Latent Variable Models for Visual Question Answering

Zixu Wang[1], Yishu Miao[1], Lucia Specia[1,2]

[1]Department of Computing, Imperial College London, UK
[2]Department of Computer Science, University of Sheffield, UK

{zixu.wang, y.miao20, l.specia}@imperial.ac.uk

Abstract

Current work on Visual Question Answering (VQA) explore deterministic approaches conditioned on various types of image and question features. We posit that, in addition to image and question pairs, other modalities are useful for teaching machine to carry out question answering. Hence in this paper, we propose latent variable models for VQA where extra information (e.g. captions and answer categories) are incorporated as latent variables, which are observed during training but in turn benefit question-answering performance at test time. Experiments on the VQA v2.0 benchmarking dataset demonstrate the effectiveness of our proposed models: they improve over strong baselines, especially those that do not rely on extensive language-vision pre-training.

1. Introduction

As a classic multi-modal machine learning problem, Visual Question Answering (VQA) [2] systems are tasked with providing a correct textual answer given an image and a textual question. Current VQA models [14, 8, 1] 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, *i.e.*, they are modelled as a classification problem. The majorities of popular VQA [7, 9, 18] models are created in a deterministic manner and explore solely information from the given image-question pair. There are other approaches attempting to incorporate extra information, such as image captions [17] and mutated inputs [6, 3]. However, it in turn restricts the practical applications as the extra information is required to be explicitly available during testing.

In this paper, we propose an approach to explore additional information as latent variables in VQA: we 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. cap-

tions – may only be available during the training phase. To that end, we introduce a continuous latent variable as the caption representation to capture the essential information from this modality. Moreover, the answer category is modelled as a discrete latent variable, which acts as an inductive bias to benefit the learning of answer prediction, and can be integrated out during testing. The motivation is that the generative framework is able to incorporate many other types of information as continuous or discrete latent variables, and as such it effectively leverages additional resources to constrain the original image-question distribution while omitting them in testing. This grants the models with stronger generalisation ability compared to its deterministic counterparts, which generally require off-the-shelf pipelines to model 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. In our work, a continuous latent variable is employed for capturing the caption distributions and constraining the generative distribution conditioned on image and question pairs. In this way, the joint multimodal representations from images and question can benefit from the caption modality during training, and it requires no explicit caption inputs in testing. Similarly, there exists a strong connection between a question and answer pair when the question provides informative signals on its type or the category of possible answers. For example, "How many", "Where is" and "what is" normally connect to numbers, locations, and objects respectively. We propose a discrete latent variable is employed for modelling answer categories and providing better inductive bias from the question and answer pairs.

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

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

- A method to incorporate additional information which does not rely on building multiple deterministic
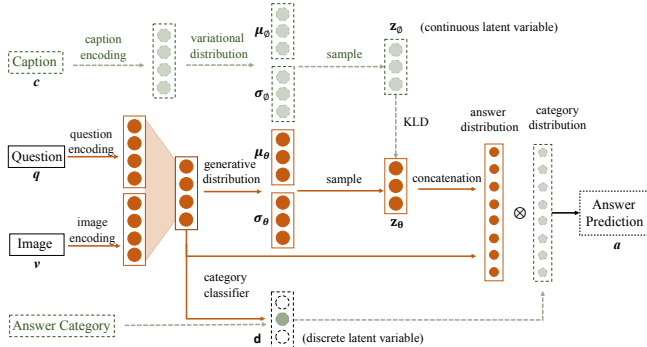
Figure 1. Architecture of our latent variable model for VQA. We use dotted lines to denote the process of proposed latent variables.

pipelines, aiming at learning the underlying compositional, relational, and hierarchical structures of multiple modalities. The models benefit from the extra information during training without providing explicit inputs in testing.

