U-CAM: Visual Explanation using Uncertainty based Class Activation Maps

Badri N. Patro Mayank Lunayach Shivansh Patel Vinay P. Namboodiri
Indian Institute of Technology, Kanpur
{ badri,mayankl,shivp,vinaypn }@iitk.ac.in

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, [15, 16]. 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 [15], it became feasible to obtain uncertainty estimates in a computationally efficient manner. This was further extended to data uncertainty and model uncertainty based estimates [23]. 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 [24].

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.

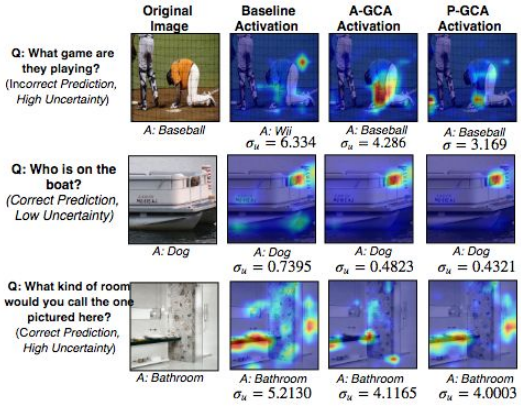

Figure 1. The figure shows the activation maps for baseline (MCB [13]) and our models (A-GCA and P-GCA). In the first example, the baseline model had predicted the wrong answer and had high uncertainty in prediction. (σ_u denotes uncertainty, see Section 3). Our model gave a correct answer while also minimizing the uncertainty (thus leading to an improved visual explanation).

Our method combines techniques from both the explanation [9] and uncertainty [23] 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.

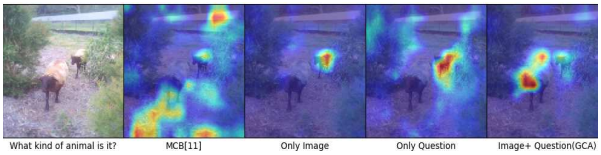

Figure 2. The first column is the original image. 2nd, 3rd, 4th, and 5th columns represent the baseline attention, attention when only image uncertainty was minimized, attention when only question uncertainty was minimized, attention when both image and question uncertainties were minimized (proposed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

2. Related work

The task of Visual question answering [34, 2, 39, 17, 35] is well studied in the vision and language community, but it has been relatively less explored for providing explanation[40] for answer prediction. Recently, lot of works that focus on explanation models, one of that is image captioning for provide basic explanation for image [5, 11, 27, 44, 47, 21, 50, 10, 7, 19, 51]. [37] has proposed an exemplar-based explanation method for generating question based on the image. Similarly, [38] has suggested a discriminator based method to obtain an explanation for paraphrase generation in text. In VQA, [54][49] have proposed interesting methods for improving attention in the question. Work that explores image and question jointly and is based on hierarchical co-attention is [32]. [41, 52, 31, 36] have proposed attention-based methods for the explanation in VQA, which use question to attend over specific regions in an image. [13, 26, 25] have suggested exciting works that advocate multimodal pooling and obtain close to state of the art in VQA. [36] has proposed an exemplar-based explanation method to improve attention in VQA. We can do systematic comparison of image-based attention while correlating with human attention maps as shown by [8]. Other methods [28, 30] explore the correlation between the distributions of the dataset. The computational efficiency of these deep learning models can be improved by [42].

Recently a lot of researchers have focused on estimating uncertainty in the deep models. [6] has first proposed a method to learn uncertainty in the weights of the neural network. Kendall *et.al.* [22] has proposed method to measure model uncertainty for image segmentation task. They observed that softmax probability function approximates relative probability between the class labels, but does not provide information about the model's uncertainty. The work by [15, 12] estimates model uncertainty of the deep network (CNN, RNN) with the help of dropout [45]. [46] has estimated uncertainty for batch normalized deep networks. [23, 24, 43] have mainly decomposed predictive uncertainty into two major types, namely aleatoric and epistemic uncertainty, which capture uncertainty about the predicted model and uncertainty present in the data itself. [33] suggested a method to measure predictive uncertainty with the help of model and data uncertainty. Recently, [29] proposed a certainty method to bring two data distributions close for the domain adaption task. Here, our objective is to analyze and minimize the uncertainty in attention mask to predict answer in VQA. In our approach, We are proposing a gradient-based certainty explanation mask which minimizes uncertainty in attention regions to improve the correct answer's 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 σ_i^2 as shown in Figure-3. To capture Aleatoric uncertainty [23], we learn the observational noise parameter σ_i for each input point x_i. Then, Aleatoric uncertainty, $(\sigma_a^2)_i$ is estimated by applying softplus function on the output logit variance. This is given by,

