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

CVPR '19

Noah Lee (noahlee357@korea.ac.kr)
Donghui Lim (ehdgnl101@korea.ac.kr)
Data Intelligence Lab, Korea University
2020 01. 27

01 Problem Statement

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

02 Motivation

Recognition → **Cognition**

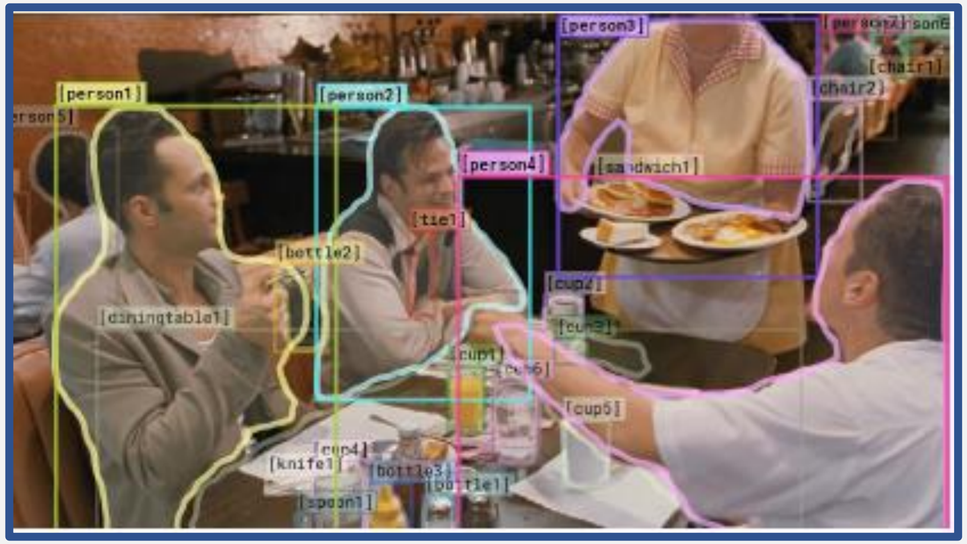
- Objects
- Attributes

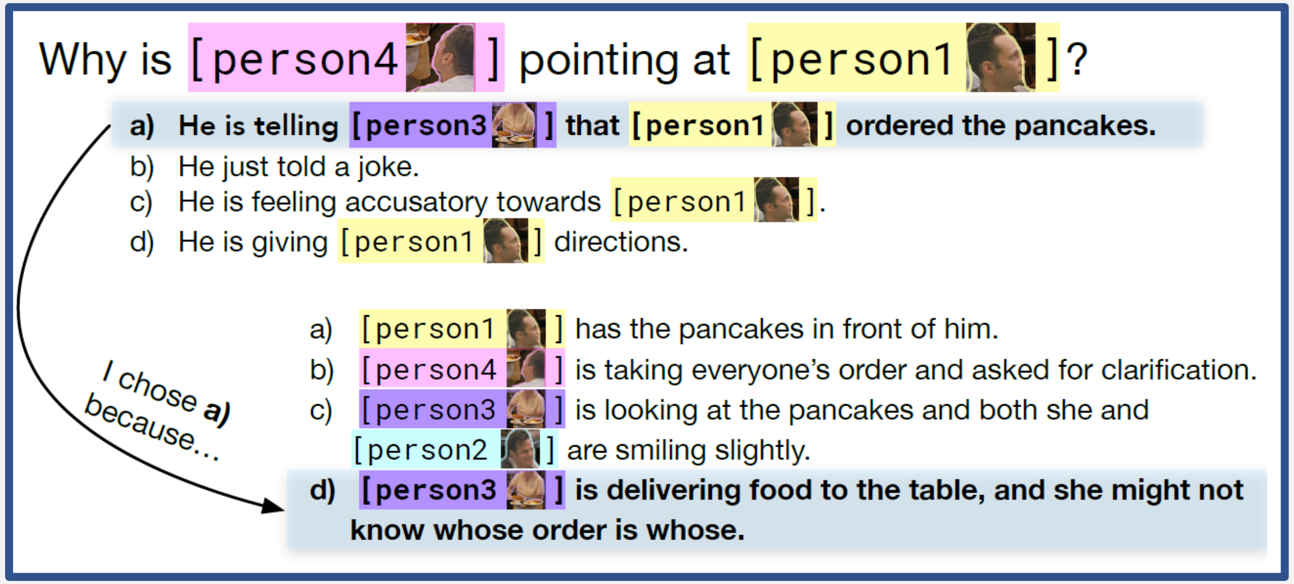
- Rationale
- Intents
- Social Dynamics

What? — Why?

