The GUESSWHAT?! and FiLM Stories

Presented by L. Celotti, for IGLU

Master Class HLU, Paris 11 Apr. 2018

IGLU
Interactive Grounded Language Understanding

Université de Lille

Université de Montréal

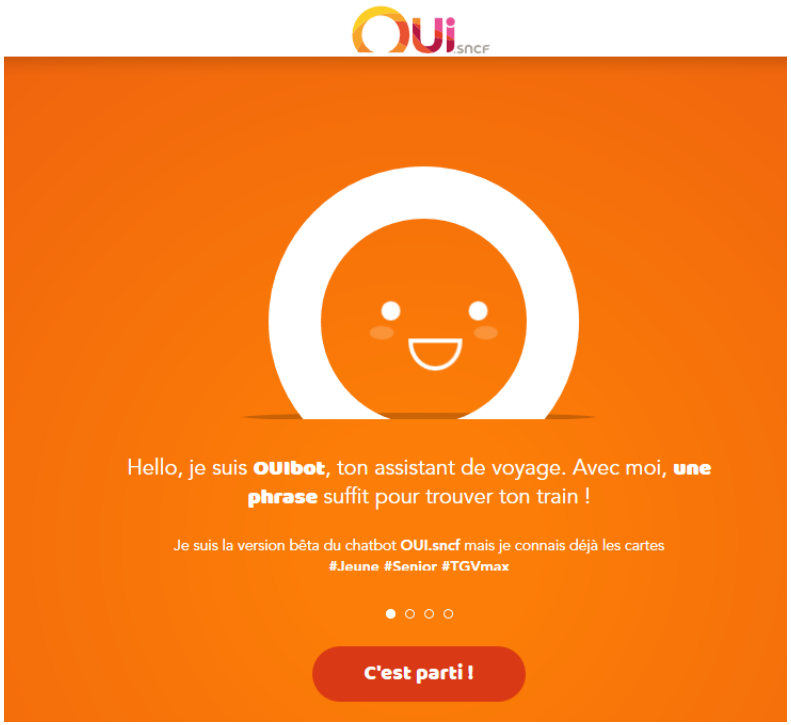
criteo

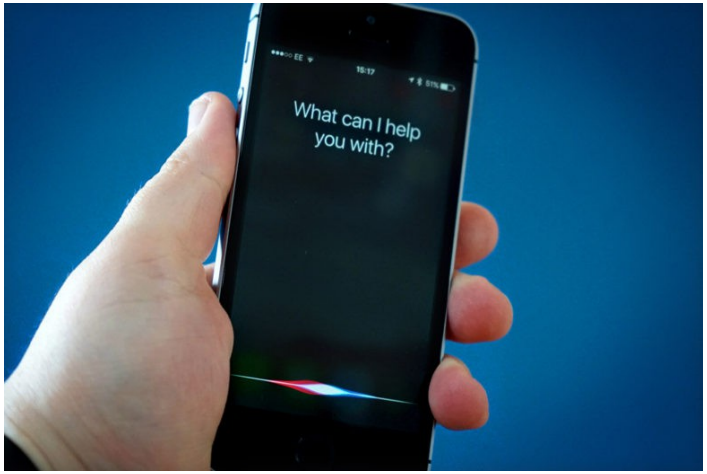
Research at Google

Motivation

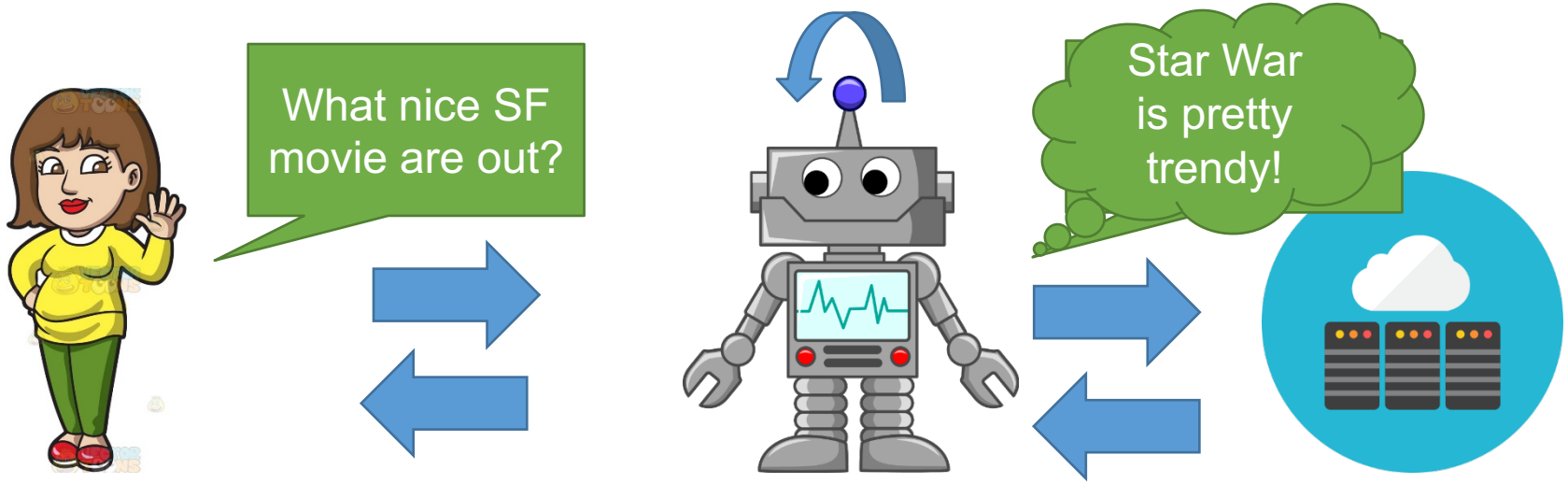
Dialogue System

Dialogue System





Dialogue System



Step 1 : Speech processing

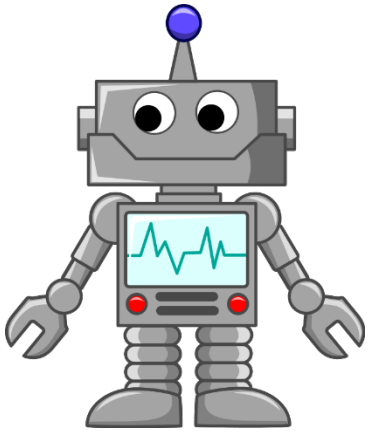
Step 2 : Language understanding

Step 3 : Query information

Step 4 : Language generation

Step 5 : Speech generation

Dialogue System



Expert system
(rule based)

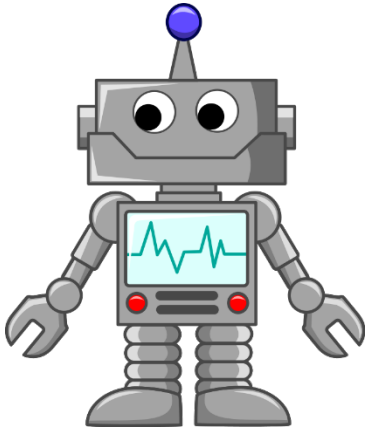
```
If user.say() == “Hello”:
  -> “Say Hello”

if user.say() == “What ${type}
movie are out?”:
  -> Browse: $type in IMD
  -> Say: “Do you know ${res}?”
```

Generative model
(machine learning)

```
- Project question into high-
dimensional space
- Generate word by word answer from
this high-dimensional space
```

Dialogue System



Chit-chat dialogue

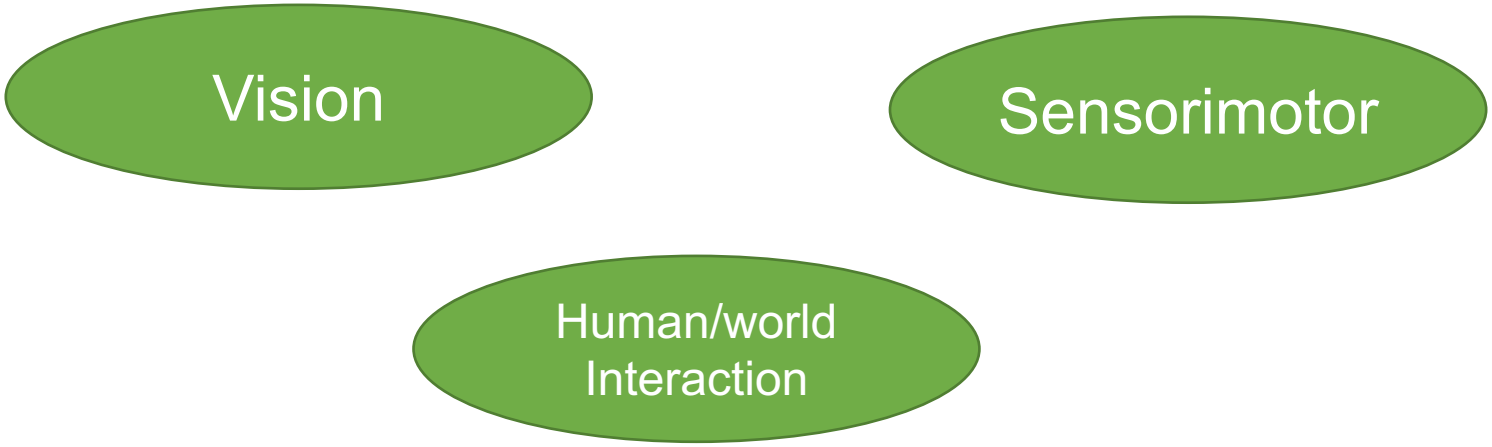
```
B: How are you today?
H: Fine! Thank you!
B: Did you hear the news? A new
baby panda is born in the city zoo!
H: ohhhh, so lovely!
```

Goal-Oriented dialogue

```
H: I want to book a plane to London.
B: Sure, when do you want to leave?
H: Tomorrow, in the morning
B: There is plane at 9am etc...
```

Grounding Language

The grounding problem is related to the problem of how words (symbols) get their meanings. [1]

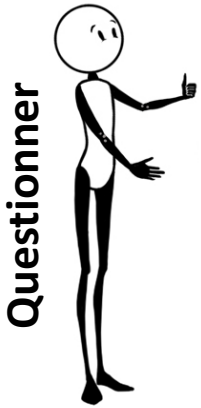


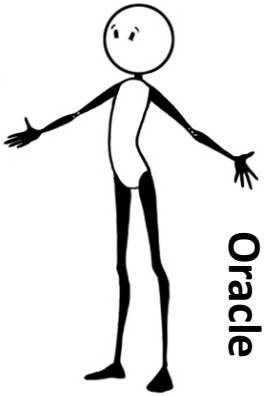
[1] H. Clark and E. Schaefer. Contributing to discourse. Cognitive Science , 13(2):259–294, 1989

GuessWhat?!

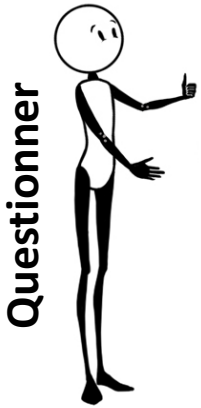
The Game

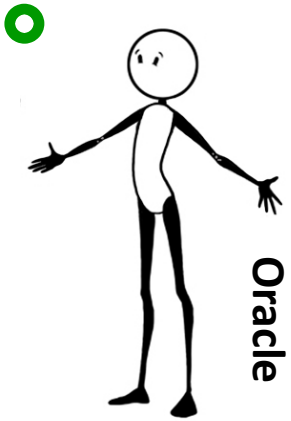
GuessWhat?! Game



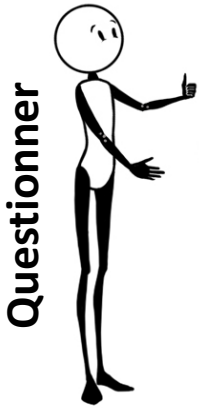


GuessWhat?! Game

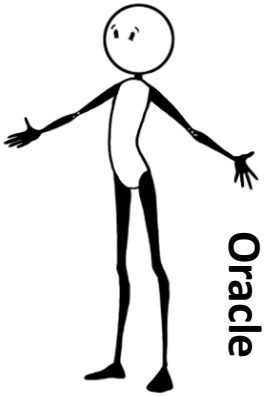




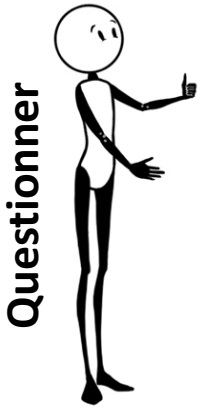
GuessWhat?! Game



Is it a vase ?

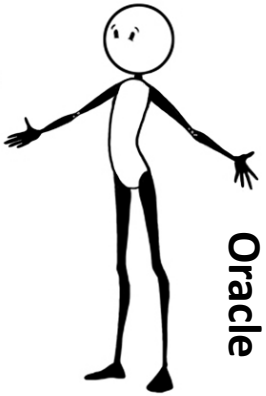


GuessWhat?! Game

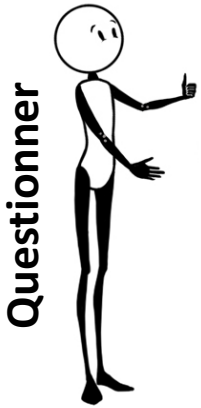


Is it a vase ?

Yes



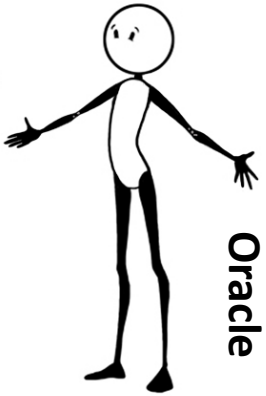
GuessWhat?! Game



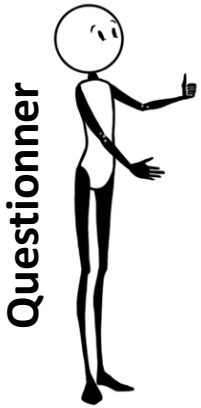
Is it a vase ?

Yes

Is it in the front row?

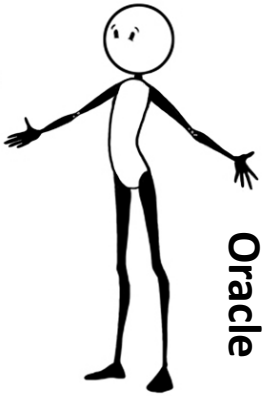


GuessWhat?! Game

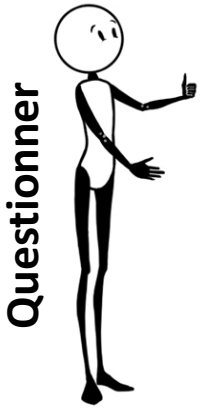


Is it a vase ? Yes

Is it in the front row? Yes



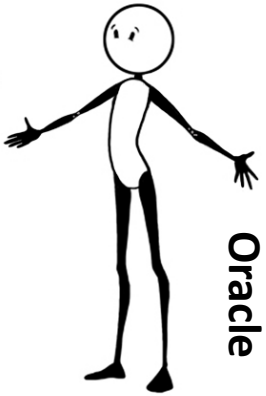
GuessWhat?! Game



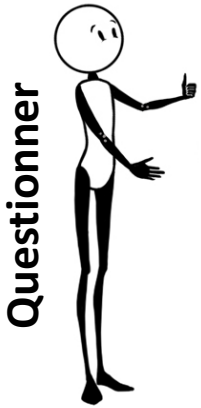
Is it a vase ? Yes

Is it in the front row? Yes

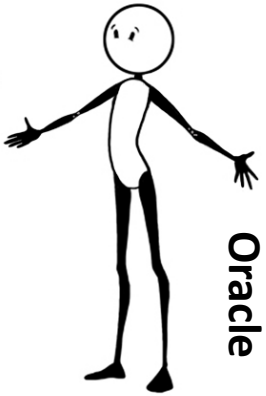
Does it have some red on it?



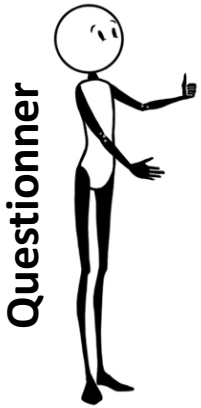
GuessWhat?! Game



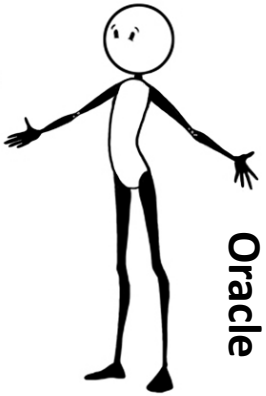
Is it a vase ?	Yes
Is it in the front row?	Yes
Does it have some red on it?	No



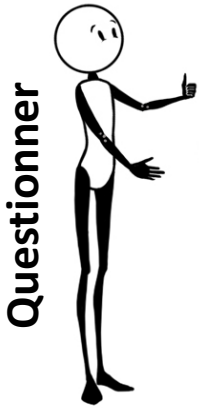
GuessWhat?! Game



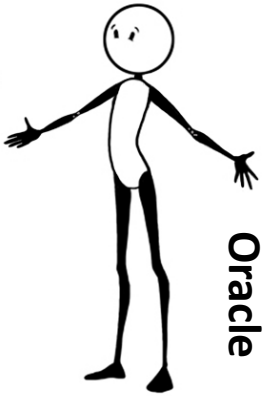
Is it a vase ?	Yes
Is it in the front row?	Yes
Does it have some red on it?	No
Is it the second vase from the right?	



