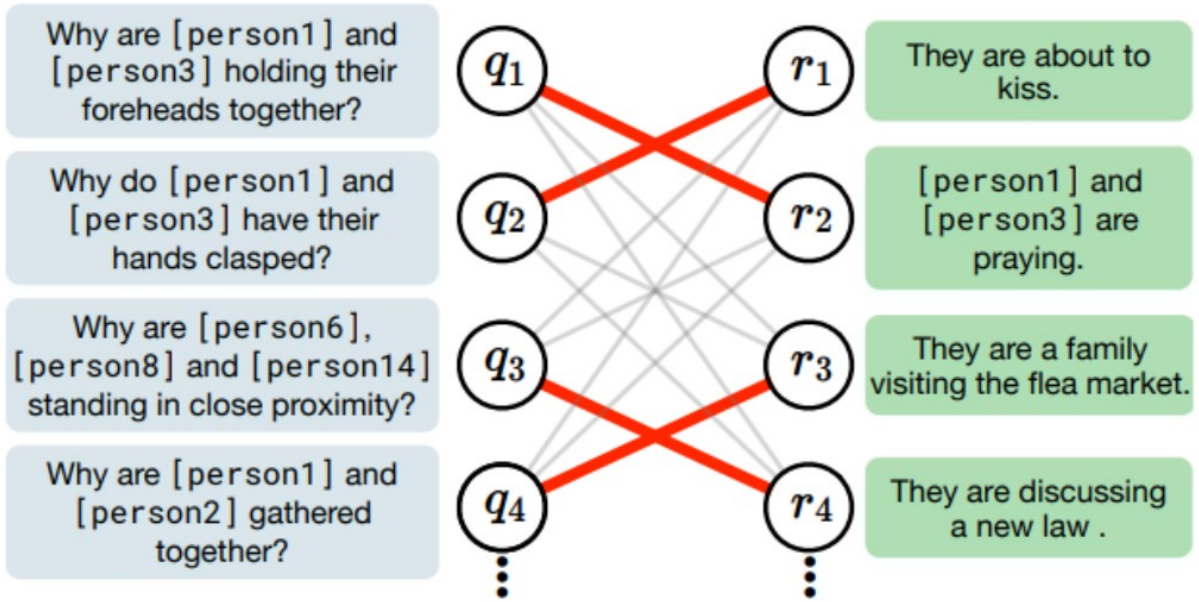


Figure 3: An overview of the construction of VCR. Using a state-of-the-art object detector [30, 25], we identify the objects in each image. The most interesting images are passed to crowd workers, along with scene-level context in the form of scene descriptions (MovieClips) and video captions (LSMDC, [68]). The crowd workers use a combination of natural language and object tags to ask and answer challenging visual questions, also providing a rationale justifying their answer.

Adversarial Matching



Adversarial Matching

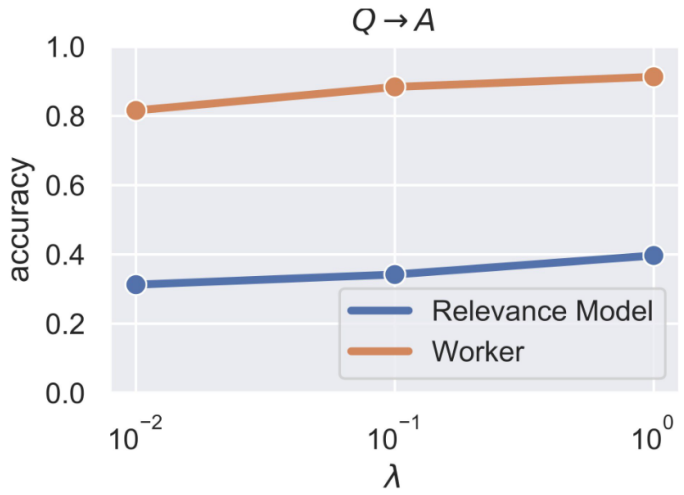
Wrong answers must be

Relevant to question yet **different from correct answer**

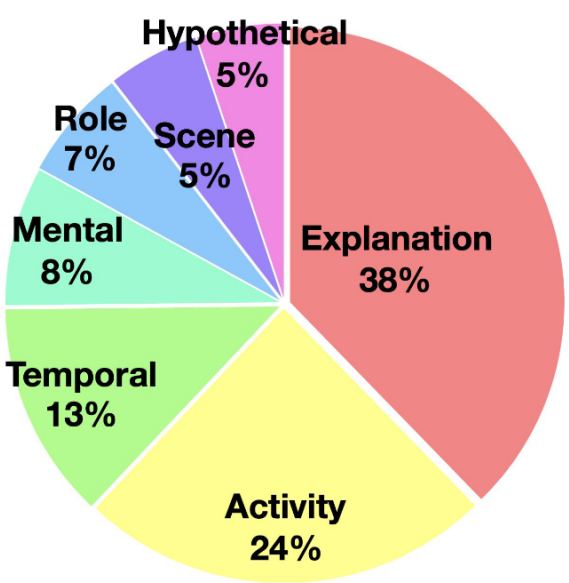
Q, A' Question Relevance **A, A' Entailment**

$$W_{i,j} = log(P_{rel}(q_i, r_j) + \lambda \; log(1 - P_{sim}(r_i, r_j))$$

$W_{i,j}$ = element of weight matrix | P_{rel} = relevance score



290K Questions | 110K images



Why is [person11] wearing sunglasses inside?

