Abstract

Understanding and explaining deep learning models is an imperative task. Towards this, we propose a method that obtains gradient-based certainty estimates that also provide visual attention maps. Particularly, we solve for visual question answering task. We incorporate modern probabilistic deep learning methods that we further improve by using the gradients for these estimates. These have two-fold benefits: a) improvement in obtaining the certainty estimates that correlate better with misclassified samples and b) improved attention maps that provide state-of-the-art results in terms of correlation with human attention regions. The improved attention maps result in consistent improvement for various methods for visual question answering. Therefore, the proposed technique can be thought of as a recipe for obtaining improved certainty estimates and explanation for deep learning models. We provide detailed empirical analysis for the visual question answering task on all standard benchmarks and comparison with state of the art methods.

1. Introduction

To interpret and explain the deep learning models, many approaches have been proposed. One of the approaches uses probabilistic techniques to obtain uncertainty estimates, [16][17]. Other approaches aim at obtaining visual explanations through methods such as Grad-CAM [9] or by attending to specific regions using hard/soft attention. With the recent probabilistic deep learning techniques by Gal and Ghahramani [16], it became feasible to obtain uncertainty estimates in a computationally efficient manner. This was further extended to data uncertainty and model uncertainty based estimates [24]. Through this work, we focus on using gradients uncertainty losses to improve attention maps while also enhancing the explainability leveraging the Bayesian nature of our approach. The uncertainties that we use are aleatoric and predictive [25].

For the estimated uncertainties, we calculate gradients using the approach similar to gradient-based class activation maps [9]. This provides “certainty maps” which helps in attending to certain regions of the attention maps. Doing this, we report an improvement in attention maps. This is illustrated in the Figure 1.

Our method combines techniques from both the explanation [9] and uncertainty [24] estimation techniques to obtain improved results. We have provided an extensive evaluation. We show that the results obtained for uncertainty estimates show a strong correlation with misclassification, i.e., when the classifier is wrong, the model is usually uncertain. Further, the attention maps provide state of the art correlation with human-annotated attention maps. We also show that on various VQA datasets, our model provides results comparable to the state-of-the-art while significantly improving the performance of baseline methods on which we incorporated our approach. Our method may be seen as a generic way to obtain Bayesian uncertainty estimates, visual explanation, and as a result, improved accuracy for Visual Question Answering (VQA) task.

Our contributions, therefore, lie in, a) unifying approaches for understanding deep learning methods using uncertainty estimate and explanation b) obtaining visual attention maps that correlate best with human attention regions and c) showing that the improved attention maps re-
sult in consistent improvement in results. This is particularly suited for vision and language-based tasks where we are interested in understanding visual grounding, i.e., for instance, if the answer for a question is ‘dog’ (Corresponding question: ‘Who is on the boat?’), it is important to understand whether the model is certain and whether it is focusing on the correct regions containing a dog. This important requirement is met by the proposed approach.

**Data uncertainty in a multi-modal setting.** Uncertainty in VQA task is two-fold. In the example below, Question, "Which kind of animal is it?" when asked (irrespective of image), may not be concretely answered. Also, seeing the image alone, in the given setting, the animal (especially the one behind) could easily be mis-classified as a dog or some other animal. These kinds of data uncertainties are tapped & hence minimized best when we consider uncertainties of the fused input (image+question). In Figure 2 we show the resultant attention maps of baseline (not minimizing uncertainty) & when we tried to minimize only-image, only-question & the fused uncertainty respectively.

![Figure 2](image.png)

**Figure 2.** The first column is the original image. 2nd, 3rd, 4th, and 5th columns represent the baseline attention, attention when only image uncertainty was minimized, attention when only question uncertainty was minimized, attention when both image and question uncertainties were minimized (proposed model) respectively.

### 2. Related work

The task of Visual question answering [35, 2, 40, 18, 36] is well studied in the vision and language community, but it has been relatively less explored for providing explanation [41] for answer prediction. Recently, lot of works that focus on explanation models, one of that is image captioning for providing basic explanation for image [5, 12, 28, 45, 48, 22, 51, 11, 7, 20, 52]. [38] has proposed method to measure predictive uncertainty with the help of uncertainty class activation maps, predicted probability in VQA. Our method also provides visual explanation based on uncertainty class activation maps, capturing and visualizing the uncertainties present in the attention maps in VQA.

