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

In this paper we aim to obtain improved attention for visual question answering (VQA) task. It is challenging to provide supervision for attention. An observation we make is that visual explanations as obtained through class activation mappings (specifically Grad-CAM) that are meant to explain the performance of various networks could form a means of supervision. However, as the distributions of attention maps and that of Grad-CAMs differ, it would not be suitable to directly use these as a form of supervision. Rather, we propose the use of a discriminator that aims to distinguish samples of visual explanation and attention maps. The use of adversarial training of the attention regions as a two player game between attention and explanation serves to bring the distributions of attention maps and visual explanations closer. Significantly, we observe that providing such a means of supervision also results in attention maps that are more closely related to human attention resulting in substantial improvement over baseline stacked attention network (SAN) models that are closely related and also result in a good improvement in rank correlation metric on the VQA task. This method can also be combined with recent MCB based methods and result in consistent improvement. We also provide comparisons with other means for learning distributions such as based on Correlation Alignment (Coral), Maximum Mean Discrepancy (MMD) and Mean Square Error (MSE) losses and observe that the adversarial loss outperforms the other forms of learning the attention maps. Visualization of the results also confirms our hypothesis that attention maps improve using this form of supervision.

1 Introduction

When asked a question based on an image, a human invariably focuses on the part of the image that aids in answering the question. This fact is commonly known in cognitive science and an extreme example that depicts perceptual blindness was demonstrated by [Simons and Chabris, 1999] where two groups of participants are passing balls. When asked to count the balls, viewers ignore a gorilla in the video as it is not pertinent to the task of counting. However, the deep networks that solve for semantic tasks such as visual question answering do not have such attentive mechanisms. The fact that the existing deep networks do not attend to the areas that humans do was shown by the work of [Das *et al.*, 2016]. While there have been some works that aim to improve the attended regions, it is challenging as obtaining supervision for attention is tedious and may not always be possible for all the semantic tasks that we would like to use deep networks. In this paper we propose a simple method to obtain self-supervision to guide attention as shown figure- 1.

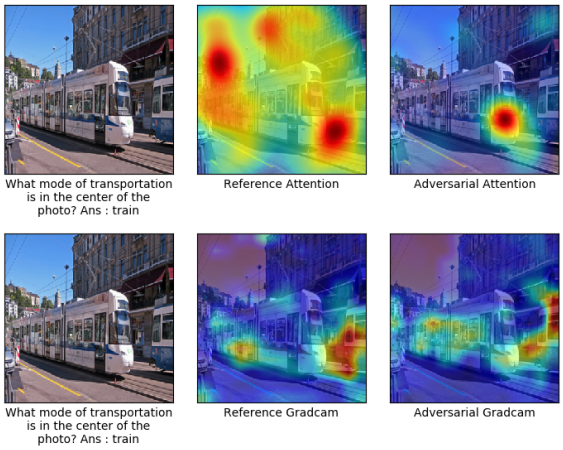

Figure 1: First row shows improvement in attention mask as compared to reference attention mask (SAN). Second row shows improvement in explanation mask (Grad-CAM) as compared to reference explanation mask. Last column shows improvement in both.

The main idea is that given the task of solving visual question answering (VQA), there exists methods based on obtaining visual explanations such as Grad-CAM [Selvaraju *et al.*,] that obtain class activation mappings from gradients that allow us to understand the areas that a network focuses while solving the task for the correct class label. As during training, class labels are available for the VQA task, it is easy to obtain such supervision. Using this it is possible to obtain surrogate supervision for supervising attention. Using the ground-truth label, one can obtain the visual explanation for a deep network that solves the visual question answering task. As the

*Contact Author

network is provided the actual label, the corresponding activation maps do aid in solving the task. Therefore, we hypothesize that this supervision can aid in obtaining better attention maps and this is evident from the results that we obtain .

The next challenge is to consider how the surrogate supervision obtained from Grad-CAM can be used to obtain better attention regions. Directly using these as supervision is not optimal as the distributions for the visual explanation differs from that of the attention maps as the attention maps are also supervised by the task loss. We show that just using the mean-square error loss for the two maps is sub-optimal. We show in this paper that a very simple way of using a two-player game between a discriminator that tries to discriminate between Grad-CAM results and attention maps and a generator that generates attention maps serves to obtain substantially improved attention maps. We show that this method performs much better and also provides state of the art results in terms of attention maps that correlates well with human attention maps. To summarize, through this paper we provide the following contributions :

- We propose a means for obtaining surrogate supervision for obtaining better attention maps based on visual explanation in the form of Grad-CAM results.
- We show that this surrogate supervision can be best used through a variant of adversarial learning to obtain attention maps that correlate well with visual explanation. Further, we observe that this performs better as against other means of supervision such as MMD [Tzeng *et al.*, 2014] or Coral [Sun and Saenko, 2016] losses.
- We provide various comparisons and results to show that we obtain better attention maps that correlate well with human attention maps and outperform other techniques for VQA. Further, we show that obtaining better attention maps also aids in obtaining better accuracies while solving for VQA. Detailed empirical analysis for the same is provided.

1.1 Motivation

In VQA, given an image & query, the attention model aims to learn the regions in an image pertinent to the answer. [Das *et al.*, 2016] has proposed Human Attention (HAT) dataset for VQA task where human annotators have annotated the regions attended in the image to mark its answer based on the question. The regions pointed by humans for answering visual question are more accurate as compared to machine pointed regions. This can be concluded through an experiment on HAT dataset where we replace human attention with attention obtained using stacked attention network with one stack. We observe that the prediction accuracy for original stacked attention network is 52.26% and 68.17% with ground truth human attention. We believe that human attention cannot be directly used as supervision as there are not enough examples of human attention (58K/215K). Further, such a method would not generalize to novel tasks. However, we are motivated by this result and have therefore developed self-supervision based method to improve attention. Given an image & its question, we obtain attention and gradient mask of the answer class. We play an adversarial game between these two players viz., attention and gradient. The aim of the game is to improve attention mask based on its gradient mask. We provide final result of this game mechanism as shown in the figure - 1. To ensure whether our approach is prudent we evaluate whether using grad-CAM as self supervision is beneficial. We do this by an experiment that replaces attention mask with Grad-CAM mask and we observed that the classification accuracy of the VQA (stacked attention network) model increases by 4.2%,i.e, from 52.05% to 56.26% on validation set. This provides a strong intuition to consider using Grad-CAM as self-supervision for the attention module.

2 Related work

Visual question answering was first proposed by [Malinowski and Fritz, 2014]. Subsequently, [Geman *et al.*, 2015] proposed a "visual Turing test" where a binary question is generated from a given test image. This is in contrast to modern approaches in which model is trying to answer free-form open-ended questions. A seminal contribution here has been standardizing the dataset used for Visual Question Answering [Antol *et al.*, 2015]. The methods for VQA can be categorized into joint embedding approaches and attention based approaches. Joint embedding based approaches have been proposed by [Antol *et al.*, 2015; Ren *et al.*, 2015; Goyal *et al.*, 2017a] where visual feature is combined with question feature to predict the answer. We are in this paper more interested in attention based methods.

There has been significant interest in including attention to solve the VQA problem. Attention based models comprises of image based, question based and some that are both image and question based attention. In image based attention approach, the aim is to use the question in order to focus attention over specific regions in an image [Shih *et al.*, 2016]. An interesting recent work [Yang *et al.*, 2016] has shown that it is possible to repeatedly obtain attention by using stacked attention over an image based on the question. Our work is closely related to this work. There has been further work [Li and Jia, 2016] that considers a region based attention model over images. The image based attention has allowed systematic comparison of various methods as well as enabled analysis of the correlation with human attention models as shown by [Das *et al.*, 2016]. In our approach, we focus on improving image based attention using adversarial game between Grad-CAM and attention mask, and show that it correlates better with human attention. There has been a number of interesting works on question based attention as well. [Zhu *et al.*, 2016][Xu and Saenko, 2016]. Work that explores joint image and question includes that based on hierarchical co-attention [Lu *et al.*, 2016]. There has been an interesting work by [Fukui *et al.*, 2016; Kim *et al.*, 2017; Kim *et al.*, 2018] that advocates multimodal pooling and obtains close to state of the art in VQA. Interestingly, we show that by combining it with the proposed method further improves our results.

3 Method

The main focus in our approach for solving visual question answering (VQA) is to use supervision obtained from visual

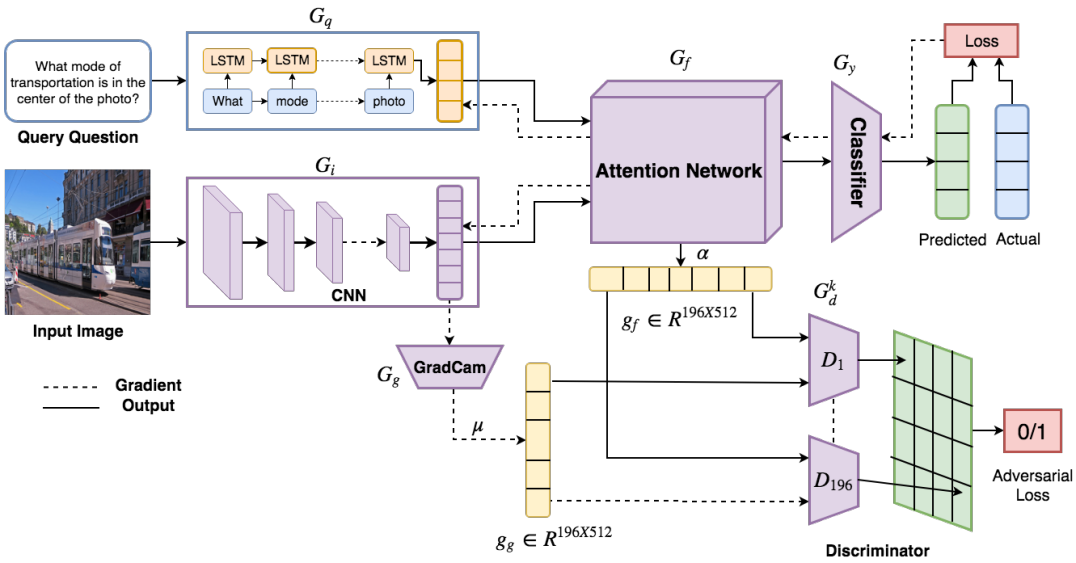

Figure 2: Illustration of model PAAN and its attention mask. Image feature and question feature is obtained using CNN and LSTM respectively. Attention mask is then obtained using these features and classification of the answer is done based on the attended feature. We have improved the attention mask with the visual explanation approaches based on Grad-CAM

explanation methods such as Grad-CAM to improve attention. As mentioned earlier, using Grad-CAM as attention shows improved performance in comparison to just using attention alone. Therefore, we believe that Grad-CAM or any other visual explanation method can be used in this setting. Further, by learning both visual explanation and attention jointly in an adversarial setting we observe improvements in both as shown empirically.

The key differences in our architecture as compared to an existing VQA architecture is the use of visual explanation and attention blocks in an adversarial setting. This is illustrated in figure 2. The other aspects of VQA are retained as is. In particular, we adopt a classification based approach for solving VQA where an image embedding is combined with the question embedding to solve for the answer. This is done using a softmax function in a multiple choice setting: $\hat{A} = \underset{A \in \Omega}{argmax} P(A|I, Q, \theta)$, where Ω is a set of all possible answers and θ represents the parameters in the network.

3.1 Overview

The three main parts of our method as illustrated in figure 2 are:

- **Attention representation**: We obtain attention embedding g_f by attending with a query question embedding g_q on an input image embedding feature g_i. This attention embedding helps to predict the answer for the query.
- **Explanation representation**: We obtain Grad-CAM feature for the attended image which visually explains the important regions responsible for the the predicted answer based on the query question.
- **Adversarial Game:** We formulate a game between Attention vs Explanation using adversarial mechanism. Through this game, we observe that we obtained improved attention regions which lead to improved prediction and therefore also results in better regions obtained through visual explanation. Thus, improving attention using Grad-CAM results in an improvement in Grad-CAM too.

Attention Representation

Initially, we obtain an embedding g_i for an image X_i using a convolution neural network(CNN). Similarly, we obtain a question feature embedding g_q for the query question X_Q using an LSTM network. These are input to an attention network that combines the image and question embeddings using a weighted softmax function and produces a weighted output attention vector g_f. There are various ways for modeling the attention network. In this paper, we have evaluated the network proposed in SAN [Yang *et al.*, 2016] and MCB [Fukui *et al.*, 2016].

Explanation Representation

One of the ways for understanding a result obtained by a deep network is to use visualization strategies. One such strategy that has gained acceptance in the community is based on Grad-CAM [Selvaraju *et al.*,]. Grad-CAM uses the gradient information of the last convolutional layer to visualize the contribution of each pixel in predicting the results. Note that Grad-CAM uses ground-truth class information and finds the gradient of the score for a class c in a convolution layer. It averages the gradient values to find the averaged μ values for each of the channels of the layer. We follow this approach and further details are provided in [Selvaraju *et al.*,].

Adversarial Game

A zero-sum adversarial game between two players is used with one set of players being Generator network and the other a set of Discriminator network. The aim is to obtain a Nash equilibrium where the Discriminator is unable to distinguish the generations of the Generator network from the 'real' distribution. In our case, the attention network is the Generator

network and the 'real' distribution is the output of Grad-CAM network. We term the resultant network as 'Adversarial Attention Network' (AAN). Specifically, the discriminator is a set of CNN layers followed by linear layer that uses a binary cross entropy loss function.

$$\min_G \max_D L_1(G,D) = E_{g_{g_i}\sim G_g(x_i)}[\log D(g_{g_i}/x_i)] + E_{g_{f_i}\sim G_f(x_i)}[\log(1 - D(G(g_{f_i}/x_i)))] \quad (1)$$

Here g_{g_i} is the output of grad-cam network G_g for a sample, x_i and g_{f_i} is the output of the attention network. In case we have access to ground-truth attention obtained from humans, we can directly use this in our framework. Here, we assume that we do not have access to such ground-truth as it is challenging to obtain this and is being used only for evaluation.

The final cost function for the network combines the loss obtained through adversarial loss for the attention network along with the cross-entropy loss while solving for VQA. The final cost function used for obtaining the parameters θ_f of the attention network, θ_y of the classification network and θ_d for the discriminator is as follows:

$$C(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{j=1}^{n}(L_c^j(\theta_f, \theta_y) + \eta L_1^j(\theta_f, \theta_d))$$

where n is number of examples, and η is the hyper-parameter, fine-tuned using validation set and L_c is standard cross entropy loss. We train the model with this cost function till it converges so that the parameters $(\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_d)$ deliver a saddle point function

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\max_{\theta_f,\theta_y}(C(\theta_f, \theta_y, \hat{\theta}_d))$$
$$(\hat{\theta}_d) = \arg\min_{\theta_d}(C(\hat{\theta}_f, \hat{\theta}_y, \theta_d)) \quad (2)$$

Pixel-wise Adversarial Attention Network (PAAN): A variation of the adversarial attention network is to obtain a local pixel-wise discriminator for obtaining an improved attention network. The idea of pixel-wise discriminators has been studied for generative adversarial networks (GANs) and is termed patch-GAN. We show here, that doing pixel-wise (with multiple channels per pixel) attention network results in an improved attention network. We term this network a Pixel-wise Adversarial Attention Network (PAAN). Though this network uses more local discrimination, it does not increase the parameters of the network as compared to AAN. The effect of local discrimination results in improved attention as well as explanation.

The resultant min-max loss function is obtained as follows:

$$\min_G \max_{D^k} L_1^k(G,D^k) = E_{g_{g_i}\sim G_g(x_i)}[\log D^k(g_{g_i}/x_i)] + E_{g_{f_i}\sim G_f(x_i)}[\log(1 - D^k(G(g_{f_i}/x_i)))]$$

Finally, the actual cost function for training the pixel-wise discriminator, attention network and Grad-CAM is given by:

$$C(\theta_f, \theta_y, \theta_d^k|_{k=1}^K) = \frac{1}{n}\sum_{j=1}^{n}\left(L_c^j(\theta_f, \theta_y) + \eta\sum_{k=1}^{K} L_1^{j,k}(\theta_f, \theta_d^k)\right)$$

The algorithm for training the same is provided in Algorithm 1.

Algorithm 1 Training PAAN

Input: Image X_I, Question X_Q
Output: Answer X_A
repeat
  Attention features $G_f(G_i(X_I), G_q(X_Q)) \leftarrow g_a$
  Classification score $G_y(g_a) \leftarrow \hat{y}$
  Ans cross entropy $L_y \leftarrow$ *loss*$(\hat{y}, y)$
  Compute Gradient,$L_f = \frac{\partial L_y}{\partial \theta_y}$,$L_i = \frac{\partial L_f}{\partial \theta_f}$
  update $\theta_c \leftarrow \theta_c$ - $\frac{\partial L_c}{\partial \theta_c}$
  Explanation features $f_t(\theta_f, X_t) \leftarrow X_t$
  repeat
    Sample fake mini batch(Attention): $\alpha_1 \ldots \alpha_{196}$
    Sample real mini batch(Gradient): $\mu \ldots \mu_{196}$
    Discriminator: $D_k^r(\mu_k) \leftarrow d_k^r, D_k^f(\alpha_k) \leftarrow d_k^f$
    Update the discriminator by ascending its stochastic gradient
    $\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}[\log D(\mu_k) + \log(1 - D(\alpha_k))]$
  until $k = 1 : K$
  Sample fake mini batch(Attention): $\alpha_1 \ldots \alpha_{196}$
  Update the Generator by descending its stochastic gradient: $\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log(1 - D(\alpha))$
until Number of Iteration

3.2 Variations of Proposed Method

While we advocate the use of Adversarial explanation method for improving attention mask that can be used by the attention network for predicting answer, we also evaluate several other explanation method of this architecture. Our intuition is that, if we can learn that attention mask which minimizes the distance between attention probability distribution and the gradient class activation map, then we train our VQA classifier module to classify correct answer for given image based on the question. To minimize these distances we have used various methods.

- **Maximum Mean Discrepancy(MMD) Net**: In this variant, we minimize this distance using MMD [Tzeng *et al.*, 2014] based standard distribution distance metric. We have computed this distance with respect to a representation $\psi(.)$. In our case, we obtain representation feature $\psi(\alpha)$ for attention & $\psi(\mu)$ for Grad-CAM map.
- **Coral Net**: In this variant, we minimize distance between second-order statistics(covariances) of attention and Grad-CAM mask using CORAL loss [Sun and Saenko, 2016] based standard distribution distance metric. Here, both (μ) and (α) are the d-dimensional deep layer activation feature for attention and Grad-CAM maps. We have computed feature co-variance matrix of attention feature and Grad-cam feature represented by $C(\alpha)$ and $C(\mu)$ respectively.

4 Experiment

We evaluate the proposed method i.e. PAAN in a number of ways which includes both quantitative analysis and qualitative analysis. Quantitative analysis includes ablation analysis with other variants that we tried using metrics such as Rank

Model	RC(↑)	EMD(↓)
SAN [Das *et al.*, 2016]	0.2432	0.4013
CoAtt-W[Lu *et al.*, 2016]	0.246	–
CoAtt-P [Lu *et al.*, 2016]	0.256	–
CoAtt-Q[Lu *et al.*, 2016]	0.264	–
MMD (ours)	0.2573	0.3895
Coral (ours)	0.2563	0.3851
MSE (ours)	0.2681	0.3824
AAN (ours)	0.2826	0.3821
PAAN (ours)	**0.3071**	**0.3801**
Human [Das *et al.*, 2016]	0.623	–

Table 1: Ablation analysis and SOTA between HAT attention and generated attention mask

correlation(RC) score [Das *et al.*, 2016], Earth Mover Distance (EMD) [Arjovsky *et al.*, 2017], and VQA accuracy etc. as shown in table 1 and 2 respectable. Also we compare of our proposed method with various state of the art models, is provided in table 3 and 4. Qualitative analysis includes visualization of improvement in attention maps for some images as we move from our base model to the PAAN model. We also provide visualization of Grad-CAM maps for all the models. As there is simultaneous improvement in Grad-CAM images also, our results also corroborate this fact.

Models	All	Yes/No	Number	Others
Baseline-ATT	56.7	78.9	35.2	36.4
MMD + SAN	58.9	80.3	37.0	43.7
Coral + SAN	59.4	80.8	36.5	45.1
MSE + SAN	60.8	80.0	36.8	47.1
AAN + SAN	62.3	80.4	37.2	49.8
PAAN + SAN	63.6	81.1	36.9	50.9
AAN + MCB	66.4	84.6	37.8	54.7
PAAN + MCB	**67.1**	**85.0**	**38.4**	**55.9**

Table 2: Ablation analysis for Open-Ended VQA1.0 accuracy on test-dev

4.1 Ablation analysis

We provided comparison of our proposed model PAAN and other variants along with base model using various metrics in the table 1 and table 2. Rank correlation, EMD score are calculated for each model against human attention map [Das *et al.*, 2016]. Each model's generated attention map is used for this purpose. Rank correlation has an increasing trend. Increase in rank correlation indicates about the dependency of the both attention maps that are compared. As rank correlation increases, attention map generated from the model and human attention map becomes more dependent. In other words, higher rank correlation shows similarity between the maps. EMD also increases towards PAAN. To verify our intuition, that we can learn better attention mask by minimising the distance between attention mask and explanation mask, we start with MMD and observe that both rank correlation and answer accuracy increase by 1.42 and 1.2 % from baseline respectively. Also, we observe that with Coral and MSE based distance minimisation technique, both RC and EMD improves as shown in the table- 1. Instead of the predefined distance minimisation technique, we adapt an heuristic method to improvement in attention. We use one of the best heuristic method is adversarial learning method. So, our proposed AAN method, to improve our attention globally with respect to Grad-CAM. AAN improves 3.9% in-terms of RC and 9.5% on VQA accuracy. Finally,our proposed PAAN, which consider local pixel-wise discriminator improves 6.4% in RC and 10.4% in vqa accuracy as mentioned in the table 1 and table 2. Since, human attention map [Das *et al.*, 2016] is only available for VQA-v1 dataset, for VQA accuracy we preform ablation for VQA-v1 only. However, we provide state of arts results for both dataset(VQA-v1 and VQA-v2).

Models	All	Y/N	Num	Oth
Baseline-ATT	56.7	78.9	35.2	36.4
DPPnet [Noh *et al.*, 2016]	57.2	80.7	37.2	41.7
SMem[Xu and Saenko]	58.0	80.9	37.3	43.1
SAN [Yang *et al.*, 2016]	58.7	79.3	36.6	46.1
DMN[Xiong *et al.*, 2016]	60.3	80.5	36.8	48.3
QRU(2)[Li and Jia, 2016]	60.7	82.3	37.0	47.7
HieCoAtt [Lu *et al.*, 2016]	61.8	79.7	38.9	51.7
MCB [Fukui *et al.*, 2016]	64.2	82.2	37.7	54.8
MLB [Kim *et al.*, 2017]	65.0	84.0	37.9	54.7
DVQA[Patro, 2018]	65.4	83.8	38.1	55.2
AAN + SAN (ours)	62.3	80.4	37.2	49.8
PAAN + SAN (ours)	63.6	81.1	36.9	50.9
AAN + MCB (ours)	66.4	84.6	37.8	54.7
PAAN + MCB (ours)	**67.1**	**85.0**	**38.4**	**55.9**

Table 3: SOTA: Open-Ended VQA1.0 accuracy on test-dev

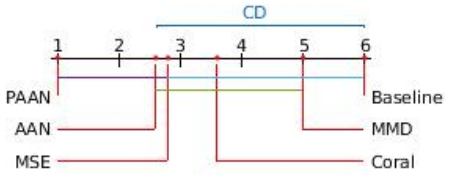

Figure 3: The mean rank of all the models on the basis of all scores are plotted on the x- axis. CD=3.3722, p=0.0003461. Here our PAAN model and others variants are described in section 4.1. The colored lines between the two models represents that these models are not significantly different from each other.

