

---

# From Recognition to Cognition: Visual Commonsense Reasoning

---

Rowan Zellers    Yonatan Bisk    Ali Farhadi    Yejin Choi    Paul G. Allen  
School of Computer Science & Engineering, University of Washington  
Allen Institute for Artificial Intelligence



How can we further **cognition-level reasoning** towards **complete** visual understanding?

**Cognition**

- Objects
- Attributes

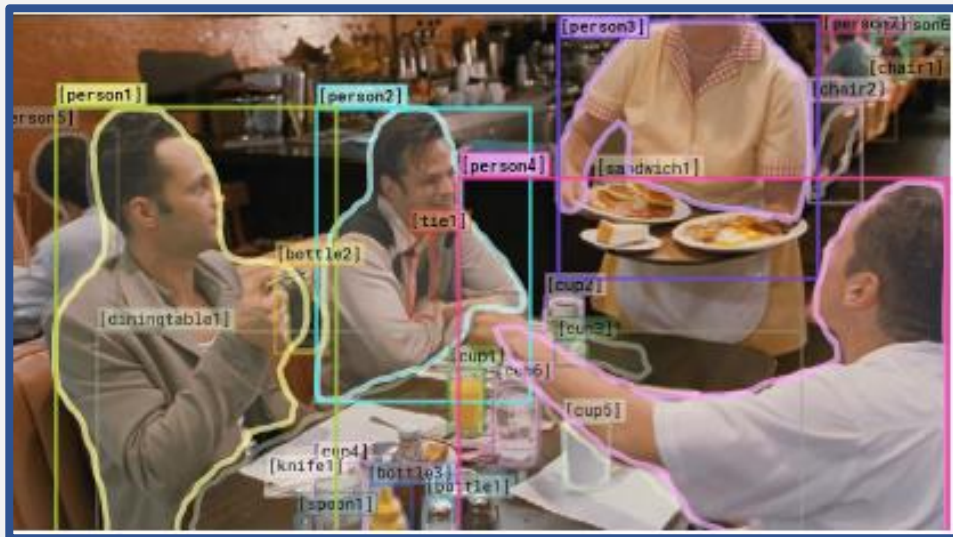
**Recognition**

- Rationale
- Intents
- Social Dynamics

What?

Why?

## Recognition to Cognition



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...

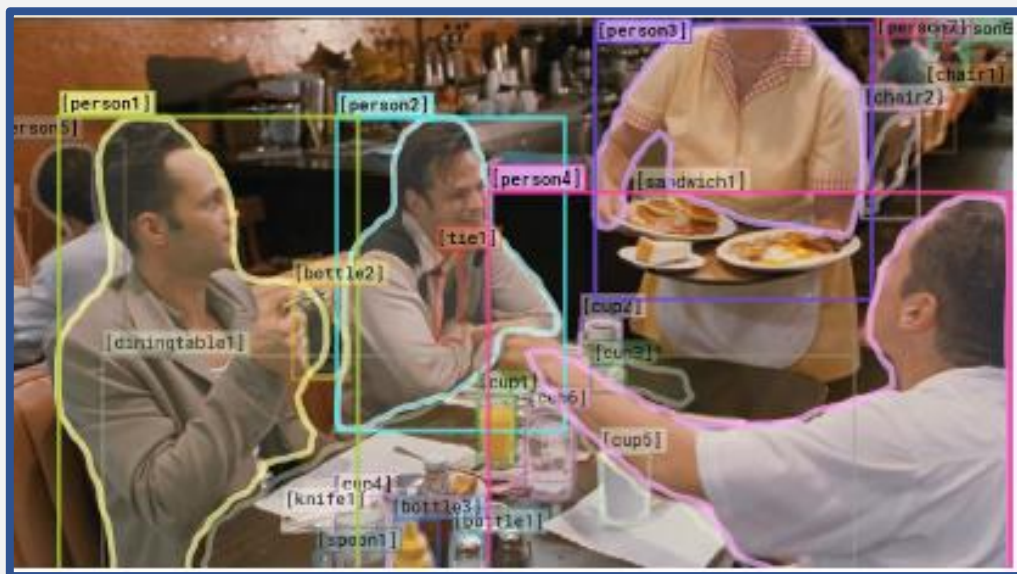
- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

Achieving beyond recognition level...

