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, Hong Kong, China [2]Microsoft Research Asia, China
[3]Massachusetts Institute of Technology, USA [4]University of Sydney, Australia

Abstract

Visual question answering (VQA) and visual question generation (VQG) are two trending topics in the computer vision, but they are usually explored separately despite their intrinsic complementary relationship. In this paper, we propose an end-to-end unified model, the Invertible Question Answering Network (iQAN), to introduce question generation as a dual task of question answering to improve the VQA performance. With our proposed invertible bilinear fusion module and parameter sharing scheme, our iQAN can accomplish VQA and its dual task VQG simultaneously. By jointly trained on two tasks with our proposed dual regularizers (termed as Dual Training), our model has a better understanding of the interactions among images, questions and answers. After training, iQAN can take either question or answer as input, and output the counterpart. Evaluated on the CLEVR and VQA2 datasets, our iQAN improves the top-1 accuracy of the prior art MUTAN VQA method by 1.33% and 0.88% (absolute increase) respectiely. We also show that our proposed dual training framework can consistently improve model performances of many popular VQA architectures. [1]

1. Introduction

Question answering (QA) and question generation (QG) are two fundamental tasks in natural language processing [24, 25]. Recently, two homogenous tasks, Visual Question Answering (VQA) [38, 36, 2, 21] and Visual Question Generation (VQG) [27, 37], have been introduced to the computer vision field as cross-modality learning tasks. VQA refers to answering questions based on the image, while VQG aims at generating reasonable questions given the image contents. Both VQA and VQG involve reasoning between the question text and the answer text based on the content of the given image.

In previous works, VQA and VQG are studied independently. As shown in Fig. 1, the VQA model usually encodes the question sentence as an embedding $\mathbf{q}$, then fuses $\mathbf{q}$ with the image feature $\mathbf{v}$ to infer the answer embedding $\hat{\mathbf{a}}$, which is decoded as the distribution over the answer vocabulary. In this work, we consider answer-based visual question generation (termed as *VQG* for simplicity) as an inverse form of VQA, which is to generate a question corresponding to the given image and answer. The VQG model merges the answer embedding $\mathbf{a}$ and the image feature $\mathbf{v}$ to get the question embedding $\hat{\mathbf{q}}$. Then $\hat{\mathbf{q}}$ is decoded to generate the question sentence. We can see that these two tasks are intrinsically correlated, *i.e.* sharing visual input and taking encoder-fusion-decoder pipeline with inverse input and output. Thus, we refer them as "Dual" tasks.

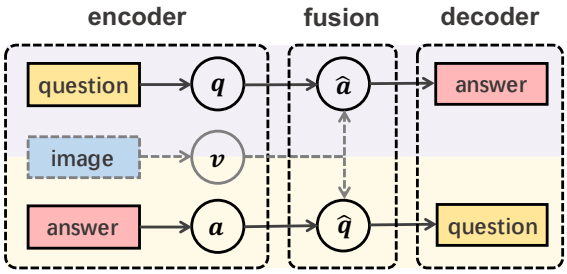

Figure 1. Problem solving schemes of VQA (top) and VQG (bottom), both of which utilize the ⟨*encoder-fusion-decoder*⟩ pipeline with Q and A in inverse order. $\mathbf{v}$, $\mathbf{q}$ and $\mathbf{a}$ respectively denote the encoded features of input image, question, and answer, while $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ represent the predicted answer/question features.

Duality reflects the inherent complementary relation between question answering and generation. Intuitively, learning to answer questions may boost the question generation and vice versa, as both of them require similar abilities: image recognition, question reasoning, cross-modal information association, etc. Taking an image with Q/A pair "What is around the man's neck? Tie" as an example. Inspired by the previous works on visual relationships [17, 18], the image can be viewed as a visual relationship ⟨*tie-around-neck*⟩. VQA is to infer ⟨*tie*⟩ given the description ⟨*around-neck*⟩, while VQG is to generate a question describing the visual information most related to the answer ⟨*tie*⟩. VQG and VQA can be viewed as two inverse reasoning processes on related semantics. Thus, joint learning through these two

[1]This work was done during the first author's internship in MSR Asia.

tasks can utilize the training data in a more efficient way, and bring mutual improvements to both VQA and VQG. So we formulate the dual training of VQA and VQG as learning an invertible cross-modality fusion model that can infer Q or A when given the counterpart conditioned on the given image.

From this perspective, we derive an invertible fusion module, *Dual MUTAN*, based on a popular VQA model MUTAN [3]. The module can complete the feature inference in a bidirectional manner, *i.e*. it can infer the answer embeddings from image+question and the question embeddings from image+answer. Furthermore, by sharing the visual encoder as well as the encoder and decoder of the question and answer, VQG and VQA models can be viewed as the two inverse forms of one model. When the model is jointly trained on the two tasks, the invertibility brought by our parameter sharing schemes can help to regularize the training process to learn more general representations. In addition, besides the label-level matching, we also introduce the similarity of the question/answer embeddings of the two tasks as extra regularizations to guide the training process.

Contribution: In this work, 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. Correspondingly, we derive a unified model that can accomplish both VQA and VQG with different forms, called Invertible Question Answering Network (iQAN). The model is jointly trained with VQA and VQG tasks and can be deployed for either task in the testing stage. Additionally, 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.

