Deep Bayesian Network for Visual Question Generation

Badri N. Patro Vinod K. Kurmi Sandeep Kumar Vinay P. Namboodiri
Indian Institute of Technology, Kanpur
{badri,vinodkk,sandepkr,vinaypn}@iitk.ac.in

Abstract

Generating natural questions from an image is a semantic task that requires using vision and language modalities to learn multimodal representations. Images can have multiple visual and language cues such as places, captions, and tags. In this paper, we propose a principled deep Bayesian learning framework that combines these cues to produce natural questions. We observe that with the addition of more cues and by minimizing uncertainty in the among cues, the Bayesian network becomes more confident. We propose a Minimizing Uncertainty of Mixture of Cues (MUMC), that minimizes uncertainty present in a mixture of cues experts for generating probabilistic questions. This is a Bayesian framework and the results show a remarkable similarity to natural questions as validated by a human study. We observe that with the addition of more cues and by minimizing uncertainty among the cues, the Bayesian framework becomes more confident. Ablation studies of our model indicate that a subset of cues is inferior at this task and hence the principled fusion of cues is preferred. Further, we observe that the proposed approach substantially improves over state-of-the-art benchmarks on the quantitative metrics (BLEU-n, METEOR, ROUGE, and CIDEr). Here we provide project link for Deep Bayesian VQG https://delta-lab-iitk.github.io/BVQG/.

1. Introduction

The interaction of humans and automated systems is an essential and increasingly active area of research. One such aspect is based on vision and language-based interaction. This area has seen many works related to visual question answering [1] and visual dialog [11]. Current dialog systems as evaluated in [9] show that when trained between bots, AI-AI dialog systems show improved performance, but that does not translate to actual improvement for Human-AI dialog. This is because, the questions generated by bots are not natural and therefore do 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, in [20] the authors show that unanswered questions can be used for improving VQA, Image captioning and Object Classification. So the generation of natural questions will further improve performance on these tasks. While not as well studied as the other tasks of answering questions or carrying a conversation, there has been work aimed at generating natural and engaging questions from an image [38, 23] which is the VQG task. The underlying principle for all these methods is an encoder-decoder formulation. We argue that there are underlying cues that motivate a natural question about an image. It is essential to incorporate these cues while generating questions. For each image, there may be a different underlying cue that is most pertinent. For some images, the place may be important ('Is it a cowshed?') whereas for others the subject and verb may provide more context ('Are the horses running?'). Our work solves this problem by using a principled approach for multimodal fusion by using a *mixture of experts (MoE)* model to combine these cues. We hypothesize that the joint distribution posterior based on the cues correlates with natural semantic questions.

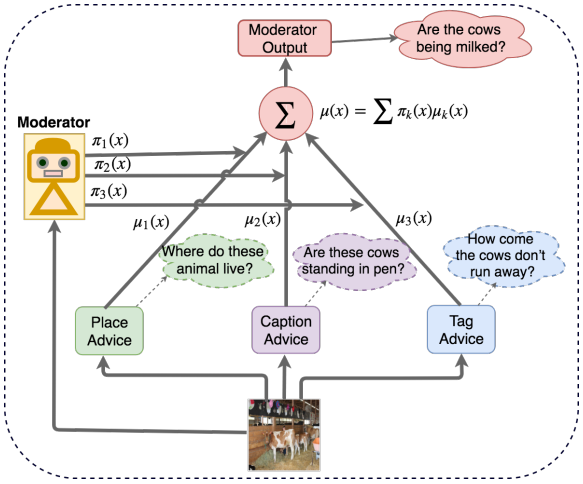

Figure 1. Here we give an overview of our network. We have three experts which provide us with information (advice) related to different cues. These are shown as Place Expert, Caption Expert and Tag Expert respectively. Then we have a moderator which weighs these advices and passes the resultant embedding to the decoder to generate natural question.

To verify our hypothesis, we systematically consider ap-

proaches to extract and combine descriptors from an image and its caption. We argue that some of the critical descriptors that could provide useful context are: a) Location description, b) Subject and Verb level description and c) Caption level description.

- *Location description*: For certain kinds of images that involve locations such as train-stations or bus-stations, the context is dominated by location. For instance, natural questions may relate to a bus or a train and hence could be more related to the destination or time related information. In such scenarios, other cues may be secondary cues. In our work, we obtain a posterior probability distribution that captures the probability of the location cue by training a Bayesian deep CNN.

- *Subject and Verb level description*: In certain images, the main context may relate to the subject and verb (for instance, food and eating). In such cases, subject-verb combinations dominate the context. Given an image we obtain a posterior probability distribution over the set of tags.

- *Caption*: For a set of natural questions, an important context could be obtained from an image caption. We can now use state-of-the-art image captioners to generate descriptive captions of an image, which is useful information for generating questions pertinent to the same image. We use this information by obtaining a posterior distribution on the caption generator.

We show the GradCAM [46] visualisations for the questions generated on the basis of single and multiple cues in Figure 2. We see that the model focuses on different regions when provided single cues (Place and Caption in the second and third image in Figure 2) and asks poor questions, but when we provide both the Place and Caption cues to the model, it focuses on correct regions which results in sensible question. So incorporating multiple cues through a principled approach in our model should lead to more natural questions.

