# Deep Modular Co-Attention Networks for Visual Question Answering — Supplementary Material

Zhou Yu[1] Jun Yu[1]* Yuhao Cui[1] Dacheng Tao[2] Qi Tian[3]

[1]Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, China.

[2]UBTECH Sydney AI Centre, School of Computer Science, FEIT, University of Sydney, Australia

[3]Noah's Ark Lab, Huawei, China

{yuz, yujun, cuiyh}@hdu.edu.cn, dacheng.tao@sydney.edu.au, tian.qi1@huawei.com

Table 1: Accuracies of **model ensembling** on the *test-standard* split to compare with the best solutions in VQA-Challenge 2018. $R$ denotes the ranking of the corresponding team. # denotes the number of used models for ensembling.

| $R$ | Team Name | # | All | Y/N | Num | Other |
|---|---|---|---|---|---|---|
| 5 | MIL-UT | - | 71.16 | 87.00 | 52.6 | 61.62 |
| 4 | CASIA-IVA | - | 71.31 | 86.98 | 51.05 | 62.31 |
| 3 | SNU-BI | 15 | 71.84 | 87.22 | 54.37 | 62.45 |
| 2 | HDU-UCAS-USYD | 12 | 72.09 | 87.61 | 51.92 | 63.19 |
| 1 | FAIR A-STAR | 30 | 72.25 | 87.82 | 51.59 | **63.43** |
| | MCAN (Ours) | 4 | **72.45** | **88.29** | **54.38** | 62.80 |

## A. Model Ensembling

To compare MCAN to the best results on VQA-v2 leaderboard[1], we train 4 MCAN$_{\text{ed}}$-6 models with slightly different hyper-parameters for ensemble. The comparative results in Table 1 indicate that MCAN surpasses the top most solutions on the leaderboard. It is worth noting that our solution only use the basic bottom-up attention visual features [1] and much fewer models for ensemble.

## B. Comparisons of Model Stability and Computational Costs

We compare MCAN$_{\text{ed}}$-6 with the best two approaches (MFH [4] and BAN-8 [3]) in Table 2 in terms of overall accuracy ±std, number of parameters and FLOPs, respectively. The accuracies are reported on the *val* split, and the standard deviation for each method is calculated by training three models with the same architecture but different initializations. The FLOPs are calculated for one testing sample. We can see that MCAN$_{\text{ed}}$-6 outperforms the counterparts in both accuracy and stability, and is more parameteric- and computational-efficient at the same time.

Table 2: Comparison of model stability and computational costs to the state-of-the-art on *val* split of VQA-v2.

| | MFH [4] | BAN-8 [3] | MCAN$_{\text{ed}}$-6 |
|---|---|---|---|
| Acc. ± std. (%) | 65.65±0.05 | 66.04±0.08 | 67.23±0.01 |
| #Params ($\times 10^6$) | 116 | 79 | 56 |
| FLOPs ($\times 10^9$) | 4.4 | 3.3 | 2.8 |

## C. More Visualized Results

Similar to Figure 7 in the main text, We visualize the learned attentions of two more examples from MCAN$_{\text{ed}}$-6 in Figure 1. For each example, we visualize the attention maps from three attention units (SA(X), SA(Y), GA(X,Y)) and from two layers (1st and 6th). For each unit, we show the attention maps from 2 parallel heads (8 heads in total). From the results, we have the similar observations and explanations to those in the main text. The visualized attentions can well explain the reasoning process of MCAN to predict the correct answers. Furthermore, we find that different heads may provide complementary information to benefit VQA performance, which is similar to the 'multi-glimpses' strategy in existing VQA approaches [2, 4].

