

iVQA: Inverse Visual Question Answering

Feng Liu¹ Tao Xiang² Timothy M. Hospedales³ Wankou Yang¹ Changyin Sun¹

¹Southeast University, China

²Queen Mary University of London, UK ³University of Edinburgh, UK

{liufeng,wkyang,cysun}@seu.edu.cn, t.xiang@qmul.ac.uk, t.hospedales@ed.ac.uk

Abstract

We propose the inverse problem of Visual question answering (iVQA), and explore its suitability as a benchmark for visuo-linguistic understanding. The iVQA task is to generate a question that corresponds to a given image and answer pair. Since the answers are less informative than the questions, and the questions have less learnable bias, an iVQA model needs to better understand the image to be successful than a VQA model. We pose question generation as a multi-modal dynamic inference process and propose an iVQA model that can gradually adjust its focus of attention guided by both a partially generated question and the answer. For evaluation, apart from existing linguistic metrics, we propose a new ranking metric. This metric compares the ground truth question's rank among a list of distractors, which allows the drawbacks of different algorithms and sources of error to be studied. Experimental results show that our model can generate diverse, grammatically correct and content correlated questions that match the given answer.

1. Introduction

As conventional object detection and recognition approach solved problems, we see a surge of interest in more challenging problems that should require greater ‘understanding’ from computer vision systems. Image captioning [31], visual question answering [2], natural language object retrieval [20] and ‘visual Turing tests’ [11] provide multi-modal AI challenges that are expected to require rich visual and linguistic understanding, as well as knowledge representation and reasoning capabilities. As interest in these grand challenges has grown, so has scrutiny of the benchmarks and models that appear to solve them. Are we making progress towards these challenges, or are good results the latest incarnation of horses [29, 30] and Potemkin villages [12], with neural networks finding unexpected correlates that provide shortcuts to give away the answer?

Recent analyses of VQA models and benchmarks have found that the reported VQA success is largely due to making predictions from dataset biases and cues given away in

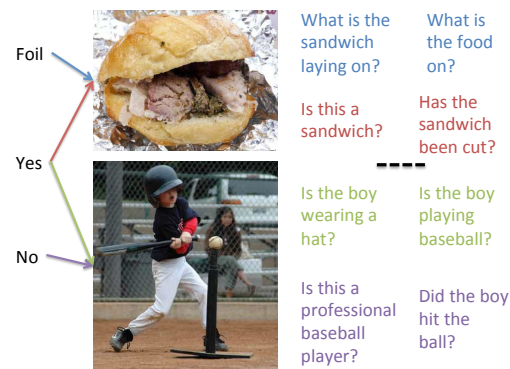


Figure 1. Illustration of iVQA task: Input answers and images along with the top questions generated by our model.

the question, with predictions being minimally dependent on understanding image content. For example it turns out that existing VQA models do not ‘look’ in the same places as humans do to answer the question [6]; they do not give different answers when the same question is asked of different images [1]; and they can perform well given no image at all [2, 17]. Moreover, VQA model predictions do not depend on more than the first few words of the question [1], and their success depends largely on being able to exploit label bias [13]. These observations have motivated renewed attempts to devise more rigorous VQA benchmarks [13].

In this paper we take a different approach, and explore whether the task of inverse VQA provides an interesting benchmark of multi-modal intelligence. The inverse VQA (iVQA) task is to input a pair of image and answer, and then ask (output) a suitable question for which the given answer holds in the context of the given image. We conjecture that iVQA, as illustrated in Fig. 1, is an interesting challenge for several reasons: (i) There may be less scope for an iVQA model to take advantage of question bias than for VQA to score highly through answer bias (there is less question bias, and exploiting it is harder than for categorical answers). (ii) The answers themselves provide a very sparse cue in iVQA compared to questions in VQA. So there may be less opportunity to deduce the question from the answer alone in

iVQA than there is to deduce the answer from the question alone in VQA. Thus the iVQA task relies more heavily on understanding image content. (iii) From a knowledge representation and reasoning perspective, iVQA may provide the opportunity to test more complex inference strategies such as counterfactual reasoning [4].

Although closely related to VQA, existing VQA models do not provide a solution to the iVQA problem. This is because much less information can be inferred from an answer than from a question. In addition, although an answer is often short consisting of a phrase or even a single word, an iVQA model-generated question is a complete sentence composed of a long sequence of words. The key to effective iVQA is thus to attend selectively and dynamically to different regions of the image as the model progresses to generate the next word. This dynamic attention mechanism has to be conditioned on both the answer and the partial sentence generated so far. To this end, a novel dynamic multi-modal attention-based iVQA model is proposed which is capable of generating diverse, grammatically correct and content correlated questions that match the given answer.

