Information Maximizing Visual Question Generation

Ranjay Krishna, Michael Bernstein, Li Fei-Fei
Stanford University
{ranjaykrishna, msb, feifeili}@stanford.edu

Abstract

Though image-to-sequence generation models have become overwhelmingly popular in human-computer communications, they suffer from strongly favoring safe generic questions ("What is in this picture?"). Generating uninformative but relevant questions is not sufficient or useful. We argue that a good question is one that has a tightly focused purpose — one that is aimed at expecting a specific type of response. We build a model that maximizes mutual information between the image, the expected answer and the generated question. To overcome the non-differentiability of discrete natural language tokens, we introduce a variational continuous latent space onto which the expected answers project. We regularize this latent space with a second latent space that ensures clustering of similar answers. Even when we don't know the expected answer, this second latent space can generate goal-driven questions specifically aimed at extracting objects ("what is the person throwing"), attributes, ("What kind of shirt is the person wearing?"), color ("what color is the frisbee?"), material ("What material is the frisbee?"), etc. We quantitatively show that our model is able to retain information about an expected answer category, resulting in more diverse, goal-driven questions. We launch our model on a set of real world images and extract previously unseen visual concepts.

1. Introduction

The task of transforming visual scenes to language, questions [47, 44], answers [64, 2] or captions [56, 30], has widely adopted an image-to-sequence architecture that encodes the image through a convolutional neural network (CNN) [33] and then decodes the language with a recurrent neural network [42]. The whole framework can be efficiently trained by maximum likelihood estimation (MLE) and has demonstrated state-of-the-art performance in various tasks [60, 8, 10, 11, 38]. However, this training procedure is not suitable for generating questions or enabling discovery of new concepts. In fact, most MLE-based training schemes have shown to produce generic questions that

Figure 1. Our new architecture generates goal-driven visual questions that maximize the likelihood of receiving an expected answer. When attempting to learn about objects or their attributes, it can generate questions aimed at attaining such answer categories.

result in uninformative answers (e.g. "yes") [22], questions (e.g. "What is the person doing?") [23], captions (e.g. "A clear day with a blue sky" [60]) or dialogue (e.g. "I don't know") [35, 51]. Simply generating a generic question is not sufficient or useful for discovering new concepts.

Instead of generating generic questions, question generation models should be goal-driven — we show how they can be trained to ask questions aimed at extracting specific answer categories. Visual question generation is not a bijection, i.e. multiple correct questions can be generated from the same image. Previous research moved away from a supervised approach to question generation to a variational approach that can generate multiple questions by sampling a latent space [23] (see Figure 2). However, previous approaches are not goal-driven — they do not guarantee that the question will result in a specific type of answer. To remedy their problem, we could encode the answer along with the image before generating the question. While such an approach allows the model to condition its question on the answer, it is neither technically feasible nor practical. Technical infeasibility arises because variational models often lead to the posterior collapsing problem [4], where the model can learn to ignore the answer when generating questions. Impracticality arises because the main purpose of asking questions is to attain an answer, implying that knowing the answer defeats the purpose of generating the question.

To tackle the first challenge, we design a visual question generation architecture that maximizes the mutual information between the generated question with the image as well as with the expected answer (see Figure 2). We call our
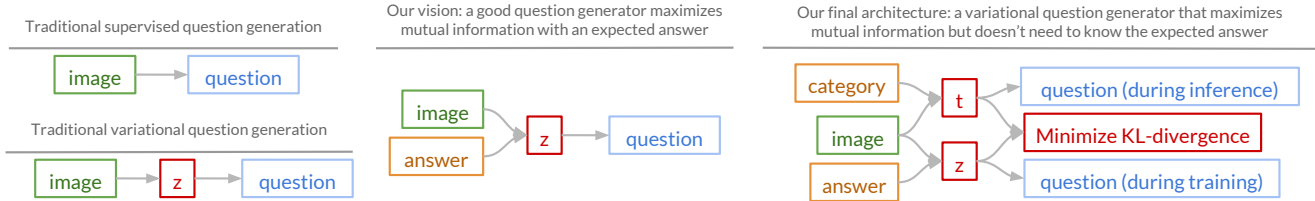
Figure 2. Since multiple questions are possible for any image, previous approaches moved away from supervised question generation to variational approaches [23]. However, this resulted in generic, uninformative questions. We argue that a good question maximizes mutual information with an expected answer. But such a model is not practical as knowing the answer defeats the purpose of generating a question. Also, such models often lead to the posterior collapsing problem [4]. Instead, we propose an architecture that maximizes mutual information between the image, answer and question while also maintaining a regularization based on answer categories. Our final model uses two latent spaces and can generate questions both in the presence and the absence of answers.

model **Information Maximizing Visual Question Generator** as it maximizes relevance with the image and expectation over the answer. Safe, generic questions that lead to uninformative answers are discouraged as they have low mutual information with either. However, optimizing for mutual information is often intractable and given the discrete tokens (words) we wish to generate, no unbiased, low variance gradient estimator exists [24, 40, 16, 52, 45, 58, 27]. We formulate our model as a variational auto-encoder that attempts to learn a joint continuous latent space between the image, question and the expected answer. Instead of directly optimizing discrete utterances, the question, image and expected answer are all trained to maximize the mutual information with this latent space. By reconstructing the image and expected answer representations, we can maximize the evidence lower bound (ELBO) and control what information the generated questions request.

