Attention Mechanisms for Visual Question Answering

Paritosh Gupta
CS17EMDS11016@iith.ac.in

Rohit Dubey
CS17EMS11019@iith.ac.in

Harish Mashetty
CS17EMS11010@iith.ac.in

Abstract

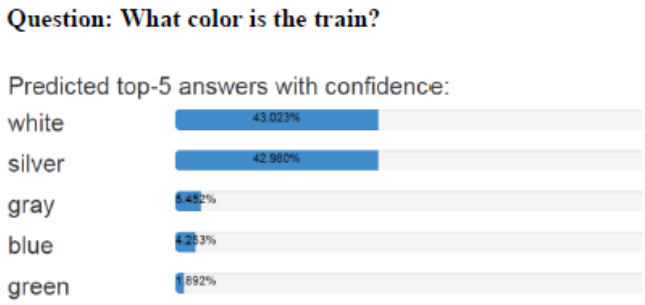
This paper presents the implementation of attention mechanisms for visual question answering (VQA) by improving upon the existing solutions. To implement the attention mechanisms, we have taken a deep learning architecture from inspired from 2017 VQA winner paper [1]. The proposed methodology implements multi modal fusion of joint embedding of images and questions, bottom-up attention to extract salient image regions, top-down one glimpse & dual top-down stacked attention for guided attention in context of the question asked. The nonlinear layers are implemented using gated tanh and sigmoid activations for final classification with cross entropy loss between soft scores from true labels and predicted scores.

1. Introduction

Visual Question Answering (VQA) is an Artificial Intelligence problem that lies at the intersection of NLP and Computer Vision. Visual Question Answering (VQA) involves an image and an open-ended natural language question, to which the machine must determine the correct answer (see Fig. 1). The model is based on a deep neural network that implements joint embedding of the input question and the given image, followed by the multi-class classifier over a set of candidate answers.

In our implementation, model takes two types of inputs- images and corresponding questions/answers. For image input we used faster R-CNN processed input for bottom-up attention available at [2] which is further normalized before feeding into the model. For questions we used 300-dimension Glove word embeddings [3] which is further encoded with last internal state of Gated recurrent unit (GRU) and feed it to the model.

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



Human annotations (Ground truth): {white, white, silver, grey, silver, silver, white, white, white, grey}

Soft scores (Class labels): {white: 1, silver 1}, soft score is calculated using equation (6).

Soft scores as targets allows multiple correct answers per question, which is often the case due to uncertainty in ground truth annotations.

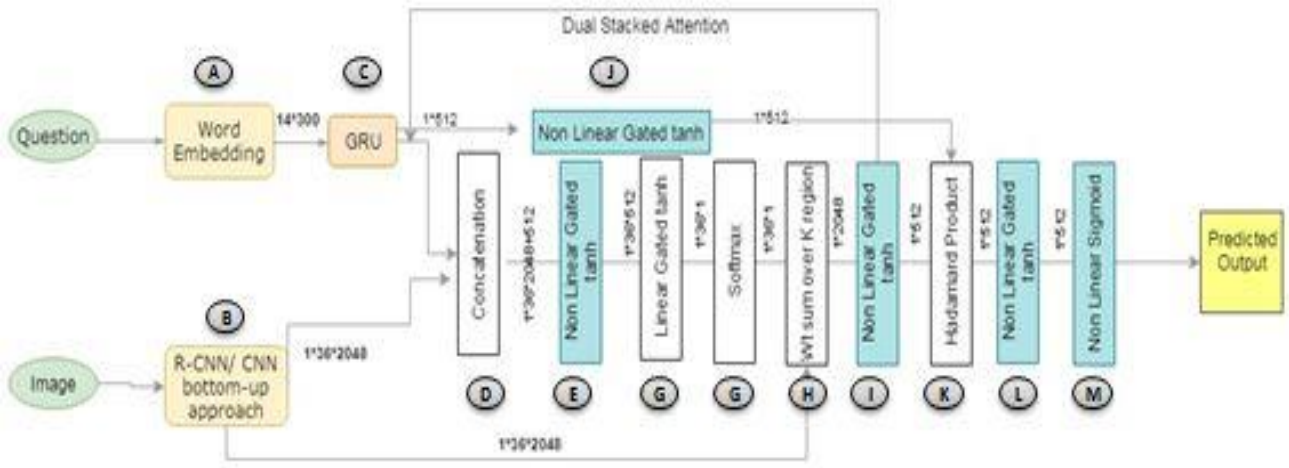


Figure2. Graphical representation of the model architecture. A deep learning network which uses a joint embedding of the input image and question followed by a multi-class classifier over a fixed set of candidate answers.

