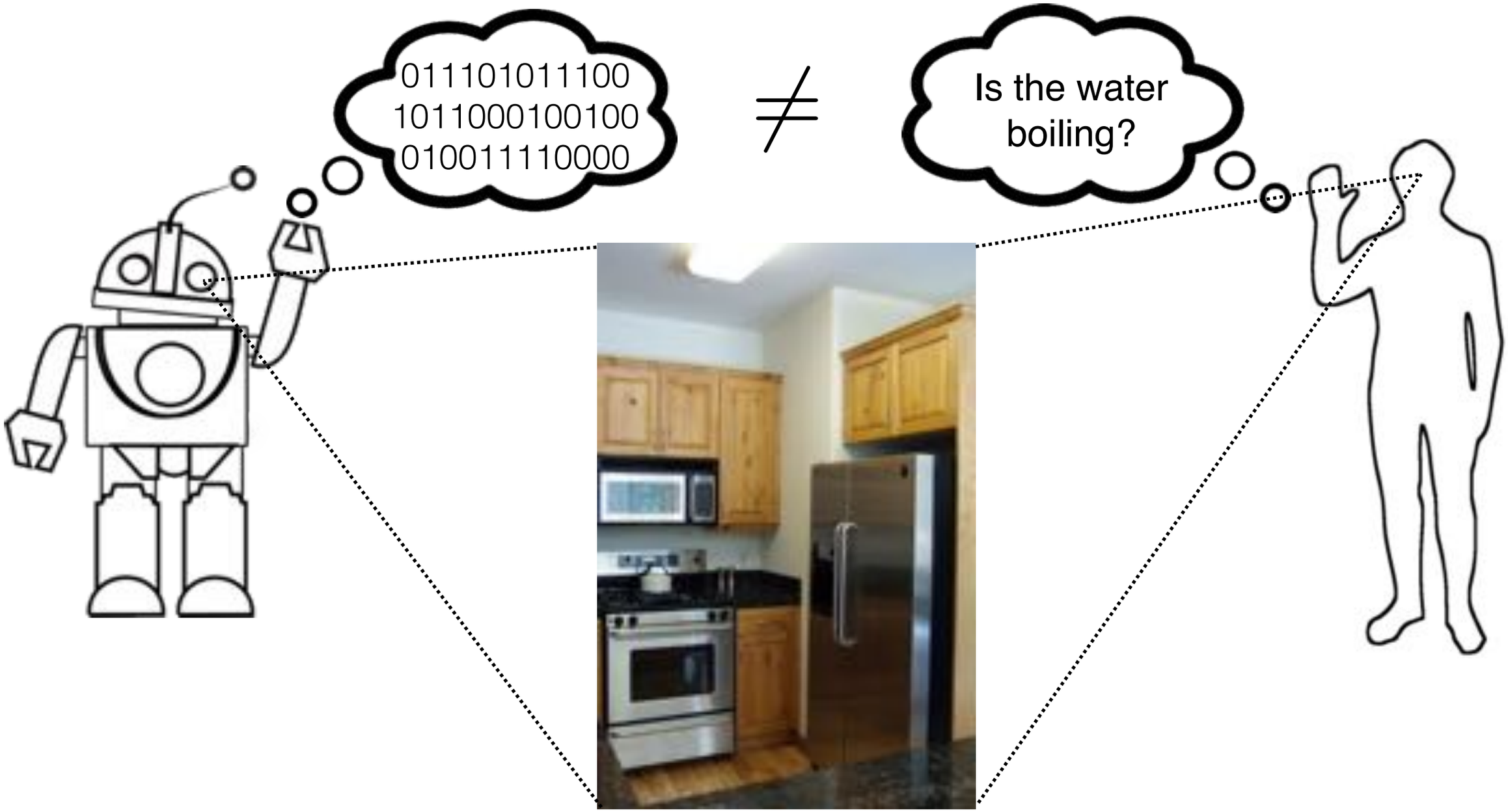
Ask Your Neurons: A Neural-based Approach to Answering Questions about Images

Mateusz Malinowski [1]
Marcus Rohrbach [2]
Mario Fritz [1]

[1] Max Planck Institute for Informatics
[2] Berkeley University of California, ICSI

Human-like Comprehension



- How far are machines from human quality understanding?
- How can we monitor progress and evaluate architectures?

Visual Turing Test (NIPS'14)

- Holistic, open-ended task
  - Visual scene understanding
  - Natural language understanding
  - Deduction
- No internal representation is evaluated
  - Challenge is open to diverse approaches
- Scalable annotation end evaluation effort
  - Only question-answer pairs

What is behind the table?
sofa

What is on the refrigerator?
magnet, paper

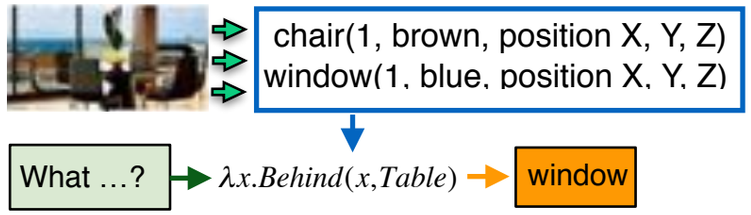
What color are the cabinets?
brown

How many lamps are there?
2

Related Work

- ## Symbolic-based Approaches

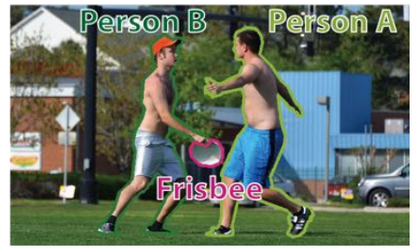
  M. Malinowski et. al. Multiworld. NIPS'14



- ## Large Scale Datasets

  S. Antol et. al. Visual QA. ICCV'15
  L. Yu et. al. al. Visual Madlibs. ICCV'15
  D. Geman et. al. Visual Turing Test. PNAS'15
  M. Ren et. al. Image QA. NIPS15
  H. Gao et. al. Are You Talking to a Machine? NIPS'15
  Y. Zhu et. al. Visual7W. arXiv'15
  L. Zhu et. al. Uncovering Temporal Context. arXiv'15

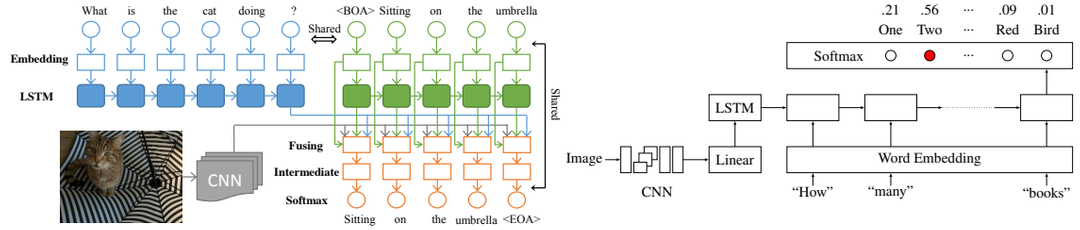
What is the mustache made of?



Person A is …

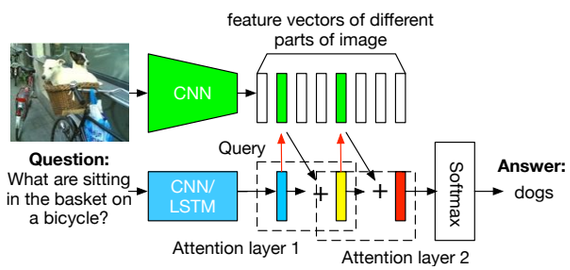
- ## Neural-based Approaches

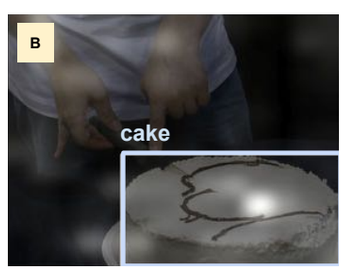
  M. Ren et. al. Image QA. NIPS'15
  H. Gao. et. al. Are You Talking to a Machine? NIPS'15
  L. Ma et. al. Learning to Answer Questions From Images. arXiv'15

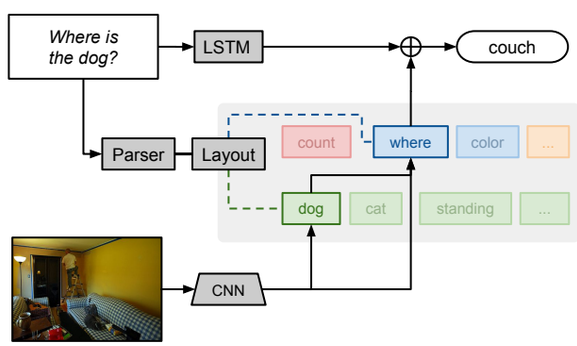


- ## Attention-based Approaches

  Z. Yang. et. al. Stacked Attention Networks. arXiv'15
  Y. Zhu et. al. Visual7W. arXiv'15
  J. Andres et. al. Deep Compositional QA. arXiv'15
  H. Xu et. al. Ask, Attend and Answer. arXiv'15
  K. Chen et. al. ABC-CNN. arXiv'15
  K. J. Shih et. al. Where To Look. arXiv'15





