# High Level Computer Vision - June 25th, 2o19

# Visual Turing Test / Visual Question Answering / Memory Networks

Bernt Schiele - schiele@mpi-inf.mpg.de
Mario Fritz - mfritz@mpi-inf.mpg.de

# Exam dates and registration

- the exam dates agreed on are: 18. + 19.07., 20. + 21. 08., 01.+02.10.

- In LSF, where the students need to register, only two dates can be entered. These will be 20.08. and 01.10.

- Exam dates 18.07., 19.07., 20.08., 21.08. should register in LSF for 20.08.

- All others for 01.10.

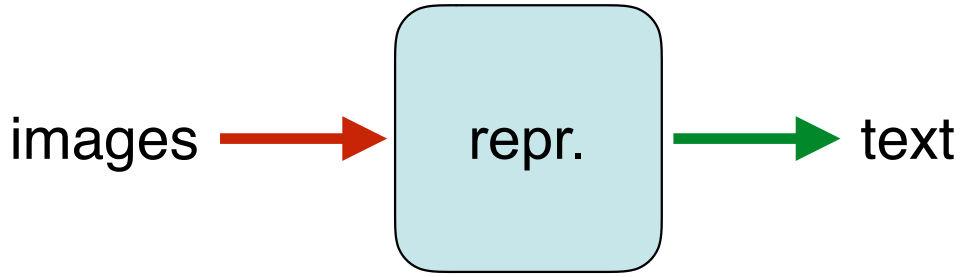
# Overview

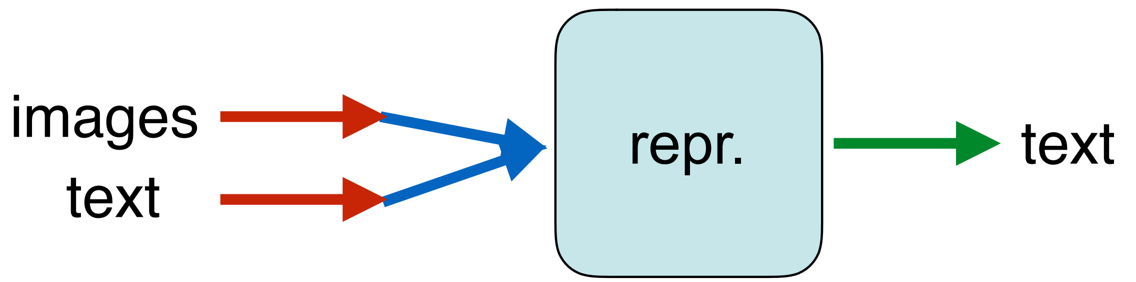
- Visual Turing Test / Visual Question Answering (VQA)
  - Motivation
  - Prior work / background
  - Overview / bigger picture
  - “Attention”-based methods

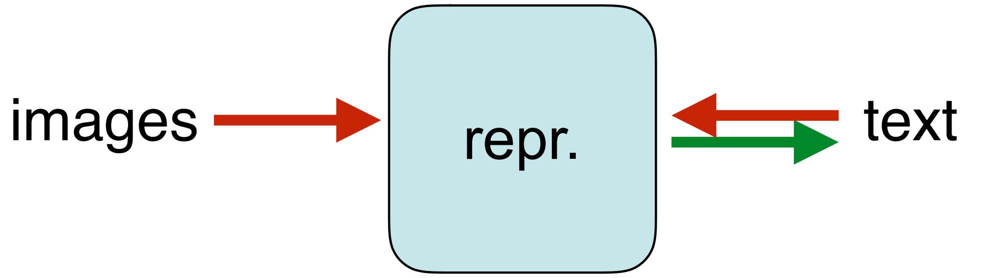
  - Relevant papers:
    - Malinowski, Fritz “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input” NIPS’14
    - Malinowski, Rohrbach, Fritz “Ask your Neurons” ICCV’15
    - Sukhbaatar “End-to-End Memory Networks” NIPS’15
    - Yang “ Stacked Attention Networks for Image Question Answering” CVPR’16

# Overview of Deep Learning Architectures

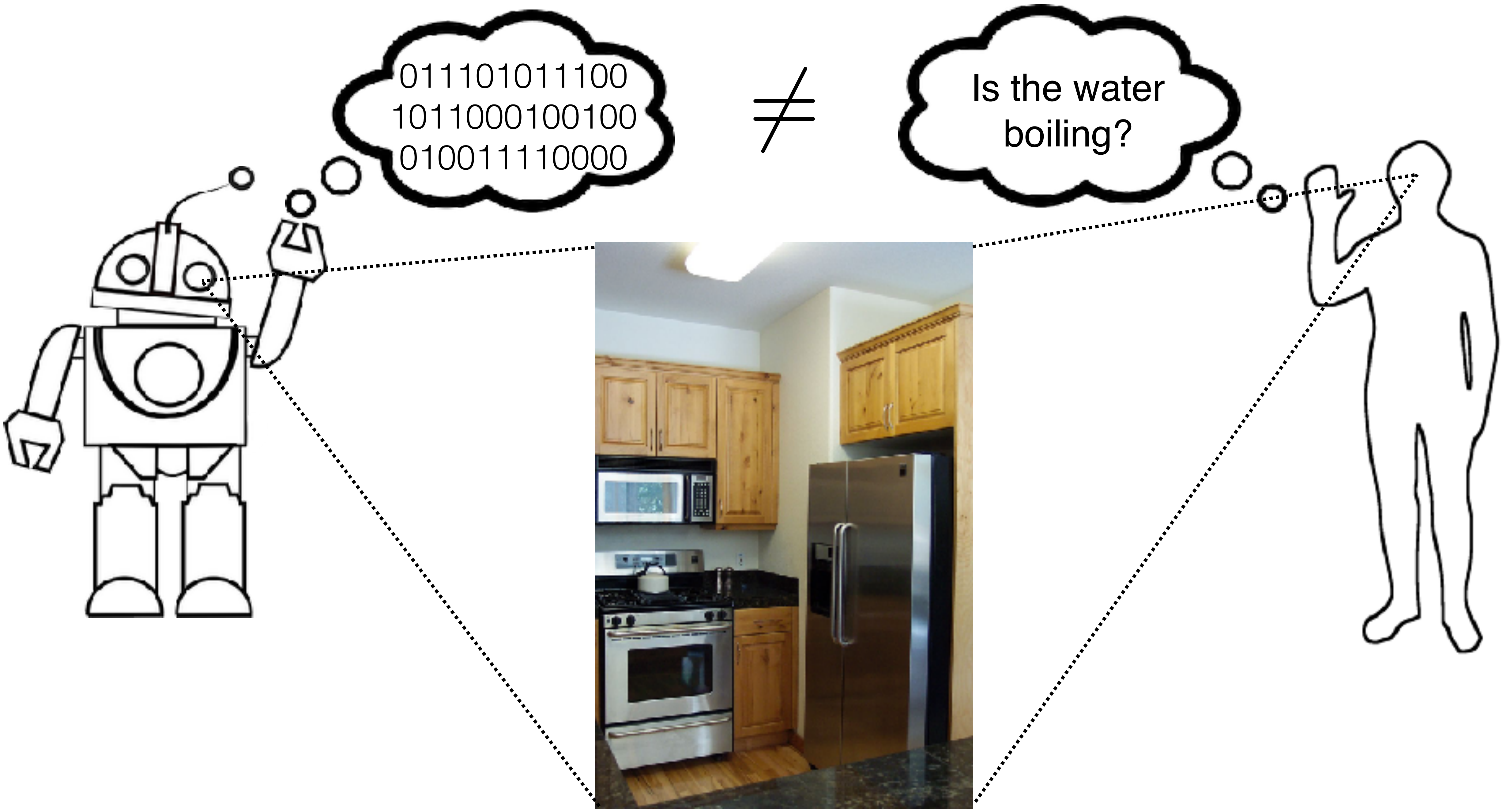
- **Encoders**
  - CNN for sequences, images, volumes
  - RNN for sequences
  - Pooling for sequences
  - Dense embedding layer (e.g. language w2v)
- **Decoders**
  - Unpooling for sequences, images, volumes
  - RNN for sequences
  - Dense regression
- **Merge**
  - Concatenate
  - Multiply
  - Sum/Average



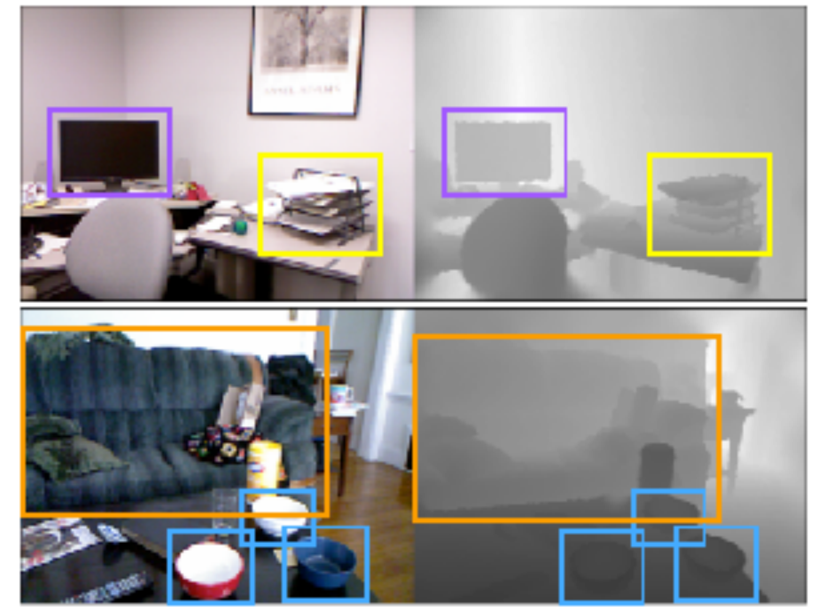



# Human-like Comprehension

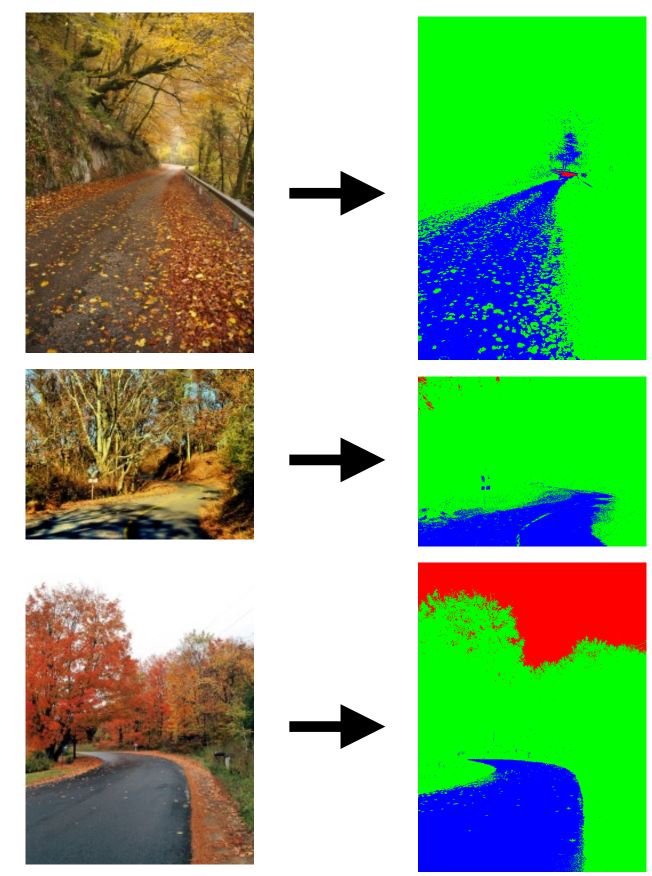



- How far are machines from human quality understanding?
- How can we monitor progress and evaluate architectures?

# Human-type Comprehension / Scene Understanding?



- Object Detection / Bounding Boxes?
- Semantic Segmentation / Pixel Annotations?
- Attributes?
- Materials?
- Spatial Relations?
- **Annotation** gets more and more challenging
- Understanding should be agnostic to some extend to the internal representation
- Scene Description -> **Evaluation** is difficult

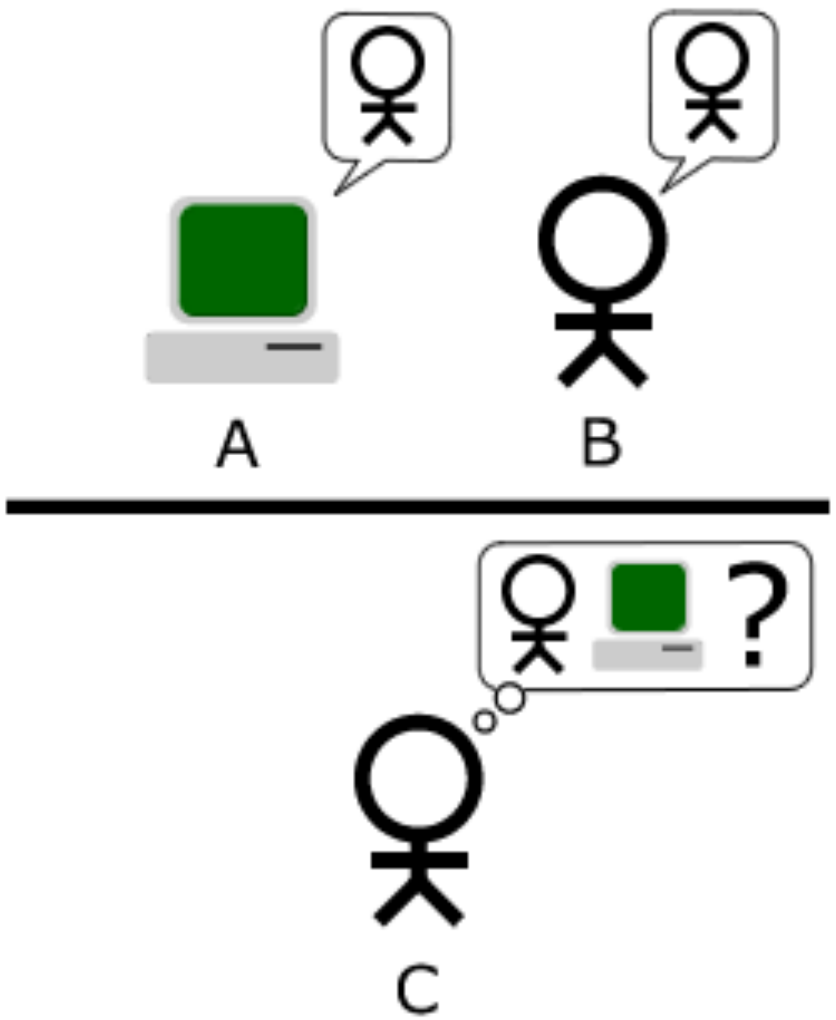


A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

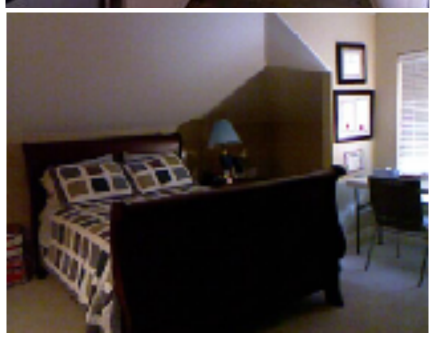
# Motivation: Turing Test

- Can a machine mimic human behavior?

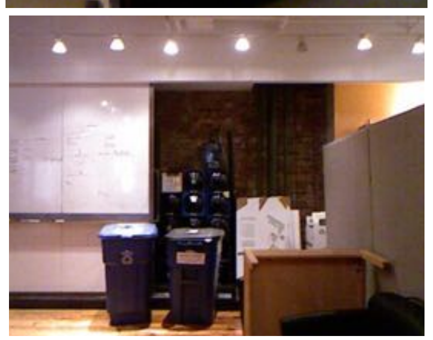


# DAQUAR: Proposed Visual Turing Challenge (NIPS'14)

Q: What is the object on the counter in the corner?
A: micro wave



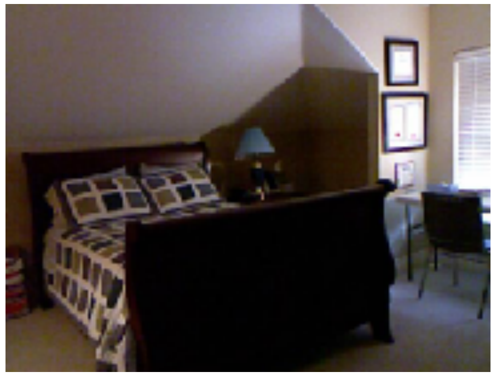
What is the color of the largest object in the scene?
A: brown



Q:How many lights are on?
A: 6

- Builds on top of NYU Depth Data set: 1449 RGBD images
- 12,5k question answer pairs (with ~ 5 answers per question)
- Answers: attributes, numbers, objects and sets of these
- Human Baselines (with and without image)
- https://www.d2.mpi-inf.mpg.de/visual-turing-challenge

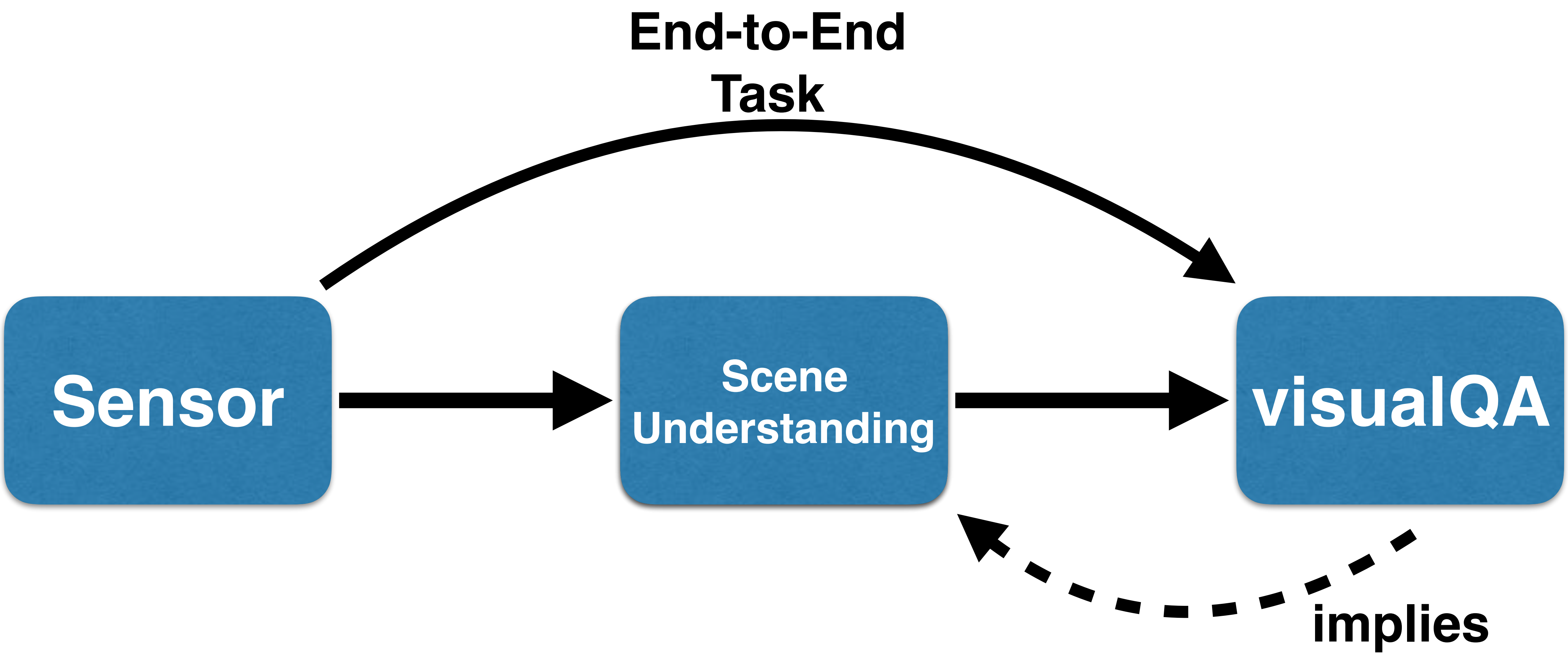
# Proposed Visual Turing Challenge



What is the color of the largest object in the scene?
A: brown

- Inspired by Turing Test:
  - Can machines answer on questions about natural images?
  - Cannot be easily be cheated like original Turing Test
- A holistic, open-ended, end-to-end task
  - Whole chain of perception, representation and deduction
- No internal representation is evaluated
  - Challenge is open to diverse approaches
- Scalable annotation effort
  - Only question-answer-pair annotations
  - Yet deep understanding of language and scenes required
- Strategies for automatic evaluation

# End-to-End Tasks




- Evaluate task that requires capability/skill (scene understanding)
- Rather than “scene understanding”
- E.g. design tasks that afford scene understanding
- Kind of facilitated by deep learning

# Our Approaches

- “Classic AI”, symbolic reasoning approach
  - A Multi-world Approach to Question Answering about Real-World Images (NIPS’14)
    Mateusz Malinowski, Mario Fritz
    NIPS’14
- Neural Network / Deep Learning / Vector Embedding (ICCV’15)

# A Multi-World to Question Answering About Real-World Images

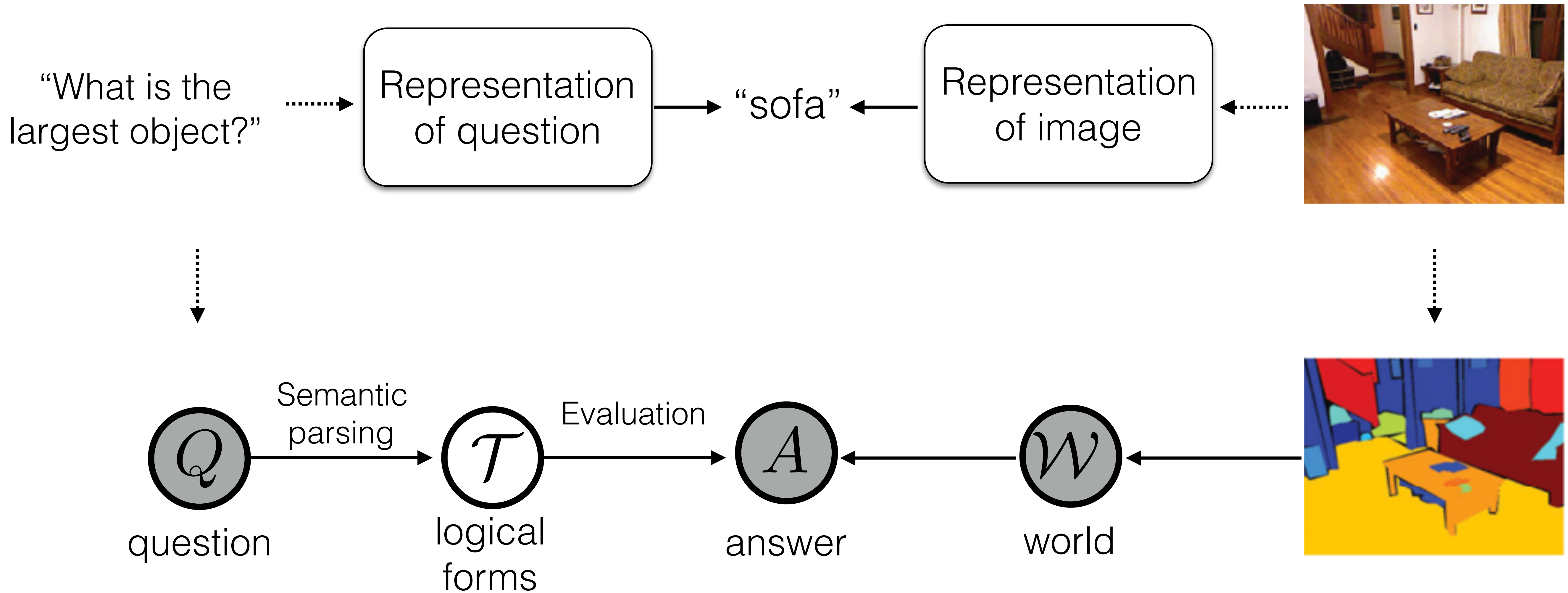
**Mateusz Malinowski, Mario Fritz**

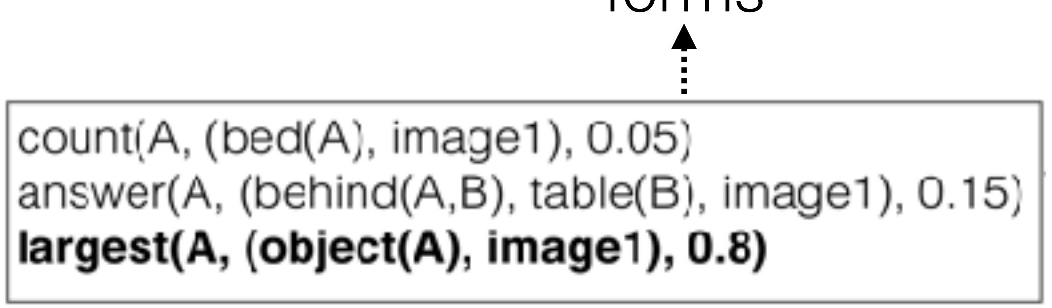
**NIPS'14**

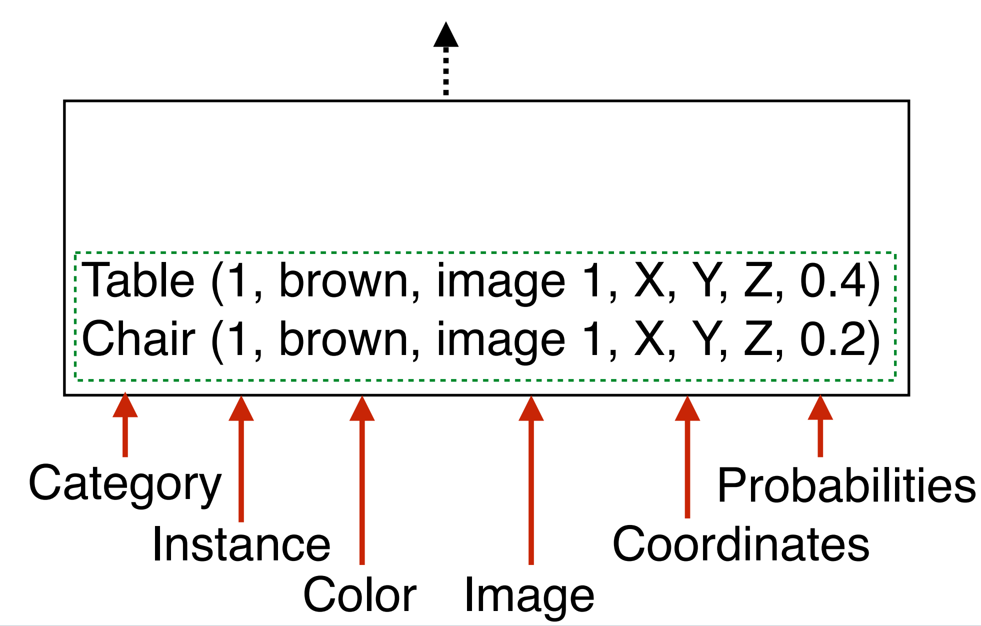
# Methods

"What is the largest object?" → Representation of question → "sofa" ← Representation of image ← 

# Method: Symbolic Approach [NIPS’14]

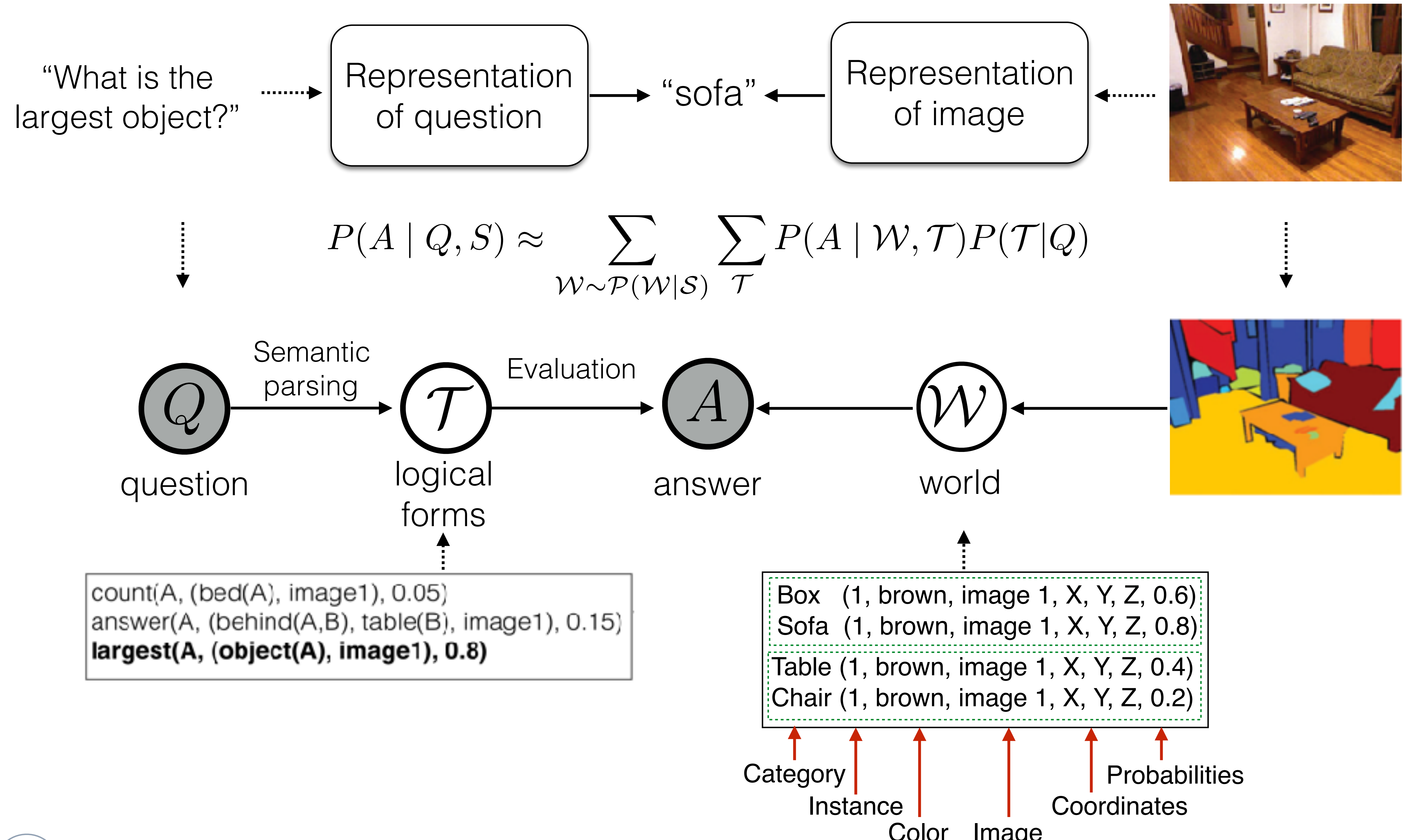



P. Liang, M. I. Jordan, D. Klein. Learning dependency-based compositional semantics. ACL 2011.

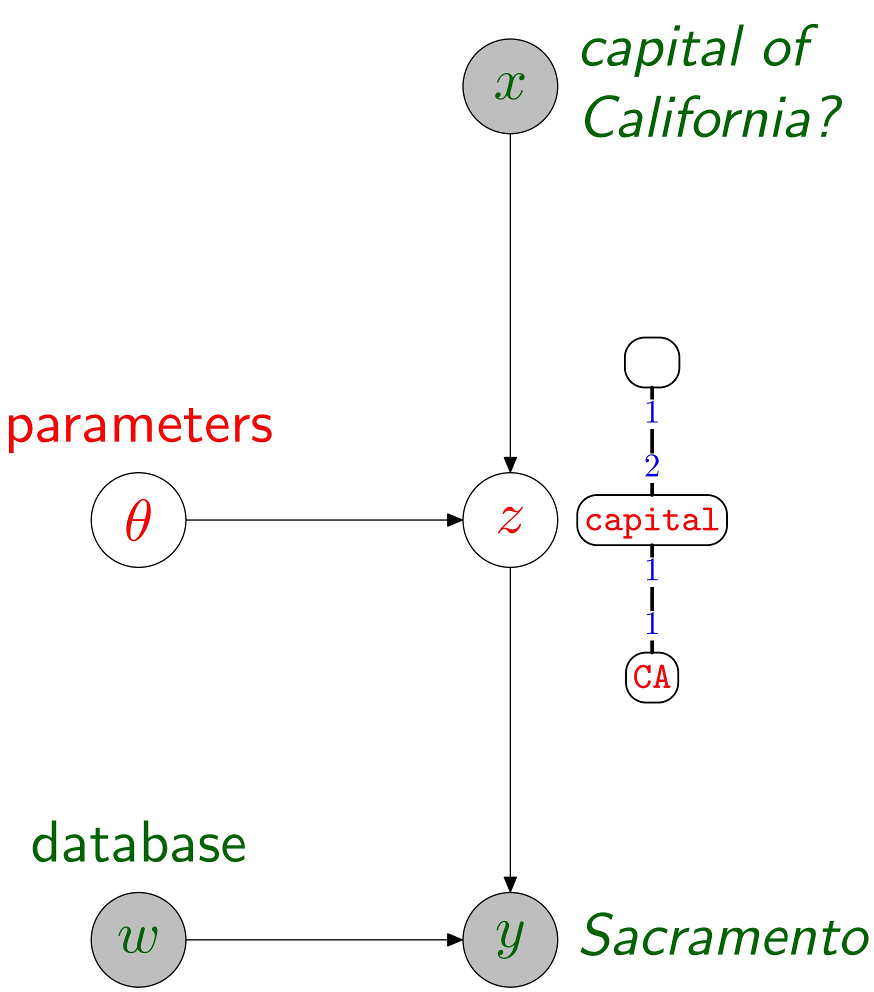
# Method: Symbolic Approach [NIPS'14]