The second challenge arises from the lack of an expected answer in real world deployments. Since we require an answer to map the image into a latent space, it is not possible to generate questions in the absence of an answer. Enumerating all possible answers is infeasible. Instead, we propose creating a second latent space that is learned from the image and the answer category instead of the answer itself. Answer categories can be objects, attributes, colors, materials time, etc. During training, we minimize the KL-divergence between these two latent spaces. Not only does this allow us to generate visual questions that maximize mutual information with the expected answer, it also acts as a regularizer into the original latent space. It prevents the learned latent spaces from overfitting to specific answers in the training set and forces them to generalize to categories of questions.

We annotate the VQA dataset [2] with 15 categories for the top 500 answers and use it to train our model, which queries for specific answer categories. We evaluate our model on relevance to the image and on its ability to expect the answer type. Finally, we run our model on 1000 real world images and discover 80 new objects, 40 new attributes, 17 new colors, and 8 new materials.

2. Related work

Visual understanding has been studied vigorously through question answering with the availability of large scale **visual question answering** (VQA) datasets [2, 64, 31, 25]. Current VQA approaches follow a traditional supervised MLE paradigm that typically relies on a CNN + RNN encoder-decoder formulation [56]. Successive models have improved performance by stacking attention [62, 37], modularizing components [1, 26, 21], adding relation networks [49], augmenting memory [59], and adding proxy tasks [13, 57]. While the performance of VQA models have been encouraging, they require a large labelled dataset with a predefined vocabulary. In contrast, we focus on the surrogate task of generating questions in the hopes of augmenting real world agents with the ability to expand it's visual knowledge by discovering new visual concepts.

In contrast to answering questions, generating questions has received little interest so far. In NLP, a few methods have attempted to automatically generate questions from knowledge bases using rule based [50] or deep learning based systems [12]. In computer vision, a few recent projects have explored the task of **visual question generation** to build curious visual agents [61, 23]. These projects have also either followed an algorithmic rule-based [54, 50] or learning-based [44, 47] approach. Newer papers have treated the generation process as a variational process [23] or placed it in a active learning [43] or reinforcement learning setting [61]. Our work draws inspiration from these previous methods and extends them by treating question generation as a process that maximizes mutual information between not just the image but also considers the expected answer's category. We believe that a good question generator should be goal driven — it should generate questions to receive a particular answer category.

There is a large body of work exploring **generative models** and learning latent representation spaces. Early work focused primarily on stacked autoencoders and then on restricted boltzman machines [55, 18, 19]. Recent successes of these applications have primarily been a result of varia-
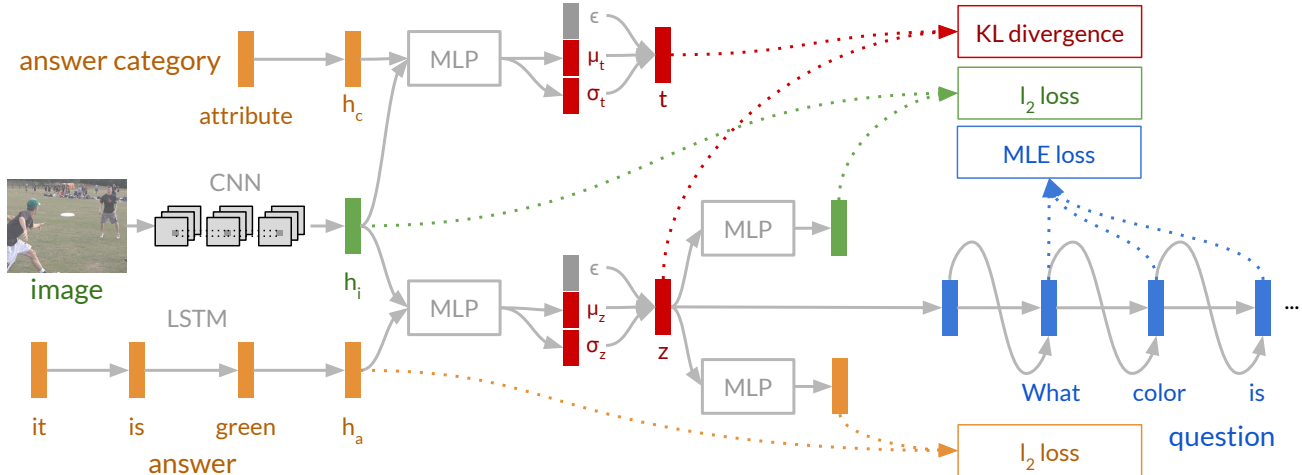
Figure 3. Training our model: we embed the image and answer into a latent space z and attempt to reconstruct them, thereby maximizing mutual information with the image and the answer. We also use z to generate questions and train it with an MLE objective. Finally, we introduce a second latent space t that is trained by minimizing KL-divergence with z. t allows us to remove the dependence on the answer when generating questions and instead grants us the ability to generate questions conditioned on the answer category.

