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

Author: Mateusz Malinowski, Marcus Rohrbach, Mario Fritz
Presenter: Hooman Shariati, Wen Xiao

1.Introduction

Q: How many chairs are on the right side of the table in the image ?

A: 3

Q: What is in front of the door and on the right of the table in the image ?

A: chair

Q: What is in front of the white board or in front of the door in the image ?

A: table, chair

Image and QA are from DAQUAR Dataset

Visual Turing Test

- It is a system that generates a random sequence of binary questions specific to the test image, such that **the answer to any question *k* is unpredictable given the true answers to the previous *k-1* questions** (also known as history of questions).
- Aim: evaluate the Image understanding of a computer system, and an important part of image understanding is the ***story line of the image***.
- Give a new direction to the computer vision research which would lead to the introduction of systems that will be one step closer to ***understanding images the way humans do***.

  - Wikipedia

1. Q: Is there a person in the blue region?	A: yes
2. Q: Is there a unique person in the blue region? (Label this person 1)	A: yes
3. Q: Is person 1 carrying something?	A: yes
4. Q: Is person 1 female?	A: yes
5. Q: Is person 1 walking on a sidewalk?	A: yes
6. Q: Is person 1 interacting with any other object?	A: no
⋮	
9. Q: Is there a unique vehicle in the yellow region? (Label this vehicle 1)	A: yes
10. Q: Is vehicle 1 light-colored?	A: yes
11. Q: Is vehicle 1 moving?	A: no
12. Q: Is vehicle 1 parked and a car?	A: yes
⋮	
14. Q: Does vehicle 1 have exactly one visible tire?	A: no
15. Q: Is vehicle 1 interacting with any other object?	A: no
17. Q: Is there a unique person in the red region?	A: no
18. Q: Is there a unique person that is female in the red region?	A: no
19. Q: Is there a person that is standing still in the red region?	A: yes
20. Q: Is there a unique person standing still in the red region? (Label this person 2)	A: yes
⋮	
23. Q: Is person 2 interacting with any other object?	A: yes
24. Q: Is person 1 taller than person 2?	A: amb.
25. Q: Is person 1 closer (to the camera) than person 2?	A: no
26. Q: Is there a person in the red region?	A: yes
27. Q: Is there a unique person in the red region? (Label this person 3)	A: yes
⋮	
36. Q: Is there an interaction between person 2 and person 3?	A: yes
37. Q: Are person 2 and person 3 talking?	A: yes

Image credit to Gemana et al.,PNAS 2015

Related Work (other tasks)

1. How can the computer ‘see’ the image?

CNN for Visual Recognition(as we did in A2)

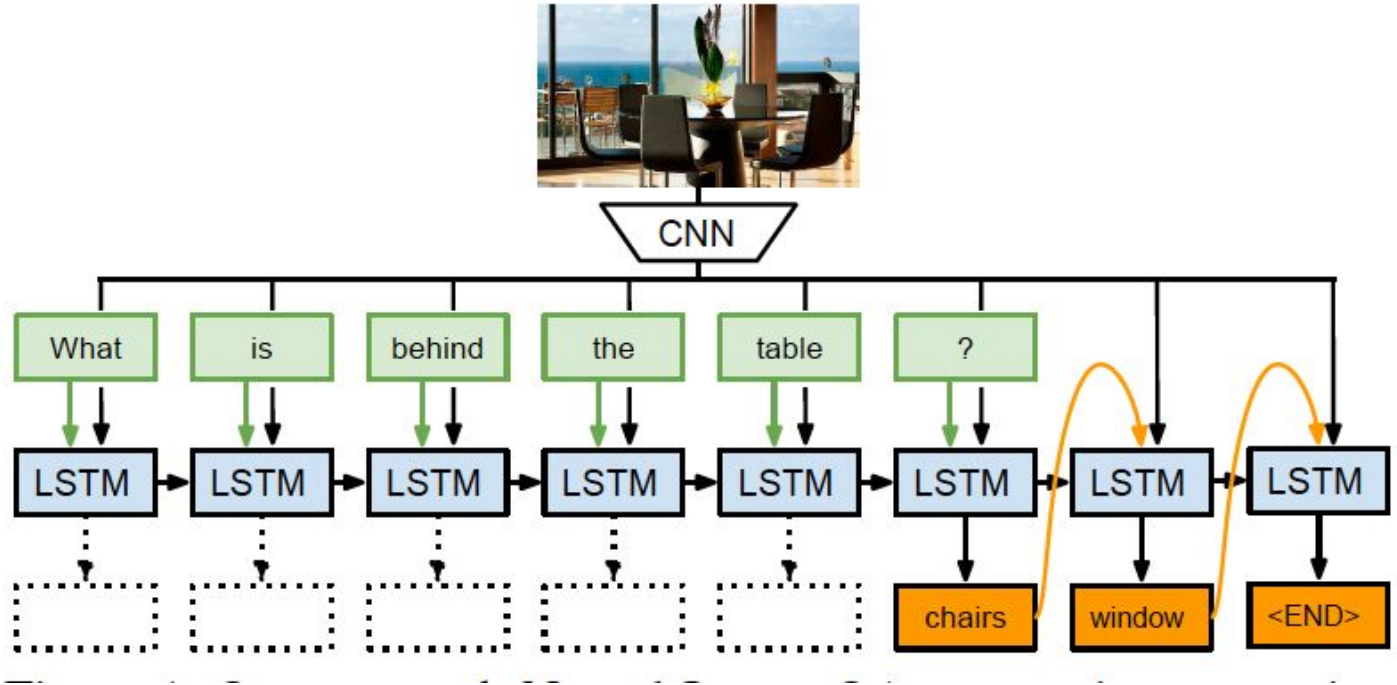
2. How can the computer ‘understand’ the question and ‘answer’ the question?