Prior evaluations of question generation methods mainly use standard machine translation metrics, *e.g.*, BLEU, METEOR, *etc.* These automatic metrics are correlated with human judgements for question generation [28]. However they provide limited power to *diagnose* question generation models in terms of when and why they succeed or fail. In this paper, we first propose an alternative and complementary ranking-based evaluation metric which is based on ranking the ground truth question among alternative distractors using an iVQA model, given the image and the answer. By controlling the types of distractors presented when using this metric, we can better understand the successes and failures of different models. Second, we perform a human study which is robust to iVQA’s one-to-many nature (multiple possible questions can have the same answer). Reassuringly our human study scores turn out to be highly correlated to our proposed new ranking metric.

The contributions of this paper are as follows: (1) The novel iVQA problem is introduced as an alternative challenge for high-level multi-modal visuo-linguistic understanding. (2) We propose a multi-modal dynamic attention based iVQA model. (3) We propose a question ranking based evaluation methodology for iVQA that is helpful to diagnose the strengths and weaknesses of different models. (4) As the dual problem of VQA, we show that iVQA has the potential to help improve VQA performance.

2. Related work

Image captioning Image captioning [31, 19] aims to describe, rather than merely recognise objects in images. It encompasses a number of classic vision capabilities as prerequisites including object [15] and action [35] recognition,

attribute description [9] and relationship inference [24]. It further requires natural language generation capabilities to synthesise open-ended linguistic descriptions. Popular benchmarks and competitions have inspired intensive research in this area. Captioning models have explicitly addressed these sub-tasks to varying degrees [19], but the most common and successful approaches use neural encoders (of images), and decoders (of captions), with little explicit knowledge representation and reasoning [31, 23]. The iVQA task investigated here is related to captioning in that we aim to produce natural language outputs, but distinct in that the outputs are sharply conditioned on the required answer, as illustrated in Fig. 1.

VQA Challenge Like captioning, VQA has gained attention as a synthesis challenge in AI, requiring both computer vision and natural language understanding to succeed [2]. Based on an image, and a natural language question about the image, a VQA system produces an answer. Unlike other vision tasks (recognition, detection, description), the question to be answered in VQA is dynamically specified at runtime. Besides visuo-linguistic grounding, many VQA examples seem to require extra information not contained in the question or image, *e.g.*, background common sense about the world. Thus VQA is hoped to provide a long term goal for AI-complete multi-modal intelligence. However increasing scrutiny has shown that learning systems excel at finding shortcuts in terms of gaming the biases in answer distributions, and giveaway correlations [1, 17], leading to doubts about the level of visuo-linguistic intelligence implied by current results [6]. Although some benchmarks in principle require open-sentence answers, most answers are simple one-word outputs, and therefore the most common approach has been to formalise answer generation as a multi-class classification problem over the most frequent answers [17, 10]. Although successful, this is somewhat unsatisfactory as it is no longer an open-world challenge. In this paper we explore a novel iVQA task as an alternative open-world benchmark for visuo-linguistic understanding.

VQA Models Existing VQA models are commonly based on two-branch neural networks, each consisting of a CNN image encoder, a LSTM question encoder which are merged before feeding to an answer decoder [2, 26]. Recently they have been enhanced through various mechanisms including better visuo-linguistic merging [21, 10], varying degrees of explicit representation [3], reasoning with external knowledge bases [32], and improving visual encoding through attention [10, 33]. With most recent models treating answer generation as a classification problem, these models cannot be directly modified for iVQA by simply swapping the answer and question encoder/decoder. The proposed iVQA model is a marriage between captioning and VQA models but with a dynamic and multi-modal attention mechanism developed specifically for iVQA.

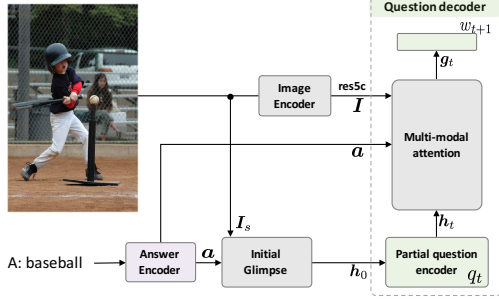


Figure 2. Overall architecture of the proposed iVQA model

Related Challenges Our proposed challenge is related to emerging task of *visual question generation* (VQG): to generate a natural question about the content of an image [28]. Introduced in [28], VQG is further studied in [36] where DenseCap [18] is used to generate region specific descriptions before being translated into questions. VQG is a pre-specified task unlike VQA and iVQA which are dynamically determined at runtime. Importantly, VQG is easier in terms of required understanding. Since VQG is not required to be answer-conditional, it often generates very general open questions that even humans cannot answer. It does not need to understand the image clearly enough – and ground the two domains richly enough – to correctly condition the generated question on the answer. Another relevant challenge is *visually grounded conversation* (VGC) which aims to generate natural-sounding conversations [7, 27]. VGC typically starts with VQG, but the following responses and further questions are generated primarily following conversational patterns mined from social media text, and only loosely grounded on the context of image content. In contrast, the image content grounding is much tighter in iVQA.

