Indian Institute of Technology, Hyderabad **EMDS - 2017**

Advanced Topics in Machine Learning (CS6360)

Instructor: Dr. Vineeth N Balasubramanian

Project Proposal
Visual Question Answering

Team Members:

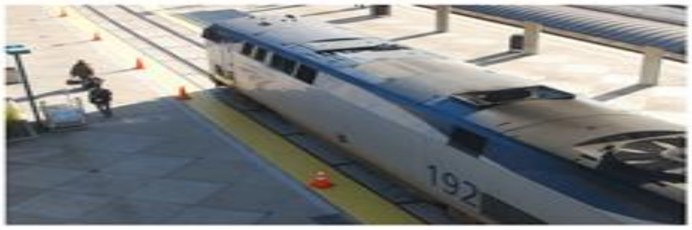
- **Rohit Dubey (CS17EMDS11019)**
- **Paritosh Gupta (CS17EMDS11016)**
- **Harish Mashetty (CS17EMDS11010)**

Project objective

Visual Question Answering (VQA) is an Artificial Intelligence problem that lies at the intersection of NLP and Computer Vision. VQA answers open ended natural language question about a given image

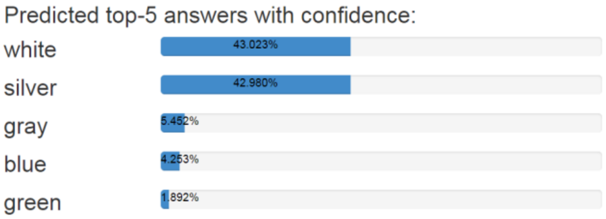
Figure 1: Example Visual questions and answers, generated using CloudCV, http://vqa.cloudcv.org/

Input Image



Questions Asked

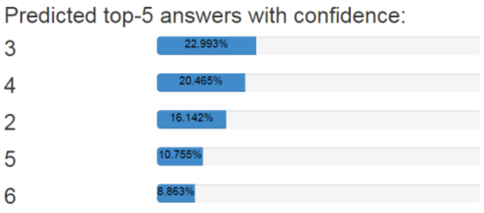
Question: What color is the train?



Predicted top-5 answers with confidence:

white	43.023%
silver	42.980%
gray	5.452%
blue	4.253%
green	1.892%

Question: How many passengers near the train?



Predicted top-5 answers with confidence:

3	22.993%
4	20.465%
2	16.142%
5	10.755%
6	8.863%

Proposed methodology

Several Neural Networks based approaches have been proposed, mostly using Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN). These techniques used CNN features for encoding images, and LSTM for encoding the question, as well as for generating the answers.

We propose to begin with implementing (LSTM + CNN + LSTM) - LSTM to extract question representation, CNN to extract image representation, another LSTM for storing context in an answer, combining the three to generate answers. We will use pre-trained CNN such as GoogleNet to extract visual features and Google's Word2Vec to extract word Embedding.

We then propose to enhance this model by exploring techniques such as Visual attention mechanisms, and modern neural network architectures.

What is expected to be novel in your work

We aim to beat the existing performance benchmarks [1], to achieve that we propose to explore techniques including (but not limited to)

1. Visual attention mechanisms to focus on parts of image which are relevant to the question.
2. Applying different variants of RNN and CNN, and modern neural network architectures.
3. Advanced techniques to learn interactions between image and question, and ways to combine joint representation.
4. Other techniques learnt as part of the course.

Datasets, Planned experiments

Datasets

1. Visual QA dataset from https://visualqa.org/
2. Common Objects in Context dataset from http://cocodataset.org

Planned Experiments

1. Begin with train, evaluate the existing approaches [[1], [2]] using combination of CNN and LSTM to establish a benchmark. We plan to use a pre-trained CNN such as GoogLeNet to extract visual features and Google's Word2Vec to learn word Embedding.
2. Improve the performance by exploring modern neural network architectures including but not limited to following approaches:

a. Experiment with choice of deep learning methods and their combinations to improve the performance.
b. Exploring visual attention mechanisms to improve the performance, for example, a question asking number of passengers near train, should pay more attention on the platform.
c. Bidirectional LSTM to better learn interactions between image and content of the question.

Performance metrics and expected results

The predicted results obtained from the resultant model will be the short answers of the asked questions and will be evaluated with standard evaluation techniques used in Visual Question Answer world.

1. One of the evaluation approaches is to use the evaluation metrics used in VQA challenge (https://visualqa.org/evaluation.html)

$$\mathrm{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

Above evaluation metric is robust to inter-human variability in phrasing the answers. In order to be consistent with 'human accuracies', machine accuracies are averaged over all 10 choose 9 sets of human annotators

2. **BLEU- The Bilingual Evaluation Understudy Score**, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.
The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of order

3. If time permits, we'll compete in VQA 2 challenge 2019 and submit our model to get competition score and rank.

Any appropriate references

[1] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, Devi Parikh Facebook AI Research Pythia v0.1: The Winning Entry to the VQA Challenge 2018

[2] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering, (arXiv:1505.05612[cs.AI]).