GuessWhat?! Game

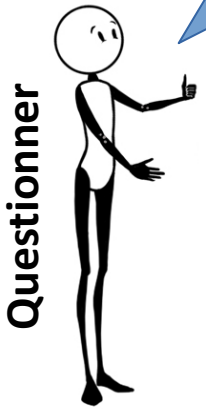


Is it a vase ?	Yes
Is it in the front row?	Yes
Does it have some red on it?	No
Is it the second vase from the right?	Yes

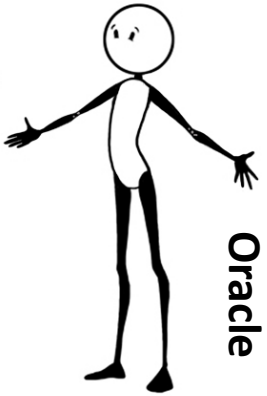


GuessWhat?! Game

I found it!

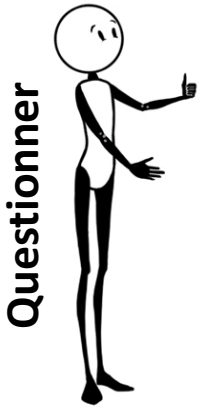


Is it a vase ?	Yes
Is it in the front row?	Yes
Does it have some red on it?	No
Is it the second vase from the right?	Yes

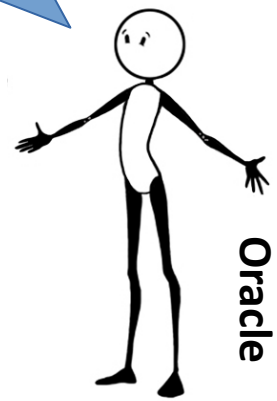


GuessWhat?! Game

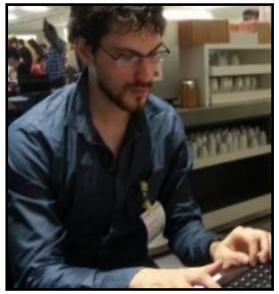
Correct!



Is it a vase ?	Yes
Is it in the front row?	Yes
Does it have some red on it?	No
Is it the second vase from the right?	Yes



The Team



Florian Strub
Univ. of Lille, Inria

Harm de Vries
Univ. of Montreal

Sarath Chandar
Univ. of Montreal

Bilal Piot
Univ. of Lille, Inria

Hugo Larochelle
Google Brain

Jérémie Mary
Criteo

Olivier Pietquin
Univ. of Lille, Inria

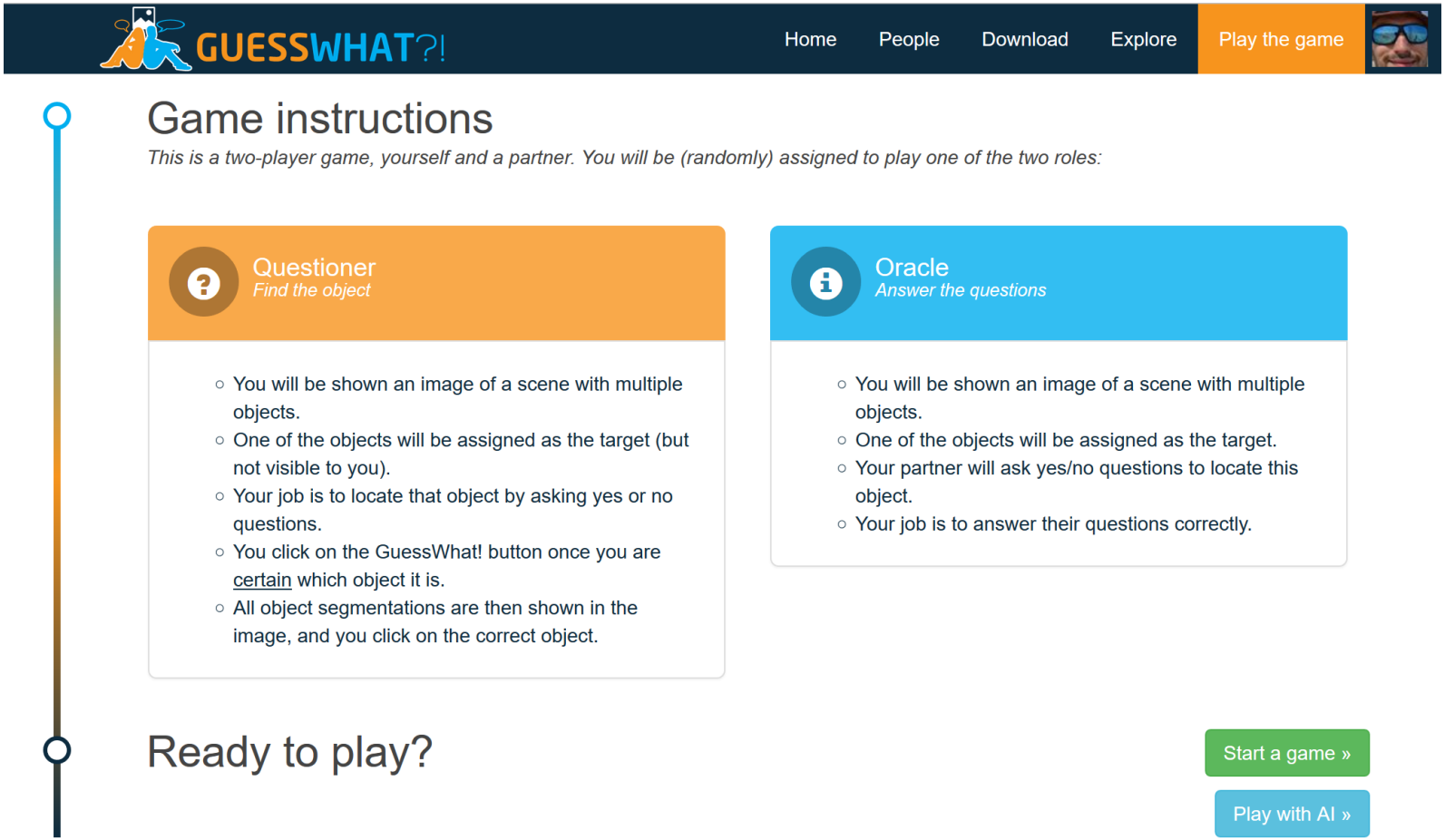
Aaron Courville
Univ. of Montreal

GuessWhat?! Game

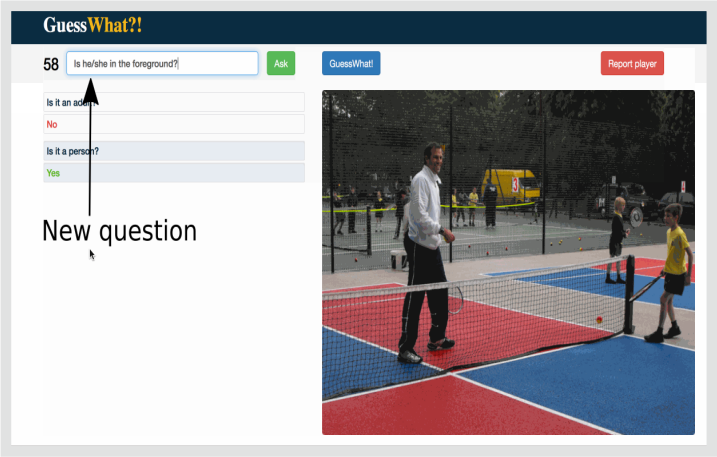
Wait! Is it really a difficult task?

- No rule-based
  - High dimensional input/output spaces
- Contextual
  - Each game has a new image
- End-2-end:
  - No image processing: input = raw pixel
  - No language processing: input/output = words
- Multi-objective
  - Asking meaningful questions
  - Asking coherent sequence of questions

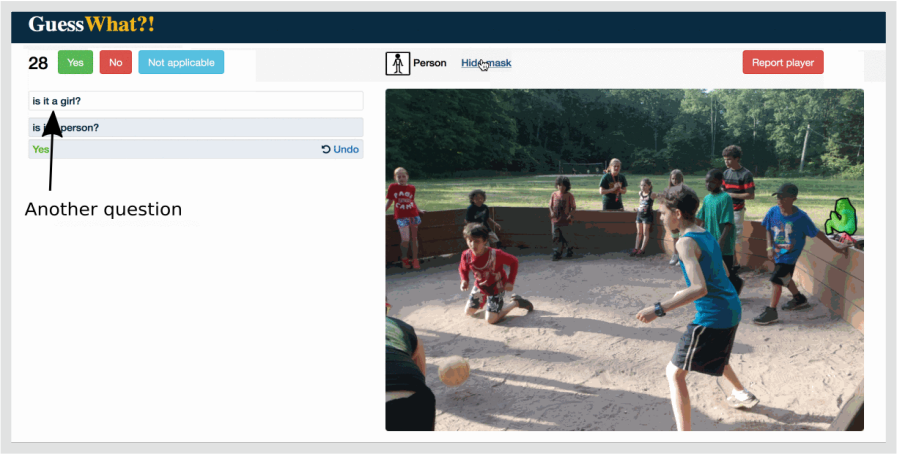
GuessWhat?! Dataset



GuessWhat?! Dataset



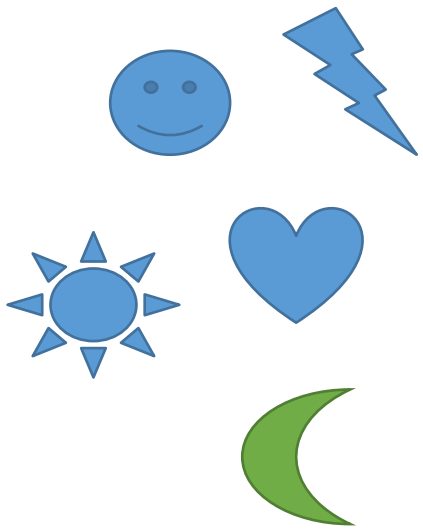
Questioner



Oracle

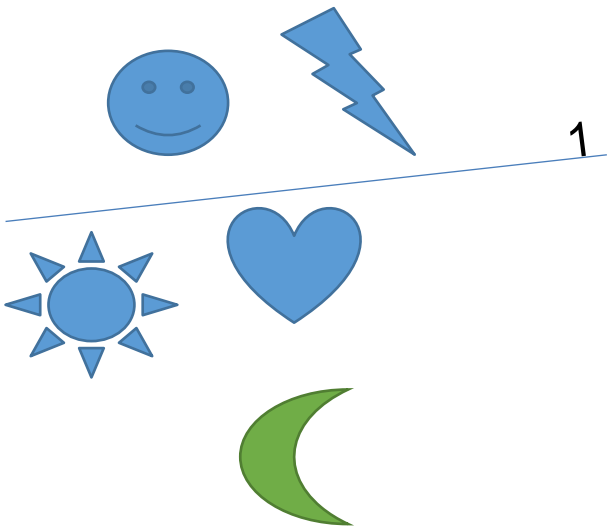
- More than 10k players
- Top 10 players completed more than 2500 games
- Highest game time was 70+ hour

GuessWhat?! Dataset



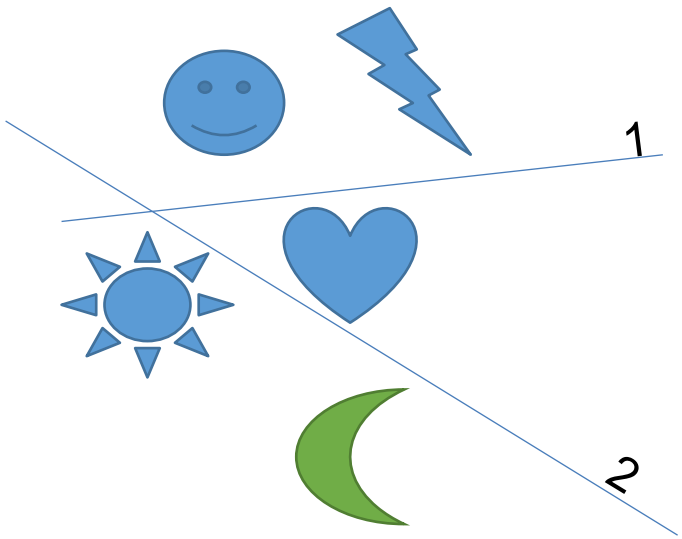
Optimal search in the object space is a binary search i.e. removing half of the objects at each step (log(n) complexity).

GuessWhat?! Dataset



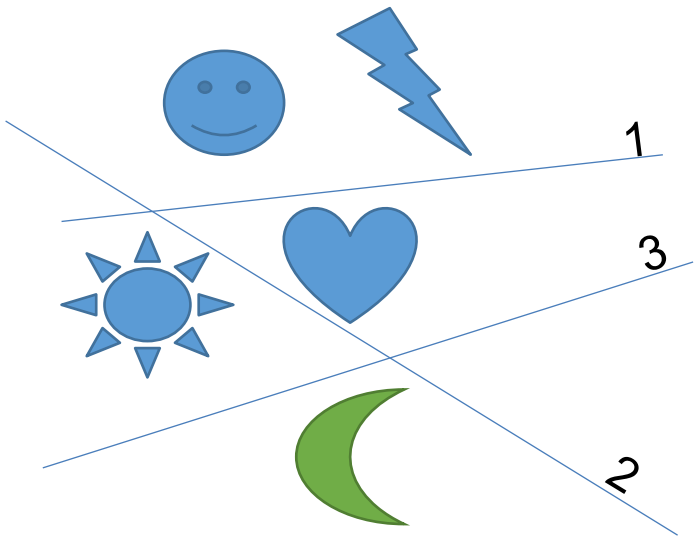
Optimal search in the object space is a binary search i.e. removing half of the objects at each step (log(n) complexity).

GuessWhat?! Dataset



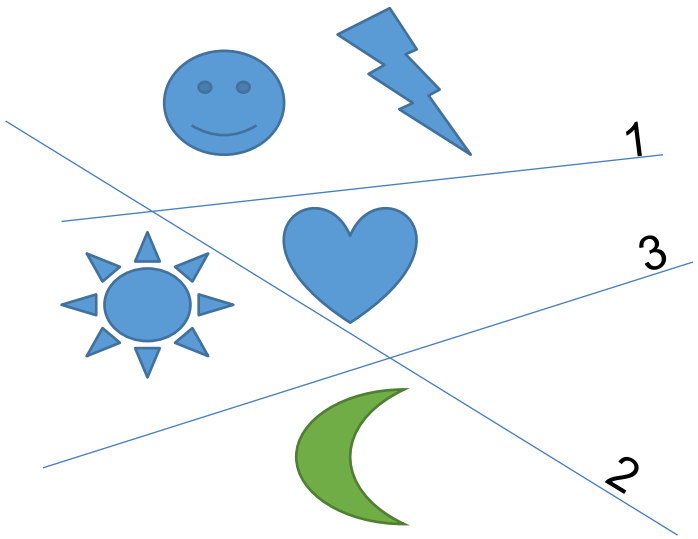
Optimal search in the object space is a binary search i.e. removing half of the objects at each step (log(n) complexity).

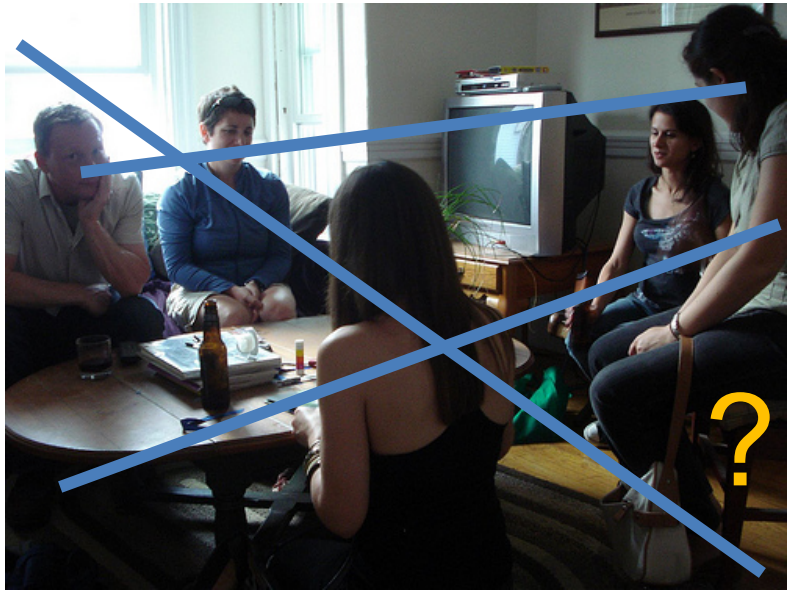
GuessWhat?! Dataset



Optimal search in the object space is a binary search i.e. removing half of the objects at each step (log(n) complexity).

