Visual Query Response

Group 18

Team Members

Dhiraj Srivastava
- MEng, Computer Science
- Email: dhirajsrivastava@vt.edu
- Research Interests: Computer Vision

Sindhura Kommu
- MS, Computer Science
- Email: sindhura@vt.edu
- Research Interests:
  - Multi-modal learning
  - Vision + Language

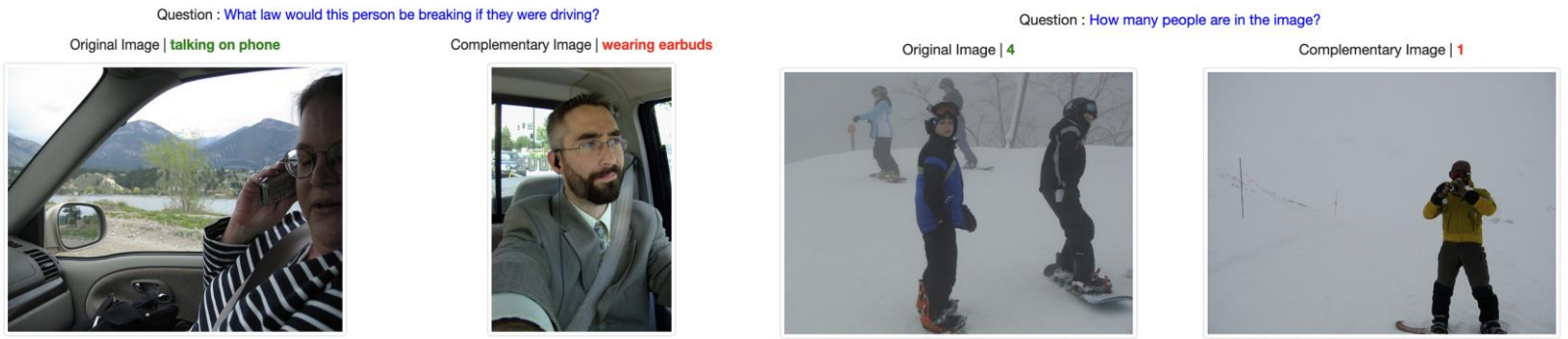
Sulakna Karunaratna
- MS, Computer Science
- Email: sulaknak@vt.edu
- Research Interests: HCI

Mahima Singh
- MEng, Computer Science
- Email: mahimasingh@vt.edu
- Concentration : DA and Artificial Intelligence

Introduction

Visual Question Answering(VQA) is a unique problem that involves both visual and natural language. In this task, the input is an image and an open-ended question about the image, and the output is an open-ended answer to the question with respect to the image.

This can help visually impaired users as well as intelligent analysts who often use visual information for analysis.



Objective

The objective of this project is to create a machine learning model that understands the content of the image and interprets the meaning of the question in order to generate an accurate and relevant answer.

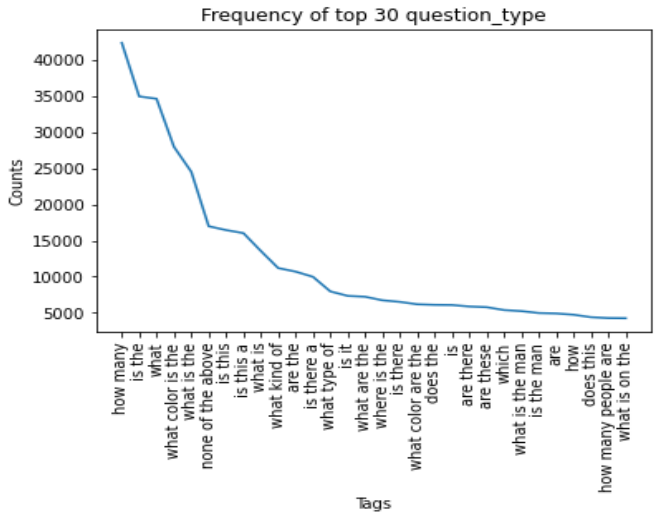
The aim is to perform a comparative study between 2 approaches 1) **CNN+LSTM** and 2) **Hierarchical Question-Image Co-attention**. Experiments have been done with several baseline models and Co-attention models. Different pre-trained models and image sizes have also been used to obtain a model with best results.