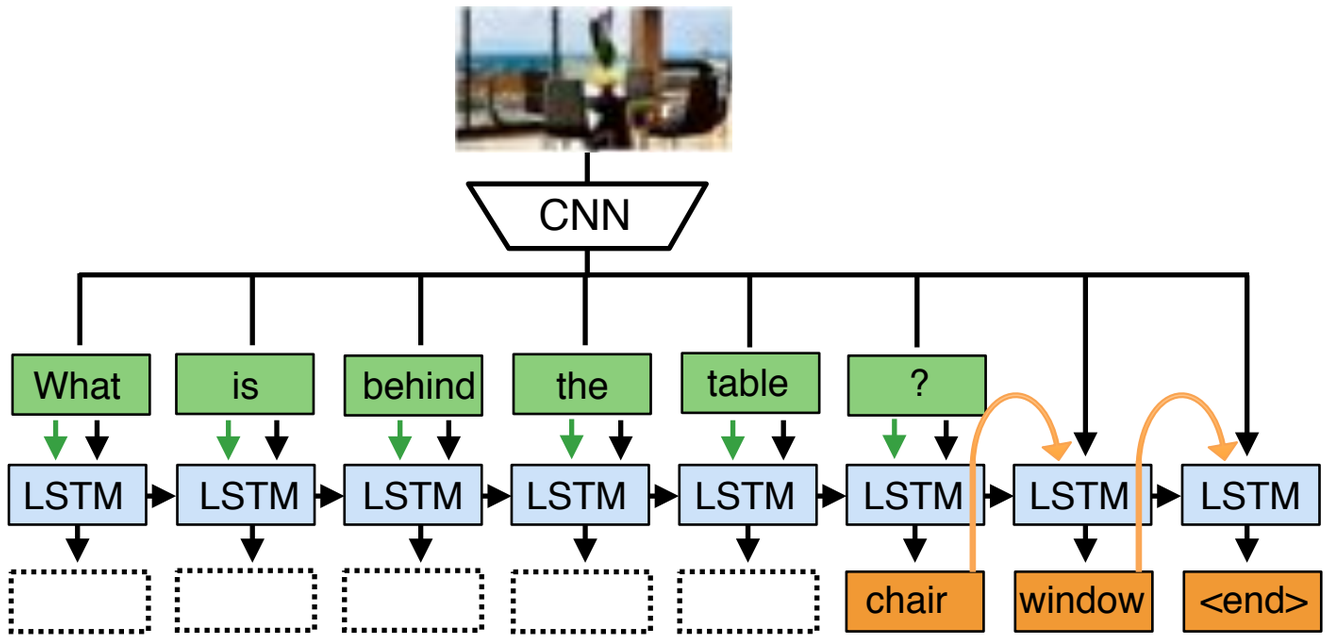


- ## Hybrid Approaches

  H. Noh et al. Dynamic Parameter Prediction. arXiv'15
  J. Andres et al. Deep Compositional QA. arXiv'15

Outline

- Neural approach to answer questions about images

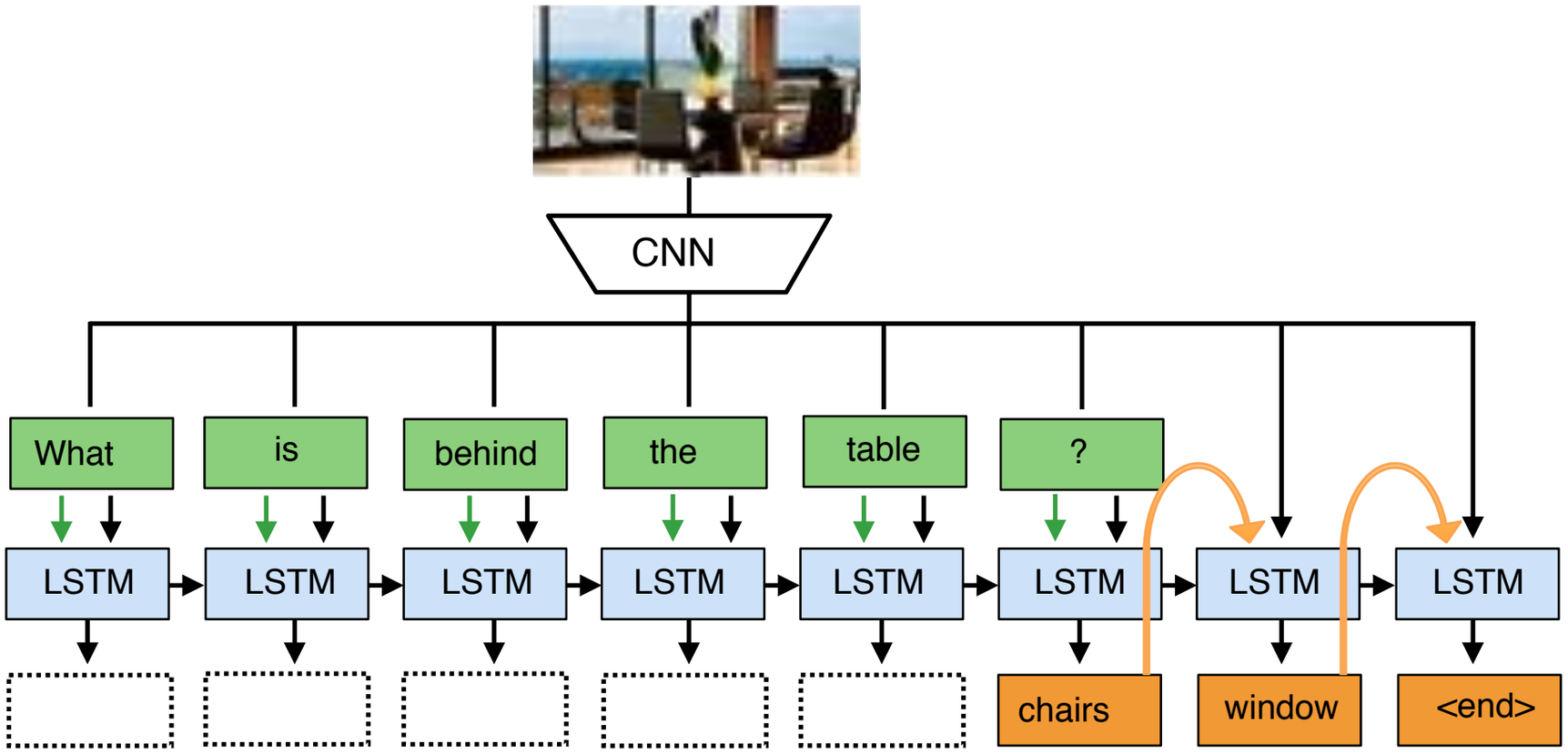


- Performance metrics based on additional annotations

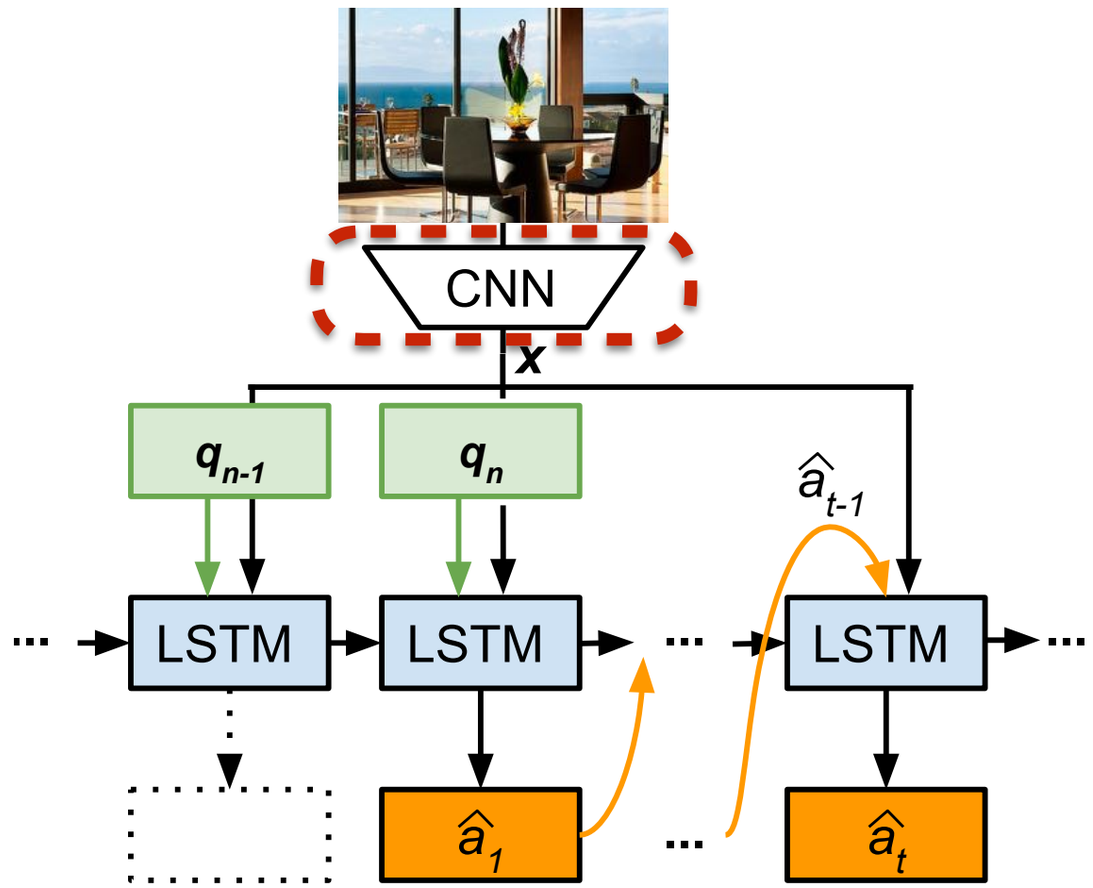
What is the object on the floor in front of the wall?

