Latent Variable Models for Visual Question Answering

Zixu Wang, Yishu Miao, Lucia Specia
Department of Computing
Imperial College London
{zixu.wang, y.miao20, l.specia}@imperial.ac.uk

Abstract

Conventional models for Visual Question Answering (VQA) explore deterministic approaches with various types of image features, question features, and attention mechanisms. However, there exist other modalities that can be explored in addition to image and question pairs to bring extra information to the models. In this work, we propose latent variable models for VQA where extra information (*e.g.* captions and answer categories) are incorporated as latent variables to improve inference, which in turn benefits question-answering performance. Experiments on the VQA v2.0 benchmarking dataset demonstrate the effectiveness of our proposed models in that they improve over strong baselines, especially those that do not rely on extensive language-vision pre-training.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) has been proposed as the task where given an image and a textual question, systems are expected to provide a textual answer. VQA models (Bigham et al., 2010; Shih et al., 2016; Kazemi and Elqursh, 2017; Anderson et al., 2018; Yang et al., 2015) are trained to learn the relationship between areas in an image and the question, and to choose the correct answer from a vocabulary of answer candidates – as a classification task. Solving VQA would entail solving a range of challenges spanning various areas such as computer vision, natural language processing, knowledge representation, and reasoning. In most current approaches to VQA (Goyal et al., 2017; Kim et al., 2018; Yu et al., 2019; Jiang et al., 2020) models are created in a deterministic manner. They focus on exploring various image and question features. We posit that there are other types of information that could be useful to assist and improve the answering accuracy, such as image captions (Wu et al., 2019; Kim and Bansal, 2019) and mutated inputs (Gokhale et al., 2020; Chen et al., 2020; Abbasnejad et al., 2020).

In this paper, we propose to employ latent variables for VQA to exploit extra information (*i.e.* image captions and answer categories) to complement limited textual information from image and question pairs. We assume a realistic setting where this information – esp. captions – may only be available during the training phase. To that end, we introduce a continuous latent variable as the representation of captions, which can still capture the essential information from this modality. Moreover, we model the question category as a discrete latent variable, which acts as an inductive bias for better learning the prediction of answers, and can be integrated out during testing. Our generative framework is able to incorporate many other types of modalities or information as continuous or discrete latent variables, while maintaining a simple architecture for testing. This grants the models with stronger generalisation ability compared to its deterministic counterparts, which generally require explicit pipelines to generate the information from external modalities.

Intuitively, image captions describe diverse aspects of an image and include attributes and relations of objects in a more informative way. As captions are often present in VQA datasets, they can be used as complementary textual information. Current approaches incorporating image captions into VQA (Karpathy and Fei-Fei, 2015; Wu et al., 2019; Kim and Bansal, 2019) are mainly focused on deterministic structures with well-designed attention models, which requires to provide explicit captions generation pipelines for testing. In our work, a continuous latent variable is employed for capturing the caption distributions, which is a generative distribution conditioned on image and question pairs. During training, a variational distribution is constructed to condition on caption inputs, and optimised the generative distribution by Kullback-Leibler (KL) divergence. In this way, the joint multimodal representations (images and question) benefits from the caption modality in training phase, and it requires no explicit caption inputs at test time.

Similarly, a discrete latent variable is employed for modelling answer categories. In the question-answering scenario, there exists a strong connection between a question and answer pair when the question provides informative signals on its type. For example, "How many", "Where is" and "what is" normally connect to numbers, locations, and objects respectively. Therefore, using a category discrete latent variable can provide better inductive bias from the question and answer pairs. Although the category definitions cannot cover all types of answers, and false prediction during testing might be observed, the latent variable can still maintain the robustness in predicting correct answers by integrating out the latent variable (summing over all the probabilities of predicted categories).

In summary, our **main contributions** are:

- A generative VQA approach combining the modularity of probabilistic latent variables with the flexibility to introduce extra continuous and/or discrete information.

- A method to incorporate additional information which does not rely on building multiple deterministic pipelines, aiming at learning the underlying compositional, relational, and hierarchical structures of multiple modalities so that the models can benefit from the extra information during inference without providing explicit inputs during testing. With the help of (Rezende et al., 2014; Kingma and Welling, 2013; Miao et al., 2016), we can relatively easily carry out efficient inference for the latent variable models with various designs.