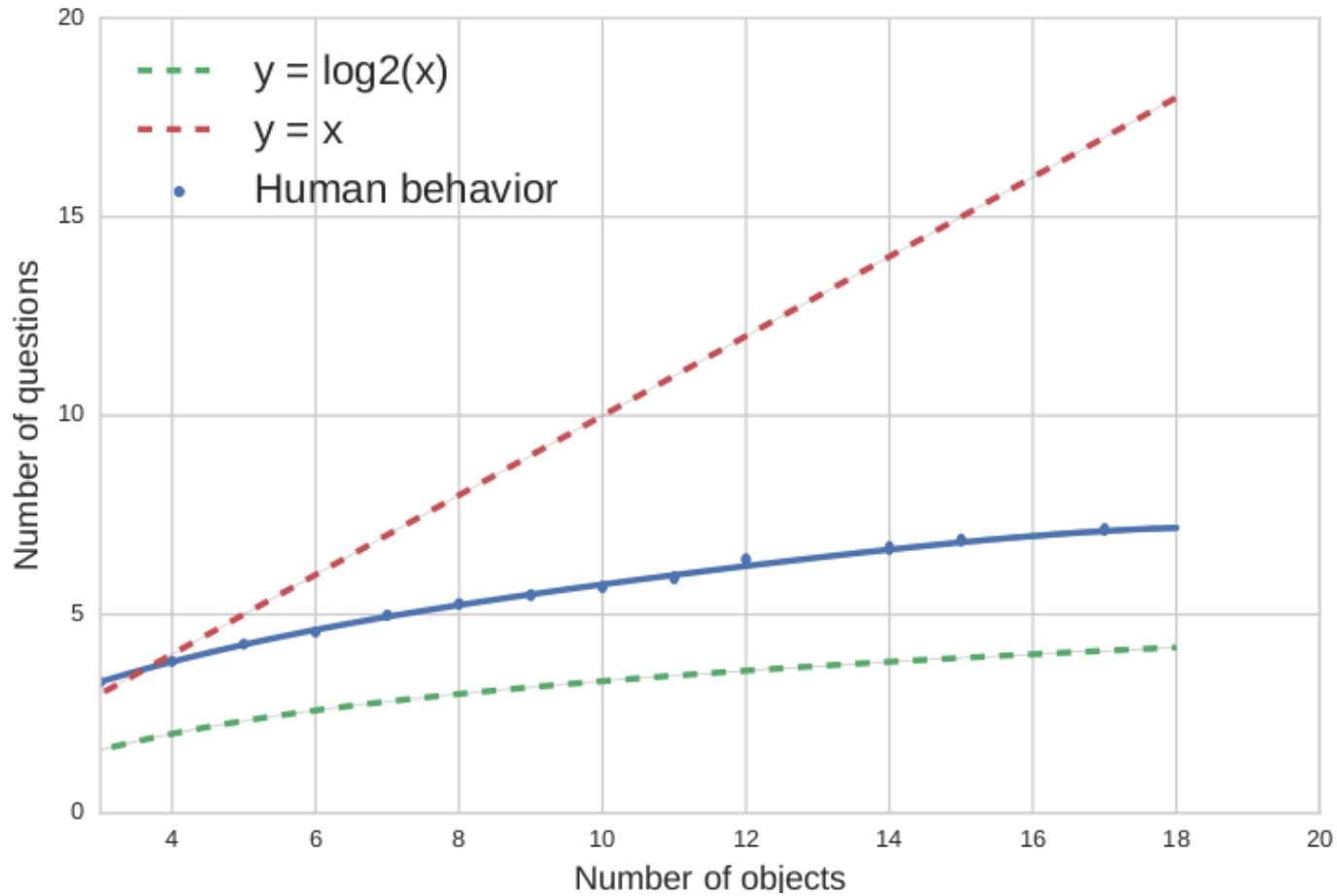
GuessWhat?! Dataset

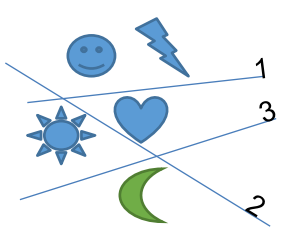




Optimal search in the object space is a binary search i.e. removing half of the objects at each step (log(n) complexity).

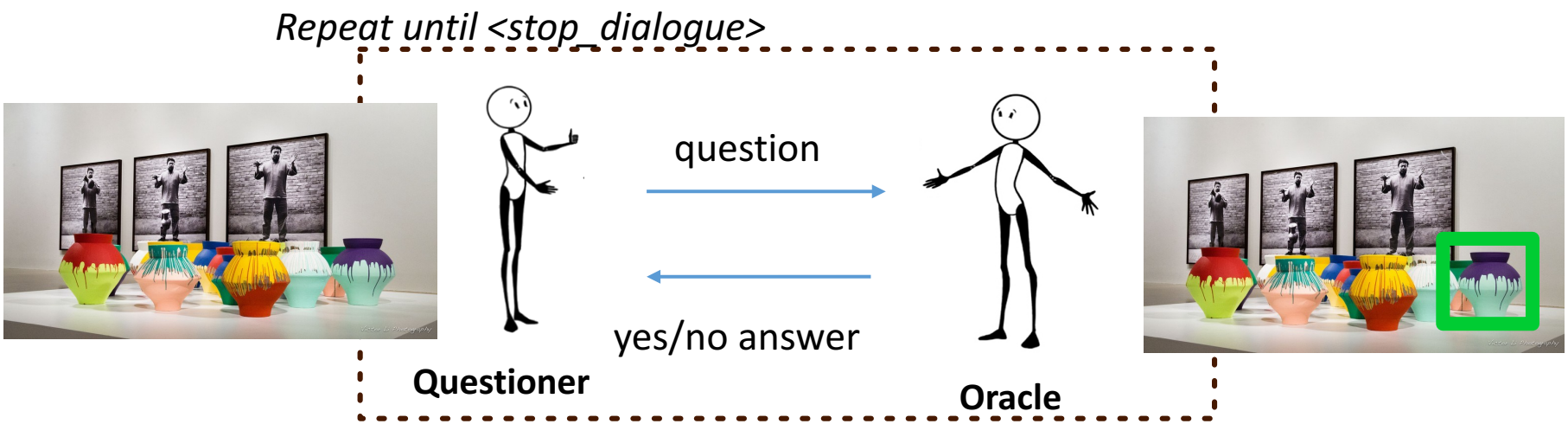
GuessWhat?! Dataset



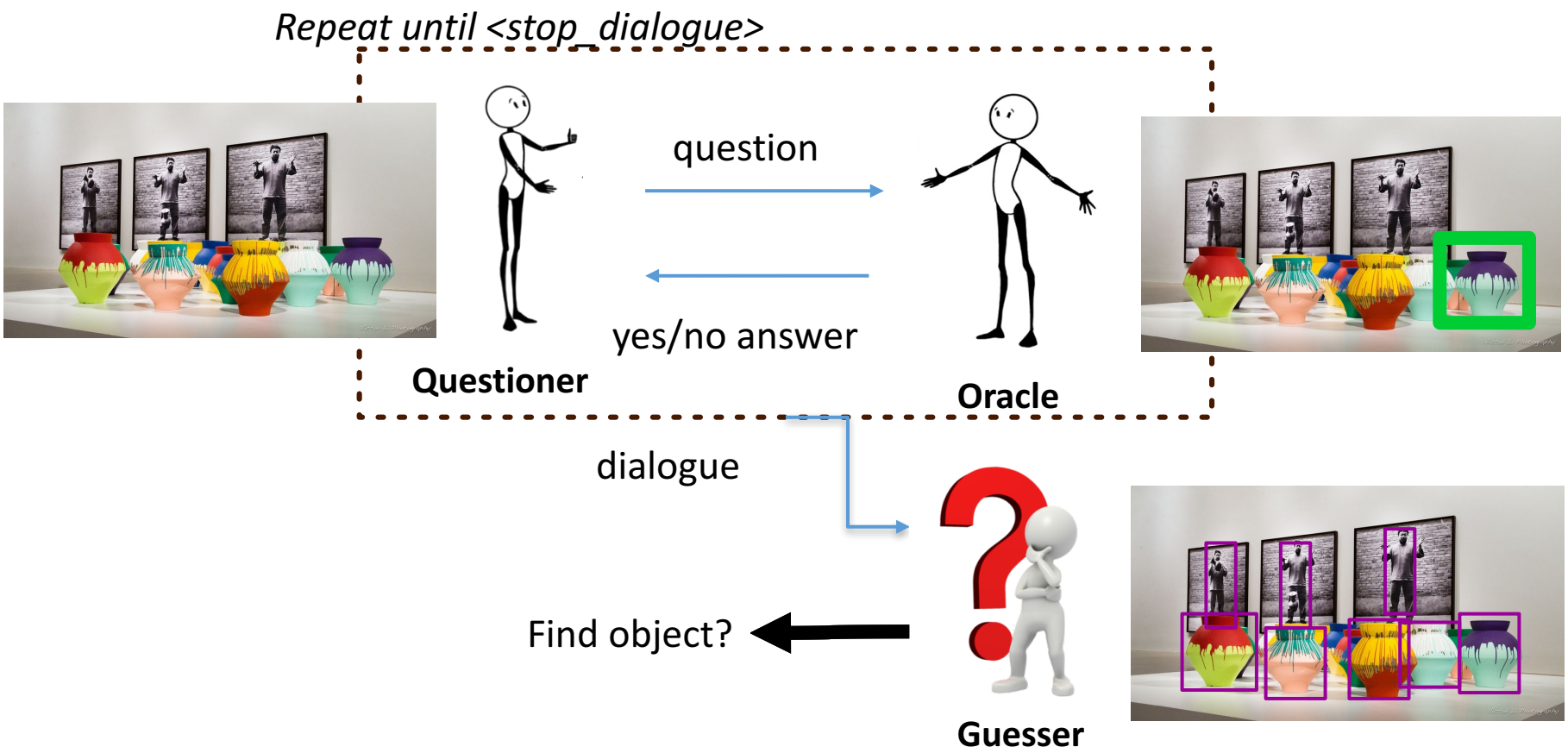


The Models

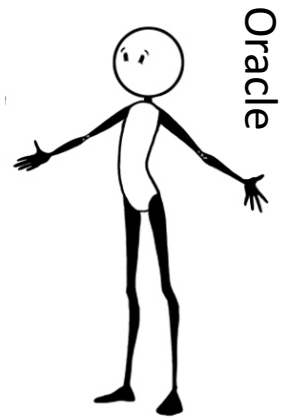
Models

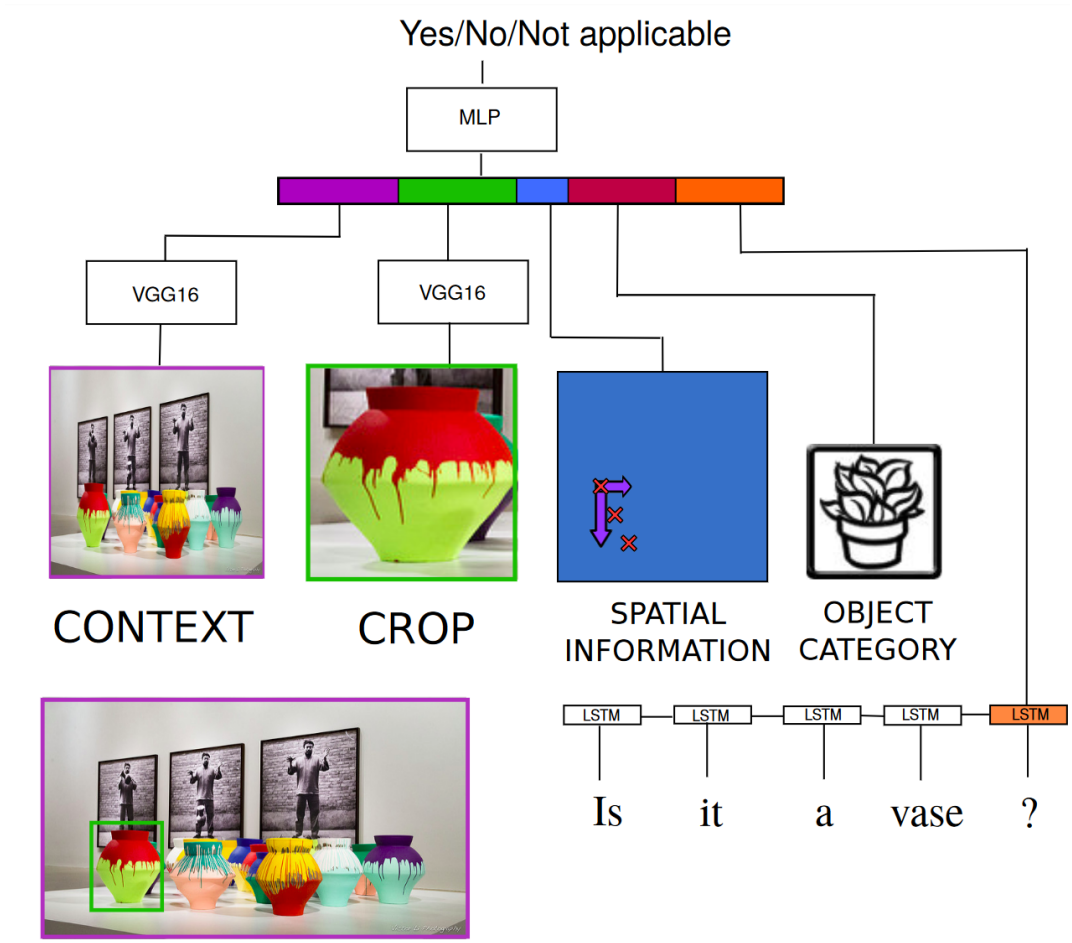

Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle

Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
dialogue
Find object?
Guesser

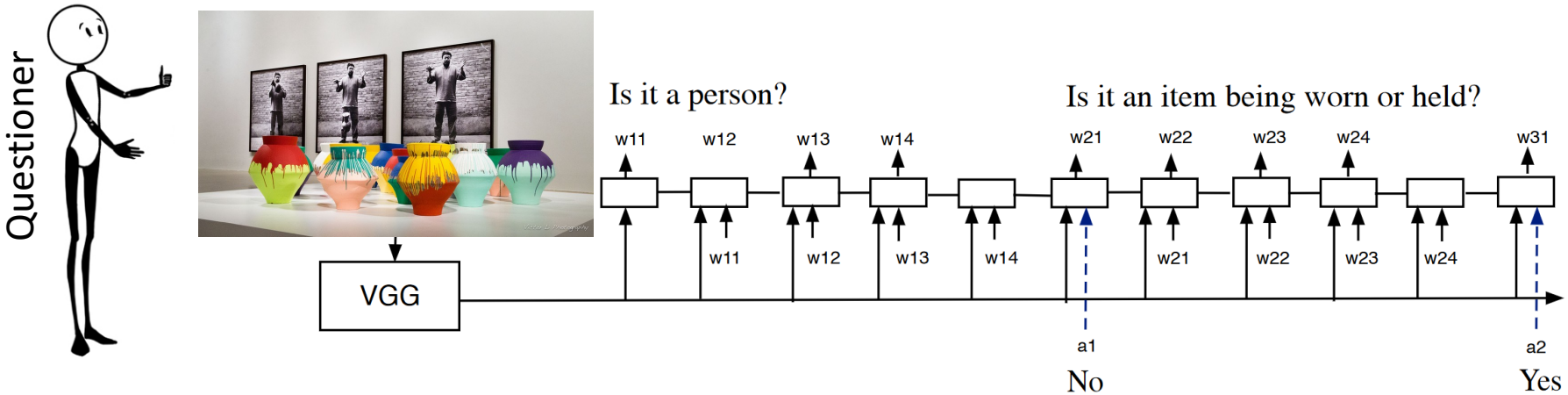
Models





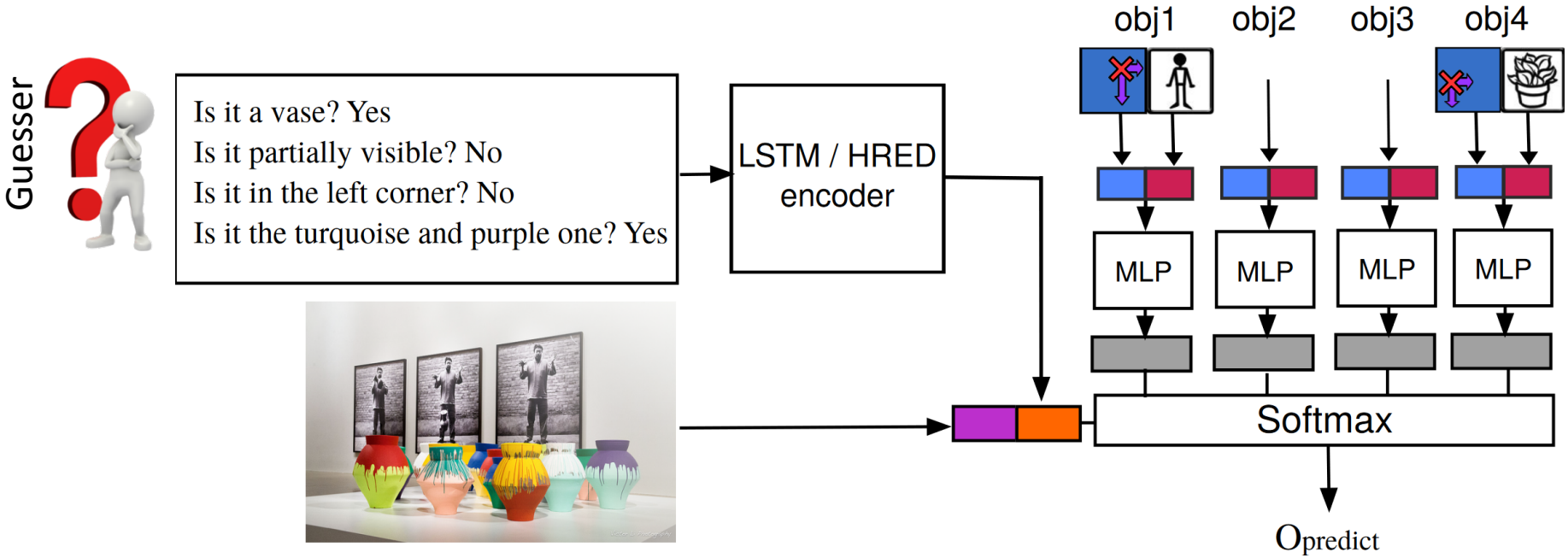
79.5% accuracy

Models



47.1% accuracy

Models



63.8% accuracy

Where SL fails and RL not

Image	Beam Search	REINFORCE
	Is it a person ? no Is it a ball ? no Is it a ball ? no Is it a ball ? no Is it a ball ? no Failure (blue bat)	Is it a person ? no Is a glove ? no Is an umbrella ? no Is in the middle ? no On a person? no is it on on far right? yes Success (red chair)
	Is it a person ? yes Is it the one in front ? yes Is it the one on the left ? no Is it the one in the middle with the red umbrella ? yes Is it the one to the right of the girl in ? no Failure (umbrella)	Is it a person ? yes Is it girl in white ? yes Success (girl)

Where RL fails too

- Tend to loose grammatical structure: Is it a cat? → cat ?
- Environment use limited modalities → limits the number of strategy
- Stopping far from optimal !

