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
In recent years, visual question answering (VQA) has become topical as a long-term goal to drive computer vision and multi-disciplinary AI research. The premise of VQA’s significance, is that both the image and textual question need to be well understood and mutually grounded in order to infer the correct answer. However, current VQA models perhaps ‘understand’ less than initially hoped, and instead master the easier task of exploiting cues given away in the question and biases in the answer distribution (Goyal et al. 2017).

In this paper we propose the inverse problem of VQA (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. 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.

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 (Vinyals et al. 2015), visual question answering (Agrawal et al. 2016), natural language object retrieval (Kazemzadeh et al. 2014) and ‘visual Turing tests’ (Geman et al. 2015) 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 (Pfungst 1911) and Potemkin villages (Goodfellow, Shlens, and Szegedy 2015), with neural networks finding unexpected correlates that provide shortcuts to give away the answer?

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

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 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 (Das et al. 2016); they do not give different answers when the same question is asked of different images (Agrawal, Batra, and Parikh 2016); and they can perform well given no image at all (Agrawal et al. 2016; Jabri, Joulin, and van der Maaten 2016). Moreover, VQA model predictions do not depend on more than the first few words of the question (Agrawal, Batra, and Parikh 2016), and their success depends largely on being able to exploit label bias (Goyal et al. 2017). These observations have motivated renewed attempts to devise more rigorous VQA benchmarks (Goyal et al. 2017).

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 (Bottou et al. 2013).

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 (Mostafazadeh et al. 2016), and they enable evaluating unconstrained open-ended sentence generation. However they provide limited power to diagnose question generation models in terms of when and why they succeed or fail. In this paper, we 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.

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.

Related work

**Image captioning** Image captioning (Vinyals et al. 2015; Karpathy and Fei-Fei 2015) aims to describe, rather than merely recognise objects in images. It encompasses a number of classic vision capabilities as prerequisites including object recognition, attribute description (Farhadi et al. 2009) and relationship inference (Lu et al. 2016). 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 (Karpathy and Fei-Fei 2015), but the most common and successful approaches use neural encoders (of images), and decoders (of captions), with little explicit knowledge representation and reasoning (Vinyals et al. 2015; Liu et al. 2017). 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 (Agrawal et al. 2016). 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 (Agrawal, Batra, and Parikh 2016; Jabri, Joulin, and van der Maaten 2016), leading to doubts about the level of visuo-linguistic intelligence implied by current results (Das et al. 2016). 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 (Jabri, Joulin, and van der Maaten 2016; Fukui et al. 2016). 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 (Agrawal et al. 2016; Malinowski, Rohrbach, and Fritz 2015). Recently they have been enhanced through various mechanisms including better visuo-linguistic merging (Kim et al. 2017; Fukui et al. 2016), varying degrees of explicit representation (Andreas et al. 2016), reasoning with external knowledge bases (Wang et al. 2016), and improving visual encoding through attention (Fukui et al. 2016; Xu and Saenko 2016). 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.

Related Challenges

Our proposed challenge is related to the recently studied task of visual question generation (VQG): to generate a natural question that is related to the context of an image (Mostafazadeh et al. 2016). Introduced in (Mostafazadeh et al. 2016), VQG is further studied in (Zhang et al. 2016) where DenseCap (Johnson, Karpathy, and Fei-Fei 2016) is used to generate regional specific description 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 (Das et al. 2017). VGC typically starts with VQG, but the following responses and further questions are generated primarily following conversational patterns mined from social media textual data, and only loosely grounded on the image content as context. In contrast, the grounding on image content is much tighter in iVQA.

Methodology

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:

\[
q^* = \max_q p(q|I, a; \Theta),
\]

where \( q \) is a sentence with words \( (w_1, w_2, ..., w_n) \) and \( \Theta \) 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:

\[
q^* = \max_q \prod_t p(w_t|w_{t-1}, ..., w_1, I, a).
\]

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.

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 for the image content. Both global features and local features are exploited for image representation. The res5c features computed using the ResNet-152 model (He et al. 2016) 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 x 14 spatial locations. To extract the local features, we resize the image to 448 x 448 before feeding it to the feature extractor as in (Fukui et al. 2016). As for the global feature, the semantic concept (Liu et al. 2017) feature \( I_s \in \mathbb{R}^{1000} \) is used. These 1,000 semantic concepts are miner 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 (Graves, Mohamed, and Hinton 2013) 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.

