Belief State for Visually Grounded, Task-Oriented Neural Dialogue Model

Master Thesis Defense | Tim Baumgärtner | University of Amsterdam

Why we need (better) Dialogue Agents

Dialogue Agents are **ubiquitous, but weak**	Dialogue Interfaces have **high information bandwidth**	Dialogue Interfaces provide **high usability**
"75-80% of the time users only employ 4 skills: 'play music', 'set a timer', 'set a reminder', and 'what is the weather'." [Zitouni, 2019]	"The interface all necks down to this tiny straw, which is, particularly in terms of output, it's like poking things with your meat sticks" [Musk, 2017]	"People already have extensive communication skills through their own native or natural language [...]. Natural language interfaces can provide the most useful and efficient way for people to interact with computers. [...] The goal [...] is to provide an interface that minimizes the training required for users." [Ogden and Bernick, 1997]

Dimensions of Dialogue

Chatbots vs Task-Oriented	Grounding	I/O
Task-Oriented: Achieve a *goal* for the user. → Book hotel or restaurant, retrieve a support article, route call to expert → Straight-forward Evaluation Chatbot: Engage in conversation with user	Additional sensory input, e.g. visual perception or audio → More realistic and improved semantic understanding [Barsalou 2008, Harnad 1990] Many other applications in NLP: Image Captioning, MT, Q&A	Spoken vs Text Based Rule Based vs Retrieval vs NLG

Dialogue System **Pipeline**

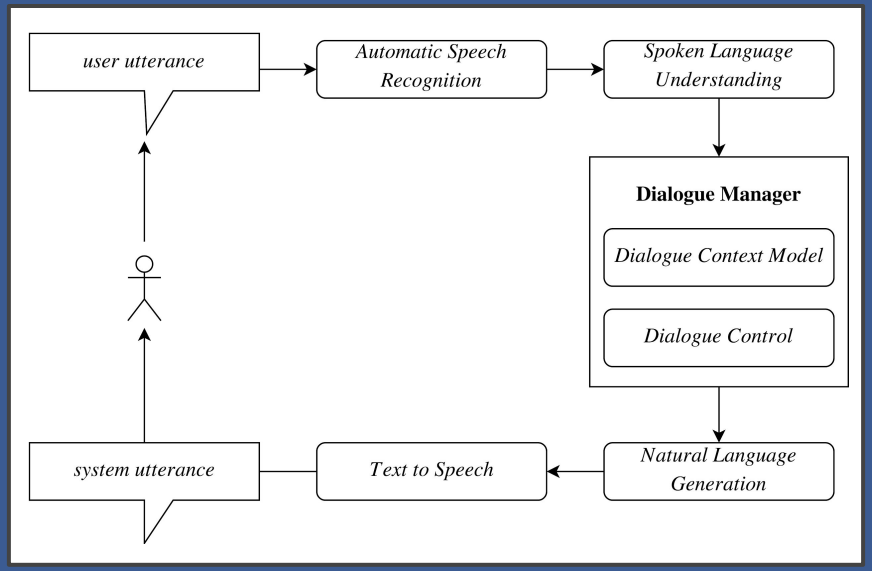
Divides problem into manageable sub-modules

Annotations for sub-modules required

Dialogue Manager steers natural language generation:

Dialogue Context Model: Keeps tracks of information provided in **Belief State**

Dialogue Control: Decides what action to take next



[Jokinen and McTear, 2009]

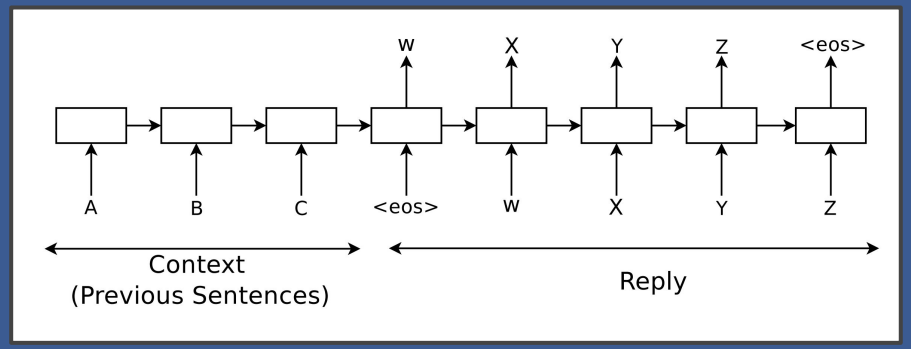
End-to-End Dialogue Systems

Learn in data-driven manner, directly from dialogues
→ Other annotations not required
→ Intermediate representations need to be learned from dialogues

Often lack coherent and diverse answers:

Q: *what is your job ?* **A**: *i 'm a lawyer .*
Q: *what do you do ?* **A**: *i 'm a doctor .*
[Vinyals & Le, 2015]

Q: *What are you doing?*
Top Answers: *I don't know.*
I don't know!
Nothing.
[Li et al., 2016]



[Vinyals and Le, 2015]

Belief State in Task-Oriented, End-to-End Dialogue Model

Dialogue System Pipeline approach requires **intermediate annotations**

End-to-End approach has sub-optimal language generation due to **weak intermediate representations**

→ **Add Belief State to End-to-End approach**

Requirements:

Summarize established information
Represent uncertainty
Without intermediate labels

Implementation:

Evaluate intermediate dialogues and use p_{task} as belief state
Condition NLG on belief state