3. Methodology

3.1. Problem formulation

The problem of inverse visual question answering (iVQA) is to infer a question q for which a given answer a holds, in the context of a particular image I . Formally:

$$q^* = \max_q p(q|I, a; \Theta), \quad (1)$$

where q is a sentence with words (w_1, w_2, \dots, w_n) and Θ is the model parameters. As a sequence generation problem, we can use a recurrent neural network language model, to generate the sentence by maximising the likelihood:

$$q^* = \max_q \prod_t p(w_t | w_{t-1}, \dots, w_1, I, a). \quad (2)$$

Since the task is conditioned on both image I and answer a , the visual information has to be integrated with the answers appropriately to generate questions.

3.2. Model overview

The architecture of our iVQA model is shown in Fig. 2. It is a deep neural network with three subnets: an image encoder, an answer encoder, and a question decoder. The two encoders provide inputs for the decoder to generate a sentence which fits to the conditioned answer and image content. A multi-modal attention module (detailed later) is also a key component that directs image attention dynamically given the outputs of both encoders and a partial question encoder. We first describe the three subnets.

The image encoder is a CNN that generates a feature representation of the image. Both global and local features are exploited for image representation. The `res5c` features computed using the ResNet-152 model [15] are utilised as local features. More specifically, The local feature collection $I = \{v_{ij}\}$ is defined as local feature $v_{ij} \in \mathbb{R}^{2048}$ over all 14×14 spatial locations. To extract the local features, we resize the image to 448×448 before feeding it to the feature extractor as in [10]. As for the global feature, the semantic concept [23] feature $I_s \in \mathbb{R}^{1000}$ is used. These 1,000 semantic concepts are mined from the most frequent words in a set of image captions. A concept classifier is learned to predict I_s as classification scores for the concepts.

For the answer encoder, a long-short memory (LSTM) network with 512 cells [14] is used, and the concatenation of the final hidden state and cell state provides the answer representation $a \in \mathbb{R}^{1024}$. With the described CNN image encoder and LSTM based answer encoder as input, a LSTM question decoder can be used to generate questions conditioned on both the images and answer. The detailed decoding processing together with the proposed attention module will be detailed next.

3.3. Dynamic multi-modal attention

Given the sparse information contained in the answer, having an effective attention model to focus on the right region of the image is critical for iVQA. Attention models have been widely studied in image captioning [25] and VQA [33]. However, our iVQA problem has some unique characteristics, and thus needs a tailor-made attention module. Specifically, compared with VQA, iVQA requires multiple decoding steps and the focus of attention therefore needs to be dynamically changed accordingly. Also unlike image captioning, the generation process has multi-modal conditions: *i.e.*, image and answer, both of which need to be integrated in every decoding step in a dynamic manner. Consider the following question-answer pair: “Q: What colour is the dress the girl is wearing?”; “A: Pink”. Given the answer, the model can infer that the question is about colour. After the model has predicted type specific partial question $q_t = \{\text{what, colour, is, the}\}$, the attention network will integrate the partial question q_t with the answer $a = \{\text{pink}\}$, and search for all objects with the pink colour and output attended features. Based on these attend-

ed features the next word w_{t+1} is predicted. Motivated by these unique characteristics, we propose an attention model that can perform inference based on image, partial question and answer jointly and dynamically. It is composed of the the following sub-modules:

Initial glimpse The initial glimpse should provide an overview cue of the input image-answer pair, to establish a good starting point for the decoding process. We use semantic concept prediction \mathbf{I}_s as a global visual cue, which captures 1-gram information that may be relevant to the question [23]. The encoded answer \mathbf{a} is taken as the textual cue, which determines the set of likely initial words of the target question. The two cues are integrated as

$$\mathbf{h}_0 = \delta(\mathbf{W}_{ih}\mathbf{I}_s + \mathbf{W}_{ah}\mathbf{a}), \quad (3)$$

where \mathbf{W}_{ih} and \mathbf{W}_{ah} are embedding weights¹, and $\delta(\cdot)$ is a tanh activation function. This joint representation is directly used as the initial memory of the decoding network.

Encoding of partial question The partial question encoder sequentially encodes the partial question generated thus far $q_t = \{w_1, w_2, \dots, w_t\}$ to a hidden representation \mathbf{h}_t . A LSTM network with 512 cells is used to encode the partial question to a hidden representation \mathbf{h}_t as

$$\begin{aligned} \mathbf{x}_t &= \mathbf{E}w_t, \\ \mathbf{h}_t, \mathbf{m}_t &= \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1}), \end{aligned} \quad (4)$$

where w_t is the one-hot coding of word w_t ; \mathbf{E} is the word embedding matrix; \mathbf{x}_t is the embedded word vector which serves as an input to the LSTM. $\text{LSTM}(\cdot)$ takes previous states $(\mathbf{h}_{t-1}, \mathbf{m}_{t-1})$ and \mathbf{x}_t as input to generate the next states. For the computation of the LSTM, readers are referred to [16] for details.

