Visual Question Answer Diversity

Chun-Ju Yang
Electrical and Computer Engineering Dept.
University of Texas at Austin

Kristen Grauman
Computer Science Dept.
University of Texas at Austin

Danna Gurari
School of Information
University of Texas at Austin

Abstract

Visual questions (VQs) can lead multiple people to respond with different answers rather than a single, agreed upon response. Moreover, the answers from a crowd can include different numbers of unique answers that arise with different relative frequencies. Such answer diversity arises for a variety of reasons including that VQs are subjective, difficult, or ambiguous. We propose a new problem of predicting the answer distribution that would be observed from a crowd for any given VQ; i.e., the number of unique answers and their relative frequencies. Our experiments confirm that the answer distribution can be predicted accurately for VQs asked by both blind and sighted people. We then propose a novel crowd-powered VQA system that uses the answer distribution predictions to reason about how many answers are needed to capture the diversity of possible human responses. Experiments demonstrate this proposed system accelerates capturing the diversity of answers with considerably less human effort than is required with a state-of-art system.

Introduction

The goal of a visual question answering (VQA) system is to empower people to learn the answer to any question about any image (Antol et al. 2015; Bigham et al. 2010; Malinowski, Rohrbach, and Fritz 2015). For example, a VQA system could enable blind people to address daily visual challenges, such as learning whether a pair of socks match or learning what type of food is in a can. VQA services could also facilitate the creation of smarter environments, such as investigating the reason for an observed crowd behavior in public places.

A challenge for designing VQA systems in practice is that different people can arrive at the same answer or diverse answers when answering a visual question (VQ). This is exemplified in Figure 1, where we show a variety of answer distributions that arose in response to six VQs. In some cases, an anonymous crowd of on-line workers (the typical individuals to provide answers) agreed upon a single answer (e.g., Figure 1; column 1), at the other extreme all individuals disagreed with each other (e.g., Figure 1; column 6), and in between these two extremes were different numbers of people clustering around a different number of answers (e.g., Figure 1; columns 2–5); i.e., different numbers of unique answers arise with differing relative frequencies. Different answer distributions arise for a numerous reasons including because VQs are ambiguous, subjective, or difficult as well as because answers are synonyms. Currently, a person cannot know the level of answer diversity that will arise in response to a VQ without collecting answers from a crowd. The best one can achieve today is to predict whether a crowd will disagree when answering a VQ (Gurari and Grauman 2017). Yet, a more flexible understanding of the fine-grained answer distribution that will arise for a VQ could be valuable for a variety of purposes.

One motivation for anticipating the answer distribution is such information could be valuable to efficiently collect all unique answers from a crowd. The motivating assumption is that capturing all plausible answers from a crowd is a feature rather than a bug (e.g., spam); i.e., an open call to crowd workers should be sufficiently large to capture all plausible perspectives for ambiguous VQs, to capture all plausible perspectives for subjective VQs, and to include a person with the rare domain expertise needed to correctly answer difficult VQs. Existing crowd-powered systems are inefficient because they choose the number of people to provide answers either based on the crowdsourcing conditions (Bigham et al. 2010) or pre-determined numbers; e.g., crowd size is one for (Malinowski, Rohrbach, and Fritz 2015), three for (Antol et al. 2015), and one or five for (Gurari and Grauman 2017). Consequently, existing approaches accrue extra costs and delays by soliciting answers from extra members of the crowd when all unique answers have already been collected (e.g., collecting five answers when only two answers are needed). Existing approaches also sacrifice on quality by not soliciting answers from enough members of the crowd to capture all unique answers (e.g., collecting only two answers when four answers are needed). Fine-grained predictions to reason about how many members of the crowd is just enough to ask for each VQ would reduce costs and delays to collect all unique answers for each VQ.

Anticipating the answer distribution from a crowd could also enable the design of more helpful automated VQA systems. This is because automated systems currently often return a single answer per VQ. Such systems could be more valuable if they informed users about the extent to which independent members of the crowd might agree on the re-

