Deep Learning for Advanced Robot Perception

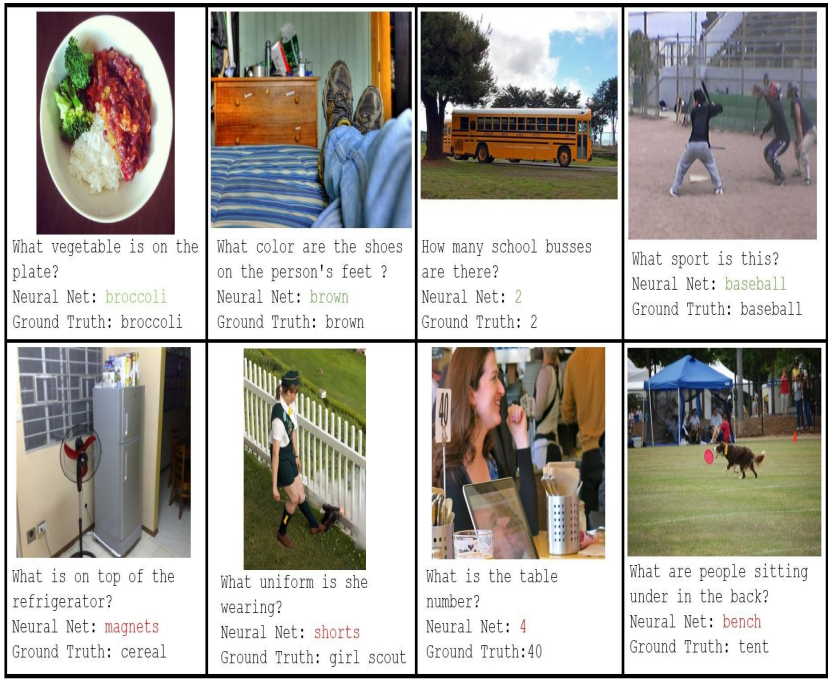
RBE 595

Fall 2016

Visual Question Answering(VQA)

Shanmuga perumal

What is VQA?