Belief State in Task-Oriented End-to-End Dialogue Model

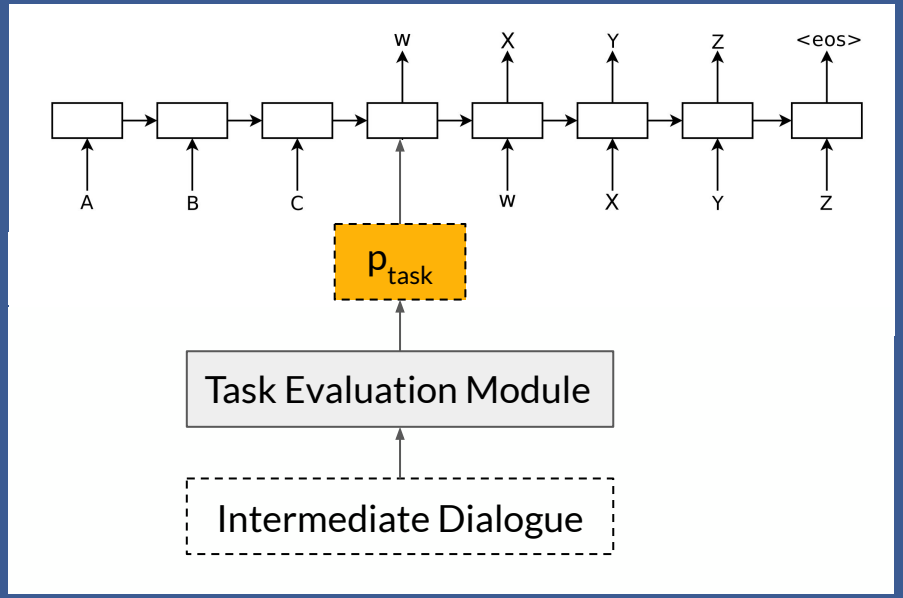
Pipeline approach requires **intermediate annotations**

End-to-End approach has weaknesses in NLG due to **implicit intermediate representations**

→ **Add Belief State to End-to-End approach**

Summarize established information
Represent uncertainty
Without intermediate labels

Evaluate intermediate dialogues and use p_{task} as belief state
Condition NLG on belief state



GuessWhat?! Task

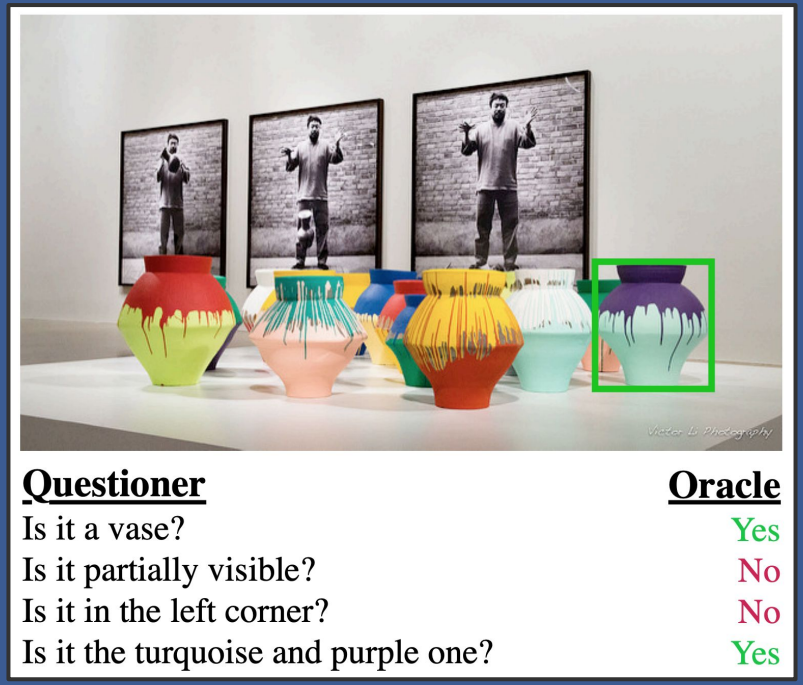
Task: **Identify an object in visual scene through series of yes/no questions**

Task-Oriented Dialogue (~155k)

Visually Grounded, MS COCO [Lin et al., 2014]

2 Agents: Questioner and Oracle

Evaluation:
Questioner generates Q - Oracle answers
After **n** Questions: Questioner choses Object



[de Vries et al., 2017]

The Questioner

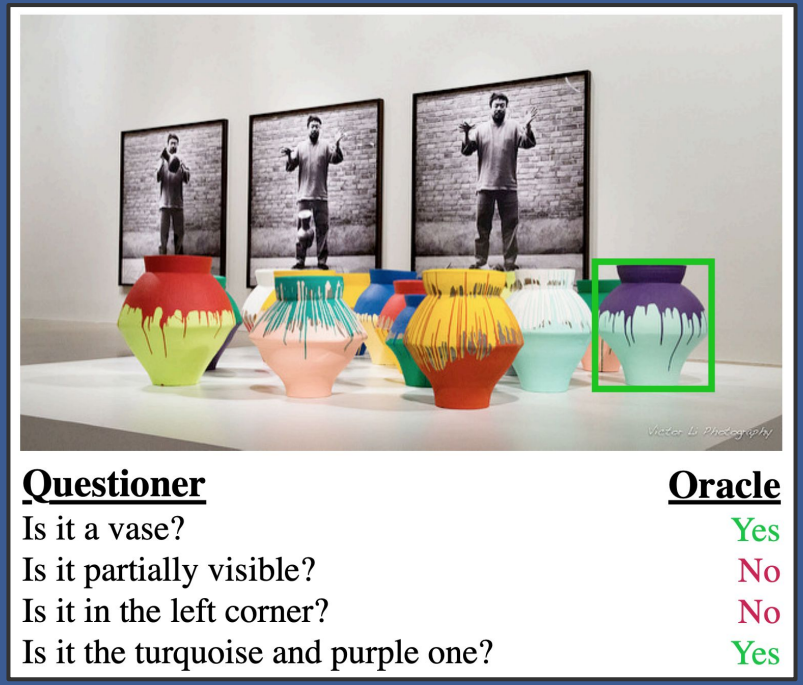
Division in two modules:
Question Generator & Guesser