# QA by Percy Liang (2011)

## Words to Predicates (Lexical Semantics)

| What | is | the | most | populous | city | in | CA | ? |
|---|---|---|---|---|---|---|---|---|
| | | | argmax | city<br>state<br>river<br>population | city<br>state<br>river<br>population | | CA | |




### Objective

$$\max_{\theta} \sum_{z} p(y \mid z, w)\, p(z \mid x, \theta)$$

**Interpretation** **Semantic parsing**

### Learning

parameters $\theta$

$(0.2, -1.3, \ldots, 0.7)$

enumerate/score DCS trees →

← numerical optimization (L-BFGS)

$k$-best list

- tree1 ✗
- tree2 ✗
- tree3 ✓
- tree4 ✗
- tree5 ✗

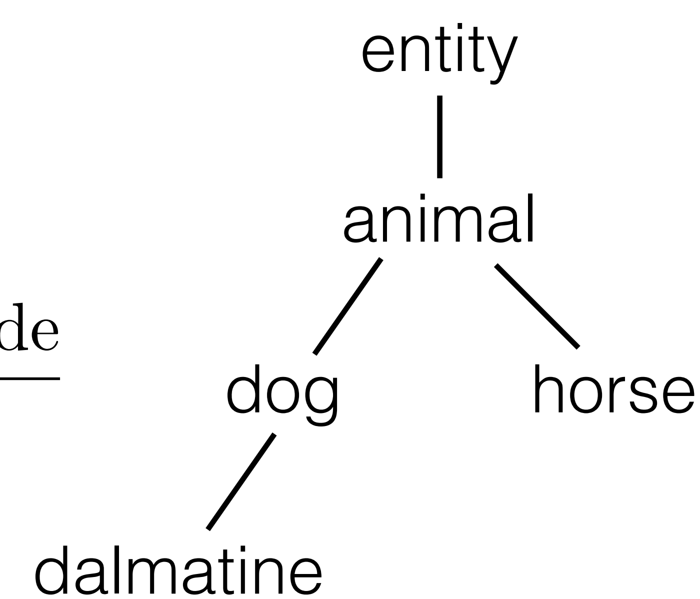
# Evaluation Criterion

- All measures can be evaluated automatically
- Less error prone than BLEU score

- Different metrics:
  - accuracy
  - WU Palmer Similarity

$$\mathrm{WUP}(w1, w2) = 2 * \frac{\text{depth most specific ancestor node}}{\mathrm{depth}(w1) + \mathrm{depth}(w2)}$$

WUP(horse, dalmatine) = 2*2/(4+3) = 4/7 = 0.57




  - WUPS: Wu Palmer extended to sets

$$\mathrm{WUPS}(A, T) = \frac{1}{N} \sum_{i=1}^{N} \min\{ \prod_{a \in A^i} \max_{t \in T^i} \mathrm{WUP}(a, t), \prod_{t \in T^i} \max_{a \in A^i} \mathrm{WUP}(a, t)\} \cdot 100$$

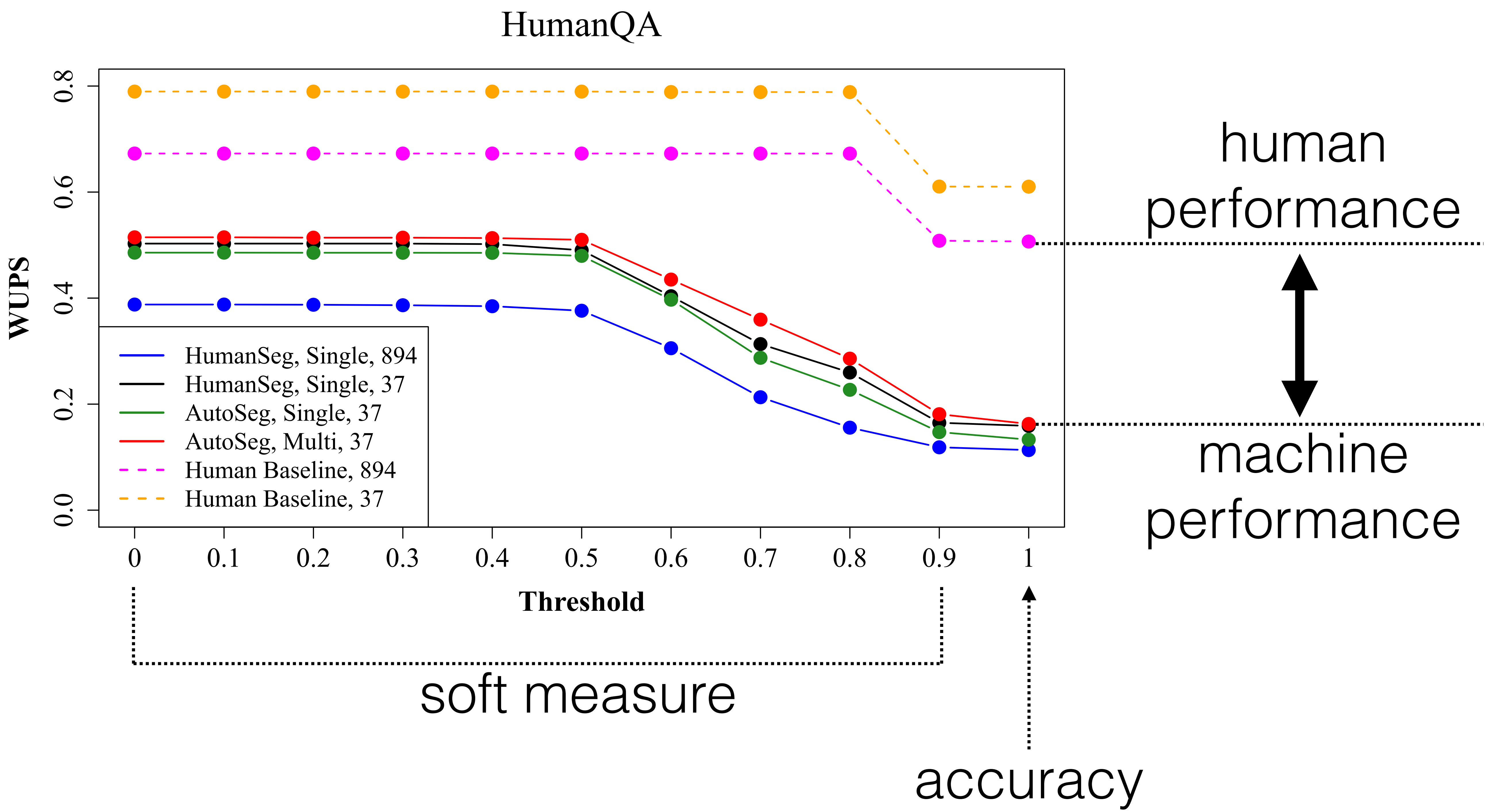
  - Additional consensus metrics over 5 annotators

# Evaluation: WUPS

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

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

# Quantitative Results



# Qualitative Results

Q: How many red chairs are there?
H: ()
M: 6
C: blinds

Q: How many chairs are at the table?
H: wall
M: 4
C: chair

Q: What is on the right side of cabinet?
H: picture
M: bed
C: bed

Q: What is on the wall?
H: mirror
M: bed
C: picture

# Conclusions

- Pros
  - First proposal of Visual Turing Challenge based on diverse real-world images
  - Multi-world for learning to answer questions about scenes
  - Bridging between symbolic reasoning and uncertainty in perception
  - Requires deep understanding of scenes at low annotation effort
- Cons
  - Poor scalability
  - Some hand crafting of ontology and predicates

# Our Approaches

- Classic AI, symbolic reasoning approach

- Neural Network / Deep Learning / Vector Embedding (ICCV’15)

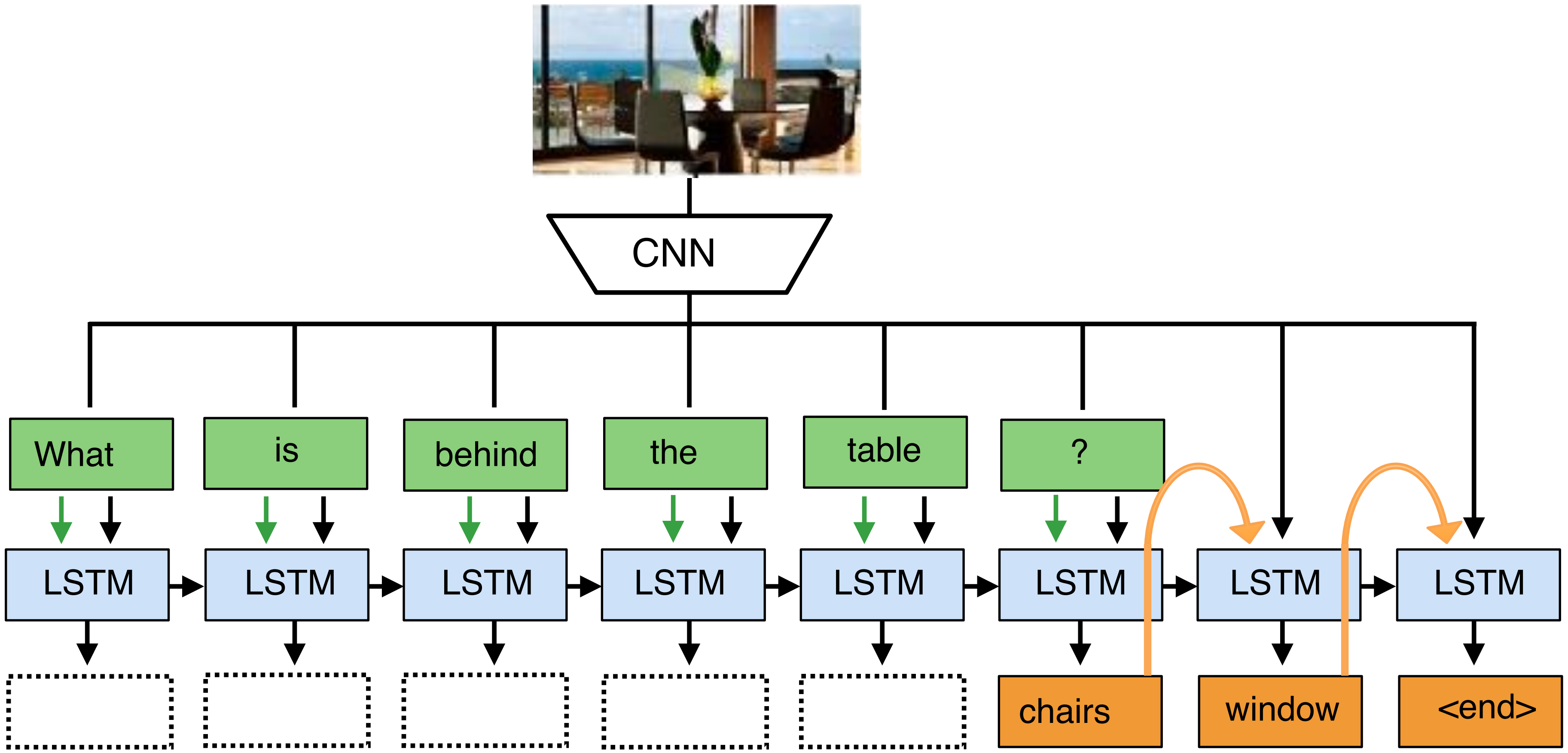
  Ask your Neurons: A Neural-based Approach to Answering Questions about Image
  Mateusz Malinowski, Marcus Rohrbach, Mario Fritz

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

**Mateusz Malinowski, Marcus Rohrbach, Mario Fritz**

**ICCV'15**

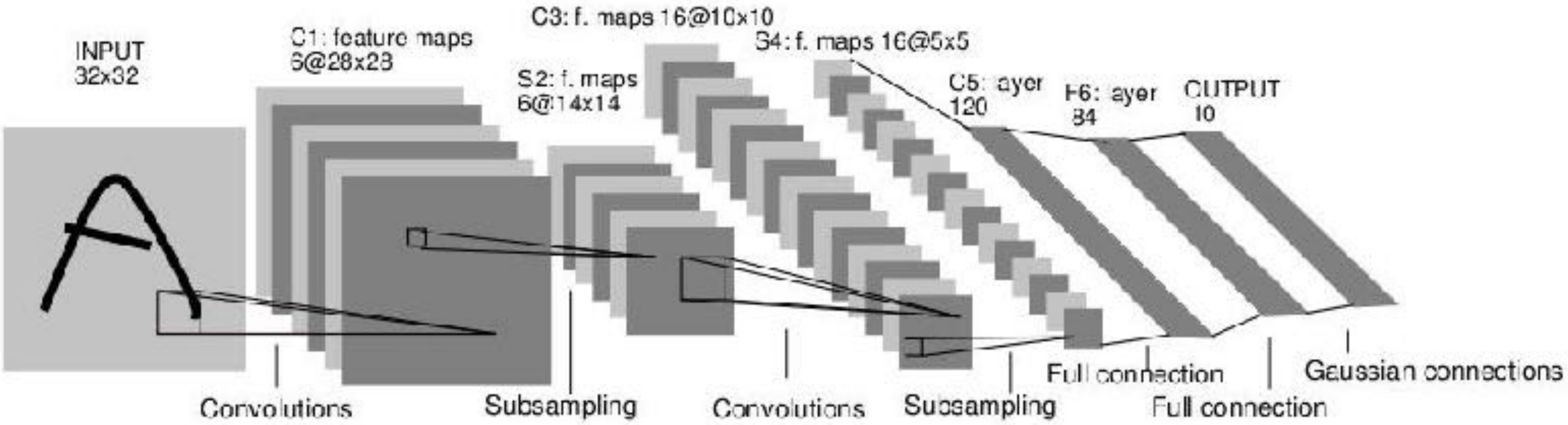
# Method: Ask Your Neurons

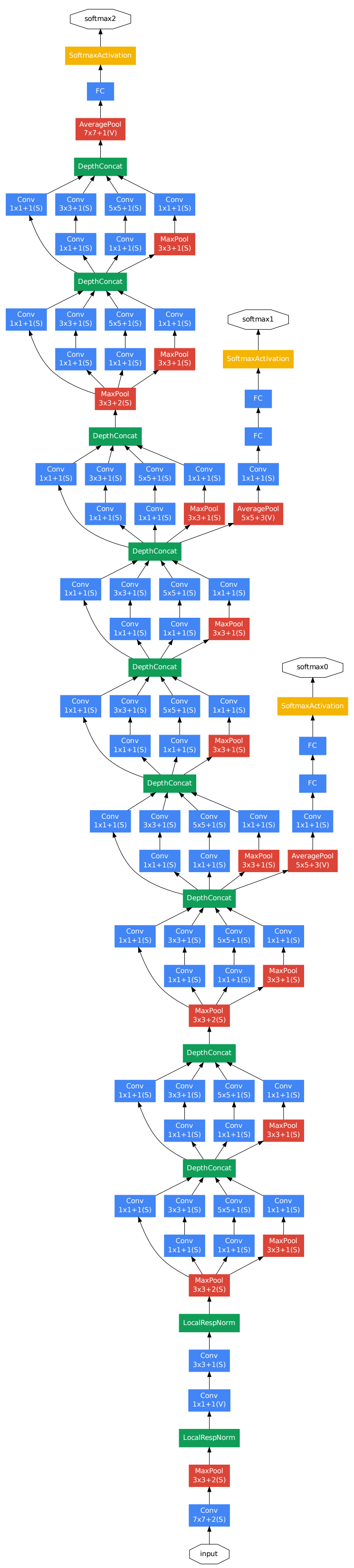


# Two Key Ingredients

- Convolutional Neural Network
- Long Short Term Memory Recurrent Neural Network

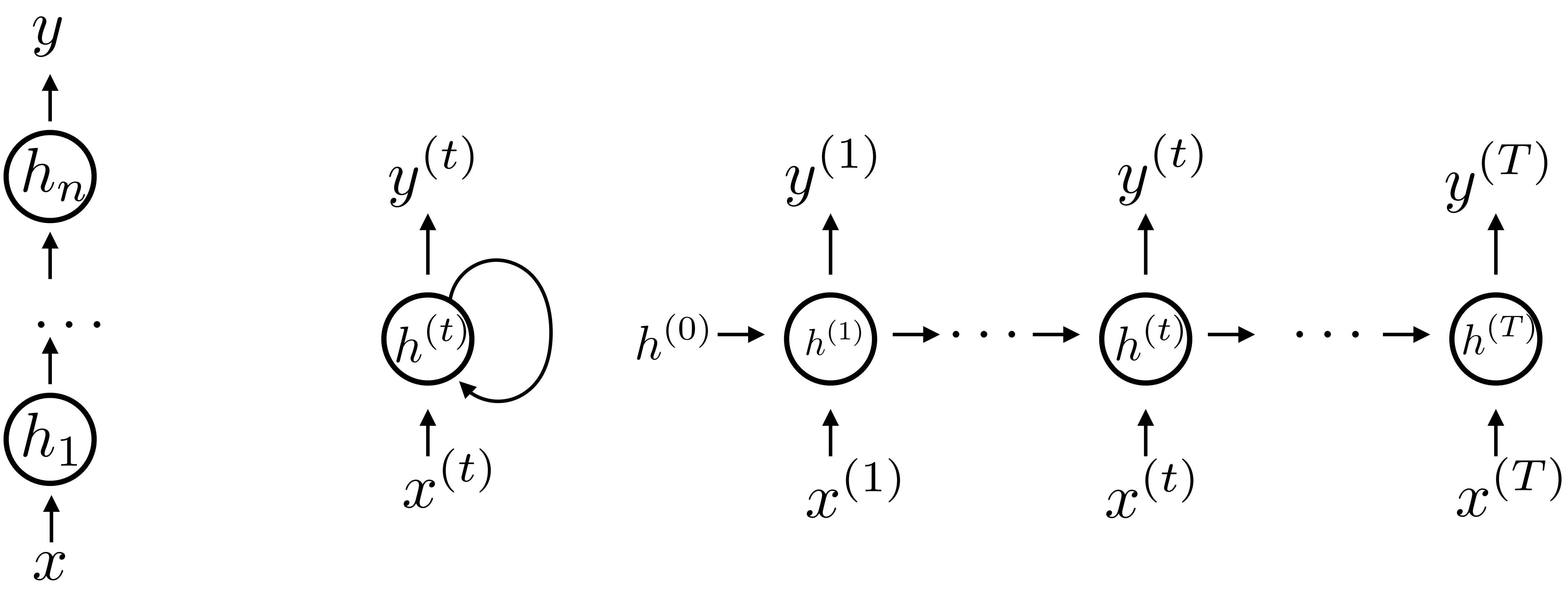
# Convolutional Neural Networks



- LeCun et al. 1989
- Neural network with specialized connectivity structure
- GoogleNet in our experiments

# Recurrent Neural Network



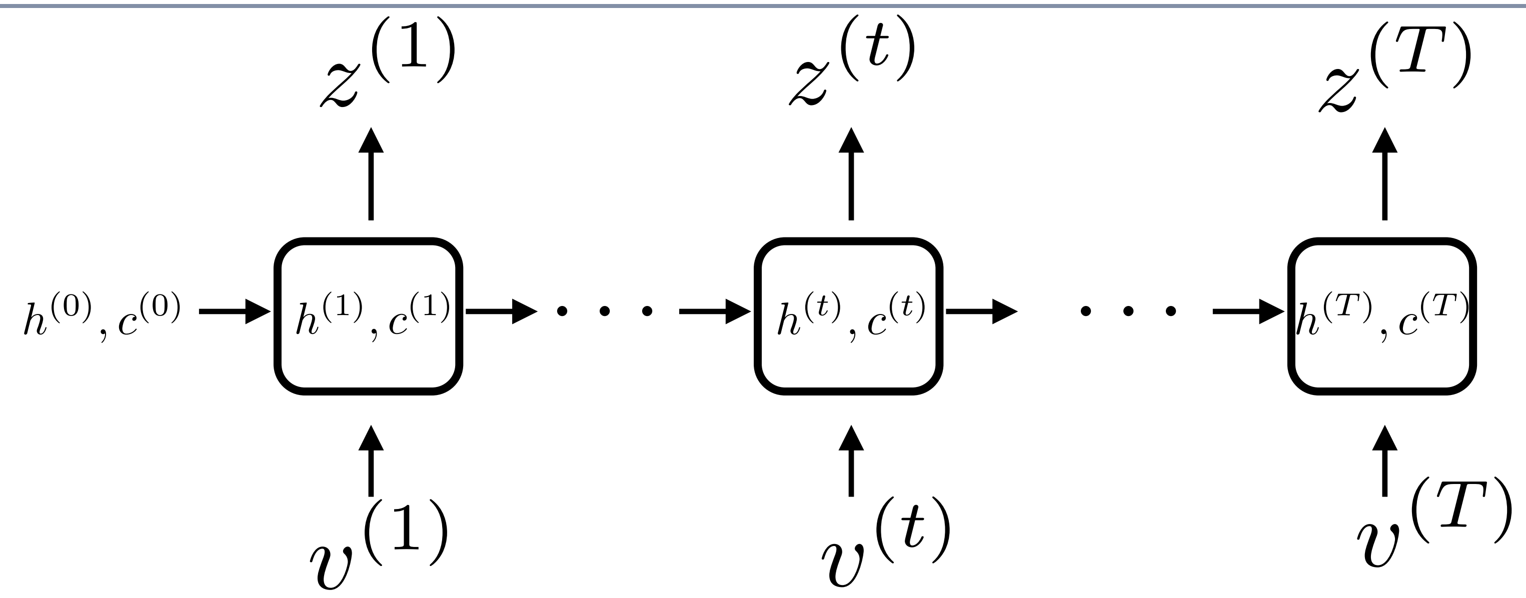

multi-layer deep feedforward network

recurrent neural network

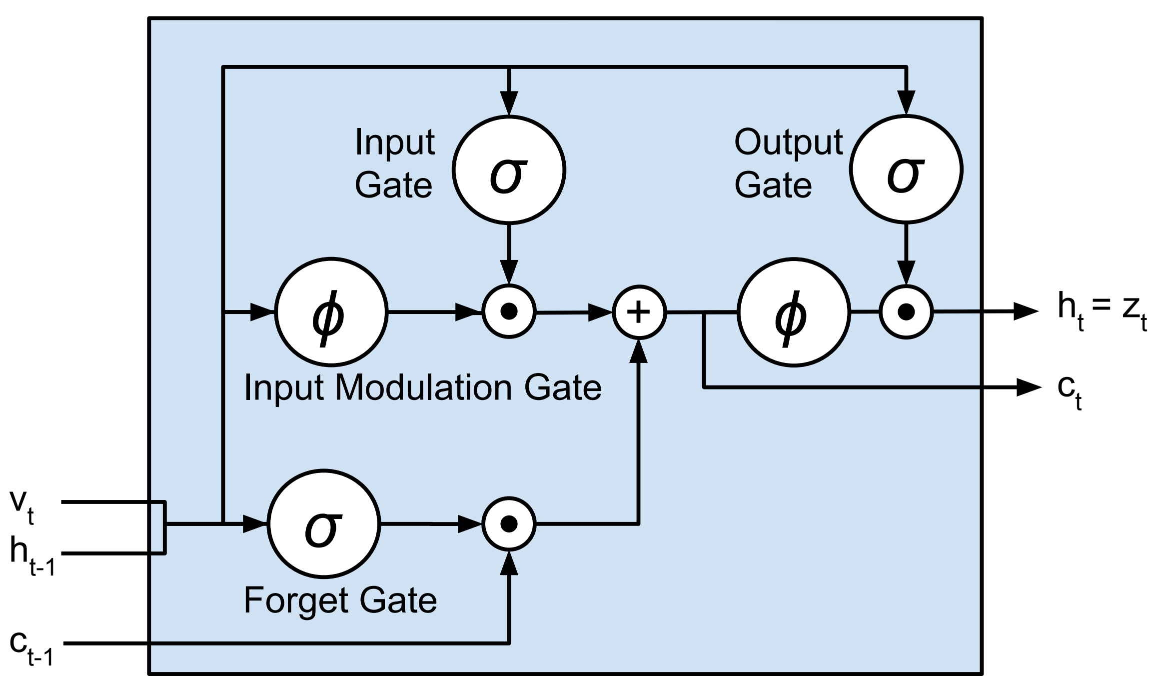
unrolled recurrent neural network

- Extension of neural networks to sequence modelling and prediction
- Training is problematic due to vanishing/exploding gradient

# Long Short Term Memory Networks (Schmidhuber)




**LSTM Unit**




$$[\boldsymbol{x}, \hat{\boldsymbol{q}}_t]$$

$$\boldsymbol{i}_t = \sigma(W_{vi}\boldsymbol{v}_t + W_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i)$$

$$\boldsymbol{f}_t = \sigma(W_{vf}\boldsymbol{v}_t + W_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f)$$

$$\boldsymbol{o}_t = \sigma(W_{vo}\boldsymbol{v}_t + W_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o)$$

$$\boldsymbol{g}_t = \phi(W_{vg}\boldsymbol{v}_t + W_{hg}\boldsymbol{h}_{t-1} + \boldsymbol{b}_g)$$

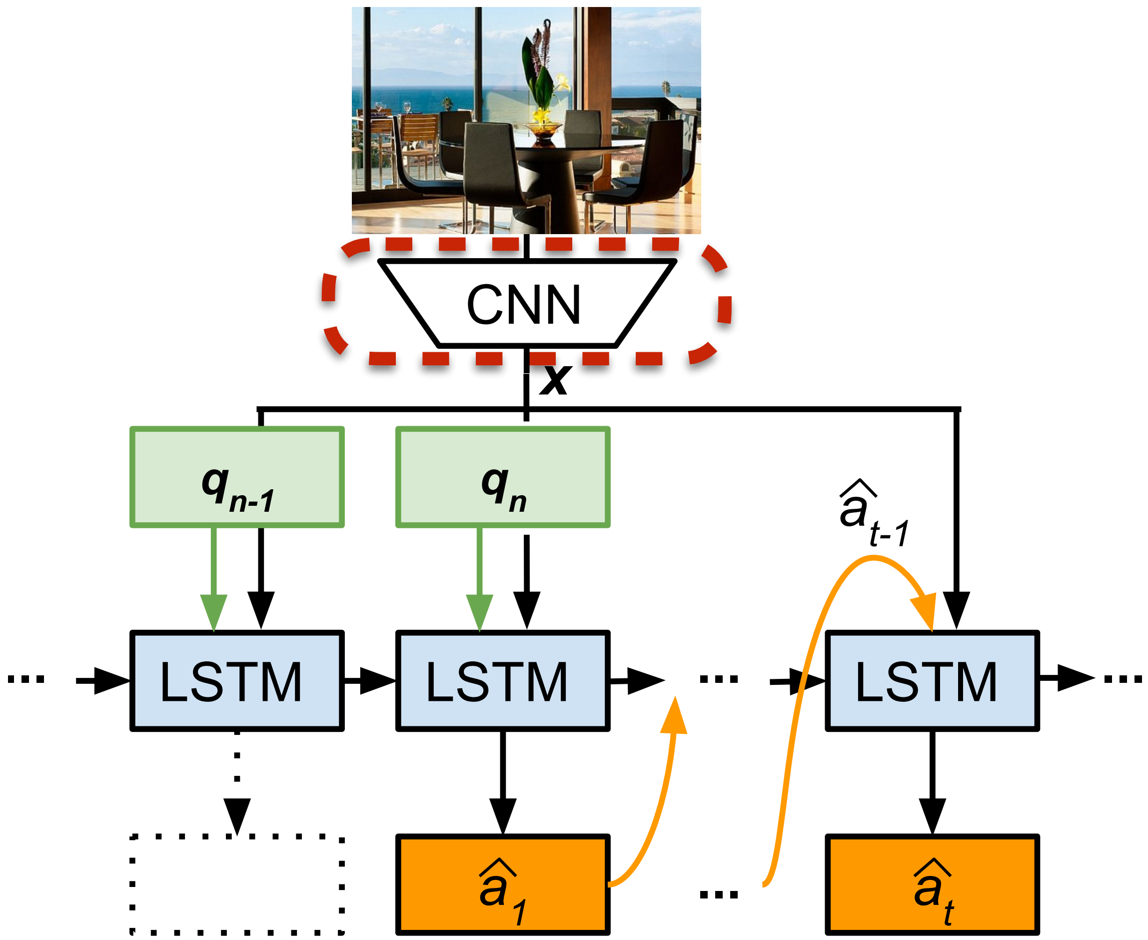
$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t) = z_t$$

*sigmoid* nonlinearity $\sigma : \mathbb{R} \mapsto [0,1]$, $\sigma(v) = (1 + e^{-v})^{-1}$

*hyperbolic tangent* nonlinearity $\phi : \mathbb{R} \mapsto [-1,1]$, $\phi(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} = 2\sigma(2v) - 1$