- Consistent improvements over state-of-the-art deterministic baseline models (*e.g.* VL-BERT (Su et al., 2020)) in experiments with the VQA v2.0 dataset. Our qualitative analysis also indicates that the latent variables capture interesting information from different modalities.

2 Model

We first present an overview of our general model structure, followed by multimodal encoders, and proposed latent variables.

2.1 General Model Structure

In a VQA task, the images normally take the leading position while the questions act as guidance, providing textual information that can be useful for answer predictions. We postulate that these joint representation can be improved by other multimodal information. Hence, we introduce captions and answer categories to our VQA model as continuous and discrete latent variables respectively to encourage a better learning in the joint distribution of image and question pairs during training. A notable advantage of the latent variable models is that they do not explicitly require captions or answer categories during testing, and therefore can be easily extended to condition on any other useful information.

Firstly we introduce the notations used in the general VQA model. V, Q, A are used to denote the input image, question, and answer instances respectively. The image feature v, question representation q, and answer representation a are extracted from the image encoder, question encoder, and answer encoder. The VQA task is constructed as a classification problem to output the most likely answer $\hat{a}$ from a fixed set of answers based on the content of the image v and question q.

$$\hat{a} = \text{argmax } p(a|v, q)$$

In our latent variable model, we introduce image captions C into the training phase. Similarly, we extract the caption features c by a caption encoder. However, instead of directly feeding in the caption features c into to the model, we employ a continuous latent distribution z to be the caption representations. Here $z \sim q(z|c)$ is modelled as variational distribution. We then build a generative distribution $z \sim p(z|v, q)$ to infer the caption information by conditioning on image and question pairs, which is optimised during training via neural variational inference.

Besides, we introduce a discrete latent variable d for modelling answer category inferred via $d \sim p(d|v, a)$, which is also conditioned on image and question pairs.

Hence, the training of the latent variable model is carried out by the samples (v, q, a, c, d); while during testing, the answer a is predicted by the image and question pair (v, q):