What are [person1] and [person2] doing?

What will [person6] do after unpacking the groceries?

What is [person3] thinking while [person5] shakes his hand?

What is [person1]'s relation to [person4]?

Where is [person1] now?

What would happen if [person3] fell asleep?

38% - Why and how | 24% cognition level activities | 13 % temporal reasoning

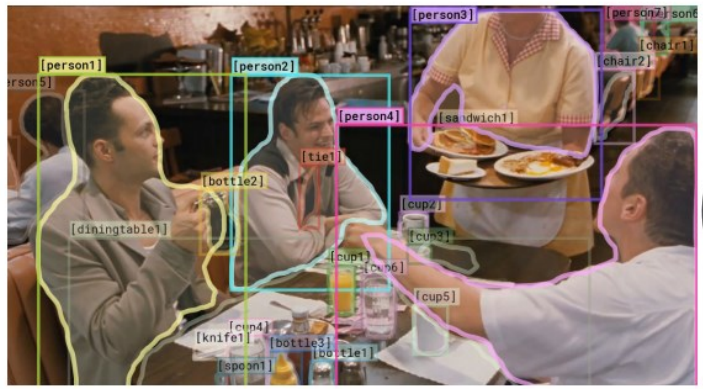
Definition:

I - image

sequence of object detections (o): bounding box (b), segmentation mask (m), class label.

Query (q) : either word in vocab or tag referring to an object.

Responses(N) : same structure as query.

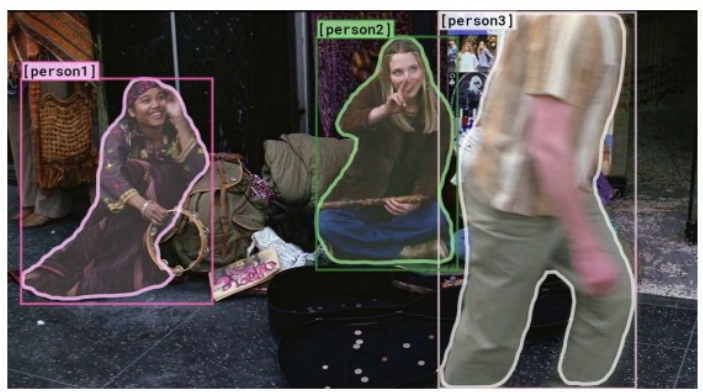


Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

*I chose **a)** because...*

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.



How did [person2] get the money that's in front of her?

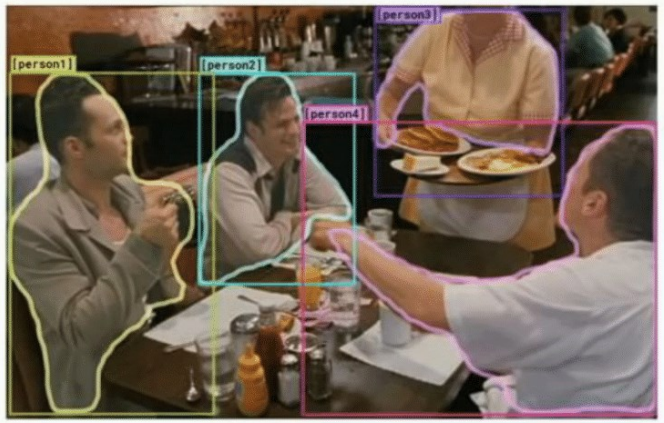
a) [person2] is selling things on the street.
b) [person2] earned this money playing music.
c) She may work jobs for the mafia.
d) She won money playing poker.

*I chose **b)** because...*

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
c) [person2] and [person1] are both holding instruments, and were probably busking for that money.
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

How to reach at the correct answer and reasoning?

1. Figure out meanings of query and responses wrt image and each other.
2. Do some inference on this representation.



Part 1: Grounding

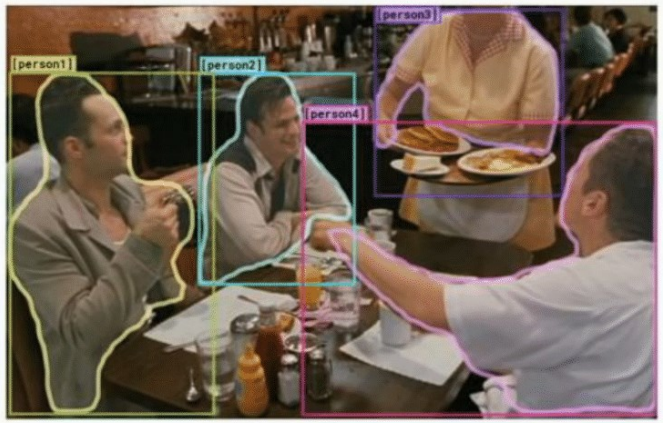
Objects p1 p2 p2 p3 ...

Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.



ResNet

Part 1: Grounding

Objects



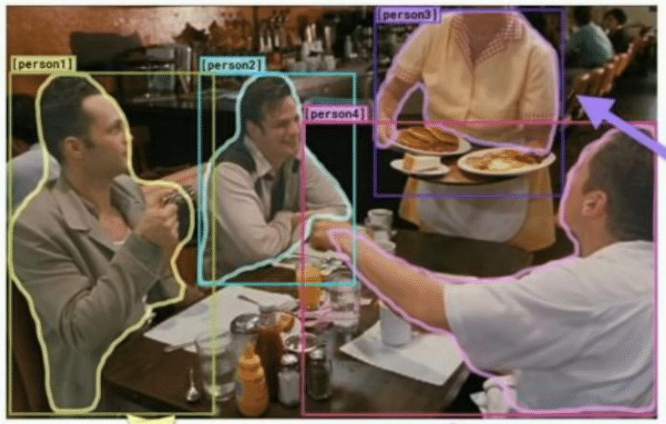

...

Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.



ResNet

Part 1: Grounding

Objects  ...

Query

Why is [person4 ]

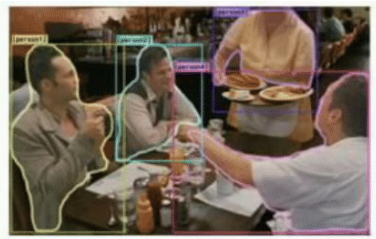
pointing at [person1 ]?

Response Choice

He is telling [person3 ]

that [person1 ]

ordered pancakes.



Objects  ...

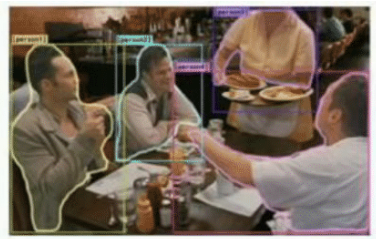
Part 1: Grounding

Query

Why is [person4 ] pointing at [person1 ]?

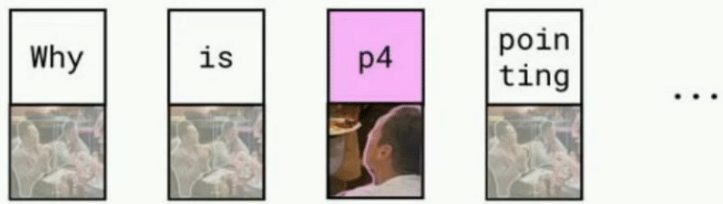
Response Choice

He is telling [person3 ] that [person1 ] ordered pancakes.



Objects  ...

Part 1: Grounding



Query

Why is [person4 ]

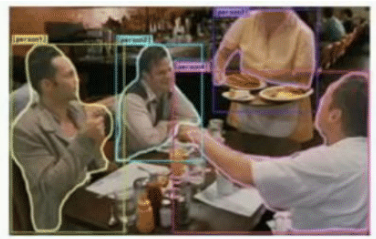
pointing at [person1 ]?

Response Choice

He is telling [person3 ]

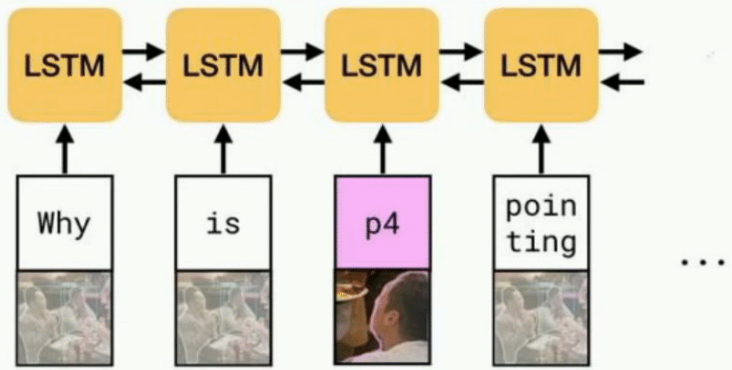
that [person1 ]

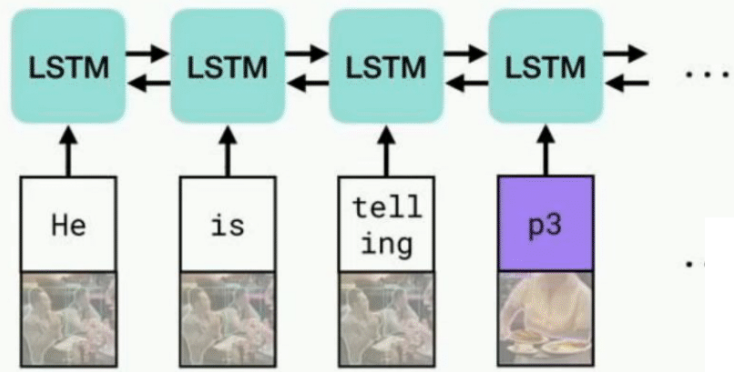
ordered pancakes.