Question Generator Challenges
Generate questions that Oracle can understand

Generate coherent dialogue such that Guesser can identify object

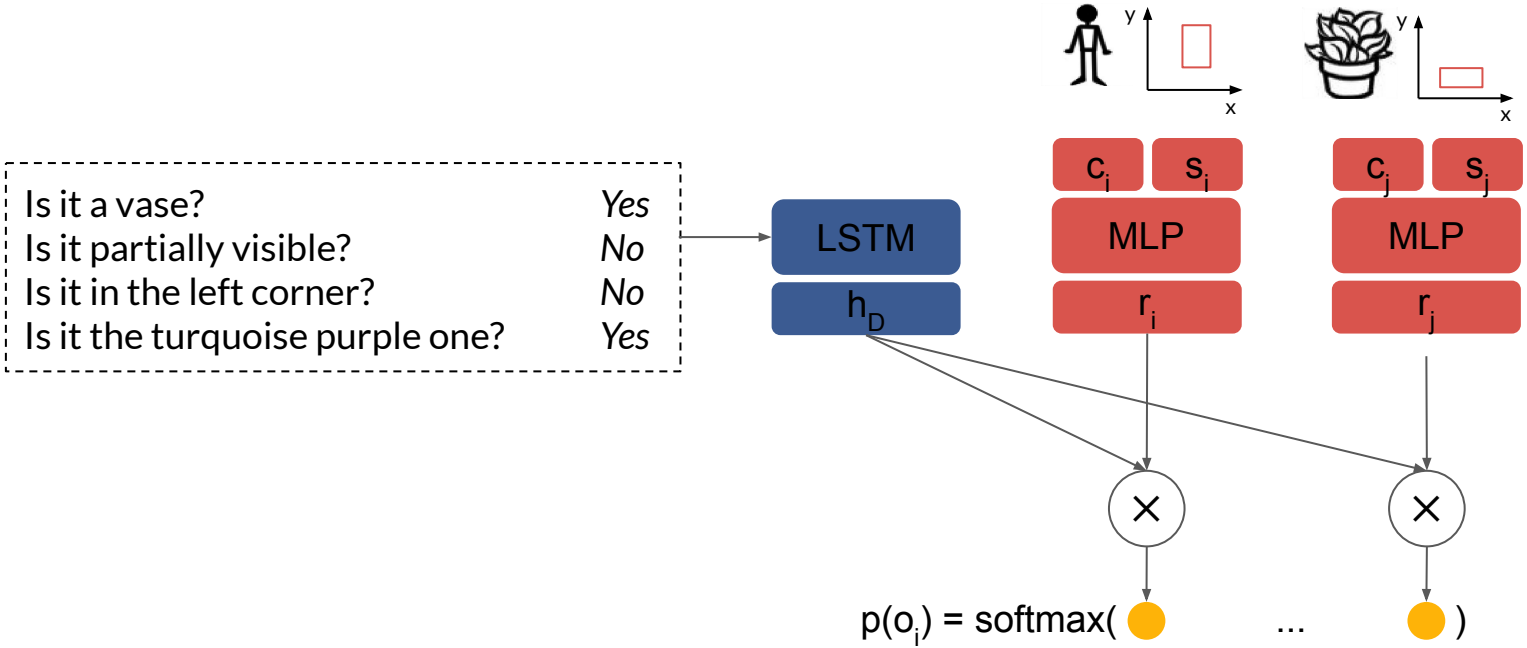
Required Skills

- Visual Understanding
- Natural Language Generation
- Natural Language Understanding



[de Vries et al., 2017]

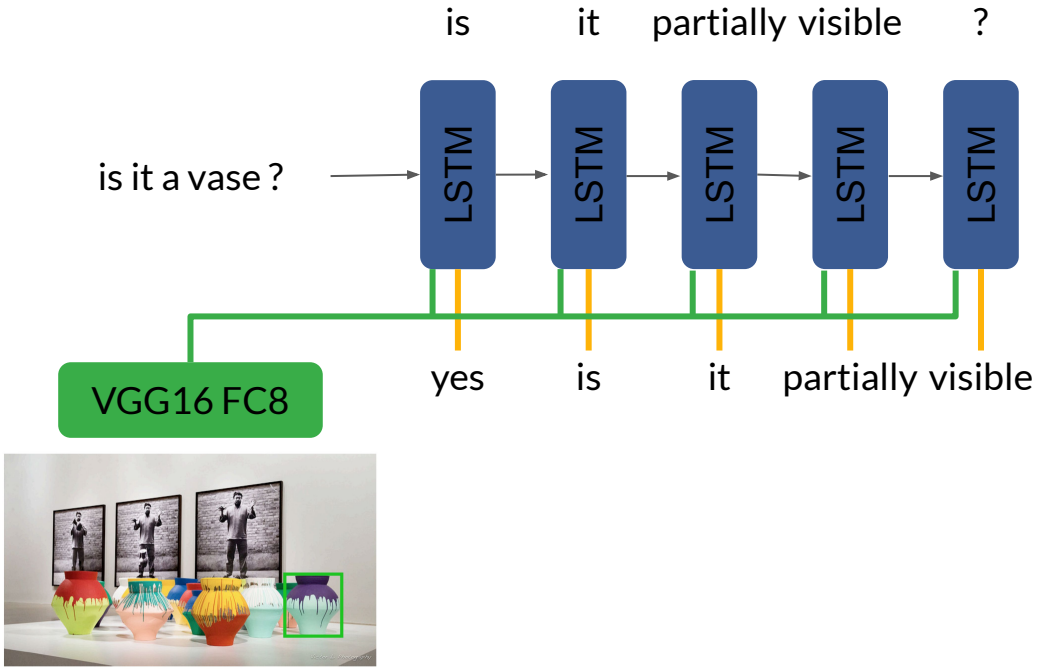
Guesser Architecture



[de Vries et al., 2017]

