# Being Negative but Constructively: Lessons Learnt from Creating Better Visual Question Answering Datasets

**Wei-Lun Chao**[*], **Hexiang Hu**[*], and **Fei Sha**
University of Southern California
Los Angeles, California, USA
weilunc@usc.edu, hexiang.frank.hu@gmail.com, feisha@usc.edu

## Abstract

Visual question answering (QA) has attracted a lot of attention lately, seen essentially as a form of (visual) Turing test that artificial intelligence should strive to excel. In this paper, we study a crucial component for this task: how can we design good datasets for the task? We focus on the design of multiple-choice based datasets where the learner has to select the right answer from a set of candidate ones including the target and the decoys. Through careful analysis of the results given by state-of-the-art learning models and human annotators on existing datasets, we show the design of the decoy answers has a significant impact on how and what the learning models learn from the datasets. In particular, the resulting learner can ignore the visual information, the question, or the both while still doing well on the task. Inspired by this, we propose automatic procedures of how to remedy such design deficiencies. We apply the procedures to re-construct decoy answers for two popular visual QA datasets as well as to create a new visual QA dataset from the Visual Genome project, resulting in the largest dataset for this task. Extensive empirical studies show that the design deficiencies have been alleviated in the remedied datasets and the performance on them is likely a more faithful indicator of the difference among learning models. The datasets will be released and publicly available via http://www.teds.usc.edu/website_vqa/.

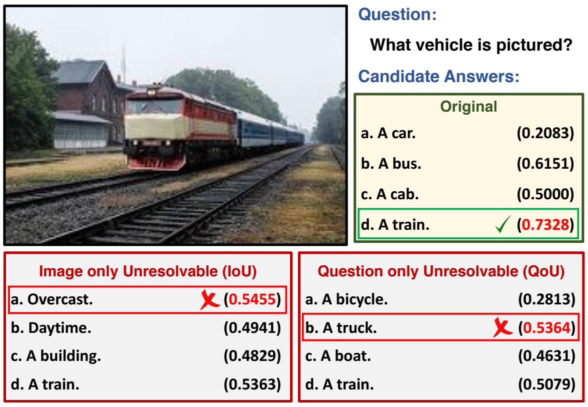


Figure 1: An illustration of how the shortcuts in the Visual7W dataset (Zhu et al., 2016) should be remedied. In the original dataset, the correct answer "A train" is easily selected by a machine as it is far often used as the correct answer than other decoy answers. (The numbers in the brackets are probability scores computed using eq.(2)). Our two procedures — IoU and QoU — create alternative decoys such that both the correct answer and the decoys are highly likely by examining either the image or the question **alone**. In these cases, machines make mistakes unless they consider all information **together**. Thus, the alternative decoys suggested our procedures are better designed to gauge how well a learning algorithm can understand all information equally well.

## 1 Introduction

Recently, multimodal information processing tasks such as image captioning (Xu et al., 2015) and visual question answering (visual QA) (Antol et al., 2015) have gained a lot of attention. A number of significant advances in learning algorithms have bee made, along with the development of nearly two dozens of datasets in this very active research domain. Among them, popular ones include MSCOCO (Lin et al., 2014; Chen et al., 2015), Visual Genome (Krishna et al., 2016), VQA (Antol et al., 2015), and several others. The overarching objective is that a learning machine needs to go beyond understanding differ-

[*] Equal contributions

ent modalities of information separately (such as image recognition alone) and to learn how to correlate them in order to perform well on those tasks.

To evaluate the progress on those complex and more AI-like tasks is however a challenging topic. For tasks involving language generation, developing an automatic evaluation metric is itself an open problem (Anderson et al., 2016; Kilickaya et al., 2016; Liu et al., 2016; Kafle and Kanan, 2016). Thus, many efforts have concentrated on tasks such as *multiple-choice* visual QA (Antol et al., 2015; Zhu et al., 2016; Jabri et al., 2016) or selecting the best caption (Hodosh et al., 2013; Hodosh and Hockenmaier, 2016; Ding et al., 2016; Lin and Parikh, 2016), where the accuracy of selection is a natural evaluation metric.

In this paper, we study how to design high-quality multiple choices for the visual QA task. In this task, the machine (or the human annotator) is presented with an image, a question and a list of candidate answers. The goal is to select the correct answer through a consistent understanding of the image, the question and each of the candidate answers. As in any multiple-choice based tests (such as GRE), designing what should be presented as negative answers — we refer them as *decoys or negatives* — is as important as deciding the questions to ask. We all have had the experience of exploiting the elimination strategy: *This question is too easy — none of the three answers could be right so the remaining one must be correct!*

While a clever strategy for taking exams, such "shortcuts" prevent us from studying faithfully how different learning algorithms grasp the meanings in images and languages (eg, the quality of the embeddings of both images and languages in the right semantic spaces). It has been noted that machines can achieve very high accuracies of selecting the correct answer without the visual input (ie, the image), the question, or neither (Jabri et al., 2016; Antol et al., 2015). Clearly, the learning algorithms have overfit on incidental statistics in the datasets. For instance, if the decoy answers have infrequently been used as correct answers (to any question), then the machine can rule out a decoy answer with a binary classifier that determines whether the answers are in the set of the correct answers — note that this classifier does not need to examine the image and it just needs to memorizes the list of the correct answers in the training dataset. See Fig. 1 for an example, and Sec. 3 for more and detailed analysis.

We focus on minimizing the impacts of exploiting such shortcuts. We suggest a set of principles for creating decoy answers. In light of the amount of human efforts in curating existing datasets for the visual QA task, we propose two procedures that revise those datasets such that the decoy answers are better designed. In contrast to some earlier works, the procedures are fully automatic and do not incur additional human annotator efforts. We apply the procedures to revise both Visual7W (Zhu et al., 2016) and VQA (Antol et al., 2015). Additionally, we create a multiple-choice based dataset from the recently released Visual Genome dataset (Krishna et al., 2016), resulting in the largest multiple-choice dataset for the visual QA task, with more than one million image-question-candidate answers triplets.