02 Motivation

Recognition to Cognition





Achieving beyond recognition level…

03 Contribution

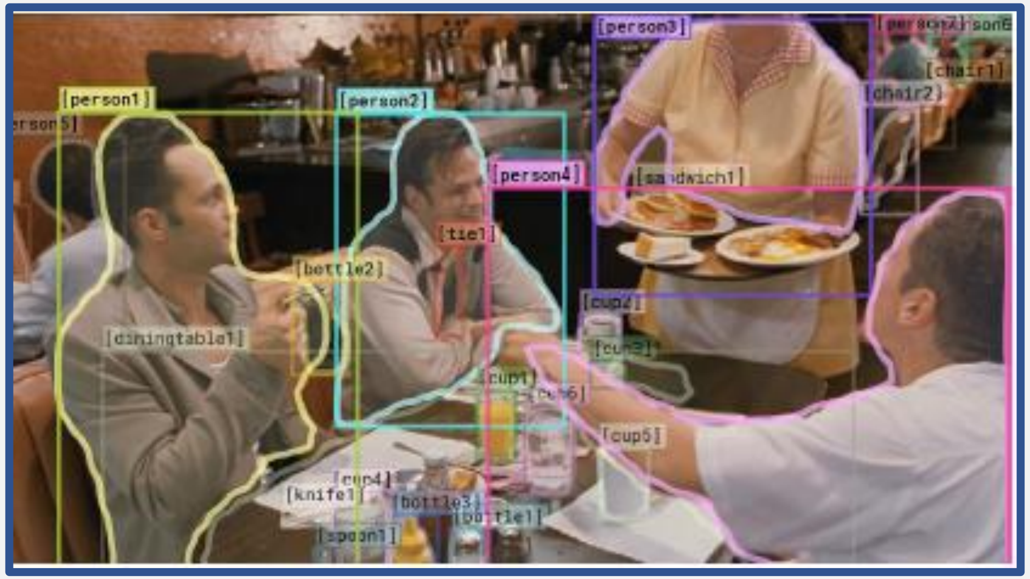
1) Proposes a new task & a large-scale multiple choice QA dataset, **VCR.**

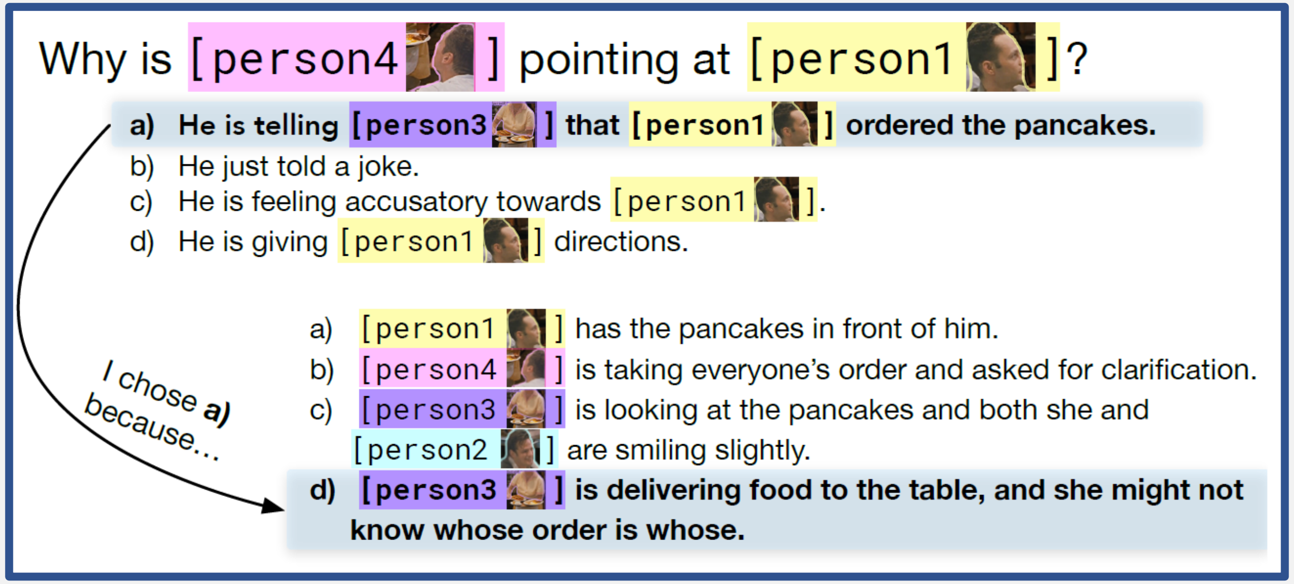
2) Presents **Adversarial Matching**, a new algorithm for robust multiple-choice dataset creation.

