VISUAL QUESTION ANSWETRING (VQA)

Foundations of Machine Learning – Fall 2018

Daniel Rivera Ruiz

New York University

Outline

1) What is VQA?
2) The VQA Dataset
3) LSTM I+Q
4) Hierarchical Co-Attention
5) Stacked Attention Networks
6) Multimodal Residual Learning
7) Focused Dynamic Attention
8) Dynamic Memory Networks
9) Multimodal Compact Bilinear
10) DualNet Architecture
11) Results Comparison
12) The VQA 2.0 Dataset
13) Future work

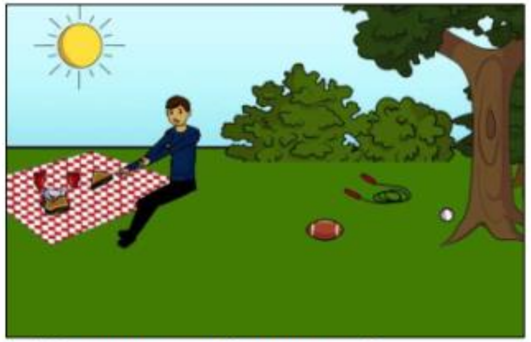
What is VQA?

Antol et al. (2015)

- Image + Question = Answer
- Intersection of Computer Vision and NLP
- "AI-complete" task:
  - ✓ multimodal knowledge
  - ✓ Evaluation metric

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?

The problem

Antol et al. (2015)

Predict the most likely answer $\hat{a}$ for a given image **x** and a question or phrase **q**:

$$\hat{a} = \underset{a \in A}{\operatorname{argmax}}\ p(a|\boldsymbol{x}, \boldsymbol{q}; \theta)$$

with model parameters θ and a set of possible answers A

The dataset

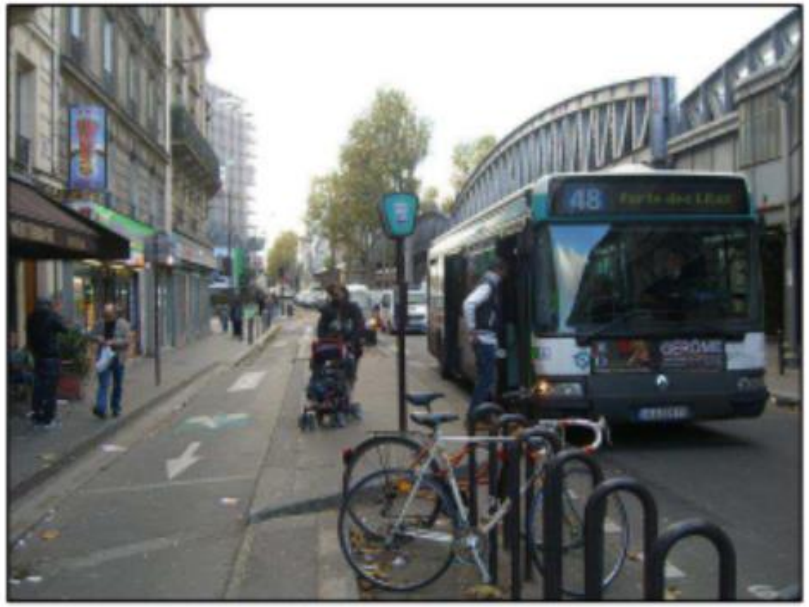
Antol et al. (2015)

- 204,721 images from the MS COCO dataset
- 50,000 abstract scenes
- 3 questions per image (over 760,000)
- 10 answers per question
- 2 tasks: open-ended and multiple-choice
- Open-ended accuracy metric: $\min\left(\frac{\text{\# humans that provided the answer}}{3}, 1\right)$

The dataset

Antol et al. (2015)

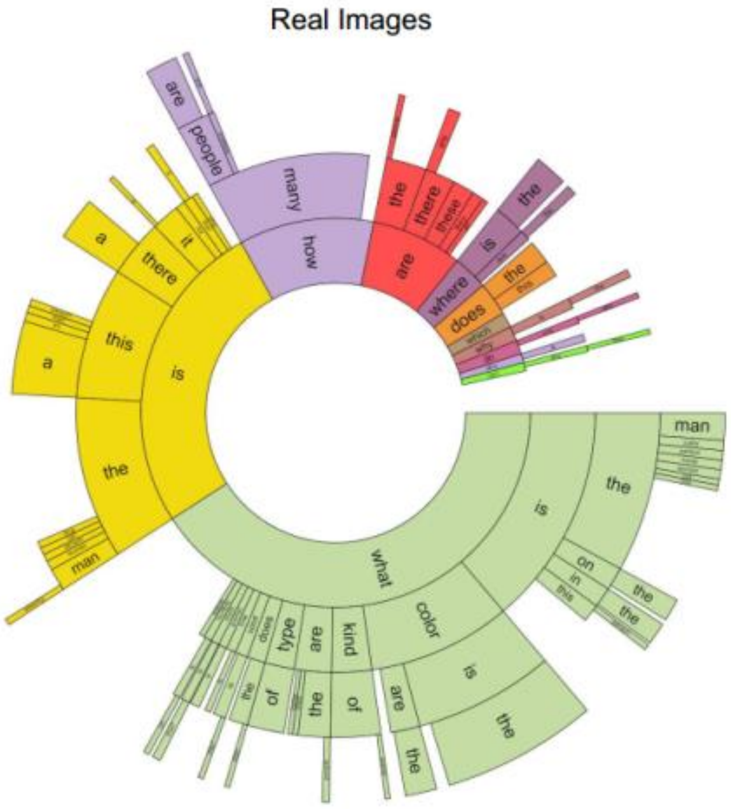
Plausible answers (blue) are collected for the multiple-choice task by having subjects answer the question without looking at the image

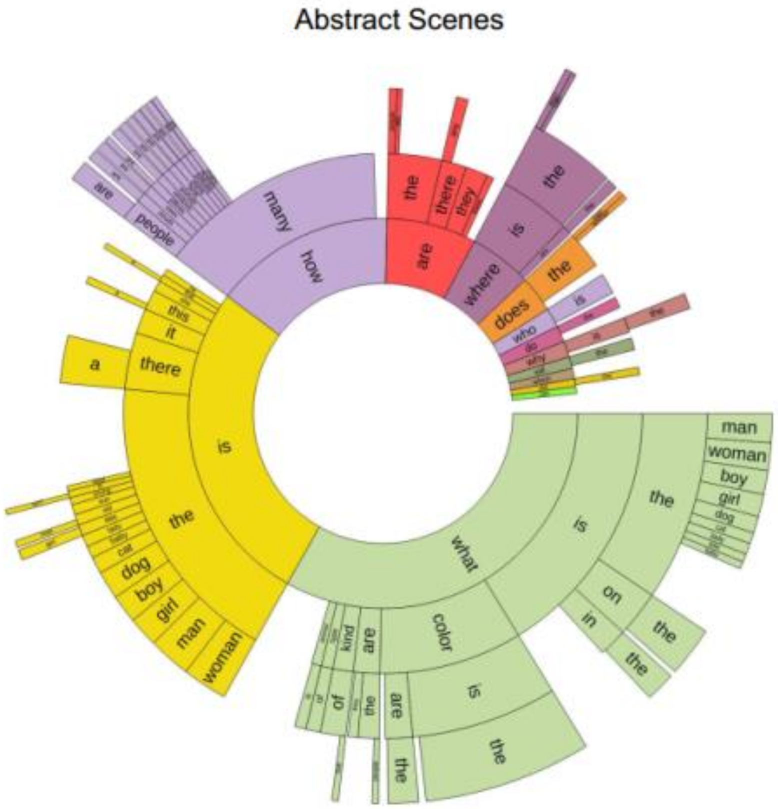


How many bikes are there?	2 2 2	3 4 12
What number is the bus?	48 48 48	4 46 number 6

The dataset

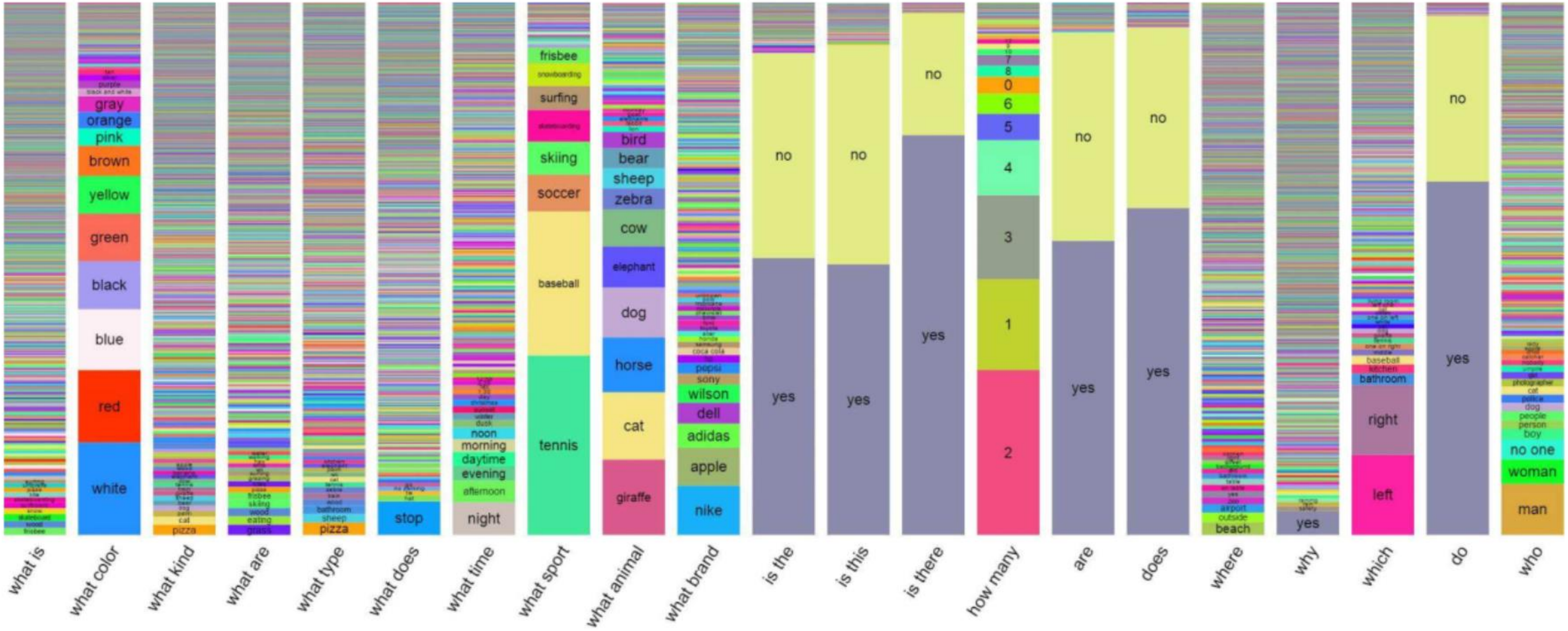
Antol et al. (2015)





The dataset

Antol et al. (2015)



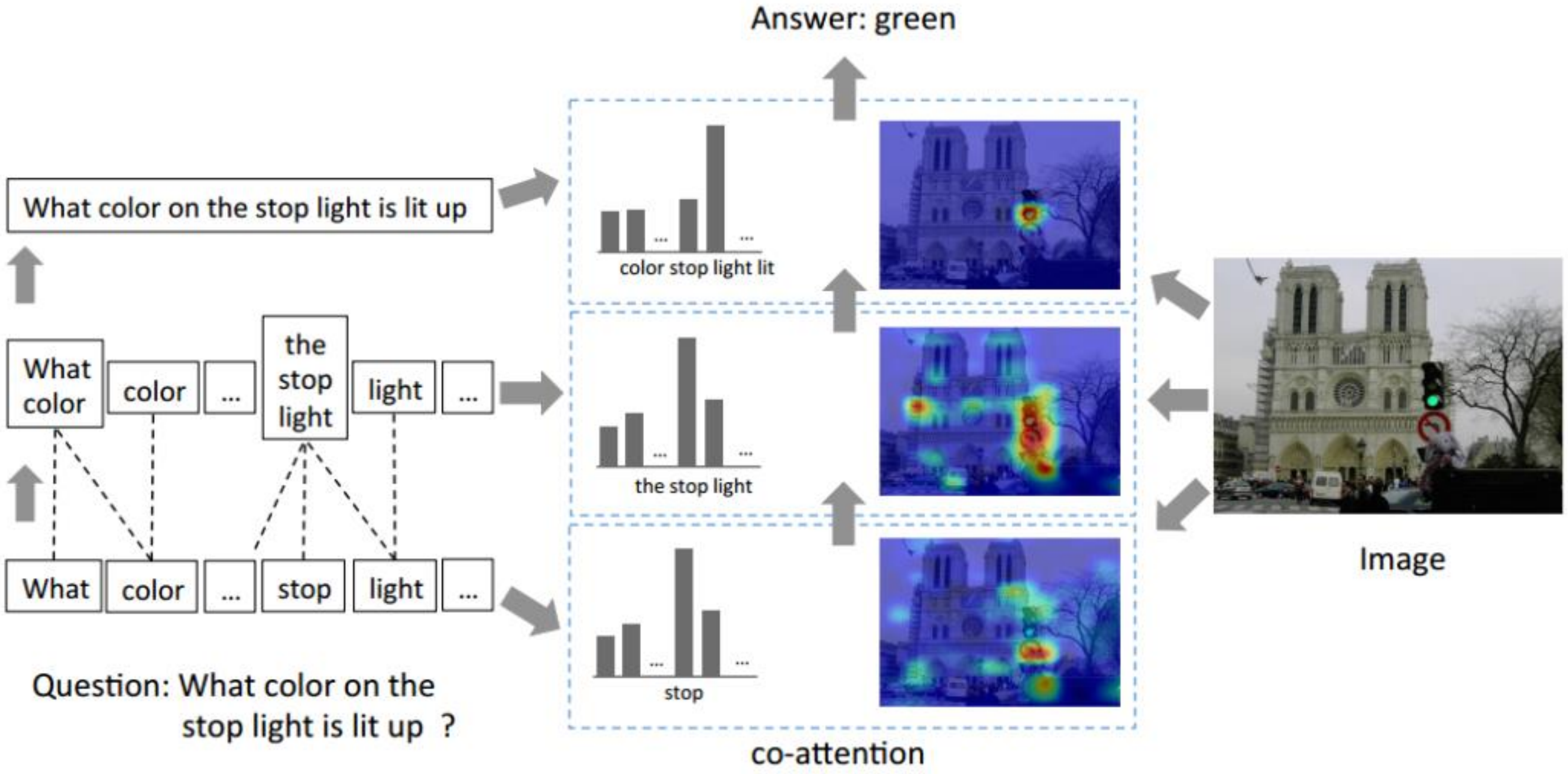
LSTM I+Q

Antol et al. (2015)