Is it a suitcase? Yes
Is it in the left side? Yes

is it a person ? yes
is it in left ? no
is it in right ? no
is it in middle ? yes
is it in front ? yes
is it in middle ? yes

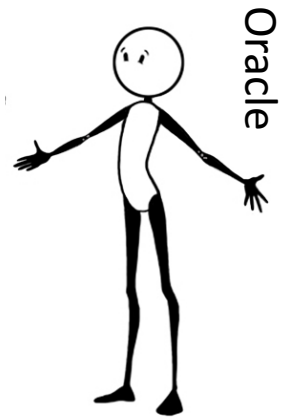
Language Bias

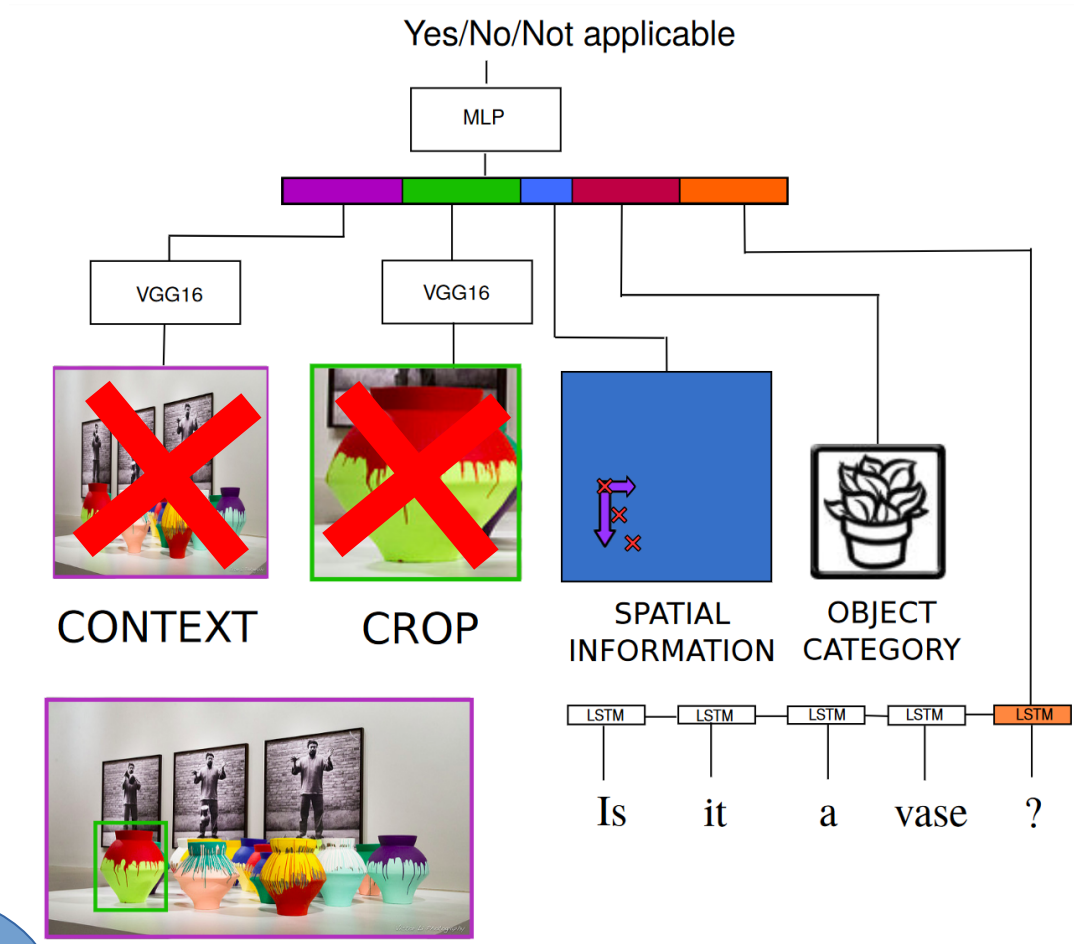
What is Language Bias?

- How many zebra are there in the image? 2 – 3
- Is the man wearing glasses? Yes
- What color is the banana? Yellow

Most of the SOTA models learn (with more or less success) language bias!!

