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

Mateusz Malinowski[1], Marcus Rohrbach[2], Mario Fritz[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany [2]UC Berkeley EECS and ICSI, Berkeley, CA, United States

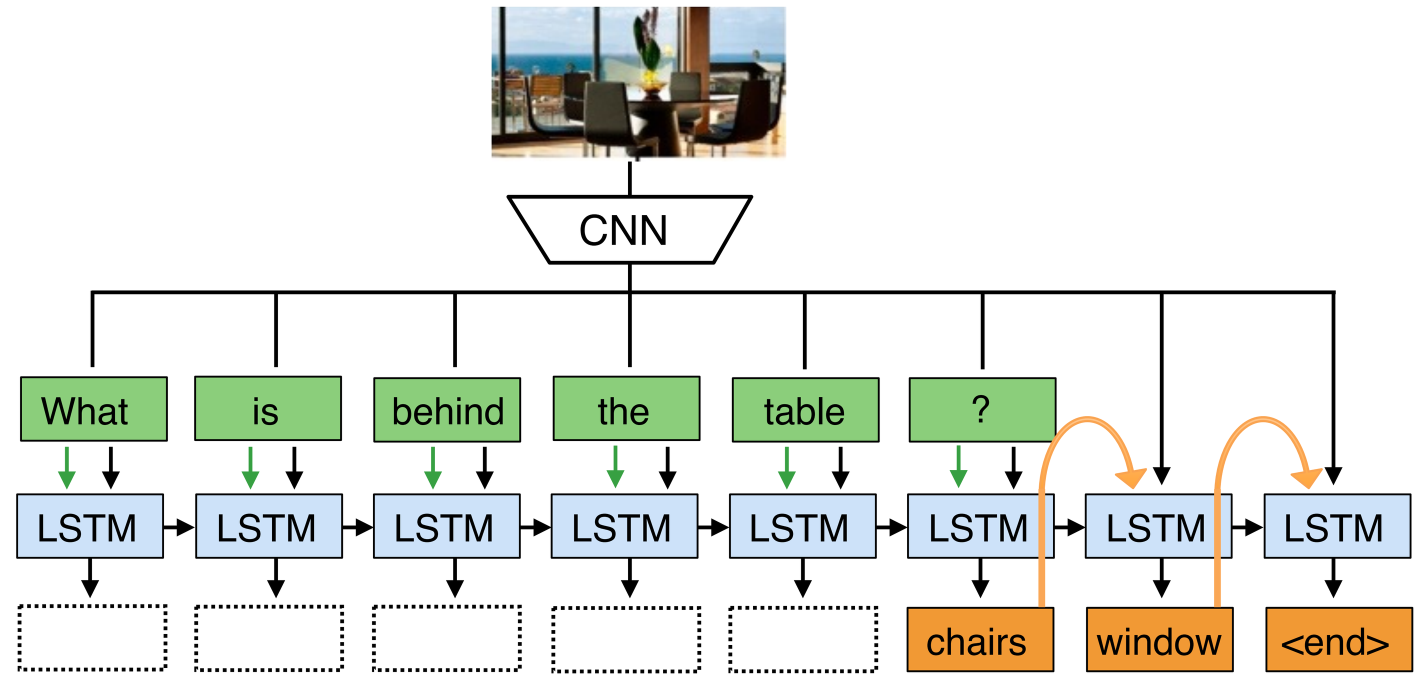
www.d2.mpi-inf.mpg.de/visual-turing-challenge

What is hanged on the chair? clothes

What is on the refrigerator? magnet, paper

What color are the cabinets? brown

How many lamps are there? 2



How many burner knobs are there?
Vision + Language: 4
Language Only: 6

What objects are found on the bed?
Vision + Language: bed sheets, pillow
Language Only: doll, pillow

What are around dining table?
Vision + Language: chair
Language Only: chair

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

Summary

Motivation

- Defining a task that benchmarks visual comprehension
  - Easy for humans, challenging for machines
  - Easy to automatically evaluate
  - Agnostic to an internal representation
  - Scalable annotation effort
- Can machines answer questions about images?
  - Meaning of a scene depends on the task (question)