Figure 2. Here we visualize the GradCAM maps corresponding to single and multiple cues for question generation.

We combine these distributions (cues) to estimate latent distributions which are then mixed through a moderator network and used by a decoder module to generate questions. On obtaining these distributions, we then obtain the combination of the cues that provides us with a combined latent distribution that is used by a decoder module that generates the question. The approach is illustrated in figure 1. The main aspect that we focus on this paper is to investigate a number of cues that can provide us with the necessary semantic correlation that can guide generation of natural questions and the ways in which these cues can be combined. The contributions of this paper are as follows:

- We provide Bayesian methods for obtaining posterior distributions by considering the advice of various experts that capture different cues embedding and aid in generating more natural questions.

- We propose a method to capturing and minimizing uncertainty (aleatoric and epistemic) in question generation task.

- We show that by Minimizing Uncertainty in Multiple Cues (MUMC) method with the help of Gaussian cross-entropy and variance minimizing loss, improves the score.

- We also analyze the different ablations of our model and show that while each of these cues does affect the generation, a probabilistic combination of these improves the generation in a statistically significant way.

2. Related Work

The task of automatically generating questions is well studied in the NLP community, but it has been relatively less explored for generating image related questions. On the other hand, there has been extensive work done in the Vision and Language domain for solving image captioning [6, 15, 30, 48, 56, 25, 57, 14, 10, 24, 58], Visual Question Answering (VQA) [37, 33, 1, 45, 34, 41, 16, 60, 28, 44] and Visual Dialog [11, 2, 54, 55, 61]. However, Visual Question Generation (VQG) is the task aimed at generating 'natural and engaging' questions for an image and was proposed by Mostafazadeh *et al.* [38]. It focuses more on questions which are interesting for a person to answer and not on those which can be answered simply by looking at the image and hence could be used to evaluate a computer vision model. One of the works in this area is [59] where the authors proposed a method for continuously generating questions from an image and subsequently answering the questions being generated. In [38], the authors used an encoder-decoder based framework that has been further adopted in our work by considering various contexts. In [23], the authors extend it by using a Variational Autoencoder based sequential routine to obtain natural questions by performing sampling of the latent variable. In a very recent work by [43], the authors use an exemplar based multimodal encoder-decoder approach to generate natural questions. Our work extends our previous work [43] by proposing a deep Bayesian multimodal network that can generate multiple questions for an image.

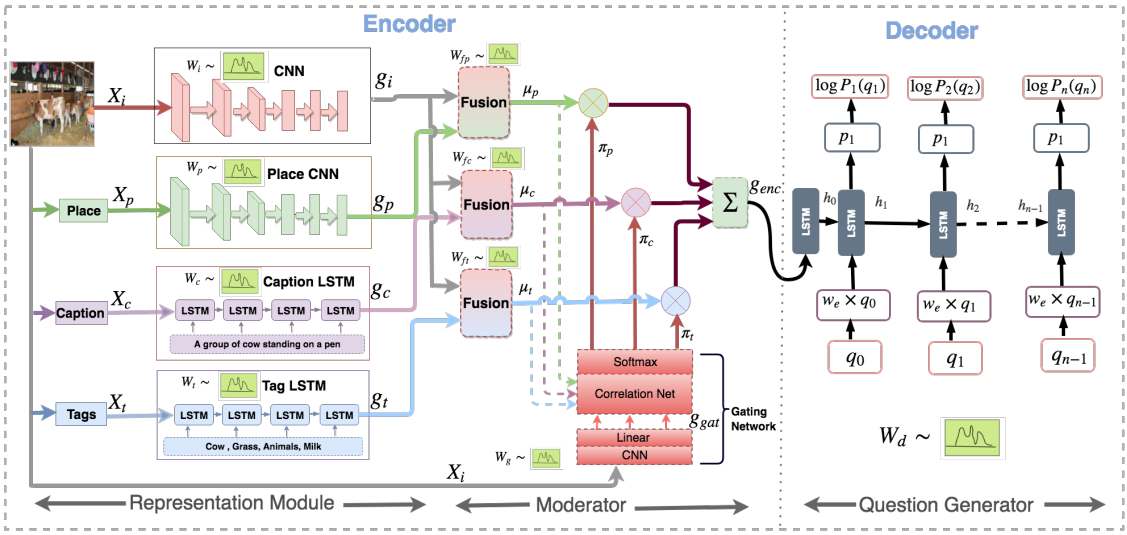

Figure 3. Multi-Cue Bayesian Moderator Network. We first use a Bayesian CNN/LSTM to obtain the embeddings g_i, g_p, g_c, g_t and then fuse those using the Fusion Module to get μ_p, μ_c, μ_t. These embeddings are then passed to the Moderator network. These are then fed to the decoder to get the questions for each image.