### 3. Modeling Uncertainty

We consider two type of uncertainties present in the deep network, one due to uncertainty present in the data (Aleatoric), and the other due to model (Epistemic uncertainty).

#### 3.1. Modeling Aleatoric Uncertainty

Given an input \( x_i \) the model \( G \) predicts the logit output \( \hat{y}_i \) which is then an input to uncertainty network \( U \) for obtaining the variance \( \sigma_i^2 \) as shown in Figure 3. To capture Aleatoric uncertainty [24], we learn the observational noise parameter \( \sigma_e \) for each input point \( x_i \). Then, Aleatoric uncertainty, \( (\sigma_i^2) \), is estimated by applying softplus function on the output logit variance. This is given by,

\[
(\sigma_i^2) = \text{Softplus}(\sigma_e^2) = \log(1 + \exp(\sigma_e^2)) \tag{1}
\]

For calculating the aleatoric uncertainty loss, we perturb the logit value \((\hat{y}_i)\) with Gaussian noise of variance \((\sigma_i^2)\), (diagonal matrix with one element corresponding to each logits value) before the softmax layer. The logits reparameterization trick [25] and [15] combines \(\hat{y}_{i,c} \) and \( \sigma_i \) to give
\(N(\hat{y}_{i,c}, \sigma^2_p)\). We then obtain a loss with respect to ground truth. It is expressed as:

\[
\hat{y}_{i,c,t} = y_{i,c} + \epsilon_t \star \sigma^2_i, \text{where} \quad \epsilon_t \sim N(0, I) \tag{2}
\]

\[
\mathcal{L}_a = \sum_i \log \frac{1}{T} \sum_t \exp (\hat{y}_{i,c,t} - \log \sum_c \exp \hat{y}_{i,c,t}) \tag{3}
\]

where \(\mathcal{L}_a\) is the aleatoric uncertainty loss (AUL), \(T\) is the number of Monte Carlo simulations. \(c\) is a the class index of the logit vector \(y_{i,t}\) which is defined for all the classes.

### 3.2. Modeling Predictive Uncertainty

To obtain the model uncertainty, we measure epistemic uncertainty. However, estimating epistemic uncertainty \([34]\) is computationally expensive, and thus we measure the predictive uncertainty, having both aleatoric and epistemic uncertainties present in it. To estimate it, we sample weights in the Bayesian networks \(G\) and then perform Monte Carlo simulations over the model to obtain the predicted class probabilities \(p(\hat{y}_{i,t})\). That is,

\[
O(\hat{y}_{i,t}) = G^t(x_i) \quad \nu^t_{i,t} = \text{Softplus}(U^t(\hat{y}_{i,t}))
\]

\[
p(\hat{y}_{i,c} | x_i, X_I) = \left( \frac{1}{T} \sum_{t=1}^T \text{Softmax}O(\hat{y}_{i,t}) \right)_c
\]

where \(c\) is the answer class, \(G^t \sim G, U^t \sim U\) and \(\nu^t_{i,t}\) is the aleatoric variance of each logit in the \(t^{th}\) MC Simulation. The entropy of the sampled logit’s probabilities can be calculated as:

\[
H(\hat{y}_i) = -\sum_{c=1}^{C} p(\hat{y}_{i,c}) \star \log p(\hat{y}_{i,c}) \tag{4}
\]

The predictive uncertainty contains entropy and aleatoric variance when its expectation is taken across \(T\) number of Monte Carlo simulations:

\[
\sigma^2_p = H(\hat{y}_i) + \frac{1}{T} \sum_{t=1}^T \nu^a_{i,t} \tag{5}
\]

where \(H(\hat{y}_i)\) is the entropy of the probability \(p(\hat{y}_i)\), which depends on the spread of class probabilities while the variance (second term in the above equation) captures both the spread and the magnitude of logit outputs, \(\hat{y}_{i,t}\). In Equation \([\text{2}]\) we can replace \(\sigma^2_a\) with predictive uncertainty \(\sigma^2_p\) (mentioned above in Equation \([\text{5}]\)) to get the predictive uncertainty loss (PUL).