$$\hat{a} = \text{argmax} \sum_{d,z} p(a|v, q, d, z)p(d|v, q)p(z|v, q)$$

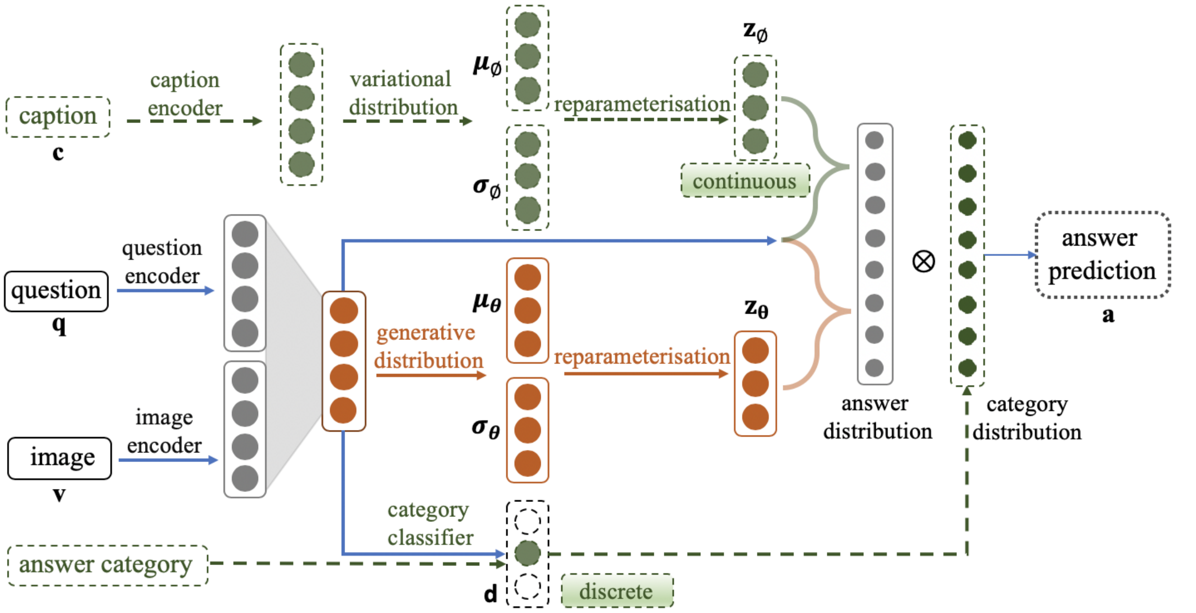


Figure 1: Architecture of our latent variable model for VQA. First the image and question (v, q) pairs are encoded through image encoder and question encoder, which are then fused into a multimodal representation. The captions (c) and answer categories (d) are introduced as continuous latent variable and discrete latent variable respectively. The model is first trained to approximate the variational distribution $(z_\phi \sim \mathcal{N}(\mu_\phi, \sigma_\phi))$ of captions (c) and the generative distribution $(z_\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta))$ of multimodal representations; then the Kullback-Leibler divergence (D_{KL}) between the two distributions is minimised to regularise the generative distribution. An answer category classifier is added to output the probabilistic distribution of categories and dot-product with the answer distribution for final answer predictions.

where the discrete latent variable d is directly integrated out, and the z is the Monte-Carlo sample from $p(z|v, q)$.

Image Encoder The input images are encoded as regional visual feature representations. These features are extracted from a bottom-up approach (Anderson et al., 2018), and pre-trained on the MSCOCO 2015 dataset (Lin et al., 2014). We represent each input image as extracted from the Faster R-CNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017). As in the standard VQA model using an attention mechanism, attention from the question features to the image features is used to output a weight for the feature vector at each spatial position in the feature map; which is first normalised and then used for performing a weighted sum over the spatial positions to produce a single feature vector to represent the image.

Question Encoder We pre-process and tokenise a question into a sequence of words, and each word in the question is further embedded into a 300-dimensional vector representation learnt along other parameters during training. The resulting dimension of word embeddings is of size (max_question_length × 300) and it is passed through a LSTM network (Hochreiter and Schmidhuber, 1997). The final question embeddings/features q are borrowed from the final state of the last LSTM unit.

Answer Encoder In VQA, the classifier usually takes fused (concatenated or element-wise multiplicated) multimodal representation of image and question pairs as input to generate a softmax distribution over output answer classes. We instead consider an energy-based representation to map the multimodal representation of image and question pairs and the answers to a shared space; where first all the answer candidates are represented as the answer embeddings and these embeddings are used to project the multimodal representation to the manifold of the answers.

2.2 Continuous Latent Variable: Caption

As captions are modelled by a continuous latent variable, we only have explicit captions during training. Here we present the generative distribution that infers caption modalities during testing, and the variational distribution that is conditioned on explicit captions during training. Therefore, the caption encoder is only used in the training phase.

Generative Distribution - $p_\theta(z|v,q)$. We use a latent distribution $p_\theta(z|v,q)$ to model the joint multimodal distributions of images and questions. Compared to its deterministic counterpart using concatenated multimodal features, we parameterise the stochastic distribution with $\mathcal{N}(z|\mu_\theta(v,q),\sigma^2_\theta(v,q))$. For each image and question pair (v,q), the model calculates the latent distribution with:

$$\begin{aligned} \pi_\theta &= u(\text{concat}(v,q)) \\ \mu_\theta &= \mathcal{L}_1(\pi_\theta) \\ \log\sigma^2_\theta &= \mathcal{L}_2(\pi_\theta) \\ p_\theta(z|v,q) &= \mathcal{N}(z|\mu_\theta(v,q),\sigma^2_\theta(v,q)) \end{aligned} \tag{1}$$

Variational Distribution - $q_\phi(z|c)$. We first apply a RNN model to embed the caption inputs C and a latent variable $q_\phi(z|c)$ to model the caption semantics and distributions, where $z \sim \mathcal{N}(z|\mu_\phi(c),\sigma^2_\phi(c))$. In our model, the variational distribution $q_\phi(z|c)$ is from:

$$\begin{aligned} \pi_\phi &= f(c) \\ \mu_\phi &= \mathcal{L}_3(\pi_\phi) \\ \log\sigma^2_\phi &= \mathcal{L}_4(\pi_\phi) \\ q_\phi(z|c) &= \mathcal{N}(z|\mu_\phi(c),\sigma^2_\phi(c)) \end{aligned} \tag{2}$$

where $\mathcal{L}_i$ are linearisation functions. For each caption c, the inference network calculated the variational parameters μ_ϕ and σ^2_ϕ to parameterise latent variational distribution of captions.