$$(\sigma_a^2)_i = Softplus(\sigma_i^2) = log(1 + exp(\sigma_i^2)) \qquad (1)$$

For calculating the aleotoric uncertainty loss, we perturb the logit value (y_i) with Gaussian noise of variance $(\sigma_a^2)_i$ (diagonal matrix with one element corrosponding to each logits value) before the softmax layer. The logits reparameterization trick [24] and [14] combines $\hat{y}_{i,c}$ and σ_i to give

$\mathcal{N}(\hat{y}_{i,c}, \sigma_i^2)$. We then obtain a loss with respect to ground truth. It is expressed as:

$$\hat{y}_{i,c,t} = y_{i,c} + \epsilon_t * \sigma_i^2, where \quad \epsilon_t \sim \mathcal{N}(0, I) \tag{2}$$

$$\mathcal{L}_a = \sum_i \log \frac{1}{T} \sum_t \exp\left(\hat{y}_{i,c,t} - \log \sum_{c^{'}} \exp \hat{y}_{i,c^{'},t}\right) \tag{3}$$

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 [33] 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 $p(y_{i,t})$. That is,

$$O(\hat{y}_{i,t}) = G^t(x_i) \quad v_{i,t}^a = Softplus(U^t(\hat{y}_{i,t}))$$
$$p(\hat{y}_{i,c}|x_i, X_I) = (\frac{1}{T}\sum_{t=1}^{T} SoftmaxO(\hat{y}_{i,t}))_c$$

where c is the answer class, $G^t \sim G, U^t \sim U$ and $v_{i,t}^a$ 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:

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

The predictive uncertainty contains entropy and aleatoric variance when it's expectation is taken across T number of Monte Carlo simulations:

$$\sigma_p^2 = H(\hat{y}_i) + \frac{1}{T}\sum_{t=1}^{T} v_{i,t}^a \tag{5}$$

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 2, we can replace σ_a^2 with predictive uncertainty σ_p^2 (mentioned above in Equation 5) to get the predictive uncertainty loss (PUL).

4. Method

Task: We solve for VQA [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- 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, $\hat{A} = \underset{A \in \Omega}{argmax} P(A|I, Q, \theta)$, where Ω is a set of all possible answers, I is the image, Q is the corresponding question, and θ is representing the parameters of the network.

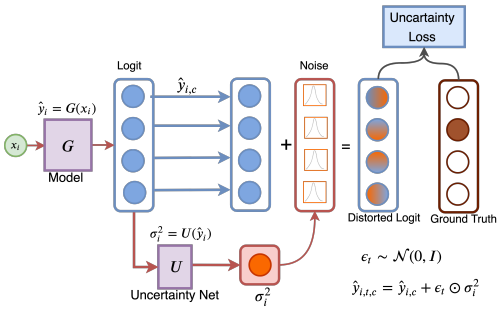

Figure 3. Illustration of Uncertainty Loss

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 θ_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 θ_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 4. Various kinds of attention networks have been proposed in the literature. In this paper, we tried with SAN [52] and MCB [13]. 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.