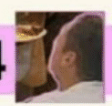
Objects  ...

Part 1: Grounding





Query

Why is [person4 ]

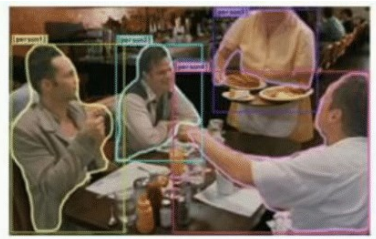
pointing at [person1 ]?

Response Choice

He is telling [person3 ]

that [person1 ]

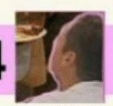
ordered pancakes.



Part 2: Contextualization

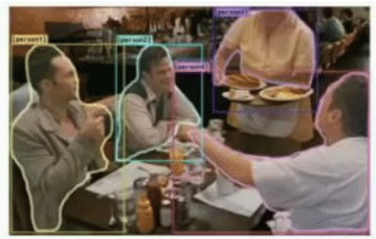
Objects     ...

Query

Why is [person4 ] pointing at [person1 ]?

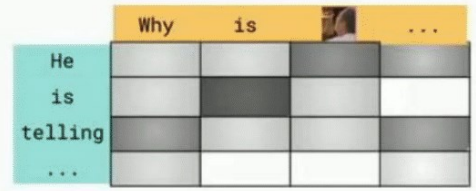
Response Choice

He is telling [person3 ] that [person1 ] ordered pancakes.

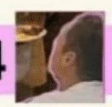


Part 2: Contextualization

Objects  ...

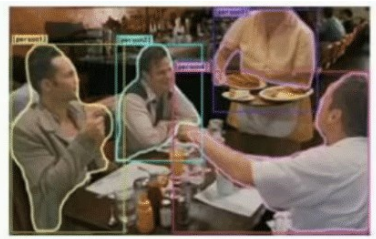


Query

Why is [person4 ] pointing at [person1 ]?

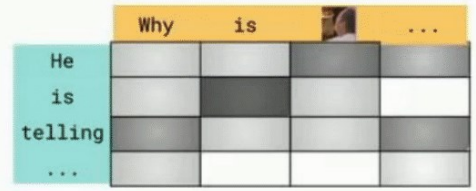
Response Choice

He is telling [person3 ] that [person1 ] ordered pancakes.

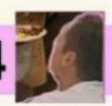


Part 2: Contextualization

Objects     ...



Query

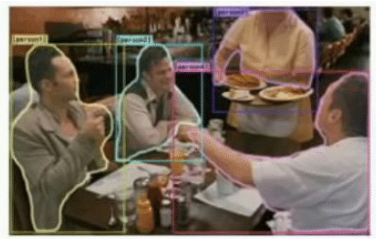
Why is [person4 ] pointing at [person1 ]?

Response Choice

 is telling [person3 ]

that [person1 ]

ordered pancakes.



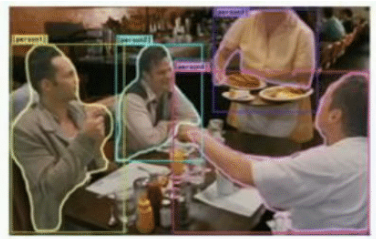
Part 3: Reasoning

Response Choice

 is telling [person3 ]

that [person1 ]

ordered pancakes.