# Method: Ask Your Neurons



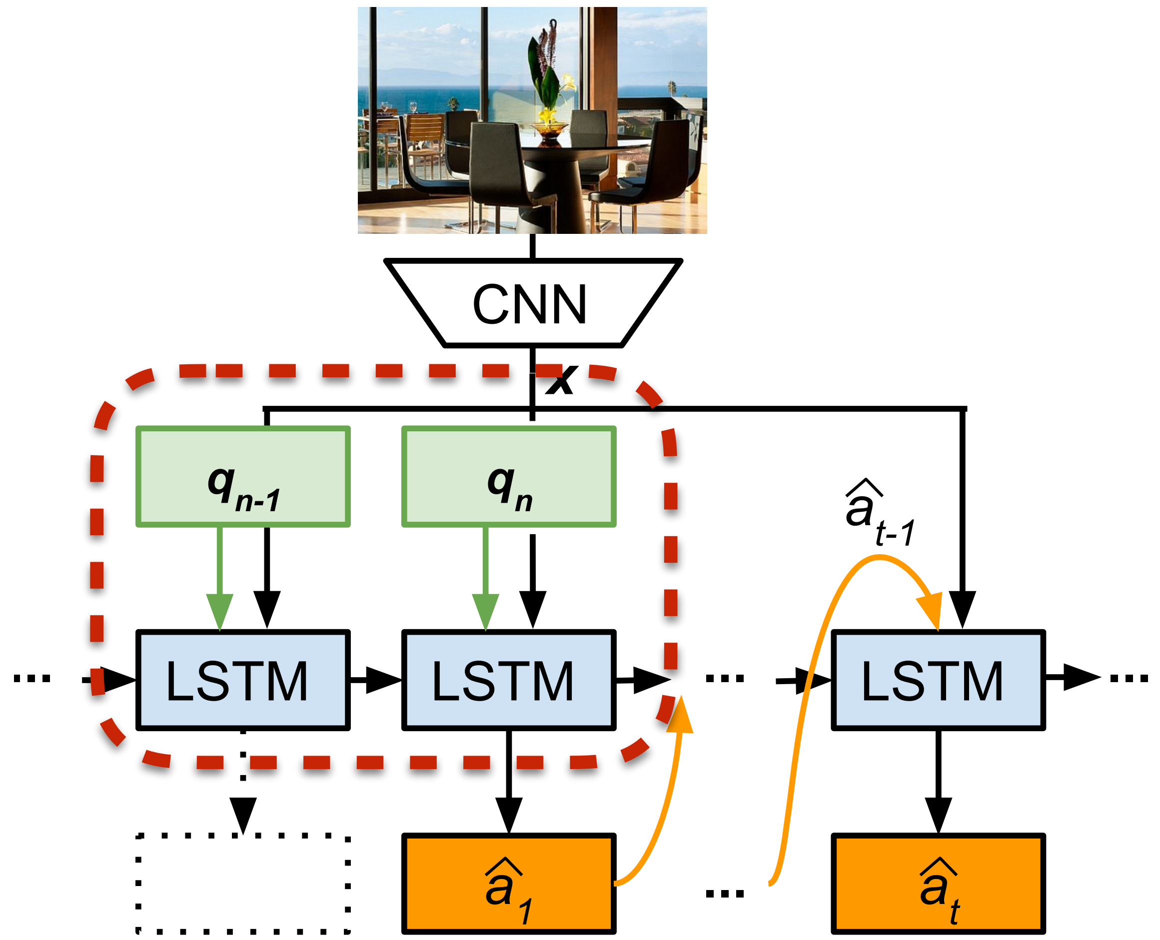

- Predicting answer sequence
  - Recursive formulation

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

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

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

# Method: Ask Your Neurons



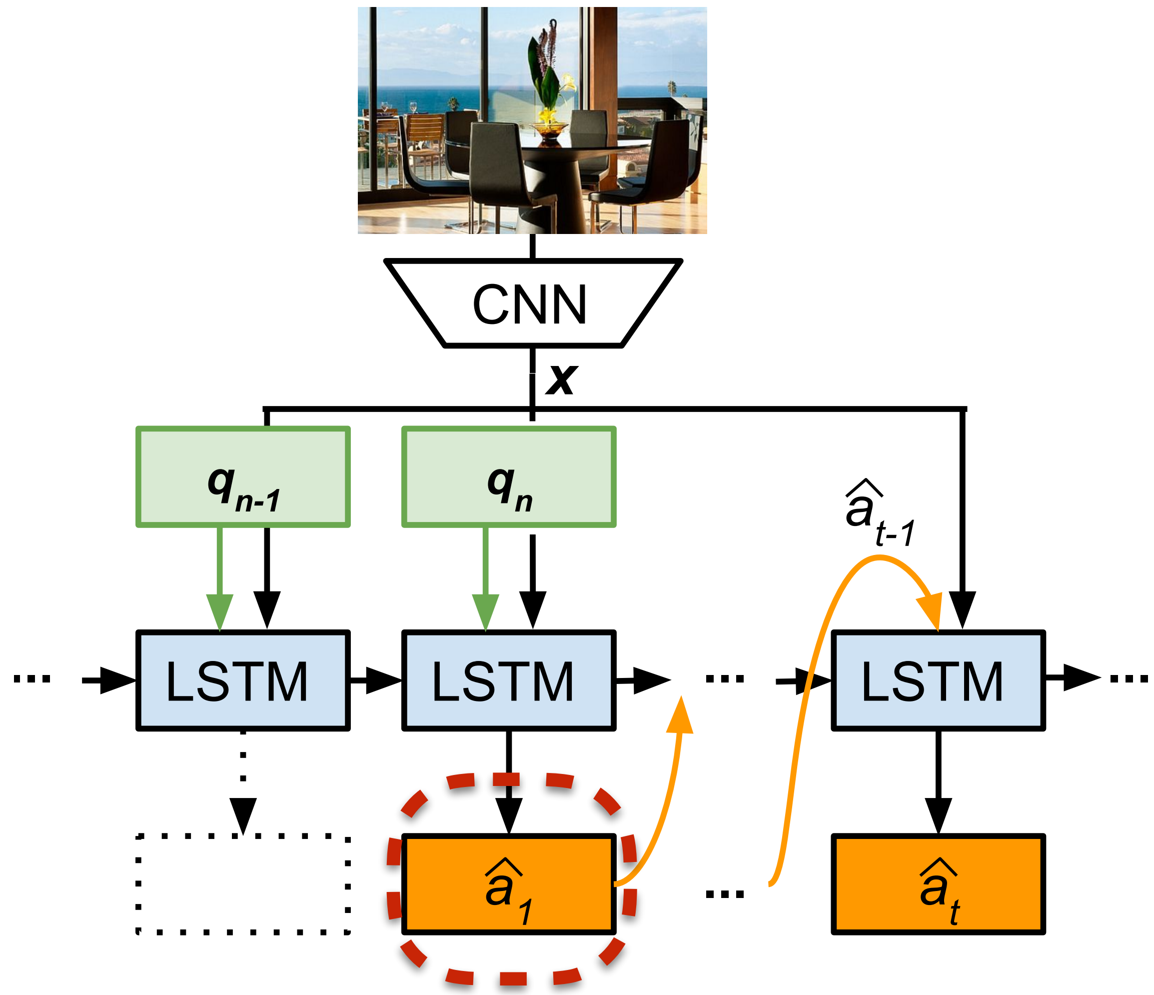

- Predicting answer sequence
  - Recursive formulation

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

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

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

# Method: Ask Your Neurons



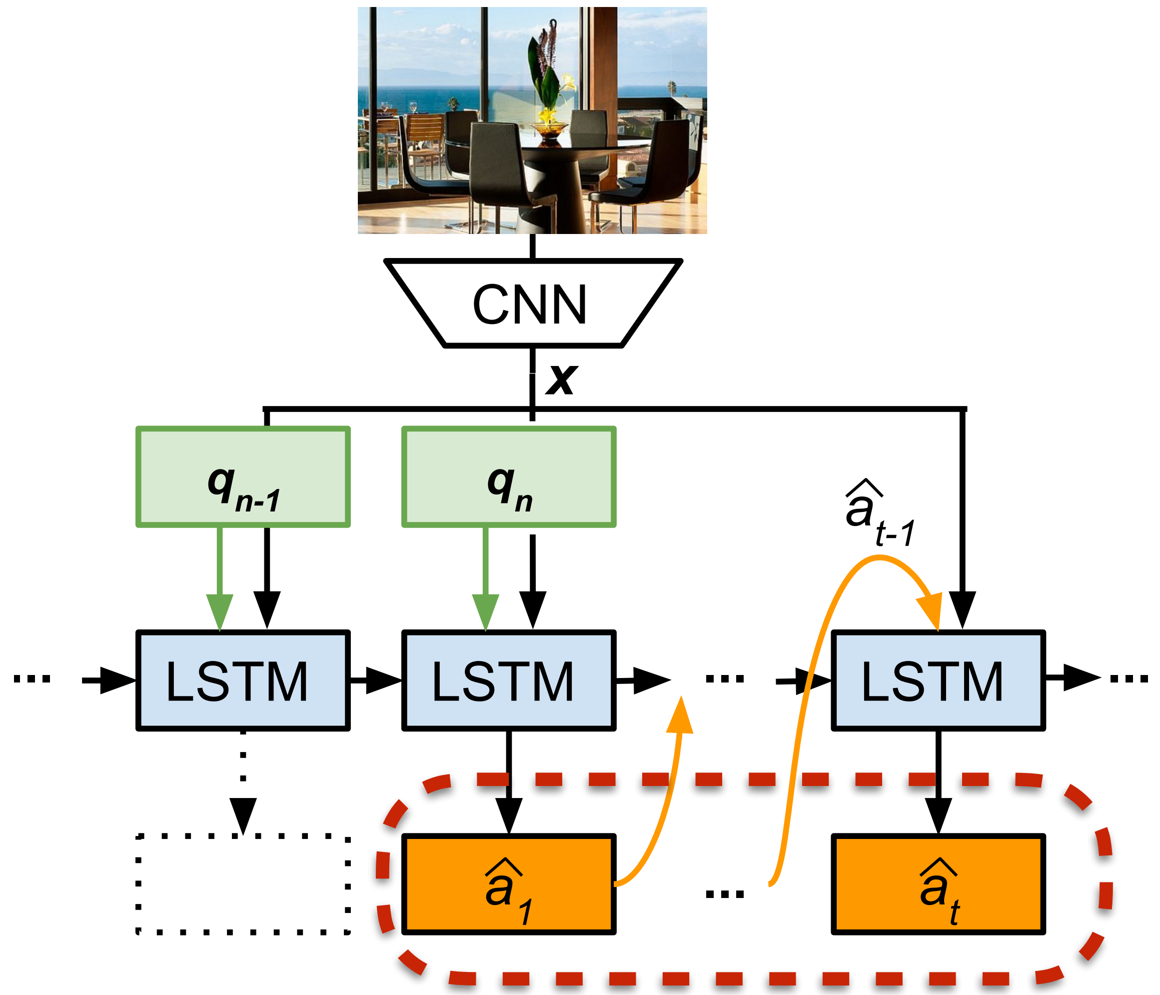

- Predicting answer sequence
  - Recursive formulation

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

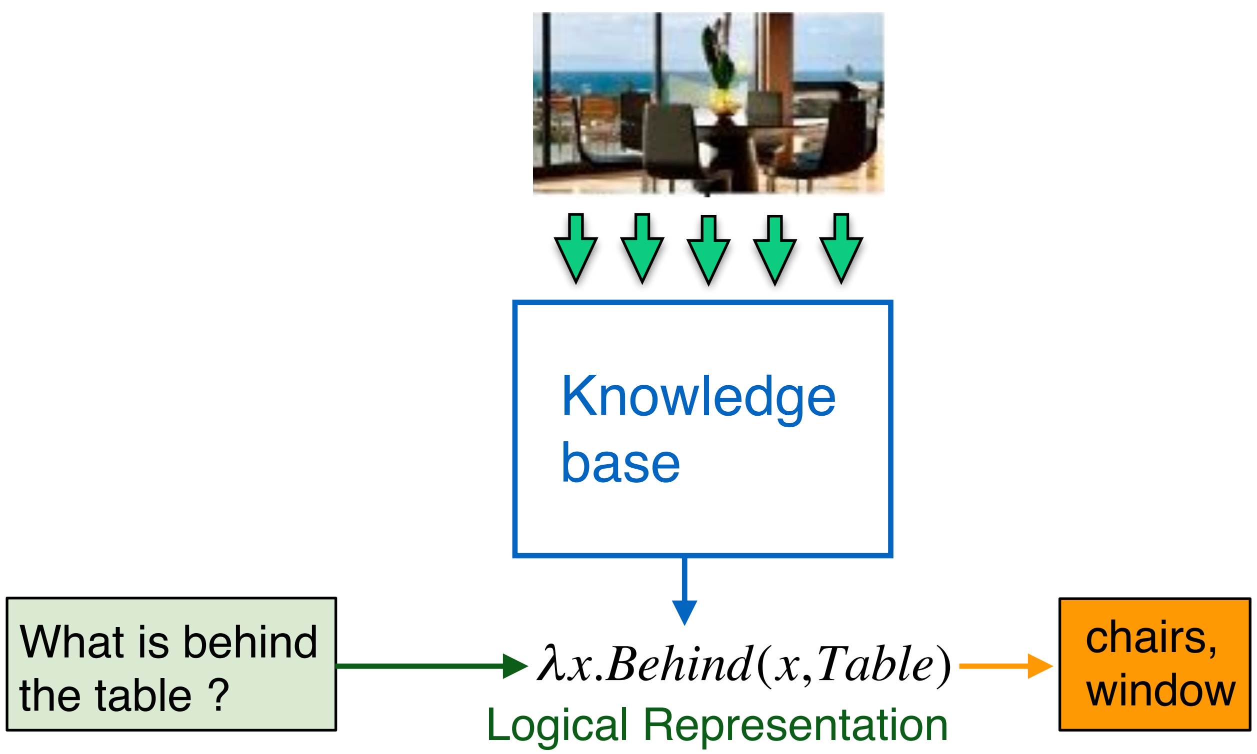
$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a} \in \mathcal{V}} p(\boldsymbol{a} | \boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

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

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

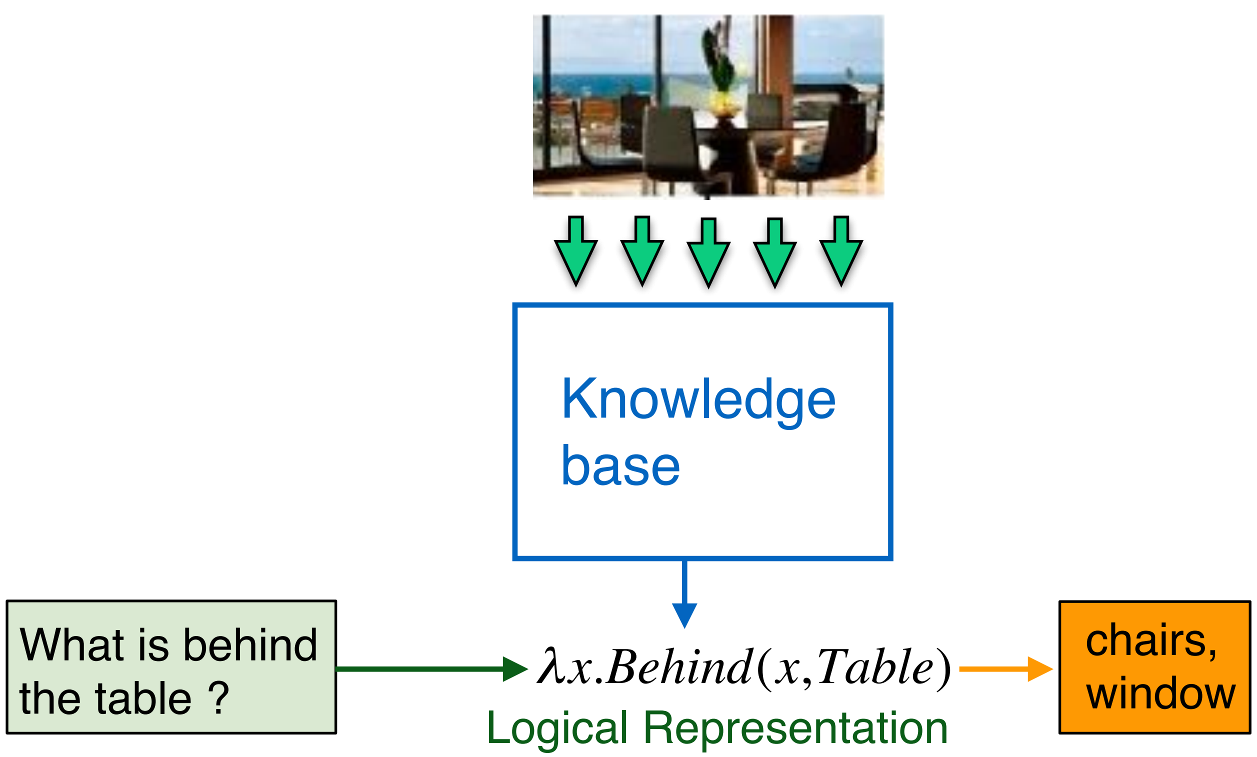



M. Malinowski, et. al. “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”. NIPS’14

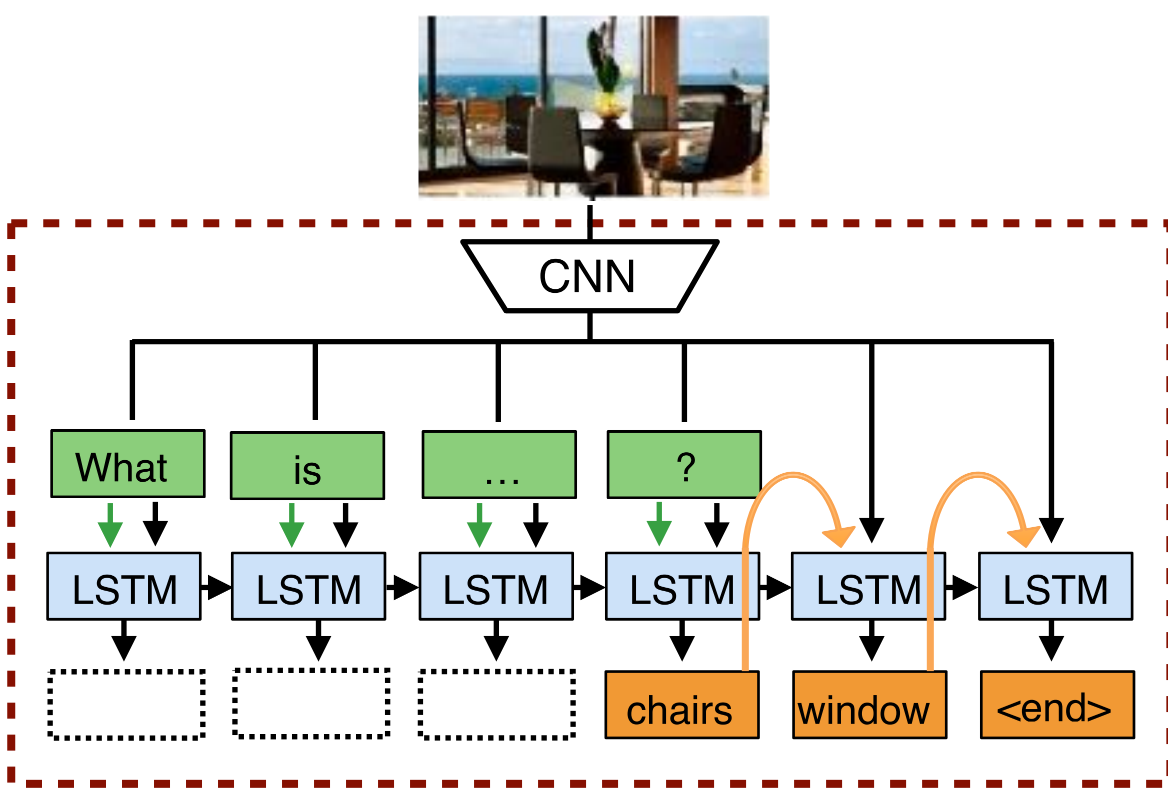
# Symbolic vs Neural-based Approaches

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

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



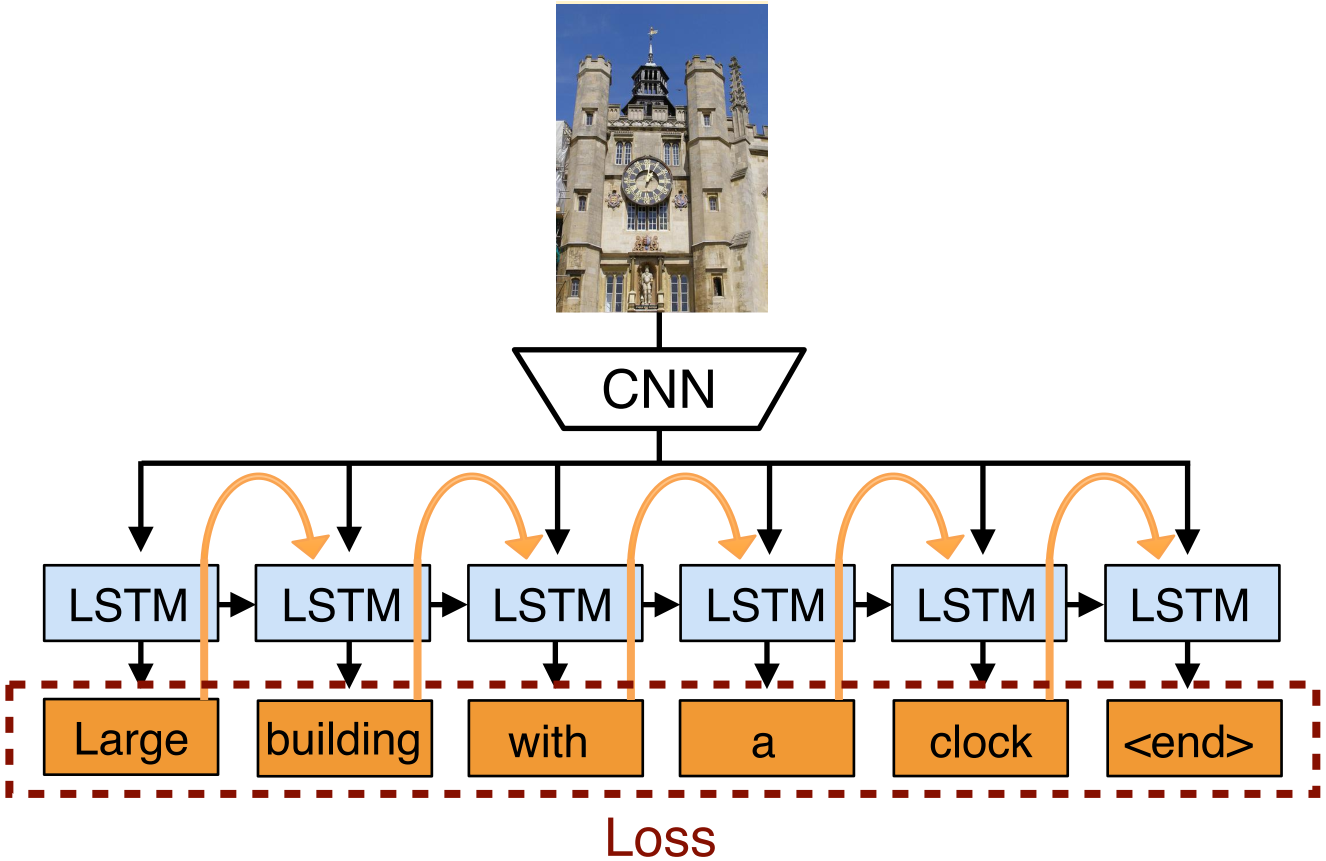

M. Malinowski, et. al. "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". NIPS'14




End-to-end, jointly trained architecture

# Neural Visual QA vs Neural Image Description

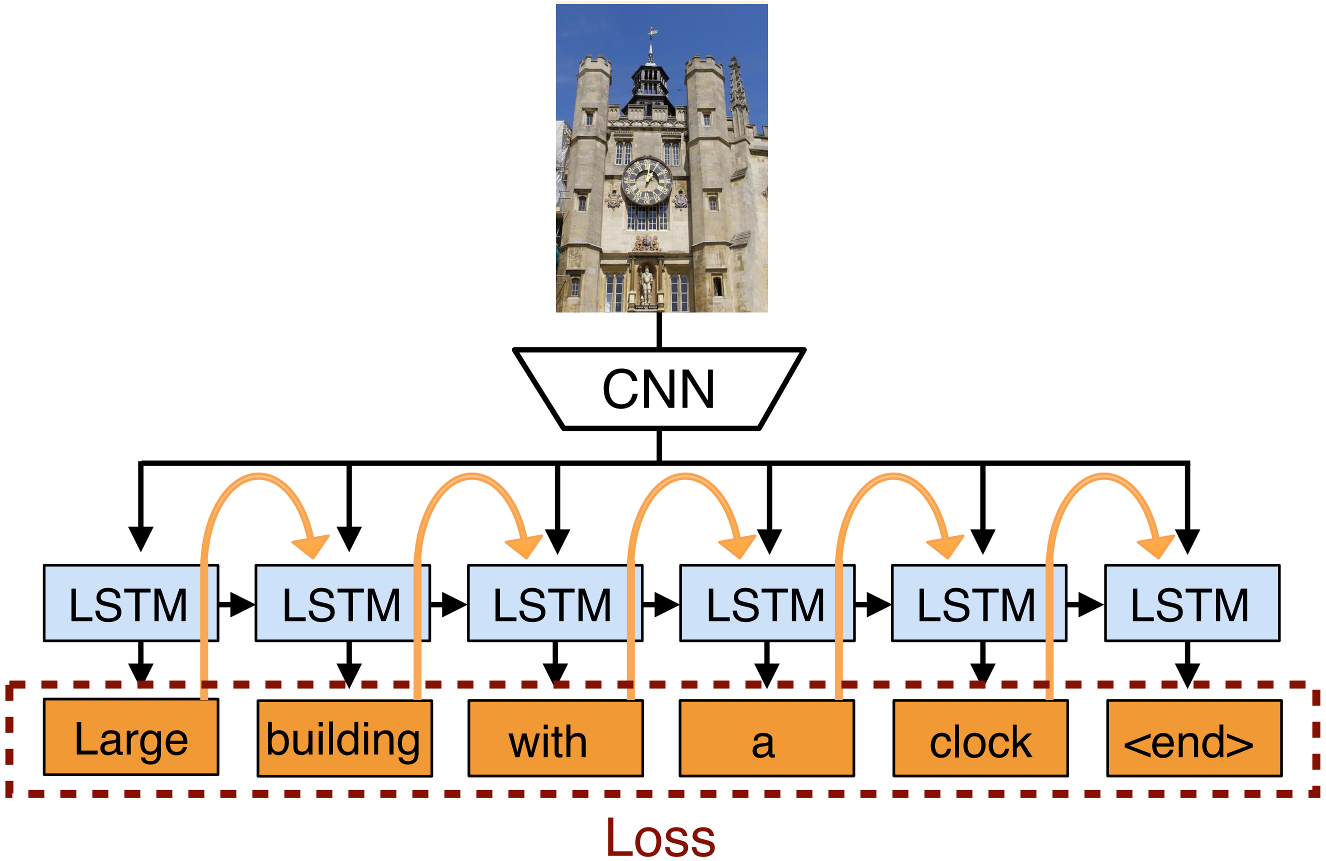
- Neural Image Description
  - Conditions on an image
  - Generates a description
    - Sequence of words
  - Loss at every step




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

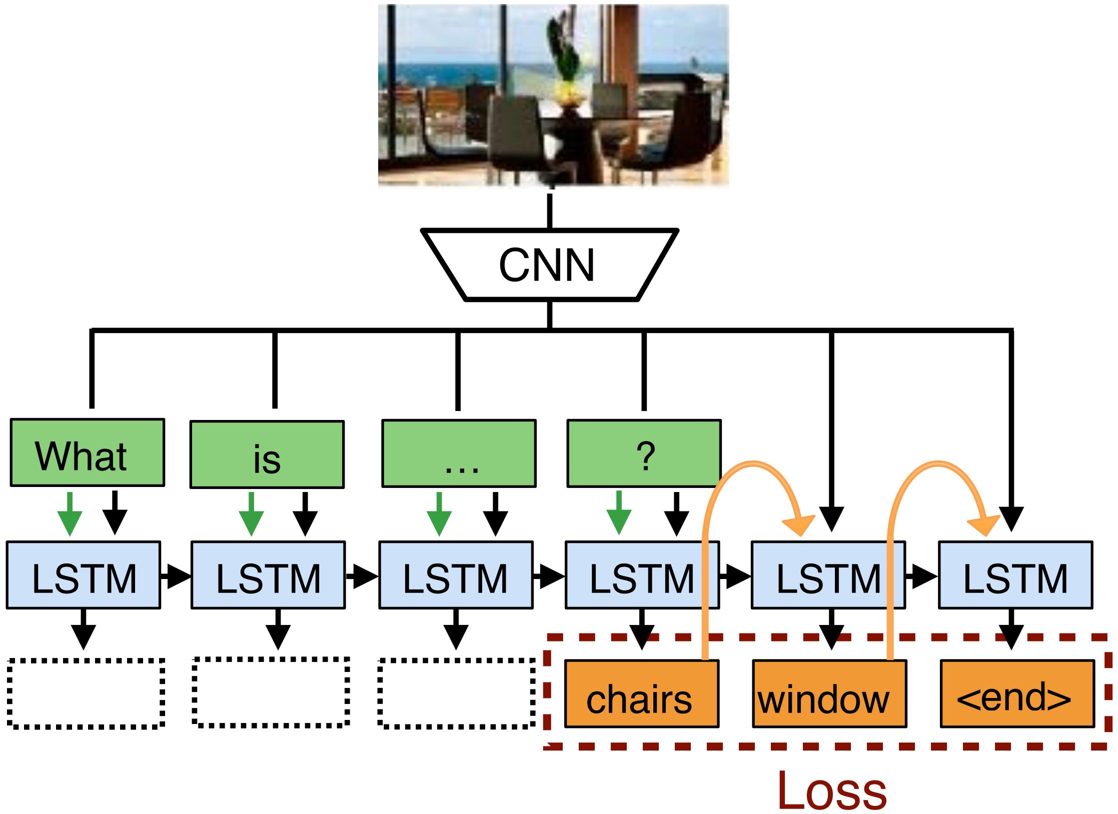
# Neural Visual QA vs Neural Image Description

