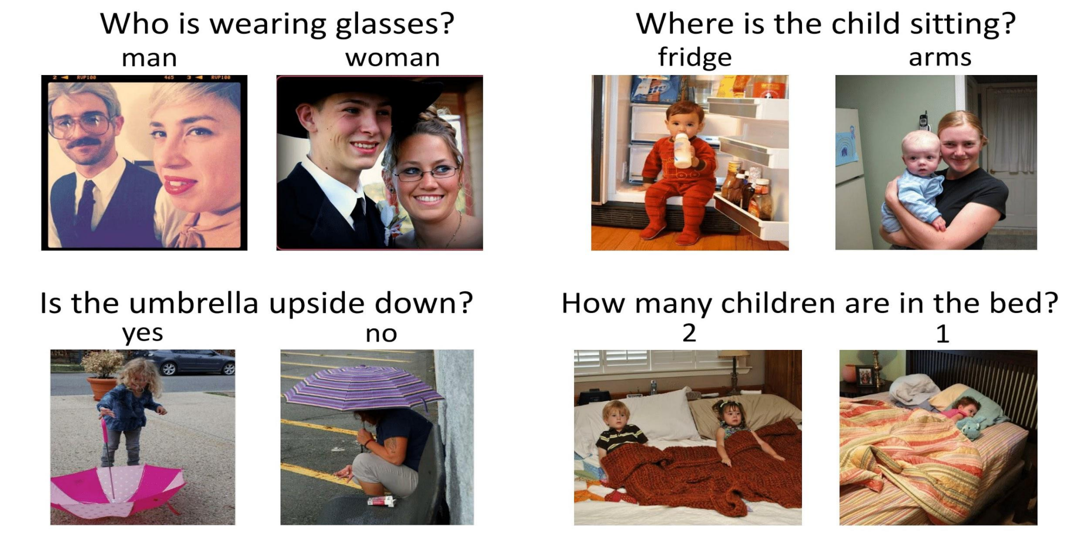
Visual Question Answering System

CLIQUE INTERESTSHIP 1.0


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

ML-T3

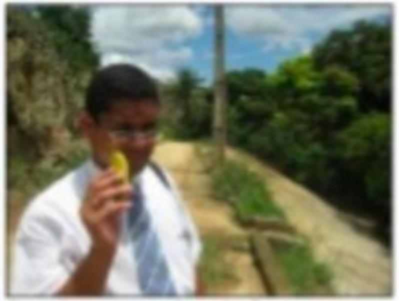
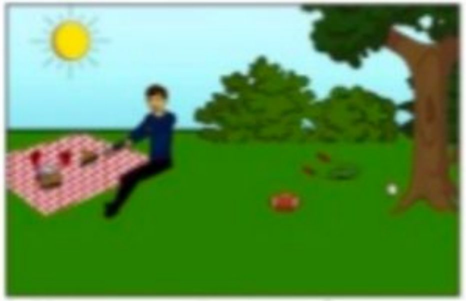
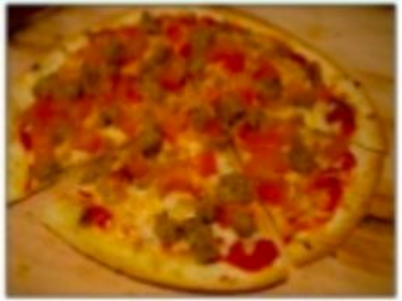
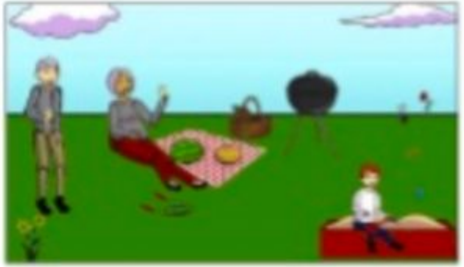
Devshree Patel

Yesha Shastri

Walkthrough

1. Explored EASY-VQA dataset and implemented it.
2. Tinkered with CloudCV VQA demo.
3. Went through : “A Convolutional Neural Network Based Approach For Visual Question Answering” for proper understanding of different models and their usage in VQA domain.
4. Implemented VQA with Bag Of Words approach with VGG Model.
5. Implemented VQA with LSTM.
6. Tried to work with “Hierarchical Question-Image Co-Attention for Visual Question Answering”.