Human 1: **bed**
Human 2: **shelf**
Human 3: **bed**
Human 4: **bookshelf**

Method: Ask Your Neurons



Method: Ask Your Neurons



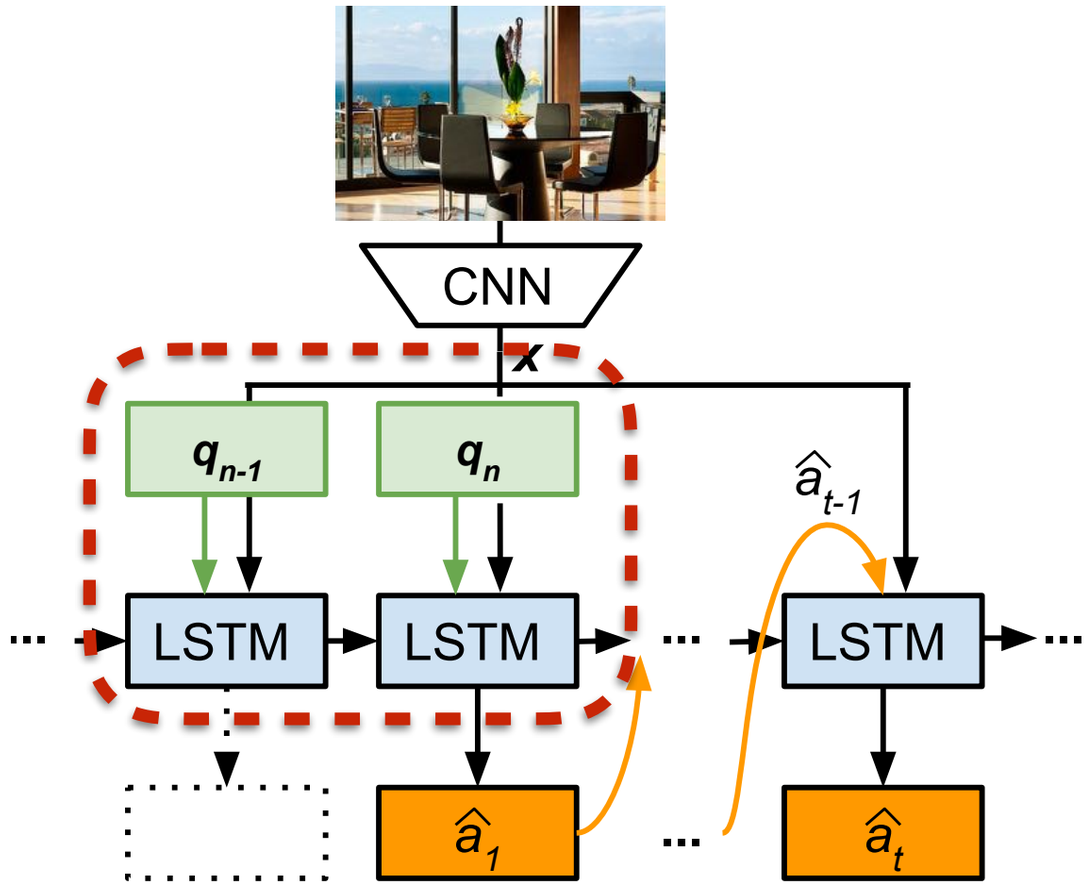
- Predicting answer sequence
  - Recursive formulation

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

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Method: Ask Your Neurons



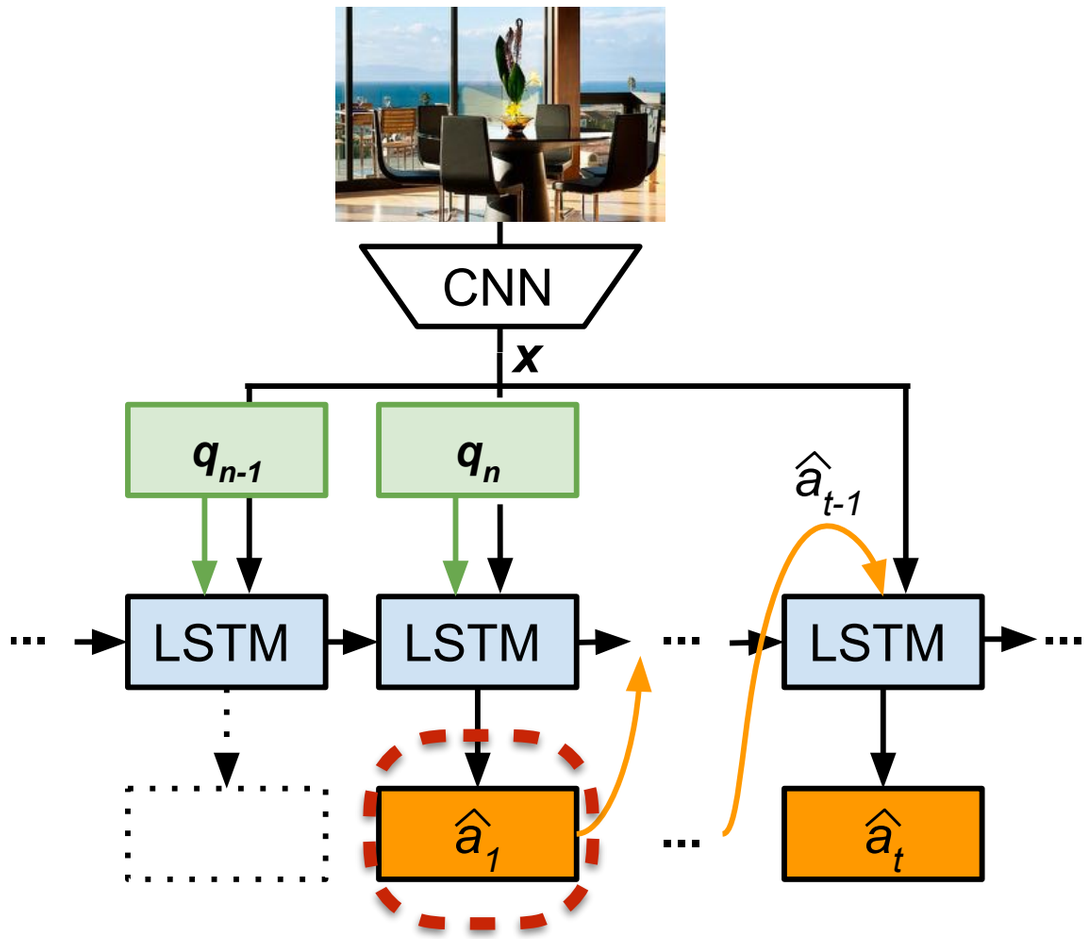
- Predicting answer sequence
  - Recursive formulation

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

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Method: Ask Your Neurons



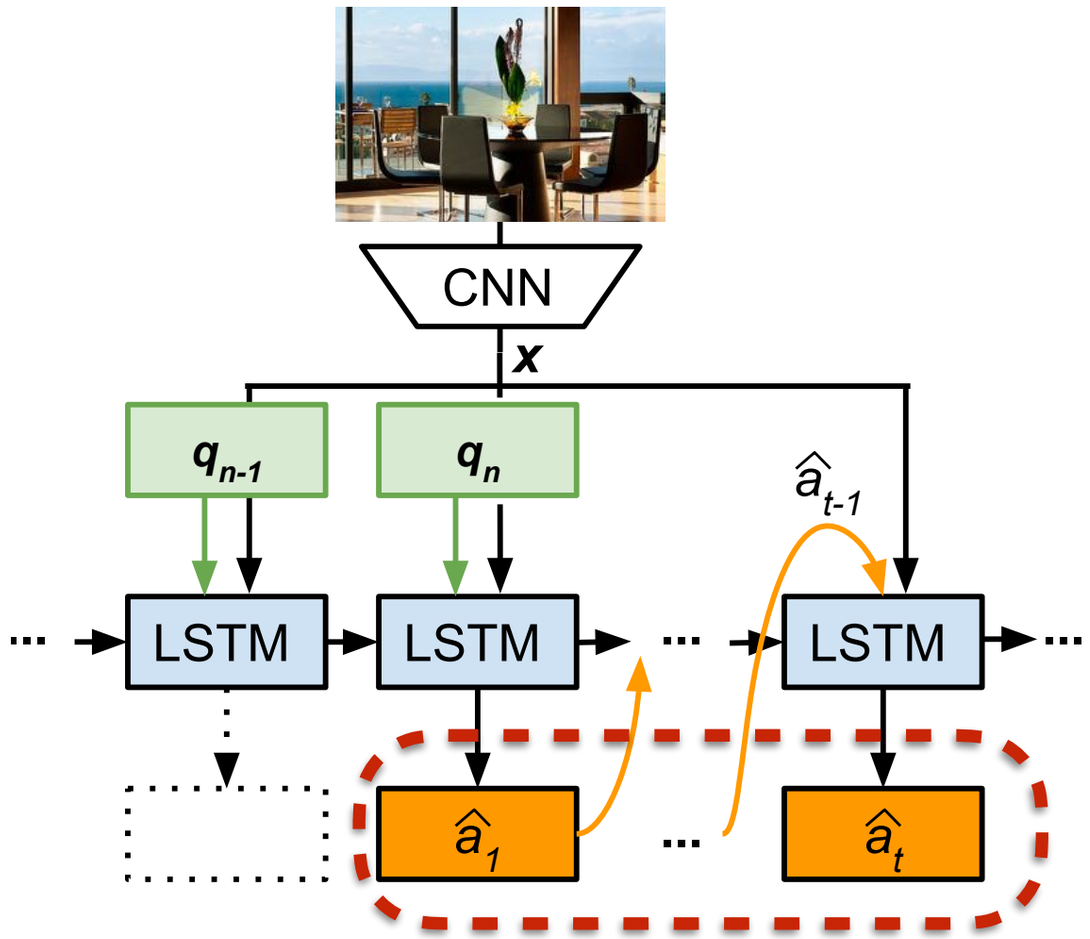
- Predicting answer sequence
  - Recursive formulation

