VISUAL QUESTION ANSWERING

Krishna Chaitanya Gudipati Benu Changmai Mayuresh Anand
{krishnachaitanya, benu, mayuresh}@ucsb.edu

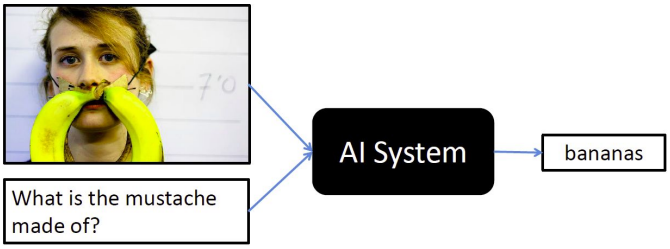
Abstract

We study and implement the work on Visual Question Answering (VQA) presented by Lu et al [2] . The challenge here is to answer the questions by generating spatial maps that highlight image regions relevant to questions. Previous works have focused on the question of visual attention or "where to look" but in this model also considers simultaneously the issue of question attention i.e. "what words to listen to". This model jointly reasons about image and question attention by modelling them using co-attention method. This implemented code achieves an accuracy of 50.2% on the VQA dataset using ResNet model.

1. INTRODUCTION

Historically, it has been one of the most pondered and imagined thought to be able to construct or build a system that can answer natural language questions. We come across many movies and series where an android can communicate to humans, can answer questions, can reason and construct judgement. What we want is that our model must be intelligent enough either by training or use of an external knowledge base so that it can answer some questions like:

- What is the color of the bird?
- Are there any humans?
- What sport is being played?
- Where is the woman while her child is sleeping?
- How many players are in the image?
- What type of wine is being served with
- Is it raining? ...

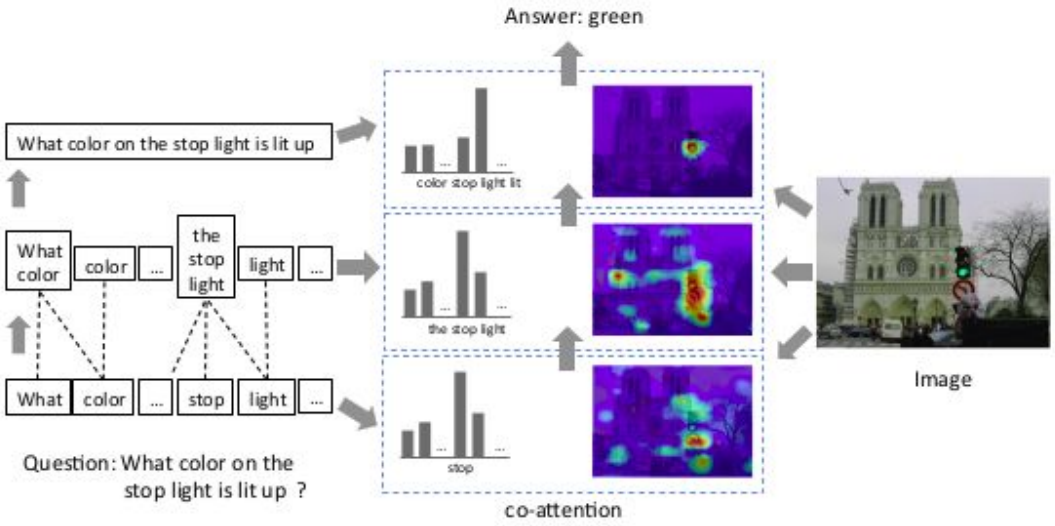


Many models have been proposed which specifically focuses on the problem of identifying "where to look" or *visual attention* i.e. where in the image can we look to find the answer to the question. While it is important to find the region that can accurately answer the question it is even more important to understand what information has been asked about in the question. Consider for example: "How many players are in the image?" and "How many players can you see in the image?". It is important to know that "How many players" is the key here to answer the question and a machine that can capture the essence of this question by looking at the three words would be able to answer many variations of the question.

