Visual Question Answering

Geraldine Versfeld, Yifan Ding, and Boqing Gong
Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA

Introduction

Visual Question Answering (VQA) is a multi-disciplinary problem that requires the intersection of many facets of machine learning such as natural language processing (NLP), computer vision, and transfer learning. Most previous works in VQA have used pre-trained networks to extract image and question features followed by a softmax classifier to predict an answer from a selection of popular answers[3]. Recent advances in VQA rely on soft attention mechanisms that selectively focus on image features, decreasing the impact of noise and unnecessary information[1,3]. The method suggested by Fukui et al. successfully adapted compact bilinear pooling and soft attention to win the 2016 VQA challenge[1]. Using the multimodal compact bilinear pooling (MCB) model as a baseline, we propose a part of speech (POS) based weighting scheme to increase the impact of more important question words and increase tolerance for different phrasings of the same query. In addition, we explore the use of different pre-trained networks for question and image feature extraction.

VQA Dataset

- 204,721 images, 1,105,904 questions, 2,581,110 annotations[3] .
- derived from MSCOCO[5].
- each question has 10 hand-made annotations.
- VQA-v2 balanced by including an additional image for each question for which the answer was different[3].



What color are her eyes?
What is the mustache made of?

Method

Using the pre-trained MCB model as a baseline, we fine-tuned the model on the VQA-v2 dataset and used different word and image embeddings to explore their role in question answering. In addition, we propose using a Part of Speech based weighting system to attend important question words and a regression based answer prediction scheme.

- Pre-trained MCB baseline was fine-tuned for 200,000 epochs using an adagrad solver and a decreasing learning rate (0.000001-0.0000001).
- POS weights were applied to the question embedding and the model was further fine-tuned for 120,000 epochs using an adagrad solver and a learning rate of 0.00000001.
- Image features are extracted using different pre-trained models including ResNet, Inception-v3, and VGG19.
- Question features extracted using Glove and word2vec
- The Compact Bilinear Pooling algorithm[1,2]:

Algorithm 1 Multimodal Compact Bilinear

input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
procedure MCB(v_1, v_2, n_1, n_2, d)
  for $k \leftarrow 1 \dots 2$ **do**
    if h_k, s_k not initialized **then**
      for $i \leftarrow 1 \dots n_k$ **do**
        sample $h_k[i]$ from $\{1, \dots, d\}$
        sample $s_k[i]$ from $\{-1, 1\}$
    $v'_k = \Psi(v_k, h_k, s_k, n_k)$
  $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
  return Φ
procedure $\Psi(v, h, s, n)$
  $y = [0, \dots, 0]$
  for $i \leftarrow 1 \dots n$ **do**
    $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
  return y

Table 1: Part of Speech Weights. Applied with Spacy POS tagger.

Part of Speech	Weight
Verb	1.4
Noun	1.4
Determiner	0.7
Adverb	1.4
Adjective	1.3
Proper Noun	1.3
Pronoun	1.4
Participle	0.9
Number	1.3
Interjection	0.8
Coordinating conjunction	0.8
Auxiliary	0.8
Adposition	0.8

Method cont'd

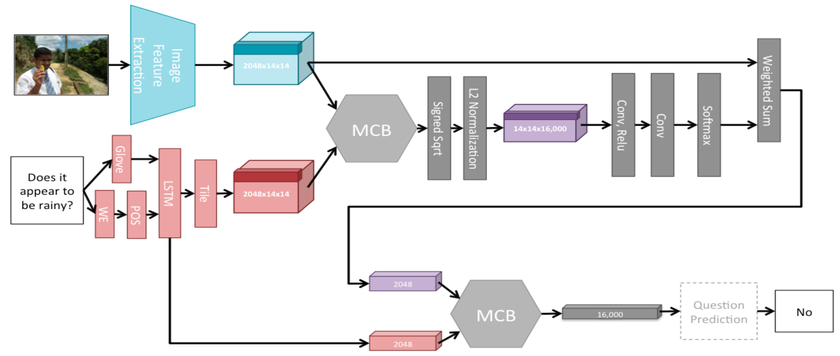


Figure 1: Our VQA Model. Basic structure of MCB baseline[3] with addition of POS question weighting and regression for question prediction.

Quantitative Results

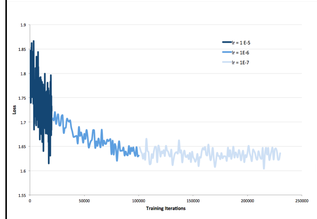


Figure 3.a: Fine-tuning training loss.

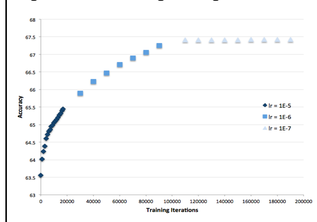


Figure 3.b: Fine-tuning validation accuracy.

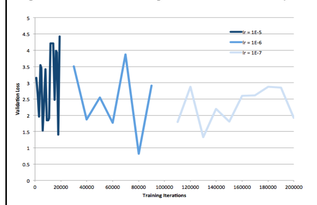


Figure 3.c: Fine-tuning loss.