2. Related Work

Visual Question Answering is one of the most popular cross-discipline tasks aiming at understanding both the image, question and their interactions. Malinowski *et al*. propose an encoder-decoder framework to merge the visual and textual information for answering prediction [23]. Shih *et al*. introduce visual attention mechanism to highlight the image regions relevant to answering the question [30]. Lu *et al*. further apply attention to the language model, called co-attention, to jointly reason about images and questions [20]. Apart from proposing new frameworks, some focus on designing effective multimodal fusion schemes [5, 13]. The bilinear model MUTAN proposed by Ben-younes *et al*. is one of the state-of-the-art methods to model interactions between two modalities [3]. Additionally, several benchmark datasets are proposed to facilitate the VQA research [22, 2]. VQA2 is the most popular open-ended Q-A dataset [6] where each question is associated with a pair of similar images that result in different answers. CLEVR was recently proposed by Johnson *et al*. with rendered images and automatically-generated questions to mitigate answer biases and diagnose the reasoning ability of VQA models [10]. In the experiment part, we will evaluate our method on these two datasets.

Visual Question Generation. Question generation has been investigated for years in natural language processing [1, 12, 29]. Recently, it has been introduced to computer vision to generate image-related questions. Mora *et al*. propose a CNN-LSTM model to simultaneously generate image-related questions and corresponding answers [26]. Mostafazadeh *et al*. collect the first VQG dataset, where each image is annotated with several questions [27]. Zhang *et al*. propose a model to automatically generate visually grounded questions [37], where Densecap [11] is used to generate region captions as extra information to guide the question generation. Jain *et al*. combine the variational autoencoder and LSTM to generate diverse questions [9]. Different from the existing works to generate question solely based on images, we provide an annotated answer as an additional cue. Therefore, VQG can be modeled as a multi-modal fusion problem like VQA.

Dual Learning. Utilizing cycle consistency to regularize the training process has a long history. It has been used as a standard trick for years in visual tracking to enforce forward-backward consistency [31]. He *et al*. formulate the idea as *Dual Learning* in machine translation [7], which uses two translation models, *A-to-B* and *B-to-A*, to form two closed translation loops *A-B-A* and *B-A-B*, and force them translate the output of each other back to the original input. So the models could learn the translation functions between A and B from large quantities of unlabeled data. Tang *et al*. introduce the idea to QA area, where question generation is modeled as dual task of QA, and leverage the probabilistic correlation between QA and QG to guide the training [33]. Zhu *et al*. use the thought in the computer vision area and propose CycleGAN to learn image-to-image translation functions in an unsupervised manner[39]. Different from one-to-one translation problems, where there exists large quantities of available unpaired data, visual question answering is a multimodal fusion problem, which is hard to model as an unsupervised learning problem. The most critical thing is to make full use of labeled data. Therefore, we introduce VQG as a dual task of VQA and leverage their inherent connections to boost VQA by training the model on the two tasks.

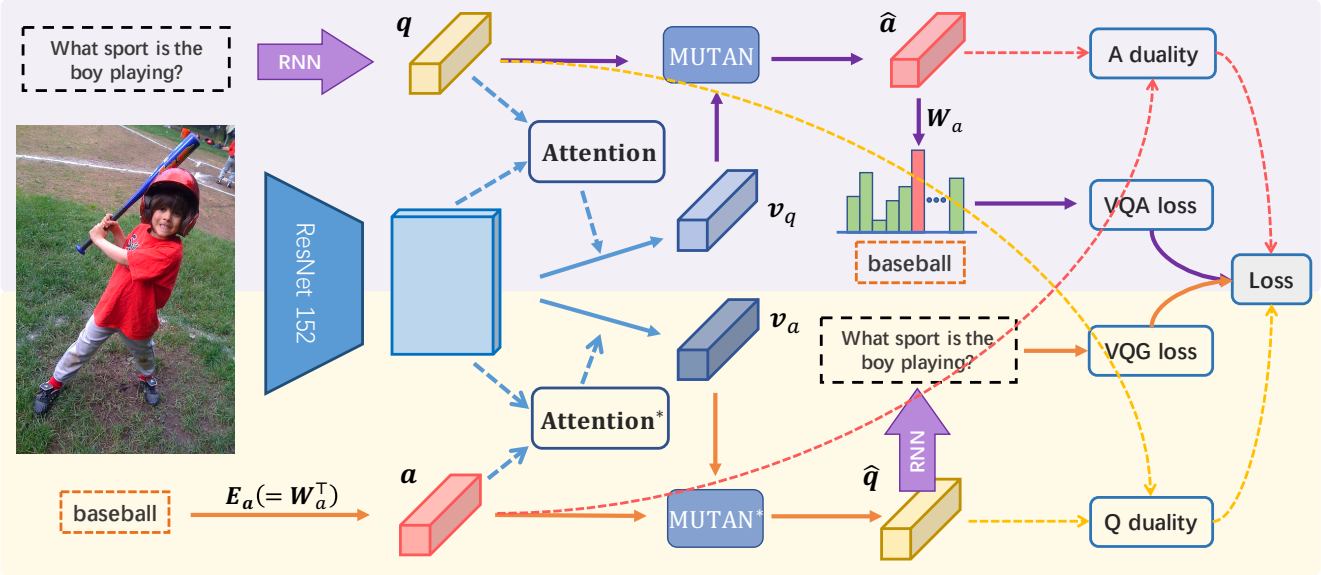

Figure 2. Overview of Invertible Question Answering Network (iQAN), which consists two components for VQA and VQG respectively. The upper component is MUTAN VQA component [3], and the lower component is its dual VQG model. Input questions and answers are encoded respectively by an **RNN** and a lookup table $\mathbf{E}_a$ 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 $\mathbf{W}_a$ 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. $(\cdot)^*$ denotes the dual form. $\mathbf{E}_a$ also shares parameters with $\mathbf{W}_a$.