We conduct extensive empirical and human studies to demonstrate the effectiveness of our procedures in creating high-quality datasets for the visual QA task. In particular, we show that machines need to use all three information (image, questions and answers) to perform well — any missing information induces a large drop in performance. Furthermore, we show that humans dominate machines in the task. However, given the revised datasets are likely reflecting the true gap in human and machine understanding of multimodal information, we expect that advances in learning algorithms likely focus more on the task itself instead of overfitting to the idiosyncrasies in the datasets.

The rest of the paper is organized as follows. In Sect. 2, we describe related work. In Sect. 3, we analyze and discuss the design deficiencies in existing datasets. In Sect. 4, we describe our automatic procedures for remedying those deficiencies. In Sect. 5 we conduct experiments and analysis. We conclude the paper in Sect. 6.

## 2 Related Work

(Kafle and Kanan, 2016; Wu et al., 2016) provide recent overviews of the status quo of the visual QA task. There are about two dozens of datasets for the task. Most of them use real images, while some are based on synthetic ones. Usually, for each image, multiple questions and their corresponding answers are generated. This can be achieved either by human annotators, or with an automatic procedure that uses captions or question templates and detailed annotations such as objects.

We concentrate on 3 datasets VQA (Antol et al., 2015), Visual7W (Zhu et al., 2016), and Visual Genome (Krishna et al., 2016). All of them use images from MSCOCO (Lin et al., 2014).

Besides the pairs of questions and correct answers, VQA, Visual7W, and visual Madlibs (Yu et al., 2015) provide decoy answers for each pair so that the task can be evaluated in multiple-choice selection accuracy. *What decoy answers to use* is the focus of our work.

In VQA, the decoys consist of human-generated plausible answers as well as high-frequency and random answers from the datasets. In Visual7W, the decoys are all human-generated plausible ones. Note that, humans generate those decoys by *only looking at the questions and the correct answers but* ***not*** *the images*. Thus, the decoys might be unrelated to the corresponding images. A learning algorithm can potentially examine the image alone and be able to identify the correct answer.

In visual Madlibs, the questions are generated with a limited set of question templates ("fill-in-the-blank") and the detailed annotations (eg, objects) of the images. Thus, similarly, a learning model can examine the image alone and deduce the correct answer.

We propose automatic procedures to revise VQA and Visual7W (and to create one based on Visual Genome) such that the decoy generations are carefully considered to prevent learning algorithms from exploiting the shortcuts in the datasets by overfitting on incident statistics. In particular, our design goal is that a learning machine needs to understand all the 3 components of an image-question-answers triplet in order to make the right choice — ignoring either one or two components will result in drastic degradation in performance.

Our work is inspired by the experiments in (Jabri et al., 2016) where they observe that machines without looking at images or questions can still perform well on the visual QA task. Others have also reported similar issues (Goyal et al., 2016; Zhang et al., 2016; Johnson et al., 2016; Agrawal et al., 2016) but are not for the multiple-choice setting. Our work extends theirs by providing more detailed analysis *as well as automatic procedures* to remedy those design deficiencies.

Besides visual QA, (Ding et al., 2016) and VisDial (Das et al., 2016) also propose automatic ways to generate decoys for the tasks of selecting the best visual caption and dialog, respectively.

| Information used | Machine | Human |
|---|---|---|
| random | 25.0 | 25.0 |
| A | 52.9 | - |
| I + A | 62.4 | 75.3 |
| Q + A | 58.2 | 36.4 |
| I + Q + A | 65.7 | 88.4 |

Table 1: Accuracy of selecting the right answers out of 4 choices (%) on the visual QA task on Visual7W-test.

## 3 Analysis of Decoy Answers' Effects

In this section, we examine in detail on the dataset Visual7W (Zhu et al., 2016), a popular choice for the visual QA task. We demonstrate how the deficiencies in designing decoy questions impact the performance of learning algorithms.

In multiple-choice visual QA datasets, a training or test example is a triplet that consists of an image I, a question Q, and a candidate answer set A. The set A contains a target T (the correct answer) and $K$ decoys (incorrect answers) denoted by D. An IQA triplet is thus $\{\mathrm{I}, \mathrm{Q}, \mathrm{A} = \{\mathrm{T}, \mathrm{D}_1, \cdots, \mathrm{D}_K\}\}$. We use C to denote either the target or a decoy.

### 3.1 Visual QA models

We investigate how well a learning algorithm can perform when supplied with different type of modalities of information. We concentrate on the one hidden-layer MLP model proposed in (Jabri et al., 2016), which had achieved state-of-the-art results on the dataset Visual7W. The model computes a scoring function $f(c, i)$

$$f(c, i) = \sigma(\boldsymbol{U} \max(0, \boldsymbol{W} g(c, i)) + b) \qquad (1)$$

over a candidate answer $c$ and the multimodal information $i$, where $g$ is the joint feature of $(c, i)$ and $\sigma(x) = 1/(1 + exp(-x))$. The information can be the image (I) alone, the question (Q) alone, or the combination of the both (I+Q).

We use the penultimate layer of ResNet-200 (He et al., 2016) as visual features to represent I and the average WORD2VEC embeddings (Mikolov et al., 2013) as text features to represent Q and C. To form the joint feature $g(c, i)$, we just concatenate the features together. The candidate that has the highest confidence in prediction is selected as the model output.

We use the standard training, validation and test splits of Visual7W, where each contains 69,817, 28,020, and 42,031 examples respectively. Each question has 4 candidate answers. Details are in Sect. 5 and the Supplementary Material.

### 3.2 Analysis results

**Machines find shortcuts** Table 1 summarizes the performance of the learning models, together with the human studies we perform on a subset of 1,000 triplets (c.f. Sect. 5 for details). There are a few interesting observations.

First, in the row of "A" where only the candidate answers (and whether they are right or wrong) are used to train a learning model, the model performs significantly better than random guess (52.9% vs. 25%) — humans will deem each of the answers equally likely *without* looking at either the image or the question! Clearly, the model learns the specific statistics of the candidate answers in the dataset and exploits those.

