

# Robust Explanations for Visual Question Answering

Badri N. Patro   Shivansh Patel   Vinay P. Namboodiri  
 Indian Institute of Technology, Kanpur  
 {badri , shi vp, vi naypn}@i i tk. ac. i n

## Abstract

*In this paper, we propose a method to obtain robust explanations for visual question answering(VQA) that correlate well with the answers. Our model explains the answers obtained through a VQA model by providing visual and textual explanations. The main challenges that we address are i) Answers and textual explanations obtained by current methods are not well correlated and ii) Current methods for visual explanation do not focus on the right location for explaining the answer. We address both these challenges by using a collaborative correlated module which ensures that even if we do not train for noise based attacks, the enhanced correlation ensures that the right explanation and answer can be generated. We further show that this also aids in improving the generated visual and textual explanations. The use of the correlated module can be thought of as a robust method to verify if the answer and explanations are coherent. We evaluate this model using VQA-X dataset. We observe that the proposed method yields better textual and visual justification that supports the decision. We showcase the robustness of the model against a noise-based perturbation attack using corresponding visual and textual explanations. A detailed empirical analysis is shown.*

## 1. Introduction

In this paper, we solve for obtaining robust explanations for visual question answering. Visual question answering is a semantic task that aims to answer questions based on an image. The practical implication for this task is that of an agent answering questions asked by, for instance, a visually impaired person that wants to know answers. An important aspect related to this model is the ability to reason whether the model is able to really understand and provide explanations for its answer. This aspect was investigated, for instance, by a recent method by [20], where the authors proposed a method to generate textual explanations and also provide localizations that contribute to the explanation for their answer. This was obtained by their approach which generates explanations based on the answer attention. How-

Figure 1: Illustration of proposed method: In case coherence of explanation and answer generation of VQA network is not enforced, a noise based perturbation will result in diverse answer and explanation being generated. This is shown in first row. In second row we illustrate that the proposed method ensures coherence and therefore is able to be robust to noise based perturbation.

ever, one drawback we observe for such an approach is that the explanation need not be correct. For instance, using a noise based perturbation on the image, we can have instances of answer and explanation being different. We solve this by jointly generating the answer and explanation. We further improve over this method by proposing a novel method that enhances the correlation by verifying that the answer and explanation do agree with each other. This is obtained through a collaborative correlated network. This concept is illustrated in figure 1 where we show that current methods can generate answers and explanations. However, these may diverge for an image corrupted using noise based perturbation. Our proposed method aids in generating robust explanations that are tolerant of such perturbation though our methods are not trained using such noise at the time of training.

We investigate various other ablations of the system, such as verifying whether the answer is correct or the explanation is correct or separately verifying that the answer and explanations are correct. These variations are illustrated in figure 2. We observe that jointly verifying that the answer and explanations are correct and agree with each other is

better than all the other variations. This is termed as a collaborative correlation module as the answer and explanation collaboratively are checked. This module not only aids in generating better textual explanations but also helps to generate better localization for the visual explanation. It also ensures the coherence of answer and explanation generation.

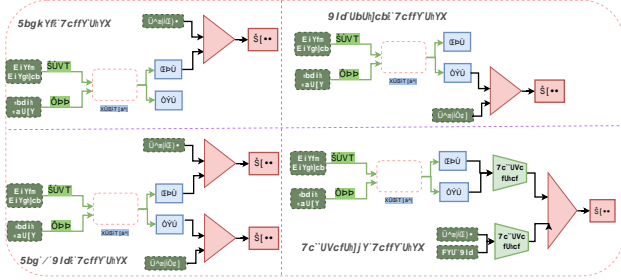


Figure 2: This figure shows variations of our methods. Answer Correlated only corrects the altered answers. Explanation Correlated only corrects the altered explanations. Answer and Explanation Correlated corrects answer and explanation separately. Collaborative Correlated Module jointly corrects both altered answer and explanation.