tional auto-encoders (VAEs) [29] and generative adversarial networks (GANs) [14]. With the reparameterization trick, VAEs can be trained to learn a semi-supervised latent space to generate images [29]. They have also been extended to continuous state space [32, 3] and sequential models [15, 9]. GANs, on the other hand, can learn image representations that support basic linear algebra [46] and even enable one-shot learning by using probabilistic inference over Bayesian programs [34]. Both VAEs and GANs have disentangled their representations based on class labels or other visual variations [28, 41]. While we do not explicitly disentangle the representation, we will demonstrate later how the second latent space regularizes the original space and disentangles the representations of different answer categories.

Generative models often require a series of tricks for successful training [48, 46, 5, 4]. And even with these tricks, training them with discrete tokens is only possible by using **gradient estimators**. As we mentioned earlier, these estimators often suffer from one of two problems: high bias [27, 27] or high variance [58]. Low variance methods like Gumbel-Softmax [24], CONCRETE distribution [40], semantic hashing [27] or vector quantization [53] result in biased estimators. Similarly, low bias methods like REIN-FORCE [58] with Monte Carlo rollouts, result in high variance [16, 52, 45]. We overcome this issue by introducing a continuous latent space that maximizes mutual information with encodings of the image, question and answer. This latent space can be trained using existing VAE training procedures that attempt to reconstruct the image and answer representations. We further extend this model with a second latent space conditioned on the answer category that removes the need for an actual answer when generating questions.

3. Information Maximizing Visual Question Generator

Our aim is to generate questions that have a tightly focused purpose — questions with the aim to learn something specific about the image. Agents with the capability to request specific categories of information can extract new concepts more effectively from the real world. In this section, we detail how we design an Information Maximizing Visual Question Generator. Recall that the goal of our model is to generate questions given an image and an answer category. For example, if we want to understand materials or binary answers, our model should generate questions "What material is that desk made out of?" or "Is the desk on the right of the chair?", respectively. Our two challenges are (1) technical infeasibility caused by non-differentiable discrete tokens and variational posterior collapse and (2) impracticality of requiring answers to generate questions. We start off with a formal definition of the problem, explain why current methods fail and then detail our training and inference process.

3.1. Problem formulation

Let q denote the question we want to generate for an image i. This question should result in the an answer a of category c. For example, the question "What is the person in red doing with the ball?" should result in the answer "kicking", which belong to category "activity". Our final goal is to define a model $p(q|i, c)$. But first, let's attempt to define a simpler model $p(q|i, a)$ that maximizes the mutual information between the image and the question $I(i, q)$ and between the expected answer and the question $I(a, q)$. This
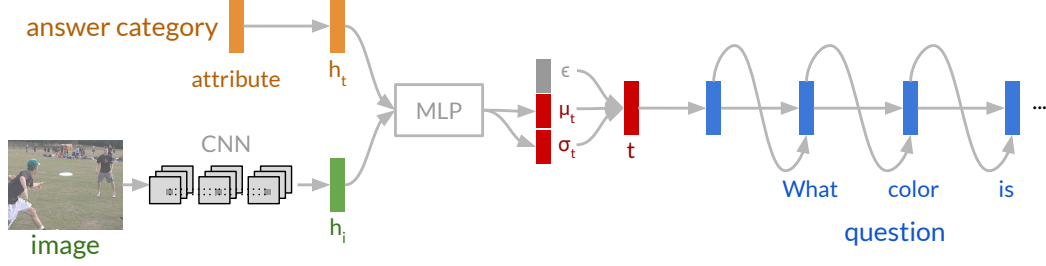
Figure 4. Inference on our model: Given an image input and an answer category (e.g. attribute), we encode both into a latent representation t, parameterized by mean μ_t and σ_t. We sample from t with noise ϵ to generate questions that are relevant to the image and who's answers result in the given answer category.

objective can be written as:

$$\max I(i, q) + \lambda I(a, q) \tag{1}$$
$$\text{s. t. } q \sim p(q|i, a)$$

where λ is a hyperparameter that adjusts for their relative importance in the optimization.

3.2. Continuous latent space

As already mentioned, directly optimizing this objective is infeasible because the exact computation of mutual information is intractable. Additionally, optimizing by estimating gradients between discrete steps is difficult as the estimator needs to have both low bias and low variance. To overcome this challenge, we introduce a continuous, dense, latent z-space. We learn a mapping $p_\theta(z|i, a)$, parameterized by θ, from the image and the expected answer to this latent space.