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 (Lu et al. 2017) and VQA (Xu and Saenko 2016). 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
Attention network will integrate the partial question specific partial question \( q \) the attention network will integrate the partial question \( q \) with the answer \( a = \{\text{pink} \} \), and search for all objects with the pink colour and output attended features. Based on these attended 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 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 \( I_o \) as a global visual cue, which captures 1-gram information that may be relevant to the question \( (Liu \ et \ al. \ 2017) \). The encoded answer \( a \) is taken as the textual cue, which determines the set of likely initial words of the target question. The two cues are integrated as

\[
h_t = \delta(W_{ih} I_o + W_{ah} a),
\]

where \( W_{ih} \) and \( W_{ah} \) are embedding weights\(^1\) and \( \delta() \) 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 \( q_t = \{w_1, w_2, \ldots, w_t\} \) to a hidden representation \( h_t \). A LSTM network with 512 cells is used to encode the partial question to a hidden representation \( h_t \) as

\[
x_t = E w_t,
\]

\[
h_t, m_t = \text{LSTM}(x_t, h_{t-1}, m_{t-1}),
\]

where \( w_t \) is the one-hot coding of word \( w_t \); \( E \) is the word embedding matrix; \( x_t \) is the embedded word vector which serves as an input to the LSTM. LSTM\((\cdot)\) takes previous states \((h_{t-1}, m_{t-1})\) and \( x_t \) as input to generate the next states. For the computation of the LSTM, readers are referred to \( \text{Hochreiter \ and \ Schmidhuber} \ (1997) \) for details.

**Multi-modal attention network** The attention network takes local features \( I \), partial question coding \( h_t \), and answer coding \( a \) as input and outputs the joint embedding of attended visual features \( c_t \) specified by the partial question-answer context \( z_t \). To obtain the partial question-answer context, the partial question coding \( h_t \) and answer coding \( a \) are fused as

\[
z_t = \text{ReLU}(W_q h_t + W_a a)
\]

where \( W_q \) and \( W_a \) are the embedding weights for question and answer respectively.

Then the visual features \( I \) and textual context \( z_t \) are spatially matched via soft attention: The visual features \( w_{ij} \) and context vector \( z_t \) are fused by a multi-modal low rank bilinear pooling (MLB) \( (Kim \ et \ al. \ 2017) \), and then the fused feature \( f_{ij} \) is used for attention map computation as follows:

\[
f_{ij} = \delta(U\delta(W_v w_{ij}) \odot \delta(W_z z_t))
\]

\[
\alpha_{ij} = \text{softmax}(p^T[f_{ij}])
\]

\[
c_t = \sum_{ij} \alpha_{ij} v_{ij},
\]

where \( c_t \) is the attended visual feature; \( \alpha_{ij} = [\alpha_{ij}^T] \) is the attention map; and \( U, W_v, W_z, p \) are the corresponding embedding weights.

The attended visual feature \( c_t \) is further fused with the textual context \( z_t \) via MLB, which can be interpreted as co-attention between vision and language \( (Zhou \ et \ al. \ 2016) \).

\[
g_t = \delta(U'\delta(W_v c_t) \odot \delta(W_z' z_t)),
\]

where \( U', W_v', W_z' \) 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 \( g_t \):

\[
w_{t+1} \sim \text{softmax}(W_o g_t),
\]

where \( W_o \) are the classifier weights. The next word \( w_{t+1} \) is sampled from the softmax classifier’s distribution.

**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 \( \text{Chen} \ et \ al. \ (2015) \) 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 theses 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 3. This is related to the multiple-choice setting of VQA \( (Agrawal \ et \ al. \ 2016) \). 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 \[^2\] Contrastive questions (CT): these are questions associated with visually similar images to \( I \) (including \( I \)) but having different answers.

\(^1\) In all equations, we omit the bias term for simplicity.

\[^2\]There can be multiple correct questions corresponding to a given answer, since multiple questions may have the same answer for one image.
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 questions. **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 [Goyal et al. 2017]. We manually checked all generated distractor questions, and removed any which were also correct for their corresponding image answer pairs.