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

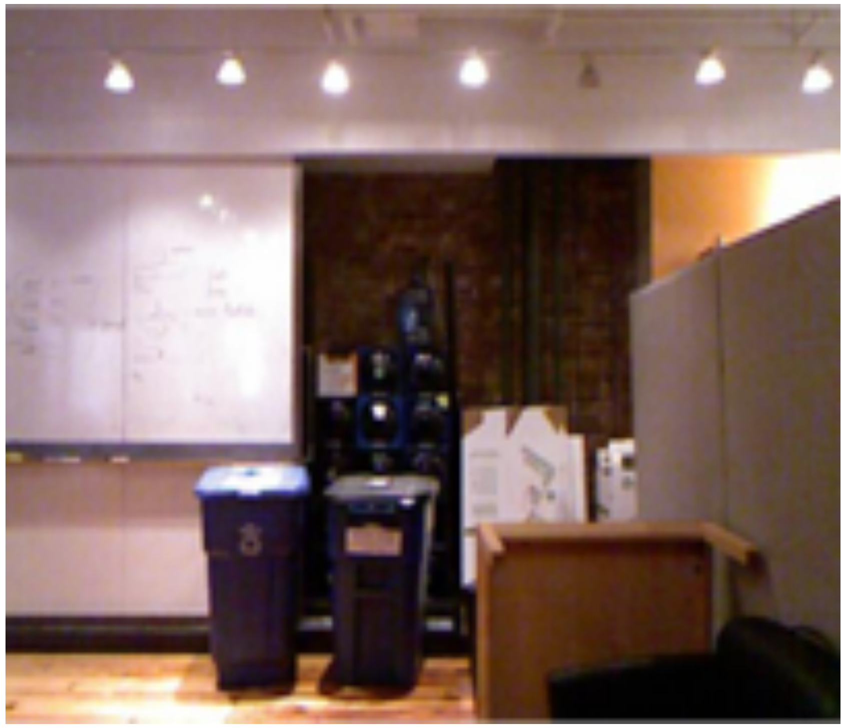



J. Donahue, et. al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”. CVPR15

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

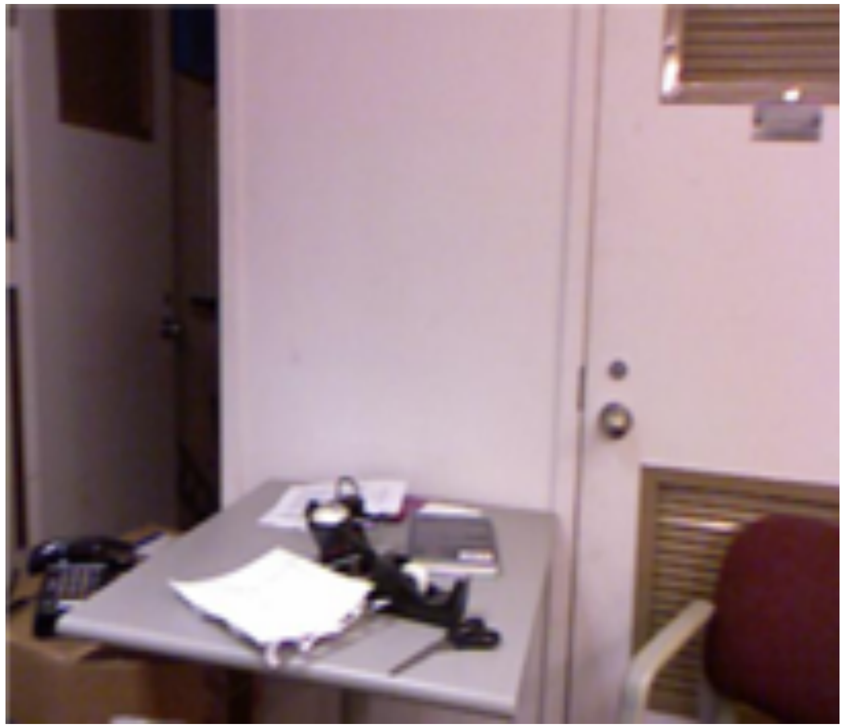


# Visual Turing Test: DAQUAR (NIPS'14)



**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

# 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**

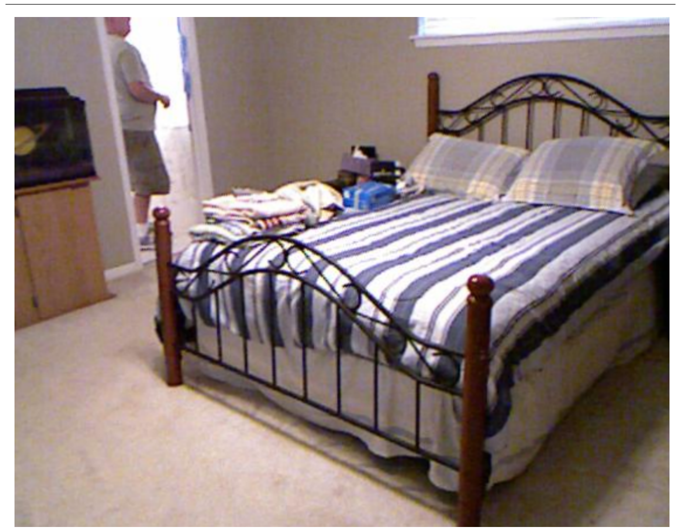
**What is the largest object?**
**bed**

# 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% |

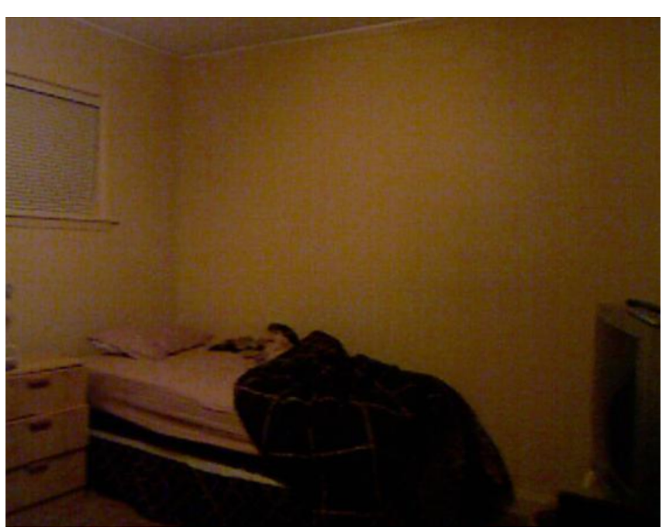
**What is on the refrigerator?**

**magnet, paper**



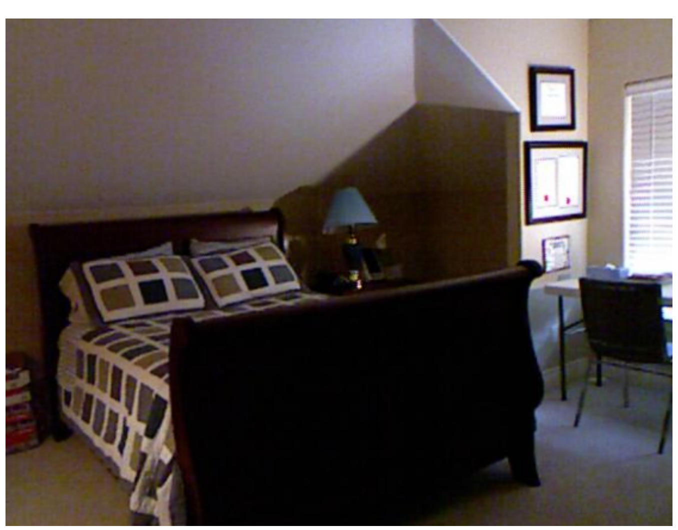
**What is the color of the comforter?**

**blue, white**



**How many drawers are there?**

**3**

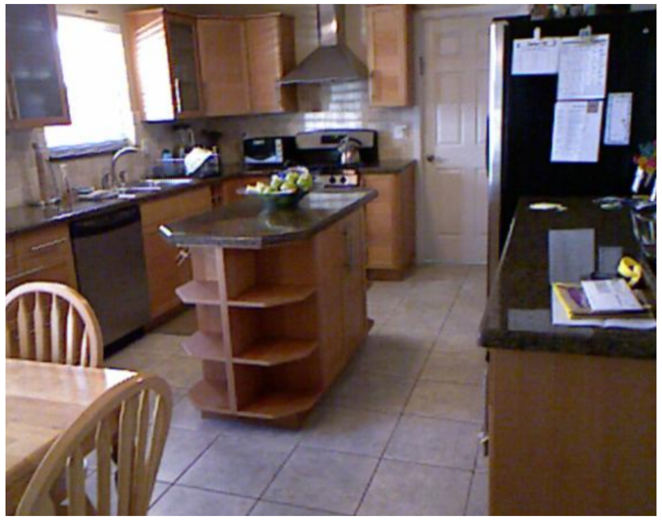


**What is the largest object?**

**bed**

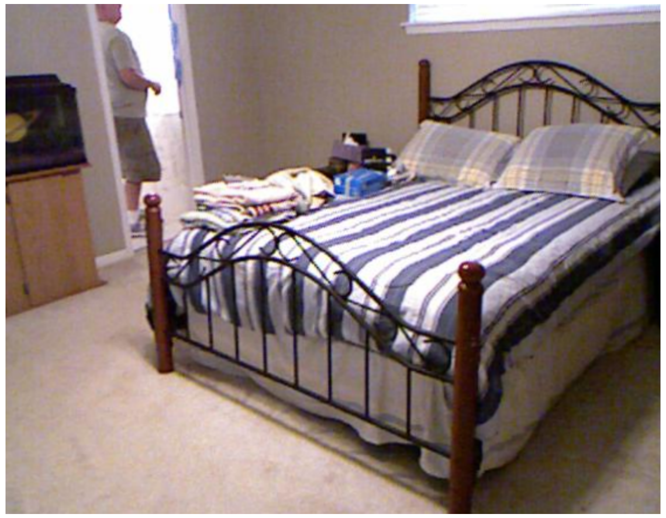
# Results on Full DAQUAR

| 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% |



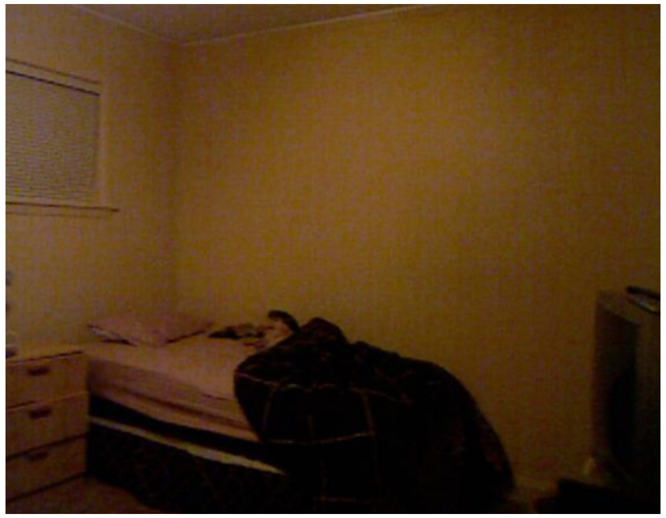
What is on the refrigerator?

magnet, paper



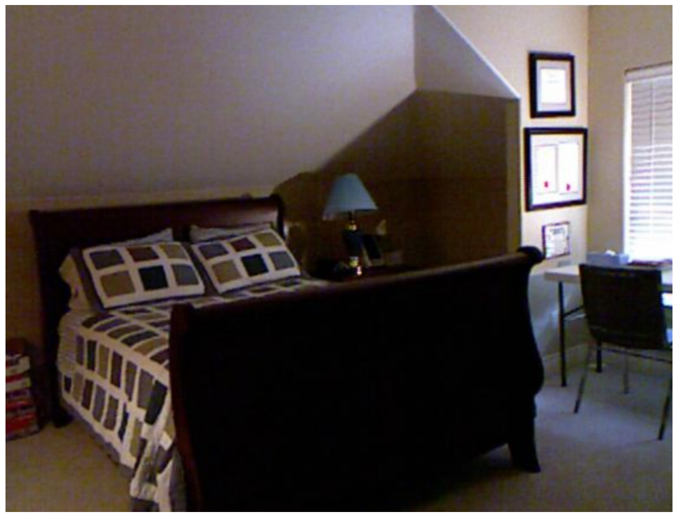
What is the color of the comforter?

blue, white



How many drawers are there?

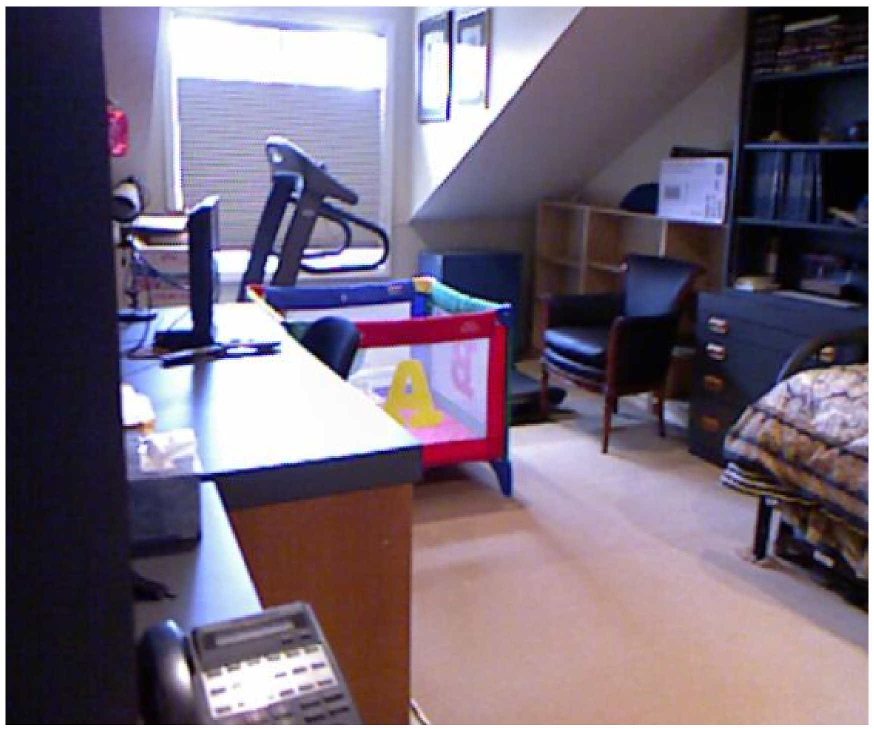
3



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**

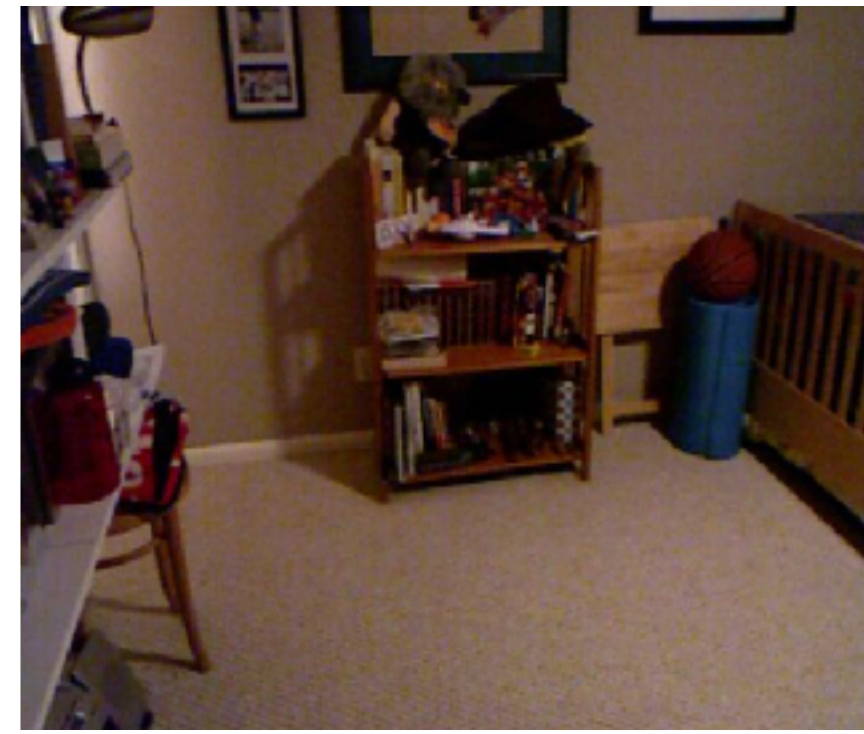
# Qualitative Results: Failure Cases

**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?**

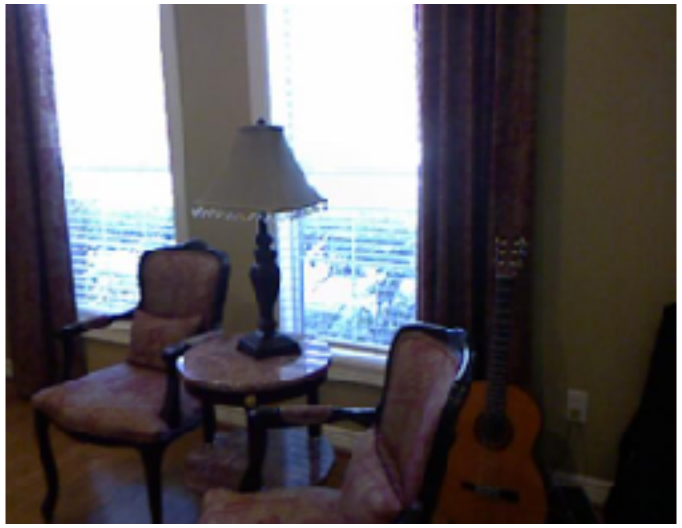
Human 1: **bed**
Human 2: **shelf**
Human 3: **bed**
Human 4: **bookshelf**

# Results on DAQUAR-Consensus

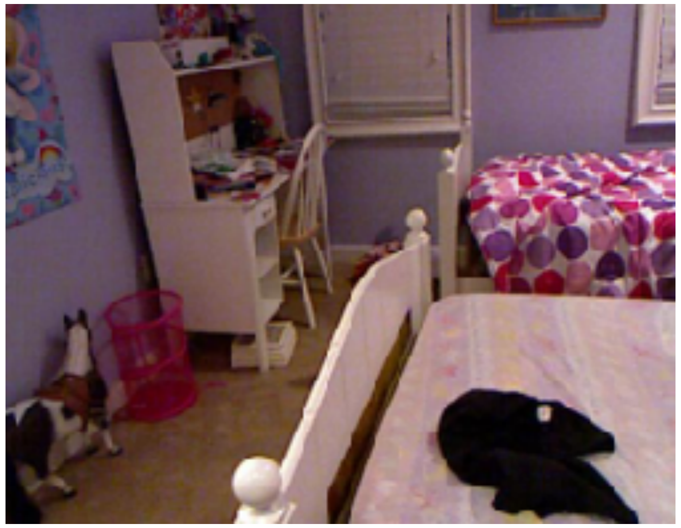
| Methods (Old Metric) | Accuracy | WUPS @0.9 |
|---|---|---|
| **Language Only (Our)** | 17.15% | 22.8% |
| **Vision + Language (Our)** | **19.43%** | **25.28%** |
| **Human performance (NIPS'14)** | 50.2% | 50.82% |

| 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

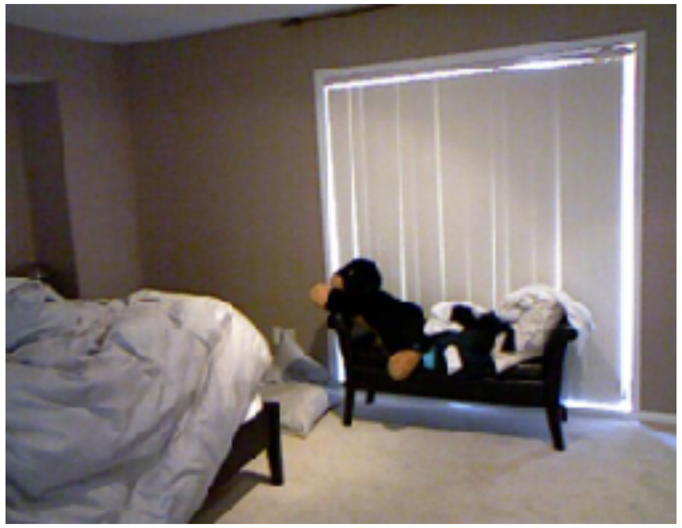


**What is in front of the curtain?**
**Model:** chair
**Human 1:** guitar
**Human 2:** chair



**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?**

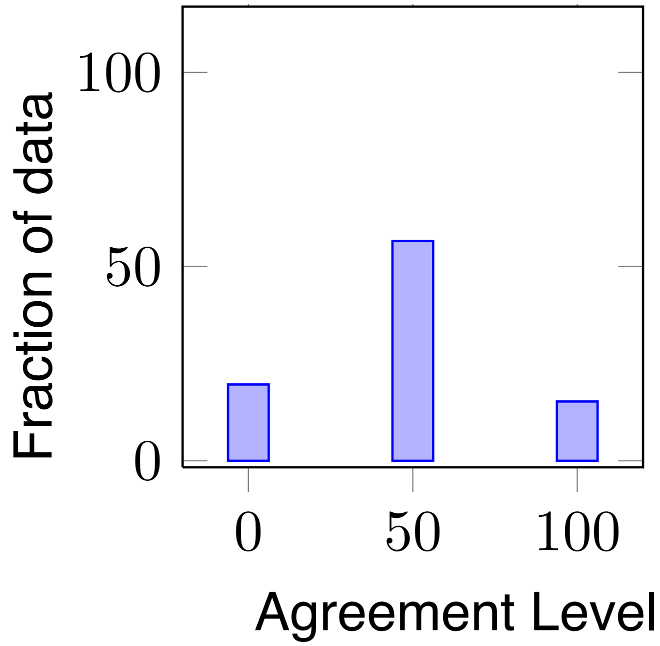
Human 1: **chair**
Human 2: **chair**
Human 3: **chair, bag**
Human 4: **wall**

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



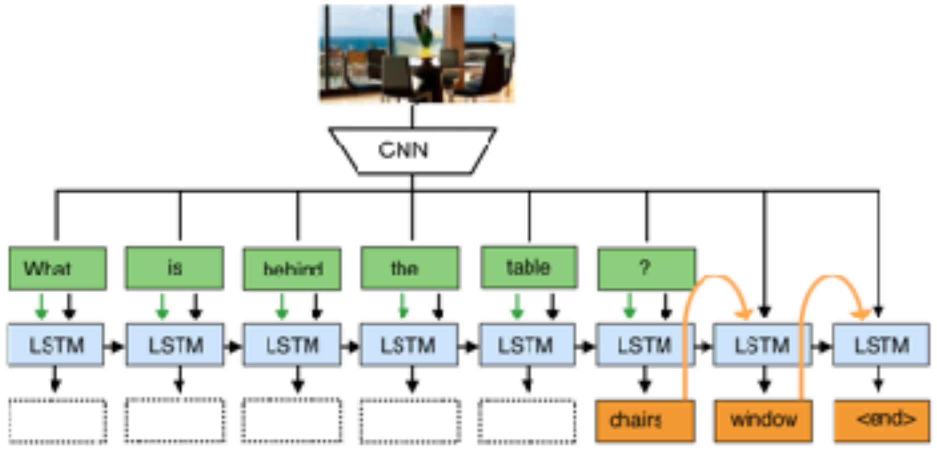
# Ask Your Neurons: A Deep Learning Approach to Visual Question Answering

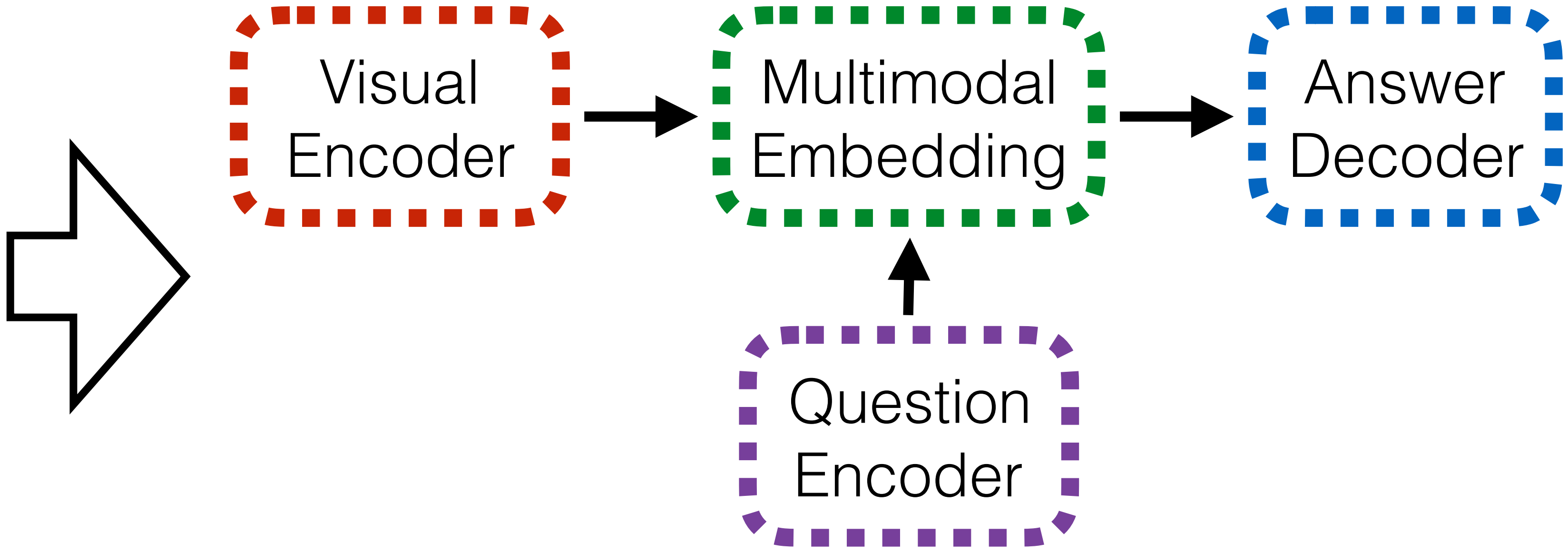
Mateusz Malinowski, Marcus Rohrbach, Mario Fritz

IJCV'17

# “Ask your neurons" again: Latest Results

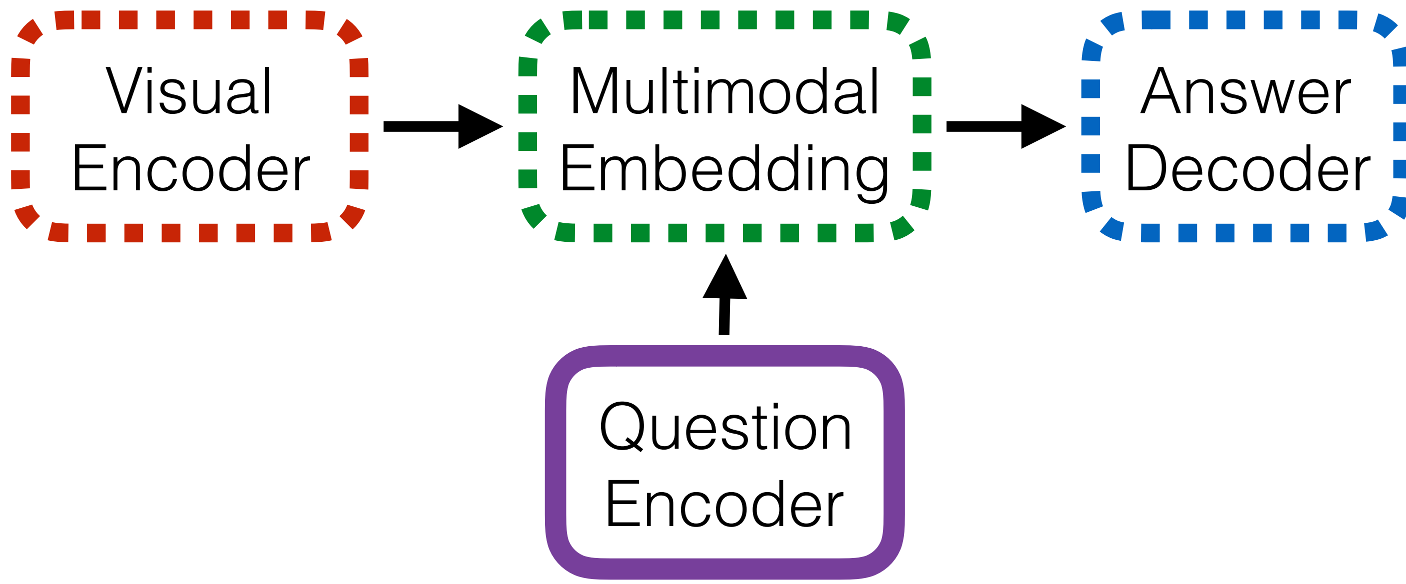
- Limit of global/holistic image representations?



# Results on VQA

| Question encoder | Word embedding learned | GLOVE |
|---|---|---|
| BOW | 47.41 | 47.91 |
| CNN | 48.26 | 48.53 |
| GRU | 47.60 | 48.11 |
| LSTM | 47.80 | **48.58** |

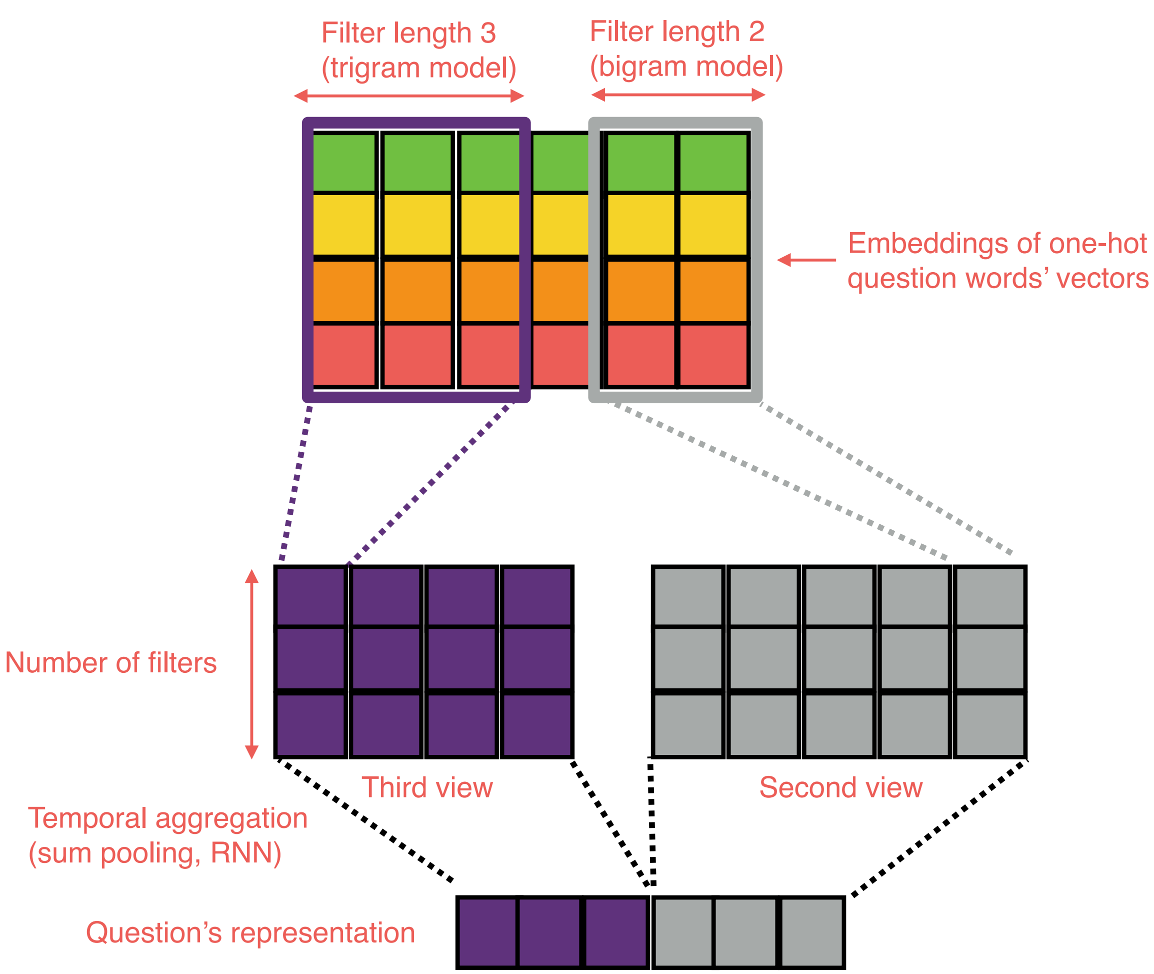



- Orderless models are very competitive
- GLOVE embedding improves results
- CNN and LSTM are often the best choices

# CNN Language Encoder

- Unifies vision and language model
- Fast (parallel) forward pass
- Relationship to n-gram models

What is behind the table?

