Language and Vision at UniTN

Raffaella Bernardi
University of Trento

UNIVERSITÀ DEGLI STUDI
DI TRENTO

LaVi @ UniTn

Learning the meaning of Quantifiers from Language, Vision (and Audio): https://quantit-clic.github.io/

none
almost none
few
the smaller part
some
many
most
almost all
all

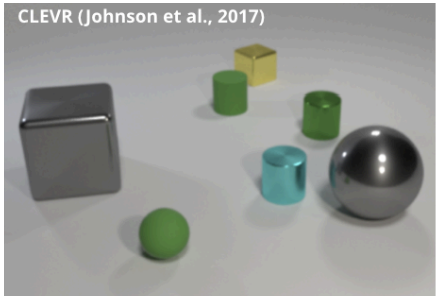
Sandro Pezzelle
(now post-doc at UvA)

Diagnostic analysis of LV models: https://foilunitn.github.io/

People riding bikes down the road approaching a dog

Ravi Shekhar
(now post-doc at QMUL)

Transfer Learning in (I)VQA:
https://continual-vista.github.io/



Multimodal tasks:

Wh-q: y ∈ {metal, blue, sphere,..,large}
Q: *What is the material of the large object that is the same shape as the tiny yellow thing?* ***A:*** *metal*

Yes/No-q: y ∈ {Yes, No}
Q: *Does the cyan ball have the same material as the large object behind the green ball?* ***A:*** *Yes*

Current Focus: Dialogues between Speakers with different background

Claudio Greco (CIMeC)

Visually Grounded Talking Agents

(in collaboration with UvA:
https://vista-unitn-uva.github.io/)

Current Focus: Multimodal Pragmatic Speaker

Alberto Testoni (DISI)

Computational Models of Language Cognitive and Language Evolution

Stella Frank

(CIMeC)

LaVi@ UniTN on going collaborations

Be Different to Be Better:

If I am feeling alone

- ❑ I cry
- ❑ I join the group
- ❑ …

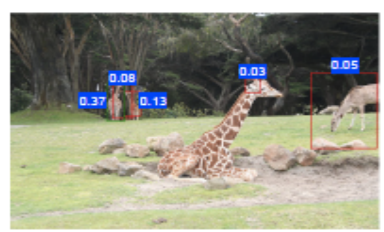
In collaboration with Uva

https://sites.google.com/view/bd2bb/home

Visually Grounded Spatial Reasoning

is it the bus on the left? No

is it the boat next to a car? No



is it one of the two in the back? Yes

In collaboration with Cordoba University

https://github.com/albertotestoni/unitn_unc_splu2020

Visual Dialogue Games

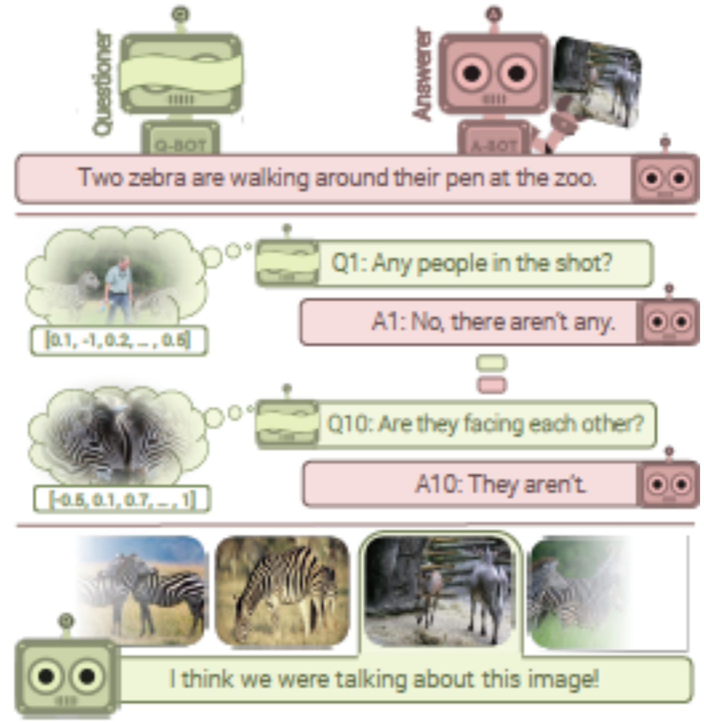
GuessWhat?!

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

De Vries et al CVPR 2017

Strub et al IJCAI 2017

GuessWhich



Das et al IEEE 2017

Das et al ICCV 2017

Murahari et al EMNLP 2019

Visually Grounded Talking Agents

GuessWhat?!

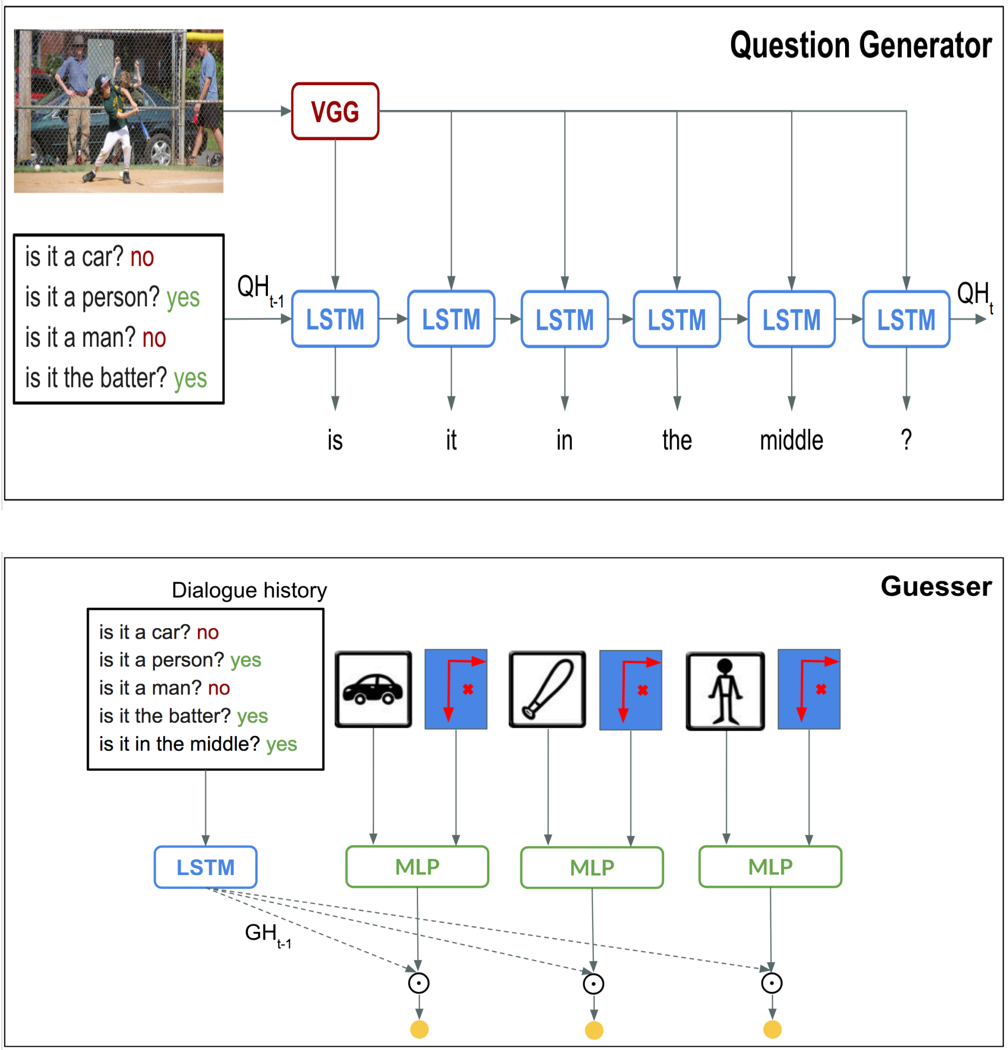
Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