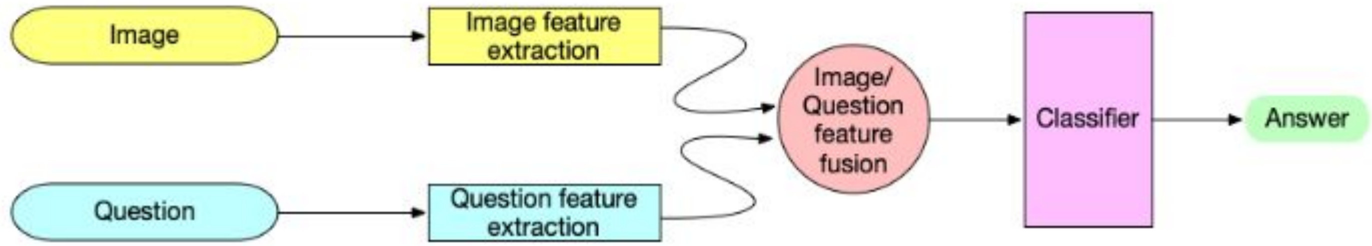
Visual Question-Answering: Types

	Real images		Abstract scenes	
Open-ended		Q: Does it appear to be rainy? A: no		Q: What is just under the tree? A: a ball
Multi-Choice		Q: How many slices of pizza are there? A: 1, 2, 3, 4		Q: What is for desert? A: cake, ice cream, cheesecake, pie