- 1) Proposes a new task & a large-scale multiple choice QA dataset, **VCR**.
- 2) Presents **Adversarial Matching**, a new algorithm for robust multiple-choice dataset creation.
- 3) Proposes **R2C**, that aims to mimic the layered inference from recognition to cognition.

# 04 Task

## Task Overview



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.

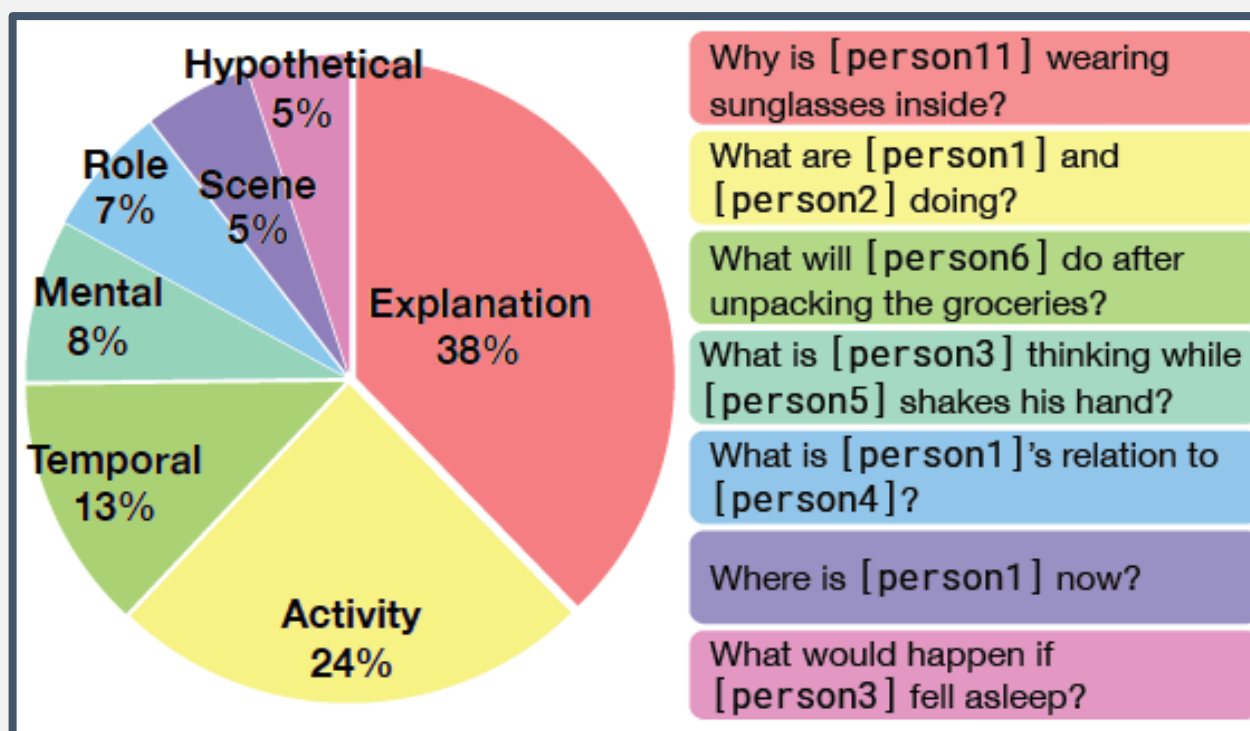
- Holistically + Cognitive understanding
- Staged answering & justification
  - Q (query) -> A (response)
  - QA (query\_concat) -> R (response)

Correct = Correct Answer + Correct Rationale

a)

d)

## Task Overview



Types of Inferences

## Dataset/Code

[Dataset]

<https://visualcommonsense.com/>

[Code]