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

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Method: Ask Your Neurons



- Predicting answer sequence
  - Recursive formulation

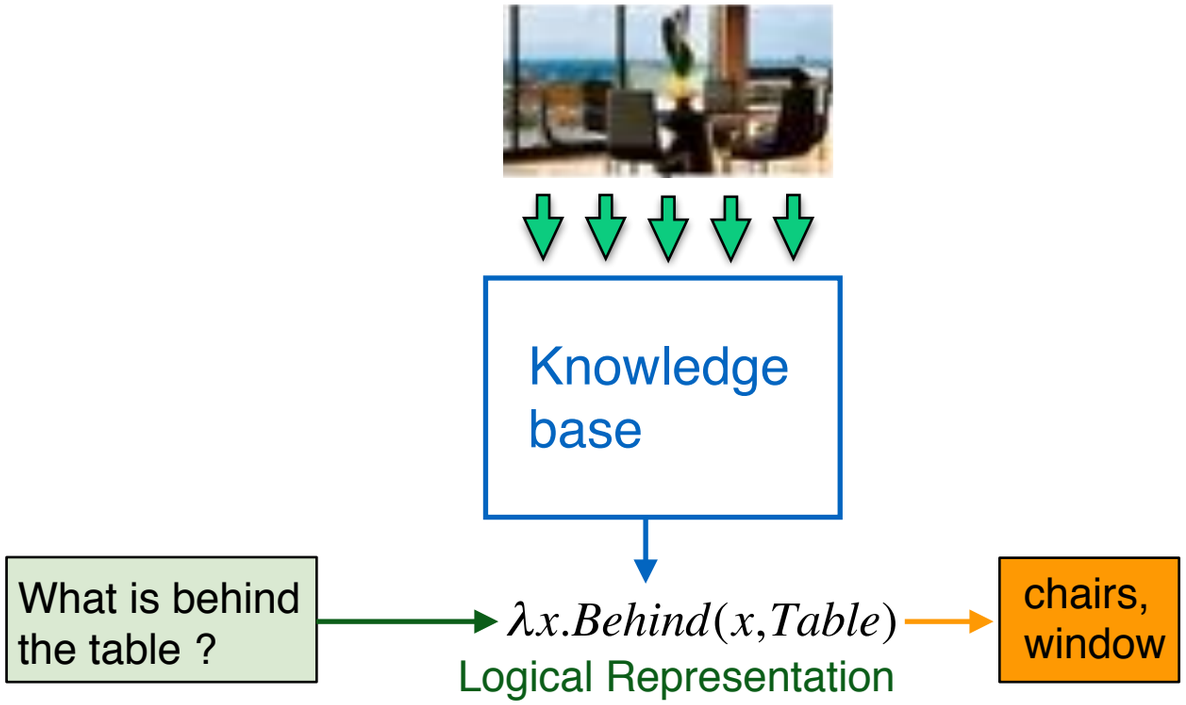
$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

$$\boldsymbol{q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Symbolic vs Neural-based Approaches

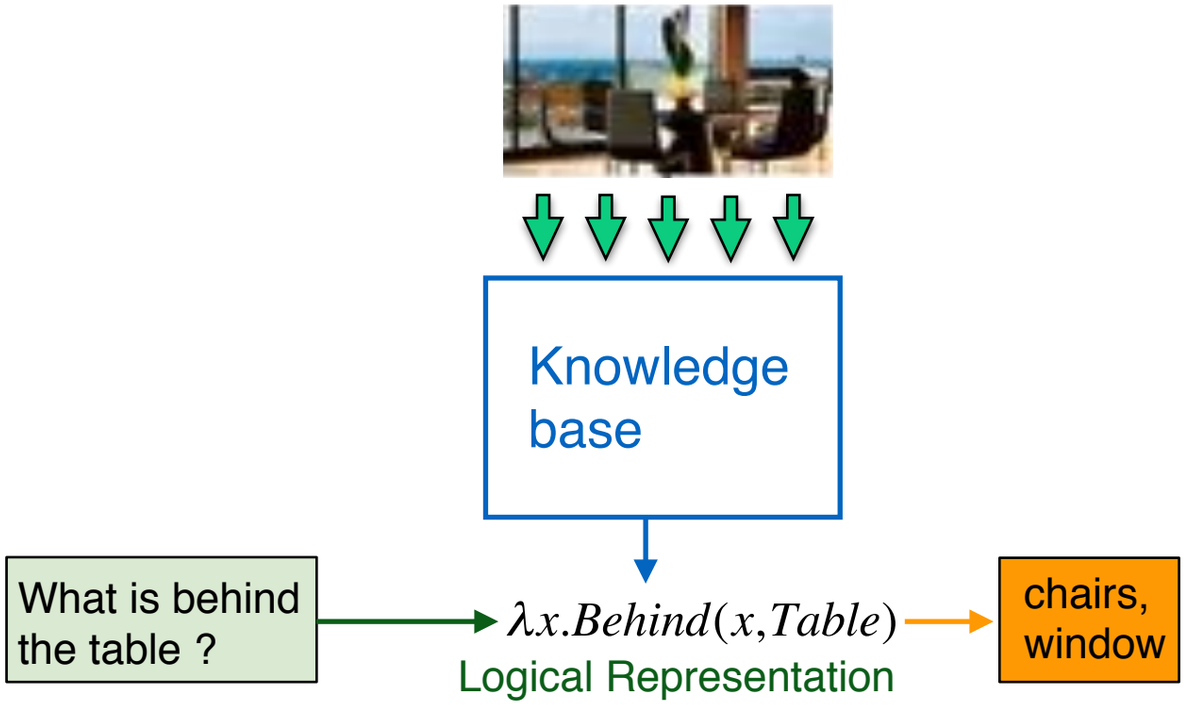
- Symbolic approach (NIPS'14)
  - Explicit representation
  - Independent components
    - Detectors, Semantic Parser, Database
  - Components trained separately
  - Many 'hard' design decisions



M. Malinowski, et. al. "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". NIPS'14

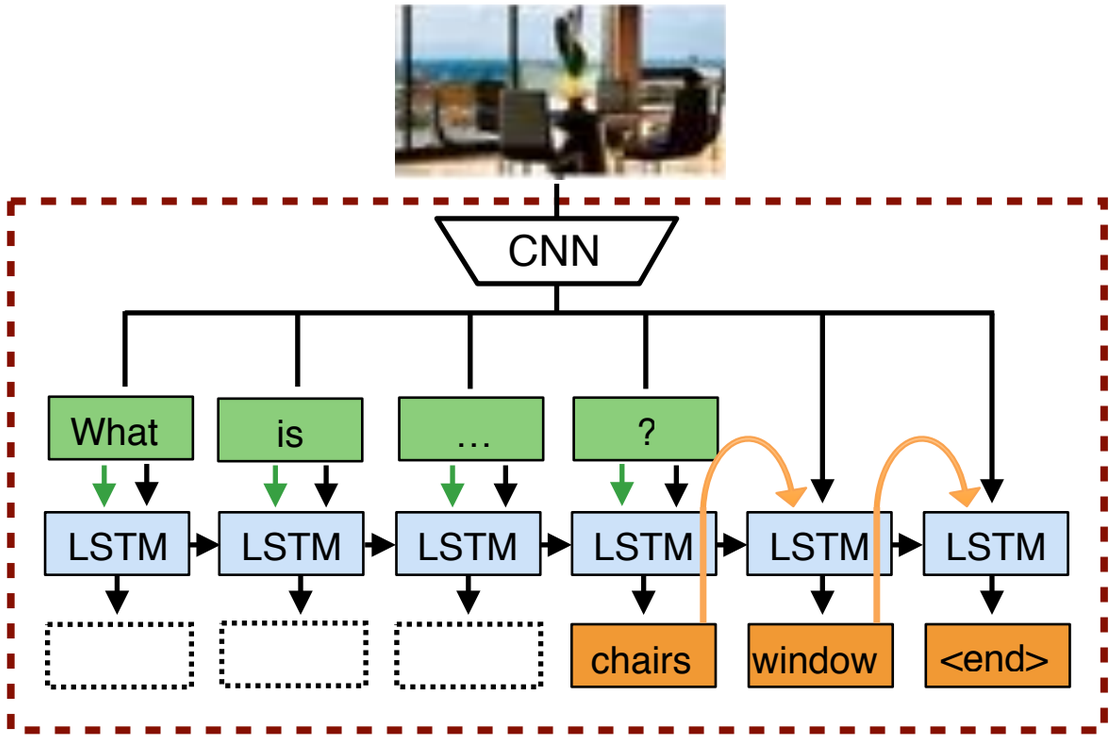
Symbolic vs Neural-based Approaches

- Symbolic approach (NIPS'14)
  - Explicit representation
  - Independent components
    - Detectors, Semantic Parser, Database
  - Components trained separately
  - Many 'hard' design decisions