2. Proposed Model

Our proposed solution takes VQA as a classification problem over a set of candidate answers. Questions are open-ended questions about images, with mostly one- or two-word answers. Our model is based on a deep neural network that implements a joint embedding of the image and of the question. The two inputs are mapped into fixed-size vector representations which are derived from Faster RCNN and GRU, respectively. Multimodal fusion of non-linear output of attended image and question encoding can be interpreted as projections into a joint "semantic" space. They are combined by concatenation of element-wise multiplication, before feeding to the classifier described above.

Let's understand the importance and working of each part of model architecture one by one.

Word Embedding: Questions are tokenized and trimmed to a maximum of 14 words and the questions with less than 14 words are zero padded. We find that not many questions' length exceeds 14 words (only 0.25% questions are greater than 14 words). After that, each word is converted into 300-Dimensional GloVe embedding vector, words missing in Glove embedding are initialized with a zero vector. We used Wikipedia/Gigaword pre-trained GloVe embedding which is available publicly. The resulting sequence of word embedding of a question is of size 1*14*300 and it is passed to GRU encoder (Fig 2- part C). The Recurrent Gated hidden unit is of dimension 512 and we used its final state as question embedding of size 1*512.

Image Bottom-up attention features: The bottom up attention image feature map taken from a publicly available Faster R-CNN framework [2] which was trained on Visual Genome Dataset and ResNet. The resulting features can be interpreted as ResNet features centered on the top-K objects in the image.

We choose K=36 for fast processing considering the available resources. Each K region is thus represented by 2048-Dimensional vector that encodes the appearance of the salient image region. Each image is of size 1*36*2048 tensor.

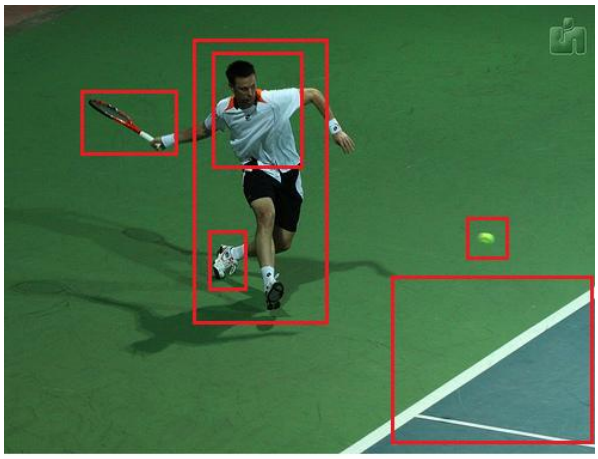
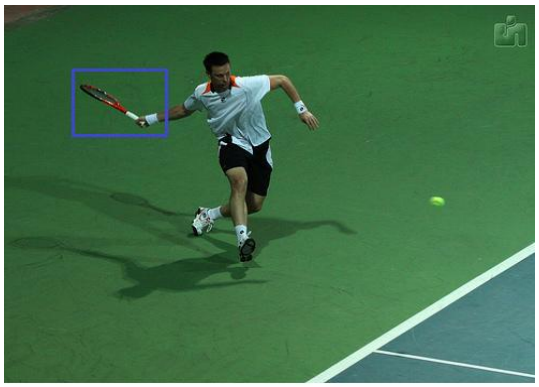


Fig: Provides salient image regions, each region represented by a CNN pooling layer feature vector of Faster R-CNN.

One-glimpse Top-down Image attention:

This model implements a standard question-guided attention mechanism used in modern VQA models. **This question-guided attention mechanism is termed as top-down attention**.