## References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 1

[2] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact

*Jun Yu is the corresponding author

[1]https://visualqa.org/roe.html

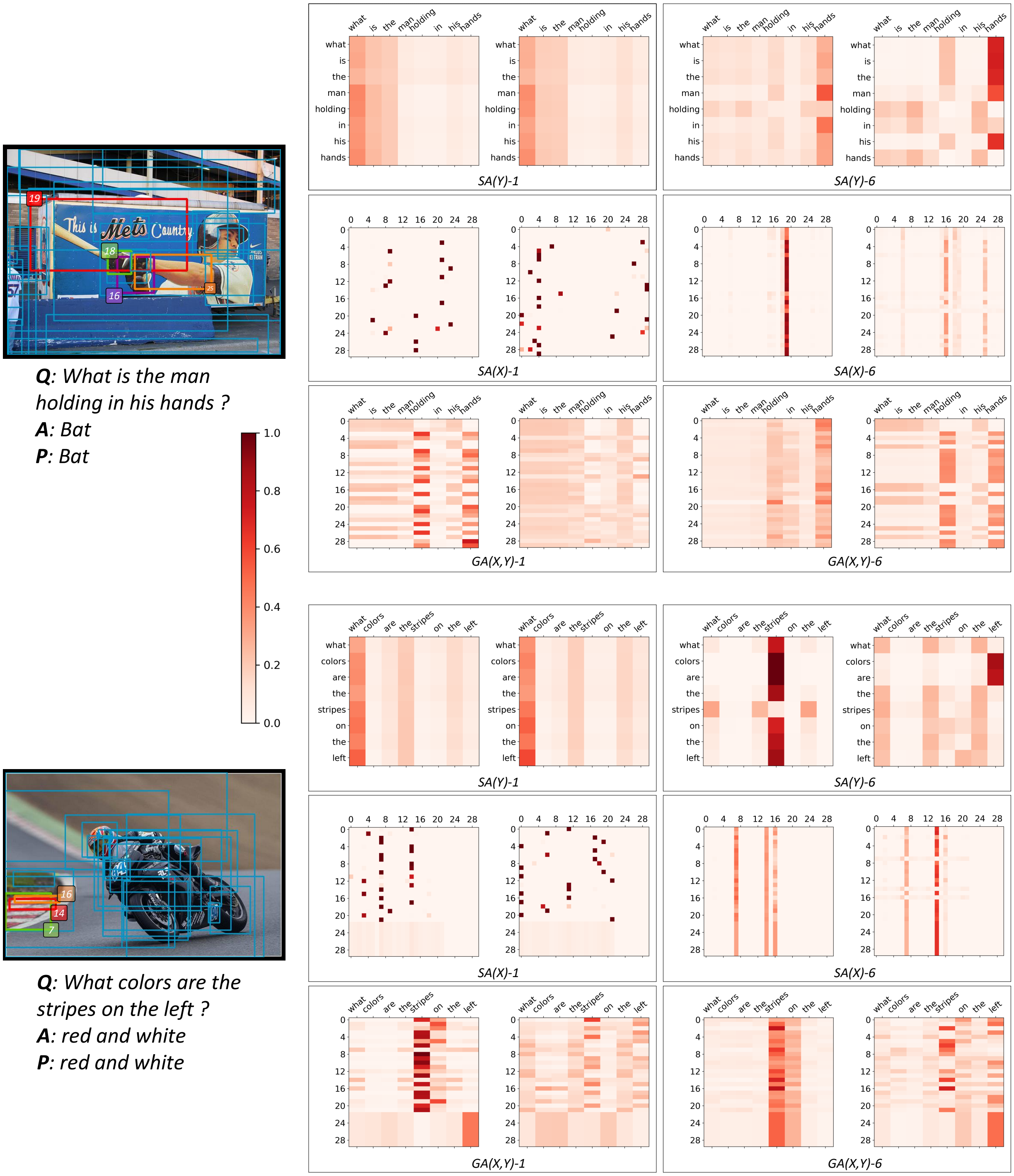


Figure 1: Two examples of the learned attention maps from typical attention units and layers. For each attention unit (within the box), we show two attention maps from different heads.

bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[3] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems (NIPS)*, 2018. 1

[4] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018. 1