Multimodal Differential Network for Visual Question Generation

Anonymous EMNLP submission

Abstract

Generating natural questions from an image is a semantic task that requires using visual and language modality to learn multimodal representations. Images can have multiple visual and language contexts that are relevant for generating questions namely places, captions, and tags. In this paper, we propose the use of exemplars for obtaining the relevant context. We obtain this by using a Multimodal Differential Network to produce natural and engaging questions. The generated questions show a remarkable similarity to the natural questions as validated by a human study. Further, we observe that the proposed approach substantially improves over state-of-the-art benchmarks on the quantitative metrics (BLEU, METEOR, ROUGE, and CIDEr).

1 Introduction

To understand the progress towards multimedia vision and language understanding, a visual Turing test was proposed by (Geman et al., 2015) that was aimed at visual question answering (Antol et al., 2015). Visual Dialog (Das et al., 2017) is a natural extension for VQA. Current dialog systems as evaluated in (Chattopadhyay et al., 2017) show that when trained between bots, AI-AI dialog systems show improvement, but that does not translate to actual improvement for Human-AI dialog. This is because, the questions generated by bots are not natural (human-like) and therefore does not translate to improved human dialog. Therefore it is imperative that improvement in the quality of questions will enable dialog agents to perform well in human interactions. Further, (Ganju et al., 2017) show that unanswered questions can be used for improving VQA, Image captioning and Object Classification.

An interesting line of work in this respect is the work of (Mostafazadeh et al., 2016). Here the authors have proposed the challenging task of generating *natural* questions for an image. One aspect that is central to a question is the context that is relevant to generate it. However, this context changes for every image. As can be seen in Figure 1, an image with a person on a skateboard would result in questions related to the event. Whereas for a little girl, the questions could be related to age rather than the action. How can one have widely varying context provided for generating questions? To solve this problem, we use the context obtained by considering exemplars, specifically we use the difference between relevant and irrelevant exemplars. We consider different contexts in the form of Location, Caption, and Part of Speech tags.

(a) Is this a skateboard competition ?

(b) Did he knock over any cones ?

(a) Is this her first time skiing?

(b) How old is that little girl?

Figure 1: Can you guess which among the given questions is human annotated and which is machine generated? [0]

Our method implicitly uses a differential context obtained through supporting and contrasting exemplars to obtain a differentiable embedding. This embedding is used by a question decoder to decode the appropriate question. As discussed further, we observe this implicit differential context to perform better than an explicit keyword

---

[0]The human annotated questions are (b) for the first image and (a) for the second image.

based context. The difference between the two approaches is illustrated in Figure 2. This also allows for better optimization as we can backpropagate through the whole network. We provide detailed empirical evidence to support our hypothesis. As seen in Figure 1 our method generates natural questions and improves over the state-of-the-art techniques for this problem.
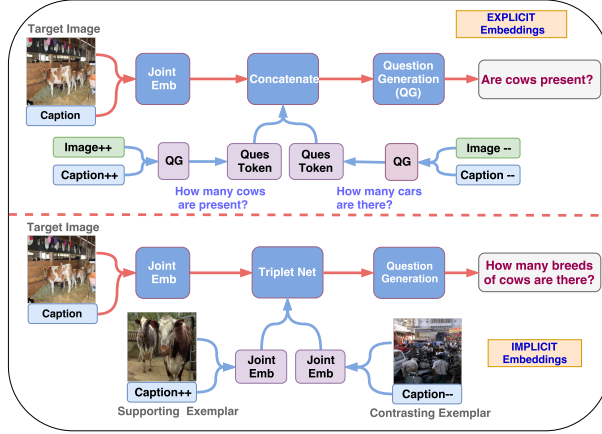

Figure 2: Here we provide intuition for using implicit embeddings instead of explicit ones. As explained in section 1, the question obtained by the implicit embeddings are natural and holistic than the explicit ones.