Goal

- End-to-end, jointly trained neural approach for answering questions about images
- Automatic performance measures that account for many scene and question interpretations

Approach

- Novel neural-based architecture with results on language-only model
  - Doubles the performance of the prior symbolic method
  - Global image representation (CNN)
  - Capable of multi-word answer generations
- Consensus metrics to measure performance

References

[1] M. Malinowski et. al. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. NIPS'14.
[2] M. Malinowski et. al. Towards a Visual Turing Challenge. NIPS'14 Workshop.
[3] M. Malinowski et. al. Hard to Cheat: A Turing Test based on Answering Questions about Images. AAAI'15 Workshop.
[4] N. Silberman et. al. Indoor segmentation and support inference from RGBD images. ECCV'12.
[5] S. Gupta et. al. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. CVPR'13.
[6] J. Van De Weijer et. al. Learning Color Names From Real-World Images. CVPR'07.
[7] P. Liang et. al. Learning Dependency-based Compositional Semantics. Computational Linguistics'13.
[8] J. Donahue et. al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR'15.
[9] C. Szegedy et. al. Going Deeper with Convolutions. CVPR'15.

Dataset

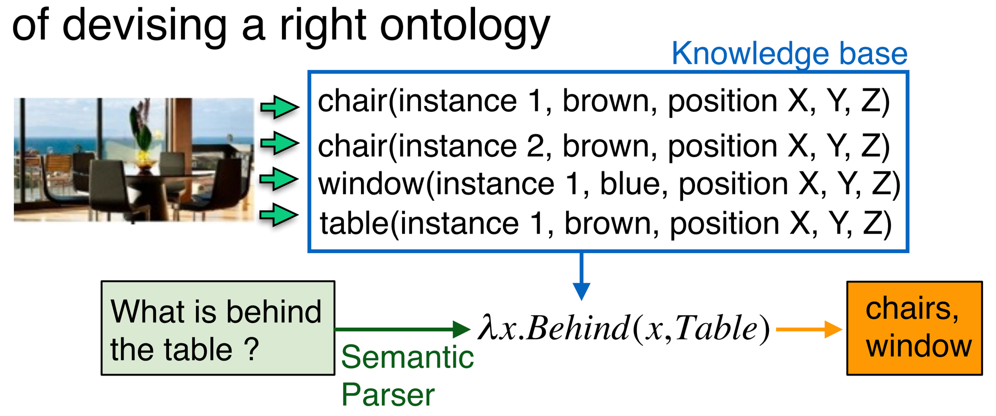
DAQUAR [1]

- Indoor images
  - Based on NYU-Depth V2 dataset [4]
- 1449 RGBD images
- 12.5k Image-Question-Answer triples
  - Around 9 QA pairs per image
- Questions about objects, set of objects, colors, numbers, and sizes of the objects
- Subjectivity is dominant in the dataset
  - Spatial relations exhibit different reference frames
  - Same objects are referred by multiple names
    - Night stand, stool, cabinet
  - Subjective objects saliency

Prior Symbolic Approach

Symbolic-based Approach [1]

- Symbolic chain of perception, knowledge representation and formal deduction system
- Scene analysis techniques such as semantic segmentation [5] and color detector [6] extract a visual 'knowledge' from images
- Semantic parser [7] transforms a question into its meaning using hand-designed predicates
  - Formal language of meaning
- Many design choices, poor scalability, problem of devising a right ontology



Ask Your Neurons

Overview

- Neural-based approach that conditions on an image and a question, generates an answer
- Implicit representation
- End-to-end formulation
- Joint training
- Natural and weak supervision
  - Architecture is directly trained on the image-question-answer triples
- A few design choices

Language-Only (Neural Blind)

- Trained only on question-answer pairs, without seeing images
- Competitive performance
  - Some answers can be decoded solely based on questions (e.g. chairs often surrounds a table)
  - To achieve a good performance handling language is important
- Answers of similar questions don't change
- Around 17.5 Acc and 23.3 WUPS@0.9