- Baseline model published with the VQA dataset
- Image features: last hidden layer (4096) of VGGNet
- Question features: LSTM (1024)
- Multimodality: element-wise multiplication (EWM)
- Attention mechanism: none
- Test-dev accuracy: **53.7%**
- Test-std accuracy: **54.1%**

Hierarchical Co-Attention

Lu et al. (2016)



Hierarchical Co-Attention

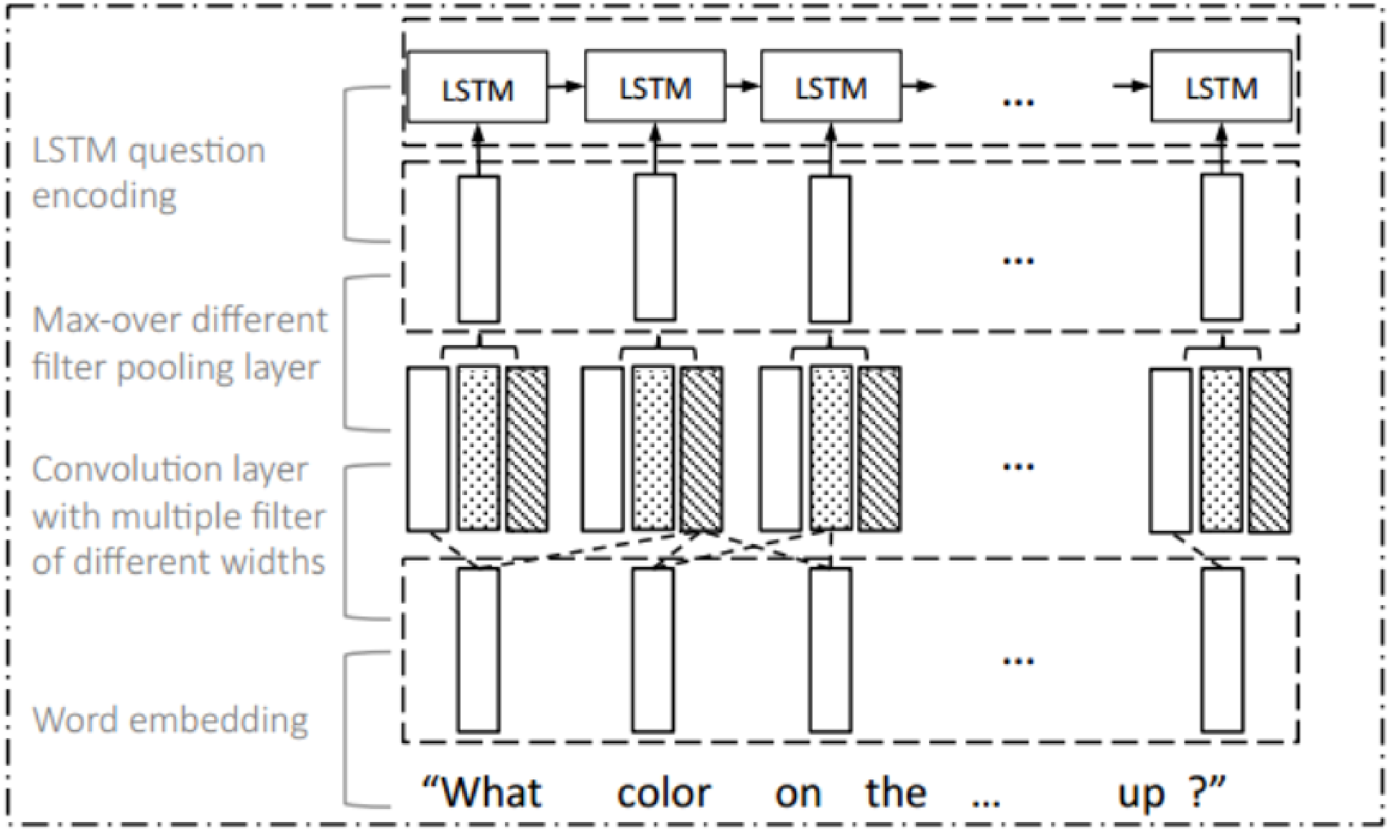
Lu et al. (2016)

Question hierarchy

$$\boldsymbol{Q}^w = \{\boldsymbol{q}_1^w, \boldsymbol{q}_2^w, \ldots, \boldsymbol{q}_T^w\}$$

$$\hat{\boldsymbol{q}}_{s,t}^p = \tanh(\boldsymbol{W}_c^s \boldsymbol{q}_{t:t+s-1}^w),$$

$$\boldsymbol{q}_{s,t}^p = \max(\hat{\boldsymbol{q}}_{1,t}^p, \hat{\boldsymbol{q}}_{2,t}^p, \hat{\boldsymbol{q}}_{3,t}^p)$$


LSTM question encoding
Max-over different filter pooling layer
Convolution layer with multiple filter of different widths
Word embedding
LSTM
LSTM
LSTM
...
LSTM
"What
color
on
the
...
up
?"

Hierarchical Co-Attention

Lu et al. (2016)

Parallel

$$\boldsymbol{C} = \tanh(\boldsymbol{Q}^T \boldsymbol{W}_b \boldsymbol{V})$$

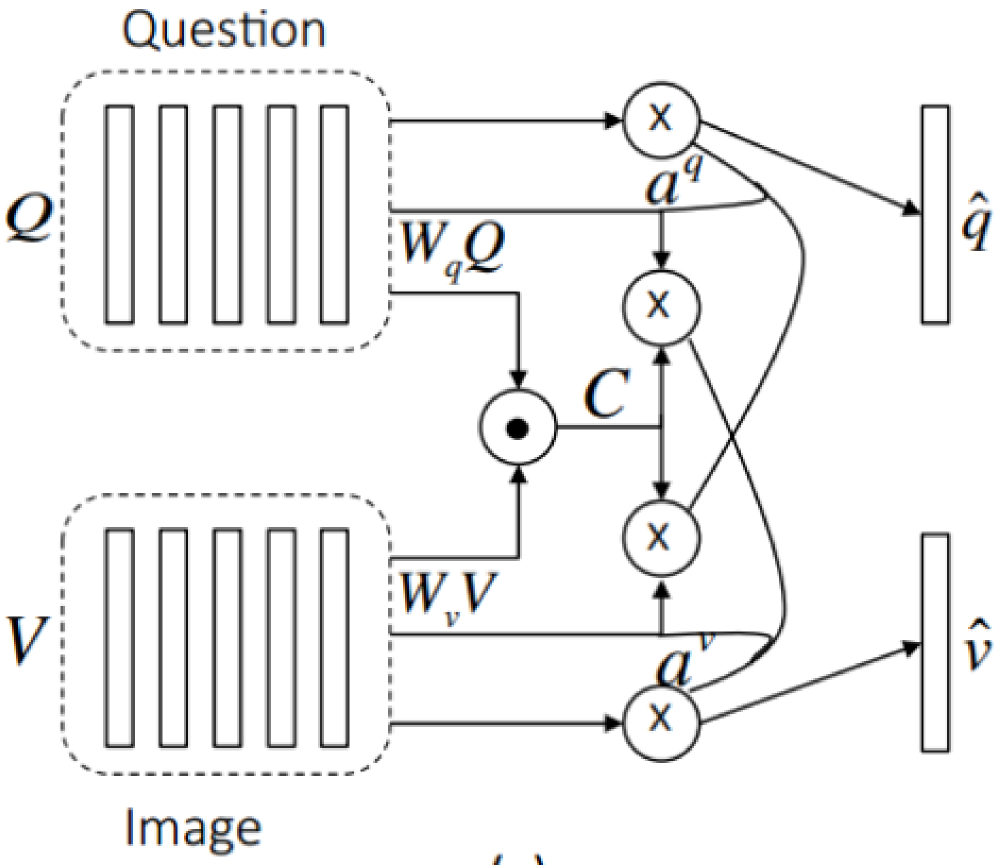
$$\boldsymbol{H}^v = \tanh(\boldsymbol{W}_v \boldsymbol{V} + (\boldsymbol{W}_q \boldsymbol{Q})\boldsymbol{C})$$

$$\boldsymbol{a}^v = \text{softmax}(w_{hv}^T \boldsymbol{H}^v)$$

$$\boldsymbol{H}^q = \tanh(\boldsymbol{W}_q \boldsymbol{Q} + (\boldsymbol{W}_v \boldsymbol{V})\boldsymbol{C}^T)$$

$$\boldsymbol{a}^q = \text{softmax}(w_{hq}^T \boldsymbol{H}^q)$$

$$\hat{\boldsymbol{v}} = \sum_{n=1}^{N} a_n^v \boldsymbol{v}_n \qquad \hat{\boldsymbol{q}} = \sum_{t=1}^{T} a_t^q \boldsymbol{q}_t$$



Hierarchical Co-Attention

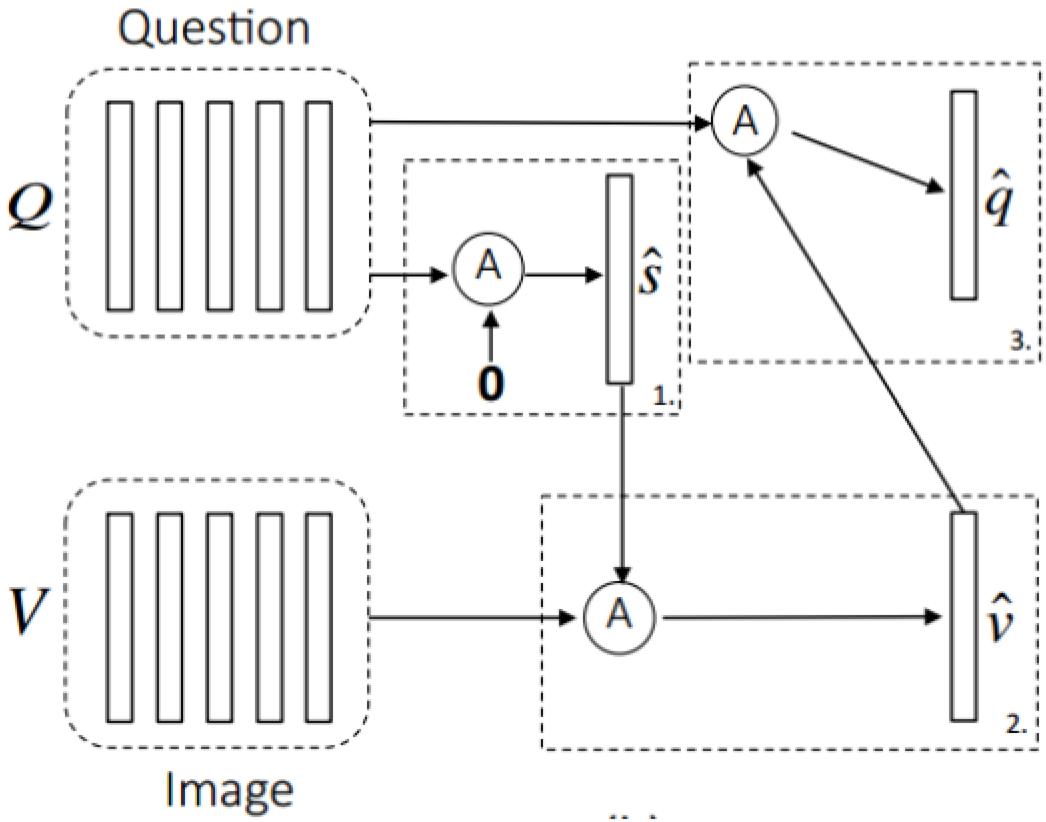
Lu et al. (2016)

Alternating

$$\boldsymbol{H} = \tanh(\boldsymbol{W}_x \boldsymbol{X} + (\boldsymbol{W}_g \boldsymbol{g})\mathbf{1}^T)$$

$$\boldsymbol{a}^x = \text{softmax}(w_{hx}^T \boldsymbol{H})$$

$$\hat{\boldsymbol{x}} = \sum a_i^x \boldsymbol{x}_i$$

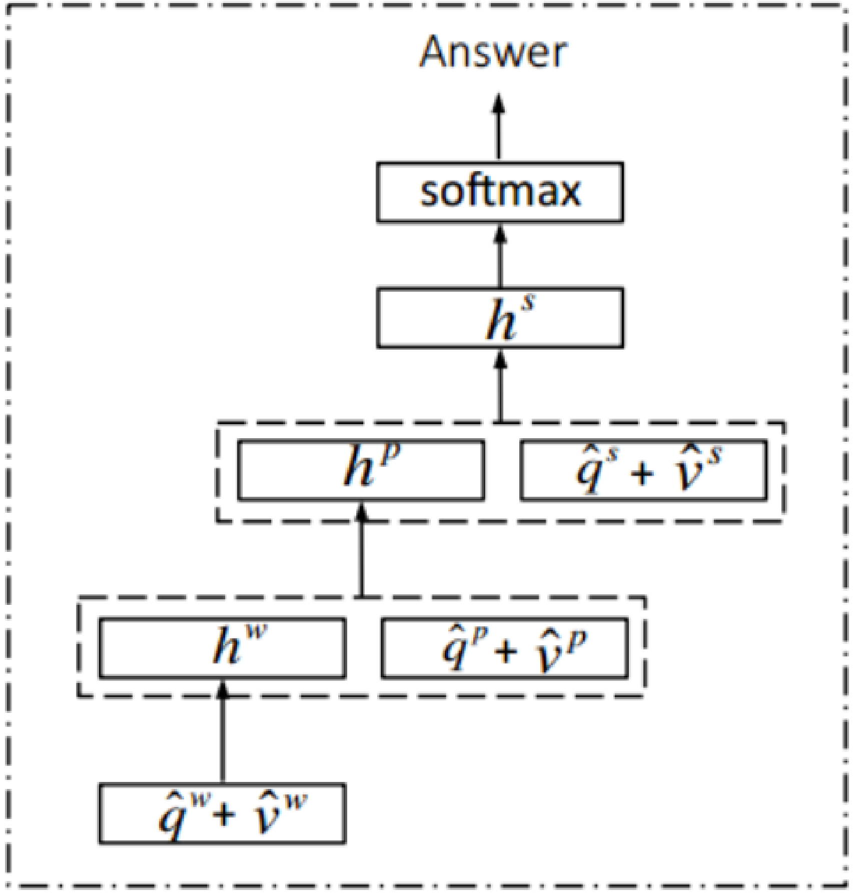


Hierarchical Co-Attention

Lu et al. (2016)