M. Malinowski, et. al. "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". NIPS'14

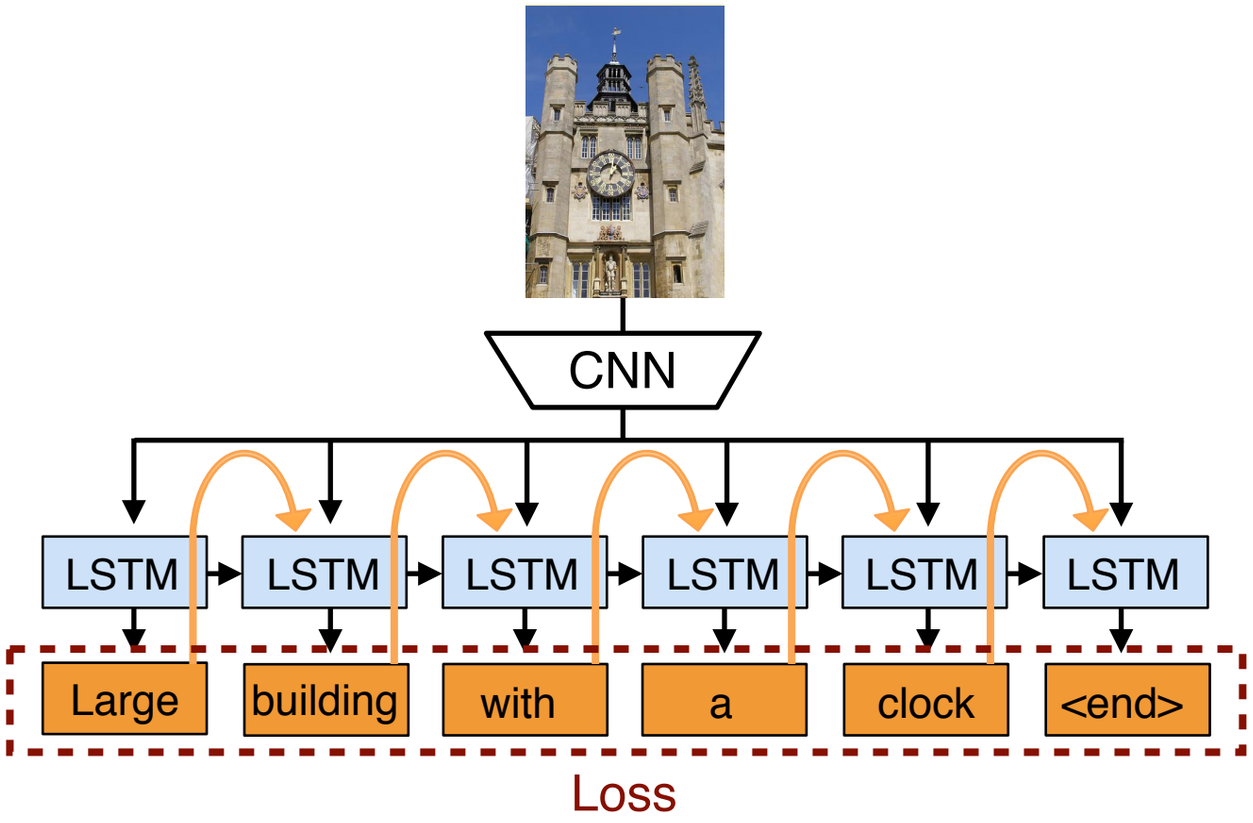
- **Ask Your Neurons (Our)**
  - Implicit representation
  - End-to-end formula
    - From images and questions to answers
  - Joint training
  - Fewer design decisions



End-to-end, jointly trained architecture

Neural Visual QA vs Neural Image Description

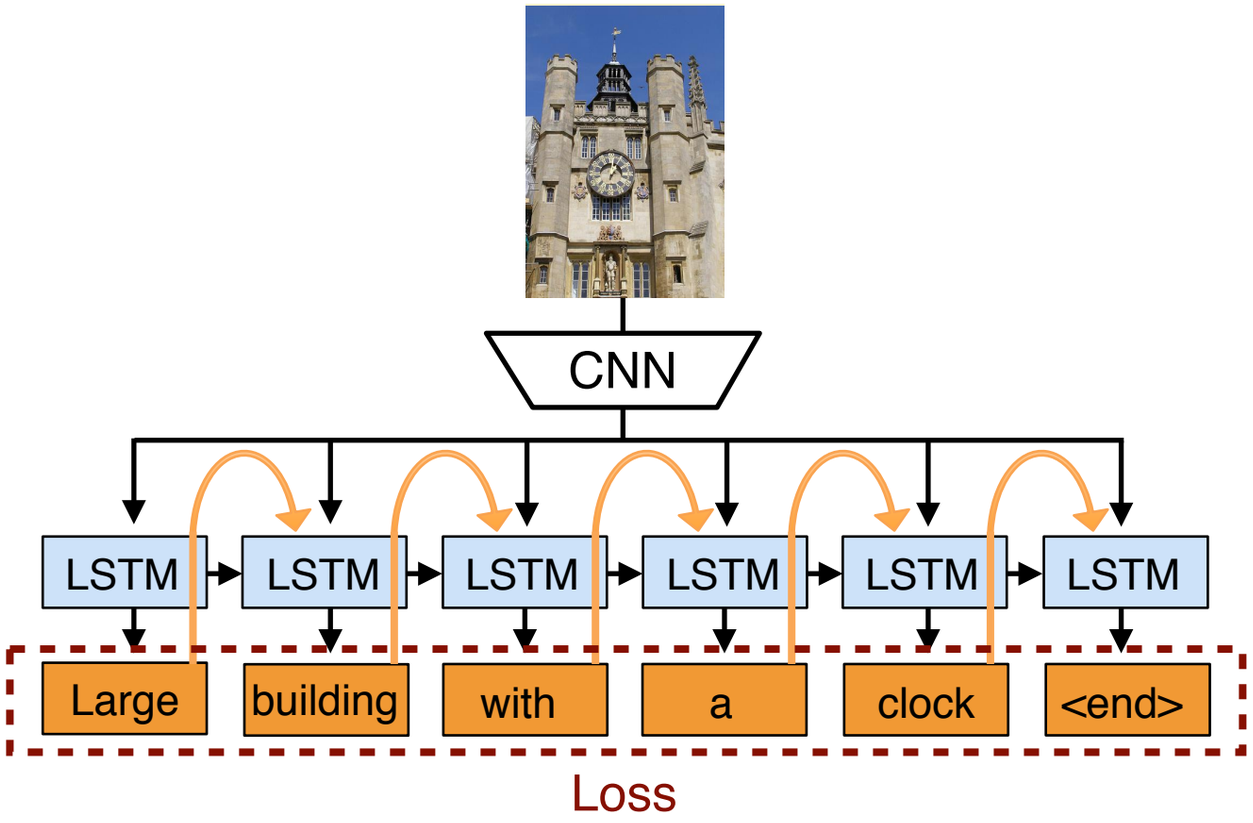
- Neural Image Description
  - Conditions on an image
  - Generates a description
    - Sequence of words
  - Loss at every step



J. Donahue, et. al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”. CVPR15

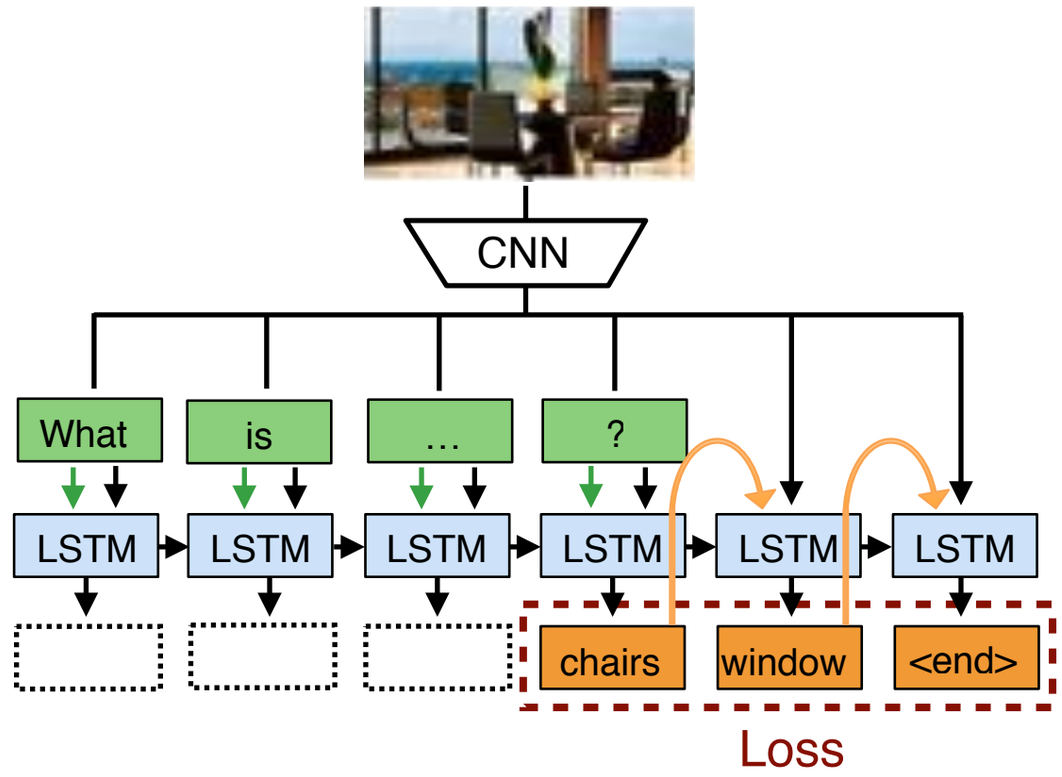
Neural Visual QA vs Neural Image Description

- Neural Image Description
  - Conditions on an image
  - Generates a description
    - Sequence of words
  - Loss at every step



J. Donahue, et. al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”. CVPR15

- **Ask Your Neurons (Our)**
  - Conditions on an image and a question
  - Generates an answer
    - Sequence of answer words
  - Loss only at answer words



Visual Turing Test: DAQUAR (NIPS'14)

What is behind the table?
sofa

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

How many doors are open?
1

- Dataset for Question Answering on Real-world images
- 1449 RGBD indoor images (NYU-Depth V2 dataset)
- 12.5k question-answer pairs about colors, numbers, objects
- Human-type subjectivity is common in the dataset

Evaluation: WUPS (NIPS'14)

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

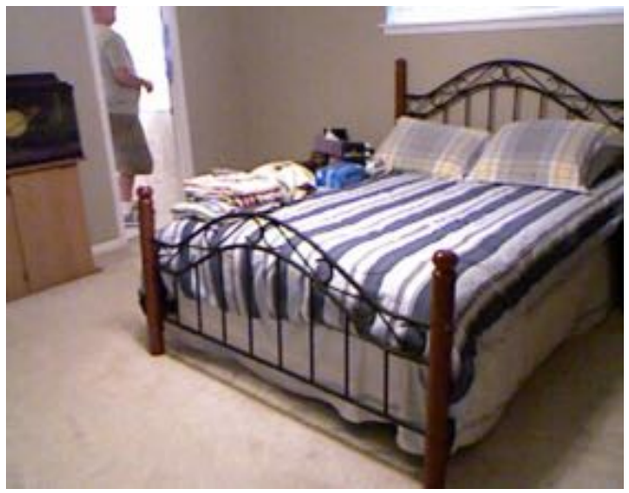
[1] Wu, Z., Palmer, M.: Verbs semantics and lexical selection. ACL. 1994.

