









Deep Learning for Advanced Robot Perception

RBE 595

Fall 2016

Visual Question Answering(VQA)
Shanmuga perumal

What is VQA?

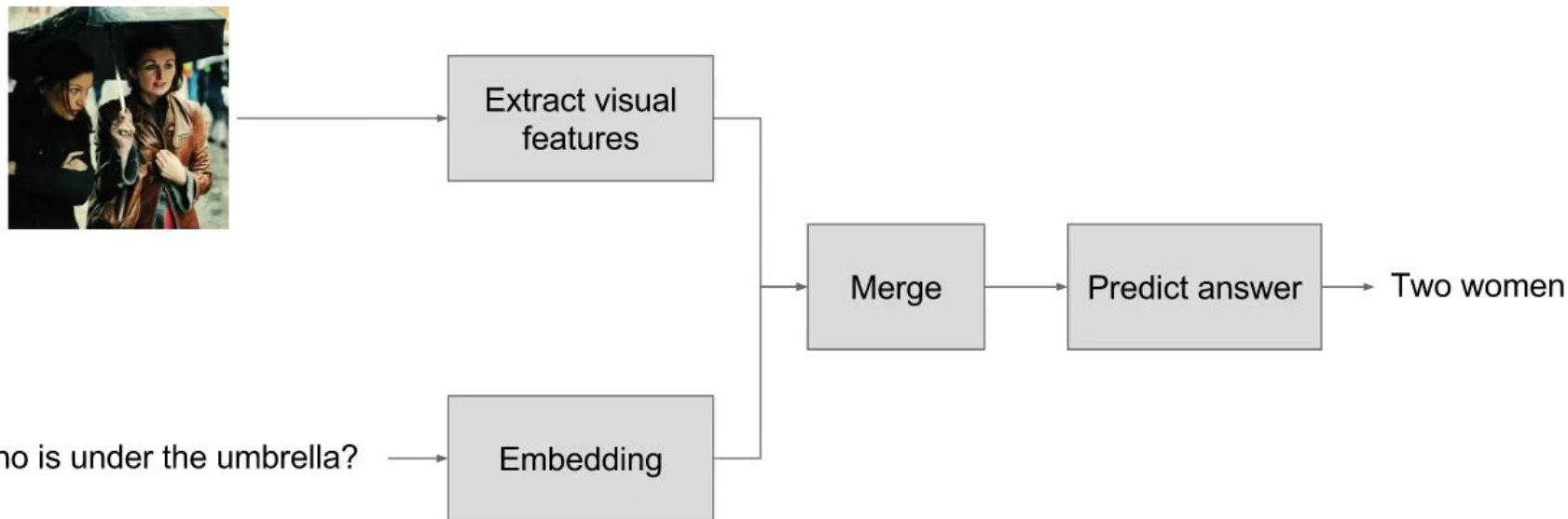
 <p>What vegetable is on the plate?</p> <p>Neural Net: broccoli</p> <p>Ground Truth: broccoli</p>	 <p>What color are the shoes on the person's feet ?</p> <p>Neural Net: brown</p> <p>Ground Truth: brown</p>	 <p>How many school busses are there?</p> <p>Neural Net: 2</p> <p>Ground Truth: 2</p>	 <p>What sport is this?</p> <p>Neural Net: baseball</p> <p>Ground Truth: baseball</p>
 <p>What is on top of the refrigerator?</p> <p>Neural Net: magnets</p> <p>Ground Truth: cereal</p>	 <p>What uniform is she wearing?</p> <p>Neural Net: shorts</p> <p>Ground Truth: girl scout</p>	 <p>What is the table number?</p> <p>Neural Net: 4</p> <p>Ground Truth: 40</p>	 <p>What are people sitting under in the back?</p> <p>Neural Net: bench</p> <p>Ground Truth: tent</p>

- 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

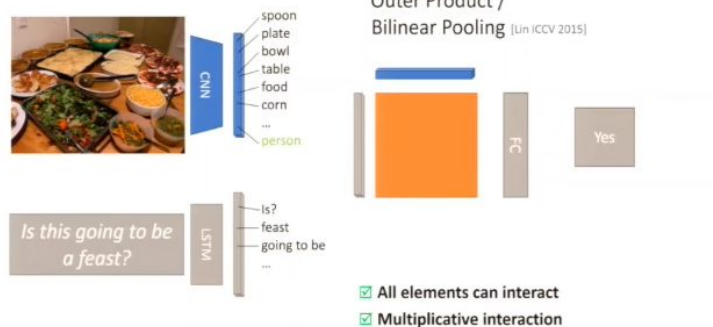
Common approach



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

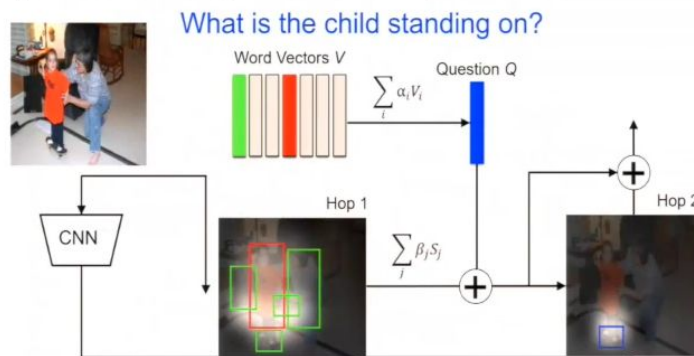
Multimodal Compact Bilinear Pooling



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

Spatial Memory Network VQA



Huijuan Hu and Kate Saenko

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Free form visual augmentation

Example



Attributes:

umbrella
beach
sunny
day
people
sand
laying
blue
... ..