De Vries et al CVPR 2017

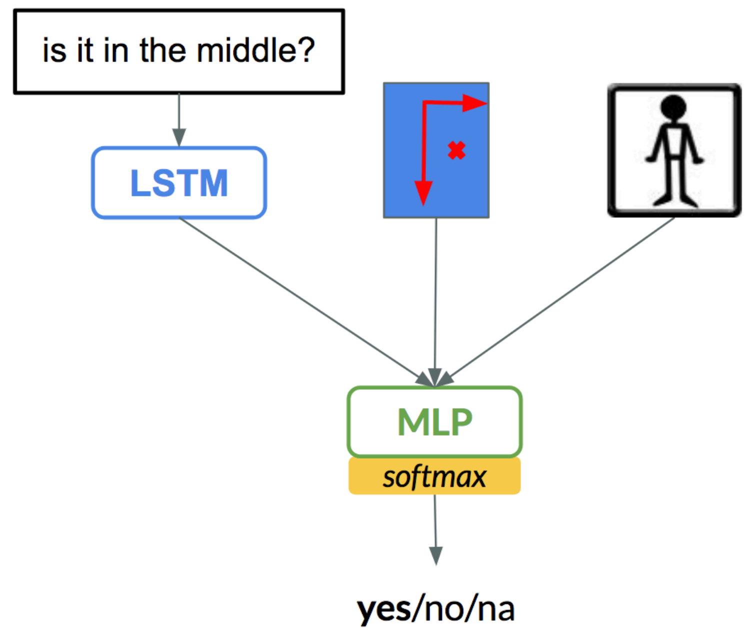
Strub et al IJCAI 2017

Guess What?! baseline

Questioner

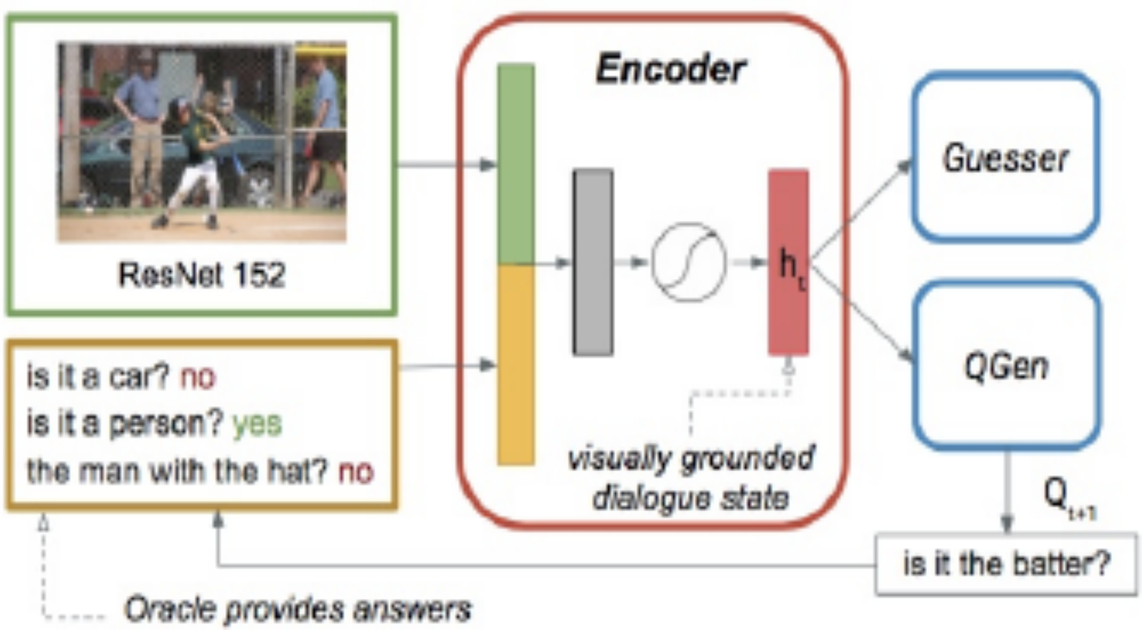


Oracle



de Vries et al 2017

Grounded Dialogue State Encoder



Raffaella Bernardi   Elia Bruni   Raquel Fernández   Ravi Shekhar   Aashish Venkatesh   Tim Baumgärtner   Barbara Plank

COLING 2018   NAACL 2019   Clic-it 2019

https://vista-unitn-uva.github.io

Learning Approaches

- **Supervised Learning** (SL) (Baseline - *de Vries et al 2017*, Our-GDSE-SL): Trained on human data

- **Reinforcement Learning** (RL) (SoA - *Strub et al. 2017*): Trained on generated data

- **Cooperative Learning** (CL) (Our-GDSE-CL): Trained on generated data and human data

Results: GuessWhat?!

	5Q	**8Q**
Baseline (de Vries et al 2017)	41.2	40.7
GDSE-SL (our)	47.8	49.7
GDSE-CL (our)	53.7(∓0.83)	**58.4**(∓0.12)

- Our best is with 10Q: 60.8(∓0.51)

Results: GuessWhat?!

	5Q	**8Q**
Baseline (de Vries et al 2017)	41.2	40.7
GDSE-SL (our)	47.8	49.7
GDSE-CL (our)	53.7(∓0.83)	**58.4**(∓0.12)
RL (Strub et al. (2017))	**56.2**(∓0.24)	56.3(∓0.05)

Our best result is with 10Q:
60.8(∓0.51)

Beyond Task Success

Model	5Q	8Q
Baseline	41.2	40.7
GDSE-SL	47.8	49.7
GDSE-CL	53.7 (±.83)	58.4 (±.12)
RL	56.2 (±.24)	56.3 (±.05)

	Lexical diversity	Question diversity	% Games with repeated Q's
BL	0.030	1.60	93.50
SL	0.101	13.61	55.80
CL	0.115 (±.02)	14.15 (±3.0)	52.19 (±4.7)
RL	0.073 (±.00)	1.04 (±.03)	96.47 (±.04)
H	0.731	47.89	—

Humans	*[success]*
1. does it have cereal on it?	no
2. does it have pink frosting?	no
3. does it have cookies?	no
4. is it a donut?	yes
5. does it have maple frosting?	no
6. is there candy?	yes
7. is it the doughnut with M&Ms?	no
8. is it dark brown?	no

GDSE-CL	*[success]*
1. is it a donut?	yes
2. is it on the left?	no
3. is it on the right?	yes
4. is it the whole donut?	yes
5. does it have pink icing?	no
6. is it touching the donut with the sprinkles?	yes
7. does it have chocolate icing?	no
8. is it the third donut?	yes

RL	*[failure]*
1. is it food?	yes
2. is it a donut?	yes
3. is it in left?	no
4. is it in top?	yes
5. is it in top?	yes
6. is it in top?	yes
7. is it in top?	yes
8. is it top?	yes

Question Type

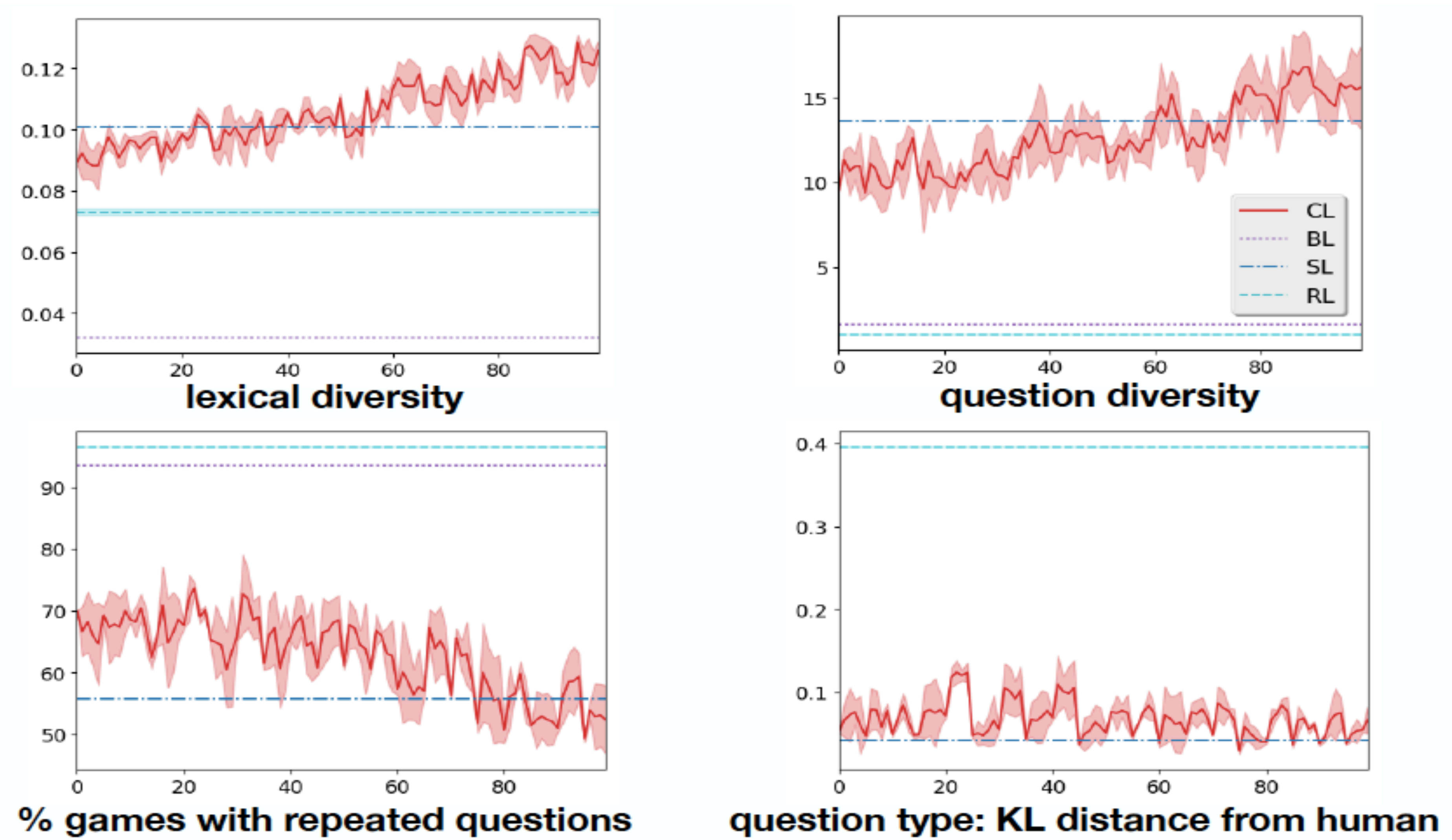
Question type	BL	SL	CL	RL	H
ENTITY	**49.00**	**48.07**	**46.51**	**23.99**	**38.11**
SUPER-CAT	19.6	12.38	12.58	14.00	14.51
OBJECT	29.4	35.70	33.92	9.99	23.61
ATTRIBUTE	**49.88**	**46.64**	**47.60**	**75.52**	**53.29**
COLOR	2.75	13.00	12.51	0.12	15.50
SHAPE	0.00	0.01	0.02	0.003	0.30
SIZE	0.02	0.33	0.39	0.024	1.38
TEXTURE	0.00	0.13	0.15	0.013	0.89
LOCATION	47.25	37.09	38.54	74.80	40.00
ACTION	1.34	7.97	7.60	0.66	7.59
Not classified	**1.12**	**5.28**	**5.90**	**0.49**	**8.60**
KL (wrt human)	0.953	0.042	0.038	0.396	0.0

