CSC 2523

Deep Learning in Computer Vision

Winter 2016

Neural-based Image Question Answering

Yunpeng Li

Faculty of Information

2016.03.01

Question Answering

Traditional Approaches

Textual question answering tasks

- Semantic parsing
- Symbolic representation
- Deduction system

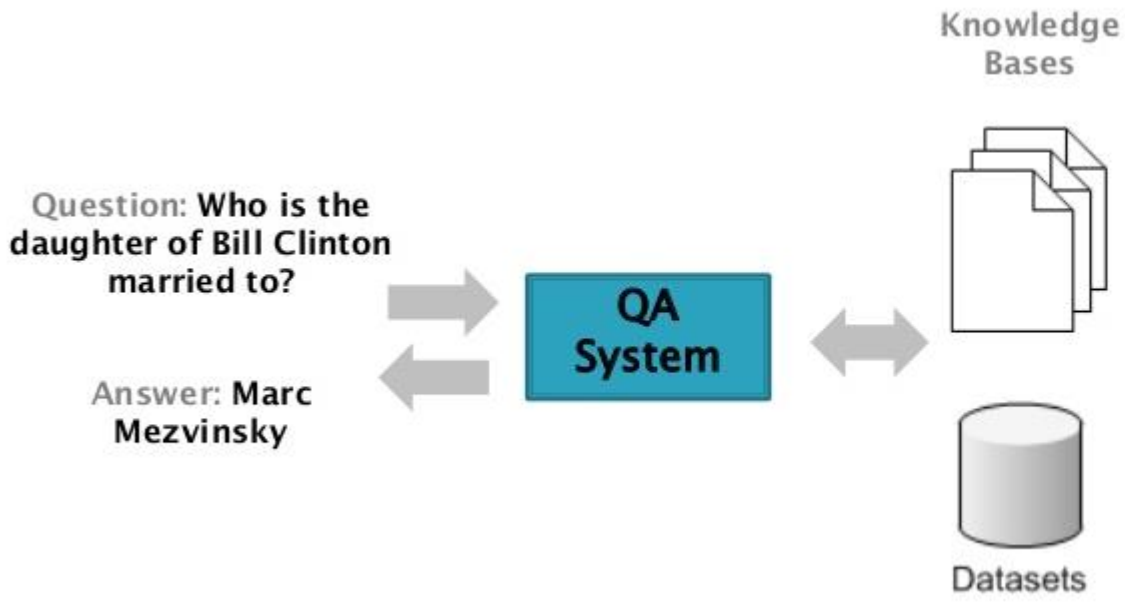
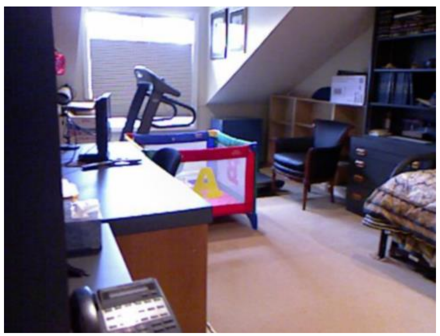


Image Credit: Question Answering over Linked Data: Challenges, Approaches & Trends (Tutorial @ ESWC 2014)

Image Question Answering

Multi-modal problem



What is on the right side of the cabinet?

Image Representation

Natural Language Processing

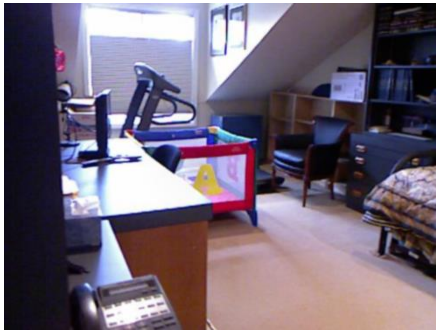
QA

Bed

Using both visual & natural language inputs

Image Question Answering

Neural-based Approach



What is on the right side of the cabinet?

