Visual Question Generation as Dual Task of Visual Question Answering

Yikang Li[1], Nan Duan[2], Bolei Zhou[3], Xiao Chu[1], Wanli Ouyang[4], Xiaogang Wang[1], Ming Zhou[2]

[1]The Chinese University of Hong Kong, China [2]Microsoft Research Asia, China

[3]Massachusetts Institute of Technology, USA [4]University of Sydney, Sydney, Australia

香港中文大學 The Chinese University of Hong Kong

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Motivation

- Is the sun to the left of the clock tower?
- How many sides does clock tower have?
- Does the clock tower cast a shadow?

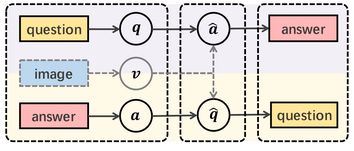
- Is the bike a two seater?
- What numbers and letters are the license plate of the motorcycle?
- Could these be family members?

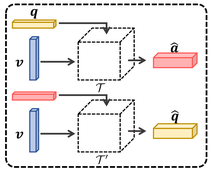
Q: What is the man holding? A: Frisbee.

VQA: person + hold → Frisbee

VQG: Frisbee → person + hold

- Visual Questions Answering covers a wide range of visual tasks: e.g. image classification, object counting, optical character recognition, etc. Current VQA dataset contains limited training data compared to the specific tasks.
- Both VQA and VQG involve reasoning between the question text and the answer text based on the content of the given image.





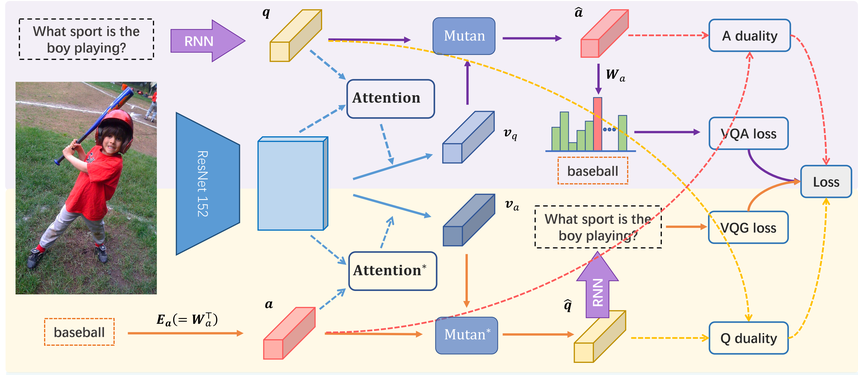
- VQG and VQA share the visual input and taking encoder-fusion-decoder pipeline with inverse order of input and output.
- Jointly learning through these two tasks can utilize the training data in a more efficient way, and bring mutual improvements to both VQA and VQG

Contribution

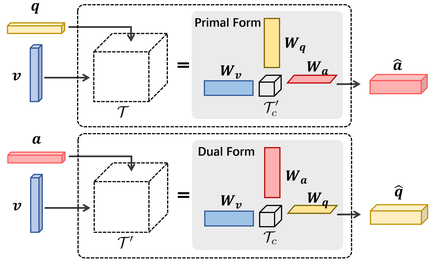
By considering VQG and VQA as dual tasks we propose a novel training framework to introduce VQG as an auxiliary task to improve VQA model performance.

- We derive a unified model, Invertible Question Answering Network (iQAN), that can accomplish both VQA and VQG with different forms.
- A novel parameter sharing scheme and duality regularization are proposed to explicitly leverage the intrinsic connections between the two tasks.
- Evaluated on VQA2 and CLEVR datasets, our proposed model achieves better results on both VQA and VQG tasks than MUTAN VQA method.
- Experimental results show that our framework can also generalize to some other popular VQA models and consistently improve their performances.

Methodology



Overview of Invertible Question Answering Network (iQAN), which consists two components for VQA and VQG respectively. The upper component is MUTAN VQA component and the lower component is its dual VQG model. Input questions and answers are encoded respectively by an RNN and a lookup table, E_a into fixed-length features. With attention and MUTAN fusion module, predicted features are obtained. The predict features are used for obtaining output (by LSTM and W_a for questions and answers respectively). A duality and Q duality are duality regularizers to constrain the similarity between the answer and question representations in both models. Two components share the MUTAN and Attention Modules. (*) denotes the dual form. E_a also shares parameters with W_a.



The Dual MUTAN in the primal form~(for VQA) and dual form~(for VQG). The two forms share one parameter set: the core tensor T_c, projection matrices of images, questions, and answers. In our experiment, W_a at the top part and W_q at the bottom part are merged with decoders.