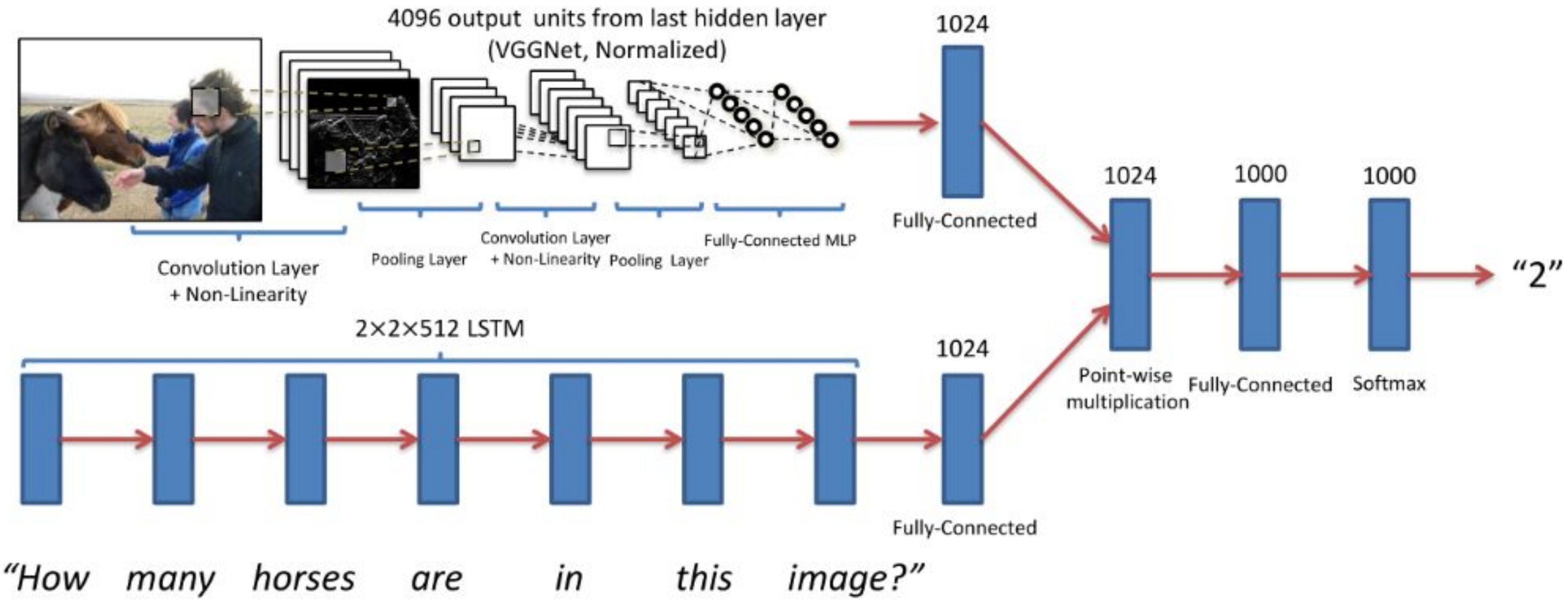
Building Blocks of the System

There are three main steps involved :

1. Feature extraction - This block is used to extract relevant features from the input image and question separately.
2. Feature fusion - This block takes both features and provides a fused multi-modal feature as output.
3. Classifier - This block takes the fused feature vector as input and classifies it into answer labels.



Feature extraction models

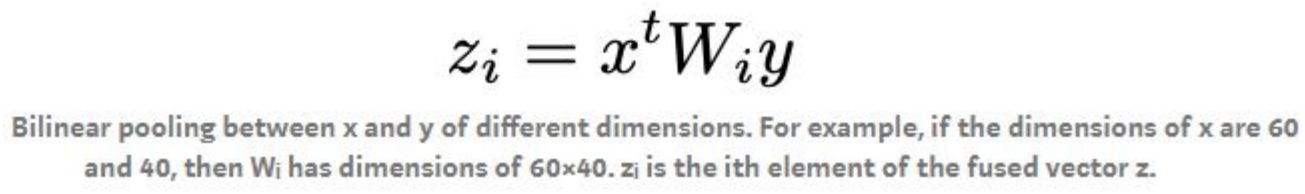


Feature Fusion Methods

1. Element-wise multiplication of image feature vectors and question feature vectors

   Disadvantage: Cannot fuse feature vectors from different dimensions

2. Bilinear Pooling :



$$z_i = x^t W_i y$$

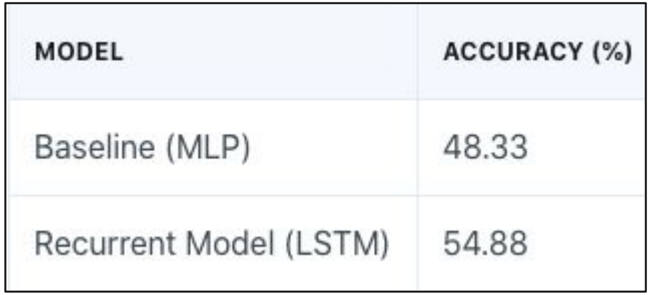
Bilinear pooling between x and y of different dimensions. For example, if the dimensions of x are 60 and 40, then W_i has dimensions of 60×40. z_i is the ith element of the fused vector z.

