Image Question Answering Using Deep Learning

Shantanu Kumar | 2013EE10798
Barun Patra | 2013CS10773

Problem Definition

The task is of answering a natural language question in the context of an image.

What is sitting on the handlebar ?

BIRD

What are sitting in the basket on a bicycle ?

DOGS

How many people are going up the mountain ?

FOUR

What is the animal doing ?

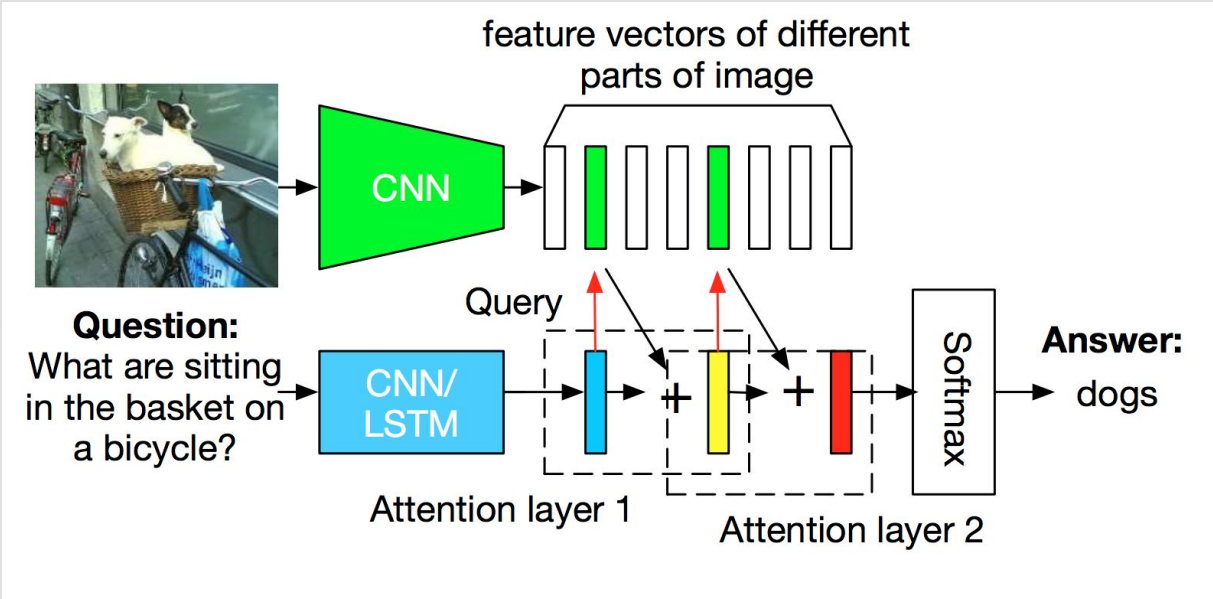
RELAXING

Visual-QA Dataset

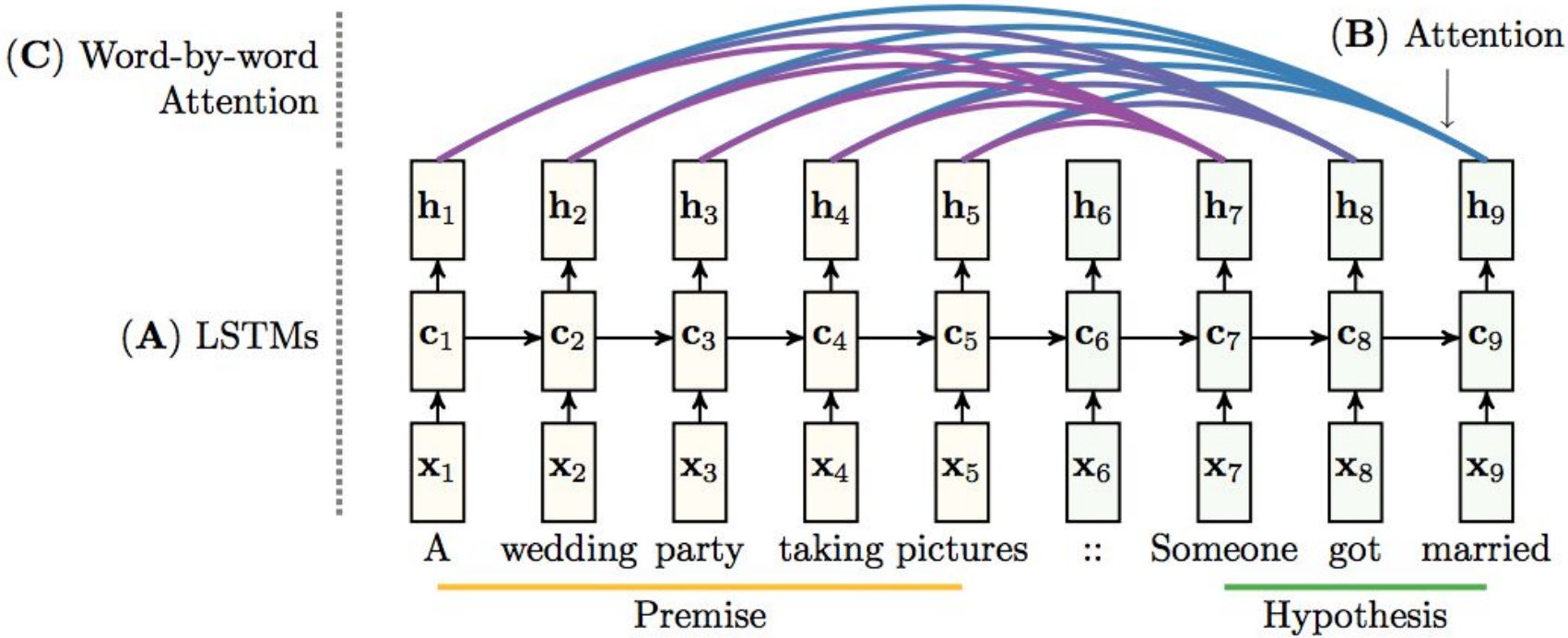
- Image-Question pair generated automatically using image captions
- 82,783 training images
- 3 Questions per image
- 248, 349 train questions and 121,512 validation questions
- One-word Answers
- Multiple choice questions and Open Ended questions

Baseline Model