3. Invertible QA Network (iQAN)

In this section, we present the dual training framework of the VQA and VQG, Invertible Question Answering Network (iQAN). The overview of our proposed iQAN is shown in Fig. 2, which consists of two components, VQA component (top) and VQG component (bottom).

In the VQA component, given a question, an RNN is used for obtaining the embedded feature $\mathbf{q} \in \mathbb{R}^{d_q}$, and CNN is used to transform the input image into a feature map $\mathbf{v}$. A MUTAN-based attention module is used to generate a question-aware visual feature $\mathbf{v}_q \in \mathbb{R}^{d_v}$ from the image and the question. Then another MUTAN fusion module is used for obtaining the answer features $\hat{\mathbf{a}} \in \mathbb{R}^{d_a}$ by fusing $\mathbf{v}_q$ and $\mathbf{q}$. Finally, a linear classifier $\mathbf{W}_a$ is used to predict the answer.

In the VQG component, given an answer, a lookup table $\mathbf{E}_a$ is used for obtaining the embedded feature $\mathbf{a} \in \mathbb{R}^{d_a}$. CNN with attention module is used for obtaining the visual feature $\mathbf{v}_a \in \mathbb{R}^{d_v}$ from the input image and the answer feature $\mathbf{a}$. Then the MUTAN in the dual form, which shares parameters with VQA MUTAN but in a different structure, is used for obtaining the predicted question features $\hat{\mathbf{q}} \in \mathbb{R}^{d_q}$. Finally, an LSTM-based decoder is employed to translate $\hat{\mathbf{q}}$ to the question sentence.

We formulate the VQA and VQG components as inverse process to each other by introducing a novel parameter sharing scheme and the duality regularizer. Consequently, we could jointly train one model with two tasks to learn the dependencies between questions and answers in a bidirectional way. The invertibility of the model could serve as a regular term to guide the training process.

3.1. The VQA component

The VQA component of our proposed iQAN is based on one of the state-of-the-art VQA models, MUTAN. We will briefly review the core part, MUTAN fusion module, which takes an image feature $\mathbf{v}_q$ and a question feature $\mathbf{q}$ as input, and predicts the answer feature $\hat{\mathbf{a}}$.

3.1.1 Review on MUTAN fusion module

Since language and visual representations are in different modalities, merging visual and linguistic features is crucial in VQA. Bilinear models are recently used in the multimodal fusion problem, which encodes bilinear interactions between $\mathbf{q}$ and $\mathbf{v}_q$ as follows:

$$\hat{\mathbf{a}} = (\mathcal{T} \times_1 \mathbf{q}) \times_2 \mathbf{v}_q \tag{1}$$

where the tensor $\mathcal{T} \in \mathbb{R}^{d_q \times d_v \times d_a}$ denotes the fully-parametrized operator for answer feature inference, and $\times_i$ denotes the *mode-i* product between a tensor $\mathcal{X}$ and a matrix

$\mathbf{M}$:

$$(\mathcal{X} \times_i \mathbf{M})[d_1, ...d_{i-1}, j, d_{i+1}...d_N] = \sum_{d_i=1}^{D_i} \mathcal{X}[d_1...d_N]\mathbf{M}[d_i, j] \tag{2}$$

To reduce the complexity of the full tensor $\mathcal{T}$, Tucker decomposition [3] is introduced as an effective way to factorize $\mathcal{T}$ as a tensor product between factor matrices $\mathbf{W}_q$, $\mathbf{W}_v$ and $\mathbf{W}_a$, and a *core tensor* $\mathcal{T}_c$:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 \mathbf{W}_q) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_a \tag{3}$$

with $\mathbf{W}_q \in \mathbb{R}^{t_q \times d_q}$, $\mathbf{W}_v \in \mathbb{R}^{t_v \times d_v}$ and $\mathbf{W}_a \in \mathbb{R}^{t_a \times d_a}$, and $\mathcal{T}_c \in \mathbb{R}^{t_q \times t_v \times t_a}$. Consequently, we can rewrite Eq. 1 as:

$$\hat{\mathbf{a}} = ((\mathcal{T}_c \times_1 (\mathbf{W}_q\mathbf{q})) \times_2 (\mathbf{W}_v\mathbf{v}_q)) \times_3 \mathbf{W}_a \tag{4}$$

where matrices $\mathbf{W}_q$ and $\mathbf{W}_v$ transform the question features $\mathbf{q}$ and image features $\mathbf{v}_q$ into dimensions t_q and t_v respectively. The squeezed bilinear core $\mathcal{T}_c$ models the interactions among the transformed features and projects them to the answer space of size t_a, which is used to infer the per-class score by $\mathbf{W}_a$.