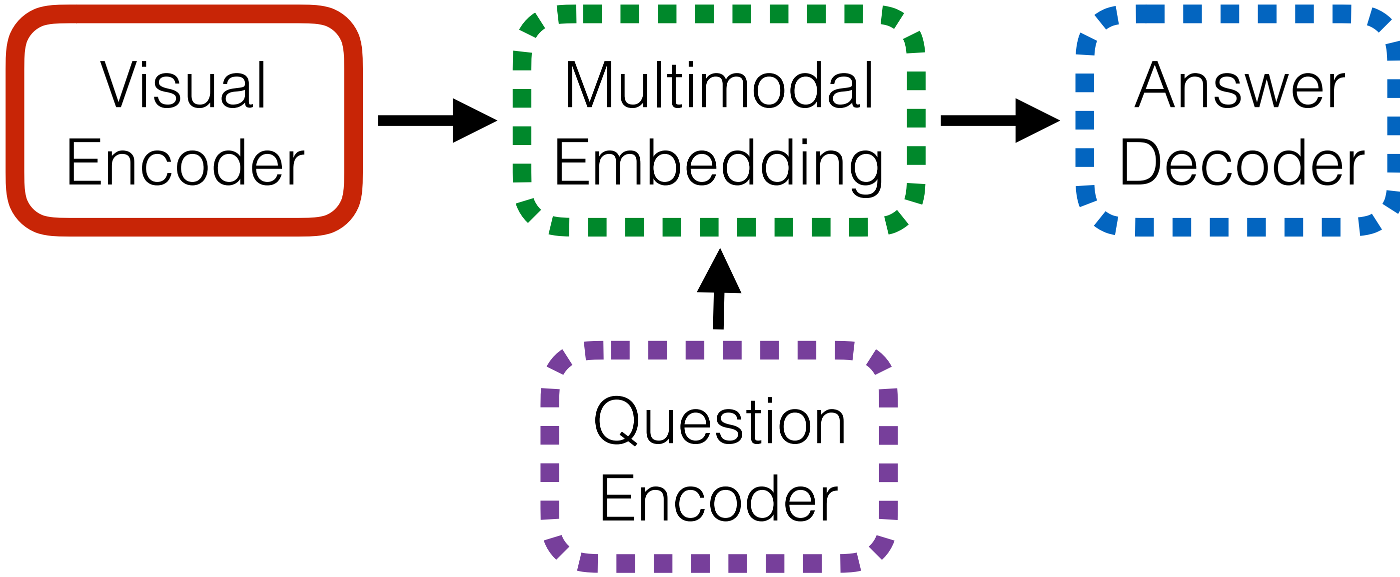



| kernel length $k$ | single view $= k$ | multi view $\leq k$ |
|---|---|---|
| 1 | 47.43 | 47.43 |
| 2 | 48.11 | 48.06 |
| 3 | **48.26** | 48.09 |
| 4 | **48.27** | 47.86 |

Kim'14 ; Kalchbrenner'14

# Results on VQA

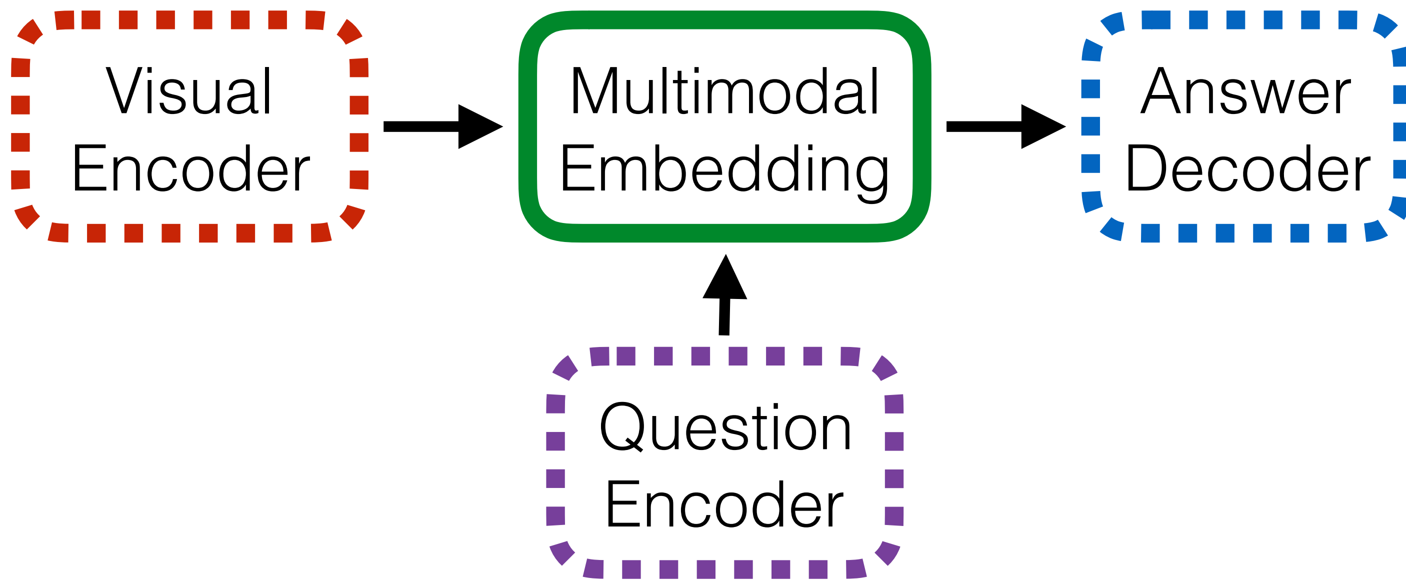
| Method | Accuracy |
|---|---|
| AlexNet | 53.69 |
| GoogLeNet | 54.52 |
| VGG-19 | 54.29 |
| ResNet-152 | **55.52** |




- Deeper and better recognition architectures improves the results on visual question answering
- We use LSTM as the question encoder

# Results on VQA

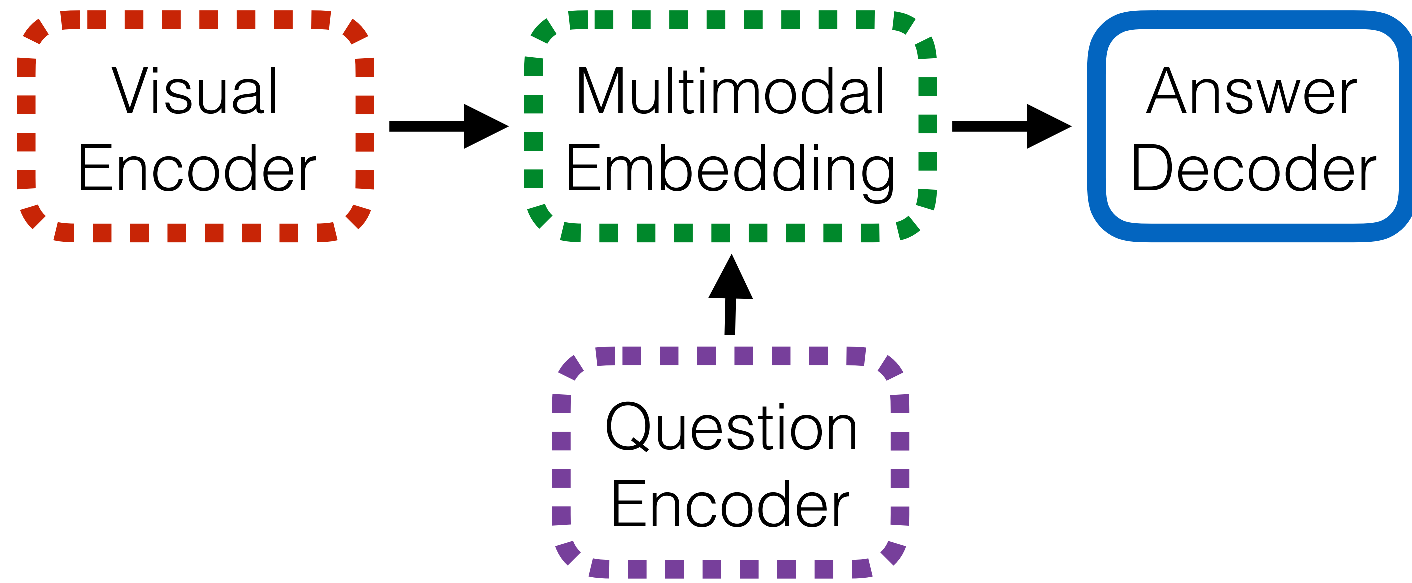
| | no norm | L2 norm |
|---|---|---|
| Concatenation | 47.21 | 52.39 |
| Summation | 40.67 | **53.27** |
| Piece-wise multiplication | 49.50 | 52.70 |




- Normalization of the visual features is important
  - We normalize by dividing by l2-norm of the feature vector
- Summation works the best

# Results on VQA

| Encoder | top frequent answers 1000 | 2000 | 3000 |
|---|---|---|---|
| BOW | 47.91 | 48.13 | 47.94 |
| CNN | 48.53 | 48.67 | 48.57 |
| LSTM | 48.58 | **48.86** | 48.65 |

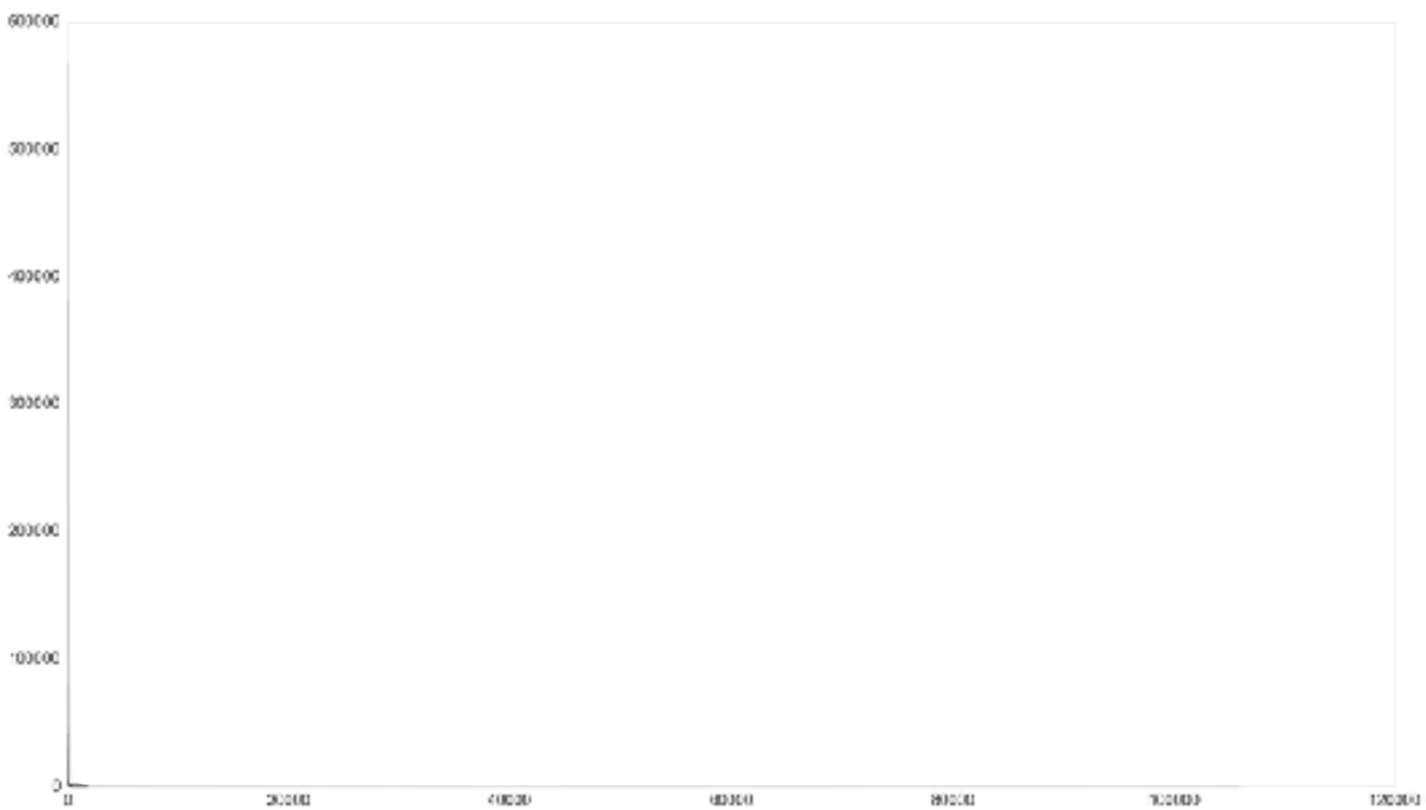



- The performance of the methods is dependent on the number of answers considered
- Many answers don't have enough examples for learning good representation
- Architectures often decide to model only top frequent answers

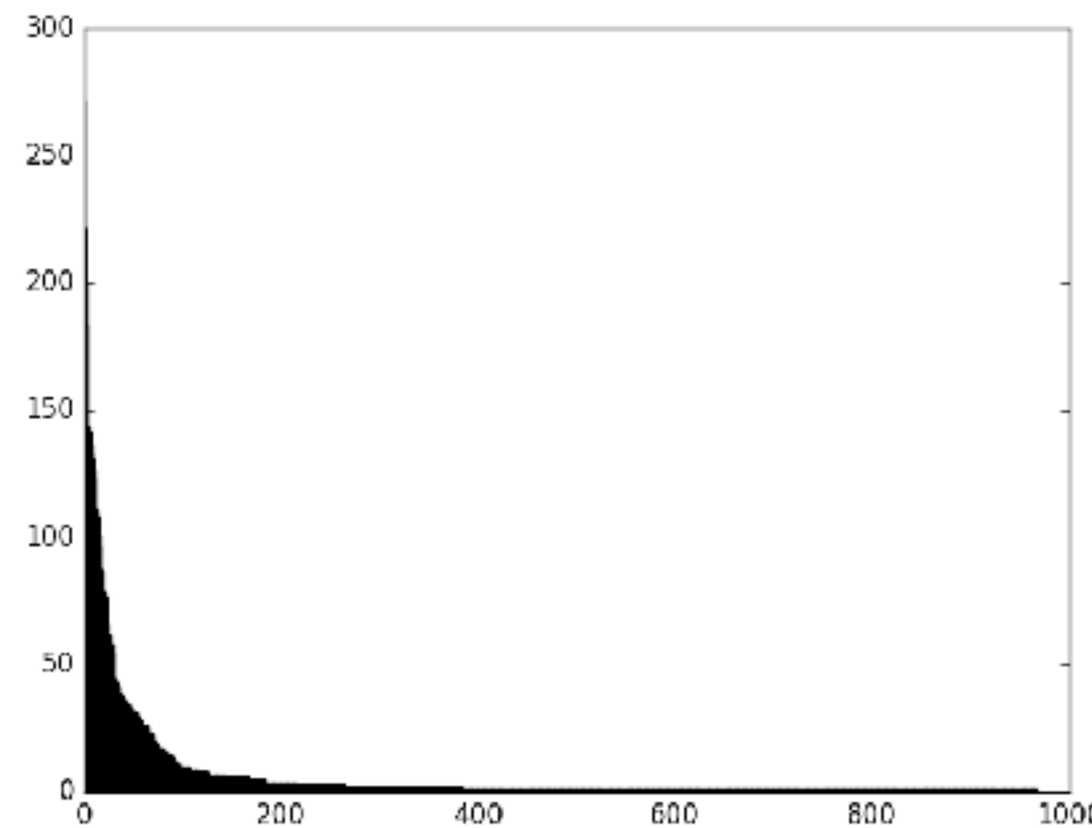
# Answer Statistic: Rare World Issue

- Highly unbalanced problem
- Strong results for method that focus on subset (e.g. restricted output space, single word answers)
- Issue of dataset? Issue of metric?

**VQA**



**DAQUAR**




- Interesting read:
  Simple Baseline for Visual Question Answering
  Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, Rob Fergus

# “Ask your neurons" again: How far goes global vision?

VQA

| | Test-dev Yes/No | Test-dev Number | Test-dev Other | Test-dev All | Test-standard Yes/No | Test-standard Number | Test-standard Other | Test-standard All |
|---|---|---|---|---|---|---|---|---|
| DMN+ (Xiong et al, 2016) | 80.5 | 36.8 | 48.3 | 60.3 | - | - | - | 60.4 |
| FDA (Ilievski et al, 2016) | 81.1 | 36.2 | 45.8 | 59.2 | - | - | - | 59.5 |
| AMA (Wu et al, 2016) | 81.0 | 38.4 | 45.2 | 59.2 | 81.1 | 37.1 | 45.8 | 59.4 |
| SAN(2, CNN) (Yang et al, 2015) | 79.3 | 36.6 | 46.1 | 58.7 | - | - | - | 58.9 |
| **Refined Ask Your Neurons** | 78.4 | 36.4 | 46.3 | 58.4 | 78.2 | 36.3 | 46.3 | 58.4 |
| SMem-VQA (Xu and Saenko, 2015) | 80.9 | 37.3 | 43.1 | 58.0 | 80.9 | 37.5 | 43.5 | 58.2 |
| D-NMN (Andreas et al, 2016a) | 80.5 | 37.4 | 43.1 | 57.9 | - | - | - | 58.0 |
| DPPnet (Noh et al, 2015) | 80.7 | 37.2 | 41.7 | 57.2 | 80.3 | 36.9 | 42.2 | 57.4 |
| iBOWIMG (Zhou et al, 2015) | 76.5 | 35.0 | 42.6 | 55.7 | 76.8 | 35.0 | 42.6 | 55.9 |
| LSTM Q+I (Antol et al, 2015) | 78.9 | 35.2 | 36.4 | 53.7 | - | - | - | 54.1 |
| Comp. Mem. (Jiang et al, 2015) | 78.3 | 35.9 | 34.5 | 52.7 | - | - | - | - |

global_vision

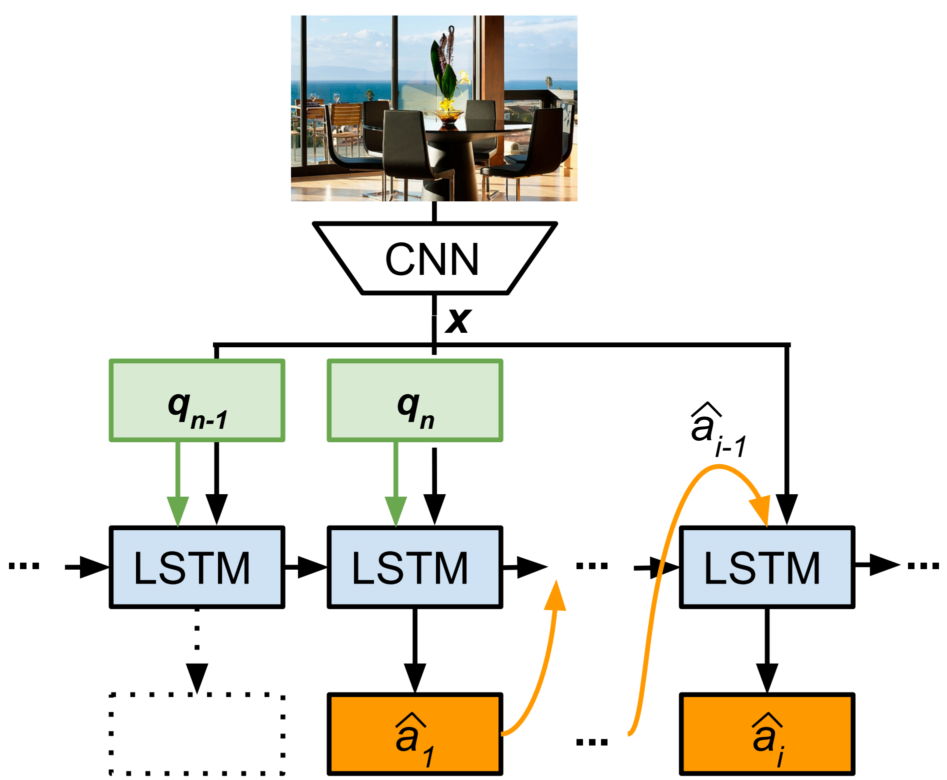
DAQUAR

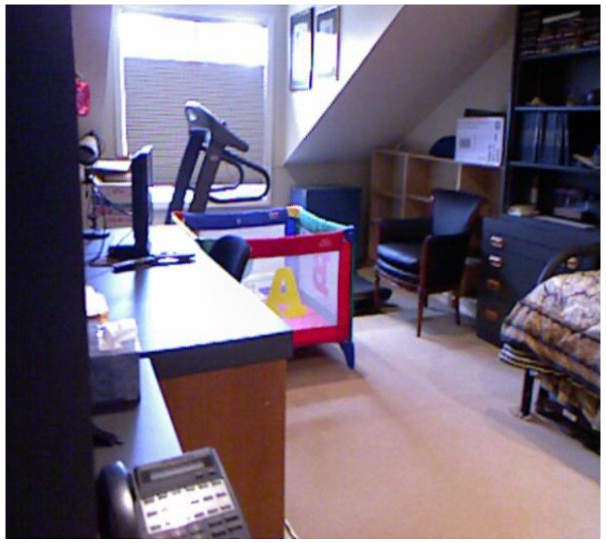
| | Accuracy on subset all | Accuracy on subset single word | WUPS@0.9 on subset all | WUPS@0.9 on subset single word | WUPS@0 on subset all | WUPS@0 on subset single word |
|---|---|---|---|---|---|---|
| Global | | | | | | |
| Ask Your Neurons | 19.43 | 21.67 | 25.28 | 27.99 | 62.00 | 65.11 |
| **Refined Ask Your Neurons** | 24.48 | 26.67 | 29.78 | 32.55 | 62.80 | 66.25 |
| **Refined Ask Your Neurons** ∗ | 25.74 | 27.26 | 31.00 | 33.25 | 63.14 | 66.79 |
| IMG-CNN (Ma et al, 2016) | 21.47 | 24.49 | 27.15 | 30.47 | 59.44 | 66.08 |
| Attention | | | | | | |
| SAN (2, CNN) (Yang et al, 2015) | - | 29.30 | - | 35.10 | - | 68.60 |
| DMN+ (Xiong et al, 2016) | - | 28.79 | - | - | - | - |
| ABC-CNN (Chen et al, 2015) | - | 25.37 | - | 31.35 | - | 65.89 |
| Comp. Mem. (Jiang et al, 2015) | 24.37 | - | 29.77 | - | 62.73 | - |

Malinowski, Rohrbach, Fritz: Arxiv’16 “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering”

# 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



**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

# Spectrum between Symbolic and Vector-based Approaches

**classic/symbolic (NIPS'14)**

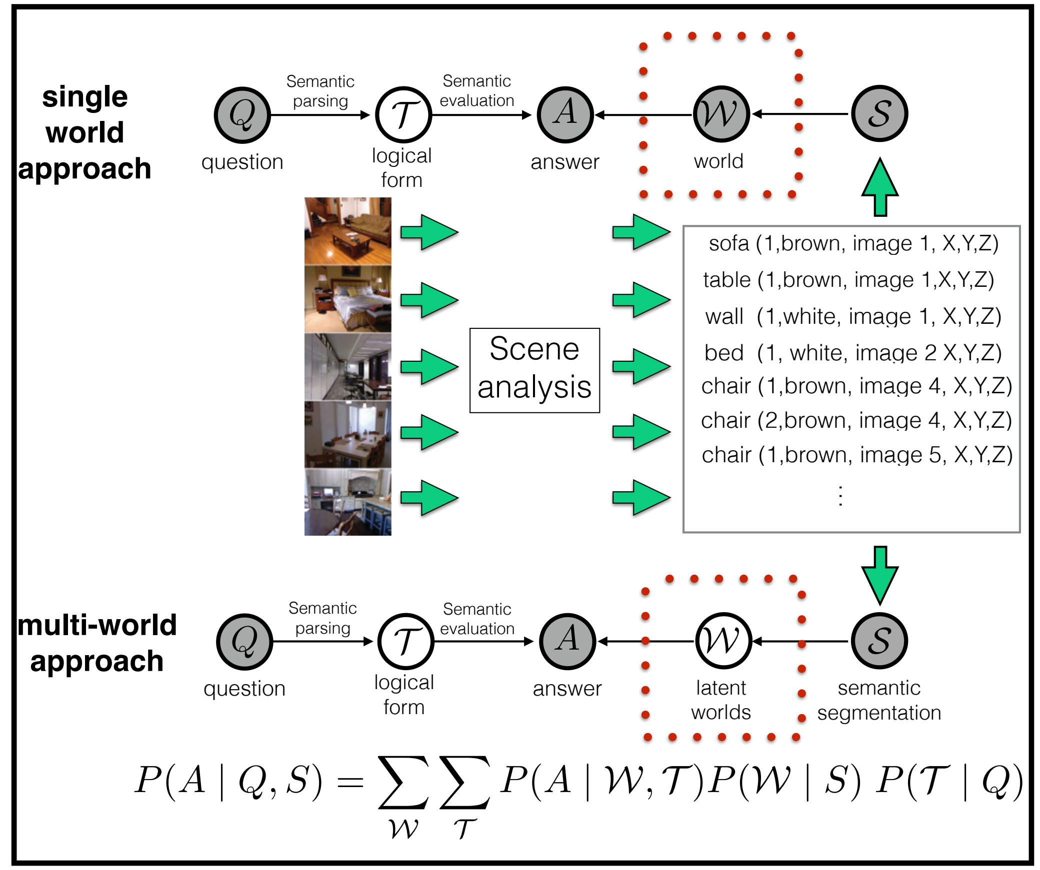
- symbolic representation
- high level of introspection
- disjoint modules
- "detailed" visual representation
- limit coverage of concepts; semantic parsing can be fragile

**deep learning (ICCV'15)**

- vector representation
- nebulous - but some hope
- end to end learning
- global CNN representation
- continuous embedding of concepts

# Methods

Detectors Classes




Ours, NIPS’14

Explicit Vision

Vectors / Neurons

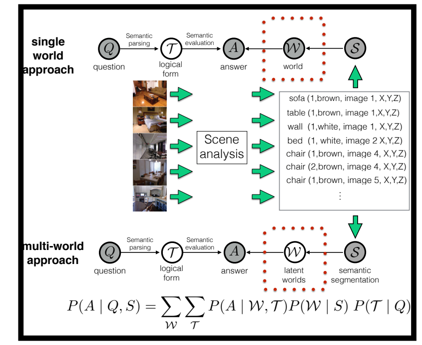
Vector / Neurons

Explicit Language

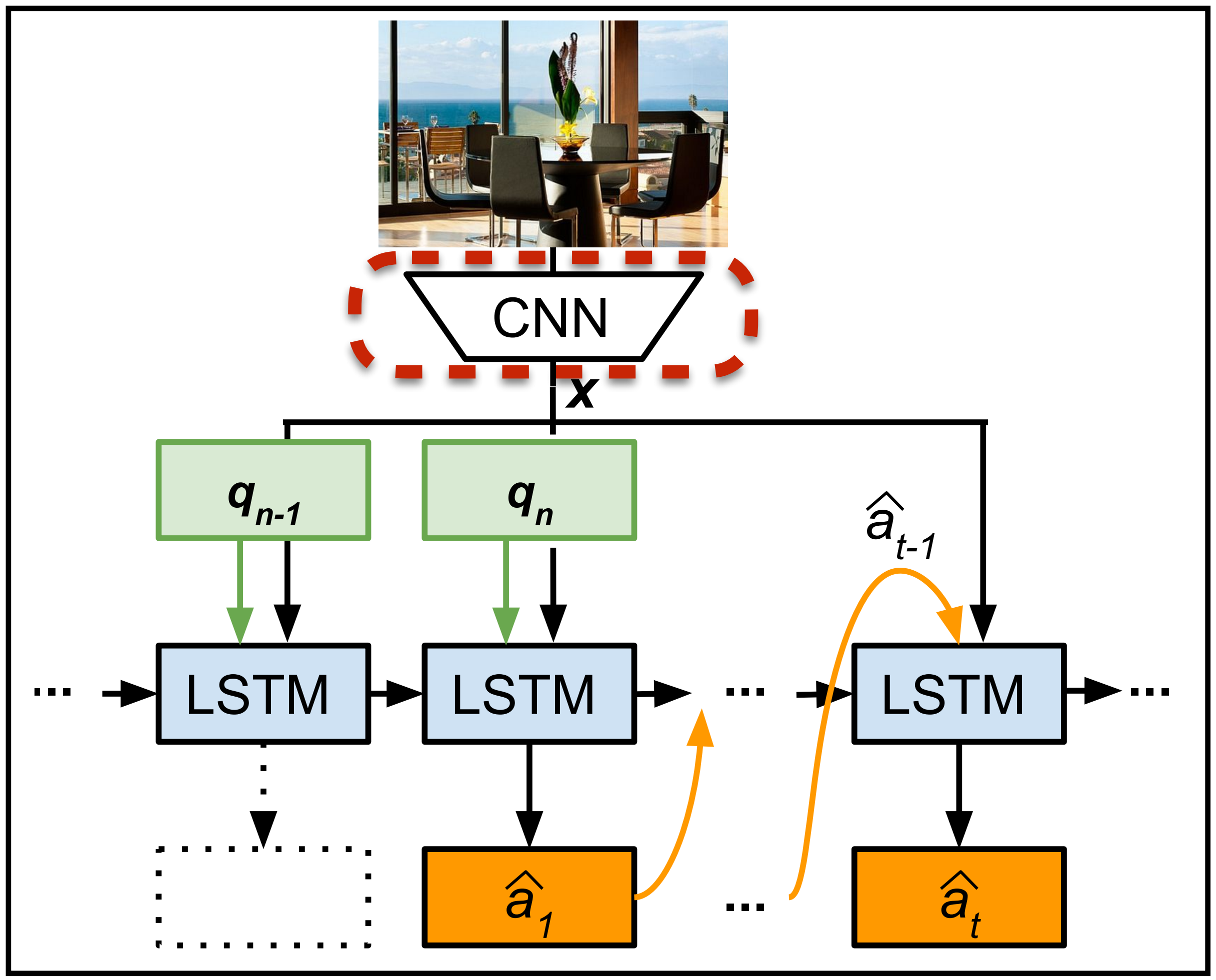
Syntax / Semantics

# Methods

Ours, ICCV'15
Antol et. al. ICCV'15
Ren et.al. NIPS'15
Gao et. al. NIPS'15
Ma et. al. AAAI 2016




NIPS'14




Detectors Classes

Explicit Vision

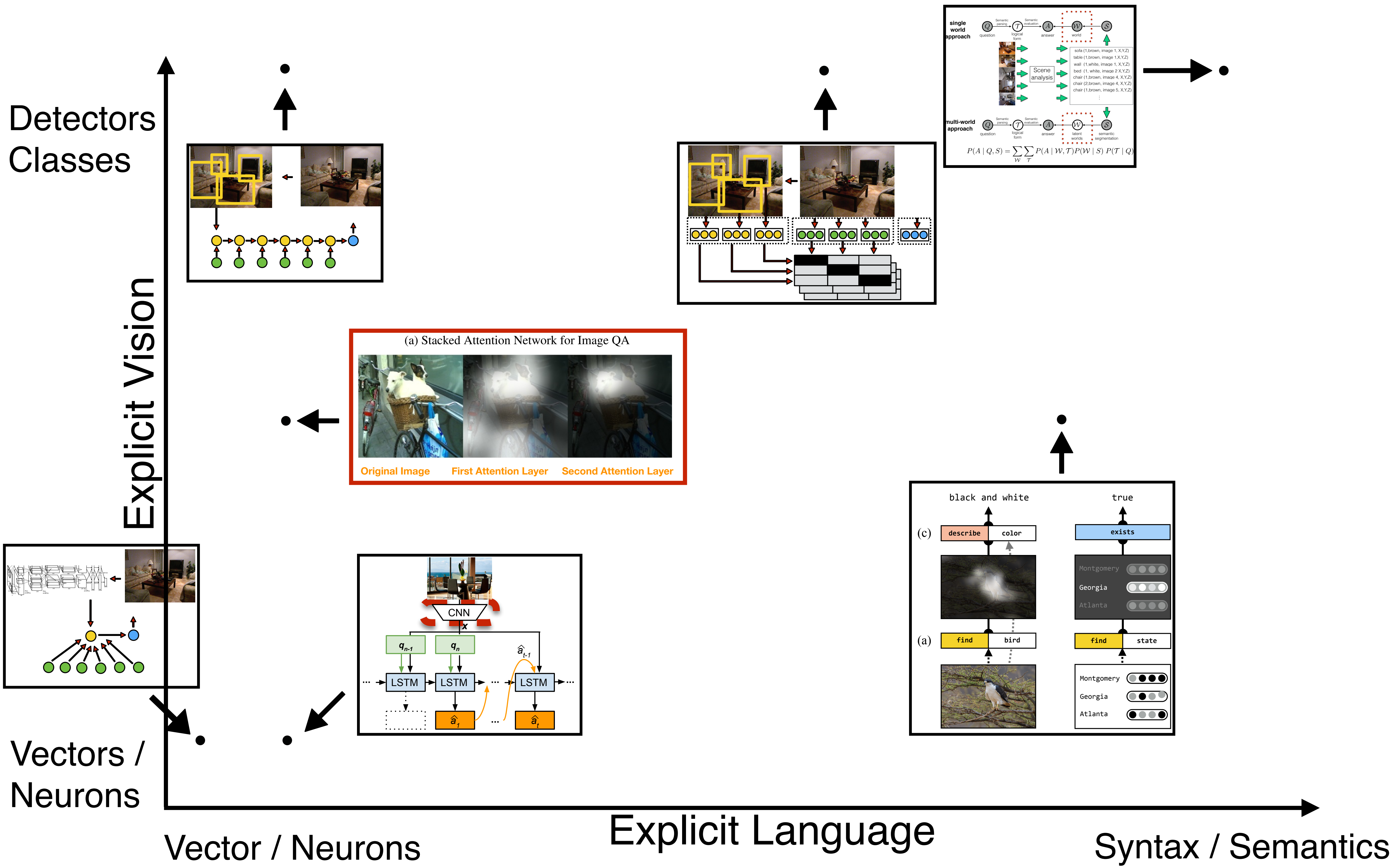
Vectors / Neurons

Vector / Neurons

Explicit Language

Syntax / Semantics

# Methods



Detectors
Classes
Explicit Vision
Vectors /
Neurons
single world approach
multi-world approach
Scene analysis
(a) Stacked Attention Network for Image QA
Original Image
First Attention Layer
Second Attention Layer
CNN
LSTM
LSTM
LSTM
black and white
true
(c)
describe
color
exists
Montgomery
Georgia
Atlanta
(a)
find
bird
find
state
Montgomery
Georgia
Atlanta
Vector / Neurons
Explicit Language
Syntax / Semantics

# Recent Related Work

- ## Symbolic 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

- ## Neural-based Approaches

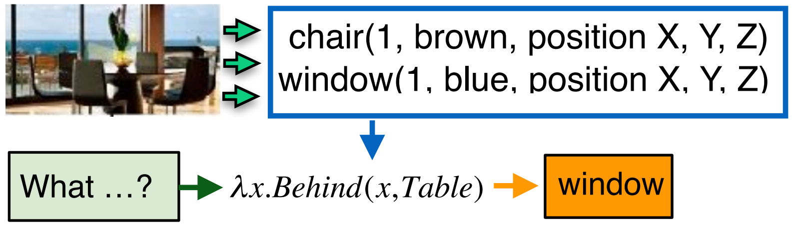
  M. Ren et. al. Image QA. NIPS'15
  H. Gao. et. al. Are You Talking to a Machine? NIPS'15
  L. Ma et. al. Learning to Answer Questions From Images. arXiv'15

- ## Attention-based Approaches

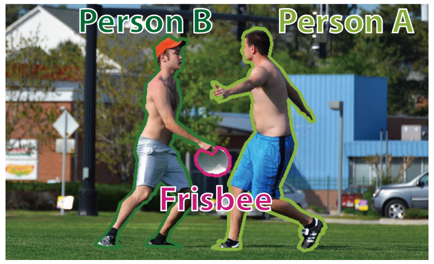
  Z. Yang. et. al. Stacked Attention Networks. arXiv'15
  Y. Zhu et. al. Visual7W. arXiv'15
  J. Andres et. al. Deep Compositional QA. arXiv'15
  H. Xu et. al. Ask, Attend and Answer. arXiv'15
  K. Chen et. al. ABC-CNN. arXiv'15
  K. J. Shih et. al. Where To Look. arXiv'15

- ## Hybrid Approaches

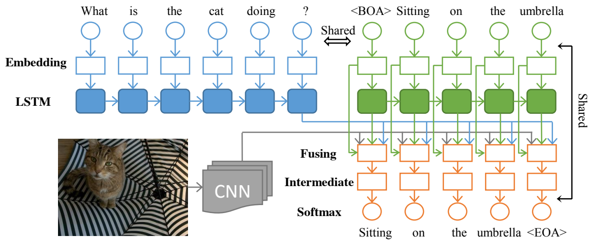
  H. Noh et al. Dynamic Parameter Prediction. arXiv'15
  J. Andres et al. Deep Compositional QA. arXiv'15

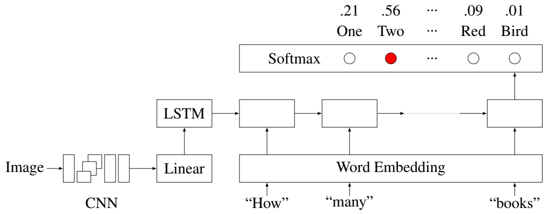


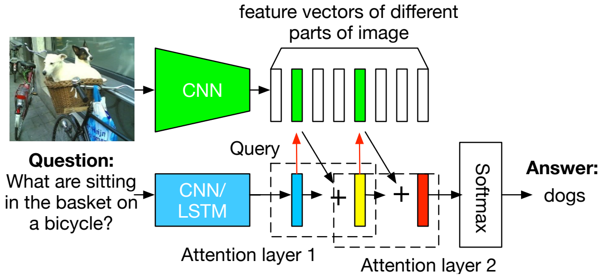

What is the mustache made of?