3. Other methods : concatenation, attention-based pooling, Bayesian-based methods and compositional approaches

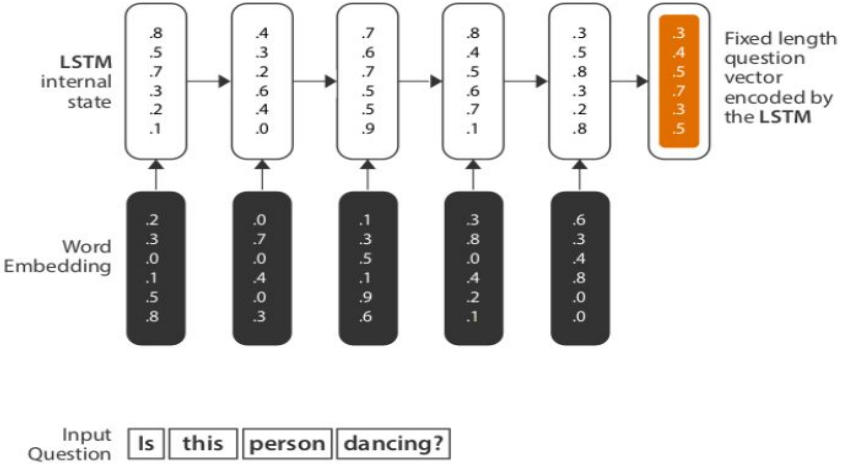
About Baseline Models

- Baseline models gauge the complexity of a dataset.
- For VQA , guessing the answers with high frequencies or guessing randomly are the simplest baseline models.
- One of the most used baseline approach is applying either a nonlinear or a linear multi-layer perceptron (MLP) to the vector that is formed by fusing the question and image features.
- Some of the methods to combine the question and image features are element-wise multiplication, element-wise addition and concatenation.
- It is proved that when question features were represented using BOW and image features were represented using CNN features from GoogLeNet it performed much better than the previous baseline model for COCO-VQA, which used an LSTM for question features.

MLP vs LSTM results on COCO VQA

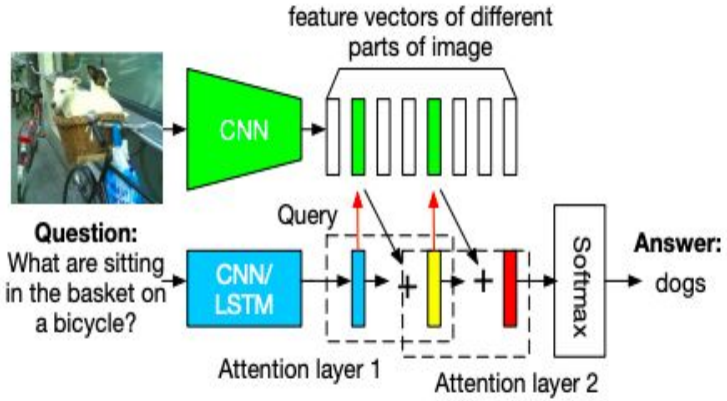


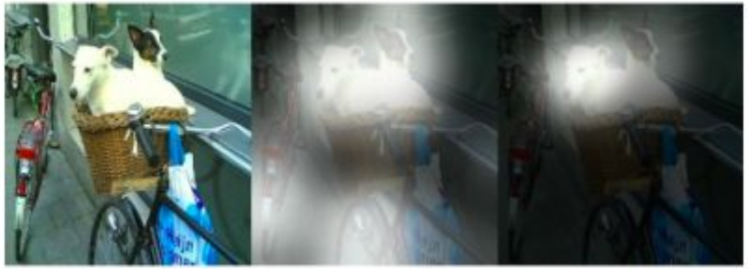
MODEL	ACCURACY (%)
Baseline (MLP)	48.33
Recurrent Model (LSTM)	54.88



- The results depict that memory based models are more efficient in comparison to Baseline Models involving Multi Layer Perceptron.
- In a Recurrent Model, the sequence of words is preserved. This nature of preserving long sequences is what makes RNNs perfect for NLP related tasks.
- We choose to go ahead with LSTMs to avoid a fundamental limitations of vanilla RNNs: the Vanishing Gradient Problem.