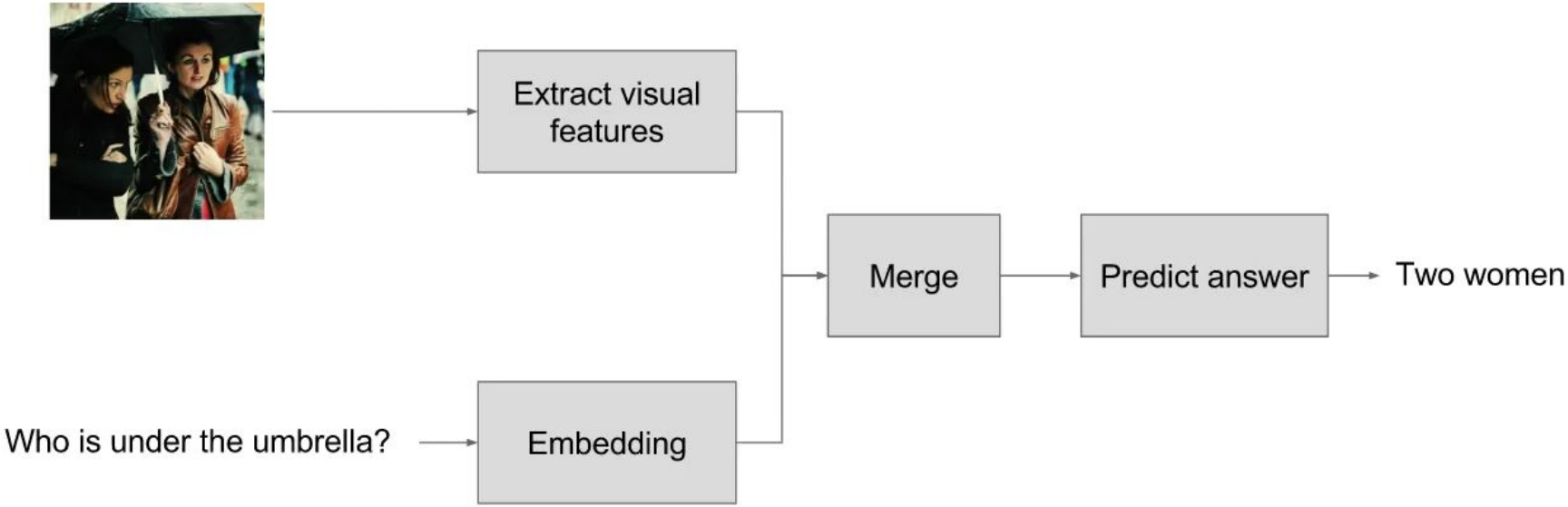


- Answer questions asked by the user based on an input image
- Requires the machine to understand the image and also the context in which the question was asked.
- Scene,object, common sense recognition

Motivation:Why VQA?

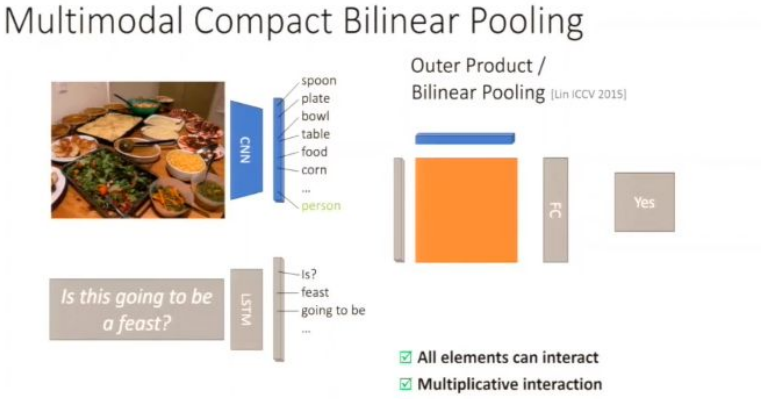
- Help blind people become aware of their surroundings
- Making robots better aware of the environment to make better decisions (domestic helper robots)
- Can help us build better machines that can pass the turing test

Common approach



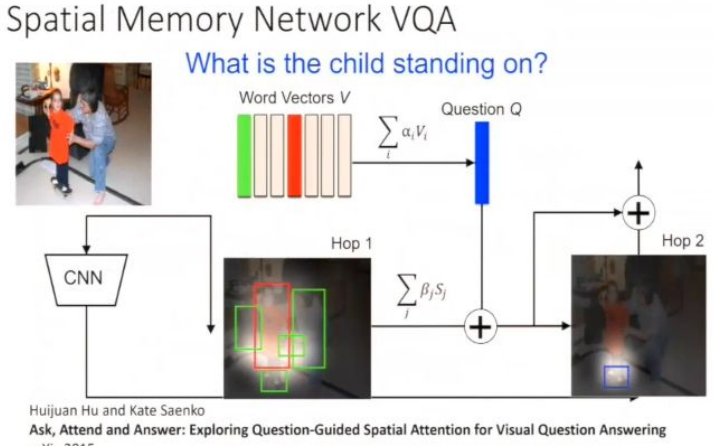
How to combine visual and linguistic vectors?

- Concatenate-no full interaction between words and image parts
- Element wise multiplication-needs perfect alignment between words and image vectors
- Outer product

