Where to Look and How to Answer: Using Regions and Answer Types

{ Ashwini Venkatesh and Surbhi Goel } The University of Texas at Austin

Introduction

We attempt to solve the task of Visual Question Answering. We propose an architecture where we leverage regions of the image along with the question. We also evaluate a modification to the output layer gradient updates which takes into account the type of the answer predicted and promotes correct answer types. We evaluate the system against the VQA existing baselines and other models on the VQA dataset.

Related Work

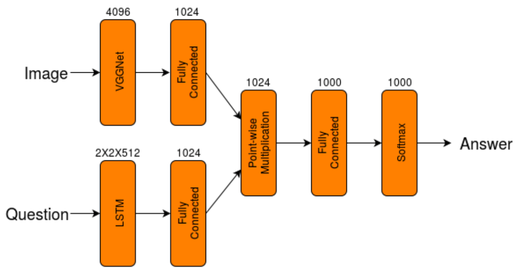

Figure 5: LSTM+CNN model from [1]

- Uses LSTM to capture structure of question
- Does not exploit image information as performances only marginally better with image input

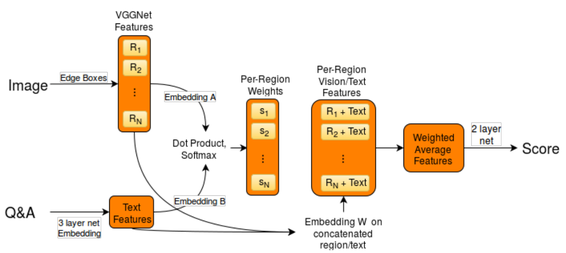

Figure 6: Attention-based model from [2]

- Uses Attention to focus on regions in the image relevant for answering the question
- Only for Multiple Choice questions: score generated per question-answer pair
- Uses Bag-of-Words approach to get text features

Answer Types

Figure 1: Types of answers represented as word clouds

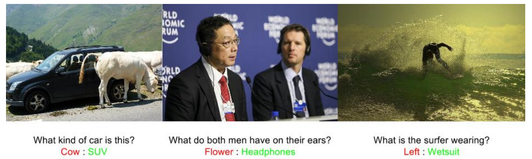

Figure 2: Qualitative results of updated gradient model

The new gradient update scales the existing gradient update depending on the type of the actual answer. It penalizes the right type of answer less harshly than the wrong types, to enable the model to learn the type of answers that a question fits.

$$\hat{\Delta}_i(x,y) = \begin{cases} \alpha\Delta_i(x,y), & \text{if } T(x_i) = T(y), x \neq y \\ \Delta_i(x,y), & \text{otherwise} \end{cases}$$

We experiment two approaches:

- Learn types in an unsupervised manner using clustering on word2vec embedding of the top 1000 answers
- Consider three main types: binary ('yes/no'), numeric ('one', 'two', ...) and other

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. *CoRR*, abs/1511.07394, 2015.

Attention Model

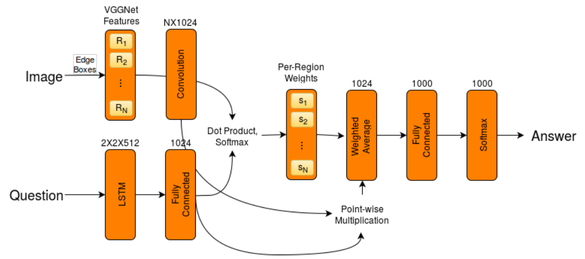

Figure 3: Proposed model with Attention

- Extract the top-ranked 3 Edge Boxes from the image after performing non-max suppression with a 0.2 intersection over union overlap criterion. Use these and the whole image itself as the input.
- Get per region weights by taking a dot product of question embedding and region embeddings and applying a softmax over it.
- Use region weights to take a weighted average of pointwise multiplication products of question and region embeddings followed by a FC layer with softmax to produce a distribution across top 1000 answers

Figure 4: Qualitative results of attention model

Results

	All	Y/N	Number	Other
[1]	54.19	79.77	33.42	40.26
3 Labels	54.24	79.75	33.17	40.45
W2V Labels	54.37	79.96	33.12	40.54

Table 1: Comparison of performance on different word clusters. 3 Labels uses Binary, Number and Other answer categories to update the gradients. W2V labels uses clusters obtained by running word2vec on top 1000 answers. This uses 17 labels to update the gradients.

	All	Y/N	Number	Other
[1]	54.19	79.77	33.42	40.26
[2]*	62.44	77.62	34.28	55.84
Ours	53.14	78.11	32.92	39.53

Table 2: Comparison of performance on VQA. * Evaluated on MCQ. Others are evaluated on Open Ended questions. Our model ran only 56000 iterations in comparison to 150000 iterations for the VQA model.

Challenges

- Restricted to train with low number of regions per image due to file number limit. Intended to train ≈ 100 regions per image
- Slow training hence unable to do parameter tuning

Future Work

- Use higher number of regions per image
- Use region specific captions using Dense captioning
- Try different approaches to concatenate region and questions embeddings. In this project we have experimented only with dot product of the vectors. We can experiment with concatenating or adding the vectors.