# Visual Question Answering

DIP Project

Sahaj Agarwal (S20160010077)
Anubhav Ujjawal (S20160010005)
Anurag Gupta (S20160010006)
Garvit Kataria (S20160010028)
Laisha Wadhawa (S20160010048)

# What is VQA?

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

You get the idea...

# Solution to VQA

A key solution to visual question answering (VQA) exists in how visual and language features extracted from an input image and question are fused together to generate context relevant answers.

# Previous Work

- Using visual features from some region proposals, which are generated by Edge Boxes or Region Proposal Network.
- Methods that employ question-guided attention on image regions.
- Stacked attention network that produces multiple attention maps on the image in a sequential manner.
- Structured attention model encoding cross-region relation that properly answers questions that involve complex inter-region relations.
- Models proposing image-guided attention on question words.
- Co-Attention mechanism by Dhruv Batra et al. that generates cross attention between image regions and question words and considers attention at three different levels, i.e, word level, phrase level and sentence level.

# Contd..

- Combination of Co-Attention architecture with a novel multi-modal feature fusion of image and question attended feature maps to learn complex relationships between image and question.

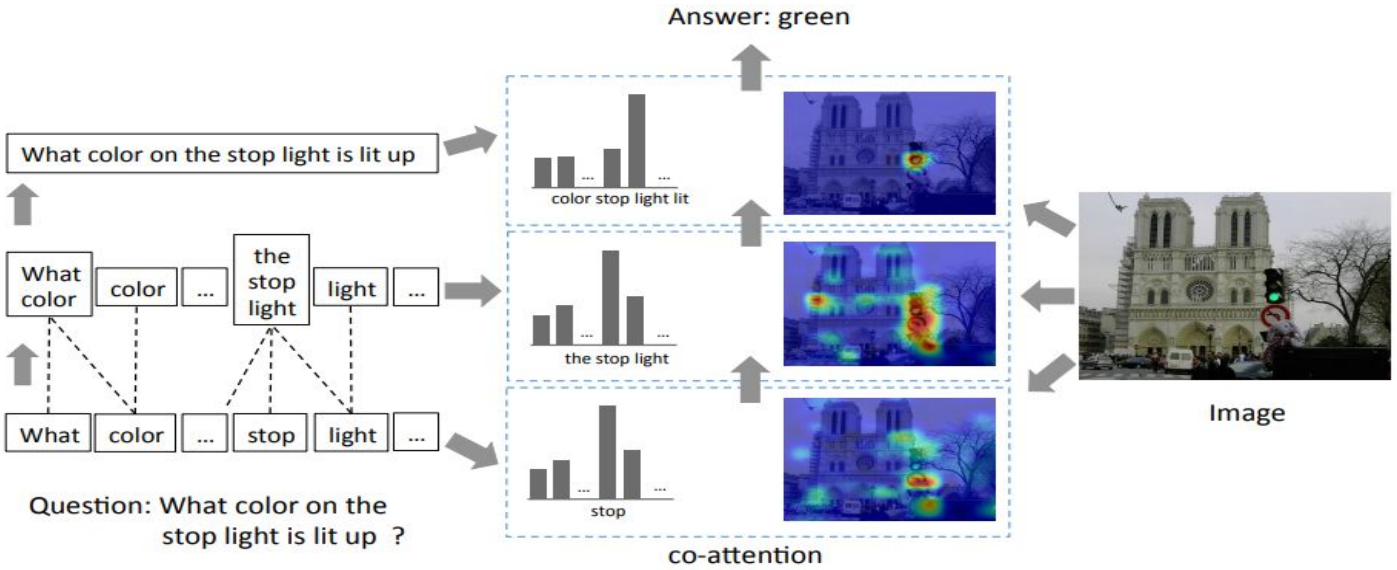



Image courtesy of paper "Hierarchical Question-Image Co-Attention for Visual Question Answering" by Dhruv Batra et al.

# Limitations of previous methods

There are essentially two key limitations of the models described earlier -

- The attention mechanism used in those models only considers a limited amount of possible interactions between image regions and question words. Thus, learning complex and unknown relationships between the two becomes difficult.
- Co-Attention additionally considers attention on question words due to image feature maps but it is created from the whole image. This does not allow the model to build accurate spatial relations with the questions.

# What is different in our approach?

- Our approach counters the problem of missing spatial relationships between images and questions by generating Dense co-attention maps per image and question pair.
- The proposed mechanism can deal with every interaction between any image region and any question word. This enables the architecture to model unknown complex relationships between attended image and question features and generalize well for new images and questions.