CNN

LSTM

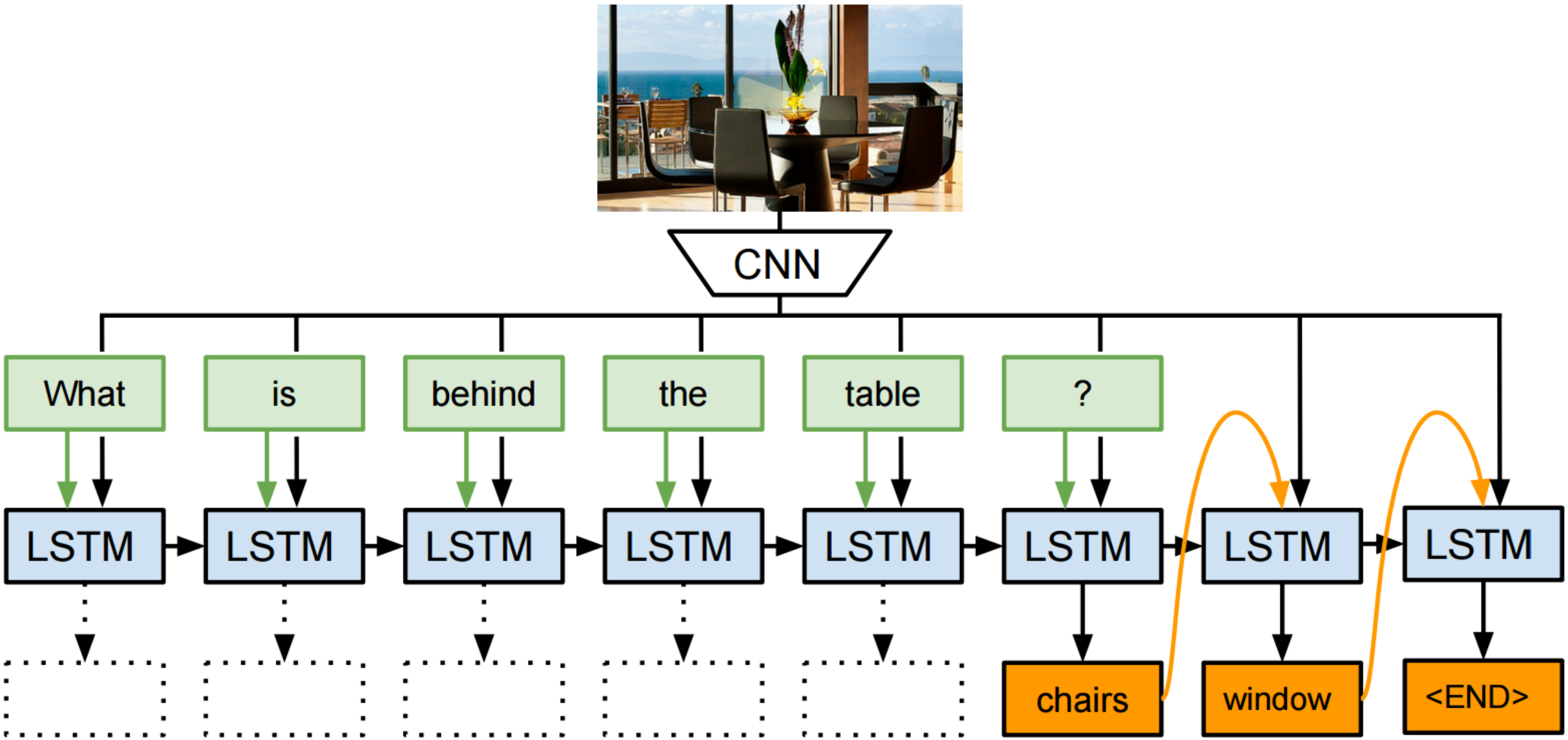
QA

Bed

Both can be processed with deep neural networks

Neural-based Question Answering

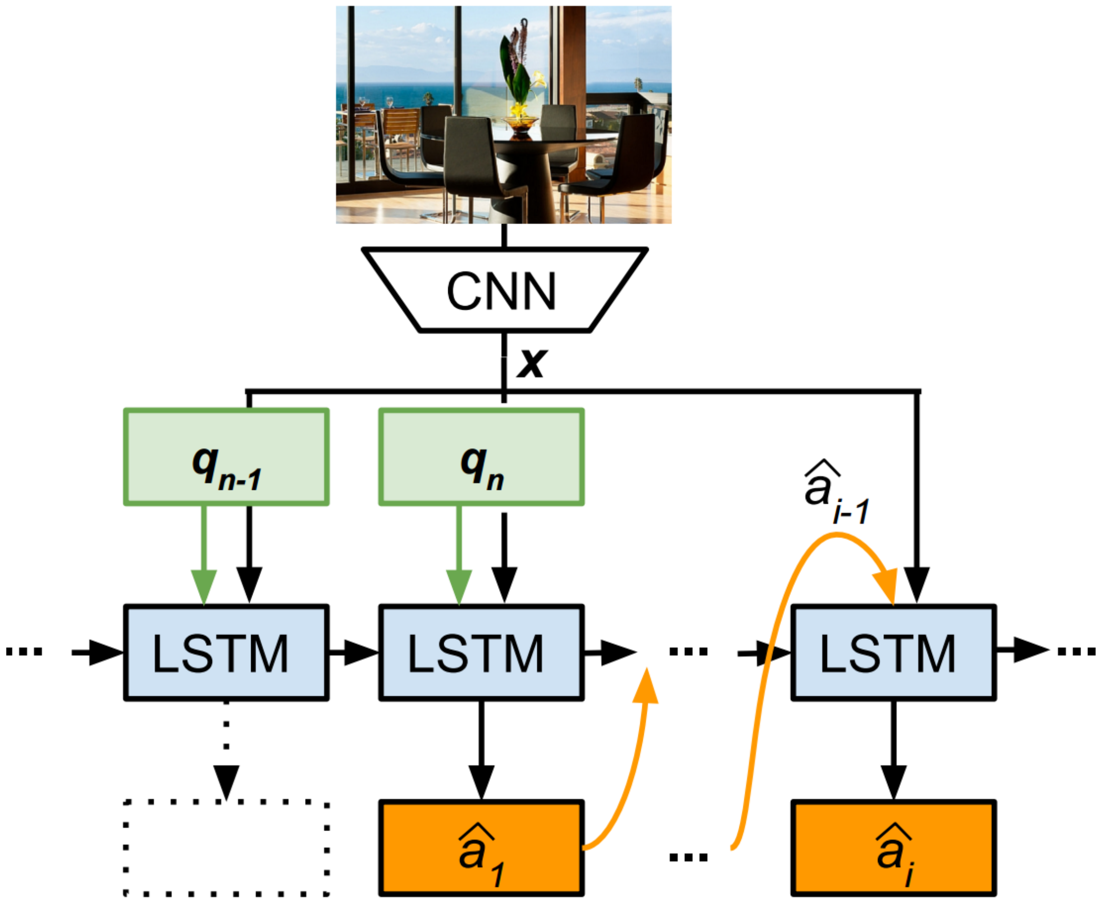
Architecture



Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Neural-based Question Answering

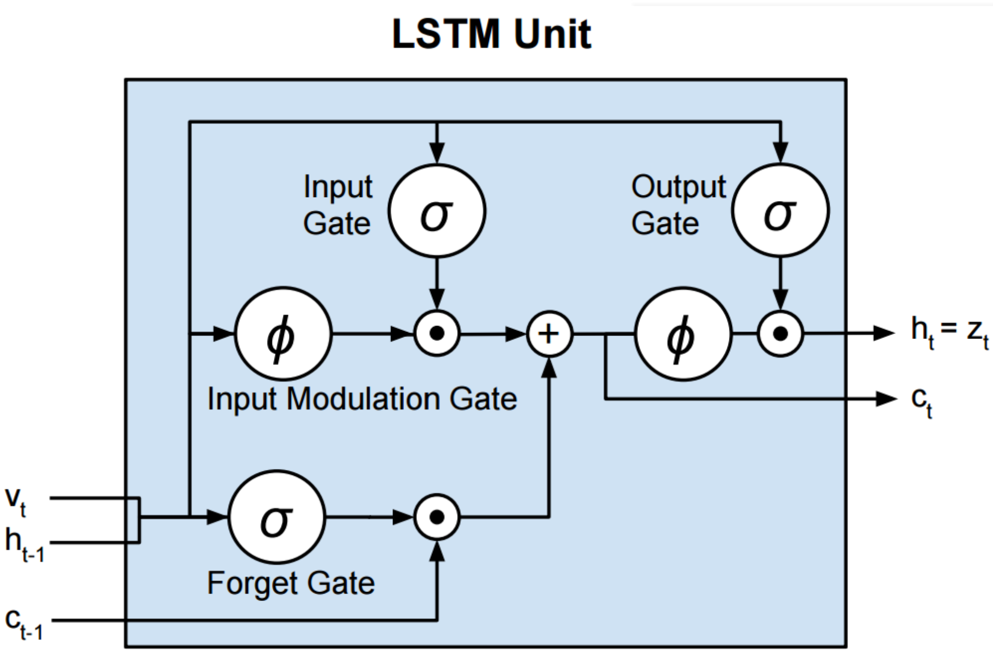
Architecture



Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Neural-based Question Answering

Architecture



$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma(W_{vi}\boldsymbol{v}_t + W_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma(W_{vf}\boldsymbol{v}_t + W_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma(W_{vo}\boldsymbol{v}_t + W_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{g}_t &= \phi(W_{vg}\boldsymbol{v}_t + W_{hg}\boldsymbol{h}_{t-1} + \boldsymbol{b}_g) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t)
\end{aligned}
$$