- **Stacked Attention Networks for Image Question Answering** (Yang et al., 2016)
- CNN model for Image and LSTM model for sentence
- Implemented Stacked attention model using **Theano & Derivatives**.



Our Approach

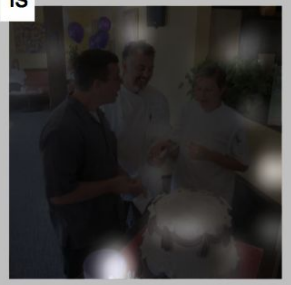


Rocktäschel, Tim, et al. "Reasoning about entailment with neural attention." *arXiv preprint arXiv:1509.06664* (2015).

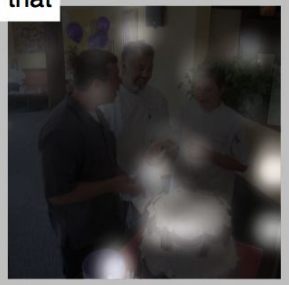
Results

	Stacked Attention Network	Word by Word Attention Network
Number of Parameters	30,096	19,758
Raw Accuracy (On Validation)	52.4 %	52.1 %
Annotator Agreed Accuracy (On Validation)	54.7 %	54.4 %

is



that



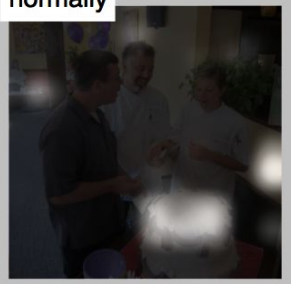
a

type

of

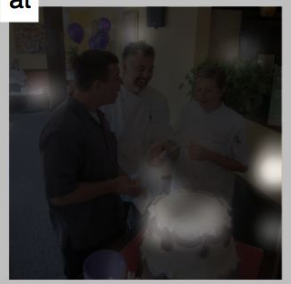
food

normally



seen

at



weddings

Question :
Is that a type of food normally seen at weddings?

