Category Consistent Cyclic Visual Question Generation

Shagun Uppal[*1], Anish Madan[*1], Sarthak Bhagat[*1], Yi Yu[2], Rajiv Ratn Shah[1]

[1] IIIT-Delhi, [2] NII Japan, [*] Equal Contribution

May 24, 2020

Visual Question Generation (*VQG*) is the task of generating natural questions given an image.

Challenges in constructing a VQG system:

- Capturing various concepts in images.
- Relevance of generated questions to the image.
- Many-to-one mapping between the image and generated questions since multiple questions are possible for an image.
- Avoid questions which invoke generic answers like "yes" / "I don’t know".

Possible Category-Question pairs:

SPATIAL: Where are the pictures hanging?

ACTIVITY: What is the little girl doing?

BINARY: Is the lamp on?

COUNT: How many pillows are there on the bed?

COLOR: What is the color of the girl’s dress?

 Shagun Uppal*[1], Anish Madan*[1], Sarthak Bhagat*[1], Yi Yu[2], Rajiv Ratn Shah[1]

Contributions

- Weaken supervision by removing the need for answers.
- Variational training using a single combined latent space for image and category by maximizing mutual information.
- Category consistency using cyclic training in two disjoint steps.
- Center loss for category-wise clustering.
- Hyper-prior on latent space for encapsulation of independent features.

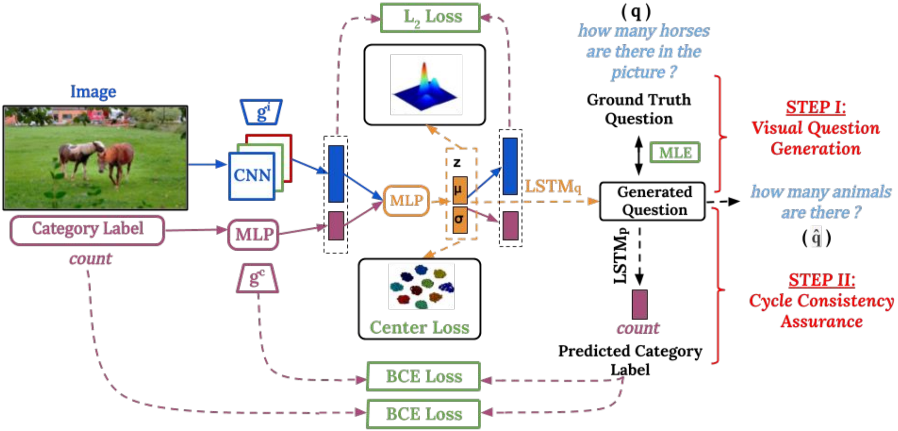

Figure 1: C3VQG Training

Inference Framework

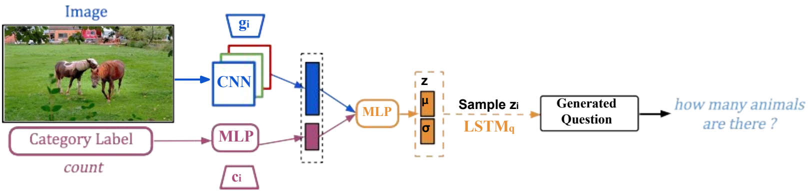

Figure 2: C3VQG Inference

Center Loss

$$\mathcal{L}_{center} = \|z_k - c_k\|_2^2$$

- Helps distinguish inter-category latent features by enforcing clustering.
- Centers are obtained by averaging the features of the corresponding classes updated based on mini-batches instead of the entire training data due to computational time constraints
- Update of these centers are scaled by a constant (< 1) to avoid sudden fluctuations.

Hyper-prior

$$\mathcal{L}_{bayes} = \sum_{j=1}^{d} \mathbb{E}_{pd(x_k^{cc})} \left[KL(f(z_{k,j} | x_{k,j}^{cc}) || \mathcal{N}(z_{k,j}; 0, \alpha^{-1})) \right] + \lambda_{reg} \sum_{j=1}^{d} (\alpha_j^{-1} - 1)^2$$

- A hyper-prior on learning the inverse variance of the variational latent prior
- Helps to capture intrinsically independent visual features within the combined latent space.
- This helps us in generating more diverse questions.