To summarize, we propose a multimodal differential network to solve the task of visual question generation. Our contributions are: (1) A method to incorporate exemplars to learn differential embeddings that captures the subtle differences between supporting and contrasting examples and aid in generating natural questions. (2) We provide Multimodal differential embeddings, as image or text alone does not capture the whole context and we show that these embeddings outperform the ablations which incorporate cues such as only image, or tags or place information. (3) We provide a thorough comparison of the proposed network against state-of-the-art benchmarks along with a user study and statistical significance test.

2   Related Work

Generating a natural and engaging question is an interesting and challenging task for a smart robot (like chat-bot). It is a step towards having a natural visual dialog instead of the widely prevalent visual question answering bots. Further, having the ability to ask natural questions based on different contexts is also useful for artificial agents that can interact with visually impaired people. While the task of generating question automatically is well studied in NLP community, it has been relatively less studied for image-related natural questions. This is still a difficult task (Mostafazadeh et al., 2016) that has gained recent interest in the community.

Recently there have been many deep learning based approaches as well for solving the text-based question generation task such as (Du et al., 2017). Further, (Serban et al., 2016) have proposed a method to generate a factoid based question based on triplet set {subject, relation and object} to capture the structural representation of text and the corresponding generated question.

These methods, however, were limited to text-based question generation. There has been extensive work done in the Vision and Language domain for solving image captioning, paragraph generation, Visual Question Answering (VQA) and Visual Dialog. (Barnard et al., 2003; Farhadi et al., 2010; Kulkarni et al., 2011) proposed conventional machine learning methods for image description. (Socher et al., 2014; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Xu et al., 2015; Fang et al., 2015; Chen and Lawrence Zitnick, 2015; Johnson et al., 2016; Yan et al., 2016) have generated descriptive sentences from images with the help of Deep Networks. There have been many works for solving Visual Dialog (Chappell et al., 2004; Das et al., 2016, 2017; De Vries et al., 2017; Strub et al., 2017). A variety of methods have been proposed by (Malinowski and Fritz, 2014; Lin et al., 2014; Antol et al., 2015; Ren et al., 2015; Ma et al., 2016; Noh et al., 2016) for solving VQA task including attention-based methods (Zhu et al., 2016; Fukui et al., 2016; Gao et al., 2015; Xu and Saenko, 2016; Lu et al., 2016; Shih et al., 2016). However, Visual Question Generation (VQG) is a separate task which is of interest in its own right and has not been so well explored (Mostafazadeh et al., 2016). This is a vision based novel task aimed at generating natural and engaging question for an image. (Yang et al., 2015) proposed a method for continuously generating questions from an image and subsequently answering those questions. The works closely related to ours are that of (Mostafazadeh et al., 2016) and (Jain et al., 2017). In the former work, the authors used an encoder-decoder based framework whereas in the latter work, the authors extend it by using a variational autoencoder based sequential routine to obtain natural questions by performing sampling of
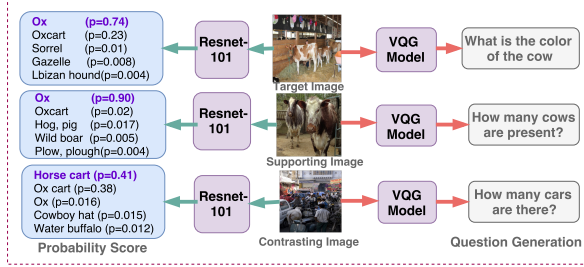
the latent variable.

3   Approach

Figure 3: An illustrative example shows the validity of our obtained exemplars with the help of an object classification network, RESNET-101. We see that the probability scores of target and supporting exemplar image are similar. That is not the case with the contrasting exemplar. The corresponding generated questions when considering the individual images are also shown.

In this section, we clarify the basis for our approach of using exemplars for question generation. To use exemplars for our method, we need to ensure that our exemplars can provide context and that our method generates valid exemplars.

We first analyze whether the exemplars are valid or not. We illustrate this in figure 3. We used a pre-trained RESNET-101 (He et al., 2016) object classification network on the target, supporting and contrasting images. We observed that the supporting image and target image have quite similar probability scores. The contrasting exemplar image, on the other hand, has completely different probability scores.