With this z-space, our new optimization becomes:

$$\max_\theta I(q, z|a, i) + \lambda_1 I(a, z) + \lambda_2 I(i, z)$$
$$\text{s. t. } z \sim p_\theta(z|i, a) \tag{2}$$
$$q \sim p_\theta(q|z)$$

where λ_1 and λ_2 are hyperparameters that relatively weight the mutual information terms in the optimization.

3.3. Variational mutual information maximization

So far, we have avoided discrete tokens. However, this mutual information maximization is still intractable as it requires knowing the true posteriors $p(z|i)$ and $p(z|a)$. Fortunately, we can opt to maximize its ELBO:

$$I(z, i) = \mathbb{H}(i) - \mathbb{H}(i|z)$$
$$= \mathbb{H}(i) + \mathbb{E}_{z \sim p(z,i)}[\mathbb{E}_{\hat{i} \sim p(i|z)}[\log p(\hat{i}|z)]]$$
$$= \mathbb{H}(i) + \mathbb{E}_{i \sim p(i)}[D_{KL}[p(\hat{i}|z)||p_\theta(\hat{i}|z)]$$
$$+ \mathbb{E}_{\hat{i} \sim p(i|z)}[\log p_\theta(\hat{i}|z)]] \tag{3}$$
$$\geq \mathbb{H}(i) + \mathbb{E}_{i \sim p(i)}[\mathbb{E}_{\hat{i} \sim p(i|z)}[\log p_\theta(\hat{i}|z)]]$$

where $\mathbb{H}(\cdot)$ is the entropy function and $\mathbb{E}$ is expectation. $p_\theta(\cdot)$ is a function parameterized by θ. This optimization is often referred to as variational information maximization [6]. Similarly,

$$I(z, a) \geq \mathbb{H}(a) + \mathbb{E}_{a \sim p(a)}[\mathbb{E}_{\hat{a} \sim p(a|z)}[\log p_\theta(\hat{a}|z)]] \tag{4}$$

The third and final conditional mutual information term $I(q, z|a, i)$ can also be bounded by:

$$I(z, q|a, i) \geq \mathbb{H}(q) +$$
$$\mathbb{E}_{q \sim p(q|i,a)}[\mathbb{E}_{\hat{q} \sim p(q|z,a,i)}[\log p_\theta(\hat{q}|z, i, a)]]]$$
$$\text{s. t. } p(q|z, a, i) = p(q|z)p(z|a, i) \tag{5}$$

Putting Eq. 3, 4 and 5 together in Eq. 2:

$$\max_\theta \mathbb{E}_{p_\theta(q,i,a)}[\log p_\theta(q|i, a, z) + \lambda_1 \log p_\theta(a|z)$$
$$+ \lambda_2 \log p_\theta(i|z)] \tag{6}$$
$$\text{s. t. } p_\theta(q, i, a) = p_\theta(q|z)p_\theta(z|i, a)p(i, a)$$

Note that we ignore the entropy terms associated with the training data as it doesn't involve the parameter θ we are trying to optimize. Therefore, optimizing Eq. 6 can be accomplished by maximizing the reconstruction of the image and answer representations while maximizing the MLE objective of generating the question.

3.4. Question generation by reconstructing image and answer representations

To functionalize the optimization presented above, we begin by first encoding the image using a CNN as a dense vector h_i (see Figure 3). Similarly, we encode the answer a using a long short term memory network (LSTM) [20], which is a variant of RNNs, into another dense vector h_a. Next, we feed h_i and h_a into a VAE that embeds both into a latent z-space. In practice, we assume that z follows a multivariate Gaussian distribution with diagonal covariance. We use the reparameterization trick [29], to generate means μ_z and standard deviations σ_z, combine it with a sampled unit Gaussian noise ϵ to generate $z = \mu_z + \sigma_z \epsilon$.

From z, we reconstruct $\hat{h}_i$ and $\hat{h}_a$ and optimize the first two terms in Eq. 6 by minimizing the following l_2 losses:

$$L_i = ||h_i - \hat{h}_i||_2, \quad L_a = ||h_a - \hat{h}_a||_2 \quad (7)$$

Next, we use a decoder LSTM to generate the question $\hat{q}$ from z-space. We minimize the MLE objective L_{MLE} between $\hat{q}$ and the true question in our training set q, which results in the third and final term in Eq. 6.

3.5. Regularizing with a second latent space

So far, we have proposed building a model that maximizes the lower bound of mutual information between a latent space, the image and the expected answer. This allows us to generate questions if we know what the expected answer should be. This is not conducive to our original goal of deploying our model in real world situations where it does not know the answer a priori. If we already know the answer to a question, there is no point in generating a question.