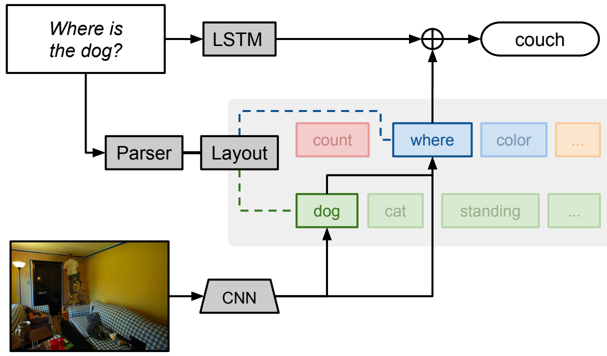

Person A is …







# Datasets

- DAQUAR (NIPS'14, ours)
  - 1449 indoor images
  - ~12.5k question-answer pairs
  - ~600 answer words (output space)
  - Many words answers (set of objects)
- DAQUAR-Reduced (NIPS'14, ours)
  - A subset of DAQUAR with 37 answer words
- Toronto COCO-QA (NIPS'15, M. Ren et. al.)
  - ~123k images
  - ~118k question-answer pairs (semi-synthetic)
  - Only one-word answers
- VQA (ICCV'15, S. Antol et. al.)
  - ~205k images
  - ~614k questions with 10 answers per question
  - Open-ended answers (in practice ignored)
- Visual Madlibs (ICCV'15)
  - Filling in blanks

What is on the refrigerator?

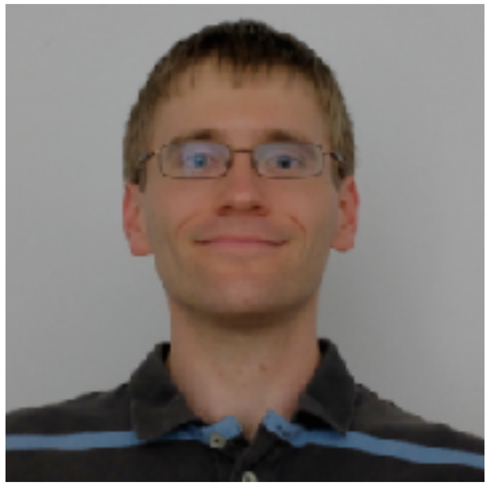
How many leftover donuts is the red bicycle holding?

What is the mustache made of?

VQA Visual Question Answering

# Overview of Challenge


Aishwarya Agrawal
(Virginia Tech)



Stanislaw Antol
(Virginia Tech)

Larry Zitnick
(Facebook AI Research)

Dhruv Batra
(Virginia Tech)

Devi Parikh
(Virginia Tech)


**http://www.visualqa.org**

# Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

# VQA Task

# VQA Task


What is the mustache made of?

# VQA Task

What is the mustache made of?

AI System

# VQA Task


What is the mustache made of?

AI System

bananas

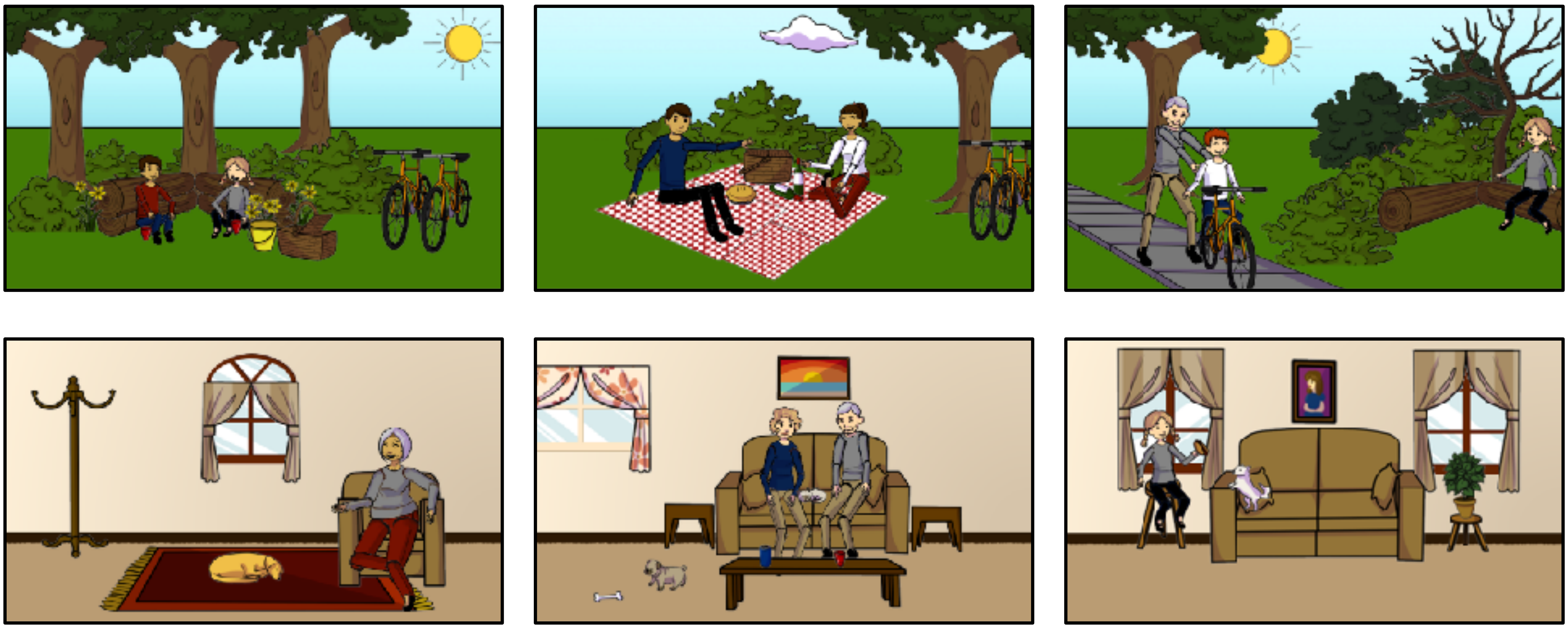
# Real images (from COCO)






Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in COntext.” ECCV 2014.
http://mscoco.org/

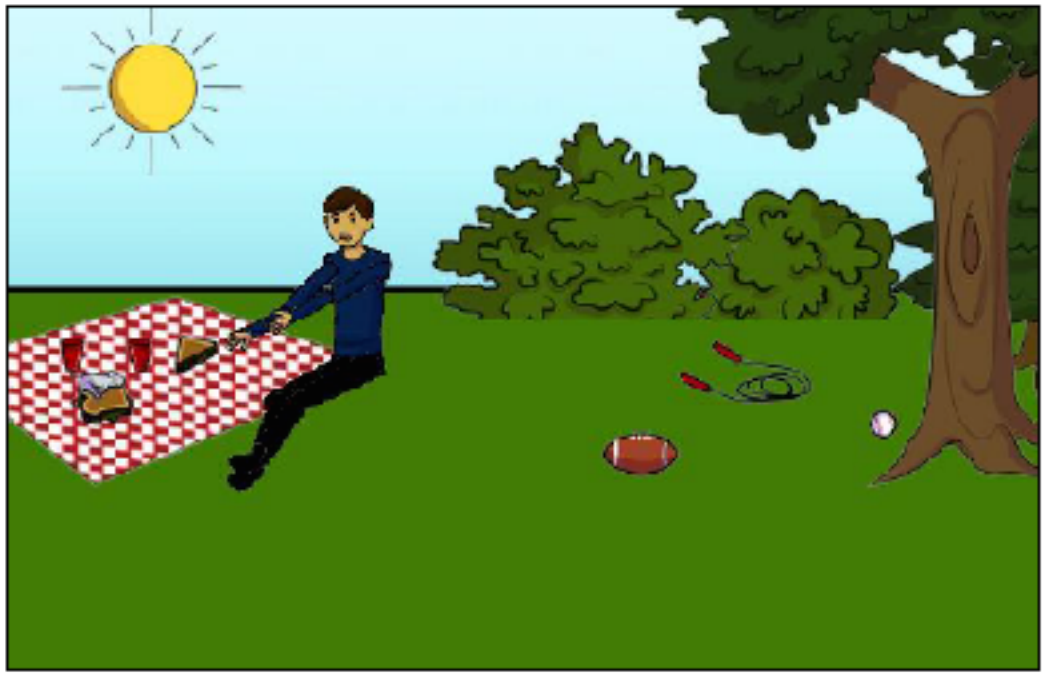
and abstract scenes.



# VQA Dataset


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

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



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

# Dataset Stats

- >250K images (COCO + 50K Abstract Scenes)
- >750K questions (3 per image)
- ~10M answers (10 w/ image + 3 w/o image)

# Two modalities of answering

- Open Ended
- Multiple Choice
  - 1 correct answer
  - 3 plausible choices
  - 10 most popular answers
  - Rest random answers

# Accuracy Metric

$$\mathrm{Acc}(\textcolor{red}{ans}) = \min\left\{\frac{\#\text{humans that said } \textcolor{red}{ans}}{3}, 1\right\}$$

1940. COCO_train2014_000000012015

Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

(1) television
(2) tv
(3) tv
(4) tv
(5) television
(6) television
(7) television
(8) tv
(9) tv
(10) television

Q: How old is this TV?

Ground Truth Answers:

(1) 20 years
(2) 35
(3) old
(4) more than thirty years old
(5) old
(6) old
(7) 80 s
(8) 30 years
(9) 15 years
(10) very old

Q: Is this TV upside-down?

Ground Truth Answers:

(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

# Human Accuracy (Real)

| | Overall | Yes/No | Number | Other |
|---|---|---|---|---|
| Open Ended | 83.30 | 95.77 | 83.39 | 72.67 |

# Human Accuracy (Real)

| | Overall | Yes/No | Number | Other |
|---|---|---|---|---|
| Open Ended | 83.30 | 95.77 | 83.39 | 72.67 |
| Multiple Choice | 91.54 | 97.40 | 86.97 | 87.91 |

# Human Accuracy (Abstract)

| | Overall | Yes/No | Number | Other |
|---|---|---|---|---|
| Open Ended | 87.49 | 95.96 | 95.04 | 75.33 |

# Human Accuracy (Abstract)

| | Overall | Yes/No | Number | Other |
|---|---|---|---|---|
| Open Ended | 87.49 | 95.96 | 95.04 | 75.33 |
| Multiple Choice | 93.57 | 97.78 | 96.71 | 88.73 |

NEW YORK UNIVERSITY

Facebook AI Research

# End-To-End Memory Networks

Sainbayar Sukhbaatar[1], Arthur Szlam[2], Jason Weston[2] and Rob Fergus[2]

[1]New York University [2]Facebook AI Research

# Motivation

- Good models exist for some data structures
  - RNN for temporal structure
  - ConvNet for spatial structure
- But we still struggle with some type of dependencies
  - out-of-order access
  - long-term dependency
  - unordered set

# Ex) Question & Answering on story

Sam moved to the garden.
Mary left the milk.
John left the football.
Daniel moved to the garden.
Sam went to the kitchen.
Sandra moved to the hallway.
Mary moved to the hallway.
Mary left the milk.
Sam drops the apple there.

out-of-order

Q: Where was the apple after the garden?

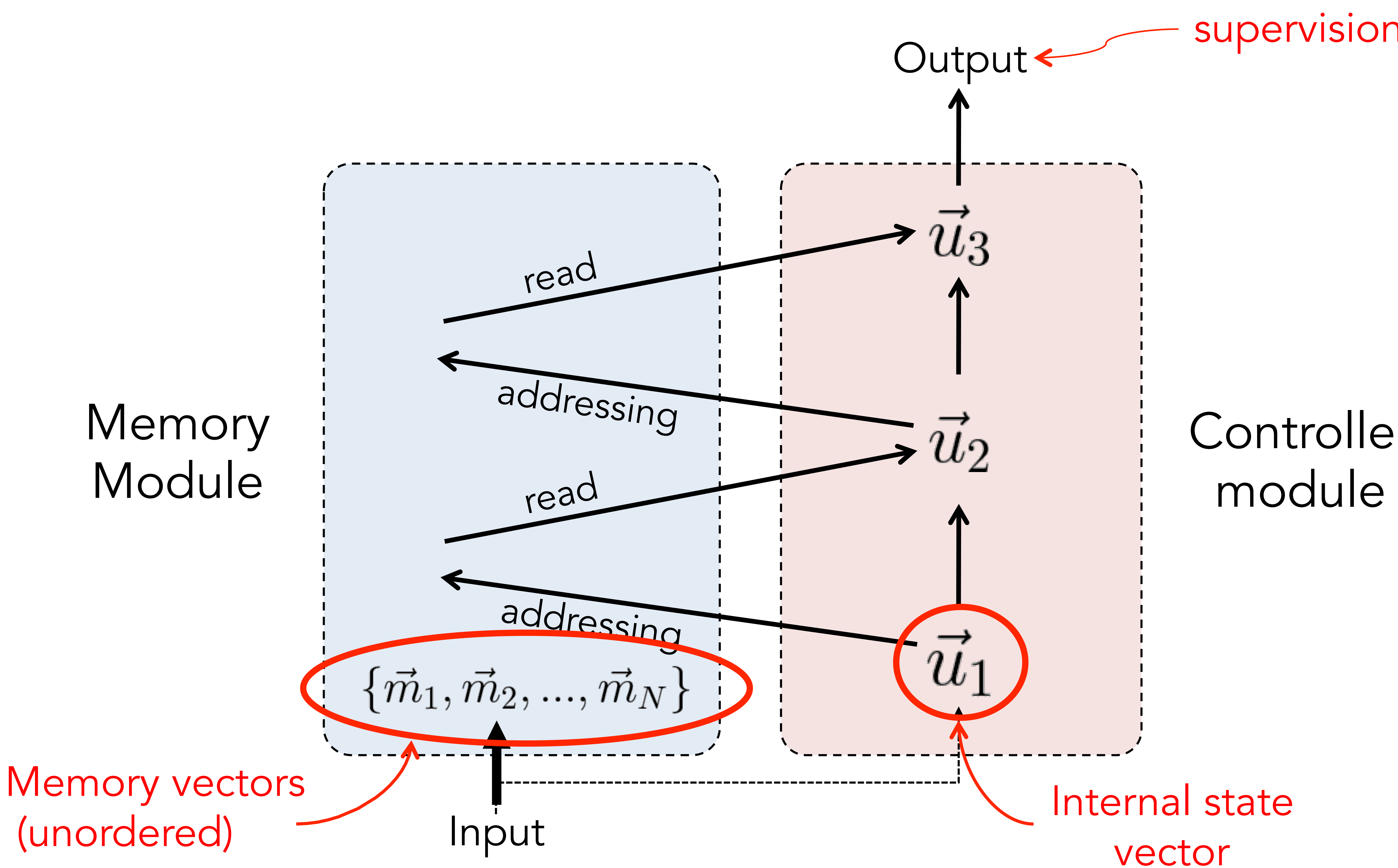
# Overview

- We propose a neural network model with external memory
  - Reads from memory with **soft attention**
  - Performs **multiple lookups** (hops) on memory
  - End-to-end training with **backpropagation**

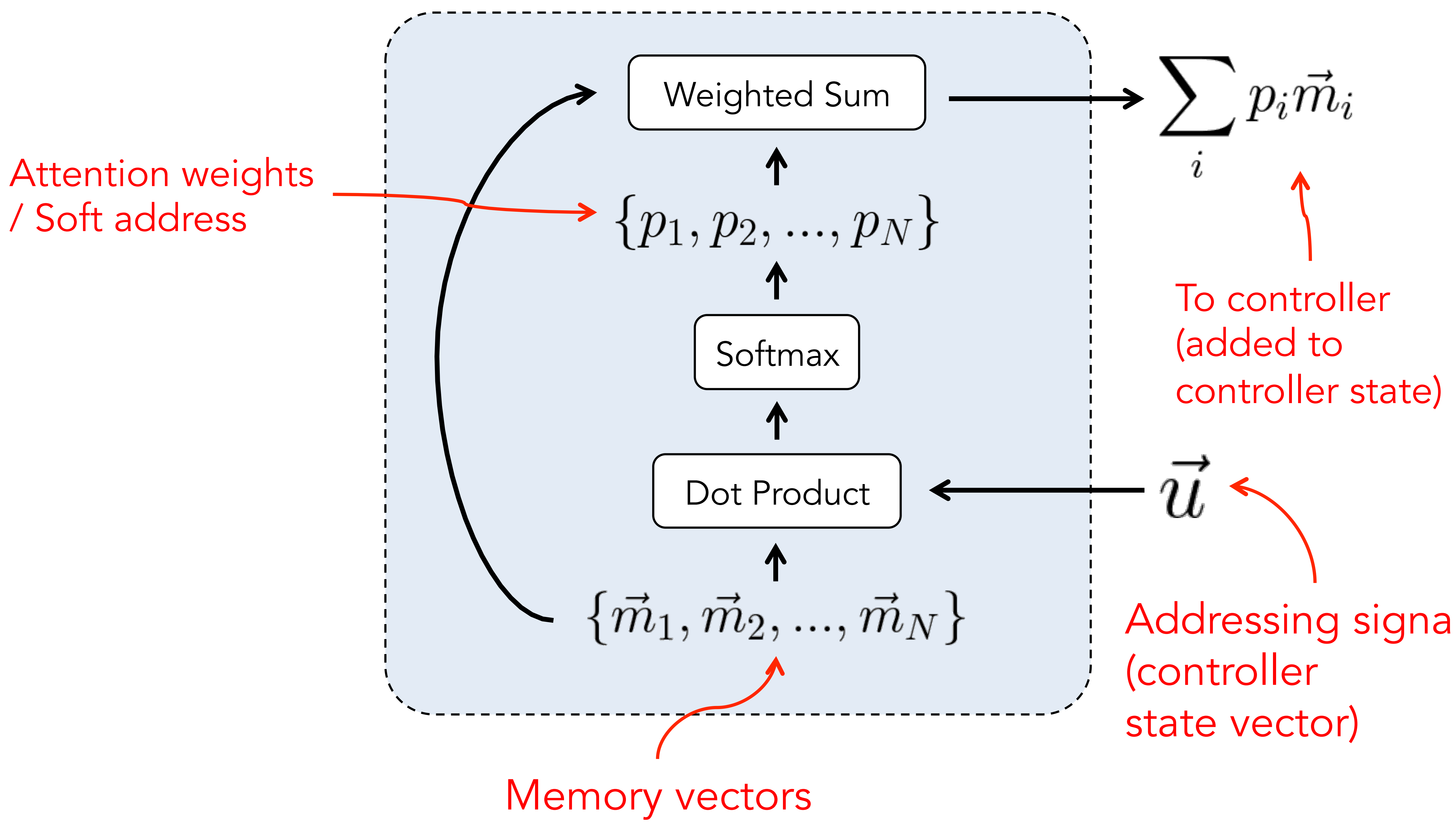
- **End-to-end Memory Network (MemN2N)**

- It is based on “Memory Networks” by [Weston, Chopra & Bordes ICLR 2015]
  - Hard attention
  - requires explicit supervision of attention during training
  - Only feasible for simple tasks
  - Severely limits application of the model
- MemN2N is **soft** attention version
- Only need supervision on the final output

# MemN2N architecture



# Memory Module



# Memory Vectors

E.g.) constructing memory vectors with Bag-of-Words (BoW)

1. Embed each word
2. Sum embedding vectors

$$\text{“Sam drops apple”} \rightarrow \underbrace{\vec{v}_{\text{Sam}} + \vec{v}_{\text{drops}} + \vec{v}_{\text{apple}}}_{\text{Embedding Vectors}} = \vec{m}_i$$

Memory Vector

E.g.) **temporal structure:** special words for time and include them in BoW

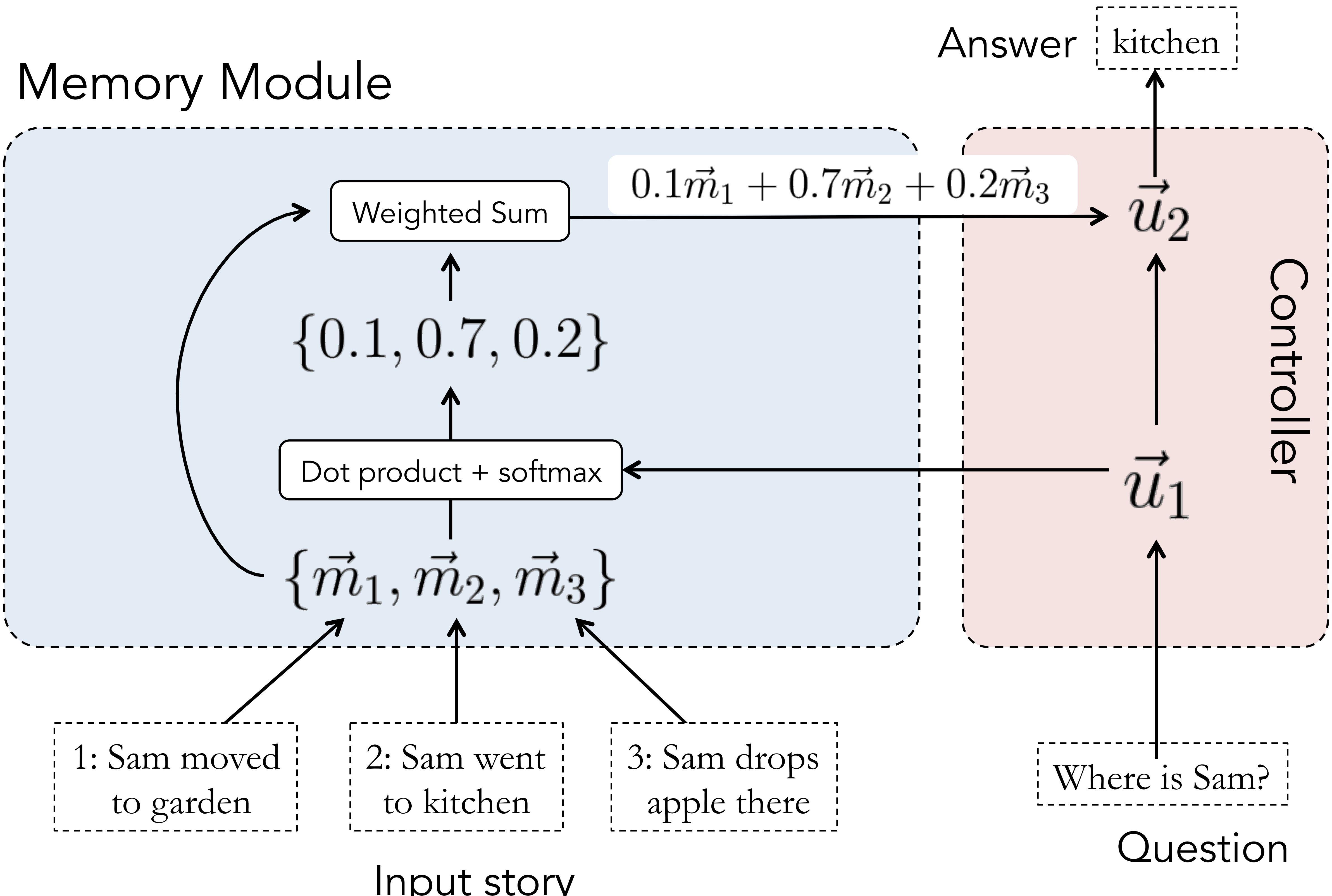
1: “Sam moved to garden”

2: “Sam went to kitchen”

3: “Sam drops apple” $\rightarrow v_{\text{Sam}} + v_{\text{drops}} + v_{\text{apple}} + v_3 = m_3$

