Visual Question Answering

Liang-Wei Chen, Shuai Tang

Outline

- ❑ Problem statement
- ❑ Common VQA benchmark datasets
- ❑ Methods
- ❑ Dataset bias problem
- ❑ Future Development

Problem Statement

What is VQA?

Given an image, can our machine answer the corresponding questions in natural language?

VQA Demo

CloudCV: Visual Question Answering (VQA), https://cloudcv.org/vqa/

How to see and how to read

Question: What can the red object on the ground be used for ?
Answer: Firefighting

Natural Language Understanding

Wang, Peng, et al. "Fvqa: Fact-based visual question answering." *arXiv preprint arXiv:1606.05433* (2016).

How to see and how to read

Wang, Peng, et al. "Fvqa: Fact-based visual question answering." *arXiv preprint arXiv:1606.05433* (2016).

How to see and how to read

Computer Vision

Natural Language Understanding

Knowledge Representation and reasoning

Question: What can the red object on the ground be used for ?

Answer: Firefighting

Support Fact: Fire hydrant can be used for fighting fires.

Wang, Peng, et al. "Fvqa: Fact-based visual question answering." *arXiv preprint arXiv:1606.05433* (2016).

Multiple choices V.S. Open-ended settings

How many cats are there?

(a) One
(b) Two
(c) Three
(d) Four

Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

Common VQA Benchmark Datasets

VQA dataset

VQA-real

Q: Where are the magazines in this picture ?

A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

VQA-abstract

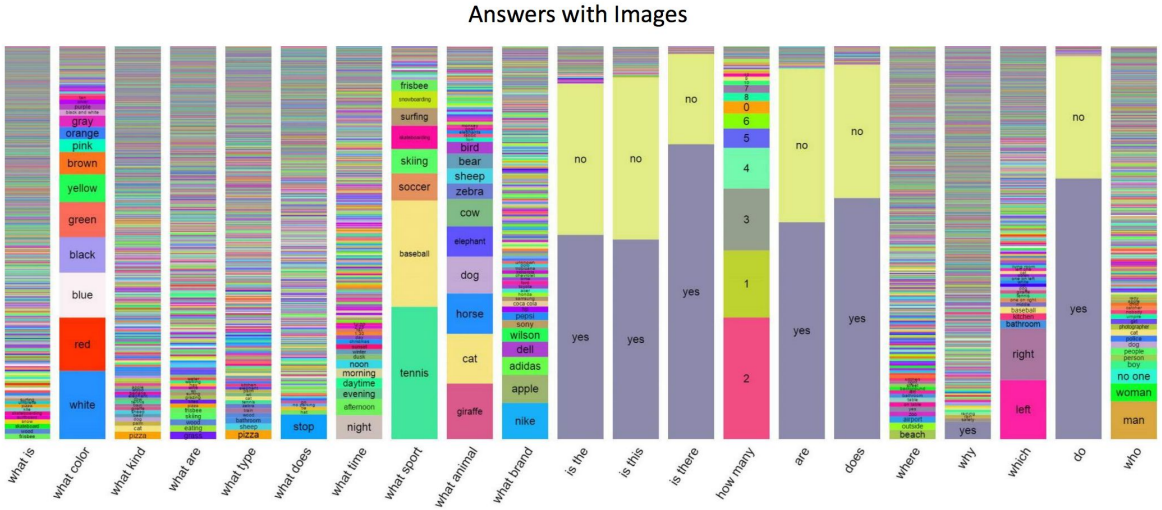
Q: Where are the flowers ?

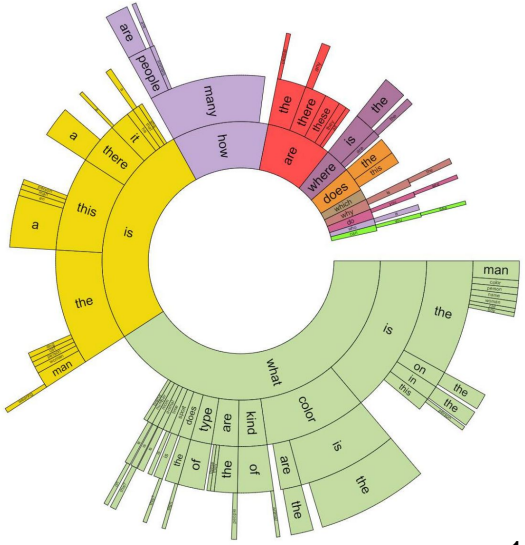
A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree

Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

VQA dataset

- Offer open-ended answers and multiple choice answers
- 250k images(MS COCO + 50k abstract images)
- 750k questions , 10M answers
- Each question is answered by 10 human annotators





Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

Microsoft COCO-QA

- Automatically generate QA pairs with MS COCO captions
- 123,287 images (72,783 for training and 38,948 for testing) and each image has one QA pair.
- 4 types of question templates: What object, How many, What color, Where

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three

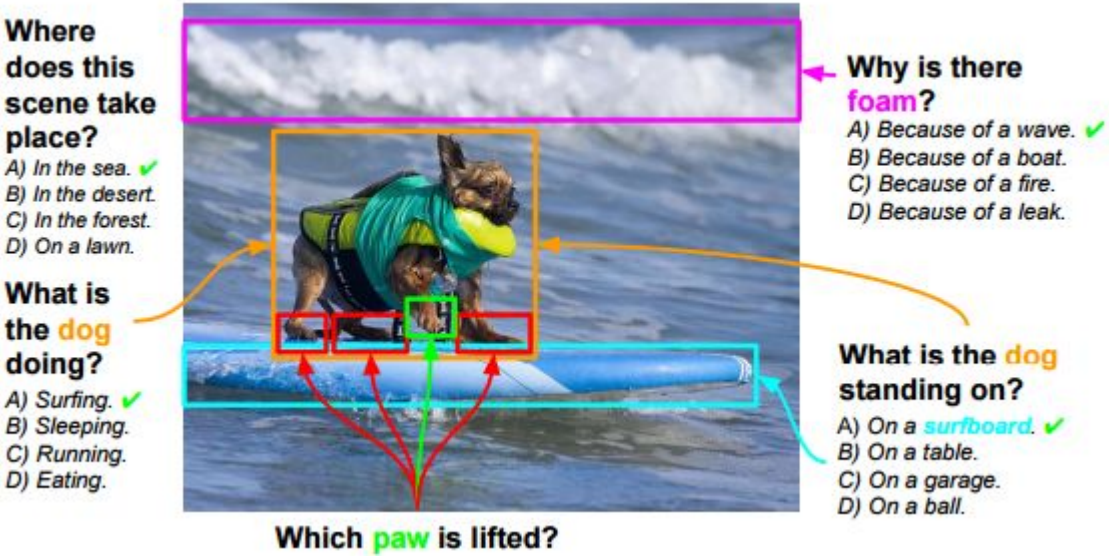
COCOQA 1238
What is the color of the tee-shirt?
Ground truth: blue

Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." *Advances in Neural Information Processing Systems*. 2015.

Visual7W

- 47,300 COCO images, 327,939 QA pairs, and 1,311,756 human-generated multiple-choices
- 7W stands for what, where, when, who, why, how and which

Telling *Pointing*



Zhu, Yuke, et al. "Visual7w: Grounded question answering in images." *Proceedings of the IEEE CVPR*. 2016.

Evaluation metrics

Recall metrics for image captioning, BLEU, etc.

- ❑ Emphasize on similarity between sentences.

Major VQA datasets use Accuracy:

- ❑ Exact string matching.
- ❑ Most answers are 1 to 3 words.

VQA dataset (with 10 human annotations):

$$\text{accuracy} = \min\left(\frac{\#\text{ humans provided that answer}}{3}, 1\right)$$

A brief summary over VQA datasets

Dataset	Number of QA pairs	Annotation	Question Diversity	Answer Diversity (top-1000 answers coverage)
COCO-QA	117,684	Open-ended	Low (generated from captions)	100%
VQA	614,163	Open-ended +Multiple choices	Encourage to have diverse annotations	82.7%
Visual 7W	327,939	Multiple choices	At least 3 W for each image	63.5%

Zhu, Yuke, et al. "Visual7w: Grounded question answering in images." *Proceedings of the IEEE CVPR*. 2016.

Methods

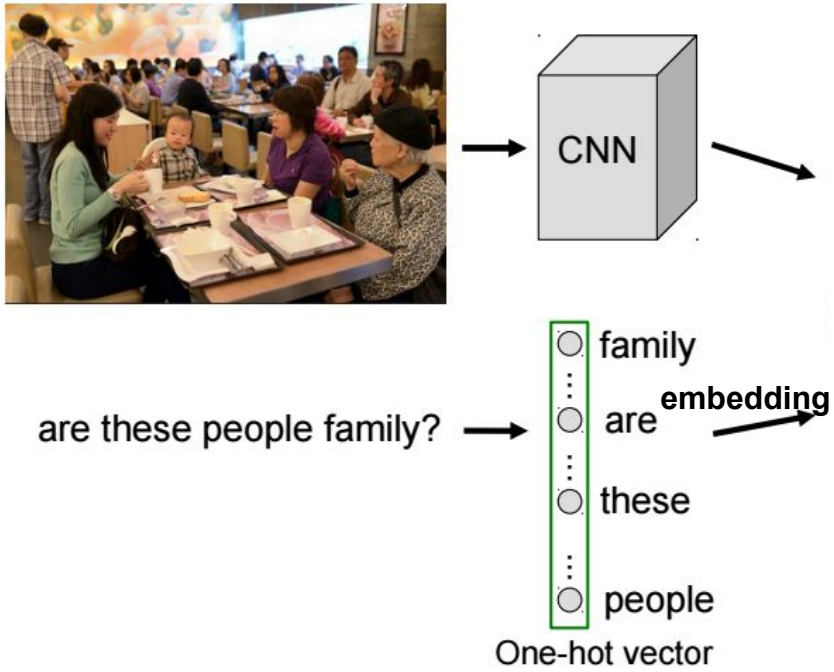
Language-Image Embedding

Bag-of-words + Image feature (iBOWIMG)

are these people family?

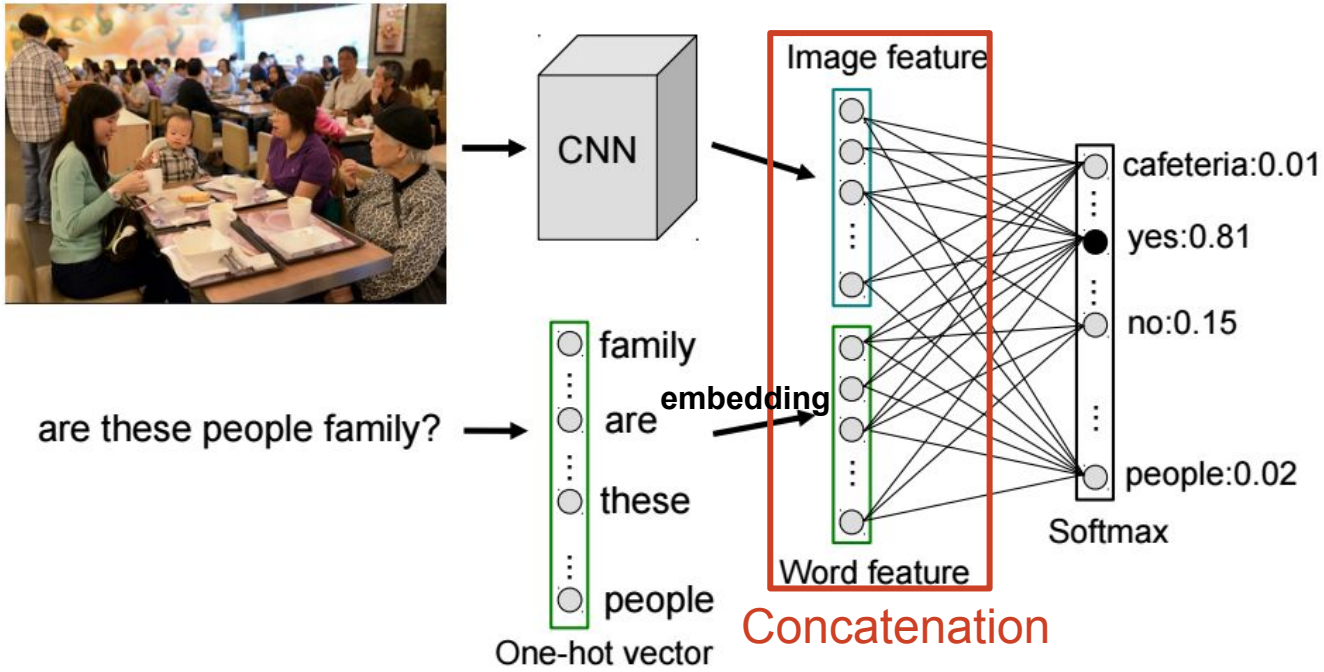
Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167* (2015).