2.3 Discrete Latent Variable: Answer Category

Assume each image and question pair (v,q) can be projected to an answer category to look for a correct answer, we are able to encourage the algorithm to focus on the confusable candidates (*i.e.* within the same category) instead of only the spurious relationships between questions and answers via simple linguistic features. In other words, it is easy to infer the answer category, but difficult to predict the correct answer by actually making use of the grounding features from multimodal information. Therefore, in order to leverage this useful inductive bias, we propose a discrete latent variable to model the answer category given an image and question pair (v,q). In particular, for each answer category d, we have a conditional independent distribution $p(a|v,q,d)$ over the answers in the certain answer category.

$$p(a|v,q) = \sum_d p(a|v,q,d)\cdot p(d|v,q) \tag{3}$$

In our case, we make use of some answer categories provided by Krishna et al. (2019). Since the answer categories do not cover all the answers in the VQA v2.0 dataset, we simply group the non-categorised answers into one separate category. Therefore, without loss of generality, we can easily introduce a hierarchy by the discrete latent variable, which can be integrated out during test time. More details can be found in Section 4.1.

3 Inference

We apply neural variational inference for learning the proposed latent variable models, which apply two different networks for training and testing.

During training, the captions and answer categories are given, so the model has access to full information of v,q,c,d,a. The variational distribution $q_\phi(z|c)$ provides the samples $z_v \sim \mathcal{N}(z|\mu_\phi(c),\sigma^2_\phi(c))$ that are combined with the deterministic multimodal features (v,q) for predicting the correct answers. Similarly, the answer categories are given during the inference, so we do not require to integrate out d, but simply make use of the projection of $p(a|v,q,d,z)$. Concurrently the distribution over categories $p(d|v,q)$ is updated.

During testing, the model can only access the image and question pairs (v,q) and it takes the samples $z_g \sim \mathcal{N}(z|\mu_\theta(v,q),\sigma^2_\theta(v,q))$ from the generative distribution $p_\theta(z|v,q)$, which is then combined with the deterministic multimodal features (v,q) and then integrate out answer categories by $p(d|v,q)$. Therefore, the training and testing processes can be described as:

Training:

$$\begin{aligned} z_v &\sim \mathcal{N}(z|\mu_\phi(c),\sigma^2_\phi(c)) \\ h_v &= f(z_v,v,q) \\ \mathcal{L} &= \mathbb{E}[\log p_\theta(a|h_v,d)] - D_{\text{KL}}[q_\phi(z)||p_\theta(z)] \\ &\quad -d\cdot\log p(d|v,q) \end{aligned} \tag{4}$$

Testing:

$$\begin{aligned} z_g &\sim \mathcal{N}(z|\mu_\theta(v,q),\sigma^2_\theta(v,q)) \\ h_g &= f(z_g,v,q) \\ \hat{a} &= \text{argmax} \sum_d p(a|h_g,d)p(d|v,q) \end{aligned} \tag{5}$$

4 Datasets & Setup

4.1 Datasets

VQA v2.0 We use the VQA v2.0 dataset [1] (Antol et al., 2015) for our proposed latent variable model. The answers are balanced in order to minimise the effectiveness of dataset priors. This dataset contains over 1.1M questions with 11.1M answers from the over 200K images in the MSCOCO 2015 dataset (Lin et al., 2014). We split the dataset with the official partition, *i.e.* 443.8K questions from 82.8K images for training, and 214.4K questions from 40.5K images for validation. Additionally, there are two test subsets called test-dev and test-standard (test-std) to evaluate model performance online. The results consist of three per-type accuracies ("Yes/No", "Number", and "Other") and an overall accuracy. We report the results on validation set and test-standard set through the official evaluation server.

Image Captions The source of image captions in our work is the MSCOCO dataset [2] (Lin et al., 2014). We map the captions with the VQA v2.0 dataset by Image_IDs, and use the official configuration in which 82.4K images for training and 40.5K for validation in order to maintain consistency.