Multi-modal attention network The attention network takes local features \mathbf{I} , partial question coding \mathbf{h}_t , and answer coding \mathbf{a} as input, and outputs the joint embedding of attended visual features \mathbf{c}_t specified by the partial question-answer context \mathbf{z}_t . To obtain the partial question-answer context, the partial question coding \mathbf{h}_t and answer coding \mathbf{a} are fused as

$$\mathbf{z}_t = \text{ReLU}(\mathbf{W}_q\mathbf{h}_t + \mathbf{W}_a\mathbf{a}) \quad (5)$$

where \mathbf{W}_q and \mathbf{W}_a are the embedding weights for question and answer respectively.

Then the visual features \mathbf{I} and textual context \mathbf{z}_t are spatially matched via soft attention: The visual features $\mathbf{v}_{i,j}$ and context vector \mathbf{z}_t are fused by a multi-modal low rank bilinear pooling (MLB) [21], and then the fused feature \mathbf{f}_{ij}

is used for attention map computation as follows:

$$\begin{aligned} \mathbf{f}_{ij} &= \delta(\mathbf{U}\delta(\mathbf{W}_v\mathbf{v}_{i,j}) \odot \delta(\mathbf{W}_z\mathbf{z}_t)) \\ \alpha_{ij}^t &= \text{softmax}(\mathbf{p}^T[\mathbf{f}_{ij}]) \\ \mathbf{c}_t &= \sum_{ij} \alpha_{ij}^t \mathbf{v}_{ij}, \end{aligned} \quad (6)$$

where \mathbf{c}_t is the attended visual feature; $\alpha_t = [\alpha_{ij}^t]$ is the attention map; and \mathbf{U} , \mathbf{W}_v , \mathbf{W}_z , \mathbf{p} are the corresponding embedding weights.

The attended visual feature \mathbf{c}_t is further fused with the textual context \mathbf{z}_t via MLB, which can be interpreted as co-attention between vision and language [37].

$$\mathbf{g}_t = \delta(\mathbf{U}'\delta(\mathbf{W}_c\mathbf{c}_t) \odot \delta(\mathbf{W}_z'\mathbf{z}_t)), \quad (7)$$

where \mathbf{U}' , \mathbf{W}_c , \mathbf{W}_z' are embedding weights of the pooling.

Word predictor The next-word predictor is a softmax classifier, which generates a distribution over the next words, leveraging the multi-modal attention network's output \mathbf{g}_t :

$$w_{t+1} \sim \text{softmax}(\mathbf{W}_o\mathbf{g}_t), \quad (8)$$

where \mathbf{W}_o are the classifier weights. The next word w_{t+1} is sampled from the softmax classifier's distribution.

4. iVQA Evaluation

We explore three of iVQA evaluation metrics including standard language-generation metrics, a new ranking-based metric, and a human validation study.

Linguistic Metrics Standard linguistic measures [5] including CIDEr, BLEU, METEOR and ROGUE-L can be used to evaluate the generated questions. Given an image, question, answer tuple, we use the ground-truth question as the reference sentence, and compare the generated question based on the given image and answer. The similarity between the machine generated questions and the reference questions can be measured by these metrics. Even though generating humanlike questions is relatively easy, doing so in a correct image+answer conditional way to get a high score is challenging, since the model has to capture all the semantic concepts and high-order interactions.

Ranking Metric We also develop a ranking based evaluation metric for the iVQA problem. For an image-answer pair (I, a) , and a candidate question q , the conditioning score $p(q|I, a; \Theta)$ is used for ranking. If one of the correct (ground truth) questions is ranked the highest then this image-answer pair is regard as correct at Rank-1. In this way, accuracy over a testing set can be computed as the percentage of the times that correct questions are ranked at the top (denoted Acc.@1). Similarly, we can measure cumulative ranking accuracy at other ranks, e.g., Acc.@3 measures the percentage of times correct questions are ranked in top

¹In all equations, we omit the bias term for simplicity.

3. This is related to the multiple-choice setting of VQA [2]. However, the difference is that we can explore specific choices of candidate question subsets to form a question pool in order to reveal insights about model strengths and weaknesses.