Adding the information about the image (ie, the row of "I+A"), the machine improves significantly and gets close to the performance when all information is used (62.4% vs. 65.7%). There is a weaker correlation between the question and the answers as "Q+A" improves over "A" only modestly. This is expected. In the Visual7W dataset, the decoys are generated by human annotators as plausible answers to the questions without being shown the images — thus, many decoy answers do not have visual groundings. For instance, a question of "what animal is running?" elicits equally likely answers such as "dog", "tiger", "lion", or "cat", while an image of a dog running in the park will immediately rule out all 3 but the "dog", see Fig. 1 for similar examples. Thus, the performance of "I+A" implies that many IQA triplets can be solved by object, attribute or concept detection on the image, without understanding the questions. This is indeed the case also for humans — humans can achieve 75.3% by considering "I+A" and not "Q". Note that the difference between machine and human on "I+A" are likely due to the difference in understanding visual information between the two.

Note that human improves significantly from "I+A" to "I+Q+A" with "Q" added, while the machine does so only marginally. The difference can be attributed to the difference in understanding the question and correlating with the answers between the two. Since each image corresponds to multiple questions or have multiple objects, solely relying on the image itself will not work well in principle. Such difference clearly indicates that in the visual QA model, the language component is weak.

**Shortcuts are due to design deficiencies** We probe deeper on how the decoy answers have impacted the performance of learning models.

As explained above, the decoy answers are drawn from all plausible answers to a question, irrespective whether they are visually grounded or not. We have also discovered that the targets (ie, correct answers) are infrequently used as decoys.

Specifically, among the 69,817 training samples, there are 19,503 unique correct answers and each one of them is used about 3.6 times as correct answers to a question. However, among all the $69,817 \times 3 \approx 210K$ decoys, each correct answer appears 7.2 times on average, far below a chance level of 10.7 times ($210K \div 19,503 \approx 10.7$). This disparity exists in the test samples too. Consequently, the following rule, computing each answer's likelihood being correct or decoy,

$$P(\text{correct}|\text{C}) = \begin{cases} 0.5, & \text{if C is never seen in training,} \\ \frac{\text{\# times C as Target}}{\text{\# times C as Target}+(\text{\# times C as Decoys})/K}, & \text{otherwise,} \end{cases} \tag{2}$$

should perform well. Indeed, it attains an accuracy of 48.73% on the test data, far better than random guess and is close to the learning model using the answers information only (the "A" row in Table 1).

**Good rules for designing decoys** Based on our analysis, we summarize the following guidance rules to design decoys: (1) **Question only Unresolvable (QoU)**. The decoys need to be equally plausible to the question. Otherwise, machines can rely on the correlation between the question and candidate answers to tell the target from decoys, even without the images. Note that this is a principle that is being followed by most datasets. (2) **Neutrality**. The decoys answers should be equally likely used as the correct answers. (3) **Image only Unresolvable (IoU)**. The decoys need to be plausible to the image. That is, their contents should appear in the image, or there exist questions so that the decoys can be treated as targets to the image. Otherwise, visual QA can be resolved by objects, attributes, or concepts detection in images, even without questions.

Ideally, each decoy in an IQA triplet should meet the three principles. **Neutrality** is comparably easier to achieve by *reusing terms in the whole set of targets as decoys*. On the contrary, a decoy may hardly meet **QoU** and **IoU** simultaneously[1]. However, as long as all decoys of an IQA triplet meet **Neutrality** and some meet **QoU** and others meet **IoU**, the triplet as a whole still achieves the three principles — a machine ignoring images or questions will likely perform poorly.

## 4 Create Better Visual QA Datasets

In this section, we describe our approaches of remedying design deficiencies in the existing datasets for the visual QA task. We introduce two automatic procedures to create new decoy answers that can prevent learning models from exploiting incident statistics in the datasets.

### 4.1 Methods

**Main Ideas** Our procedures operate on a dataset that already contains image-question-target (IQT) triplets, ie, we do not assume it has decoys already. For instance, we have used our procedures to create a multiple-choice dataset from the Visual Genome dataset which has no decoy. We assume that each image in the dataset is coupled with "multiple" QT pairs, which is the case in nearly all the existing datasets. Given an IQT triplet (I, Q, T), we create two sets of decoy answers,

- **QoU-decoys**. We search among all other triplets that have similar questions to Q. The targets of those triplets are then collected as the decoys for T. As the targets to similar questions are likely plausible for the question Q, QoU-decoys likely follow the rules of **Neutrality** and **Question only Unresolvable (QoU)**. We compute the average WORD2VEC (Mikolov et al., 2013) to represent a question, and use the cos similarity to measure the similarity between questions.

- **IoU-decoys**. We collect the targets from other triplets of the *same* image to be the decoys for T. The resulting decoys thus definitely follow the rules of **Neutrality** and **Image only Unresolvable (IoU)**.

We then combine the triplet (I, Q, T) with QoU-decoys and IoU-decoys to form an IQA triplet as a training or test sample.

**Resolving ambiguous decoys** One potential drawback of automatically selected decoys is that they may be semantically similar, ambiguous, or rephrased terms to the target (Zhu et al., 2016). We utilize two filtering steps to alleviate it. First, we perform string matching between a decoy and the target, deleting those decoys that contain or are covered by the target (e.g., "daytime" vs "during the daytime" and "ponytail" vs "pony tail").

Secondly, we utilize the WordNet hierarchy and the Wu-Palmer (WUP) score (Wu and Palmer, 1994) to eliminate semantically similar decoys. The WUP score measures how similar two *word senses* are (in the range of $[0, 1]$), based on the depth of their two word senses in the taxonomy and that of their least common subsumer. We compute the similarity of two strings according to the WUP scores in a similar manner to (Malinowski and Fritz, 2014), in which the WUP score is used for the evaluation of visual QA performance. We eliminate decoys that have higher WUP-based similarity to the target. We use NLTK toolkit (Bird et al., 2009) to compute the similarity. See the Supplementary Material for more details.

**Other details** For QoU-decoys, we sort and keep for each triplet the top $N$ (eg, 10,000) similar triplets from the entire dataset according to the question similarity. Then for each triplet, we compute the WUP-based similarity of each potential decoy to the target successively, and accept those with similarity below 0.9 until we reach $K$ decoys. We also perform such a check among selected decoys to ensure their diversity. For IoU-decoys, the potential decoys are sorted randomly. The WUP-based similarity with a threshold of 0.9 is then applied to remove ambiguous decoys.

