A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input

Motivation
- Can machines answer questions about images?
- Evaluating chain of perception, representation, deduction

Contributions
- An approach and a dataset for question-answering about real-world scenes that is directly trained from question-answer pairs
- Combine language with perception in a multi-world Bayesian framework: connecting discrete reasoning and uncertain representations form perception
- Establish first results on complex question-answering task on sizable dataset of real-world scenes

Overview
- Bayesian approach to uncertainty in language and images
- Two approaches: single and multi-world
- Each world is a set of facts derived from scene segmentation

DAQUAR - Towards a Visual Turing Challenge
- NYU-Depth V2 dataset with textual question-answer pairs
- Human question-answer pairs (HumanQA)
- Single with Neg. 3

Method
- Question-answering task [2]
- Dataset of geographical facts: "What is the total population of the ten largest capitals in the US?"
- Trained from textual question-answer pairs
- Bayesian integration over latent logical representations
- Answers are retrieved from the dataset of the facts
- Log-linear model $P(A|Q,S) > e^xP(Q)$
- Features are counts of templates:
  - A thing is under a scene predicate
  - Two predicates are linked via a relations
  - A predicate has a b/s

Results
- WUPS Scores
- With more categories the semantic boundaries are becoming more fuzzy
- Propose a family of soft measures [3] and introduce WUPS scores

Questions
- How many doors are open?
- What is behind the table?
- What is the largest object in image 1?

References