To summarize, through this paper we provide the following contributions:

- We investigate the tolerance of VQA systems against noise based perturbation and propose a method that is robust to such perturbation.
- We propose joint explanation and answer generation systems that are improved using a novel collaborative correlated network that ensures that the answer and explanations are coherent and correct.
- We illustrate the role of the correlated networks by providing corresponding attention maps that correlate well with human annotated attention maps.
- A detailed empirical analysis of variants of our model provides quantitative evidence of improvements over the current state of the art methods in terms of improved answers being tolerant to adversarial attacks and generating coherent textual and visual explanations.

## 2. Related Work

Extensive work has been done in the Vision and Language domain for solving image captioning [4, 12, 24, 47, 51, 23, 53, 11, 6, 22, 54], Visual Question Answering (VQA) [33, 30, 1, 44, 32, 36], Visual Question Generation (VQG) [35, 21, 40] and Visual Dialog [7, 2, 49, 50, 55]. Malinowski et al. [33] has proposed Visual question answering task, which answer natural language question based on the image. [1, 17] generalize this task with a large bias free dataset. Joint

embedding approach was proposed by [1, 30, 1, 44, 32, 36] where they combine image features with question features to predict answers. Attention-based approach comprises image-based attention, question-based attention and both image and question based attention. Recent work from [60, 14, 15, 52, 31, 45, 27, 38, 41] considers region-based image attention.

**Explanation:** Early textual explanation models spanned a variety of applications such as medicine [46], feedback for teaching [25] and were generally template based. Some methods find discriminative visual patches [9], [5] whereas others aim to understand intermediate features which are important for the end decisions [57], [10], [59]. Recently, the author [41] has proposed a new paradigm of providing visual explanation using uncertainty based class activation map. Furthermore, a variety of works proposed methods to visually explain decisions. The method that is closest to our work is the recent work by [20]. Their work aims at generating explanations and also providing visual justification for their answer. None of the methods checks robustness. To the best of our knowledge, we are raising a new novel issue related to the robustness of the prediction and explanation that has not been previously considered in the literature. Our work is inspired by this effort and aims to provide a more robust and coherent explanation generation, which is verified experimentally.

**Adversarial Methods:** Generative adversarial networks (GAN) [16] is an adversarial methods that consists of a generative model, G, which captures the data distribution, and a discriminative model, D, which estimates the probability of a sample as to whether it came from the training data or not. GANs are widely used to explain data distribution and various other tasks in vision domain [34, 42]. Also, there has been a lot of work on GANs in the field of natural language processing [34, 58, 19, 56, 18, 28]. Reed et al. [43] have proposed a model which combines vision with language to generate image from text. Li et al. [26] has proposed an advanced method to generate sequence of conversion about an image. Patro et al. [39] have proposed an adversarial method to improve explanation and attention using surrogate supervision method.

In this work, we propose a collaborative correlated module to generate both answer and textual explanation of that answer which will be tightly correlated with each other. We show that this module ensures that even if we do not train for noise based attacks, the enhanced correlation can ensure that the right explanation and answer can be generated.

## 3. Method

Interpretability of models through explanation does not consider the robustness and consistency of the provided explanations. While providing interpretability for understanding models is important, ensuring that the same are robust and consistent is also crucially important. We are specifically

