ASK YOUR NEURONS:

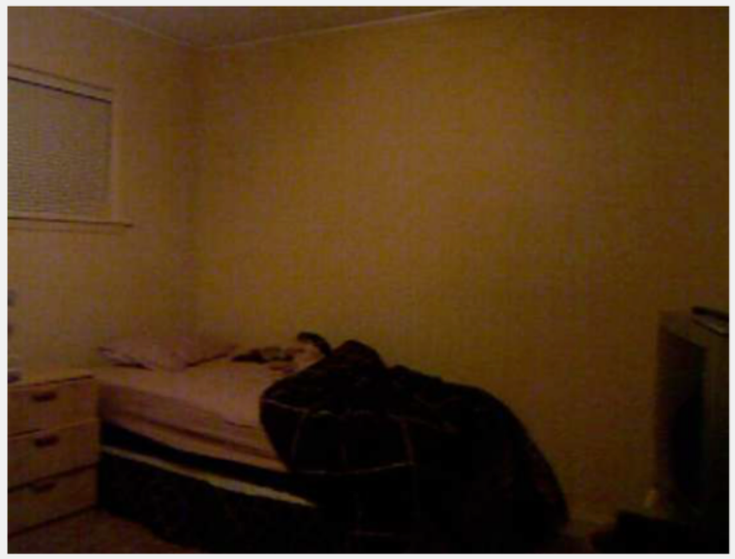
A NEURAL-BASED APPROACH TO QUESTION AND ANSWERING

Malinowski, Rohrbach, Fritz

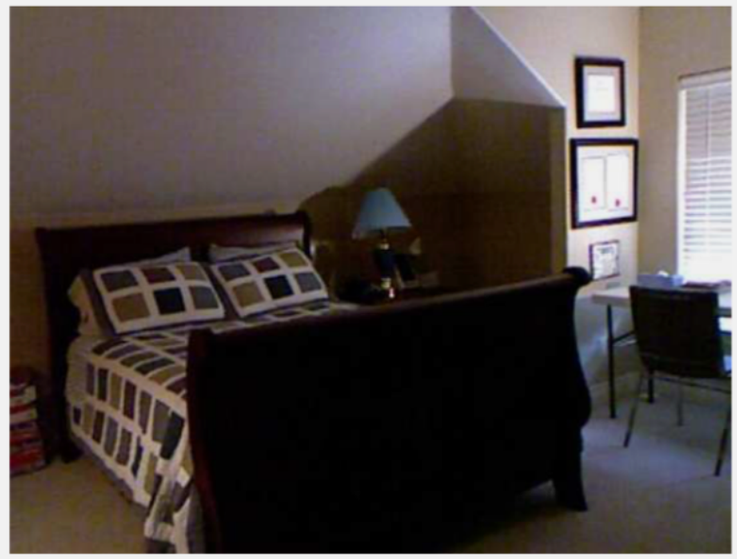
Presenter - Tyler Folkman

IMAGE QUESTION AND ANSWERING

What is on the right side of the cabinet?



How many drawers are there?



What is the largest object?