Approach

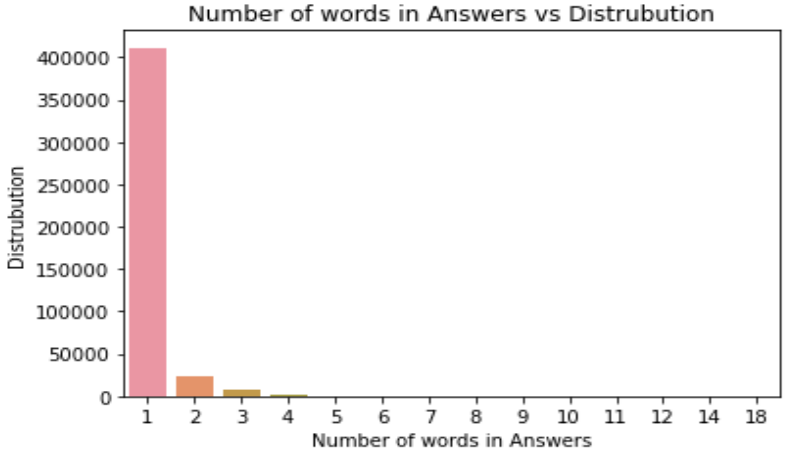
1. Exploratory Data Analysis
    a. VQA 2.0 dataset (https://visualqa.org/)
    b. Images from COCO dataset along with question-answer pairs
    c. Each image has at least 3 questions
    d. Each question has 10 answers from unique people
2. Feature Engineering
3. Models:
    a. Baseline Models
    b. Hierarchical Question-Image Co-attention
4. Experimental Results

Exploratory Data Analysis

To gain an understanding of the types of questions asked and answers provided, we start with analyzing the questions and answers in this VQA train dataset.



The most frequent question types are “how many”, “is the”, “What”, “what is the” .

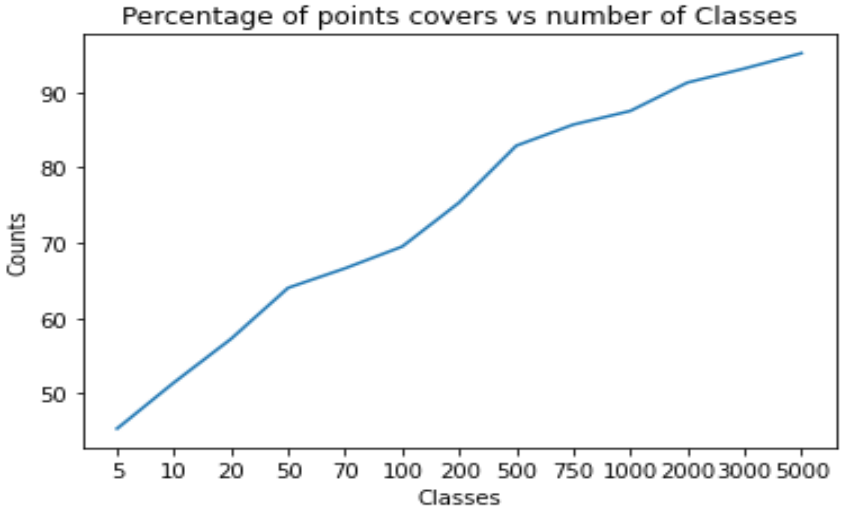


Most answers consist of a single word.