Answer : Yes

What

fruit

is

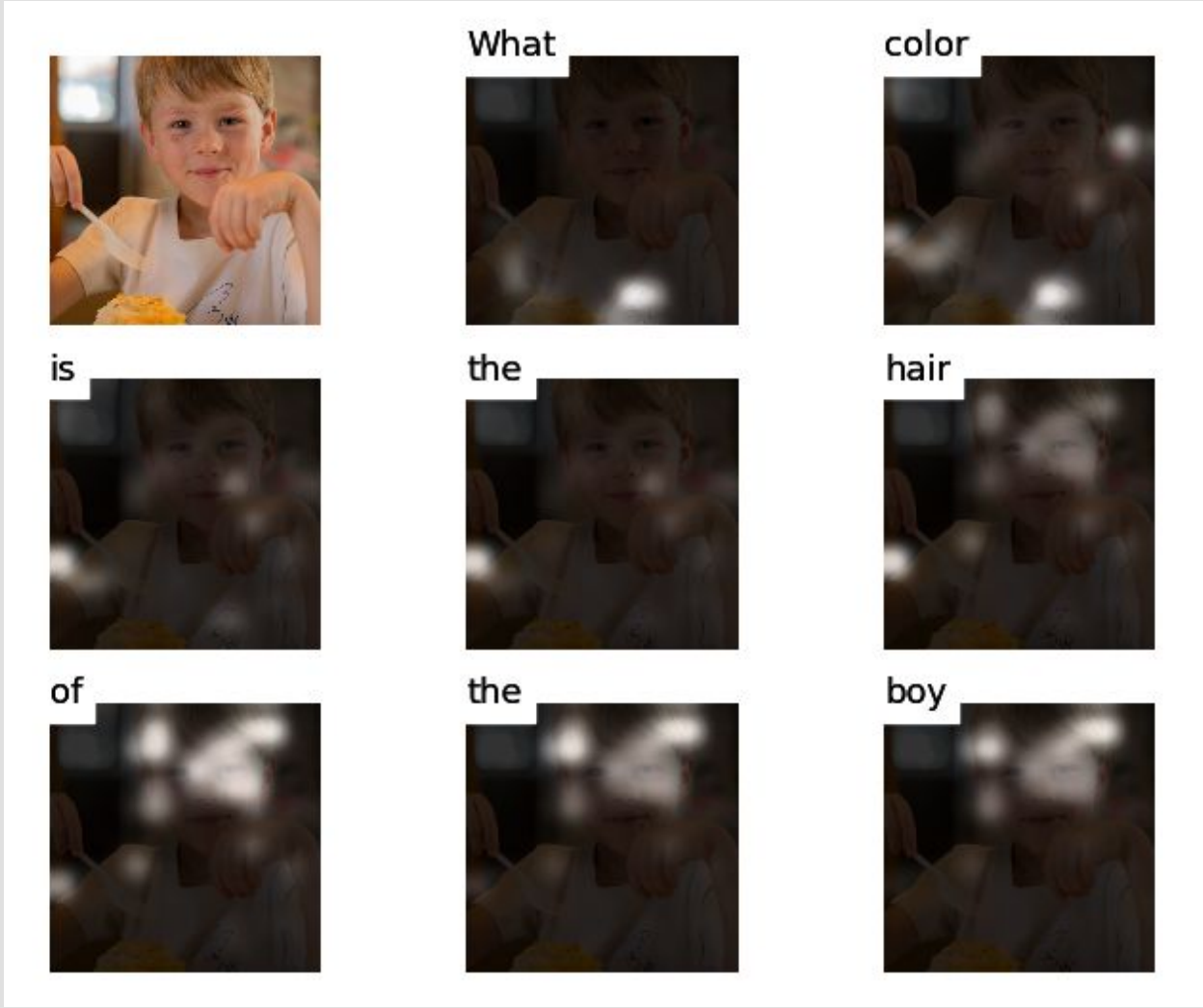
on

the

table

Question :
What fruit is on the table?

Answer : Apple



Question : What color is the hair of the boy?

Answer : Blonde

Demo