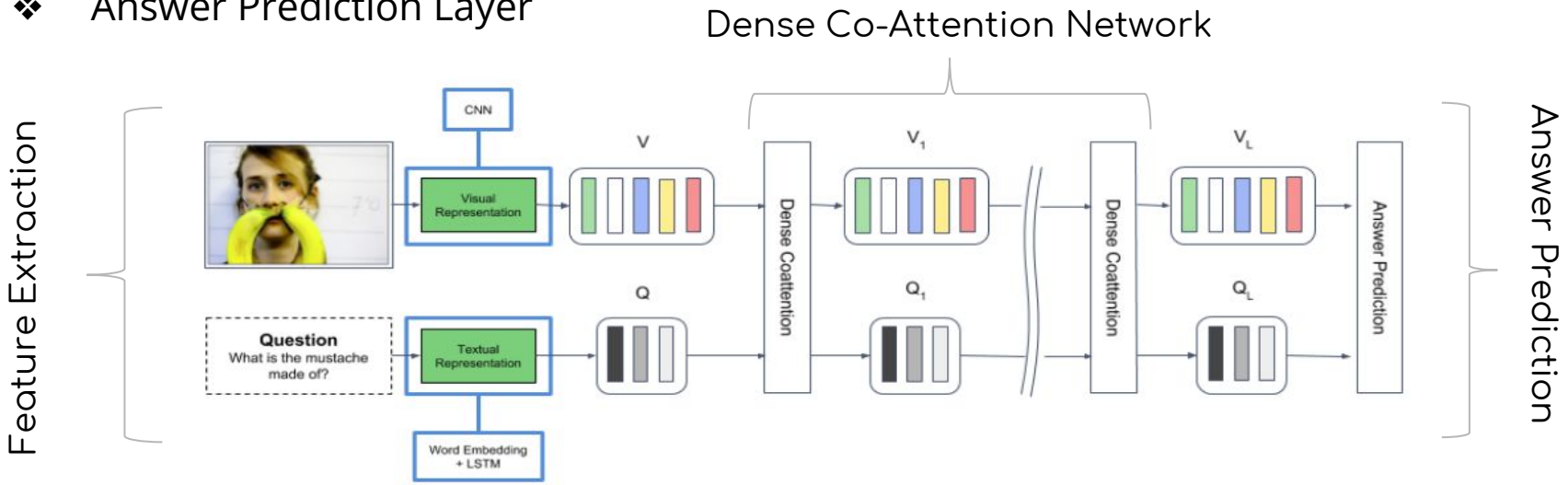
# Our Approach..

- Generate an attention map on every image region for each question word and an attention map on every question word for each image region.
- Redo the process in a hierarchical fashion for phrase level and sentence level for each question.
- With the obtained image and question feature maps perform computation of attended features, concatenation of multimodal representations, and their transformation by a fully connected network with ReLU and a residual connection.
- Finally, using a softmax nonlinearity, generate the score for each possible answer and output the answers with the top scores.

# Proposed Architecture

We describe the proposed architecture as a sequential combination of 3 different layers -

- Feature Extraction Layer
- Dense Co-Attention Network
- Answer Prediction Layer



# About Architecture

- Feature Extraction Layer
  - Bidirectional LSTM with residual embeddings for language feature extraction.
  - ResNet50 pre trained on ImageNet for Image feature extraction.
  - Multimodal Feature Fusion to map image and question features to lower-dimension fused feature representation.
- Dense Co-Attention Network
  - Parallel Attention
  - A sequence of dense co-attention layers stacked onto each other updating fused features at every level.
- Answer Prediction Layer
  - Fully Connected layer for feature  transformation and softmax scores for answer prediction.

# Key Concepts Used

Attention - *At a higher-level, an attention mechanism enables your model to focus on relevant parts of your input more than the irrelevant parts when doing a prediction task.*