Results on Full DAQUAR

Methods	Accuracy	WUPS @0.9
Baseline: Symbolic (NIPS'14)	7.86%	11.86%
Language Only (Our)	17.15%	22.80%
Vision + Language (Our)	**19.43%**	**25.28%**
Human performance (NIPS'14)	50.20%	50.82%

What is on the refrigerator?

magnet, paper

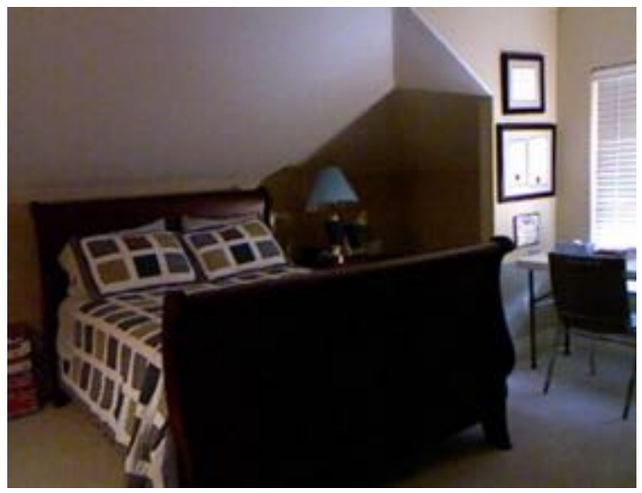


What is the color of the comforter?

blue, white

How many drawers are there?

3



What is the largest object?

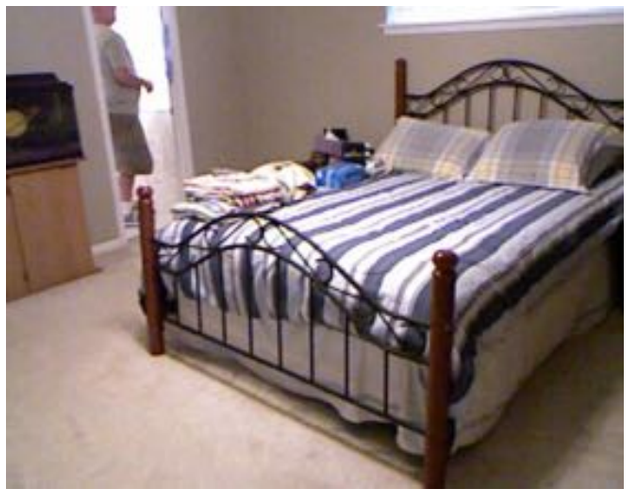
bed

Results on Full DAQUAR

Methods	Accuracy	WUPS @0.9
Baseline: Symbolic (NIPS'14)	7.86%	11.86%
Language Only (Our)	17.15%	22.80%
Vision + Language (Our)	**19.43%**	**25.28%**
Human performance (NIPS'14)	50.20%	50.82%

What is on the refrigerator?

magnet, paper

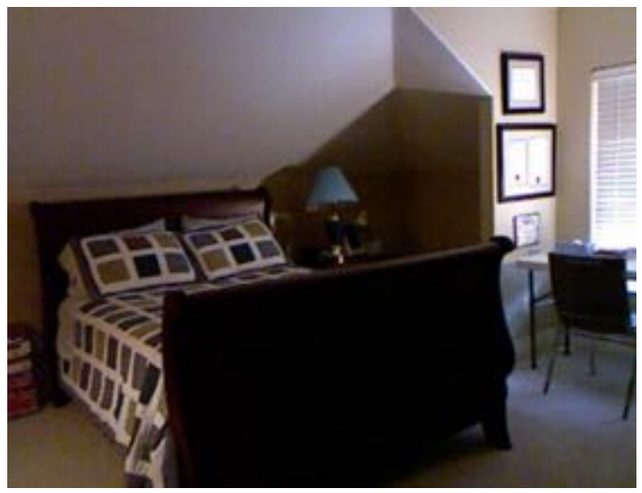


What is the color of the comforter?

blue, white

How many drawers are there?

3



What is the largest object?

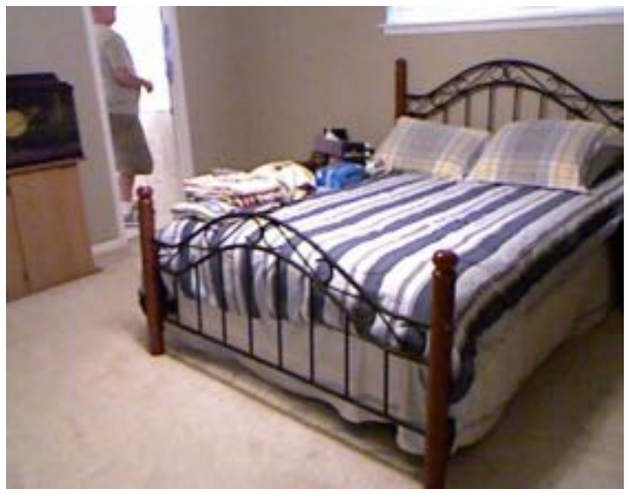
bed

Results on Full DAQUAR

Methods	Accuracy	WUPS @0.9
Baseline: Symbolic (NIPS'14)	7.86%	11.86%
Language Only (Our)	17.15%	22.80%
Vision + Language (Our)	**19.43%**	**25.28%**
Human performance (NIPS'14)	50.20%	50.82%

What is on the refrigerator?

magnet, paper

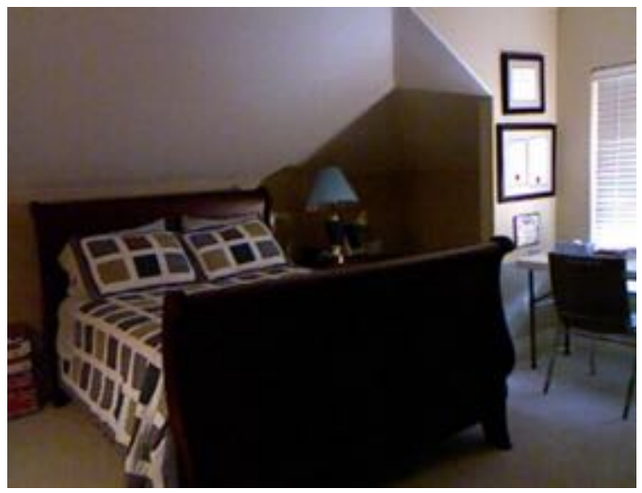


What is the color of the comforter?

blue, white

How many drawers are there?

3



What is the largest object?

bed

Qualitative Results

What is on the right side of the cabinet?

Vision + Language: **bed**

Language Only: **bed**

What objects are found on the bed?

Vision + Language: **bed sheets, pillow**

Language Only: **doll, pillow**

How many burner knobs are there?

Vision + Language: **4**

Language Only: **6**

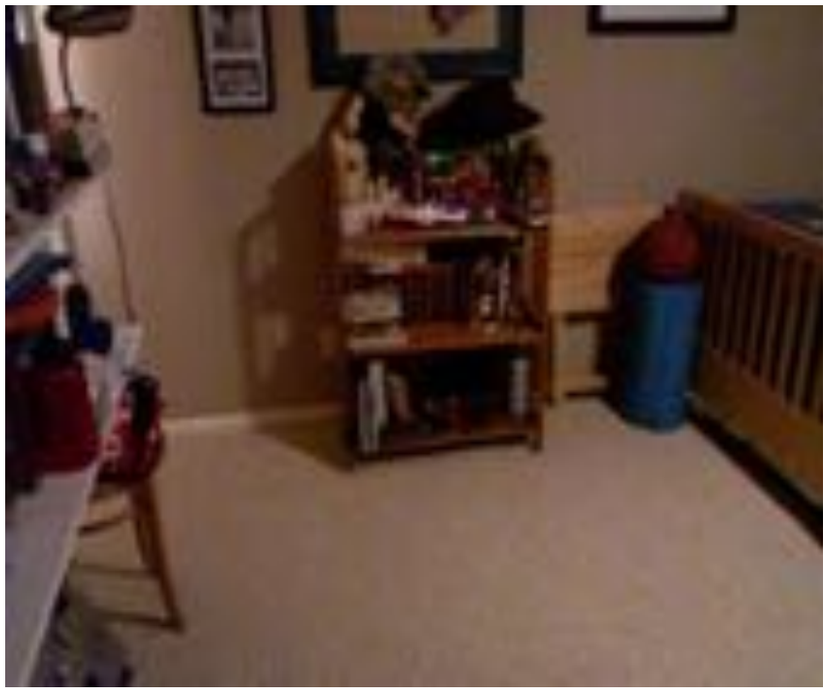
Qualitative Results: Failure Cases

How many chairs are there?