Question Generator: **Baseline** Architecture

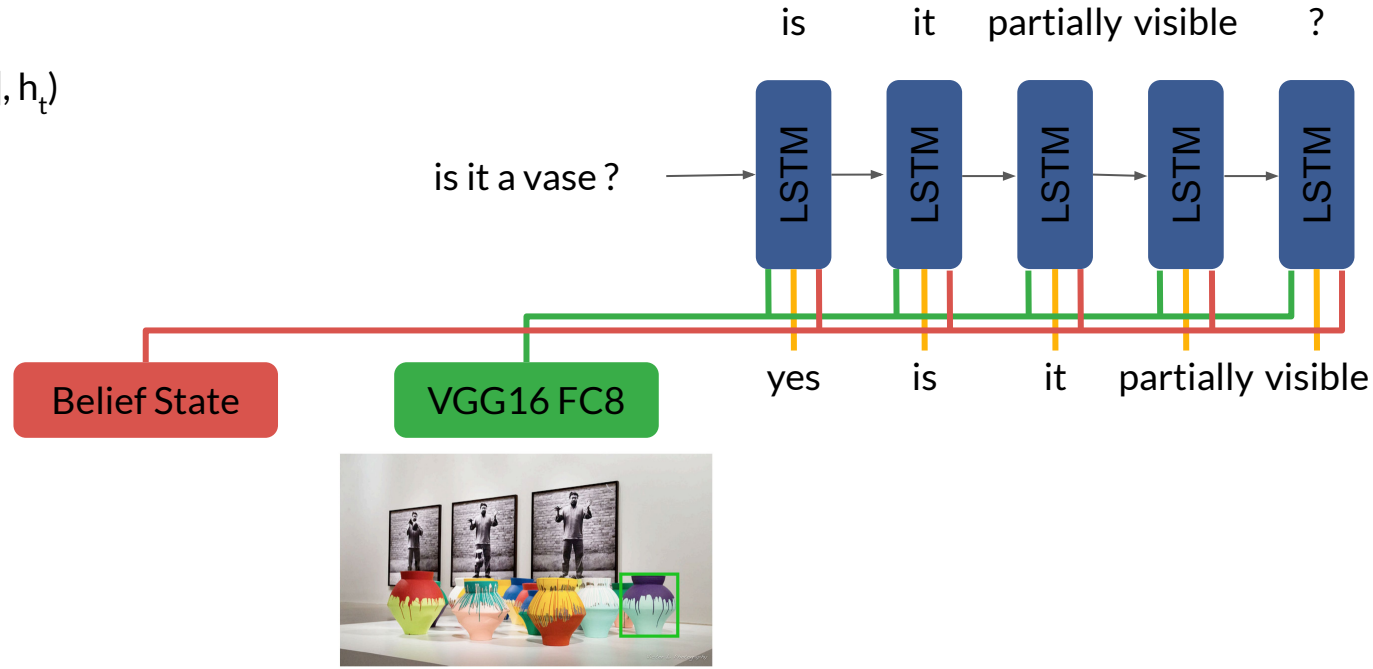
$h_{t+1} = \text{LSTM}([w_t, v], h_t)$



[de Vries et al., 2017]

Question Generator: **Belief** Architecture

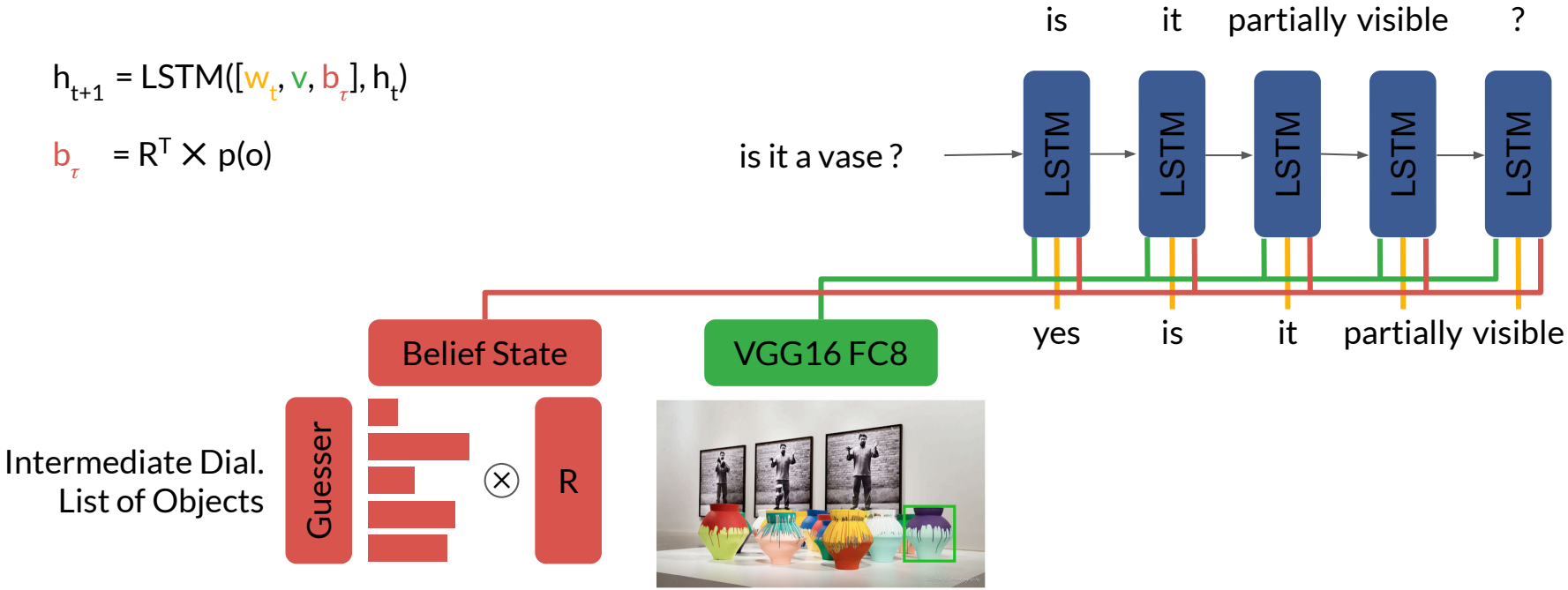
$h_{t+1} = \text{LSTM}([w_t, v, b_\tau], h_t)$