<https://github.com/rowanz/r2c/>

- 290k multiple choice QA problems  
(pairs of Q,A,R)
- 110k unique movie scenes

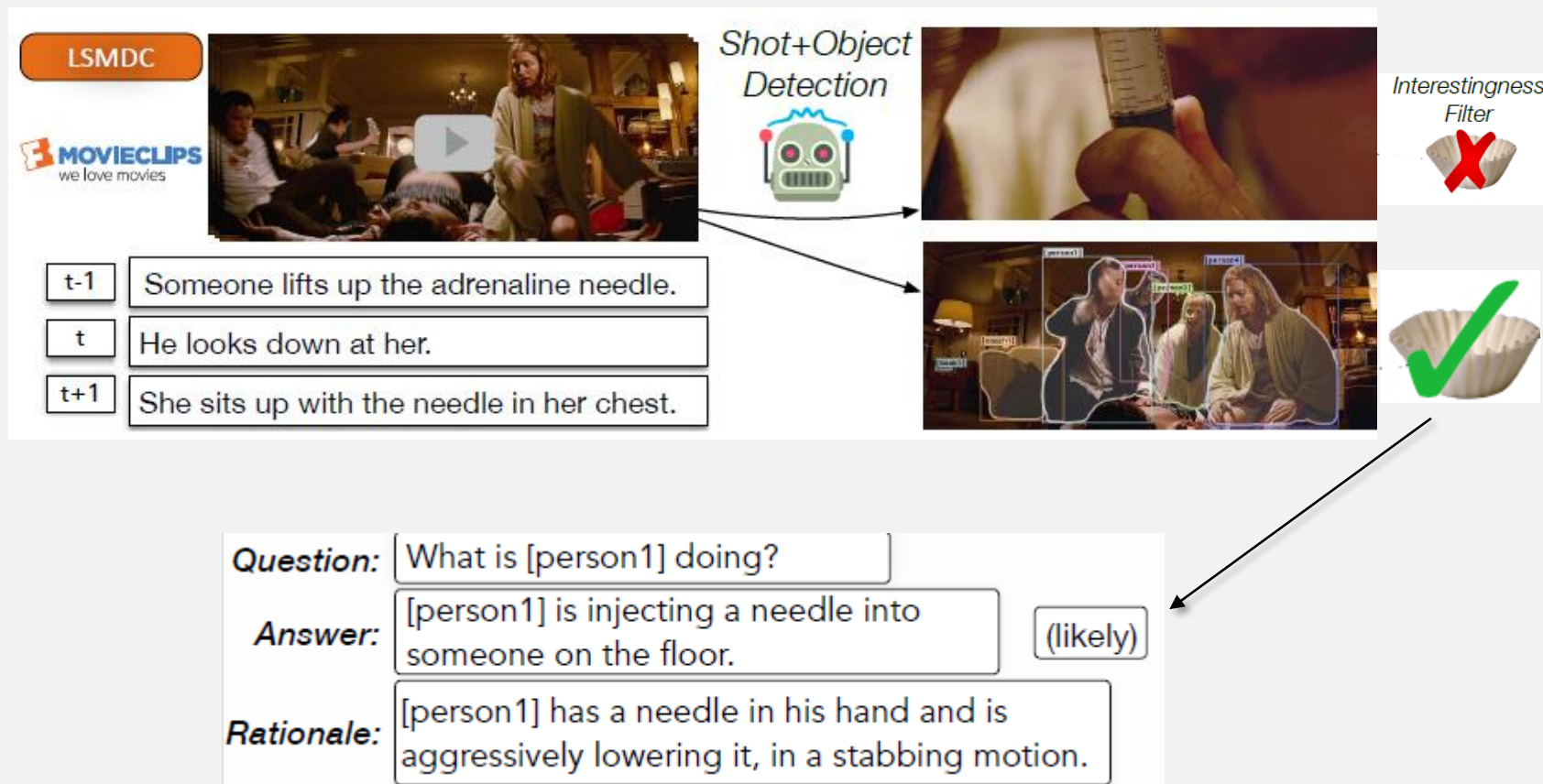




# 05 Dataset

## Collection

1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing



## Collection

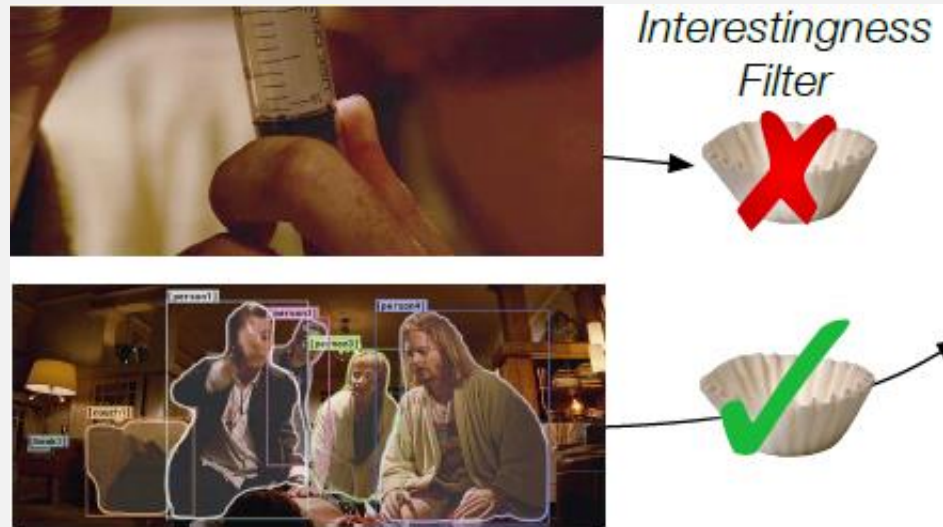
1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing



- Iterate through video clip (1 fps)
- Register shot boundary w/ 30 pixel mean difference
- Apply Mask-RCNN (threshold 0.7) -> 3 confidence tags