Dialogue Strategy

Question Type Shift after getting “YES” answer

	BL	SL	CL	RL	Human
SUPER-CAT → OBJ/ATT	89.05	92.61	89.75	95.63	89.56
OBJECT → ATTRIBUTE	67.87	60.92	65.06	**99.46**	88.70

Evolution of linguistic factors over 100 training epochs



Summing up

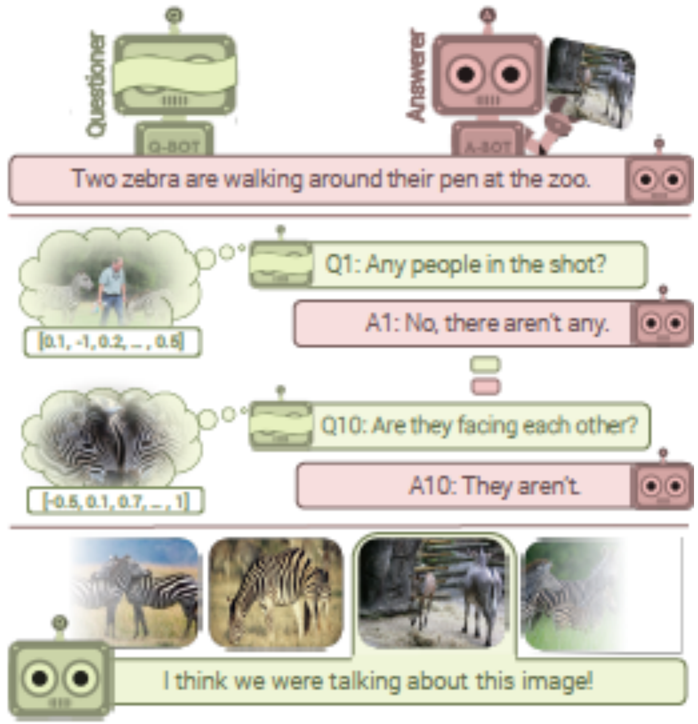
Take-home message:

➔Don’t stop at the task accuracy, quality of the dialogue is also important.

Next:

➔ how flexible is our architecture?

GuessWhich Game



Das et al IEEE 2017
Das et al ICCV 2017
Murahari et al EMNLP 2019

The Dialogues

A room with a couch, tv monitor and a table

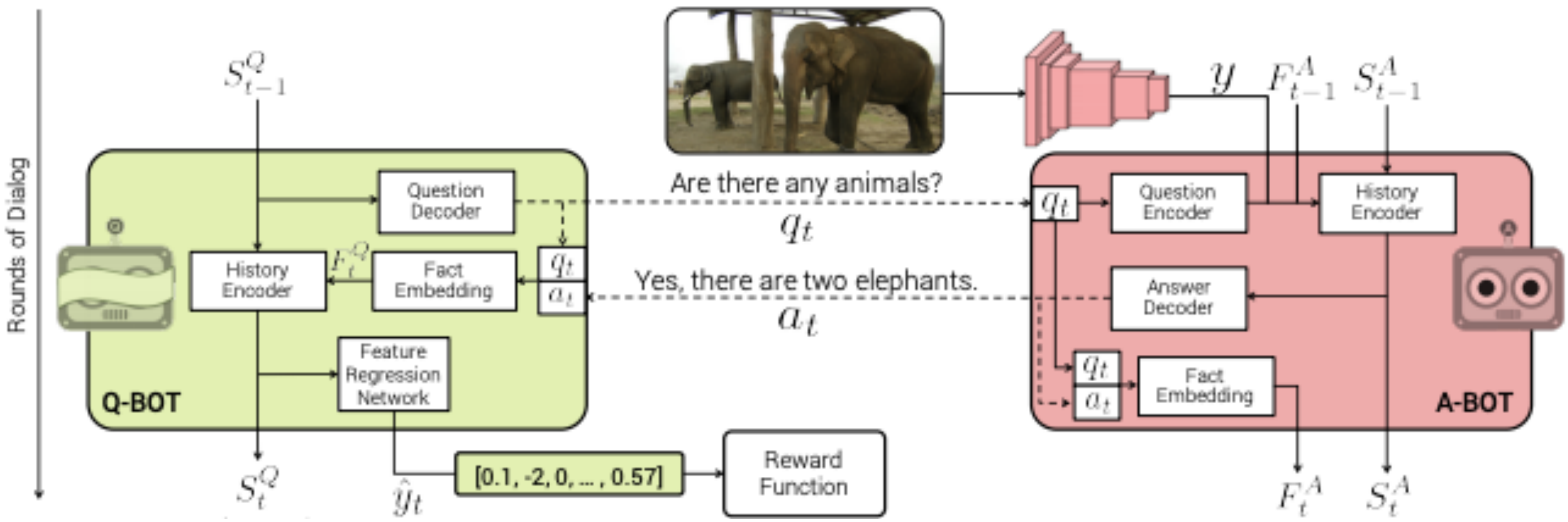
ReCap	
1. is tv flat screen?	no, it is small screen of some sort
2. does table have wood table?	yes, it does
3. any people?	no people
4. is tv on?	no
5. what color is couch?	brown
6. are there any boots?	no, there aren't any
7. can you see wintows?	no
8. does table have wook on table?	no it doesn't
9. what color are walls?	white
10. does table look clean?	not really