Question Generator: Belief Architecture

$h_{t+1} = \text{LSTM}([w_t, v, b_\tau], h_t)$

$b_\tau = R^T \times p(o)$

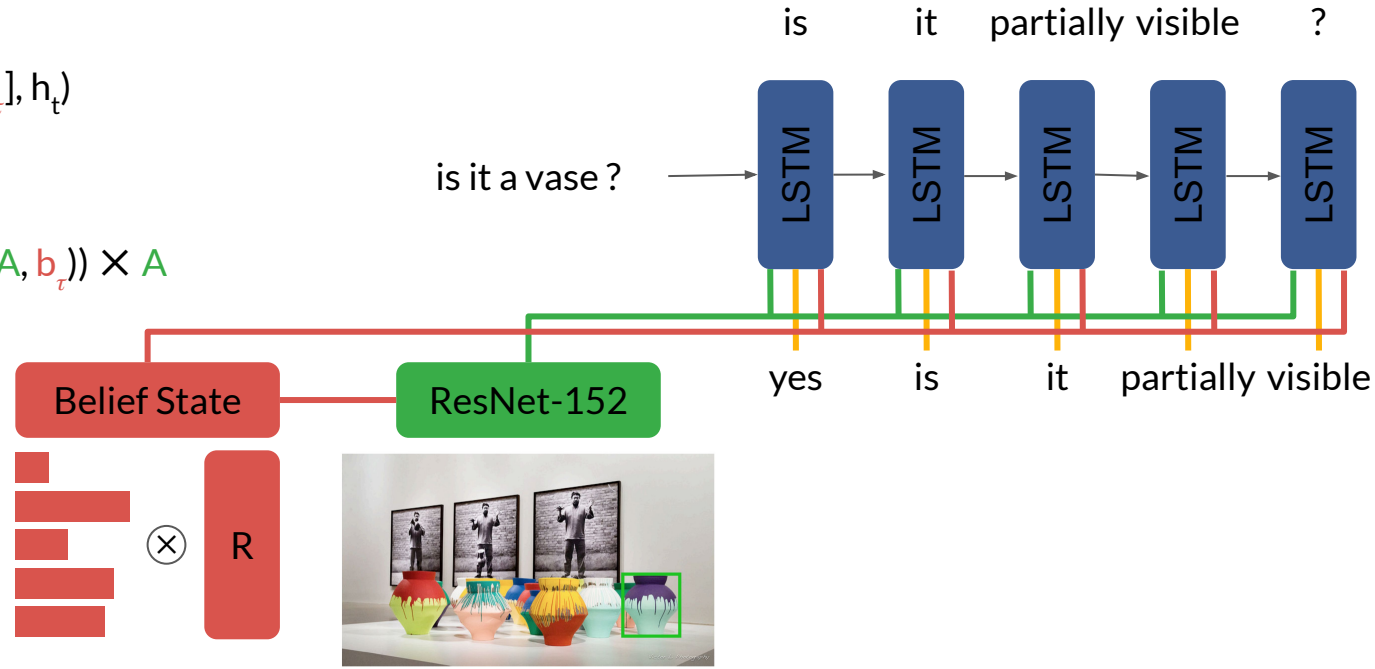


Question Generator: Belief w/ Visual Attention

$h_{t+1} = \text{LSTM}([w_t, v_\tau, b_\tau], h_t)$

$b_\tau = R^T \times p(o)$

$v_\tau = \text{softmax}(\text{score}(A, b_\tau)) \times A$



Experiments & Results

Object Representation for Belief State

Category: $\mathbf{R} = c \times W_{category}$

Category+Spatial: $\mathbf{R} = MLP([c \times W_{category}, s])$

Guesser Obj. Rep.: $\mathbf{R} = MLP_{Guesser}([c \times W_{category}, s])$

Belief State Fine-Tuning

Freeze “Guesser” Parameters

Update “Guesser” Parameters through Question Generator Loss

Experiments & Results: **Object Representations**

Belief State Representation	Cross Entropy	Test Accuracy (n=5)	Test Accuracy (best n)
Baseline [de Vries, 2017]	1.475	42.55%	42.55% (n=6)
Category	1.443	48.30%	49.60% (n=8)
Category+Spatial	**1.433**	**49.49%**	**50.23% (n=8)**
Guesser Obj. Rep	1.436	48.57%	49.06% (n=8)

