

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

Wei-Lun Chao\*, Hexiang Hu\*, Fei Sha  
University of Southern California  
Los Angeles, California, USA

weilunchao760414@gmail.com, hexiang.frank.hu@gmail.com, feisha@usc.edu

## Abstract

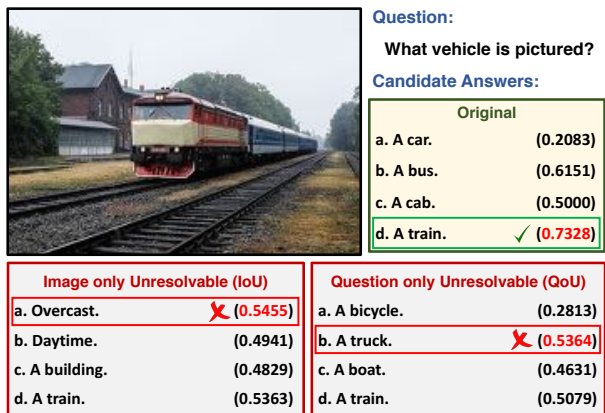
Visual question answering (QA) has attracted a lot of attention lately, seen as a form of (visual) Turing test that AI research should strive to achieve. In this paper, we study a crucial component of this task: how to design good datasets? We focus on multiple-choice datasets where the learner has to select the right answer from a set of candidate ones. Through analysis of the results attained by state-of-the-art learning models on existing datasets, we show the design of the candidate answers has a significant impact on how and what the models learn. In particular, the resulting learner can ignore the visual information, the question, or both while still doing well on the task. We propose automatic procedures to remedy such design deficiencies, re-constructing the candidate answers for two popular datasets as well as creating a new dataset from Visual Genome<sup>1</sup>. We conduct extensive empirical studies to demonstrate the effectiveness of our methods in creating better visual QA datasets.

## 1. Introduction

Recently, multimodal information processing tasks such as image captioning [27] and visual question answering (visual QA) [3] have gained a lot of attention. A number of significant advances in learning algorithms have been made, along with the development of nearly two dozens of datasets in this very active research domain. Among those datasets, popular ones include MSCOCO [18, 5], Visual Genome [16], VQA [3], and several others. The overarching objective is that a

\* Equal contributions

<sup>1</sup>The datasets are released and publicly available via [http://www.teds.usc.edu/website\\_vqa/](http://www.teds.usc.edu/website_vqa/).



Question: What vehicle is pictured?

Candidate Answers:

Original	
a. A car.	(0.2083)
b. A bus.	(0.6151)
c. A cab.	(0.5000)
d. A train.	✓ (0.7328)

Image only Unresolvable (IoU)	
a. Overcast.	✗ (0.5455)
b. Daytime.	(0.4941)
c. A building.	(0.4829)
d. A train.	(0.5363)

Question only Unresolvable (QoU)	
a. A bicycle.	(0.2813)
b. A truck.	✗ (0.5364)
c. A boat.	(0.4631)
d. A train.	(0.5079)

Figure 1. An illustration of how the shortcuts in the Visual7W dataset [30] 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 the other decoy (negative) answers. (The numbers in the brackets are probability scores computed using eq.(2)). Our two procedures — QoU and IoU (cf. sec. 4) — 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.

learning machine needs to go beyond understanding different 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 [2, 15, 20, 14]. Thus, many efforts have concentrated on tasks such as *multiple-choice* visual QA [3, 30, 12] or selecting the

best caption [11, 10, 7, 19], where the selection accuracy 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* — is as important as deciding the questions to ask. We all have had the experience of exploiting the elimination strategy: *This question is 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 comprehend the meanings in images and languages (*e.g.*, the quality of the embeddings of both images and languages in a semantic space). It has been noted that machines can achieve very high accuracies of selecting the correct answer without the visual input (*i.e.*, the image), the question, or both [12, 3]. Clearly, the learning algorithms have overfit on incidental statistics in the datasets. For instance, if the decoy answers have rarely been used as the correct answers (to any questions), 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 memorize 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 [30] and VQA [3]. Additionally, we create a multiple-choice based dataset from the recently released Visual Genome dataset [16], 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 between the human and the 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. Our datasets are released and publicly available via [http://www.teds.usc.edu/website\\_vqa/](http://www.teds.usc.edu/website_vqa/).

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

[14, 25] 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-world 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 [3], Visual7W [30], and Visual Genome [16]. All of them use images from MSCOCO [18].

Besides the pairs of questions and correct answers, VQA, Visual7W, and visual Madlibs [28] 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 generation is carefully orchestrated 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 [12] 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 [8, 29, 13, 1], though not in the multiple-choice setting. Our work extends theirs by providing more detailed analysis *as well as automatic procedures* to remedy those design deficiencies.