The Dialogues

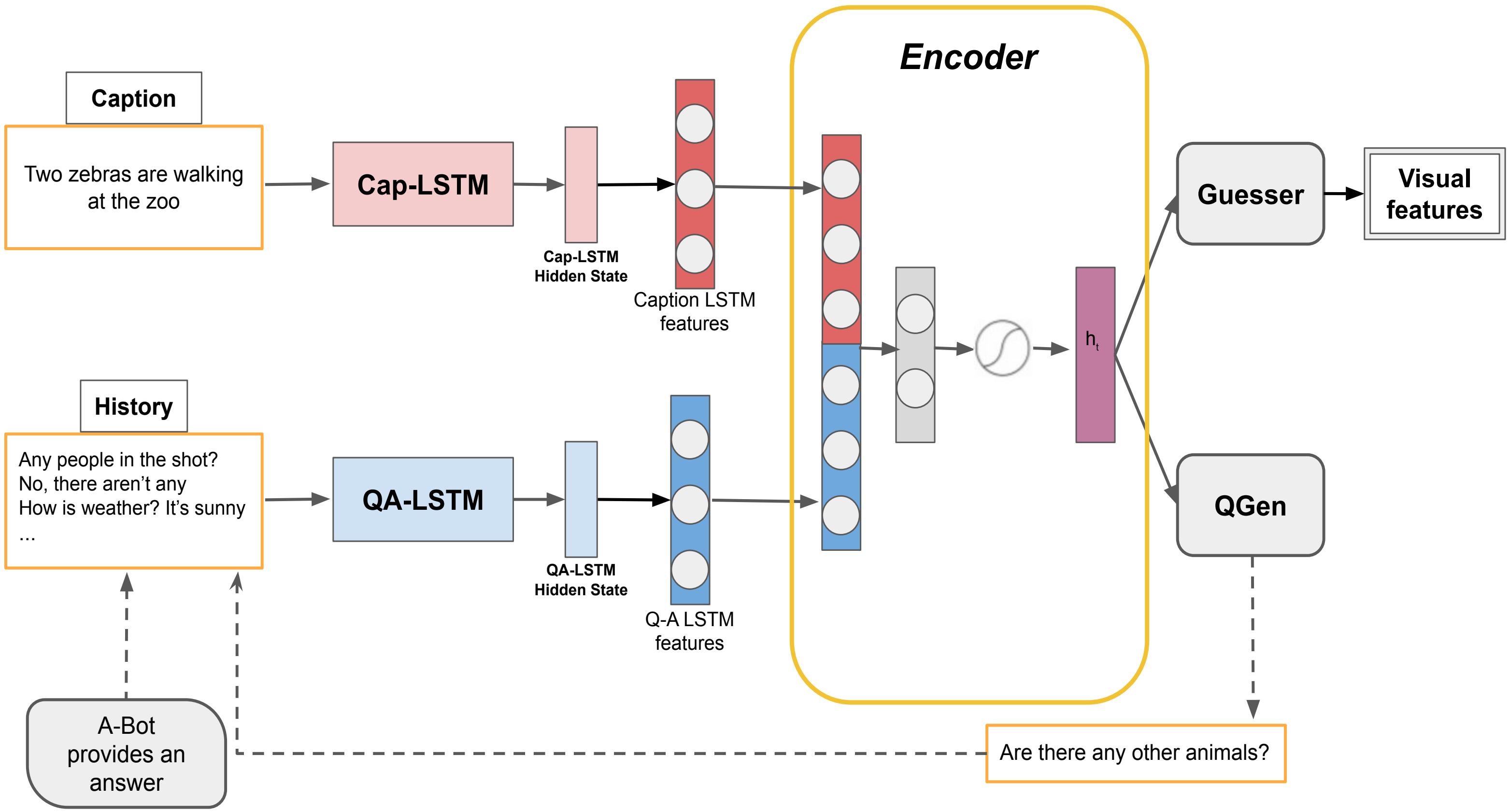
A room with a couch, tv monitor and a table

ReCap	
1. is tv flat screen?	no, it is small screen of some sort
2. does table have wood table?	yes, it does
3. any people?	no people
4. is tv on?	no
5. what color is couch?	brown
6. are there any boots?	no, there aren't any
7. can you see wintows?	no
8. does table have wook on table?	no it doesn't
9. what color are walls?	white
10. does table look clean?	not really

Q-Bot and A-BoT



A simple Model of the Questioner



ca. 10K candidates images

ReCap: it re-reads the caption at each turn

Results

Mean Percentile Rank (MPR): 95% means that, in average, the target image is closer to the one chosen by the model more than the 95% of the candidate images.

With 9628 candidates, 95% MPR corresponds to a Mean Rank of 481.4
A difference of +/– 1% MPR corresponds to –/+ 100 mean rank.

	MPR
Chance	50.00
Qbot-SL	91.19
Qbot-RL	94.19
AQM+/indA	94.64
AQM+/depA	97.45
ReCap	95.54

GT dialogues	MPR
Guesser + QGen	94.84
ReCap	95.65
Guesser caption	49.99
Guesser dialogue	49.99
Guesser caption +dialogue	94.92
Guesser-USE caption	96.90

The dialogues work as a language incubator. They don't provide info to identify the image

The Role of the Dialogue

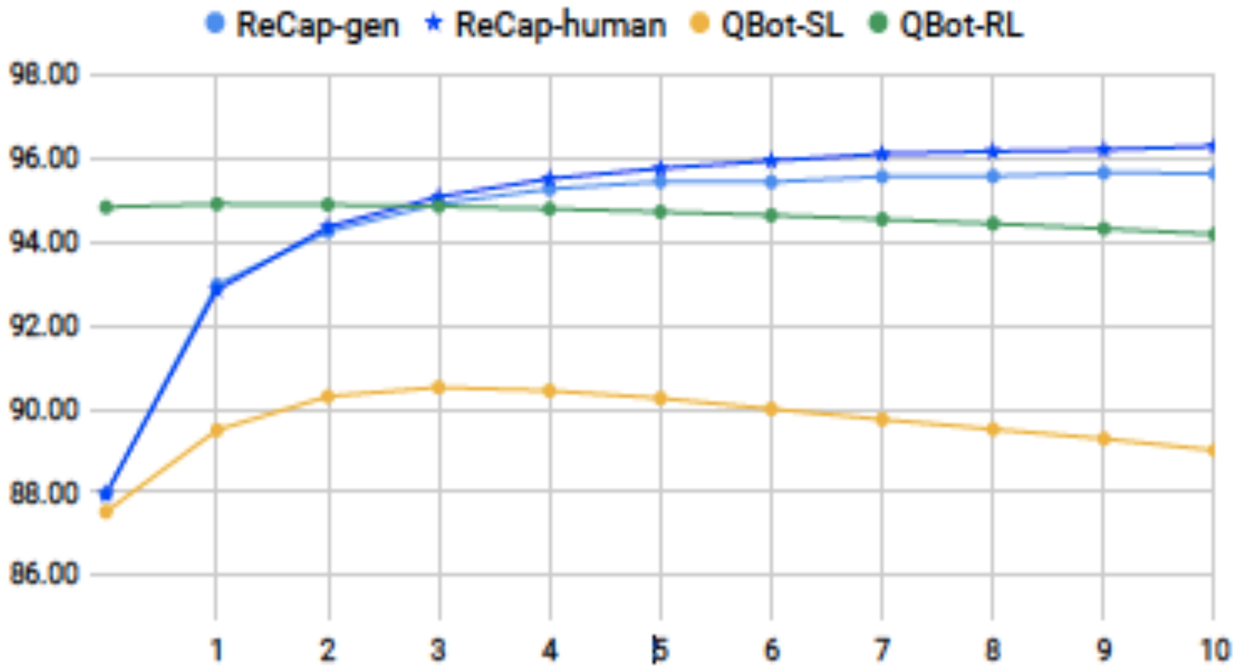
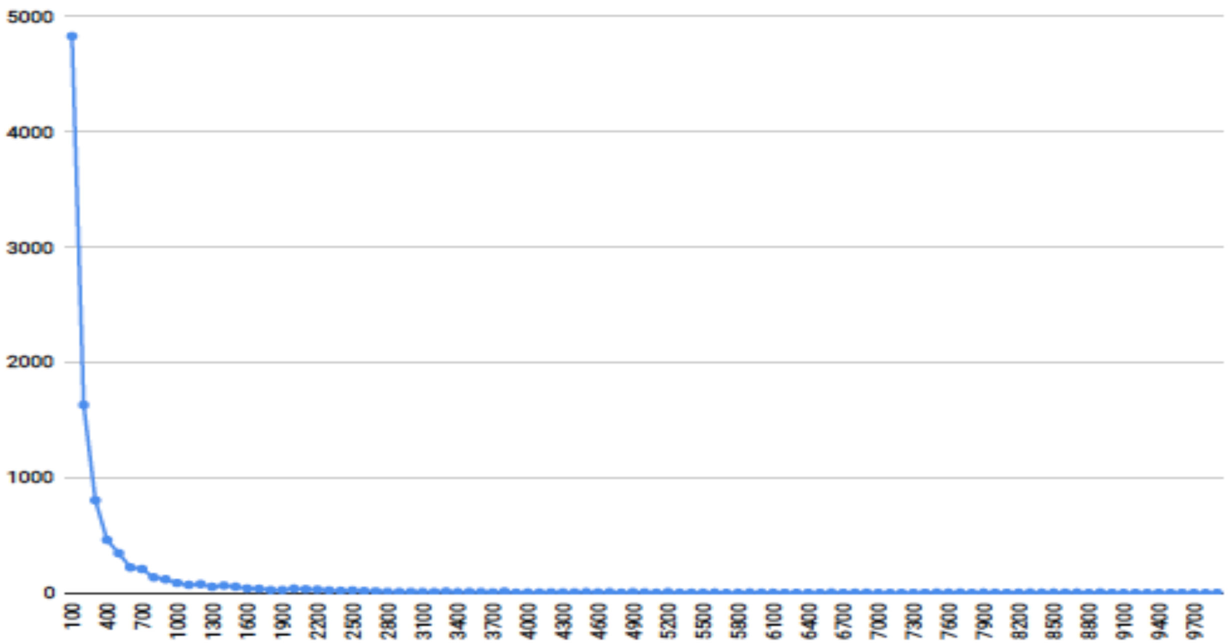


Figure 3: MPR distribution per dialogue round: comparison of ReCap model tested on human dialogues vs. ReCap and QBot models tested on generated dialogues.

Analysis of the Test Set



Distribution of rank assigned to the target image by ReCap

A room with a couch, tv monitor and a table.