Exemplars aim to provide appropriate context. To better understand the context, we experimented by analysing the questions generated through an exemplar. We observed that indeed a supporting exemplar could identify relevant tags (cows in Figure 3) for generating questions. We improve use of exemplars by using a triplet network. This network ensures that the joint image-caption embedding for the supporting exemplar are closer to that of the target image-caption and vice-versa. We empirically evaluated whether an explicit approach that uses the differential set of tags as a one-hot encoding improves the question generation, or the implicit embedding obtained based on the triplet network. We observed that the implicit multimodal differential network empirically provided better context for generating questions. Our understanding of this phenomenon is that both target and supporting exemplars generate similar questions whereas contrasting exemplars generate very different questions from the target question. The triplet network that enhances the joint embedding thus aids to improve the generation of target question. These are observed to be better than the explicitly obtained context tags as can be seen in Figure 2. We now explain our method in detail.

4   Method

The task in visual question generation (VQG) is to generate a natural language question $\hat{Q}$, for an image I. We consider a set of pre-generated context C from image I. We maximize the conditional probability of generated question given image and context as follows:

$$\hat{\theta} = \arg\max_{\theta} \sum_{(I,C,Q)} \log P(Q \mid I, C, \theta) \qquad (1)$$

where θ is a vector for all possible parameters of our model. Q is the ground truth question. The log probability for the question is calculated by using joint probability over $\{q_0, q_1, ..., q_N\}$ with the help of chain rule. For a particular question, the above term is obtained as:

$$\log P(\hat{Q}|I, C) = \sum_{t=0}^{N} \log P(q_t | I, C, q_0, .., q_{t-1})$$

where N is length of the sequence, and q_t is the tth word of the question. We have removed θ for simplicity.

Our method is based on a sequence to sequence network (Sutskever et al., 2014; Vinyals et al., 2015; Bahdanau et al., 2014). The sequence to sequence network has a text sequence as input and output. In our method, we take an image as input and generate a natural question as output. The architecture for our model is shown in Figure 4. Our model contains three main modules, (a) Representation Module that extracts multimodal features (b) Mixture Module that fuses the multimodal representation and (c) Decoder that generates question using an LSTM-based language model.

During inference, we sample a question word q_i from the softmax distribution and continue sampling until the end token or maximum length for the question is reached. We experimented with both sampling and *argmax* and found out that *argmax* works better. This result is provided in the supplementary material.
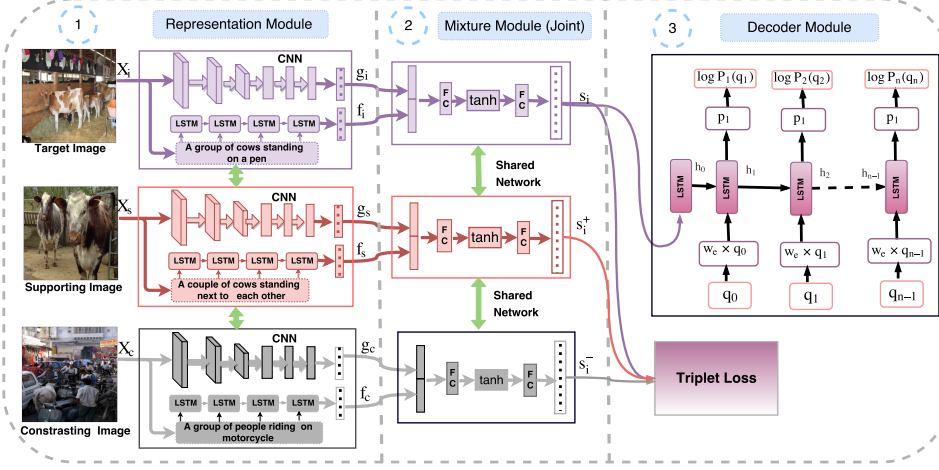
3

Figure 4: This is an overview of our Multimodal Differential Network for Visual Question Generation. It consists of a Representation Module which extracts multimodal features, a Mixture Module that fuses the multimodal representation and a Decoder that generates question using an LSTM based language model. In this figure, we have shown the Joint Mixture Module. We train our network with a Cross-Entropy and Triplet Loss.

