C3VQG: Category Consistent Cyclic Visual Question Generation Supplementary Material

Shagun Uppal^{1*}, Anish Madan^{1*}, Sarthak Bhagat^{1*}, Yi Yu², Rajiv Ratn Shah¹

¹IIT-Delhi, India; ²NII, Japan

{shagun16088, anish16223, sarthak16189, rajivrtn}@iitd.ac.in, yiyu@nii.ac.jp

In this document, we begin by showing the final formulation of the loss functions to train our network, along with the detailed algorithm. We then derive the variational lower bound of the objective function of step 1 of our approach. We also provide an illustration to depict the inference procedure of C3VQG. This is followed by listing the hyperparameters values and the details of VQA dataset used for training and evaluation of the C3VQG model. Lastly, we show some more qualitative results as well as the complete results of diversity-based performance of our model which showcases the strength of using our proposed hyper-prior on the latent space.

1 IMPLEMENTATION DETAILS

Using the individual losses defined earlier, we aim to optimize a combined loss $\mathcal{L}_{total}$ that is the weighted sum of individual loss terms. Combining the individual losses defined earlier, we obtain the optimization objective as follows:

$$\min_{\mathbf{W}} \mathcal{L}_{total} = \min_{\mathbf{W}} \left[\mathcal{L}_Q + \lambda_I \mathcal{L}_I + \lambda_C \mathcal{L}_C + \lambda_{cons} \mathcal{L}_{cons} + \lambda_{center} \mathcal{L}_{center} + \lambda_{bayes} \mathcal{L}_{bayes} \right] \quad (1)$$

where, $\mathbf{W}$ represents the combination of all learnable parameters in the complete model and λ s are the hyperparameters depicting the weight of each loss in the combined objective.

For training our model using Algorithm 1, we use gradient descent algorithm with Adam optimizer. We train the model for 15 epochs on a machine with single GeForce GTX 1080 GPU using the PyTorch framework.

2 DERIVATION

We re-iterate the objective function optimized in the main paper. The objective function is parameterized by ϕ which is optimized as follows:

$$\max_{\phi} I(q, z|i, C) + \lambda_1 I(i, z) + \lambda_2 I(C, z) \quad (2)$$

$$s.t \quad z \sim p_{\phi}(z|i, C) \quad (3)$$

$$q \sim p_{\phi}(q|z) \quad (4)$$

where λ_1 and λ_2 are the weights for the mutual information terms. The mutual information in Equation 2 is intractable as we do not know true values of the posteriors $p(z|i)$ and $p(z|C)$. So we instead try to minimize its variational lower bound (ELBO). In the equations below, $\mathbb{H}$ stands for entropy while $\mathbb{E}$ for expectation.

Algorithm 1: Training Algorithm for C3VQG with all components.

```

1 Input:  $dset$  containing  $n$  training tuples of form  $\langle i, q, C \rangle$ ,
   multi-task loss weights for all individual losses:  $\lambda$ s,
   gradient descent learning rate  $\alpha_{LR}$ .
2 Output: Optimal weights for all the individual components
   of the model  $\mathbf{W}$ .
3 Initialize  $\mathbf{W}$  with Kaiming initialization [2].
4 for  $epoch \leftarrow 1$  to  $num\_epochs$  do
5   for  $k \leftarrow 1$  to  $n$  do
6      $i_k, q_k, C_k \leftarrow dset[k]$ 
7     Get  $h_k^i$  and  $h_k^c$  using Equation 3.
8     Concatenate  $h_k^i$  and  $h_k^c$  to get  $z_k$  using Equation 4.
9     Use  $z_k$  to predict  $h_k^i$  and  $h_k^c$  using networks  $p^i$  and
        $p^c$  and compute  $\mathcal{L}_I$  and  $\mathcal{L}_C$  using Equation 7.
10    Generate question  $\hat{q}_{k, C_k}$  using Equation 5 and
       compute  $\mathcal{L}_Q$  using Equation 6.
11    Predict category  $C_k^{pred}$  from generated question
       using Equation 8 and compute  $\mathcal{L}_{cons}$  using
       Equation 9.
12    Compute  $\mathcal{L}_{center}$  and  $\mathcal{L}_{bayes}$  using Equation 10
       and 12 respectively.
13    Find gradient of the total loss w.r.t.  $\mathbf{W}$ , i.e.  $\nabla_{\mathbf{W}} \mathcal{L}_{total}$ .
14    Take gradient descent step,
        $\mathbf{W} \leftarrow \mathbf{W} - \alpha_{LR} \nabla_{\mathbf{W}} \mathcal{L}_{total}$ .
15   end
16 end