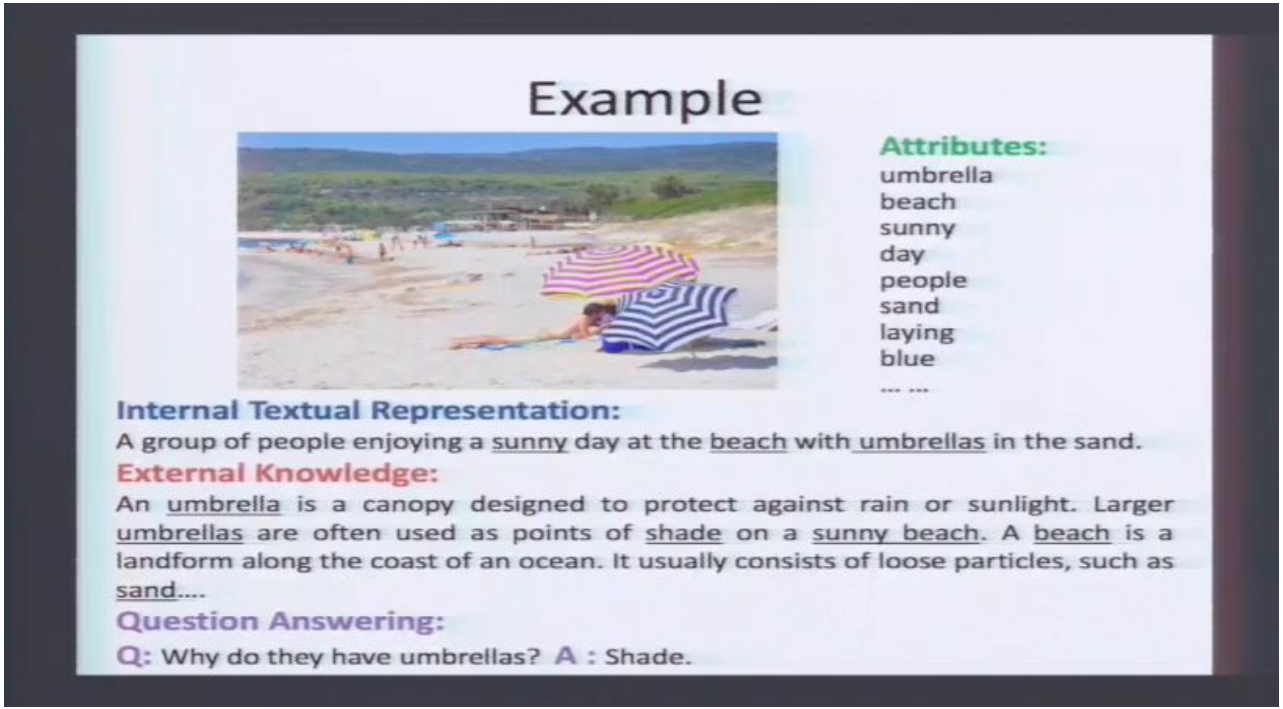


Grounding

- Questions mostly related to specific parts of the image
- Exploit relation between image to noun
- Attention based models- creates masks to look for the area of image that needs to be looked at to answer the question