Part 3: Reasoning

Response

Attended Query

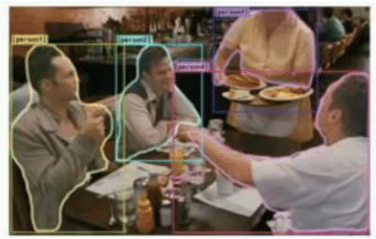
Attended Objects

Response Choice

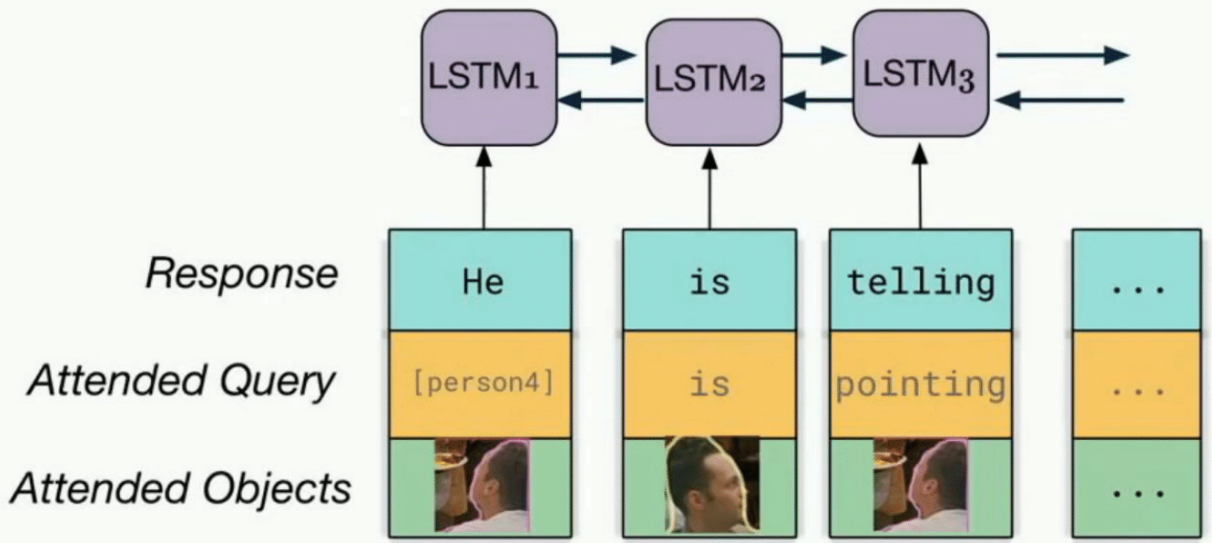
 is telling [person3 ]

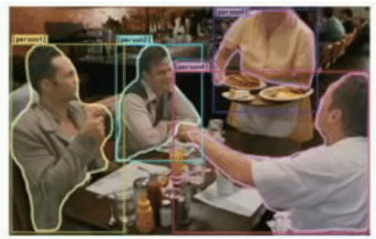
that [person1 ]

ordered pancakes.