### 4. Method

**Task:** We solve for VQA \([\text{2}]\) task. The key difference in our architecture as compared to the existing VQA models is the introduction of gradient-based certainty maps. A detailed figure of the model is given in the Figure \([\text{4}]\) We keep other aspects of the VQA model unchanged. In a typical open-ended VQA task, we have a multi-class classification task. A combined (image and question) input embedding is fed to the model. Then, the output logits are fed to a softmax function, giving probabilities of the predictions in the multiple-choice answer space. That is, \(\hat{A} = \arg\max P(A | I, Q, \theta)\), where \(\Omega\) is a set of all possible answers, \(I\) is the image, \(Q\) is the corresponding question, and \(\theta\) is representing the parameters of the network.

#### 4.1. U-CAM Approach

The three main parts of our method are Attention Representation, Uncertainties Estimation, and computing gradients of uncertainty losses. In the following sections, we explain them in detail.

##### 4.1.1 Attention Representation

We obtain an embedding, \(g_i \in \mathcal{R}^{u \times v \times C}\) where \(u\) is width, \(v\) is height of the image and \(C\) represents the number of applied filters on the image \(X_I\) in the convolution neural network (CNN). The CNN is parameterized by a function \(G_i(X_i, \theta_i)\), where \(\theta_i\) represents the weights. Similarly, for the query question \(X_Q\), we obtain a question feature embedding \(g_q\) using a LSTM network. This network is parameterized by a function \(G_q(X_q, \theta_q)\), where \(\theta_q\) represents the weights. Both \(g_i\) and \(g_q\) are fed to an attention network that combines the image and question embeddings using a weighted softmax function and produces a weighted output attention vector, \(g_f\) as illustrated in Figure \([\text{4}]\). Various kinds of attention networks have been proposed in the literature. In this paper, we tried with SAN \([\text{53}]\) and MCB \([\text{14}]\). Finally, we obtain attention feature \(f_i\) using attention extractor network \(G_f : f_i = G_f(g_i, g_q)\). The attended feature \(f_i\) is passed through a classifier and the model is trained using the cross-entropy loss. Many a times, model is not certain about the answer class to which the input belongs, which sometimes leads to decrease in accuracy. To tackle this, we have proposed a technique to reduce the class uncertainty by increasing the certainty of the attention mask. Additionally, we also incorporate a loss based on the uncertainty which is described next.
4.1.2 Estimating Uncertainties: Aleatoric & Predictive

The attention feature, \( f_i \), obtained from the previous step is fed to the classifier \( G_c \). The output of the classifier is fed to \( G_v \), which produces class probabilities, \( y_i \). \( G_c \)'s output is also fed to a variance predictor network, \( G_v \), which outputs the logits’ variance, \( \sigma_i \) as mentioned in the Equation 1. For calculating the aleatoric uncertainty loss, we perturb the logit value (\( y_i \)) with Gaussian noise of variance \( (\sigma_i)^2 \) before the softmax layer. The Gaussian likelihood for classification is given by \( p(y_i|f_i, w) = \mathcal{N}(y_i; G_y(G_c(f_i; w), \tau^{-1}(f_i; w))) \), where \( w \) represents model’s parameters, \( \tau \) is the precision, \( f_i \) is the attended fused input, and \( G_y(G_c(\cdot)) \) is the output logit producing network as shown in the Figure 4. The above setting represents the perturbation of model output with the variance of the observed noise, \( \tau^{-1} \). We make sure that \( \tau(\cdot) \) is a positive or positive definite matrix (in case of Multivariate) by using the logit reparameterization trick [25] [15].

Finally, we then obtain an aleatoric loss, \( L_a \) with respect to ground truth as mentioned in the Equation 2. Our proposed model, which uses this loss as one of the components of its uncertainty loss, is called Aleatoric-GCA (A-GCA).