Shagun Uppal* [1], Anish Madan* [1], Sarthak Bhagat* [1], Yi Yu[2], Rajiv Ratn Shah[1]

Experimental Results - Qualitative Generations

Figure 3: Question generated for each image from multiple answer categories using our approach.

Experimental Results - Qualitative Generations

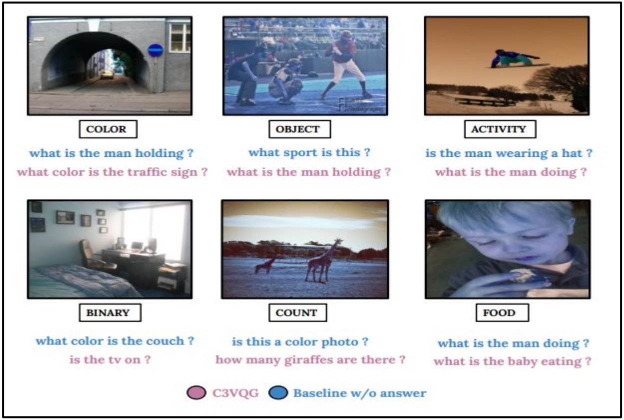

Figure 4: Qualitative results for C3VQG and Krishna et. al[1] without answers.

[1]Krishna, Bernstein, and Fei-Fei, "Information Maximizing Visual Question Generation".

We evaluate the efficacy of our approach using a set of evaluation metrics.

- Language Modelling Metrics: BLEU, METEOR, CIDEr, ROUGE-L
- Diversity Based Metrics
- Relevance Based Metrics (Crowd Sourced Metrics)

Experimental Results - Quantitative metrics

Supervision	Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	CIDEr	ROUGE-L
Supervised (w A)	IA2Q [24]	32.43	15.49	9.24	6.23	11.21	36.22	-
	V-IA2Q [9]	36.91	17.79	10.21	6.25	12.39	36.39	-
	Krishna *et al.* [14]	47.40	28.95	19.93	14.49	18.35	85.99	49.10
Weakly Supervised (w/o A)	IC2Q [24]	30.42	13.55	6.23	4.44	9.42	27.42	-
	V-IC2Q [9]	35.40	**25.55**	14.94	**10.78**	13.35	42.54	-
	Krishna *et al.* [14] w/o A	31.20	16.20	11.18	6.24	12.11	35.89	40.27
	I	38.44	19.83	12.02	7.69	13.27	45.19	40.90
	I + II	38.80	20.12	12.32	7.96	13.40	46.42	41.27
	I + CL	38.81	20.14	12.30	7.91	13.41	46.96	41.21
	I + II + CL	38.94	20.30	12.47	8.10	13.47	**47.32**	41.27
	I + II + Bayes	38.71	19.89	12.14	7.87	13.23	42.47	41.32
	I + CL + Bayes	38.64	20.06	12.28	7.95	13.32	45.83	41.16
	I + II + CL + Bayes	**41.87**	22.11	**14.96**	10.04	**13.60**	46.87	**42.34**

Table 1: Ablation study for different components of C3VQG using different language modeling quantitative metrics against other baselines in VQG. We compare our approach against previous works using answers as well as without answers.

Experimental Results - Quantitative metrics

Categories	V-IC2Q [9]		Krishna *et al.* [14]		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 2: Quantitative evaluation of C3VQG against other baselines using diversity-based metrics.

Shagun Uppal* [1], Anish Madan* [1], Sarthak Bhagat* [1], Yi Yu[2], Rajiv Ratn Shah[1]

Experimental Results - Quantitative metrics

Model	Relevance	
	Image	Category
V-IC2Q [9]	90.10	39.00
Krishna *et al.* [14] w/o A	**98.10**	42.70
C3VQG w/o Bayes, CL	98.00	58.40
C3VQG	97.80	**60.50**

Table 3: Quantitative evaluation of C3VQG against other weakly supervised baselines using crowd-sourced metrics.

Shagun Uppal* [1], Anish Madan* [1], Sarthak Bhagat* [1], Yi Yu[2], Rajiv Ratn Shah[1]

Thank You !

For more details, please check our paper:
[C3VQG: Category Consistent Cyclic Visual Question Generation]

Shagun Uppal* [1], Anish Madan* [1], Sarthak Bhagat* [1], Yi Yu[2], Rajiv Ratn Shah[1]