4.1 Multimodal Differential Network

The proposed Multimodal Differential Network (MDN) consists of a representation module and a joint mixture module.

4.1.1 Finding Exemplars

We used an efficient KNN-based approach (k-d tree) with Euclidean metric to obtain the exemplars. This is obtained through a coarse quantization of nearest neighbors of the training examples into 50 clusters, and selecting the nearest as supporting and farthest as the contrasting exemplars. We experimented with ITML based metric learning (Davis et al., 2007) for image features. Surprisingly, the KNN-based approach outperforms the latter one. We also tried random exemplars and different number of exemplars and found that $k = 5$ works best. We provide these results in the supplementary material.

4.1.2 Representation Module

We use a triplet network (Hoffer and Ailon, 2015) in our representation module. The triplet network consists of three subparts: target, supporting, and contrasting networks. All three networks share the same parameters. Given an image x_i we obtain an embedding g_i using a CNN parameterized by a function $G(x_i, W_c)$ where W_c are the weights for the CNN. The caption C_i results in a caption embedding f_i through an LSTM parameterized by a function $F(C_i, W_l)$ where W_l are the weights for the LSTM. This is shown in part 1 of Figure 4. Similarly we obtain image embeddings g_s & g_o

and caption embeddings f_s & f_o.

$$g_i = G(x_i, W_c) = CNN(x_i)$$
$$f_i = F(C_i, W_l) = LSTM(C_i) \quad (2)$$

4.1.3 Mixture Module

The Mixture module brings the image and caption embeddings to a joint feature embedding space. The input to the module is the embeddings obtained from the representation module. We have evaluated four different approaches for fusion viz., joint, element-wise addition, hadamard and attention method. Each of these variants receives image features g_i & the caption embedding f_i, and outputs a fixed dimensional feature vector s_i. The Joint method concatenates g_i & f_i and maps them to a fixed length feature vector s_i as follows:

$$s_i = W_j^T * \tanh(W_{ij}g_i \frown (W_{cj}f_i + b_j)) \quad (3)$$

where g_i is the 4096-dimensional convolutional feature from the FC7 layer of pretrained VGG-19 Net (Simonyan and Zisserman, 2014). W_{ij}, W_{cj}, W_j are the weights and b_j is the bias for different layers. $\frown$ is the concatenation operator.

Similarly, We obtain context vectors s_i^+ & s_i^- for the supporting and contrasting exemplars. Details for other fusion methods are present in supplementary.The aim of the triplet network (Schroff et al., 2015) is to obtain context vectors that bring the supporting exemplar embeddings closer to the target embedding and vice-versa. This is obtained as follows:

$$D(t(s_i), t(s_i^+)) + \alpha < D(t(s_i), t(s_i^-))$$
$$\forall(t(s_i), t(s_i^+), t(s_i^-)) \in M, \quad (4)$$

4

where $D(t(s_i), t(s_j)) = ||t(s_i) - t(s_j)||_2^2$ is the euclidean distance between two embeddings $t(s_i)$ and $t(s_j)$. M is the training dataset that contains all set of possible triplets. $T(s_i, s_i^+, s_i^-)$ is the triplet loss function. This is decomposed into two terms, one that brings the supporting sample closer and one that pushes the contrasting sample further. This is given by

$$T(s_i, s_i^+, s_i^-) = \max(0, D^+ + \alpha - D^-) \quad (5)$$

Here D^+, D^- represent the euclidean distance between the target and supporting sample, and target and opposing sample respectively. The parameter α controls the separation margin between these and is obtained through validation data.