Besides the visual QA task, [7] and VisDial [6] also propose automatic ways to generate decoys for the tasks of selecting the best visual caption and dialog, respectively.

### 3. Analysis of Decoy Answers’ Effects

In this section, we examine in detail the dataset Visual7W [30], 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  $\{I, Q, A = \{T, D_1, \dots, D_K\}\}$ . 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 modalities of information. We concentrate on the one hidden-layer MLP model proposed in [12], which had achieved state-of-the-art results on the dataset Visual7W. The model computes a scoring function  $f(c, i)$

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

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  $i$  can be null, the image ( $I$ ) alone, the question ( $Q$ ) alone, or the combination of both ( $I+Q$ ).

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

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

Given an IQA triplet, we use the penultimate layer of ResNet-200 [9] as visual features to represent  $I$  and the average WORD2VEC embeddings [22] as text features to represent  $Q$  and  $C$ . To form the joint feature  $g(c, i)$ , we just concatenate the features together. The candidate  $c \in A$  that has the highest  $f(c, i)$  score 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. The parameters of  $f(c, i)$  are learned by minimizing the binary logistic loss of predicting whether or not a candidate  $c$  is the target of an IQA triplet. 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 performed 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 guessing and humans (52.9% vs. 25%) — humans will deem each of the answers equally likely *without* looking at both the image and the question! Note that in this case, the information  $i$  in eq. (1) contains nothing. Thus, the model learns the specific statistics of the candidate answers in the dataset and exploits those.

Adding the information about the image (*i.e.*, 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 between the two in understanding visual information.

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 as the model cannot fully exploit the information in “Q”, making a smaller relative improvement 3.3% (from 62.4% to 65.7%) where humans improved relatively 17.4%.

**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 (*i.e.*, 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 of being correct,

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

should perform well. Essentially, it measures how unbiased  $C$  is used as the target and the decoys. Indeed, it attains an accuracy of 48.73% on the test data, far better than the random inguess 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, they 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 the 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<sup>2</sup>. 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 either the images or the 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, *i.e.*, 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,

<sup>2</sup>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.

- **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 [22] 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 [30]. 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 [26] 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 [21], 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 [4] 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 have  $K$  decoys. We also perform such a check among selected decoys to ensure they are not very similar to each other. 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” [3, 28, 30, 6]. 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. [28, 6, 7] automatically find decoys from similar questions or captions based on question templates and annotated objects, trigrams and GLOVE embeddings [23], and paragraph vectors [17] and linguistic surface similarity, respectively. The later two are for different tasks from visual QA, and only [7] considers removing semantically ambiguous decoys like ours. [3, 30] 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 [3]** The dataset uses images from MSCOCO [18] 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) [30]** The dataset uses 47,300 images from MSCOCO [18] and contains 139,868 IQA triplets. Each has 3 decoys generated by humans.

**Visual Genome (VG) [16]** The dataset uses 101,174 images from MSCOCO [18] 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

Table 2. Summary of visual QA datasets

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

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.

## 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) for multimodal information, respectively. The hidden-layer has 8,192 neurons. We use a 200-layer ResNet [9] to compute visual features which are 2,048-dimensional. The ResNet is pre-trained on ImageNet [24]. The `WORD2VEC` feature [22] for questions and answers are 300-dimensional, pre-trained on Google News. The parameters of the MLP models are learned by minimizing the binary logistic loss of predicting whether or not a candidate answer is the target of the corresponding IQA triplet. We use stochastic gradient descent with mini-batch size of 100, momentum of 0.9, and the stepped learning rate policy in optimization. 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 selected by our QoU-decoys generating procedure, (3) **IoU: Orig** replaced with ones selected by our IoU-decoys generating procedure (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 [30]. We thus conduct a user study via Amazon Mechanic Turk (AMT) to test humans’ performance on the datasets after they are remedied by our automatic procedures. We select 1,000 IQA triplets from each dataset. 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 from exploiting 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 inadequate to resolve), 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 MLP-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 the 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.

Table 3. Test accuracy (%) on Visual7W.

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 4. Accuracy (%) on the validation set in VQA

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 [3]

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.

**Differences across 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 noticeable improvement. In Supplementary Material, however, we show that once we remove such questions/answers pairs, the degree of improvements increases substantially.

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

Table 5. Test accuracy (%) on qaVG

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 6. Use models trained on qaVG to improve Visual7W and VQA (Accuracy in %).