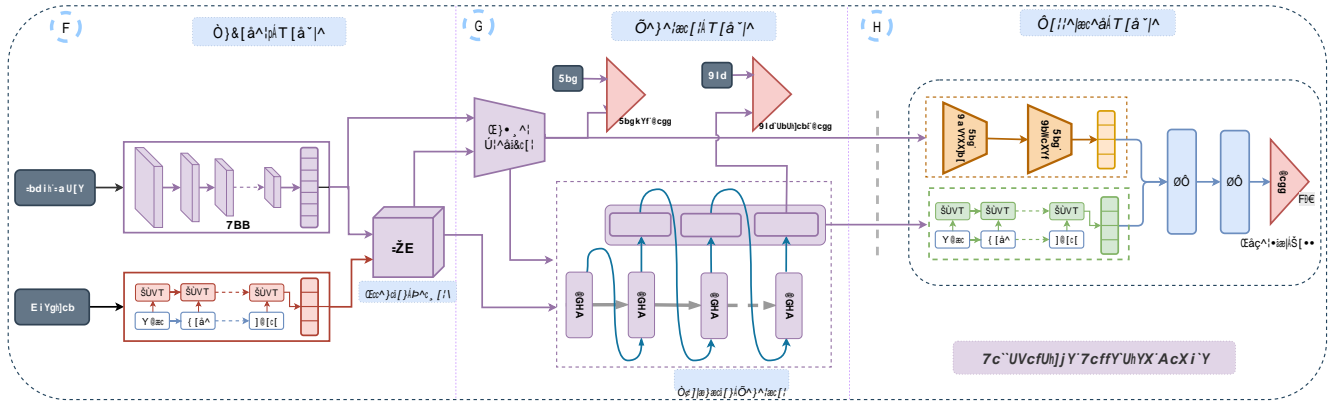


Figure 3: Illustration of Collaborative Correlated Model(CCM). The model receives input image feature and its question feature using CNN and LSTM respectively. then the model predict answer and generate explanation for the predicted answer. During training we ensure that the model collaborates with both answer and explanation features and learns jointly using adversarial fashion.

investigating this problem and provide a simple framework to solve for the same. The main focus of our approach is to ensure correctness and correlation in visual question answering(VQA). We propose a correlated network that learns joint embedding by collaborating with textual explanation and answer embedding. The key difference between our architecture and other existing VQA architectures is in the use of a mutually collaborating module for explanation and answering blocks in a joint adversarial mechanism. This is illustrated in figure 3. The other aspects of VQA and explanation are retained as it 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:

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A|I, Q; \theta)$$

where  $\mathcal{A}$  is a set of all possible answers,  $I$  and  $Q$  are image and question respectively and  $\theta$  represents the parameters in the network. Also we adopt generation based approaches for explanation generation

$$\hat{E} = \underset{E}{\operatorname{argmax}} P(e_t|(I, Q), A, e_0, \dots, e_{t-1}; \theta_e)$$

where  $\mathcal{E}$  is the explanation vocabulary.

We provide four variants of our model, as shown in figure 2. Correlated Answer Module (CAM) only corrects altered answers, Correlated Explanation Module (CEM) only corrects altered explanations, Answer and Explanation Correlated Module (AECM) corrects answers and explanations separately and Collaborative Correlated Module (CCM) corrects both answers and explanations jointly.

CCM comprises of three modules: Encoder Module, Generator Module, and Correlated Module. The Encoder Module encodes images and questions using CNN and LSTM. We combine these two using an attention mechanism to obtain

an attention feature. Using the attention feature, the generator module predicts answers and explanations that justify the predicted answer. The Correlated module defends the model against perturbations in the input, which facilitates the model to predict correct answers even under perturbation though we do not perturb the input at the time of training. Details of each module are as follows:

### 3.1. Encoder Module

Given an input image  $X_i$ , we obtain an embedding  $g_i \in \mathbb{R}^{W \times H \times C}$  using CNN which is parameterized by a function  $G_i(X_i, \theta_i)$ , where  $\theta_i$  represents the weights for the image embedding module. Similarly, for the query question  $X_q$ , we obtain question feature embedding  $g_q$  after passing it through an LSTM, which is parameterized using a function  $G_q(X_q, \theta_q)$ , where  $\theta_q$  represents the weights for the question embedding module. The image embedding  $g_i$  and question embedding  $g_q$  are used in an attention network, which is parameterized using a function  $G_f(g_i, g_q, \theta_f)$  where  $\theta_f$  are the parameters of the attention network that combines the image and question embeddings with a weighted softmax function and produces an output attention weighted vector  $g_f$ . The corresponding attention expressions are as follows:

$$\begin{aligned} f_j &= \tanh(\theta_f g_i \odot g_q) \\ f_s &= ||(\text{signed\_sqrt}(f_j))||_2 \\ &= \text{softmax}(\theta_a (a f_s + b_a)) \\ g_f &= (\odot g_i) \odot f_q \end{aligned} \quad (1)$$

where  $\odot$  represents element-wise multiplication and  $\sigma(\cdot)$  represents the sigmoid operator.