### 4.2 Comparison to other datasets

Several authors have noticed the design deficiencies in the existing databases and have proposed "fixes" (Antol et al., 2015; Yu et al., 2015; Zhu et al., 2016; Das et al., 2016). No dataset has used a procedure to generate IoU-decoys. We empirically show that how the IoU-decoys significantly remedy the design deficiencies for the decoys in the datasets.

Several previous efforts have generated decoys that are similar in spirit to our **QoU**-decoys. (Yu et al., 2015; Das et al., 2016; Ding et al., 2016) automatically find decoys from similar questions or captions based on question templates and an-

[1] E.g., in Fig 1, for the question "What vehicle is pictured?", the only answer that meets both principles is "train", which is the correct answer instead of being a decoy.

notated objects, tri-grams and GloVe embeddings (Pennington et al., 2014), and paragraph vectors (Le and Mikolov, 2014) and linguistic surface similarity, respectively. The later two are for different tasks from visual QA, and only (Ding et al., 2016) considers removing semantically ambiguous decoys like ours. (Antol et al., 2015; Zhu et al., 2016) ask humans to create decoys, given the questions and targets. As shown previously, such decoys may fail the rule of **Neutrality**.

## 5 Empirical Studies

### 5.1 Dataset

We examine our automatic procedures of creating decoys on the following three datasets. Table 2 summarizes their characteristics.

**VQA Real (Antol et al., 2015)** The dataset uses images from MSCOCO (Lin et al., 2014) under the same training/validation/testing splits to construct IQA triplets. Totally 614,163 IQA triplets are generated for 204,721 images. Each question has 18 candidate answers: in general 3 decoys are human-generated, 4 are randomly sampled, and 10 are randomly sampled frequent-occurring targets. *As the test set does not indicate the targets, our studies focus on the training and validation sets.*

**Visual7W Telling (Visual7W) (Zhu et al., 2016)** The dataset uses 47,300 images from MSCOCO (Lin et al., 2014) and contains 139,868 IQA triplets. Each has 3 decoys generated by humans.

**Visual Genome (VG) (Krishna et al., 2016)** The dataset uses 101,174 images from MSCOCO (Lin et al., 2014) and contains 1,445,322 IQT triplets. No decoys are provided. Human annotators are asked to write diverse pairs of questions and answers freely about an image or with respect to some regions of it. On average an image is coupled with 14 question-answer pairs. We divide the dataset into non-overlapping 50%/20%/30% for training/validation /testing. Additionally, we partition such that each portion is a “superset” of the corresponding one in Visual7W, respectively.

**Creating decoys** We create 3 QoU-decoys and 3 IoU-decoys for every IQT triplet in each dataset, following the steps in Sect. 4.1. In the cases we cannot find 3 decoys, we include random ones from the original set of decoys for VQA and Visual7W; for VG, we randomly include those from the top 10 frequently-occurring targets.

| Dataset | # of Images | | | # of triplets | | | # of decoys |
|---|---|---|---|---|---|---|---|
| Name | train | val | test | train | val | test | per triplet |
| VQA | 83k | 41k | 81k | 248k | 121k | 244k | 17 |
| Visual7W | 14k | 5k | 8k | 69k | 28k | 42k | 3 |
| VG | 49k | 19k | 29k | 727k | 283k | 433k | - |

Table 2: Summary of visual QA datasets

### 5.2 Setup

**Visual QA models** We utilize the MLP models mentioned in Sect. 3 for all the experiments. We denote **MLP-A**, **MLP-QA**, **MLP-IA**, **MLP-IQA** as the models using A (Answers only), Q+A (Question plus Answers), I+A (Image plus Answers), and I+Q+A (Image, Question and Answers), respectively. The hidden-layer has 8,192 neurons. We use a 200-layer ResNet (He et al., 2016) to compute visual features which are 2,048-dimensional. The ResNet is pre-trained on ImageNet (Russakovsky et al., 2015). The word2vec feature (Mikolov et al., 2013) for questions and answers are 300-dimensional, pre-trained on Google News. We use stochastic gradient descent with mini-batch size of 100, momentum of 0.9, and the stepped learning rate policy to optimize the MLP. We tune the number of iterations and the step size using the validation set. Details are in the Supplementary Material.

**Evaluation Metric** For Visual7W and VG, we compute the accuracy of picking the target from multiple choices. For VQA, we follow its protocol by comparing the picked answer to 10 human-generated targets. The accuracy is computed based on the number of exactly matched targets (divided by 3 and clipped at 1).

**Decoy sets to compare** For each dataset, we can thus derive several variants: (1) **Orig**: the original sets of decoys from the datasets, (2) **QoU**: **Orig** replaced with ones generated by our generated QoU-decoys, (3) **IoU**: **Orig** replaced with our generated IoU-decoys (4) **QoU+IoU**: **Orig** replaced with ones combining **QoU** and **IoU** (5) **All**: combining **Orig**, **QoU**, and **IoU**.

**User studies** Automatic decoy generation may lead to ambiguous decoys as mentioned in Sect. 4 and (Zhu et al., 2016). We thus conduct a user study via Amazon Mechanic Turk (AMT) to test humans’ performance on the datasets after remedied by our automatic procedures. We select for each dataset 1,000 IQA triplets. Each triplet is answered by three workers and in total 169 workers get involved. We report the average human performance and compare it to the learning models'. See the Supplementary Material for details.

### 5.3 Results

The performances of learning models and humans on the 3 datasets are reported in Table 3, 4, and 5.

**Effectiveness of new decoys** A better set of decoys will force learning models to integrate all 3 pieces of information — images, questions and answers — to make the correct selection from multiple-choices. In particular, they should prevent learning algorithms to exploit shortcuts such that partial information is sufficient for performing well on the visual QA task.

Table 3 clearly indicates that those goals have been achieved. With the Orig decoys, the relatively small gain from MLP-IA to MLP-IQA suggests that the question information can be ignored to attain good performance. However, with the IoU-decoys which require questions to help to resolve (as image itself is unresolvable), the gain is substantial (from 27.3% to 84.1%). Likewise, with the QoU-decoys (question itself is not adequate to resolve), including images information improves from MLIP-QA (40.7%) substantially to MLP-IQA's 57.6%. Note that with the Orig decoys, this gain is smaller (58.2% vs 65.7%).