Visual Input worsened results

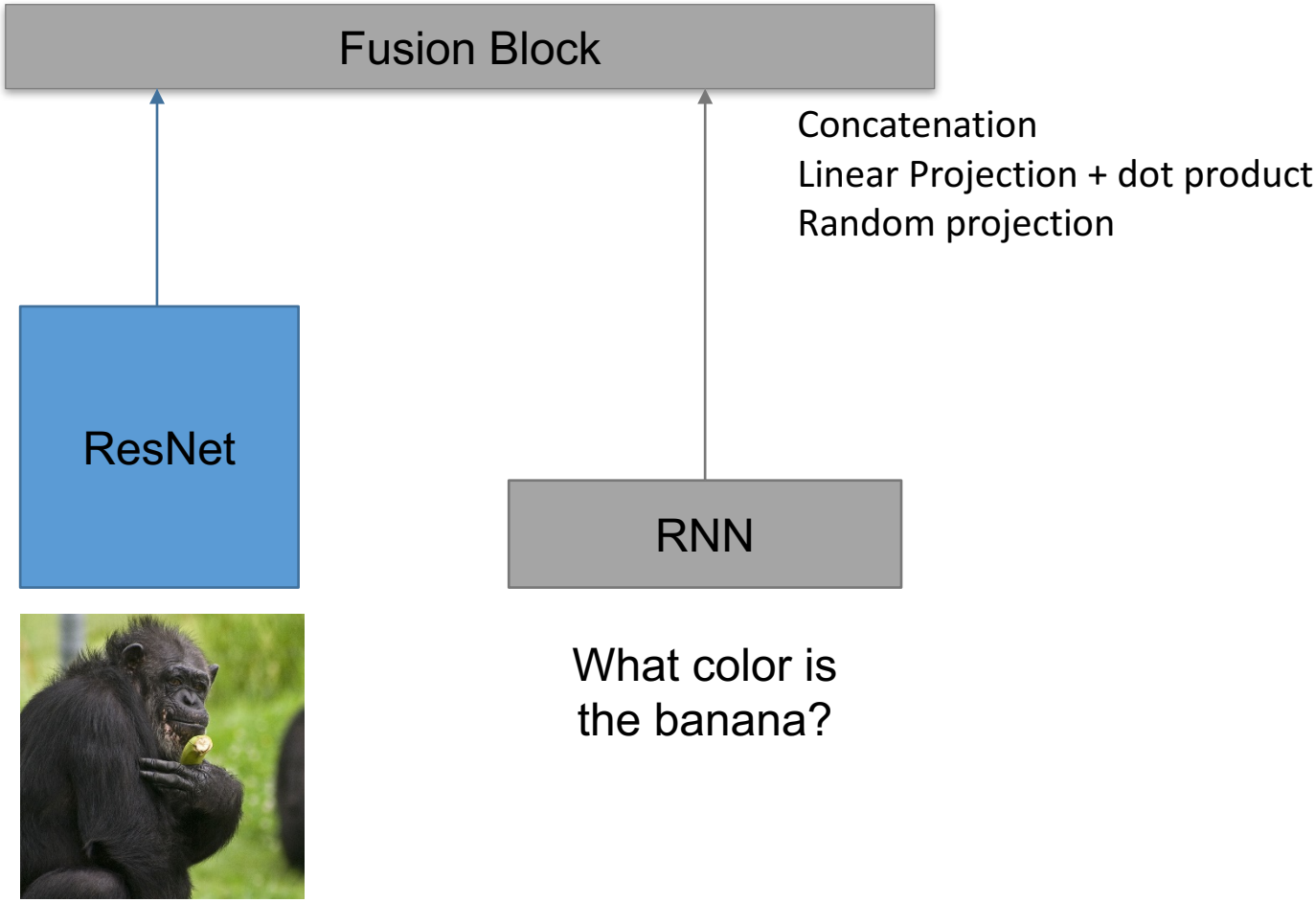




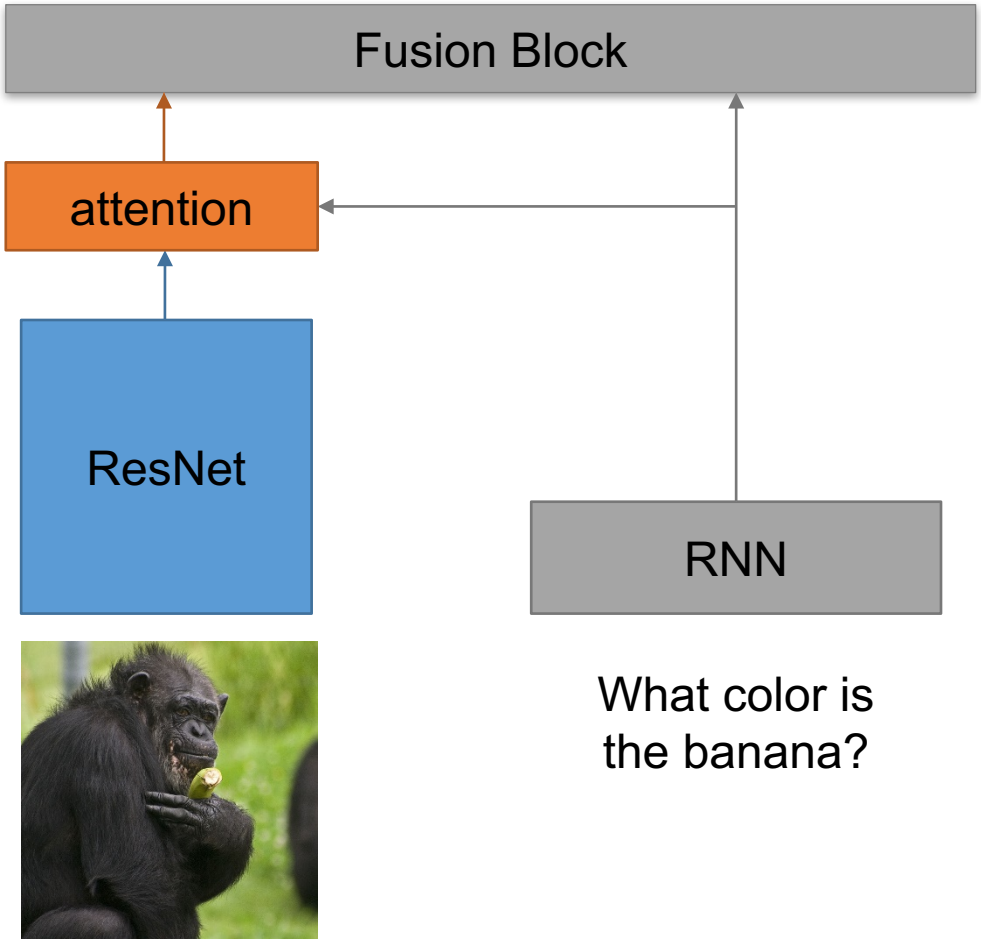
78.9% accuracy

79.5% accuracy

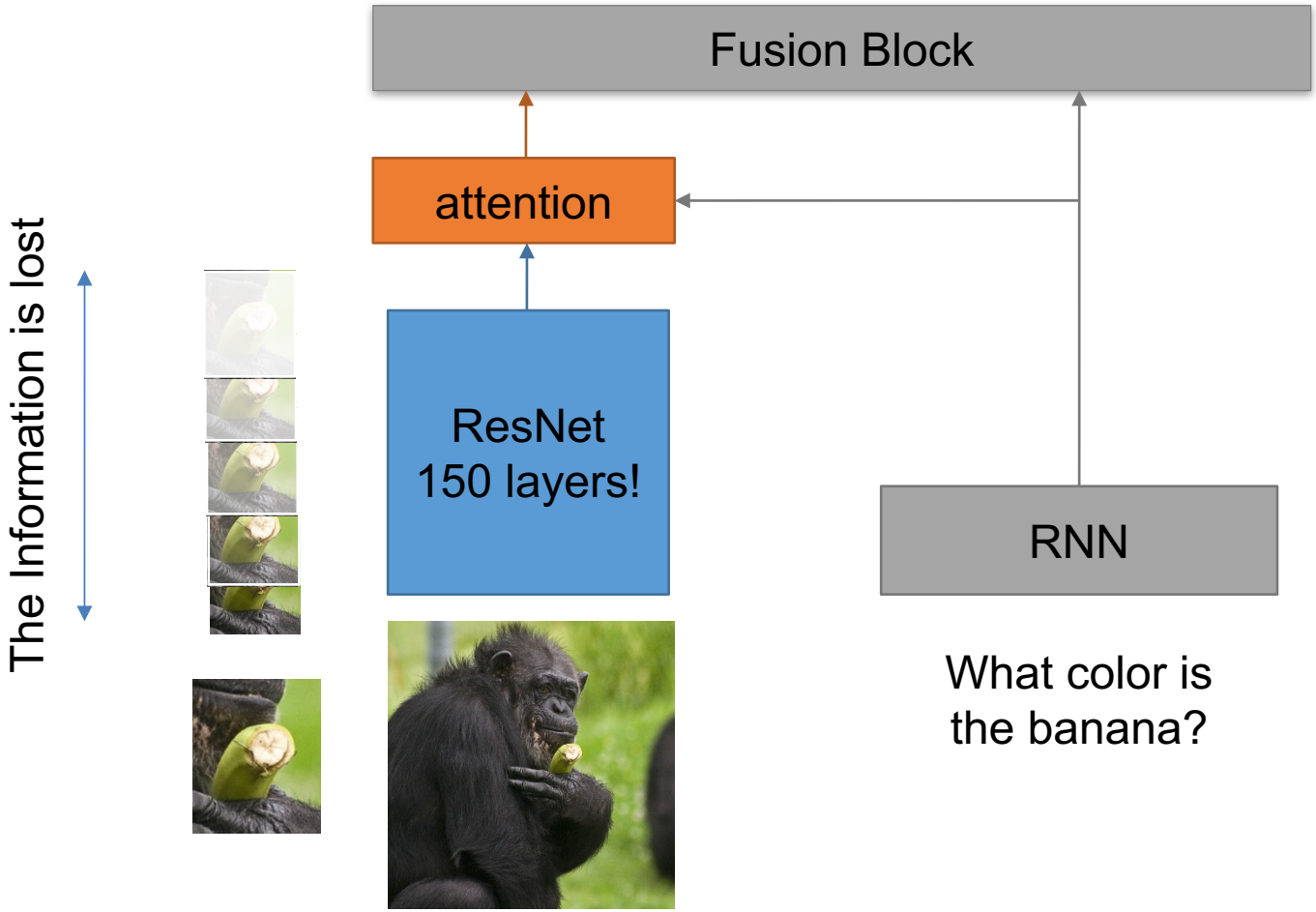
Concatenation wasn't enough



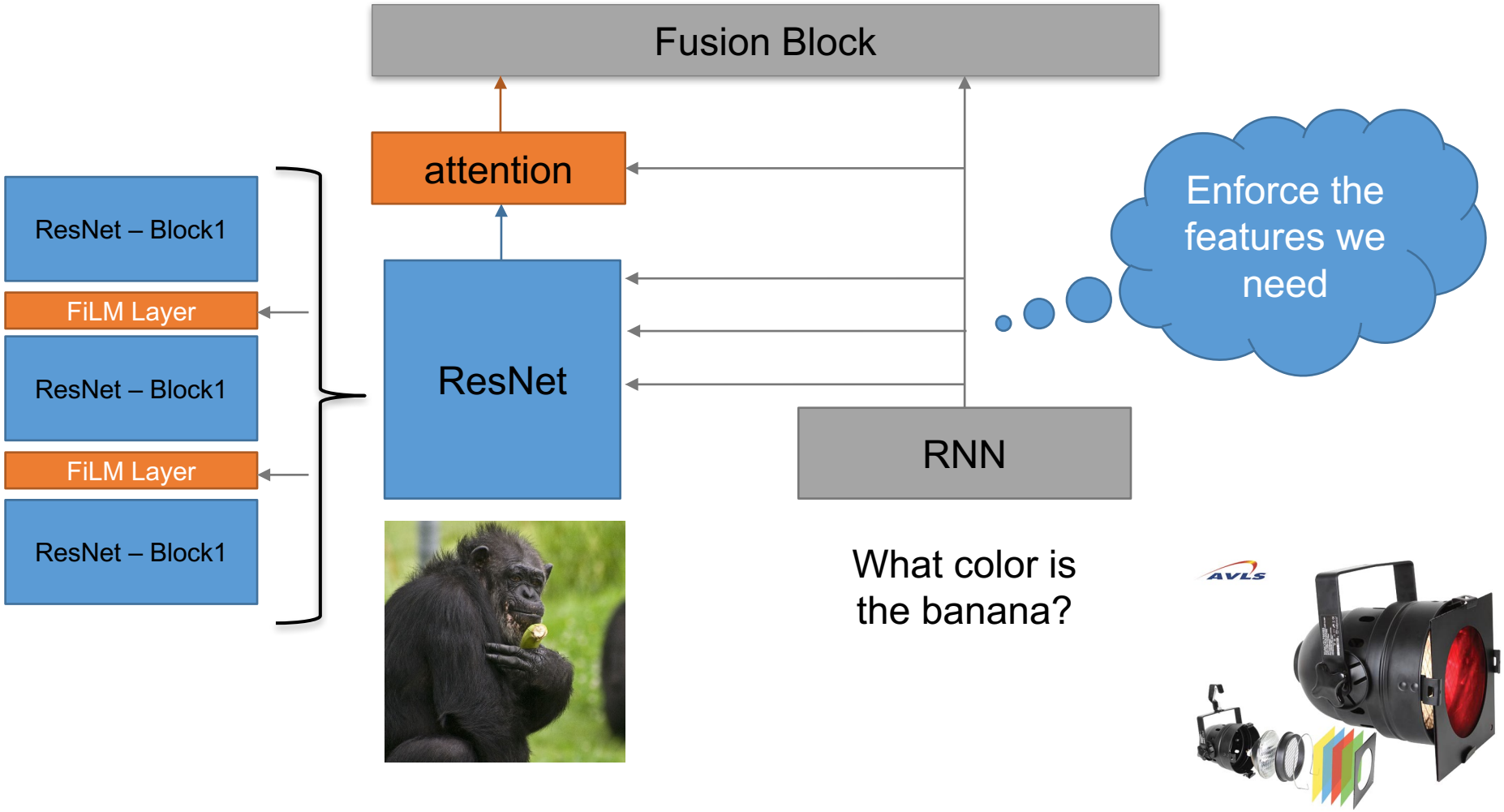
Attention isn't enough



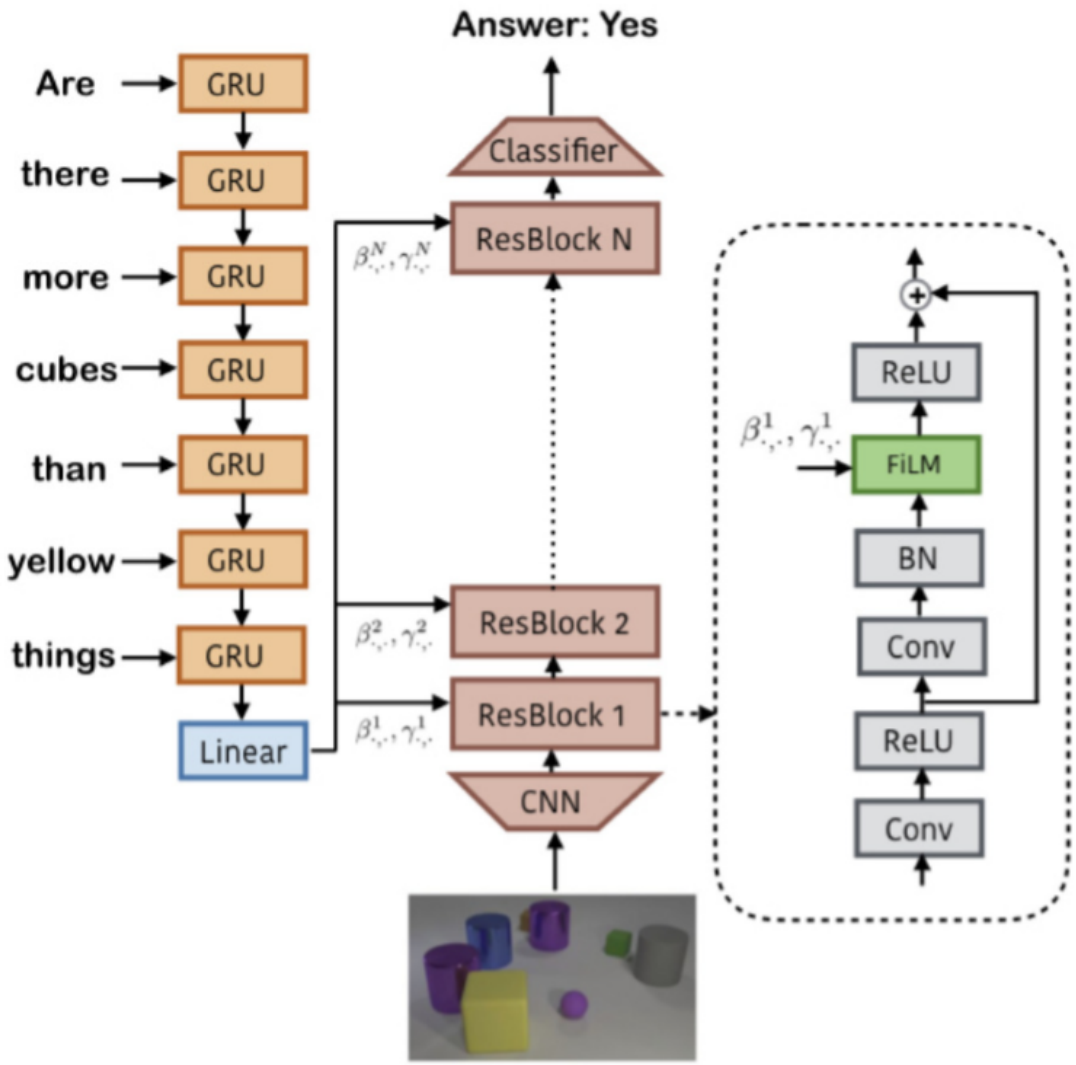
Information Lost



Distributed Attention and FiLM

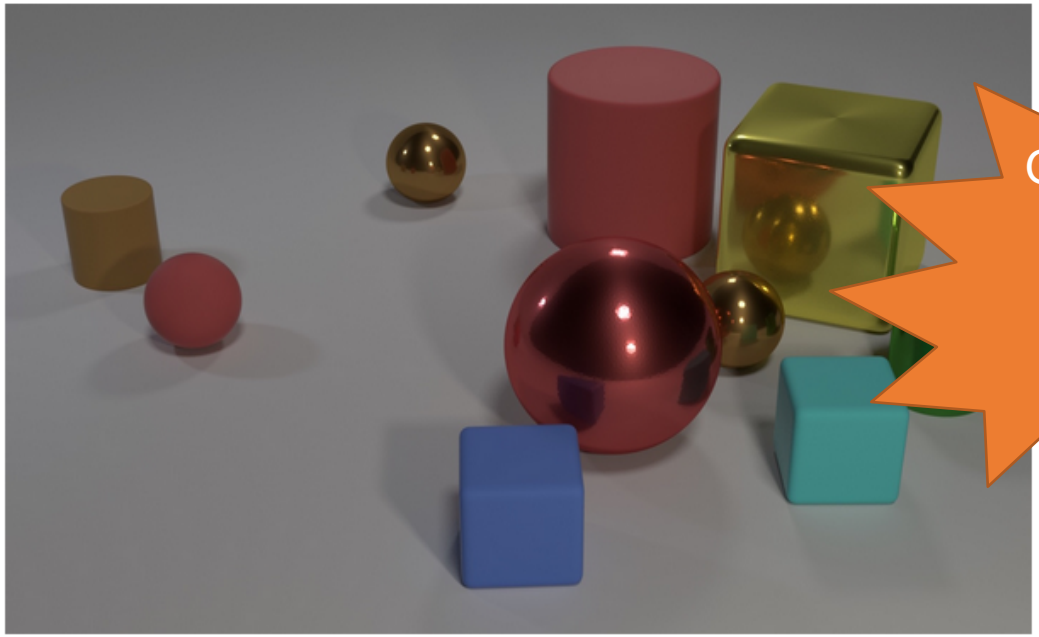


Distributed Attention and FiLM



Unbiased data: CLEVR

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



Guaranteed without language bias

Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?

Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Q: **How many** objects are **either** **small cylinders** or **red** things?

FiLM on CLEVR

Model	**Overall**
Human (Johnson et al.)	92.6
CNN+LSTM (Johnson et al.)	52.3
Stacked Attention (Santoro et al.)	76.6
End-to-End Module Networks* (Hu et al.)	83.7
Prog. Generator+Execution Engine* (Johnson et al.)	96.9
Relation Networks (Santoro et al.)	95.5
FiLM (Ours)	**97.7**

* = Uses additional program-level supervision.

Conclusions

Conclusions

We exaplained how

- **Guesswhat?!** proposes an interesting framework to study purposeful language interaction constrained by a visual scene
- To stay alert against **language bias**
- Benefit from placing **attention mechanisms all along** the convolutional pipeline
    - It will enable us to modulate from low level to high level features
- **FiLM** as a new performant attention mechanism

Thank you for Listening

guesswhat.ai

iAdvize dictanova EXTERNATIC CogniTalk EPITECH. L'ECOLE DE L'INNOVATION ET DE L'EXPERTISE INFORMATIQUE ML WEEK THE MACHINE LEARNING WEEK

Supplementary Material

Language Grounding

How to train a computer to acquire *natural* language?

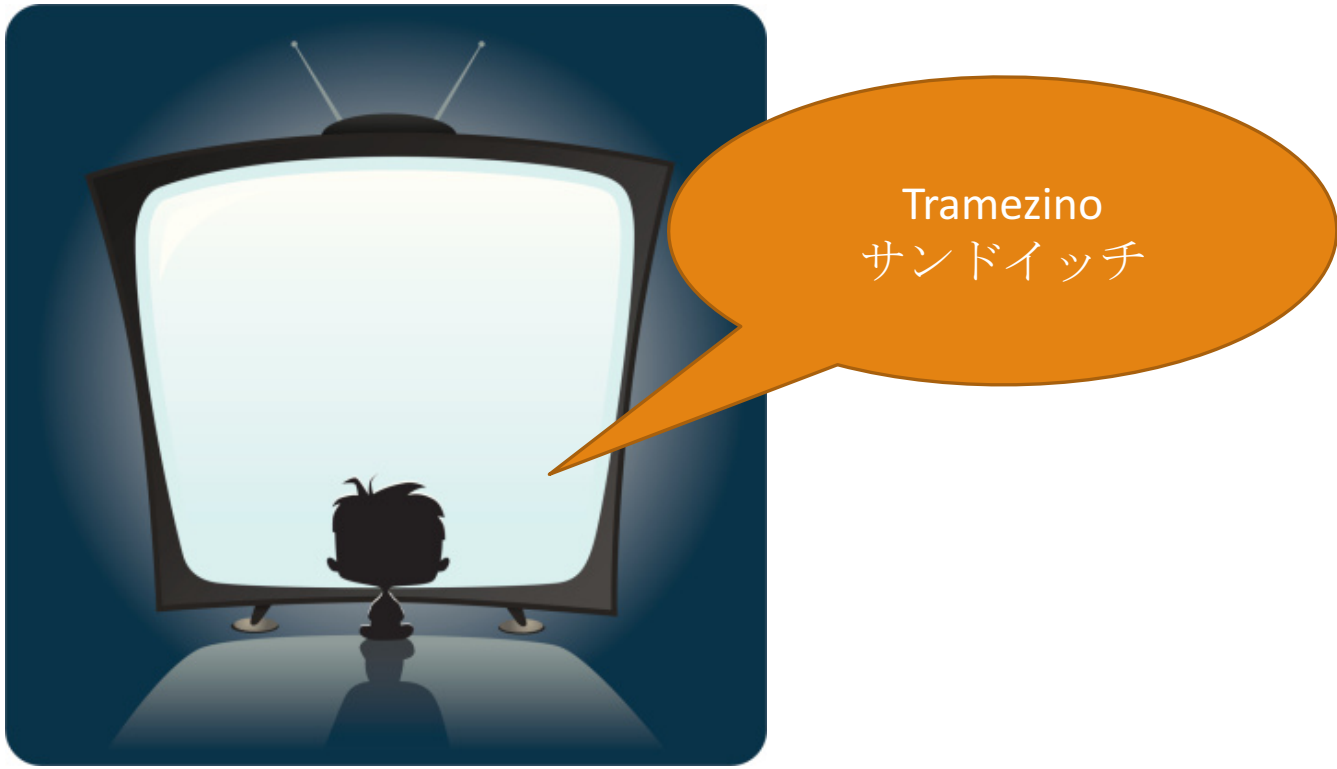
J. Weizenbaum. Eliza. a computer program for the study of natural language communication between man and machine. Commun. ACM, 9(1):36–45, 1966

Language Grounding

How to train a computer to acquire *natural* language?

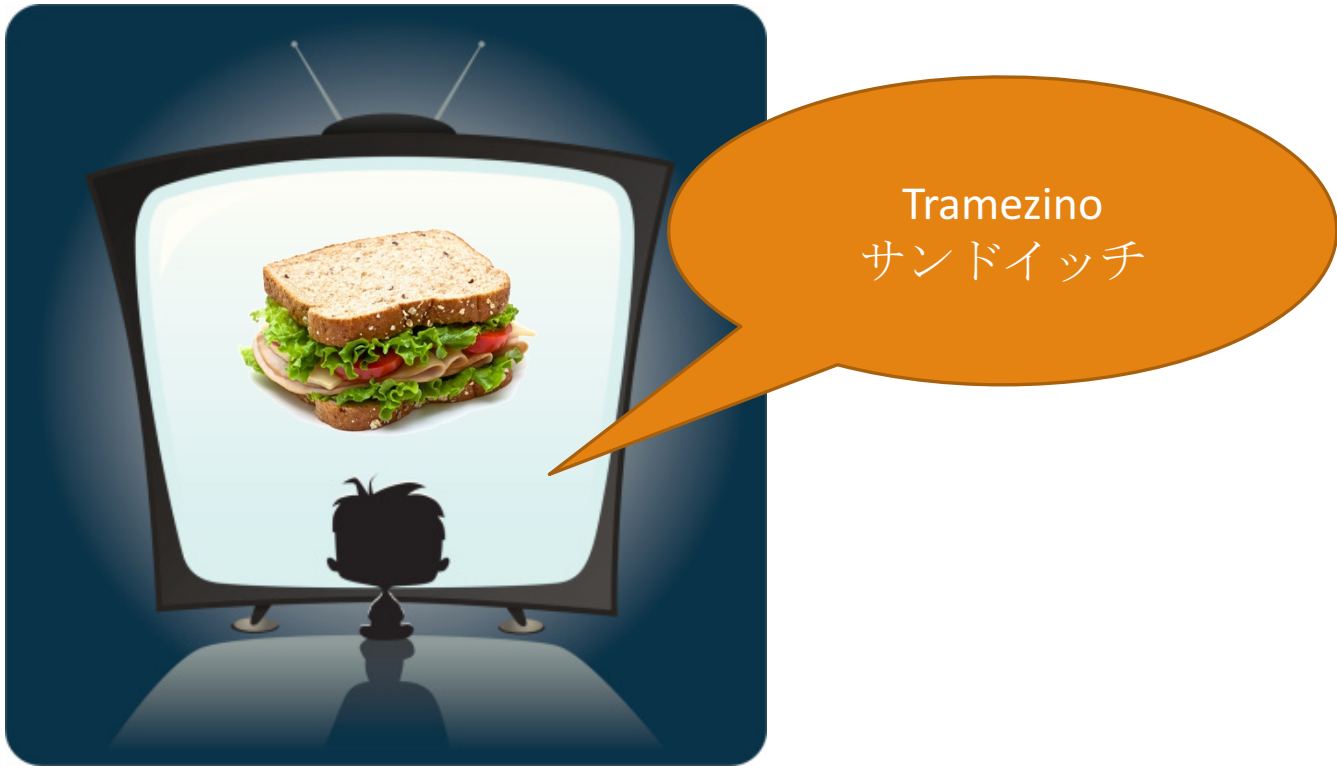
Language Grounding

How to train a computer to acquire *natural* language?



Language Grounding

How to train a computer to acquire *natural* language?



Language Grounding

How to train a computer to acquire *natural* language?

Language Grounding

How to train a computer to acquire *natural* language?

Starting point

Issue : How to verify this hypothesis!

Game features:

- Dialogue (for interaction)
- Visually grounded
- Collaborative
- Goal-oriented with a clear reward

End-to-end Optimization of Goal-driven and Visually Grounded Dialogue Systems

GuessWhat?! Dataset

#64374

- is it an animal? Yes
- one of the two in the bottom right corner? Yes
- the one most to the right? No
- the one to the left of it? Yes

Success

Dataset
It's rich

- **155,280** played games
- **821,889** questions+answers
- **66,537** images
- **134,073** objects

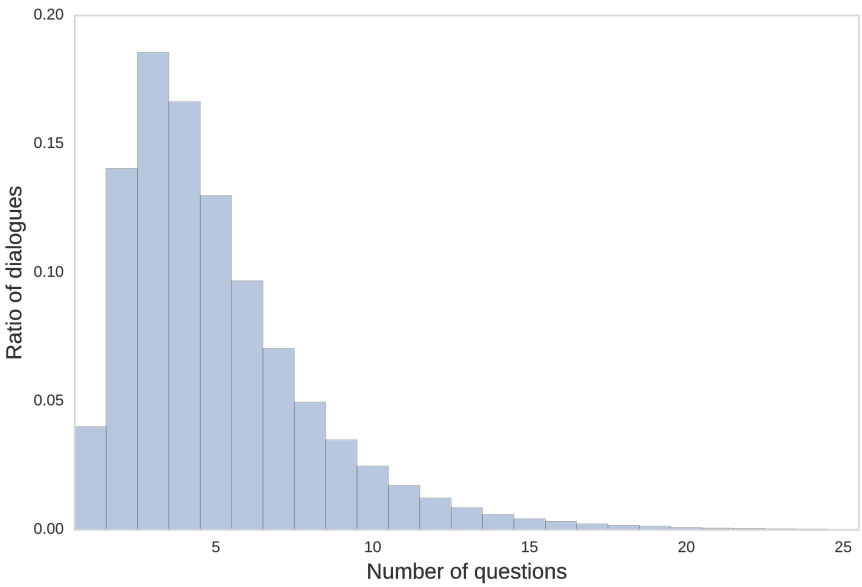
Download the dataset.