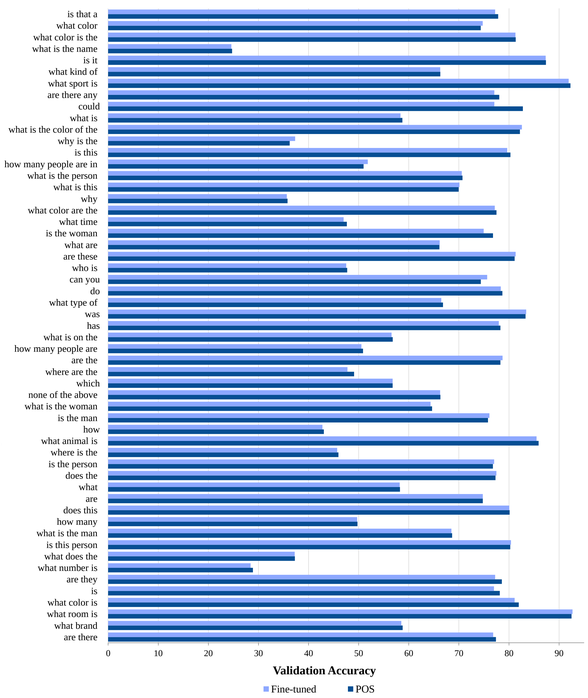


Figure 4: Validation Accuracy by Question Type.

Table 2: Accuracy of models on Validation, Test-dev, and Test Standard Data Splits.

	Validation				Test-dev				Test-standard			
	Y/N	No.	Other	All	Y/N	No.	Other	All	Y/N	No.	Other	All
Pre-trained	73.01	43.77	61.62	63.56	69.19	36.55	53.20	57.90	–	–	–	–
Fine-tuned	78.61	46.43	64.75	67.55	72.20	36.35	53.85	59.42	–	–	–	–
POS	78.97	46.54	64.89	67.77	73.89	36.83	54.85	60.65	74.08	36.43	54.84	60.81
Ensemble	–	–	–	–	74.43	37.25	55.27	61.11	–	–	–	–

Qualitative Results

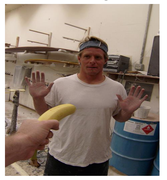
Figure 2.a: Examples of correctly predicted answers.

Question: What is the woman holding over her head?



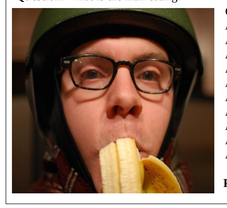
Ground truth answers:
Answer 1: bananas
Answer 2: bowl of bananas
Answer 3: plate of bananas
Answer 4: fruit
Answer 5: bananas
Answer 6: fruit
Answer 7: bananas
Answer 8: fruit
Answer 9: serving platter of bananas
Answer 10: plate
Predicted answer: banana

Question: What is in the person's hand?



Ground truth answers:
Answer 1: banana
Answer 2: banana
Answer 3: banana
Answer 4: banana
Answer 5: banana
Answer 6: banana
Answer 7: banana
Answer 8: banana
Answer 9: banana
Answer 10: banana
Predicted answer: banana

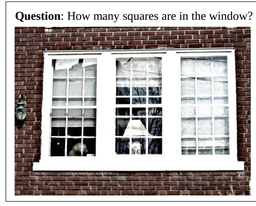
Question: What is the man eating?



Ground truth answers:
Answer 1: banana
Answer 2: banana
Answer 3: banana
Answer 4: banana
Answer 5: yes
Answer 6: banana
Answer 7: banana
Answer 8: banana
Answer 9: banana
Answer 10: banana
Predicted answer: banana

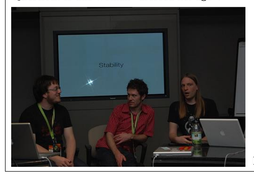
Figure 2.a: Examples of incorrectly predicted answers. Some inaccuracy could be as a result of poorly framed or ambiguous questions. Additionally, some questions could not be reasonably answered based on the given image.

Question: How many squares are in the window?



Ground truth answers:
Answer 1: 24
Answer 2: 27
Answer 3: 18
Answer 4: 15
Answer 5: 15
Answer 6: many
Answer 7: 15
Answer 8: many
Answer 9: 45
Answer 10: 45
Predicted answer: 24

Question: Is this woman "dressed her age"?



Ground truth answers:
Answer 1: yes
Answer 2: yes
Answer 3: yes
Answer 4: yes
Answer 5: na
Answer 6: no
Answer 7: yes
Answer 8: yes
Answer 9: yes
Answer 10: yes
Predicted answer: no

Question: Is all the fruit citrus?



Ground truth answers:
Answer 1: no
Answer 2: no
Answer 3: no
Answer 4: no
Answer 5: no
Answer 6: no
Answer 7: yes
Answer 8: no
Answer 9: no
Answer 10: no
Predicted answer: yes

Discussion & Conclusions

- The POS weighted model resulted in a moderate improvement in performance across all answer types. Although POS weighting appears to have improved performance in simpler question types, in more complex question types that require logic or counting the improvement was quite small or or slight worsening in accuracy was observed.
- The example question and answer pairs highlighted some of the challenges in the VQA dataset. Some questions are phrased ambiguously resulting in a lack of consensus among the ground truth answers. Additionally, some questions have no visual grounding in the image about which they are being asked.
- In future, low-scoring question types such as "how many" or "why" could be addressed by creating a model more capable of basic reasoning and counting.
- The lower accuracy of "what is the name" type questions could be because the answers to these questions are proper nouns which may not be included in the subset of popular answers from which the answer is predicted. This could be addressed in future by using a regression based answer generator instead of a softmax classifier to predict answers.

References

1. Fukui, Akira and Park, Dong Huk and Yang, Daylen and Rohrbach, Anna and Darrell, Trevor and Rohrbach, Marcus. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. arXiv:1606.01847. September, 2016.
2. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra,Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. arXiv:1612.00837 . May 2017
3. Yang Gao, Oscar Beijbom, Ning Zhang, Trevor Darrell. Compact Bilinear Pooling. arXiv:1511.06062. April, 2016
4. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. VQA: Visual Question Answering. arXiv:1505.00468. May, 2015.

Acknowledgements

This work was supported by NSF grant number 100461121.