If we define $\tilde{\mathbf{q}} = \mathbf{W}_q\mathbf{q} \in \mathbb{R}^{t_q}$ and $\tilde{\mathbf{v}}_q = \mathbf{W}_v\mathbf{v}_q \in \mathbb{R}^{t_v}$, then we have:

$$\tilde{\mathbf{a}} = (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}}_q \in \mathbb{R}^{t_a} \tag{5}$$

Thus, $\tilde{\mathbf{a}}$ can be viewed as the *answer feature* where $\hat{\mathbf{a}} = \tilde{\mathbf{a}}^\top \times \mathbf{W}_a$.

To balance the complexity and expressivity of the interaction modeling, the low rank assumption is introduced, and $\mathcal{T}_c[:, :, k]$ can be expressed as a sum of R rank-1 matrices:

$$\mathcal{T}_c[:, :, k] = \sum_{r=1}^{R} \mathbf{m}_r^k \otimes {\mathbf{n}_r^k}^\top \tag{6}$$

with $\mathbf{m}_\mathbf{r}^\mathbf{k} \in \mathbb{R}^{t_q}$, $\mathbf{n}_\mathbf{r}^\mathbf{k} \in \mathbb{R}^{t_v}$ and $\otimes$ denoting the outer product. Then each element of $\tilde{\mathbf{a}}$ can be written as:

$$\tilde{\mathbf{a}}[k] = \sum_{r=1}^{R} (\tilde{\mathbf{q}}^\top \mathbf{m}_r^k)(\tilde{\mathbf{v}}_q^\top \mathbf{n}_r^k) \tag{7}$$

We can define R matrices $\mathbf{M}_r \in \mathbb{R}^{t_q \times t_a}$ and $\mathbf{N}_r \in \mathbb{R}^{t_v \times t_a}$ such that $\mathbf{M}_r[:, k] = \mathbf{m}_r^k$ and $\mathbf{N}_r[:, k] = \mathbf{n}_r^k$. Therefore, with the low rank constraint, Eq. 5 is further simplified as:

$$\tilde{\mathbf{a}} = \sum_{r=1}^{R} (\tilde{\mathbf{q}}^\top \mathbf{M}_r) \odot (\tilde{\mathbf{v}}_q^\top \mathbf{N}_r) \tag{8}$$

where $\odot$ denotes the element-wise product.

With MUTAN, low computational complexity and strong expressivity of the model are both obtained for visual question answering part.

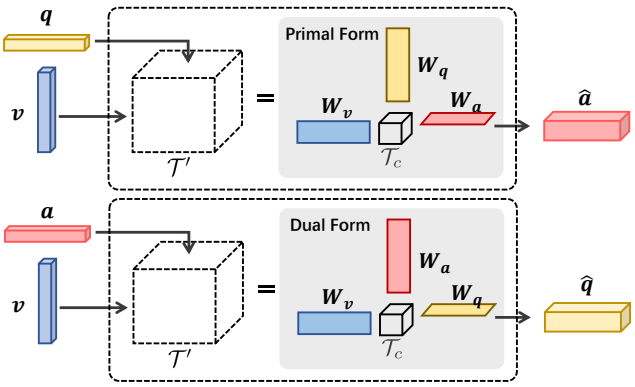

Figure 3. The Dual MUTAN in the primal form (for VQA) and dual form (for VQG). The two forms share one parameter set: the core tensor $\mathcal{T}^c$, projection matrices of images, questions, and answers. In our experiment, $\mathbf{W}_a$ at the top part and $\mathbf{W}_q$ at the bottom part are merged with decoders.

3.2. The VQG component

The VQG component of our proposed iQAN is formulated as generating a question (word sequence) given an image and an answer label.

During training, our target is to learn a model such that the generated question $\hat{q}$ is similar to the referenced one q^*. The generation of each word of the question can be written as:

$$\hat{w}_t = \arg\max_{w \in \mathbb{W}} p\left(w \middle| v, w_0, ..., w_{t-1}\right) \tag{9}$$

where $\mathbb{W}$ denotes the word vocabulary. $\hat{w}_t$ is the predicted word at t step. w_i represents the i-th ground-truth word. Beam search will be used during inference.

VQG shares the visual CNN with VQA part. The answer feature $\mathbf{a} \in \mathbb{R}^{d_a}$ is directly retrieved from the answer embedding table $\mathbf{E}_a$. MUTAN is also utilized for visual attention module and visual & answer representations fusion at VQG. Similar to Eq. 8, the inference of question features $\tilde{\mathbf{q}}$ can be formulated as:

$$\tilde{\mathbf{q}} = \sum_{r=1}^{R} (\tilde{\mathbf{a}}^\top \mathbf{M}'_r) \odot (\tilde{\mathbf{v}}_a^\top \mathbf{N}'_r) \tag{10}$$

with $\tilde{\mathbf{a}} = \mathbf{W}_a\mathbf{a} \in \mathbb{R}^{t_a}$ and $\tilde{\mathbf{v}}_a = \mathbf{W}_v\mathbf{v}_a \in \mathbb{R}^{t_v}$. $\mathbf{M}'_r$ and $\mathbf{N}'_r$ are defined as Eq. 8. The predicted question feature $\tilde{\mathbf{q}}$ is fed into an RNN-based model to generate the predicted question.