Along with aleatoric loss \( L_a \), we combine \( L_{VE} \) and \( L_{UDL} \) as mentioned in the Equation 10 and 11 respectively to get total uncertainty loss \( L_u \). The classifier is trained by jointly minimizing both the classification loss, \( L_y \) and the uncertainty loss, \( L_u \). In Equation 2 we can replace \( \sigma_i^2 \) with predictive uncertainty \( \sigma_i^2 \) (mentioned above in Equation 5) to get the predictive uncertainty loss (PUL). Accordingly, the model which uses this loss as one of the constituents of its uncertainty loss is called Predictive-GCA (P-GCA).

Next, we compute the gradients of standard classification loss and uncertainty loss with respect to attended image feature, \( f_i \). Besides training, we also use these gradients to obtain visualizations describing important regions responsible for answer prediction, as mentioned in the qualitative analysis section (Section 5.6).

4.1.3 Gradient Certainty Explanation for Attention

Uncertainty present in the attention maps often leads to uncertainty in the predictions and can be attributed to the noise in data and the uncertainty present in the model itself. We improve the certainty in these cases by adding the certainty gradients to the existing Standard Cross-Entropy (SCE) loss gradients for training the model during backpropagation.

Our objective is to improve the model’s attention in the regions where the classifier is more certain. The classifier will perform better by focusing more on certain attention regions, as those regions are more suited for the classification task. We can get an explanation for the classifier output as done in the existing Grad-CAM approach \( \frac{\partial L_y}{\partial f_i} \). But that explanation does not take the model and data uncertainties into the account. We improve this explanation using the certainty gradients \( \frac{\partial L_u}{\partial f_i} \). If we can minimize uncertainty in the VQA explanation, then uncertainties in the image and question features, and thus uncertainties in the attention regions would be subsequently reduced. It is the uncertain regions which are a primary source for errors in the prediction, as shown in Figure 4.

In our proposed method, we compute the gradient of the Standard Classification (cross entropy) loss \( L_y \) with respect to attention feature i.e. \( \frac{\partial L_y}{\partial f_i} \) and also the gradient of the uncertainty loss \( L_u \), i.e. \( \frac{\partial L_u}{\partial f_i} \). The obtained uncertainty gradients are passed through a gradient reversal layer, giving us the certainty gradients, i.e., \( -\frac{\partial L_u}{\partial f_i} \).

\[
\nabla_y^t = -\lambda \frac{\partial L_u}{\partial f_i} \frac{\partial L_y}{\partial f_i}
\]
The positive sign of gradient $\nabla_y''$ indicates that the attention certainty is activated on these regions and vice-versa. It can be expressed as:

$$\nabla_y'' = \text{ReLU}(\nabla_y') + \gamma \text{ReLU}(-\nabla_y') \tag{7}$$

We apply a ReLU activation function on the product of gradients of the attention map and the gradients of certainty as we are only interested in attention regions that have a positive influence on interested answer class, i.e. attention regions whose intensity should be increased in order to increase answer class probability $y_c$, whereas negative values are multiplied by $\gamma$ (large negative number) as the negative attention regions are likely to belong to other categories in image. As expected, without this ReLU, localization maps sometimes highlights more than just the desired class and achieves lower localization performance. Then we normalize $\nabla_y''$ to get attention regions which are highly activated and giving more weight to certain regions and is expressed as:

$$\nabla_y'' = \frac{\nabla_y''}{\sum_u \sum_v (\nabla_y'')_{uw}} \tag{8}$$

Images with higher uncertainty are equivalent to having lower certainty, so the certain regions of these images should have lower attention values. We use residual gradient connection to obtain the final gradient, which is the sum of gradient mask of $L_y$ (with respect to attention feature) and the gradient certainty mask $\nabla_y''$ and is given by:

$$\frac{\partial L_y}{\partial f_i} = \frac{\partial L_y}{\partial f_i} + \nabla_y'' \tag{9}$$

Where $\frac{\partial L_y}{\partial f_i}$ is the gradient mask of $L_y$ when gradients are taken with respect to attention feature. More details are given in the Algorithm 1.

