Towards a Visual Turing Challenge

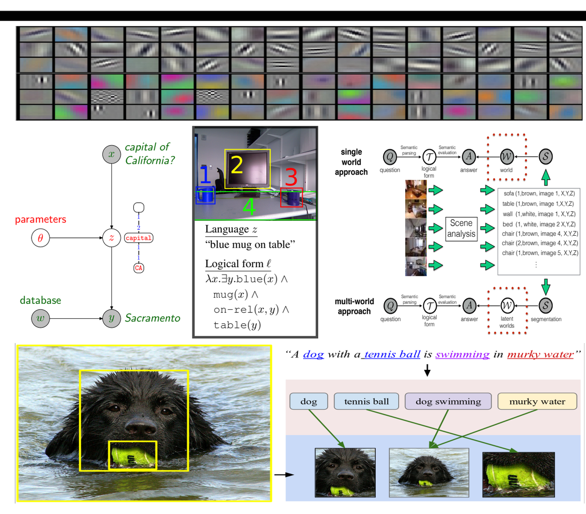
Mateusz Malinowski and Mario Fritz
{mmalinow, mfritz}@ mpi-inf.mpg.de
max planck institut informatik

Motivation

- Stronger vision and language techniques are being developed
- Can machines answer on natural questions about real-world?
  - A holistic and open-ended test that resembles the famous Turing Test
  - Understanding human intentions in the human-machine communication
  - Less subjective than Turing Test in the interpretation of the answers
  - Cheaper annotations as logical forms are not required
- Benchmarking holistic tasks that test chain of perception, representation and deduction
- Maintain tractable annotation effort
- Shape a benchmark that applies to many approaches: Don't impose strong constraints on the methods

Related work

- Machine perception
- Machine language understanding
- Grounding
- Image-to-sentence alignment
- Question-answering problem



Overview

- Introduce a holistic Visual Turing Challenge
- Discuss associated challenges in Vision and NLP
- Introduce and discuss performance measures
  - Social consensus to benchmark different architectures

Challenges

- Vision and language
  - Joint treatment of both modalities
    - 'Which hand of the teacher is on her chin?'
    - Ideally closing the loop for improved perception
  - Richness of the concepts
    - Object categories
    - Attributes (e.g. genders, colors, states)
    - Unknown human notion of spatial relations
  - Ambiguities in the reference frame
    - Object-centric
    - Observer-centric
    - World-centric
  - Contextualization of the concepts
    - White in 'white elephant' and 'white snow'
- Common sense knowledge
  - Narrows down likely options or locations
    - 'Which object on the table is used for cutting?'
    - 'What is in front of scissors?'
- Defining a benchmark
  - End-to-end system that learns from textual question-answer pairs
  - Internal representation of architectures is irrelevant
  - Easy to collect a dataset
  - Hard to define automatic performance measures

Challenges in DAQUAR

- Unconstraint questions and defined but large answer space
- Vision and language
  - Many categories with fuzzy semantic boundaries
    - Nouns such as tool, night stand, cabinet may refer to the same thing
  - Human notion of spatial concepts
  - Different reference frames
  - Questions of substantial length (10.5 words in average)
  - Possible language errors
- Common sense knowledge
  - Strong non-visual cues for predicting an object
    - 'Which object on the table is used for cutting?'
- Pragmatism of the question answering task
  - Understanding hidden intentions of the questioner
  - Grounding of the meaning as a latent sub-goal

Metrics

- Automatic Evaluation by Design
- Ambiguity
  - Cultural bias
  - Fined grained categorization
  - Reference frame
- 'Soft' Accuracy