Experiments & Results: **Belief State Fine-Tuning**

Belief State Representation	Cross Entropy	Test Accuracy (n=5)	Test Accuracy (best n)
Baseline [de Vries, 2017]	*1.475*	*42.55%*	*42.55% (n=6)*
Category+Spatial (frozen)	*1.433*	*49.49%*	*50.23% (n=8)*
Category	1.428	53.83%	54.65% (n=8)
Category+Spatial	**1.432**	**54.75%**	**55.63% (n=7)**
Guesser Obj. Rep	1.437	54.46%	55.22% (n=7)

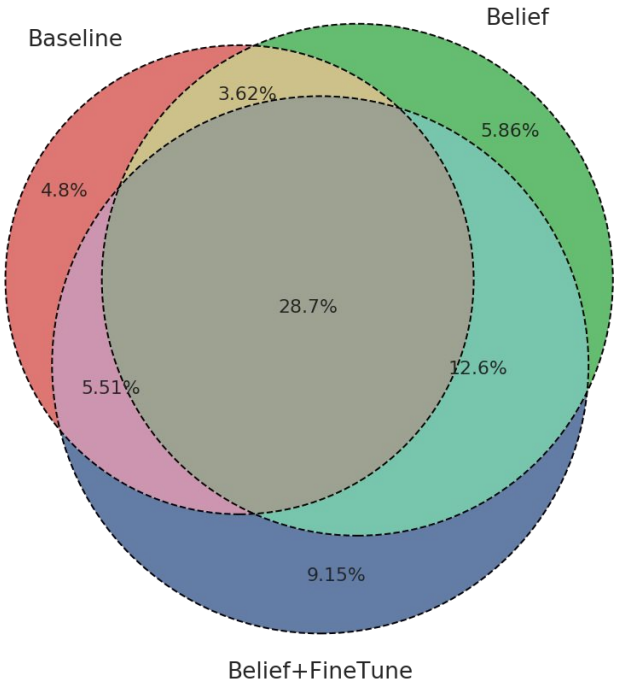
Experiments & Results: **Visual Attention**

Attention Query	Fine Tuning	Cross Entropy	Test Accuracy (n=5)	Test Accuracy (best n)
Hidden	n/a	1.450	43.19%	43.13% (n=6)
Category+Spatial	✗	1.445	44.38%	45.06% (n=6)
Category+Spatial	✔	1.440	45.86%	56.04% (n=6)
Category+Spatial	✗	1.430	47.66%	48.78% (n=8)
Category+Spatial	✔	1.422	54.23%	55.11% (n=7)

Analysis: **Models**

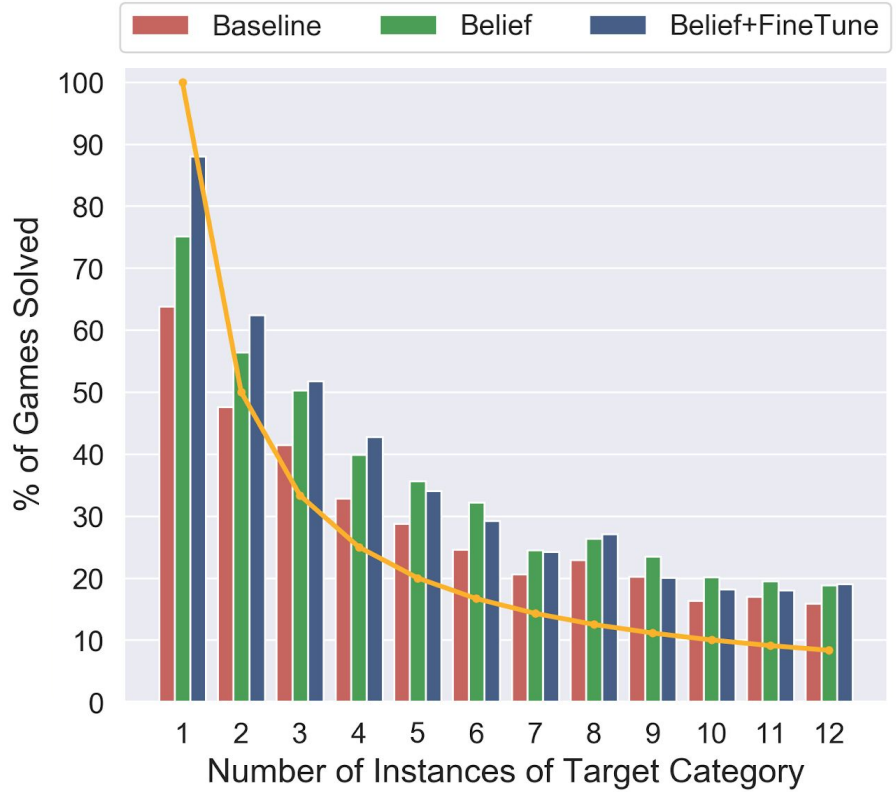
Model	Validation Accuracy (n=5)	Validation Accuracy (best n)
Baseline	42.90%	43.05% (n=6)
Belief	50.00%	50.78% (n=8)
Belief+FineTune	55.08%	56.15% (n=7)