4.2 Decoder: Question Generator

The role of decoder is to predict the probability for a question, given s_i. RNN provides a nice way to perform conditioning on previous state value using a fixed length hidden vector. The conditional probability of a question token at particular time step q_t is modeled using an LSTM as used in machine translation (Sutskever et al., 2014). At time step t, the conditional probability is denoted by $P(q_t|I, C, q_0, ...q_{t-1}) = P(q_t|I, C, h_t)$, where h_t is the hidden state of the LSTM cell at time step t, which is conditioned on all the previously generated words $\{q_0, q_1, ...q_{N-1}\}$. The word with maximum probability in the probability distribution of the LSTM cell at step k is fed as an input to the LSTM cell at step $k+1$ as shown in part 3 of Figure 4. At $t = -1$, we are feeding the output of the mixture module to LSTM. $\hat{Q} = \{\hat{q_0}, \hat{q_1}, ...\hat{q_{N-1}}\}$ are the predicted question tokens for the input image I. Here, we are using $\hat{q_0}$ and $\hat{q_{N-1}}$ as the special token START and STOP respectively. The softmax probability for the predicted question token at different time steps is given by the following equations where LSTM refers to the standard LSTM cell equations:

$$x_{-1} = S_i = \text{Mixture Module}(g_i, f_i)$$
$$h_0 = \text{LSTM}(x_{-1})$$
$$x_t = W_e * q_t, \forall t \in \{0, 1, 2, ...N-1\}$$
$$h_{t+1} = \text{LSTM}(x_t, h_t), \forall t \in \{0, 1, 2, ...N-1\}$$
$$o_{t+1} = W_o * h_{t+1}$$
$$\hat{y}_{t+1} = P(q_{t+1}|I, C, h_t) = \text{Softmax}(o_{t+1})$$
$$Loss_{t+1} = loss(\hat{y}_{t+1}, y_{t+1})$$

Where $\hat{y}_{t+1}$ is the probability distribution over all question tokens. $loss$ is cross entropy loss.

4.3 Cost function

Our objective is to minimize the total loss, that is the sum of cross entropy loss and triplet loss over all training examples. The total loss is:

$$L = \frac{1}{M} \sum_{i=1}^{M} (L_{cross} + L_{triplet}) \quad (6)$$

where M is the total number of samples, $L_{triplet}$ is the triplet loss function 5. L_{cross} is the cross entropy loss between the predicted and ground truth questions and is given by:

$$L_{cross} = \frac{-1}{N} \sum_{t=1}^{N} y_t \log P(\hat{q_t}|I_i, C_i, \hat{q_0}, ..\hat{q_{t-1}})$$

where, N is the total number of question tokens, y_t is the ground truth label. Anonymized version of the code is provided [1].

4.4 Variations of Proposed Method

While, we advocate the use of multimodal differential network for generating embeddings that can be used by the decoder for generating questions, we also evaluate several variants of this architecture. These are as follows:

Tag Net: In this variant, we consider extracting the part-of-speech (POS) tags for the words present in the caption and obtaining a Tag embedding by considering different methods of combining the one-hot vectors. Further details and experimental results are present in the supplementary. This Tag embedding is then combined with the image embedding and provided to the decoder network.

Place Net: In this variant we explore obtaining embeddings based on the visual scene understanding. This is obtained using a pre-trained PlaceCNN (Zhou et al., 2017) that is trained to classify 365 different types of scene categories. We then combine the activation map for the input image and the VGG-19 based place embedding to obtain the joint embedding used by the decoder.

Differential Image Network: Instead of using multimodal differential network for generating embeddings, we also evaluate differential image network for the same. In this case, the embedding does not include the caption but is based only on the image feature. We also exeperimented with using multiple exemplars and random exemplars.

---

[1]The anonymized link for code is https://github.com/MDN-VQG/EMNLP-MDN-VQG

5

Figure 5: These are some examples from the VQG-COCO dataset which provide a comparison between our generated questions and human annotated questions. (a) is the human annotated question for all the images. More qualitative results are present in the supplementary material.

Further details, pseudocode and results regarding these are present in the supplementary material.

