Explanation vs Attention: A Two Player Game to obtain Attention for VQA

First Author[1*], **Second Author**[2], **Third Author**[2,3] and **Fourth Author**[4]
[1]First Affiliation
[2]Second Affiliation
[3]Third Affiliation
[4]Fourth Affiliation
{first, second}@example.com, third@other.example.com, fourth@example.com

Abstract

1 Introduction

This is the supplementary material for the paper 'Explanation vs Attention: A Two Player Game to obtain Attention for VQA'. We provide relevant theoretical background, additional qualitative results, and detailed ablation analysis for better understanding of our paper.

2 Experiment

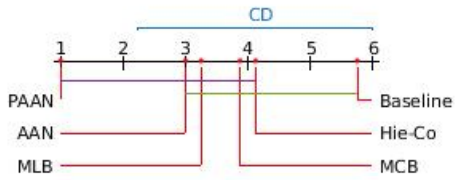

Figure 1: The mean rank of all the models on the basis of all scores are plotted on the x- axis. CD = 3.7702, p = 0.0015655. Here Joint refers to our CDM model and others are the different variations described in section 4.3 The colored lines between the two models represents that these models are not significantly different from each other.

2.1 Statistical Significance Analysis

We analyze Statistical Significance [Demšar, 2006] of our PAAN model against the variants mentioned in section 4.3 of our method as well as different methods for explanation generation. The Critical Difference (CD) for Nemenyi [Fišer *et al.*, 2016] test depends on given α (confidence level, which is 0.05 in our case) for average ranks and N(number of tested datasets). Low difference in ranks for two model implies that they are significantly less different. Otherwise, they are statistically different. Figure 1 visualizes the post hoc analysis using the CD diagram. It is clear that PAAN works best and is significantly different from other methods.

[*]Contact Author

2.2 Evaluation methods

This answer can be evaluated using accuracy metric provided by [Antol *et al.*, 2015] as follows :

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\min\left(\frac{\sum_{t\in T^i} I[a_i = t]}{3}, 1\right) \quad (1)$$

where a_i is the predicted answer and t is the annotated answer in the target answer set T^i of the i^{th} example and $I[.]$ is the indicator function. The predicted answer a_i is correct if at least 3 annotators agree on the predicted answer.

2.3 Attention Visualization

Attention visualization means displaying the attention probability distribution matrix which is the most prominent part of a given image based on the query question. In this section, we visualize the attention of different layers of the stacked attention network. The size of attention probability distribution is 14×14 and the size of preprocessed image from COCO-QA is 448×448. In order to visualize the attention, we need to make attention probability distribution same size as given COCO-QA Image. To do this, first scale the attention probability distribution to 448×448 using bi-cubic method. Then, apply Gaussian filter of size 31×31 with mean 0 and variance 1 and finally multiply or mask attention probability distribution on original image. Also, we visualize the MAN for different category like Object, Numbers, Color and Location. For each type we provide two example along with its supporting and contrasting attention distribution. We follow the sequence like original image, attention probability of first stack and then attention probability of second stack.

2.4 VQA Dataset

We have conducted our experiments on VQA benchmark VQA-v1[Antol *et al.*, 2015] dataset, which contains human annotated question and answer based on images on MS-COCO dataset. This dataset contains 204721 images in total, out of which 82783 images are for training, 40504 images for validation and 81434 images for testing. Each image is associated with 3 questions and each question has 10 possible answer. There are 248349 Question-Answer pairs for training, 121512 pairs for validation and 244302 pairs for testing.

2.5 Qualitative Result

We provide attention map visualization of all models for 3 example images as shown in Figure 2. We can vividly see how attention is improving as we go from our baseline model (SAN) to the proposed adversarial model (PAAN). For example, in the second row, SAN is not able to focus on any specific portion of the image but as we go towards right, it is able to focus near the bus. Same can be seen for other images also.

We have also visualized Grad-CAM maps for the same images to corroborate our hypothesis that Grad-CAM is better way of visualization of network learning as it can focus on correct portions of the image even in our base line model (SAN), hence, can be used as a tutor to improve attention maps. Also, Grad-CAM is simultaneously improving according to our assumption and can also be seen in the Figure 2. For example, in SAN it tries to focus on correct portions but only along with other points of focus too. But in our proposed model, visualization is improved to focus only on required portion.