Datasets	Decoys	Best w/o using qaVG	qaVG model	
			initial	fine-tuned
Visual7W	Orig	65.7	60.5	<b>69.1</b>
	IoU +QoU	52.0	58.1	<b>58.7</b>
	All	45.1	48.9	<b>51.0</b>
VQA	Orig	64.6	42.2	<b>65.6</b>
	IoU +QoU	63.7	47.9	<b>64.1</b>
	All	58.9	37.5	<b>59.4</b>

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. So there is no mismatch in distribution between images (and their features).

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 [12] where a model pre-trained on VQA is fine-tuned on Visual7W.

## 5.4. Qualitative Results

In Fig. 2, we present examples of 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 examples. We categorized those failure cases into two situations.

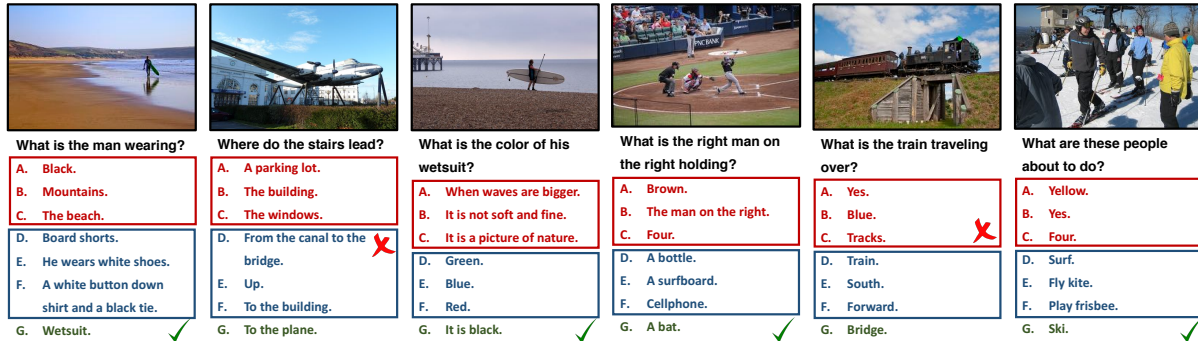


Figure 2. Example image-question-target triplets from Visual7W, VQA, and VG, together with our IoU-decoys (A, B, C.) and QoU-decoys (D, E, F). G is the target. Machine’s selections are denoted by green ticks (correctly answered) or red crosses (wrongly answered).

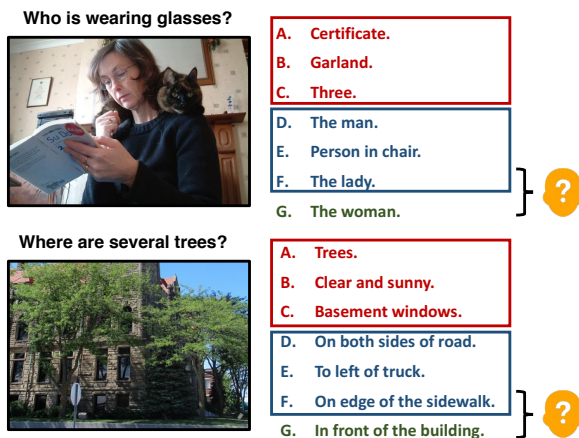


Figure 3. Ambiguous examples by our IoU-decoys (A, B, C) and QoU-decoys (D, E, F). G is the target. Ambiguous decoys F are marked.

- Our filtering steps in Sect. 4 fail, as observed in the top example. 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 [6, 7] 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 an 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 provide “shortcuts” for machines to exploit to perform well on the task. We describe several principles of constructing good decoys and propose 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 are released and available at [http://www.teds.usc.edu/website\\_vqa/](http://www.teds.usc.edu/website_vqa/).

## References

- [1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.
- [2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
- [4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- [5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. *arXiv preprint arXiv:1611.08669*, 2016.



- [7] N. Ding, S. Goodman, F. Sha, and R. Soicuc. Understanding image and text simultaneously: a dual vision-language machine comprehension task. *arXiv preprint arXiv:1612.07833*, 2016.
- [8] 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. *arXiv preprint arXiv:1612.00837*, 2016.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [10] M. Hodosh and J. Hockenmaier. Focused evaluation for image description with binary forced-choice tasks. In *ACL Workshop*, 2016.
- [11] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [12] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016.
- [13] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016.
- [14] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. *arXiv preprint arXiv:1610.01465*, 2016.
- [15] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*, 2016.
- [16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.
- [17] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.
- [18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [19] X. Lin and D. Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.
- [20] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2016.
- [21] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.
- [22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.
- [23] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- [25] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016.
- [26] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, 1994.
- [27] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [28] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015.
- [29] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, 2016.
- [30] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.