From the formulation in Eq. 8 and Eq. 10, the VQG MUTAN could be viewed as the conjugate form of the VQA MUTAN. We will investigate the connection between the two MUTAN modules in the next section.

3.3. Dual MUTAN

To leverage the duality of questions and answers, we derive a *Dual MUTAN* from the original MUTAN to fin-

ish the primal (question-to-answer) and its dual (answer-to-question) inference on the feature level with one fusion kernel.

First we rewrite Eq. 5 and its dual form:

$$\begin{aligned}\tilde{\mathbf{a}}^* &= (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}} \\ \tilde{\mathbf{q}}^* &= (\mathcal{T}_c' \times_1 \tilde{\mathbf{a}}) \times_2 \tilde{\mathbf{v}}'\end{aligned} \tag{11}$$

where $\mathcal{T}_c \in \mathbb{R}^{t_q \times t_v \times t_a}$, $\mathcal{T}_c' \in \mathbb{R}^{t_a \times t_v \times t_q}$, $\tilde{\mathbf{q}} = \mathbf{W}_q \mathbf{q}$, $\tilde{\mathbf{a}} = \mathbf{W}_a \mathbf{a}$, $\tilde{\mathbf{v}} = \mathbf{W}_v \mathbf{v}$, and $\tilde{\mathbf{v}}' = \mathbf{W}_v' \mathbf{v}$. For simplicity, it is assumed that both VQA and VQG adopt $\mathbf{v}$ as visual input, which can be replaced by the post-attention feature $\mathbf{v}_a$ or $\mathbf{v}_q$. Noticing both $\mathcal{T}_c$ and $\mathcal{T}_c'$ depict the interactions among the image, question, and answer embeddings, but with different dimension arrangements, we assume the relationship between $\mathcal{T}_c'$ and $\mathcal{T}_c$ as follows:

$$\mathcal{T}_c'[:, i, :] = \mathcal{T}_c^\top[:, i, :] \tag{12}$$

Additionally, the transform matrices for visual information $\mathbf{W}_v'$ and $\mathbf{W}_v$ can also be shared. Therefore, we can unify the question and answer embedding inference with single three-way operator $\mathcal{T}_c$:

$$\begin{aligned}\tilde{\mathbf{a}}^* &= (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}} \\ \tilde{\mathbf{q}}^* &= (\mathcal{T}_c \times_3 \tilde{\mathbf{a}}) \times_2 \tilde{\mathbf{v}}\end{aligned} \tag{13}$$

Furthermore, since $\mathcal{T}_c[:, i, :]$ models the correlation between the re-parameterized question and answer embeddings, considering the duality of Q and A, we introduce the symmetry as an additional constraint for $\mathcal{T}_c[:, i, :]$:

$$\begin{cases} t_a = t_q = t \\ \mathcal{T}_c[:, i, :] = \mathcal{T}_c^\top[:, i, :], \ i \in \{1, 2, ..., t_v\} \end{cases} \tag{14}$$

Correspondingly, Eq. 13 could be written as:

$$\begin{aligned}\tilde{\mathbf{a}}^* &= (\mathcal{T}_c \times_1 \tilde{\mathbf{q}}) \times_2 \tilde{\mathbf{v}} \\ \tilde{\mathbf{q}}^* &= (\mathcal{T}_c \times_1 \tilde{\mathbf{a}}) \times_2 \tilde{\mathbf{v}}\end{aligned} \tag{15}$$

That is to say, we could infer $\tilde{\mathbf{a}}$ or $\tilde{\mathbf{q}}$ by just alternating the *mode-1* input of the kernel.

By introducing the low rank constraint like Eq. 8, the inference of answer and question features $\tilde{\mathbf{a}}^*$ and $\tilde{\mathbf{q}}^*$ can be reformulated as:

$$\begin{aligned}\tilde{\mathbf{a}}^* &= \sum_{r=1}^{R} \left(\tilde{\mathbf{q}}^\top \mathbf{M}_r\right) \odot \left(\tilde{\mathbf{v}}^\top \mathbf{N}_r\right) \\ \tilde{\mathbf{q}}^* &= \sum_{r=1}^{R} \left(\tilde{\mathbf{a}}^\top \mathbf{M}_r\right) \odot \left(\tilde{\mathbf{v}}^\top \mathbf{N}_r\right)\end{aligned} \tag{16}$$

And the target answer and question embeddings are provided by:

$$\begin{aligned}\hat{\mathbf{a}} &= \tilde{\mathbf{a}}^{*\top} \times \mathbf{W}_a \\ \hat{\mathbf{q}} &= \tilde{\mathbf{q}}^{*\top} \times \mathbf{W}_q\end{aligned} \tag{17}$$

As shown in Fig. 3, we unify the two MUTAN modules by sharing parameters $\mathbf{W_a}$, $\mathbf{W}_q$, $\mathbf{W}_v$, and $\mathcal{T}_c$. And we call this invertible module *Dual MUTAN*.