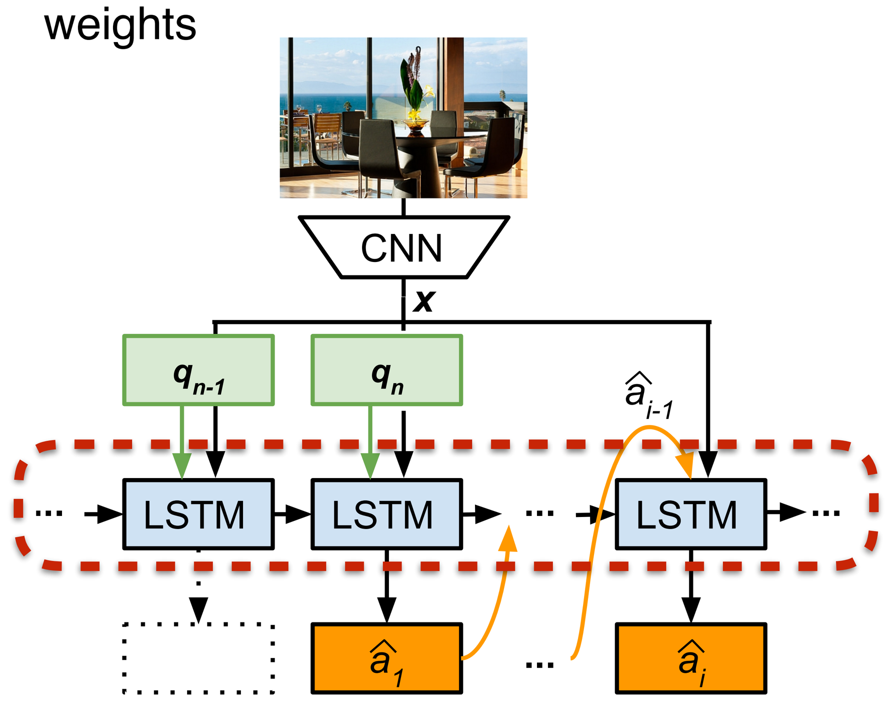
Vision + Language (Neural Image)

- Multimodal
  - Conditions on both language and image
  - Uses LSTM for language modeling
  - Uses CNN for image modeling
- Global visual representation
- Best performance: around 19.4 Acc and 25.3 WUPS@0.9

LSTM and CNN

Multiple-words Answer Generation

- Our architecture is trained to generate multiple words answers
- Answers at each step are fed back to LSTM
- Can be seen as an encoder-decoder architecture with two LSTM [8] and shared weights



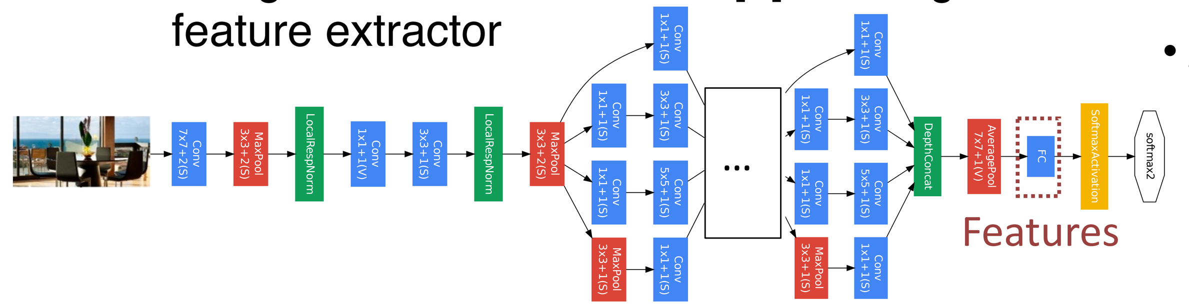
$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta})$, $\boldsymbol{x}$ - image representation, $\mathcal{V}$ - vocabulary