It is expected that MLP-IA matches better QoU-decoys but not IoU-decoys, and MLP-QA is other way around. Thus it is natural to combine these two decoys. What is particularly appealing is that MLP-IQA improves noticeably over models learned with partial information, on the combined IoU+QoU-decoys (and "All" decoys). Furthermore, using answer information only (MLP-A) attains about chance-level accuracy.

On the VQA dataset (Table 4), the same observations hold, though to a lesser degree. On any of the IoU or QoU columns, we observe substantial gains when the complementary information is added to the model (such as MLP-IA to MLP-IQA). All these improvements are much more visible than those observed on the original decoy sets.

Combining both Table 3 and 4, we notice that the improvements from MLP-QA to MLP-IQA tend to be lower when facing IoU-decoys. This is also expected as it is difficult to have decoys that are simultaneously both IoU and QoU — such answers tend to be the target answers. Nonetheless, we deem this as a future direction to explore.

| Method | Orig | IoU | QoU | IoU+QoU | All |
|---|---|---|---|---|---|
| MLP-A | 52.9 | 27.0 | 34.1 | 17.7 | 15.6 |
| MLP-IA | 62.4 | 27.3 | 55.0 | 23.6 | 22.2 |
| MLP-QA | 58.2 | 84.1 | 40.7 | 37.8 | 31.9 |
| MLP-IQA | 65.7 | 84.1 | 57.6 | 52.0 | 45.1 |
| Human | 88.4 | - | - | 84.1 | - |
| Random | 25.0 | 25.0 | 25.0 | 14.3 | 10.0 |

Table 3: Accuracy (%) on Visual7W-test set.

| Method | Orig | IoU | QoU | IoU+QoU | All |
|---|---|---|---|---|---|
| MLP-A | 31.2 | 39.9 | 45.7 | 31.2 | 27.4 |
| MLP-IA | 42.0 | 39.8 | 55.1 | 34.1 | 28.7 |
| MLP-QA | 58.0 | 84.7 | 55.1 | 54.4 | 50.0 |
| MLP-IQA | 64.6 | 85.2 | 65.4 | 63.7 | 58.9 |
| Human | 88.5* | - | - | 89.0 | - |
| Random | 5.6 | 25.0 | 25.0 | 14.3 | 4.2 |

*:taken from (Antol et al., 2015)

Table 4: Accuracy (%) on VQA-val2014 set

**Differences in datasets** Contrasting Visual7W to VQA (on the column IoU+QoU), we notice that Visual7W tends to have bigger improvements in general. This is due to the fact that VQA has many questions with "Yes" or "No" as the targets — the only valid decoy to the target Yes is No, and vice versa. As such decoys are already captured by **Orig** of VQA (Yes and No are both top frequency targets), adding other decoy answers will not make any improvement. In Supplementary Material, however, we show that once we remove such questions/answers pairs, the range of improvements increased.

For completeness, we include the results on the Visual Genome dataset in Table 5. This dataset has no "Orig" decoys, and we have created a multiple-choice based dataset **qaVG** from it for the task — it has over 1 million triplets, the largest dataset on this task to our knowledge.

With **qaVG**, we also investigate whether it is possible to use it to improve the performances on the other two datasets — note that the images in both VQA and Visual7W are derived from MSCOCO.

We use the MLP-IQA trained on qaVG with both IoU and QoU decoys. This model initializes the models for the Visual7W and VQA datasets. We report the accuracies before and after fine-tuning, together with the best results learned solely from those two datasets respectively. As shown in Table 6, fine-tuning improves the performance on those datasets. In particular, the result on the original Visual7W (the row with "Orig") attains the state-of-the-art — previously the best performance on this dataset was reported as 68.5% by (Jabri et al., 2016) where a model pre-trained on

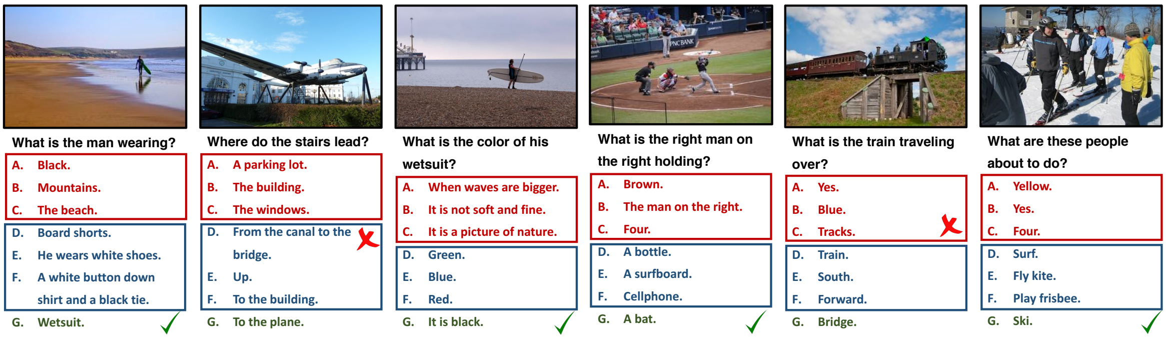


Figure 2: Example image-question-target triplets from Visual7W, VQA, and VG, together with our IoU-decoys (A, B, C.) and QoU-decoys (D, E, F). G is the target. Machine's selections are denoted by green ticks (correctly answered) or red crosses (wrongly answered).

| Method | Orig | IoU | QoU | IoU+QoU |
|---|---|---|---|---|
| MLP-A | - | 29.1 | 36.2 | 19.5 |
| MLP-IA | - | 29.5 | 60.2 | 25.2 |
| MLP-QA | - | 89.3 | 45.6 | 43.9 |
| MLP-IQA | - | 89.2 | 64.3 | 58.5 |
| Human | - | - | - | 82.5 |
| Random | - | 25.0 | 25.0 | 14.3 |

Table 5: Accuracy (%) on qaVG-test set.

| Datasets | Decoys | Prior | qaVG model | |
|---|---|---|---|---|
| | | best | initial | fine-tuned |
| Visual7W | Orig | 65.7 | 60.5 | **69.1** |
| | IoU+QoU | 52.0 | 58.1 | **58.7** |
| | All | 45.1 | 48.9 | **51.0** |
| VQA | Orig | 64.6 | 42.2 | **65.6** |
| | IoU+QoU | 63.7 | 47.9 | **64.1** |
| | All | 58.9 | 37.5 | **59.4** |

Table 6: Model trained on qaVG-train to improve Visual7W-test and VQA-val2014 (Accuracy in %).