Part 3: Reasoning

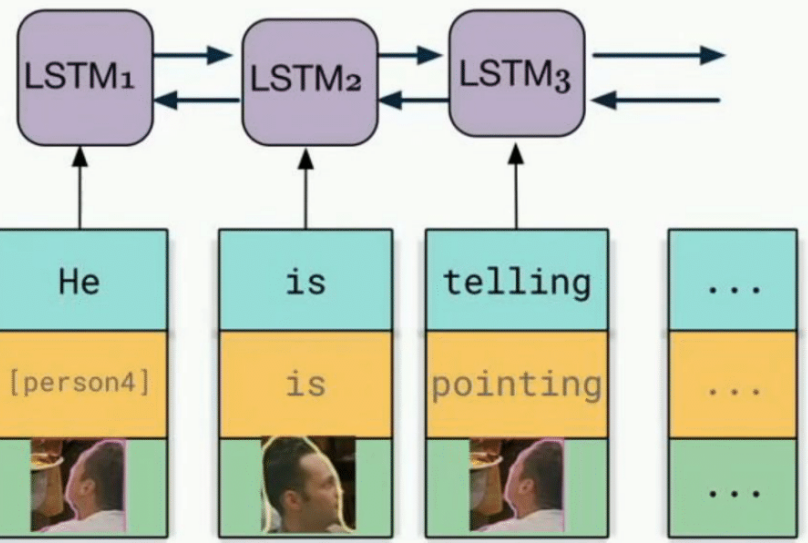




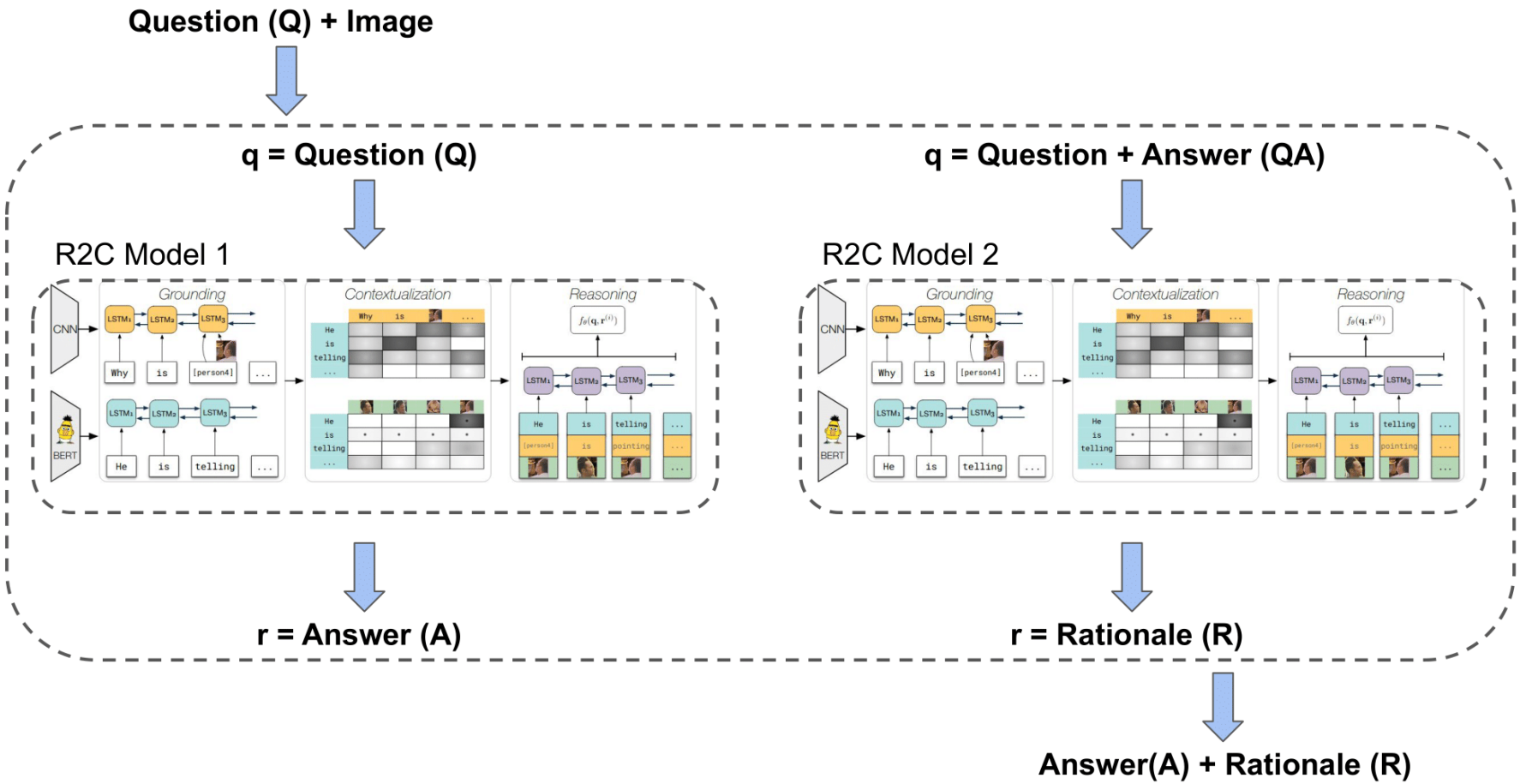
Part 3: Reasoning

$f_\theta(\mathbf{q}, \mathbf{r}^{(i)})$

Max pool+multilayer perceptron



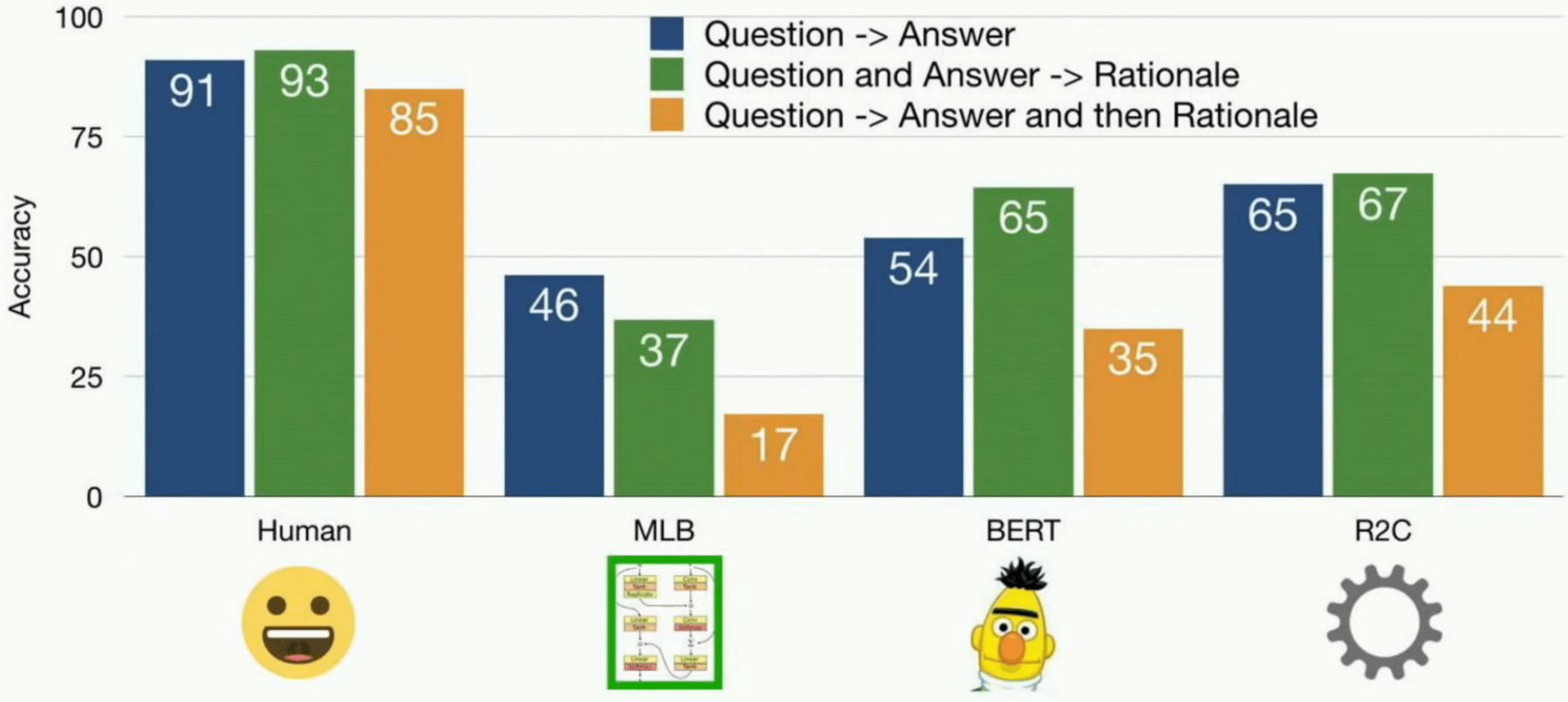
Final architecture



Results

	Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
Text Only	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
Text Only	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
Text Only	LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [38]	39.4	40.5	34.0	33.7	13.5	13.8
VQA	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
VQA	MLB [42]	45.5	46.2	36.1	36.8	17.0	17.2