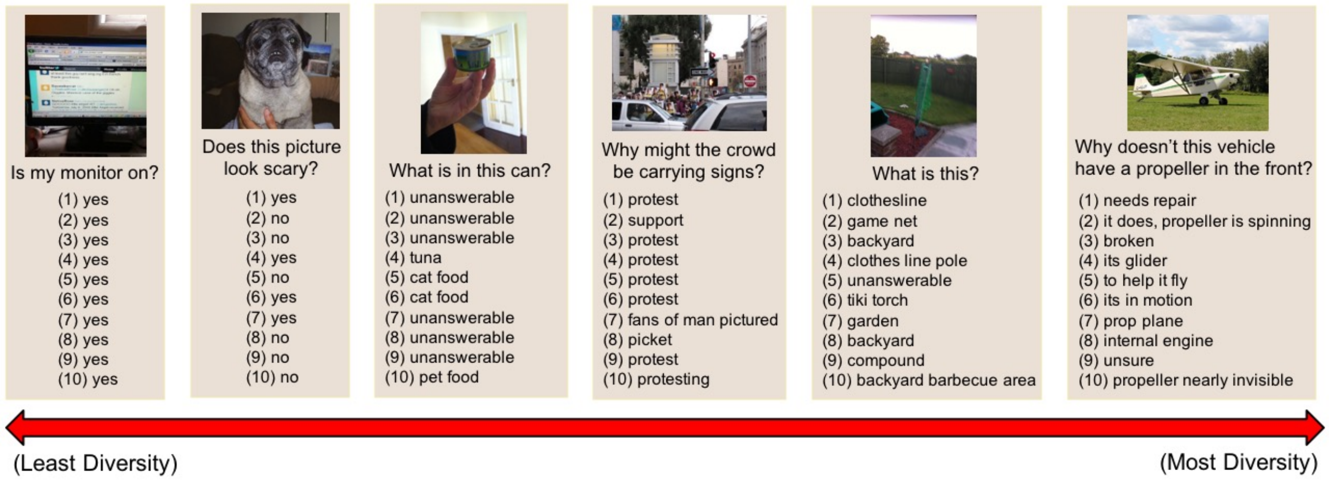


Figure 1: Examples of visual questions and corresponding answers from 10 different people. The examples include visual questions asked by both blind individuals and sighted individuals. As observed, the answer distribution can vary from unanimous agreement on a single answer (first column) to uniform disagreement on different answers (last column).

turned answer; e.g., 50% versus 100% agreement for "Yes".

Furthermore, anticipating the answer distribution from a crowd could provide valuable, real-time feedback so that those asking questions could quickly modify their VQ to yield the desired answer diversity. For example, a teacher may want to gauge how active a debate will be in response to a question (s)he is considering asking about a picture or painting; will it shut down conversation versus encourage a simple controversy (two answers) versus encourage an open-ended, exploration of many plausible answers? Additionally, a blind person who is choosing what to eat for lunch may initially fail to ask a VQ that elicits unanimous agreement about "What is in this can?"; e.g., see column 2 in Figure 1.

Accordingly, our goal is to design a VQA framework that can account for the diversity of answers inherent in crowd intelligence. We propose a new problem of predicting the answer distribution directly from the VQ. This task is challenging because it necessitates designing a framework that can simultaneously model and synthesize different individuals' (potentially conflicting) perceptions of images and language for the many possible causes of disagreement (e.g., ambiguity, subjectivity). Our findings show multiple prediction systems yield promising predictive power for this task on VQs asked by both sighted and blind people.

We also propose a novel crowdsourcing system for efficiently collecting the diversity of plausible answers for a set of VQs. Our experiments demonstrate the benefit of employing fine-grained predictions to reason about how many members of the crowd is just enough to ask for each VQ in order to reduce human involvement when collecting the unique answers for a collection of VQs.

Related Works

Automated Answer Prediction An automated VQA system typically returns a single answer by identifying the option that has the highest probability of being correct (Andreas et al. 2016; Antol et al. 2015; Malinowski, Rohrbach, and Fritz 2015; Goyal et al. 2017; Zhang et al. 2016b). This confidence score provided by an algorithm is determined by many factors such as the biases of the observed training data, embedded algorithm assumptions, and process used for training the algorithm (e.g., overfitting, underfitting). Our goal is distinct. While different algorithms can lead to different confidences in their predictions for the same visual question, we instead are seeking a consistent single prediction that reveals the answer distribution expected from a crowd for a given visual question; e.g., will members of a crowd all return the same answer, have a split opinion between two answers, or all return different answers? Our experiments confirm that it is possible to predict the answer distribution one would observe from a crowd.

Predicting Distributions Our work relates to prior work that predicts the distribution of emotions evoked by an image (Peng et al. 2015; Zhao et al. 2016; Yang, Sun, and Sun 2017; Ali et al. 2017). For example, (Peng et al. 2015) predicts the distribution that occurs for six pre-defined categories of emotions. To our knowledge, our work is the first to predict the distribution that would be observed from a crowd for the domain of VQA. Furthermore, our work is not constrained to a pre-defined set of categories. Rather, we propose models that predict the answer distribution without knowing the categories of answers that will be observed.

Measuring Difficulty to Answer a Visual Question Our work relates to prior works that examine the difficulty to answer a VQ. One set of approaches aim to understand the difficulty of a VQ for a computer. For example, one approach measures difficulty based on the availability of a similar matching VQ in an existing database; i.e., whether a VQ can be answered using the knowledge in a given database (Yeh, Lee, and Darrell 2008). Another approach measures difficulty based on the algorithm's consistency in returning the same answer when it's repeatedly queried to answer the VQ, with different words occluded in the question at different iterations (Goyal et al. 2016). A further approach shows that a VQA algorithm's chance for success is related with the level of crowd agreement (i.e., the accuracy is higher when crowd agreement is higher) (Malinowski, Rohrbach, and Fritz 2015). Unlike such work, we are interested in the difficulty of a VQ for a human. Thus, our work more closely aligns with prior work that tries to identify the difficulty of a VQ by directly asking the crowd to indicate the minimum age required for a human to successfully answer the VQ (Antol et al. 2015). Unlike this work, we use answer distribution as a possible indicator of VQ difficulty for a person; i.e., people tend to agree on fewer answers for easier VQs and disagree more for more complex VQs (e.g., large counting problems). Moreover, rather than focus solely on difficulty, we consider difficulty as one possible reason out of many (e.g., subjectivity, ambiguity, answer granularity) that can explain answer diversity from a crowd.