### 4.2. Cost Function

We estimate aleatoric uncertainty in logits space by perturbing each logit using the variance obtained from data. The uncertainty present in the logits value can be minimized using cross-entropy loss on Gaussian distorted logits, as shown in the Equation 10. The distorted logit is obtained using a Gaussian multivariate function, having positive diagonal variance. To stabilize the training process we add an additional term to the uncertainty loss, calling it Variance Equalizer (VE) loss, $L_{VE}$.

$$L_{VE} = \exp(\sigma_f^2) - \exp(\sigma_0^2) \tag{10}$$

where $\sigma_0$ is a constant. The uncertainty distorted loss (UDL) is the difference between the typical cross-entropy loss and the aleatoric/predictive loss estimated in the Equation 10. The scalar difference is passed to an activation function to enhance the difference in either direction and is given by:

$$L_{UDL} = \left\{ \begin{array}{l} \alpha(\exp[L_p - L_\gamma] - 1), \quad \text{if } [L_p - L_\gamma] < 0. \\ \left[L_p - L_\gamma\right], \quad \text{otherwise.} \end{array} \right. \tag{11}$$

By putting this constraint, we ensure that the predictive uncertainty loss does not deviate much from the actual cross-entropy loss. The total uncertainty loss is the combination of Aleatoric (or prediction uncertainty loss), Uncertainty Distorted Loss, and Variance equalizer loss.

$$L_u = L_p + L_{VE} + L_{UDL} \tag{12}$$

The final cost function for the network combines the loss obtained through uncertainty (aleatoric or predictive) loss $L_u$ for the attention network with the cross-entropy.

The cost function used for obtaining the parameters $\theta_f$ of the attention network, $\theta_c$ of the classification network, $\theta_y$ of the prediction network and $\theta_u$ for uncertainty network is as follows:

$$C(\theta_f, \theta_c, \theta_y, \theta_u) = \frac{1}{n} \sum_{j=1}^{n} L_y(\theta_f, \theta_c, \theta_y) + \eta L_u(\theta_f, \theta_c, \theta_u)$$

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

$$\langle \hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u \rangle = \arg \max_{\theta_f, \theta_c, \theta_y, \theta_u} \left( C(\theta_f, \theta_c, \theta_y, \theta_u) \right) \tag{13}$$
5. Experiments

We evaluate the proposed GCA methods and have provided both quantitative analysis and qualitative analysis. The former includes: i) Ablation analysis of proposed models (Section 5.2), ii) Analysis of uncertainty effect on answer predictions (Figure 5(a,b)), iii) Differences of Top-2 softmax scores for answers for some representative questions (Figure 5(c,d)) and iv) Comparison of attention map of our proposed uncertainty model against other variants using Rank correlation (RC) and Earth Mover Distance (EMD) [3] as shown in Table 3 for VQA-HAT 8 and in Table 2 for VQA-X 19. Finally, we compare PGCA with state of the art methods as mentioned in Section 5.4. Qualitative analysis includes visualization of certainty activation maps for some representative images as we move from our basic model to the P-GCA model. (Section 5.6)

5.1. Datasets

VQA-v1 [2]: We conduct our experiments on VQA benchmark VQA-v1 2 dataset, which contains human-annotated questions and answers based on images on MS-COCO dataset. This dataset includes 2,04,721 images in total, out of which 82,783 images are for training, 40,504 images for validation, and 81,434 images for testing. Each image is associated with three questions, and each question has ten possible answers. There are 248349 Question-Answer pairs for training, 121512 pairs for validation, and 244302 pairs for testing.

VQA-v2 [18]: We provide benchmark result on VQA-v2 18 dataset. This dataset removes bias present in VQA-v1 by adding a conjugate image pair. It contains 443,757 image-question pairs on the training set, 214,354 pairs on the validation set and 447,793 pairs on the test set, which is more than twice the first version. All the questions and answers pairs are annotated by human annotators. The benchmark results on VQA-v2 dataset is presented in Table 5.