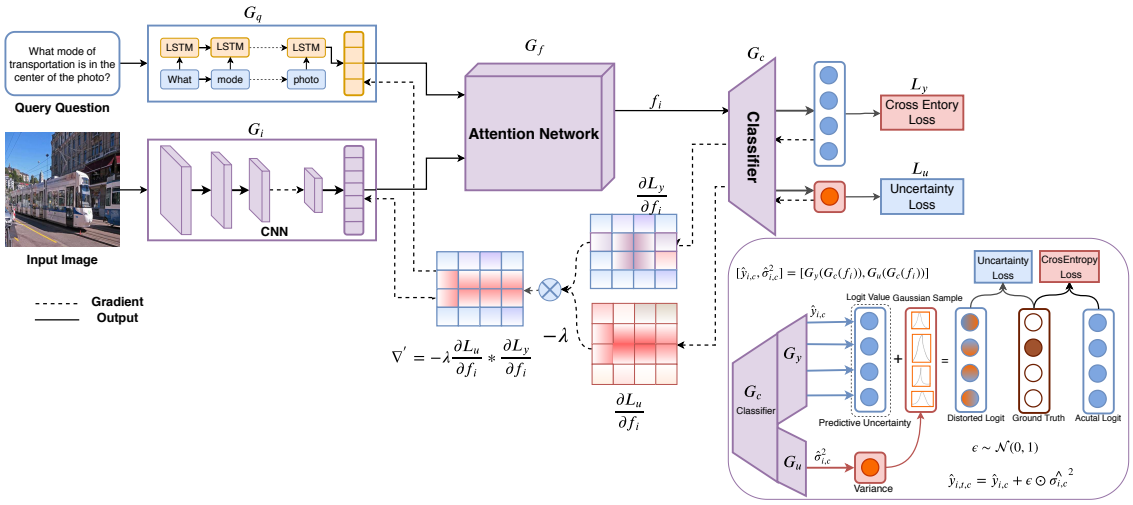

Figure 4. Illustration of model Gradient-based Certainty Attention Mask (GCA) and its certainty mask. We obtain image feature and question feature using CNN and LSTM, respectively. We then obtain attention mask using these features, and classification of the answer is done based on the attended feature.

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_y, 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, σ_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 $(\sigma_a^2)_i$ 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, τ is the precision, f_i is the attended fused input, and $G_y(G_c(.))$ 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, τ^{-1}. We make sure that $\tau(.)$ is a positive or positive definite matrix (in case of Multivariate) by using the logit reparameterization trick [24, 14]. Finally, we then obtain an aleatoric loss, $\mathcal{L}_a$ with respect to ground truth as mentioned in the Equation- 3. 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 $\mathcal{L}_a$, we combine $\mathcal{L}_{VE}$ and $\mathcal{L}_{UDL}$ as mentioned in the Equation 10 and 11 respectively to get total uncertainty loss $\mathcal{L}_u$. The classifier is trained by jointly minimizing both the classification loss, $\mathcal{L}_y$ and the uncertainty loss, $\mathcal{L}_u$. In Equation 2, we can replace σ_a^2 with predictive uncertainty σ_p^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 ($\frac{\partial \mathcal{L}_y}{\partial f_i}$). But that explanation does not take the model and data uncertainties into the account. We improve this explanation using the certainty gradients ($-\frac{\partial \mathcal{L}_u}{\partial f_i}$). 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 1.

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 $\mathcal{L}_u$ i.e. $\frac{\partial \mathcal{L}_u}{\partial f_i}$. The obtained uncertainty gradients are passed through a gradient reversal layer, giving us the certainty gradients, i.e., $-\frac{\partial \mathcal{L}_u}{\partial f_i}$.

$$\nabla_y^{'} = -\lambda \frac{\partial \mathcal{L}_u}{\partial f_i} * \frac{\partial \mathcal{L}_y}{\partial f_i} \tag{6}$$

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:

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

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 γ (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:

$$\nabla_y^{'''} = \frac{(\nabla_y^{''})_{u,v}}{\sum_u \sum_v (\nabla_y^{''})_{uv}} \tag{8}$$

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 $\mathcal{L}_y$ (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 $\mathcal{L}_y$ 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 3. The distorted logit is obtained using a Gaussian multivariate function, having positive diagonal variance. To stabilize the training process [14], we add an additional term to the uncertainty loss, calling it Variance Equalizer(VE) loss, $\mathcal{L}_{VE}$.

$$\mathcal{L}_{VE} = \exp(\sigma_i^2) - \exp({\sigma_0}^2) \tag{10}$$

where σ_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 3. The scalar difference is passed to an activation function to enhance the difference in either direction and is given

Algorithm 1 Gradient Certainty base Attention (GCA)