This is a close up picture of a roosters. neck

Summing up

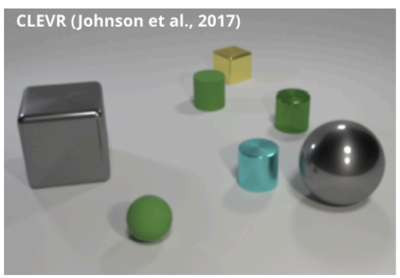
- The metric used is too coarse
- The dataset too skewed

What we have learned so far about Visually Grounded Talking Agents

- They are interesting and challenging.
- There are good “baselines” available.
- Advantage of using cooperative learning within the model’s modules.
- It might be good to use pre-trained language embedding.
- Let’s not forget to evaluate the dialogues.

Continual Learning

Continual Learning in VQA:
https://continual-vista.github.io/



Multimodal tasks:

Wh-q: **y** ∈ {metal, blue, sphere,..,large}
Q: *What is the material of the large object that is the same shape as the tiny yellow thing?* ***A:*** *metal*

Yes/No-q: **y** ∈ {Yes, No}
Q: *Does the cyan ball have the same material as the large object behind the green ball?* ***A:*** *Yes*

Claudio Greco
(CIMeC)

Modeling Human Learning

- *Transfer learning*: the situation where what has been learned in one setting is exploited to improve generalization in another setting (Holyoak and Thagard, 1997)

- *Lifelong Learning* systems should be able to learn from a stream of tasks
 (Thrun and Mitchell, 1995)

- *Curriculum Learning* a learning strategy which starts from easy training examples and gradually handles harder ones (Elman 1993)

Our Work on VQA

We ask whether MM models:

1. *benefit* from learning question types of incremental difficulty
2. *forget* how to answer question types previously learned

Learning to answer questions

Moradlou and Ginzburg 2018:
Children learn to answer Wh-Q before learning to answer polar questions

Wh answered by child:

a. MOT: what’s that? CHI: yyy dog. MOT: that’s a little dog.
b. MOT: where’d [: where did] it go? CHI: down. MOT: down.

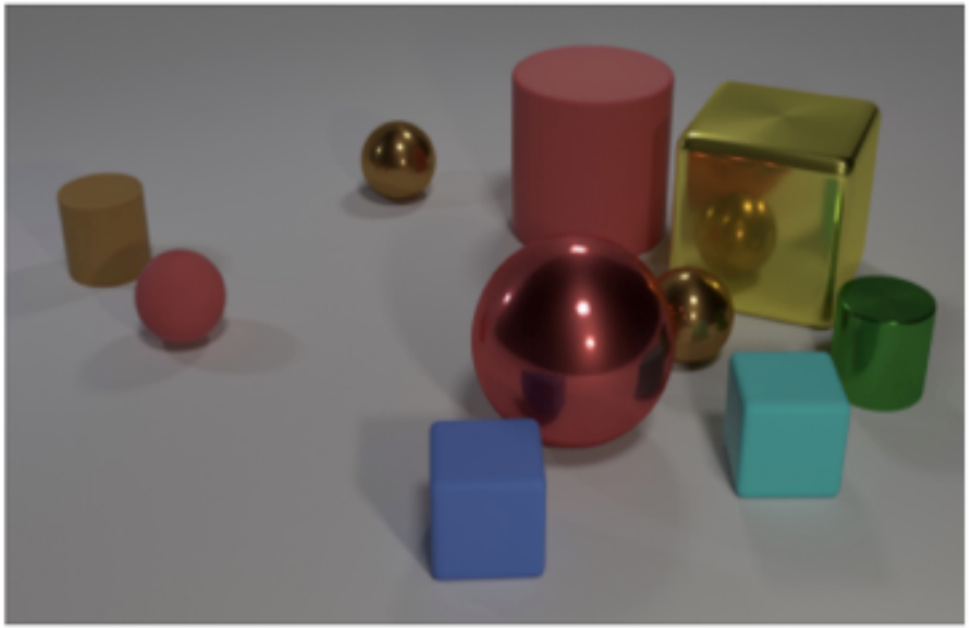
Polar not answered:

MOT: who’s that? is that the doctor?

Polar questions answered were request polars:
MOT: you want some rice? Child: (reaches out with bowl)

“the answer that can be provided to such questions in “training sessions” between parent and child is <u>easier to ground perceptually</u> than the abstract entities expressed by propositional answers required for polar questions.”

A diagnostic Dataset for VQA models



attribute, counting, comparison, spatial relationships, logical operations

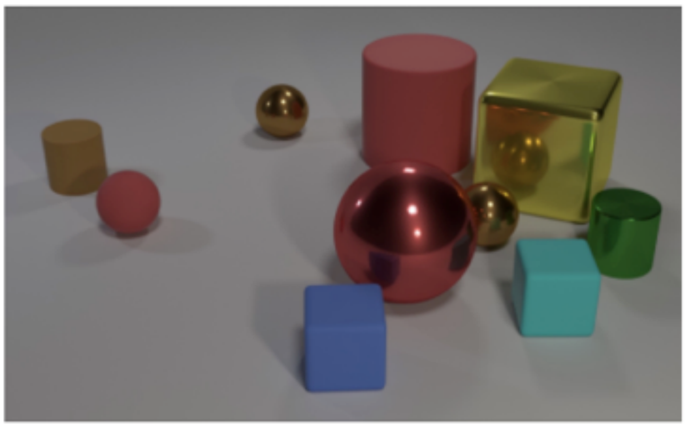
attribute q → Wh
(color, shape, material and size)

comparison q → Y/N

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?

Johnson et al 2017

Experiments



1. Does the model benefit from learning Y/N-Q after having learned Wh-Q?
2. Does the model forget Wh-Q after having learned Y/N-Q?
3. What if the order of the two tasks is reversed?

Task Wh-Q

Q: *What size is the cylinder that is left to the yellow cube*?

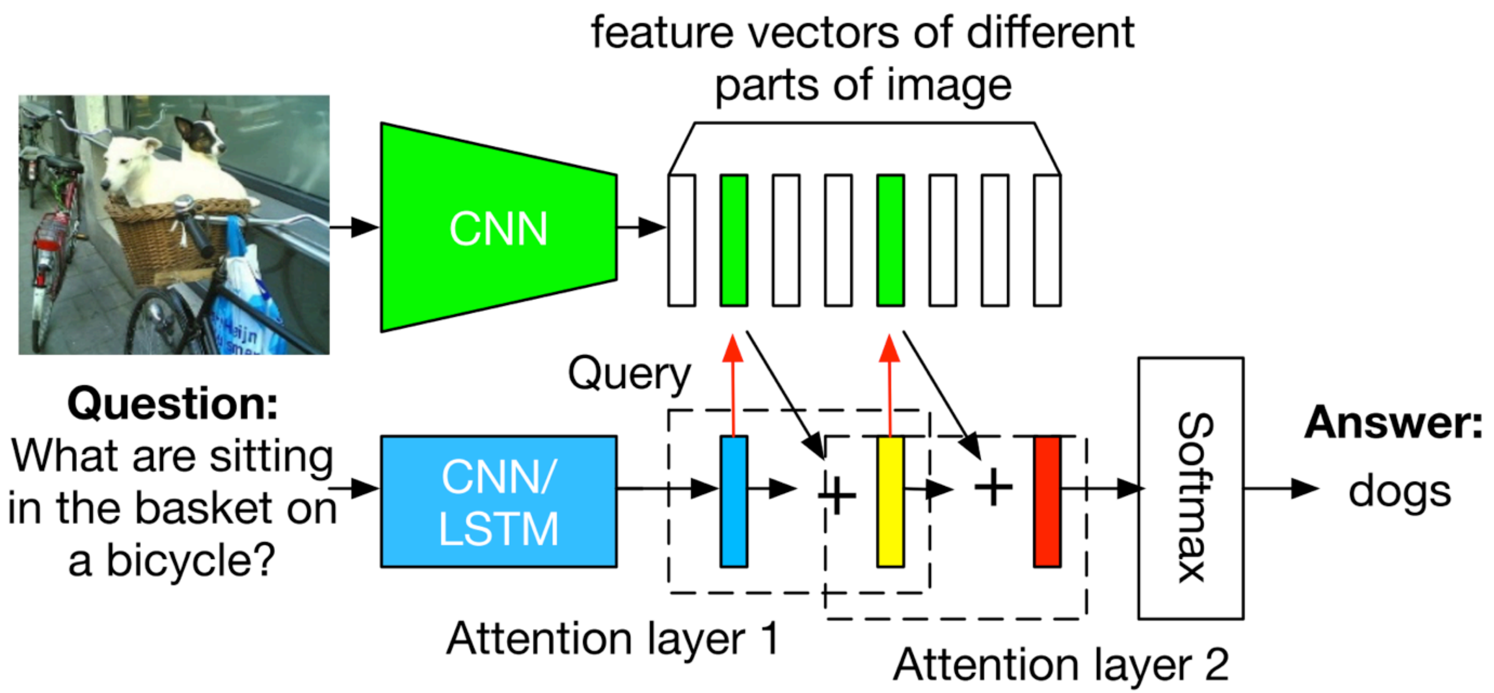
A: Large

Task Y/N-Q

Q: *Does the red bal have the same material as the large yellow cube*? **A**: Yes