Solving Problems with a Limited Human Budget Various prior works explore efficiently allocating limited human resources in order to optimize an outcome. For example, (Jain and Grauman 2013) decide how much human effort to allocate per image from three choices in order to accurately segment a batch of images. Another work similarly focuses on segmenting a batch of images, but does so by distributing the work between more costly crowd workers and less expensive algorithms (Gurari et al. 2016). Another work examines how to perform biomedical citation screening more efficiently by distributing the work between domain experts and less costly crowd workers (Nguyen, Wallace, and Lease 2015). Most closely related to our work is (Gurari and Grauman 2017), which decides whether to employ one or five crowd workers to answer a VQ in order to collect all plausibly valid answers. Unlike (Gurari and Grauman 2017), which offers a binary predictor, we propose a prediction system that has the fine-grained understanding to measure *to what extent* a crowd will disagree and, thus, how many answers to collect to accelerate the collection of unique answers when given a limited human budget.

Approach

In this section, we describe our datasets, models for predicting answer diversity, and then our experiments for evaluating our prediction models.

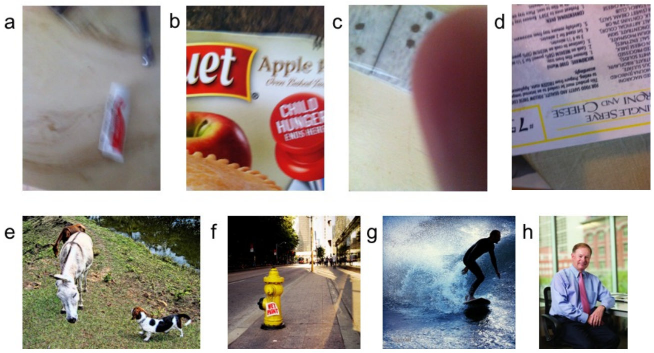


Figure 2: Examples of images in the (a–d) VizWiz and (e–f) VQA datasets. As shown, VizWiz often contains many poor quality images due to numerous phenomena: (a) not focused, (b) object partly in frame, (c) no salient object, (d) object upside down. In contrast, (e)-(h) exemplifies that VQA often contains higher quality images.

Datasets

We now describe two VQA datasets that we use for our experiments and how we use these datasets to produce a diversity score for each VQ.

VQA Real Images (VQA) (Antol et al. 2015): We train with 250,000 and test with 214,354 VQAs from the VQA-Real Images version 2 dataset. The images of this dataset come from Microsoft COCO (MSCOCO) (Lin et al. 2014) and consist of common objects in everyday scenes. Many images show non-iconic views, meaning there is not a single salient object centered in the image (e.g., Figure 2e). The questions come from online crowd workers who were shown the images and instructed to provide questions. Each VQ is paired with 10 crowdsourced answers.

VizWiz (Gurari et al. 2018): We train with 6,408 and test with 1,601 VQAs. Each VQ was asked by a blind person who took a picture and recorded a question about it with a mobile phone application (Bigham et al. 2010). The questions often address challenges blind users face in their daily lives, e.g., ”What is this?”, ”What is on the screen?”, ”How to cook this?”, ”What is the expiration date?”. Because the images are taken by blind users, the quality is often poor; e.g., out of frame or not focused. Examples of such images are shown in Figure 2(a-d). Each VQ is paired with 10 crowdsourced answers.

Diversity Measure (Ground Truth Values)

Our definition of diversity is inspired by information theory, where the goal is to measure the amount of information contained in a message. In particular, we are interested in how much information is observed when aggregating answers collected from multiple, independent people. We compute entropy, using the follow equation:

$$E = \sum_{i=1}^{N} -p_i \mathrm{log} p_i \quad (1)$$

where M represents the number of people providing answers, N represents the number of unique answers observed from M people, and p_i represents the fraction of M people who provided the i-th answer from the N possible answers. When all M people agree on one answer, the diversity score is 0 (i.e., $E = -1\mathrm{log}_2 1$), which indicates few bits are required to code the answers. On the other hand, if $M = 10$ with all people offering different answers, the diversity score is 3.32 (i.e., $E = 10 * -0.1\mathrm{log}_2 0.1$), which indicates many bits are required to code the answers.

For both datasets, each VQ has 10 open-ended answers that were collected from independent crowd workers on Amazon Mechanical Turk (AMT). We calculated diversity scores using these 10 answers with equation 1.

Prediction Models

We propose two learning frameworks. We first describe a regression framework based on handcrafted features for the question and image. We then describe a classification framework that uses deep learning to learn directly from the question and image.

Handcrafted Features We first propose a regression model to capture that visual questions elicit different amounts of information from a crowd, ranging from one agreed upon answer to N distinct answers from N independent viewers. We use concatenated image and question features with our regression model. Our model is inspired in part by the findings of prior work which shows handcrafted features outperform a modern deep learning based system in predicting whether a crowd will disagree on the answer to a visual question (Gurari and Grauman 2017). Thus, a key focus here is in selecting a richer feature set that will capture the fine-grained nuances that determine the diversity of responses from a crowd, rather than simply detecting if a crowd will disagree (Gurari and Grauman 2017).

We represent the *question* using the following features:

– *Question length*: we use the number of words in the question, inspired by the hypothesis that additional information (i.e., words) offers the precision needed for a crowd to agree on a single answer while less information leaves greater ambiguity/space for a greater answer distribution.