### 3.2. Generator Module

Our generator module aims to predict an answer and generate an explanation. Both the answer prediction and textual explanation rely on the attention feature  $g_f$ . Answer generation is through a classifier network  $g_y$  using a fully

connected network. Our generator module aims to predict answer and generate explanation. First, it helps to obtain the probability of the predicted answer class with the help of a softmax classifier. The answer prediction relies on the attention feature  $g_f$ .  $g_f$  is projected into  $V_a$  dimensional answer space using a fully connected layer. Answer classifier network is defined as follows:  $g_y = h(y_1, h(y_2, g_f))$

$$g_y = h(y_1, h(y_2, g_f)) \quad (2)$$

where  $h$  is the ReLU activation function,  $y_1 \in \mathbb{R}^{V_a \times 1}$  and  $y_2 \in \mathbb{R}^{1 \times I_f}$ . At training time, we minimize the cross entropy loss between the predicted and the ground truth answers.

$$L_y(f, y) = L_y(G_y(G_f(g_i, g_q)); y) \quad (3)$$

For generating textual explanation, we condition attention feature  $g_f$  on answer embedding  $g_y$ . This is generated using an LSTM based sequence generator for generating textual explanation  $g_e$ . At training time, we generate meaningful explanations by minimizing the cross entropy loss between generated explanation and ground truth explanation.

$$L_e(f, y, e) = L(G_e(G_y, G_f); e) \quad (4)$$

### 3.3. Correlated Module

We introduce an Adversarial Correlated Network. Using the adversarial mechanism, we develop a Collaborative Correlated Module which simultaneously checks whether the predicted answer and the corresponding explanation are correct or not. Each time the model predicts a wrong answer, the correlated module tries to correct it by comparing it with the real answer and the corresponding real explanation. Our correlated module includes Correlated Answer Module, Correlated Explanation Module and Collaborative Correlated Module. Each of the correlated modules are explained as follows:

**Correlated Answer Module:** We use a  $V_a$  dimensional one-hot vector representation for every answer word and transform it into a real valued word representation  $f_{da}$  by matrix  $_{aw} \in \mathbb{R}^{I_{aw} \times V_a}$ . We pass the obtained  $I_{aw}$  dimensional word embedding through a fully connected layer to obtain  $I_a$  dimensional answer encoding  $g_{da}$ .  $g_{da}$  is represented as follows:  $g_{da} = h(d_a, h(_{aw}, A))$ , where  $h$  is a nonlinear function ReLU.

**Correlated Explanation Module :** For explanation, we obtain representation  $g_{de}$  using an LSTM. The hidden state representation of the last word of the LSTM network provides a semantic representation of the whole sentence conditioned on all the previously generated words  $e_0, e_1, \dots, e_t$ . The model can be represented as  $g_{de} = \text{LSTM}(E)$

**Collaborative Correlated Module:** We design a collaborative network by concatenating the answer and explanation

#### Algorithm 1 Training CCM

---

```

1: Input: Image  $X_I$ , Question  $X_Q$ 
2: Output: Answer  $y$ , Explanation  $e$ 
3: repeat
4:   Answer Generator  $G_y(X_I, X_Q) \rightarrow g_y$ 
5:   Explanation Generator  $G_e(X_I, X_Q) \rightarrow g_e$ 
6:   Ans cross entropy  $L_{ans} = \text{loss}(\hat{y}, y)$ 
7:   Exp cross entropy  $L_{exp} = \text{loss}(\hat{e}, e)$ 
8:   repeat
9:     Sample mini batch of fake Ans and Exp:
10:     $y_f^1 \dots y_f^m$  and  $e_f^1 \dots e_f^m$ 
11:    Sample mini batch of real Ans and Exp:
12:     $y_r^1 \dots y_r^m$  and  $e_r^1 \dots e_r^m$ 
13:    Discriminator:  $D_j(D_y(y_r^i), D_e(e_r^i)) \rightarrow D(Y, E)$ 
14:    Update the discriminator using stochastic gradient ascent
15:     $\frac{1}{m} \sum_{i=1}^m [\log D(Y, E) + \log(1 - D(G_y, G_e))]$ 
16:    until  $k = 1 : K$ 
17:    Sample mini batch of Real Ans and Exp:
18:     $y_f^1 \dots y_f^m$  and  $e_f^1 \dots e_f^m$ 
19:    Update the Generator by descending its stochastic gradient:
20:     $\frac{1}{m} \sum_{i=1}^m \log(1 - D(G_y, G_e))$ 
21:  until Number of iterations