#113037

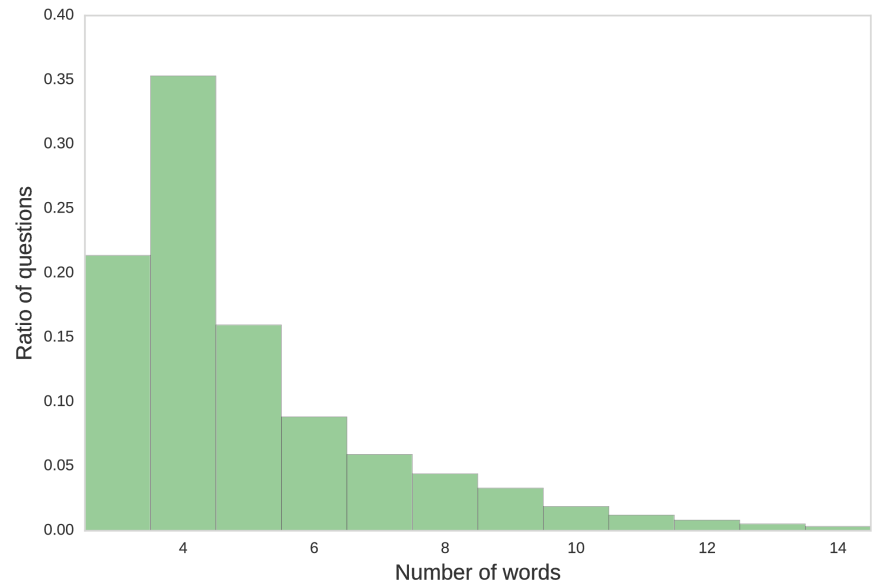
- is it a person? Yes
- are they sitting in the front row? No
- are they in the next row? No
- are they in the back row? Yes
- are they on the left? Yes
- is it the guy with the pink shirt? Yes

Success

GuessWhat?! Dataset

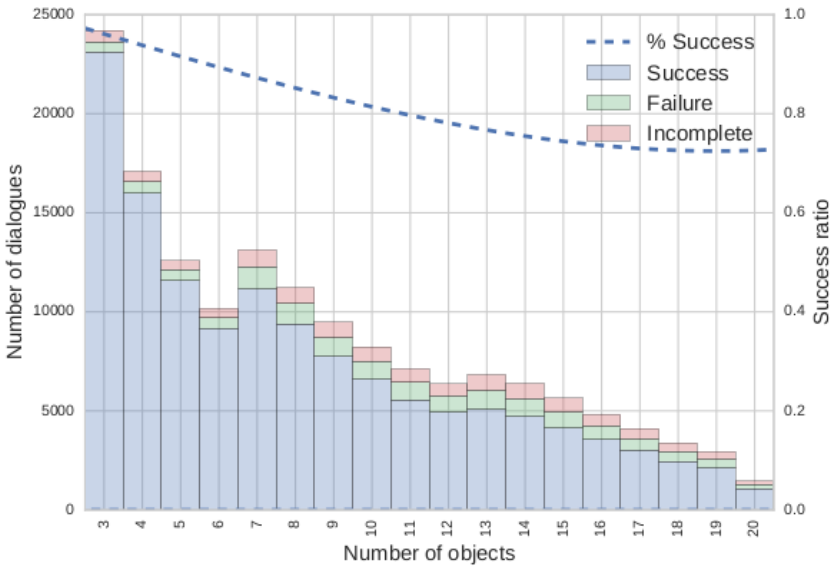


Average: 5+ questions

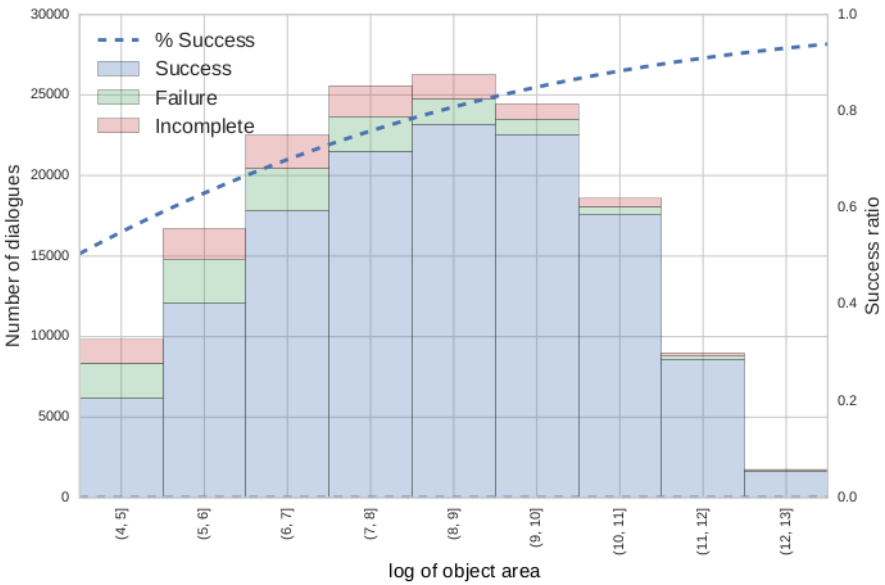


Average: ~5 words

GuessWhat?! Dataset



The more object there are, the lower is the success ratio



The bigger the object is, the higher is the success ratio

GuessWhat?! Dataset

Models

Grounding the language requires to interact with the environment

Limitation of supervised learning:

- Action/state space of dialogues is large: Hard to generalize
- Imitation miss the planning aspect of dialogue
- Does supervised learning really manage to ground vision/language?

#113037

is it a person? Yes

are they sitting in the front row? No

are they in the next row? No

are they in the back row? Yes

are they on the left? Yes

is it the guy with the pink shirt? Yes

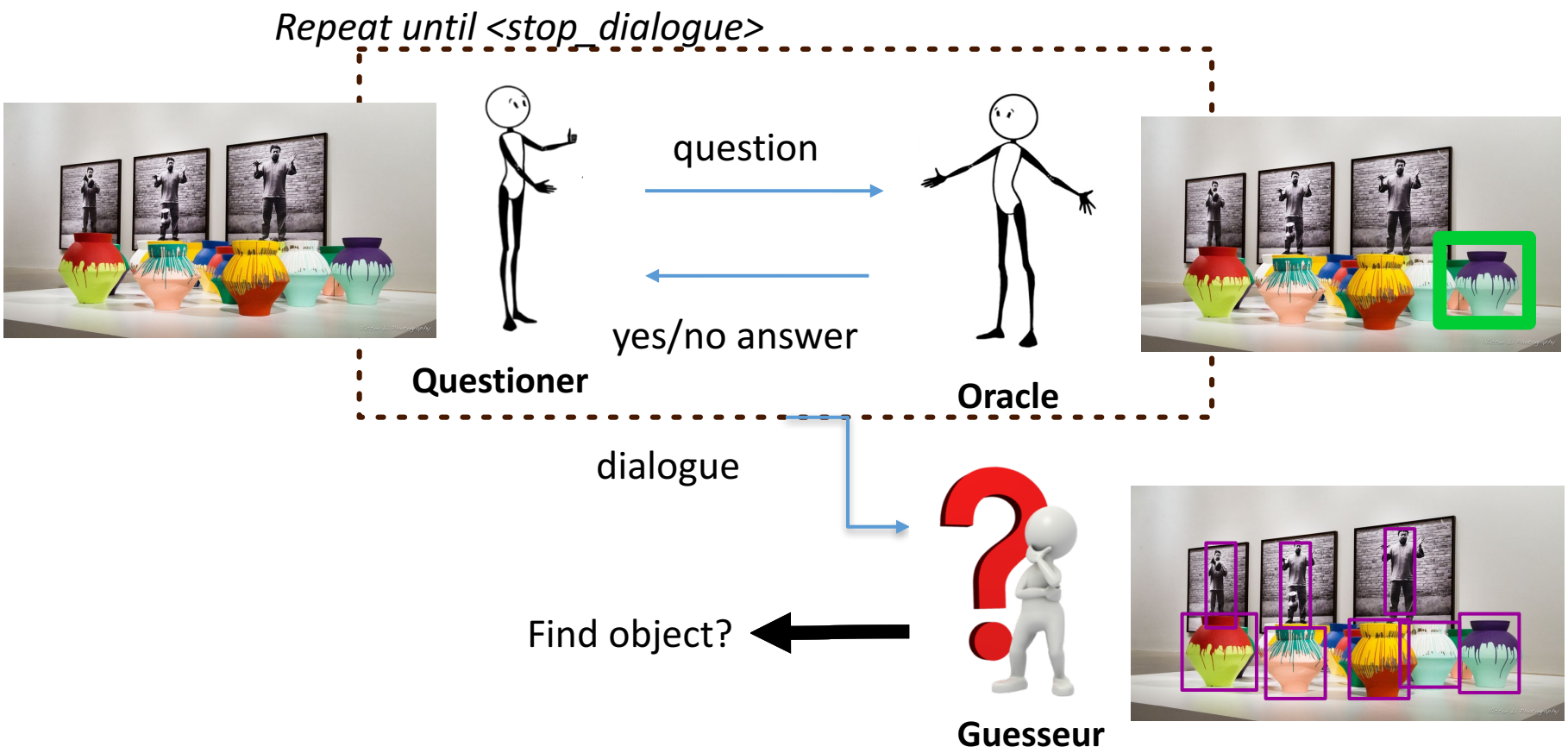
Success

What if the answer would have been no?

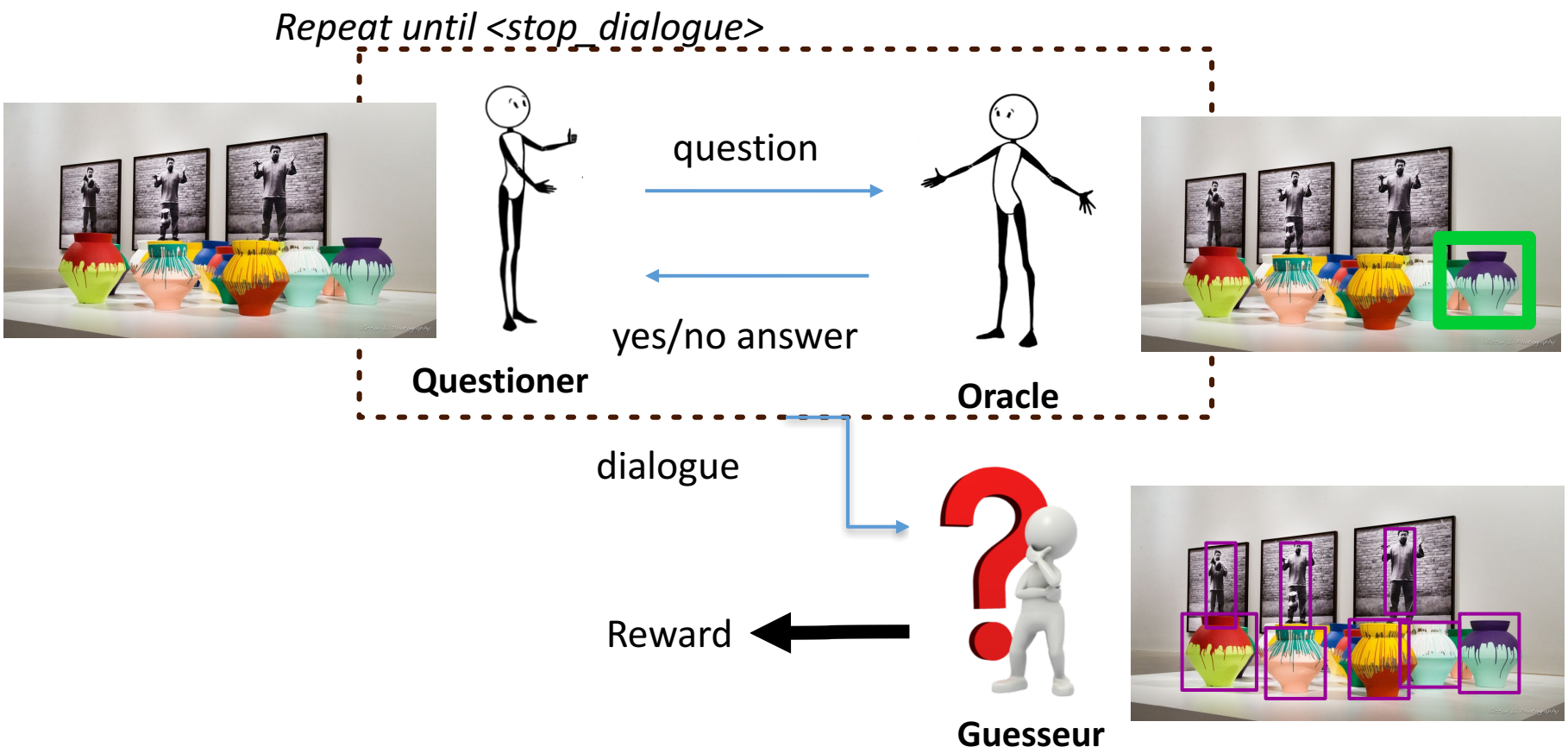
Is it the best question?

Can we phrase it differently?

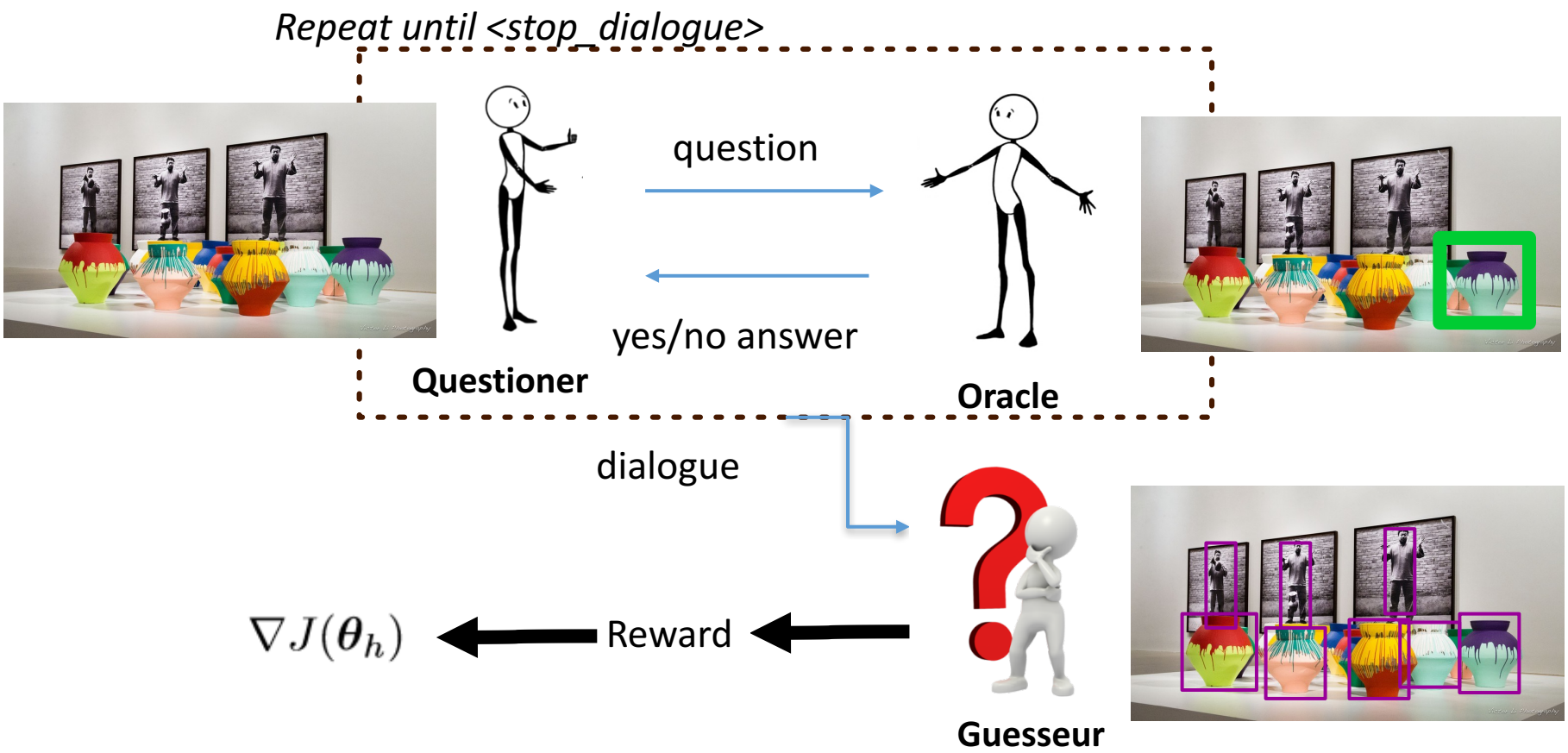
Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
dialogue
Find object?
Guesseur