To remedy this, we propose a second latent t-space. Instead of using both a and i to encode h_a and h_i into z-space, we discard the answer and instead only use its category c. We classify answers as being one of a few predefined categories, such as objects (e.g. "cat"), attributes (e.g. "cold"), color (e.g. "brown"), relationship (e.g. "ride"), counting (e.g. "1"), etc. These categories are cast as a one hot vector and encoded as h_c and used, along with h_i to embed into the variational t-space. We train t-space by minimizing the KL-divergence with z-space:

$$
\begin{aligned}
L_t &= D_{KL}(p_\theta(z|i,a), p_\phi(t|i,c)) \\
&= \log \sigma_t - \log \sigma_z + \frac{\sigma_z + (\mu_t - \mu_z)^2}{2\sigma_t} - 0.5
\end{aligned} \quad (8)
$$

where ϕ are the parameters used to embed into t-space. This allows us to now utilize $p_\phi(t|i,c)$ to embed into a space that closely resembles z-space. Since we assume that both z-space and t-space follow a multivariate Gaussian with diagonal covariance, the KL term has the analytical form shown above. We no longer need to know the answer a to embed and generate questions. Intuitively, the t-space can be also thought of as a regularizer on z-space, preventing the model from overfitting to the answers in the training data and relying instead on utilizing the answer categories.

Putting them together, the final loss for our model is:

$$L = L_{MLE} + \lambda_1 L_a + \lambda_2 L_i + \lambda_3 L_t \quad (9)$$

where λ_1 and λ_2 have already been introduced and λ_3 is a hyperparameter that controls the amount of regularization used in our model. Note that we are omitting the KL-loss with respect to a unit normal centered at zero that maintains the two latent spaces' priors.

3.6. Inference

During inference, we are given an image i and answer category c and are expected to generate questions. We encode the inputs into the second latent t-space and sample from it to generate questions, as shown in Figure 4. This allows us to generate goal-driven questions for any image, focused towards extracting its objects, its attributes, etc.

3.7. Implementation details

We implement our model using PyTorch and plan on releasing all our code. We use ResNet18 [17] as our image encoder and do not fine-tune its weights. h_i, h_a and h_t are all 512 dimensional vectors. z-space and t-space are 100 dimensions. The encoders for the image and answer are trained only from L_{MLE} and not L_i, L_a or L_t to prevent the encoders from simply optimizing for the reconstruction loss at the cost of not being able to generate questions. We optimized the hyperparameters such that $\lambda_1 = 0.01$, $\lambda_2 = 0.001$, $\lambda_3 = 0.005$ with a learning rate of 0.001 that decays every 4 epochs for a total of 10 epochs.

4. Experiments

To test our visual question generation model, we perform a series of experiments and evaluate the model along multiple dimensions. We start by discussing the dataset and evaluation metrics used. We then showcase examples of our model's generated questions when conditioned on the answer. Next, we demonstrate its ability when conditioned only on the answer category. We compare both these cases against a series of baselines and ablations. We analyze the diversity of questions produced within each answer category. Finally, we report a small proof of concept deployment of our model on real world images found online and show that it can learn new concepts.

4.1. Experimental setup

Dataset. To enable the kind of interaction where we can specify input answer categories, we need a VQA dataset that categorizes its answers. The VQA dataset [2] has a few basic categorizations of questions but not their answers. We annotate the VQA [2] dataset answers with a set of 15 categories and label their top 500 answers. These categories include objects (e.g. "cat", "person"), attributes (e.g. "cold", "old"), color (e.g. "brown", "red"), relationship (e.g. "ride", "jump"), counting (e.g. "1", "10"), etc. The top 500 answers make up the 82% of the VQA dataset, resulting in $367K$ training+validation examples. We treat their validation set as our test set as the answers in their test set are not publicly available. We break the training set up into a 80-20% train-validation split.

Evaluation metrics. All past question generation papers have used a variety of evaluation metrics to calcu-

Table 1. We report our model's efficacy with multiple metrics. We use language modeling metrics to measure its capability to generate questions similar to the ground truth. Next, we measure the model's ability to maximize mutual information by predicting the answer or its category from the latent space embedding. Finally, we measure the relevance of the question with the image. Note that language modeling scores are multiplied by 100 to show more significant digits and mutual information and relevance scores are reported in percentages.