2.6 Visual Dialog Task

We have conducted same experiment for Visual dialog task, which is introduced by [Das *et al.*, 2017]. The visual dialog task is defined as :
Given Image I, a caption C, a dialog history H till $t-1$ rounds i.e. $H = \{H_0 = C, H_1 = (q_1, a_1), ...H_{t-1} = (q_{t-1}, a_{t-1})$ and the following question q_t at round t. The objective of visual dialog agent is to predict a natural language answer to the question q_t.
We conduct an experiment to visualize the attention provided by the Q-Bot and A-Bot. It is observed that the attention mask provide by the Q and A Bot is not good as compared to the Grad-CAM mask in each turn of the dialog as shown in first row of Figure 3. Our objective is to improve Attention mask in the visual dialog. In order to tackle this problem, we used same technique(PAAN) to improved attention mask in VQA as like in adversarial game section. The attention and Grad-CAM is improved as shown in the second row of the Figure 3.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[Fišer *et al.*, 2016] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. Janes v0. 4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina*, 2(4):2, 2016.

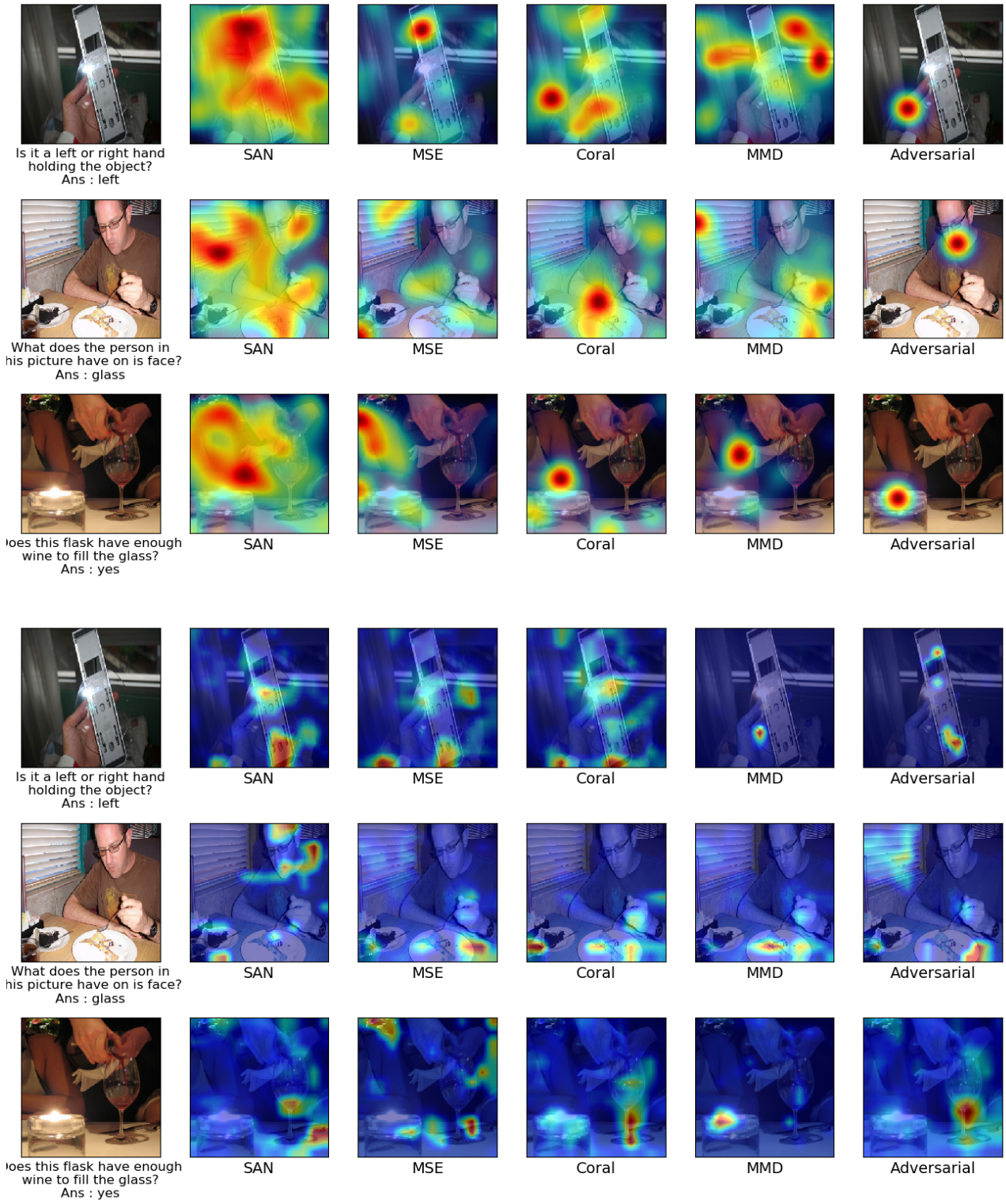

Figure 2: Examples with different approaches in each column for improving attention using explanation in a self supervised manner. The first column indicates the given target image and its question and answer. Starting from second column, it indicates the Attention map(above 3) / Grad-CAM map(below 3) for Stacked Attention Network, MSE based approach, Coral based approach, MMD based approach, Adversarial based approach respectively.

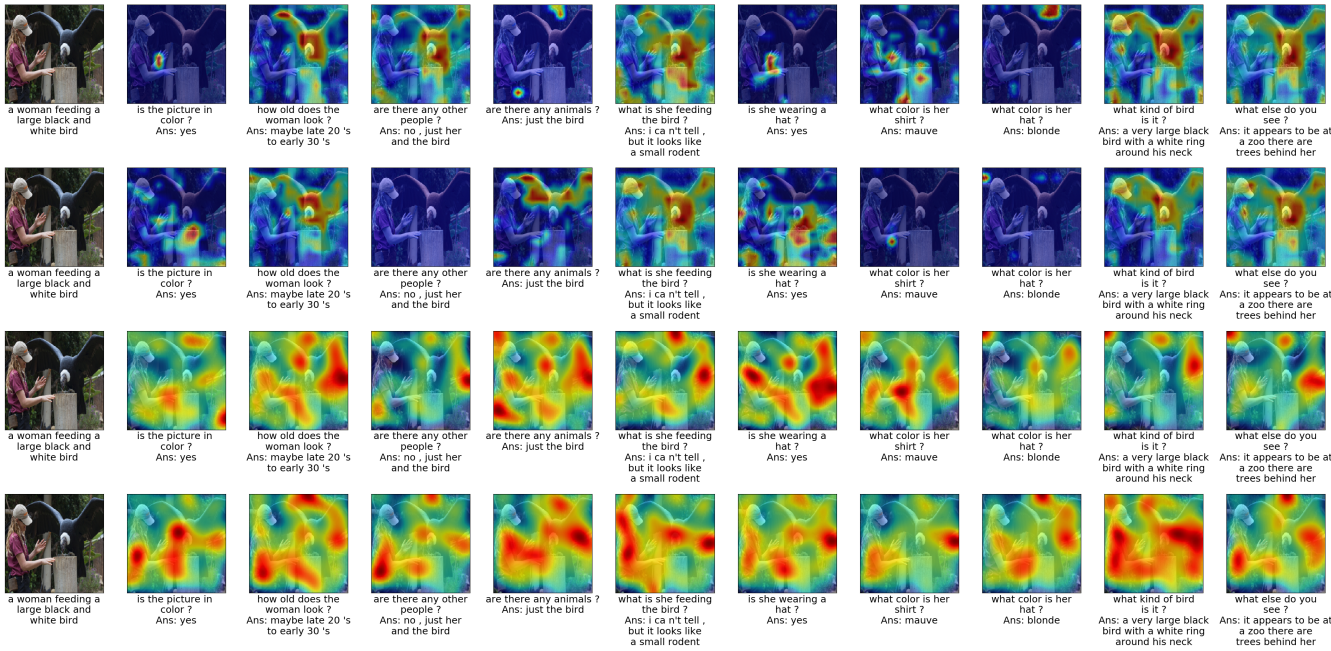

Figure 3: This figure shows visual explanation and attention map for Visual Dialog task. The first row contains grad cam results and second contains attention results for Adversarial approaches. We have shown all the dialog turns present in the dataset.