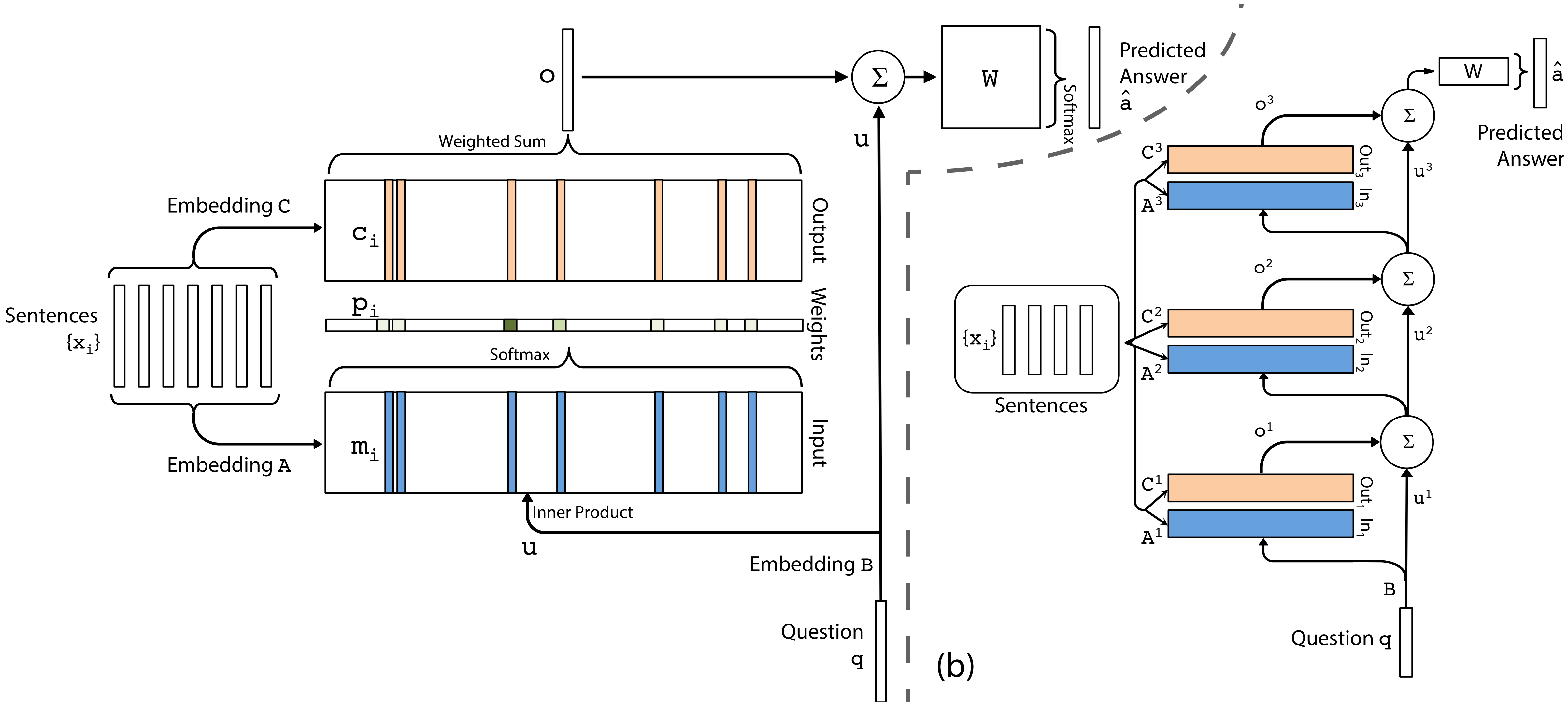
Time embedding

# Question & Answering



# Attention Mechanism and Memory Networks

- Architecture



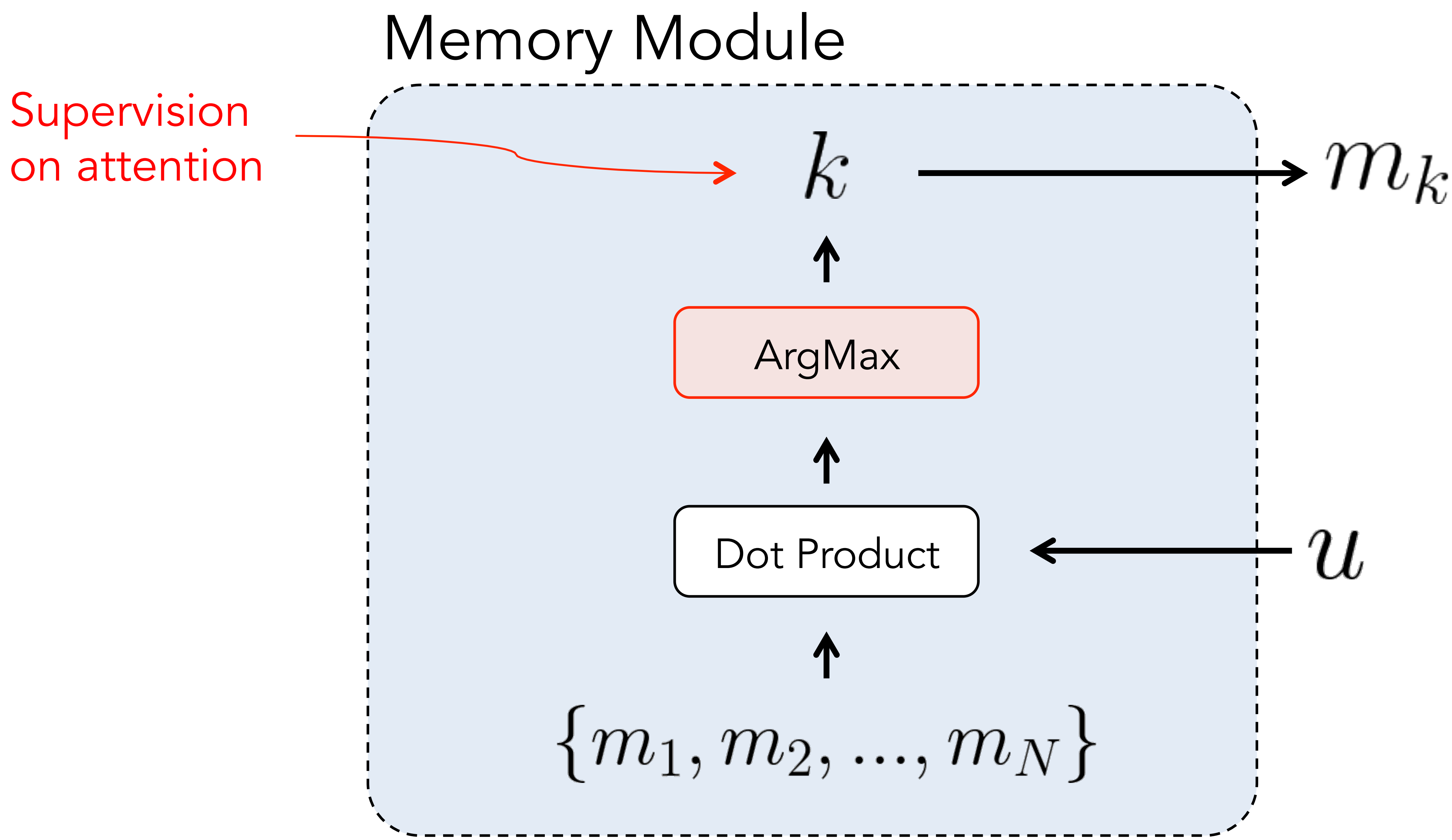

- Example results:

| | | |
|---|---|---|
| Sam walks into the kitchen.<br>Sam picks up an apple.<br>Sam walks into the bedroom.<br>Sam drops the apple.<br>Q: Where is the apple?<br>A. Bedroom | Brian is a lion.<br>Julius is a lion.<br>Julius is white.<br>Bernhard is green.<br>Q: What color is Brian?<br>A. White | Mary journeyed to the den.<br>Mary went back to the kitchen.<br>John journeyed to the bedroom.<br>Mary discarded the milk.<br>Q: Where was the milk before the den?<br>A. Hallway |

Sukhbaatar, Szlam, Weston, Fergus: End-To-End Memory Networks ArXiv 2015

# Related Work (I)

Hard attention Memory Network [Weston et al. ICLR 2015]



# Related Work (II)

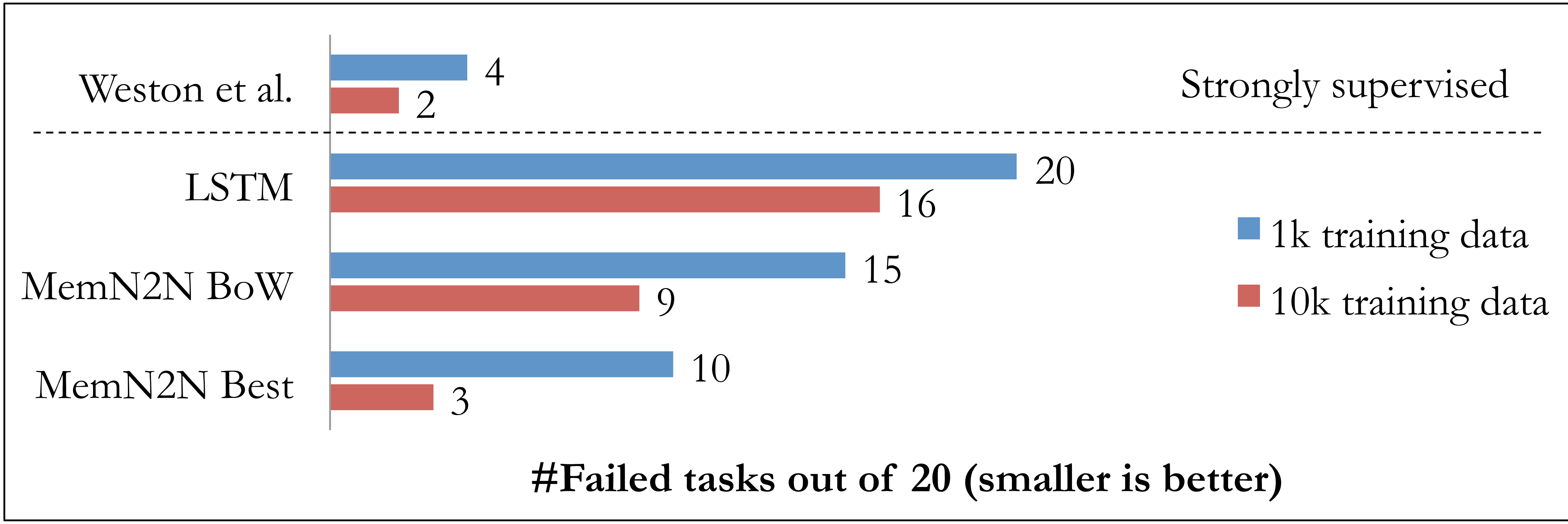
- RNNsearch [Bahdanau et al. 2015]
  - Encoder-decoder RNN with attention
  - Our model can be considered as an attention model with multiple hops
- Recent works on external memory
  - Stack memory for RNNs [Joulin & Mikolov. 2015]
  - Neural Turing Machine [Graves et al. 2014]
- Early works on neural network and memory
  - [Steinbuch & Piske. 1963]; [Taylor. 1959]
  - [Das et al. 1992]; [Mozer et al. 1993]
- Concurrent works
  - Dynamic Memory Networks [Kumar et al. 2015]
  - Attentive reader [Hermann et al. 2015]
  - Stack, Queue [Grefenstette et al. 2015]

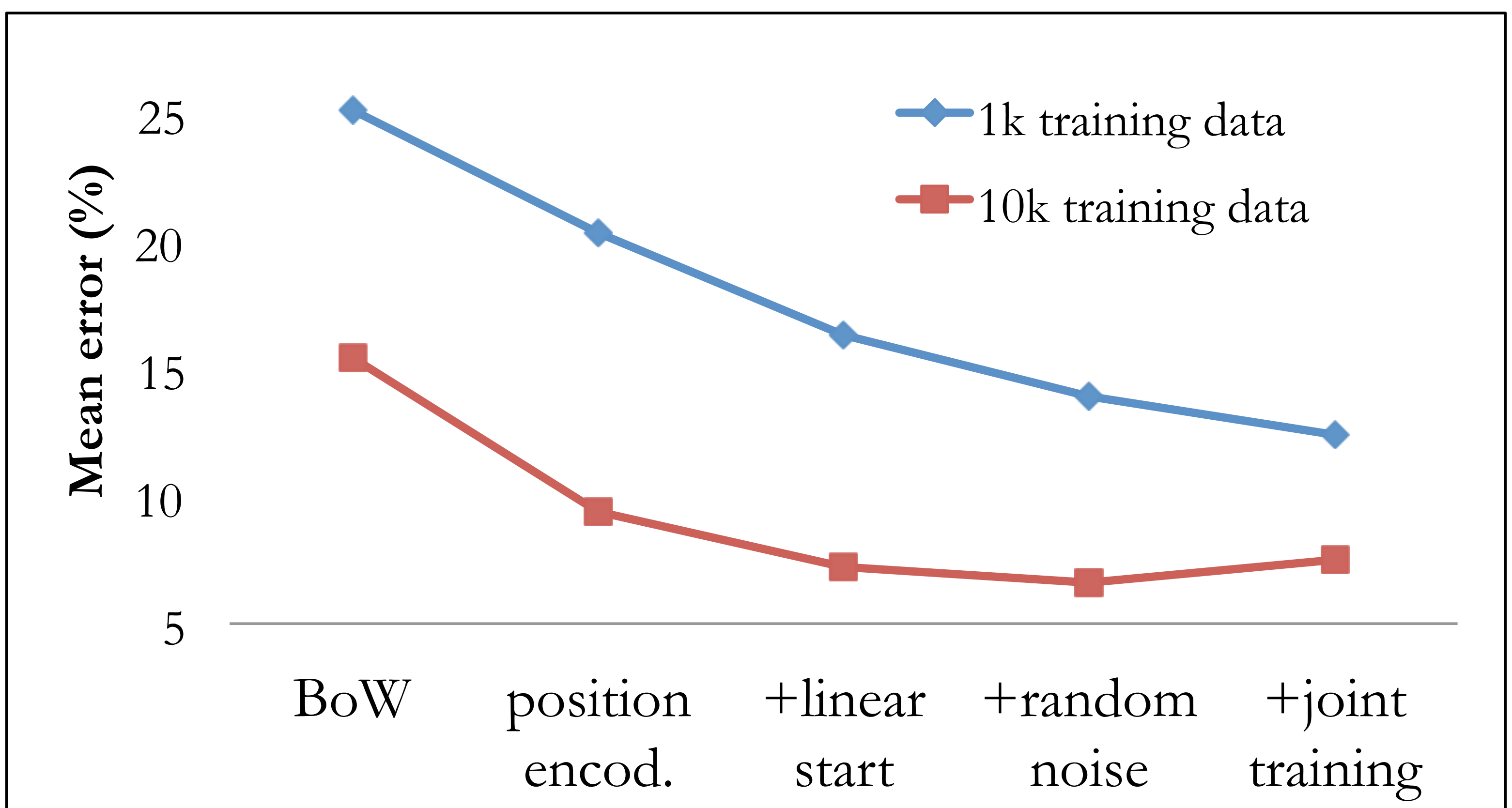
# Experiment on bAbI Q&A data

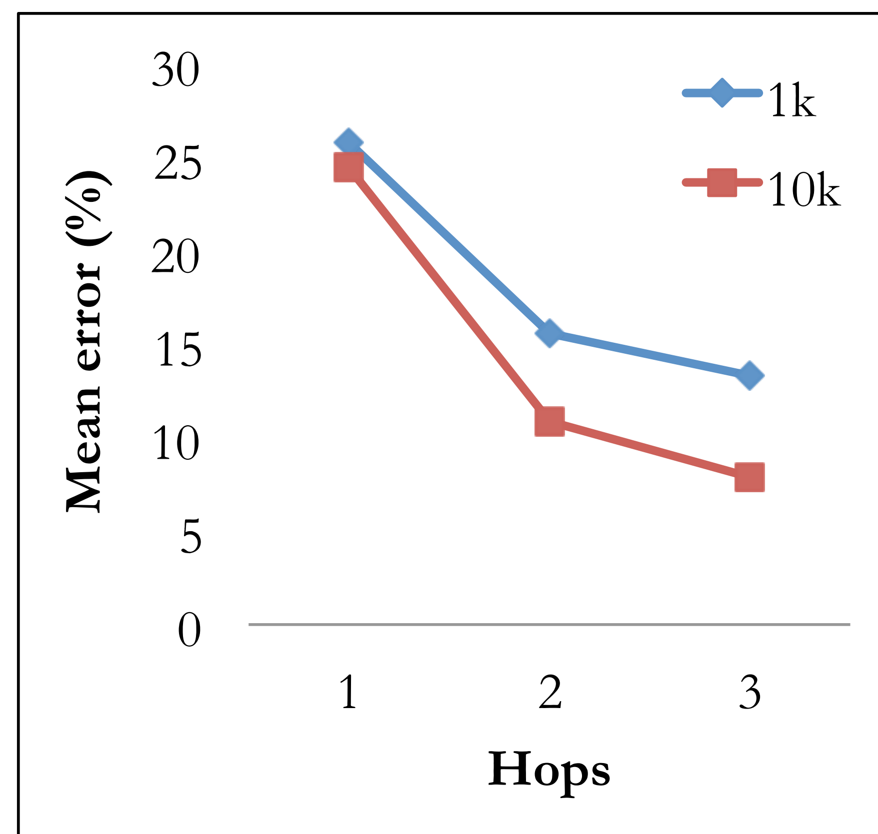
- Data: 20 bAbI tasks [Weston et al. arXiv: 1502.05698, 2015]
- Answer questions after reading short story
- Small vocabulary, simple language
- Different tasks require different reasoning
- Training data size 1K or 10K for each task

| | |
|---|---|
| Sam walks into the kitchen.<br>Sam picks up an apple.<br>Sam walks into the bedroom.<br>Sam drops the apple.<br>Q: Where is the apple?<br>A. Bedroom | Brian is a lion.<br>Julius is a lion.<br>Julius is white.<br>Bernhard is green.<br>Q: What color is Brian?<br>A. White |

# Performance on bAbI test set



# Examples of Attention Weights

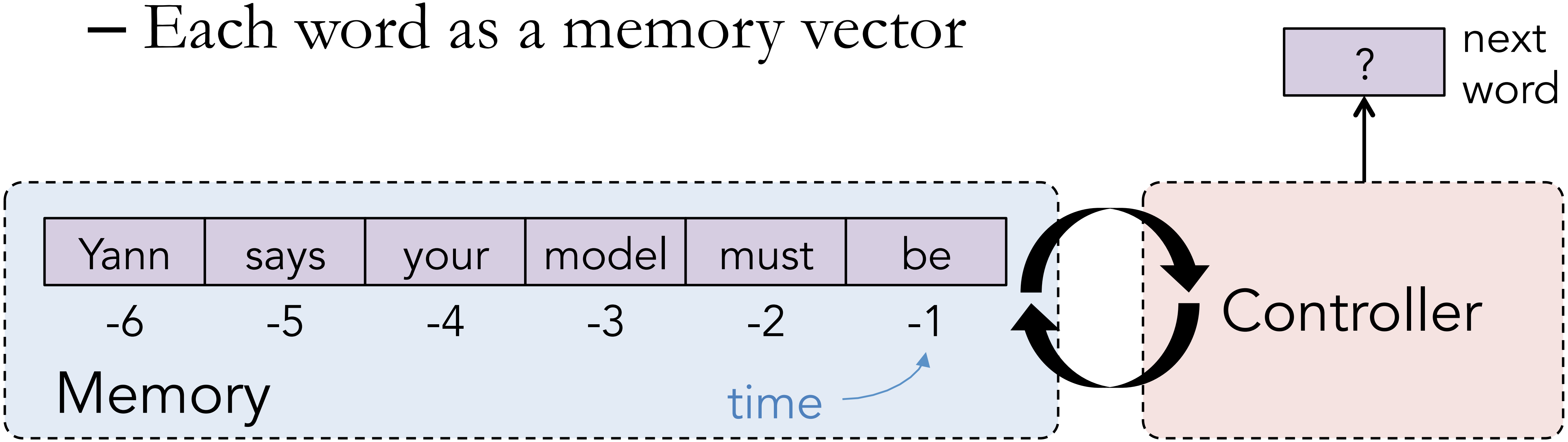
- 2 test cases:

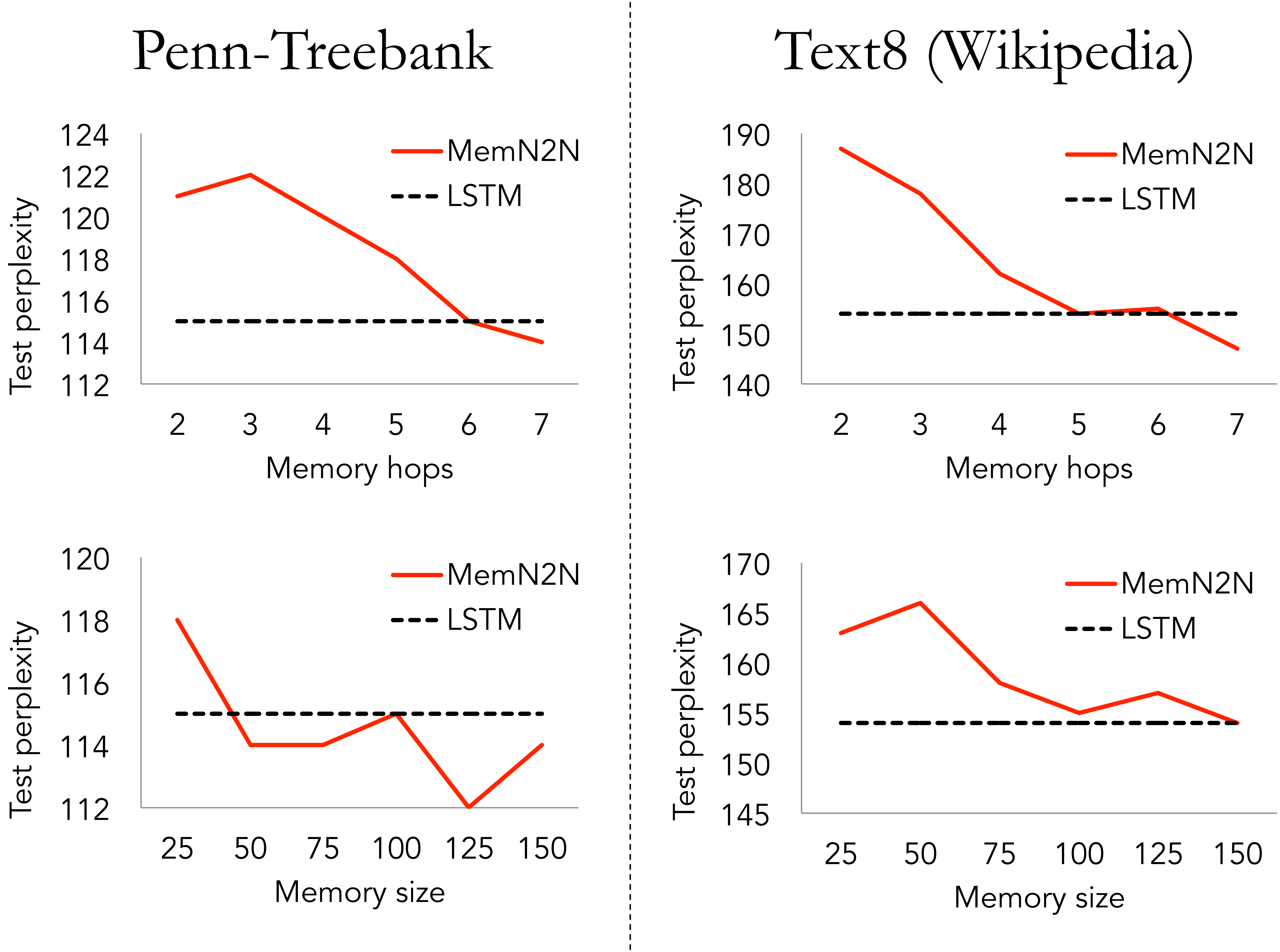
| **Story (2: 2 supporting facts)** | **Hop 1** | **Hop 2** | **Hop 3** |
|---|---|---|---|
| John dropped the milk. | 0.06 | 0.00 | 0.00 |
| John took the milk there. | 0.88 | 1.00 | 0.00 |
| Sandra went back to the bathroom. | 0.00 | 0.00 | 0.00 |
| John moved to the hallway. | 0.00 | 0.00 | 1.00 |
| Mary went back to the bedroom. | 0.00 | 0.00 | 0.00 |
| **Where is the milk? Answer: hallway Prediction: hallway** | | | |

| **Story (16: basic induction)** | **Hop 1** | **Hop 2** | **Hop 3** |
|---|---|---|---|
| Brian is a frog. | 0.00 | 0.98 | 0.00 |
| Lily is gray. | 0.07 | 0.00 | 0.00 |
| Brian is yellow. | 0.07 | 0.00 | 1.00 |
| Julius is green. | 0.06 | 0.00 | 0.00 |
| Greg is a frog. | 0.76 | 0.02 | 0.00 |
| **What color is Greg? Answer: yellow Prediction: yellow** | | | |

# Experiment on Language modeling

- Data
  - Penn Treebank: 1M words 10K vocab
  - Text8 (Wikipedia): 16M words 40K vocab
- Model
  - Controller module: linear + non-linearity
  - Each word as a memory vector





Penn-Treebank
Text8 (Wikipedia)
MemN2N
LSTM
Test perplexity
124
122
120
118
116
114
112
2
3
4
5
6
7
Memory hops
190
180
170
160
150
140
120
25
50
75
100
125
150
Memory size
170
165
155
145

# Conclusion

- Proposed a neural net model with external memory
  - Soft attention over memory locations
  - End-to-end training with backpropagation
- Good results on a toy QA tasks
- Comparable to LSTM on language modeling
- Versatile model: also apply to writing and games

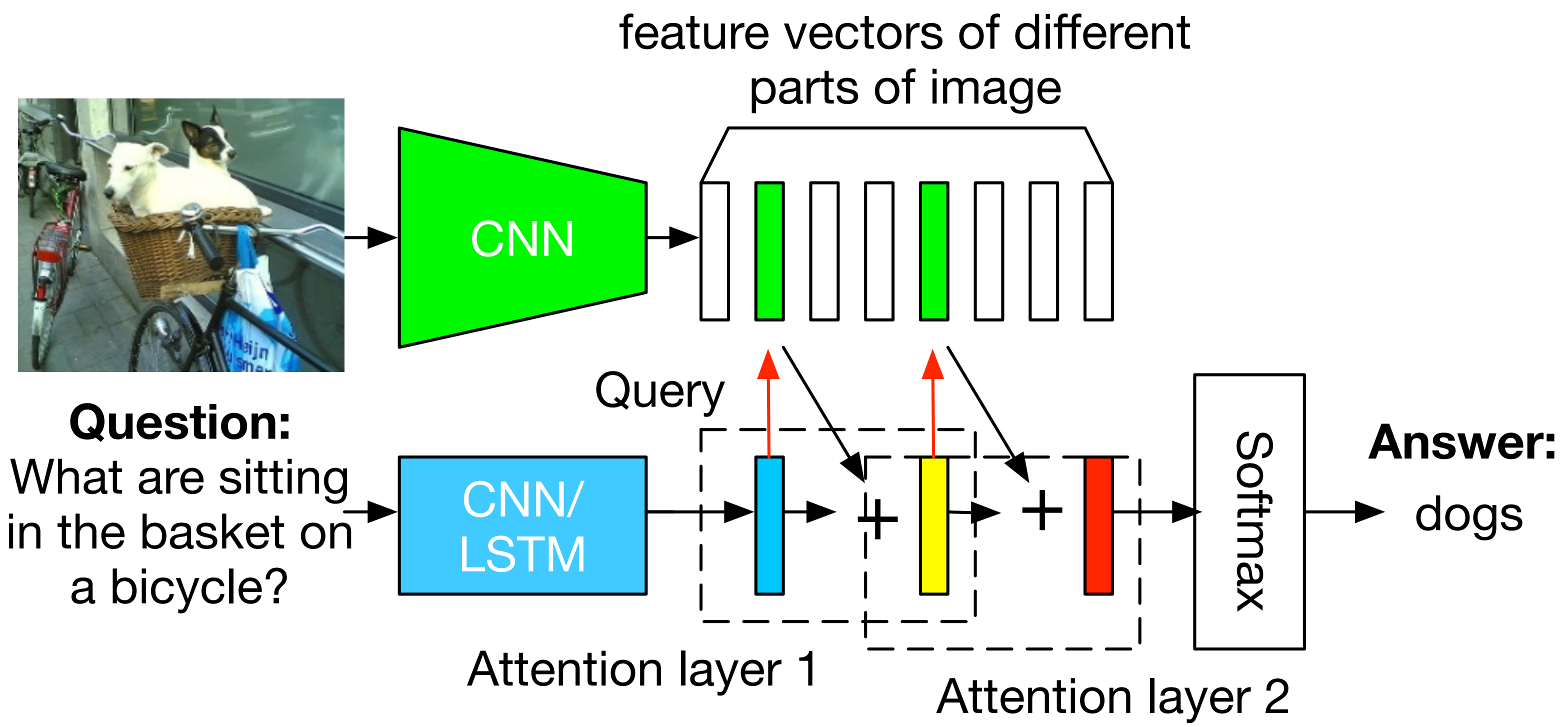
# Stacked Attention Network for Image Question Answering

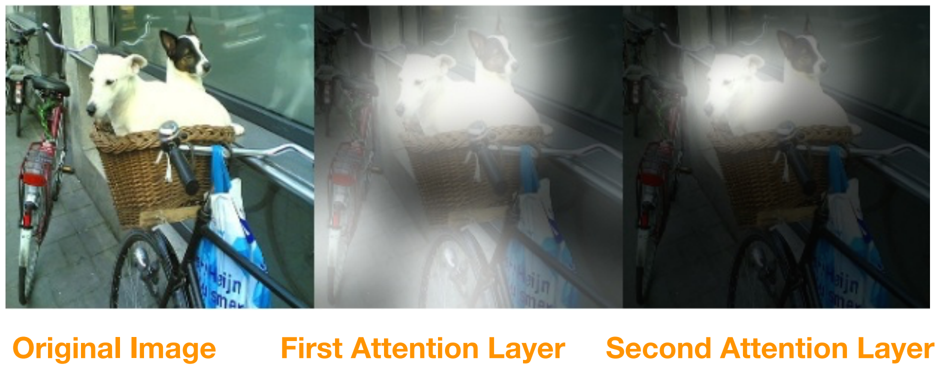
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

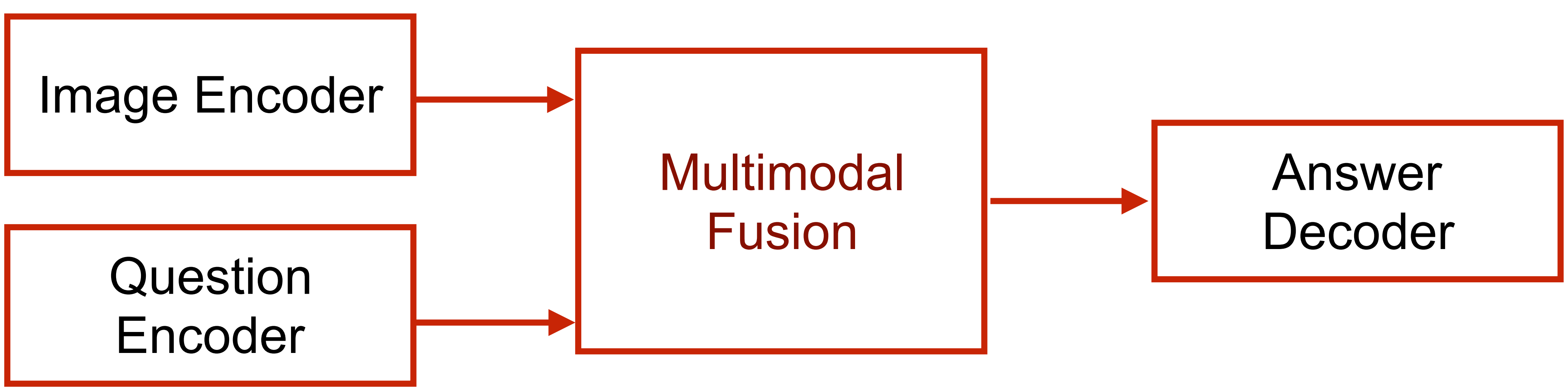
CVPR'16

slide credits: Mateusz Malinowski

# Stacked Attention Networks - Multimodal Fusion

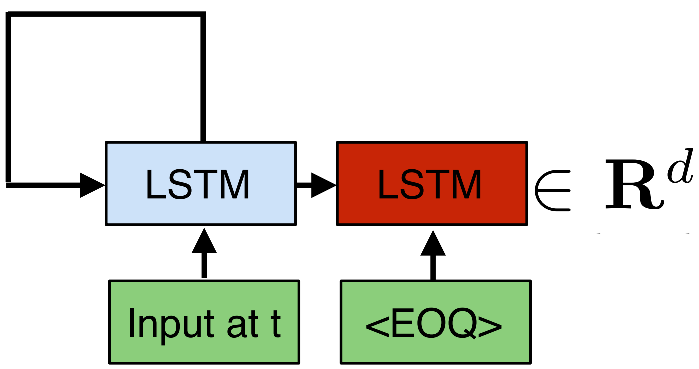


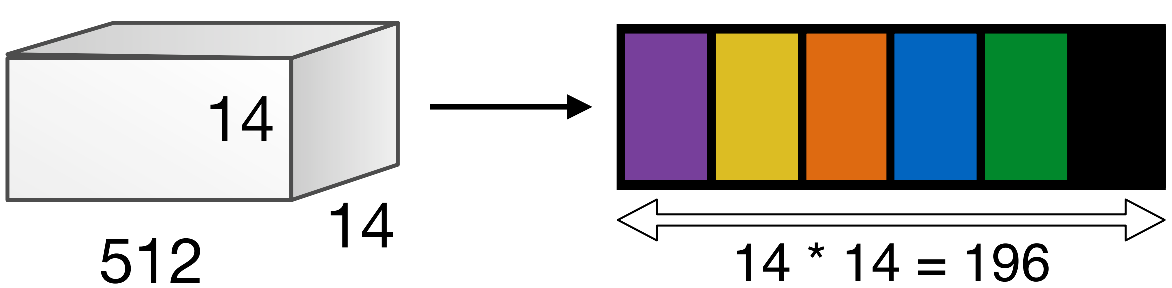



# Stacked Attention Networks - Multimodal Fusion

- More informative representation
    - Model can place higher weights at regions