Furthermore, when the decoder after the dual MUTAN-MUTAN are considered, the predicted answer embedding $\hat{\mathbf{a}}$ will be fed into another linear transform layer to get the per-class score, and the question embedding $\hat{\mathbf{q}}$ will be decoded by LSTM, both of which have linear transforms afterwards. So the linear transforms in Eq. 17 can be skipped for efficiency. And we can directly use $\tilde{\mathbf{a}}^*$ and $\tilde{\mathbf{q}}^*$ as the predicted features to feed into decoders.

3.4. Weight Sharing between Encoder and Decoder

Considering the duality of VQA and VQG, the encoder and decoder of Q/A can be viewed as inverse transformation to each other. Hence, we could employ these properties to propose corresponding weight sharing scheme to learn better representations through two processes.

In the VQG component, the input answer is embedded into features $\mathbf{a}$ by the matrix $\mathbf{E}_a$, which stores the embeddings of each answer. For the answer generation in the VQA component, the predicted feature $\hat{\mathbf{a}}$ is decoded for obtaining the answer through a linear classifier $\mathbf{W}_a$, which can be regarded as a set of per-class templates for the feature matching. Thus, we can directly share the weights of $\mathbf{E}_a$ and $\mathbf{W}_a$, where $\mathbf{E}_a = \mathbf{W}_a^T$, to reflect their intrinsic connections.

For input questions in the VQA component, RNN is applied to encode the question sentence into a fixed-size feature vector $\mathbf{q}$. For the question generation in the VQG component, RNN is also utilized to decode the vector back to a word sequence. Sharing the weights of two RNNs can be an option. But it makes no sense to use one RNN for two different purposes. However, since question encoder and decoder use identical word vocabulary, we can share their word embeddings. So the two tasks could help to learn more general word representations.

3.5. Duality Regularizer

With Dual MUTAN, we have reformulated the feature fusion part of VQA and VQG (ϕ and ϕ^*) as the inverse process to each other. ϕ and ϕ^* are expected to form a closed cycle on the feature level. Consequently, given a question/answer pair $(\mathbf{q}, \mathbf{a})$, the predicted answer/question representations are expected to have the following form:

$$\mathbf{a} \approx \hat{\mathbf{a}} = \phi(\mathbf{q}, \mathbf{v}) \text{ and } \mathbf{q} \approx \hat{\mathbf{q}} = \phi^*(\mathbf{a}, \mathbf{v}). \tag{18}$$

To leverage the property above, we propose the Duality Regularizer, $\text{smooth}_{L1}(\hat{\mathbf{q}} - \mathbf{q})$ and $\text{smooth}_{L1}(\hat{\mathbf{a}} - \mathbf{a})$, where the loss function smooth_{L1} is defined as:

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5 * x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{19}$$

By minimizing Q/A duality loss, primal and dual question/answer representations are unified, and VQG and VQA are linked with each other. Moreover, the Duality Regularizer could also provide soft targets for the question/answer features.

3.6. Dual Training

With our proposed weight sharing schema (Dual MUTAN and Sharing De-/Encoder), our VQA and VQG models can be reconstructed to each other with parameters shared. Hence, joint training on VQG and VQA tasks introduces the *invertibility* of the model as an additional regular term to regularize the training process. The overall training loss including our proposed Q/A duality is as below:

$$\begin{aligned} Loss = & L_{(vqa)}\left(a, a^*\right) + L_{(vqg)}\left(q, q^*\right) \\ & + \text{smooth}_{L1}\left(\mathbf{q} - \hat{\mathbf{q}}\right) + \text{smooth}_{L1}\left(\mathbf{a} - \hat{\mathbf{a}}\right) \end{aligned} \tag{20}$$

where $L_{(vqa)}\left(a, a^*\right)$ and $L_{(vqg)}\left(q, q^*\right)$ adopt the multinomial classification loss [3] and sequence generation loss [35] as the unary loss for VQA and VQG components respectively, and the latter two terms are our proposed Q/A duality losses in Sec. 3.5. As every operation is differentiable, the entire model can be trained in an end-to-end manner.

4. Experiments

Model implementation details, data preparation and experiment results will be introduced in this section. Besides, we evaluate the effectiveness of the cycle-consistency in VQA and show that our proposed dual training scheme is more suitable for the supervised learning problem.

4.1. Implementation Details

Our iQAN is based on PyTorch implementation of MUTAN VQA [3]. We directly use the ImageNet-pretrained ResNet-152 [8] as our base model, whose block_5c output without the final average pooling is used as the visual features. All images are resized to 448×448. Newly introduced parameters are randomly initialized. Adam [14] with fixed learning rate 0.0001 is used to update the parameters. The training batch size is 512. All models are trained for 50 epochs.

4.2. Data Preparation

We evaluate the proposed method on two large-scale VQA datasets, VQA2 [6] and CLEVR [10], both of which provide images and labeled ⟨Q,A⟩ pairs. However, these two datasets contain some of the questions with less informative answers as *yes/no* or *number*. It is nearly impossible for a model to generate expected questions from an answer like *yes*. Therefore, we filter out these question-answer pairs for both the VQA2 and the CLEVR to fairly explore the duality of Q and A: For VQA2, we only select the questions with annotated question type starts with "what", "where" or "who". For CLEVR, the questions starting with "what" and whose answer is not a *number* are selected. Additionally, for VQA2, the answer vocabulary only contains the top-2000 most frequent answers as in [3]. The ⟨Q,A⟩ pairs whose answer is not in the vocabulary will be removed. Detailed statistics of cleansed datasets are shown in Tab. 2.

