Generating Gold Questions for Difficult Visual Recognition Tasks

Diane Larlus, Florent Perronnin
Xerox Research Center Europe
Meylan, France
firstname.name@xrce.xerox.com

Pramod Kompalli, Vivek Mishra
Xerox Research Center India
Bangalore, India
firstname.name@xerox.com

Abstract

Gold questions are a standard mechanism to detect insincere workers on crowdsourcing platforms. They usually rely on the assumption that workers should obtain perfect accuracy on the task. In this work, we are interested in crowdsourcing difficult multi-class visual recognition tasks, for which this assumption is not met, and we propose a novel method for generating gold questions in this context.

1. Introduction

This work considers the problem of crowdsourcing multi-class visual recognition tasks: given a query image, we wish a set of workers to annotate it with the correct label. We assume that the task is provided as a multiple choice question: given the image, and candidate classes, the worker should select one class within this restricted set (see Fig. 1). These candidate classes are selected using the output of an algorithm, some prior data or any complementary information (for instance, the meta-data of an image).

We focus on classification problems that are too difficult to be solved by humans or computers alone (*e.g.* because they involve a large number of classes) and we wish to combine their complementarities in a hybrid system [2]. An example of such a difficult task is fine-grained visual classification, which involves recognizing sub-ordinate categories of a base-level category. In this work, we experiment with bird species classification, on the 200 classes of the Caltech-UCSD-Birds (CUB) dataset [2]. We consider a sequential pipeline where a computer vision algorithm predicts the 5 most likely classes, and the annotator selects the final label from among them. This pipeline aims to improve the accuracy of the fully automatic classification system with a human in the loop. Short-listing is advantageous as browsing the complete list of classes would be unmanageable. However the task is still challenging as, quite often, short listed classes are also difficult to distinguish.

One standard issue with crowdsourcing is that unreliable crowdworkers or bots could easily compromise the results if not properly detected and filtered out. Hence, a challenge

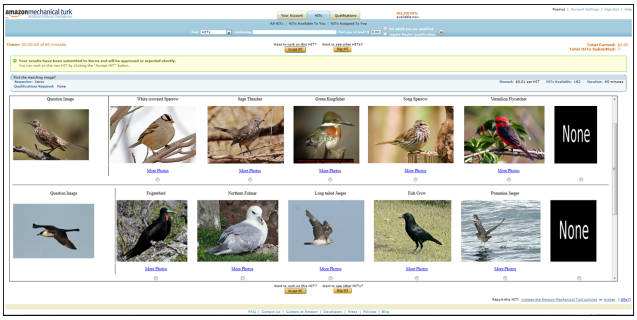

Figure 1. Screenshot of the AMT HIT as displayed to the worker. The query image (on the left) has to be assigned to one of the 5 proposed classes (or to the "none" option).

is to detect insincere workers automatically. A common mechanism is to place a gold question in each HIT (Human Intelligence Task), *i.e.* a question for which the answer is known *apriori*. The assumption is that, if a worker provides the correct answer for the gold question, then he/she is sincere and his/her other input can be trusted. This is such a common mechanism that crowdworkers are aware of the presence of a gold question. Lazy workers could search and answer only this one, earning money without doing actual work on the other questions. Consequently, a good gold question should be difficult to spot.

Designing gold questions is trivial for simple visual recognition problems (*e.g.* bird vs car). For such tasks one can expect 100% accuracy from the workers (or very close to). Therefore, the gold questions may be sampled randomly from the standard questions – one just needs to annotate the corresponding images. However, this random sampling strategy does not apply to our scenario where we wish to crowdsource tasks which are difficult for humans. This problem was addressed in other domains [3], but it is unclear how to apply this technique to visual data.

We propose a novel approach to designing gold questions for difficult multi-class visual classification problems. It involves two key concepts: (i) the popularity of the classes and (ii) a measure of distance between classes. Experiments on Amazon mechanical Turk (AMT) validate our approach.