RNN/LSTM for sequence modeling(as we did in A3)

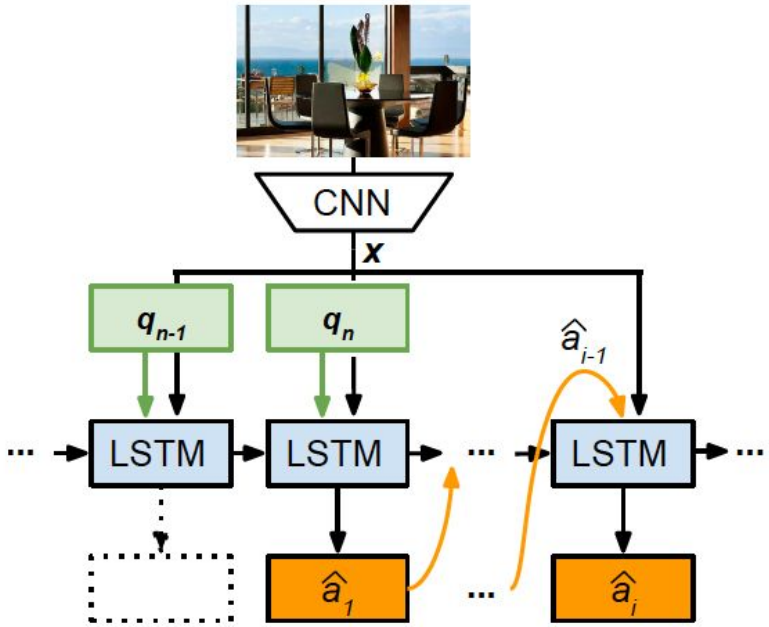
3. Other task to combine the image and text?

Image caption(as we did in A4), Image grounding

Model: Ask Your Neurons(End-to-end)



Model: Ask Your Neurons



Predict the answer:

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

$\boldsymbol{x}$ - image representation

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

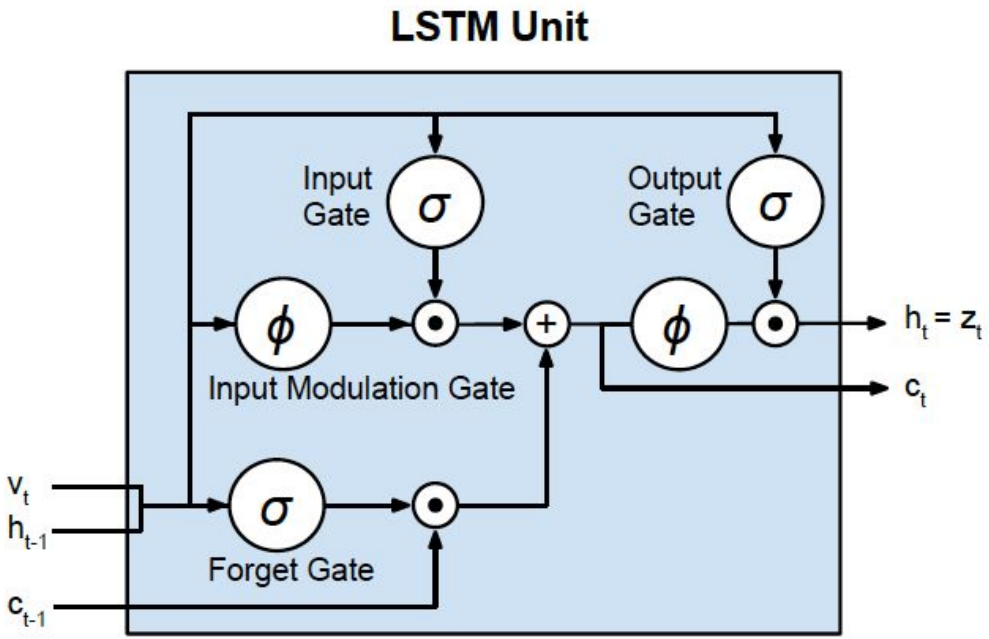
- question word sequence

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

- the set of previous words

Notice: Loss only at answer words

LSTM



Non-linearity:

$$\sigma(v) = (1 + e^{-v})^{-1}$$

$$\phi(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} = 2\sigma(2v) - 1$$

LSTM equations:

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

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

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

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

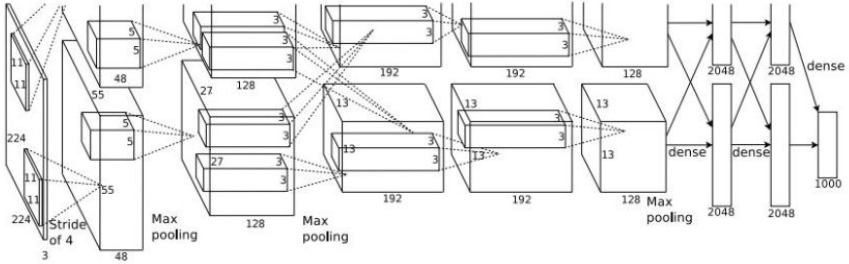
$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t$$

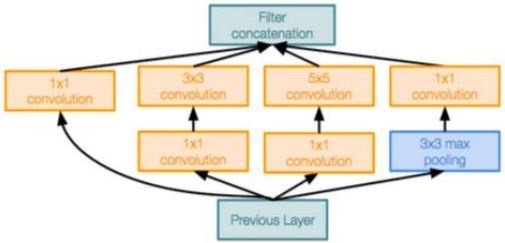
$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t)$$

Loss: Cross-entropy

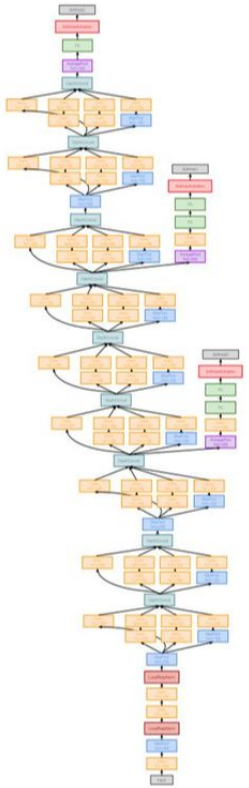
CNN

1. CNN models are pre-trained on ImageNet dataset
2. GoogleNet consistently outperforms AlexNet





Inception module

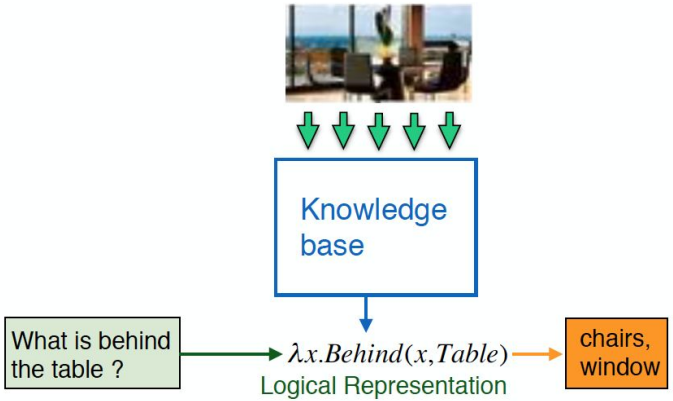


Images from Leonid's slides

Neural model vs Symbolic model

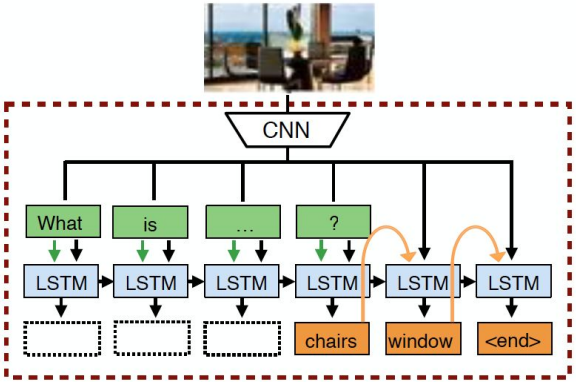
Symbolic approach (NIPS'14)

- Explicit representation
- Independent components
  - Detectors, Semantic Parser, Database
  - Segmented pictures
- Components trained separately
- Many 'hard' design decisions



Ask Your Neurons

- Implicit representation
- End-to-end formula
  - From images and questions to answers
- Joint training
- Fewer design decisions



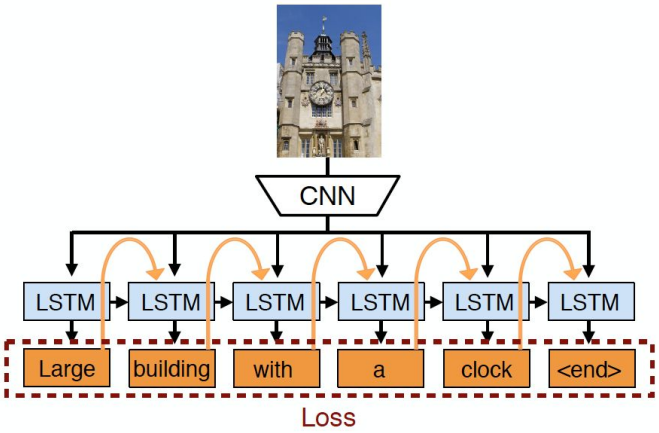
End-to-end, jointly trained architecture

Slides credit to Malinowski et al., oral presentation on ICCV,2015

Neural Visual QA vs Neural Image Caption

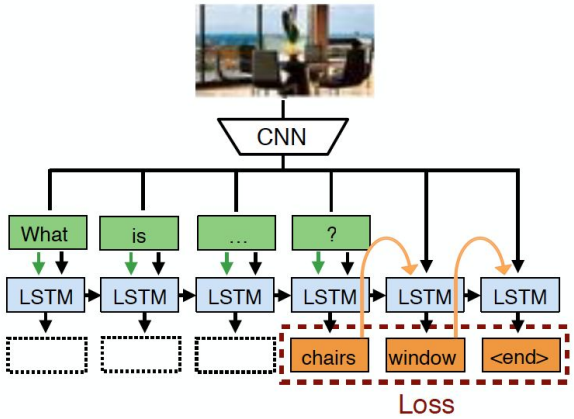
Neural Image Description

- Conditions on an image
- Generates a description
- Sequence of words
- Loss at every step
- Hard to validate
  - Diversity of description



Ask Your Neurons (Our)

- Conditions on an image and a question
- Generates an answer
  - Sequence of answer words
- Loss only at answer words
- Easy to validate
  - Generally, questions has unique answers



Slides credit to Malinowski et al., oral presentation on ICCV,2015

Training:

GoogleNet pretrained on ImageNet.

Default hyper parameters for LSTM and CNN.

Randomly initialized the last FC layer of CNN, trained together with LSTM.

Train, validate and test on the same dataset DAQUAR as their previous work

No information on the training/validation/test set split of their data, or on the definition of their accuracy metrics.

We assumed it was the same as in their previous work.

Dataset

DAtaset for QUestion Answering on Real-world images (DAQUAR)

795 training images 6795 question-answer pairs

653 test images 5673 question-answer pairs

Asked 5 humans to provide questions and answers the only instructions were:

“Provide valid questions and answers related to basic colors, numbers, or types of both objects and sets of objects”

some biases showing humans tend to focus on a few prominent objects. For instance we have more than 400 occurrences of table and chair in the answers.

QA: (What is behind the table?, window)
Spatial relation like 'behind' are dependent on the reference frame. Here the annotator uses observer-centric view.

QA: (what is beneath the candle holder, decorative plate)
Some annotators use variations on spatial relations that are similar, e.g. 'beneath' is closely related to 'below'.

QA: (what is in front of the wall divider?, cabinet)
Annotators use additional properties to clarify object references (i.e. wall divider). Moreover, the perspective plays an important role in these spatial relations interpretations.

The annotators are using different names to call the same things. The names of the brown object near the bed include 'night stand', 'stool', and 'cabinet'.

Some objects, like the table on the left of image, are severely occluded or truncated. Yet, the annotators refer to them in the questions.

QA: (What is in front of toilet?, door)
Here the 'open door' to the restroom is not clearly visible, yet captured by the annotator.

QA: (what is behind the table?, sofa)
Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view
QA: (how many lights are on?, 6)
Moreover, some questions require detection of states 'light on or off'

Q: what is at the back side of the sofas?
Annotators use wide range spatial relations, such as 'backside' which is object-centric.

QA1: (what is in front of the curtain behind the armchair?, guitar)

QA2: (what is in front of the curtain?, guitar)

Spatial relations matter more in complex environments where reference resolution becomes more relevant. In cluttered scenes, pragmatism starts playing a more important role

QA1:(How many doors are in the image?, 1
QA2:(How many doors are in the image?, 5
Different interpretation of 'door' results in different counts: 1 door at the end of the hall vs. 5 doors including lockers

QA: (How many drawers are there?, 8)
The annotators use their common-sense knowledge for amodal completion. Here the annotator infers the 8th drawer from the context

QA: (What is the shape of the green chair?, horse shaped)
In this example, an annotator refers to a "horse shaped chair" which requires a quite abstract reasoning about the shapes.

QA: (What is the object on the counter in the corner?, microwave)
References like 'corner' are difficult to resolve given current computer vision models. Yet such scene features are frequently used by humans.

QA: (How many doors are open?, 1)
Notion of states of object (like open) is not well captured by current vision techniques. Annotators use such attributes frequently for disambiguation.

QA: (Where is oven?, on the right side of refrigerator)
On some occasions, the annotators prefer to use more complex responses. With spatial relations, we can increase the answer's precision.

Figure credit to Malinowski et al., Multi Question ICCV,2014

Evaluation metrics:

1. Strict string matching: $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{A^i = T^i\} \cdot 100$

2. Semantic matching using WUP to account for word-level ambiguities:

    a. I.e ‘carton’ and ‘box’ can be associated with similar concepts, so the model should not be strongly penalized for this type of mistakes.