Vision + Language: **1**
Language Only: **4**
Human: **2**

How many glass cups are there?

Vision + Language: **2**
Language Only: **6**
Human: **4**

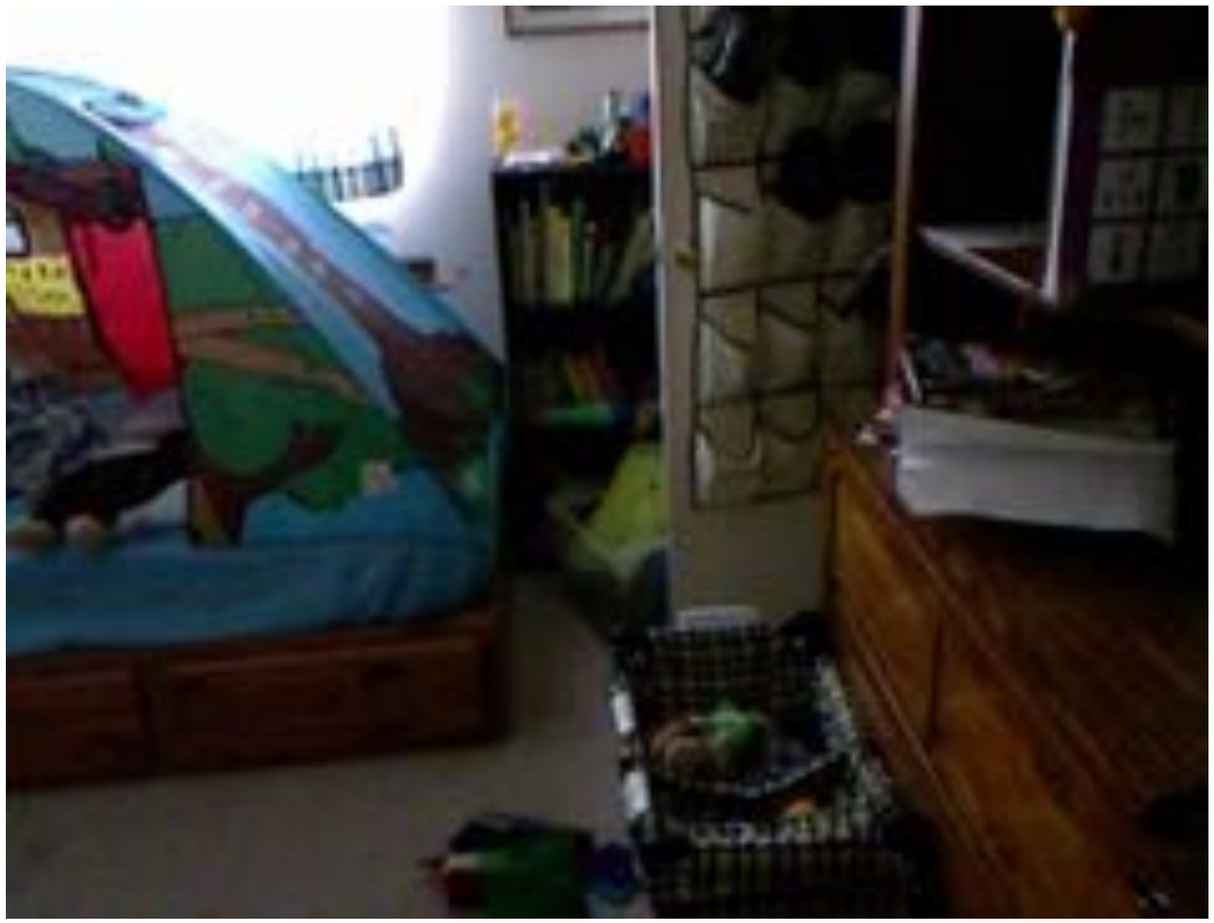


What is on the left side of the bed?

Vision + Language: **night stand**
Language Only: **night stand**
Human: **ball**

1. New Performance Metric: Min Consensus

- WUPS handle word-level ambiguities
- But how to embrace many possible interpretations of both a question and a scene?



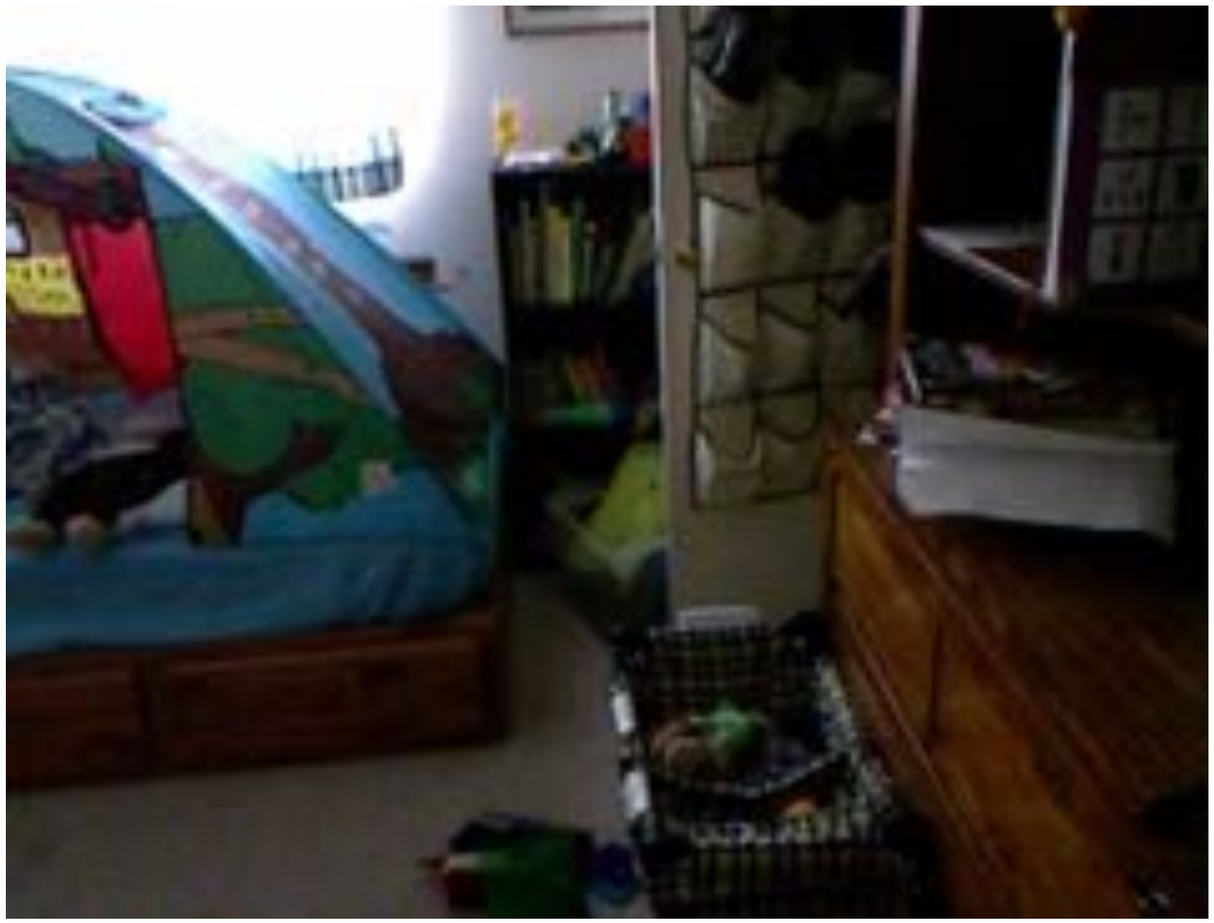
What is the object on the floor in front of the wall?

Human 1: **bed**
Human 2: **shelf**
Human 3: **bed**
Human 4: **bookshelf**

1. New Performance Metric: Min Consensus

- We extend WUPS scores by Min Consensus
  - Finding at least one human answer that matches with the predicted one
  - Treat all possible interpretations equal

$$\frac{1}{N}\sum_{i=1}^{N}\max_{k=1}^{K}\left(\min\{\prod_{a\in A^i}\max_{t\in T_k^i}\mu(a,t),\ \prod_{t\in T_k^i}\max_{a\in A^i}\mu(a,t)\}\right)$$



What is the object on the floor in front of the wall?

Human 1: **bed**
Human 2: **shelf**
Human 3: **bed**
Human 4: **bookshelf**

Results on DAQUAR-Consensus

Methods (Old Metric)	Accuracy	WUPS @0.9
Language Only (Our)	17.15%	22.8%
Vision + Language (Our)	**19.43%**	**25.28%**
Human performance (NIPS'14)	50.2%	50.82%

Methods (Min Consensus)	Accuracy	WUPS @0.9
Language Only (Our)	22.56%	30.93%
Vision + Language (Our)	**26.53%**	**34.87%**
Human performance (Our)	60.50%	69.65%

Results on DAQUAR-Consensus

What is in front of the curtain?

Model: chair

Human 1: guitar

Human 2: chair

What color are the beds?

Model: white

Human 1: white

Human 2: pink

How many steel chairs are there?

Model: 4

Human 1: 2

Human 2: 4

What is the largest object?