The proposed model is a multi-modal attention model for VQA with the following unique features:

- **Co-attention:** this mechanism jointly reasons about visual attention and question attention, which we are referred to as co-attention. Here, the image representation is used to guide the question attention and the question representation(s) are used to guide image attention.

- **Question Hierarchy:** A hierarchical architecture co-attends to the image and question at three levels: (a) word level (b) phrase level and (c) question level. Word level embeds a vector space through an embedding matrix, phrase level uses 1D convolutional neural network to capture the information contained in unigrams, bigrams and trigrams. This is achieved by convolving word representations with temporal filters of varying support which are then pooled into single phrase level. At question level recurrent neural network is used to encode the entire question and joint question and image co-attention are constructed and combined recursively to predict a distribution over the answers.



Hierarchical Co-attention on question and image features

2. DATASET

We use the VQA v2 dataset provided in the VQA challenge. Details of which are the following:
The input images were taken from the MS COCO dataset. Each image has 5-6 open ended questions.

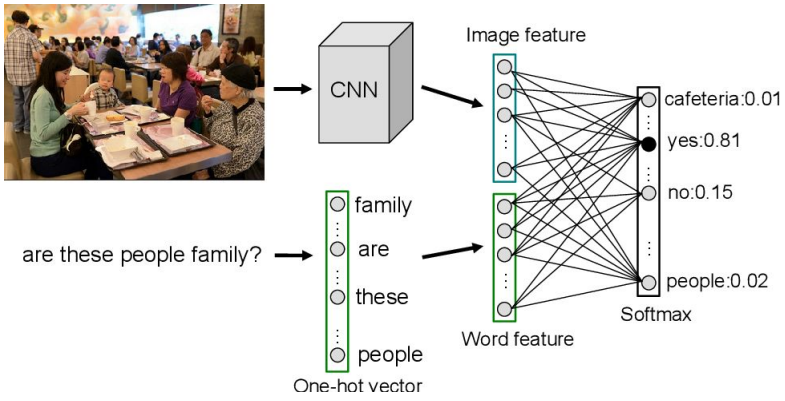
VQA ANNOTATIONS	VQA INPUT IMAGES
• Training : 21,50,000 answers • Validation : 510,000answers **VQA INPUT QUESTIONS** • Training : 215,000 questions • Validation : 51,000 questions	• Training : 40,000 images • Validation : 10,000 images

We treat the VQA as a classification problem. So we extract the top 1000 most frequent annotations from the training data and treat it as a 1000 class classification problem. We then filter the training questions based on the extracted annotations and create vocabulary by tokenizing with the NLTK tokenizer.

3. MODEL DESCRIPTION

Baseline Model:

- In baseline model, we extract the 1024 size feature vector using GoogleNet and also a 1024 size feature vector from question using the Bag of Words model. Finally, we concatenate these two feature vectors and apply softmax to predict the answer.



Hierarchical Co-attention model:

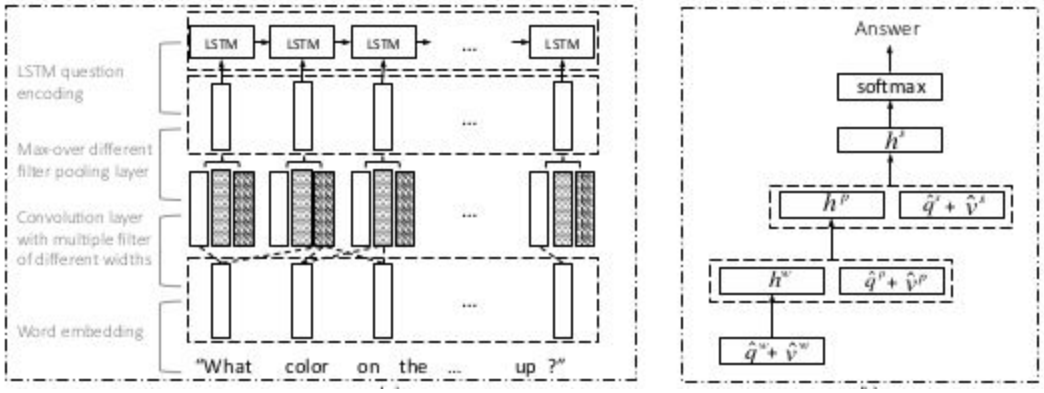
- **Question Hierarchy:**

  $Q = \{q_1, ..., q_T\}$ is 1-hot encoding of question words, we embed the words vector space to get $Q^w = \{q_1, ..., q_T\}$ (word embedding). Apply 1D convolution on this embedding vector, this is done by computing the inner product of the word vectors with filters of three window sizes: unigram, bigram and trigram.

  $\hat{q}^p{}_{s,t} = tanh(W^s{}_c q^w{}_{t:t+s-1})$, $s \in \{1, 2, 3\}$ where $W^s{}_c$ is the weight parameters. Q^w is now 0-padded appropriately and max-pooling is applied across different n-grams at each word location

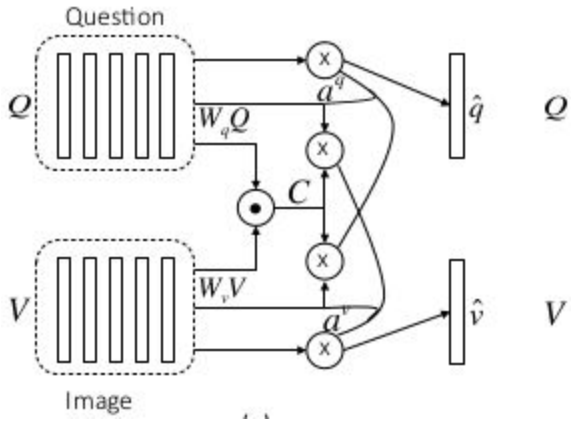
  $q_t = max(\hat{q}^p{}_{1,t}, \hat{q}^p{}_{2,t}, \hat{q}^p{}_{3,t})$, $t \in \{1, 2, \ ... \ , T\}$.

  Pooling here is adaptive while selecting different gram features at each timestep. An LSTM is used to encode the sequence $q_t{}^p$ after max pooling operation and the corresponding question-level feature $q_t{}^s$ is the LSTM hidden vector at time *t*.



- **Co-attention:** We use parallel co-attention mode. Images and questions are connected to each other by calculating the affinity matrix $C = tanh(Q^T W_b V)$ where $V \in R^{d \times N}$, $Q \in R^{d \times T}$, a $C \in R^{T \times N}$, *and* $W_b \in R^{d \times d}$. Now consider affinity matrix as feature and learn to predict image and question maps via following:

  $H^v = tanh(W_v V + (W_q Q)C)$, $H^q = tanh(W_q V + (W_v V)C^T)$, $a^v = softmax(w^T{}_{hv} H^v)$, $a^q = softmax(w^T{}_{hq} H^q)$



  Where $W_q, W_v \in R^{k \times d}$, $w_{hv}, w_{hq} \in R^k$, $a^v \in R^N$ *and* $a^q \in R^T$ are attention probabilities of each image region v_n *and* q_t. Image and attention vectors are calculated as the weighted sum of the image features and question features, i.e.,

$$\hat{v} = \sum_{n=1}^{N} a_n{}^v v_n, \ \hat{q} = \sum_{t-1}^{T} a_t{}^q q_t$$

  The parallel co-attention is done at each level in the hierarchy, leading to $\hat{v}^r$ and $\hat{q}^r$ where $r \in \{w, p, s\}$