Feature Engineering

Types of input

1. Images
    a. Convert all images into vectors of same size
    b. Feature vectors of train and validation datasets are stored in pickle file format
2. Questions
    a. Convert inputs into fixed size vectors
    b. Encode answer vectors using label-binarizer
    c. Store final outputs as train.csv and val.csv



Choosing top 1000 most frequent answers as possible output for building the models will be a good choice as it covers more than 85% of answers present in dataset

Image Feature Extraction

We are using two architectures to observe if model performance would depend on the type of pre-trained model used for feature extraction. Both the architectures are trained on Imagenet and the last layer of classification is removed

Feature extraction is done on two image sizes 224*224 and 448*448 separately. Hence for each image a dimensional vector representation of 7*7*512 is used for the image size 224*224 and 14*14*512 for image size of 448*448.

Pre-trained Models used: VGG-19, Resnet-50

Image Sizes used: 224*224, 448*448

Question Feature Extraction

Question Vectors are extracted by the following steps:

1. Tokenizing the text
2. Sequence Padding

The shape of the question features is [24,].

Shape of train answer vectors is (310649,1000) and val answer vectors is (8872,1000) as we are considering only the top 1000 frequent answers.

Annotations and question json file of train and val are combined and stored as train.csv and val.csv to avoid frequent use of processing the annotations and questions json files.

Baseline Models

The scaled image is fed into a convolutional neural network (CNN) such as VGG-19 which outputs a feature vector encoding the contents of the image and is referred to as an *image embedding*.

The question is fed into an embedding layer, resulting in a *question embedding*. (LSTM)

These embedding vectors are then projected into the same number of dimensions using corresponding fully connected layers (a linear transformation)

Then combined using pointwise multiplication (multiplying values at corresponding dimensions).

The final stage is a multilayer perceptron at the end that outputs a score distribution over each of the top 1000 answers.

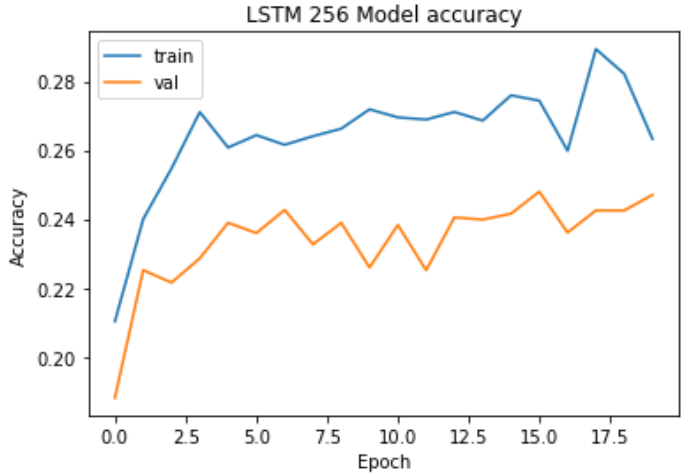
Cross-entropy loss and Adam optimizer is used for training models

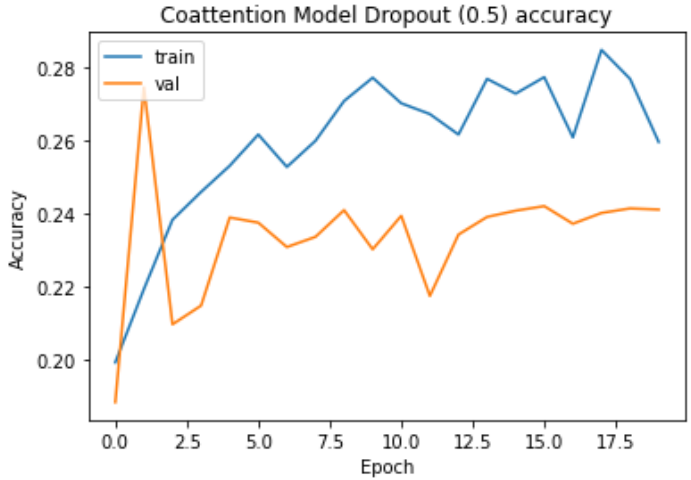
Hierarchical Question-Image Co-attention