Question Pool For a particular image-answer pair, the candidate ranking questions are collected from the following subsets. **Correct questions (GT)**: given image-answer pair (I, a) , the correct (ground truth) questions are defined as all the questions with answer a in image I ². **Contrastive questions (CT)**: these are questions associated with visually similar images to I (including I) but having different answers. The similarity of the images is measured using the image CNN feature. **Plausible questions (PS)**: These test whether the model can tell the subtle difference between questions and maintain grammar correctness. They are obtained by randomly replacing one of the key words (e.g., verbs, nouns, adjective, and adverb) in the ground truth question. **Popular questions (PP)**: Popular questions are chosen to be the most popular questions with the same answer type as a across the whole dataset. These diagnose the extent to which the model is relying on label-bias. **Answer-related (RN)**: These are chosen to be the random questions having answer a but from other images. These diagnose the extent to which the model is relying on visual features, which did not always happen in VQA [13]. We manually checked all generated distractor questions, and removed any which were also correct for their corresponding image answer.

Human study iVQA is open-ended and one-to-many in that there can be many correct questions for one image and answer. Therefore, given an image and an answer, the correct questions may not be annotated exhaustively in existing datasets. The proposed ranking metric computed on the selected question pool ameliorates the effects of the open-ended/one-to-many questions generated by an iVQA model. However, it does not measure directly how ‘correct’ the generated questions are, when they differ from the human annotations originally provided. To evaluate iVQA in a way that awards ‘credit’ for correct questions that are not annotated originally, we perform a human evaluation study. Image-answer pairs are randomly selected from the test set, and annotators assess the generated questions, scoring them from 0 (complete nonsense) to 4 (perfect). The mean score is used as the metric.

5. Experiments

5.1. Datasets and settings

Dataset: We repurpose the VQA dataset [2] to investigate the iVQA task. The VQA dataset uses images from M-

²There can be multiple correct questions corresponding to a given answer, since multiple questions may have the same answer for one image.

S COCO [5], including 82,783 training, 40,504 validation and 81,434 test images. Three question-answer pairs are collected for each image. Since the test set answers are not available, we adopt the commonly used off-line data split in [19, 25] for image captioning: 82,783 images are used for training, and 5,000 for validation and test each.

Training: The model is trained by minimising the cross entropy loss between machine generated questions and ground truth questions. The Adam [22] optimiser is employed with a batch size 32 for 30 epochs. The initial learning rate is set to be $5e-4$, and it is annealed 0.83 times per epoch with an exponential decay.

Inference: Beam search with max sentence length 20 is used to generate questions. We use beam size 3 for quantitative and 10 for qualitative results.

Question Pool: For the proposed ranking accuracy metric, given each image-answer pair, the question pool contains 24 questions, of which 1-3 are GT, 3-5 are CT (so that the total of GT+CT is 6), 6 are PP, 6 are PS, and 6 are RN.

5.2. Baseline models

Answer only (A): It uses a LSTM encoder to encode tokenised answers to a fixed 512-dimensional representation, then a LSTM decoder to generate questions.

Image only (I, VQG): The visual only model is similar to the GRNN model in [28], however, we use a more powerful image feature: the same `res5c` feature of ResNet-152 [15] used in our model. This feature is fed into a LSTM decoder as the initial state.

Image+Answer Type (I+AT): VQG models [28] generate questions purely based on visual cues. To make VQG more competitive in our answer-conditional iVQA setting, we also provide one-hot encoding of the answer *type*. This hint helps a VQG model generate the right question type (e.g., ‘is...’, ‘what...’).

NN: We adapt the nearest neighbour (NN) image captioning method [8] to our problem. As iVQA is conditioned on both image and answer, we averaged the distance computed from both modalities for NN computation.

SAT: Show attend and tell (SAT) [34] is a strong attentional captioning method. To provide a strong competitor to our approach, we modify SAT to take input from both modalities by setting the initial state of the decoding LSTM as the joint embedding of image and answer.

VQG+VQA: The VQG+VQA baseline uses the VQG model above to generate question proposals from the image, and then uses VQA to select the question with maximum conditioning score. We use VQG to generate 10 candidates for each image for VQA re-ranking, and the retrained multi-modal low-rank bilinear attention network [21] is used as the VQA model.

Ours: Our model processes images with local and global semantic features, and dynamic multi-modal attention

(I+A+Att+ I_s). The global semantic feature is obtained following [23] by learning a concept predictor on the training split using the 1,000 most frequent caption words.