3) Proposes **R2C**, that aims to mimic the layered inference from recognition to cognition.

04 Task

Task Overview





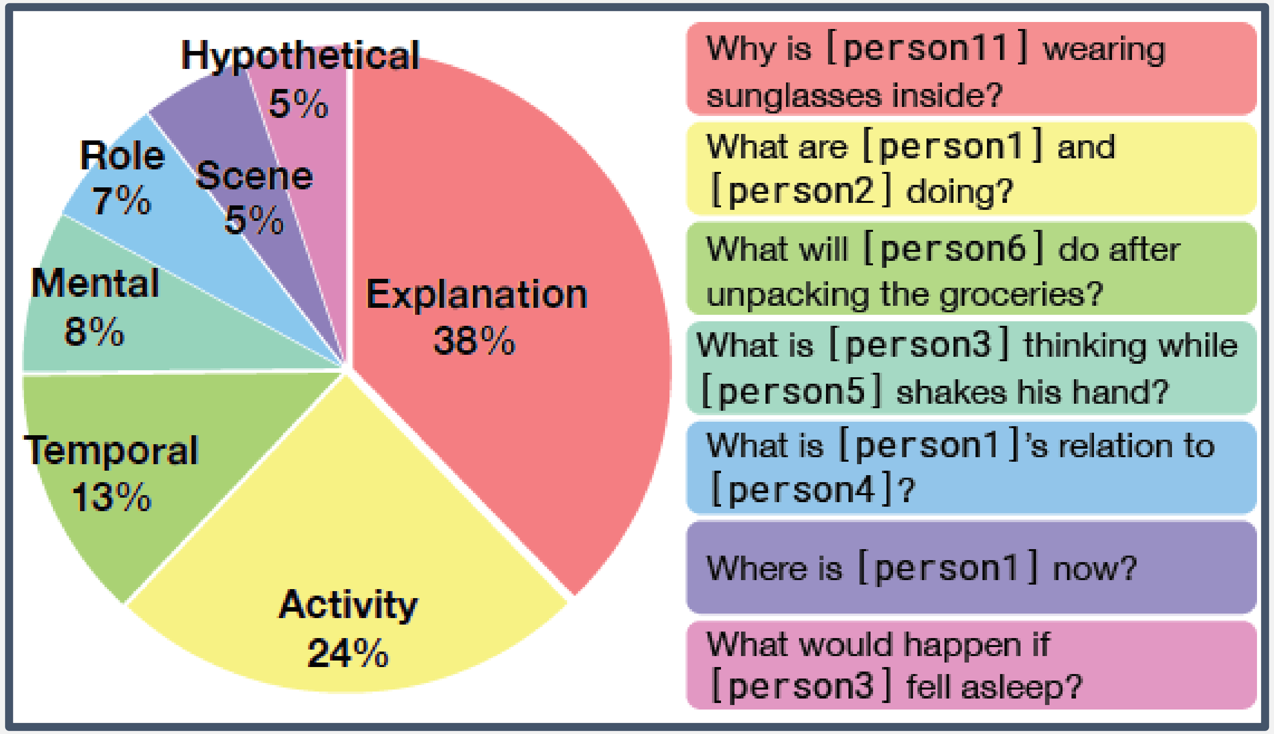
- Holistically + Cognitive understanding
- Staged answering & justification
  - Q (query) -> A (response)
  - QA (query_concat) -> R (response)

Correct = Correct Answer + Correct Rationale

a) d)