Prediction

$$\boldsymbol{h}^w = \tanh(\boldsymbol{W}_w(\hat{\boldsymbol{q}}^w + \hat{\boldsymbol{v}}^w))$$
$$\boldsymbol{h}^p = \tanh(\boldsymbol{W}_p[(\hat{\boldsymbol{q}}^p + \hat{\boldsymbol{v}}^p); \boldsymbol{h}^w])$$
$$\boldsymbol{h}^s = \tanh(\boldsymbol{W}_s[(\hat{\boldsymbol{q}}^s + \hat{\boldsymbol{v}}^s); \boldsymbol{h}^p])$$
$$p = \text{softmax}(\boldsymbol{W}_h \boldsymbol{h}^s)$$



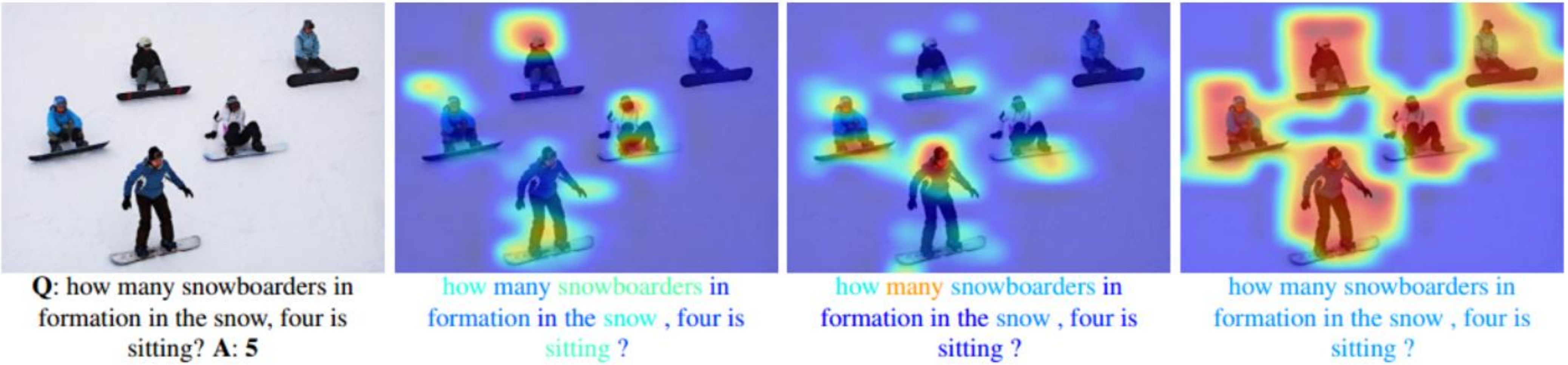
Hierarchical Co-Attention

Lu et al. (2016)

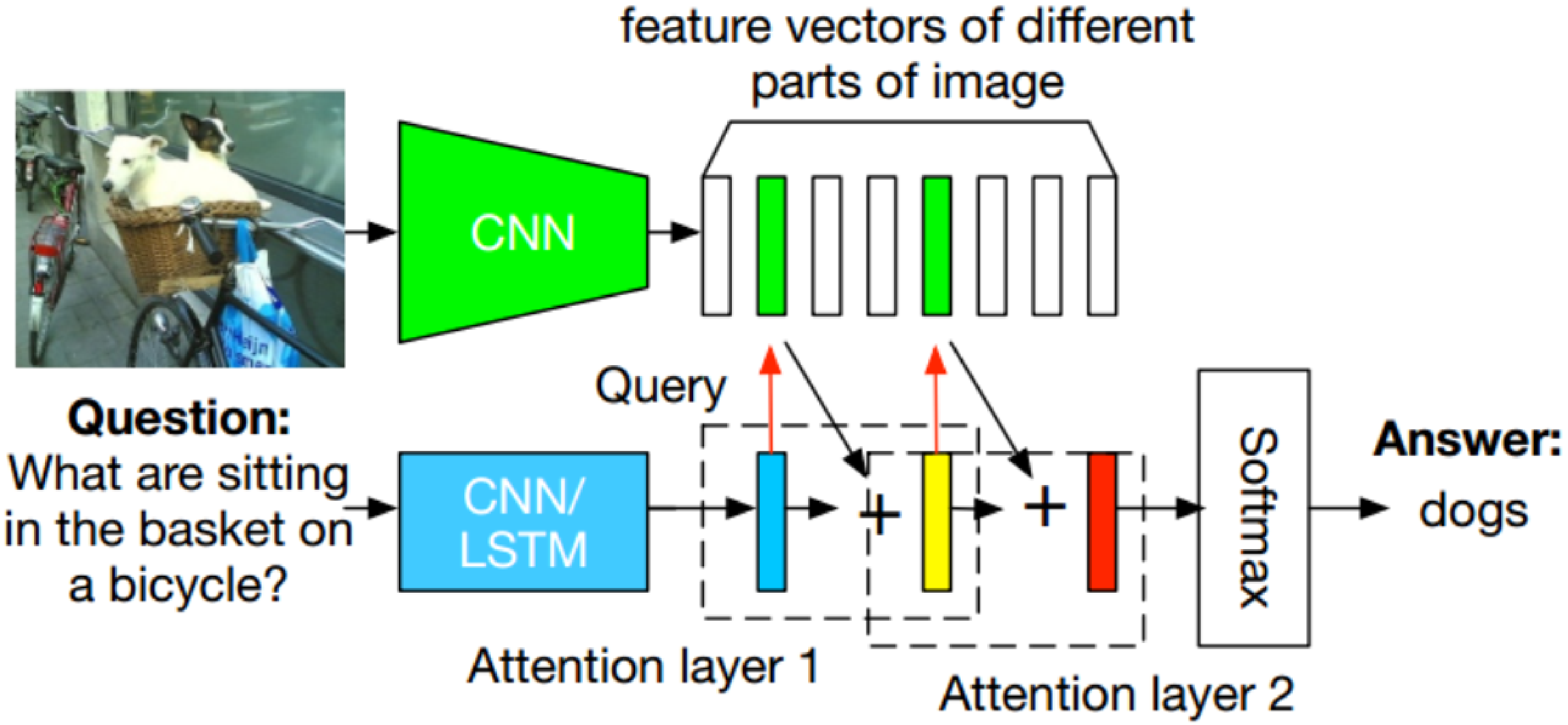
- Image features: last hidden layer (4096) of VGGNet
- Question features: LSTM (1024)
- Multimodality: EWM
- Attention mechanism: Image + Question
- Test-dev accuracy: **61.8%**
- Test-std accuracy: **62.1%**

Hierarchical Co-Attention

Lu et al. (2016)


Q: how many snowboarders in formation in the snow, four is sitting? A: 5
how many snowboarders in formation in the snow , four is sitting ?
how many snowboarders in formation in the snow , four is sitting ?
how many snowboarders in formation in the snow , four is sitting ?

Stacked Attention Networks (SAN) Yang et al. (2015)


feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

Stacked Attention Networks (SAN) Yang et al. (2015)

$$\tilde{h}_c = \max[h_{c,1}, h_{c,2}, \ldots, h_{c,T-c+1}]$$

$$h = [\tilde{h}_1, \tilde{h}_2, \tilde{h}_3]$$

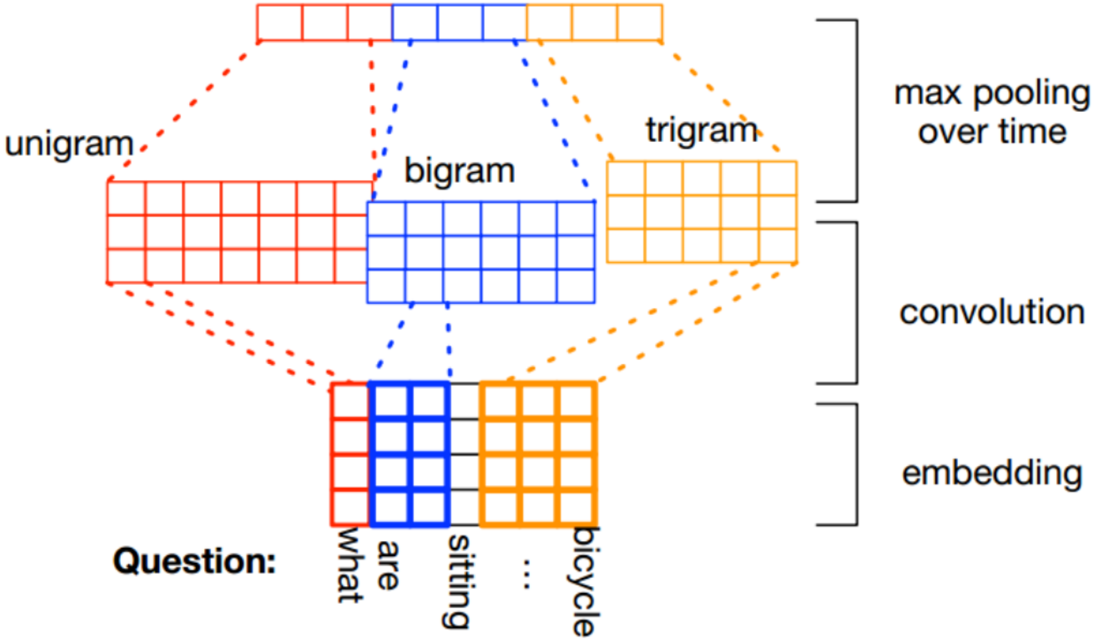
$$h_A^k = \tanh(W_{I,A}^k v_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)$$

$$p_I^k = softmax(W_p^k h_A^k + b_p^k)$$

$$\tilde{v}_I^k = \sum p_i^k v_i$$

$$u^k = \tilde{v}_I^k + u^{k-1}$$

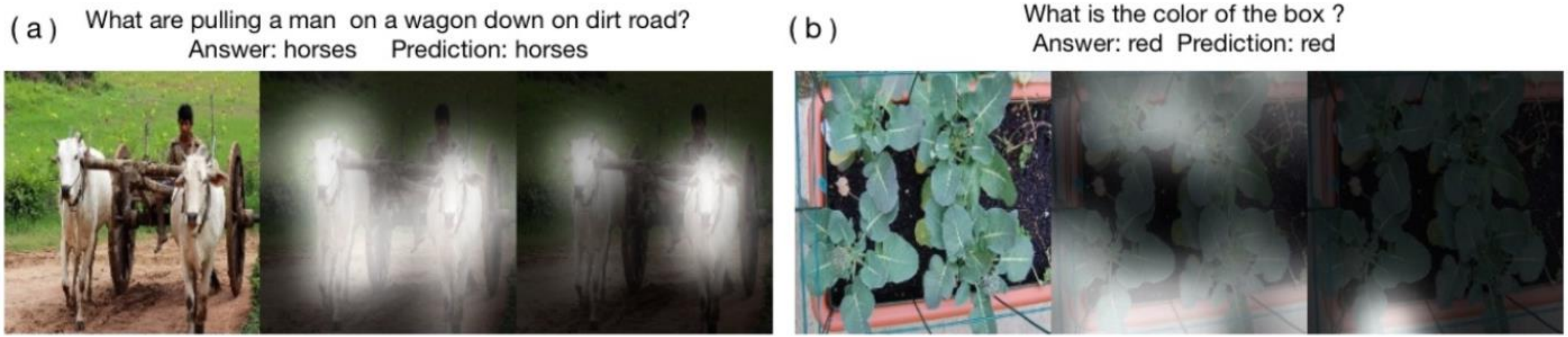
$$p_{ans} = \text{softmax}(W_u u^K + b_u)$$



Stacked Attention Networks (SAN) Yang et al. (2015)

- Image features: VGGNet (512)
- Question features: CNN (1280)
- Multimodality: EWM
- Attention mechanism: Question
- Test-dev accuracy: **58.7%**
- Test-std accuracy: **58.9%**

Stacked Attention Networks (SAN) Yang et al. (2015)

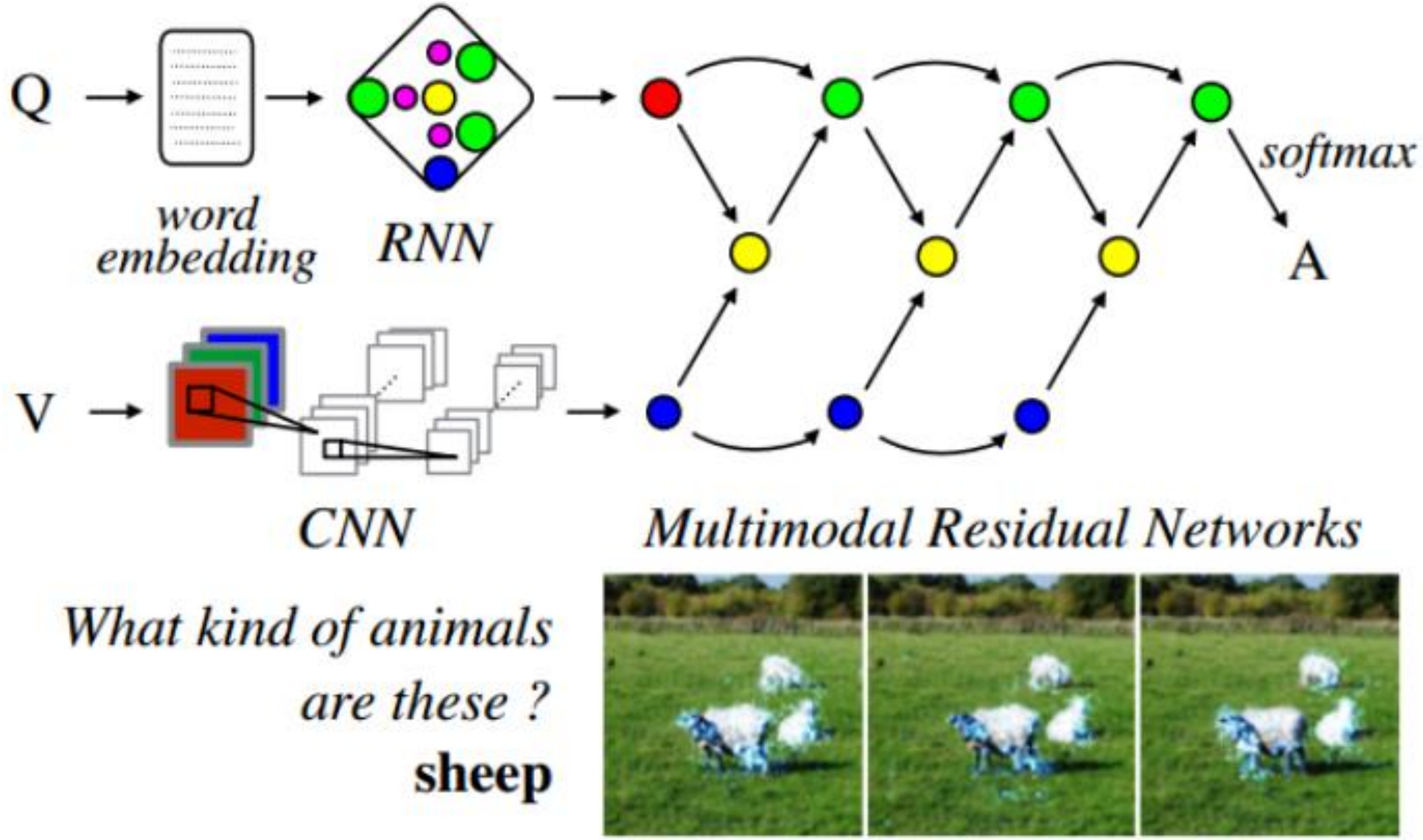


What are pulling a man on a wagon down on dirt road?
Answer: horses Prediction: horses