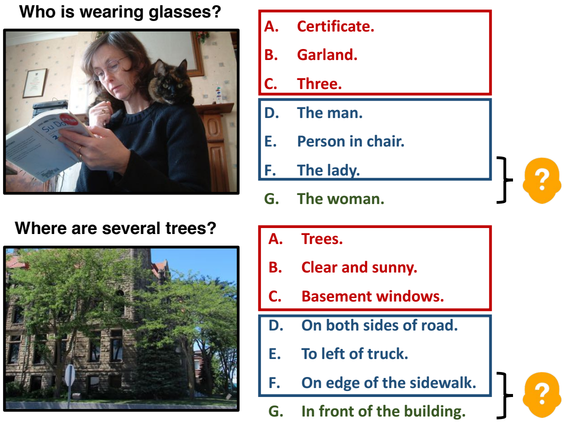


Figure 3: Ambiguous examples by our IoU-decoys (A, B, C) and QoU-decoys (D, E, F). G is the target. Ambiguous decoys F are marked.

VQA is fine-tuned on Visual7W.

### 5.4 Qualitative Results

We present in Fig. 2 example image-question-target triplets from **V7W**, **VQA**, and **VG**, together with our IoU-decoys (A, B, C) and QoU-decoys (D, E, F). G is the target. The predictions by the corresponding MLP-IQA are also included. Ignoring information from images or questions makes it extremely challenging to answer the triplet correctly, even for humans.

Our automatic procedures do fail at some triplets, resulting in ambiguous decoys to the targets. See Fig. 3 for example. We categorized those failure cases into two situations.

- Our filtering steps in Sect. 4 fail. The WUP-based similarity relies on the WordNet hierarchy. For some semantically similar words like "lady" and "woman", the similarity is only 0.632, much lower than that of 0.857 between "cat" and "dog". This issue can be alleviated by considering alternative semantic measures by WORD2VEC or by those used in (Das et al., 2016; Ding et al., 2016) for searching similar questions.

- The question itself is ambiguous to answer. In the bottom example in Fig. 3, both candidates D and F seem valid as a target. Another representative case is when asked about the background of a image. In images that contain sky and mountains in the distance, both terms can valid.

## 6 Conclusion

We perform detailed analysis on existing datasets for multiple-choice visual QA. We found that the design of decoys can inadvertently offers "shortcuts" for machines to exploit to perform well on the task. We describe several principles of constructing good decoys and present automatic procedures to remedy existing datasets. We also created a new dataset by applying our procedures to the Visual Genome, resulting in the largest multiple-choice dataset for the task with over 1 million image-question-candidate answers triplets. We conduct extensive empirical studies to demonstrate the effectiveness of our methods in creating better visual QA datasets. The remedied datasets and the Visual Genome based dataset will be released and publicly available via `http://www.teds.usc.edu/website_vqa/`.

## References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *EMNLP* .

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, pages 382–398.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. pages 2425–2433.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *arXiv preprint arXiv:1611.08669* .

Nan Ding, Sebastian Goodman, Fei Sha, and Radu Soricut. 2016. Understanding image and text simultaneously: a dual vision-language machine comprehension task. *arXiv preprint arXiv:1612.07833* .

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837* .

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. pages 770–778.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *ACL Workshop*. volume 3.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47:853–899.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *ECCV*. Springer, pages 727–739.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890* .

Kushal Kafle and Christopher Kanan. 2016. Visual question answering: Datasets, algorithms, and future challenges. *CVPR* .

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2016. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600* .

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* .

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. volume 14, pages 1188–1196.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, pages 740–755.

Xiao Lin and Devi Parikh. 2016. Leveraging visual question answering for image-caption ranking. In *ECCV*. Springer, pages 261–277.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *EMNLP* .

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*. pages 1682–1690.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910* .

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *ACL*. Association for Computational Linguistics, pages 133–138.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. volume 14, pages 77–81.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*. pages 2461–2469.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*. pages 5014–5022.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *CVPR*. pages 4995–5004.

# Supplementary Material

In this Supplementary Material, we provide details omitted in the main text.

- Sect. A: Details on the MLP-based models (Sect. 3.1 and 5.2 of the main text).
- Sect. B: WUPS-based similarity for filtering out ambiguous decoys (Sect. 4.1 of the main text).
- Sect. C: Details on user studies (Sect. 5.2 of the main text)
- Sect. D: Detailed results on VQA (Antol et al., 2015) w/o question-answer (QA) pairs that have Yes/No as the targets (Sect. 5.3 of the main text).

## A Details on the MLP models

The one hidden-layer MLP model used in our experiments has 8,192 hidden units, exactly following (Jabri et al., 2016). It contains a batch normalization layer before ReLU, and a dropout layer after ReLU. We set the dropout rate to be 0.5.

The input to the model is the concatenated features of images, questions, and answers. We change all characters to lowercases and all integer numbers within $[0, 10]$ to words before computing WORD2VEC. We perform $\ell_2$ normalization to features of each information before concatenation.

We train the model using stochastic gradient descent with mini-batch size of 100, momentum of 0.9, and the stepped learning rate policy: the learning rate is divided by 10 after every $M$ iterations (mini-batches). We set the initial learning rate to be 0.01 (we further consider 0.001 for the case of fine-tuning in Sect. 6.2 of the main text). We tune $M$ as a hyper-parameter of training. We validate the hyper-parameters on the validation set. For each model, we train with at most 600,000 iterations and select the best model using validation set.

Within each mini-batch, we sample 100 IQA triplets. For each triplet, we randomly choose to use QoU-decoys or IoU-decoys when training on IoU+QoU, or QoU-decoys or IoU-decoys or Orig when training on All. We then take the target and **3** decoys for each triplet to train the binary classifier. Specifically on VQA, which has 17 Orig decoys for A triplet, we randomly choose 3 decoys out of them. That is, 100 triplets in the mini-batch corresponds to 400 examples with binary labels. This procedure is to prevent *unbalanced training*, where machines simply learn to predict the dominant label, as suggested in (Jabri et al., 2016).

We note that in all the experiments in the main text, we use the *same decoys set for training and testing*.