equal # datapoint per task

Model: Stacked Attention Network

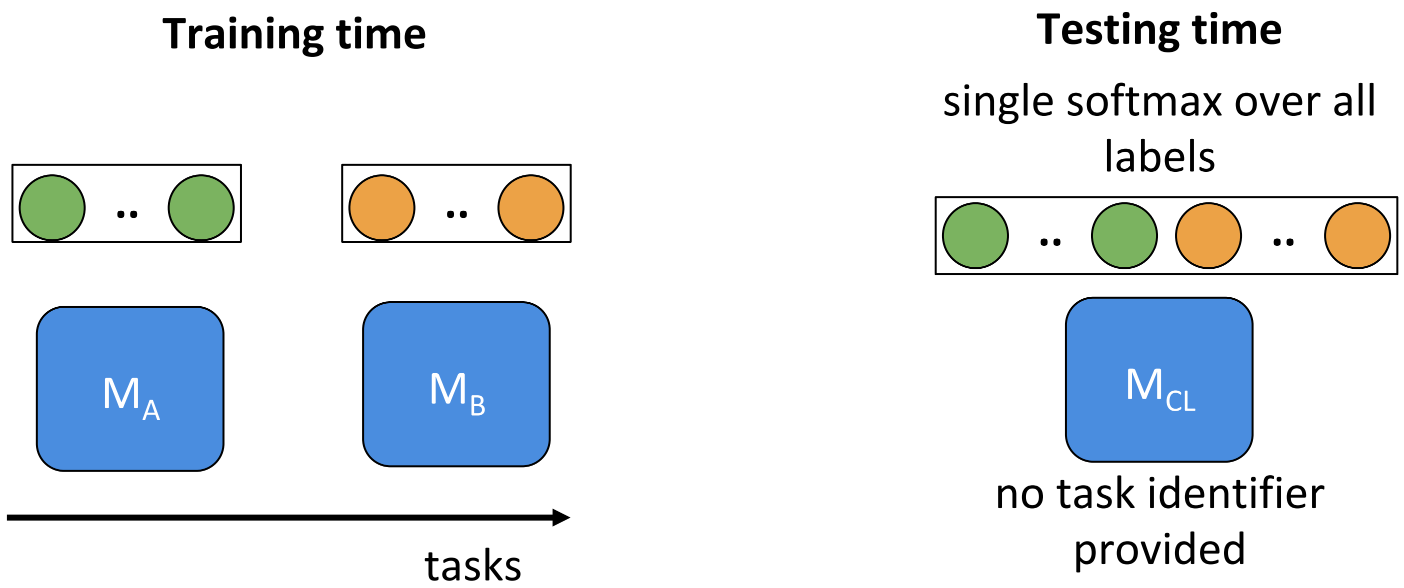


Yang et al. 2015

Wh-Q easier than Y/N-Q

	Wh-Q	Y/N Q
Random baseline	0.09	0.50
LSTM-CNN-SA	0.81	0.52

Training Setup:



Training Methods

Naïve: trained on Task A and then *fine-tuned* on Task B

Cumulative: trained on the training sets of *both* tasks

Continual Learning methods

	Wh-Q	Y/N Q
LSTM-CNN-SA	0.81	0.52

Naïve:
trained on Task A, then *finetuned* on Task B

Cumulative:
trained on the training sets of *both* tasks

- The model improves on Y/N -Q if trained first/ together with Wh-Q
- The model forgets about Wh-Q after having learned Y/N-Q

Wh → Y/N		
	Wh	Y/N
Random (both task)	0.04	0.25
Naïve	0.00	0.61
Cumulative	0.81	0.74

Vs.

The model does not improve on Wh-Q after having learned Y/N-Q

The model forgets Y/N-Q after having learned Wh-Q

Y/N → Wh		
	Y/N	Wh
Random (both task)	0.25	0.4
Naïve	0.00	0.81
Cumulative	0.74	0.81

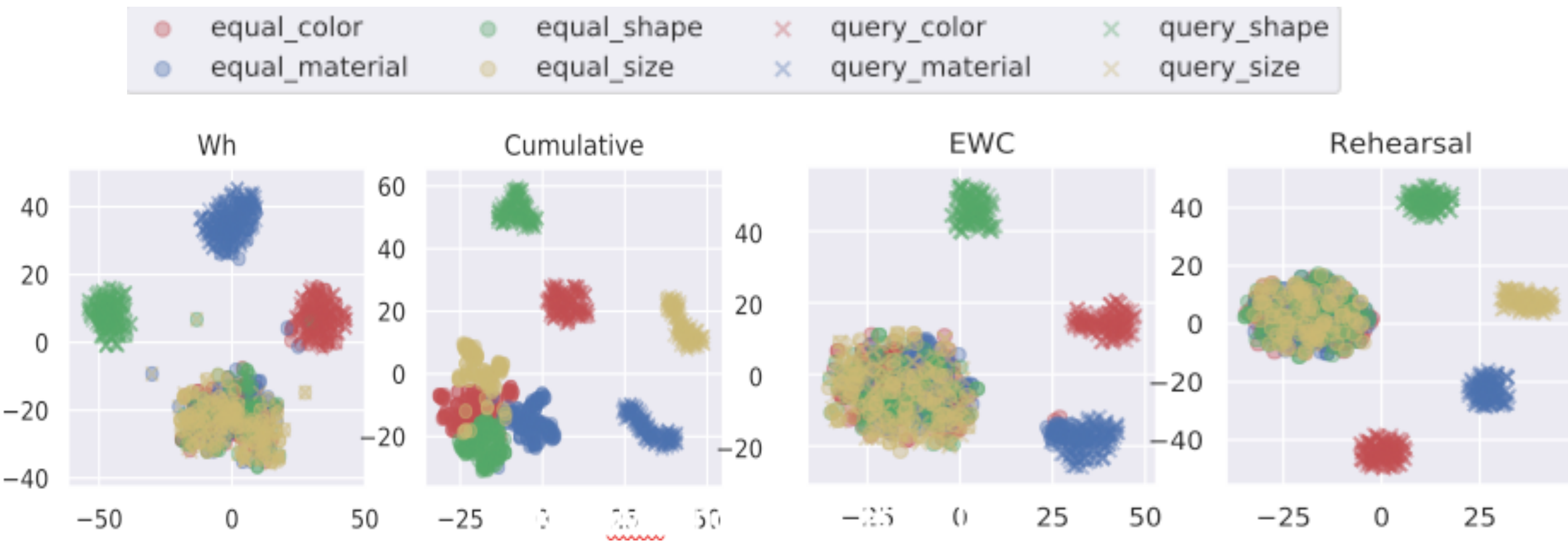
Note: training on both types of questions together improves Y/N

Continual Learning training methods

- *Elastic Weight Consolidation (EWC), (Kirkpatrick et al 2017):* has a parameter that should help the model to reduce error for both tasks.
- *Rehearsal (Robins 1995)*: trained on Task A, then fine-tuned through batches taken from a dataset of Task B and rehearsed on small number of examples from Task A.

Analysis

Analysis of the neuron activations on the penultimate hidden layer



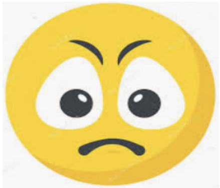
Task A: Wh and Task B: Y/N

Conclusion

1. Do VQA models benefit from learning question types of incremental difficulty?

Yes: 

2. Do they forget how to answer question types previously learned?

Yes: 

These results call for studies on how it is possible to enhance visually-grounded models with continual learning methods

➔ See T. L. Hayes et al in arXiv

They Are Not All Alike: Answering Different Spatial Questions Requires Different Grounding Strategies

Alberto Testoni[1], Claudio Greco[1], Tobias Bianchi[3], Mauricio Mazuecos[2], Agata Marcante[4], Luciana Benotti[2], Raffaella Bernardi[1]

[1] University of Trento, Italy
[2] Universidad de Córdoba, Conicet Argentina
[3] ISAE-Supaero, France
[4] Université de Lorraine, France

Third International Workshop on Spatial Language Understanding, SpLU 2020

Spatial Reasoning

Do VQA models apply different strategies when answering different types of spatial questions?

Does the attention of the models differ when answering different types of questions?