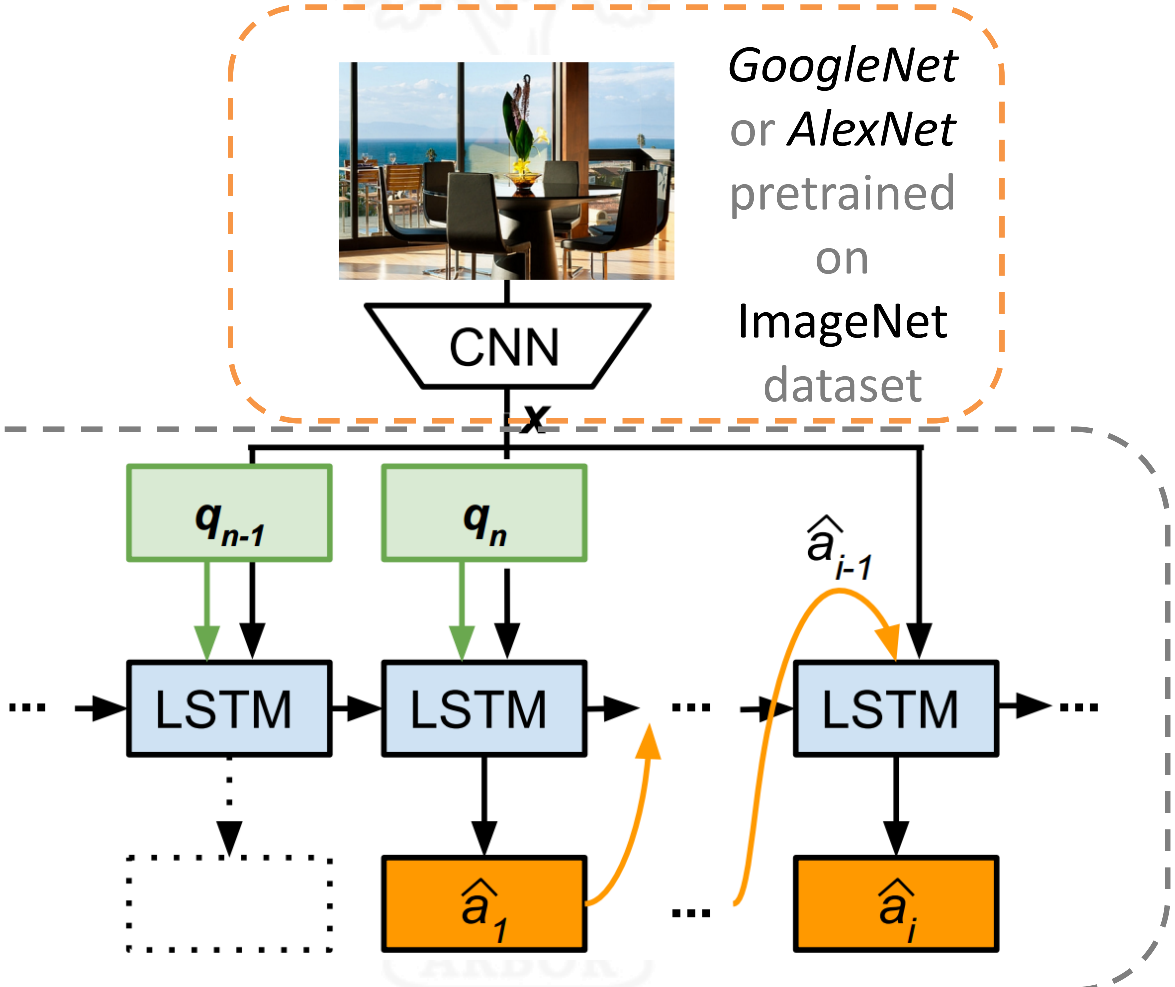
Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Neural-based Question Answering

Training

Loss function:

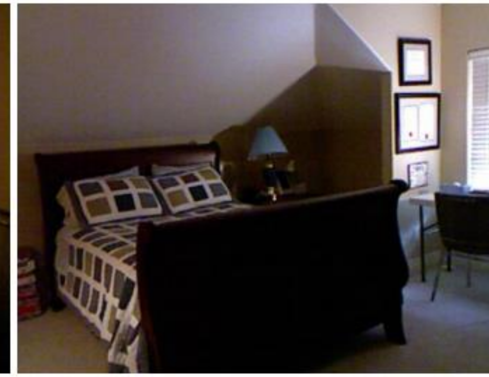
Cross entropy



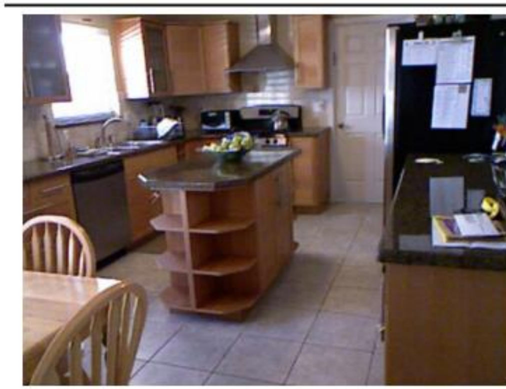
Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Result