VQA	MUTAN [6]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C	**63.8**	**65.1**	**67.2**	**67.3**	**43.1**	**44.0**
	Human		91.0		93.0		85.0

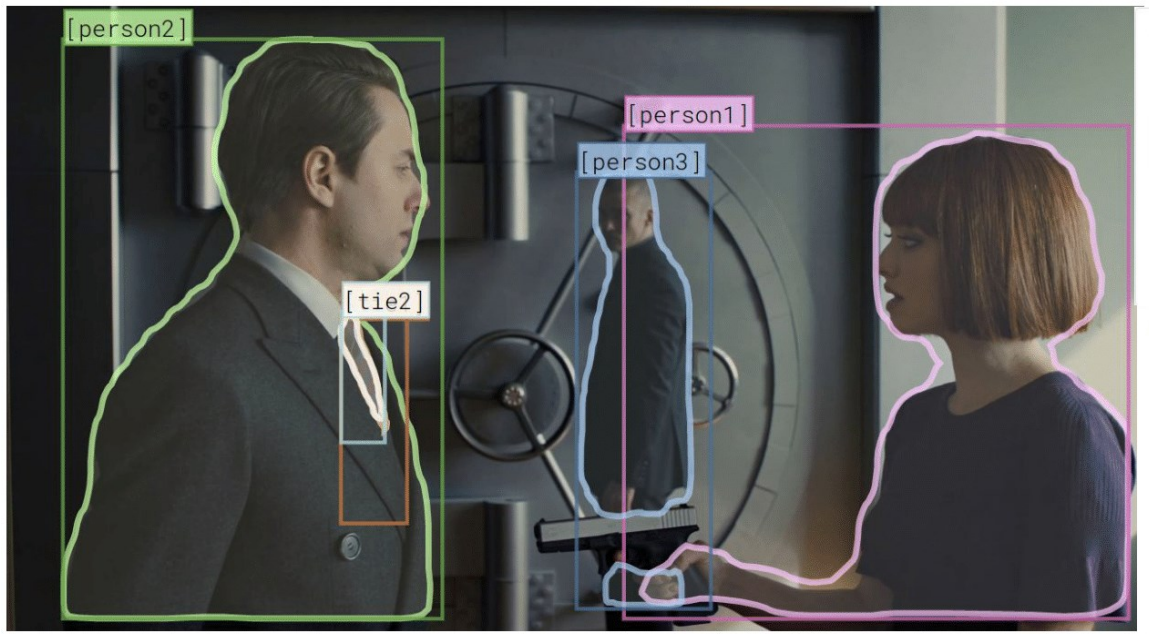
VCR Results


Accuracy
100
75
50
25
0
Question -> Answer
Question and Answer -> Rationale
Question -> Answer and then Rationale
91
93
85
46
37
17
54
65
35
65
67
44
Human
MLB
BERT
R2C

Ablation Results

Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
R2C	**63.8**	**67.2**	**43.1**
No query	48.3	43.5	21.5
No reasoning module	63.6	65.7	42.2
No vision representation	53.1	63.2	33.8
GloVe representations	46.4	38.3	18.3

Sample Results (Correct)