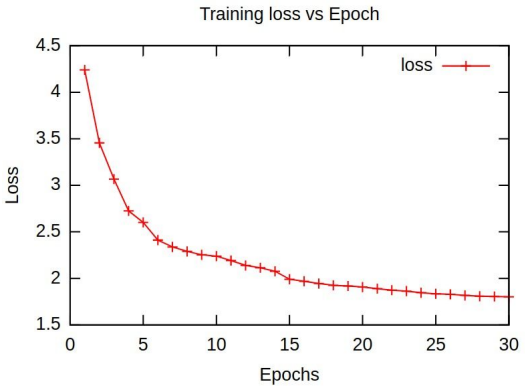
4. EXPERIMENTAL DETAILS

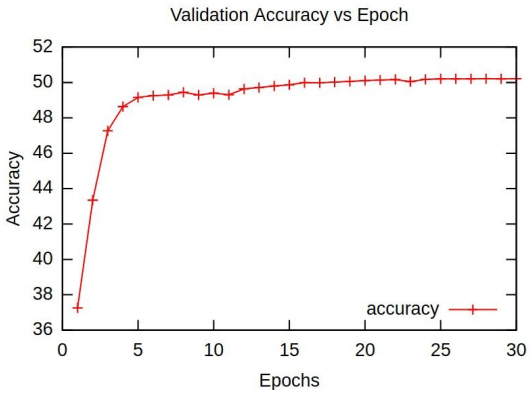
Below are the various hyper-parameters we used for training the model:

- Learning rate: 0.001
- Optimizer: Adam optimizer
- Batch size: 128
- #Epochs: 30

We performed our training on GPU-aided XSEDE Comet cluster with the following configurations:

- GPU: K80
- #GPU cores: 4
- RAM size: 12GB
- GPU memory: 100GB





The total training time was around 20 mins/epoch. We calculated validation accuracy at the end of each epoch. The final validation scores for the baseline model was ~42.1% while for the co-attention model we achieved nearly ~50.2% accuracy. Figures below display how the loss and validation accuracy shifted along the course of training for the co-attention model.

5. VISUALIZATIONS

Upon testing the baseline and co-attention models on different visual scenarios we could observe that both the models fare pretty well when it comes to simple images with fewer components which are distinctly separated from each other.

Question: What sport is being played in the picture?
Baseline: **Skateboarding**
Coattention: **Skateboarding**

Question: What room is this used as
Baseline: **Bathroom**
Coattention: **Bathroom**

Question: Is there a baby in the picture?
Baseline: **Yes**
Coattention: **Yes**

For questions involving images which required the model to focus locally on the concerned features, the co-attention model clearly outperforms the baseline.

Question: How many zebras are in the picture?
Baseline: 1
Coattention: 2

Question: Is there a car in the picture?
Baseline: No
Coattention: Yes

Question: How many people are in the picture?
Baseline: 3
Coattention: 4

Both the models fail in cases where the objects are vague and many. Due to the lack of definite localized features, the co-attention model is not able to guide the question attention and vice versa

Question: How many boats are in the picture?
Baseline: 2
Coattention: 3

Question: Is the cat sitting on a person?
Baseline: No
Coattention: No

Question: How many umbrellas are in the picture?
Baseline: 2
Coattention: 2

6. CONCLUSION

We explored the problem statement of Visual Question Answering which served as an effective way to get hands-on experience with some interesting concepts of both Computer Vision and Natural Language Processing. The experiment involved a classification problem where text-based questions to visual inputs were answered by finding the most probable answer label. We implemented 2 different models, namely a Simple Baseline and Hierarchical Co-attention model. In the process we experimented with image features provided by well-known network structures like ResNet, GoogleNet and VGG. The textual features were generated at word-level, phrase-level and sentence-level which is the hierarchical structure of questions that the Co-attention model uses. The experimental results revealed interesting insights into both the models which resonated with our theoretical understanding of the limitations and capabilities of each of the models. We were also able to achieve performances of the models in terms of validation accuracy which were at par with the results as stated in the papers that we followed.

REFERENCES

[1] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167.

[2] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." In Advances In Neural Information Processing Systems, pp. 289-297. 2016.

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).