Bag-of-words + Image feature (iBOWIMG)



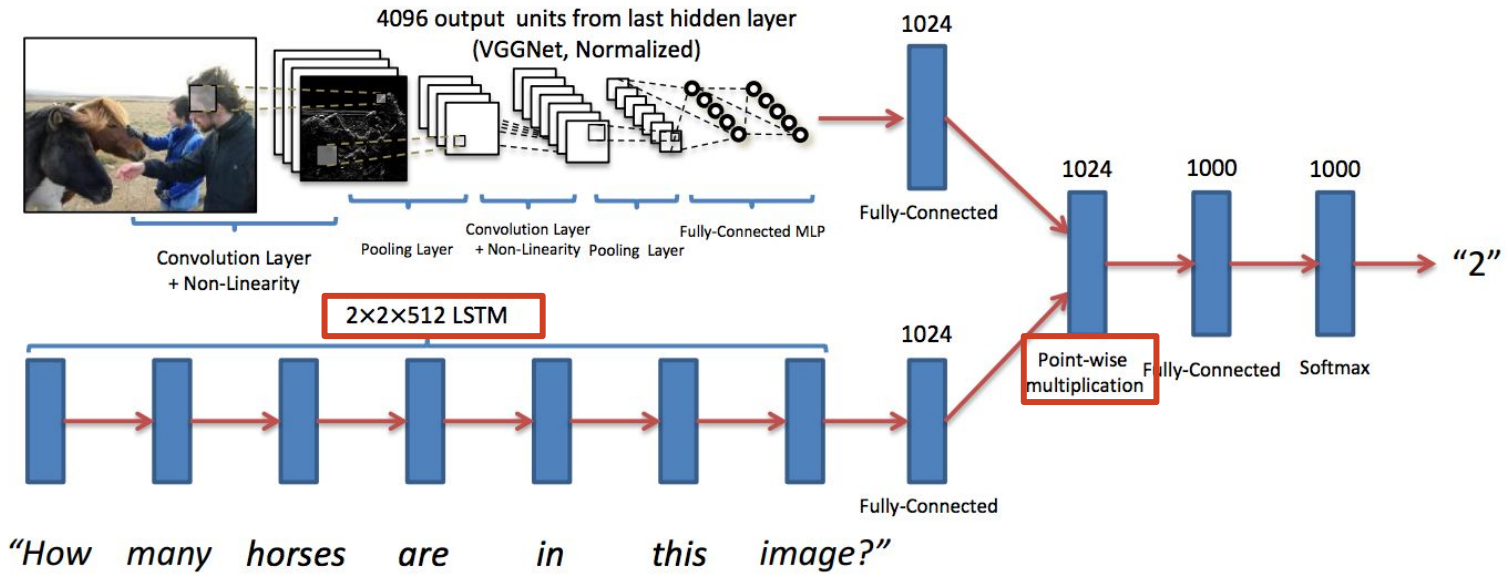
Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167* (2015).

Bag-of-words + Image feature (iBOWIMG)



Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167* (2015).

LSTM + Image feature (LSTM Q + I)



Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

BoW V.S. LSTM ?

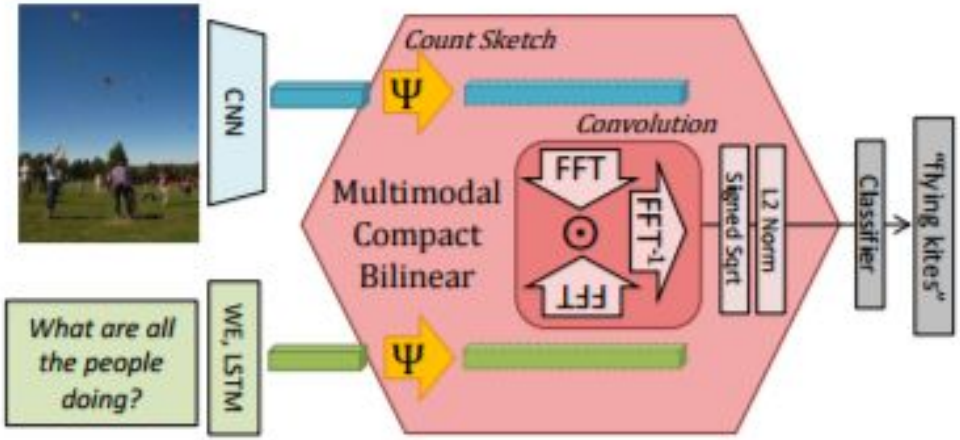
Performances on the VQA-real dataset (test-dev)

	Open-Ended				**Multiple-Choice**			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	other
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44
LSTM Q+I	57.75	80.50	36.77	43.08	62.70	80.52	38.22	53.01

Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167* (2015).

Multimodal Compact Bilinear Pooling (MCB)

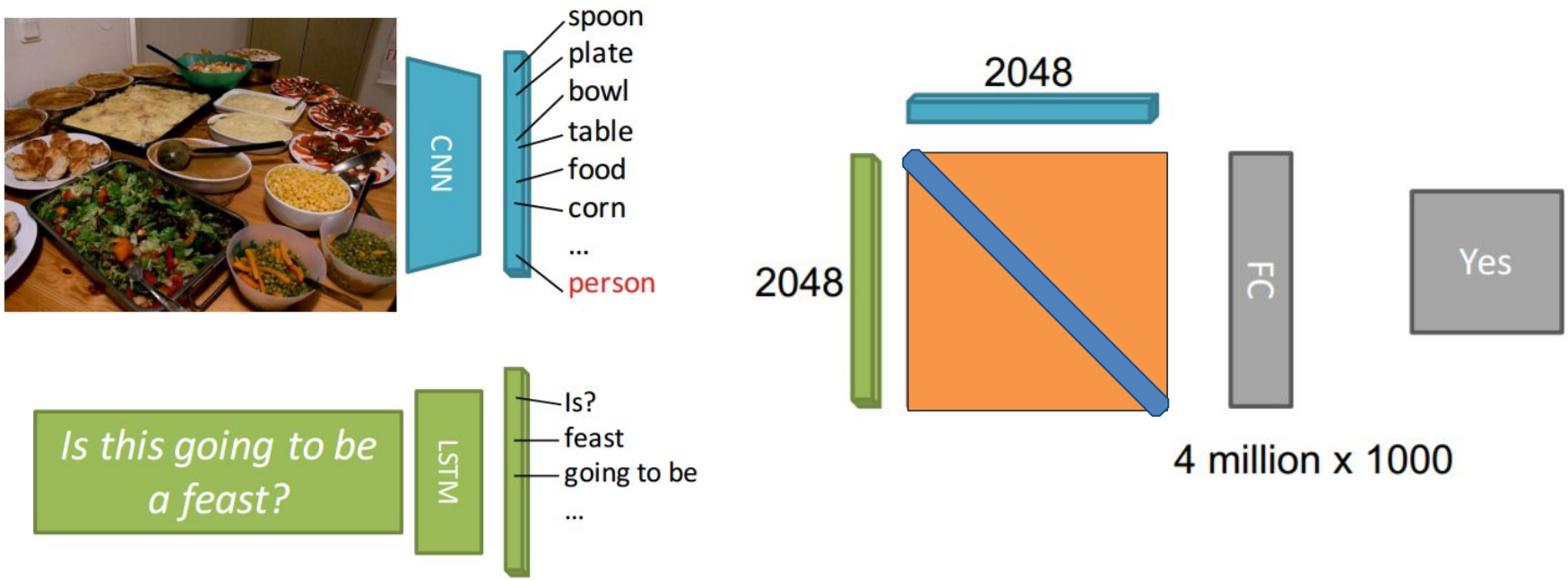
- Use outer product for embedding.
- Results in high dimensional features.
- Then approximate it with low dimensional features.



Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

Bilinear Pooling

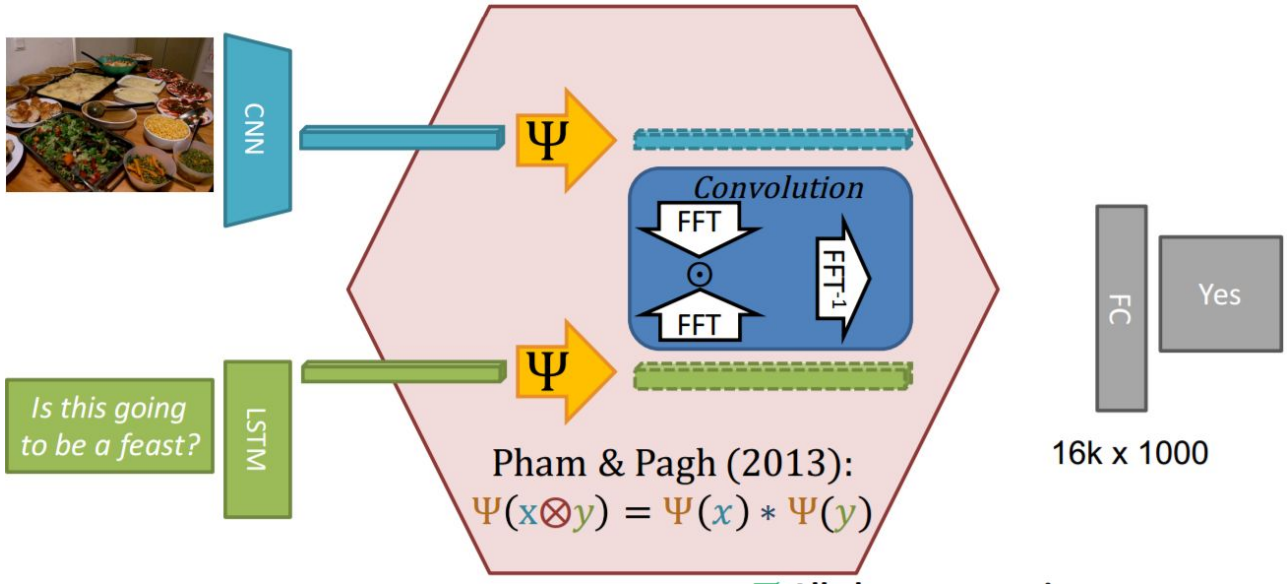


Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Figure source

Multimodal Compact Bilinear Pooling (MCB)