## Collection

1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing




- At least two people
- Additional filtering process (uninteresting or blurry)
  - 1) Logistic regression (Interesting vs Uninteresting)
  - 2) ResNet 50 (Interesting vs Moderate vs Boring)

## Collection: Crowdsourcing

1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing

Crowd workers ask and answer questions 🤖🤖🤖



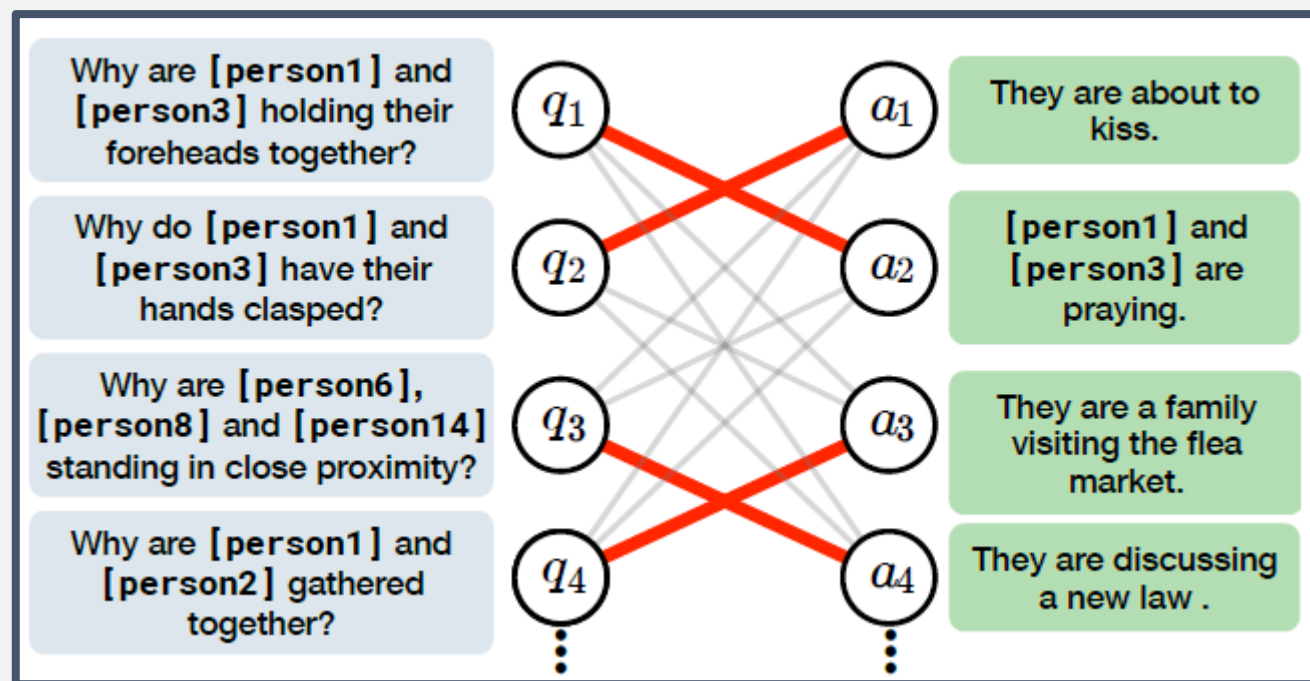
**Question:** What is [person1] doing?