04 Task

Task Overview



Types of Inferences

05 Dataset

Dataset/Code

[Dataset]
https://visualcommonsense.com/

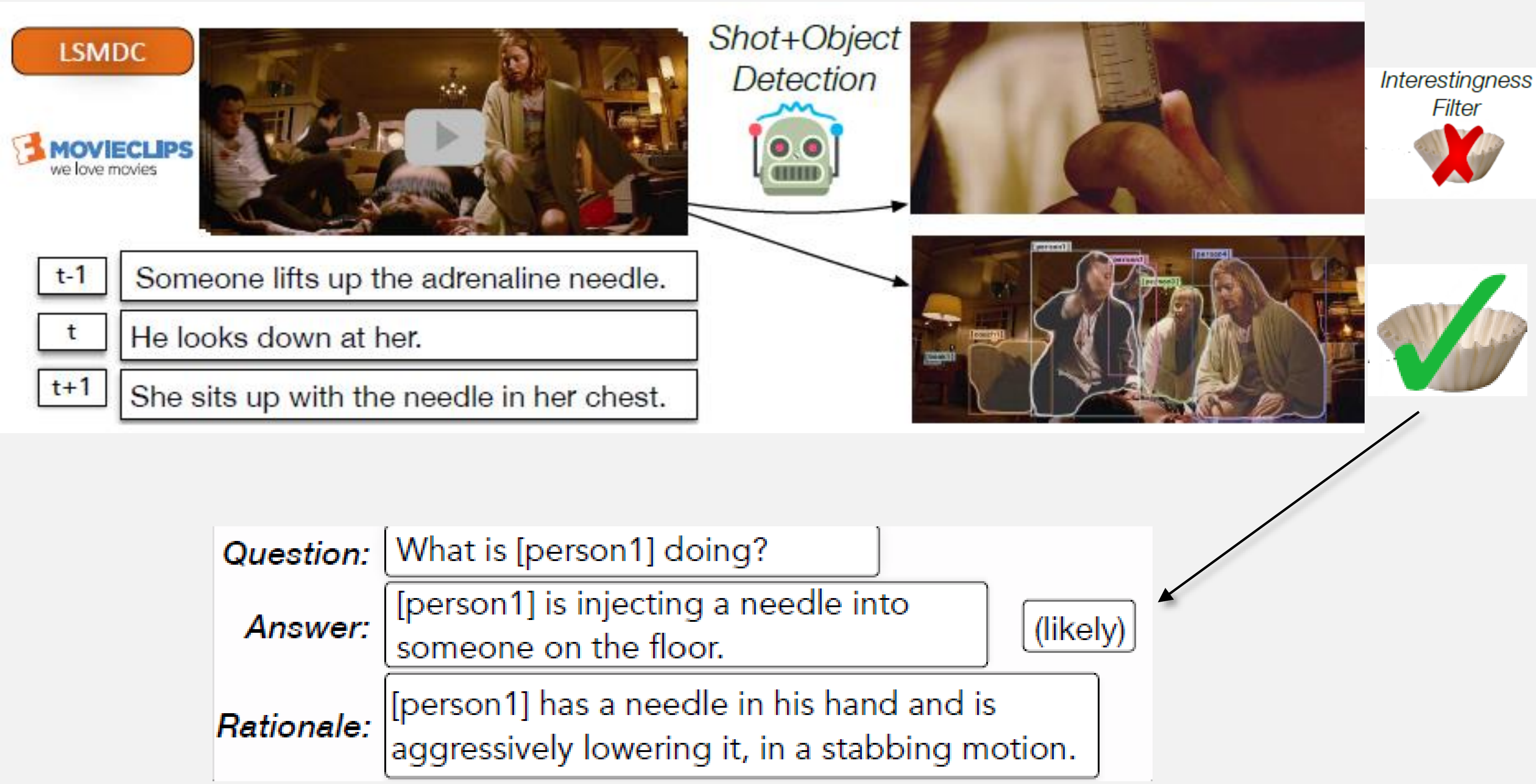
[Code]
https://github.com/rowanz/r2c/

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

05 Dataset

Collection

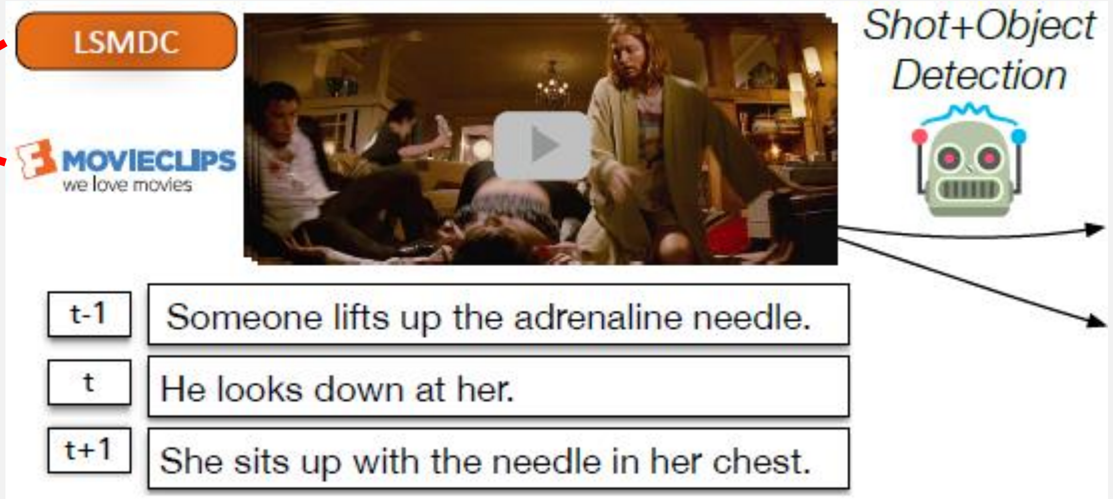
1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing



05 Dataset

Collection

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

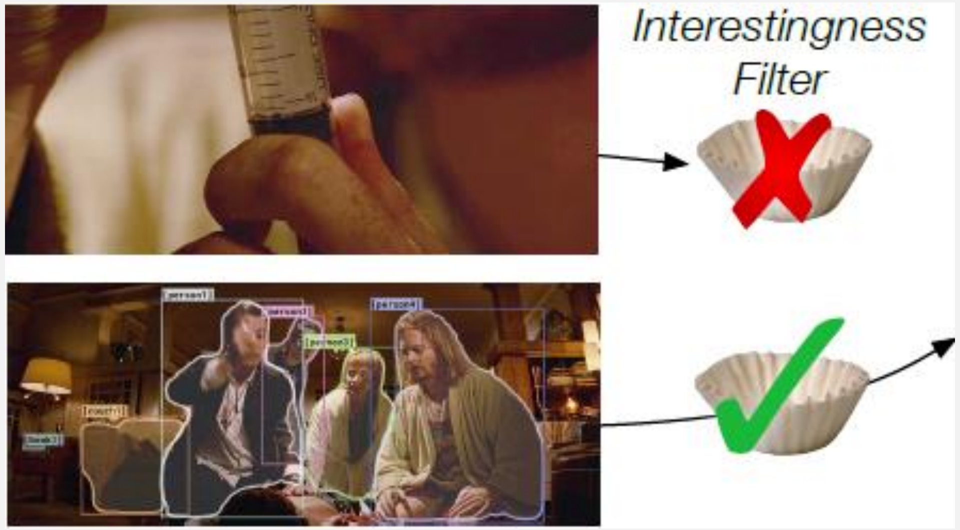


- Iterate through video clip (1 fps)
- Register shot boundary w/ 30 pixel mean difference in HSV
- Apply Mask-RCNN (threshold 0.7) -> 3 or more confidence tags

05 Dataset

Collection

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

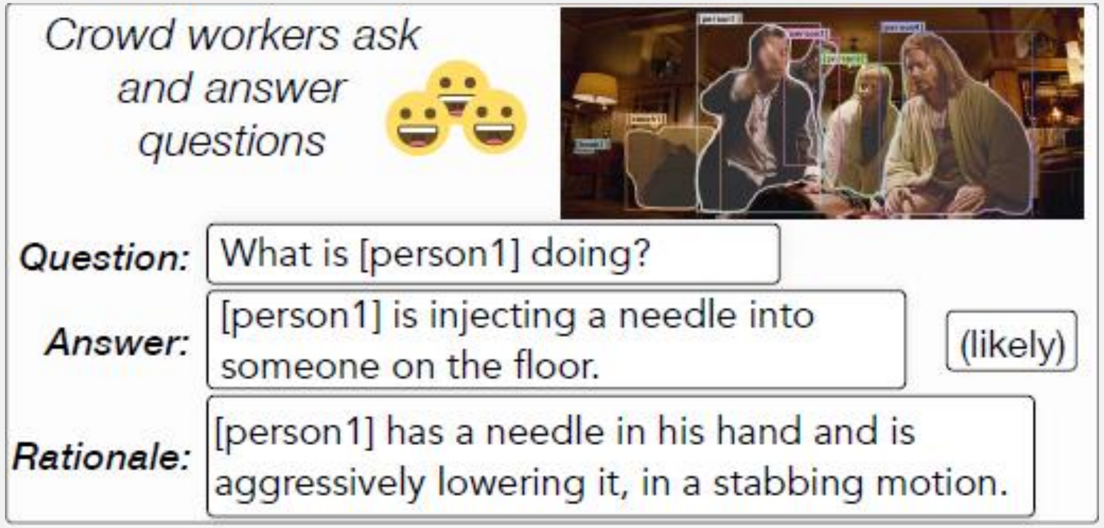


- At least two people
- Additional filtering process (uninteresting or blurry)
  1) Logistic regression (Interesting vs Uninteresting)
  2) ResNet 50 (Interesting vs Moderate vs Boring)