Compact Bilinear Pooling via Count Sketch

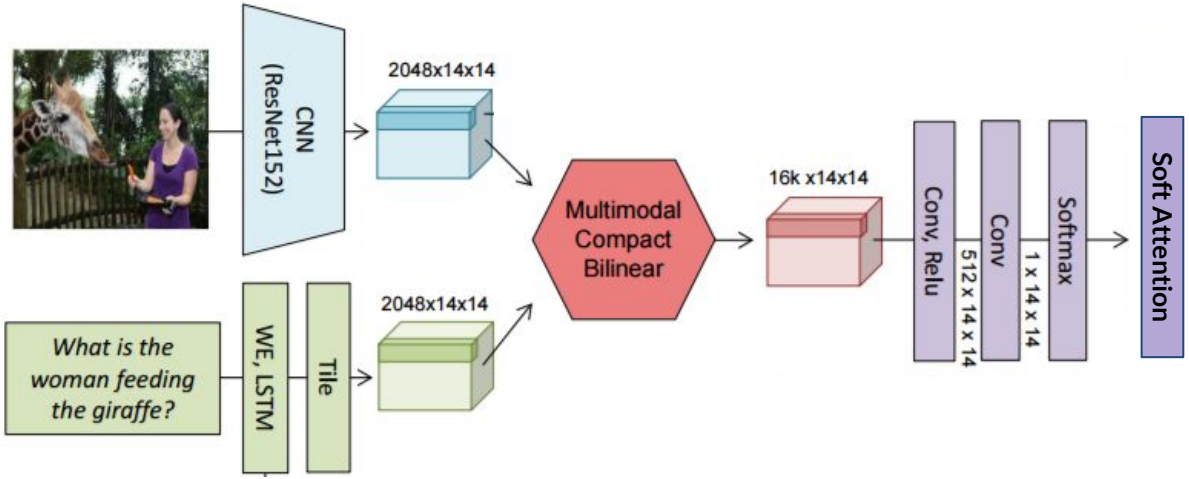


Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Figure source

Multimodal Compact Bilinear Pooling (MCB)

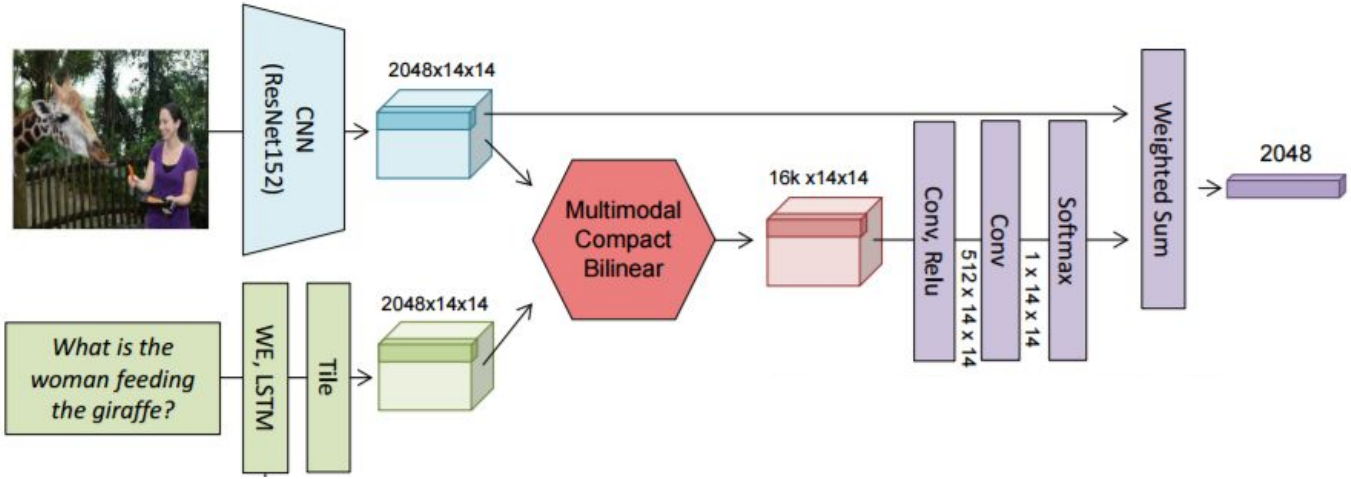
MCB+soft attention for VQA open-ended questions:



Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

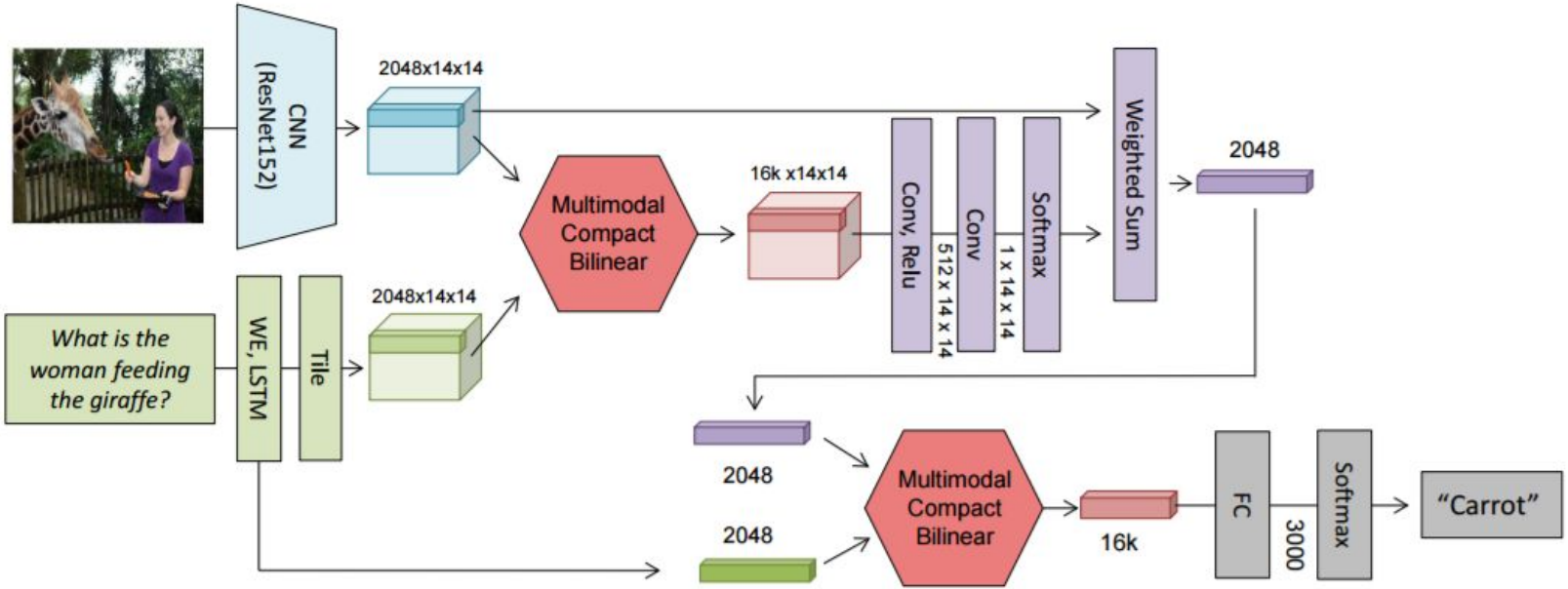
MCB+soft attention for VQA open-ended questions:



Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

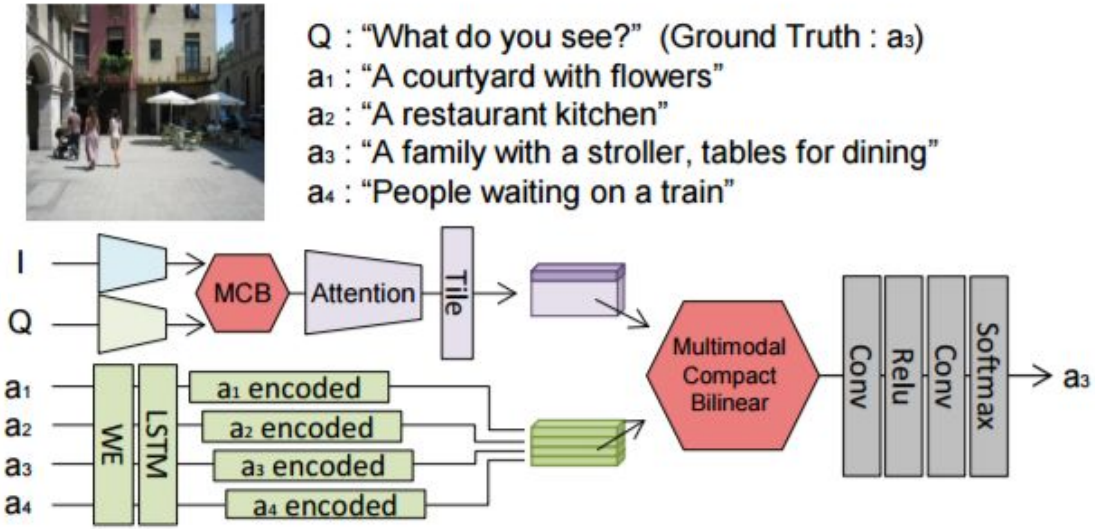
MCB+soft attention for VQA open-ended questions:



Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

Answer Encoding
(for multiple choice questions):



Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

Comparison of multimodal pooling methods

Method	Accuracy
Element-wise Sum	56.50
Concatenation	57.49
Concatenation + FC	58.40
Concatenation + FC + FC	57.10
Element-wise Product	58.57
Element-wise Product + FC	56.44
Element-wise Product + FC + FC	57.88
MCB ($2048 \times 2048 \rightarrow 16K$)	**59.83**

Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Multimodal Compact Bilinear Pooling (MCB)

Additional Experiment results

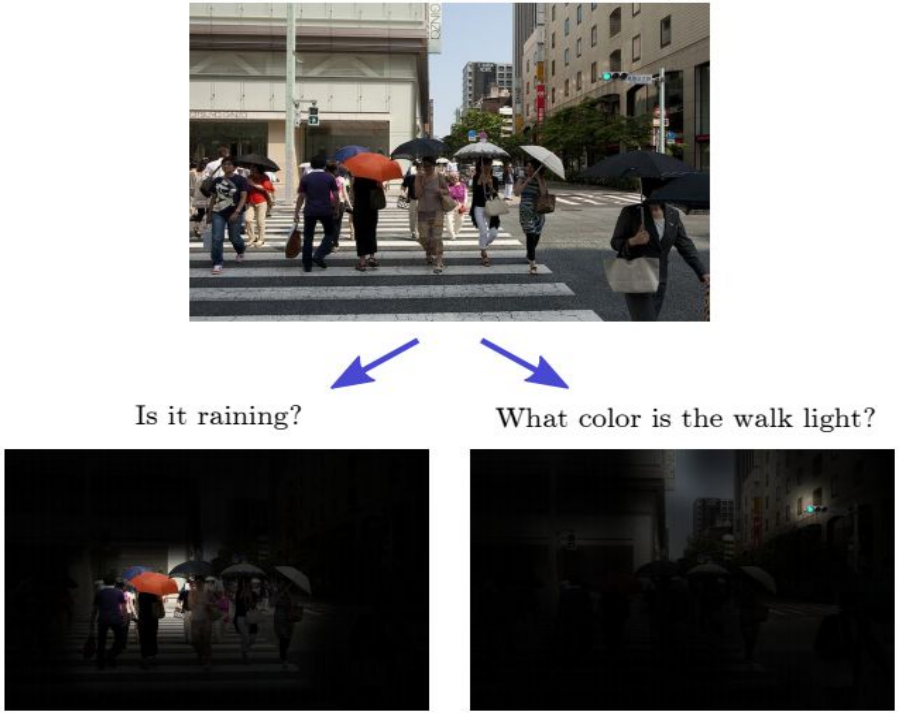
Method	Accuracy
Full Bilinear ($128 \times 128 \rightarrow 16K$)	58.46
MCB ($128 \times 128 \rightarrow 4K$)	58.69
Element-wise Product with VGG-19	55.97
MCB ($d = 16K$) with VGG-19	**57.05**
Concatenation + FC with Attention	58.36
MCB ($d = 16K$) with Attention	**62.50**

Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

Methods

1. **Region-based Image Attention**

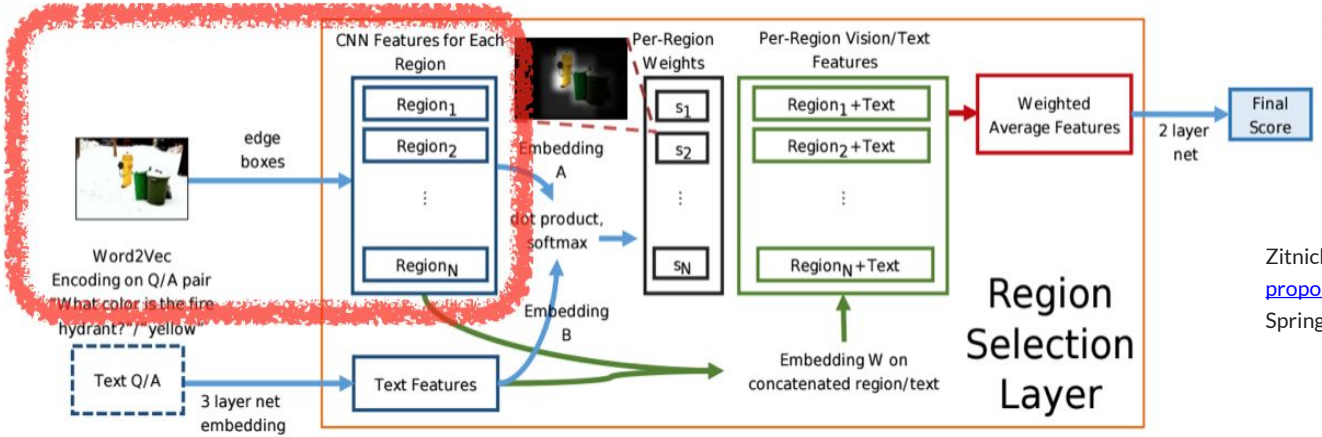
Focus on image regions to answer questions?

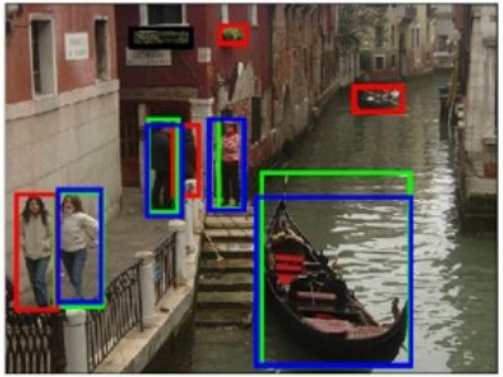


Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Image features

- Extract top ranked 99 regions from edge box + (1 whole image)
- Features are from ImageNet by VGGnets





Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges." *European Conference on Computer Vision.* Springer International Publishing, 2014.

Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Co-attention of image and question

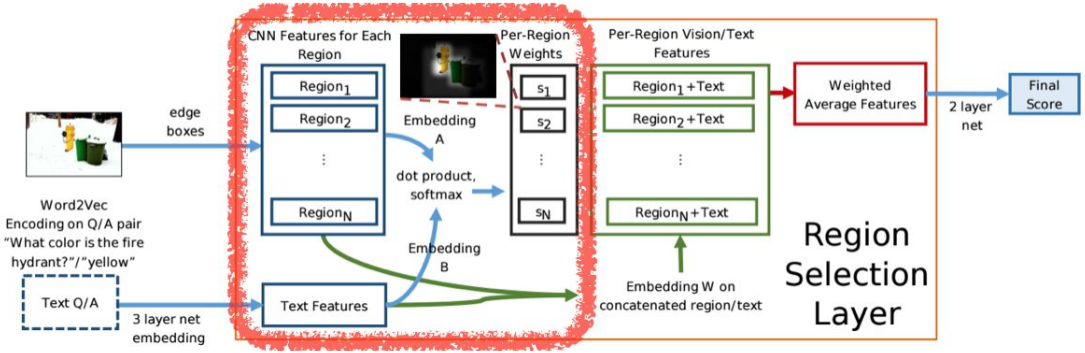
$$X_r = \begin{pmatrix} region1 \\ region2 \\ \dots \\ regionN \end{pmatrix}^T \longrightarrow G_r = AX_r + \vec{b}_r$$

$$\vec{x}_l = \text{text features} \longrightarrow \vec{g}_l = B\vec{x}_l + \vec{b}_l$$

(word embedding)

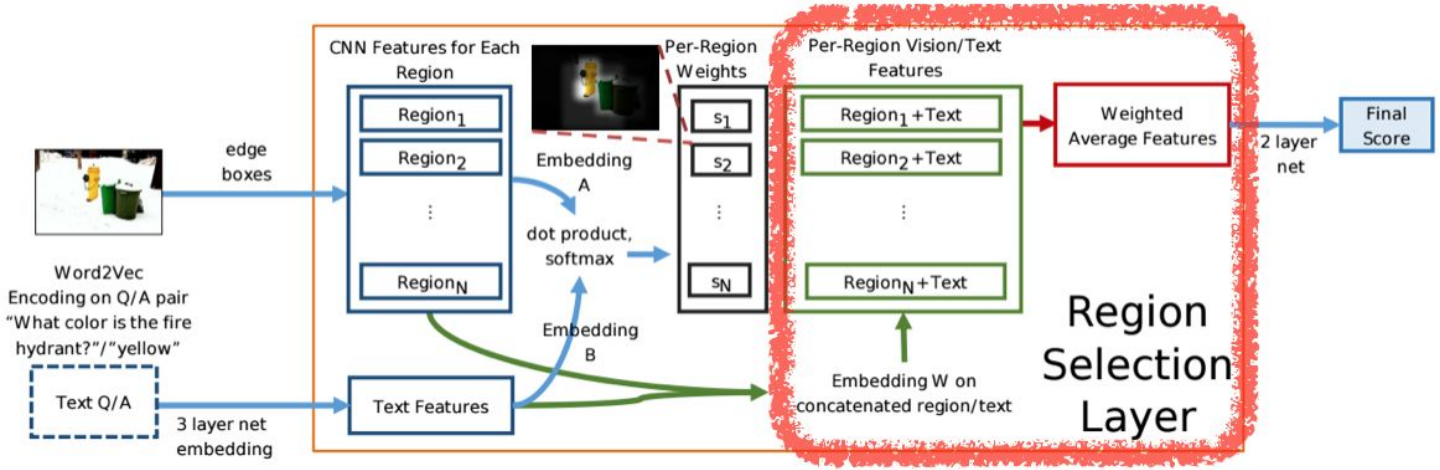
$$\vec{s}_{l,r} = \sigma(G_r^T \vec{g}_l)$$

(softmax)



Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Linearly combine the region features



Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

What color on the stop light is lit up?

L: red (-0.1)
I: red (-0.8)
R: green (1.1)
Ans: green

What color is the light?

L: red (1.0)
I: red (0.3)
R: red (1.7)
Ans: red

What color is the street sign?

L: gray (-0.2)
I: gray (-0.4)
R: yellow (0.4)
Ans: yellow

Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Performances on the VQA-real dataset (test-dev)

	Open-Ended				**Multiple-Choice**			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	other
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44
LSTM Q+I	57.75	80.50	36.77	43.08	62.70	**80.52**	38.22	53.01
Region attention	-	-	-	-	62.44	77.62	34.28	55.84

Accuracies by type of question

- On the VQA-real dataset (validation set)

	region	image	text	freq
overall	**58.94**	57.83	53.98	100.0%
is/are/was	**75.42**	74.63	75.00	33.3%
identify: what kind/type/animal	**52.89**	52.10	45.11	23.8%
how many	33.38	**36.84**	34.05	10.3%
what color	**53.96**	43.52	32.59	9.8%
interpret: can/could/does/has	**75.73**	74.43	**75.73**	4.6%
none of the above	45.40	44.04	**48.23**	4.1%
where	42.11	**42.43**	37.61	2.5%
why/how	26.31	28.18	**29.24**	2.2%

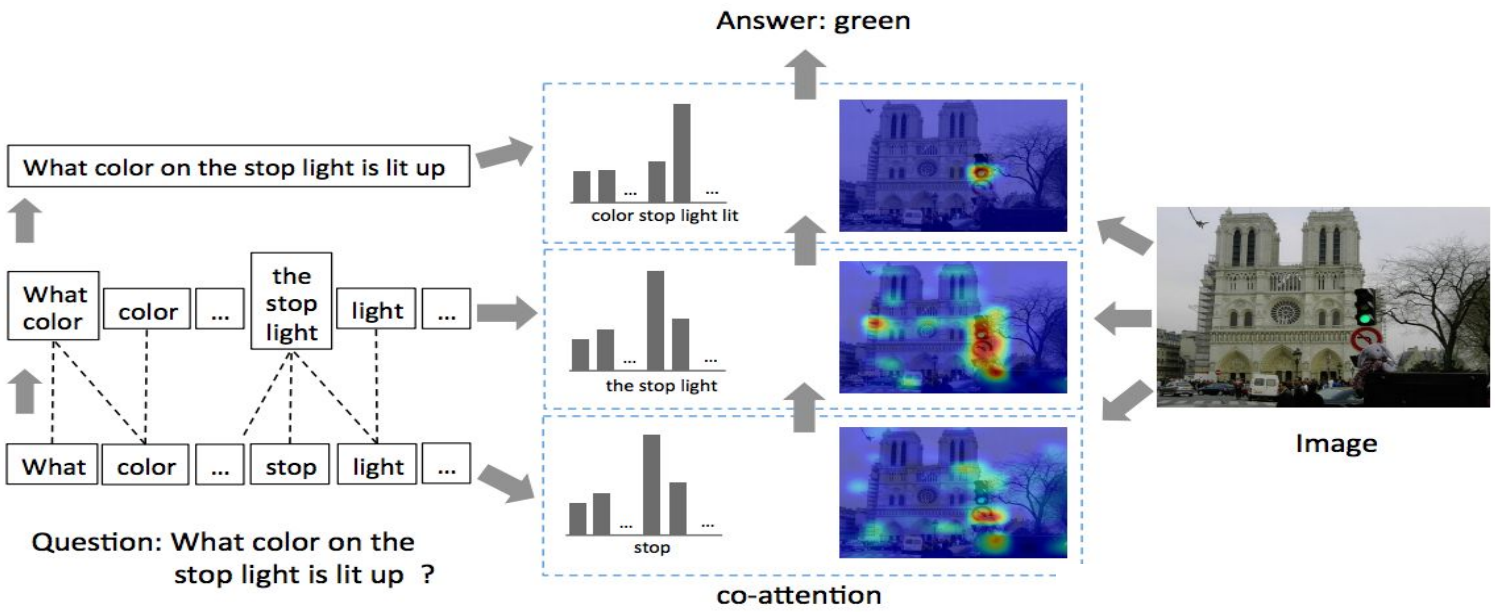
- Region : region weighted features
- Image : whole image features
- Text : no image feature

Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Methods

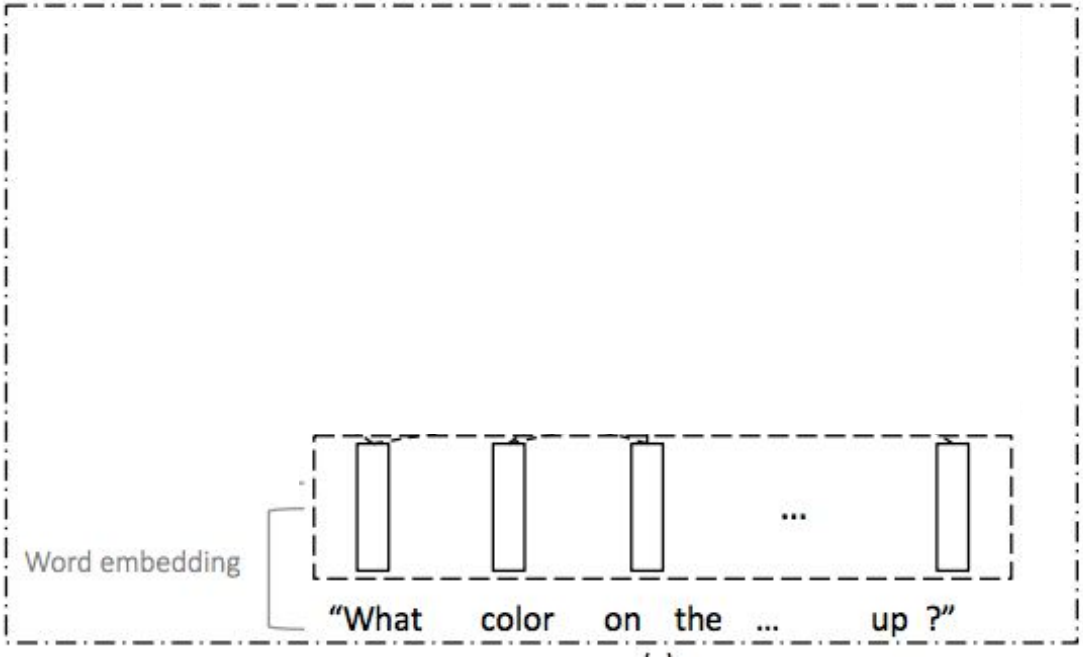
2. Hierarchical Question Attention

Hierarchical Question-Image Co-Attention (HieCoAtt)



Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

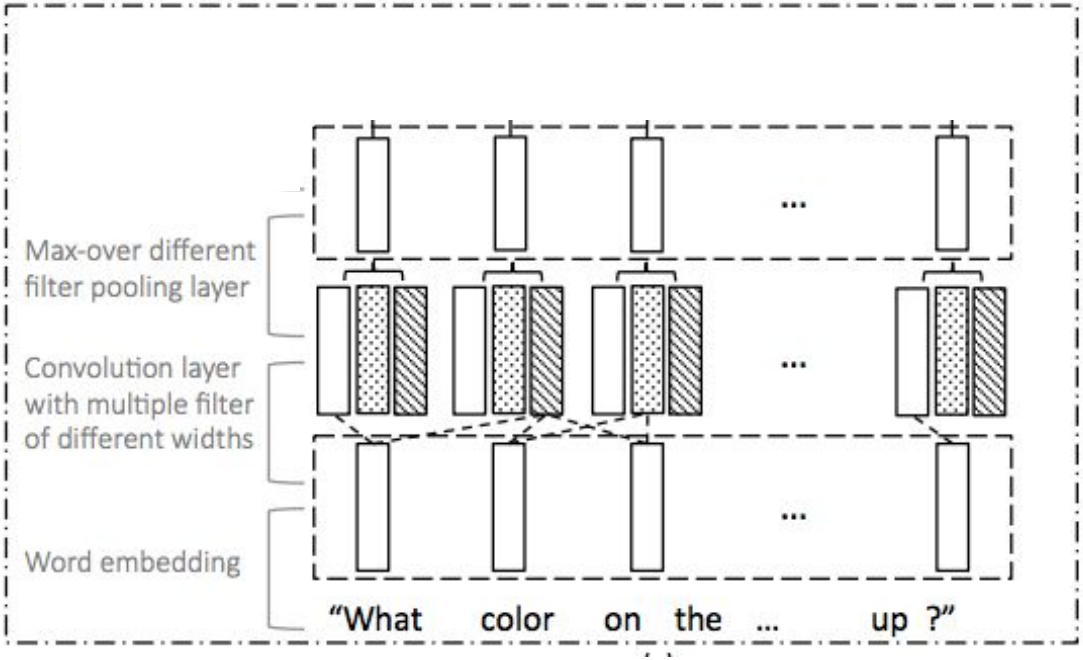
Question hierarchy



Word Level

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

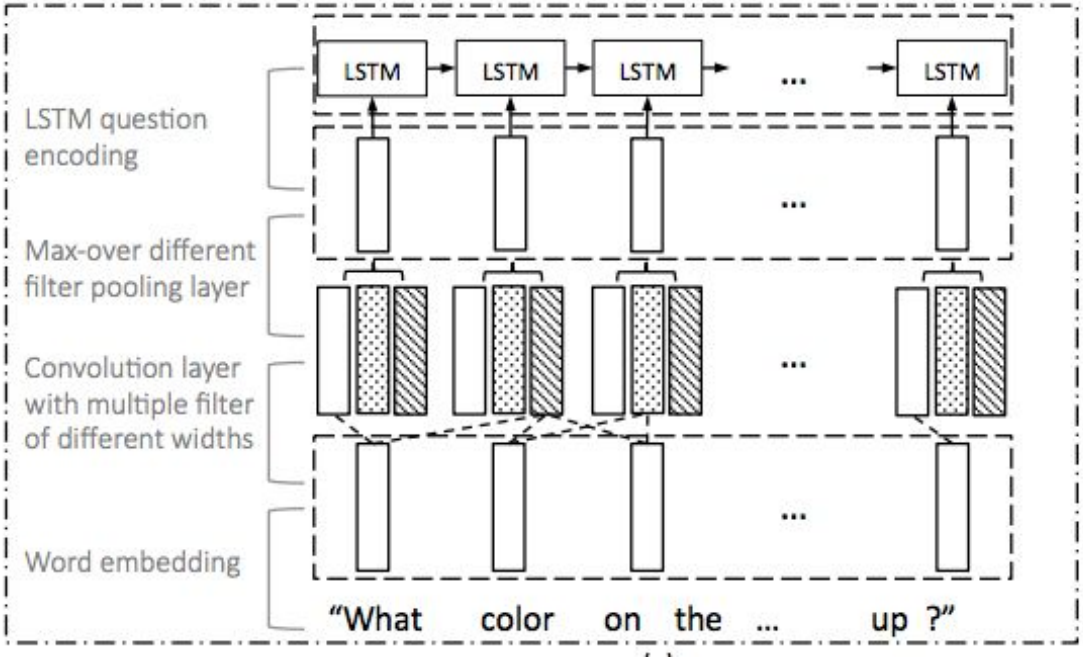
Question hierarchy



Phrase Level

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems.* 2016.

Question hierarchy



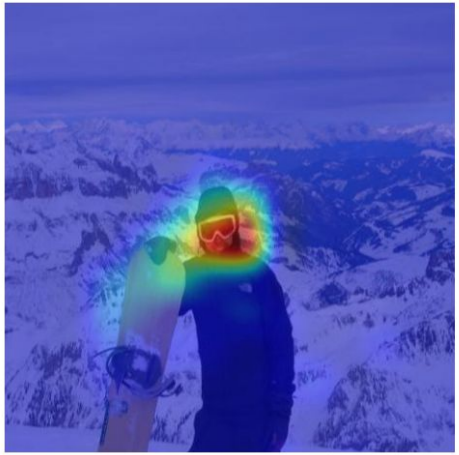
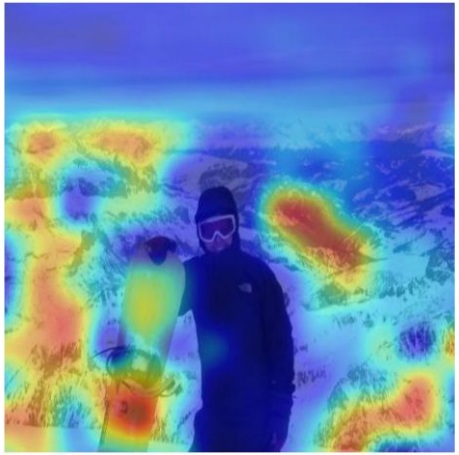
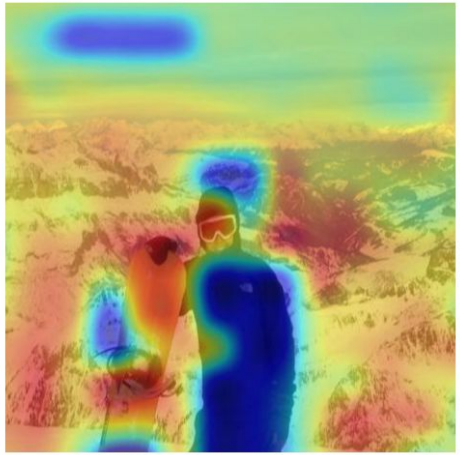
Question Level

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

Word Level

The colored words are those with higher weights.

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

	Word Level	Phrase Level	Question Level
			
Q: what is the man holding a snowboard on top of a snow covered? **A**: **mountain**	what is the man holding a snowboard on top of a snow covered	what is the man holding a snowboard on top of a snow covered ?	what is the man holding a snowboard on top of a snow covered ?

The colored words are those with higher weights.

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

Ablation study on the VQA-real dataset (validation set)

- The attention mechanisms closest to the 'top' of the hierarchy matter most

	validation			
Method	Y/N	Num	Other	All
LSTM Q+I	**79.8**	32.9	40.7	54.3
Image Atten	79.8	33.9	43.6	55.9
Question Atten	79.4	33.3	41.7	54.8
W/O Q-Atten	79.6	32.1	42.9	55.3
W/O P-Atten	79.5	34.1	45.4	56.7
W/O W-Atten	79.6	34.4	45.6	56.8
Full Model	79.6	**35.0**	**45.7**	**57.0**

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." *Advances In Neural Information Processing Systems*. 2016.