**Answer:** [person1] is injecting a needle into someone on the floor. (likely)

**Rationale:** [person1] has a needle in his hand and is aggressively lowering it, in a stabbing motion.

- Ask 1~3 questions, answers, & rationales to each image
- Produce **unique** answer and rationale

## Adversarial Matching



25%

||

25%

||

25%

||

25%

Robust to  
answer-only  
biases

Alleviating "Annotation Artifacts"

## Adversarial Matching Details

Aligning Detection

Rematch detection tags

Semantic Categories

Utilizing "buckets"

Relevance Model

Training relevance btw R & Q

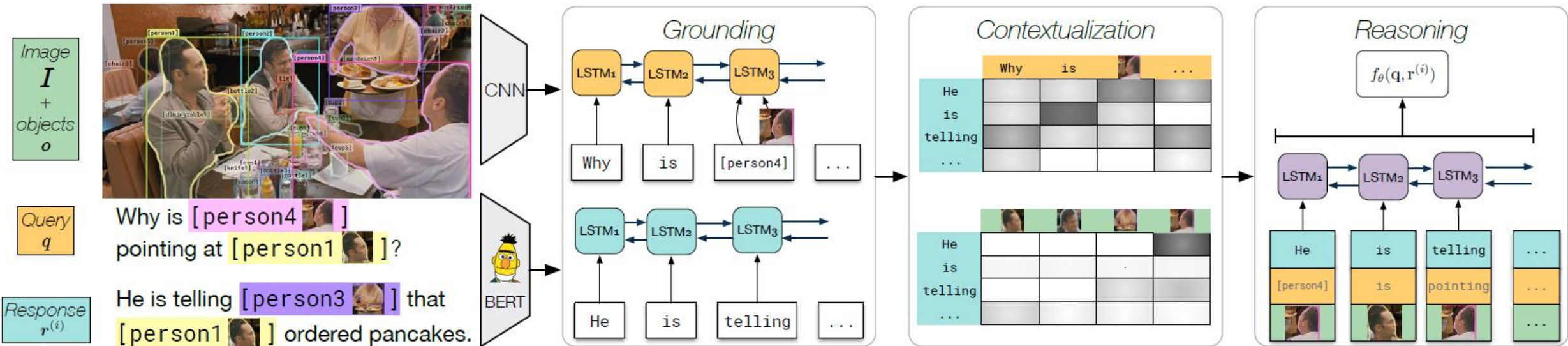
Similarity Model

Avoiding conflation



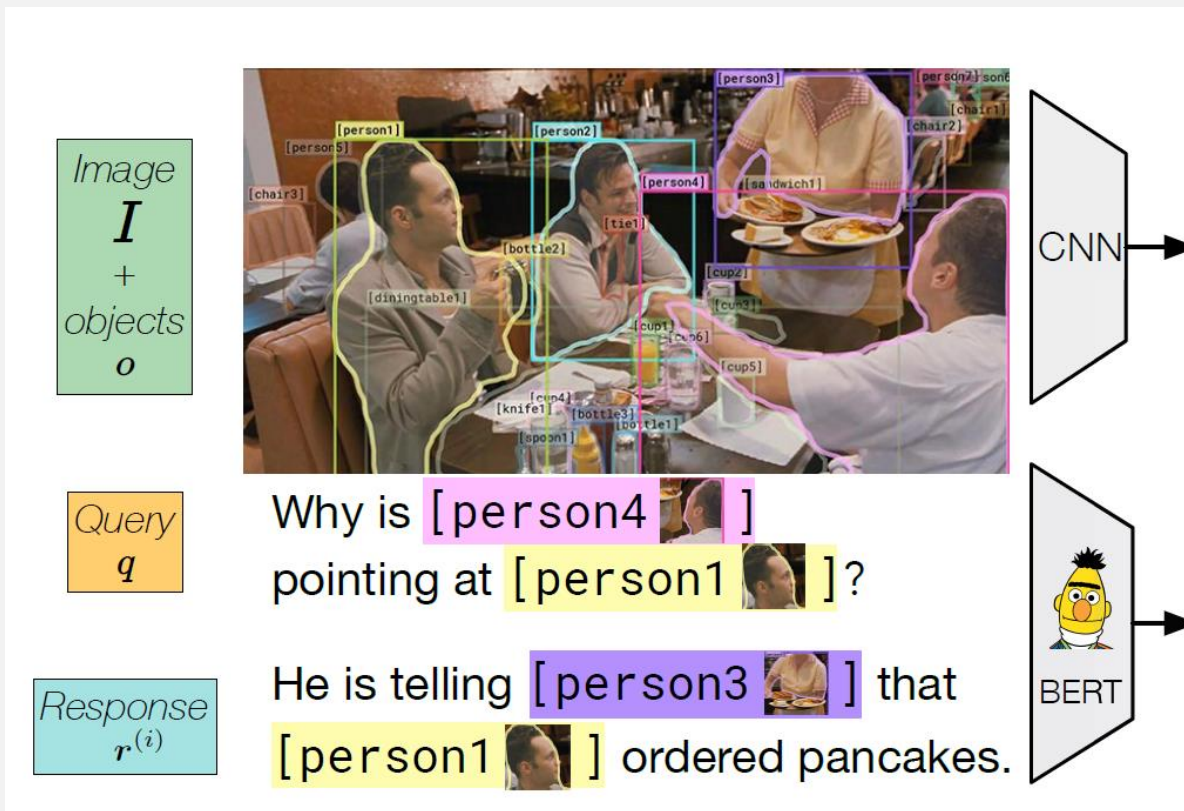
06

# R2C overview



Input : Query, Response  
Output :  $f(q, r)$

## Feature embedding



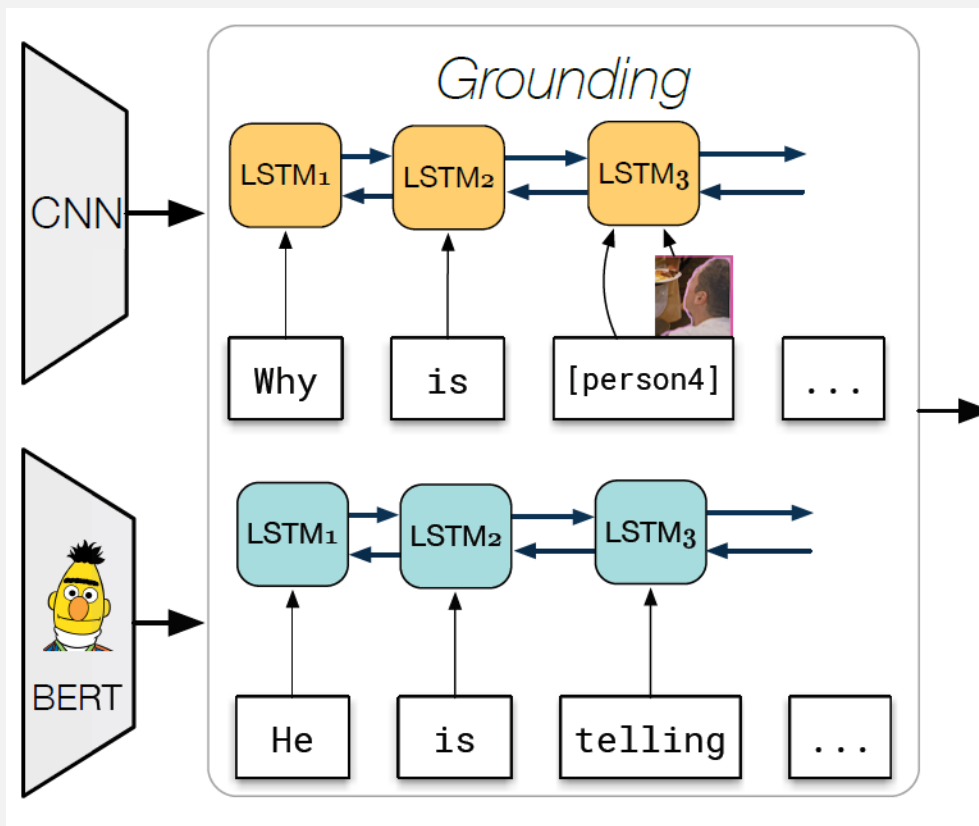
Word  $\rightarrow$  representation by Bert