**Human study** 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 questions. **Popular questions (PP):** Popular questions are chosen to be the most popular questions with the same answer type 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 [Goyal et al. 2017]. We manually checked all generated distractor questions, and removed any which were also correct for their corresponding image answer pairs.

**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 [Mostafazadeh al. 2016], however, we use a more powerful image feature: the same res5c feature of ResNet-152 [He et al. 2016] used in our model. This feature is fed into a LSTM decoder as the initial state.

**Image+Answer Type (I+AT):** VQG models [Mostafazadeh al. 2016] 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 [Devlin et al. 2015] 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) [Xu et al. 2015] 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 \( p(a | q^*, I) \). The VQG for question proposal is necessary because there must be sufficiently large number of proposals to cover the correct questions. As it is an open-ended problem, exhaustive ranking is infeasible. We use VQG to generate 10 candidates for each image for VQA re-ranking. The retrained multi-modal low-rank bilinear attention network [Kim et al. 2017] 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+Ix). The global semantic feature is obtained following [Liu et al. 2017] by learning a concept predictor on the training split with the 1,000 most frequent caption words as concepts.

**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 [Goyal et al. 2017], the ‘V’ does matter in iVQA. (ii) The margin between the image + answer type (I+AT) and image-only (I) setting exists, but is not too 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 the (I) and (I+AT) cases 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) The captioning models adapted to iVQA perform well, but underperform our purpose-designed approach.

**Human study** The human study is applied on a subset of 300 samples, and evaluates the models in a way that is robust to open-ended question generation. Results in Table 1 show.

![Figure 3: Qualitative results of iVQA. Larger numbers in brackets mean higher confidence. The attention of generating the questions in purple is further visualized in Fig. 4.](image)

![Figure 4: Dynamic attention maps generated by the proposed model. Input answer: ‘Bowtie’ (top), ‘Couch’ (bottom).](image)

| Model       | 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 | 2.00   |
| I           | 0.652  | 0.086  | 0.121  | 0.179  | 0.280  | 0.310   | 0.117  | 13.012 | 28.644 | 2.10   |
| I+AT        | 0.904  | 0.122  | 0.164  | 0.234  | 0.350  | 0.397   | 0.151  | 20.277 | 36.134 | 2.70   |
| NN          | 1.372  | 0.175  | 0.223  | 0.294  | 0.404  | 0.428   | 0.183  | 26.783 | 47.955 | 3.11   |
| SAT         | 1.533  | 0.192  | 0.241  | 0.311  | 0.417  | 0.456   | 0.195  | 29.722 | 48.118 | 3.30   |
| VQG+VQA     | 1.110  | 0.147  | 0.193  | 0.261  | 0.371  | 0.396   | 0.165  | 16.529 | 41.655 | 2.85   |
| Ours        | 1.714  | 0.208  | 0.256  | 0.326  | 0.430  | 0.468   | 0.205  | 32.899 | 51.418 | 3.39   |

Table 1: Overall question generation performance on the testing set.
How dynamic attention helps
To illustrate our dynamic attention model, the contribution comes from the proposed dynamic attention model. 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 [Agrawal, Batra, and Parikh 2016], and the same images generate very different questions when paired with different answers, showing richer reasoning than in VQG (Mostafazadeh et al. 2016). (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 the true question, rather than whether their open-world settings, then a second score \( p(q|I, a^*) \) is computed from the iVQA model. The two scores \( p(a^*|I, q) \) and \( p(q|I, a^*) \) 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, and more principled ways of exploiting this dual relationship are part of our future work.

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. It is also amenable to diagnosis of model behaviour via selection of distractors in a question-ranking task, and via sampling the distribution of predicted questions. In the future we will investigate jointly solving VQA and iVQA in terms of making coherent attentional fixations, and question/answer predictions.

References
[Agrawal, Batra, and Parikh 2016] Agrawal, A.; Batra, D.; and Parikh, D. 2016. Analyzing the behavior of visual question answering models. In EMNLP.
[Agrawal et al. 2016] Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2016. Vqa: Visual question answering. IJCV.
[Andreas et al. 2016] Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In CVPR.
[Bottou et al. 2013] Bottou, L.; Peters, J.; Quinonero Candela, J.; Charles, D. X.; Chickering, D. M.; Portugalay, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. J. Mach. Learn. Res. 14(1):3207–3260.