- The improvements over deterministic baseline models (*e.g.* UpDn [1] and VL-BERT [15]) in experiments with the VQA v2.0 dataset demonstrate the effectiveness of our proposed latent variable models. Our qualitative analysis also indicates that using extra resources (*i.e.* captions and answer categories) as latent variables captures complementary information during training and benefits the VQA performance in testing.

2. Model

We first present an overview of our general model structure, followed by the encoders for different modalities, and the proposed corresponding latent variables.

2.1. General Model Structure

In a VQA task, images and questions are normally used to learn a joint multimodal distribution for answer predictions. We postulate that the 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:

$$\hat{a} = \operatorname{argmax} p(a|v, q) \tag{1}$$

In our latent variable model, we introduce image captions C to 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. We originally experimented using $q(z|v, q, c)$ as the vairational distribution. However, this distribution is quite close (*i.e.* small KL divergence) to the generative distribution $p(z|v, q)$, which weakens the learning signal from KL divergence.

In addition, we introduce a discrete latent variable d for modelling answer category inferred via $d \sim p(d|v, q)$, 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). During testing, the answer a is predicted from the image and question pair (v, q):

$$\hat{a} = \operatorname{argmax} \sum_{d,z} p(a|v, q, d, z) p(d|v, q) p(z|v, q) \tag{2}$$

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

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 is conditioned on images and questions 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 $\mathcal{N}(z|\mu_\theta(v, q), \sigma_\theta^2(v, q))$.

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 $z \sim \mathcal{N}(z|\mu_\phi(c), \sigma_\phi^2(c))$.

2.3. Discrete Latent Variable: Answer Category

Assuming that each image and question pair (v, q) can be projected to an answer category to help find a correct answer, we are able to encourage the model to distinguishing candidates across answer categories instead of only the spurious relationships between questions and answers via simple linguistic features. 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.

$$p(a|v, q) = \sum_d p(a|v, q, d) \cdot p(d|v, q) \qquad (3)$$

We trained an answer category classifier using joint image-question pairs as the input, given the true labels as shown at left bottom in Figure 1. We then use the category distribution to modify the answer distribution through element-wise production to get more precise answer distribution.

3. Datasets & Setup

3.1. Datasets

We use the VQA v2.0 dataset [2] for our proposed latent variable model. The answers are balanced in order to minimise the effectiveness of dataset priors. We report the results on validation set and test-standard set through the official evaluation server. The source of image captions in our work is the MSCOCO dataset [12].

We use answer categories from the annotations of [10]. 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% [1] of the VQA v2.0 dataset; and the other answers are treated as an additional category.

4. 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.

4.1. Quantitative Analysis

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

---

[1] 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 summing over all the probabilities of predicted categories.

	VQA v2.0 test-dev (%)				test-std (%)
	All	Yes/No	Num	Other	All
Caption [17]	-	-	-	-	68.37
DFAF [13]	70.22	86.09	53.32	60.49	70.34
MLIN [5]	71.09	87.07	53.39	60.49	71.27
UpDn [1]	65.32	81.82	44.21	56.05	65.67
+ latent (ours)	<u>66.01</u>	82.96	44.58	55.94	<u>66.29</u>
VL-BERT$_{large}$ [15]	71.79	-	-	-	72.22
+ 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.

related VQA models; where [17] uses generated captions to assist answer predictions and [13, 5] 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 [17] (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 question pairs are required for answer prediction. [13] and [5] both achieve 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

	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.
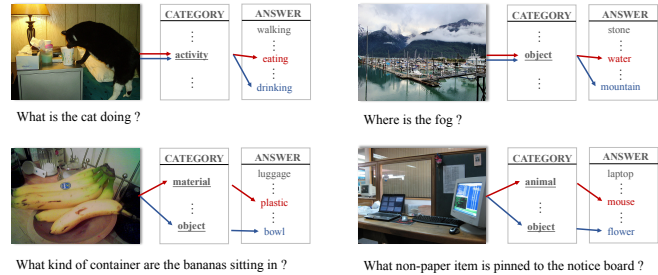