```

$$\begin{aligned}
I(C, z) &= \mathbb{H}(C) - \mathbb{H}(C|z) \\
&= \mathbb{H}(C) + \mathbb{E}_{z \sim p(z, C)} [\mathbb{E}_{\hat{C} \sim p(C|z)} [\log p(\hat{C}|z)]] \\
&= \mathbb{H}(C) + \mathbb{E}_{C \sim p(C)} [D_{KL}[p(\hat{C}|z) || p_{\phi}(\hat{C}|z)]] \\
&\quad + \mathbb{E}_{\hat{C} \sim p(C|z)} [\log p(\hat{C}|z)] \\
&\geq \mathbb{H}(C) + \mathbb{E}_{C \sim p(C)} [\mathbb{E}_{\hat{C} \sim p(C|z)} [\log p_{\phi}(\hat{C}|z)]]
\end{aligned} \quad (5)$$

We similarly compute the expression for $I(i, z)$:

$$I(i, z) \geq \mathbb{H}(i) + \mathbb{E}_{i \sim p(i)} [\mathbb{E}_{\hat{i} \sim p(i|z)} [\log p_{\phi}(\hat{i}|z)]] \quad (6)$$

The expression for $I(q, z|i, C)$ then follows as:

$$I(q, z|i, C) \geq \mathbb{H}(q) + \mathbb{E}_{q \sim p(q|i, C)} [\mathbb{E}_{\hat{q} \sim p(q|z, C, i)} [\log p_{\phi}(\hat{q}|z, i, C)]] \quad (7)$$

$$\text{where } p(q|z, i, C) = p(q|z)p(z|i, C) \quad (8)$$

*Equal contribution. Ordered Randomly.

We substitute equations 5, 6, and 7 in equation 2:

$$\max_{\phi} \mathbb{E}_{p_{\phi}(q,i,C)} [\log p_{\phi}(q|z,i,C) + \lambda_1 \log p_{\phi}(i|z) + \lambda_2 \log p_{\phi}(C|z)] \quad (9)$$

where $p_{\phi}(q,i,C) = p_{\phi}(q|z)p_{\phi}(z|i,C)p_{\phi}(i,C)$

Hence, we can optimize the variational lower bound by maximizing the image and category reconstruction whilst also maximizing the MLE of question generation.

3 INFERENCE FRAMEWORK

We illustrate the inference flow using Figure 1. During inference, given an image conditioned on the category label, z_i is sampled from the combined generative latent representation z of the inputs learnt by the model. This representation is then passed through the temporal network $LSTM_q$, thereby, outputting the generated question. While the training of C3VQG requires images and their corresponding ground-truth questions from different answer categories, the inference only requires the images with answer categories of the questions to be generated.

4 HYPERPARAMETERS

We present all a list of all the hyperparameter values used in training the C3VQG model.

Hyperparameter	Symbol	Value
Image Recon. Weight	λ_I	1.0
Category Recon. Weight	λ_C	2.0
Question Recon. Weight	λ_Q	3.0
Category Consistency Weight	λ_{cons}	2.0
Center Loss Weight	λ_{center}	3.0
Hyper-prior KL-Divergence Weight	λ_{bayes}	3.0
Hyper-prior Regularisation Weight	λ_{reg}	2.0
Dimension of combined latent space	d	64
Learning Rate	α_{LR}	1e-3

Table 1: Hyperparameters values used for training C3VQG.

5 DATASET DETAILS

We list the details about the VQA dataset [1] used for the training and evaluation of C3VQG against the state-of-the-art [3–5] in VQG.