5.3. Results

Overall In the first experiment we report the overall iVQA performance on the test split. The results are shown in Table 1 with both the standard linguistic metrics, as well as our ranking accuracy metric. From the table, we can make the following observations: (i) Unlike VQA the margin between the no-image case (A), and the full model (Ours) is dramatic. The ranking accuracies are more than doubled, and the language metrics show similarly striking improvements. This demonstrates that unlike conventional VQA [13], the ‘V’ does matter in iVQA. (ii) The margin between the image + answer type (I+AT) and image-only (I) setting exists, but is less significant. This shows that while it is a useful hint for an iVQA model to know the question type, it still really needs the actual semantic answer to generate the right questions. E.g., rather than just knowing that it was counting something (answer type), the model does need to know *how many objects were counted* (answer) in order to generate the right question specifying what object type needs to be counted – as there may be other objects that could be counted. (iii) The margins between (I) and (I+AT) and the full model are also striking. This demonstrates that as a test of multi-modal intelligence, iVQA reassuringly requires both modalities in order to do well. (iv) VQG+VQA indeed performs better than the vanilla VQG model by making the generated question more answer conditional, but it is still weaker than the captioning adapted models (NN and SAT) or our proposed model. The reason is that the VQG model has too low *sensitivity*, thus is unable to provide the right question candidates for VQA model to select. Also, VQA model can be easily fooled by similar questions not relevant to the image because VQA models are trained to distinguish different answers rather than questions. (v) The captioning adapted models (NN and SAT) perform well, but are still inferior to the proposed model which is specifically designed for iVQA.

Human study The human study is applied on a subset of 1000 samples, and evaluates the models in a way that is robust to open-ended question generation. Results in Table 1 show that (i) our model performs the best among all competitors; (ii) the human study scores are highly correlated with the proposed ranking metric. Specifically, the Pearson correlation coefficient between manually labelled scores and the proposed acc@1 and acc@3 metrics are 0.917 and 0.981 separately, while the best performing linguistic measure (CIDEr) can only reach 0.898, which demonstrates the effectiveness of the proposed ranking metric. Since human evaluation is expensive, the proposed ranking metric is a reasonable and cost-effective alternative.

Qualitative Results Examples of questions generated by our model are shown in Fig. 3. The results illustrate a few interesting points: (i) The generated questions are highly conditional on both images and answers. Particularly, the same answers generate different questions for different images, unlike the situation in VQA [1]; and the same images generate very different questions when paired with different answers, showing richer reasoning than in VQG [28]. (ii) Unlike VQA, there are multiple reasonable questions that correspond to one image-answer pair. This is both due to alternative phrasing of the same question (*‘where is the bear?’*, *‘where is the teddy bear?’*, *‘what is the teddy bear sitting on?’*), as well as multiple semantically distinct questions having the same answer (e.g., *‘are the children eating?’*, *‘is the child wearing a shirt?’*, *‘is the child wearing a hat?’*). Since the annotation is not exhaustive, the standard linguistic metrics could be misleading: the generated questions can be correct but just have never been asked by the annotators. Our proposed ranking metric is more robust to this, as models are only scored according to how plausibly they rate the true question, rather than whether their open-world estimate of the question matches annotated ground-truth. Our human study evaluates the methods in a way that credits open-world question generation. The open-ended question generation formulation of our model means it is straightforward to *sample* the distribution over questions given images and answers, in order to explore the model’s beliefs.

5.4. Further analysis

Analysis by Failure Type The proposed new evaluation metric enables us to understand the mistakes each model makes. The results in Fig. 4 show the Rank-1 predictions of each model (highest scoring question) broken down according to the category of that prediction. It again shows the much superior performance of our iVQA model (I+A+Att+ I_s): 32.67% of the top ranked prediction is correct, which almost doubles that of the VQA model (16.4%). But the main objective here is to analyse which distractor types are mistakenly ranked high: (i) Without being able to condition on the answers, the image-only method (I) makes predictions dominated by contrastive (CT) distractor questions taken from similar looking images. (ii) The answer-only (A) and image+answer type (I+AT) methods make predictions that are dominated by distractor questions of the popular (PP) type. This suggests that these models are trying to rely on unconditional distribution of label statistics in order to solve iVQA. (iii) The VQA-based baseline instead is dominated by answer-related (RN) distractor. This suggests that the VQA model fails to correctly take into account image context (the ‘V’ is not being accounted for [1, 13]), and is simply picking questions that generate the corresponding answer independently of the required conditioning on image context. (iv) The nearest neighbour (NN)



(GT) are the children playing?

A: Yes (GT)
Q:
are the children happy?
(-4.75)
are the children eating?
(-5.20)
are the children having
fun? (-5.54)
is there a child in the
picture? (-5.73)
are there any children in
the picture? (-5.88)

A: Purple
Q:
what color is the child 's
shirt? (-2.97)
what color is the girls shirt?
(-3.36)
what color is the girl 's
shirt? (-3.49)
what color is the kids shirt?
(-3.89)
what color is the little boy
's shirt? (-4.54)

A: Red
Q:
what color is the child 's
shirt? (-3.72)
what color is the boy 's
shirt? (-4.31)
what color is the child 's
hat? (-4.58)
what color is the baby 's
shirt? (-4.69)
what color is the girl 's
shirt? (-4.71)

A: No
Q:
are the children eating?
(-5.44)
are these children in the
wild? (-6.87)
is the child wearing a shirt?
(-7.01)
are these children in a
park? (-7.44)
is the child wearing a hat?
(-7.52)



(GT) what color is the couch
the bear is on?