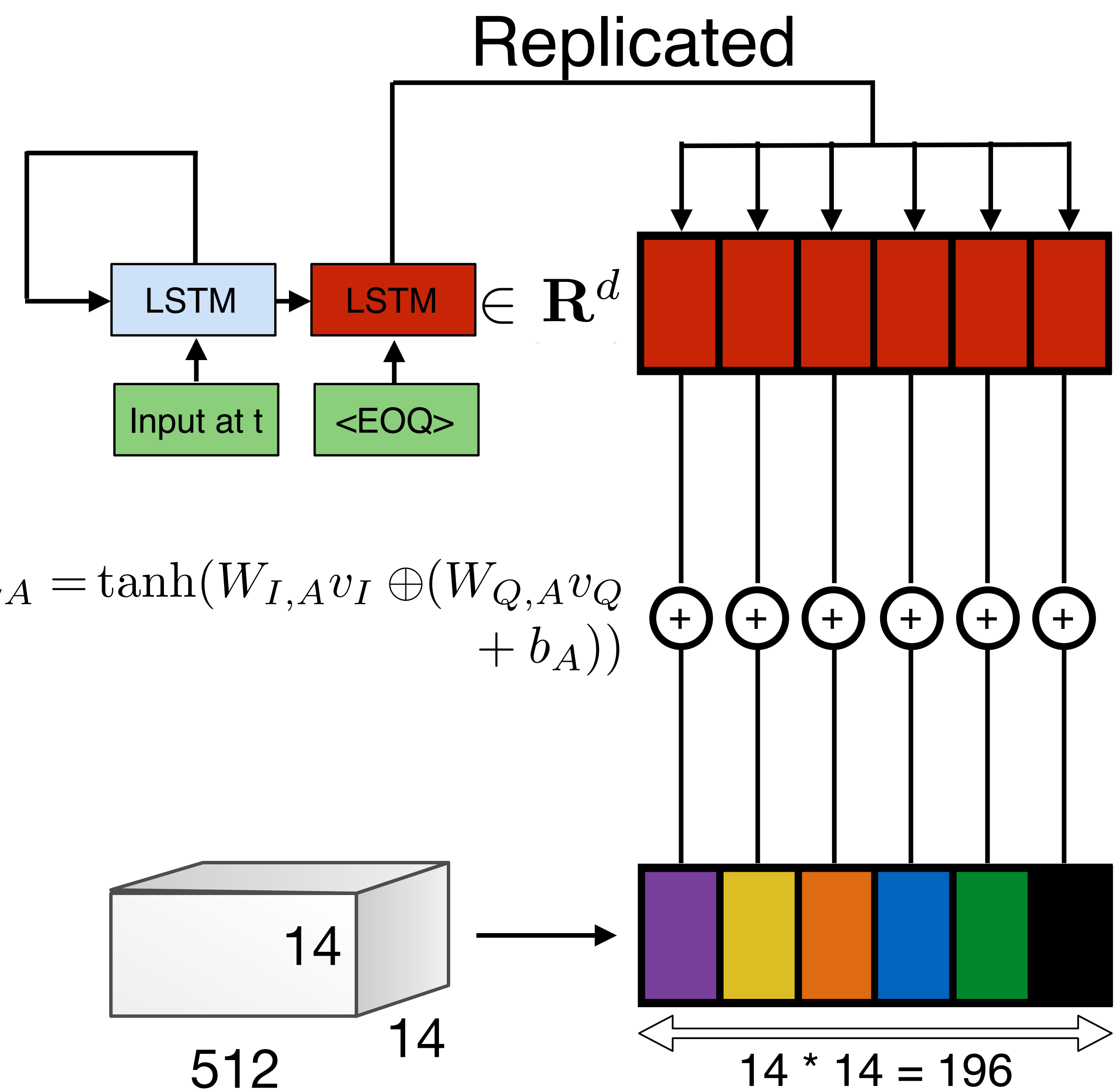




$$v_I = \tanh(W_I f_I + b_I) \in \mathbf{R}^{d \times m}$$

$$f_I = \mathrm{CNN}_{vgg}(I)$$

# Stacked Attention Networks - Multimodal Fusion

- More informative representation
  - Model can place higher weights at regions



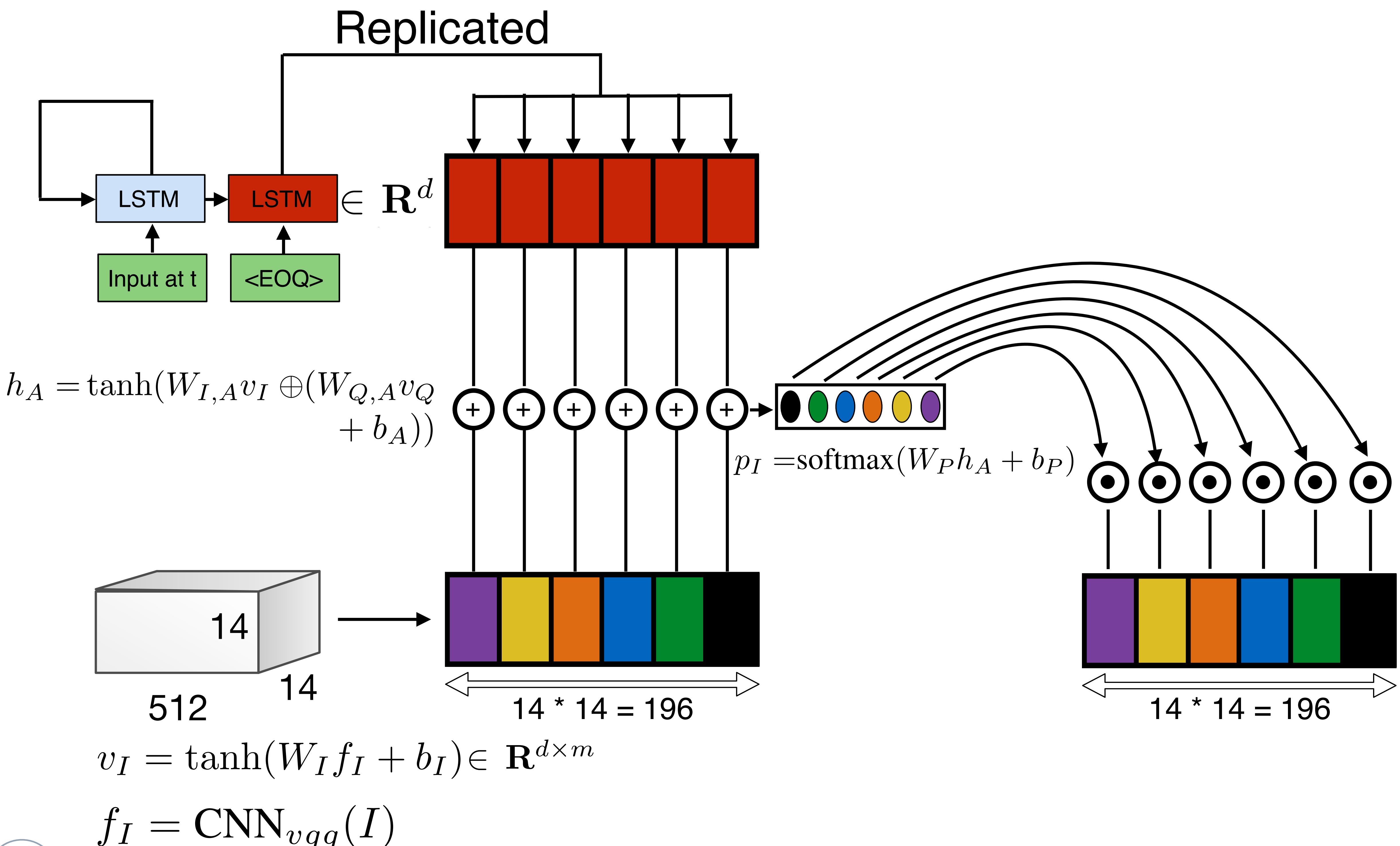

$$h_A = \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A))$$

$$v_I = \tanh(W_I f_I + b_I) \in \mathbf{R}^{d \times m}$$

$$f_I = \text{CNN}_{vgg}(I)$$

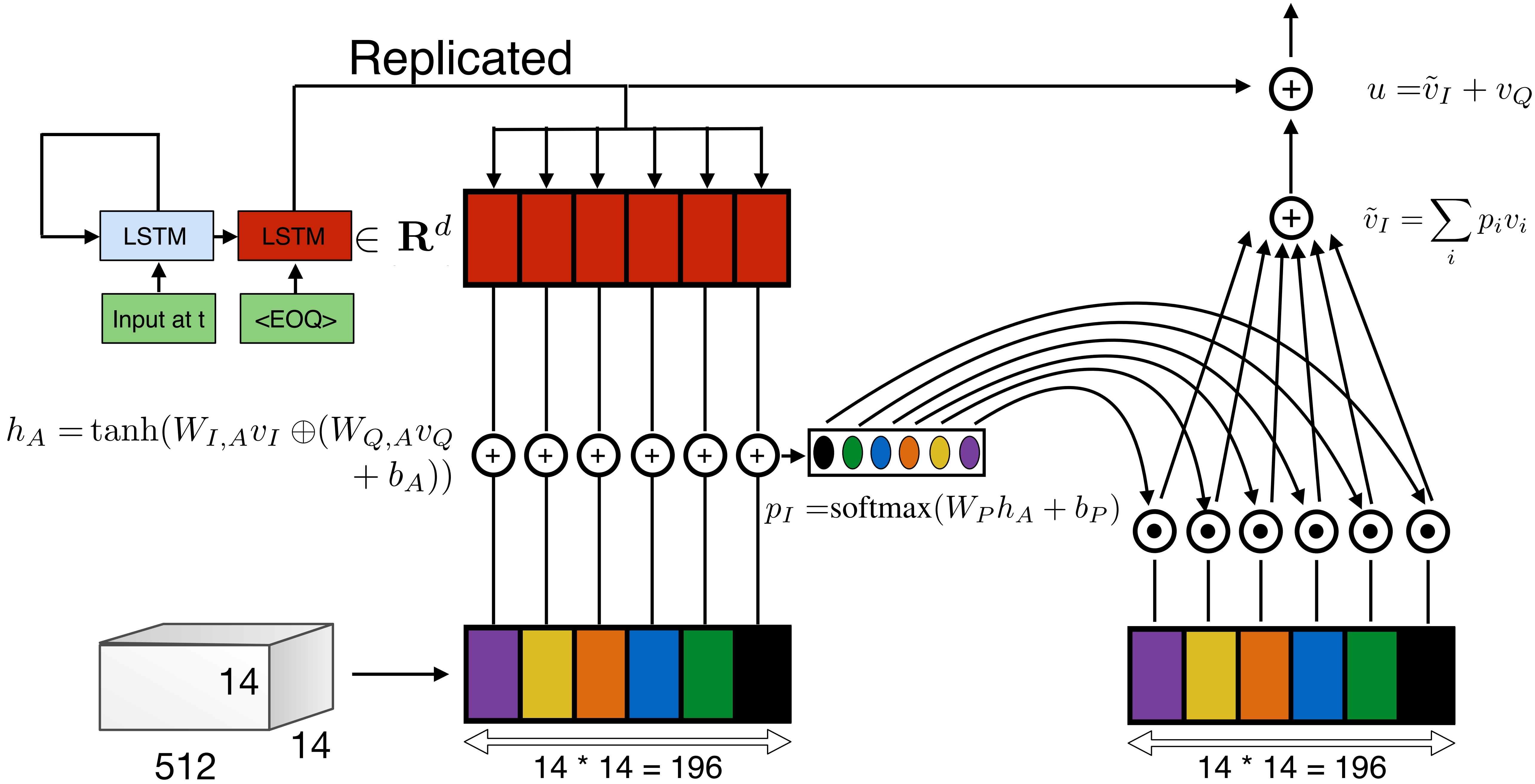
# Stacked Attention Networks - Multimodal Fusion

- More informative representation
  - Model can place higher weights at regions



# Stacked Attention Networks - Multimodal Fusion

- More informative representation
  - Model can place higher weights at regions

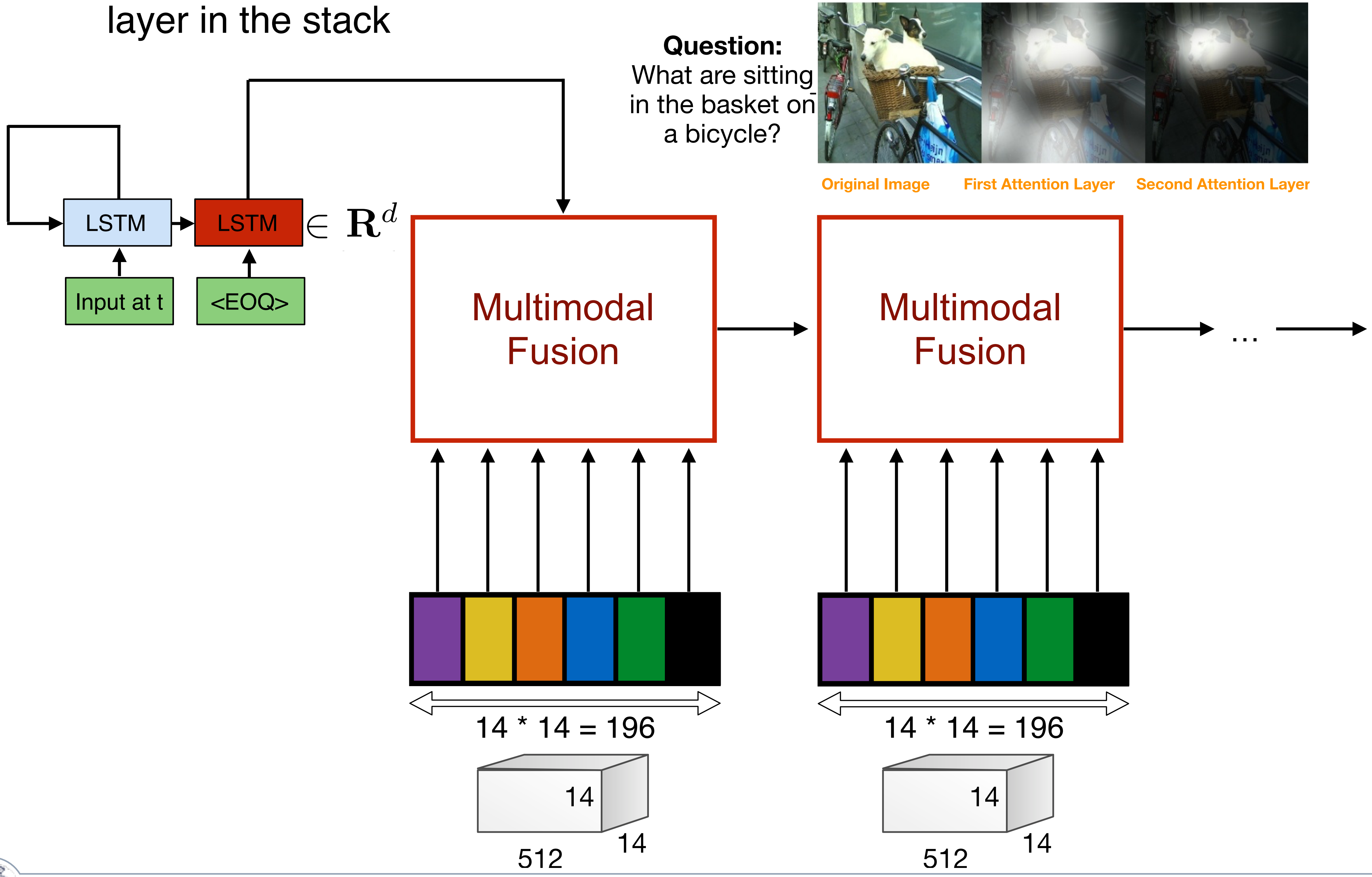



$v_I = \tanh(W_I f_I + b_I) \in \mathbf{R}^{d \times m}$

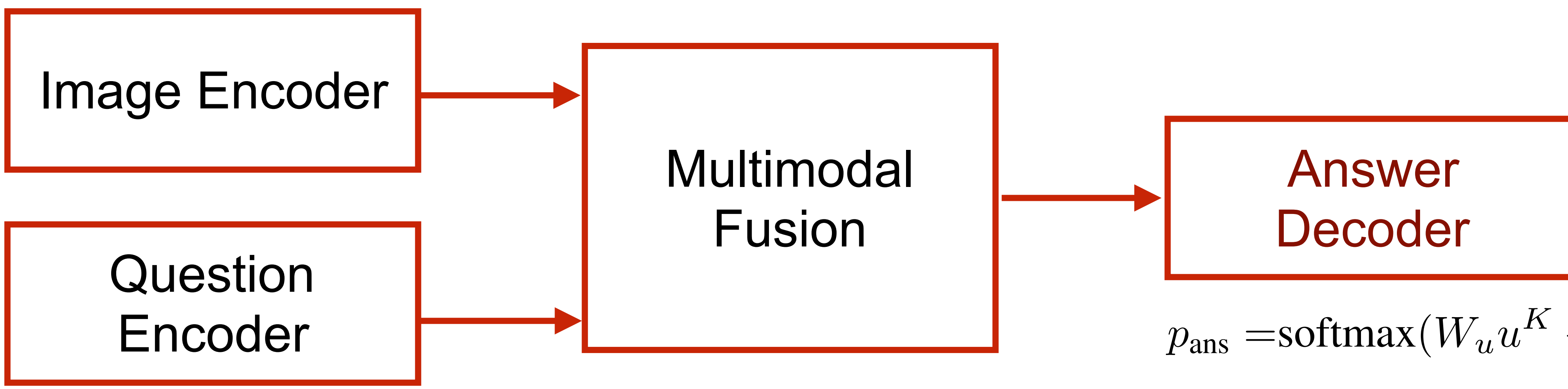
$f_I = \mathrm{CNN}_{vgg}(I)$

# Stacked Attention Networks - Many stacks

- Many stacks for many phases of reasoning
  - The output of the fusion module can be treated as a language vector for the next layer in the stack



# Stacked Attention Networks - Answer Decoder




$$p_{\text{ans}} = \text{softmax}(W_u u^K + b_u)$$

# Stacked Attention Networks - Results

- Significantly improves results over all Visual Turing Test datasets

| Methods | Accuracy | WUPS0.9 | WUPS0.0 |
|---|---|---|---|
| **Multi-World:** [18] | | | |
| Multi-World | 7.9 | 11.9 | 38.8 |
| **Ask-Your-Neurons:** [19] | | | |
| Language | 17.2 | 22.8 | 58.4 |
| Language + IMG | 19.4 | 25.3 | 62.0 |
| **CNN:** [17] | | | |
| IMG-CNN | 23.4 | 29.6 | 63.0 |
| **Ours:** | | | |
| SAN(1, LSTM) | 28.9 | 34.7 | 68.5 |
| SAN(1, CNN) | 29.2 | 35.1 | 67.8 |
| SAN(2, LSTM) | **29.3** | 34.9 | 68.1 |
| SAN(2, CNN) | **29.3** | **35.1** | **68.6** |
| **Human :**[18] | | | |
| Human | 50.2 | 50.8 | 67.3 |

DAQUAR

| Methods | Accuracy | WUPS0.9 | WUPS0.0 |
|---|---|---|---|
| **VSE:** [21] | | | |
| GUESS | 6.7 | 17.4 | 73.4 |
| BOW | 37.5 | 48.5 | 82.8 |
| LSTM | 36.8 | 47.6 | 82.3 |
| IMG | 43.0 | 58.6 | 85.9 |
| IMG+BOW | 55.9 | 66.8 | 89.0 |
| VIS+LSTM | 53.3 | 63.9 | 88.3 |
| 2-VIS+BLSTM | 55.1 | 65.3 | 88.6 |
| **CNN:** [17] | | | |
| IMG-CNN | 55.0 | 65.4 | 88.6 |
| CNN | 32.7 | 44.3 | 80.9 |
| **Ours:** | | | |
| SAN(1, LSTM) | 59.6 | 69.6 | 90.1 |
| SAN(1, CNN) | 60.7 | 70.6 | 90.5 |
| SAN(2, LSTM) | 61.0 | 71.0 | 90.7 |
| SAN(2, CNN) | **61.6** | **71.6** | **90.9** |

Toronto COCO-QA

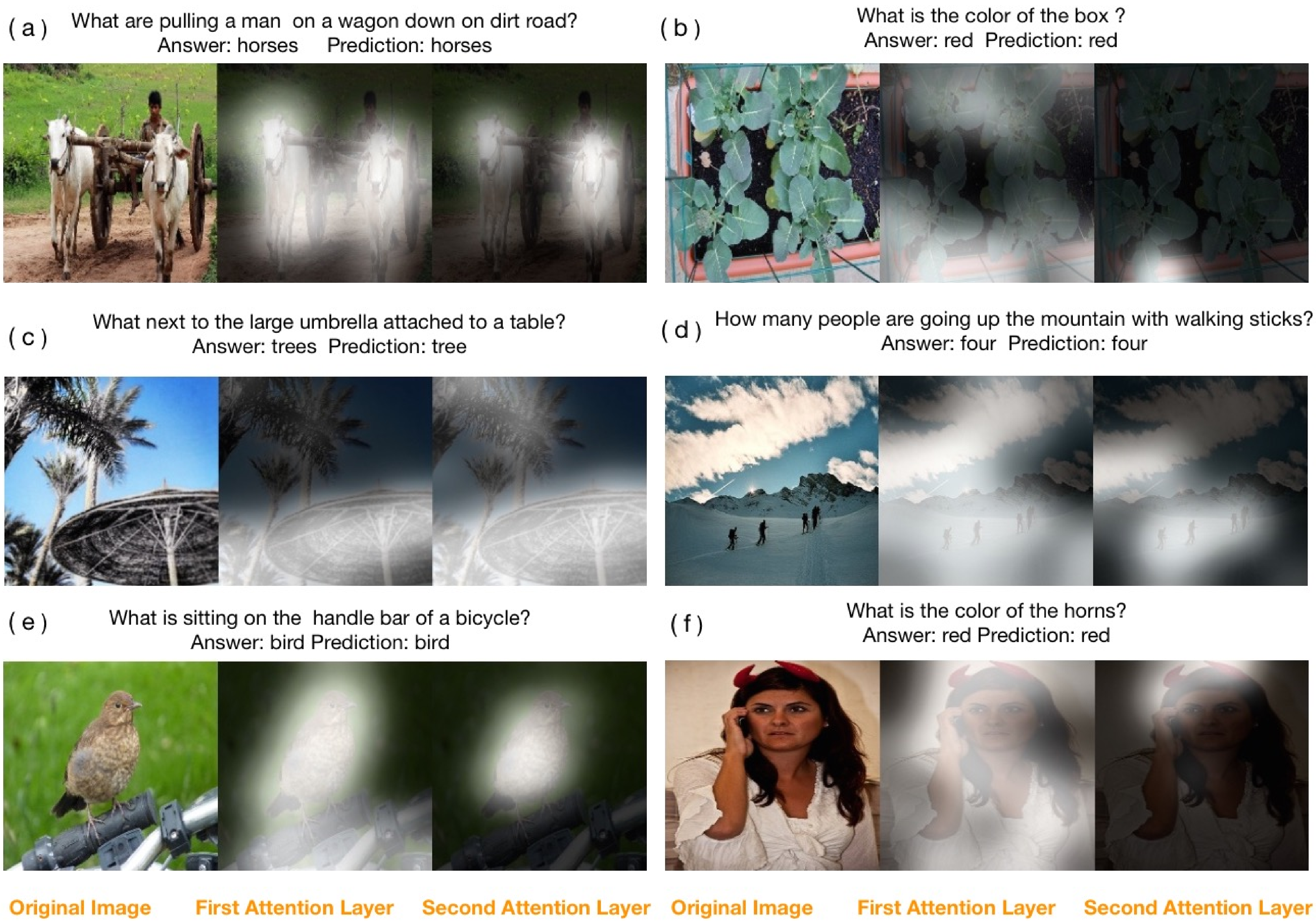
# Stacked Attention Networks - Results

- Significantly improves results over all Visual Turing Test datasets

| Methods | All | Yes/No 36% | Number 10% | Other 54% |
|---|---|---|---|---|
| **VQA:** [1] | | | | |
| Question | 48.1 | 75.7 | 36.7 | 27.1 |
| Image | 28.1 | 64.0 | 0.4 | 3.8 |
| Q+I | 52.6 | 75.6 | 33.7 | 37.4 |
| LSTM Q | 48.8 | 78.2 | 35.7 | 26.6 |
| LSTM Q+I | 53.7 | **78.9** | 35.2 | 36.4 |
| **Ours:** | | | | |
| SAN(1, LSTM) | 56.6 | 78.1 | 41.6 | 44.8 |
| SAN(1, CNN) | 56.9 | 78.8 | 42.0 | 45.0 |
| SAN(2, LSTM) | 57.3 | 78.3 | **42.2** | 45.9 |
| SAN(2, CNN) | **57.6** | 78.6 | 41.8 | **46.4** |
| **Human:** [1] | | | | |
| Human | 83.3 | 95.8 | 83.4 | 72.7 |

VQA

# Examples (good)

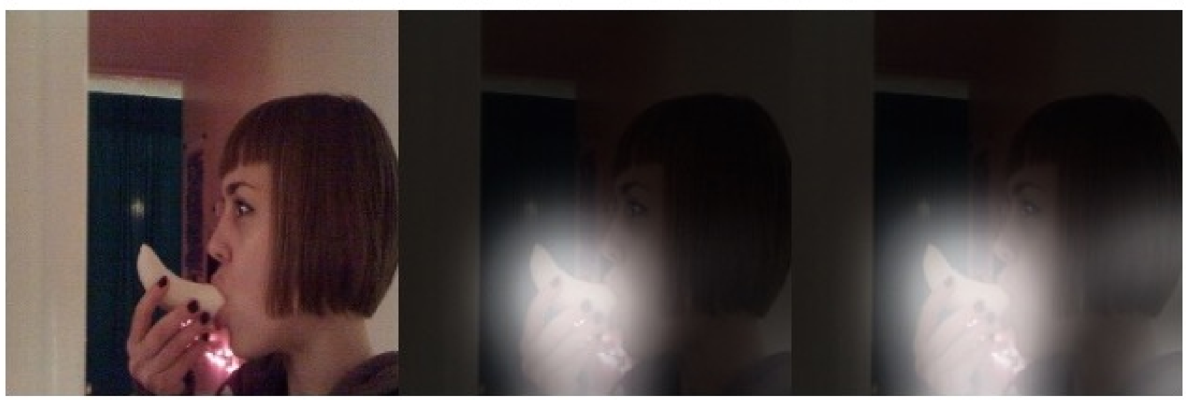


# Examples (bad)

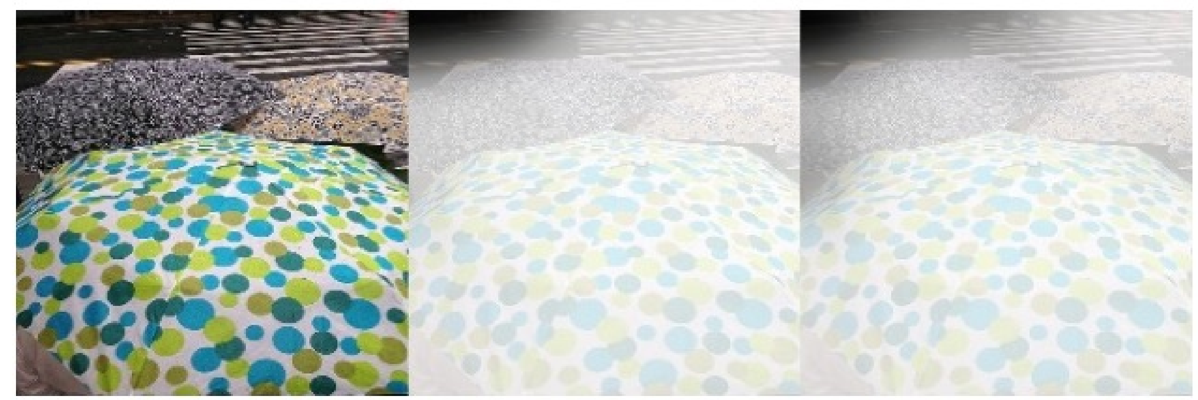
( a ) What swim in the ocean near two large ferries?
Answer: ducks Prediction: boats

( b ) What is the color of the shirt?
Answer: purple Prediction: green

( c ) What is the young woman eating?
Answer: banana Prediction: donut



( d ) How many umbrellas with various patterns?
Answer: three Prediction: two



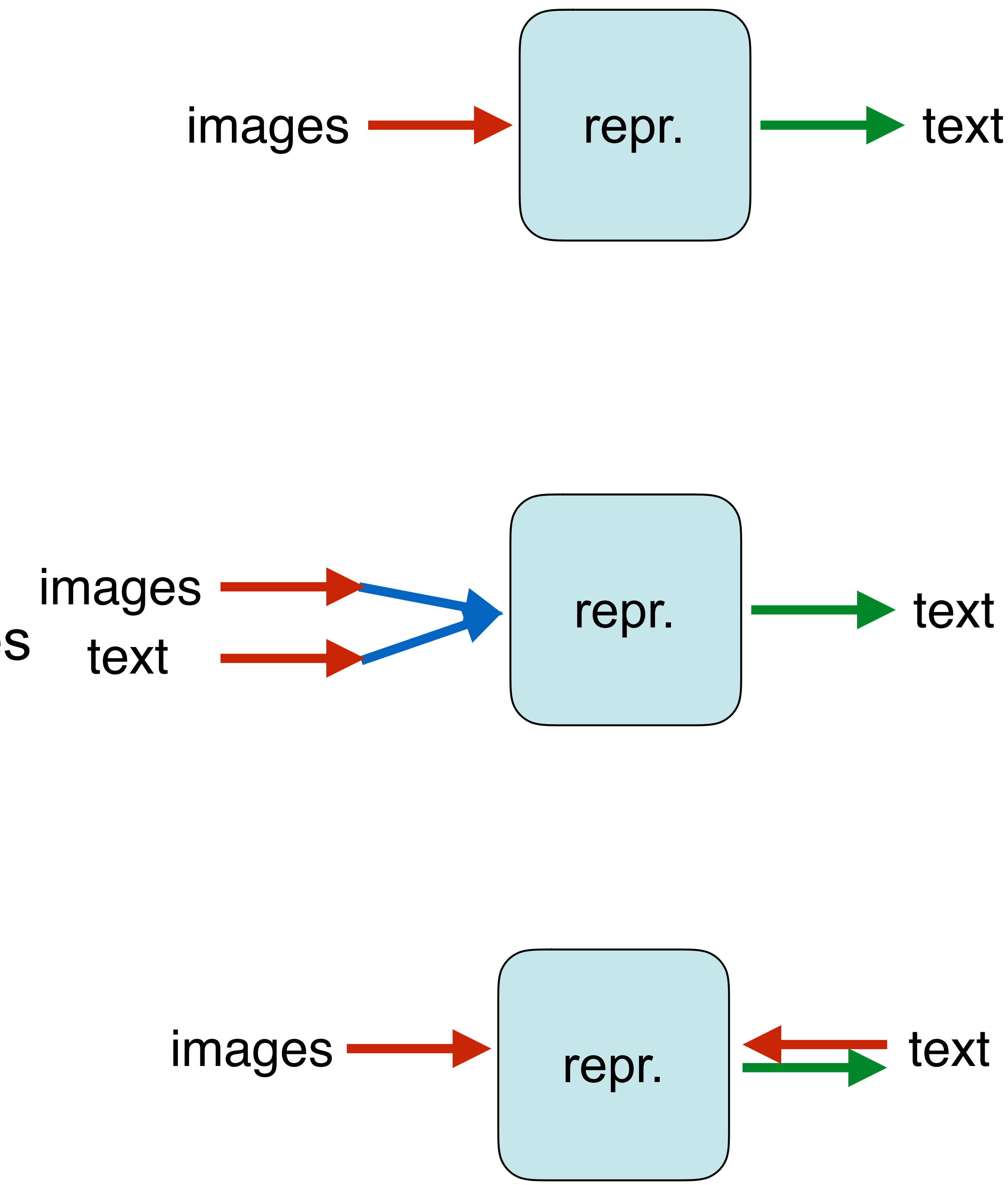
( e ) The very old looking what is on display?
Answer: pot Prediction: vase

( f ) What are passing underneath the walkway bridge?
Answer: cars Prediction: trains

**Original Image** **First Attention Layer** **Second Attention Layer** **Original Image** **First Attention Layer** **Second Attention Layer**

# Overview of Deep Learning Architectures

- **Encoders**
  - CNN for sequences, images, volumes
  - RNN for sequences
  - Pooling for sequences
  - Dense embedding layer (e.g. language w2v)
- **Decoders**
  - Unpooling for sequences, images, volumes
  - RNN for sequences
  - Dense regression
- **Merge**
  - Concatenate
  - Multiply
  - Sum/Average



max planck institut
informatik

# Thank you for your attention