What is the color of the box ?
Answer: red Prediction: red

Multimodal Residual Network (MRN)

Kim et al. (2016)


Q
word embedding
RNN
softmax
A
V
CNN
Multimodal Residual Networks
What kind of animals are these ?
sheep

Multimodal Residual Network (MRN)

Kim et al. (2016)

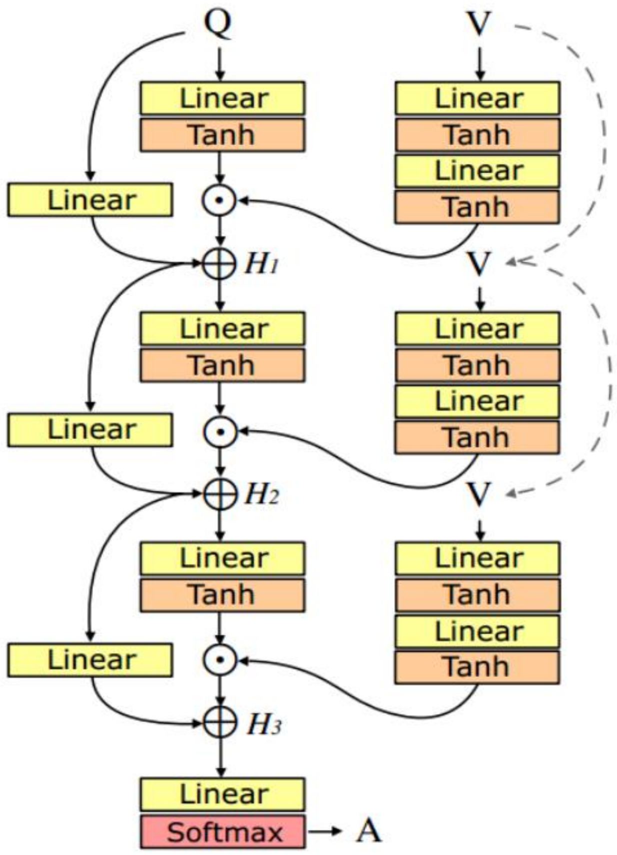
$$\boldsymbol{q}^k = F(\boldsymbol{q}^{k-1}, \boldsymbol{V}) + \boldsymbol{q}^{k-1}$$

$$F^{(k)}(\boldsymbol{q}, \boldsymbol{v}) = \tanh(\boxed{W_{\boldsymbol{q}}^{(k)} \boldsymbol{q}}) \odot \tanh(W_2^{(k)} \tanh(W_1^{(k)} \boldsymbol{v}))$$

$$H_L(\boldsymbol{q}, \boldsymbol{v}) = W_{\boldsymbol{q}'} \boldsymbol{q} + \sum_{l=1}^{L} W_{F^{(l)}} F^{(l)}(H_{l-1}, \boldsymbol{v})$$

$$H_0 = \boldsymbol{q}$$

$$W_{\boldsymbol{q}'} = \prod_{l=1}^{L} W_{\boldsymbol{q}'}^{(l)} \qquad W_{F^{(l)}} = \prod_{m=l+1}^{L} W_{\boldsymbol{q}'}^{(m)}$$



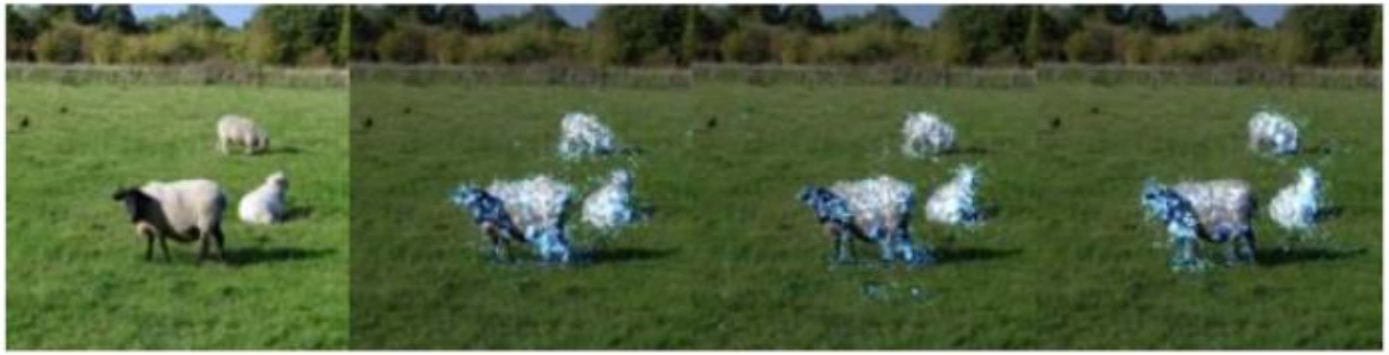
Multimodal Residual Network (MRN)

Kim et al. (2016)

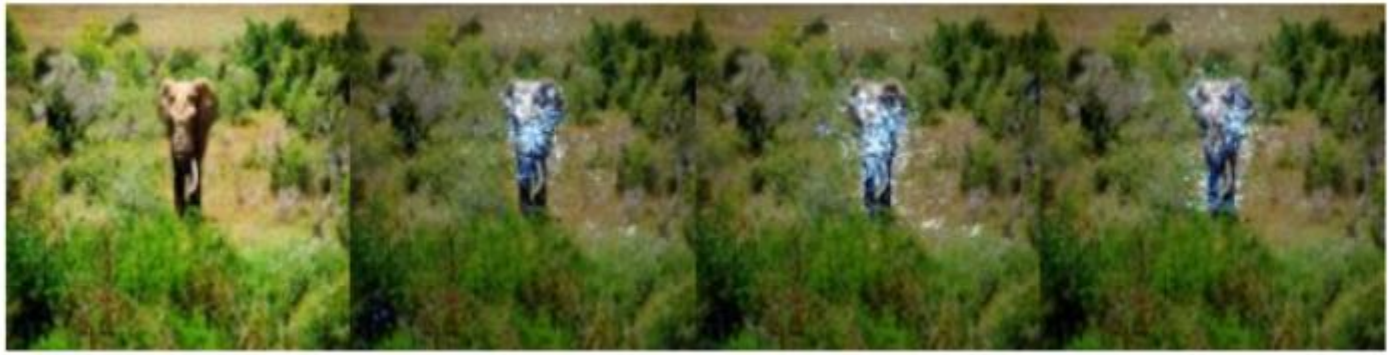
- Image features: last hidden layer (4096) of VGGNet
- Question features: GRU (2400)
- Multimodality: EWM
- Attention mechanism: Implicit
- Test-dev accuracy: **60.5%**
- Test-std accuracy: **61.8%**

Multimodal Residual Network (MRN)

Kim et al. (2016)



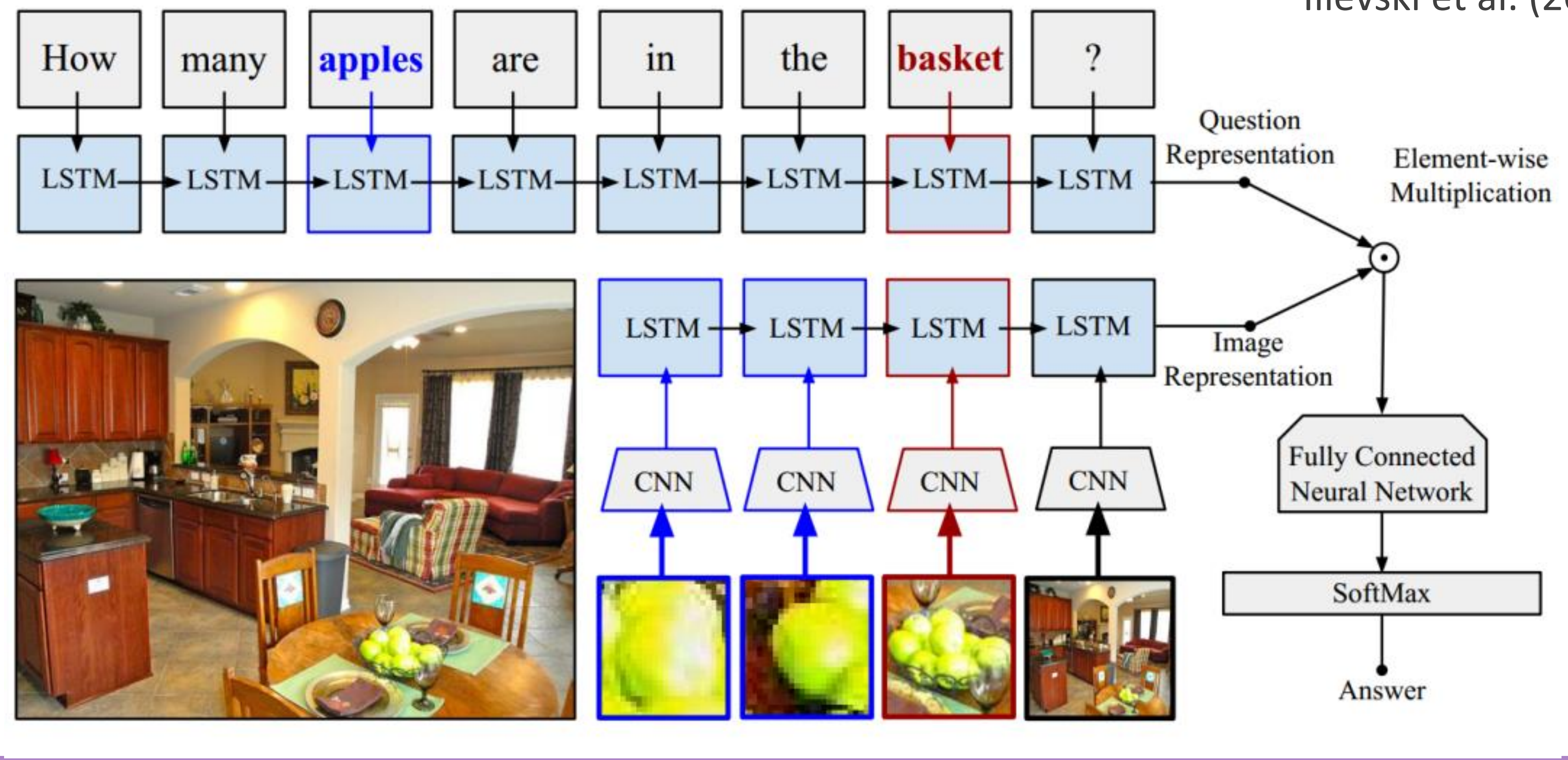
(a) *What kind of animals are these ?* **sheep**



(b) *What animal is the picture ?* **elephant**

Focused Dynamic Attention (FDA)

Ilievski et al. (2016)



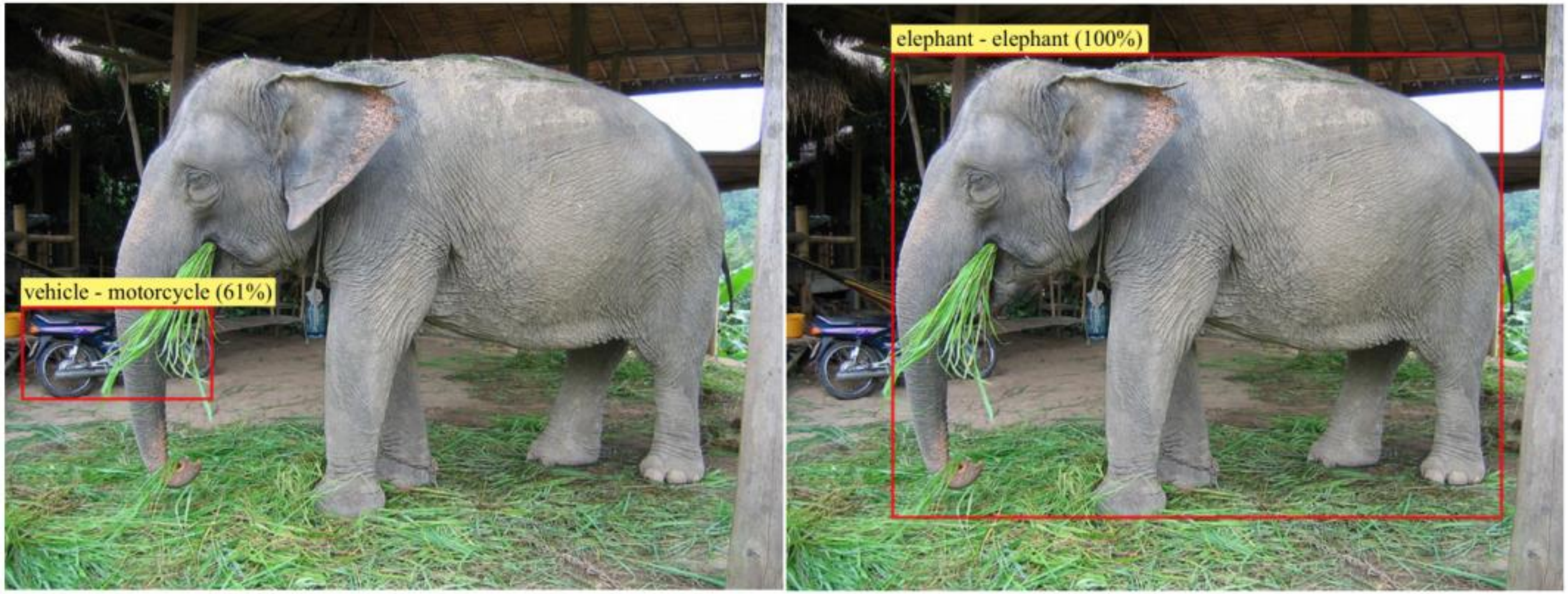
Focused Dynamic Attention (FDA)