Baseline Oracle Accuracy per Question Type

	Frequency (%)	Accuracy (%)
Entity	44.38	93.37
Spatial	33.73	67.30
Color	8.07	61.64
Action	3.46	64.32
Size	0.60	60.41
Texture	0.61	69.92
Shape	0.19	68.44
Not classified	8.96	75.02
Total	100	75.94

Attribute Questions

Q-classification scheme from Shekhar R. et al., 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In Proceedings of NAACL 2019

Experiment 1 - Accuracy per Question Type

	LSTM (%)	V-LSTM (%)	LXMERT-S (%)	LXMERT (%)
Entity	93.37	83.24	88.64	91.09
Spatial	67.30	66.40	71.31	77.00
Color	61.64	68.06	70.51	76.42
Action	64.32	65.44	70.23	77.16
Size	60.41	62.76	67.23	75.44
Texture	69.92	66.15	71.92	77.47
Shape	68.44	64.12	70.76	74.42
Not classified	75.02	70.45	74.94	82.18
Total	75.94	72.70	77.41	82.21

Better than LSTM

Worse than LSTM

Spatial Question Classification

Manual observations of patterns:
- Relational questions: PP NP (PRO/ENTITY)
- Absolute questions: location word
- Group questions: number (group/order)

Automatic classification:
by identifying nouns, prepositions, and numbers using PoS Stanza (Qui et al 2020)

	Freq. %	Example
Relational	31.9	Is it the pen **behind** the **PC**?
Absolute	31.8	Is it the one on the **left**?
Group	17.3	Is it among the **4** women?
Other	19.0	Can you sleep on it?

Experiment 2 – Accuracy on Spatial Questions

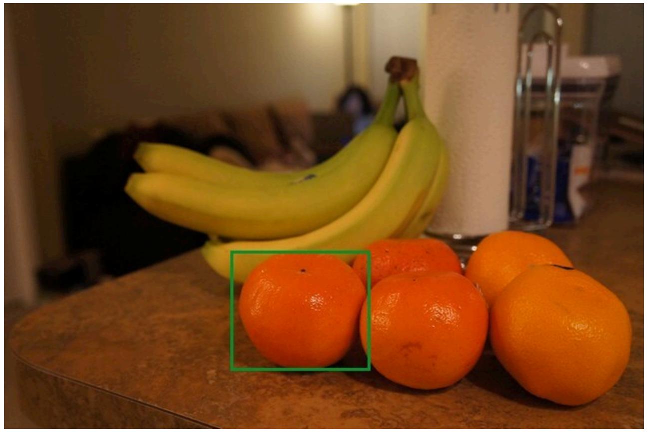
	LSTM	V-LSTM	LXMERT-S	LXMERT
Absolute	76.4	75.2	80.5	83.4
Relational	67.1	63.5	69.6	77.2
Group	63.3	62.8	68.4	71.6

Worse than LSTM

Better than LSTM

Error Analysis: the Role of the Dialogue History

Manual error analysis of 20% of LXMERT errors on spatial questions.
For absolute and group questions, ~50% of errors are related to missing dialogue history.



1. Is it a fruit?	Yes
2. Is it an orange?	Yes
3. Is it on our right?	No
4. In the middle?	No
5. The last single one?	Yes

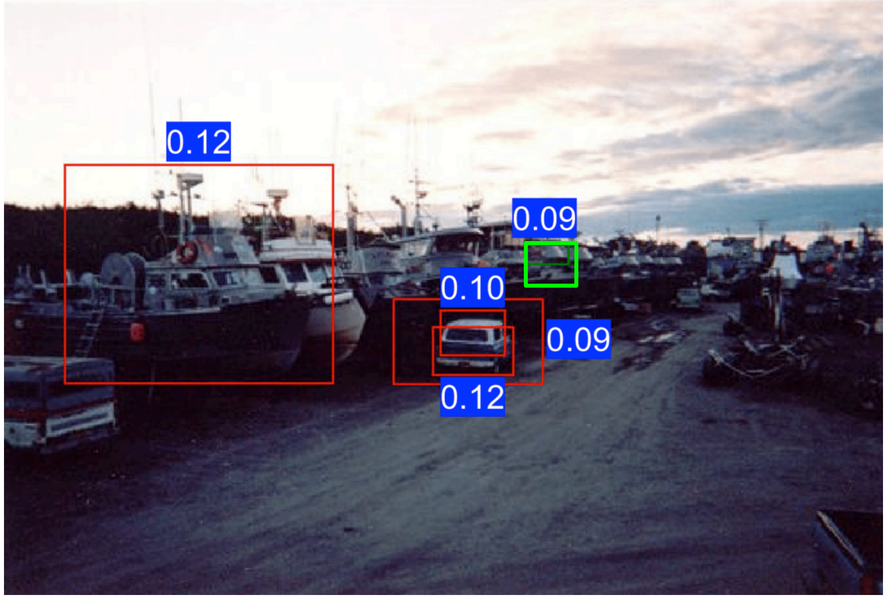
LXMERT Attention Analysis

Is it the bus on the left?

Absolute Question

LXMERT Attention Analysis

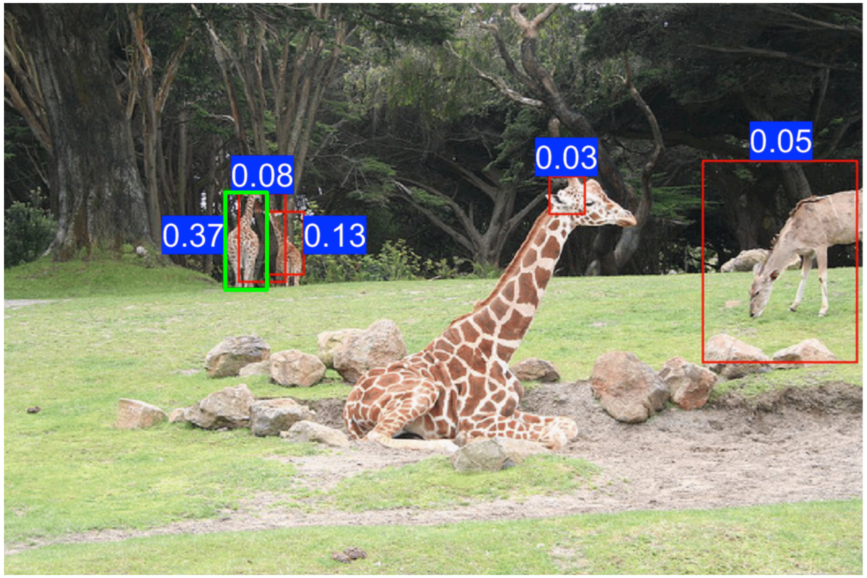
Is it the boat next to a car?



Relational Question

LXMERT Attention Analysis

Is it one of the two in the back?



Group Question

Summary of Contributions and Conclusion

- We adapted LXMERT to play the role of the Oracle of the GuessWhat?! Game, obtaining an overall accuracy of 82.21% (+6.27% with respect to the usual baseline).
- LXMERT improves over the baseline also on spatial questions (+9.70%), but they remain a large source of errors also for this model – with 77.00% accuracy.
- We propose a new classification method for spatial questions. The fine-grained evaluation shows that the hardest spatial questions are the relational and group ones.
- Our qualitative analysis shows that LXMERT's attention shows different patterns for absolute and relational questions as expected. Moreover, we found that some spatial questions need the dialogue history to be interpreted correctly.

(Internship) Projects

- Multimodal Spatial Reasoning Dota
- Ensemble Models for GuessWhat?! Daniel

Be Different to Be Better

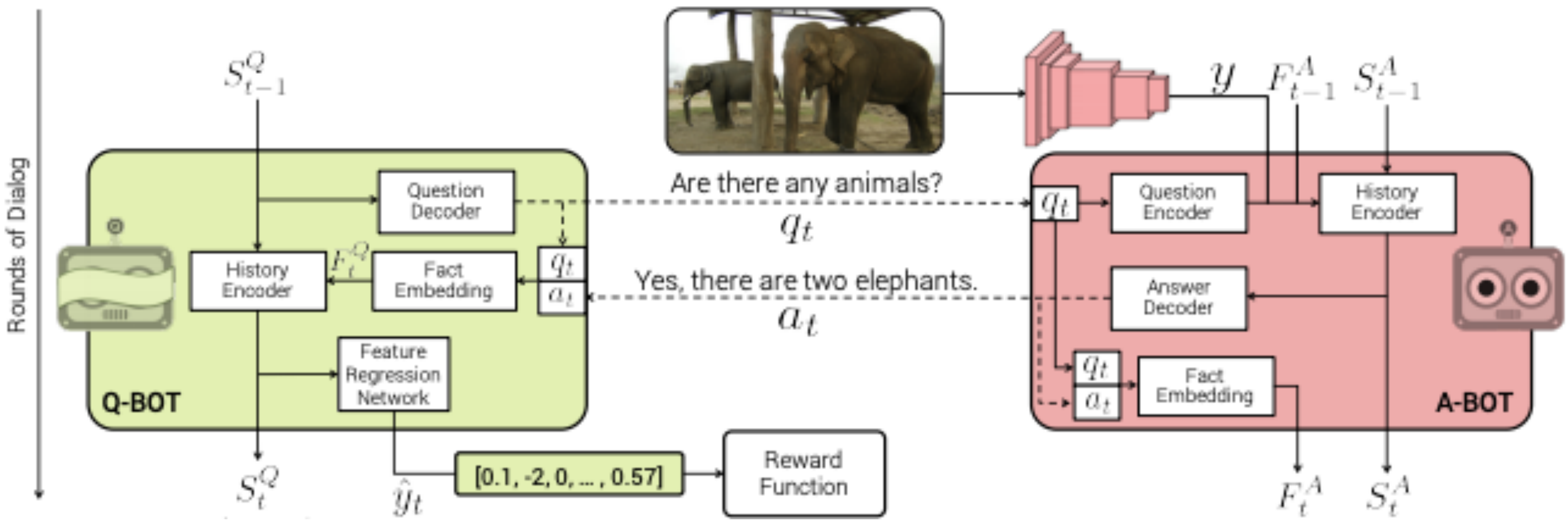
If I am feeling alone

- ❑ I cry
- ❑ I join the group
- ❑ …

We have collected the data and cleaned them.
We are building the data to train and evaluate the models on the task.