	Models	Language modeling						Mutual information		Relevance	
		Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	CIDEr	Answer	Category	Image	Category
z-space	IA2Q [57]	32.43	15.49	9.24	6.23	11.21	36.22	11.48	35.33	91.10	36.80
	V-IA2Q [23]	36.91	17.79	10.21	6.25	12.39	36.39	11.13	36.91	90.10	39.00
	Ours w/o A	38.88	20.74	12.75	6.29	12.78	40.13	10.02	40.44	98.10	42.70
	Ours w/o AC	38.99	21.48	12.73	6.57	13.01	42.13	10.10	60.00	96.80	42.80
	Ours w/o C	**50.09**	**32.32**	**24.61**	**16.27**	**20.58**	**94.33**	**33.44**	61.04	**98.00**	82.40
	Ours	48.09	29.76	20.71	15.17	18.78	92.13	30.23	**91.02**	97.10	**91.20**
t-space	IC2Q	30.42	13.55	6.23	4.44	9.42	27.42	9.88	40.23	90.00	38.80
	V-IC2Q	35.40	25.55	14.94	10.78	13.35	42.54	10.11	60.23	92.20	45.00
	Ours w/o A	31.20	16.20	11.18	6.24	12.11	35.89	9.35	68.23	**98.00**	52.50
	Ours	**47.40**	**28.95**	**19.93**	**14.49**	**18.35**	**85.99**	**28.23**	**99.02**	97.20	**98.00**

Figure 5. TSNE [39] visualization of the latent encodings. When we don't reconstruct the answer, the embedding show no separation between answers or their categories, confirming the posterior collapse. Meanwhile, by reconstructing the answer, both the z-space and the t-space encodings are visually separable. Different colors represent categories of answers and we only show 8 categories for aesthetics.

late the quality of a question. While some have focused on maximizing diversity [54, 23, 63], others have treated it as a proxy task to improve question answering [36, 47, 57]. Diversity measures have included using variants of beam search [54], measuring novel questions or unique tri-grams [23] or creating rule-based datasets [63]. Proxy tasks have typically used accuracy of multiple-choice answers to measure the performance of question generation.

We too report a variety of different evaluation metrics to highlight different components of our model. First, we use *language modeling* evaluation metrics like BLEU, ME-TEOR and CIDEr [7] to calculate how well our generated questions match the ground truth questions in our test set. Next, we measure the *mutual information* retained in the latent space by training a classifier to classify answer categories encoded in the latent space. This metric sheds light on how well our method retains information about the input answers or answer categories. Next, we measure *relevance* of the question, ensuring that the questions are valid for the given image and result in the expected answer category. Relevance results are calculated from majority vote conducted by hiring 3 crowd-workers that vote on whether a question can be answered given its corresponding image. Finally, we report diversity scores for each category, which measures the number of unique questions generated.

Baselines. We adapt a series of past CNN-RNN models to accept answer or answer types when generating questions. The first model IA2Q is a supervised, non-variational model that takes an image and answer as input and generates a question [57]. This model is reminiscent of the VQA models often used to answer questions [57], except the answers are now inputs and the questions outputs [2, 64]. Next, V-IA2Q is a variational version of IA2Q, which embeds the answer and question to a latent space before generating the question [23]. We also train versions of these models that accept the answer categories instead of the answer: IC2Q and V-IC2Q. When generating from a variational model, we set $z = \mu_z$ or $t = \mu_t$ to keep its outputs consistent for all measures except diversity.

We refer to our full model as Ours and can generate questions from either the answer latent space z or the category latent space t. We perform ablations on this model by removing specific components. Ours w/o A doesn't maximize mutual information with respect to the expected answer but can also generate questions from both the z and t spaces. Ours w/o C doesn't include the t- space and can only generate questions from answers. Finally, Ours w/o AC doesn't train with the reconstruction loss nor does it have a second latent space t. Our evaluations empirically demonstrate how these ablations justify our model designs.

	counting	binary	object	color	attribute	materials	spatial	shape	location	time	activity
	how many people are in the photo?	is this a busy street?	what is the man riding?	what color is the traffic light?	how talented is the person?	what is the road made out of?	what direction is the road sign in?	what shape is the fire hydrant?	what is the man riding?	what time of day is this?	what is the man doing?
	how many skateboards are there?	is the man wearing a hat?	what is the person holding?	what color is the boys shirt?	how tall is the person?	what is the man standing on?	is the sun to the left or right of the picture?	what is the shape of the pole?	is the man going to the right of the girl?	Is it day or night?	what are people watching?
	how many types of vegetables are there?	is the food in the foreground prickly?	what is in the bowl?	what color is the plate?	what is the name of the bowl?	what is the table made of?	is the fork in the foreground of the photo?	what shape is the plate?	where is this food?	what time of day is this?	what is the name of the bowl?
	how many types of fruits are there?	is the food healthy?	what is that food?	what color is the mat?	how does the food taste?	what is the container made of?	what is the name of the bowl?	what shape is the plate?	where is in the pan?	when was the food eaten?	what is the name of the bowl?
	how many people are in the photo?	is this a city scene?	what is the name of the bus?	what color is the bus?	how does the food taste?	what is the fence made of?	which side is the driver?	what shape is the building?	where is this bus probably?	what time of day is this?	what is the person doing?
	how many people are in the photo?	is this a busy street?	what is in front of the bus?	what color is the car?	what is the weather like in this picture?	what is the road made of?	which side is the person on?	what shape is the road?	where is this bus going?	when was this picture taken?	what is the man doing?
	how many horses are there?	are there animals in the photo?	what animal is this?	what color is the barn?	is this zoo clean or dirty?	what is the fence made of?	what kind of animal is this?	what shape is the fence?	what room is this?	what season is this?	what is the cow doing?
	how many cows are there?	is that a cow in the photo?	what are the animals behind?	what color is the horse?	how is the cow?	what is the fence made out of?	where is the cow lying?	what shape is the house?	what kind of animal is this?	what time is this?	what are the animals doing?
	how many people are in this photo?	is this person wearing a helmet?	what is the man standing on?	what color is the car?	what is the gender of the person?	what is the bench made of?	what direction is the person looking?	what is the gender of the person?	where is the boy?	is it night time?	what is the man doing?
	how many balls are in this photo?	is this person holding a bat?	what is the person standing on?	what color is the bat?	what is the weather like?	what is the bat made of?	which hand is the man holding?	what is the shape of the bat?	where is the game?	what time of day is this?	what is the person doing?