It has been shown that for small datasets, Bayesian Neural Networks [17] are robust to overfitting and weights are easily learned. The earliest works in Bayesian Neural networks by [39, 40, 35, 12, 13, 51, 8] focused on the idea that model weights come from a random distribution and tried to approximate the posterior distribution of the weights given the data. To approximate the intractable posterior distribution, variational inference is one of the existing approaches introduced by [22, 5, 21, 7]. Gaussian distribution is a popular choice for the variational distribution, but it is computationally expensive [7]. This can be overcome by using a Bernoulli distribution which we also use in our work. There has been some recent work which applies these concepts to CNNs [17] (Bayesian CNN) and LSTMs [19] (Bayesian LSTM) for obtaining probabilistic representations of images and sequential data respectively. These methods show that using Dropout [49] training in deep neural networks (DNN) can be interpreted as an approximate Bayesian inference in deep Gaussian processes and can be used to represent uncertainty in DNNs. Recently Kurmi *et al.* [31] has proposed a method to minimise uncertainty in source and target domain and Patro *et al.* [44] has proposed an gradient based method to minimise uncertainty in the attention regions for solving VQA task. To the best of our knowledge, the usage of Bayesian fusion of cues for end-to-end inference setting has not been considered previously for a deep learning setting. Having a principled approach for fusing multiple cues will be beneficial even in other settings such as autonomous robots, cars, etc. We compare our work with the some related works for question generation in the experimental section and show that considering different contexts and combining them using a product of experts setup can improve the task of natural question generation.

3. Method

We adopt a generation framework that uses an image embedding combined with various cues namely, place, caption and tag embeddings to generate natural questions. We propose a Multi Cue Bayesian Moderator Network (MC-BMN) to generate questions based on a given image.

3.1. Finding Cues

As location is one of an important cue, we used different scene semantic categories present in the image as a place-based cue to generate natural questions. We use pre-trained PlaceCNN [64] which is modeled to classify 365 types of scene categories. Captions also play a significant role in providing semantic meaning for the questions for an image. Tags provide information relevant to various topics in an image. We are using parts-of-speech (POS) tagging for captions to obtain these. The tags are clustered into three categories namely, Noun tag, Verb tag and Question tags. Noun tag consists of all the noun & pronouns present in the caption, and similarly, the Verb tag includes verb & adverbs present in the caption sentence whereas the Question tags consist of (Why, How, What, When, Where, Who and Which). Each tag token is represented as a one-hot vector of the dimension of vocabulary size. For generalization, we have considered five tokens from each category of the tags.

3.2. Representation module

Given an input image x_i, we obtain its embedding g_i using a Bayesian CNN [17] that we parameterize through a function $G(x_i, W_i)$ where W_i are the weights of the Bayesian CNN. We have used a pretrained VGG-19 [47] CNN trained on ImageNet for image classification task as the base CNN which was also used by the previous state-

of-the-art methods like [38] and [23]. To make Bayesian CNN [17], We use pretrained CNN layers and put Dropout layer with dropout rate p, before each CNN layer to capture Epistemic Uncertainty. Then, we extracted g_i, a d-dimensional image feature from the Bayesian CNN network as shown in figure 3. Similarly we obtain place embeddings g_p using a Bayesian PlaceCNN $G(x_p, W_p)$ for place input x_p. The Bayesian PlaceCNN is the pretrained PlaceCNN with similar placement of dropout layer as the VGG-19 CNN.

To generate caption and tag embeddings, we use a V (size of vocabulary) dimensional one-hot vector representation for every word in the Caption & Tags and transform them into a real valued word embedding X_{we} for each word using a matrix $W_C \in \mathcal{R}^{E_C \times V}$. Then the E_C dimensional word embeddings are fed to the Bayesian LSTM to obtain the required representations for the caption and tag inputs. Bayesian LSTM is designed by adding dropout layer into each gate of the LSTM and output layer of the LSTM as done in [19]. So we obtain g_c, g_t using a Bayesian LSTMs $F(x_c, W_c)$ and $F(x_t, W_t)$ for caption input x_c, and tag input x_t respectively.

3.3. Bayesian Fusion Module

There have been some works for VQA which use a projection of multiple modalities to a common space with the help of a fusion network to obtain better results [1, 65]. We use a similar fusion network to combine multiple modalities, namely caption, tag and place with the image. The fusion network can be represented by the following equations:

$$\mu_p = W_{pp} * \tanh(W_i g_i \otimes W_p g_p + b_p)$$

$$\mu_c = W_{cc} * \tanh(W_i g_i \otimes W_c g_c + b_c)$$

$$\mu_t = W_{cc} * \tanh(W_i g_i \otimes W_t g_t + b_t)$$

where, $g_\star$ is the embedding for corresponding cues, $W_\star$ and $b_\star$ are the weights and the biases for different cues($\star$ represent$\{p, c, t\}$). Here $\otimes$ represent element-wise multiplication operation. We use a dropout layer before the last linear layer for the fusion network. We also experimented with other fusion techniques like addition, attention, and concatenation but element-wise multiplication performed the best for all the metrics.

3.4. Bayesian Moderator Module

We propose a Moderator Module to combine the fused embeddings. The proposed model is similar to the work of [52, 3, 62]. The Moderator module receives input image x_i and obtains a gating embedding g_{gat} using a Bayesian CNN that we parametrize through a function $G(x_i, W_g)$. Then, a correlation network finds the correlation between gating embedding g_{gat} and μ_B to obtain scaling factors π_B, where $B \in \{p, c, t\}$. Finally, Moderator combines the fused embeddings μ_B with the scaling factors π_B to obtain the final embedding g_{enc}.

$$g_{gat} = BayesianCNN(x_i; W_g)$$

$$\pi_B = softmax(g_B * g_{gat}) \forall B \in \{p, c, t\}$$

$$g_{enc} = \sum_{B \in \{p,c,t\}} \pi_B * \mu_B$$

3.5. Decoder: Question Generator

The decoder's task is to predict the whole question sentence given an image I and its cues (C). The probability for a question word depends on the previously generated words. This conditional probability $P(q_{t+1}|I, C, q_0, ..., q_t)$ is modeled with a LSTM for sequential tasks such as machine translation [50]. We use a Bayesian LSTM similar to the one used in our Representation Module for this question generation task. At t =-1, we feed the moderator advice g_{enc} to the LSTM. The output of the word with maximum probability in the distribution $P(q_t|g_{enc}, h_t)$ in the LSTM cell at time step t is fed as input to the LSTM cell at step t+1 as mentioned in the decoder in figure 3. At time steps $t = 0 : (T-1)$, the softmax probability is given by:

$$\begin{aligned} x_{-1} &= g_{enc} \\ x_t &= W_C * q_t, \forall t \in \{0, 1, 2, ...T-1\} \\ h_{t+1} &= LSTM(x_t, h_t), \forall t \in \{0, 1, 2, ...N-1\} \\ o_{t+1} &= W_o * h_{t+1} \\ \hat{y}_{t+1} &= P(q_{t+1}|g_{enc}, h_t) = \text{softmax}(o_{t+1}) \\ Loss_{t+1} &= \text{loss}(\hat{y}_{t+1}, y_{t+1}) \end{aligned} \tag{1}$$

where h_t is the hidden state and o_t is the output state for LSTM.

3.5.1 Uncertainty in Generator Module

The decoder module is generating diverse words which lead to uncertainty in the generated sentences. The uncertainty present in the model can be captured by estimating Epistemic uncertainty [26], and the uncertainty present in the data can be captured by estimating Aleatoric uncertainty [18]. The predictive uncertainty [36] is the total uncertainty which is the combination of both uncertainties. The predictive uncertainty measures the model's capability for generating question word token by focusing on various cues (caption, tag, and place) networks. We use the similar Bayesian decoder network to capture predictive uncertainty by approximating the posterior over the weights of Bayesian decoder using MC-dropout as described in [27, 31, 44]. The uncertainty in these cues moderators occurs mainly due to either noise or lack of data to learn mixture of cues. We proposed a method Minimising Uncertainty for mixture of Cue (MUMC), which enhances model performance by minimizing uncertainty.

(a) MC-BMN (b) MC-BMN-2 (c) MC-SMix

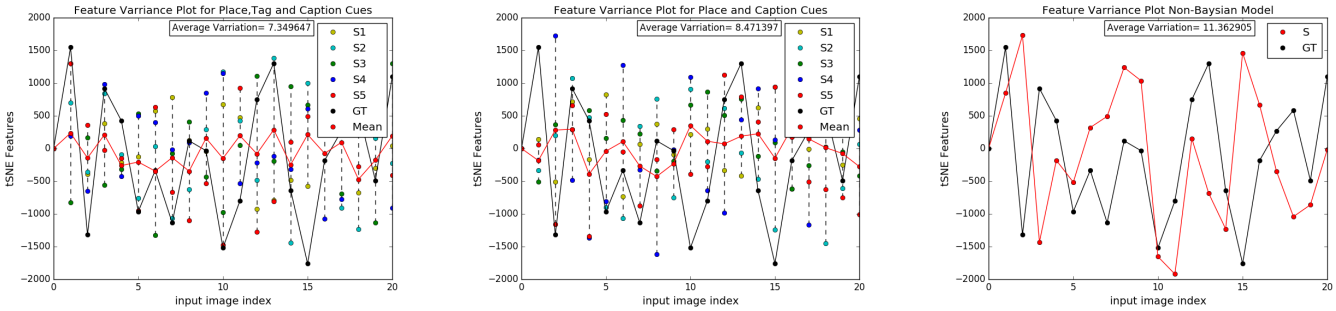

Figure 4. Variance plots for Bayesian and Non-Bayesian networks for a toy example of 20 images. We have drawn 5 samples of each image using Monte-Carlo sampling from a distribution (this is predictive posterior distribution for the Bayesian case) and then plot the mean features of these 5 samples along with the ground truth features. MC-BMN (3 cues) reduces normalized variance (difference in mean feature value & ground truth feature value) as compared to two cues(MC-BMN-2). Whereas for MC-SMix(Non-Bayesian network), the variance is too high as compared to MC-BMN.

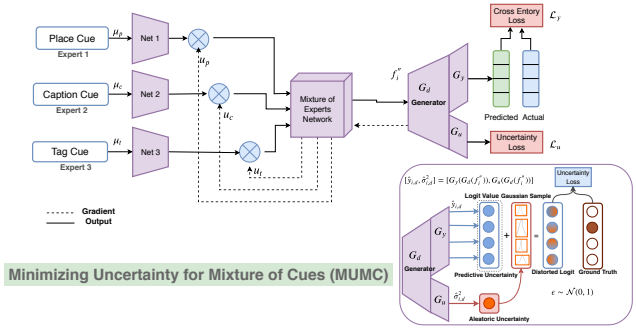

Figure 5. Model architecture for minimizing uncertainty for mixture of Cues

MUMC: The decoder generates a logit out $y_{i,g}$ and variance network predict variance in each generated word token.