Figure 2. Examples of our latent variable model outperforming the baseline UpDn model from the introduction of answer category as a discrete latent variable. The answer predicted by UpDn is highlighted in red and the answer from our model is in blue. We also show other sample answer candidates within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

can still improve 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 LXMERT, Oscar and Uniter [16, 11, 4]. They have been extensively pre-trained using massive datasets on languages and vision tasks (including VQA) in a multi-task learning fashion. Our work is not directly comparable, and is not aimed at improving and beating the state-of-the-art performance. Instead, it is focused on exploring the potential of latent variable models to represent additional useful information in multimodal learning and to contribute to pre-trained vision-language frameworks. We draw attention to the advantage of using generative framework on the VQA task. In this case, we can employ more information during training (which is omitted in testing) to regularise the original multimodal distribution. This can be demonstrated by the improvements on VL-BERT brought by the latent variables.

4.2. Qualitative Analysis

We perform an 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.

4.2.1 Effect of Captions

The introduction of captions as a continuous latent variable improves the classification performance, with an additional modality as input 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 for All questions altogether. The combined strategy reaches 0.94 which indicates that the benefits from the two latent variables are complementary. Note that neither the captions nor the answer categories is available during testing; we only make use of these modalities during training.

To further investigate potential benefit of the captions, we design an experiment that feed in ground truth captions via variational distribution for caption representations instead of inferring them 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, which means that if we were to use instead automatically generated captions from an image captioning pipeline, the numbers might drop due to the possible captioning errors. Primarily, our proposed model ('Ours w/o caption') achieves the performance on par with with the model with ground truth captions, which demonstrates the effectiveness of the strategy that incorporates extra modality by latent variables.

4.2.2 Effect of Answer Category

As it can be observed from Table 2, after introducing answer category as an additional discrete latent variable, our proposed model can also be improved over the UpDn baseline, where the largest improvement can be observed for the "Yes/No" type. For the question types "Num" and "Other", the results of UpDn+category are lower than the baseline. This may be due to the multiple answer candidates under the two categories. For example, although the category classier can accurately predict the answer category (*e.g.*, "count", "color", *etc.*), it can be still difficult to distinguish among the answers - {"9", "20", ..., "many" for "count"; "black", "brown", ..., "black and white" for "color"}. We highlight that the contribution of answer categories as a discrete latent variable is to introduce an inductive bias which helps predict the correct answer categories and answers given a specific image and question.

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 do so, as shown in Figure 2. For the two cases in the top row, both models predict answers under the same and correct answer categories, hence the answer space is similar; however, our latent variable model can effectively distinguish and learn the difference among the answers which fall within the same category. The bottom row of Figure 2 shows two cases where the two models predict answers in different answer categories, and therefore they are also very different in meaning. Our model not only outputs the highest probability for the correct answer category, but also makes the correct final prediction.

5. Conclusions

In this paper, we propose to tackle VQA under the framework of latent variable models, employing captions and answer categories as the continuous and the discrete latent variables respectively to constrain the original image-question distribution while omitting the extra information during the test phase. Our experimental results and qualitative analysis show the effectiveness of the latent variables in boosting answering performance at test time when only image and question pairs are available. This framework could be easily generalised to incorporate other types of information or modalities to enhance VQA and other tasks.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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, 2018. 1, 2, 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 3

[3] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 4

[5] Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. Multi-modality latent interaction network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[6] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*, 2020. 1

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1

[8] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 1

[9] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 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., 2018. 1

[10] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[11] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 4

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[13] Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. *arXiv preprint arXiv:1812.05252*, 2018. 3

[14] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Computer Vision and Pattern Recognition*, 2016. 1

[15] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2, 3

[16] Hao Tan and Mohit Bansal. 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, Nov. 2019. Association for Computational Linguistics. 4

[17] Jialin Wu, Zeyuan Hu, and Raymond Mooney. 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, July 2019. Association for Computational Linguistics. 1, 3

[18] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 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, 2019. 1