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

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

CNN

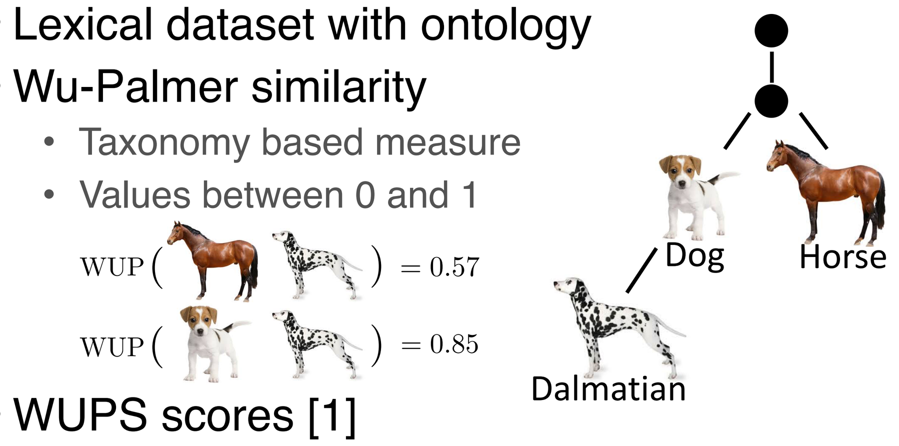
- Global visual representation
- GoogleNet-like architecture [9] as image feature extractor



Performance Metrics

WUPS

- Limitations of Accuracy
  - Acc(Dalmatian, Dog) = Acc(Horse, Dog)
- Lexical dataset with ontology
- Wu-Palmer similarity
  - Taxonomy based measure
  - Values between 0 and 1

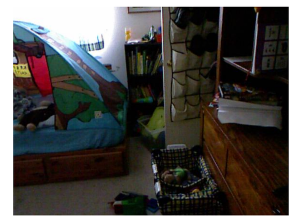


- WUPS scores [1]
  - Embrace word-level ambiguities
  - Soft, set-based generalization of Accuracy

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

Consensus

- Limitations of WUPS
  - Doesn't account for many question and scene interpretations



What is the object on the floor in front of the wall?
Human 1: bed
Human 2: shelf
Human 3: bed
Human 4: bookshelf

- Min Consensus
  - Scores for at least one matching ground truth

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

- Average Consensus
  - Measures agreement of the answers
  - Down-weight 'controversial' answers

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

Quantitative Results

Standard Metrics

Method	Accuracy	WUPS 0.9
Symbolic QA [2]	7.86	11.86
Neural Image QA (single-word)	19.43	25.28
Neural Image QA (multi-words)	17.49	23.28
Neural Blind QA (single-word)	17.15	22.80
Neural Blind QA (multi-words)	17.06	22.30
Human QA	50.20	50.82
Human QA; Blind	7.34	13.17

Agreement

Level: Neural Image single-word	Accuracy	WUPS 0.9
No agreement	9.13	13.06
>= 50% agreement	24.10	30.94
Full agreement	29.62	37.71

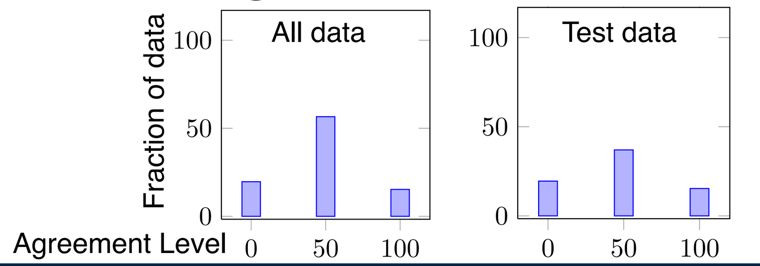
Min Consensus

Method	Accuracy	WUPS 0.9
Neural Blind QA (single-word)	22.56	30.93
Neural Image QA (single-word)	26.53	34.87

Average Consensus

Method	Accuracy	WUPS 0.9
Neural Blind QA (single-word)	11.57	18.97
Neural Image QA (single-word)	13.51	21.36

Human Agreement



Acknowledgements. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD).