The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Explanation vs Attention: A Two-Player Game to Obtain Attention for VQA

Badri N. Patro, Anupriy, Vinay P. Namboodiri
Indian Institute of Technology, Kanpur
{badri, anupriy, vinaypn}@iitk.ac.in

Abstract

In this paper, we aim to obtain improved attention for a 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 a substantial improvement over baseline stacked attention network (SAN) models. It also results in a good improvement in rank correlation metric on the VQA task. This method can also be combined with recent MCB based methods and results 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 is a commonly known fact in cognitive science. 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 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.

The main idea is that given the task of solving visual question answering (VQA), there exist methods based on obtaining visual explanations such as Grad-CAM (Selvaraju et al. 2017) 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. One can obtain the visual explanation using the ground-truth label for a deep network that solves the visual question answering task. As the 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. In this paper, we show that a very simple way of using a two-player game between a discriminator that tries to discriminate between Grad-CAM results & attention maps and a generator that generates attention maps serves to provide 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 the visual ex-

planation in the form of Grad-CAM results. This method performs better as compared to other forms of surrogate supervision such as using RISE (Petsiuk, Das, and Saenko 2018).

- We show that this surrogate supervision can be best used through a variant of adversarial learning to obtain attention maps that correlate well with the 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. A detailed empirical analysis for the same is provided.

1.1 Motivation

In VQA, given an image & a 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 the answer based on the question. The regions pointed by humans for answering the visual question are more accurate as compared to those obtained by other techniques. 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 increases with ground truth human attention map for the stacked attention network (Yang et al. 2016). 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 well to novel tasks. However, we are motivated by this result and have therefore developed in this paper a self-supervision based method to improve attention. We formulate a game between Attention vs. Explanation using adversarial mechanism. Through this game, we observe that we obtain improved attention regions, which lead to improved prediction and therefore, also results in better regions obtained through visual explanation as shown in the figure- 2. Thus, improving attention using Grad-CAM results in an improvement in Grad-CAM too. 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 (SAN) model increases by 4% on the 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 (VQA) 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 the 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, Kiros, and Zemel 2015; Goyal et al. 2017; Noh, Hongsuck Seo, and Han 2016) where visual features are combined with question features to predict the answer. Attention based approaches are the other category of methods for solving VQA. It comprises of image based, question based and some that are both image and question based attention. (Shih, Singh, and Hoiem 2016) has proposed an image based attention approach, the aim is to use the question in order to focus attention over specific regions in an image. (Yang et al. 2016) has proposed a method to repeatedly obtain attention by using stacked attention over an image based on the question. Our work uses this as one of the baselines. (Li and Jia 2016) has proposed a region based attention model over images. Similarly, (Zhu et al. 2016; Xu and Saenko 2016; Bao et al. 2018) have proposed interesting method for question based attention. A work that explores joint image and question includes that is based on hierarchical co-attention is (Lu et al. 2016). There has been interesting work by (Fukui et al. 2016; Kim et al. 2017; Kim, Jun, and Zhang 2018; Patro et al. 2018) that advocates multimodal pooling and obtains close to state of the art results in VQA.

The task of VQA is well studied in the vision and language community, but it has been relatively less explored for providing explanation (Selvaraju et al. 2017; Goyal et al. 2017) for answer prediction. We start with image captioning (Socher et al. 2014; Vinyals et al. 2015; Karpathy and Fei-Fei 2015; Xu et al. 2015; Fang et al. 2015; Chen and Lawrence Zitnick 2015; Johnson, Karpathy, and Fei-Fei 2016; Yan et al. 2016) to provide a basic explanation for an image. The next level of challenging task is to provide an explanation for the visual question answering system. The attention-based model provides some short of basic explanation for VQA. This is observed that models (Das et al. 2016) are not looking at the same regions as humans are looking. So we need to improve the attention of the model and its explanation for answer prediction. (Patro et al. 2018) has proposed an exemplar-based method to improve the attention map for the VQA task. (Jain and Wallace 2019) has proposed a method to evaluate how attention weights can provide a correct explanation in language prediction task. Recently, Uncertainty based explanation method (Patro et al. 2019) is proposed to improve the attention mask for VQA. There are very interesting methods to provide visual explanations such as Grad-CAM (Selvaraju et al. 2017), RISE (Petsiuk, Das, and Saenko 2018), U-CAM (Patro et al. 2019). In contrast to the above-mentioned approaches, we focus on improving image-based attention using an adversarial game between visual explanation mask (Grad-CAM) and attention mask and show that it correlates better with human attention. Our approach allows the use of visual explanation as a means for obtaining surrogate supervision for attention.

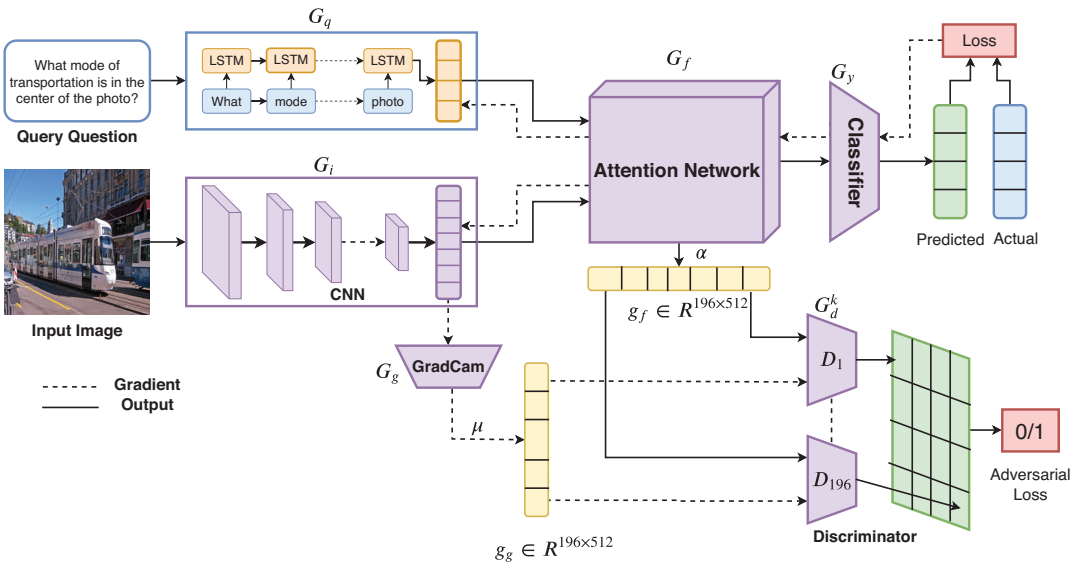

Figure 1: Illustration of model PAAN and its attention mask. Image feature and question feature are obtained using CNN and LSTM respectively. Attention mask is then obtained using these features and classification of the answer is done based on the attended feature. We have improved the attention mask with the visual explanation approaches based on Grad-CAM

3 Method

The main focus in our approach for solving visual question answering (VQA) is to use supervision obtained from visual 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 the 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 1. 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 Our Approach

The three main components of our approach, as illustrated in figure 1 are 1) Attention representation, 2)Explanation representation, 3) Adversarial Game. The details of our method are provided in the following sub-sections:

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 of 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. 2017). 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. 2017). We have also evaluated with another such approach termed as RISE (Petsiuk, Das, and Saenko 2018). We observed better results using Grad-CAM.

Adversarial Game A zero-sum adversarial game between two players (P1, P2) is used with one set of players being a Generator network and the other being a discriminator network. They choose a decision from their respective decision sets $\mathcal{K}_1$ and $\mathcal{K}_2$. 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). A game objective $\mathcal{L}(G, D) : \mathcal{K}_1 \times \mathcal{K}_2 \in \mathcal{R}$, sets the utilities of the players. Concretely, by choosing a proper strategy $(G, D) \in \mathcal{K}_1 \times \mathcal{K}_2$ the utility of P1 is $-\mathcal{L}(G, D)$, while the utility of P2 is $\mathcal{L}(G, D)$. The goal of either P1/P2 is to maximize their worst

case utilities; thus,

$$\begin{aligned} &\min_{G\in\mathcal{K}_1}\max_{D\in\mathcal{K}_2} L(G,D) \quad \text{(Goal of P1)},\\ &\max_{D\in\mathcal{K}_2}\min_{G\in\mathcal{K}_1} L(G,D) \quad \text{(Goal of P2)} \end{aligned} \tag{1}$$

The above formulation raises the question of whether there exists a solution(G^*, D^*) to which both players may jointly converge. The solution to this question is to obtain a Nash equilibrium where the Discriminator is unable to distinguish the generations of the Generator network from the 'real' distribution *i.e.* $[\max_{D\in\mathcal{K}_2} L(G^*, D) = \min_{G\in\mathcal{K}_1} L(G, D^*)]$.

Since pure equilibrium does not always exist (Nash and others 1950), there exists an approximate solution for this issue as a Mixed Nash Equilibrium,i.e.

$$\max_{D\in\mathcal{K}_2} \mathbb{E}_{G\sim D_1} L(G,D) = \min_{G\in\mathcal{K}_1} \mathbb{E}_{D\sim D_2} L(G,D) \tag{2}$$

Where D_1 is the distribution over K_1, and D_2 is the distribution over K_2. In zero-sum adversarial game, the sum of the generator's loss and the discriminator's loss is always zero, i.e. the generator's loss is: $\mathcal{L}^G = -\mathcal{L}^D$. The solution for a zero-sum game is called a minimax solution, where the goal is to minimize the maximum loss. We can summarize the entire game by stating that the loss function is L^G (which is the discriminator's payoff), so that the minimax objective is

$$\begin{aligned} \min_G \max_D L_1(G,D) = E_{g_{g_i}\sim G_g(x_i)}[\log D(g_{g_i}/x_i)] + \\ E_{g_{f_i}\sim G_f(x_i)}[\log(1 - D(G(g_{f_i}/x_i)))] \end{aligned} \tag{3}$$

For simplicity, we remove subscript i. Here g_g is the output of Grad-cam network G_g for a sample, x_i and g_f is the output of the attention network. The discriminator wants to maximize the objective (i.e., its payoff) such that $D(g_g/x)$ is close to 1 and $D(G(g_f/x))$ is close to zero. The generator wants to minimize the objective (i.e., its loss) so that D(G(z)) is close to 1. Specifically, the discriminator is a set of CNN layers followed by a linear layer that uses a binary cross entropy loss function. 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 an 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:

$$C(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{j=1}^{n} (L_c^j(\theta_f, \theta_y) + \eta L^j(\theta_f, \theta_d)) \tag{4}$$

Where n is the number of examples, and $\eta = 10$ 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.

$$\begin{aligned} (\hat{\theta}_f, \hat{\theta}_y) &= \arg\max_{\theta_f,\theta_y}(C(\theta_f, \theta_y, \hat{\theta}_d))\\ (\hat{\theta}_d) &= \arg\min_{\theta_d}(C(\hat{\theta}_f, \hat{\theta}_y, \theta_d)) \end{aligned} \tag{5}$$

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 $G_y(g_a) \leftarrow \hat{y}$
  Answer 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): $\mu_1 \ldots \mu_{196}$
    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

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 algorithm for training the same is provided in Algorithm 1. The resultant min-max loss function is obtained as follows:

$$\begin{aligned} \min_G \max_{D^k} L_1^k(G, D^k) = E_{g_{g_i}\sim G_g(x_i)}[\log D^k(g_{g_i}/x_i)] + \\ E_{g_{f_i}\sim G_f(x_i)}[\log(1 - D^k(G(g_{g_i}/x_i)))] \end{aligned} \tag{6}$$

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

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

The main problem we face is the model convergence issue where the model parameter oscillates and does not converge using gradient descent in a minimax game. To handle convergence issue, we add JS-divergence (Fuglede and Topsoe 2004) to the cost function, which penalizes a poor generated mask badly as compared to a good one, which is different from KL-divergence (Kullback and Leibler 1951). The second issue faced is "vanishing gradient", when discriminator is successful (which can well distinguish between generated and discriminator sample), then the generator gradient

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_RISE	0.2591	0.3992
Coral_RISE	0.2609	0.3978
MSE_RISE	0.2622	0.3921
AAN_RISE	0.2683	0.3900
PAAN_RISE	0.2754	0.3894
MMD (ours)	0.2573	0.3895
Coral (ours)	0.2563	0.3851
MSE (ours)	0.2681	0.3814
AAN (ours)	0.2896	0.3721
PAAN (ours)	**0.3071**	**0.3701**
PAAN_Ran_07	0.1213	0.6700
PAAN_Ran_20	0.1746	0.5872
Human (Das et al. 2016)	0.623	–

Table 1: Attention mask comparison for SOTA & Ablation Methods

vanishes and learns nothing. To handle this issue, we add Pearson-χ^2 divergence (Mao et al. 2017) to the GAN cost function.

3.2 Variations of Proposed Method

While we advocate the use of Adversarial explanation method for improving the attention mask, we also evaluate several other explanation methods for this architecture. Our intuition is that, if we learn an attention mask that minimizes the distance between attention probability distribution and the gradient class activation map, then we are more easily able to train our VQA classifier module to provide correct answer. 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 (co-variances) 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.

We trained our variants MMD and CORAL directly without adversarial loss to bring Grad-CAM based pseudo distribution close to attention distribution. Finally we replace MMD and CORAL with adversarial loss.

Models	All	Yes/No	Num	Oth
Baseline-ATT	56.7	78.9	35.2	36.4
MMD_SAN_RISE	56.9	79.1	35.8	38.1
Coral_SAN_RISE	57.4	79.8	36.0	39.6
MSE_SAN_RISE	58.2	80.1	36.4	40.2
AAN_SAN_RISE	59.3	80.4	36.9	42.5
PAAN_SAN_RISE	60.1	80.8	37.3	44.2
MMD_SAN_GCAM	58.9	80.3	37.0	43.7
Coral_SAN_GCAM	59.4	80.8	36.5	45.1
MSE_SAN_GCAM	60.8	80.0	36.8	47.1
AAN_SAN_GCAM	62.3	80.4	37.2	49.8
PAAN_SAN_GCAM	63.6	81.1	36.9	50.9
AAN_MCB_GCAM	66.4	84.6	37.8	54.7
PAAN_MCB_GCAM	**67.1**	**85.0**	**38.4**	**55.9**
PAAN_SAN_Ran_07	55.2	77.2	35.1	36.2
PAAN_SAN_Ran_20	57.3	78.7	35.6	39.7

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

Models	All	Y/N	Num	Oth
Baseline-ATT	56.7	78.9	35.2	36.4
DPPnet (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 (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 et al. 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

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 correlation (RC) score (Das et al. 2016), Earth Mover Distance (EMD) (Arjovsky, Chintala, and Bottou 2017), and VQA accuracy etc. as shown in table 1 and 2 respectable. We also compare our proposed method with various state of the art models, as 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.

4.1 Ablation analysis on model parameter

We provide comparisons of our proposed model PAAN and other variants along with base model using various metrics in the table 1 and table 2. Rank correlation and EMD score are calculated for each model against human attention map (Das et al. 2016). Each model's generated attention map is

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
DVQA (Patro et al. 2018)	65.9	82.4	43.2	56.8
MUTAN (Ben et al. 2017)	66.0	82.8	44.5	56.5
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 (2018)	68.0	83.1	51.6	58.9
GCA (Patro et al. 2019)	69.2	85.4	50.1	59.4
BAN (2018)	69.5	85.3	**50.9**	60.2
BU (Anderson et al. 2018)	**70.34**	**86.6**	48.64	**61.15**
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

used for this purpose. The rank correlation has an increasing trend. Increase in rank correlation indicates the dependency of the 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 improves for 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 adversarial learning method. The proposed AAN method improves 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 considers local pixelwise 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 perform ablation for VQA-v1 only. However, we provide state of the art results for both datasets (VQA-v1 and VQA-v2).

4.2 Ablation on Explanation: Why do we select Grad-CAM?

While calculating Grad-CAM one uses the "true" class labels in obtaining activation maps. When observing attention, one just infers these for a sample without using the ground-truth label. At test time, Grad-CAM results cannot be used as true class labels would not be available. By using Grad-CAM as supervision, the aim is to obtain dense supervision for the attention module that will guide the attention methods as against the sparse rewards that are available based on the correct classification prediction. To validate this we conduct an experiment with another kind of visual explanation, i.e., RISE (Petsiuk, Das, and Saenko 2018) in a similar way to Grad-CAM(Selvaraju et al. 2017). In RISE, we use the true label to obtain RISE based activation maps, instead of Grad-CAM, that corresponds to the true prediction. This, as surrogate supervision, is observed by us to provide better results as compared to using only attention without supervision. We evaluate the rank correlation of the attention mask for RISE supervision and observe that it is much lower than Grad-CAM supervision, as shown in table-1. This method results in an improvement of 3.22% in terms of rank correlation over the baseline SAN (Das et al. 2016) method while we obtain an improvement of 6.39% using Grad-CAM supervision. Similarly, we observe that the Earth Mover Distance of RISE based model is higher than the Grad-CAM based model. We believe that the suggested framework can always be improved by any other surrogate supervision technique that can be developed.

4.3 Why adversarial learning rather than supervised learning?

Attention and Grad-CAM distributions differ as has been pointed out. However, the Grad-CAM results are based on the true labels. Therefore, if the distributions are close, then it would serve the purpose. This is because, the attention maps need not exactly correspond to the gradient of the class activations. By using adversarial learning and trying to fool the discriminator, we are able to serve our purpose. This is ensured also by providing comparisons against explicitly using Grad-CAM as supervision with MSE loss results in lower performance. Therefore, adversarial learning is a good method for solving this problem (better even than other distribution matching techniques such as MMD or CORAL). To validate this, we conduct an experiment on distribution matching between the generated attention mask and the ground truth explanation mask. One of the simplest ways to measure the overlapping distributions is the Wasserstein (Arjovsky and Bottou 2017) distance between them. We observe that for a perfect adversarial game, the model achieves pure or Mixed Nash Equilibrium, the joint distribution between p (explanation) distribution and q (attention) distribution should be diagonal, that is p & q distribution are highly overlapped. And the EMD should be very small. So, using Grad-CAM supervision for attention mask helps to achieve more close towards Mixed Nash Equilibrium in two player game as compared to random and RISE based supervision. We also observe that if the overlapping region between p-distribution and q-distribution is very low, then KL-divergence in our adversarial game completely fails and JS-divergence works well. In this experiment, we consider three types of explanations mask, Grad-CAM (Selvaraju et al. 2017), RISE (Petsiuk, Das, and Saenko 2018) and a random mask. We start to observe that with a random explanation mask the accuracy is not improving; rather it is decreasing, when the overlap of the distribution increases, the performance in terms of rank correlation and accuracy is also increasing. We show the experiment result for $PAAN\_SAN\_Ran\_07$ whose distribution overlapping is 7% and $PAAN\_SAN\_Ran\_20$ distribution overlapping is 20% as shown in table-2 and second last row of table -1.

4.4 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

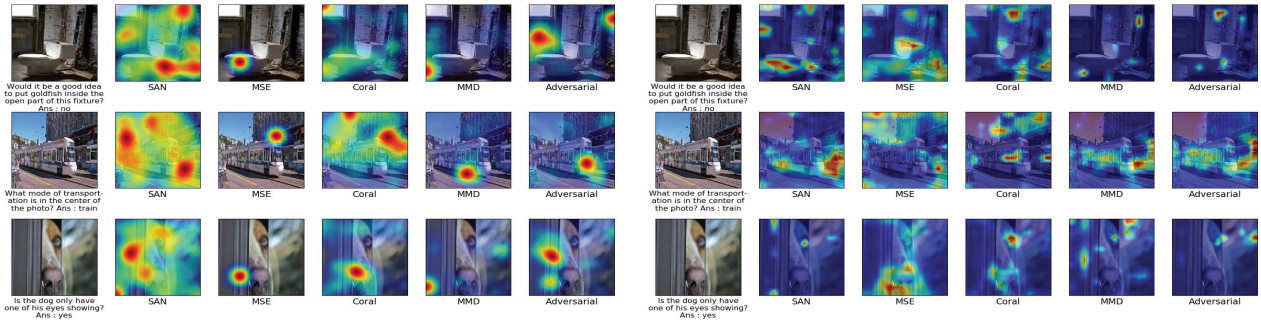

Figure 2: Examples with different approaches in each column for improving attention using explanation in a self supervised manner. The first column indicates the given target image and its question and answer. Starting from second column, it indicates the Attention map (left) / Grad-CAM map (right) for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