$$y_{i,g} = G_y(G_o(f_i)), \quad v_{i,g} = G_v(G_o(f_i)) \tag{2}$$

where $f_i = g_{gen}$ is the output feature of the Bayesian Moderator Module. G_o is the decoder network, G_y is the final word token classifier and G_v is the variance predictor network. In order to capture uncertainty in the data, we learn observational noise parameter $\sigma_{i,g}$ for each input point x_i and its cues. This can be achieved by corrupting the logit value ($y_{i,g}$) with the Gaussian noise with variance $\sigma_{i,g}$ (diagonal matrix with one element for each logits value) before the softmax layer. We defined a Logits Reparameterization Trick (LRT), which combines two outputs $y_{i,g}, \sigma_{i,g}$ and then we obtain a loss with respect to the ground truth. That is, after combining we get $\mathcal{N}(y_{i,g}, (\sigma_{i,g})^2)$ which is expressed as:

$$\hat{y}_{i,t,g} = y_{i,g} + \epsilon_{t,g} \odot \sigma_{i,g}, \quad where \quad \epsilon_{t,g} \sim \mathcal{N}(0,1) \tag{3}$$

$$\mathcal{L}_u = \sum_i \log \frac{1}{T} \sum_t \exp\left(\hat{y}_{i,t,g} - \log \sum_{M'} \exp \hat{y}_{i,t,M'}\right) \tag{4}$$

Where M is the total word tokens, $\mathcal{L}_u$ is minimized for true word token M, and T is the number of Monte Carlo simulations. $M^{'}$ is the element in the logit vector $y_{i,t}$ for all the classes. $\sigma_{i,g}$ is the standard deviation, ($\sigma_{i,g} = \sqrt{v_{i,g}}$).

We compute gradients of the predictive uncertainty σ_g^2 of our generator with respect to the features f_i. We first compute gradient of the uncertainty loss $\mathcal{L}^v$ with respect to cues moderator feature $f_i = g_{gen}$ i.e. $\frac{\partial \mathcal{L}_v}{\partial f_i}$ Now we pass the uncertainty gradient through a gradient reversal layer to reverse the gradient of the all the cues is given by

$$\nabla_y = -\gamma * \frac{\partial \mathcal{L}_u}{\partial f_i}$$

We perform a weighted combination of forward cues moderator feature maps μ_p, μ_c, μ_t with the reverse uncertainty gradients i.e.

$$\nabla^{'}_{g_{enc}} = \sum_{B \in \{p,c,t\}} -\gamma * \frac{\partial \mathcal{L}_u}{\partial f_i} * \mu_B$$

We use residual connection to obtain the final moderator cue feature by combining original cue moderator feature with the gradient certainty mask $\nabla^{''}_y$ and is given by:

$$g^{'}_{enc} = g_{enc} + \sum_{B \in \{p,c,t\}} \nabla^{'}_{g_{enc}} * g_{enc}$$

From this moderator feature we are generating question word tokens.

3.6. Cost Function

We estimate aleatoric uncertainty in logit space by distorting each logit value by the variance obtained from data. The uncertainty present in each logit value can be minimized using cross-entropy loss on Gaussian distorted logits as shown in equation- 3. The distorted logits is obtained using Gaussian multivariate function with positive diagonal variance. The uncertainty distorted loss is the difference between actual cross entropy loss and the uncertainty loss

(a) COCO (b) BING (c) FLICKR

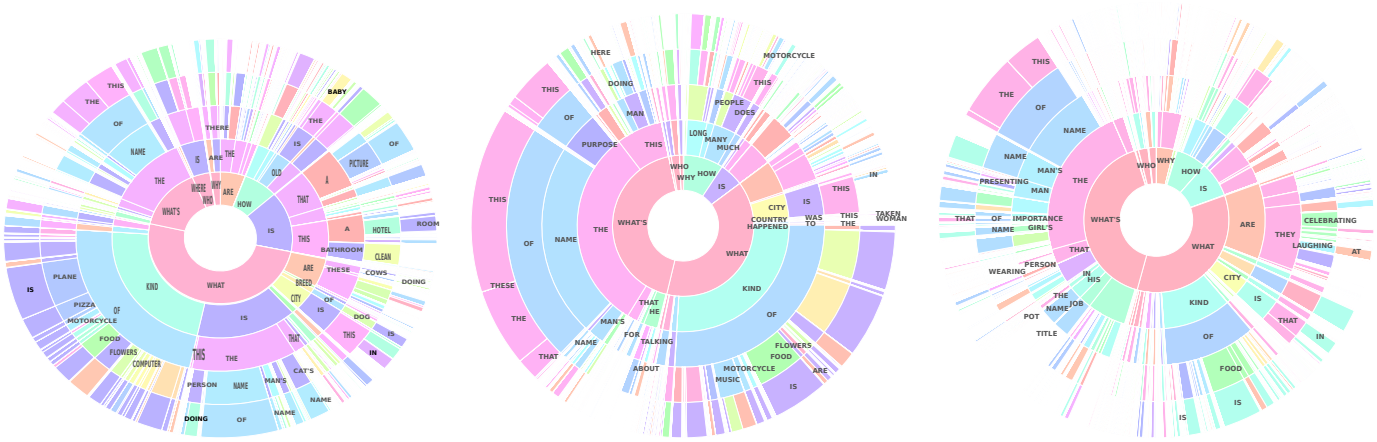

Figure 6. Sunburst plot of generated questions for MC-BMN on VQG-COCO dataset,VQG-Bing dataset, VQG-Flickr dataset are shown in Fig-a, Fig-b, Fig-c respectably : The i^{th} ring captures the frequency distribution over words for the i^{th} word of the generated question. While some words have high frequency, the outer rings illustrate a fine blend of words.