4.3. Performance Metrics

VQA is commonly formulated as the multinomial classification problem while VQG is a sequence generation problem. Therefore, we use top-1 accuracy (Acc@1) and top-5 accuracy (Acc@5) as VQA metrics. CIDEr [34] is used to indicate the quality of generated questions. Detailed evaluation results with other metrics including BLEU [28], METEOR [16] and ROUGE-L [19] are shown in supplementary materials.

4.4. Component Analysis

We compare our proposed Dual Training scheme with the baseline MUTAN model on the filtered VQA2 and CLEVR datasets. Tab. 1 shows our investigation on different settings. Model 1 is the baseline model with separated VQA and VQG models.

First, we focus on the VQA2 dataset. By comparing the model 1 and 2 in Tab. 1, our proposed Dual MUTAN can help to improve VQA but not significantly. This is because Dual MUTAN module may learn unreasonable parameters if not appropriately regularized. Therefore, with the regularization from the duality regularizer and the encoder & decoder weight sharing, the model performance is further improved, and the full model (model 5 Tab. 1) outperforms the baseline model by 0.88% absolute increase on top-1 accuracy, which is a significant improvement for VQA. Experiment results on the full VQA2 and CLEVR datasets are shown in Sec. 4.6.

We also evaluate our proposed method on the CLEVR dataset, which is designed to diagnose the reasoning ability of VQA models. Compared to VQA2, CLEVR dataset have simpler rendered images and much harder questions which require a strong reasoning ability, e.g."What size is the cylinder that is left of the brown metal thing that is left of the big sphere". By comparing our full model and baseline model, 1.33% gains on overall Acc@1 show that our dual training scheme could help to improve the reasoning ability of the VQA model.

4.5. Dual Training for Other VQA Models

Although the dual training method is derived from MUTAN, the core idea of dual training can be applied to other latest VQA methods [38, 13] (shown in Tab. 3).

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**

Table 1. Ablation study on the cleansed dataset. **Dual MUTAN**: our proposed sharing MUTAN scheme. **Duality Regularizer**: an additional regular term defined in Eq. (19) and (20) to guarantee the similarity of dual pairs ($\mathbf{q} \approx \hat{\mathbf{q}}$ and $\mathbf{a} \approx \hat{\mathbf{a}}$). **Sharing De- & Encoder**: parameter sharing scheme for decoders and encoders of Q and A. Model 1 is the baseline model with VQA and VQG models separated. Additionally, the per-question-type top-1 accuracies on CLEVR are also listed.

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

Table 2. Statistics of the filtered-version and full VQA2 [6] and CLEVR [10] datasets.

iBOWIMG is a simple VQA model with bag-of-words (BOW) question encoder which simply concatenates image and question embeddings to predict the answer. Correspondingly, we implement a dual VQG model with identical feature concatenation fusion. Since there is no parameter for fusion part, Dual Training only requires decoder & encoder weight sharing and duality regularizers. Experiment results show that jointly training VQG and VQA could bring mutual improvements to both, especially for VQA model (1.39% on Acc@1). However, the improvement for VQG is not significant, because iBOWIMG VQA uses BOW to encode questions while the VQG model uses LSTM to decode question features, where the compulsive similarity of predicted features for LSTM and BOW-encoded feature will be too strong as a regularizer.

MLB is another latest bilinear VQA model that can be viewed as the special case of MUTAN which sets the core bilinear operator $\mathcal{T}^c$ to identity. Therefore, the derived dual training scheme can be applied to MLB model directly and can bring mutual improvements on VQG and VQA tasks.

MUTAN: The original MUTAN model in [3] utilizes the pretrained skip-thought model [15] as question encoder, so we change that to LSTM (trained from scratch) to make it sharable with decoder. For both versions, the dual training could consistently bring gains to VQA. Nevertheless, the worse VQG performance of MUTAN + Sharing LSTM shows that using one LSTM to finish decoding and encoding may deteriorate the question generation result.

Experimental results on three latest VQA models show that our proposed dual training strategy can be used for other VQA models and bring concordant improvements.

4.6. Dual Training on the Full VQA Dataset

As we discussed in Sec. 4.2, generating questions from less informative answers like *yes/no* or *numer* is almost impossible, where the training loss from the VQG part will dominate the and deteriorate the model training. Thus, we propose to apply the dual training only for the selected QA pairs in Sec. 4.2, and employ the rest only for the VQA training (**Ours-Selective** in Tab. 4). In addition, we also list the results with baseline MUTAN [3] VQA model (**MUTAN**) and applying our proposed dual training directly to the full dataset (**Ours-Full**).

From the experiment results in Tab. 4 we can see that the gain brought by our proposed Dual Training mainly comes from the Selected Questions, where the gains on the less informative questions are rather marginal. Moreover, more significant improvements are observed on **Ours-Selective**, which shows that our Dual Training scheme is more suitable for the QA pairs with comparable information.