Analysis: **Solved Games**



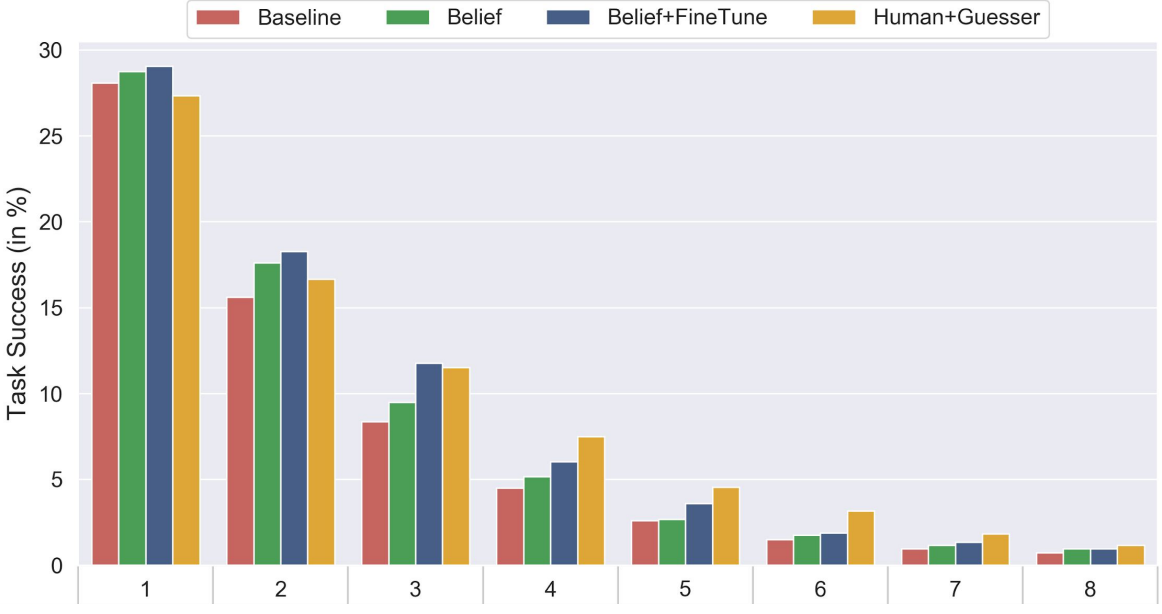
	All Games	Baseline	Belief	Belief +FineTune
Num Games	23,739	10,122	12,056	13,284
Num Objects	8.54 (±4.67)	6.84 (±4.07)	7.02 (±4.19)	7.07 (±4.14)
Num Object Categories	3.49 (±1.72)	3.28 (±1.49)	3.38 (±1.56)	3.51 (±1.61)
Num Instances of Target Cat.	3.99 (±3.59)	2.66 (±2.62)	2.68 (±2.63)	2.52 (±2.50)
Log of Target Object Area	8.64 (±2.00)	9.13 (±2.04)	9.17 (±2.02)	9.08 (±2.02)

Analysis: **Number of Objects**



Analysis: **Knowing when to stop**

	Baseline	Belief	Belief+ FineTune	Human+ Guesser	Human
Task Success	62.30%	67.58%	72.88%	72.13%	90.80%
Avg. (StD.) Num Q	2.19 (±1.55)	2.26 (±1.59)	2.35 (±1.60)	2.61 (±1.96)	5.07 (±3.23)



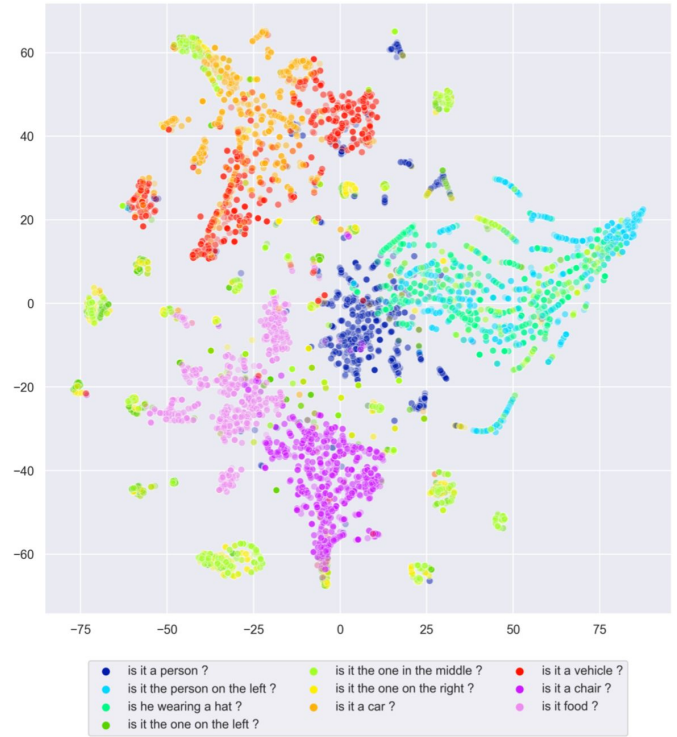
Analysis: **Influence of Belief State**

Does the generated question contain the category of the argmax of the belief probabilities?

Games	Up to 1st correct Guess	Baseline	Belief	Belief+ FineTune
All Games	✗	29.99%	50.04%	42.43%
All Games	✓	26.14%	43.63%	39.63%
Single Target Cat. Instance	✗	26.06%	47.70%	52.11%
Single Target Cat. Instance	✓	17.16%	33.70%	49.72%

Analysis: **Influence of Belief State**

(a) *Belief*



(b) *Belief+FineTune*

Analysis: **Ablation Studies**

Bag of Objects: Belief State with uniform probabilities
→ performance from providing the list of objects vs. belief probabilities

All Categories: Belief State over all 81 MS COCO Categories
→ performance from improved perception