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



	What is on the refrigerator?	What is the colour of the comforter?	What objects are found on the bed?
Neural-Image-QA:	magnet, paper	blue, white	bed sheets, pillow
Language only:	magnet, paper	blue, green, red, yellow	doll, pillow

Evaluation

WUPS

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

The best weighted match between answer & truth

Similarity based on the depth of two words in WordNet

WUP(curtain, blinds) = 0.94
WUP(carton, box) = 0.94
WUP(stove, fire extinguisher) = 0.82

WUPS @0.0 Smaller threshold

WUPS @0.9 More forgiving metric

Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).
Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In Advances in Neural Information Processing Systems (pp. 1682-1690).

Evaluation

Result

	Accu-racy	WUPS @0.9	WUPS @0.0
Malinowski et al. [20]	12.73	18.10	51.47
Neural-Image-QA (ours)			
- multiple words	29.27	36.50	79.47
- single word	**34.68**	**40.76**	79.54
Language only (ours)			
- multiple words	32.32	38.39	80.05
- single word	31.65	38.35	**80.08**

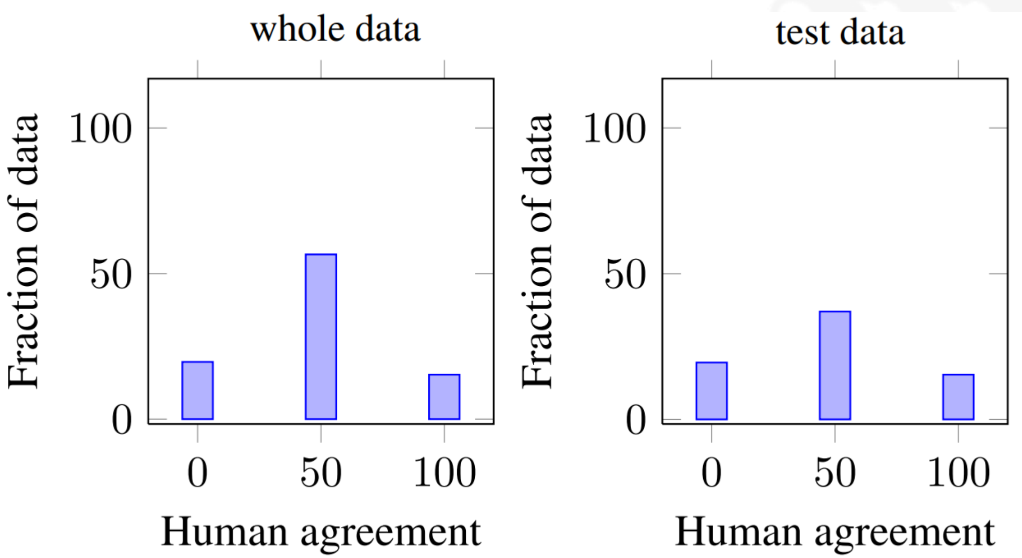
DAQUAR dataset

12, 468 human question answer pairs on images of indoor scenes

Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Evaluation

Consensus



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

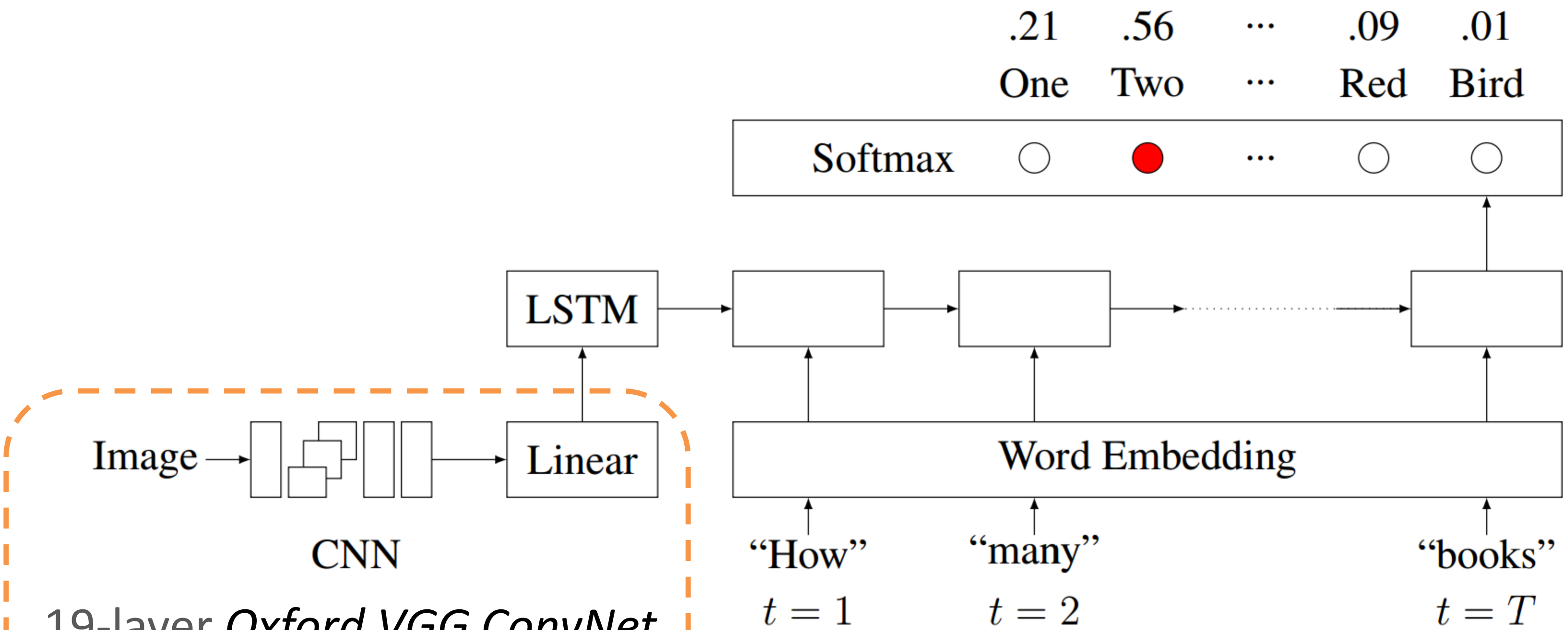
How plausible are the "ground truths"?

	Accu-racy	WUPS @0.9	WUPS @0.0
Subset: No agreement			
Language only (ours)			
- multiple words	8.86	12.46	38.89
- single word	8.50	12.05	40.94
Neural-Image-QA (ours)			
- multiple words	**10.31**	**13.39**	40.05
- single word	9.13	13.06	**43.48**
Subset: $\geq 50\%$ agreement			
Language only (ours)			
- multiple words	21.17	27.43	66.68
- single word	20.73	27.38	67.69
Neural-Image-QA (ours)			
- multiple words	20.45	27.71	67.30
- single word	**24.10**	**30.94**	**71.95**
Subset: Full Agreement			
Language only (ours)			
- multiple words	27.86	35.26	78.83
- single word	25.26	32.89	79.08
Neural-Image-QA (ours)			
- multiple words	22.85	33.29	78.56
- single word	**29.62**	**37.71**	**82.31**

Malinowski et al. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1-9).

Exploring Models and Data for Image Question Answering

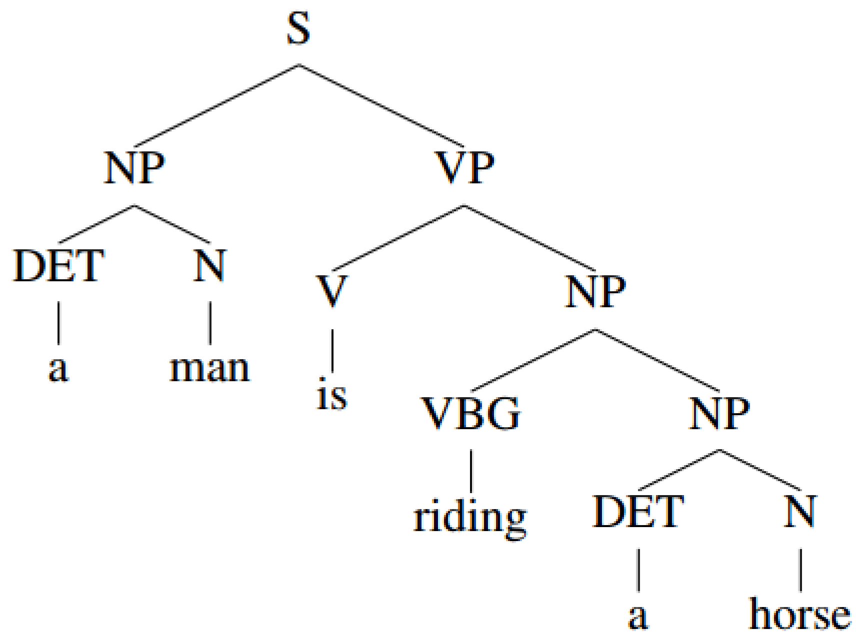
Architecture



19-layer ***Oxford VGG ConvNet***
trained on ***ImageNet***
Frozen during training