Answer Categories We use answer categories from the annotations of Krishna et al. (2019). The answers in the VQA v2.0 dataset are annotated with a set of 15 categories for the top 500 answers that makes up the 82% of the VQA v2.0 dataset; and the other answers are treated as an additional category. Especially, the annotated answer categories include objects (*e.g.* "mountain", "flower"), attributes (*e.g.* "cold", "old"), color (*e.g.* "white", "blue"), counting (*e.g.* "5", "2"), *etc.*

4.2 Baseline Models

Bottom-up Top-Down (UpDn) The Bottom-up Top-Down (UpDn) model (Anderson et al., 2018; Teney et al., 2018) uses visual features from the salient areas (region proposals) in an image (bottom-up) and gives them weights using attention mechanism (top-down) with features from question encoding. Especially, for each image, UpDn uses an image encoder [3] to output a set of object features. Similarly, for each question, a question encoder is used to output a set of word features. Consequently both features are fed into an attention module (where question representation is used as context to weight image object features) to predict answer distributions.

VL-BERT In order to compare with state-of-the-art methods, we further conduct the experiments using recently proposed Visual-Linguistic BERT (VL-BERT) framework (Su et al., 2020). It is implemented with the Transformer model (Vaswani et al., 2017) as the backbone and inspired from BERT (Devlin et al., 2019), by taking both visual and language features as input to better align the visual-linguistic clues and to better exploit the generic multi-modal representation. Notably, the input contains a sequence of words and region-of-interests (RoI) with certain special elements to disambiguate different input formats. VL-BERT is pre-trained mainly on two tasks – Masked Language Modelling and Masked RoI Classification – to improve the detailed alignment between visual and linguistic contents.

4.3 Training and Implementation Details

Our latent variable model is implemented based on and compared to both the UpDn baseline architecture and the pre-trained VL-BERT framework (see Appendix).

For our UpDn-based model, we use common feature extraction, pre-processing, and loss function. The image features are extracted from the first 36 region proposals with dimension of 2,048 for the VQA v2.0 dataset, which are fixed during training. For questions, we first pad and trim all the questions to a maximum length of 14 words, which are then tokenised and embedded to a 300 dimensional vector; we further extract question features using a two-layer bi-directional LSTM (state size of LSTM layer is set to 1,024.) with random initialisation. The size of question features is 1,024, which is then combined with the image features through a fusion layer to output a joint representation of size 1,024. The output is fed to a linear layer of size 3,000 followed by softmax to produce probabilities over the answer classes. We optimise this model with Adam optimizer (Kingma and Ba, 2014) for 50 epochs with a batch size of 256. The learning rate is initially set to $1.5e^{-3}$ and exponentially decayed, with $\beta_1 = 0.900$ and $\beta_2 = 0.999$.

[1] https://visualqa.org/

[2] https://cocodataset.org/

[3] Faster R-CNN or other region proposal networks, used as an object detection model to identify and to localise objects belonging to certain classes.

	VQA v2.0 test-dev (%)				test-std (%)
	All	Yes/No	Num	Other	All
Caption (Wu et al., 2019)	-	-	-	-	68.37
DFAF (Peng et al., 2018)	70.22	86.09	53.32	60.49	70.34
MLIN (Gao et al., 2019)	71.09	87.07	53.39	60.49	71.27
UpDn (Anderson et al., 2018)	65.32	81.82	44.21	56.05	65.67
UpDn + latent (ours)	66.01	82.96	44.58	55.94	66.29
VL-BERT$_{\text{large}}$ (Su et al., 2020)	71.79	-	-	-	72.22
VL-BERT$_{\text{large}}$ + latent (ours)	**72.03**	88.03	54.16	62.42	**72.37**

Table 1: Experimental results on VQA v2.0 test-dev and test-standard (test-std) set. Accuracies are reported in percentage (%) terms. The state-of-the-art scores are in bold; underlined scores are best among (baseline *vs.* latent variable extension); and both underlines and bold scores are the overall best results.

5 Experiments

In this section, we first describe the experimental results of our latent variable model, compared with both a UpDn (Bottom-up Top-down) baseline model and a state-of-the-art pre-trained visual-linguistic model (VL-BERT); then we conduct qualitative analysis to validate the effectiveness of proposed components.