– *First two words of a question*: we use two one-hot encoding vectors for the first two words in the question, built using vocabularies learned during training to define all possible words at the first and second word location of the question respectively. Intuitively, the first two words can be a strong indicator of the level of answer diversity; e.g., “is the... ?” likely will lead to at most two answers, ”Yes” or ”No”, whereas “why is... ?” may lead to many different answers.

– *Lexical categories*: we tally for all words in the question how many belong to each of 15 parts of speech tags. Part of the motivation is to count the amount of descriptive language (e.g., number of nouns, adjectives, and prepositions) based on the intuition that greater descriptive language offers more leading clues/landmarks for different people to arrive at a single answer. Part of the motivation is to also to detect the tense of a question based on the hypothesis that asking about future and past events relies less on grounded visual content in an image and rather more on a person's imagination, thereby leaving more opportunity for a distribution of answers from a crowd than when asking questions about the present tense. The lexical categories we consider are as follows:

- Determiner
- Singular Noun (e.g., desk)
- Plural noun (e.g., desks)
- Preposition/subordinating conjunction
- Existential there (e.g., ”there is”)
- Adjective (e.g., big)
- Comparative adjective (e.g., bigger)
- Superlative adjective (e.g., biggest)
- Modal (e.g., could, will)
- Base form verb, (e.g., take)
- Past tense verb (e.g., took)
- Gerund/present participle verb (e.g., taking)
- Third person singular present verb (e.g., takes)
- Wh-determiner (e.g., which)
- Wh-pronoun (e.g., who, what)

We chose the following *image-based* features to capture the perceived diversity of an image:

– *GoogleNet Features*: We extract inception-v3 (Szegedy et al. 2015b) features resulting in a 2048-dimension output features. As one of the recent ImageNet Large Scale Visual Recognition Competition (ILSVRC) (Russakovsky et al. 2015) winners, GoogleNet (Szegedy et al. 2015b) aims to recognize objects from 1000 classes that cover a wide variety range of objects for image extraction. We applied principal component analysis afterward to reduce the dimensionality to 100.

– *Convolutional neural network (CNN) based salient object subitizing (SOS) model*: SOS model (Zhang et al. 2016a) predicts the number of salient objects in the image. This CNN based SOS model is fine-tuned from GoogleNet (Szegedy et al. 2015a). The output fully connected layer has a 5-dimensional score vector corresponding to the probability of the image belonging to 5 categories; i.e., 0, 1, 2, 3 and 4+ salient objects in the image.

We use a linear regression model to predict the diversity score from the image and question features. We chose linear regression over other models because we observed better results from this model during initial testing.

Deep Learning System Similar to prior work (Gurari and Grauman 2017), we also train a deep learning system from scratch using the architecture described in (Antol et al. 2015; Lu et al. 2015). The question is encoded with a 1024-dimensional Long Short Term Memory (LSTM) model that

	VQA - Real Images			VizWiz		
	RC	CC	MAE	RC	CC	MAE
Status Quo	-0.01	-0.01	0.3	0.03	0.03	0.3
CrowdVerge (Gurari and Grauman 2017)	0.46	0.46	0.38	0.12	0.12	0.23
Ours: Deep Learning	0.53	0.59	0.45	**0.36**	**0.36**	**0.22**
Ours: Linear Regression	**0.63**	**0.63**	**0.18**	0.29	0.29	0.19

Table 1: Comparison of methods for predicting answer diversity in a single dataset setting. Specifically, we compare our linear regression system and deep learning system with the "Status Quo" that randomly assigns a score and the CrowdVerge system (Gurari and Grauman 2017) that predicts whether a crowd will agree or disagree on a single answer. Higher rank coefficient (RC) scores, higher correlation coefficient (CC) scores, and lower mean absolute error (MAE) scores are better.

takes in a one-hot descriptor of each word in the question. The image is described with the 4096-dimensional output from the last fully connected layer of the Convolutional Neural Network, VGG16 (Simonyan and Zisserman 2014).

Experiments

Baselines We compare our system against the following:

- **Status Quo**: this predictor randomly assigns a diversity score to illustrate what occurs from random guessing.
- **CrowdVerge** (Gurari and Grauman 2017): this model predicts whether a VQ will lead to answer (dis)agreement from a crowd. We use its confidence in its prediction, which ranges from 0 to 1, as the diversity score.

Evaluation Metrics We evaluate each model using:

- **Pearson's correlation coefficient (CC)**: CC indicates how strong the prediction correlates to ground truth for all VQs. Values range between +1 and -1, with values further from 0 indicating stronger predictive power.
- **Spearman rank correlation (RC)**: RC is similar to CC, except the prediction scores are ranked.
- **Mean Absolute Error (MAE)**: measures the absolute difference between prediction and ground truth for a VQ, and then averages all the difference values.

Single Dataset Prediction Performance We first train and test using the same dataset, i.e., train with VQA training data and test on a disjoint VQA testing dataset as well as train with VizWiz training data and test on a disjoint VizWiz testing dataset. We evaluate both our linear regression system and deep learning system.

Table 1 shows the results. Overall, our models outperform all baselines for both datasets across all evaluation metrics. For example, for the VQA dataset, our top-performing model improves the CC compared to the best-performing baseline by 0.17 (Table 1; Ours: Linear Regression versus CrowdVerge). For the VizWiz dataset, we see our top-performing model improves the CC by 0.24 (Table 1; Ours: Deep Learning versus CrowdVerge). For VQA, linear regression model even beats the deep learning system (Table 1; VQA-Real Images; Ours: Linear Regression versus Ours: Deep Learning).