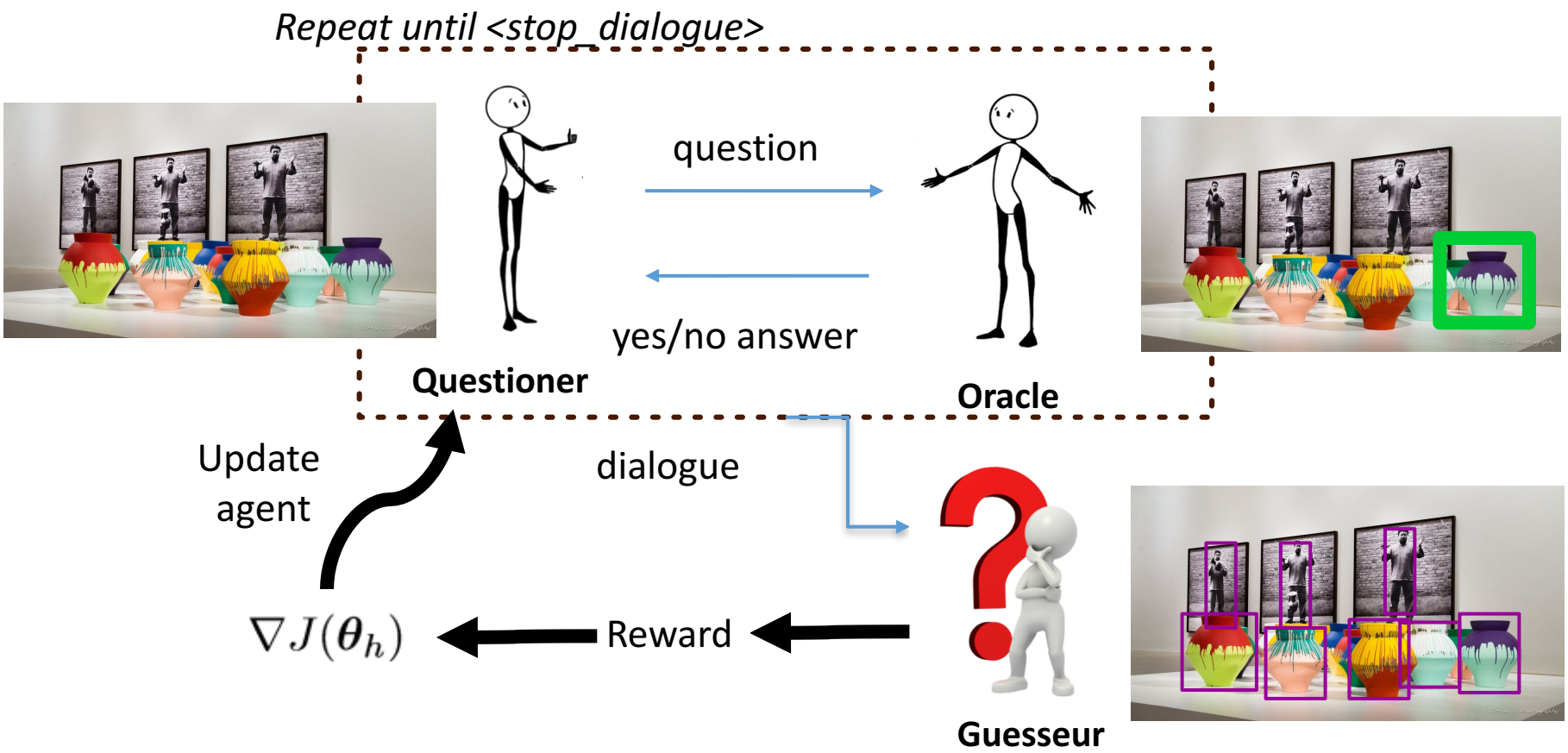
Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
dialogue
Reward
Guesseur

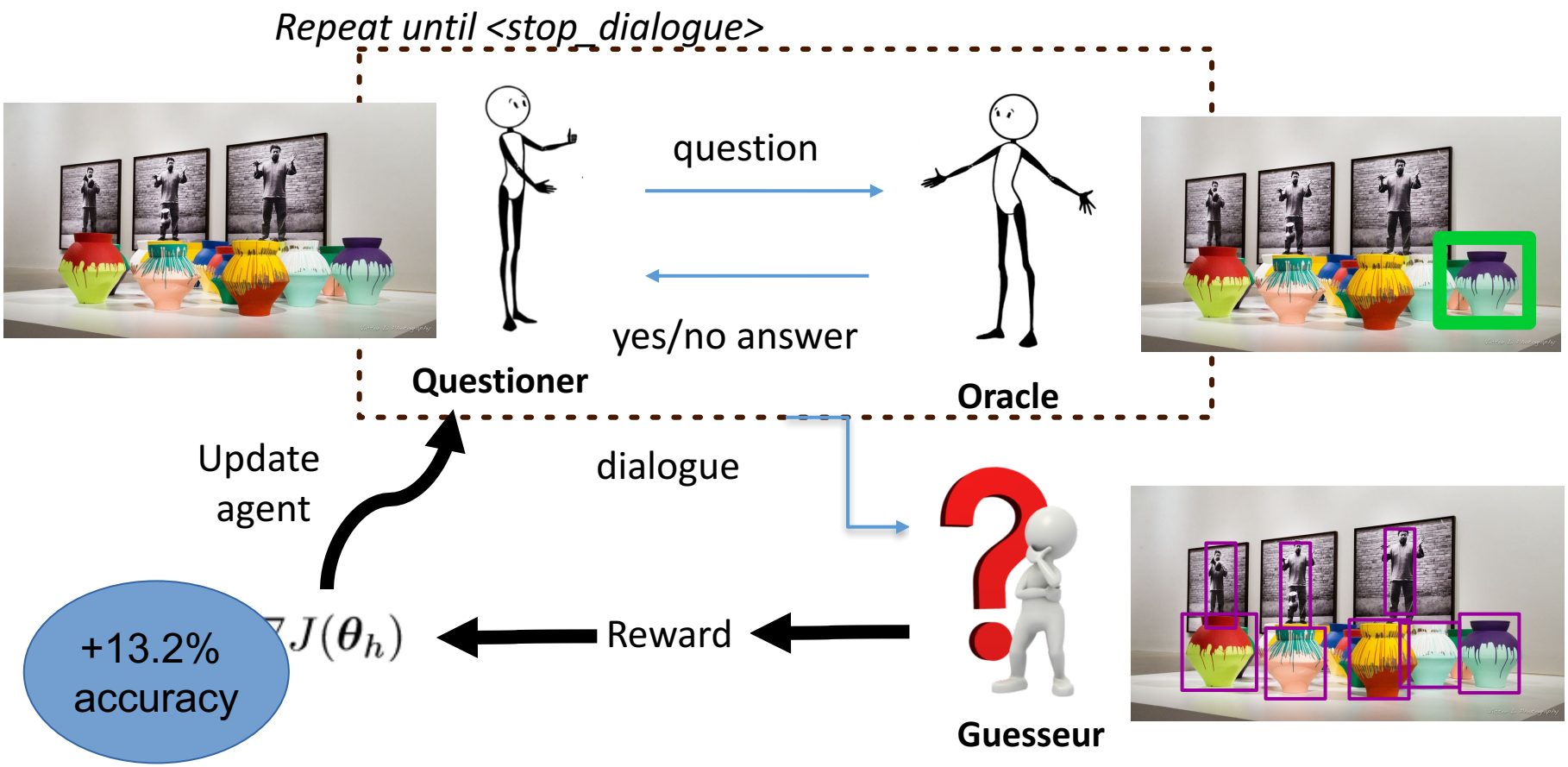
Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
dialogue
Reward
$\nabla J(\theta_h)$
Guesseur

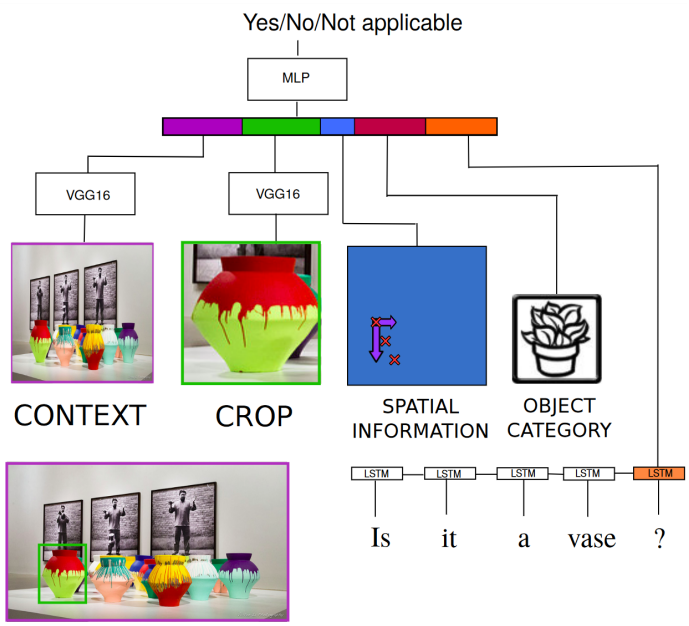
Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
Update agent
dialogue
$\nabla J(\theta_h)$
Reward
Guesseur

Models


Repeat until <stop_dialogue>
question
yes/no answer
Questioner
Oracle
Update agent
dialogue
+13.2% accuracy
$\nabla J(\theta_h)$
Reward
Guesseur

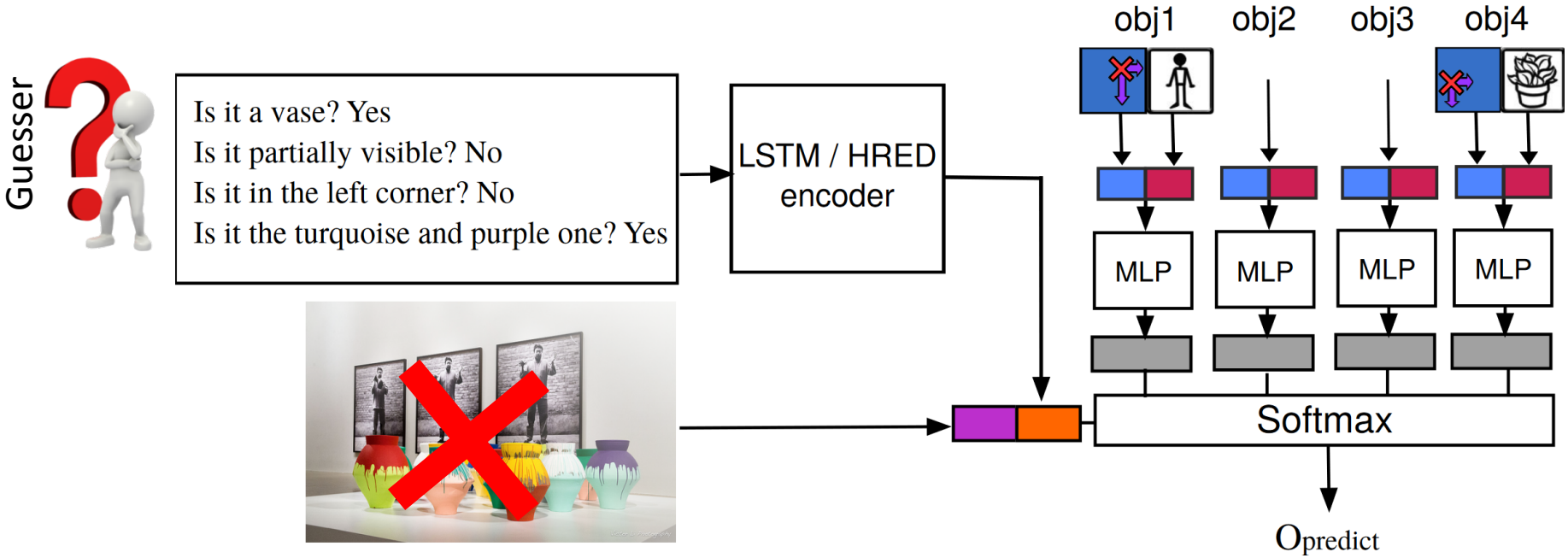
Multimodal Learning



Model	Test err
Dominant class (no)	50.9%
Question	41.2%
Image	46.7%
Crop	43.0%
Question + Crop	29.2%
Question + Image	39.8%
Question + Category	25.7%
Question + Spatial	31.3%
Question + Category + Spatial	**21.5%**
Question + Category + Crop	24.7%
Question + Spatial + Crop	26.2%
Question + Category + Spatial + Crop	22.1%
Question + Category + Spatial + Image	23.5%

- Image features are not helpful !!!

Multimodal Learning



63.8% accuracy

Multimodal Learning

The RL quickly find the oracle/guesser limitations and focus and the question they can use

is it a person ? yes
is it in left ? no
is it in right ? no
is it in middle ? yes
is it in front ? yes
is it in middle ? yes

We need to improve the Oracle/Guesser to improve the language!

Multimodal Learning

Our models fail to fuse two modalities:
- First modality : Vision
- Second Modality: Language

Multimodal Learning

MS COCO [1]

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

VQA [2]

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

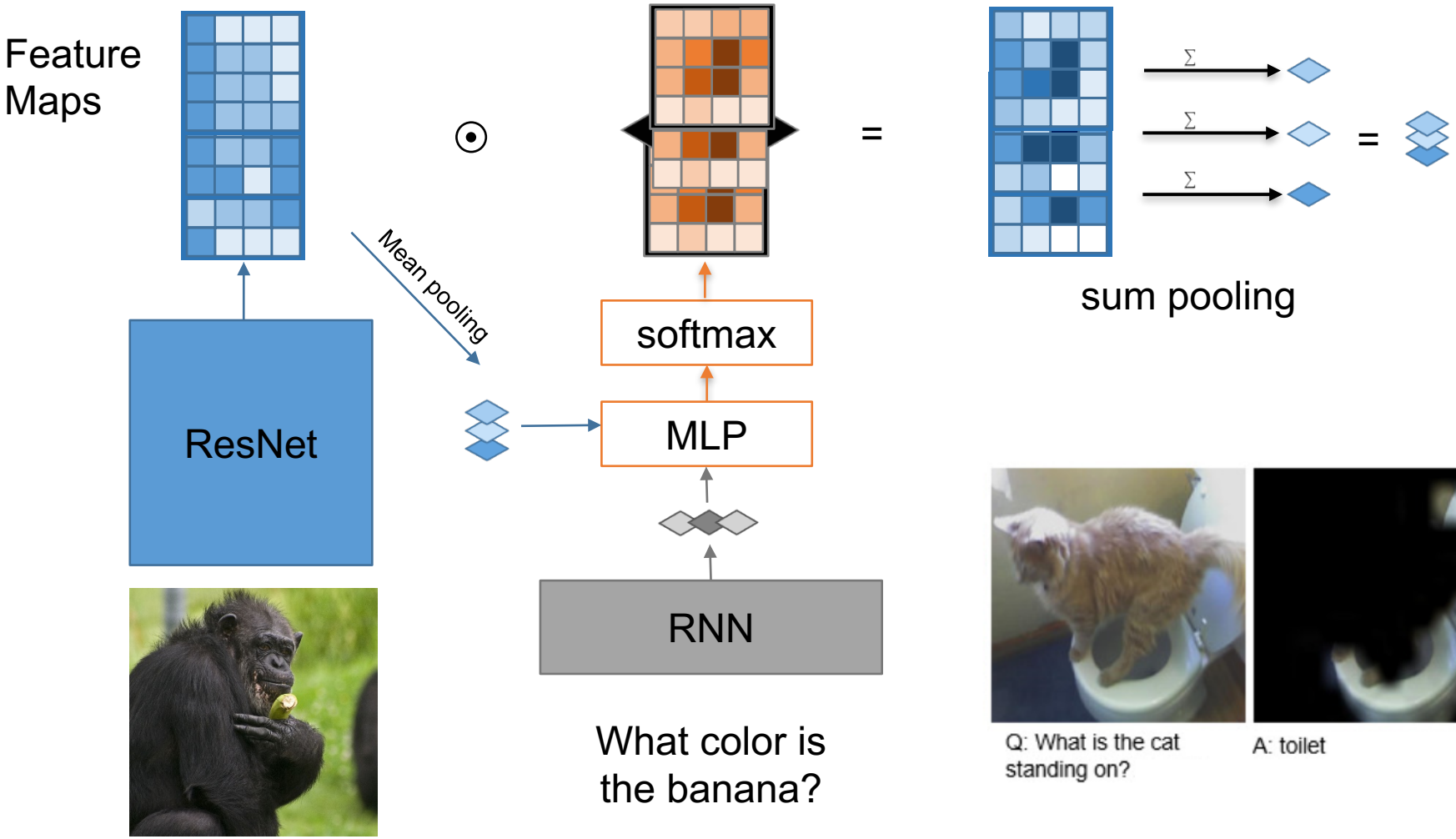
RefeRit [3]

- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

Sample referring expressions for an object in a natural scene.

[1] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. Microsoft coco: Common objects in context. In Proc of ECCV, 2014.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, Z. Lawrence, and D. Parikh. Vqa: Visual question answering. In Proc. of ICCV, 2015
[3] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In Proc. of EMNLP, 2014.