Wu-Palmer (WUP) word similarity measure

1. Based on edge counting in a taxonomy like WorldNet or Ontology
2. WUP also weights the edges based on distance in the hierarchy.
    a. Ex: Going from inanimate to animate is a larger distance than going from Felid to Canid.
3. WordNet:
    a. Large lexical database of English words grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept
    b. Synsets are interlinked by means of conceptual-semantic and lexical relations

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

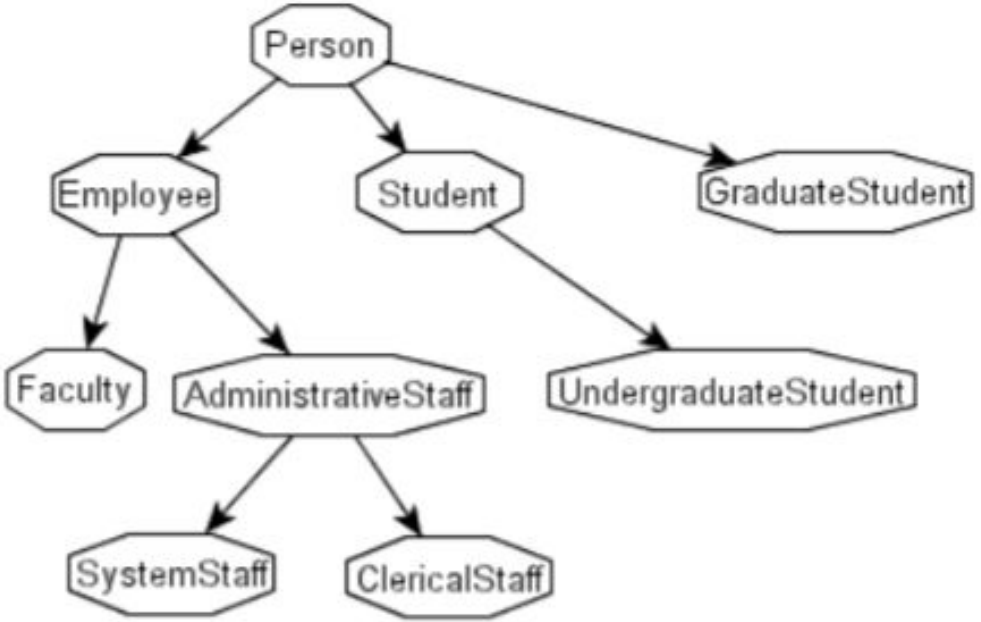


Figure credit to Malinowski et al., Multi Question ICCV,2014

WUPS (WUP Set)

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

Multiply WUP(a, b) with 0.1 whenever WUP(a, b) < t

For precise answers, consider to words similar if WUP(a, b) > 0.9

t = 1, is same as string matching

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

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

Table credit to Malinowski et al., oral presentation on ICCV,2015

Evaluation

- Comparison with previous approach based on semantic parsing
- Comparison with how well questions can be answered without images
- Tried different subsets of the dataset and different accuracy metrics

  (to boost their score?)

	Accu-racy	WUPS @0.9	WUPS @0.0
Malinowski et al. [20]	7.86	11.86	38.79
Neural-Image-QA (ours)			
- multiple words	17.49	23.28	57.76
- single word	**19.43**	**25.28**	**62.00**
Human answers [20]	50.20	50.82	67.27
Language only (ours)			
- multiple words	17.06	22.30	56.53
- single word	17.15	22.80	58.42
Human answers, no images	7.34	13.17	35.56

Table credit to Malinowski et al., oral presentation on ICCV,2015

- Their performance drops dramatically with longer answers.

- They mention dataset bias:
  - 90% of the answers contain a single word

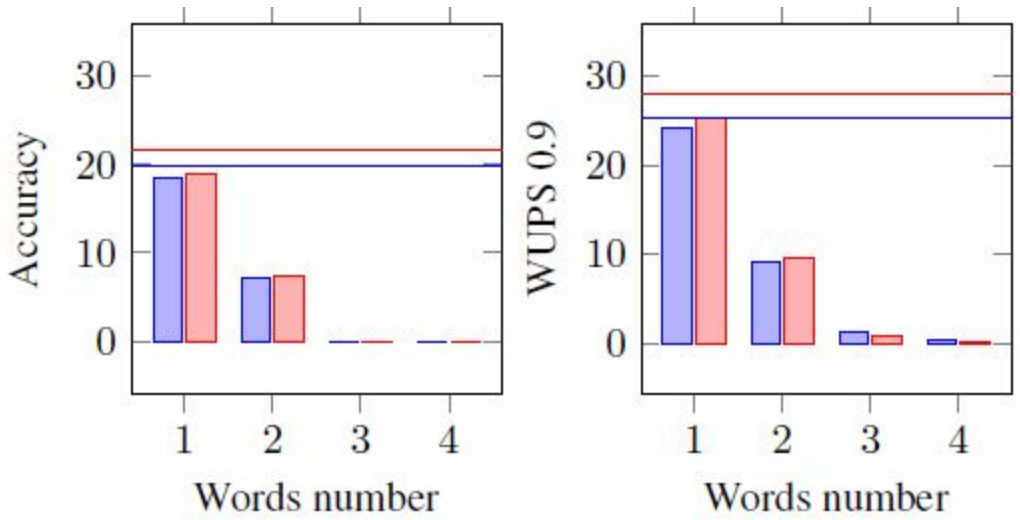


Table credit to Malinowski et al., oral presentation on ICCV,2015

- 5 additional test answers for each image-question pair (by 5 additional people).
- Same directions as before.
- Their explanation as to why the benchmark performance of humans was 50%.