For each input region i=1 to K in the image, the feature v_i is concatenated with question embedding q (Fig 2- part D). Then both passed through a non-linear layer f_a (Fig 2- part E) and a linear layer (Fig 2- part F) followed by a SoftMax to obtain a scalar attention weight α associated with that location.



Question: what is the man holding in hand?
Fig: Task specific, in this example, question is related to man holding an object, thus the corresponding image region is given more importance.

$$a_i = w_a f_a([v_i, q]) \quad (1)$$
$$\alpha = softmax(a) \quad (2)$$
$$\hat{v} = \sum_{i=1}^{k} \alpha_i v_i \quad (3)$$

Where w_a is a learned parameter vector. The attention weights are normalized over all locations with a SoftMax function (eq 2) (Fig 2- part G). Then image feature from all regions/locations are weighted by the normalized values and summed together (eq 3) (Fig 2- part H) to get a single 2048-sized vector $\hat{v}$representing the attended image.
The weighted sum of all the image regions/locations is a basic attention mechanism known as simple one-glimpse, one-way attention.

Dual Top-down Stacked attention:

For complicated questions, a single attention layer is not sufficient to identify the correct region for answer prediction. Dual stacked attention adds the attended image vector back to the question vector and performs one more iteration of attention. We implemented dual top down stacked attention from the paper [4] (eq 4), which queries an image two times to focus attention to the specific region relevant to the context of asked question.

$$\hat{v} = q + \hat{v} \quad (4)$$

(a) Stacked Attention Network for Image QA

Original Image First Attention Layer Second Attention Layer

Question: What is sitting in the basket on a bicycle?

Multimodal Fusion:

The image representation $\hat{v}$ we got for one-glimpse or for stacked attention and question (q) are passed through non-linear layers (gated tanh) and then combined with simple Hadamard product (element-wise multiplication):

$$h = f_q(q) \circ f_v(\hat{v}) \quad (5)$$

The resulting vector h is called as the joint embedding of the image and question. This h is then fed to the output classifier.

Output Classifier:

We treat VQA as a multi-class classification task. A set of candidate answers (output vocabulary), is a predetermined set of all correct answers that appeared more than 8 times in training dataset. This result in N=3129 candidate answers and acted as classes in classification task. In VQA2 dataset, each question has been asked to multiple people, and their response recorded, thus in many cases there is no one correct answer, for example some people can identify color of an object as silver, while others might say white. Thus, instead of taking one correct answer we create soft score using formula in equation (6), thus any answer which was given by three or more persons is taken as correct. This approach of soft scoring takes care of any ambiguity/disagreement between human annotators.

$$Acc(ans) = \min\left\{\frac{\#\text{ humans that said ans}}{3}, 1\right\} \quad (6)$$

The multi-class classifier passes the joint embedding h through a non-linear layer f_o and then through a linear mapping w_o to predict as score $\hat{s}$ for each

N candidates:

$$\hat{s} = \sigma\big(w_o f_o(h)\big) \qquad (7)$$

where sigma is a sigmoid (logistic) activation function and w_o is learned weights. The loss is binary cross entropy with logits w.r.t soft scores. This final stage is acting as logistic regression that predicts the correctness of each candidate answer. The advantage of this approach is that sigmoid outputs allow optimization for multiple correct answers per questions and use of soft scores as targets provides a somewhat better training signal than binary targets, as they capture the occasional uncertainty in ground truth observations.

The classification accuracy is calculated by comparing predicted and actual set of answers per question for given sample.

Non-Linear Layers:

Our model used non-linear layers in multiple occasions. We tried ReLU and tanh and selected tanh after considering the trade-off between time complexity vs performance. In our model, each non-linear layer uses gated tanh activation. The tanh function is defined as-

$$\hat{\boldsymbol{y}} = \tanh(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \qquad (8)$$
$$\boldsymbol{g} = \boldsymbol{\sigma}(\boldsymbol{W}'\boldsymbol{x} + \boldsymbol{b}') \qquad (9)$$
$$\boldsymbol{y} = \hat{\boldsymbol{y}} \circ \boldsymbol{g} \qquad (10)$$

This formulation is inspired from similar gating operations within recurrent units such as GRUs.