VQA-HAT [8]: To compare our attention map with human-annotated attention maps, we use VQA-HAT [8] dataset. This dataset is developed for image de-blurring for answering the visual question. It contains 58475 human-annotated attention maps out of 248349 training examples and includes three sets of 1374 human-annotated attention maps out of 121512 validation examples of question image pairs in the validation dataset. This dataset is developed for VQA-v1 only.

5.2. Ablation Analysis for Uncertainty

Our proposed GCA model’s loss consists of undistorted and distorted loss. The undistorted loss is the Standard Cross-Entropy (SCE) loss. The distorted loss consists of uncertain loss (either aleatoric uncertainty loss (AUL), or predictive uncertainty loss (PUL)), Variance Equalizer (VE) loss and Uncertainty Distorted loss (UDL). In the first block of the Table 1 we report the results when these losses are used individually. (Only SCE loss is there in the Baseline). We use a variant of the MCB [14] model as our baseline method. As seen, PUL, when used individually, outperforms the other 4. This could be attributed to PUL guiding the model to minimize both the data and the model uncertainty. The second block of the Table 1 depicts the results when we tried while combining two different individual losses. The model variant, which is guided using the combination of PUL and UDL loss performs best among the five variants. Then finally, after combining (AUL+UDL+VE+SCE), denoting it as A-GCA model and combining (PUL+UDL+VE+SCE), indicating it as P-GCA, we report an improvement of around 2.5% and 2.7% accuracy score respectively.

Further, we plotted Predictive uncertainty (Figure 5(a,b)) of some randomly chosen samples against the Classification error (error=log \frac{1}{1-p}, where p is the probability of misclassification). As seen, when the samples are correct, they are also certain and have less Classification Error (CE). To visualize the direct effect of decreased uncertainty, we plotted (Figure 5(c, d)). It can be seen that how similar classes like (glasses, sunglasses) and (black, gray), etc., thus leading to uncertainty, got separated more in the logit space in the proposed model.

5.3. Analysis of Attention Maps

We compare attention maps produced by our proposed GCA model, and its variants with the base model and re-
Figure 5. (a) Uncertainty vs Classification Error plots for our network for 20,000 randomly sampled images. We drew 25 samples of each image using Monte-Carlo sampling from the distribution. (b) Plots showing frequency of samples vs Uncertainty and frequency of samples vs Classification error respectively (c) Distance between the Top 2 Softmax scores for some Questions of type other than yes/no (d) Distance between the Top 2 Softmax scores for some Questions of type yes/no (Questions corresponding to (c) and (d) could be found in supplementary.)

Table 3. Ablation analysis and SOTA between HAT [8] attention and generated attention mask

<table>
<thead>
<tr>
<th>Model</th>
<th>RC(↑)</th>
<th>EMD(↓)</th>
<th>CD(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN [8]</td>
<td>0.2432</td>
<td>0.4013</td>
<td>–</td>
</tr>
<tr>
<td>CoAtt-W [33]</td>
<td>0.246</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CoAtt-P [33]</td>
<td>0.256</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CoAtt-Q [33]</td>
<td>0.264</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DVQA(K=1) [37]</td>
<td>0.328</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Baseline (MCB)</td>
<td>0.2790</td>
<td>0.3931</td>
<td>–</td>
</tr>
<tr>
<td>VE (ours)</td>
<td>0.2832</td>
<td>0.3931</td>
<td>0.1013</td>
</tr>
<tr>
<td>UDL (ours)</td>
<td>0.2850</td>
<td>0.3914</td>
<td>0.1229</td>
</tr>
<tr>
<td>AUL (ours)</td>
<td>0.2937</td>
<td>0.3867</td>
<td>0.1502</td>
</tr>
<tr>
<td>PUL(ours)</td>
<td>0.3012</td>
<td>0.3805</td>
<td>0.1585</td>
</tr>
<tr>
<td>PUL + VE (ours)</td>
<td>0.3139</td>
<td>0.3851</td>
<td>0.1631</td>
</tr>
<tr>
<td>PUL + UDLI(ours)</td>
<td>0.3243</td>
<td>0.3824</td>
<td>0.1630</td>
</tr>
<tr>
<td>A-GCA (ours)</td>
<td>0.3311</td>
<td>0.3784</td>
<td>0.1683</td>
</tr>
<tr>
<td>P-GCA (ours)</td>
<td>0.3341</td>
<td>0.3721</td>
<td>0.1710</td>
</tr>
<tr>
<td>Human [8]</td>
<td>0.623</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 4. SOTA: Open-Ended VQA1.0 accuracy on test-dev and Table-5 respectively. Also, we compare with our gradient certainty explanation with human explanation present in VQA-v2 dataset for the various model as mentioned in Table-2. This human explanation mask only available for VQA-v2 dataset. We observe that our attention (P-GCA) mask performs better than others as well. The evaluation methods for VQA dataset and HAT dataset are provided in supplementary material.