Figure 6. Example questions generated for a set of images and answer categories. Incorrect questions are shown in grey and occur when no relevant question van be generated for a given image and answer category.

4.2. Mutual information maximization

We check whether our model improves the mutual information retained in the latent space with the input answer. We freeze the weights of a trained model and embed input images, answers and categories into the latent z or t-space, depending on the model. We train a simple 3-layer MLP that attempts to classify the latent code as either one of the 15 answer categories or as one of the 500 answers. We evaluate our model on the test set with a random chance of 6.67% and 0.20%, respectively. Table 1 shows that the baseline models do a poor job of actually remembering the answer or category, justifying the need for a mutual information maximization approach. Since these models are unable to retain information about the input answers, it also explains why they often generate safe, generic, uninformative questions. Since our model can embed into both the z as well as the t space, we report how well these two spaces retain information. We find that Ours retains near perfect information about the input answer category with an accuracy of 99.02% from t-space and 32.44% from the z-space. We find that when trained without the t-space, Ours w/o C retains more information as it no longer has to constrain the z-space to regularize answers of the same category. We also visualize a TSNE [39] representation of the two spaces in Figure 5. Models that don't reconstruct the answer (e.g. in Ours w/o A, Ours w/o AC or any of the baselines) show visually inseparable categories.

4.3. Generating questions given the answers

Since our model can produce questions from both answers as well as answer categories, we evaluate both scenarios individually. The language modeling section in Table 1 showcases how the various models perform when generating questions from the z-space, i.e. generating questions from answers. We find that Ours w/o C performs the best over all the baselines and across all ablations of our model. This is likely because the latent space has more capacity when it is not also being regularized by the t-space. We find that Ours w/o A performs 6 METEOR points worse than Ours and Ours w/o C implying that forcing the model to reconstruct the answer does improve the quality of questions generated to better match the ground truth.

4.4. Generating questions with answer types

The lower half of Table 1 evaluates how well our model and the baselines perform when generating questions in the absence of the actual answer and only in the presense of the answer categories. We find that overall, all the language metrics are slightly lower than when the questions were generated from the z-space. This is expected as now the questions need to be generated with only the answer category encoded in the t-space without knowing exactly what the answer is. Therefore, the models are penalized for asking an "object" question about the "horse" when the answer expects the question to focus on the "saddle" instead. We

Table 2. Diversity measures across different answer categories. We report the generation strength, percent of unique questions generated normalized by number of unique ground truth questions and generation inventiveness, percent of unique questions generated unseen during training. All questions were generated from the t-space of our model for a fair comparison with V-IC2Q.

	V-IC2Q		Ours	
	Strength	Inventive	Strength	Inventive
counting	15.77	30.91	**26.06**	**41.30**
binary	18.15	41.95	**28.85**	**54.50**
object	11.27	34.84	**24.19**	**43.20**
color	4.03	13.03	**17.12**	**23.65**
attribute	37.76	41.09	**46.10**	**52.03**
materials	36.13	31.13	**45.75**	**40.72**
spatial	61.12	62.54	**70.17**	**68.18**
food	21.81	20.38	**33.37**	**31.19**
shape	35.51	44.03	**45.81**	**55.65**
location	34.68	18.11	**45.25**	**27.22**
people	22.58	17.38	**36.20**	**31.29**
time	25.58	15.51	**34.43**	**25.30**
activity	7.45	13.23	**21.32**	**26.53**
Overall	12.97	38.32	**26.06**	**52.11**

also qualitatively sample and report a random set of questions generated by our model in Figure 6. We see that our model often uses concepts in the image to ground the questions. It asks specific questions like "what is the bat made of?" or "is the man going to the right of the girl?". However, there are categories like "time" that have a low diversity of training questions and result in the inevitable "what time of day is this?" question. The qualitative errors we have observed often occur when the model is forced to ask a question about a category that is not present in the image; it is hard to ask about "food" when no food is present.