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

END TO END ARCHITECTURE

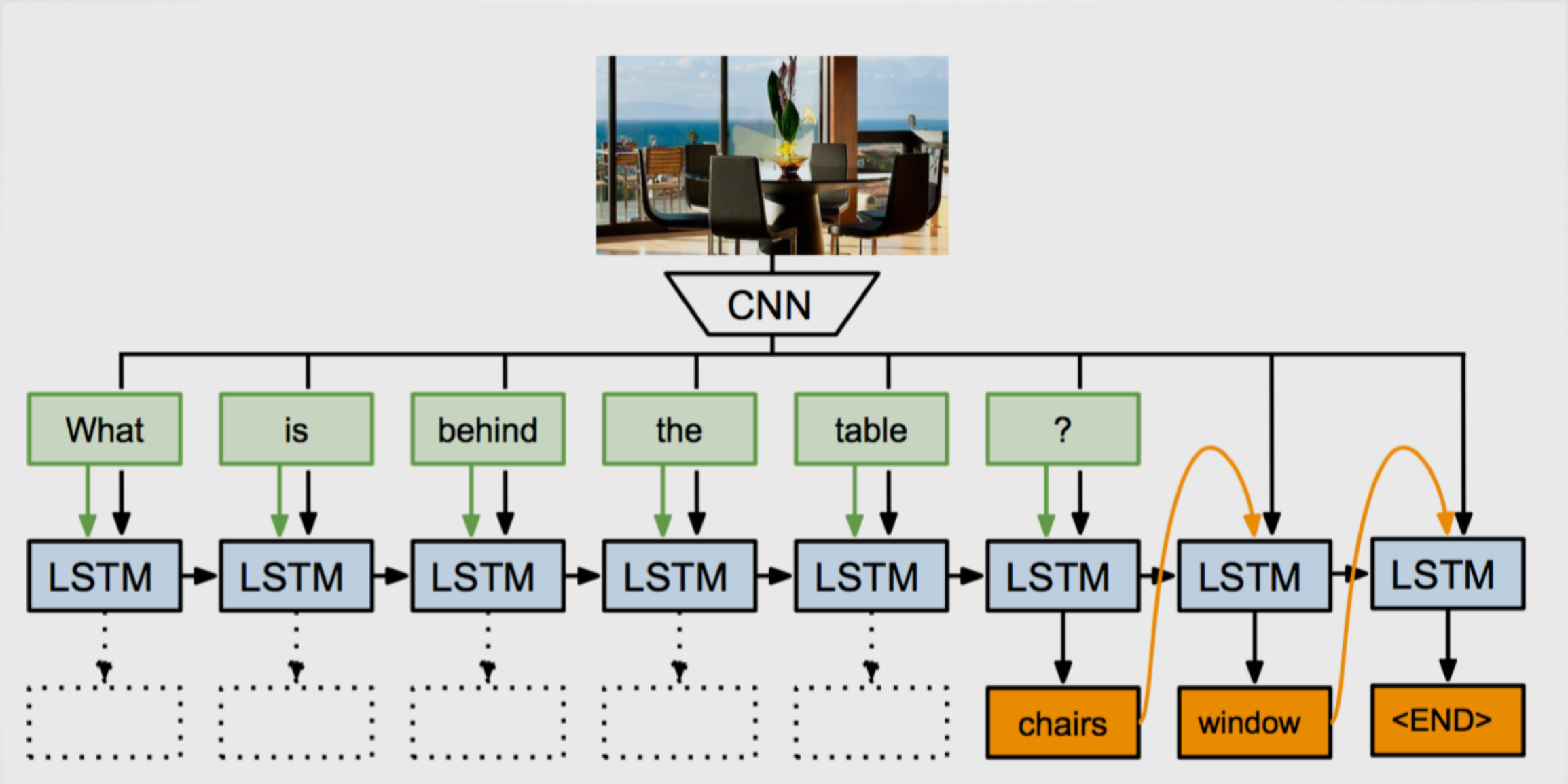


Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

PROBLEM FORMULATION

$$\hat{\boldsymbol{a}}_t = \underset{\boldsymbol{a} \in \mathcal{V}}{\arg\max}\, p(\boldsymbol{a} | \boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta})$$

- **x** is the image; **q** the question; θ the parameters to learn.
- $\hat{A}_{t-1}$ is the set of previous answer words, where $ token indicates end of answer sequence and ? end of question.
- $\hat{a}_t$ is the current answer word; ν the vocabulary.