```

---

embeddings to obtain a joint embedding. The collaborative module is trained in an adversarial setting to correct the misclassified answers. Given answer embedding  $g_{da}$  and explanation embedding  $g_{de}$ , we obtain a joint feature  $g_{dj}$  by concatenating both of them and passing it through a fully connected layer to obtain final feature embedding which is as follows:

$$g_{dj} = g_{dj} \tanh(g_{da}; g_{de})$$

where  $;$  indicates the concatenation of two modules. The complete scheme is shown in figure 3. We use these modules to make variants of our model, as shown in figure 2. We train the discriminator in an adversarial manner between the generated and ground truth embedding. The adversarial cost function is given by:

$$\min_G \max_D L_c(G, D) = E_{y, y_e \in E} [\log D(Y, E)] + E_{g_y, g_y, g_e \in G_e} [\log(1 - D(G_y, G_e))]$$

The final cost function for CCM can be formulated as follows:

$$L = L_y + L_e - L_c$$

where  $L_y$  is the loss of answer generator module,  $L_e$  is the loss of explanation generator module,  $L_c$  is the loss of collaborative correlated module and  $\alpha$  is a hyper-parameter. We trained our model by optimizing this cost function with model parameters  $(f, e, y, d)$  to deliver a saddle point

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	54.7	38.1	26.8	19.1	18.0	42.9	66.1	14.0
CAM	55.0	38.3	26.8	19.0	18.3	43.1	69.2	15.2
CEM	54.5	38.3	27.2	19.6	18.4	43.1	68.2	15.1
AECM	55.5	38.9	27.4	19.3	18.3	43.4	67.9	14.8
CCM	<b>56.7</b>	<b>40.8</b>	<b>29.2</b>	<b>21.1</b>	<b>19.7</b>	<b>44.9</b>	<b>73.9</b>	<b>16.2</b>

Table 1: Ablation Analysis of Our Model. We achieve improvements in all the metrics.

(a) Image

(b) Question

Figure 4: (a)**Blurring Images**: BLUE-4 score of our variants of model vs increasing blur in validation images. Slope with which CCM’s score decreases is less as compared to other models and hence it is robust to blur in images. (b)**Replacing Words**: BLUE-4 score of our variants of models vs increasing random exchange of question words with question vocabulary words. Slope with which CCM’s score decreases is less as compared to other models and hence it is robust to replaced question words.

function as follows:

$$\begin{aligned}
 (\hat{f}, \hat{e}, \hat{y}) &= \arg \max_{f, e, y} (C(f, e, y, d)) \\
 (\hat{d}) &= \arg \min_d (C(\hat{f}, \hat{e}, \hat{y}, d))
 \end{aligned} \tag{5}$$

## 4. Experiments

We evaluate our proposed CCM method using quantitative and qualitative analysis. The quantitative evaluation is conducted using standard metrics like BLEU [37], METEOR [3], ROUGE [29] and CIDEr [48]. We evaluate our attention maps using rank correlation [38]. We further consider the statistical significance for the many ablations as well as the state-of-the-art models. In qualitative analysis, we show the word statistics of generated explanation with a Sunburst plot in figure 7. We provide gradual improvement in visualization of attention maps for a few images as we move from our base model to CCM model. We perform ablation analysis based on various samples of noise.

### 4.1. Dataset

We evaluate our proposed model on VQA Explanation Dataset (VQA-X) [20] which contains human annotated explanations for open-ended question answer(QA) pairs. QA

Model	RC
Random Point [20]	+0.0017
Uniform [20]	+0.0003
Answering [20]	+0.2211
ME [20]	+0.3423
Baseline (ME)	+0.3425
CAM (ours)	+0.3483
CEM (ours)	+0.3589
AECM (ours)	+0.3595
CCM (ours)	<b>+0.3679</b>

Table 2: Ablation and State of the art comparison with our models for Rank Correlation(higher is better)

pairs are taken from Visual Question Answering (VQA) dataset [1]. VQA-X consists of one explanation per QA pair in train split and three explanations per QA pair in validation and test split with a total of 31,536 explanations in training split, 4,377 explanations in validation split and 5904 explanations in test split. VQA-X also consists of human annotated visual justification collected from Amazon Mechanical Turk.