5.4. 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 [8] 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-5. We use a variant of MCB [14] model as our baseline method. We obtain an improvement of around 5.2% using A-GCA model and 5.51% using P-GCA model in terms of rank correlation with human attention. From this, we justify that our attention map is more similar to human attention map. We also compare with the baselines on the answer accuracy on VQA-v1 [2] dataset as shown in Table-4. We obtain an improvement of around 2.7% over the comparable MCB baseline. Our MCB based model A-GCA and P-GCA improves by 0.9% and 1.1% ac-
Tables 5. SOTA: Open-Ended VQA2.0 accuracy on test-dev

curacy as compared to state of the art model DVQA [37] on VQA-v1. However, using a saliency-based method [21] 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), as noted by [8]. 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 the tasks. We provide state of the art results for VQA-v2 in Table-5. This table shows that using GCA method, the VQA result improves. We have provided more results for attention map visualization for both types of uncertainty methods here.

5.5. Training and Model Configuration

We trained the P-GCA model using classification loss and uncertainty 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 uncertainty 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 uncertainty model.

5.6. Qualitative Result

We provide attention map visualization of all models for 5 example images, as shown in Figure-6. The first raw, the baseline model misclassifies the answer due to high uncertainty value, that gets resolved by our methods (P-GCA). We can see how attention is improved as we go from our baseline model (MCB) to the proposed Gradient Certainty model (P-GCA). For example, in the first row, MCB is unable to focus on any specific portion of the image, but as we go towards the right, it focuses the cup bottom, (indicated by intense orange color in the map). Same can be seen for other images also. We have visualized Grad-CAM maps

<table>
<thead>
<tr>
<th>Models</th>
<th>All</th>
<th>Y/N</th>
<th>Num</th>
<th>Oth</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN-2 [53]</td>
<td>56.9</td>
<td>74.1</td>
<td>35.5</td>
<td>44.5</td>
</tr>
<tr>
<td>MCB [14]</td>
<td>64.0</td>
<td>78.8</td>
<td>38.3</td>
<td>53.3</td>
</tr>
<tr>
<td>Bottom [11]</td>
<td>65.3</td>
<td>81.8</td>
<td>44.2</td>
<td>56.0</td>
</tr>
<tr>
<td>DVQA [37]</td>
<td>56.9</td>
<td>7.42</td>
<td>35.5</td>
<td>44.5</td>
</tr>
<tr>
<td>MLB [27]</td>
<td>66.3</td>
<td>83.6</td>
<td>44.9</td>
<td>56.3</td>
</tr>
<tr>
<td>Bottom [11]</td>
<td>68.0</td>
<td>83.1</td>
<td>51.6</td>
<td>58.9</td>
</tr>
<tr>
<td>DVQA [37]</td>
<td>65.9</td>
<td>82.4</td>
<td>43.2</td>
<td>56.8</td>
</tr>
<tr>
<td>MLB [27]</td>
<td>66.3</td>
<td>83.6</td>
<td>44.9</td>
<td>56.3</td>
</tr>
<tr>
<td>Counter [54]</td>
<td>67.5</td>
<td>84.3</td>
<td>47.1</td>
<td>57.9</td>
</tr>
<tr>
<td>BAN [26]</td>
<td>69.5</td>
<td>85.3</td>
<td>50.9</td>
<td>60.2</td>
</tr>
<tr>
<td>P-GCA + SAN (ours)</td>
<td>59.2</td>
<td>75.7</td>
<td>36.6</td>
<td>46.8</td>
</tr>
<tr>
<td>P-GCA + MCB (ours)</td>
<td>65.7</td>
<td>79.6</td>
<td>40.1</td>
<td>54.7</td>
</tr>
<tr>
<td>P-GCA + Counter (ours)</td>
<td>69.2</td>
<td><strong>85.4</strong></td>
<td>50.1</td>
<td>59.4</td>
</tr>
</tbody>
</table>