NEURAL-IMAGE-QA ARCHITECTURE

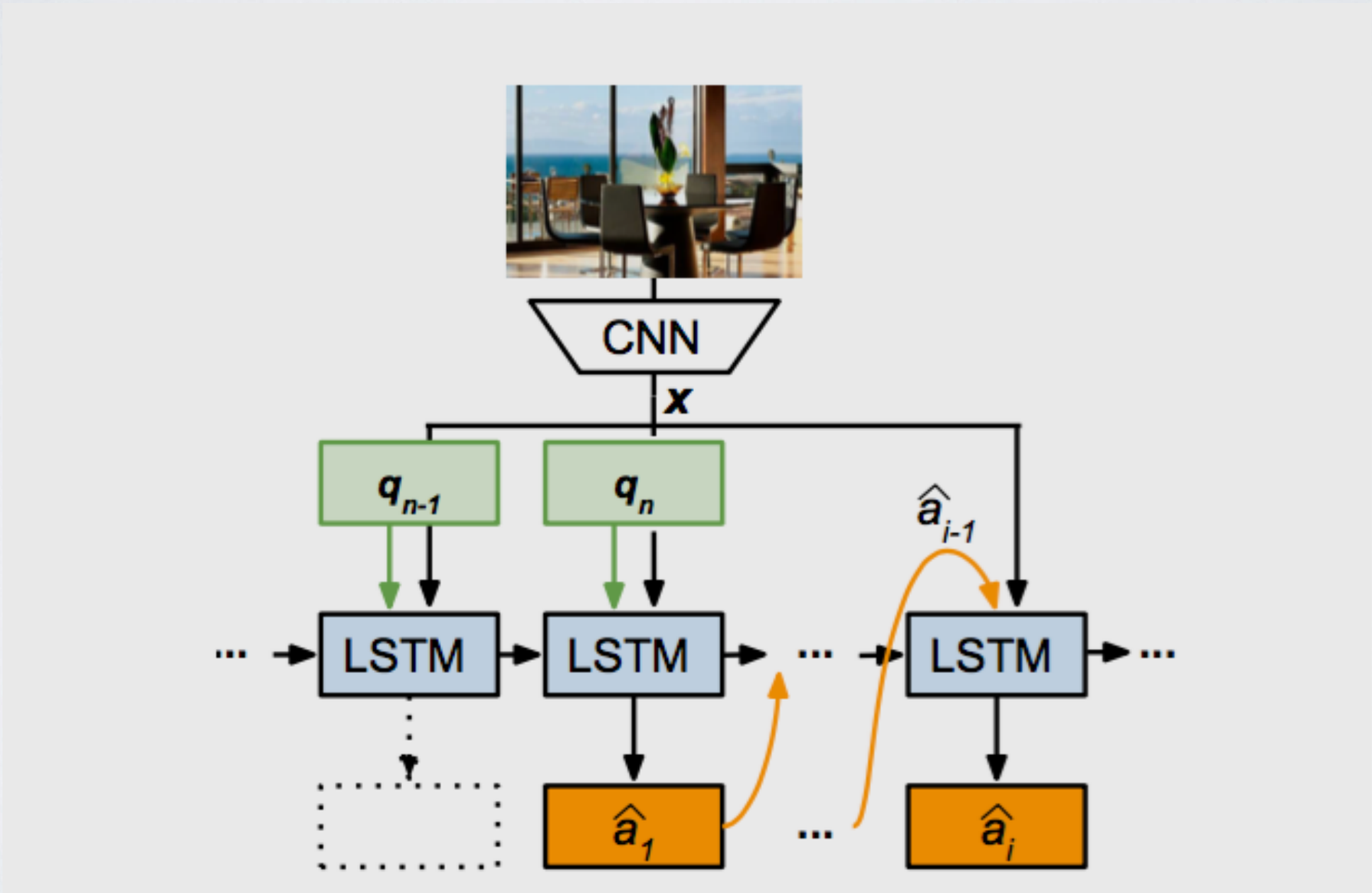


Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

NEURAL-IMAGE-QA ARCHITECTURE

- Caffe implementation of LSTM - sigmoid and hyperbolic tangent nonlinearities.
- Pre-train CNN on ImageNet with GoogleNet architecture.
- Randomly initialize and train last layer with LSTM - crucial step.
- Use default hyper-parameters for LSTM and CNN from previous work.