4.7. Discussion

From the qualitative results of VQA and VQG generated by the trained model in Fig. 4, we can see that the dual-trained iQAN has learnt the interactions among answers, questions and images in a bidirectional way. Its VQA form can associate the question and image to find the answer, while VQG form can generate questions corresponding to the given answers although they are not identical to the labeled ones.

More interestingly, attention maps of VQA and VQG also reflect the intrinsic duality of the two problems and how they work. QA pairs usually involve a set of interacted visual concepts within the image, where VQA and VQG focus on different parts. For example, the bottom-left image shows "a man wearing a tie around his neck". VQA concentrates on the "tie" given the "man's neck", while VQG captures the contents related to the "tie". The two processes are both reasoning among the objects but with different cues. Therefore, dual training on VQG and VQA can

Model	iBOWIMG [38]			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

Table 3. Evaluation of Dual Training Scheme on different VQA models. **Acc@1** and **Acc@5** are the VQA metrics, while **CIDEr** score is used to measure the question generation quality. **Baseline** models are separately trained on filtered data. **Dual Training** is to employ our proposed parameter sharing schemes and Dual Regularizer. The Dual Training version is to train one model with two tasks while Baseline is to train two different models. *Sharing LSTM* denote the question encoder and decoder share one LSTM.

Figure 4. Qualitative results of our proposed iQAN for VQA and VQG. Corresponding attention maps are also shown. Green and Red indicate the correct and false model-generated results respectively.

Model \ Dataset	VQA2 [6]			CLEVR [10]
	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**

Table 4. Experiments on the full dataset. **Ours-Full** denotes our proposed iQAN trained on the full dataset. **Ours-Selective** means we use the selected questions for dual training and the rest only to train VQA. **Other** denotes the results tested on the less informative questions like Yes/No and counting questions. **Selected** denotes the results tested on the selected informative questions. **All** denotes the overall results.

be read as helping the model learn to recognize their interactions by shading different parts. So the same set of QA pairs will provide nearly double training instances, which explains why our proposed dual training strategy could improve the model performance.

5. Conclusion

In this paper, we present the first attempt to consider answer-based visual question generation as a dual task of visual question answering and propose a generalizable dual training scheme, Invertible Question Answering Network (iQAN). The proposed method reconstructs VQA model to its dual VQG form thus we can train a single model jointly with two conjugate tasks. Experiments show that our dual trained model outperforms the prior art model on both VQA2 and CLEVR dataset. We further show the proposed dual training scheme can be applied to some other popular VQA models and brings consistent gains.

Acknowledgement

This work is supported by Hong Kong Ph.D. Fellowship Scheme, SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK419412, CUHK14203015, CUHK14207814, CUHK14208417, CUHK14202217, and CUHK14239816), the Hong Kong Innovation and Technology Support Programme (No.ITS/121/15FX), and in part by the China Postdoctoral Science Foundation under Grant 2014M552339. We also thank Jianfan Han, Xihui Liu, Hang Zhou, and Shaoshuai Shi for helpful discussions.

References

[1] H. Ali, Y. Chali, and S. A. Hasan. Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010. 2

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. 2015. 1, 2, 10

[3] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017. 2, 3, 4, 6, 7, 8, 10

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 10

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016. 2

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 2, 6, 7, 8, 10

[7] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma. Dual learning for machine translation. In *NIPS*, 2016. 2, 11

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[9] U. Jain, Z. Zhang, and A. Schwing. Creativity: Generating diverse questions using variational autoencoders. *arXiv preprint arXiv:1704.03493*, 2017. 2

[10] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*, 2017. 2, 6, 7, 8

[11] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015. 2

[12] S. Kalady, A. Elikkottil, and R. Das. Natural language question generation using syntax and keywords. In *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010. 2

[13] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*, 2017. 2, 6, 8, 10

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[15] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015. 7

[16] A. Lavie and A. Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *ACLW*, 2007. 6

[17] Y. Li, W. Ouyang, X. Wang, and X. Tang. Vip-cnn: Visual phrase guided convolutional neural network. *CVPR*, 2017. 1

[18] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 1

[19] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACLW*, 2004. 6

[20] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 2

[21] P. Lu, H. Li, Z. Wei, J. Wang, and X. Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *AAAI*, 2018. 1

[22] M. Malinowski and M. Fritz. Towards a visual turing challenge. *NIPS workshop*, 2014. 2

[23] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[24] C. D. Manning, H. Schütze, et al. *Foundations of statistical natural language processing*. MIT Press, 1999. 1

[25] J. H. Martin and D. Jurafsky. Speech and language processing. 2000. 1

[26] I. M. Mora, S. P. de la Puente, and X. Giro-i Nieto. Towards automatic generation of question answer pairs from images. *CVPRW*, 2016. 2

[27] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016. 1, 2

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[29] I. V. Serban, A. García-Durán, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, and Y. Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016. 2

[30] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016. 2

[31] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010. 2

[32] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000. 12

[33] D. Tang, N. Duan, T. Qin, and M. Zhou. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*, 2017. 2

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 6

[36] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level attention networks for visual question answering. In *CVPR*, 2017. 1

[37] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang. Automatic generation of grounded visual questions. *arXiv preprint arXiv:1612.06530*, 2016. 1, 2

[38] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 1, 6, 8, 10

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 2, 11