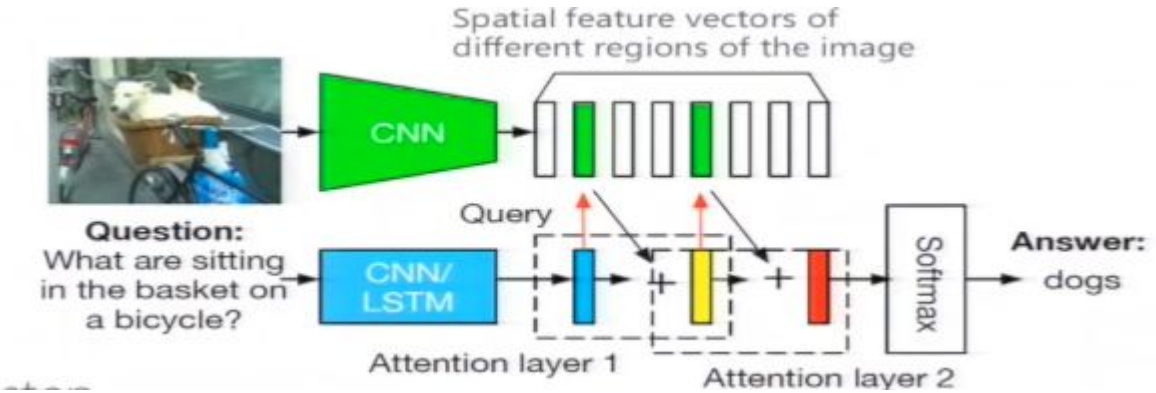


Free form visual augmentation

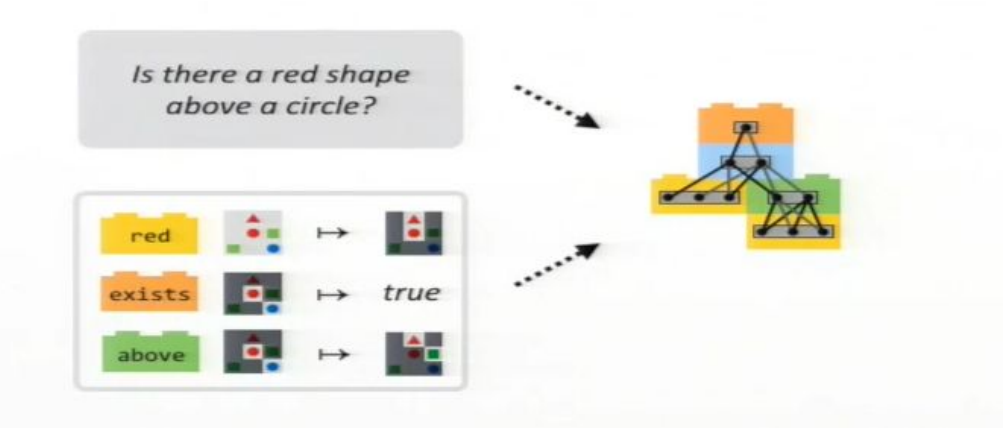


Stacked attention network

- Four parts
  - Question model
  - Image model
  - Multilayer attention model
  - Answer model



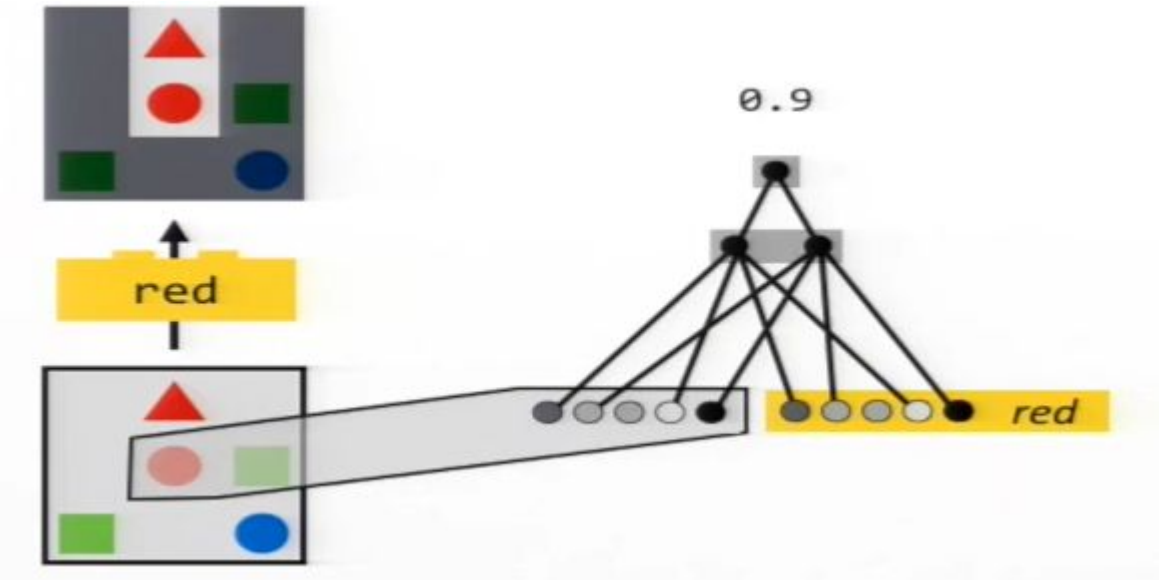
Module networks



- Building the networks on the fly
- Connect modules together based on the question
- Question specific networks
- NLP parser parses through the question
- This then passes through the network