05 Dataset

Collection: Crowdsourcing

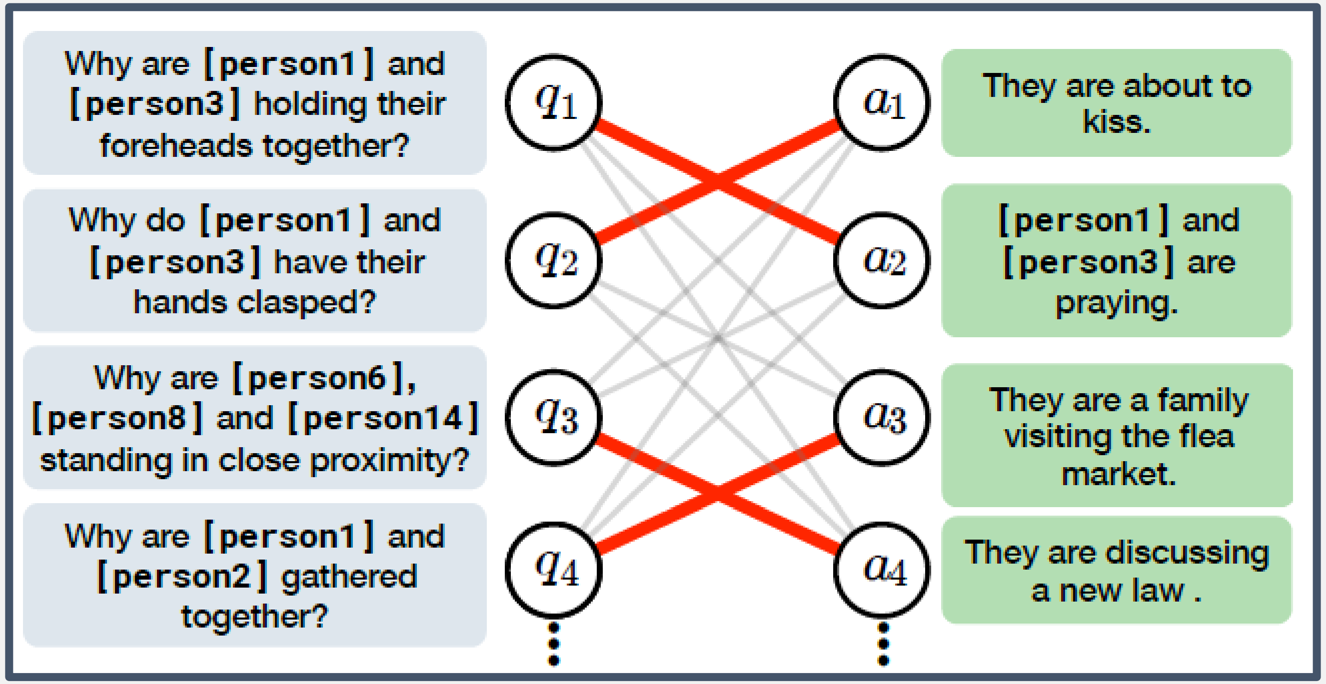
1. Shot Detection Pipeline
2. Interestingness Filter
3. Crowdsourcing