Ilievski et al. (2016)

- Image features: ResNet (2048)
- Question features: LSTM (512)
- Multimodality: EWM
- Attention mechanism: Object detection
- Test-dev accuracy: **59.2%**
- Test-std accuracy: **59.5%**

Focused Dynamic Attention (FDA)

Ilievski et al. (2016)

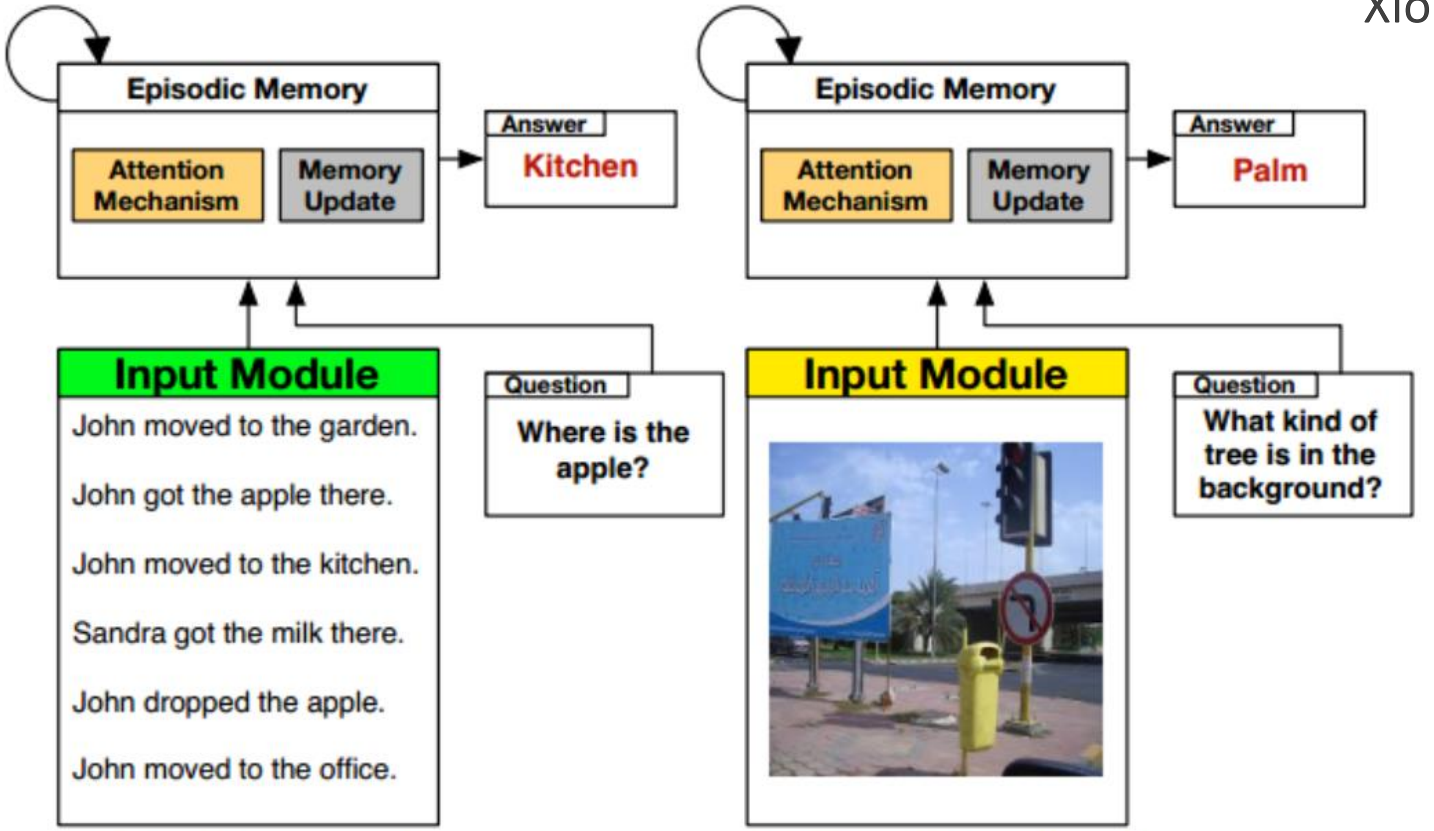


What type of **vehicle** is pictured?
- Motorcycle.

Does the **elephant** have tusks?
- No.

Dynamic Memory Networks (DMN)

Xiong et al. (2016)



(a) Text Question-Answering (b) Visual Question-Answering

Dynamic Memory Networks (DMN)

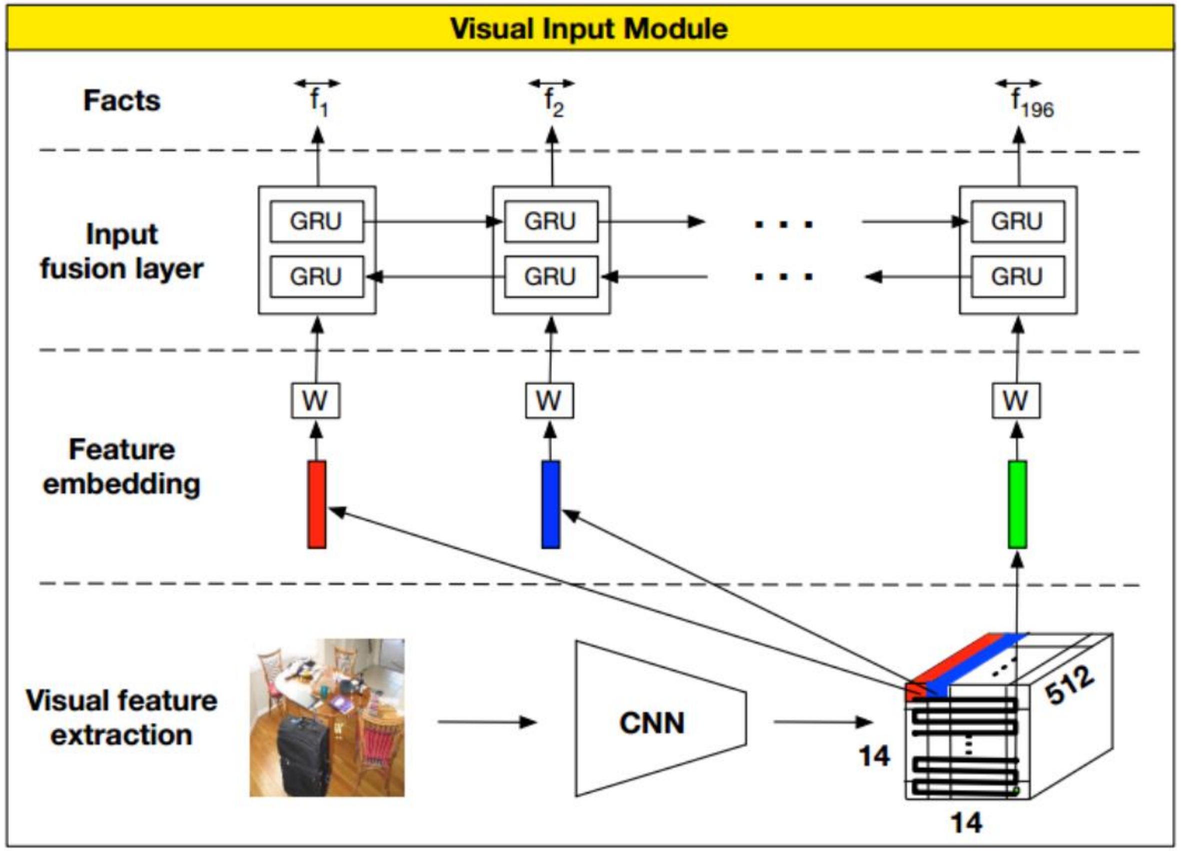
Xiong et al. (2016)

Visual Input Module

$$\overrightarrow{f_i} = GRU_{fwd}(f_i, \overrightarrow{f_{i-1}})$$

$$\overleftarrow{f_i} = GRU_{bwd}(f_i, \overleftarrow{f_{i+1}})$$

$$\overleftrightarrow{f_i} = \overleftarrow{f_i} + \overrightarrow{f_i}$$


Visual Input Module
Facts
$\overleftrightarrow{f_1}$
$\overleftrightarrow{f_2}$
$\overleftrightarrow{f_{196}}$
Input fusion layer
GRU
GRU
GRU
GRU
GRU
GRU
W
W
W
Feature embedding
Visual feature extraction
CNN
14
14
512

Dynamic Memory Networks (DMN)

Xiong et al. (2016)

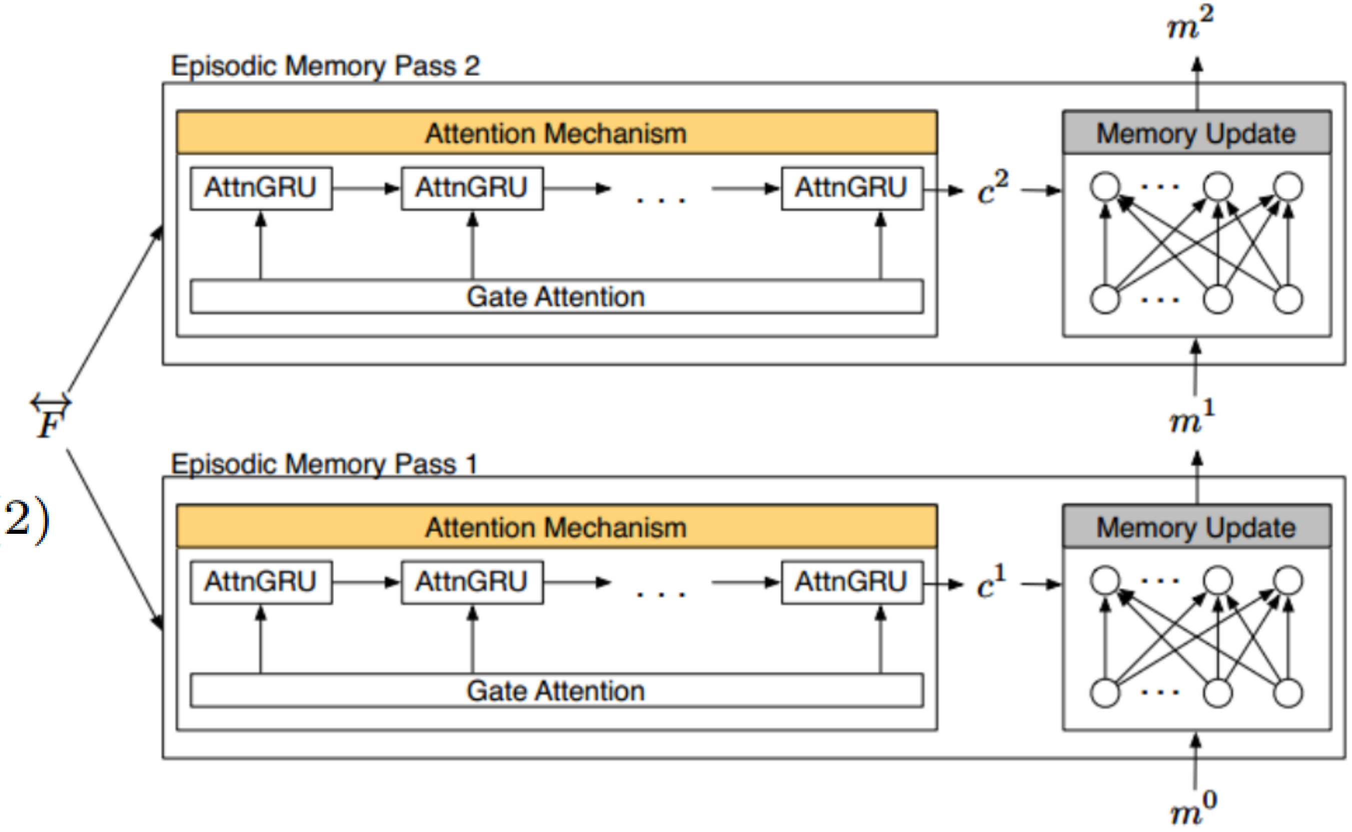
Episodic Memory

$$
\begin{aligned}
z_i^t = [\, & \overleftrightarrow{f_i} \odot q;\ \overleftrightarrow{f_i} \odot m^{t-1}; \\
& |\overleftrightarrow{f_i} - q|; |\overleftrightarrow{f_i} - m^{t-1}|]
\end{aligned}
$$