4.2 Statistical Significance Analysis

We analyze Statistical Significance [Demšar, 2006] of our PAAN model against the variants mentioned in section 4.1 of our method as well as other 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 models implies that they are significantly less different. Otherwise, they are statistically different. Figure 3 visualizes the post hoc analysis using the CD diagram. It is clear that PAAN works best and is significantly different from other methods.

4.3 Comparison with baseline and state-of-the-art

We obtain the initial comparison with the baselines on the rank correlation on human attention (HAT) dataset [Das *et al.*, 2016] that provides human attention while solving for VQA. Between humans the rank correlation is 62.3%. The comparison of various state-of-the-art methods and baselines are provided in table 1. We use variant of SAN[Yang *et al.*, 2016] model as our baseline method. We obtain an improvement of

Models	All	Y/N	Num	Oth
SAN-2[Yang *et al.*, 2016]	54.9	74.1	35.5	44.5
MCB [Fukui *et al.*, 2016]	64.0	78.8	38.3	53.3
Bottom[Anderson *et al.*]	65.3	81.8	44.2	56.0
DVQA[Patro, 2018]	65.9	82.4	43.2	56.8
MLB [Kim *et al.*, 2017]	66.3	83.6	44.9	56.3
DA-NTN [Bai *et al.*, 2018]	67.5	84.3	47.1	57.9
Counter[Zhang *et al.*, 2018]	68.0	83.1	51.6	58.9
BAN[Kim *et al.*, 2018]	69.5	85.3	50.9	60.2
AAN + SAN (ours)	60.1	76.4	35.2	51.8
PAAN + SAN (ours)	61.3	78.0	38.6	52.9
AAN + MCB (ours)	67.6	84.8	47.5	57.7
PAAN + MCB (ours)	**68.4**	**85.1**	**48.4**	**59.1**

Table 4: SOTA: Open-Ended VQA2.0 accuracy on test-dev

around 3.7% using AAN network and 6.39% using PAAN network in terms of rank correlation with human attention. Also, we compare with the baselines on the answer accuracy on VQA-v1[Antol *et al.*, 2015] and VQA-v2[Goyal *et al.*, 2017b] dataset as shown in table 3 and table 4 respectively. We obtain an improvement of around 5.8% over the comparable baseline. Further incorporating MCB improves the results for both AAN and PAAN resulting in an improvement of 7.1% over dynamic memory network and 3% improvement over MCB method on VQA-v1 and 4.2% on VQA-v2. However, as noted by [Das *et al.*, 2016], using a saliency based method [Judd *et al.*, 2009] that is trained on eye tracking data to obtain a measure of where people look in a task independent manner results in more correlation with human attention (0.49). However, this is explicitly trained using human attention and is not task dependent. In our approach, we aim to obtain a method that can simulate human cognitive abilities for solving tasks. We have provided more results for attention map visualization, training setup, dataset, and evaluation methods here [1].

4.4 Training and Model Configuration

We trained the PAAN model using classification loss and adversarial loss in an end-to-end manner. We have used ADAM optimizer to update the classification model parameter and configured hyper-parameter values using validation dataset as follows: {learning rate = 0.0001, batch size = 200, beta = 0.95, alpha = 0.99 and epsilon = 1e-8} to train the classification model. We have used SGD optimizer to update the adversarial model parameter and configured hyper-parameter values using validation dataset as follows: {learning rate = 0.004, batch size = 200, and epsilon = 1e-8} to train the adversarial model. We, also used weight clipping and learning rate decaying function to decrease learning rate on every 10 epoch.