## So, What is Co-Attention?

*Attention on question words due to image features*

+

*Attention on image regions due to question word features*

= Co-Attention

## Alright, Hierarchical Co-Attention though?

*Co-Attention*

+

*Parallel attention at 3 different question levels (Word, phrase and sentence)*

# Contd...

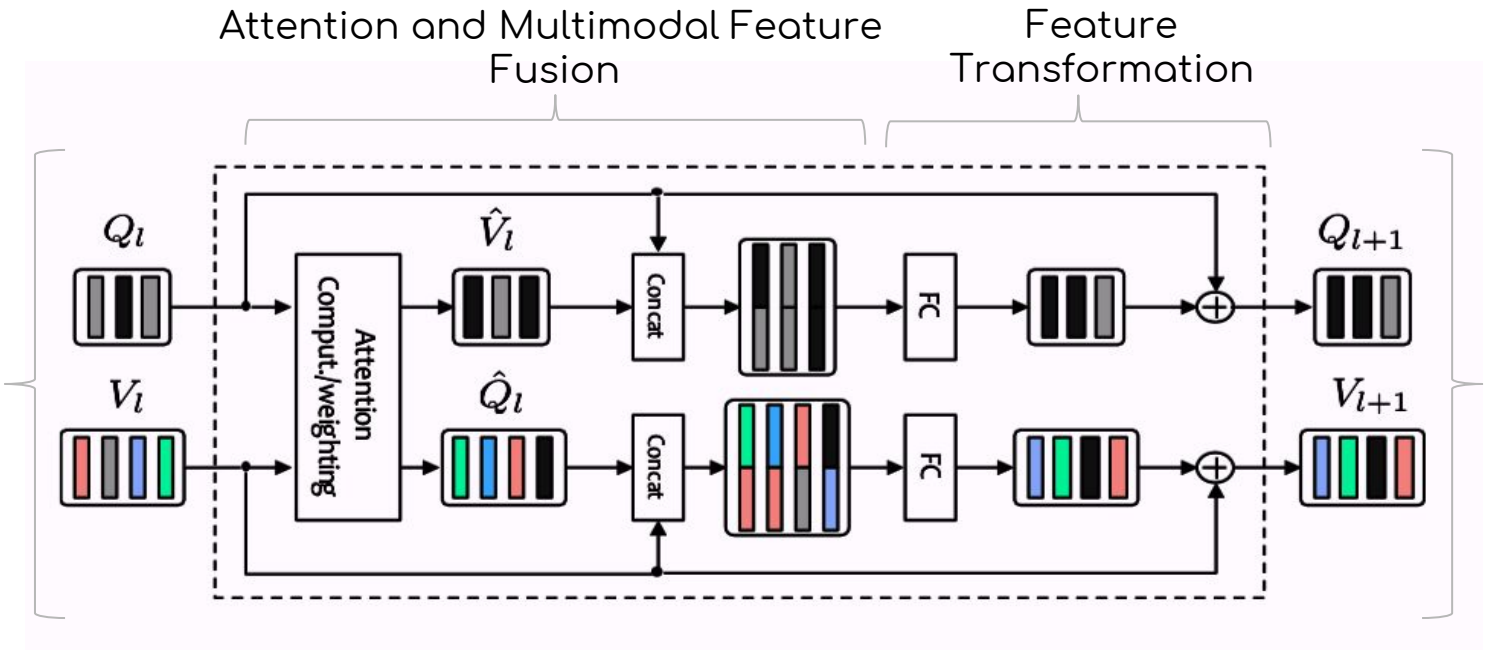
**Multimodal Feature Fusion -** *Fusion of feature maps of different modalities to obtain correlated fused maps that understand domain relationships between the two modalities is termed as Multimodal Feature Fusion.*

In our case, the two modalities are Images and Questions.

- We fuse the feature maps of these two domains by calculating an outer product of the corresponding maps followed by Multimodal Factorized Bilinear (MFB) pooling, which computes a fused feature with a matrix factorization trick to reduce the number of parameters and improve convergence rate.

# Dense Co-Attention layer

Output Features from layer **L**
Input to layer **L + 1**




Output of layer **L + 1**

These layers are stacked together in a sequence to form
Dense Co-Attention Network

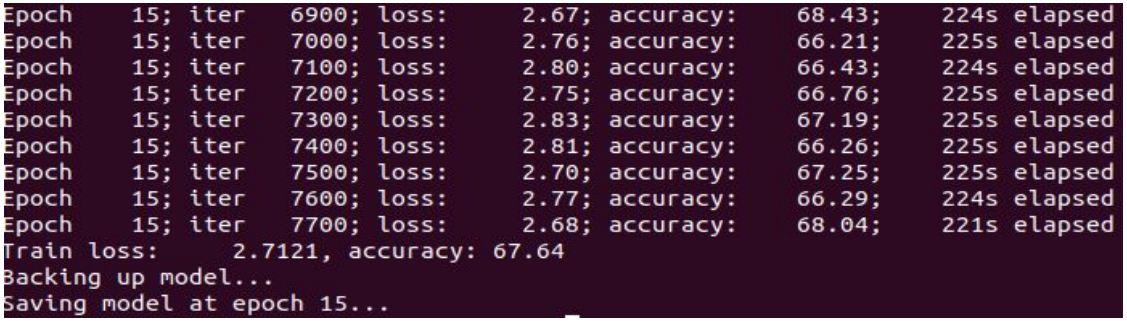
# Dataset

We are using the MSCOCO dataset available publicly at VQA official website - *http://www.visualqa.org*. Two versions of the dataset are available -

- VQA v1 (Currently used for training, validation and testing)
  - 82,783 training images, 248,349 training questions and about 2,483,490 annotations for the questions.
  - The answers are of a maximum of 6 words and include both free and MCQ type answers.
  - The test set comprises of test and test-dev dataset splits.
- VQA v2
  - Similar to VQA v1 with 443,757  training questions and about 4,437,570 training annotations.
  - Added complementary training and validation pairs.

# Our Progress so far...

- We have completed data pre-processing and loading it to feed to our model.
- We have completed implementation of model architecture and data transforms for the model.
- We have completed training and test loop implementation.
- We are in the process of training our model on the dataset.
- Currently, the model training is at epoch 15 with an average accuracy of about 67.64.



# Thank You