mentioned in equation- 4. The difference is passed through an activation function to enhance the difference in either direction and is given by :

$$\mathcal{L}_u = \begin{cases} \alpha(\exp^{[\mathcal{L}_p - \mathcal{L}_y]} - 1), & \text{if } [\mathcal{L}_p - \mathcal{L}_y] < 0. \\ [\mathcal{L}_p - \mathcal{L}_y], & \text{otherwise.} \end{cases} \quad (5)$$

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

In the question generator module, we use the cross entropy loss function between the predicted and ground truth question, which is given by:

$$L_{gen} = \frac{-1}{NM} \sum_{i=1}^{N} \sum_{t=1}^{M} y_t \log p(q_t|(g_{enc})_i, q_0, ..q_{t-1}) \quad (6)$$

where, N is the total number of training examples, M is the total number of question tokens, $\mathrm{P}(q_t|(g_{enc})_i, q_0, ..q_t)$ is the predicted probability of the question token, y_t is the ground truth label. We have provided the pseudo-code for our method in our project webpage.

4. Experiments

We evaluate the proposed method in the following ways: First, we evaluate our proposed MC-BMN against other variants described in section 4.2. Second, we further compare our network with state-of-the-art methods such as Natural [38] and Creative [23]. Third, we have shown in figure 4, the variance plots for different samples drawn from the posterior for Bayesian and Non-Bayesian methods. Finally, we perform a user study to gauge human opinion on the naturalness of the generated question and analyze the word statistics with the help of a Sunburst plot as shown in Figure 6. We also consider the significance of the various methods for combining the cues as well as for the state-of-the-art models. The quantitative evaluations are performed using standard metrics namely BLEU [42], METEOR [4], ROUGE [32] and CIDEr [53]. BLEU metric scores show strong correlations with human for the VQG task and is recommended by Mostafazadeh *et al.* [38] for further bench-marking. In the paper, we provide the comparison with respect to only BLEU-1 and METEOR metrics and the full comparison with all metrics(BLEU-n, CIDER and ROUGE) and further details are present in our project webpage[1].

Method	BLEU1	METEOR	ROUGE	CIDEr
MC-SMix	31.1	19.1	32.6	42.8
MC-BMix	36.4	22.6	40.7	46.6
MC-SMN	33.1	21.1	37.6	47.8
MC-BMN +PC	24.6	11.1	24.0	45.2
MC-BMN	**40.7**	**22.6**	**41.9**	**49.7**

Table 1. Ablation Analysis on VQG-COCO Dataset.It has the different variations of our model described in 'Comparison with State-of-the-Art and Ablation Analysis' section of the paper. As expected the performance with the generated captions is not as good as with the ground truth captions. Note that these are the max scores over all the epochs. PC tends for Predicted Caption

4.1. Dataset

We conduct our experiments on Visual Question Generation (VQG) dataset [38], which contains human annotated questions based on images of MS-COCO dataset. This dataset [38] was developed for generating natural and engaging questions. It 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 [23] also used the questions from VQA dataset [1]

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

Methods	BLEU1		METEOR	
	Max	Avg	Max	Avg
Natural [38]	19.2	-	19.7	-
Creative [23]	35.6	-	19.9	-
MDN [43]	36.0	-	**23.4**	-
Img Only (Bernoulli Dropout (BD))	21.8	19.57 $\pm$2.5	13.8	13.45$\pm$1.52
Place Only(BD)	26.5	25.36$\pm$1.14	14.5	13.60$\pm$0.40
Cap Only (BD)	27.8	26.40 $\pm$1.52	18.4	17.60$\pm$0.65
Tag Only (BD)	20.3	18.13 $\pm$2.09	12.1	12.10$\pm$0.61
Img+Place (BD)	27.7	26.96 $\pm$0.65	16.5	16.00$\pm$0.41
Img+Cap (BD)	26.5	24.43 $\pm$1.14	15.0	14.56 $\pm$0.31
Img+Tag (BD)	31.4	29.96 $\pm$1.47	20.1	18.96$\pm$1.08
Img+Place+Cap (BD)	28.7	27.86 $\pm$0.74	18.1	15.56$\pm$1.77
Img+Place+Tag (BD)	30.6	28.46 $\pm$1.58	18.5	17.60 $\pm$0.73
Img+Cap+Tag (BD)	37.3	36.43 $\pm$1.15	21.7	20.70$\pm$0.49
MC-SMN(Img+Place+Cap+Tag(w/o Dropout))	33.3	33.33 $\pm$0.00	21.1	21.10 $\pm$0.00
MC-BMN (Img+Place+Cap+Tag (Gaussian Dropout))	38.6	35.63 $\pm$2.73	22.9	21.53 $\pm$1.06
MC-BMN(Img+Place+Cap+Tag(BD)) (**Ours**)	**40.7**	**38.73** $\pm$1.67	22.6	22.03 $\pm$0.80
Humans[38]	86.0	–	60.8	–