Data Type	Training	Validation
VQA Annotations (answers)	4,437,570	2,143,540
VQA Input Questions	443,757	214,354
VQA Input Images	82,783	40,504

Table 2: Dataset details for the VQA dataset.

Krishna *et al.* [4] annotates the answers with a set of 15 categories and labels their top 500 answers. This makes up 82% of the entire VQA dataset consisting of 367K training and validation examples. We use a 80:20 training-validation split for our experiments.

6 QUALITATIVE RESULTS

We depict additional qualitative results wherein we mention four questions (from different answer categories) generated using our approach on a collection of images from the validation set. As depicted in Figure 2, the questions generated are non-generic and category-specific for each of the following examples.

7 DIVERSITY-BASED PERFORMANCE EVALUATION

Here, we depict additional comparisons in terms of the diversity-based metrics. As depicted in Table 3, the performance in terms of the diversity of generated questions achieved by our approach with all components beats the state-of-art in VQG even without the requirement of additional answer supervision. The difference in the *strength* and *inventiveness* values with and without the latent hyper-prior suggests that capturing decorrelated features in each latent dimension enables our model to generate non-generic questions from a divergent pool of categories.

REFERENCES

- [1] 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)*.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 1026–1034.
- [3] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating Diverse Questions Using Variational Autoencoders. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5415–5424.
- [4] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information Maximizing Visual Question Generation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 2008–2018.
- [5] Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A Joint Model for Question Answering and Question Generation. *ArXiv abs/1706.01450* (2017).

Figure 1: C3VQG Inference Framework

	ATTRIBUTE is the building on the right tall ? COLOR what is the color of the building ?	SPATIAL is the tree on left of the building ? MATERIAL what is the building made of ?		BINARY is this a kitchen ? COUNT how many layers is the cake ?	ATTRIBUTE how does the cake taste ? OBJECT what are the people eating ?
	COUNT how many cars are there ? COLOR what color is the truck ?	ACTIVITY what is the man in middle doing ? SPATIAL where is the car parked ?		BINARY is the train in the station ? COLOR what is the color of the train ?	ATTRIBUTE is this a modern train ? MATERIAL what are the tracks made of ?
	COUNT how many people are there ? COLOR what color is the man's shirt ?	ACTIVITY what is the man doing ? OBJECT what is the man riding ?		COUNT how many cars are there ? COLOR what color is the car ?	ATTRIBUTE is the building tall ? SPATIAL where is the car on the left parked ?

Figure 2: Questions generated for each image from multiple answer categories using C3VQG approach.

Categories	V-IC2Q [3]		Krishna <i>et al.</i> [4]		C3VQG w/o Bayes		C3VQG	
	Strength	Inventiveness	Strength	Inventiveness	Strength	Inventiveness	Strength	Inventiveness
count	15.77	30.91	26.06	41.30	58.33	55.20	65.21	61.84
binary	18.15	41.95	28.85	54.50	58.39	36.32	65.12	38.55
object	11.27	34.84	24.19	43.20	57.77	51.51	65.58	58.85
color	4.03	13.03	17.12	23.65	58.38	48.97	65.21	54.34
attribute	37.76	41.09	46.10	52.03	60.05	58.38	64.59	63.02
materials	36.13	31.13	45.75	40.72	57.93	56.79	64.87	63.48
spatial	61.12	62.54	70.17	68.18	57.90	57.80	65.18	64.96
food	21.81	20.38	33.37	31.19	58.49	55.42	65.20	62.21
shape	35.51	44.03	45.81	55.65	58.85	58.75	66.01	65.98
location	34.68	18.11	45.25	27.22	58.39	58.10	65.09	64.72
predicate	22.58	17.38	36.20	31.29	57.05	57.05	65.67	65.67
time	25.58	15.51	34.43	25.30	58.13	58.10	65.00	64.96
activity	7.45	13.23	21.32	26.53	58.00	56.78	64.98	63.67
Overall	12.97	38.32	26.06	52.11	58.23	54.99	65.24	61.55

Table 3: Quantitative evaluation of C3VQG against other baselines using diversity-based metrics.