4.5. Measuring diversity of questions

For all the $177K$ images in our test set, we generated one question per answer category, resulting in a total of $2M$ questions. We report diversity in Table 2 using two existing metrics: (1) Strength of generation: the percentage of unique generated questions normalized by the number of unique ground truth questions and (2) Inventiveness of generation: the percentage of unique questions unseen during training normalized by all unique questions generated. We compare our model with the baseline V-IC2Q which does not reconstruct the answer or the image. We find that our method results in more diverse set of questions across most categories. Questions asking for "shape" and "materials" tend to generate the most unseen questions as the model learns to generate questions like "what [shape/material] is the ___ [made out of]?" and injects objects in the given image into the missing blank. Answers agnostic to the image contents, such as "time", result in the fewest number of novel questions.

Table 3. We categorized 1000 images into one of the answer categories, generated questions and asked crowd workers to answer them. We report the number of questions asked per category and the number of new concepts discovered by our model versus a baseline. We also show examples of new discovered concepts.

Category	Questions	V-IC2Q	Ours	Examples
object	411	10	80	blackthorns, robins
attributes	205	8	40	desecrated, crowned
colors	164	12	17	burgandy, Alabaster
materials	220	4	8	polyester, spandex

4.6. Real world deployment of our model

To examine our model in a real world deployment, we generated 1 question each for 1000 images with hashtags #food, #nature, #sports, #fashion scraped from online public social media posts. Since our model needs an input answer category to ask a question, we trained a simple ResNet18 CNN [17] on the VQA images to output one of 4 categories (see Table 3). We generated answer categories using the CNN and fed it into our model to generate the questions. The questions were sent to two crowd workers: one answered the question and the other reported the relevance of the question with the image and the answer with the answer category. We found all the questions asked by both Ours and V-IC2Q to be relevant to the image while 97.2% and 56.8% were relevant to the answer category. Our methods questions led to more unseen concepts.

5. Conclusion

We believe that visual question generation should be a task that is aimed at extracting specific categories of concepts from an image. We define a good question to be one that is not only relevant to the image but is also designed to expect a specific answer category. We build Information Maximizing Visual Question Generator that maximizes the mutual information between the generated question, the input image and the expected answer. We extend this model to overcome technical challenges associated with maximizing mutual information with discrete tokens and collapsing posterior while also allowing it to generate questions when the expected answer is absent. We analyze the questions using language modeling, diversity, relevance and mutual information metrics. We further show that through a real world deployment of this system, it can discover new concepts.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016. 2

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 5, 6

[3] E. Archer, I. M. Park, L. Buesing, J. Cunningham, and L. Paninski. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*, 2015. 3

[4] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015. 1, 2, 3

[5] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015. 3

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 4

[7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[8] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015. 1

[9] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015. 3

[10] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 1

[11] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 1

[12] X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017. 2

[13] 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. 2

[14] 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. 3

[15] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 3

[16] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016. 2, 3

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 8

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 2

[19] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[21] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017. 2

[22] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016. 1

[23] U. Jain, Z. Zhang, and A. G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, pages 5415–5424, 2017. 1, 2, 6

[24] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 3

[25] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017. 2

[26] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017, 2017. 2

[27] Ł. Kaiser, A. Roy, A. Vaswani, N. Pamar, S. Bengio, J. Uszkoreit, and N. Shazeer. Fast decoding in sequence models using discrete latent variables. *arXiv preprint arXiv:1803.03382*, 2018. 2, 3

[28] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014. 3

[29] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014. 3, 4

[30] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3337–3345. IEEE, 2017. 1

[31] 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. 2

[32] R. G. Krishnan, U. Shalit, and D. Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015. 3

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[34] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 3

[35] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015. 1

[36] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6116–6124, 2018. 6

[37] 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. 2

[38] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 1

[39] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 6, 7

[40] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 2, 3

[41] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 3

[42] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010. 1

[43] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. van der Maaten. Learning by asking questions. *arXiv preprint arXiv:1712.01238*, 2017. 2

[44] 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. 1, 2

[45] R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017. 2, 3

[46] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[47] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015. 1, 2, 6

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 3

[49] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 2

[50] I. V. Serban, A. García-Durán, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, and Y. Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016. 2

[51] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784, 2016. 1

[52] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[53] A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017. 3

[54] 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. 2, 6

[55] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008. 2

[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. 1, 2

[57] T. Wang, X. Yuan, and A. Trischler. A joint model for question answering and question generation. *arXiv preprint arXiv:1706.01450*, 2017. 2, 6

[58] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 2, 3

[59] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016. 2

[60] 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. 1

[61] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Visual curiosity: Learning to ask questions to learn visual recognition. *arXiv preprint arXiv:1810.00912*, 2018. 2

[62] 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. 2

[63] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang. Automatic generation of grounded visual questions. *arXiv preprint arXiv:1612.06530*, 2016. 6

[64] 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. 1, 2, 6