$$Z_i^t = W^{(2)} \tanh\left(W^{(1)} z_i^t + b^{(1)}\right) + b^{(2)}$$

$$g_i^t = \text{softmax}(Z_i^t)$$

$$m^t = ReLU(W^t[m^{t-1}; c^t; q] + b)$$



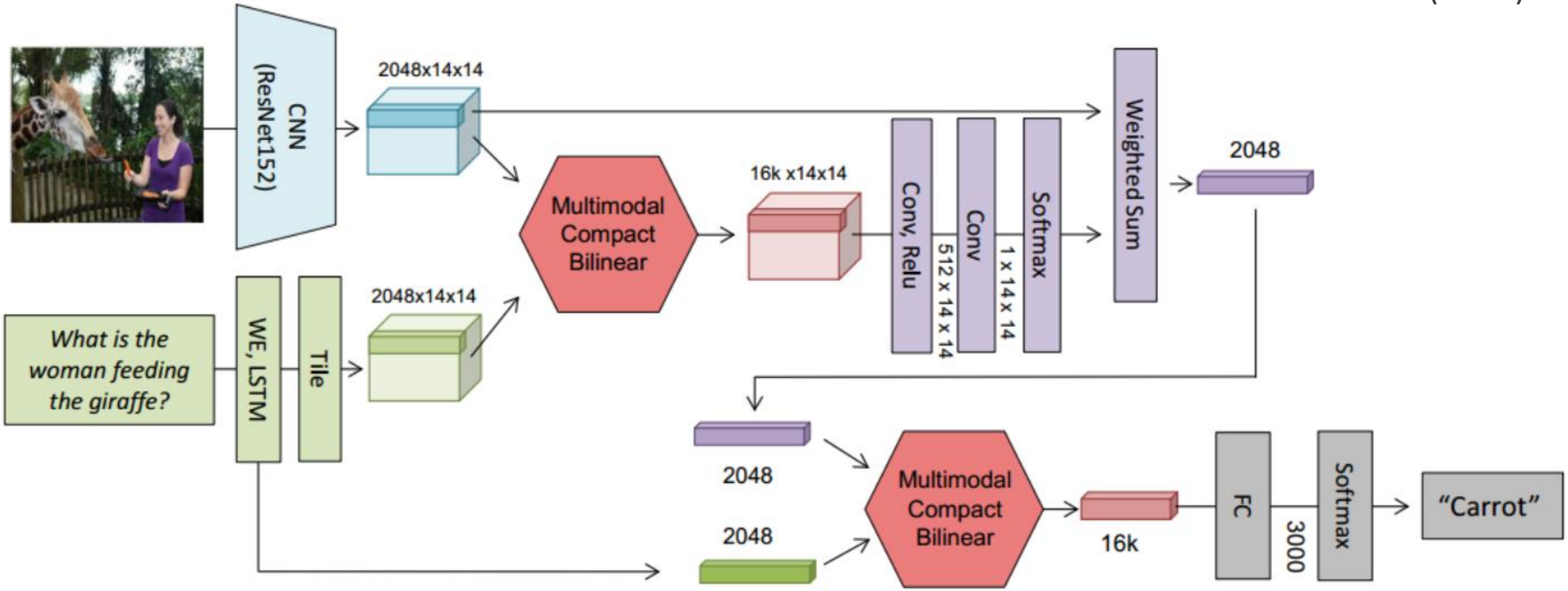
Dynamic Memory Networks (DMN)

Xiong et al. (2016)

- Image features: VGGNet (512)
- Question features: GRU (512)
- Multimodality: EWM + concatenation
- Attention mechanism: Episodic memory
- Test-dev accuracy: **60.3%**
- Test-std accuracy: **60.4%**

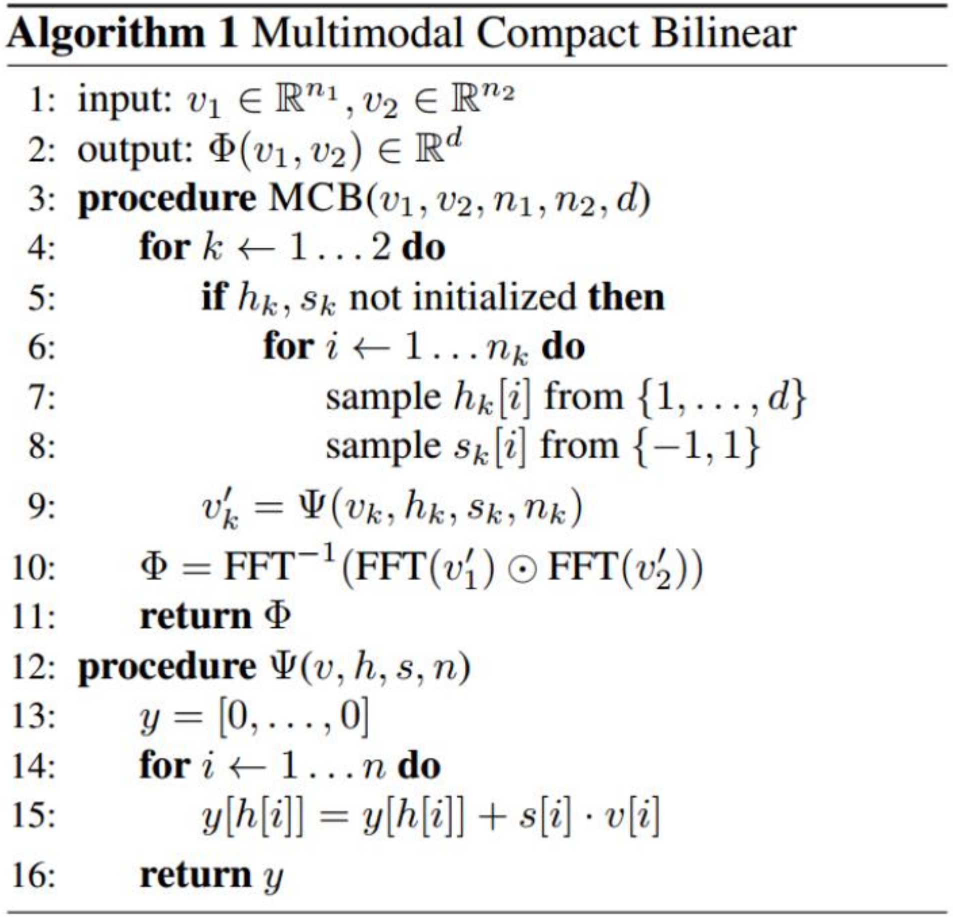
Multimodal Compact Bilinear (MCB)

Fukui et al. (2016)



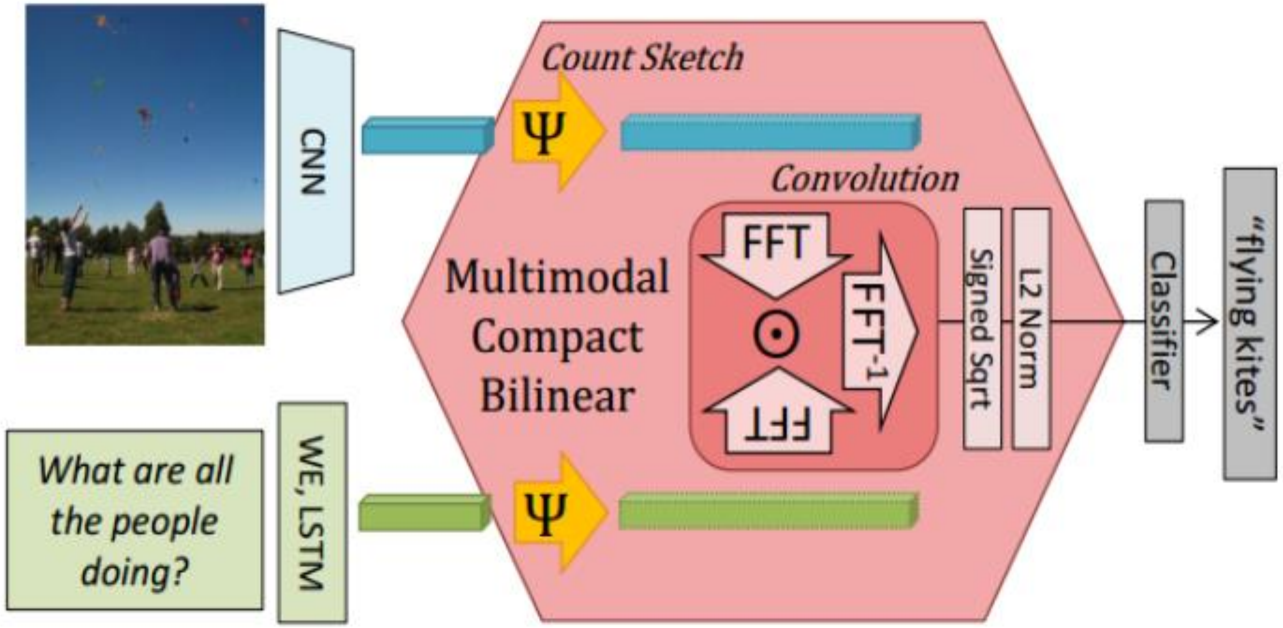
Multimodal Compact Bilinear (MCB)

Fukui et al. (2016)



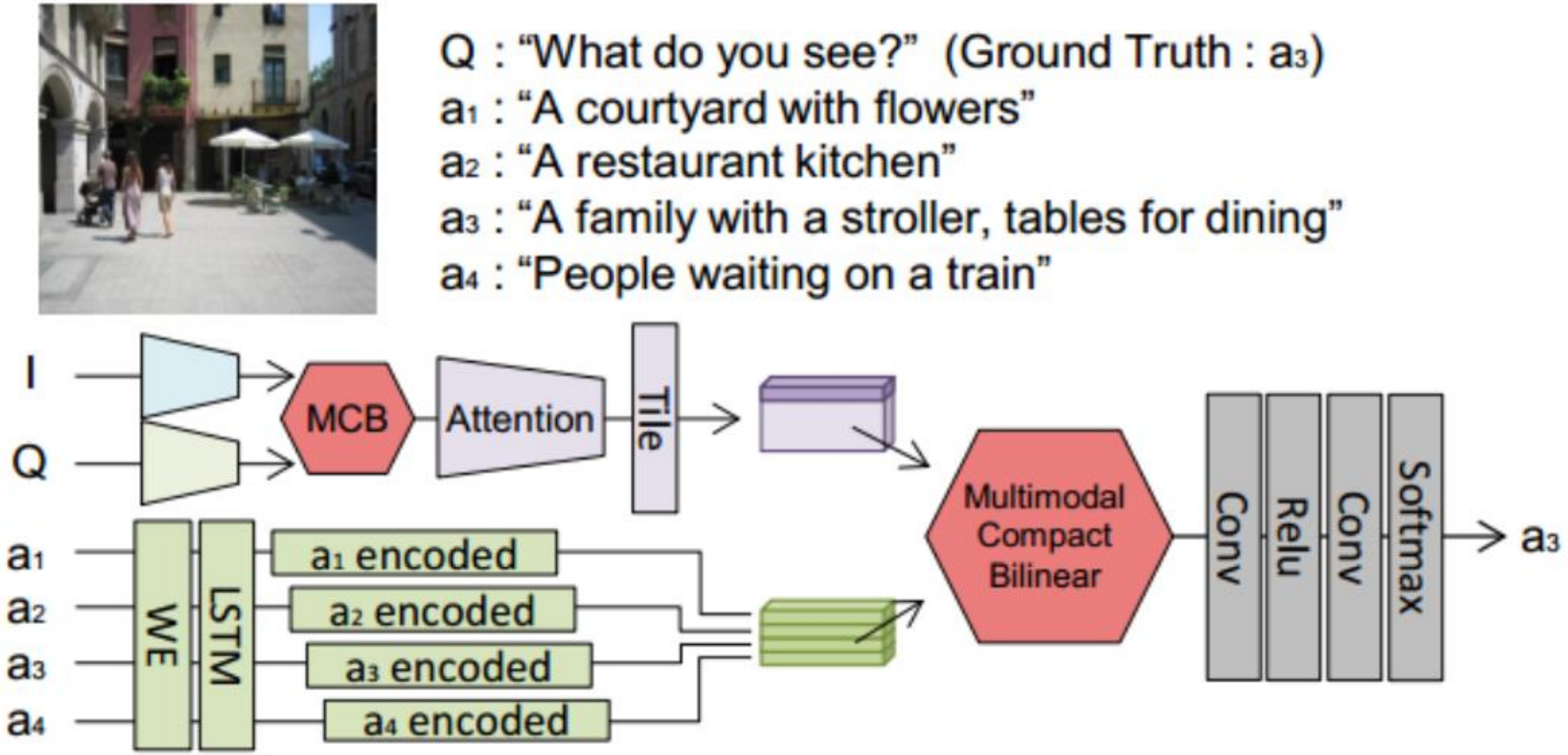
Algorithm 1 Multimodal Compact Bilinear

```
input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
procedure MCB($v_1, v_2, n_1, n_2, d$)
    for $k \leftarrow 1 \dots 2$ do
        if $h_k, s_k$ not initialized then
            for $i \leftarrow 1 \dots n_k$ do
                sample $h_k[i]$ from $\{1, \dots, d\}$
                sample $s_k[i]$ from $\{-1, 1\}$
        $v'_k = \Psi(v_k, h_k, s_k, n_k)$
    $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
    return $\Phi$
procedure $\Psi(v, h, s, n)$
    $y = [0, \dots, 0]$
    for $i \leftarrow 1 \dots n$ do
        $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
    return $y$
```



Multimodal Compact Bilinear (MCB)

Fukui et al. (2016)



Q : "What do you see?" (Ground Truth : a_3)

a_1 : "A courtyard with flowers"

a_2 : "A restaurant kitchen"

a_3 : "A family with a stroller, tables for dining"

a_4 : "People waiting on a train"