We show examples of prediction results for both datasets in Figure 3. As observed, our models can predict well in the presence of different language properties such as different first words in the questions; e.g., "What ...", "Is ...", "Who's ...", "Had ...". As shown, our models also can predict well for a variety of images such as high quality images of scenes and objects as well as lower quality images.

Interestingly, the overall performance for VizWiz is worse than VQA (Table 1; column 1-3 versus column 4-6). We hypothesize one reason for this is because of the small size of the training data. VizWiz only has 6,408 training examples, which is nearly two orders of magnitude smaller than the number of training examples in VQA.

Next, we examine the predictive power of the image and question information alone. To do so, we train the top-performing linear regression model using the question information and image information independently.

Table 2 shows the results. We observe question features are the most predictive. We also observe image information alone has predictive power for both datasets (Table 2; row 1; Ours LR: I; CC; VQA/VQA and VizWiz/VizWiz). The predictive capability of our image-based model is considerably stronger when learned on VizWiz compared to VQA; i.e., CC improves from 0.09 to 0.21 (Table 2; Ours LR: I; CC; VQA/VQA versus VizWiz/VizWiz). We suspect this is because the images in VizWiz are more consistent; e.g., many are blurry and so will lead to unanimous agreement the VQ is unanswerable. Overall, we also observe the models predict better for both datasets when using both image and question information (Table 2; rows 1 and 2 versus row 3).

Cross Dataset Prediction Performance In order to examine how well our models generalize, we next do cross dataset testing. We again examine our best performing linear regression model for this experiment.

We first train using the VizWiz training dataset and then test with the VQA test dataset. We find the VizWiz model has little change to performance across all evaluation metrics when being applied to the VizWiz test dataset rather than the VQA test dataset (i.e, (Table 3; row 3, columns 7-12; RC, CC, MAE; VizWiz/VizWiz versus VizWiz/VQA). However, when the model is trained only on the image information, it drops in performance when applied on the VQA test dataset rather than the VizWiz test dataset (Table 3; row 1, column 7-12; RC, CC, MAE; VizWiz/VizWiz versus VizWiz/VQA). This suggests there is a domain mismatch between the images in the VizWiz dataset and VQA dataset.

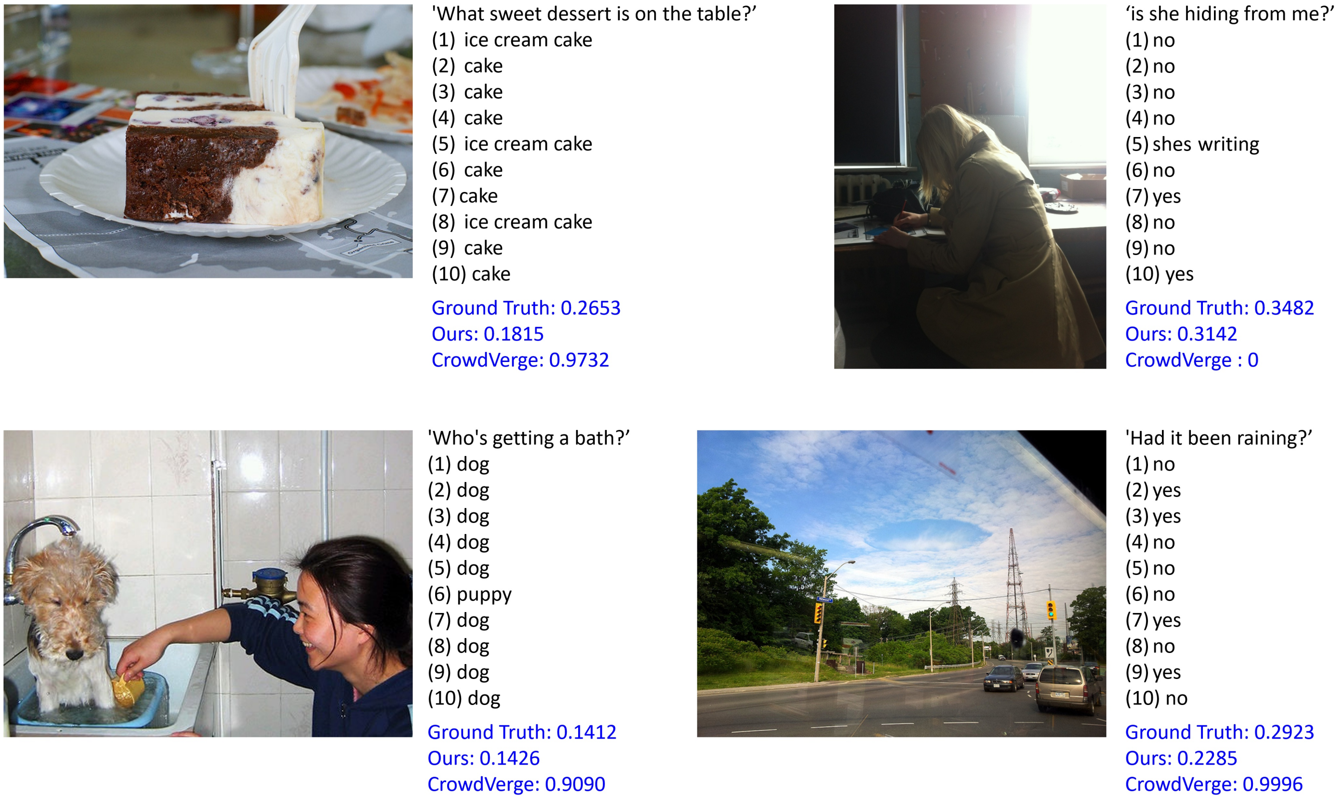


Figure 3: Examples of resulting answer diversity scores for the ground truth, our predicted entropy score, and CrowdVerge's predicted disagreement score. As observed, our approach can predict scores similar to ground truth for a diversity of questions (e.g., different first words) and images (e.g., focused on objects and scenes).