Table 2. Comparison with **state-of-the-art** and different combination of **Cues**. The first block consists of the SOTA methods, second block depicts the models which uses only a single type of information such as Image or Place, third block has models which take one cue along with the Image information, fourth block takes two cues along with the Image information. The second last block consists of variations of our method. First is MC-SMN (Simple Moderator Network) in which there is no dropout (w/o Dropout) at inference time as explained in section 4.3 and the second one uses Gaussian dropout instead of the Bernoulli dropout (BD) which we have used across all the models.

for training purpose, whereas the work by [38] uses only the VQG-COCO dataset. We understand that the size of this dataset is small and there are other datasets like VQA [1], Visual7W [66] and Visual Genome [29] which have thousands of images and questions. But, VQA questions are mainly visually grounded and literal, Visual7w questions are designed to be answerable by only the image, and questions in Visual Genome focus on cognitive tasks, making them unnatural for asking a human [38] and hence not suited for the VQG task.

4.2. Comparison with different cues

The first analysis is considering the various combinations of cues such as caption and place. The comparison is provided in table 2. The second block of table 2 depicts the models which use only a single type of information such as Image or Place. We use these models as our baseline and compare other variations of our model with the best single cue. The third block takes into consideration one cue along with the Image information, and we see an improvement of around 4% in BLEU1 and 2% in METEOR score. The fourth block takes two cues along with the Image information and obtains an improvement of around 10% in BLEU and 3% in METEOR scores. The question tags performs the best among all the 3 tags. This is reasonable as question tag can guide the type of question. The second last block consists of variations of our method. the first variation corresponds to the model in which there is no dropout at inference time and the second one uses Gaussian dropout instead of the Bernoulli dropout which we have used across all the models. As we can see, the application of dropout leads to a significant increase in the BLEU score and also Bernoulli dropout works best. We also observe that our proposed method MC-BMN gets an improvement of 13% in BLEU and 5% in METEOR score over the single cue baselines. Tags work well in general along with other cues than caption as it provides more precise information compared to the caption, but the performance drops significantly if only the tag information is provided as there is not much information for generating sensible questions. While comparing the various embedding, we also evaluated various ways of integrating the different cues to obtain joint embedding.

4.3. Comparison with state-of-the-art methods and Ablation Analysis

The comparison of our method with various state-of-the-art methods and ablation analysis is provided in table 2. We observe that in terms of METEOR score, obtain an improvement of around 3% using our proposed method over previous work by Mostafazadeh et. al [38] and Jain et. al [23]. For BLEU score the improvement is around 20% over [38], 5% over [23]. But it's still quite far from human performance.

Ablation Analysis: We consider different variants of our methods. These are use of Conventional CNN and a concatenation of the various embeddings (Multi Cue Simple Mixture (MC-SMix)), a Bayesian CNN and concatenation of the various embeddings (Multi Cue Bayesian Mixture (MC-BMix)), and the final one uses a mixture of experts

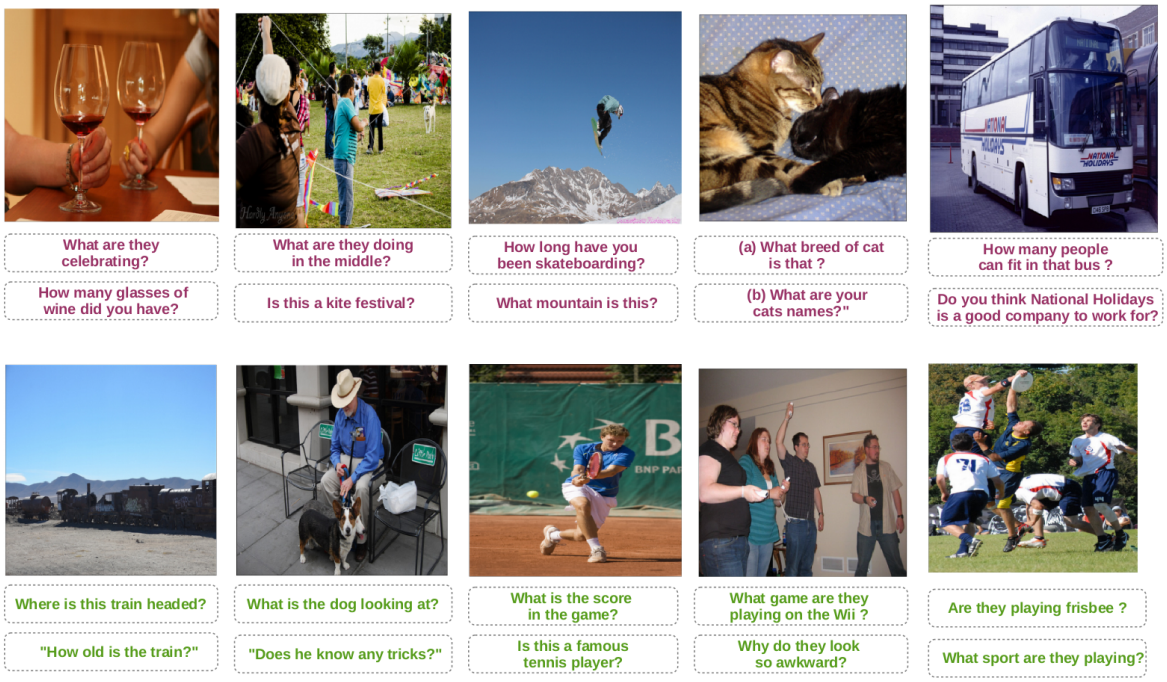

Figure 7. Examples of questions generated by our method for different images. First question in each image is generated by our method and second one is the ground truth question. More results are present in the project webpage.

along with a conventional CNN (Multi Cue Simple Moderator Network (MC-SMN)). MC-SMN actually corresponds to our MC-BMN method without dropout. Our proposed method improves upon these ablations.