A: Brown (GT)
Q:
what color is the bear?
(-2.39)
what color is the couch?
(-2.52)
what color is the teddy
bear? (-2.81)
what color is the table?
(-2.97)
what color is the
bedspread? (-3.47)

A: Bowtie
Q:
what is the bear wearing?
(-2.92)
what is the teddy bear
wearing? (-3.08)
what kind of bear is this?
(-4.20)
what is on the bear 's neck?
(-4.77)
what kind of tie is the bear
wearing? (-4.84)

A: Couch
Q:
what is the bear sitting on?
(-2.01)
where is the bear?
(-2.36)
where is the teddy bear?
(-2.61)
what is the teddy bear
sitting on? (-2.86)
where is the bear sitting?
(-2.87)

A: No
Q:
is the bear alive?
(-4.06)
is this a real bear?
(-4.09)
is the bear real?
(-4.45)
is the bear sleeping?
(-4.74)
is that a real bear?
(-4.79)

Figure 3. Qualitative results of iVQA. Larger numbers in brackets mean higher confidence. The attention of generating the questions in purple is further visualised in Fig. 5.

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	Acc@1	Acc@3	Human
A	0.952	0.146	0.192	0.265	0.371	0.408	0.161	14.589	28.795	1.92
I	0.652	0.086	0.121	0.179	0.280	0.310	0.117	13.012	28.644	2.04
I+AT	0.904	0.122	0.164	0.234	0.350	0.397	0.151	20.277	36.134	2.65
NN	1.372	0.175	0.223	0.294	0.404	0.428	0.183	26.783	48.755	3.01
SAT	1.533	0.192	0.241	0.311	0.417	0.456	0.195	29.722	48.118	3.19
VQG+VQA	1.110	0.147	0.193	0.261	0.371	0.396	0.165	16.529	41.655	2.79
Ours	1.714	0.208	0.256	0.326	0.430	0.468	0.205	32.899	51.418	3.31

Table 1. Overall question generation performance on the testing set.

approach performs well and has an evenly distributed set of error types, but it is weaker than the proposed in correctly capturing the visual and answer conditions. It is reflected on the larger portion of CT and RN errors. (v) Our full model (I+A+Att+I_s) has the largest fraction of correct predictions and manage to suppress the plausible (PS) and contrastive (CT) errors. It still makes some mistakes of ranking the popular (PP) questions at the top, but still fewer than answer (A) and image+answer type (I+AT) alternatives.

Ablation study We evaluate the contributions of our key technical components: dynamic attention, and multi-modal inference with both local and global semantic image features. The results in Table 2 verify that each of these components contributes to the final result, and the biggest contribution comes from the proposed dynamic attention model.

How dynamic attention helps To illustrate our dynamic attention in iVQA, we visualise attention maps computed during question generation in Fig. 5. From the examples, we see that focus of attention varies over time in an

	CIDEr	BLEU-4	Acc@1	Acc@3
I+A	1.541	0.196	29.929	48.356
I+Att+A	1.698	0.208	32.547	51.024
I+Att+A+I _s	1.750	0.214	33.636	52.344

Table 2. Ablation study on the contributions of key model components. The results are obtained on the validation set.

appropriate way according to the partially generated question as well as the answer. For example, Fig. 5 shows that the model focuses on regions on and below the bear at the point of generating the words ‘on’ and ‘wearing’ respectively. These examples also demonstrate that the learned iVQA model has achieved some degree of multi-modal visuo-linguistic understanding and implicit reasoning capability.

Can iVQA help VQA? The problem of iVQA can be seen as a dual problem of VQA, so we investigate whether VQA performance can be boosted by applying iVQA to obtain a second opinion. Specifically, we conduct the experiments on the more challenging VQA 2.0 dataset and MLB-att [21] trained on the training set is employed as the

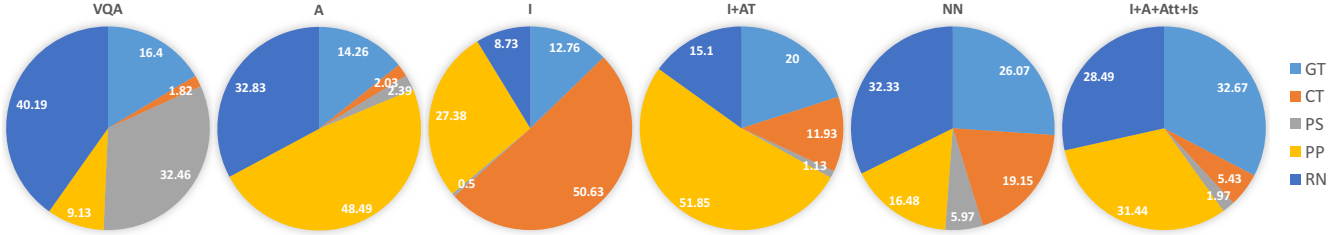


Figure 4. Comparison of effects of different distractors on different models on the test set.