2. Generating Gold Questions

While gold questions and standard questions are both multiple choice questions, a difference is that in the former case we would typically ensure that the correct class is within the possible choices. The correct class (which is known a priori) is later referred to as the *positive* class while the other classes are referred to as *negative* classes. The main challenge in designing a gold question is therefore in the choice of these positive and negative classes to bias the sampling of gold questions toward simpler questions.

The proposed approach relies on two key concepts to choose positive/negative classes: (i) *class popularity* which is used to sample positive classes and (ii) *class-to-class distance* which is used to sample negative classes. We first describe these two concepts.

Class popularity. Even for classification problems that are difficult overall, some classes are easier to recognize than others. Such classes make good candidates for the positive class. We assume that common or *popular* classes will be easier to recognize for a non-expert, where a quantitative measure of popularity must be defined. A plausible measure would be the number of hits returned by a text search engine (such as Google search) or a visual search engine (such as Google image search). In our experiments, we used the former approach. As an example, the 5 most popular classes out of the 200 bird classes of [2] are Green Kingfisher, Field Sparrow, Cardinal, Mockingbird and Mallard.

Class-to-class distance. In fine-grained problems, two classes can be very similar, but pairs of classes can be chosen to be different enough so that even a non-expert will easily distinguish them. We propose to choose negative classes so that a worker will be confident that the query image does not belong to any of them. For this purpose, we measure a distance between classes and choose the negative classes from among the most distant ones. We rely on *a priori information* such as attributes and ontologies, and use the notion of class embedding [1], *i.e.* we represent classes as vectors in a Euclidean space, to measure class-to-class distances. We choose attributes to embed classes: each class is represented as a vector of attribute relevances. In our experiments, we use the 312 attributes provided with CUB [2], meaning that each class is represented as a 312D vector.

Trade-off. Gold questions should display the right trade-off: they should not be so easy that they can be spotted by workers, but they should not be so difficult that they are poor indicators of the worker motivation. We combine our measures of popularity and class-to-class distance to find a balance. We create a pool of the p most popular classes and sample the positive class from this pool. With a small p value, easier classes are selected, so there is a higher chance that a worker will recognize the class, but this lowers the diversity of gold questions, and the gold question will be easier to spot after a few HITs. On the other hand, a larger p value increases the diversity in the gold questions (gold questions are more difficult to spot), but the resulting gold questions are more difficult. For each class, we also create a pool of the n most distant classes and sample negative classes from this pool. We have with n the same trade-off as with p. A large n makes the tasks more difficult but introduces variety. A small n makes the tasks easier but less varied. In all our experiments, we used $p = n = 10$ which seemed to provide a good trade-off.

Table 1. A summary of worker performance on the labeling task, to highlight effect of the proposed Gold question design.

accuracy on gold questions	78.3%
accuracy on test questions	42.5%

3. Experimental Validation

Experiments were conducted on the CUB-200-2011 dataset [4] using the standard training/test split. We use a sequential pipeline where a computer vision algorithm predicts the 5 most likely classes. The annotator is asked to review these 5 classes and to mark the correct one. The task is less tedious for the annotator who has to choose one class out of 5 instead of 1 out of 200. Each HIT is composed of three questions, two of which are real questions (query image sampled from the test set) and one of which is a gold question (query image sampled from the training set). The order of the three questions is randomized. These HITs were deployed on Amazon Mechanical Turk and distributed to real workers.

Results. Table 1 shows the accuracy obtained by crowdworkers on the gold questions and on the real questions. Overall, the accuracy over gold data is significantly higher. This shows that the proposed strategy generates questions that are much easier for the workers than standard test questions. Additionally, an internal study showed that gold questions were difficult to detect in most of the cases, and for most of the annotators. Finally, we looked at the accuracy on real questions, considering the subset of images for which the gold question was answered correctly, and incorrectly respectively. These accuracies (43.4% and 39.4% respectively) are comparable, which also implies that the gold questions were not that easily detected (or that there were very few attempts to cheat the system).

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2

[2] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *ECCV*, 2010. 1, 2

[3] D. Oleson, A. Sorokin, G. P. Laughlin, V. Hester, J. Le, and L. Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Human Computation*, 2011. 1

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2