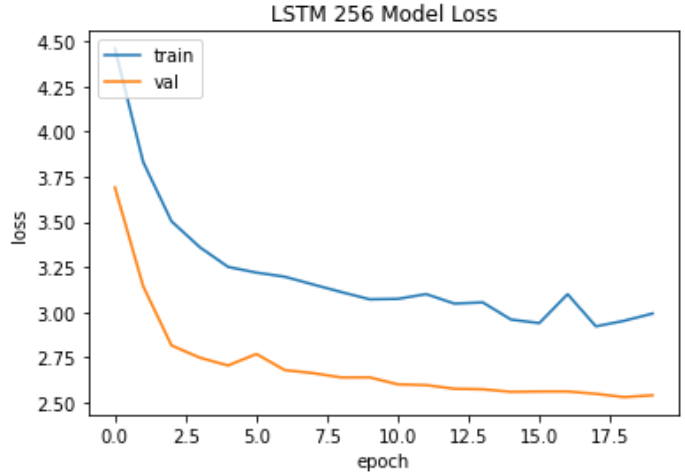
The approach involves a hierarchical attention mechanism that attends to both the image and the question at multiple levels of granularity. This allows the model to focus on relevant regions of the image and words in the question that are important for answering the question.

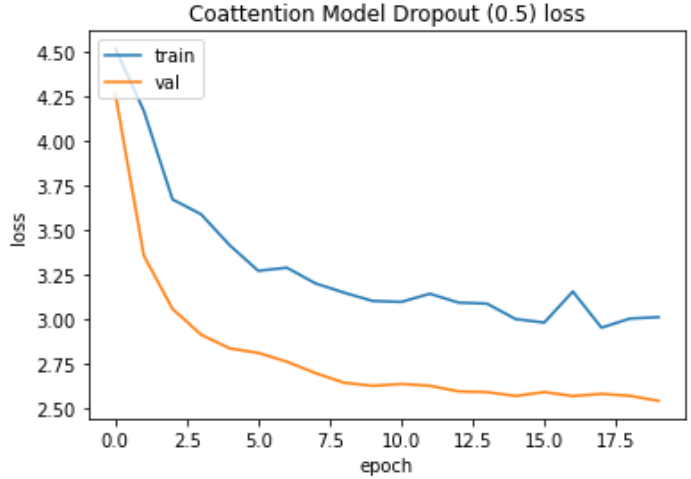
At the lowest level, the model attends to specific image regions and question words to generate a set of region and word features. These features are then used to compute co-attention scores between the image and question at the next level, which allows the model to attend to more abstract concepts such as objects and relationships.

Finally, the model attends to the entire image and question to generate a joint feature representation that is used to predict the answer.


LSTM 256 Model accuracy
train
val
Accuracy
Epoch


Coattention Model Dropout (0.5) accuracy
train
val
Accuracy
Epoch


LSTM 256 Model Loss
train
val
loss
epoch


Coattention Model Dropout (0.5) loss
train
val
loss
epoch

Experimental Results

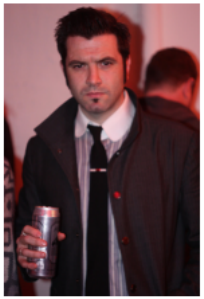
Evaluating performance on validation set -

accuracy = minimum (humans that provided the answer/3 , 1)

It means that the answer is 100% accurate if at least 3 humans gave the exact answer and it gets score of 1. If the answer predicted by model isn’t present amongst 10 humans, it gets a score of zero.