Figure 5. Dynamic attention maps generated by the proposed model. Input answer: ‘Bowtie’ (top), ‘Couch’ (bottom). Because the conditioning answer is different the model generates totally different attention maps in producing the output question.

VQA model. During testing, the top 3 answers with the largest VQA score are served as answer candidates, then a second score is computed from the iVQA model. They are further combined by a score fusion network, whose output is utilised as the confidence of final prediction. Before the VQA-iVQA fusion, the VQA model alone can achieve a validation accuracy of 57.85, while after the final model reaches an accuracy of 58.86, where the performance gain is mainly from the challenging number type (improved from 34.94 to 38.71). These results thus show that iVQA can indeed assist VQA. Actually iVQA can also be used as a diagnosis tool to extract the belief set of an existing VQA model, which is part of the ongoing work.

Contrasting VQA and iVQA as benchmarks Finally, we discuss iVQA’s interest as a benchmark compared to the conventional VQA. Two of the main kinds of bias that a VQA/iVQA model could use to cheat the benchmark are the output *Prior bias* (Ignore both inputs and predict only the most likely answer on VQA; use question frequency in iVQA), and *Language bias* (ignore the image and use only the input – question for VQA, answer for iVQA – to predict the output). A good multimodal intelligence benchmark should require understanding and mutual grounding of both modalities, and should be hard to game by exploiting those biases. To analyses these issues we compare performance on iVQA and VQA benchmarks using Prior-alone and Language-alone (LSTM Q for VQA and Answer only for iVQA) baselines versus the full multi-modal model in

each case (DeeperLSTM+Norm I for VQA [2], and I+A model for iVQA). The results in Table 3 show that for VQA the bias-based baselines approach the performance of a full multi-modal model much more closely than the corresponding baselines do for iVQA. This suggests that VQA is easier to ‘game’ (achieve an apparently high score without any image understanding or multimodal grounding), compared to iVQA. Thus we propose that iVQA makes a distinct and interesting benchmark for multimodal intelligence.

	split	Prior	Language	Language+Visual
iVQA (acc@1)	test	3.94	14.59	28.44
VQA (accuracy)	test-dev	29.66	48.76	57.75

Table 3. VQA vs. iVQA on bias-based gameability.

6. Conclusion

We have introduced the novel task of inverse VQA as an alternative multi-modal visual intelligence challenge to the popular VQA paradigm. The analyses suggest that iVQA is appealing in terms of being less game-able via exploiting label-bias, more clearly requiring the mutual grounding and understanding of both visual and linguistic modalities, and naturally providing an open-world prediction setting.

Acknowledgements: This project received support from Natural Science Foundation of China (NSFC) grant #61773117, #61473086, #61520106009, #61533008, and #U1713209.

References

- [1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016. 1, 2, 6
- [2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. Vqa: Visual question answering. *IJCV*, 2016. 1, 2, 5, 8
- [3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016. 2
- [4] L. Bottou, J. Peters, J. Quiñero Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR*, 2013. 2
- [5] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 4, 5
- [6] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, 2016. 1, 2
- [7] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *CVPR*, 2017. 3
- [8] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015. 5
- [9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2
- [10] 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. In *EMNLP*, 2016. 2, 3
- [11] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *PNAS*, 2015. 1
- [12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1
- [13] 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. 1, 5, 6
- [14] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013. 3
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 5
- [16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 4
- [17] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016. 1, 2
- [18] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 3
- [19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 5
- [20] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1
- [21] 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, 4, 5, 7
- [22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [23] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. Semantic regularisation for recurrent image annotation. In *CVPR*, 2017. 2, 3, 4, 6
- [24] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 2
- [25] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 3, 5
- [26] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2
- [27] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP*, 2017. 3
- [28] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *ACL*, 2016. 2, 3, 5, 6
- [29] O. Pfungst. *Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*. H. Holt and company, New York, 1911. 1
- [30] B. L. Sturm. A simple method to determine if a music information retrieval system is a horse. *TMM*, 2014. 1
- [31] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2
- [32] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *TPAMI*, 2017. 2
- [33] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 2, 3
- [34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 5
- [35] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *TPAMI*, 2012. 2
- [36] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang. Automatic generation of grounded visual questions. In *IJCAI*, 2017. 3
- [37] L. Zhou, C. Xu, P. Koch, and J. J. Corso. Image caption generation with text-conditional semantic attention. *arXiv preprint arXiv:1606.04621*, 2016. 4