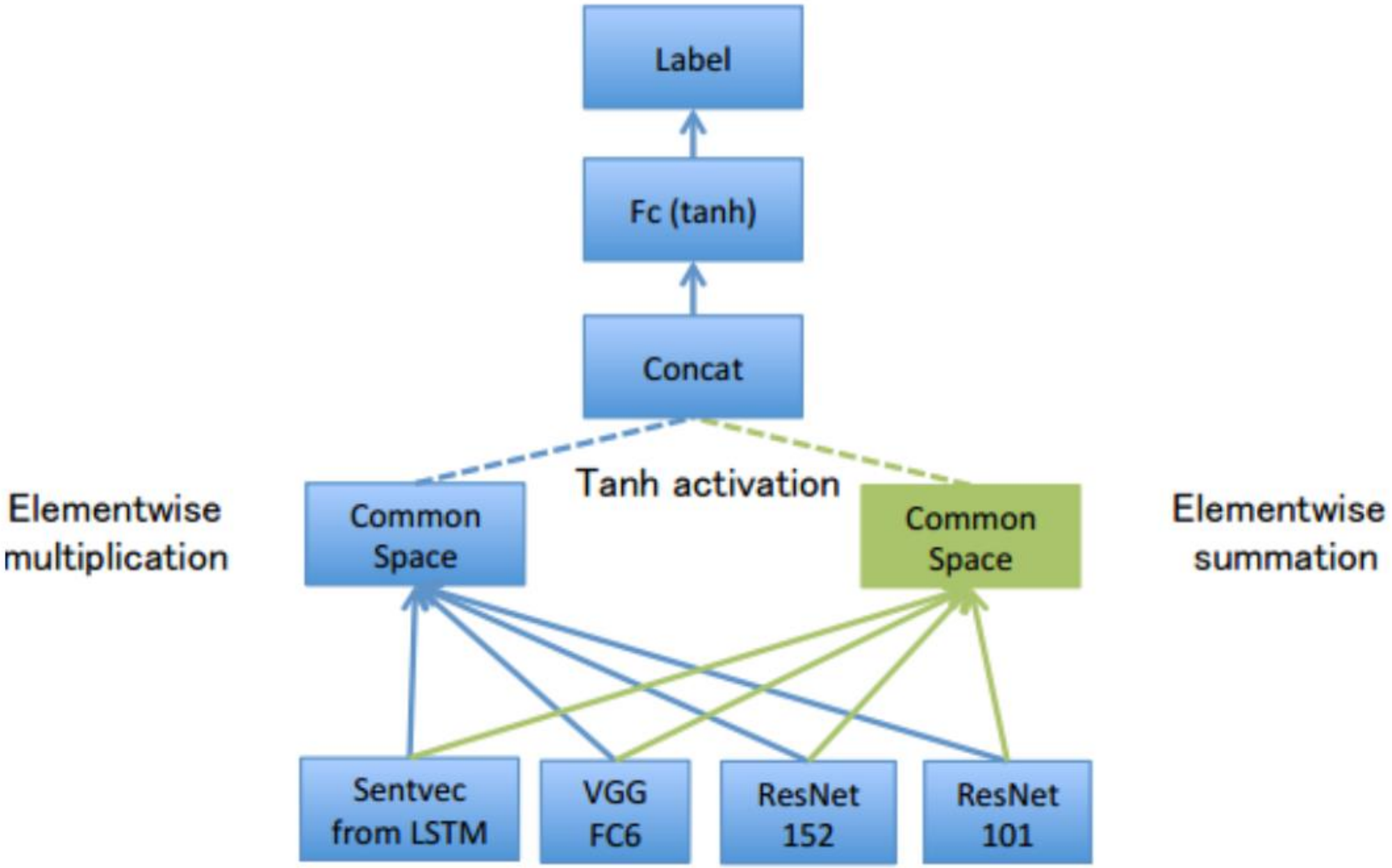
Multimodal Compact Bilinear (MCB)

Fukui et al. (2016)

- Image features: ResNet (2048)
- Question features: LSTM (1024)
- Multimodality: MCB Pooling (Count Sketch + FFT)
- Attention mechanism: MCB Pooling
- Test-dev accuracy: **66.7%**
- Test-std accuracy: **66.5%**

DualNet

Saito et al. (2016)



DualNet

Saito et al. (2016)

$$
\begin{aligned}
I_{M_1} &= \tanh(W_{M_1} I_1) \\
I_{M_2} &= \tanh(W_{M_2} I_2) \\
I_{M_3} &= \tanh(W_{M_3} I_3) \\
Q_M &= \tanh(W_{M_q} Q) \\
F_M &= I_{M_1} \odot I_{M_2} \odot I_{M_3} \odot Q_M
\end{aligned}
$$

$$
\begin{aligned}
I_{S_1} &= \tanh(W_{S_1} I_1) \\
I_{S_2} &= \tanh(W_{S_2} I_2) \\
I_{S_3} &= \tanh(W_{S_3} I_3) \\
Q_S &= \tanh(W_{S_q} Q) \\
F_S &= I_{S_1} + I_{S_2} + I_{S_3} + Q_S
\end{aligned}
$$

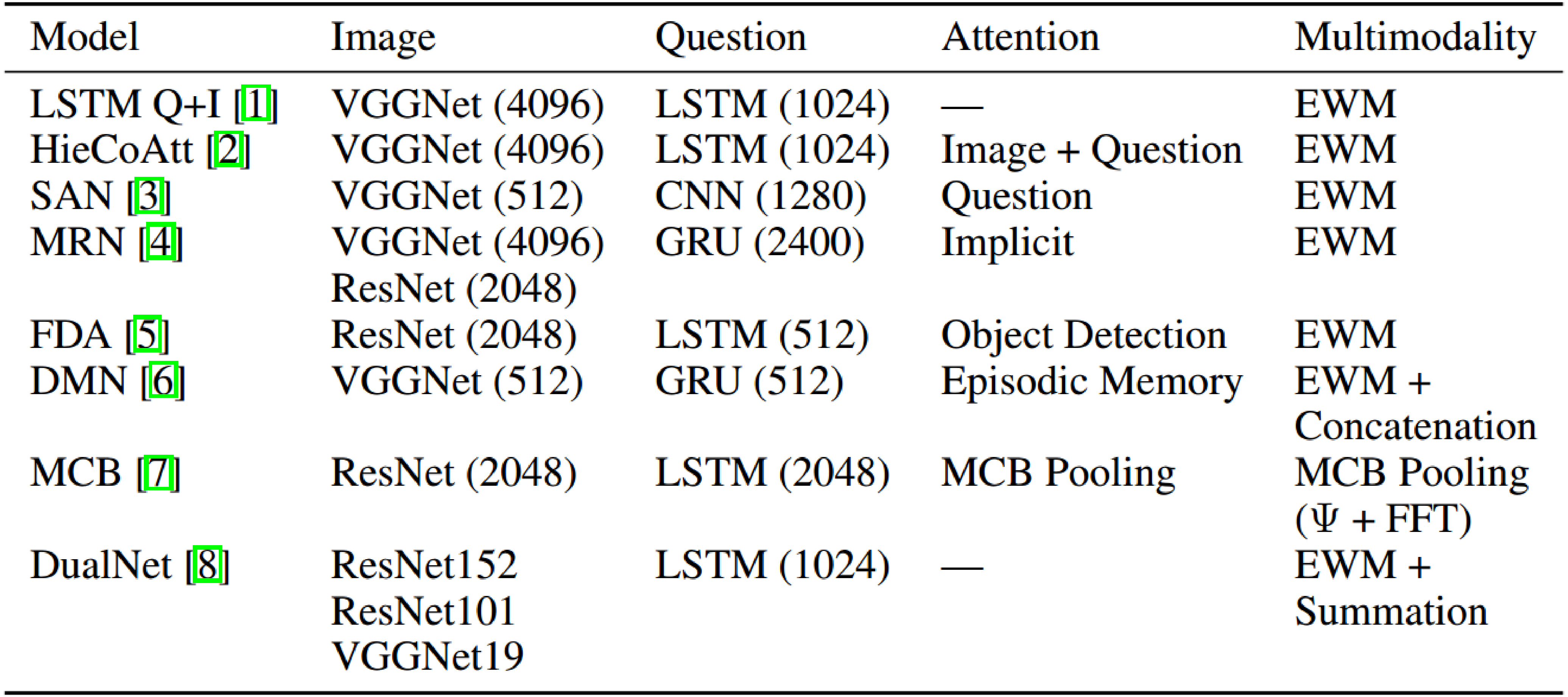
$$
\begin{aligned}
F &= [F_M, F_S] \\
Output &= W_{f_2} \tanh(W_{f_1} F)
\end{aligned}
$$

DualNet

Saito et al. (2016)

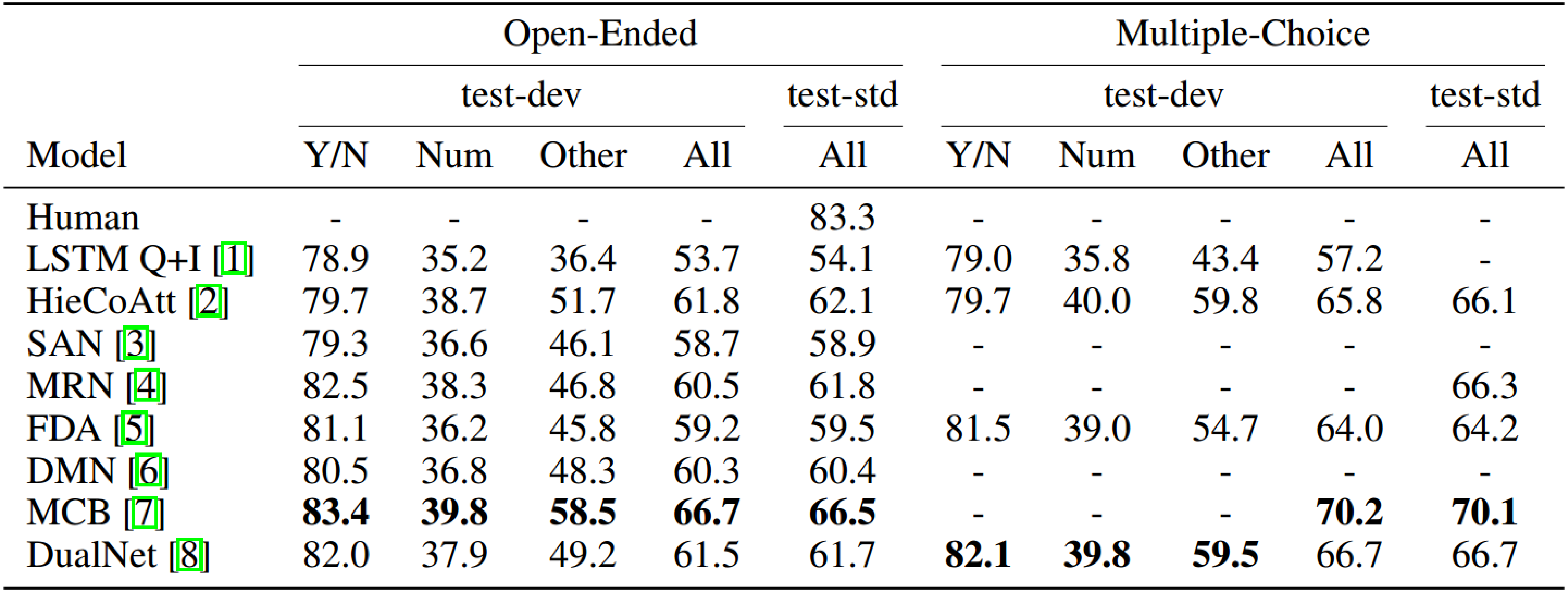
- Image features: ResNet152 + ResNet101 + VGGNet19
- Question features: LSTM (1024)
- Multimodality: EWM + Summation
- Attention mechanism: None
- Test-dev accuracy: **61.5%**
- Test-std accuracy: **61.7%**

Algorithms Comparison



Model	Image	Question	Attention	Multimodality
LSTM Q+I [1]	VGGNet (4096)	LSTM (1024)	—	EWM
HieCoAtt [2]	VGGNet (4096)	LSTM (1024)	Image + Question	EWM
SAN [3]	VGGNet (512)	CNN (1280)	Question	EWM
MRN [4]	VGGNet (4096) ResNet (2048)	GRU (2400)	Implicit	EWM
FDA [5]	ResNet (2048)	LSTM (512)	Object Detection	EWM
DMN [6]	VGGNet (512)	GRU (512)	Episodic Memory	EWM + Concatenation
MCB [7]	ResNet (2048)	LSTM (2048)	MCB Pooling	MCB Pooling (Ψ + FFT)
DualNet [8]	ResNet152 ResNet101 VGGNet19	LSTM (1024)	—	EWM + Summation

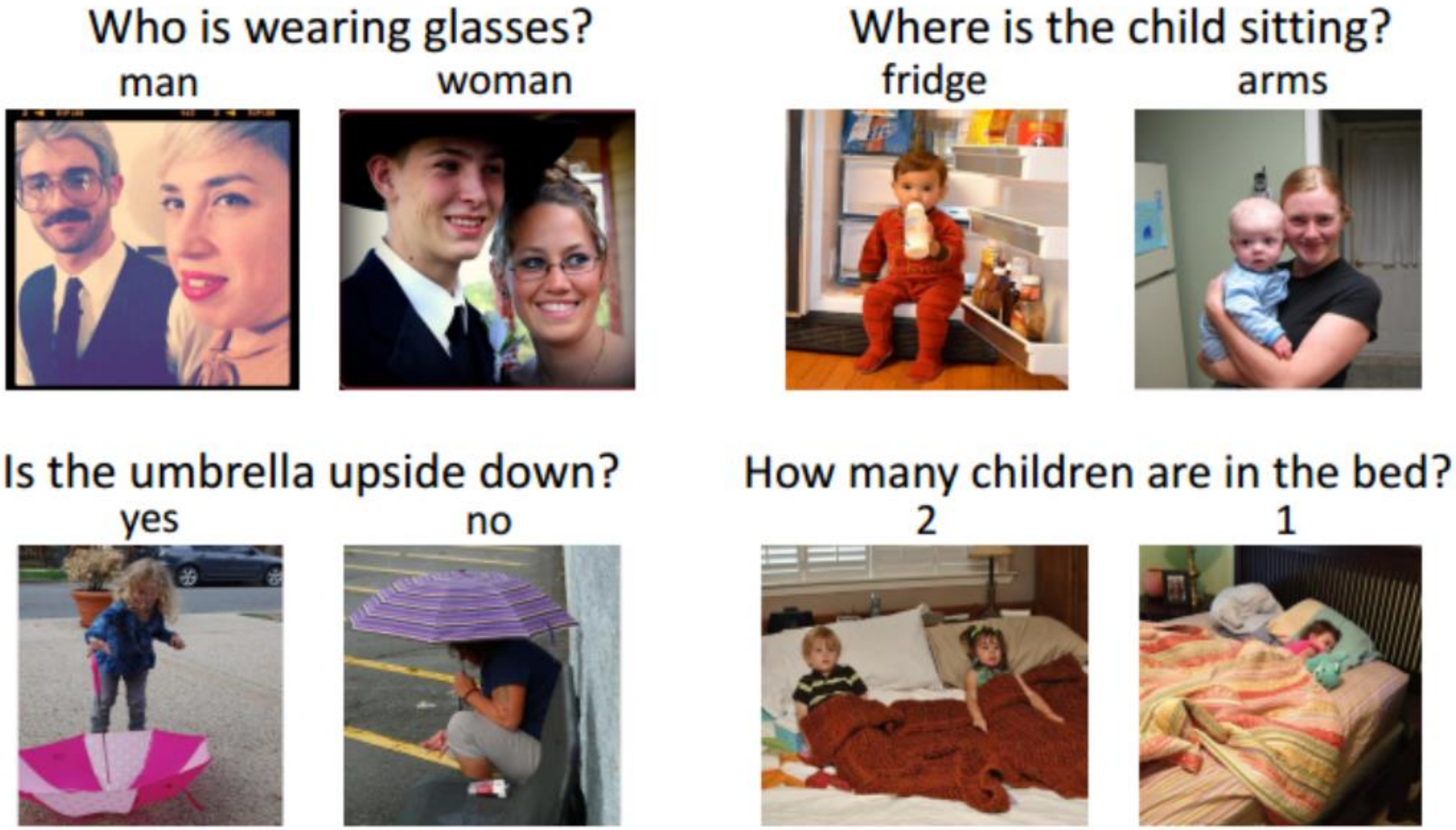
Results Comparison



	Open-Ended					Multiple-Choice				
	test-dev				test-std	test-dev				test-std
Model	Y/N	Num	Other	All	All	Y/N	Num	Other	All	All
Human	-	-	-	-	83.3	-	-	-	-	-
LSTM Q+I [1]	78.9	35.2	36.4	53.7	54.1	79.0	35.8	43.4	57.2	-
HieCoAtt [2]	79.7	38.7	51.7	61.8	62.1	79.7	40.0	59.8	65.8	66.1
SAN [3]	79.3	36.6	46.1	58.7	58.9	-	-	-	-	-
MRN [4]	82.5	38.3	46.8	60.5	61.8	-	-	-	-	66.3
FDA [5]	81.1	36.2	45.8	59.2	59.5	81.5	39.0	54.7	64.0	64.2
DMN [6]	80.5	36.8	48.3	60.3	60.4	-	-	-	-	-
MCB [7]	**83.4**	**39.8**	**58.5**	**66.7**	**66.5**	-	-	-	**70.2**	**70.1**
DualNet [8]	82.0	37.9	49.2	61.5	61.7	**82.1**	**39.8**	**59.5**	66.7	66.7