Attention Based Models in VQA

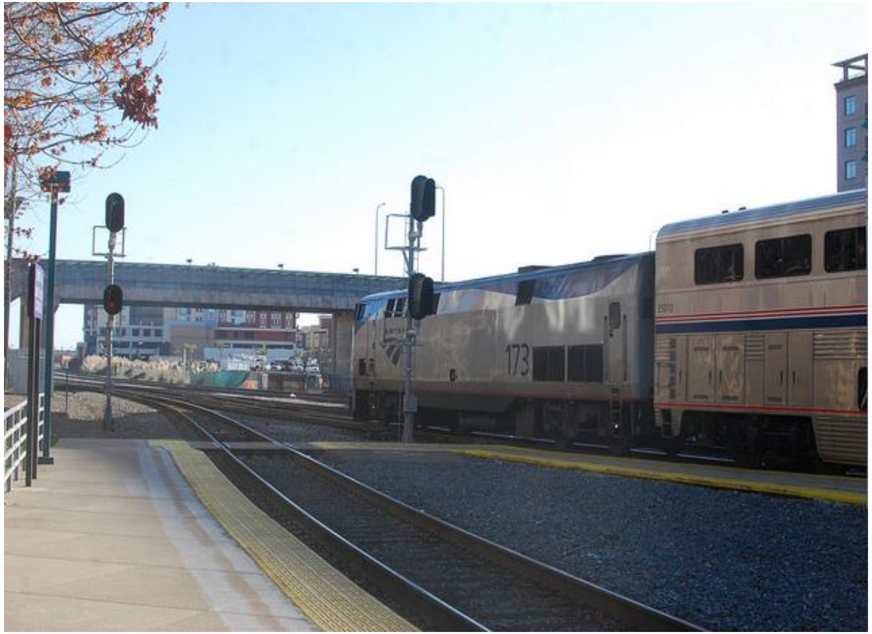

feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2



1. Stacked Attention Networks(SANs) use semantic representation of a question as query to search for the regions in an image that are related to the answer.
2. The SAN first uses the question vector to query the image vectors in the first visual attention layer, then combine the question vector and the retrieved image vectors to form a refined query vector to query the image vectors again in the second attention layer.
3. The higher-level attention layer gives a sharper attention distribution focusing on the regions that are more relevant to the answer.
4. Finally, we combine the image features from the highest attention layer with the last query vector to predict the answer.

Results

Q: What vehicle is in the picture?

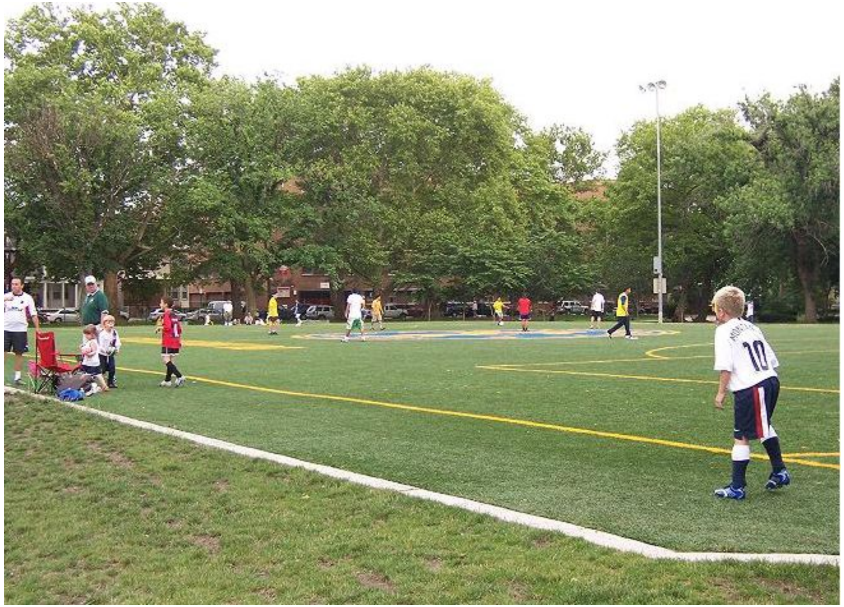


Answer:

78.32 %	train
01.11 %	truck
00.98 %	passenger
00.95 %	fire truck
00.68 %	bus

Results

Q: What are they playing?



Answer:

40.52 %	tennis
28.45 %	soccer
17.88 %	baseball
11.67 %	frisbee
00.15 %	football

THANK YOU!