Model	Combination	BLEU-4	METEOR	ROU-L	CIDEr	SPICE
ME1 [20]	+ Ans + Att + Des	6.1	12.8	26.4	36.2	12.1
ME2 [20]	- Ans - Att + Des	5.9	12.6	26.3	35.2	11.9
ME3 [20]	- Ans - Att + Exp	18.0	17.3	42.1	63.6	13.8
ME4 [20]	+ Ans - Att + Exp	18.0	17.6	42.4	66.3	14.3
ME5 [20]	- Ans + Att + Exp	19.5	18.2	43.4	71.3	15.1
ME6 [20]	+ Ans + Att + Exp	19.8	18.6	44.0	73.4	15.4
Baseline	- Ans + Att + Exp	19.1	18.0	42.9	66.1	14.0
Our (CCM)	+ Ans + Att + Exp	<b>21.1</b>	<b>19.7</b>	<b>44.9</b>	<b>73.9</b>	<b>16.2</b>

Table 3: Comparison with State of the Art models on VQA-X dataset. Ans denotes use of ground truth answers for explanation generation, Att denotes answer attention and Exp denotes explanation attention. After using all three, our model outperforms all the other models.

Figure 5: This figure shows visual and textual explanation for images picked from validation set. The 1<sup>st</sup>, 3<sup>rd</sup>, 5<sup>th</sup> and 7<sup>th</sup> columns contain original image and its question. 2<sup>nd</sup>, 4<sup>th</sup>, 6<sup>th</sup> and 8<sup>th</sup> column provide predicted answer, visual explanation mask and textual explanation.

## 4.2. Ablation Analysis

We provide comparisons of our models with the prevalent baselines. We compare our model on textual and visual explanations. We add noise during inference time to test the robustness of models. As the methods were not intended or designed to be robust to noise, normal methods perturbed more as compared to our models.

### 4.2.1 Analysis on Textual Explanations

For textual explanations, we consider different variations of our method and various ways to obtain collaborating embedding as mentioned in section 3.3. Table 1 shows the performance of variants of our model on different metrics for the VQA-X test set. It is clear that CCM outperforms all the other variants. There is a trend of increasing scores as we move from baseline to CCM. We achieve an improvement of about 2% in BLEU-4, 1.7% in METEOR, 2% in ROUGE-L,

7% in CIDEr and 2.1% in SPICE.

### 4.2.2 Analysis on Visual Explanation

We measure the similarity between our explanation and ground truth explanation by rank correlation and its results are shown in table 2. We start with the rank correlation of the baseline model and then we compare with different variants such as CAM, CEM, AECM and CCM. We achieve an improvement of about 2.56% when we move from state of the art(ME) to CCM.

### 4.2.3 Statistical Significance Analysis

We analyze Statistical Significance [8] of our model(CCM) against the variants mentioned in section 3. The Critical Difference(CD) for Nemenyi [13] test depends on given confidence level (0.05 in our case) for average ranks and number of tested datasets N. Low difference in ranks between

(a) Image

(b) Question

(c) Image-Question

(a) Image

(b) Question

(c) Image-Question

Figure 6: **Effect of Image and Question Noise** : First row, First Column shows variation of Mean of Bleu-1 score for image noise, second column shows mean of Bleu-1 score for question noise and third column shows mean of Bleu-1 score for image and question noise combined. We add noise in image features by introducing gaussian noise with increasing noise intensity. Noise in questions is added by randomly masking question words with increasing probability of masking shown in column-2. Finally we add noise in both image features and question with increasing intensity/probability shown in column-3. Purple line shows performance of our model(CCM). It can be observed that our model perturbs least as compared to other models and hence it is robust as compared to other models. Similarly second row provides the variation of Standard Deviation of Bleu-1 score for image noise, question noise and combined noise(image and question).

two models implies that they are significantly less different and vice versa. Figure 8 visualizes the post hoc analysis using the CD diagram. It is clear that CCM is significantly different from other methods.