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

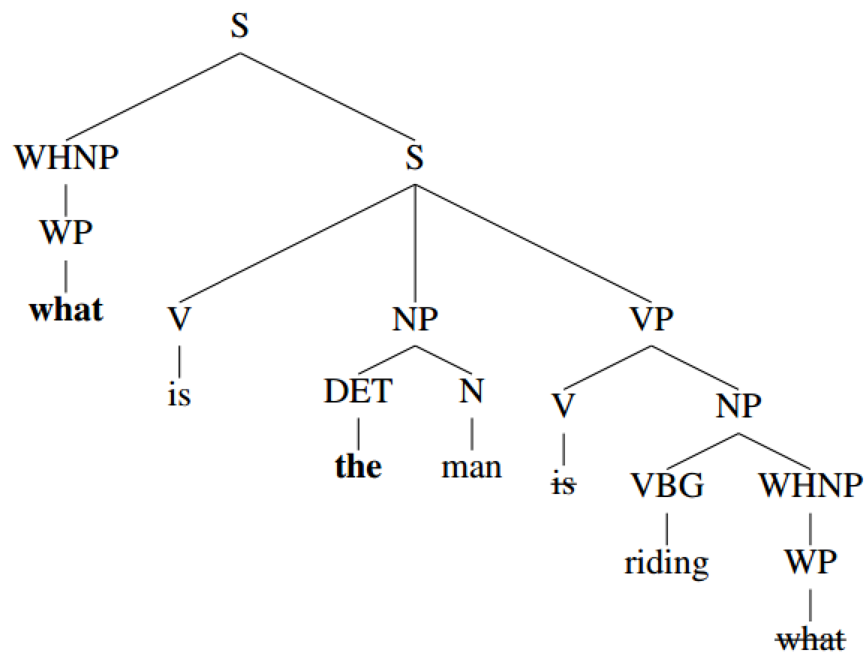
Question Generation



Generate question-answer pairs from image captions.

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

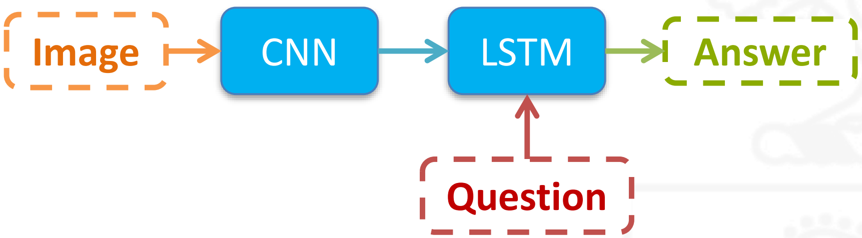
Question Generation



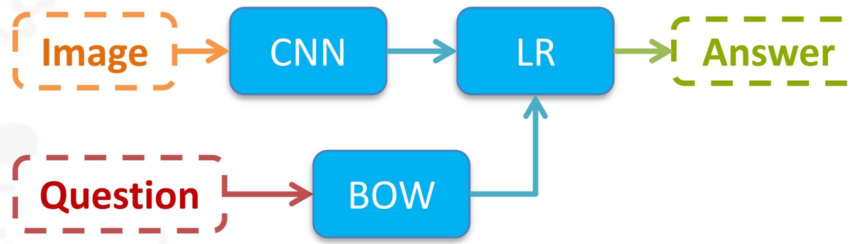
Generate question-answer pairs from image captions.