## B WUP-based similarity for filtering out ambiguous decoys

We use the Wu-Palmer (WUP) score (Wu and Palmer, 1994), which characterizes the *word sense* similarity, to filter out ambiguous decoys to the target (correct answer). The WUP score is computed based on the WordNet hierarchy. Essentially, it measures the similarity of two *nodes* (ie, synsets) in the hierarchy. As a *word* might correspond to multiple nodes, we measure the word similarity as follows:

$$WUP(w_1, w_2) = \max_{(n_1,n_2)\in N_1\times N_2} WUP(n_1, n_2), \tag{3}$$

where $N1$ and $N2$ are the sets of nodes that words $w_1$ and $w_2$ correspond to, respectively. That is, the word similarity is based on the most similar pair of nodes from both words. We consider only the NOUN and ADJ nodes for tractable computation.

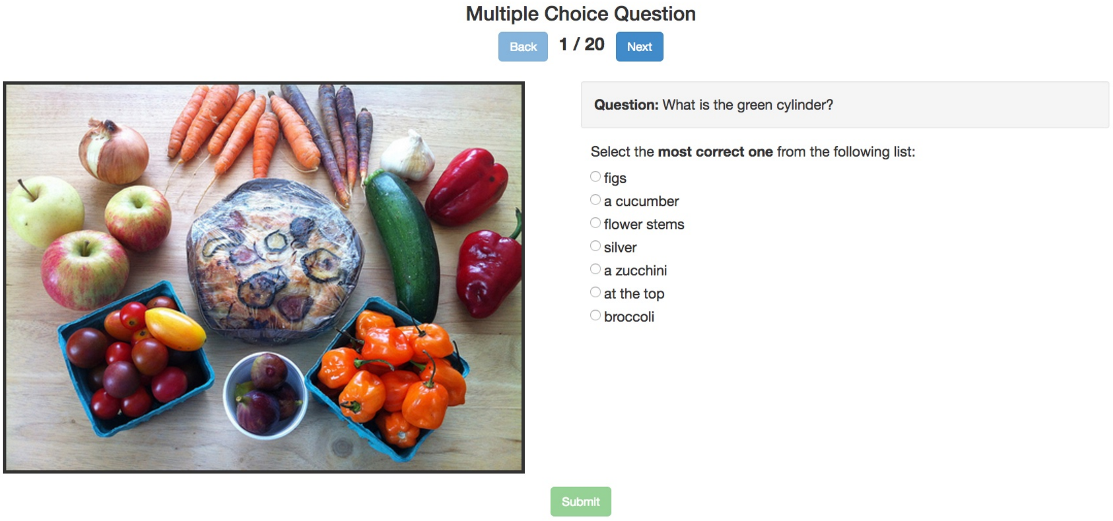
Since a candidate answer may contain more than one word (ie, a word sequence), we compute the similarity between two word sequences $WS_1$ and $WS_2$ as follows

$$\begin{aligned} &WUP(WS_1, WS_2) = \\ &\max\{\prod_{w_1\in WS_1} \max_{w_2\in WS_2} WUP(w_1, w_2) \qquad (4) \\ &\quad \prod_{w_2\in WS_2} \max_{w_1\in WS_1} WUP(w_1, w_2)\}. \end{aligned}$$

This formulations is highly similar to the one proposed in (Malinowski and Fritz, 2014) for evaluating open-ended visual QA. The main difference is that we use "max" rather than "min" to compute the final score. Note that our performance of using the WUP score is to filter out ambiguous decoys to the target. For example, we consider "a cute cat" to be ambiguous to "cat". Using eq. (4) gives a similarity 1, which can not be achieved by taking "min".

## C Details on user studies

As mentioned in Sect. 5.2 of the main text, we provide details on user studies. Fig. 4 shows our



**Question:** What is the green cylinder?

Select the **most correct one** from the following list:

- ○ figs
- ○ a cucumber
- ○ flower stems
- ○ silver
- ○ a zucchini
- ○ at the top
- ○ broccoli

Figure 4: User interface for human evaluation on Visual7W (IoU-decoys+QoU-decoys).

user interface. We perform the studies using Amazon Mechanic Turk (AMT) on Visual7W (Zhu et al., 2016), VQA (Antol et al., 2015) and Visual Genome (VG) (Krishna et al., 2016). We mainly evaluate on our IoU-decoys and QoU-decoys (combined together).

For each dataset, we randomly sample 1,000 image-question-target triplets together with the corresponding IoU-decoys and QoU-decoys to evaluate human performance. For each of these triplet, three workers are assigned to select the most correct candidate answer according to the image and question. We compute the average accuracy of these workers and report them in Table 3, 4 and 5 of the main text.

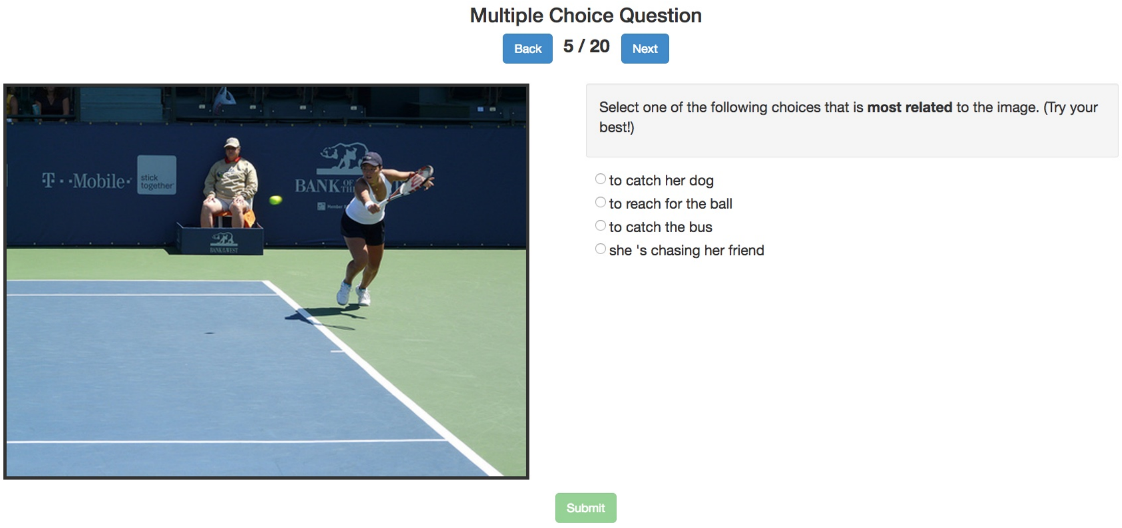
Besides our created decoys, we conduct human evaluation using the Orig decoys of Visual7W so as to investigate the difference between human-generated and automatically generated decoys. We also study how humans will perform given only partial information (ie, images + candidate answers or questions + candidate answers), again using the Orig decoys of Visual7W. The corresponding interfaces are shown in Fig. 5 and 6. For these studies, we use the same set of 1,000 triplets used to evaluate our created decoys for fair comparison. We make sure that a single worker does not work on the same triplet across the four studies on Visual7W. Results are reported in Table 1 of the main text.