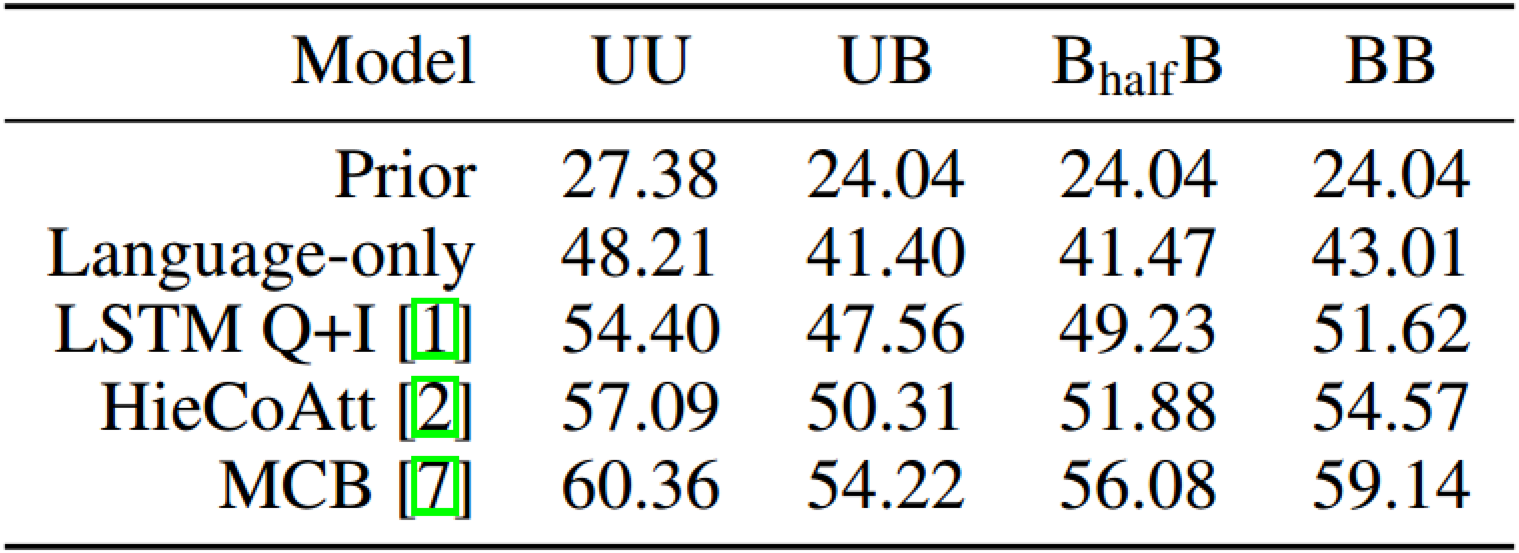
VQA 2.0

Goyal et al. (2016)


Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

VQA 2.0

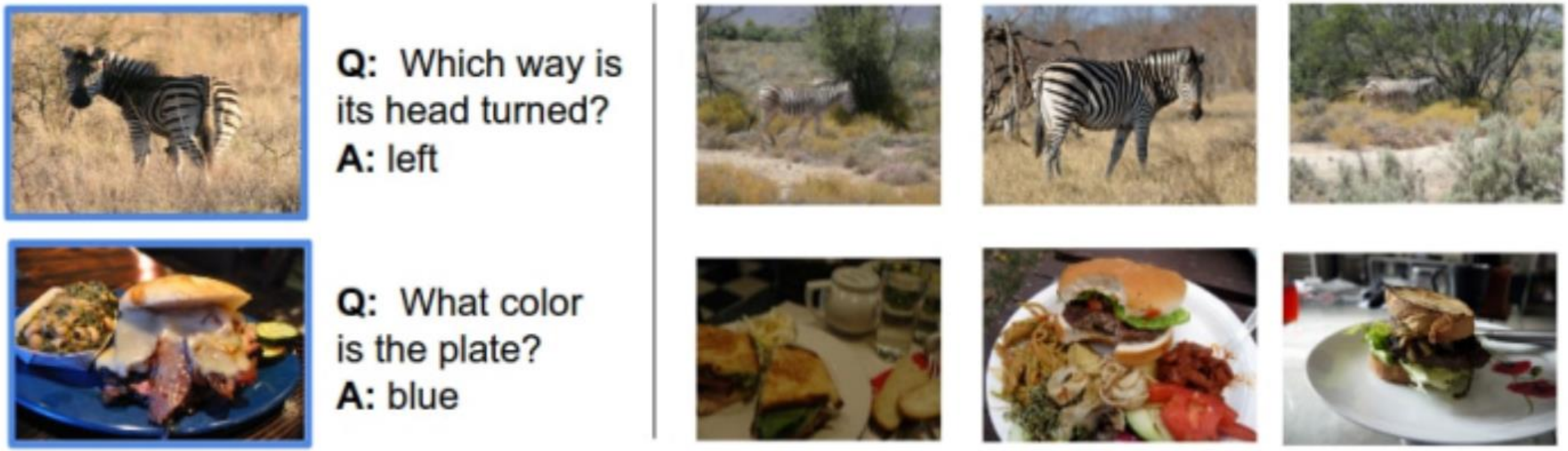
Goyal et al. (2016)



Model	UU	UB	$B_{half}B$	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
LSTM Q+I [1]	54.40	47.56	49.23	51.62
HieCoAtt [2]	57.09	50.31	51.88	54.57
MCB [7]	60.36	54.22	56.08	59.14

VQA 2.0

Goyal et al. (2016)



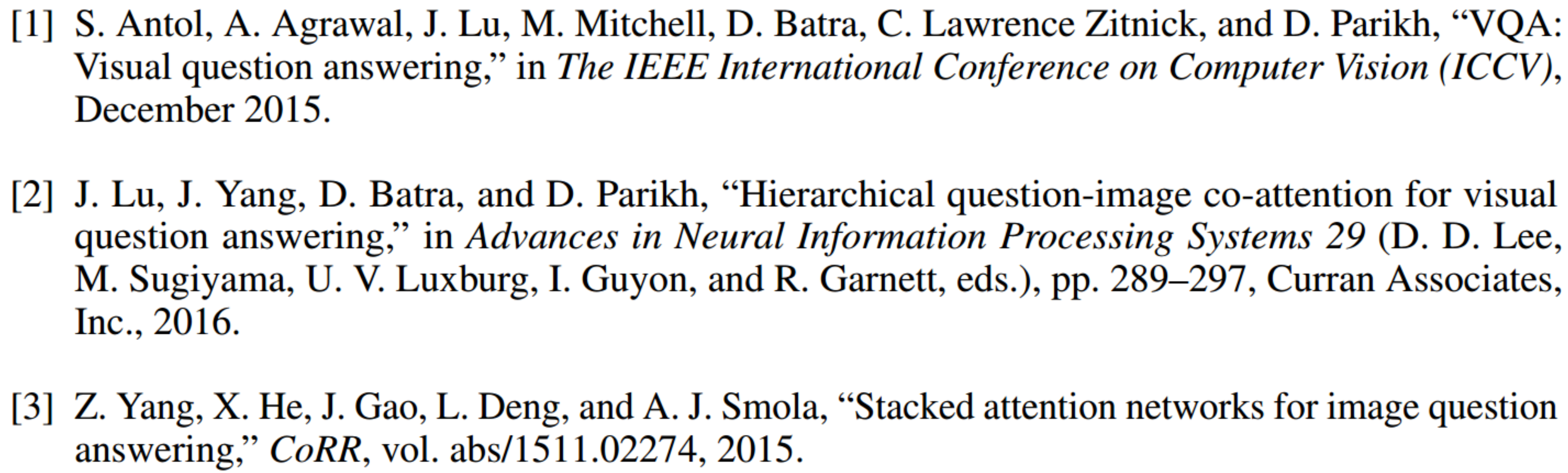
Q: Which way is its head turned?
A: left

Q: What color is the plate?
A: blue

Future Work

- Counting mechanisms
- Apply MCB to other models
- Kernel methods
- Ensemble methods
- Emphasize image understanding (VQA 2.0)

References

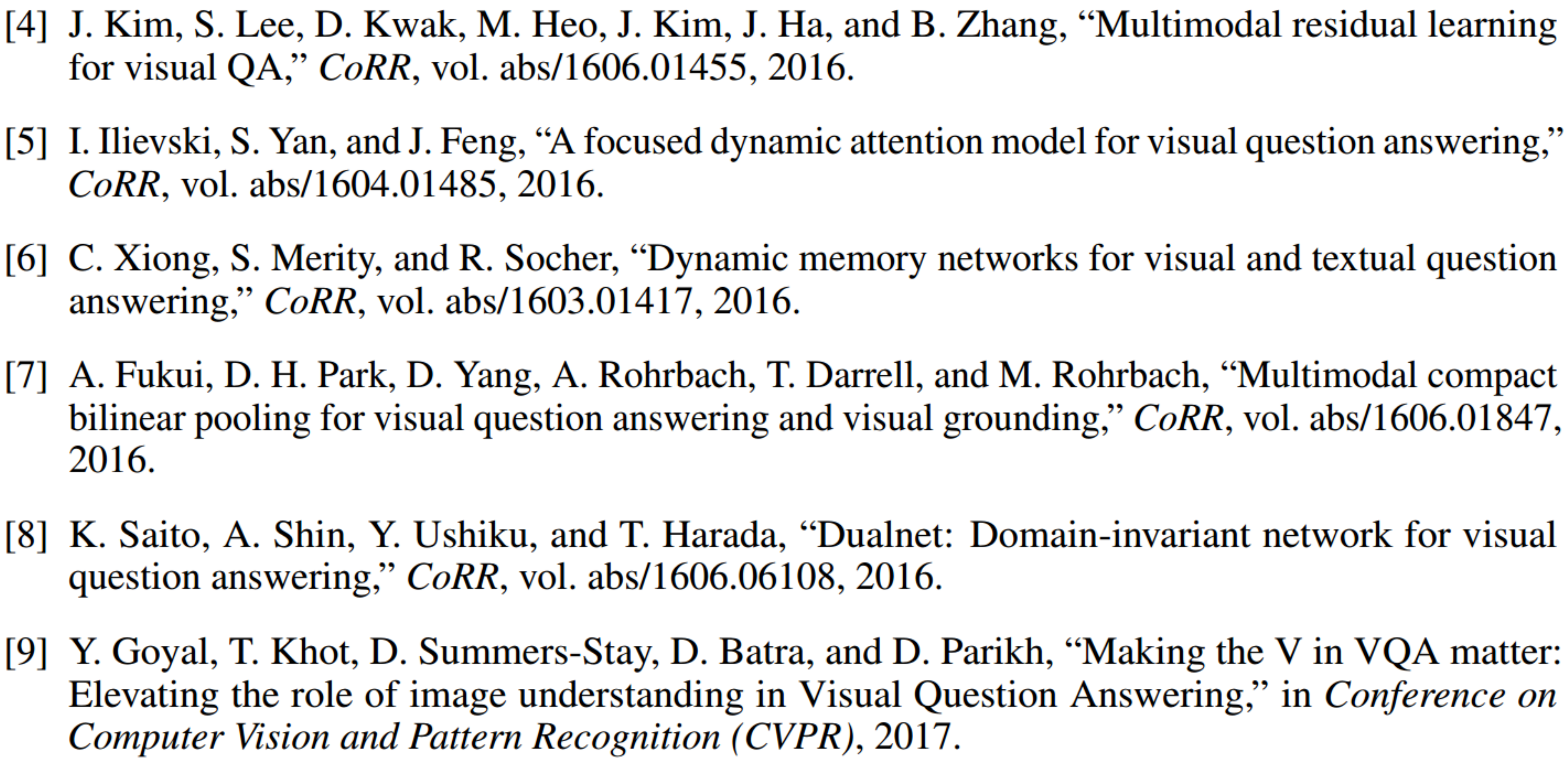


[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “VQA: Visual question answering,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[2] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 289–297, Curran Associates, Inc., 2016.

[3] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, “Stacked attention networks for image question answering,” *CoRR*, vol. abs/1511.02274, 2015.

References



[4] J. Kim, S. Lee, D. Kwak, M. Heo, J. Kim, J. Ha, and B. Zhang, “Multimodal residual learning for visual QA,” *CoRR*, vol. abs/1606.01455, 2016.

[5] I. Ilievski, S. Yan, and J. Feng, “A focused dynamic attention model for visual question answering,” *CoRR*, vol. abs/1604.01485, 2016.

[6] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” *CoRR*, vol. abs/1603.01417, 2016.

[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” *CoRR*, vol. abs/1606.01847, 2016.

[8] K. Saito, A. Shin, Y. Ushiku, and T. Harada, “Dualnet: Domain-invariant network for visual question answering,” *CoRR*, vol. abs/1606.06108, 2016.

[9] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Thank You!