4.5 Dataset

We conduct our experiments on Visual Question Generation (VQG) dataset (Mostafazadeh et al., 2016), which contains human annotated questions based on images of MS-COCO dataset. This dataset was developed for generating natural and engaging questions based on common sense reasoning. We use VQG-COCO dataset for our experiments which contains a total of 2500 training images, 1250 validation images, and 1250 testing images. Each image in the dataset contains five natural questions and five ground truth captions. It is worth noting that the work of (Jain et al., 2017) also used the questions from VQA dataset (Antol et al., 2015) for training purpose, whereas the work by (Mostafazadeh et al., 2016) uses only the VQG-COCO dataset. VQA-1.0 dataset is also built on images from MS-COCO dataset. It contains a total of 82783 images for training, 40504 for validation and 81434 for testing. Each image is associated with 3 questions. We used pretrained caption generation model (Karpathy and Fei-Fei, 2015) to extract captions for VQA dataset as the human annotated captions are not there in the dataset. We also get good results on the VQA dataset (as shown in Table 2) which shows that our method doesn't necessitate the presence of ground truth captions. We train our model separately for VQG-COCO and VQA dataset.

4.6 Inference

We made use of the 1250 validation images to tune the hyperparameters and are providing the results on test set of VQG-COCO dataset. During inference, We use the Representation module to find the embeddings for the image and ground truth caption without using the supporting and contrasting exemplars. The mixture module provides the joint representation of the target image and ground truth caption. Finally, the decoder takes in the joint features and generates the question. We also experimented with the captions generated by an Image-Captioning network (Karpathy and Fei-Fei, 2015) for VQG-COCO dataset and the result for that and training details are present in the supplementary material.

5 Experiments

We evaluate our proposed MDN method in the following ways: First, we evaluate it against other variants described in section 4.4 and 4.1.3. Second, we further compare our network with state-of-the-art methods for VQA 1.0 and VQG-COCO dataset. We perform a user study to gauge human opinion on naturalness of the generated question and analyze the word statistics in Figure 6. This is an important test as humans are the best deciders of naturalness. We further consider the statistical significance for the various ablations as
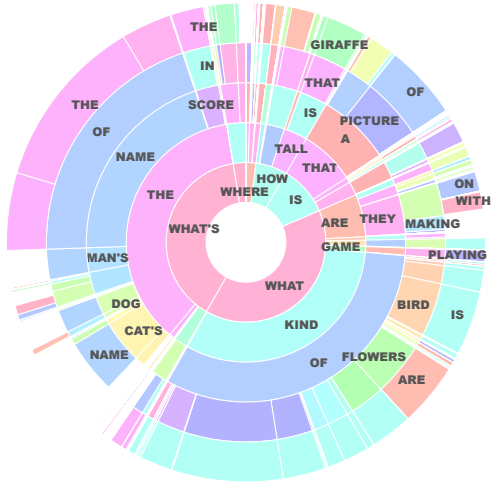
Figure 6: Sunburst plot for VQG-COCO: The ith ring captures the frequency distribution over words for the ith word of the generated question. The angle subtended at the center is proportional to the frequency of the word. While some words have high frequency, the outer rings illustrate a fine blend of words. We have restricted the plot to 5 rings for easy readability. Best viewed in color.

well as the state-of-the-art models. The quantitative evaluation is conducted using standard metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015). Although these metrics have not been shown to correlate with 'naturalness' of the question these still provide a reasonable quantitative measure for comparison. Here we only provide the BLEU1 scores, but the remaining BLEU-n metric scores are present in the supplementary. We observe that the proposed MDN provides improved embeddings to the decoder. We believe that these embeddings capture instance specific differential information that helps in guiding the question generation. Details regarding the metrics are given in the supplementary material.

5.1 Ablation Analysis

We considered different variations of our method mentioned in section 4.4 and the various ways to obtain the joint multimodal embedding as described in section 4.1.3. The results for the VQG-COCO test set are given in table 1. In this table, every block provides the results for one of the variations of obtaining the embeddings and different ways of combining them. We observe that the Joint Method (JM) of combining the embeddings works the best in all cases except the Tag Embed-

dings. Among the ablations, the proposed MDN method works way better than the other variants in terms of BLEU, METEOR and ROUGE metrics by achieving an improvement of 6%, 12% and 18% in the scores respectively over the best other variant.