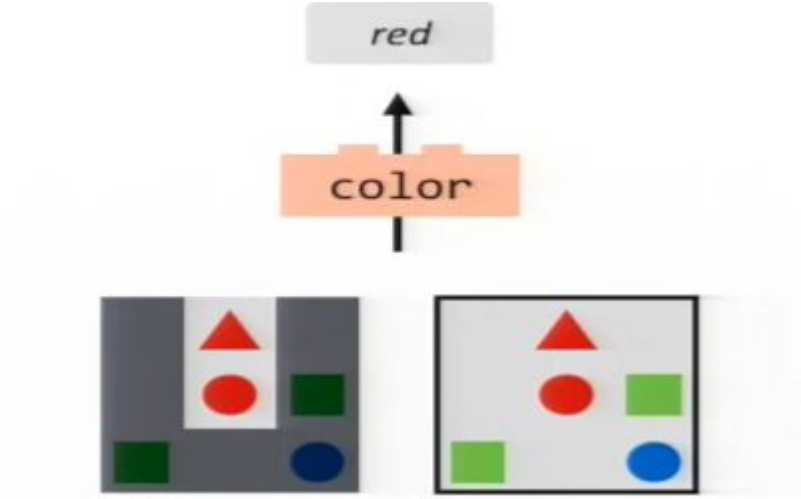
Find module

- Finds an instance of something



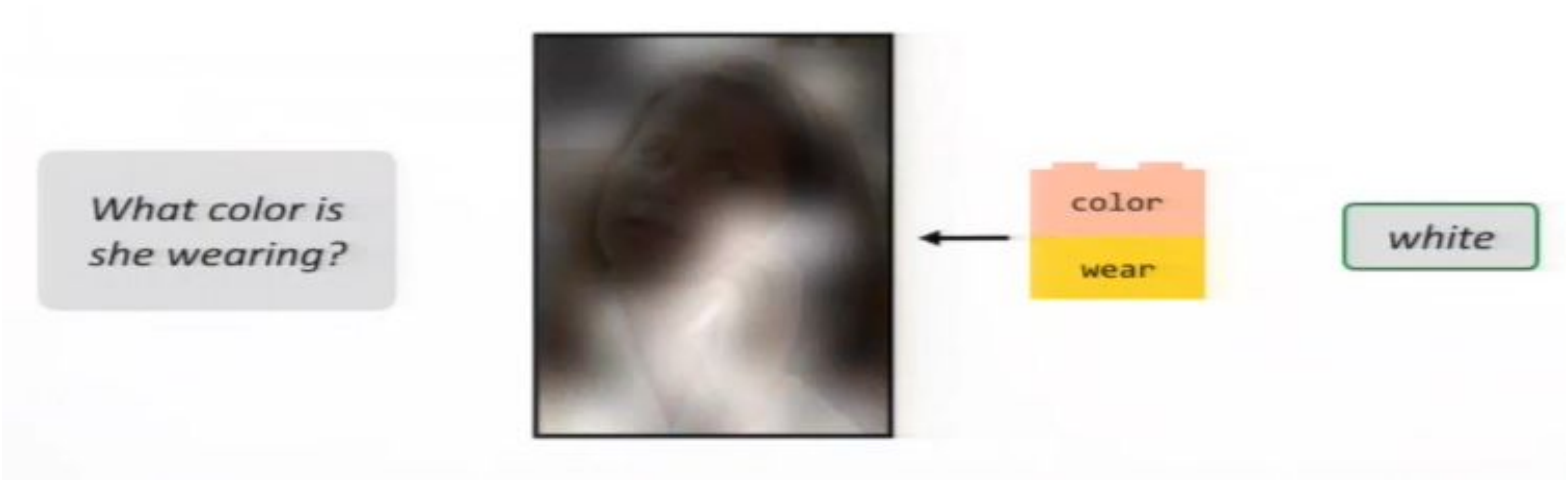
Describe module

- Describes a particular area



Context specific

- The modules are context specific



Visual explanation

- The next step in VQA
- Requires the computer to answer why it gave the answers that it gave
- Instills confidence in machines