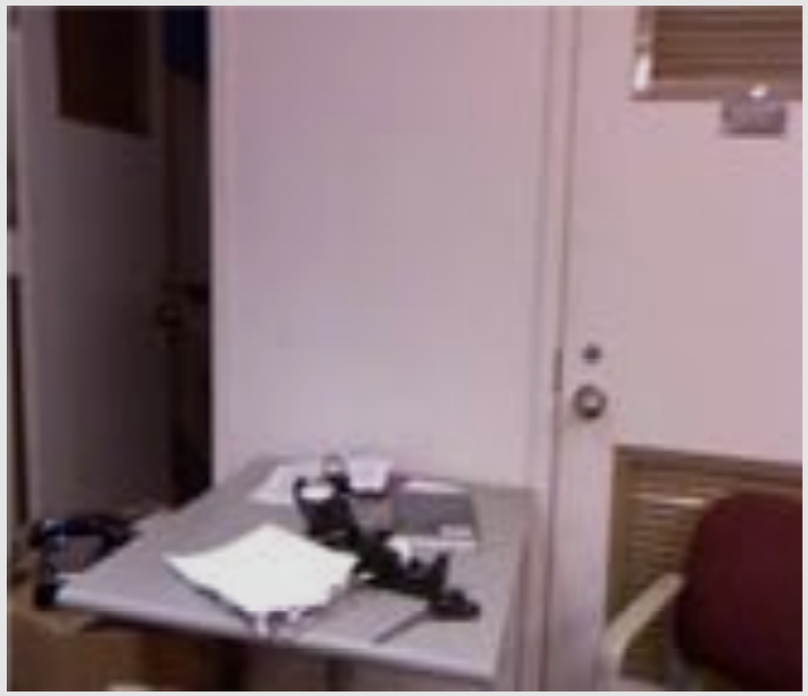
DAQUAR DATA

- 12,468 human question answer pairs on indoor scene images.
- 90% have single word answers. Longest answer: 7 entities.
- 2,483 different questions. Average number of words per question: 11.53 [7-31]
- Most frequent entities: table, chair, and lamp

DAQUAR DATA

What is behind the table?
sofa

What is the object on the counter in the corner?
microwave



How many doors are open?
1

Image credit: http://datasets.d2.mpi-inf.mpg.de/mateusz15neural_qa/ask_your_neurons-slides-low_res.pdf

WUPS SCORE

$$\mathbf{WUPS}(A,T) = \frac{1}{N}\sum_{i=1}^{N} \min\{ \prod_{a \in A^i} \max_{t \in T^i} \mu(a,t), \prod_{t \in T^i} \max_{a \in A^i} \mu(a,t) \}$$

- **a** - answer words; **t** - target words

- **u** - Wu-Palmer Similarity: calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer)

Image credit: http://datasets.d2.mpi-inf.mpg.de/mateusz15neural_qa/ask_your_neurons-slides-low_res.pdf

WUPS SCORE

Ground Truth	Predictions		
Armchair 	Wardrobe 	Chair 	
Accuracy	0	=	0
Wu-Palmer Similarity [1]	0.8	<	0.9
WUPS @0.9 (NIPS'14)	≈ 0	<<	0.9

- Generalization of accuracy that allows for word-level ambiguities. Mistakes with similar words should be less penalized. Smaller threshold more forgiving.

EXPERIMENTS

- Train three types of models: (1) Neural-Image-QA, (2) Only language features, (3) Only use data with single word answers.

- "Human answer, no image": asked participants to answer DAQUAR questions with no images.

RESULTS - ALL CLASSES

	Accu-racy	WUPS @0.9	WUPS @0.0
Malinowski et al. [20]	7.86	11.86	38.79
Neural-Image-QA (ours)			
- multiple words	17.49	23.28	57.76
- single word	**19.43**	**25.28**	**62.00**
Human answers [20]	50.20	50.82	67.27
Language only (ours)			
- multiple words	17.06	22.30	56.53
- single word	17.15	22.80	58.42
Human answers, no images	7.34	13.17	35.56

Table 1. Results on DAQUAR, all classes, single reference, in %.

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

RESULTS - ALL CLASSES

	Accu-racy	WUPS @0.9	WUPS @0.0
Malinowski et al. [20]	7.86	11.86	38.79
Neural-Image-QA (ours)			
- multiple words	17.49	23.28	57.76
- single word	**19.43**	**25.28**	**62.00**
Human answers [20]	50.20	50.82	67.27
Language only (ours)			
- multiple words	17.06	22.30	56.53
- single word	17.15	22.80	58.42
Human answers, no images	7.34	13.17	35.56

Table 1. Results on DAQUAR, all classes, single reference, in %.

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

RESULTS - ALL CLASSES

	Accu-racy	WUPS @0.9	WUPS @0.0
Malinowski et al. [20]	7.86	11.86	38.79
Neural-Image-QA (ours)			
- multiple words	17.49	23.28	57.76
- single word	**19.43**	**25.28**	**62.00**
Human answers [20]	50.20	50.82	67.27
Language only (ours)			
- multiple words	17.06	22.30	56.53
- single word	17.15	22.80	58.42
Human answers, no images	7.34	13.17	35.56

Table 1. Results on DAQUAR, all classes, single reference, in %.