Object  $\rightarrow$  Feature embedding by CNN



## 06 Model

### Grounding



Learn joint image-language representation

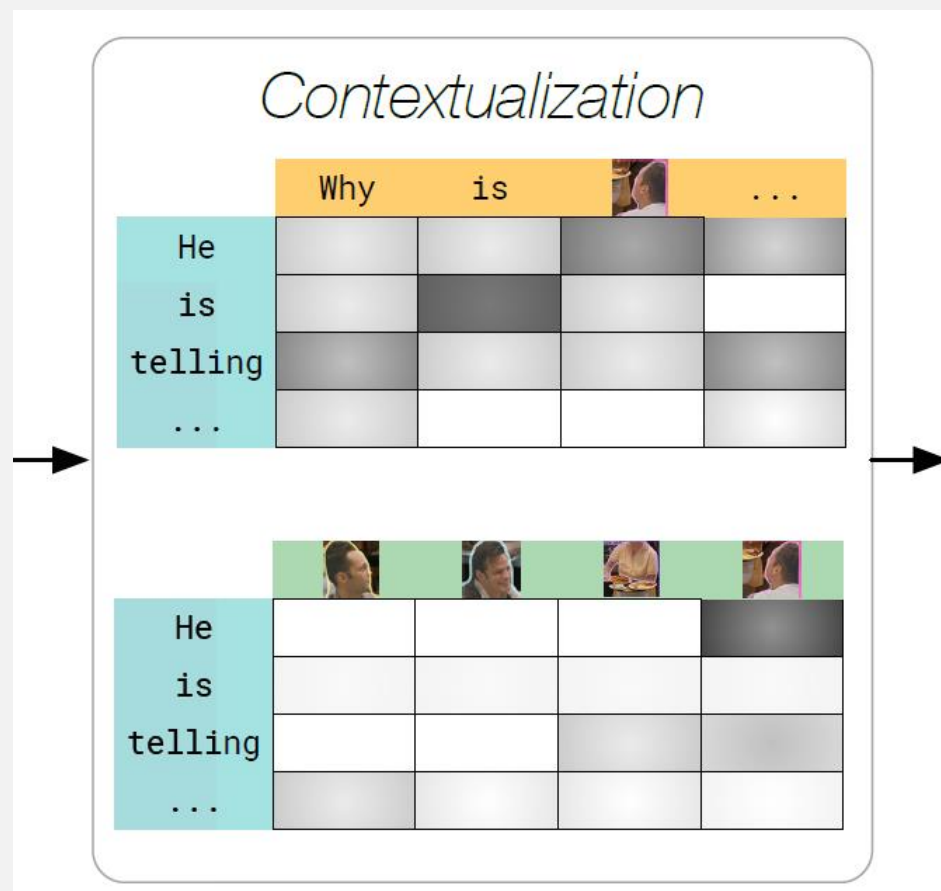
Bidirectional LSTM

Parameter share

Output : r for response, q for query

## 06 Model

### Contextualization

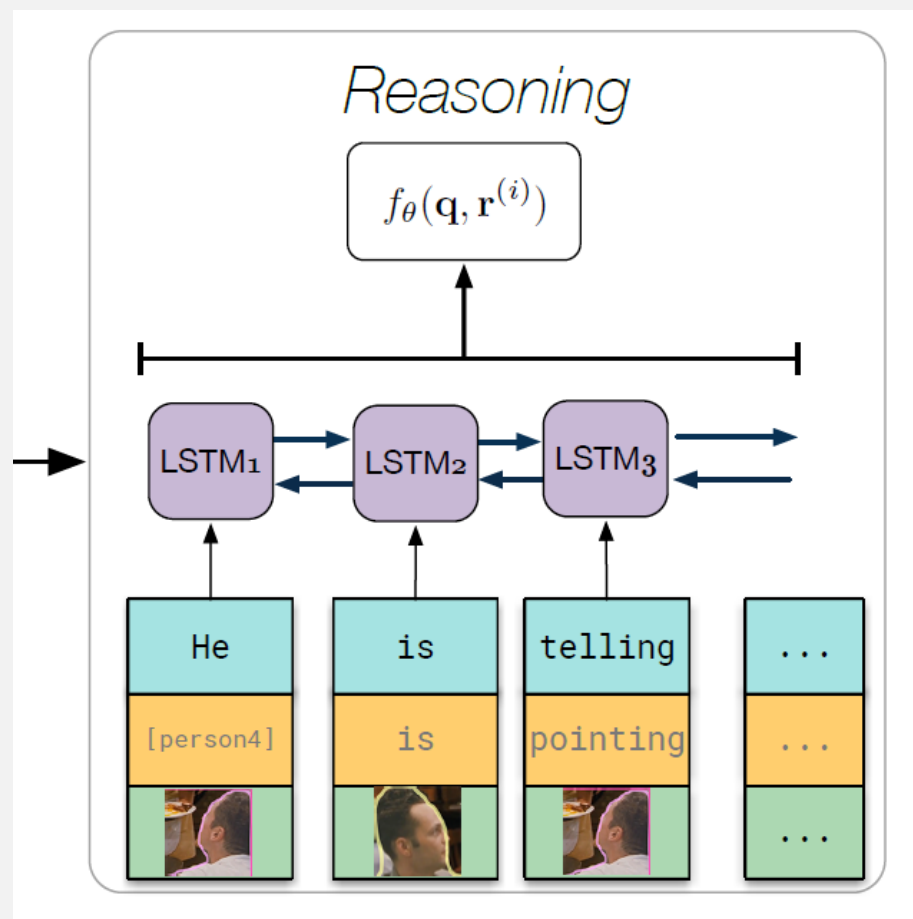


By attention mechanism

Attended query :  $\hat{q}_i$

Attended object :  $\hat{o}_i$

## Reasoning



Input :  $r_i, \hat{q}_i, \hat{o}_i$

Output : Probability

Image feature embedding : ResNet50  
Projection of image feature : 2176 to 512

Language feature embedding : BERT

LSTM hidden size : 256

Recurrent dropout = 0.3  
Optimizer : Adam  
Learning rate :  $2 \cdot 10^{-4}$  to  $10^{-4}$

Batch size : 96  
Epochs : 20

## Text-only Baselines

- 1) BERT
- 2) BERT(response only)
- 3) ESIM + ELMo
- 4) LSTM + ELMo

## VQA Baselines

- 1) Revisited VQA
- 2) Bottom-UP and Top-down attention
- 3) Multimodal Low-Rank Bilinear Attention (MLB)
- 4) Multimodal Tucker Fusion (MUTAN)

## Main Tasks

Model		$Q \rightarrow A$		$QA \rightarrow R$		$Q \rightarrow AR$	
		Val	Test	Val	Test	Val	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
	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
	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
	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
	MLB [42]	45.5	46.2	36.1	36.8	17.0	17.2
	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	

## Ablation

Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
<b>R2C</b>	<b>63.8</b>	<b>67.2</b>	<b>43.1</b>
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

Strong textual representations are crucial to VCR

- 1) Proposes a new task & a large-scale multiple choice QA dataset, **VCR**
- 2) Presents **Adversarial Matching**, a new algorithm for robust multiple-choice dataset creation.
- 3) Proposes **R2C**, that aims to mimic the layered inference from recognition to cognition



Q&A