5.1 Quantitative Analysis

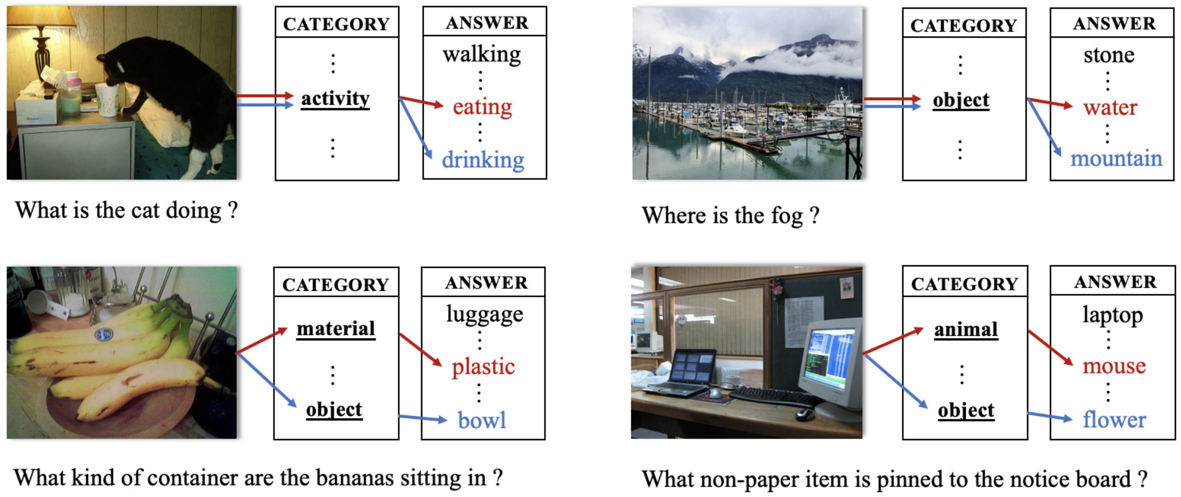
The experimental results are reported in Table 1. We compare the results of our latent variable model with the baseline model (UpDn), a state-of-the-art visual-linguistic pre-training model (VL-BERT), and three other related VQA models; where Wu et al. (2019) uses generated captions to assist answer predictions and Peng et al. (2018); Gao et al. (2019) explore the interactions between visual and linguistic inputs.

As demonstrated in Table 1, our latent variable model outperforms when acting as an extension. In particular, our latent variable model outperforms UpDn by 0.69% accuracy on test-dev set and by 0.62% accuracy on test-standard set. In addition, our model improves the performance by 0.24% accuracy than its VL-BERT counterpart on test-dev set and by 0.15% accuracy on test-standard set. These results indicate the effectiveness of including captions and answer categories as latent variables, to promote the distribution of image and caption pairs to be closer to the captions' space, and to learn a better distinction among different kinds of answers, or different answers within the same answer category.

The result of Wu et al. (2019) (68.37%) is a very strong baseline, which follows a traditional deterministic approach. However, their model is trained to generate captions that can be used at test time, while in our case only image and caption pairs are required for answer prediction. Peng et al. (2018) and Gao et al. (2019) both achieves comparable performance (70.34% and 71.27% on test-standard set, respectively) to VL-BERT (72.22%) without pre-training, by dynamically modulating the intra-modality information and exploring the latent interaction between modalities. Our latent variable model has the overall best result when combined with the strong pre-training VL-BERT, which indicates both the effectiveness of the visual-linguistic pre-training framework, and the incorporation of continuous (captions) and discrete (answer categories) latent variables.

Compared to the results on the standard baseline (UpDn), the improvements achieved by our proposed model on the VL-BERT framework is smaller. This is because VL-BERT has been pre-trained on massive image captioning data, where the learning of visual features have largely benefited from the modality of captions already. Nevertheless, based upon the strong baseline model, our proposed model can still expedite the performance slightly, which further indicates the effectiveness of the latent variable framework.

The state-of-the-art performance on VQA v2.0 among pre-training frameworks is achieved by Learning Cross-Modality Encoder Representations from Transformers (LXMERT) (Tan and Bansal, 2019). It has been extensively pre-trained using massive datasets on five languages and vision tasks in a multi-task learning fashion, which hence achieves 72.40 on test-dev and 72.54 on test-std. Our work is not directly comparable in this case,



What is the cat doing ?

Where is the fog ?

What kind of container are the bananas sitting in ?

What non-paper item is pinned to the notice board ?

Figure 2: Examples of our latent variable model outperforming the baseline UpDn model from the introduction of answer category as a discrete latent variable. The answer predicted by UpDn is highlighted in red and the answer from our model is in blue. We also show other sample answer candidates within each category.

and was not originally aimed at beating the state-of-the-art. Instead, it is focused on exploring the potentials of latent variable models in multimodal learning, which can be demonstrated by the improvements brought by the latent variables.

5.2 Qualitative Analysis

We perform ablation study to qualitatively analyse the effect of the components introduced in our work brought by the continuous (image caption) and discrete (answer category) latent variables, as shown in Table 2.

	VQA v2.0 val			
	All	Yes/No	Num	Other
UpDn	63.15	80.38	42.84	55.86
UpDn + caption	63.85	81.10	43.63	55.90
UpDn + category	63.51	81.62	42.17	55.38
Ours w/o caption	**64.09**	81.82	44.37	55.74
Ours w/ caption	<u>64.24</u>	82.36	44.52	56.02

Table 2: Ablation study to investigate the effect of each component: caption, and answer category. "Ours w/o caption" indicates our final model in which only image and question pairs are needed at test time; while "Ours w/ caption" represents the model using caption during evaluation. The result of our best model are in bold; while the best performance with captions as inputs during testing is underlined for comparison.

5.2.1 Effect of Captions

Introduction of captions as a continuous latent variable improves the classification performance, with an additional modality as inputs to benefit the learning of multimodal representations. According to the breakdown numbers in 2, the improvements brought by the latent variables of captions and answer categories are 0.70 and 0.36 respectively. The combined strategy reaches 0.94 which indicates that the benefits from the two latent variables are almost orthogonal. Note that neither the captions nor the answer categories is available during testing, and we only make use of the modalities in training.

However, to further investigate the latent variable of captions, we design an experiment that feed in ground truth captions via variational distribution for caption representations instead of inferring from question and answer pairs (*i.e.* use $q_{\phi}(z|c)$ to replace $p_{\theta}(z|v,q)$). We test this out in the validation dataset and obtain 64.24 ('Ours w/ caption') compared to 64.09 ('Ours w/o caption'). It shows that having explicit captions as input gives slightly better performance. However, the captions in these experiments are ground truth instead of the generated sentences from an image captioning pipeline, which means that the numbers might drop due to the possible errors in language generation during testing. Primarily, our proposed model ('Ours w/o caption') achieves the performance on par with it, which demonstrates the effectiveness of the strategy that incorporates extra modality by latent variable and it is sufficient enough to leverage the caption information during training without feeding any explicit captions in testing.

5.2.2 Effect of Answer Category

It can be observed from Table 2 that after introducing answer category as an additional discrete latent variable, our proposed model can also be improved over the UpDn baseline, of which the largest improvement can be observed from the "Yes/No" type.

In order to further elaborate the effectiveness of answer categories, we extract examples where our model predicted the correct answers while the UpDn baseline failed to, as shown in Figure 2. In the top two cases, both models predict answers under the same and correct answer categories, hence small distance in the answer space; however, our latent variable model can effectively distinguish and learn the difference among the answers which fall in the same category. In the bottom of Figure 2, we show two cases where the distance between the answers of two models are are relatively larger as they belong to different answer categories. Our model not only outputs the highest probability for the correct answer category, but also makes the correct final prediction. These examples indicate that the latent variable for answer categories effectively helps the model to have better predictions on answers, as the inductive bias encourages the model to devote more attention to distinguish the answers within the same category.

6 Related Work

VQA with Captions Recently, there are some work proposed to take use of image captions (Karpathy and Fei-Fei, 2015; Wu et al., 2019; Kim and Bansal, 2019) for exploiting more textual features from images that could be relevant to answer predictions. Captions are generally used in the aim of helping to answer a particular visual question; where captions contain meaningful information of the image in the VQA domain by relating the modalities, *i.e.* textual questions and textual answers. Wu et al. (2019) exploits the connection between VQA and captions by jointly generating question-relevant captions that are targeted to help in answering a specific visual question. The empirical results show that, in the cases of, using automatically generated captions would enhance the robustness and accuracy of VQA tasks. While in our work, we only introduce captions at training time and we discard the caption information at test time to promote the caption information to be learnt into (image, question) multimodal representation.