Performances on the VQA-real dataset (test-dev)

	Open-Ended				**Multiple-Choice**			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	other
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44
LSTM Q+I	57.75	80.50	36.77	43.08	62.70	**80.52**	38.22	53.01
MCB	60.8	**81.2**	35.1	49.3	65.40	-	-	-
Region attention	-	-	-	-	62.44	77.62	34.28	55.84
HieCoAtt	**61.80**	79.78	**38.78**	**51.78**	**65.80**	79.70	**40.00**	**59.80**

Dataset Bias Problem

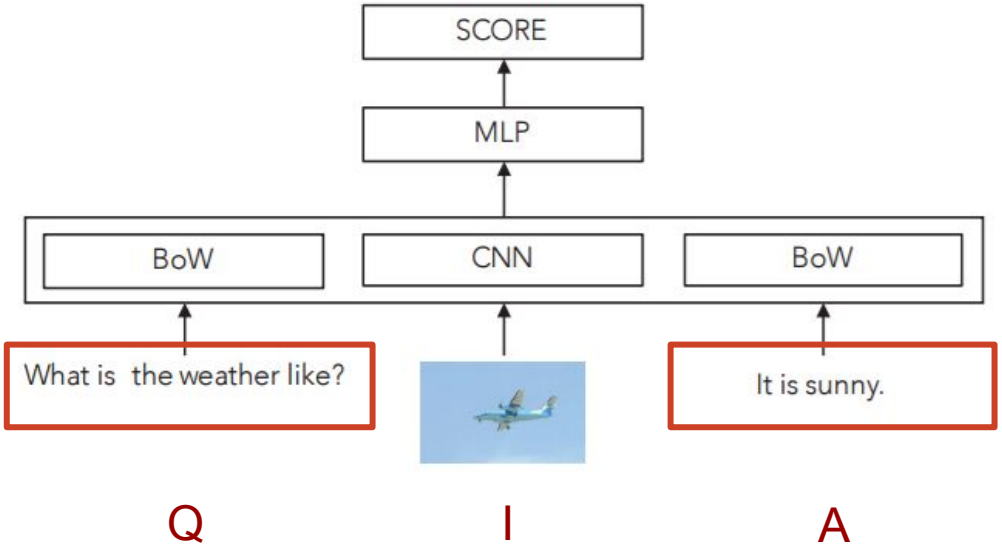
Baselines that Exploit Dataset Biases

By predicting correctness of an Image-Question-Answer triplet:

- reaches state-of-the-art performance on Visual7W Telling.
- performs competitively on VQA Real multiple choice.

Models:

$$\textbf{MLP:}\ y = \sigma(\mathbf{W}_2 \max(0, \mathbf{W}_1 \mathbf{x}_{iqa}) + b)$$



Jabri, Allan, Armand Joulin, and Laurens van der Maaten. "Revisiting visual question answering baselines." *European Conference on Computer Vision*. Springer International Publishing, 2016.

Baselines that Exploit Dataset Biases

Baselines:

- MLP(A,Q,I): use all features
- MLP(A,I): Answers + Images
- MLP(A,Q): Answers + Questions
- MLP(A): Answers

Accuracies on Visual7W Telling

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2
MLP (A)	47.3	58.2	74.3	63.6	57.1	49.6	52.9
MLP (A, Q)	54.9	60.0	76.8	66.0	64.5	54.9	58.5
MLP (A, I)	60.8	74.9	81.9	70.3	64.4	51.2	63.8
MLP (A, Q, I)	**64.5**	**75.9**	**82.1**	**72.9**	**68.0**	**56.4**	**67.1**

Jabri, Allan, Armand Joulin, and Laurens van der Maaten. "Revisiting visual question answering baselines." *European Conference on Computer Vision*. Springer International Publishing, 2016.

Baselines that Exploit Dataset Biases

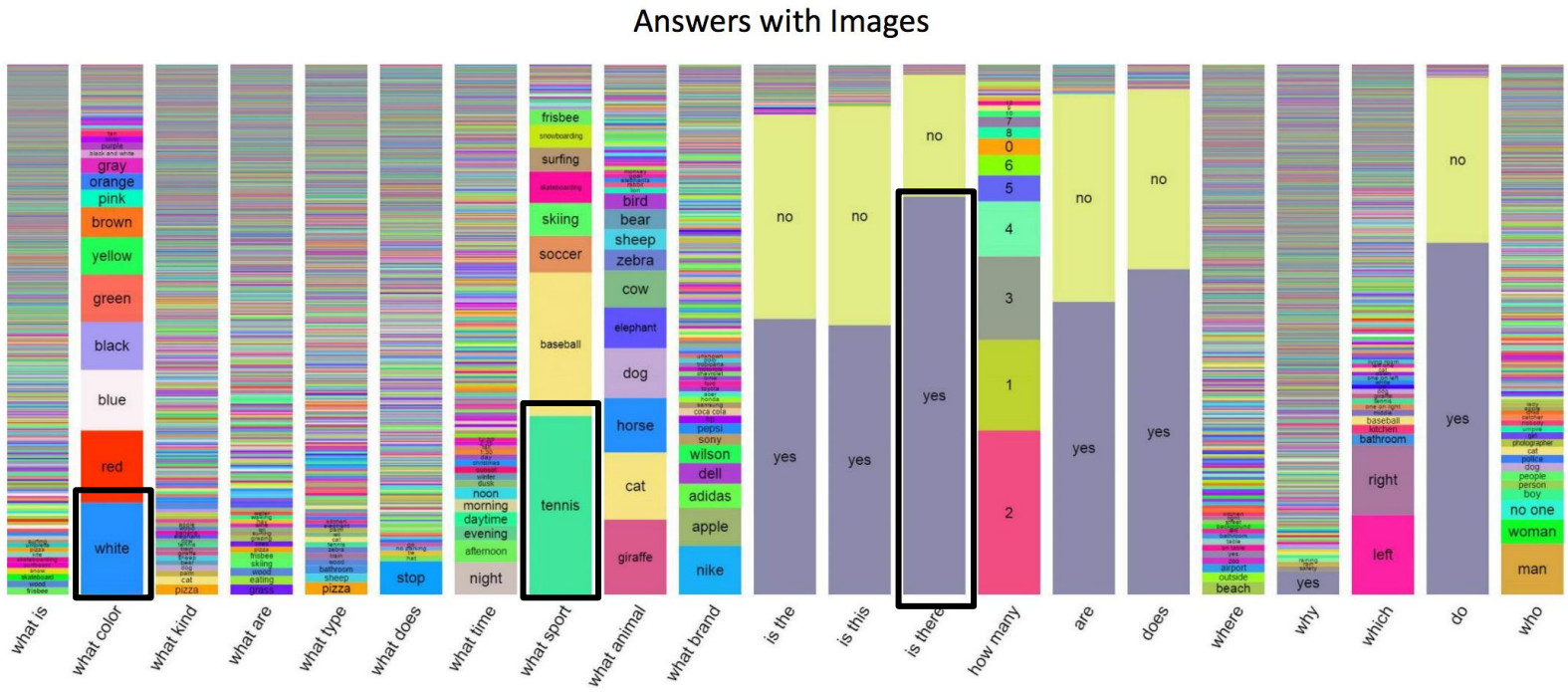
Accuracies on VQA Multiple Choice (test)

Method	Yes/No	Number	Other	All
Two-Layer LSTM [5]	80.6	37.7	53.6	63.1
Region selection [23]	77.2	33.5	56.1	62.4
MCB [21]*	–	–	–	65.4
MCB + Att + GloVe + Genome [21]*	–	–	–	**69.9**
MLP (A, Q, I)	**80.8**	17.6	**62.0**	65.2

* evaluated on test-dev set

Jabri, Allan, Armand Joulin, and Laurens van der Maaten. "Revisiting visual question answering baselines." *European Conference on Computer Vision*. Springer International Publishing, 2016.

VQA dataset

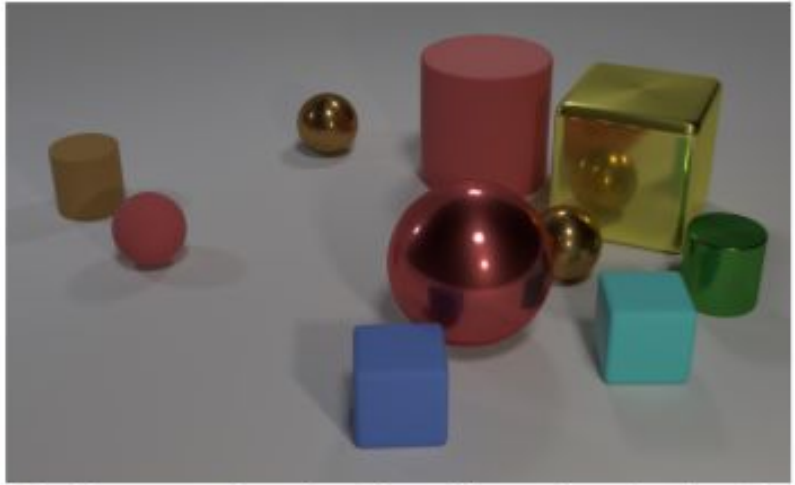


Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

Responses to Dataset Biases

1. **More Balanced Datasets**

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

A sample image and questions from CLEVR

- 100,000 computer rendered images.
- 864,986 generated questions.
- **Scene graph** representations for all images.
- **Functional program** representation for all images.

With complex questions like: counting, attribute identification, comparison, multiple attention, and logical operations.

Johnson, Justin, et al. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning." *arXiv preprint arXiv:1612.06890* (2016).

VQA 2.0: Making the V in VQA matter

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

Goyal, Yash, et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering." *arXiv preprint arXiv:1612.00837* (2016).

VQA 2.0: Making the V mater

Results:

VQA models trained/tested on unbalanced/balanced datasets

Approach	UU(%)	UB(%)	BB(%)
Language Only	48.21	40.03	39.98
LSTM(Q+I)	54.40	46.56	48.18
HieCoAtt	57.09	49.51	51.02
MCB	60.36	53.67	55.35

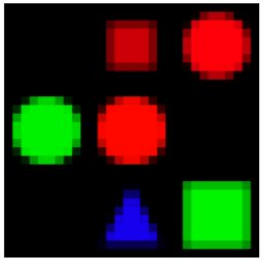
Goyal, Yash, et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering." *arXiv preprint arXiv:1612.00837* (2016).

Responses to Dataset Biases

2. Compositional Models

Neural Module Network(NMN)

Exploit compositional nature of questions:

				
how many different lights in various different shapes and sizes?	*what is the color of the horse?*	*what color is the vase?*	*is the bus full of passengers?*	*is there a red shape above a circle?*

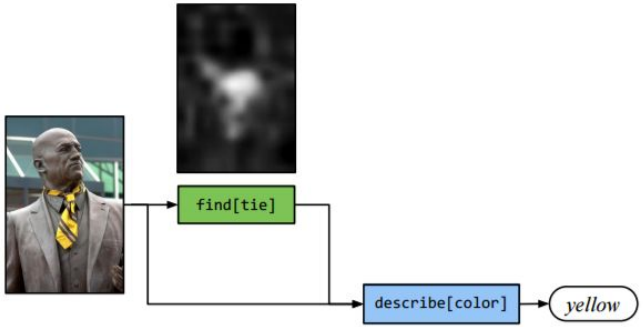
`describe[count](`
`  find[light])`
	`describe[color](`
	`  find[horse])`
		`describe[color](`
		`  find[vase])`
			`describe[is](`
			`  combine[and](`
			`  find[bus],`
			`  find[full])`
				`measure[is](`
				`  combine[and](`
				`  find[red],`
				`  transform[above](`
				`  find[circle])))`

four (four)	brown (brown)	green (green)	yes (yes)	yes (yes)

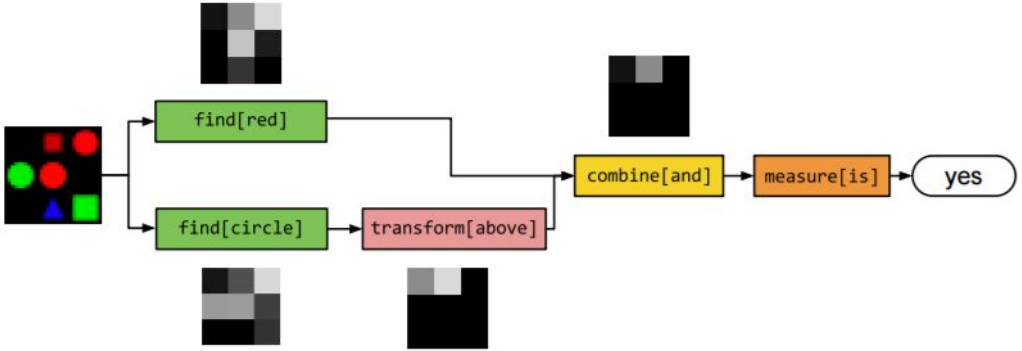
Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Neural Module Network(NMN)

Question answering in steps:



What color is his tie?