- Bilinear Fusion Module:

$$\hat{\mathbf{a}} = (\mathcal{T} \times_1 \mathbf{q}) \times_2 \mathbf{v}_q$$

- Multimodal Tucker Decomposition:

$$\hat{\mathbf{a}} = ((\mathcal{T}_c \times_1 (\mathbf{W}_q \mathbf{q})) \times_2 (\mathbf{W}_v \mathbf{v}_q)) \times_3 \mathbf{W}_a$$

- Feature fusion of VQA and VQG (where $\tilde{q} = \mathbf{W}_q \cdot q$, $\tilde{a} = \mathbf{W}_a \cdot a$, $\tilde{v} = \mathbf{W}_v \cdot q = v$, $\tilde{v}' = \mathbf{W}_v' \cdot v$)

$$\tilde{\mathbf{a}}^* = (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}}$$
$$\tilde{\mathbf{q}}^* = (\mathcal{T}_c' \times_1 \tilde{\mathbf{a}}) \times_2 \tilde{\mathbf{v}}'$$

- Dual MUTAN Assumptions:

$$\mathcal{T}_c'[:,i,:] = \mathcal{T}_c^\top[:,i,:] \qquad \mathcal{T}_c[:,i,:] = \mathcal{T}_c'^\top[:,i,:]$$

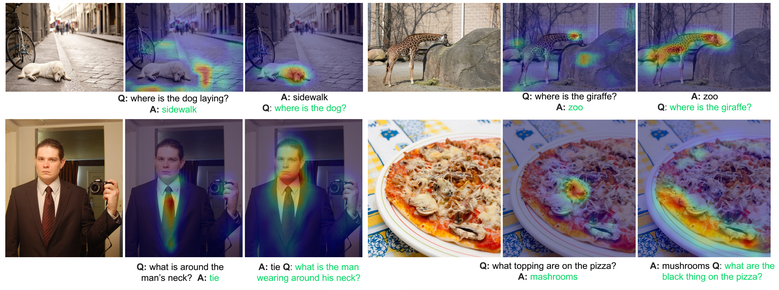
- Feature Fusion with Dual MUTAN:

$$\tilde{\mathbf{a}}^* = (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}}$$
$$\tilde{\mathbf{q}}^* = (\mathcal{T}_c \times_1 \tilde{\mathbf{a}}) \times_2 \tilde{\mathbf{v}}$$

- Duality Regularizer:

$$\mathbf{a} \approx \hat{\mathbf{a}} = \phi(\mathbf{q}, \mathbf{v}) \text{ and } \mathbf{q} \approx \hat{\mathbf{q}} = \phi^*(\mathbf{a}, \mathbf{v})$$

Experiments



Data Statistics

Dataset	Train		Validation	
	#images	#Question	#images	#Question
VQA2-Filter	68,434	163,550	33,645	78,047
VQA2-Full	82,783	443,757	40,504	214,354
CLEVR-Filter	57,656	107,132	12,365	22,759
CLEVR-Full	70,000	699,960	15,000	150,000

Experiment on the Full Dataset

Model \ Dataset	VQA2			CLEVR
	Selected	Other	All	
MUTAN	51.58	57.02	54.85	70.89
Ours-Full	52.14	57.06	55.10	73.25
Ours-Selective	**52.79**	**57.13**	**55.41**	**76.30**

Ours-Selective: apply dual training only on Filtered data.

Ablation Study

model	Dual MUTAN	Duality Regularizer	Sharing De- & Encoder	VQA2-Filter [6]		CLEVR-Filter [10]				
				acc@1	acc@5	size	material	shape	color	**Overall**
1	-	-	-	50.72	78.56	86.76	88.25	82.26	76.86	83.74
2	✓	-	-	50.99	78.71	87.29	88.10	82.82	77.62	84.13
3	✓	-	✓	51.23	78.82	87.42	88.81	82.84	77.73	84.40
4	✓	✓	-	51.42	79.00	87.75	88.48	**84.28**	77.97	84.78
5	✓	✓	✓	**51.60**	**79.16**	**87.75**	**88.91**	84.08	**78.86**	**85.07**

Generalization to other VQA Models

Model	iBOWIMG [36]			MLB [13]			MUTAN [3]			MUTAN [3] + Sharing LSTM		
	Acc@1	Acc@5	CIDEr	Acc@1	Acc@5	CIDEr	Acc@1	Acc@5	CIDEr	Acc@1	Acc@5	CIDEr
Baseline	42.05	72.79	2.224	50.23	77.64	2.236	50.72	78.56	2.203	49.91	77.47	**2.217**
Dual Training	**43.44**	**74.27**	**2.263**	**50.83**	**78.12**	**2.271**	**51.60**	**79.16**	**2.379**	**50.78**	**78.16**	2.117

Augment VQA with VQG

Methods \ Dataset	Cleansed VQA2 [6]				Full VQA2 [6]			
	0.1 (Q,A) + 0.9 A		0.5 (Q,A) + 0.5 A		0.1 (Q,A) + 0.9 A		0.5 (Q,A) + 0.5 A	
	Acc@1	CIDEr	Acc@1	CIDEr	Acc@1	CIDEr	Acc@1	CIDEr
Baseline	33.60	1.332	46.68	1.930	36.88	1.194	50.77	1.495
DT	35.23	1.540	47.63	2.101	42.14	1.104	51.51	1.467
VQG+DT	38.87	1.528	47.99	2.072	43.05	1.103	50.82	1.471
VQG+DT+FT	**39.95**	**1.739**	**48.48**	**2.281**	**44.61**	**1.237**	**52.13**	**1.488**

Github Repo: iQAN