Stacked Attention for Visual QA

Bingbin Liu, Weini Yu

Stanford Computer Science

Introduction

Lying at the intersection of NLP and computer vision, **Visual Question Answering (VQA)** refers to the task where given an image, a natural language question and a list of candidate answers, the model should predict the best matching answer.

In this project we revisit a simple **BOW** VQA baseline, and extend it with **LSTM** and **spatial attention**. Our model achieves a testing accuracy of **61.8 on Visual7W** dataset and offers interesting visualization results.

Data

Visual7W Telling dataset:

- 327939 QA pairs on 47300 COCO images
- **Multiple-choice**: 1 question about an image, 4 candidate answers, out of which 1 is correct
- **6 Question types**: what, where, when, who, why, how

Where are the trees?
A) Behind the zebras.
B) In front of the elephants.
C) Around the giraffes.
D) Next to the lions.

Who is the manufacturer of the laptop?
A) Apple.
B) Dell.
C) Hp.
D) Samsung.

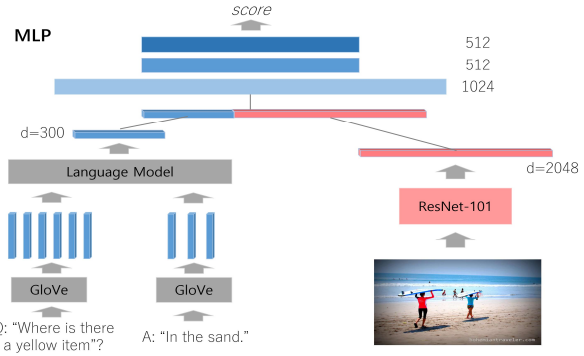
Preprocess:

- **Textual**: 300-dim GloVe word vectors pretrained on 6 billion tokens from Wikipedia 2014.
- **Visual-vector**: **2048**-dim vector by average pooling on layer 4 of ResNet-101 pretrained on ImageNet
- **Visual-spatial**: **2048 x 7 x 7** vector from layer 4 of ResNet-101 pretrained on ImageNet.

Model and Architecture

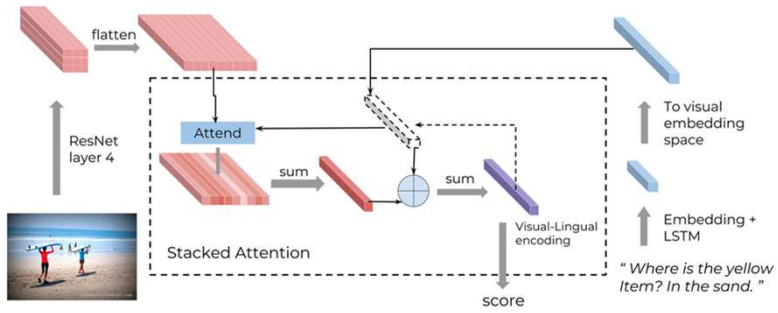
BOW / LSTM

Concatenation of vector representation of images + language encoding (using Bag-of-Words or LSTM) passing through a MLP.



LSTM with Stacked Spatial Attention

Use language encoding to attend the image. The attention process happens in multiple passes, so that the attention can be refined hierarchically.



Results and Analysis

Table 1: Accuracy with different question types (Q+A+I, BCE loss)

Model	What	Where	When	Who	Why	How	Overall
LSTM-Att	0.515	0.570	0.750	0.595	0.555	0.498	0.556
BOW	**0.585**	0.694	0.797	0.673	0.566	0.511	0.609
LSTM	0.584	**0.719**	**0.803**	**0.684**	**0.593**	**0.526**	**0.618**
LSTM-Att	0.560	0.691	0.800	0.660	0.590	0.512	0.597

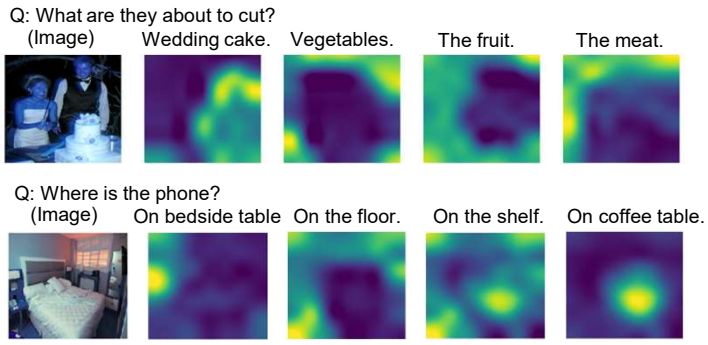
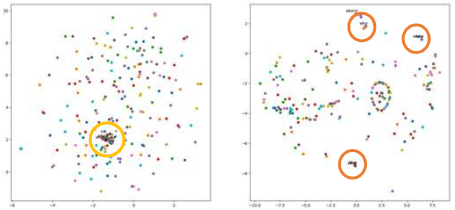


Table 4: Accuracy with different features (LSTM, BCE loss)

Feature	What	Where	When	Who	Why	How	Overall
A	0.474	0.578	0.767	0.649	0.556	0.489	0.530
Q+A	0.528	0.587	0.782	0.648	0.564	**0.531**	0.564
Q+A+I	**0.584**	**0.719**	**0.803**	**0.684**	**0.593**	0.526	**0.618**

Fine-tune over GloVe: the 6 question words shifted drastically.



Future Work

Finer spatial attention & tuning
Text attention.