Figure 6. 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 the second column, it shows the activation map for baseline (MCB) Attention Network, Aleatoric (AUL), Predictive (PUL), A-GCA, P-GCA based approach respectively.

to support our hypothesis that Grad-CAM is a very good way for visualizing what the network learns as it can focus on right portions of the image even in the baseline model (MCB), and therefore, can be used as a tutor to improve attention maps. For example, in MCB it tries to focus on the right portions but with the focus to other points as well. However, in our proposed model, visualization improves as the models focuses only on the required portion.

6. Conclusion

In this paper, we provide a method that uses gradient-based certainty attention regions to obtain improved visual question answering. The proposed method yields improved uncertainty estimates that are correspondingly more certain or uncertain, show consistent correlation with misclassification and are focused quantitatively on better attention regions as compared to other states of the art methods. The proposed architecture can be easily incorporated in various existing VQA methods as we show by incorporating the method in SAN [53] and MCB [14] models. The proposed technique could be used as a general means for obtaining improved uncertainty and explanation regions for various vision and language tasks, and in future, we aim to evaluate this further for other tasks such as ‘Visual Dialog’ and image captioning tasks.

7. Acknowledgment

We acknowledge the help provided by Delta Lab members and our family who have supported us in this research activity.
References


8. supplementary

This is the supplementary material for the paper ‘U-CAM: Visual Explanation using Uncertainty based Class Activation Maps’.
A. Questions for Figure 5(c, d)

Table 6 and 7 have the questions corresponding to ID’s present in the Figure 5 of the main paper.

<table>
<thead>
<tr>
<th>Question</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>What does the person in this picture have on his face?</td>
<td>1</td>
</tr>
<tr>
<td>How many baby elephants are there?</td>
<td>2</td>
</tr>
<tr>
<td>What is in the bowl?</td>
<td>3</td>
</tr>
<tr>
<td>Is the television on or off?</td>
<td>4</td>
</tr>
<tr>
<td>What color is the walk light?</td>
<td>5</td>
</tr>
<tr>
<td>Which way is its head turned?</td>
<td>6</td>
</tr>
<tr>
<td>How many people are riding on each bike?</td>
<td>7</td>
</tr>
<tr>
<td>What animal is in this picture?</td>
<td>8</td>
</tr>
<tr>
<td>What color is the road?</td>
<td>9</td>
</tr>
<tr>
<td>What color is the boy’s hair?</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 6. Reference for the Figure 5(c) of the main paper.

Algorithm 2: Rank Correlation Procedure

1: procedure :(Initialization)
2: \( P_H \) : Probability distribution of Human Attention Map
3: \( P_M \) : Probability distribution of our model
4: Rank:
5: Compute Rank of \( P_H \) : \( R_H \)
6: Compute Rank of \( P_M \) : \( R_M \)
7: Rank Difference :
8: Compute difference in rank between \( R_H \) & \( R_M \) :
\( Rank_{Diff} \)
9: Compute square of rank difference \( Rank_{Diff} \)
\( : S_{Rank_{Diff}} \)
10: Rank Correlation:
11: Compute Dimension of \( P_M \) : \( N \)
12: Compute Rank Correlation using :
\[ R_{Cor} = 1 - \frac{6 * S_{Rank_{Diff}}}{N^3 - N} \]
13: end procedure

B. Evaluation Methods

Accuracy: VQA dataset has 3 type of answers: yes/no, number and other. The evaluation is carried out using two test splits, i.e test-dev and test-standard. The question in corresponding test split are of two types: Open-Ended and Multiple-choice. Our model generates a single word answer on the open ended task. For each question there are 10 candidate answer provided with their respective confidence