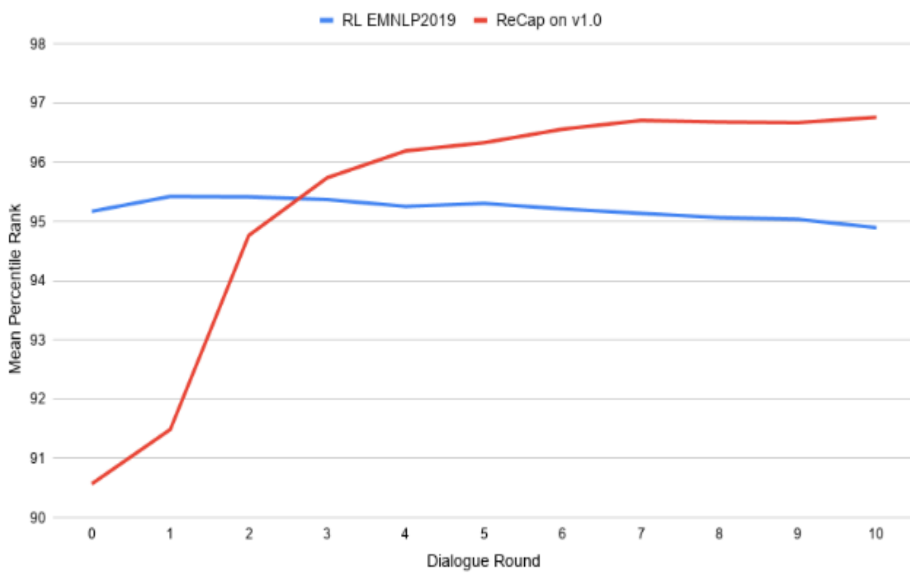
We will need to adjust baselines to be trained and evaluated.

But “new” model..
Diverse Q-BOT



Diverse Q-Bot (EMNLP 2019): receives a penality when it asks a question similar to the one asked in the previous turn

Re-Cap vs. Diverse-QBot



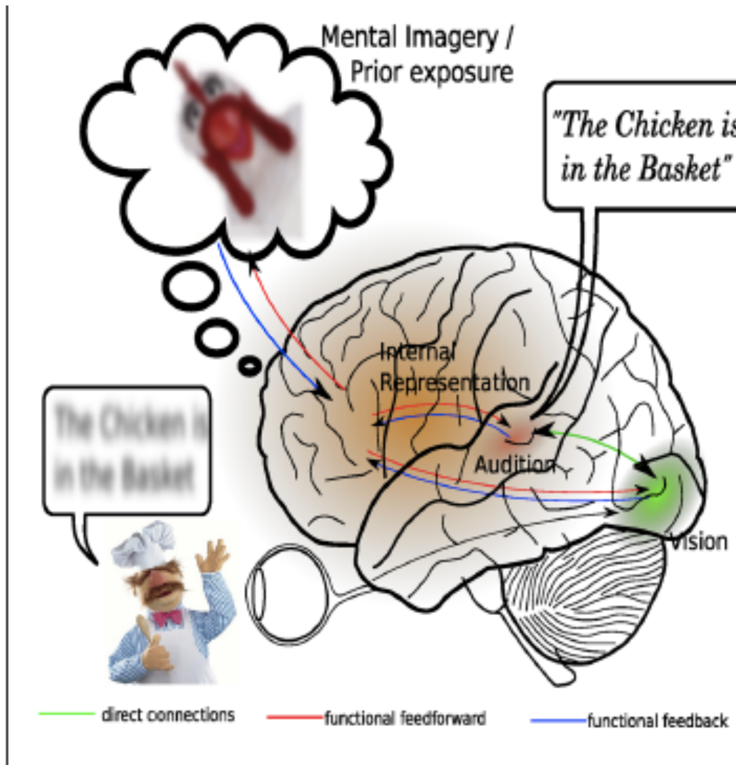
Diverse Q-Bot: 94.8
ReCap: 96.76

Training: 120K (VisDial 1)
Candidate images: 2K

	Novel questions ↑	Novel questions (no duplicates) ↑	Unique questions during dialogue ↑	Mutual overlap ↓	Games with repeated Qs ↓
EMNLP	429	376	8.22	0.41	81.17%
ReCap	**1319**	**1250**	**8.80**	**0.27**	**62.74%**

↑: higher is better
↓: lower is better

Mental Imagery module
Prior exposure

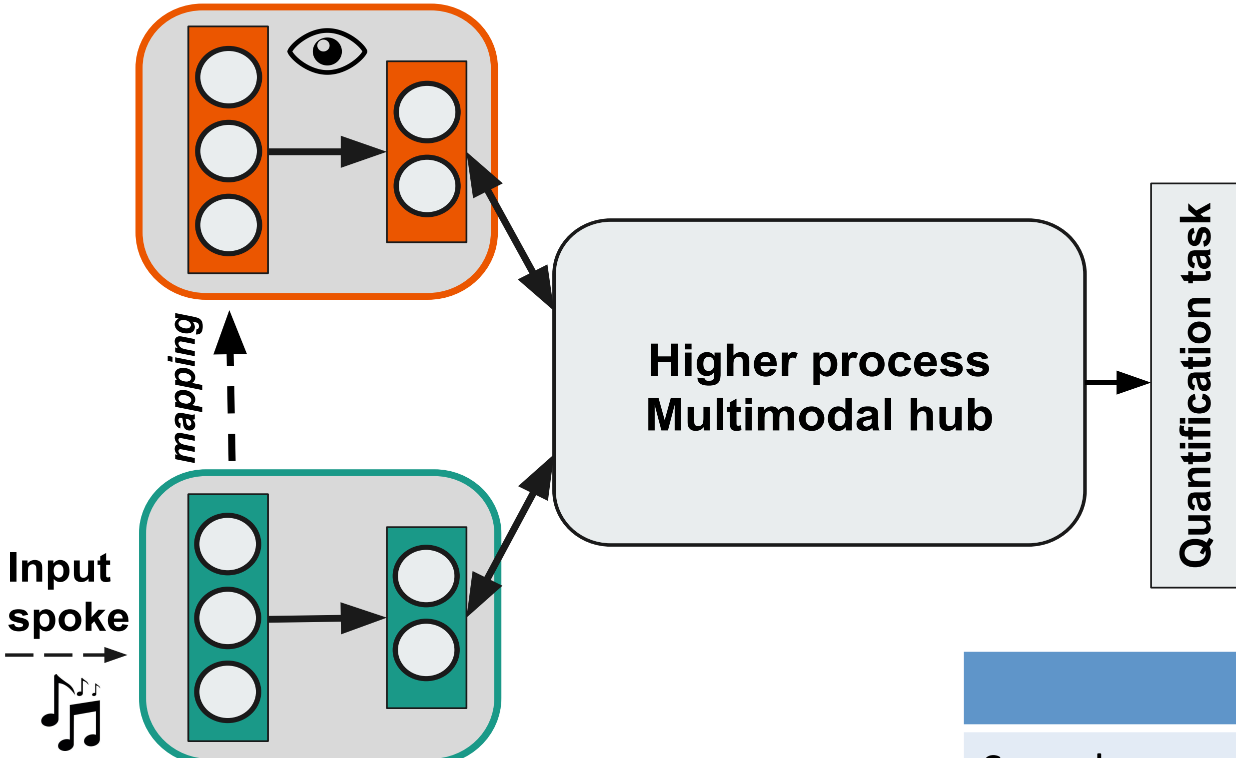


credits to Talsma 2015

Learning Quantifiers from audio-visual inputs

Audio-visual inputs aligned at the individual level

Imagining Vision from the Auditory Input



	Pearson's *r*
Sound	0.68
Vision	0.72
H&S	0.86
Audio-Vision *prior*	0.78