Emb.	Method	BLEU1	METEOR	ROUGE	CIDEr
Tag	AtM	22.4	8.6	22.5	20.8
Tag	HM	**24.4**	**10.8**	**24.3**	**55.0**
Tag	AM	24.4	10.6	23.9	49.4
Tag	JM	22.2	10.5	22.8	50.1
PlaceCNN	AtM	24.4	10.3	24.0	51.8
PlaceCNN	HM	24.0	10.4	24.3	49.8
PlaceCNN	AM	24.1	10.6	24.3	51.5
PlaceCNN	JM	**25.7**	**10.8**	**24.5**	**56.1**
Diff. Img	AtM	20.5	8.5	**24.4**	19.2
Diff. Img	HM	23.6	8.6	22.3	22.0
Diff. Img	AM	20.6	8.5	24.4	19.2
Diff. Img	JM	**30.4**	**11.7**	22.3	**22.8**
MDN	AtM	22.4	8.8	24.6	22.4
MDN	HM	26.6	12.8	30.1	31.4
MDN	AM	29.6	15.4	32.8	41.6
MDN (Ours)	JM	**36.0**	**23.4**	**41.8**	**50.7**

Table 1: Analysis of variants of our proposed method on VQG-COCO Dataset as mentioned in section 4.4 and different ways of getting a joint embedding (Attention (AtM), Hadamard (HM), Addition (AM) and Joint (JM) method as given in section 4.1.3) for each method. Refer section 5.1 for more details.

5.2 Baseline and State-of-the-Art

The comparison of our method with various baselines and state-of-the-art methods is provided in table 2 for VQA 1.0 and table 3 for VQG-COCO dataset. The comparable baselines for our method are the image based and caption based models in which we use either only the image or the caption embedding and generate the question. In both the tables, the first block consists of the current state-of-the-art methods on that dataset and the second contains the baselines. We observe that for the VQA dataset we achieve an improvement of 8% in BLEU and 7% in METEOR metric scores over the baselines, whereas for VQG-COCO dataset this is 15% for both the metrics. We improve over the previous state-of-the-art (Yang et al., 2015) for VQA dataset by around 6% in BLEU score and 10% in METEOR score. In the VQG-COCO dataset, we improve over (Mostafazadeh et al., 2016) by 3.7% and (Jain et al., 2017) by 3.5% in terms of METEOR scores.

5.3 Statistical Significance Analysis

We have analysed Statistical Significance (Demšar, 2006) of our MDN model for VQG for different variations of the mixture module mentioned in section 4.1.3 and also against the state-of-the-art methods. The Critical

7

Methods	BLEU1	METEOR	ROUGE	CIDEr
Sample(Yang,2015)	38.8	12.7	34.2	13.3
Max(Yang,2015)	59.4	17.8	49.3	33.1
Image Only	56.6	15.1	40.0	31.0
Caption Only	57.1	15.5	36.6	30.5
MDN-Attention	60.7	16.7	49.8	33.6
MDN-Hadamard	61.7	16.7	50.1	29.3
MDN-Addition	61.7	18.3	50.4	**42.6**
MDN-Joint (**Ours**)	**65.1**	**22.7**	**52.0**	33.1

Table 2: State-of-the-Art comparison on VQA-1.0 Dataset. The first block consists of the state-of-the-art results, second block refers to the baselines mentioned in section 5.2, third block provides the results for the variants of mixture module present in section 4.1.3.

Context	BLEU1	METEOR	ROUGE	CIDEr
Natural2016	19.2	19.7	-	-
Creative2017	35.6	19.9	-	-
Image Only	20.8	8.6	22.6	18.8
Caption Only	21.1	8.5	25.9	22.3
Tag-Hadamard	24.4	10.8	24.3	55.0
PlaceCNN-Joint	25.7	10.8	24.5	**56.1**
Diff.Image-Joint	30.4	11.7	26.3	38.8
MDN-Joint (**Ours**)	**36.0**	**23.4**	**41.8**	50.7
Humans2016	**86.0**	**60.8**	-	-

Table 3: State-of-the-Art (SOTA) comparison on VQG-COCO Dataset. The first block consists of the SOTA results, second block refers to the baselines mentioned in section 5.2, third block shows the results for the best method for different ablations mentioned in table 1.