Bayesian VQA A potential effective model used in VQA is a Bayesian approach (Singh et al.; Vedantam et al., 2019) to take image and question into their latent spaces, in the aim of maximising the mutual information between the image, the question, and consequently the answer. This approach would expedite efficient reasoning as well as disentangle this reasoning from perception. In general, Bayesian approach considerably outperforms the quantitative metrics in state-of-the-art benchmarks. There has been some work on exploring Bayesian and latent variable methods for Visual Question Generation (Patro et al., 2020; Krishna et al., 2019). However, in our work, we frame VQA under the variational inference framework where we approximate both the variational and generative distribution during training.

VQA with Pre-training Pre-training has been emerged recently in applying to VQA (Zhou et al., 2019; Li et al., 2019; Tan and Bansal, 2019; Lu et al., 2019). In order to learn the vision-language connections, *i.e.* to understand the visual concepts while relating the modalities between visuals and languages, one would want to generalise on the bases of both intra-modality and cross-modality relationships. One particular model which has significant contribution in this would be LXMERT. On the other hand, in order to accomplish generic VQA tasks, such as visual reasoning in a large-scale caption included dataset, one would consider a pre-training via Transformer layers, where taking visual and language embedded features as input (with or without any explicit supervisions), which associated input with attention, in order to obtain mutual or relevant information from the given image, the question, the caption (including text-only corpus), and the expected answer.

7 Conclusion

In this paper, we propose to tackle VQA under the framework of latent variables models, to employ captions and answer categories as the continuous and the discrete latent variables respectively. Our experimental results and qualitative analysis show the effectiveness of the latent variables in boosting answering performance. This framework could be easily generalised to incorporate other types of information or modalities to enhance VQA tasks.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. 2019. Multi-modality latent interaction network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

Hyounghun Kim and Mohit Bansal. 2019. Improving visual question answering by referring to generated paragraph captions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3606–3612, Florence, Italy. Association for Computational Linguistics.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1564–1574. Curran Associates, Inc.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13–23. Curran Associates, Inc.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

Badri Patro, Vinod Kurmi, Sandeep Kumar, and Vinay Namboodiri. 2020. Deep bayesian network for visual question generation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1566–1576.

Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. 2018. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. *arXiv preprint arXiv:1812.05252*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Kevin J. Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Computer Vision and Pattern Recognition*.

Gursimran Singh, Saeid Naderiparizi, and Setareh Cohan. A bayesian approach to visual question answering.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2019. Probabilistic neural-symbolic models for interpretable visual question answering. In *ICML*.

Jialin Wu, Zeyuan Hu, and Raymond Mooney. 2019. Generating question relevant captions to aid visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3585–3594, Florence, Italy. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.

A Implementation and Training Details of VL-BERT based Latent Variable Model

Modifications of Model Architecture The construction of our latent variable model based on the pre-training VL-BERT is slightly different (mainly on the encoder) from that on the baseline model, in order to have consistency within the pre-training architecture. Especially, only one Transformer encoder (Vaswani et al., 2017) is used to take the concatenated sequence of words and region-of-interests (RoI) as inputs, with a special token [MASK] in the middle to represent the predicted answer. The outputs of the Transformer encoder corresponding to the [MASK] act as the multimodal representation for answer predictions. Besides, we apply the same Transformer encoder to the captions (to replace questions), in which a well-trained attention can be learnt between the captions and region-of-interests (RoI). The outputs corresponding to the position of captions are taken as the caption features.

Training Details We train our latent variable model based on a well pre-trained visual-linguistic checkpoint and the pre-training details can be referred in Su et al. (2020). The image features are extracted similar to that of UpDn, where a Faster R-CNN detector is applied. The linguistic sequence starts with a special classification element ([CLS]), which is then embedded with WordPiece embeddings (Wu et al., 2016) with $30,000$ vocabularies. A special token is assigned to each particular element, and for the visual elements, a special [IMG] token is assigned for each one of them. A special separation ([SEP]) is inserted to separate different sentences, and the visual and linguistic elements. We fine-tune our latent variable model for 20 epochs. We applied Adam optimiser (Kingma and Ba, 2014), with base learning rate $1e^{-4}$, $\beta_1 = 0.900$, and $\beta_2 = 0.999$; weight decay of $1e^{-4}$. The learning rate is warmed up over the first 2,000 steps, and linearly decay afterwards.