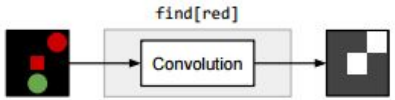
Is there a red shape above a circle?

Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Neural Module Network(NMN)

Neural Modules:

Find $\quad Image \rightarrow Attention$

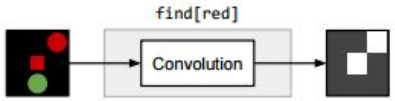


Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

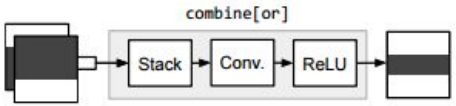
Neural Module Network(NMN)

Neural Modules:

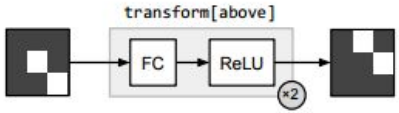
Find $\textit{Image} \rightarrow \textit{Attention}$



Combine $\textit{Attention} \times \textit{Attention} \rightarrow \textit{Attention}$



Transform $\textit{Attention} \rightarrow \textit{Attention}$

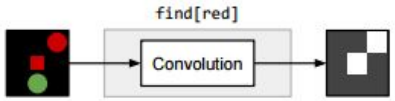


Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

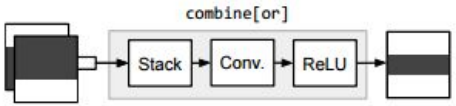
Neural Module Network(NMN)

Neural Modules:

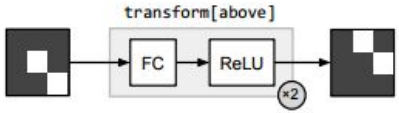
Find *Image → Attention*



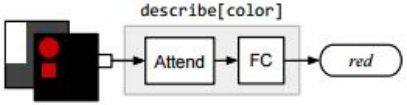
Combine *Attention × Attention → Attention*



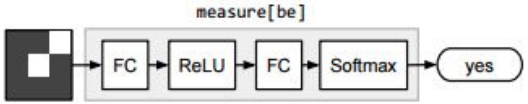
Transform *Attention → Attention*



Describe *Image × Attention → Label*



Measure *Attention → Label*



Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Neural Module Network(NMN)

From questions to networks:

- Parse questions to structured queries:

  "Is there a circle next to a square?" -> **is**(circle, **next-to**(square))

Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

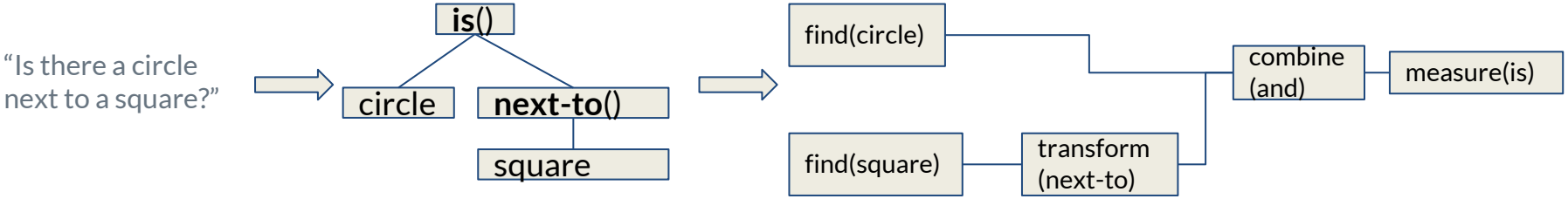
Neural Module Network(NMN)

From questions to networks:

- Parse questions to structured queries:

  "Is there a circle next to a square?" -> **is**(circle, **next-to**(square))

- Generate network Layout:



Question to layout as a deterministic process

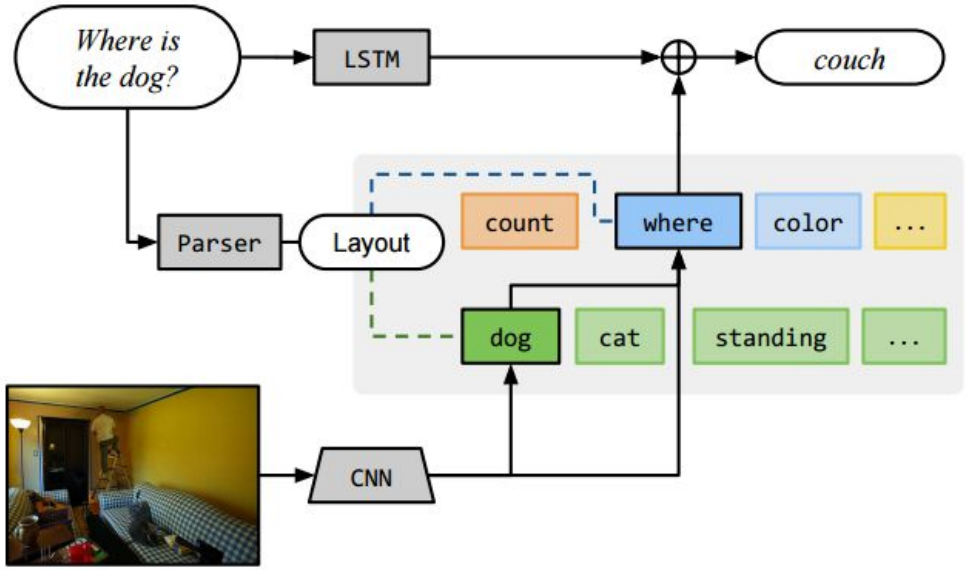
Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Neural Module Network(NMN)

Final Model:

Why using LSTM feature?

- aggressive simplification of the question, need grammar cues.
- capture semantic regularities, e.g. common sense.



Answering *"Where is the dog?"* with NMN

Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Neural Module Network(NMN)

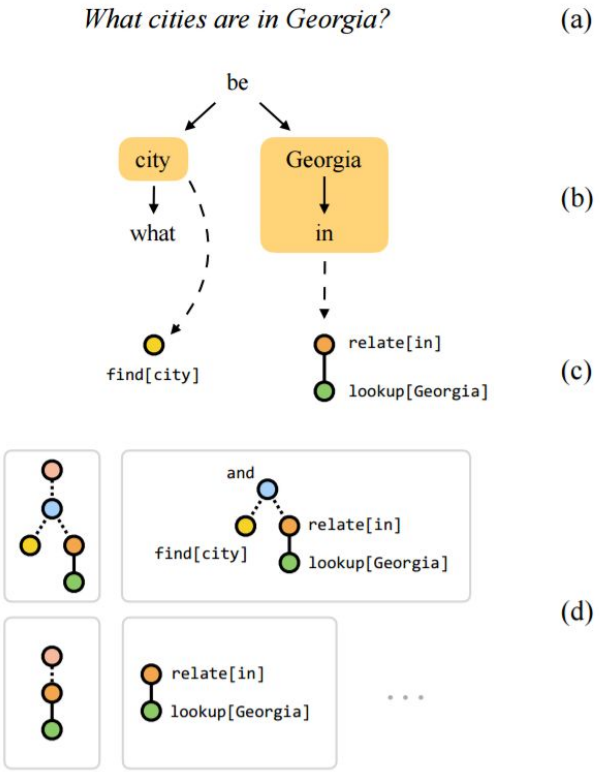
Results on VQA

Results on the VQA open ended questions

	test-dev				test
	Yes/No	Number	Other	All	All
LSTM	78.7	36.6	28.1	49.8	–
ATT+LSTM	80.6	36.4	42.0	57.2	–
NMN	70.7	36.8	39.2	54.8	–
NMN+LSTM	81.2	35.2	43.3	58.0	–
NMN+LSTM+FT	81.2	38.0	44.0	58.6	**58.7**

Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Dynamical Neural Module Network(DNMN)



Generate Layout candidates from questions

Generate layout candidates:

The input sentence (a)
is represented as a dependency parse (b).
Fragments are then associated with modules (c),
fragments are assembled into full layouts (d).

Select best candidates:

$$s(z_i|x) = a^\top \sigma(Bh_q(x) + Cf(z_i) + d)$$

$$p(z_i|x;\theta_\ell) = e^{s(z_i|x)} \Big/ \sum_{j=1}^{n} e^{s(z_j|x)}$$

What is the constraint in this setting?

Andreas, Jacob, et al. "Learning to compose neural networks for question answering." *arXiv preprint arXiv:1601.01705* (2016).

Dynamical Neural Module Network(DNMN)

- Evaluate all layouts $p(z|x;\theta_\ell)$, cheap
- Evaluate all answers given all layout $p_z(y|w;\theta_e)$, expensive

A reinforcement learning problem at heart!

State → (w, x) Action → z Reward → $\log p(y|z, w; \theta_e)$

Policy gradient : $\mathbb{E}[(\nabla \log p(z|x;\theta_\ell)) \cdot \log p(y|z, w; \theta_e)]$ (REINFORCE rule)

Adding language prior to DNMN model: $\log p_z(y|w, x) = (Ah_q(x) + B[\![z]\!]_w)_y$

Andreas, Jacob, et al. "Learning to compose neural networks for question answering." *arXiv preprint arXiv:1601.01705* (2016).

Dynamical Neural Module Network(DNMN)

Results on the VQA open ended questions

	test-dev				test-std
	Yes/No	Number	Other	All	All
Zhou (2015)	76.6	35.0	42.6	55.7	55.9
NMN	81.2	38.0	44.0	58.6	58.7
D-NMN	81.1	38.6	45.5	59.4	**59.4**

Zhou(2015) - iBOWING

Future Development

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

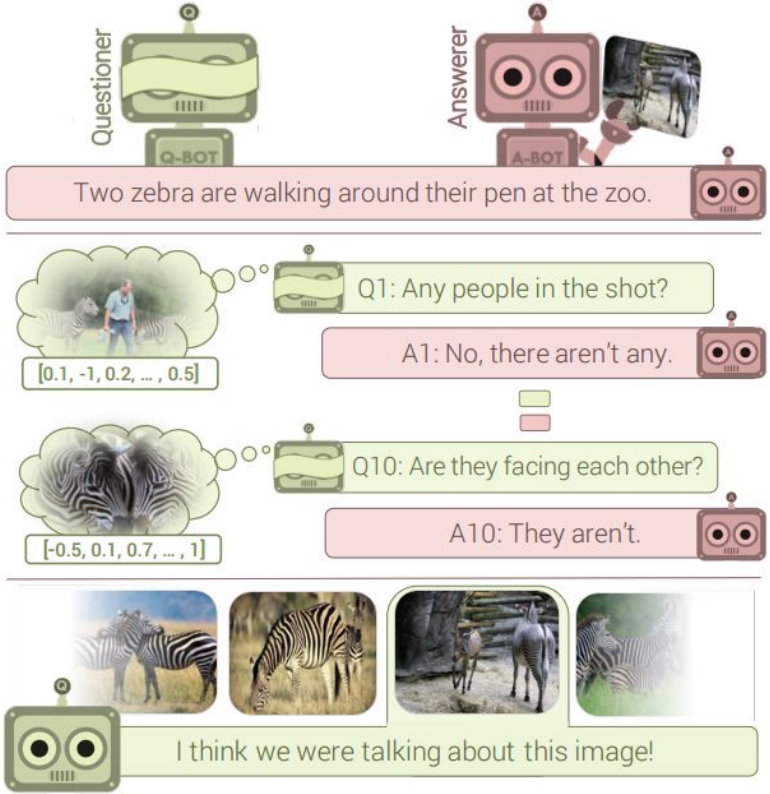
A complete cooperative game of two agents:

Q-BOT :

- ❑ Build a mental model of the unseen image purely from the natural language dialog,
- ❑ Predict an image feature and gain rewards for two agents

A-BOT:

- ❑ See the target image, generate a caption for Q-BOT
- ❑ Build a mental model of what Q-BOT understands, and answer questions.