3. Preliminary results

The model is run with following settings:

- Batch Size- 512 (training and validation)
- Hidden dimension- 512
- Epoch- 160

The train classification accuracy we got with Stacked attention was 62.24% and with one-glimpse was 61.57%.

The evaluation classification accuracy we got with Stacked attention was 57.35% and one-glimpse was 56.63%

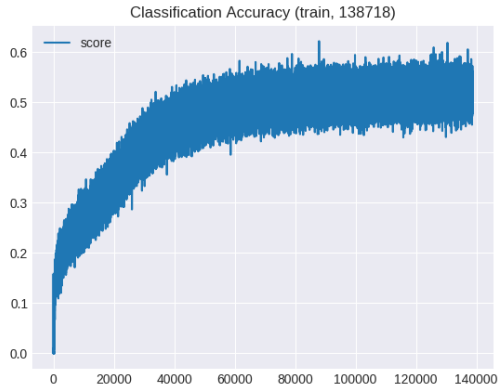


Fig 3: Train Classification Accuracy (Classification Accuracy vs number of epochs)

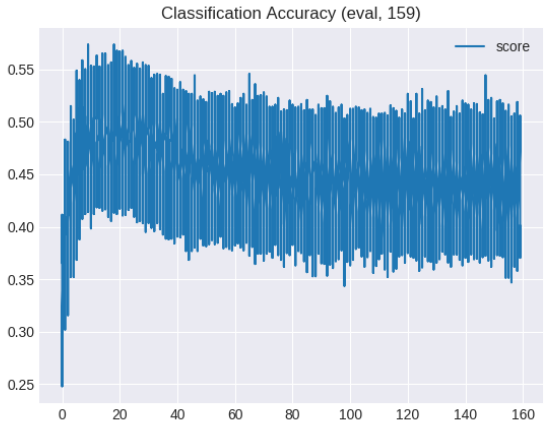


Fig 4: Evaluation Classification Accuracy (Classification Accuracy vs number of epochs)

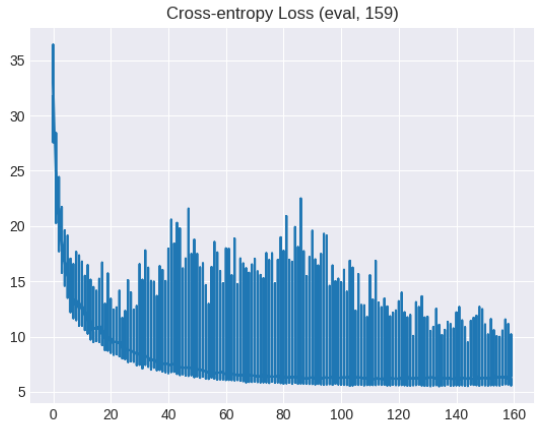


Fig 5: Evaluation Cross-entropy Loss (Cross Entropy loss vs number of epochs)

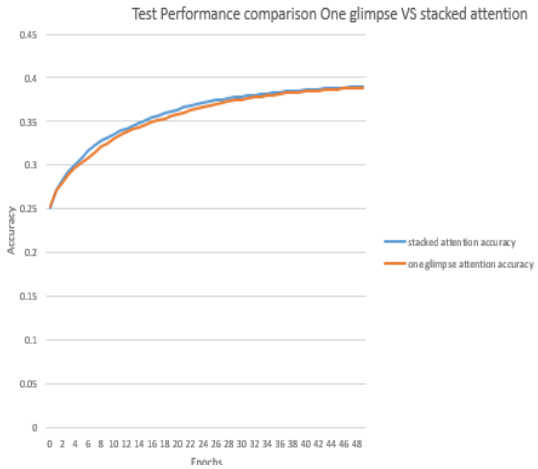


Fig 6: Stacked attention gives higher accuracy than one-glimpse in the initial epochs. This was tested on smaller dataset 13,000 images.

4. References:

[1]Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

[2]https://imagecaption.blob.core.windows.net/imagecaption/trainval_36.zip

[3]http://nlp.stanford.edu/data/glove.6B.zip

[4] https://arxiv.org/pdf/1511.02274v2.pdf