Train/Test	**VQA/VQA**			**VQA/VizWiz**			**VizWiz/VizWiz**			**VizWiz/VQA**		
	RC	**CC**	**MAE**	**RC**	**CC**	**MAE**	**RC**	**CC**	**MAE**	**RC**	**CC**	**MAE**
Ours LR: I	0.09	0.09	0.24	0.02	0.02	0.27	0.21	0.21	0.20	0.02	0.02	0.32
Ours LR: Q	0.62	0.62	0.18	0.13	0.13	0.24	0.20	0.20	0.20	0.27	0.27	0.28
Ours LR: Q+I	**0.63**	**0.63**	**0.18**	**0.13**	**0.13**	**0.24**	**0.29**	**0.29**	**0.19**	**0.28**	**0.28**	**0.27**

Table 2: For both single-dataset (column 1 and 3; VQA/VQA; Viz/Viz) and cross-dataset (column 2 and 4; VQA/Viz; Viz/VQA) settings, we train and test with image-based features alone (row 1;"Ours RL: I"), question-based features alone (row 2; "Ours LR: Q"), and both features together (row 3; "Ours LR: Q+I") for our linear regression model (Ours LR).

Train/Test	**VQA/VQA**			**VQA/VizWiz**			**VizWiz/VizWiz**			**VizWiz/VQA**		
	RC	**CC**	**MAE**	**RC**	**CC**	**MAE**	**RC**	**CC**	**MAE**	**RC**	**CC**	**MAE**
Status Quo	-0.01	-0.01	0.3	-0.01	-0.01	0.5	0.03	0.03	0.3	0	0	0.5
CrowdVerge	0.46	0.46	0.38	0.06	0.06	0.39	0.12	0.12	0.23	0.13	0.13	0.46
Ours LR	**0.63**	**0.63**	**0.18**	**0.13**	**0.13**	**0.24**	**0.29**	**0.29**	**0.19**	**0.28**	**0.28**	**0.27**

Table 3: Cross-dataset performance of our model, "Status Quo" baseline, and "CrowdVerge" (Gurari and Grauman 2017).

We next examine the performance of training with the VQA training dataset and testing on the VizWiz test dataset. We observe a large performance drop when applying the VQA model on the VizWiz test dataset rather than the VQA test dataset (Table 3; row 3, columns 1-6; RC, CC, MAE; VQA/VQA versus VQA/VizWiz). The results suggest again that there is a domain mismatch between the VQA and VizWiz datasets. We hypothesize the domain mismatch arises in part because images are significantly different in the two datasets (Figure 2). Specificially, VQA images are taken by sighted people and often are of good quality while VizWiz images are taken by blind people and often are of poor quality. We suspect these image-based issues bring out challenges for model generalization.

Collecting Answers With Budget Constraint

In this section, we examine how many people to recruit from a crowd to provide an answer in order to efficiently capture all unique answers for a set of VQs. We propose to use our diversity prediction model to address this problem. Specifically, we predict the number of answers that need to be collected for *each VQ* to maximize the number of unique answers that are captured for a set of visual questions under a given budget constraint.

Budgeted Answer Collection System

Objective and Optimization Suppose we have n VQs and a budget B which is the total number of answers we can afford to collect. We can collect up to q answers for each VQ; For an individual VQ, VQ k, let $x = [x_k^1, x_k^2, x_k^3, ..., x_k^q] \in \mathcal{R}^q$ be the selection vector indicating the number of answers that we collect; e.g., $x_k^1 = 1$ means to collect one answer. Let $c = [1, 2, 3, ...,\text{q}] \in \mathcal{R}^q$ be the cost vector, $c^T x$ is the corresponding cost upon the selection such as the cost when collecting one answer. In order to know how many unique answers we can expect when collecting 1 answer, 2 answers, ..., q answers, we let answer expectation (AE) vector, $E = [E_k^1, E_k^2, E_k^3, ..., E_k^q]$ be the expectation of a unique answer when collecting 1 answer, 2 answers, ..., q answers.

We define our annotation budget constraint as follows:

$$\begin{aligned} x = &\underset{x}{\text{argmax}} \sum_{k=1}^{n} E_k^1 x_k^1 + E_k^2 x_k^2 + E_k^3 x_k^3, ... + E_k^q x_k^q, \\ \textbf{s.t. } & c^T x \le B, \qquad\qquad\qquad\qquad\qquad\qquad (2) \\ & x_k^1 + x_k^2 + x_k^3 + ... + x_k^q = 1, \forall k = 1, 2, ..., n, \\ & x_k^1, x_k^2, x_k^3, ..., x_k^q \in \{0, 1\}, \forall k = 1, 2, ..., n. \end{aligned}$$

Equation 2 aims to find the selection vector x for each VQ that yields the maximum total number of unique answers collected for all VQs. The first constraint ensures the number of collected answers is within a certain budget, the second constraint ensures one selection of how many answers to collect per VQ, and the third constraint ensures the selection vector entries are binary. We solve this optimization equation using a mixed-integer linear programming based branch and bound method.