Model	Mean	Std Dev	Actual	Abs. Dif
Baseline	51.9	0.235	53.60	1.720
CAM	53.9	0.254	54.42	0.520
CEM	53.7	0.360	54.36	0.660
AECM	54.3	0.257	54.39	0.093
CCM	56.7	<b>0.187</b>	56.70	<b>0.007</b>

Table 4: Performance of models against image and question noise(on BLUE-1 score)

### 4.3. Analysis on Robustness of Model

We analyze the behavior of our model against increasing noise in figure 6. We sample the noise 50 times and report mean and standard deviation. Note that the model has not been trained with this noise, rather we only use it in the test time. Since Bleu-1 score varies the most(among all the scores), we choose Bleu-1 so that we can get proper estimates of standard deviation(Bleu-4 score deviates much less). First, we add gaussian noise to image features. Mean of the noise is same as that of the image features and the standard

deviation is  $\times$  (standard deviation) of the image features where  $\times = 1, \dots, 5$ . We observe that our model has the highest mean, lowest standard deviation and lowest slope for mean and standard deviation. Second, we randomly mask words in questions during validation time. We mask words with increasing probabilities of 0.05, 0.1, 0.15, 0.2, 0.25. Humans are robust to such noise since we can extract the semantic meaning of a sentence even if some words are masked or missing. So we test our model on this noise and observe that our model follows the previous trend. Then, we add both image and question noise with increasing magnitude. We again observe that our model outperforms all other models and it perturbs least.

In figure 4(a), we analyze the effect of blurring the input images and extracting features from these noisy images. We observe that our model again outperforms other models along with deviating less w.r.t. the input. In figure 4(b), we analyze the effect of replacing words in question from question vocabulary. Humans are not much affected by these since we can extract the meaning of a sentence based on the neighboring words. Again, our model outperforms all other models and perturbs the least. Since our model is learning joint distribution from two inputs(answers and explanations), it

Figure 7: Sunburst plot for Generated Explanation: The  $j^{\text{th}}$  ring captures the frequency distribution over words for the  $i^{\text{th}}$  word of the generated explanation. The angle subtended at the center is proportional to the frequency of the word. While some words have high frequency, the outer rings illustrate a fine blend of words. We have restricted the plot to 5 rings for easy readability.

Figure 8: The mean rank of all the models on the basis of scores. Here  $CD=2.531$  and  $p=0.007255$ . CDM refers to our model and others are different variations as described in section 1. The colored lines between the two models represents that these models are not significantly different from each other.

gives robust explanations for VQA.

#### 4.4. Comparison with State of the Art

We obtain the initial comparison with the baselines on the rank correlation on visual explanation (VQA-X) dataset [20] that provides a visual explanation in terms of segmentation mask while solving for VQA. We use a variant of the multimodal explanation [20] model as our baseline method. We obtain an improvement of around 1.72% using AEEM network and 2.56% using CCM network from state of the art model ME [20] in terms of rank correlation. The comparison between various state-of-the-art methods and baselines is provided in table 2. We compare our model with the state of the art model in table 3. Rows 1 to 6 denote variants of state of the art model. The last row, CCM, is our best

proposed model. We observe that for the VQA-X dataset, we achieve an improvement of 2% in BLEU-4 score and 1.7% in METEOR metric scores over the baseline. We improve over the previous state-of-the-art [20] for VQA-X dataset by around 1.3% in BLEU-4 score and 1.1% in METEOR score.

#### 4.5. Qualitative Result

We provide predicted answers, generated textual explanations and visual explanations in the form of attention maps for CCM in figure 5. It is apparent that our model points to prominent regions for answering questions as well as justifying the answer. For example, in the first row, CCM is able to focus on a specific portion of the image and provide textual explanation for the predicted answer. The same results are observed for most of the examples. Also, we have shown the distribution of generated explanations, as demonstrated in the figure-7.

#### 4.6. Training and Model Configuration

We use cross entropy loss to train Answer and Explanation Module and adversarial loss to train Correlated Module in an end-to-end manner. We use ADAM to update Answer and Explanation Module and SGD to update Correlated Module. We configure hyper-parameters of Generator module using the validation set which are as follows: learning rate = 0.0007, batch size = 64, alpha = 0.99, beta = 0.95, and epsilon =  $1e-8$ . For correlated module, the hyper-parameters are as follows: learning rate = 0.0007 and momentum = 0.9. We also use weight clipping and learning rate decaying to decrease the learning rate every ten epochs.

### 5. Conclusion

In this paper, we solve for a method that can generate textual and visual explanations for visual question answering. The proposed method introduces a collaborative correlated module that ensures that generated textual explanations and answers are coherent with each other and are robust to perturbation in image and text. We provide a detailed evaluation that compares the accuracy of the answers and various standard metrics for the generated textual and visual explanations. We observe that while our model generates comparable accuracies for answer generation, the generated textual and visual explanations are more coherent and are robust against perturbation. Our explanations change when there is a change in answer and hence signifies the importance of image and question in robustness of our model. In the future, we would consider ways for obtaining improved insights while solving challenging vision and language based tasks building upon the proposed work.

### 6. Acknowledgements

Assistance through SERB grant CRG/2018/003566 is duly acknowledged.



## References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.
- [2] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. of ACL workshop on Intrinsic and Extrinsic Evaluation measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, 2005.
- [4] K. Barnard, P. Duygulu, and D. Forsyth. N. de Freitas, d. Blei, and MI Jordan, "Matching Words and Pictures", submitted to *JMLR*, 2003.
- [5] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9–16, 2013.
- [6] X. 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.
- [7] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [8] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.
- [9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.
- [10] V. Escorcia, J. Carlos Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1256–1264, 2015.
- [11] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [12] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.
- [13] D. Fišer, T. Erjavec, and N. Ljubešić. Janes v0. 4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina*, 2(4):2, 2016.
- [14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.
- [15] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304, 2015.
- [16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. 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.
- [18] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang. Long text generation via adversarial training with leaked information. *arXiv preprint arXiv:1709.08624*, 2017.
- [19] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596, 2017.
- [20] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. 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.
- [21] U. Jain, Z. Zhang, and A. Schwing. Creativity: Generating diverse questions using variational autoencoders. *arXiv preprint arXiv:1704.03493*, 2017.
- [22] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.
- [23] A. Karpathy and L. 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.
- [24] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer, 2011.
- [25] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc. Explainable artificial intelligence for training and tutoring. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE ..., 2005.
- [26] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.
- [27] R. Li and J. Jia. Visual question answering with question representation update (qr). In *Advances in Neural Information Processing Systems*, pages 4655–4663, 2016.
- [28] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition gan for visual paragraph generation. *CoRR*, abs/1703.07022, 2, 2017.
- [29] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004.
- [30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

- [31] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.
- [32] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [33] M. Malinowski and M. 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.
- [34] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [35] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016.
- [36] H. Noh, P. Hongsuck Seo, and B. 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.
- [37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [38] B. Patro and V. P. Namboodiri. Differential attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [39] B. N. Patro, Anupriy, and V. P. Namboodiri. Explanation vs attention: A two-player game to obtain attention for vqa. *ArXiv*, abs/1911.08618, 2019.
- [40] B. N. Patro, S. Kumar, V. K. Kurmi, and V. 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.
- [41] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. *ICCV*, abs/1908.06306, 2019.
- [42] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [43] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- [44] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2953–2961, 2015.
- [45] K. J. Shih, S. Singh, and D. 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.
- [46] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379, 1975.
- [47] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014.
- [48] R. Vedantam, L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.
- [49] P. Veličković, D. Wang, N. D. Lane, and P. Liò. X-cnn: Cross-modal convolutional neural networks for sparse datasets. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–8. IEEE, 2016.
- [50] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.
- [51] O. Vinyals, A. Toshev, S. Bengio, and D. 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.
- [52] H. Xu and K. 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.
- [53] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.
- [54] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.
- [55] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2461–2469. IEEE, 2015.
- [56] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.
- [57] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [58] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin. Adversarial feature matching for text generation. *arXiv preprint arXiv:1706.03850*, 2017.
- [59] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.
- [60] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.