procedure GCA(I, Q)
  Input: Image X_I, Question X_Q
  Output: Answer y_c
  while loop **do**
    Attention features $G_f(G_i(X_I), G_q(X_Q)) \leftarrow f_i$
    Answer Logit $G_y(G_c(f_i)) \leftarrow \hat{y}$
    Data Uncertainty $G_v(G_c(f_i)) \leftarrow \sigma_A^2$
    if A-GCA **then**:
      $\sigma_W^2 = \sigma_A^2$
    else if P-GCA **then**:
      $\sigma_W^2 = \sigma_A^2 + H(\hat{y}_{i,t})$, (Ref: eq- 5)
    end if
    Ans cross entropy $\mathcal{L}_y \leftarrow loss(\hat{y}, y)$
    Variance Equalizer $\mathcal{L}_{VE} := \sum ReLU(\exp^{\sigma_w^2} - \exp^{I})$,
    while $t = 1 : \#MC - Samples$ **do**
      Sample $\epsilon_t^w \sim \mathcal{N}(0, \sigma_W^2)$
      Distorted Logits: $\hat{y}_{i,t} = \epsilon_t^w + \hat{y}_i$
      Gaussian Cross Entropy $L_p = -\sum y \log p(\hat{y_d}|F(.))$
      Distorted Loss : $\mathcal{L}_{UDL} = \exp(\mathcal{L}_y - \mathcal{L}_p)^2$
      Aleatoric uncertainty loss $\mathcal{L}_u = \mathcal{L}_p + \mathcal{L}_{VE} + \mathcal{L}_{UDL}$
    end while
    Compute Gradients w.r.t f_i, $\nabla_y = \frac{\partial \mathcal{L}_y}{\partial f_i}$, $\nabla_u = \frac{\partial \mathcal{L}_u}{\partial f_i}$
    Certainty Gradients $\nabla_u^{'} = -\lambda \nabla_u * \nabla_y$
    Certainty Activation $\nabla_u^{''} = ReLU(\nabla_u^{'}) + \gamma ReLU(-\nabla_u^{'})$
    Final Certainty Gradients $\nabla_u^{'''} = softmax(\nabla_u^{''})$
    Final Attention Gradient $\nabla_y = \nabla_y + \nabla_u^{'''}$
    update $\theta_f \leftarrow \theta_f - \eta \nabla_y$
  end while
end procedure

by :

$$\mathcal{L}_{UDL} = \begin{cases} \alpha(\exp^{[\mathcal{L}_p - \mathcal{L}_y]} - 1), & \text{if } [\mathcal{L}_p - \mathcal{L}_y] < 0. \\ [\mathcal{L}_p - \mathcal{L}_y], & \text{otherwise.} \end{cases} \tag{11}$$

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.

$$\mathcal{L}_u = \mathcal{L}_p + \mathcal{L}_{VE} + \mathcal{L}_{UDL} \tag{12}$$

The final cost function for the network combines the loss obtained through uncertainty (aleatoric or predictive) loss $\mathcal{L}_u$ for the attention network with the cross-entropy.

The cost function used for obtaining the parameters θ_f of the attention network, θ_c of the classification network, θ_y of the prediction network and θ_u for uncertainty network is as follows:

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

where n is the number of examples, and η 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

$$(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u) = \arg \max_{\theta_f, \theta_c, \theta_y, \theta_u} (C(\theta_f, \theta_c, \theta_y, \theta_u)) \tag{13}$$

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 [18] . 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 [17]: We provide benchmark result on VQA-v2 [17] 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 uncertainly 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 [13] 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.

Models	All	Yes/No	Number	Others
Baseline	63.8	82.2	37.3	54.2
VE	64.1	82.3	37.2	54.3
UDL	64.4	82.6	37.2	54.5
AUL	64.7	82.9	37.4	54.6
PUL	64.9	83.0	37.5	54.6
UDL+VE	64.8	82.8	37.4	54.5
AUL+VE	65.0	83.3	37.8	54.7
PUL+ VE	65.3	83.3	37.9	54.9
AUL +UDL	65.6	83.3	37.6	55.0
PUL + UDL	65.9	83.7	37.8	55.2
A-GCA (ours)	66.3	84.2	38.0	55.5
P-GCA (ours)	**66.5**	**84.7**	**38.4**	**55.9**

Table 1. Ablation analysis for Open-Ended VQA1.0 accuracy on test-dev

Model	RC(↑)	EMD(↓)
Baseline	0.3017	0.3825
Deconv ReLU	0.3198	0.3801
Guided GradCAM	0.3275	0.3781
Aleatoric mask	0.3571	0.3763
Predictive mask	**0.3718**	**0.3714**

Table 2. Rank Correlation for explanation mask in VQA-X [18] data with our explanation mask using Grad-Cam.

