Multimodal Explanations: Justifying Decisions and Pointing to the Evidence (Supplementary Material)

Dong Huk Park[1], Lisa Anne Hendricks[1], Zeynep Akata[2,3], Anna Rohrbach[1,3],
Bernt Schiele[3], Trevor Darrell[1], and Marcus Rohrbach[4]

[1]EECS, UC Berkeley, [2]University of Amsterdam, [3]MPI for Informatics, [4]Facebook AI Research

1. Visual Question Answering Model

Section 1.1 extends section 4 in the main paper, detailing the VQA answering model of PJ-X by providing the detailed formulas we omitted for brevity. The VQA model that we use throughout the experiments is based on the state-of-the-art MCB model [2], but trains and evaluates faster (reduction of $\sim 30\%$). The main difference between the two models is how they combine two different representations and create multimodal features. We evaluate our VQA model using the same accuracy measure as in the VQA challenge in Section 1.2.

1.1. Model Details

For spatial image features $f^I(I,n,m)$ from the last convolutional layer of ResNet-152, question Q, 2-layer $LSTM$ $f^Q(Q)$ we compute:

$$\begin{aligned}
\bar{f}^{IQ}(I,n,m,Q) =& (W_1 f^I(I,n,m) + b_1) \odot f^Q(Q) && (1)\\
f^{IQ}(I,Q) =& L2(\textit{signed_sqrt}(\bar{f}^{IQ}(I,Q))) && (2)\\
\bar{\alpha}^{pointA}_{n,m} =& f^{pointA}(I,n,m,Q) && (3)\\
=& W_3 \rho(W_2 f^{IQ}(I,Q) + b_2) + b_3 && (4)
\end{aligned}$$

with ReLU $\rho(x) = max(x,0)$. This process gives us a $N \times M$ attention map $\bar{\alpha}_{n,m}$. We apply softmax to produce a normalized soft attention map for predicting the answer:

$$\alpha^{pointA}_{n,m} = \frac{\exp(\bar{\alpha}^{pointA}_{n,m})}{\sum_{i=1}^{N}\sum_{j=1}^{M}\exp(\bar{\alpha}^{pointA}_{i,j})} \tag{5}$$

$$\begin{aligned}
\bar{f}^{y}(I,Q) =& (\sum_{x=1}^{N}\sum_{y=1}^{M} \alpha^{pointA}_{n,m} f^I(I,n,m)) \odot f^Q(Q) && (6)\\
f^y(I,Q) =& W_4 \bar{f}^y(I,Q) + b_4 && (7)\\
p(y|I,Q) =& Softmax(f^y(I,Q)) && (8)\\
\hat{y} =& \operatorname*{argmax}_{y\in Y} p(y|I,Q) && (9)
\end{aligned}$$

1.2. Results

As shown in Table 1, our VQA model leads to a considerable improvement in performance, especially in VQA v2. Our model shows an overall 0.51% improvement on VQA v1 and 2.57% on VQA v2, where the most significant gain is achieved in the "Number" category. In addition, our model is more efficient than MCB [2] in that the training and inference are $\sim 30\%$ faster.

Method	VQA v1 All	Yes/No	Number	Other	VQA v2 All	Yes/No	Number	Other
MCB [2]	62.50	79.69	31.29	**54.67**	59.14	77.37	36.66	51.23
Our VQA model	**63.01**	**82.01**	**35.47**	52.99	**61.71**	**78.68**	**38.51**	**52.53**

Table 1: OpenEnded results on VQA v1 [1] and VQA v2 [3] datasets. The models are trained on train set, validated on validation set, and reported on the test-dev accuracies. The columns indicate the accuracies of the model for each different question type. Our model achieves higher accuracy than MCB [2] while being faster at train and test time.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.