In summary, 169 workers are involved in our studies. On our IoU-decoys and QoU-decoys, humans achieve 84.1%, 89.0%, and 82.5% on Visual7W, VQA, and VG, respectively. Compared to the human performance on the Orig decoys that involve human effort in creation (ie, 88.4% on Visual7W, and 88.5% on VQA as reported in (Antol et al., 2015)), these results suggest that the ways we create the decoys and the filtering steps mentioned in Sect. 4.2 lead to high-quality datasets with limited ambiguity.

## D Detailed results on VQA w/o QA pairs that have Yes/No as the targets

As mentioned in Sect. 5.3 of the main text, the validation set of VQA contains 45,478 QA pairs (out of totally 12,1512 pairs) that have Yes or No as the correct answers. The only reasonable decoy to Yes is No, and vice versa — any other decoy could be easily recognized in principle. Since both of them are among top 10 frequent answers, they are already included in the Orig decoys — our IoU-decoys and QoU-decoys can hardly make improvement. We thus remove all those pairs (denoted as Yes/No QA pairs) to investigate the improvement on the remaining pairs, for which having multiple choices makes sense. We denote the subset of VQA as VQA$^-$ (we remove Yes/No pairs in both training and validation set).

We conduct the same experiments as in Sect. 5.3 on VQA$^-$. Table 7 summarizes the machines' as well as humans' results. Compared to Table 4 of the main text, most of the results drop, which is expected as those removed Yes/No pairs are considered simpler ones — their *effective* random chance is 50%. The exception is for the MLP-IA models,



Multiple Choice Question

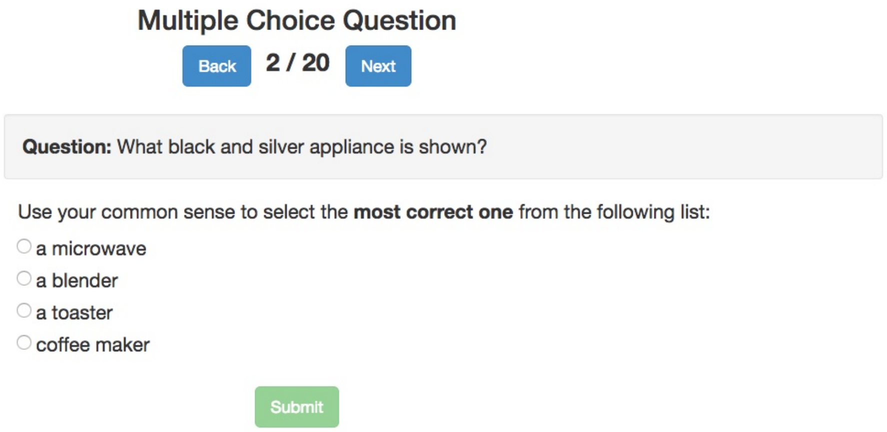
Back 5 / 20 Next

Select one of the following choices that is **most related** to the image. (Try your best!)

○ to catch her dog
○ to reach for the ball
○ to catch the bus
○ she 's chasing her friend

Submit

Figure 5: User interface for human evaluation on Visual7W (Orig decoys), where questions are blocked.



Multiple Choice Question

Back 2 / 20 Next

**Question:** What black and silver appliance is shown?

Use your common sense to select the **most correct one** from the following list:

○ a microwave
○ a blender
○ a toaster
○ coffee maker

Submit

Figure 6: User interface for human evaluation on Visual7W (Orig decoys), where images are not blocked.

which performs roughly the same or even better on VQA$^-$, suggesting that Yes/No pairs are somehow difficult to MLP-IA. This, however, makes sense since without the questions (eg, those start with "Is there a ..." or "Does the person ..."), a machine cannot directly tell if the correct answer falls into Yes or No, or others.

We see that on VQA$^-$, the improvement by our IoU-decoys and QoU-decoys becomes significant. The gain brought by images on QoU (from 39.3% to 56.6%) is much larger than that on Orig (from 45.8% to 55.6%). Similarly, the gain brought by questions on IoU (from 44.8% to 81.8%) is much larger than that on Orig (from 43.0% to 55.6%). After combining IoU-decoys and QoU-decoys as in IoU+QoU and All, the improvement by either including images to MLP-QA or including questions to MLP-IA is noticeable higher than that on Orig. Moreover, even with only 6 decoys, the performance by MLP-A on IoU+QoU is already lower than that on Orig, which has 17 decoys, demonstrating the effectiveness of our decoys in preventing machines from overfitting to the incidental statistics. Those observations together demonstrate how our proposed ways for creating decoys improve the quality of multiple-choice visual QA datasets.

| Method | Orig | IoU | QoU | IoU+QoU | All |
|---|---|---|---|---|---|
| MLP-A | 28.8 | 42.9 | 34.5 | 23.6 | 15.8 |
| MLP-IA | 43.0 | 44.8 | 53.2 | 35.5 | 28.5 |
| MLP-QA | 45.8 | 80.7 | 39.3 | 38.2 | 31.9 |
| MLP-IQA | 55.6 | 81.8 | 56.6 | 53.7 | 46.5 |
| Human-IQA | - | - | - | 85.5 | - |
| Random | 5.6 | 25.0 | 25.0 | 14.3 | 4.2 |

Table 7: Accuracy (%) on VQA$^-$-2014val, which contains 76,034 triplets.