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 it's variants with the base model and re-

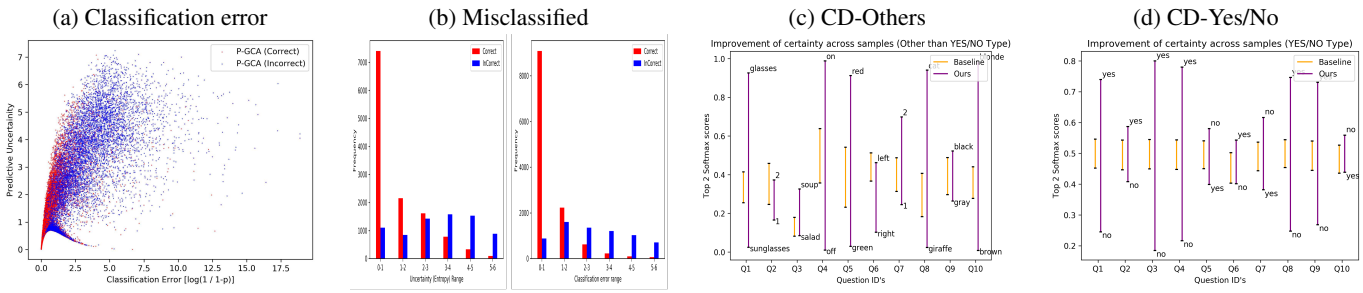

Figure 5. (a) Uncertainty vs Classification Error plots for our network for 20,000 randomly sampled images. We drew 25 samples of each image using Monte-Carlo sampling from the distribution. (b) Plots showing frequency of samples vs Uncertainty and frequency of samples vs Classification error respectively (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.)

Model	RC(↑)	EMD(↓)	CD(↑)
SAN [8]	0.2432	0.4013	–
CoAtt-W[32]	0.246	–	–
CoAtt-P [32]	0.256	–	–
CoAtt-Q[32]	0.264	–	–
DVQA(K=1)[36]	0.328	–	–
Baseline (MCB)	0.2790	0.3931	–
VE (ours)	0.2832	0.3931	0.1013
UDL (ours)	0.2850	0.3914	0.1229
AUL (ours)	0.2937	0.3867	0.1502
PUL(ours)	0.3012	0.3805	0.1585
PUL + VE (ours)	0.3139	0.3851	0.1631
PUL + UDL(ours)	0.3243	0.3824	0.1630
A-GCA (ours)	0.3311	0.3784	0.1683
P-GCA (ours)	**0.3341**	**0.3721**	**0.1710**
Human [8]	0.623	–	–

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

ports them in Table-3. Rank correlation and EMD score are calculated for the produced attention map against human-annotated attention (HAT) maps [8]. In the table, as we approach the best-proposed GCA model, Rank correlation (RC) is increasing. EMD is also decreasing (Lower the better) as we move towards GCA. To verify our intuition, that we can learn better attention mask by minimizing the uncertainty present in the attention mask, we start with VE and observe that both rank correlation and answer accuracy increase by 0.42 and 0.3 % from baseline respectively. We also observe that with UDL, AUL, and PUL based loss minimization technique, both RC and EMD improves, as shown in the Table- 3. Aleatoric-GCA (A-GCA) improves 5.21% in terms of RC and 2.5% in terms of accuracy. Finally, the proposed Predictive-GCA (P-GCA), which is modeled to consider both data and the model uncertainty improves the RC by 5.51% and accuracy by 2.7% as shown in the Table- 3 and Table- 1. Since HAT maps are only available for VQA-v1 dataset, thus, this ablation analysis has been performed only for VQA-v1. We also providing SOTA results for VQA-v1 and VQA-v2 dataset as shown in Table- 4

Models	All	Y/N	Num	Oth
DPPnet [35]	57.2	80.7	37.2	41.7
SMem[[49]]	58.0	80.9	37.3	43.1
SAN [52]	58.7	79.3	36.6	46.1
DMN[48]	60.3	80.5	36.8	48.3
QRU(2)[31]	60.7	82.3	37.0	47.7
HieCoAtt [32]	61.8	79.7	38.9	51.7
MCB [13]	64.2	82.2	37.7	54.8
MLB [26]	65.0	84.0	37.9	54.7
DVQA[36]	65.4	83.8	38.1	55.2
P-GCA + SAN (ours)	60.4	80.7	36.6	47.9
A-GCA + MCB (ours)	66.3	84.2	38.0	55.5
P-GCA + MCB (ours)	**66.5**	**84.6**	**38.4**	**55.9**

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 3. We use a variant of MCB [13] 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-

Models	All	Y/N	Num	Oth
SAN-2[52]	56.9	74.1	35.5	44.5
MCB [13]	64.0	78.8	38.3	53.3
Bottom[[1]]	65.3	81.8	44.2	56.0
DVQA[36]	65.9	82.4	43.2	56.8
MLB [26]	66.3	83.6	44.9	56.3
DA-NTN [4]	67.5	84.3	47.1	57.9
Counter[53]	68.0	83.1	51.6	58.9
BAN[25]	**69.5**	85.3	**50.9**	**60.2**
P-GCA + SAN (ours)	59.2	75.7	36.6	46.8
P-GCA + MCB (ours)	65.7	79.6	40.1	54.7
P-GCA + Counter (ours)	69.2	**85.4**	50.1	59.4

Table 5. SOTA: Open-Ended VQA2.0 accuracy on test-dev

curacy as compared to state of the art model DVQA [36] on VQA-v1. However, using a saliency-based method [20] 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[1].