Multimodal Learning



Multimodal Learning

Attention Mechanism

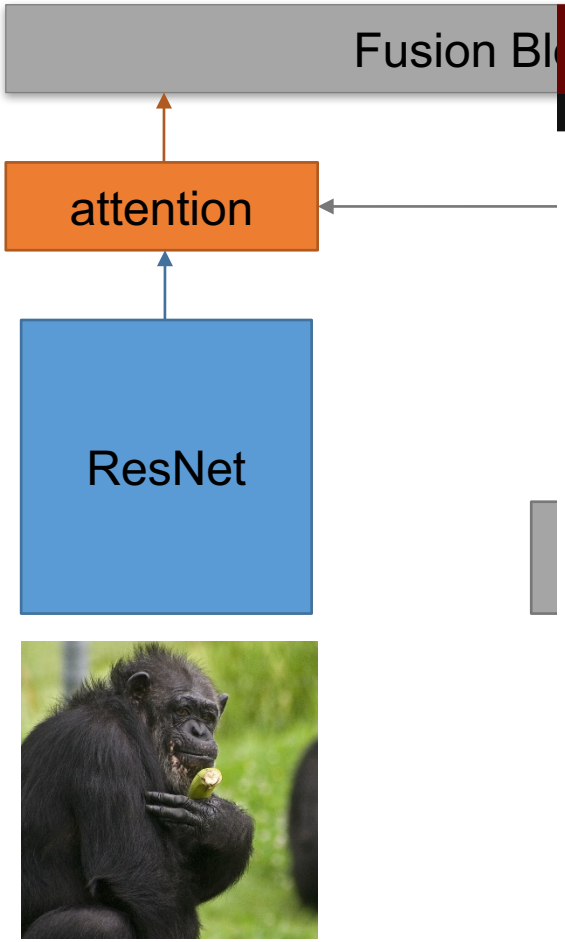
$$\xi_{w,h} = MLP([\boldsymbol{F_{i,\cdot,w,h}}; \boldsymbol{e_q}])$$

$$\alpha_{w,h} = \frac{\exp(\xi_{w,h})}{\sum_{w,h} \exp(\xi_{w,h})}$$

$$\boldsymbol{e}_v = \sum_{w,h} \alpha_{w,h} \boldsymbol{F_{i,\cdot,w,h}}$$

Concatenation
Dot product
Random projection (MCB)
Linear projection + Dot Product (MLB)
Tucker Decompostion
…

Multimodal Learning



VQA Visual Question Answering

Home | People | Code | Demo | Download | Evaluation | Challenge | Browse | Visualize | Workshop

Overview | Challenge Guidelines | Leaderboards ↓

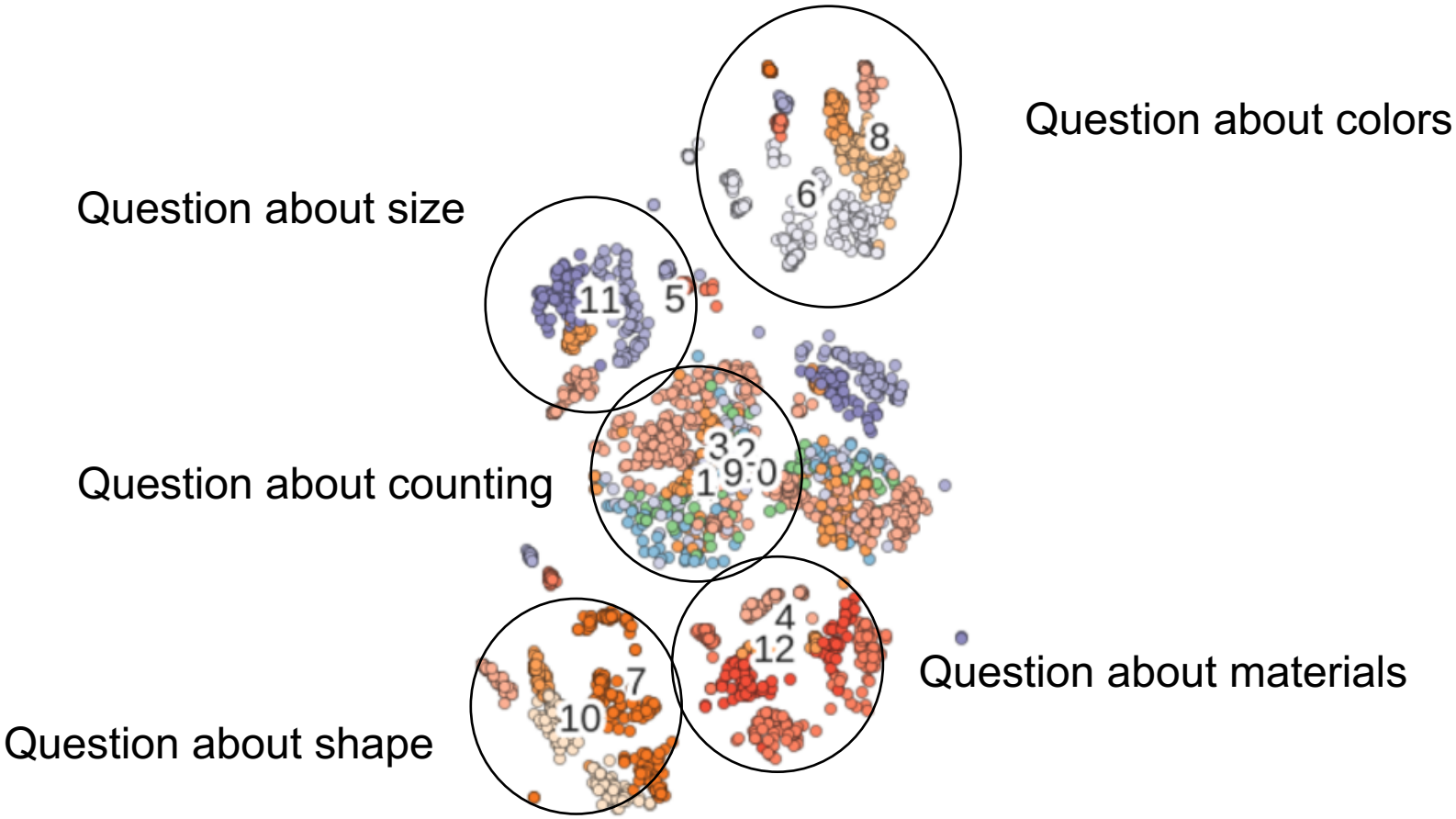
Real Open-Ended

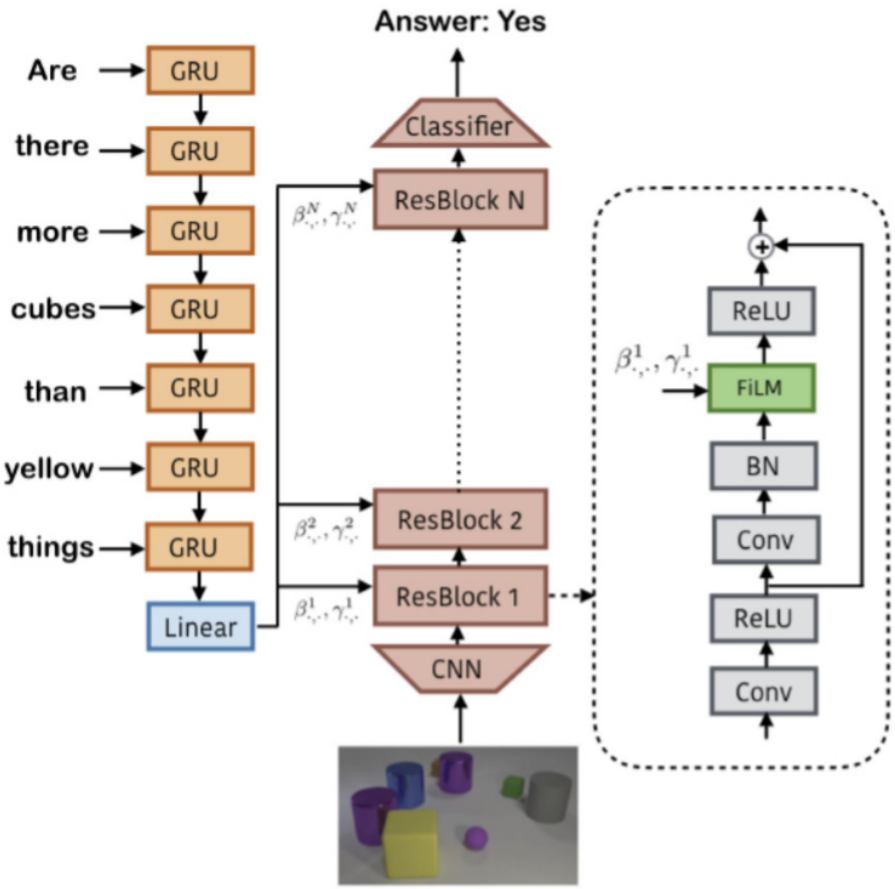
Standard | Dev | Challenge

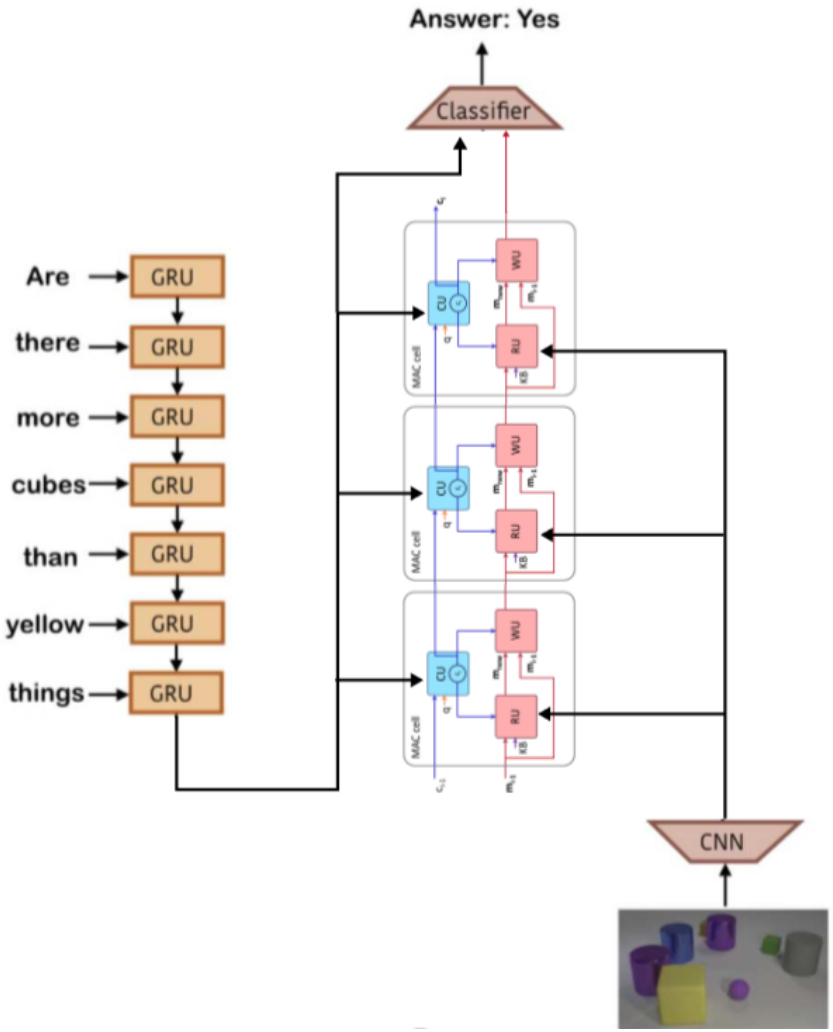
Results as of 06/2016.
For information about each test split please see the challenge page.

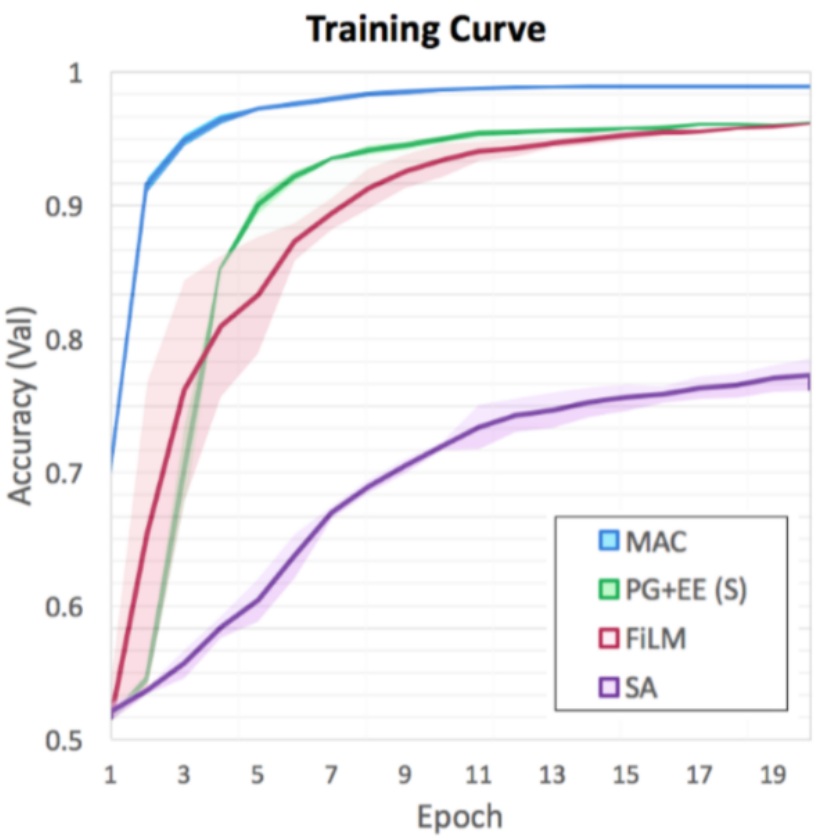
	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4
POSTECH[11]	81.85	38.02	53.12	63.35
Brandeis[3]	82.53	36.54	51.71	62.8
VTComputerVison[19]	80.31	37.87	52.16	62.23
MIL-UT[7]	82.39	36.7	49.76	61.82
klab[23]	81.9	38.85	49.23	61.59
SHB_1026[13]	82.51	36.77	48.06	61.05
MMCX[8]	81.22	35.9	48.74	60.76
VT_CV_Jiasen[20]	80.86	37.85	47.99	60.46
LV-NUS[6]	81.75	35.14	45.96	59.56
ACVT_Adelaide[1]	81.23	36.97	45.89	59.51

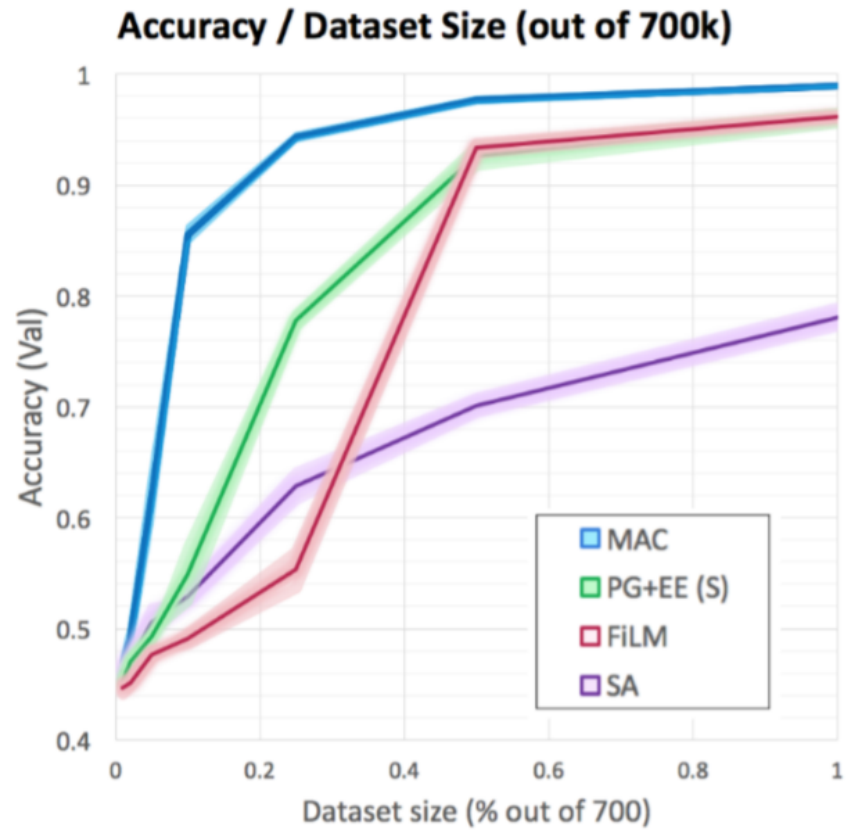
Multimodal Learning




Answer: Yes
Are
there
more
cubes
than
yellow
things
GRU
Linear
Classifier
ResBlock N
ResBlock 2
ResBlock 1
CNN
ReLU
FiLM
BN
Conv
ReLU
Conv


Answer: Yes
Classifier
Are
there
more
cubes
than
yellow
things
GRU
CNN


Training Curve
Accuracy (Val)
1
0.9
0.8
0.7
0.6
0.5
1
3
5
7
9
11
13
15
17
19
Epoch
MAC
PG+EE (S)
FiLM
SA


Accuracy / Dataset Size (out of 700k)
Accuracy (Val)
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.2
0.4
0.6
0.8
1
Dataset size (% out of 700)
MAC
PG+EE (S)
FiLM
SA