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

RESULTS - SINGLE WORD

	Accu-racy	WUPS @0.9	WUPS @0.0
Neural-Image-QA (ours)	21.67	27.99	65.11
Language only (ours)	19.13	25.16	61.51

Table 2. Results of the single word model on the one-word answers subset of DAQUAR, all classes, single reference, in %.

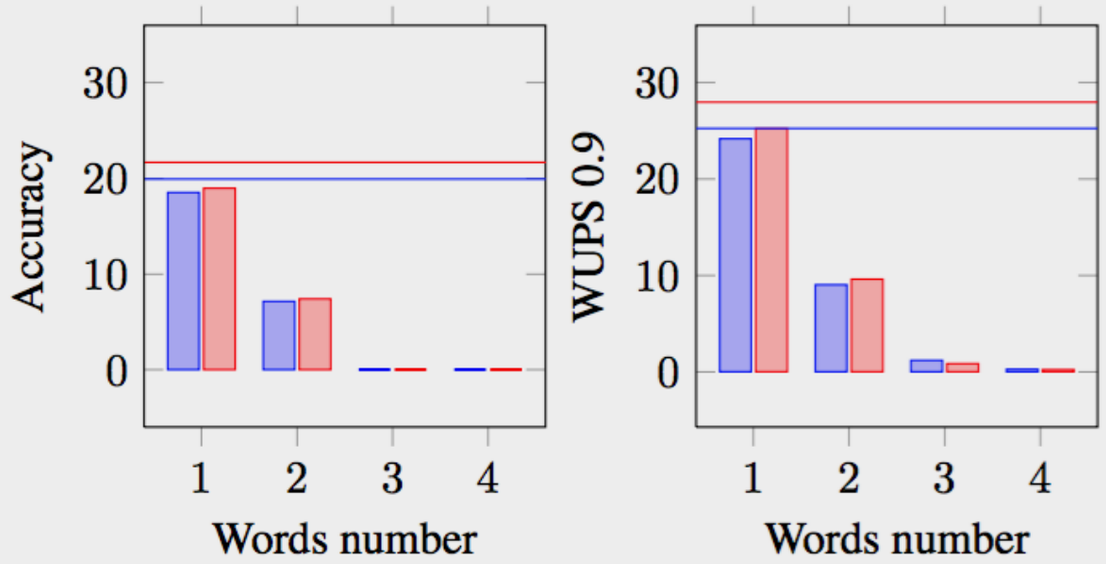


Figure 4. Language only (blue bar) and Neural-Image-QA (red bar) "multi word" models evaluated on different subsets of DAQUAR. We consider 1, 2, 3, 4 word subsets. The blue and red horizontal lines represent "single word" variants evaluated on the answers with exactly 1 word.

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

DAQUAR-CONSENSUS

- Ask multiple people to answer same question. Data has an average of 5 responses per image.

- Average Consensus Score - average WUPS score - in limit measures inter human agreement for question. Encourages predictions of most agreeable answers.

- Min Consensus Score - Replaces average with max to use human answer closest to predicted. Most optimistic evaluation.

CONSENSUS RESULTS

	Accuracy	WUPS @0.9	WUPS @0.0
WUPS [20]	50.20	50.82	67.27
ACM (ours)	36.78	45.68	64.10
MCM (ours)	60.50	69.65	82.40

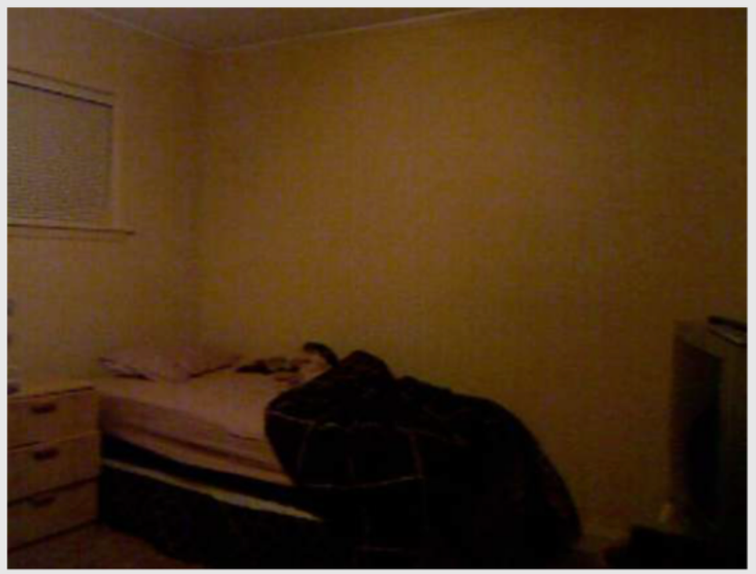
Table 6. Min and Average Consensus on human answers from DAQUAR, as reference sentence we use all answers in DAQUAR-Consensus which are not in DAQUAR, in %