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

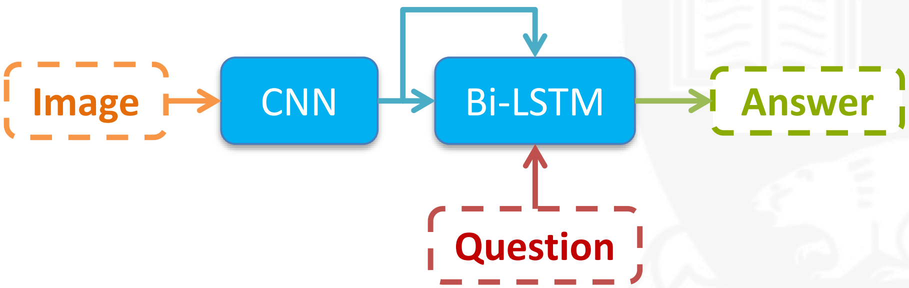
Evaluation



1. VIS+LSTM



3. IMG+BOW



2. 2-VIS+BLSTM

Average of the others

4. FULL

DAQUAR
12, 468 human question answer pairs

COCOQA
117,684 auto-generated question answer pairs

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

Evaluation

Overall

	DAQUAR			COCO-QA		
	ACC.	WUPS 0.9	WUPS 0.0	ACC.	WUPS 0.9	WUPS 0.0
MULTI-WORLD [32]	0.1273	0.1810	0.5147	-	-	-
GUESS	0.1824	0.2965	0.7759	0.0730	0.1837	0.7413
BOW	0.3267	0.4319	0.8130	0.3752	0.4854	0.8278
LSTM	0.3273	0.4350	0.8162	0.3676	0.4758	0.8234
IMG	-	-	-	0.4302	0.5864	0.8585
IMG+PRIOR	-	-	-	0.4466	0.6020	0.8624
K-NN (K=31, 13)	0.3185	0.4242	0.8063	0.4496	0.5698	0.8557
IMG+BOW	0.3417	0.4499	0.8148	**0.5592**	**0.6678**	**0.8899**
VIS+LSTM	0.3441	0.4605	**0.8223**	0.5331	0.6391	0.8825
ASK-NEURON [14]	0.3468	0.4076	0.7954	-	-	-
2-VIS+BLSTM	**0.3578**	**0.4683**	0.8215	0.5509	0.6534	0.8864
FULL	**0.3694**	**0.4815**	**0.8268**	**0.5784**	**0.6790**	**0.8952**
HUMAN	0.6027	0.6104	0.7896	-	-	-

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

Evaluation

Category

	Object	Number	Color	Location
GUESS	0.0239	0.3606	0.1457	0.0908
BOW	0.3727	0.4356	0.3475	0.4084
LSTM	0.3587	0.4534	0.3626	0.3842
IMG	0.4073	0.2926	0.4268	0.4419
IMG+PRIOR	-	0.3739	0.4899	0.4451
K-NN	0.4799	0.3699	0.3723	0.4080
IMG+BOW	**0.5866**	0.4410	**0.5196**	**0.4939**
VIS+LSTM	0.5653	**0.4610**	0.4587	0.4552
2-VIS+BLSTM	0.5817	0.4479	0.4953	0.4734
FULL	**0.6108**	**0.4766**	0.5148	**0.5028**

Ren et al. (2015). Exploring models and data for image question answering. In Advances in Neural Information Processing Systems(pp. 2935-2943).

Evaluation

COCOQA 23419
What is the black and white cat wearing?
Ground truth: hat
IMG+BOW: hat (0.50)
2-VIS+BLSTM: tie (0.34)
BOW: tie (0.60)

COCOQA 23419a
What is wearing a hat?
Ground truth: cat
IMG+BOW: cat (0.94)
2-VIS+BLSTM: cat (0.90)
BOW: dog (0.42)

DAQUAR 2136
What is right of table?
Ground truth: shelves
IMG+BOW: shelves (0.33)
2-VIS+BLSTM: shelves (0.28)
LSTM: shelves (0.20)

DAQUAR 2136a
What is in front of table?
Ground truth: chair
IMG+BOW: chair (0.64)
2-VIS+BLSTM: chair (0.31)
LSTM: chair (0.37)

COCOQA 11372
What do two women hold with a picture on it?
Ground truth: cake
IMG+BOW: cake (0.19)
2-VIS+BLSTM: cake (0.19)
BOW: umbrella (0.15)

DAQUAR 3018
What is on the right side?
Ground truth: table
IMG+BOW: tv (0.28)
2-VIS+LSTM: sofa (0.17)
LSTM: cabinet (0.22)

DAQUAR 1426
What is on the right side table?
Ground truth: tv
IMG+BOW: tv (0.25)
2-VIS+LSTM: tv (0.29)
LSTM: tv (0.14)

DAQUAR 1426a
What is on the left side of the room?
Ground truth: bed
IMG+BOW: door (0.19)
2-VIS+LSTM: door (0.25)
LSTM: door (0.13)

COCOQA 15756
What does the man rid while wearing a black wet suit?
Ground truth: surfboard
IMG+BOW: jacket (0.35)
2-VIS+LSTM: surfboard (0.53)
BOW: tie (0.30)

COCOQA 9715
What is displayed with the mattress off of it?
Ground truth: bed
IMG+BOW: bench (0.36)
2-VIS+LSTM: bed (0.18)
BOW: airplane (0.08)

COCOQA 25124
What is sitting in a sink in the rest room?
Ground truth: cat
IMG+BOW: toilet (0.77)
2-VIS+LSTM: toilet (0.90)
BOW: cat (0.83)

CSC 2523

Deep Learning in Computer Vision

Winter 2016

Thank You!

Yunpeng Li

Faculty of Information

2016.03.01