Internal Textual Representation:

A group of people enjoying a sunny day at the beach with umbrellas in the sand.

External Knowledge:

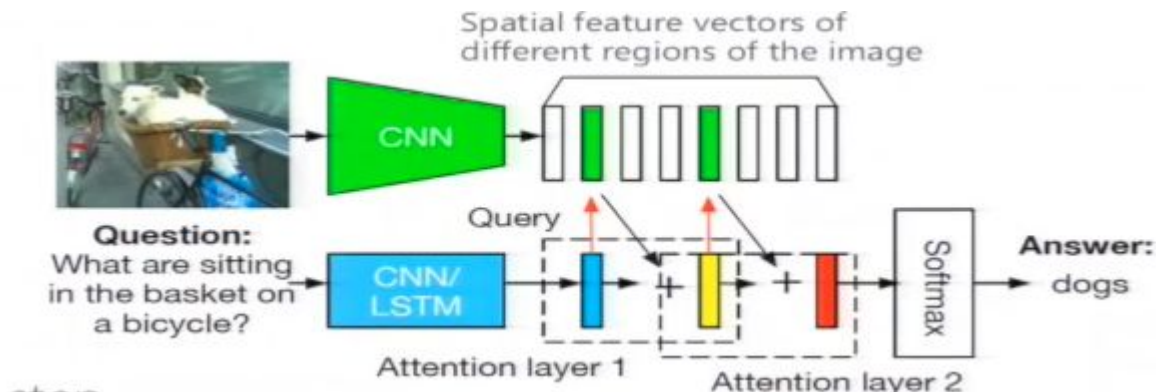
An umbrella is a canopy designed to protect against rain or sunlight. Larger umbrellas are often used as points of shade on a sunny beach. A beach is a landform along the coast of an ocean. It usually consists of loose particles, such as sand....

Question Answering:

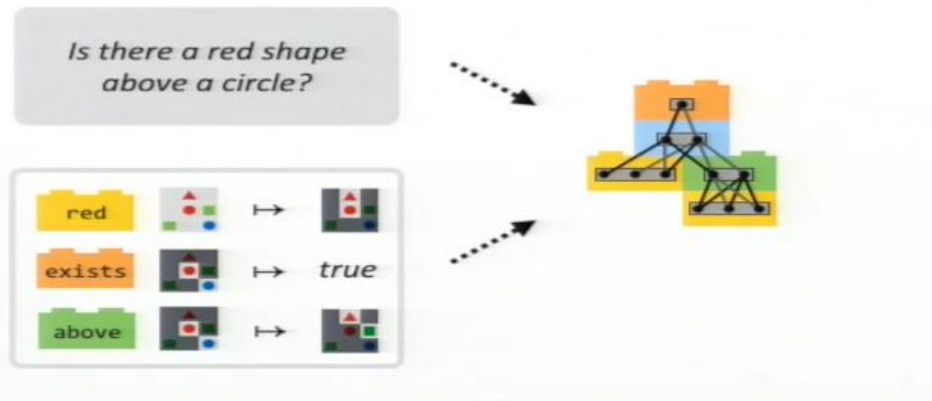
Q: Why do they have umbrellas? A : Shade.