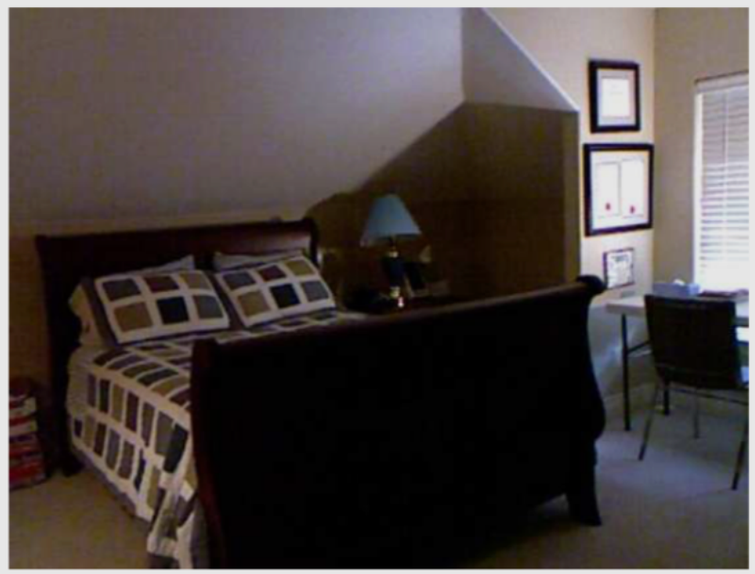
	Accu-racy	WUPS @0.9	WUPS @0.0
Average Consensus Metric			
Language only (ours)			
- multiple words	11.60	18.24	52.68
- single word	11.57	18.97	54.39
Neural-Image-QA (ours)			
- multiple words	11.31	18.62	53.21
- single word	**13.51**	**21.36**	**58.03**
Min Consensus Metric			
Language only (ours)			
- multiple words	22.14	29.43	66.88
- single word	22.56	30.93	69.82
Neural-Image-QA (ours)			
- multiple words	22.74	30.54	68.17
- single word	**26.53**	**34.87**	**74.51**

Table 5. Results on DAQUAR-Consensus, all classes, consensus in %.

Image credit: Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. Malinowski, Rohrbach, Fritz. ICCV 2015.

SINGLE ANSWER EXAMPLES





What is on the right side of the cabinet?		How many drawers are there?	What is the largest object?
Neural-Image-QA:	bed	3	bed
Language only:	bed	6	table

Table 7. Examples of questions and answers. Correct predictions are colored in green, incorrect in red.

MULTIPLE ANSWER EXAMPLES

What is on the refrigerator?

What is the colour of the comforter?

What objects are found on the bed?

Neural-Image-QA:	magnet, paper	blue, white	bed sheets, pillow
Language only:	magnet, paper	blue, green, red, yellow	doll, pillow

Table 8. Examples of questions and answers with multiple words. Correct predictions are colored in green, incorrect in red.

FAILURE EXAMPLES

	How many chairs are there?	What is the object fixed on the window?	Which item is red in colour?
Neural-Image-QA:	1	curtain	remote control
Language only:	4	curtain	clock
Ground truth answers:	2	handle	toaster

Table 9. Examples of questions and answers - failure cases.

CONCLUSION

- Errors with smaller objects, indoor scene statistics, spatial reasoning, and too few data.

- Language only model only slightly worse and outperforms human baseline.

- End to end architecture.

- Possible extensions: (1) Larger datasets, (2) Explore better pre-training techniques to better leverage images.

DISCUSSION

- Is only retraining the last layer sufficient for capturing global relationships?
- Why does two layer LSTM perform worse?
- Are there better ways to evaluate how much common sense knowledge is encoded in a QA system (other than a language only model)?
- With extra training data, maybe fine-tuning the CNN more (or even training one from scratch) could help since the network they used was designed for image recognition and not necessarily question answering.
- How well does this method generalize?