4.5 Qualitative Result

We provide attention map visualization of all models for 3 example images as shown in Figure 4. 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

[1]https://github.com/2019ijcai/explanation

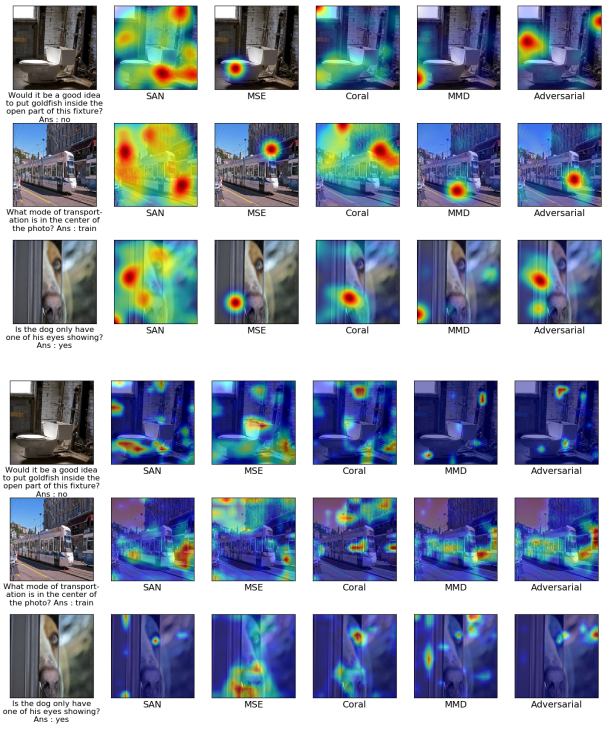

Figure 4: Examples with different approaches in each column for improving attention using explanation in a self supervised manner. The first column indicates the given target image and its question and answer. Starting from second column, it indicates the Attention map(above) / Grad-CAM map(below) for Stack Attention Network, MSE based approach, Coral based approach, MMD based approach, Adversarial based approach respectively.

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 verify our hypothesis that Grad-CAM is a better way of visualization of network. We see that it can focus on right portions of the image even in our base line model (SAN).Hence, it can be used to improve attention maps. Also, Grad-CAM is simultaneously improving according to our assumption and can also be seen in the Figure 4. For eg. in SAN it focuses on correct portions along with other points of focus too. But in our proposed model, visualization is improved to focus only on required portion.

5 Conclusion

In this paper we have proposed a method to obtain surrogate supervision for obtaining improved attention using visual explanation. Specifically, we consider the use of Grad-CAM, however, other such modules could also be considered. We show that the use of adversarial method to use the surrogate supervision performs best with the pixel-wise adversarial method (PAAN) performing better against other methods of using this supervision. The proposed method shows that the improved attention indeed results in improved results for the semantic task such as VQA or Visual dialog. Our method provides an initial means for obtaining surrogate supervision for attention and in future we would like to further investigate other means of obtaining improved attention.

References

[Anderson *et al.*, 2018] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[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.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *stat*, 1050:26, 2017.

[Bai *et al.*, 2018] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. Deep attention neural tensor network for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–35, 2018.

[Das *et al.*, 2016] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[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.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[Geman *et al.*, 2015] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, 112(12):3618–3623, 03 2015.

[Goyal *et al.*, 2017a] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2017.

[Goyal *et al.*, 2017b] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[Judd *et al.*, 2009] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[Kim *et al.*, 2017] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1571–1581, 2018.

[Li and Jia, 2016] Ruiyu Li and Jiaya Jia. Visual question answering with question representation update (qru). In *Advances in Neural Information Processing Systems*, pages 4655–4663, 2016.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[Noh *et al.*, 2016] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 30–38, 2016.

[Patro, 2018] Namboodiri Vinay P. Patro, Badri. Differential attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2953–2961, 2015.

[Selvaraju *et al.*,] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization.

[Shih *et al.*, 2016] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4621, 2016.

[Simons and Chabris, 1999] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, 28(9):1059–1074, 1999.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[Tzeng *et al.*, 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[Xiong *et al.*, 2016] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

[Xu and Saenko, 2016] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[Zhang *et al.*, 2018] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. 2018.

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.