Models	Accuracy
VGG Co-attention (448*448)	47.26
VGG Co-attention (448*448) (SGD)	31.04
VGG Co-attention (448*448) (LSTM)(256)	38.91
VGG Co-attention (448*448) Dropout (0.5)	31.40
VGG Co-Attention (224*224)	47.34
VGG Image Attention (224*224)	46.88
VGG Question Attention (224*224)	42.95
Resnet Co-Attention (224*224)	40.95
Resnet Image Attention (224*224)	43.73
Resnet Question Attention (224*224)	43.42
VGG Baseline (224*224) (LSTM)	43.99
VGG Baseline (224*224) (GLOVE)	44.05
VGG Baseline (224*224) (BERT)	42.81
Resnet Baseline (224*224) (LSTM)	44.34
Resnet Baseline (224*224) (GLOVE)	43.89
Resnet Baseline (224*224) (BERT)	42.81

Inference



```
Question : is he wearing a hat?
Actual Answer:  no
Top Predicted answers:  [('no', 55.54821), ('yes', 44.45179), ('1', 6.1360186e-07), ('2', 9.231724e-08), ('0', 3.4746705e-08)]
****************************************************************************************************
```

```
Question : what is the color of the glowing traffic light?
Actual Answer:  green
Top Predicted answers:  [('red', 66.80796), ('green', 23.41766), ('yellow', 8.173193), ('white', 1.0922873), ('red and yellow', 0.1431633)]
****************************************************************************************************
```

Question : what is the color of the glowing traffic light?
Actual Answer: green
Top Predicted answers: [('red', 66.80796), ('green', 23.41766), ('yellow', 8.173193), ('white', 1.0922873), ('red and yellow', 0.1431633)]
****

Question : is this food able to be prepared on a grill?
Actual Answer: no
Top Predicted answers: [('no', 50.234337), ('yes', 49.765656), ('1', 4.7045123e-06), ('2', 1.1151117e-06), ('cake', 3.876436e-07)]
****

Question : what is in the foreground?
Actual Answer: fence

Question : is the beach crowded?
Actual Answer: yes
Top Predicted answers: [('yes', 52.5334), ('no', 47.466602), ('1', 1.7095749e-06), ('2', 3.9817536e-07), ('afternoon', 1.2678767e-07)
****

Question : are any of the boats moving?
Actual Answer: no
Top Predicted answers: [('no', 54.454113), ('yes', 45.545788), ('1', 3.419024e-05), ('day', 2.779732e-05), ('0', 1.6353299e-05)]
****

Question : how many people are on the bike?
Actual Answer: 2
Top Predicted answers: [('1', 31.405476), ('2', 21.377087), ('0', 17.672972), ('3', 10.873758), ('4', 5.7885256)]
****

Question : What color is the bed 's comforter?
Top Predicted answers: [('white', 31.101748), ('black', 20.588829), ('brown', 15.156311), ('blue', 10.076968), ('gray', 4.9236856)]
****

Question : Is he surfing?
Top Predicted answers: [('yes', 68.25143), ('no', 25.461214), ('surfing', 5.8568053), ('wetsuit', 0.1411966), ('ocean', 0.08410803)]
****

Question : Are the numbers written in Roman numerals?
Top Predicted answers: [('0', 16.223413), ('1', 8.809862), ('2', 8.522955), ('3', 6.475351), ('5', 5.1545978)]
****

Lessons Learned

1. Using existing resources: Leveraging existing code libraries and research papers to study and learn about the ongoing research. Fine tuning the pre-trained deep learning models to save time and effort.
2. Divide and Conquer: Breaking the project into smaller steps. Setting appropriate timeline and deadlines. Dividing the workload among team members. Using code management tools such as github for effective collaboration.
3. Establishing effective communication channel for group project
4. Computation effects runtime

Future Work

Question types that have a few possible answers working very well (eg ‘what room’, ‘what sport’). And question type like ‘why’, ‘how’ will have vast possible answers so the model is not working so well.

We are using VGG which gives 512 filters at the end. If we increase the filters, the model can identify more patterns so that we can increase model performance (eg ResNet final Conv layer has 2048 filters).

Appendix

Git Hub Repository Link : here

References

- Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- S. Hassantabar. Visual question answering: Datasets, methods, challenges and oppurtunities. Princeton University, 2018.
- I. Ilievski and J. Feng. Multimodal learning and reasoning for visual question answering. In NIPS, 2017. 3
- R. C. Staudemeyer and E. R. Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks

Thank You!