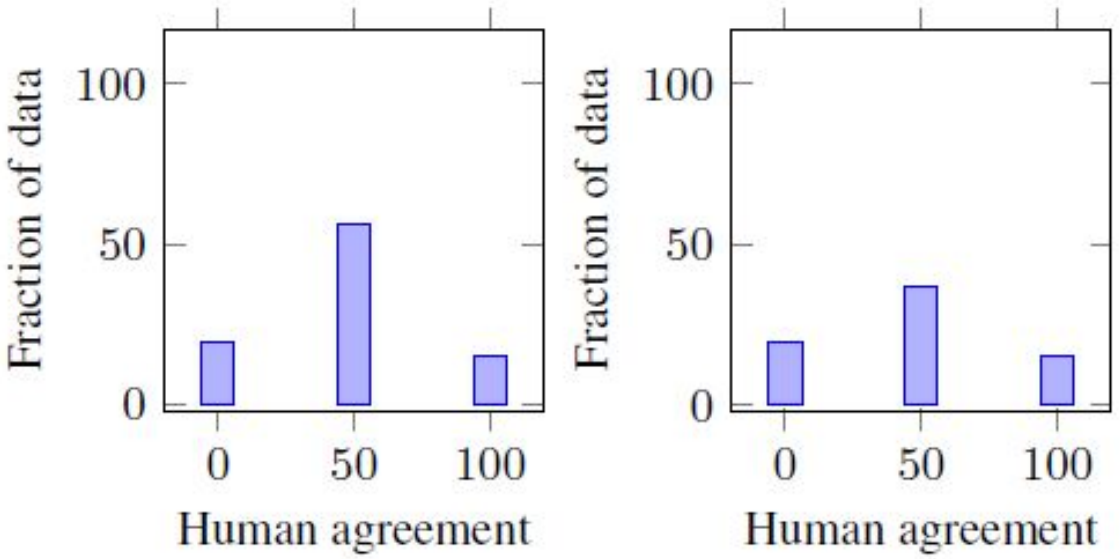


Table credit to Malinowski et al., oral presentation on ICCV,2015

Two new scores to capture consensus

- Average Consensus Metric (ACM): Prefers mainstream answers.

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

- Min Consensus Metric (MCM): Prefers closest matching answers.

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

	Accu-racy	WUPS @0.9	WUPS @0.0			
Average Consensus Metric						
Language only (ours)						
- multiple words	11.60	18.24	52.68	17.06	22.30	56.53
- single word	11.57	18.97	54.39	17.15	22.80	58.42
Neural-Image-QA (ours)						
- multiple words	11.31	18.62	53.21	17.49	23.28	57.76
- single word	**13.51**	**21.36**	**58.03**	**19.43**	**25.28**	**62.00**
Min Consensus Metric						
Language only (ours)						
- multiple words	22.14	29.43	66.88			
- single word	22.56	30.93	69.82			
Neural-Image-QA (ours)						
- multiple words	22.74	30.54	68.17			
- single word	**26.53**	**34.87**	**74.51**			

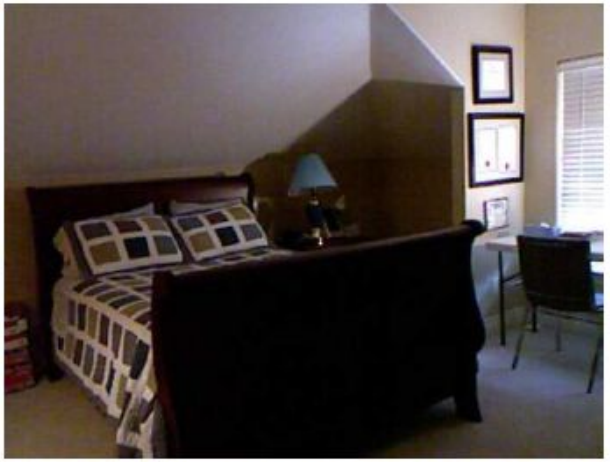
Table credit to Malinowski et al., oral presentation on ICCV,2015

	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**

Table credit to Malinowski et al., oral presentation on ICCV,2015

Some examples

Counting Questions:



	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

Table credit to Malinowski et al., oral presentation on ICCV,2015

Color Questions

	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

Table credit to Malinowski et al., oral presentation on ICCV,2015

Spatial Relationship Questions

How many chairs are there?

What is the object fixed on the window?

Which item is red in colour?

Neural-Image-QA:	1	curtain	remote control
Language only:	4	curtain	clock
Ground truth answers:	2	handle	toaster

Table credit to Malinowski et al., oral presentation on ICCV,2015

Discussion

Strength and Weaknesses:

- Novel approach to an interesting problem.
  - But weak on evaluation. And low on implementation details
- Compared only to their own previous works (without even mentioning the architecture of their previous works or making a comparison).
- Unclear about training/test/validation splits and training parameters
- Modified both the dataset and evaluation metrics to boost their scores, to no avail.
  - Changed the number of answer words
  - Changed the number of provided ground truth answers for each question
  - Used several metrics

Difficulty with Spatial Relations

- Perform relatively well on “what color” and “how many” questions, but they have difficulty with questions like “what is to the left of the fridge”.
    - Could be due to CNNs.
    - Providing more spatial information through an attention mechanism might help

- Also, difficulty with small objects, questions with negations, and shapes.
    - They attribute this to under-representation of these cases in training data

Doesn’t Learn Enough From Images

- Humans answer 7.34% without images, and 50.20 % with images.
- Their system answers 17.06% without images, but only 17.49% with images.