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

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

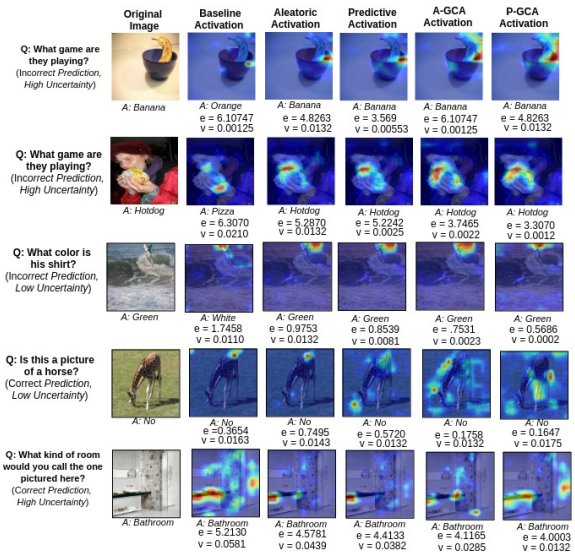

Figure 6. Examples with different approaches in each column for improving attention using explanation in a self-supervised manner. The first column indicates the given target image and its question and answer. Starting from the second column, it shows the activation map for baseline (MCB) Attention Network, 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 [52] and MCB [13] 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

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 8

[2] 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. 2, 3, 6, 7

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *stat*, 1050:26, 2017. 6

[4] 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. 8

[5] K Barnard, P Duygulu, and D Forsyth. N. de freitas, d. *Blei, and MI Jordan," Matching Words and Pictures", submitted to JMLR*, 2003. 2

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015. 2

[7] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015. 2

[8] 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. 2, 6, 7, 8

[9] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 2

[11] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010. 2

[12] Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*, 2017. 2

[13] 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. 1, 2, 3, 6, 7, 8

[14] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 2, 4, 5

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016. 1, 2

[16] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016. 1

[17] 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. 2, 6

[18] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 6

[19] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2

[20] 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. 8

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[22] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 2

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 1, 2

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. 2018. 1, 2, 4

[25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1571–1581, 2018. 2, 8

[26] 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. 2, 7, 8

[27] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer, 2011. 2

[28] Vinod Kumar Kurmi, Vipul Bajaj, Venkatesh K Subramanian, and Vinay P Namboodiri. Curriculum based dropout discriminator for domain adaptation. *arXiv preprint arXiv:1907.10628*, 2019. 2

[29] Vinod Kumar Kurmi, Shanu Kumar, and Vinay P Namboodiri. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 491–500, 2019. 2

[30] Vinod Kumar Kurmi and Vinay P Namboodiri. Looking back at labels: A class based domain adaptation technique. *arXiv preprint arXiv:1904.01341*, 2019. 2

[31] Ruiyu Li and Jiaya Jia. Visual question answering with question representation update (qru). In *Advances in Neural Information Processing Systems*, pages 4655–4663, 2016. 2, 7

[32] 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. 2, 7

[33] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018. 2, 3

[34] 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. 2

[35] 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. 2, 7

[36] Badri Patro and Vinay P. Namboodiri. Differential attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 7, 8

[37] Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay Namboodiri. Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4002–4012. Association for Computational Linguistics, 2018. 2

[38] Badri Narayana Patro, Vinod Kumar Kurmi, Sandeep Kumar, and Vinay Namboodiri. Learning semantic sentence embeddings using sequential pair-wise discriminator. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2715–2729, 2018. 2

[39] 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. 2

[40] 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. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[41] 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. 2

[42] Pravendra Singh, Vinay Kumar Verma, Piyush Rai, and Vinay P. Namboodiri. Hetconv: Heterogeneous kernel-based convolutions for deep cnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[43] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018. 2

[44] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014. 2

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 2

[46] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. *arXiv preprint arXiv:1802.06455*, 2018. 2

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 2

[48] 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. 7

[49] 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. 2, 7

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 2

[51] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 2

[52] 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. 2, 3, 7, 8

[53] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. 2018. 8

[54] 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. 2