Stacked attention network

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



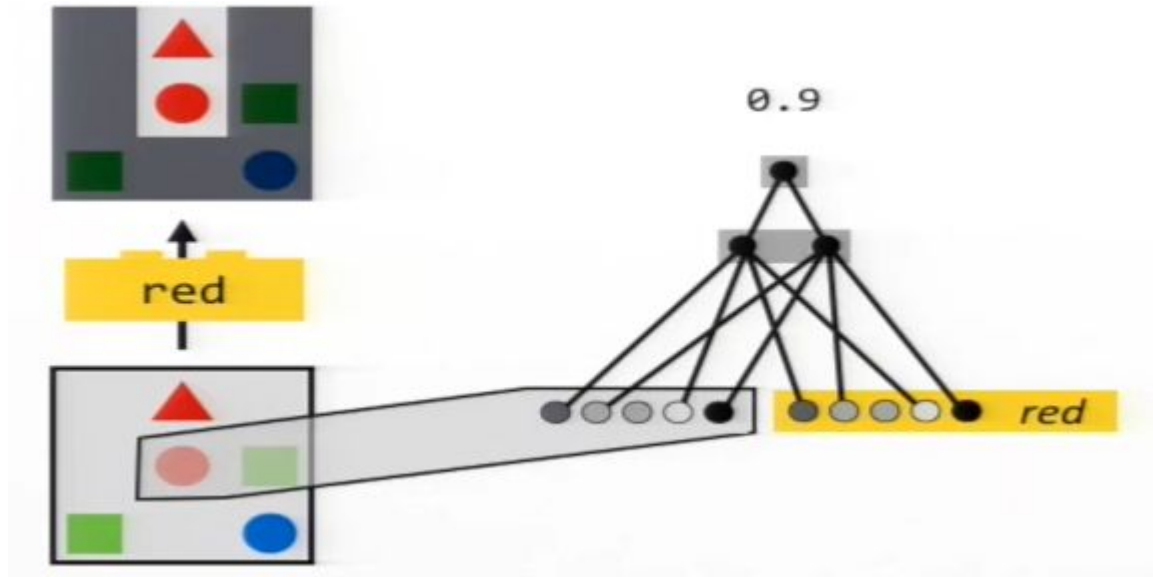
Module networks



- 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

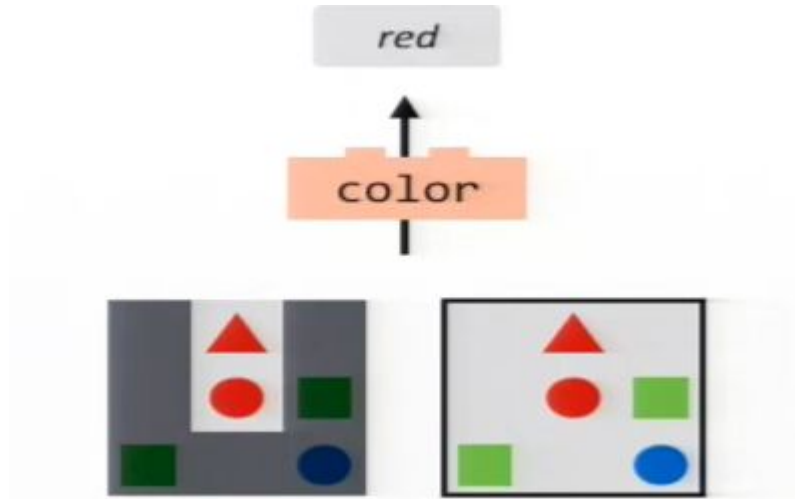
Find module

- Finds an instance of something



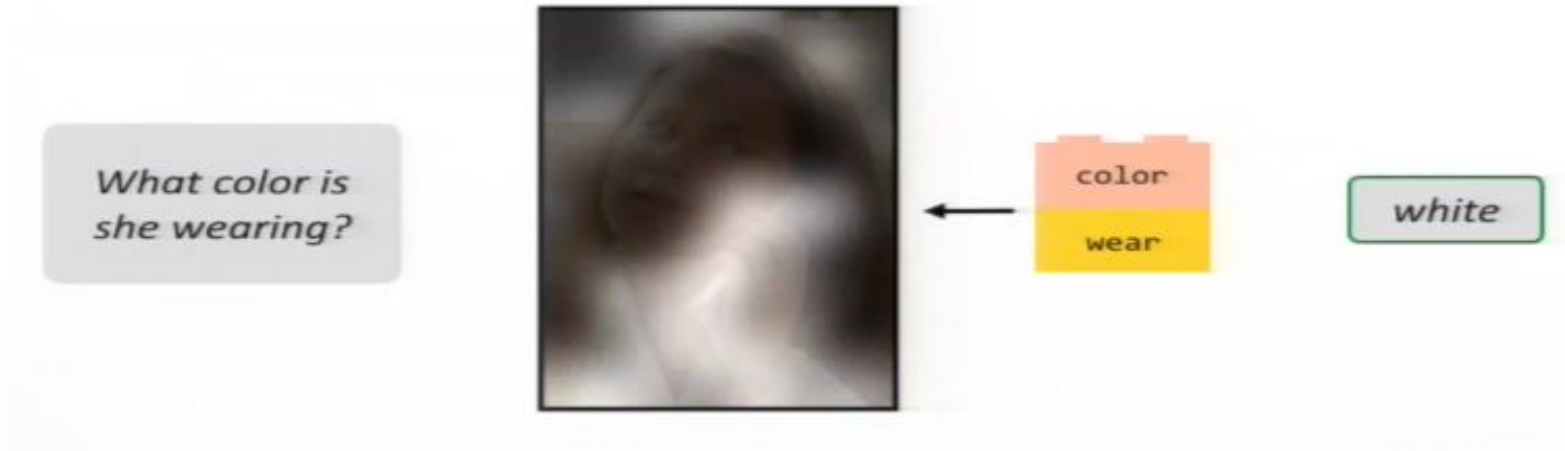
Describe module

- 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