Mapping Answer Distribution to Expectation To calculate the answer expectation vector (AE vector) for each VQ in equation 2, we generate a mapping from each possible diversity score to an AE vector.

At training time, we begin by generating a unique diversity score list that indicates all possible diversity scores that can arise from having M answers (i.e., 57 possible scores for 10 answers). Recall these diversity scores capture that a crowd can arrive at the same number of unique answers with different underlying answer distributions one would observe from a crowd. Specifically, there are different probabilities of capturing all unique answers when collecting some number of answers based on the underlying answer distribution; e.g., the answer distribution of [“yes”, “yes”, “yes”, “no”, “no”, “no”, “no”, “no”, “no”, “no”] contains two unique answers with a 70%/30% split and the probability to capture both unique answers when collecting two answers is $P_k^{2u} = \frac{C_1^3 C_1^7}{C_2^{10}}$.

Next, we calculate the AE vector for each possible diversity score. When collecting M answers from M people, we generate a probability vector that contains probabilities of capturing 1 unique answer, 2 unique answers,..., q unique answers, denoted as $P_k^M = [P_k^{1u}, P_k^{2u}, P_k^{3u}, ..., P_k^{qu}]$. q is the maximum number of answers that we can collect, and we let $q = 10$ since there are 10 answers for each VQ in both datasets we use (i.e., VizWiz and VQA). The expectation is the value times its probability, so the AE when collecting M answers is:

$$E_k^M = \sum_{j=1}^{q} j P_k^{ju} \tag{3}$$

We calculate the AE for all possible values of M (e.g., when collecting 1 answer, 2 answers, up to 10 answers) using the following equation:

$$E_k^1 = 1P_k^{1u} + 2P_k^{2u} + 3P_k^{3u} +, ..., +10P_k^{10u} \tag{4}$$

Consequently, for VQ k, we have an expectation vector $[E_k^1, E_k^2, E_k^3, ..., E_k^{10}]$ that indicates the AE when collecting 1 answer, 2 answers, and so on. We calculate this AE vector for each possible answer distribution. Thus, in our system, each answer distribution maps each of the 57 possible diversity scores that can arise from 10 answers to an AE.

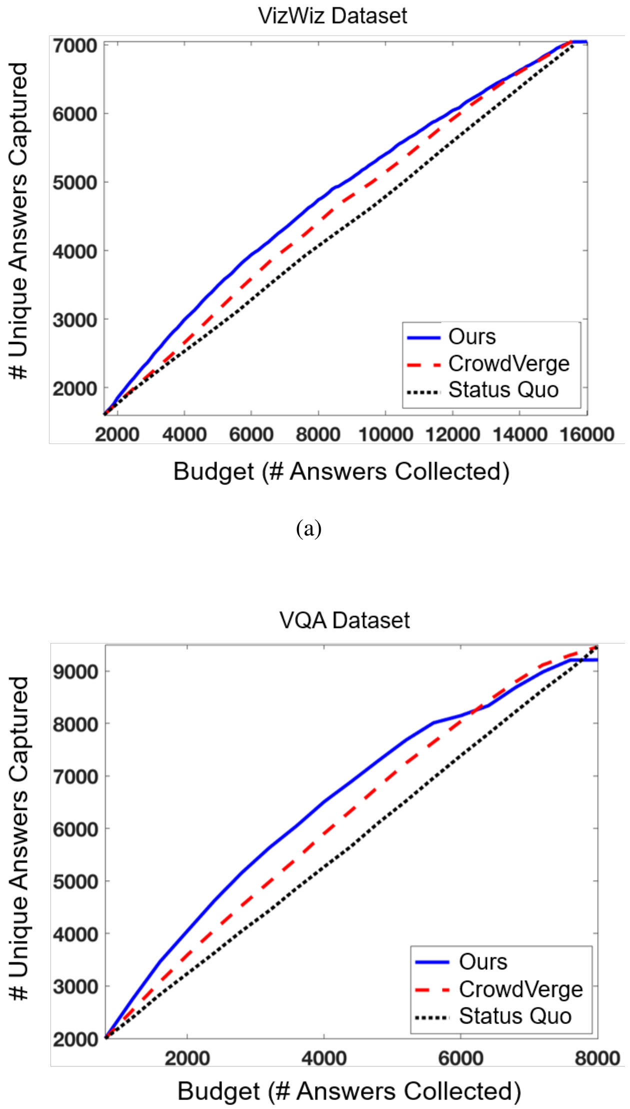
At test time, we use the predicted diversity score to identify the AE vector to use when solving our optimization equation 2. That tells us, for each VQ, what is the exact number of answers we need to collect in order to maximize the unique answers that will actually be captured over all visual questions in a given collection.

Experiments

Evaluation We select based on the selection results from equation 2 the available number of answers from the ground truth answer distribution and then count the number of unique answers that are actually captured from all ground truth answers. We do this all for all VQs in the test dataset.