Difference (CD) for Nemenyi (Fišer et al., 2016) test depends upon the given α (confidence level, which is 0.05 in our case) for average ranks and N (number of tested datasets). If the difference in the rank of the two methods lies within CD, then they are not significantly different and vice-versa. Figure 7 visualizes the post hoc analysis using the CD diagram. From the figure, it is clear that MDN-Joint works best and is statistically significantly different from the state-of-the-art methods.
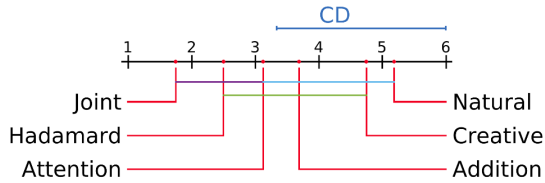

Figure 7: The mean rank of all the models on the basis of METEOR score are plotted on the x-axis. Here Joint refers to our MDN-Joint model and others are the different variations described in section 4.1.3 and Natural (Mostafazadeh et al., 2016), Creative (Jain et al., 2017). The colored lines between the two models represents that these models are not significantly different from each other.
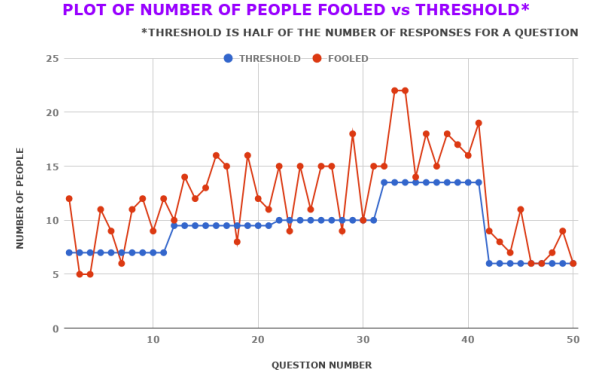

Figure 8: Perceptual Realism Plot for human survey. Here every question has different number of responses and hence the threshold which is the half of total responses for each question is varying. This plot is only for 50 of the 100 questions involved in the survey. See section 5.4 for more details.

5.4 Perceptual Realism

A human is the best judge of naturalness of any question, We evaluated our proposed MDN method using a 'Naturalness' Turing test (Zhang et al., 2016) on 175 people. People were shown an image with 2 questions just as in figure 1 and were asked to rate the naturalness of both the questions on a scale of 1 to 5 where 1 means 'Least Natural' and 5 is the 'Most Natural'. We provided 175 people with 100 such images from the VQG-COCO validation dataset which has 1250 images. Figure 8 indicates the number of people who were fooled (rated the generated question more or equal to the ground truth question). For the 100 images, on an average 59.7% people were fooled in this experiment and this shows that our model is able to generate natural questions.

6 Conclusion

In this paper we have proposed a novel method for generating natural questions for an image. The approach relies on obtaining multimodal differential embeddings from image and its caption. We also provide ablation analysis and a detailed comparison with state-of-the-art methods, perform a user study to evaluate the naturalness of our generated questions and also ensure that the results are statistically significant. In future, we would like to analyse means of obtaining composite embeddings. We also aim to consider the generalisation of this approach to other vision and language tasks.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

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

Alan R Chappell, Andrew J Cowell, David A Thurman, and Judi R Thomson. 2004. Supporting mutual understanding in a visual dialogue between analyst and computer. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 48, pages 376–380. SAGE Publications Sage CA: Los Angeles, CA.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Xinlei Chen and C Lawrence Zitnick. 2015. 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.

Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. 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)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

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

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

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. Janes v0. 4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina*, 2(4):2.

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

Siddha Ganju, Olga Russakovsky, and Abhinav Gupta. 2017. What's in a question: Using visual questions as a form of supervision. In *CVPR*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out:Proceedings of the ACL-04 workshop*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

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

Lin Ma, Zhengdong Lu, and Hang Li. 2016. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mateusz Malinowski and Mario Fritz. 2014. A multiworld approach to question answering about realworld scenes based on uncertain input. In *Advances in Neural Information Processing Systems*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1802–1813.

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*.

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

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

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

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

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

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

Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015. Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460*.

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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