Model	Belief State Representation	$\lvert\mathbf{W}_v\rvert$	$\lvert\mathbf{R}\rvert$	Cross Entropy	Validation Accuracy (n=5)	Validation Accuracy (best n)
Baseline	n/a	512	n/a	1.475	42.90%	43.05% (n=6)
Belief	Category	0	64	1.443	49.48%	49.94% (n=8)
All Categories	Category	64	64	1.456	44.52%	44.75% (n=8)
Belief	Category+Spatial	0	256	1.433	50.00%	50.78% (n=8)
Bag of Objects	Category+Spatial	64	64	1.431	45.32%	46.40% (n=7)

Analysis: **Qualitative Example**

Baseline

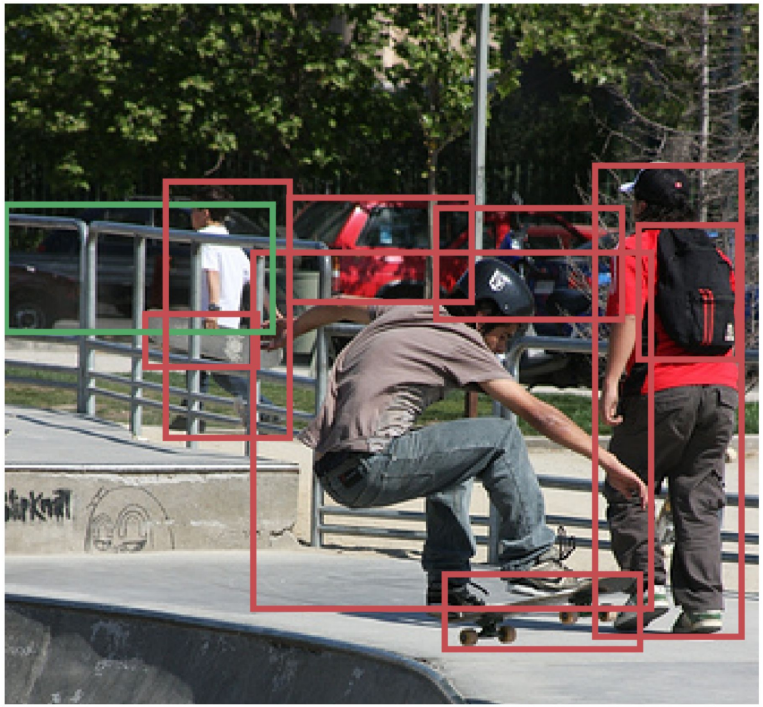
is it a person? no
is it a skateboard? no
is it a skateboard? no
is it a car? yes
is it the one on the right? no
is it the one on the right? no
is it the one on the right? no
is it the one on the right? no
→ Failure

Belief

is it a person? no
is it a skateboard? no
is it a car? yes
is it the one on the left? yes
is it the whole car? yes
is it the whole car? yes
is it the car on the left? yes
is it the whole car? yes
→ Success

Belief+FineTune

is it a person? no
is it a car? yes
is it the one in the middle? no
is it the one on the right? no
is it the one on the left? yes
the one that is cut off? yes
the whole car? yes
the whole car? yes
→ Success



Contributions

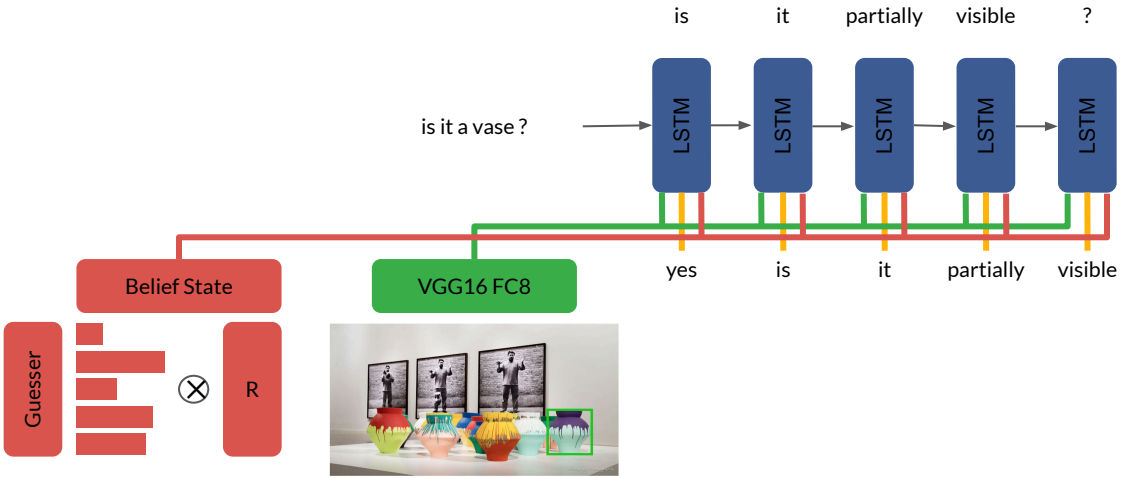
E2E trainable dialogue system with belief state

Application and experiments in GW?! scenario

SOTA for GW?! in supervised setting

Detailed linguistic and task-success analysis

Code base and web-based tool for qual. analysis



References

Zitouni I. (2019). http://ruder.io/aaai-2019-highlights/

Musk, E. (2017). https://waitbutwhy.com/2017/04/neuralink.html

Ogden, W. C., & Bernick, P. (1997). Using natural language interfaces. In Handbook of human-computer interaction (pp. 137-161). North-Holland.

Jokinen, K., & McTear, M. (2009). Spoken dialogue systems. Synthesis Lectures on Human Language Technologies, 2(1), 1-151.

Vinyals, O., & Le, Q. (2015). A neural conversational model. arXiv preprint arXiv:1506.05869.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5503-5512).

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of NAACL-HLT (pp. 110-119).