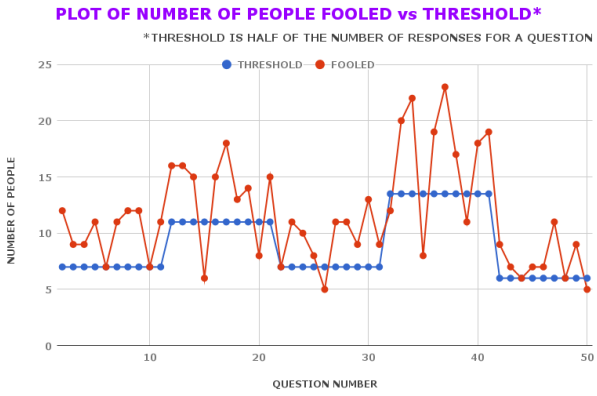

Figure 8. Perceptual Realism Plot for human survey (section 4.4). The blue and red dots represent the threshold and the number of people fooled for each question respectively. Here every question has different number of responses and hence the threshold for each question is varying. Also, we are only providing the plot for 50 of 100 questions involved in the survey.

4.4. Perceptual Realism

A human is the best judge of the naturalness of any question; we also evaluated our proposed MC-BMN method using a 'Naturalness' Turing test [63] on 175 people. People were shown an image with two questions just as in figure 7 and were asked to rate the naturalness of both the questions on a scale of 1 to 5 where one means 'Least Natural' and 5 is the 'Most Natural.' We provided them 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 61.8%, people were fooled. If we provide both questions as the ground truth ones then on an average 50 % people were fooled, and this shows that our model can generate natural questions.

5. Conclusion

In this paper, we have proposed a novel solution for the problem of generating natural questions for an image. The approach relies on obtaining the advice of different Bayesian experts that are used for generating natural questions. We provide a detailed comparison with state of the art baseline methods, perform a user study to evaluate the naturalness of the questions and also ensure that the results are statistically significant. Our work introduces a principled framework to include cues for vision and language-based interaction. We aim to further validate the generalization of the approach by extending this approach to other vision and language tasks. The resulting approach has been also analysed in terms of Conventional CNN, Bayesian LSTM with product of experts and we observe that the proposed Bayesian Expert model improved over all the other variants.

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] T. Baldacchino, E. J. Cross, K. Worden, and J. Rowson. Variational bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mechanical Systems and Signal Processing*, 66:178–200, 2016.

[4] 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.

[5] D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. pages 215–238, 1998.

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

[7] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[8] W. L. Buntine and A. S. Weigend. Bayesian backpropagation. *Complex systems*, 5(6):603–643, 1991.

[9] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2017.

[10] 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.

[11] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield. Large automatic learning, rule extraction, and generalization. *Complex systems*, 1(5):877–922, 1987.

[13] J. S. Denker and Y. Lecun. Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems*, pages 853–859, 1991.

[14] 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.

[15] 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.

[16] 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.

[17] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

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

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

[20] S. Ganju, O. Russakovsky, and A. Gupta. What's in a question: Using visual questions as a form of supervision. In *CVPR*, 2017.

[21] A. Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2348–2356, 2011.

[22] G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proc. of the Conference on Computational learning theory (COLT)*, pages 5–13. ACM, 1993.

[23] U. Jain, Z. Zhang, and A. G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6485–6494, 2017.

[24] J. Johnson, A. Karpathy, and L. 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.

[25] 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.

[26] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[27] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. 2018.

[28] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017.

[29] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[30] 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.

[31] V. K. Kurmi, S. Kumar, and V. 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.

[32] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out:Proceedings of the ACL-04 workshop*, 2004.

[33] 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.

[34] 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.

[35] D. J. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[36] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[37] 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.

[38] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. 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, 2016.

[39] R. M. Neal. Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems*, pages 475–482, 1993.

[40] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[41] 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.

[42] 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.

[43] 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, 2018.

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

[45] 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.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 618–626. IEEE, 2017.

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] 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.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[50] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.

[51] N. Tishby, E. Levin, and S. A. Solla. Consistent inference of probabilities in layered networks: Predictions and generalization. In *IJCNN International Joint Conference on Neural Networks*, volume 2, pages 403–409. IEEE New York, 1989.

[52] N. Ueda and Z. Ghahramani. Optimal model inference for bayesian mixture of experts. In *IEEE Workshop on Neural Networks for Signal Processing X*, volume 1, pages 145–154. IEEE, 2000.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460*, 2015.

[60] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[61] 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.

[62] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

[63] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[64] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[65] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[66] 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.