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 around 3.7% using AAN network and 6.39% using PAAN network in terms of rank correlation with human attention. We also compare with the baselines on the answer accuracy on VQA-v1(Antol et al. 2015) and VQA-v2(Goyal et al. 2017) 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. The method is not limited to classification alone though all the methods proposed for VQA-1 and VQA-2 datasets follow this. The proposed framework can easily be extended to generative frameworks that generate answers in terms of sentences. We use visual dialog task(Das et al. 2017) for generative framework, where we visualised improved attention map with respect to generated answer. We observe improvement of overall performance in terms of NDGC values by 1.2% and MRR values by 0.78% over the baseline dialog model (Das et al. 2017). We have provided more results of AAN and PAAN for VQA and Visual dialog, attention map visualization, dataset, and evaluation methods in our project page- 1.

4.5 Qualitative Result

We provide attention map visualization for all models 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 a better way of visualization of network learning as it can focus on the right portions of the image even in our base line model (SAN). Therefore, it can be used as a tutor to improve attention maps. Our PAAN model helps to learn the attention distribution in an adversarial manner from Grad-CAM distribution as compared to SAN and others. Also, Grad-CAM is simultaneously improved according to our assumption and can also be seen in the Figure 2. For example, in SAN the focus of Grad-CAM is spread over the image. In our proposed model, visualization is improved to focus only on the required portion. In the project website[1], we show variance in attention map for the same question to the image and its composite image in VQA2.0 dataset. We also provide statically significant analysis result for variants of our models compare with PAAN model in our project page- 1.

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 for obtaining improved attention.

[1] https://delta-lab-iitk.github.io/TwoPlayer/

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Arjovsky, M., and Bottou, L. 2017. Towards principled methods for training generative adversarial networks. *stat* 1050:17.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *stat* 1050:26.

Bai, Y.; Fu, J.; Zhao, T.; and Mei, T. 2018. Deep attention neural tensor network for visual question answering. In *ECCV*.

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving machine attention from human rationales. In *EMNLP*.

Ben, younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*.

Chen, X., and Lawrence Zitnick, C. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*.

Das, A.; Agrawal, H.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2016. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*.

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.; et al. 2015. From captions to visual concepts and back. In *CVPR*.

Fuglede, B., and Topsoe, F. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 31. IEEE.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America* 112(12):3618–3623.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Jain, S., and Wallace, B. C. 2019. Attention is not explanation. In *NAACL*, 3543–3556.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 4565–4574.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *ICCV*, 2106–2113.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3–7.

Kim, J.-H.; On, K. W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *NIPS*, 1571–1581.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1):79–86.

Li, R., and Jia, J. 2016. Visual question answering with question representation update (qru). In *NIPS*, 4655–4663.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.

Malinowski, M., and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *ICCV*.

Nash, J. F., et al. 1950. Equilibrium points in n-person games.

Noh, H.; Hongsuck Seo, P.; and Han, B. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 30–38.

Patro; Badri; Namboodiri; and P, V. 2018. Differential attention for visual question answering. In *CVPR*, 7680–7688.

Patro, B. N.; Lunayach, M.; Patel, S.; and Namboodiri, V. P. 2019. U-cam: Visual explanation using uncertainty based class activation maps. In *The IEEE International Conference on Computer Vision (ICCV)*.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NIPS*, 2953–2961.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*.

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

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL* 2(1):207–218.

Sun, B., and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 443–450. Springer.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.

Xu, H., and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 451–466. Springer.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional image generation from visual attributes. In *ECCV*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*.

Zhang, Y.; Hare, J.; and Prügel-Bennett, A. 2018. Learning to count objects in natural images for visual question answering.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *CVPR*.