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

05 Dataset

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

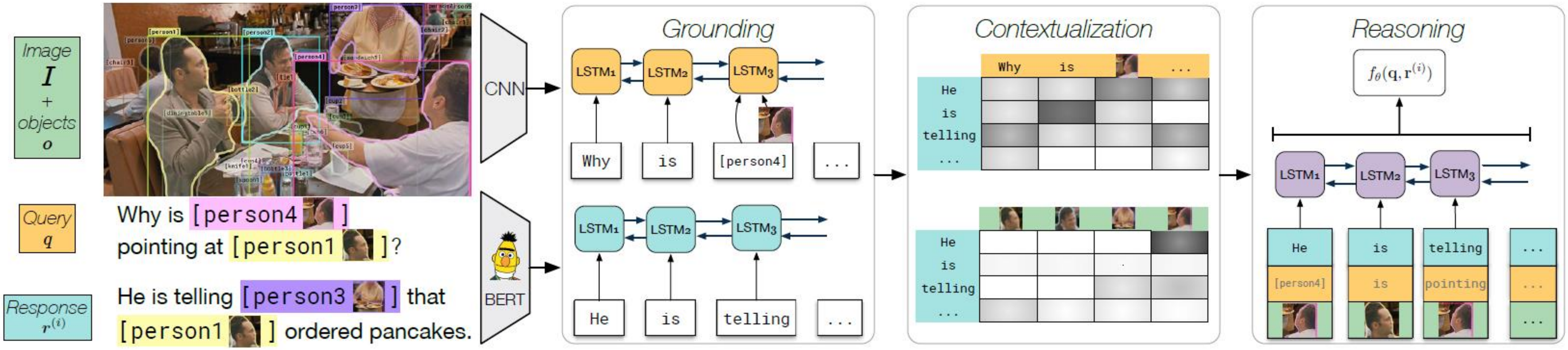
Training relevance btw R & Q

Similarity Model

Avoiding conflation

06 Model

R2C overview



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

06 Model

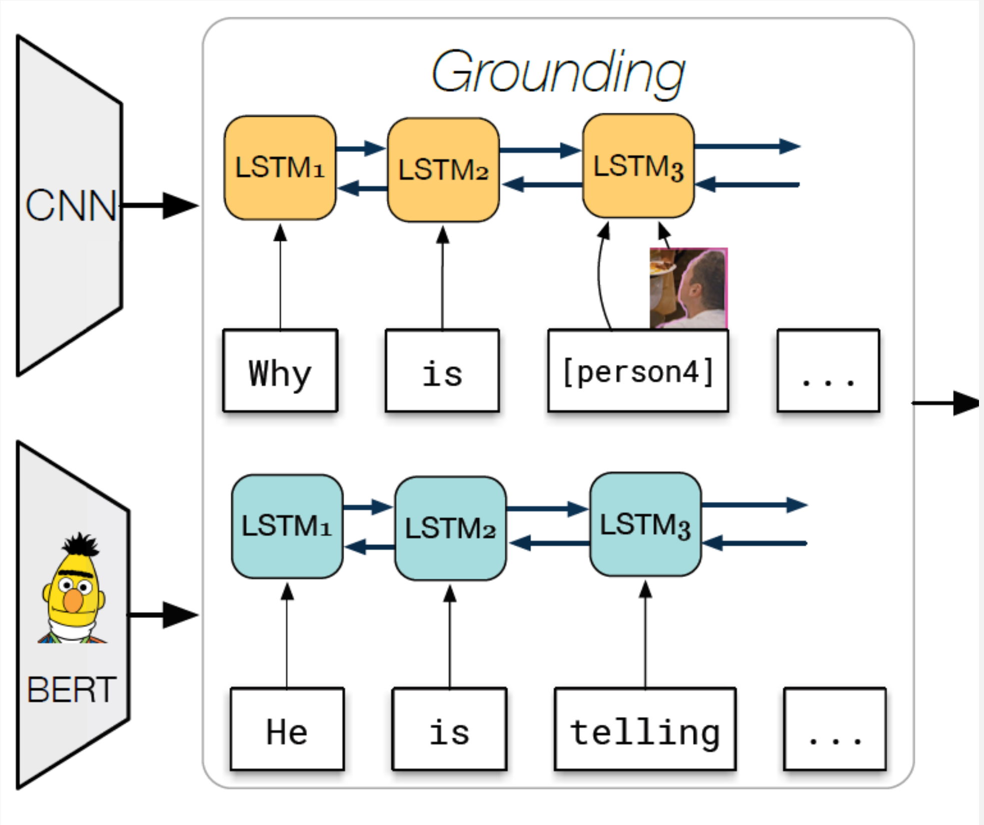
Feature embedding

Word -> representation by Bert

Object -> Feature embedding by CNN

06 Model

Grounding



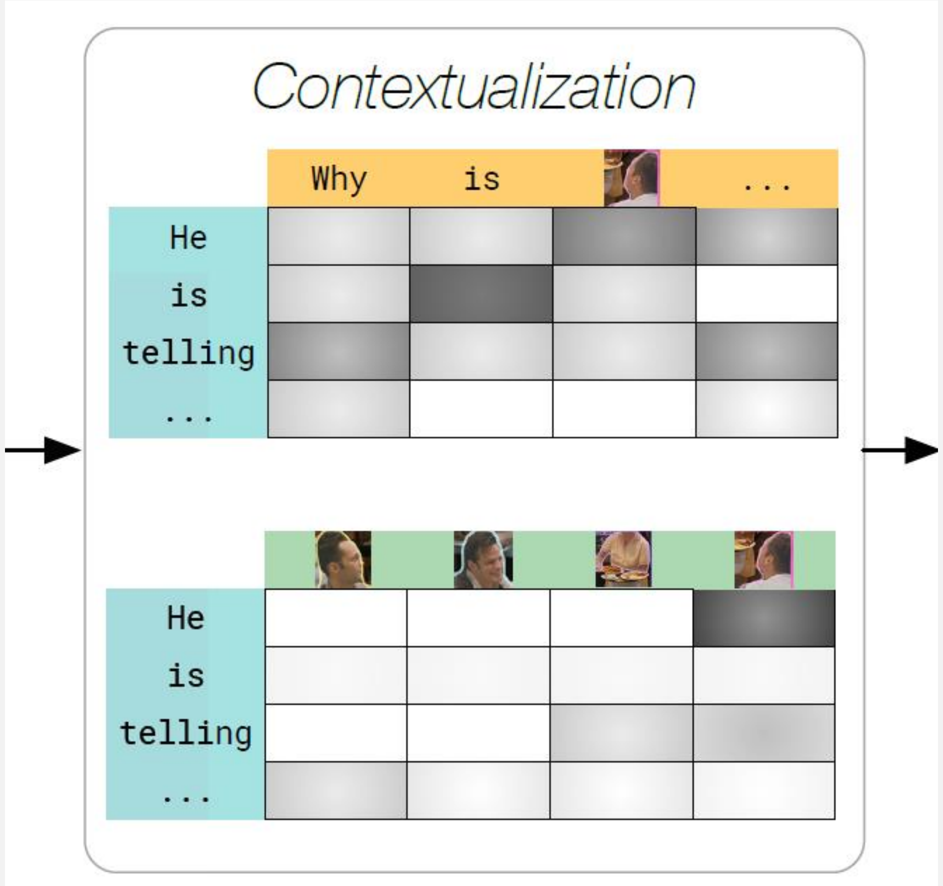
Learn joint image-language representation

Bidirectional LSTM

Parameter share

Output : r for response, q for query

Contextualization



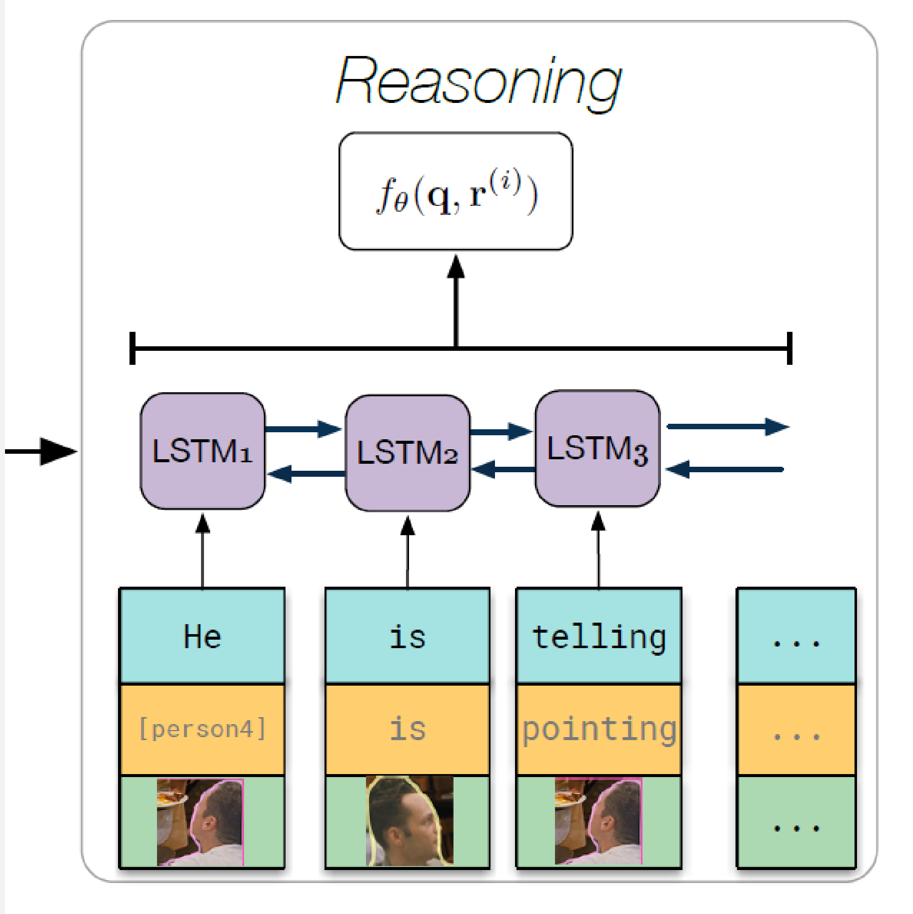
By attention mechanism

Attended query : $\hat{q}_i$

Attended object : $\hat{o}_i$

06 Model

Reasoning



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

Output : Probability

07 Settings

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*10^-4 to 10^-4

Batch size : 96
Epochs : 20

08 Baselines

Text-only Baselines

1)BERT

2)BERT(response only)

3)ESIM + ELMo

4)LSTM + ELMo

VQA Baselines

1)RevisitedVQA

2)Bottom-UP and Top-down attention

3)Multimodal Low-Rank Bilinear Attention(MLB)

4)Multimodal Tucker Fusion(MUTAN)

09 Experiment

Main Tasks

	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

09 Experiment

Ablation

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

Strong textual representations are crucial to VCR

10 Conclusion

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