Baselines We compare our system against the following:

- **Status Quo**: this system randomly collects one answer or 10 answers for each VQ to reflect the status quo for crowd-powered systems is selecting a pre-determined number of answers ranging from a minimum of one answer per VQ (Malinowski, Rohrbach, and Fritz 2015) to a maximum of 10 answers per VQ (Antol et al. 2015).



(a)

(b)

Figure 4: Total number of unique answers captured for a batch of visual questions for different human budget values for the (a) VizWiz dataset and (b) VQA dataset. As observed, our optimization approach consistently accelerates the collection of unique answers over baselines for most budget constraints.

- **CrowdVerge** (Gurari and Grauman 2017): CrowdVerge is a related state-of-art system which predicts how to best allocate a given budget of human-generated answers for a collection of VQs. This system first arranges its VQs based on its predicted score that answer disagreement will occur. It then collects one answer for all VQs and collects extra answers with the available extra budget only for visual questions most likely to lead to disagreement. In other words, it is a binary system – collect one answer or collect 10 answers per VQ.

Datasets We tested on all visual questions in the VizWiz test dataset and a subset of VQs in the VQA test dataset. For the latter dataset, we used a subset of 2,000 VQs which we curated to have an even distribution of diversity scores spanning from no diversity to complete answer diversity.

Results Results are shown in Figures 4a and 4b for both VizWiz and VQA test datasets. Shown are the number of unique answers that actually are captured across different budget constraints.

As observed in Figure 4, our budget allocation approach typically performs better than both baselines for both datasets. The advantage of our approach over the baselines is greatest in the lower budget zone and tapers off in the higher budget zone. While our approach typically captures more unique answers than the top-performing CrowdVerge baseline, an exception happens in the VQA dataset in the high budget zone where our approach performs comparable and slightly poorer (i.e., Figure 4b; range from 6,000 to 8,000 VQs). In this zone, most of the unique answers for each VQ already are captured. We hypothesize that the imperfect fine-grained predictions result in the system misallocating where to collect more answers when only a few unique answers remain to be collected.

We attribute the overall advantage of our approach over the baselines (i.e., CrowdVerge and Status Quo) to the baselines not being able to capture less than a pre-defined maximum number of answers (i.e., 10 answers) in order to collect all unique answers for a VQ. Specifically, existing crowd-powered systems greedily assign a fixed number of answers per VQ; e.g., CrowdVerge collects one answer when agreement is expected and the maximum possible number of answers when disagreement is expected. In contrast, our approach can allocate human effort in a fine-grained manner that permits best use of the budgeted answers to collect all unique answers for the entire budget by permitting a different number of answers to be collected for each VQ when trying to collect more than one unique answer for a VQ.

Conclusion

Existing VQA systems do not account for the fact that different VQs lead to different degrees of answer diversity. We propose a novel problem of predicting the answer entropy directly from a VQ and offer new methods which predict well for this task. We also demonstrate the benefit of such predictions in crowdsourcing answers with a limited manual annotation budget. Our proposed system outperforms today's state-of-art crowdsourcing system to efficiently collect the diversity of unique answers for a collection of VQs.

Acknowledgments

This work is supported by NSF awards to Danna Gurari (IIS-1755593) and Kristen Grauman (IIS-1514118). We gratefully acknowledge the computing support from Texas Advanced Computing Center (TACC).

References

Ali, A. R.; Shahid, U.; Ali, M.; and Ho, J. 2017. High-level concepts for affective understanding of images. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 679–687. IEEE.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Learning to compose neural networks for question answering. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 1545—1554.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. Vizwiz: Nearly real-time answers to visual questions. In *ACM symposium on User interface software and technology (UIST)*, 333–342.

Goyal, Y.; Mohapatra, A.; Parikh, D.; and Batra, D. 2016. Towards transparent AI systems: Interpreting visual question answering models. In *International Conference on Machine Learning (ICML) Workshop on Visualization for Deep Learning*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gurari, D., and Grauman, K. 2017. CrowdVerge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3511–3522. ACM.

Gurari, D.; Jain, S. D.; Betke, M.; and Grauman, K. 2016. Pull the plug? predicting if computers or humans should segment images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 382–391.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jain, S. D., and Grauman, K. 2013. Predicting sufficient annotation strength for interactive foreground segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1313–1320.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *IEEE European Conference on Computer Vision (ECCV)*, 740–755.

Lu, J.; Lin, X.; Batra, D.; and Parikh, D. 2015. Deeper lstm and normalized cnn visual question answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *IEEE European Conference on Computer Vision (ECCV)*, 1–9.

Nguyen, A. T.; Wallace, B. C.; and Lease, M. 2015. Combining crowd and expert labels using decision theoretic active learning. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 860–868.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015a. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015b. Rethinking the inception architecture for computer vision. *CoRR* abs/1512.00567.

Yang, J.; Sun, M.; and Sun, X. 2017. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, 224–230.

Yeh, T.; Lee, J. J.; and Darrell, T. 2008. Photo-based question answering. In *ACM International Conference on Multimedia*, 389–398.

Zhang, J.; Ma, S.; Sameki, M.; Sclaroff, S.; Betke, M.; Lin, Z.; Shen, X.; Price, B.; and Měch, R. 2016a. Salient object subitizing.

Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016b. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; Xie, W.; Jiang, X.; and Chua, T.-S. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 2016 ACM on Multimedia Conference*, 1385–1394. ACM.