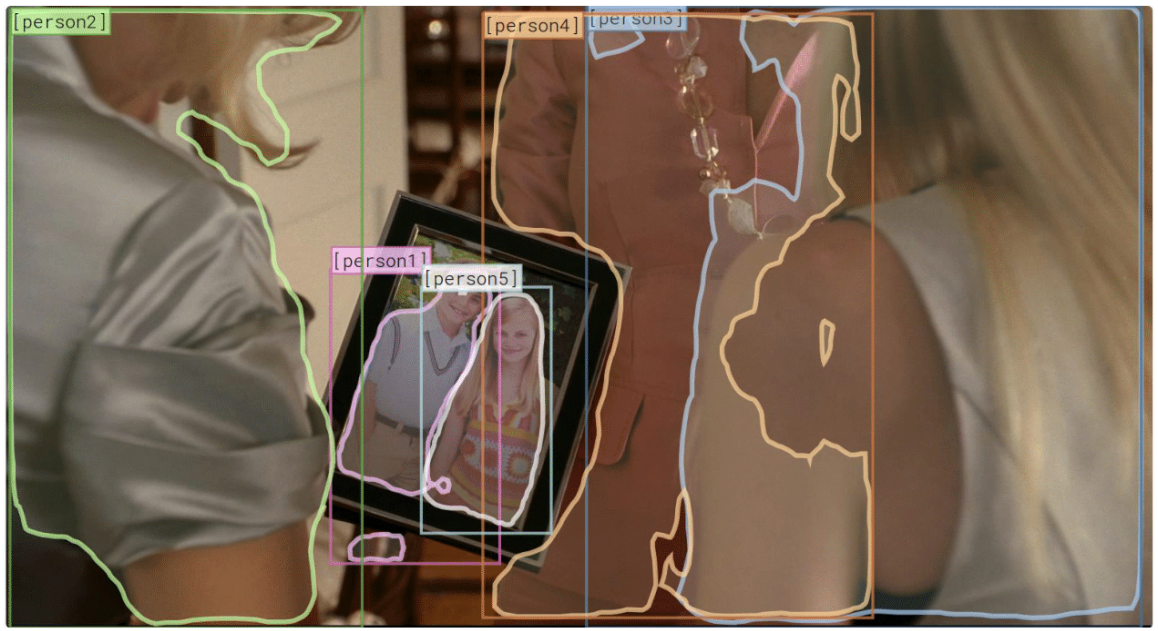
1. Why is [person1] pointing a gun at [person2] ?

a) [person1] wants to kill [person2] . **1.4%**
b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. **71.7%**
c) [person2] has done something to upset [person1] . **18.7%**
d) Because [person2] is [person1] 's daughter. [person1] wants to protect [person2] . **8.2%**

I think so because...

a) [person1] is chasing [person1] and [person3] because they just robbed a bank. **33.8%**
b) Robbers will sometimes hold their gun in the air to get everyone's attention. **5.3%**
c) The vault in the background is similar to a bank vault. [person3] is waiting by the vault for someone to open it. **49.1%**
d) A room with barred windows and a counter usually resembles a bank. **11.7%**

Sample Results (Incorrect)



1. What is going to happen next?

a) [person2] is going to walk up and punch [person4] in the face. **10.8%**
b) Someone is going to read [person4] a bed time story. **15.2%**
c) [person5] is going to fall down. **5.1%**
d) [person2] is going to say how cute [person4] 's children are. **68.9%**

I think so because...

a) They are the right age to be father and son and [person5] is hugging [person3] like they are his son. **1.5%**
b) It looks like [person4] is showing the photo to [person2] , and [person2] will want to be polite. **31.6%**
c) [person2] is smirking and looking down at [person4] . **6.2%**
d) You can see [person4] smiling and facing the crib and decor in the room. **60.7%**

Critique

The “Good”

- **Dataset**:
  - A comprehensive pipeline to select and annotate interesting images at scale. 290K questions with over 110k images.
- **Adversarial Matching**:
  - Leverages the Natural Language Inference task to come up with a robust dataset with minimized annotation artifacts. Better BERT better dataset.
- **Ablation study:**
  - Performed ablation study of all the important model components as well as alternatives not included, gives reasoning behind model design decisions.
- **Details:**
  - Main paper: 8 pages, Appendix: 17 pages
  - Detailed description of Dataset, Adversarial Matching, R2C Model, even the hyper-parameters used. Ensuring detailed insights and easy reproducibility of results.
- **Output analysis:**
  - Performed analysis of correct as well as incorrect results, giving more insight into the model’s understanding of the world.

Critique

The “Not So Good”

- **Language bias in the dataset:**
  - Abstract images help to reduce language bias, i.e things like fire hydrant being red, or sky being blue. Artificial images discourage bias due to this. Could this have been added?
- **Adversarial Matching:**
  - P_{rel} computation limits application of Bigger BERT easily.

- **R2C Model:**
  - Knowledge bases for world knowledge and automatic reasoning generation.
  - Better handling of the subtasks Q→A and QA→R rather than naive composition.
- **Evaluation Methodology:**
  - Better baseline VQA models could have been selected to ensure a fair comparison with R2C.
  - No masking-approach in the paper or masking-based results to ensure complete attention coverage of images.

Thank you!

Transformer – (Personal) Architecture Intuition

Stacked blocks: the deeper the model is, the higher the non-linearity aspect.

FFN: Perform non-linear transformation after self-attention, to obtain better representation.

Residual connections: Allow disabling Multi-Head Attention and/or FFN, throughout the whole stack. Help gradients propagation during training.

Core block of A - (Attention+Add & Norm) and B - (FFN+Add & Norm):

- For each input word w, get an intermediate embedding e that summarizes interaction between w and the whole sentence.
- Transform e by applying it non-linear transformation(s).