A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

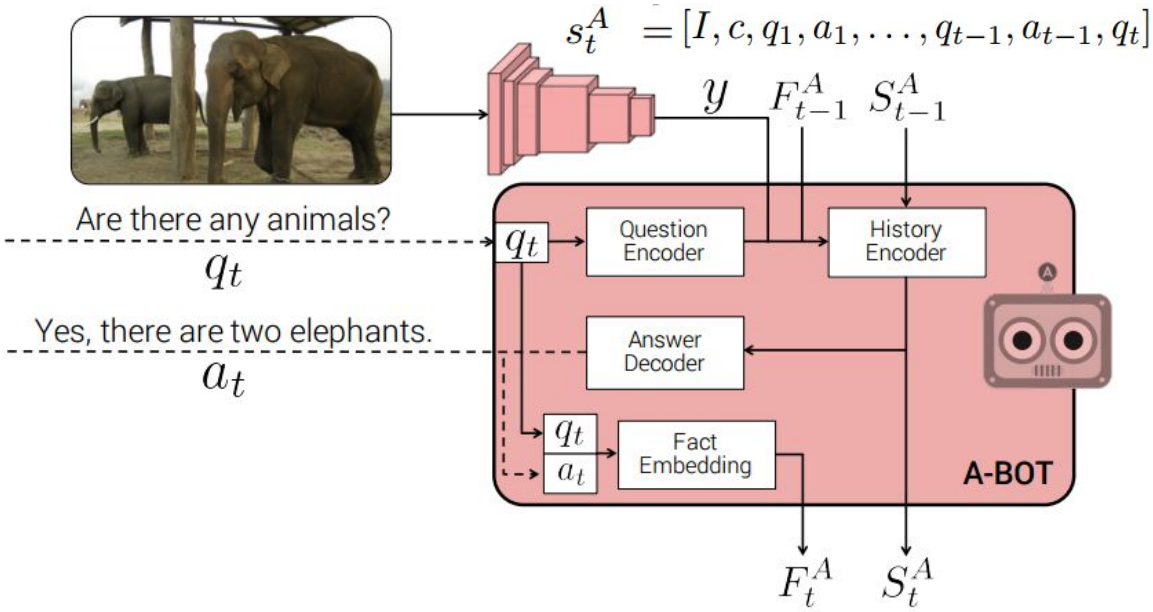
Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Visual Dialog

A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

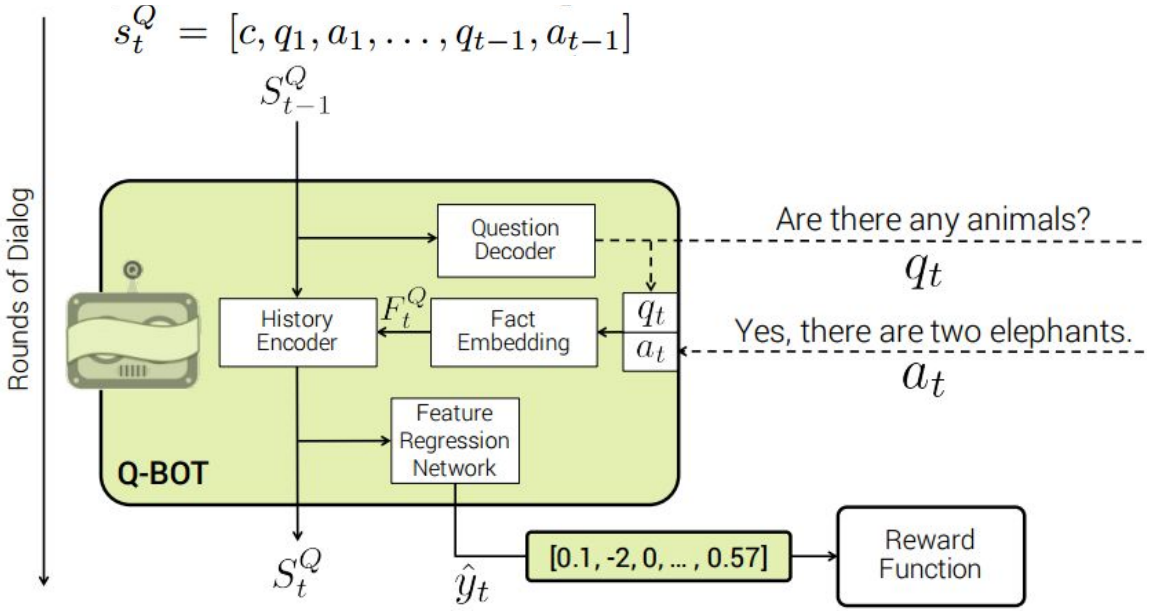
Policy Networks for Q-BOT and A-BOT:



A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

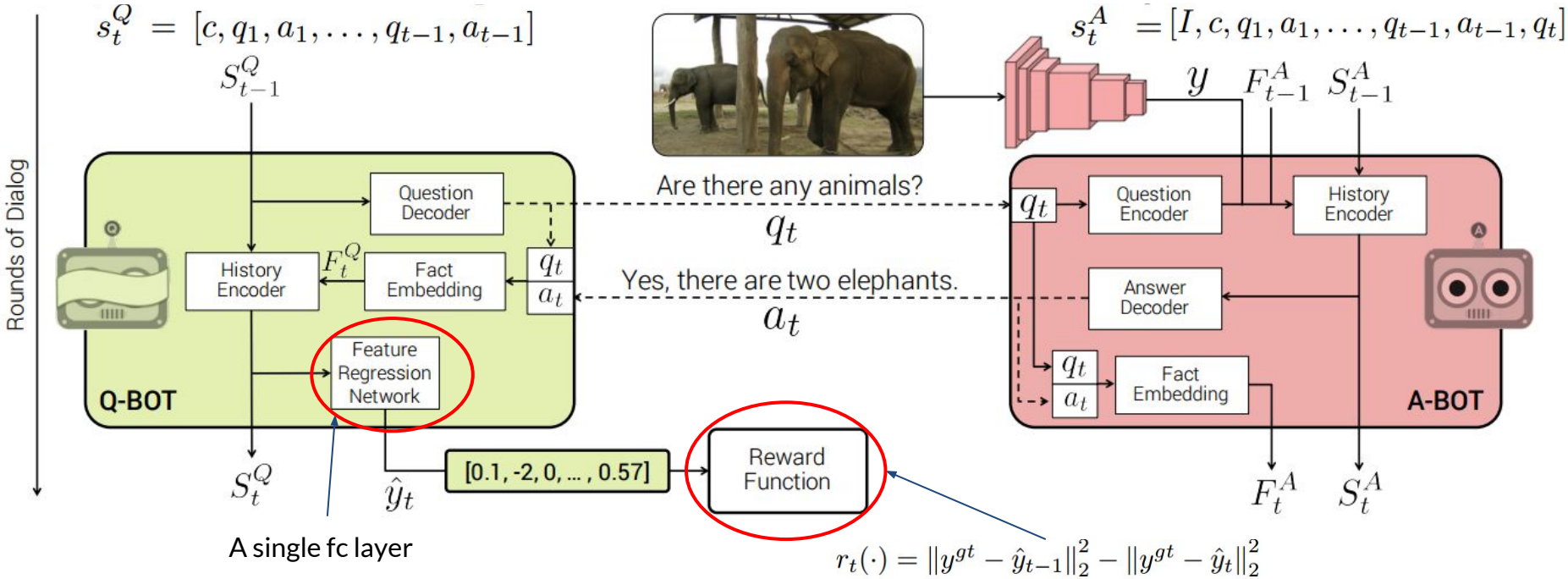
Policy Networks for Q-BOT and A-BOT:



A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

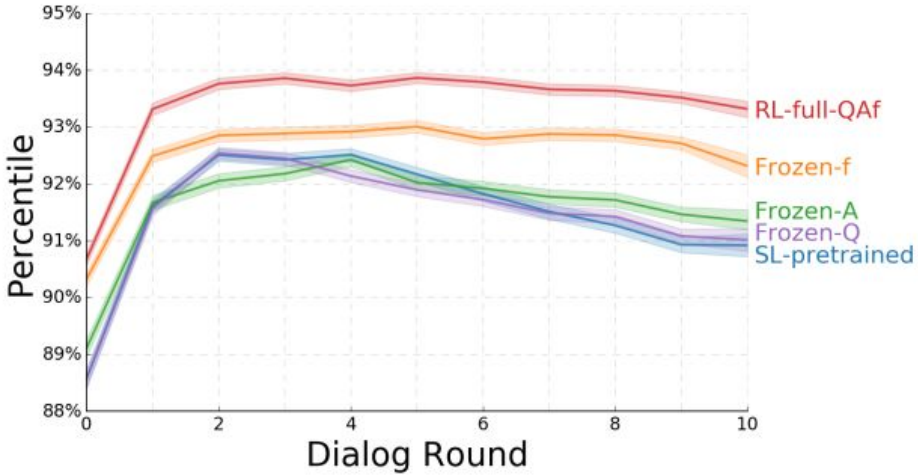
Policy Networks for Q-BOT and A-BOT:



A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Results:



(a) Guessing Game Evaluation.

A. Das, S. Kottur, et al.. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).

Summary

A new language-image task called " Visual Question Answering"

- Can machines see images, reason and answer to questions like human?

Various approaches

- language-image embedding
- Attention mechanism
- Compositional models

Dataset bias

- Can exploit answer distribution biases
- High question-answer correlation

Further development

- More Balanced dataset
- Emphasis on visual features

Thank you for your attention !

Reading list

- Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. "Visual Question Answering: A Survey of Methods and Datasets." arXiv preprint arXiv:1607.05910 (2016).
- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. "VQA: Visual Question Answering." arXiv preprint arXiv:1505.00468 (2016).
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning." arXiv preprint arXiv:1612.06890 (2016).
- Kevin J. Shih, Saurabh Singh, Derek Hoiem. "Where To Look: Focus Regions for Visual Question Answering." arXiv preprint arXiv:1511.07394 (2016).
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. "Learning to Compose Neural Networks for Question Answering." arXiv preprint arXiv:1601.01705 (2016).
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding." arXiv preprint arXiv:1606.01847 (2016).
- Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. "Hierarchical Question-Image Co-Attention for Visual Question Answering."." arXiv preprint arXiv:1606.00061 (2017).
- Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, Dhruv Batra. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).
- Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei. "Visual7W: Grounded Question Answering in Images." arXiv preprint arXiv:1511.03416 (2016).
- Allan Jabri, Armand Joulin, Laurens van der Maaten. "Revisiting Visual Question Answering Baselines." arXiv preprint arXiv:1606.08390 (2016).
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. "Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources." arXiv preprint arXiv:1511.06973 (2016).