Model: bed

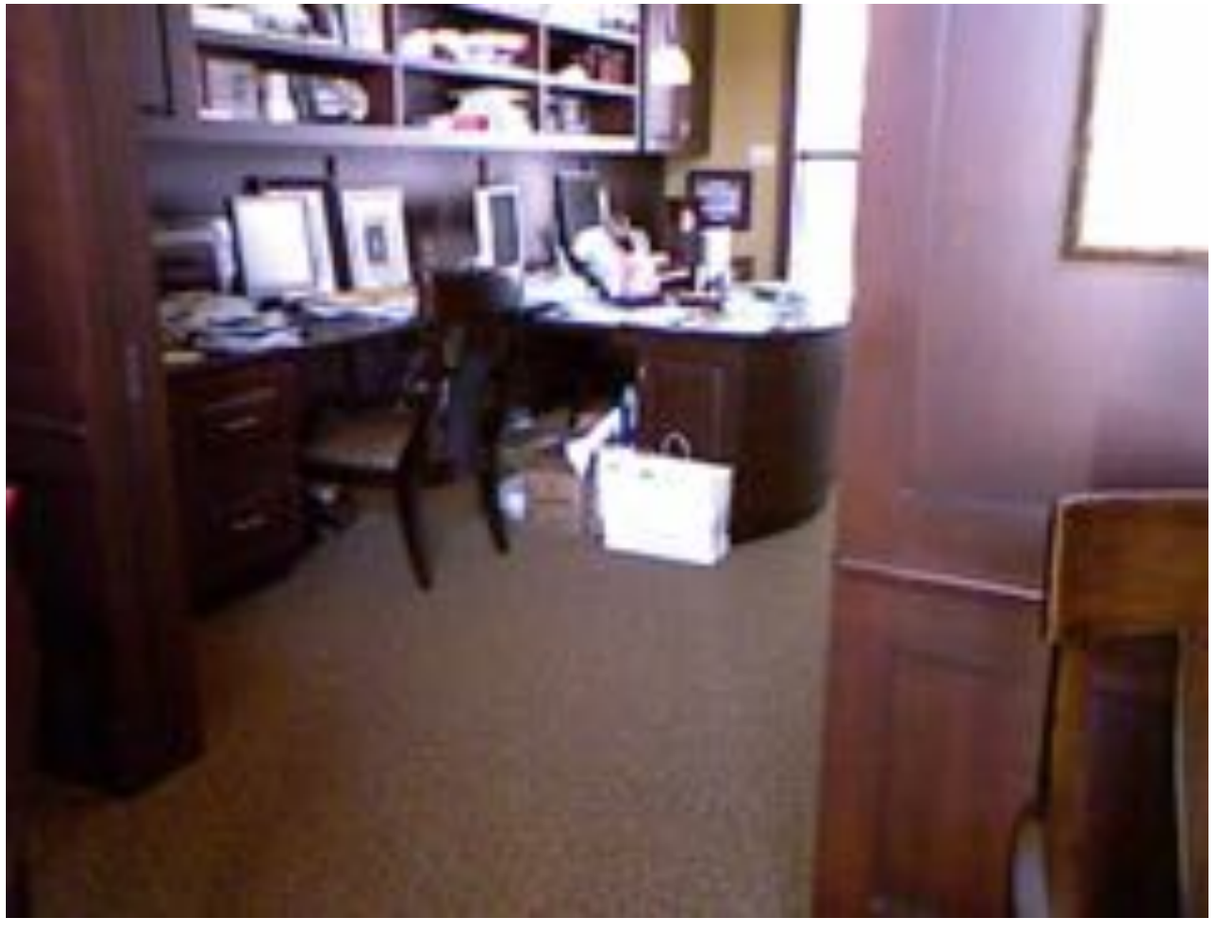
Human 1: bed

Human 2: quilt

2. New Performance Metric: Average Consensus

- We extend WUPS scores by Average Consensus
  - Averaging over multiple possible human answers
  - Encourages the most agreeable answers

$$\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\min\{\prod_{a\in A^i}\max_{t\in T_k^i}\mu(a,t),\ \prod_{t\in T_k^i}\max_{a\in A^i}\mu(a,t)\}$$



What is in front of table?

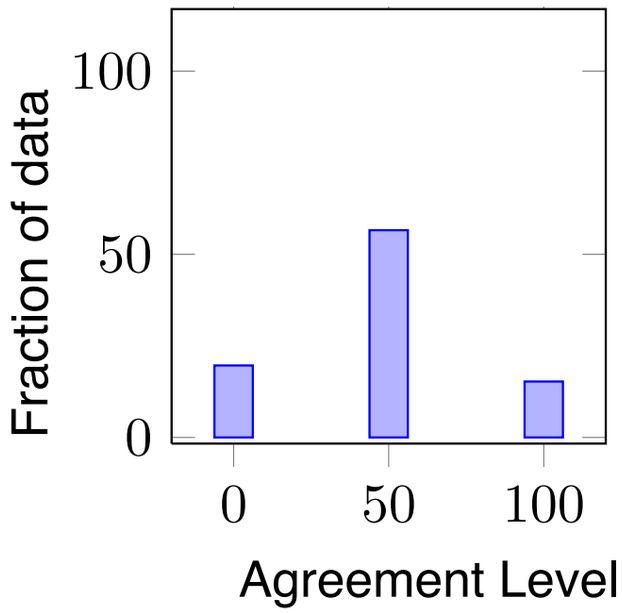
Human 1: **chair**
Human 2: **chair**
Human 3: **chair, bag**
Human 4: **wall**

For the Average Consensus:
answer chair is better than wall

Results on DAQUAR-Consensus

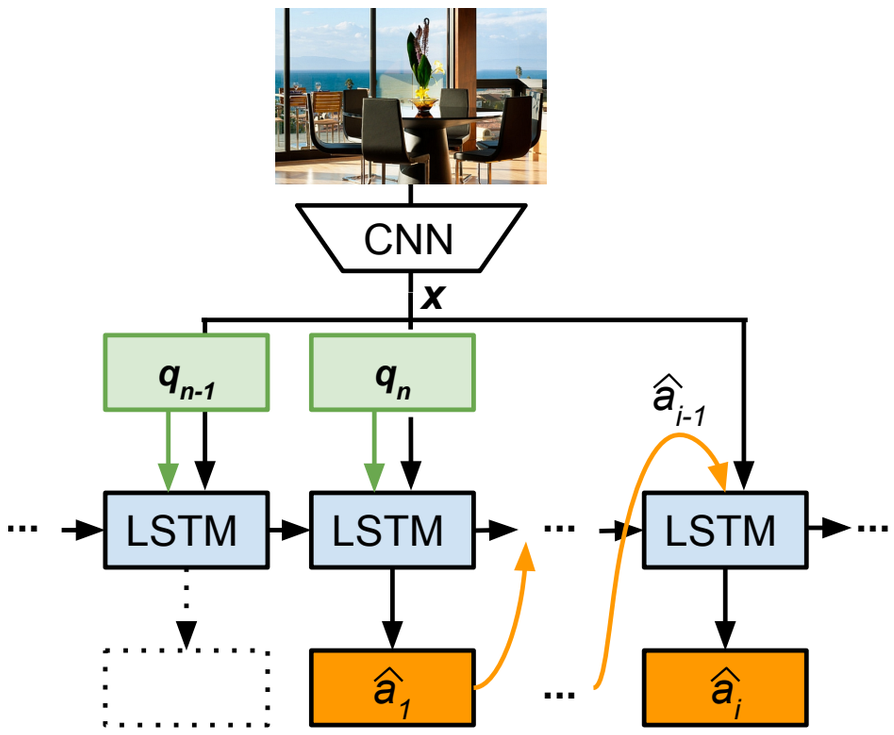
Methods (Average Consensus)	Accuracy	WUPS @0.9
Language Only (Our)	11.57%	18.97%
Vision + Language (Our)	**13.51%**	**21.36%**
Human performance (Our)	36.78%	45.68%

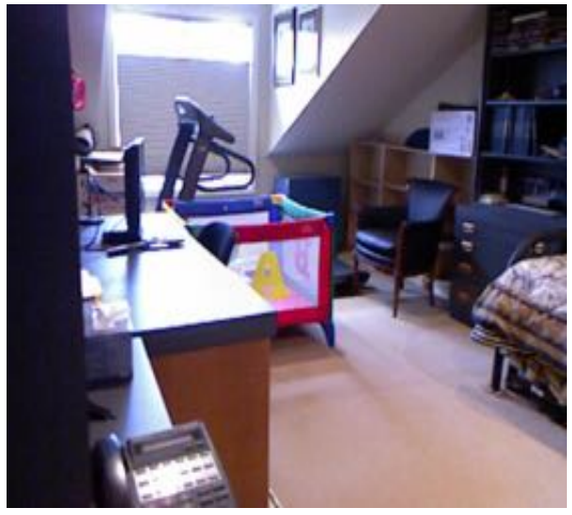
Amount of subjectivity in the task captured by the Consensus metric



Conclusions

- Towards a Visual Turing Test
  - Can machine answer questions about images?
- Novel Neural-based architecture
- End-to-end training on Image-Question-Answer triples
- Doubles the performance of the previous work on DAQUAR
- New Consensus Metrics to deal with many interpretations
- Outlook: Explore spectrum between classic AI and Deep Learning





What is on the right side of the cabinet?
Vision + Language: bed
Language Only: bed

How many burner knobs are there?
Vision + Language: 4
Language Only: 6

Thank you for your attention!

Ask Your Neurons:
A Neural-based Approach to
Answering Questions about Images

Mateusz Malinowski
Marcus Rohrbach
Mario Fritz

https://www.d2.mpi-inf.mpg.de/visual-turing-challenge

I am expecting to finish my PhD in 2016 and
looking for new opportunities.