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

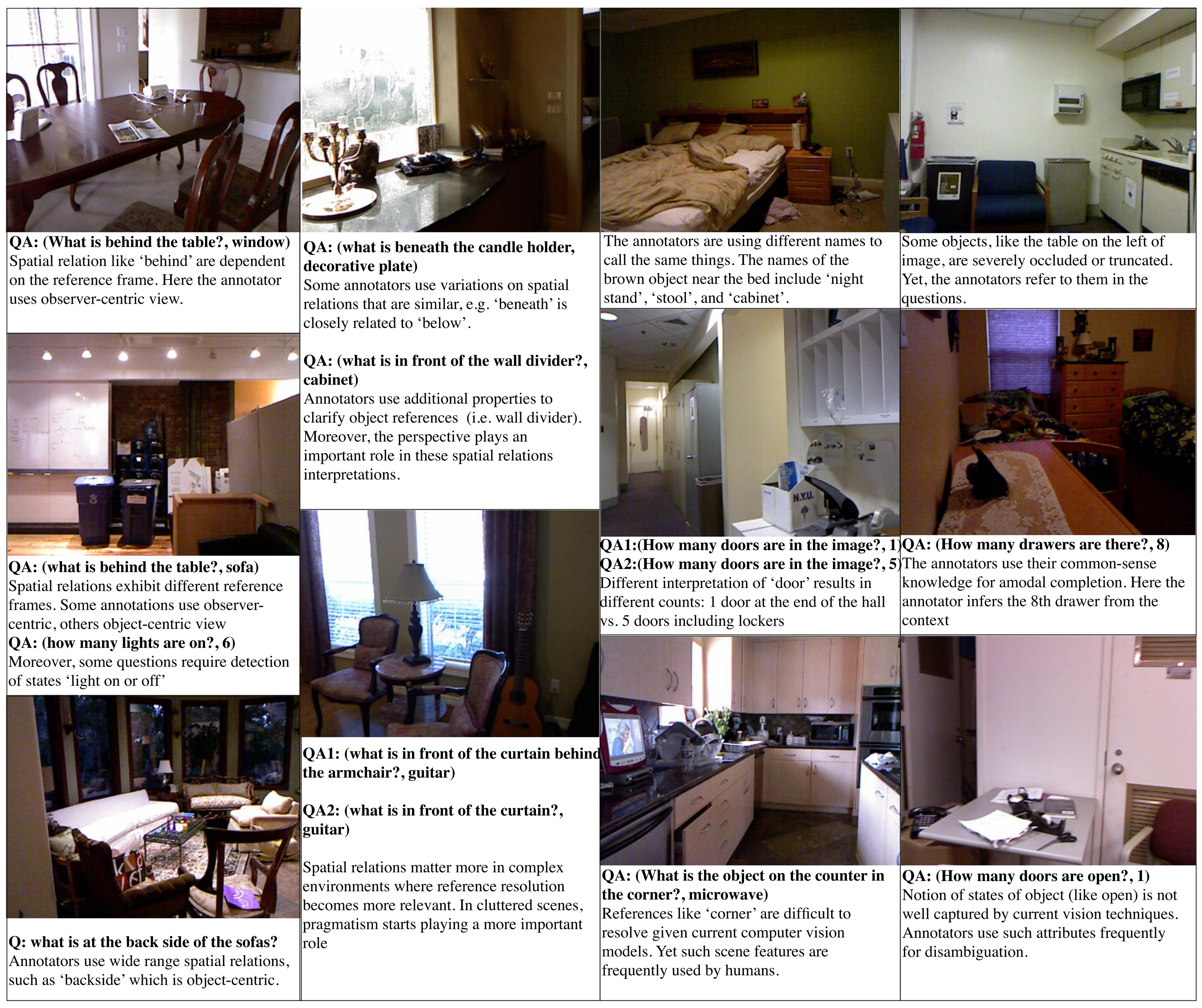
- Lacks of the coverage in the lexical databases
- Further development of the metrics
  - Consider many valid human answers
  - Interpretation metric
    - Maximal score over different human answers
  - Consensus metric
    - Average over different human answers
    - Takes an agreement between human responses into account
- Experimental scenarios
  - Controlled and open scenarios with another resources available in training

Conclusions

- Visual Turing Challenge provides a rich set of challenges in Vision and NLP - yet annotation and evaluation remain tractable
- Automatic benchmarking, but coverage can be an issue
- Cultural bias, changes in the reference frame, naming ambiguities, and unknown spatial relation are inherent to the challenge

DAQUAR

- NYU-Depth V2 dataset with textual question-answer pairs
- 1449 RGBD indoor images
- 12,5k question-answer pairs
- Annotations are: colors, numbers, objects
- Subjectivity is prominent in the dataset [1]
- About 9 question-answer pairs per image
- Object's category occurs 4 times in training set



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 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'

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

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

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

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.

[1] M. Malinowski and M. Fritz "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input" NIPS 2014