- Our suggestions:
  - Increase the learning capacity of the portion of the model that learns from images
  - Encode the entire question first and pass it along with the image to a seperate LSTM
  - Pre-train the LSTM on a different question/answers set to reduce dependence on the particular question/answers contained in the training set.

VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

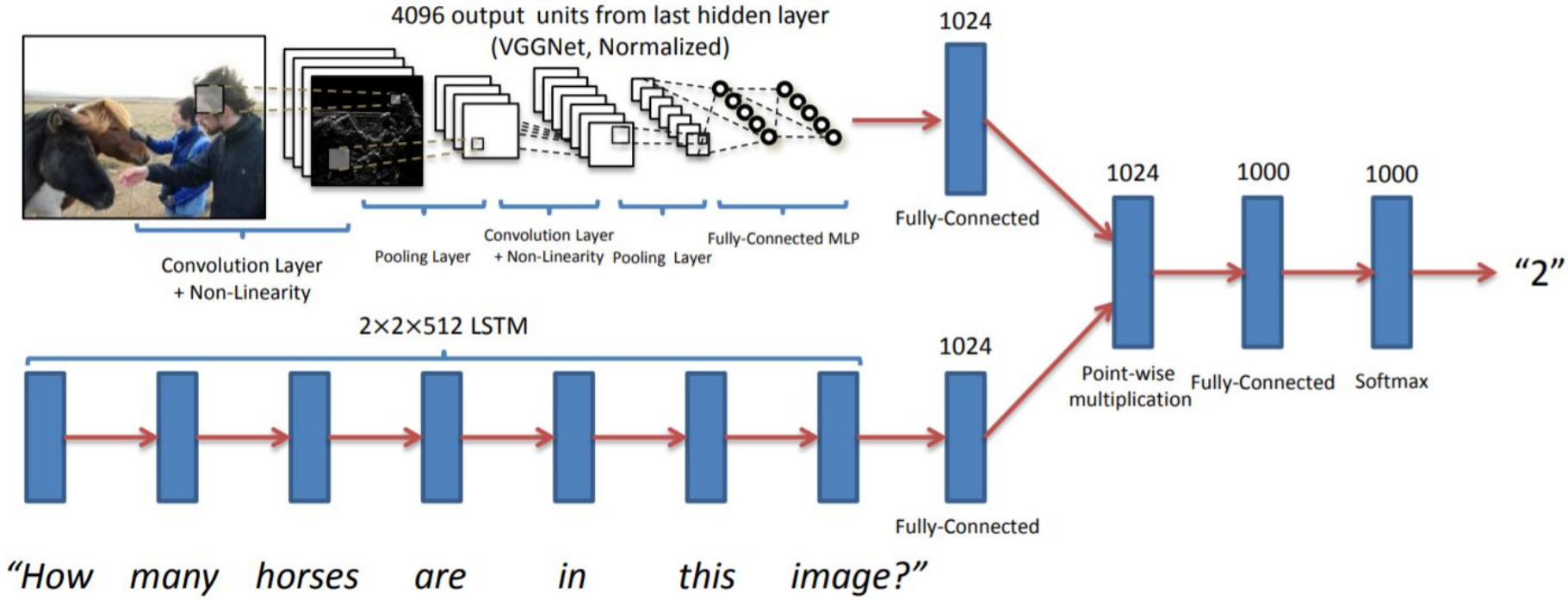
Abstract—We propose the task of *free-form* and *open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ~0.25M images, ~0.76M questions, and ~10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance.

1 Introduction

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [14], [7], [10], [36], [24], [22], [51]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word n-gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as "AI-complete" as desired.

What makes for a compelling "AI-complete" task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a single sub-domain (such as CV) and (ii) have a well-defined *quantitative evaluation metric* to track progress. For some

Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.



http://vqa.cloudcv.org/

Credit to Agrawal et al., Visual Question answering, ICCV,2016

Visual Question Answering(CVPR 2016)

- Visual Question Answering Dataset (VQA):
  - 250K images (COCO and abstract scenes)
  - 760K questions
  - 10M answers by multiple people
  - "yes/no", "number", and "object" answers; majority single word
  - Has confidence and Consensus measures (i.e. how many people agree on a given answer)

- Opens the way for automatic evaluation
  - many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format: http://visualqa.org/visualize/

- Adds human baseline performance and compares previous VQA methods

Is something under the sink broken?	yes	no
	yes	no
	yes	no
What number do you see?	33	5
	33	6
	33	7

Can you park here?	no	no
	no	no
	no	yes
What color is the hydrant?	white and orange	red
	white and orange	red
	white and orange	yellow

What kind of store is this?	bakery	art supplies
	bakery	grocery
	pastry	grocery
Is the display case as full as it could be?	no	no
	no	yes
	no	yes

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

Does this man have children?	yes	yes
	yes	yes
	yes	yes
Is this man crying?	no	no
	no	yes
	no	yes

Has the pizza been baked?	yes	yes
	yes	yes
	yes	yes
What kind of cheese is topped on this pizza?	feta	mozzarella
	feta	mozzarella
	ricotta	mozzarella

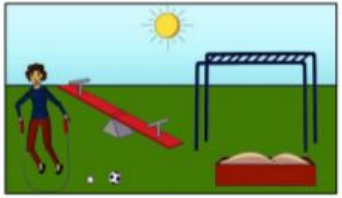
How many pickles are on the plate?	1	1
	1	1
	1	1
What is the shape of the plate?	circle	circle
	round	round
	round	round

What does the sign say?	stop	stop
	stop	stop
	stop	yield
What shape is this sign?	octagon	diamond
	octagon	octagon
	octagon	round

How many glasses are on the table?	3	2
	3	2
	3	6
What is the woman reaching for?	door handle	fruit
	glass	glass
	wine	remote

Do you think the boy on the ground has broken legs?	yes	no
	yes	no
	yes	yes
Why is the boy on the right freaking out?	his friend is hurt	ghost
	other boy fell down	lightning
	someone fell	sprayed by hose

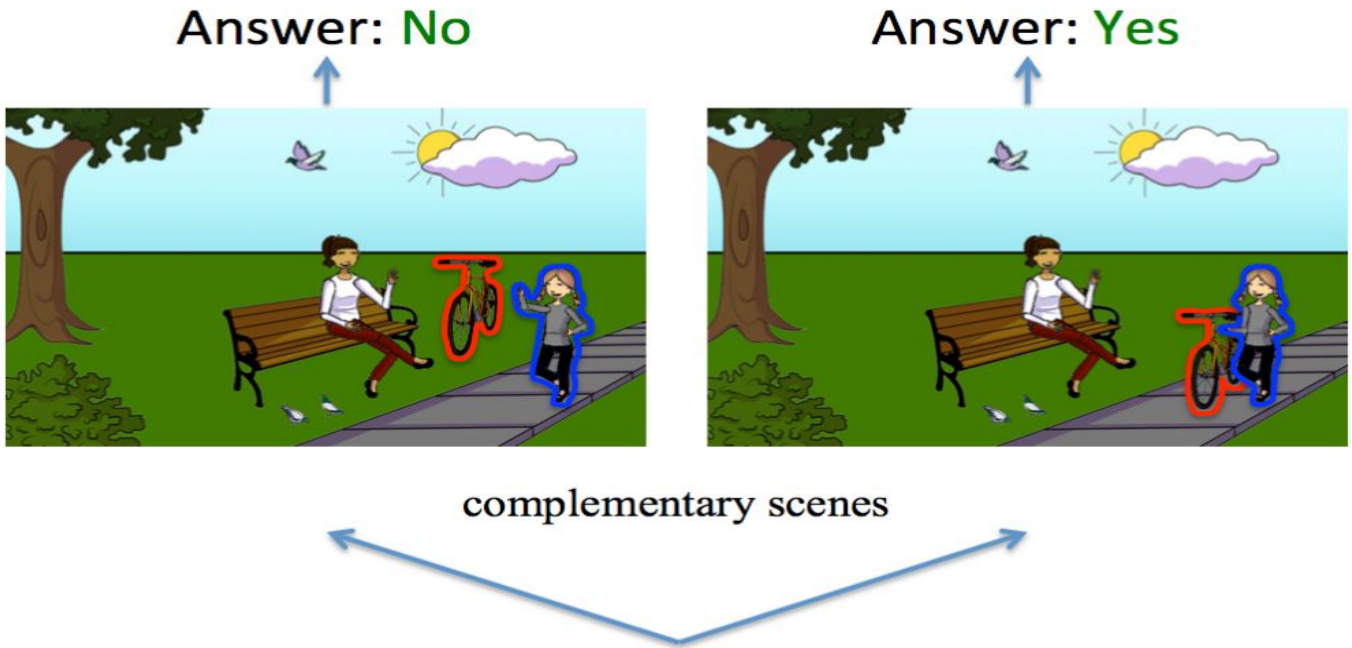
Are the kids in the room the grandchildren of the adults?	probably	yes
	yes	yes
	yes	yes
What is on the bookshelf?	nothing	books
	nothing	books
	nothing	books



How many balls are there?	2	1
	2	2
	2	3
What side of the teeter totter is on the ground?	right	left
	right	left
	right side	right side

Credit to Agrawal et al., Visual Question answering, ICCV,2016

Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)



Credit to Zhanget al., Yin and Yang, ICCV,2016

Making the V in VQA Matter: (CVPR 2017)

- Answers why models ignore visual information
  - Inherent structure in our world and bias in language are easier signals to learn from

- Suggests a way to counter language priors
  - For each question, collect complementary images such that every question is associated with pair of similar images that result in two different answers to the question.

- Balanced VQA dataset

Who is wearing glasses?

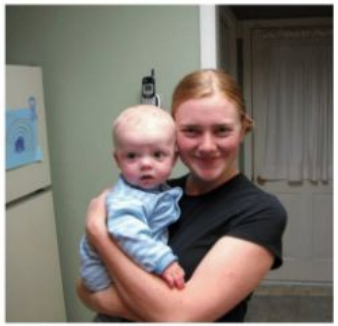
man

woman

Where is the child sitting?

fridge

arms



Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

Credit to Goyal et al., making V in VQA, ICCV,2017