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

Mateusz Malinowski, Marcus Rohrbach, Mario Fritz

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

Summary

Motivation: Defining a task that benchmarks visual comprehension - Easy for humans, challenging for machines - Easy to automatically evaluate - Approach 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

- Novel neural-based architecture with results on language-only model
- Outperforms the previous similar method
- Global image representation (CHI)
- Capable of multilingual answers generation
- Consensus metrics to measure performance

References

Dataset DAQUAR [1] - Indoor images - 1449 RGBD images - 12.5k image-Question-Answer triplets
- Questions about objects, sets of colors, numbers, and sizes of the objects
- Subjectivity is dominant in the dataset
- Spatial relations exhibit different reference frames
- Some objects are referred by multiple names: right, left, opposite
- Subjective objects saliency

- Indoor images
- 1449 RGBD images
- 12.5k image-Question-Answer triplets
- Questions about objects, sets of colors, numbers, and sizes of the objects
- Subjectivity is dominant in the dataset
- Spatial relations exhibit different reference frames
- Some objects are referred by multiple names: right, left, opposite
- 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

Language-Only (Neural Blind)

- Only trained on question-answer pairs, without seeing images
- Competitive performance

- Some answers can be decided solely based on questions (e.g., chairs often surround 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 image modeling
- Uses CNN for image modeling
- Global visual representation
- Best performance: around 19.4 Acc and 25.9 WUPS@0.9

LSTM and CNN

- Multiple-words Answer Generation

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

Performance Metrics

- WUPS: Limitations of Accuracy
- Accuracy: CNN + LSTM
- Wu-Palmer similarity
- Taxonomy-based measure
- Vocabulary between 2 and 1

Consensus

- Limitations of WUPS

- Doesn’t account for many question and scene interpretations
- Inaccurate in assigning weights

Quantitative Results

Standard Metrics

Method | Accuracy | WUPS 0.9
--- | --- | ---
Neural QA (single-word) | 19.43 | 25.28
Neural Image QA (multi-words) | 17.74 | 22.28
Neural Blind QA (single-word) | 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
- 50% agreement: 24.15
- Full agreement: 29.62

Min Consensus

- Method | Accuracy | WUPS 0.9
--- | --- | ---
Neural QA Blind (single-word) | 22.56 | 30.93
Neural Image QA (single-word) | 28.53 | 34.87

Average Consensus

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

Acknowledgements: Marcus Rohrbach was supported by a scholarship within the Frontiers Program of the German Academic Exchange Service (DAAD).

Figure: Visualization of the Answered Question Answering task in the DAQUAR Image Dataset.