Interactive Consensus Agreement Games For Labeling Images

Paul Upchurch and **Daniel Sedra** and **Andrew Mullen** and **Haym Hirsh** and **Kavita Bala**
Computing and Information Science, Cornell University

Abstract

Scene understanding algorithms in computer vision are improving dramatically by training deep convolutional neural networks on millions of accurately annotated images. Collecting large-scale datasets for this kind of training is challenging, and the learning algorithms are only as good as the data they train on. Training annotations are often obtained by taking the majority label from independent crowdsourced workers using platforms such as Amazon Mechanical Turk. However, the accuracy of the resulting annotations can vary, with the hardest-to-annotate samples having prohibitively low accuracy.

Our insight is that in cases where independent worker annotations are poor more accurate results can be obtained by having workers collaborate. This paper introduces consensus agreement games, a novel method for assigning annotations to images by the agreement of multiple consensuses of small cliques of workers. We demonstrate that this approach reduces error by 37.8% on two different datasets at a cost of $0.10 or $0.17 per annotation. The higher cost is justified because our method does not need to be run on the entire dataset. Ultimately, our method enables us to more accurately annotate images and build more challenging training datasets for learning algorithms.

Introduction

Creating large-scale image datasets has proved crucial to enabling breakthrough performance on computer vision tasks (Krizhevsky, Sutskever, and Hinton 2012). A significant barrier to the creation of such datasets has been the human labor required to accurately annotate large collections of images. Increasingly such datasets have been labeled through innovations in the area of crowdsourcing and human computation, whereby the efforts of large numbers of often inexpert Internet-based workers are used to yield data of surprising accuracy (Deng et al. 2009; Kanefsky, Barlow, and Gulick 2001; Raddick et al. 2007; Russell et al. 2008; Sorokin and Forsyth 2008; Von Ahn and Dabbish 2004; Westphal et al. 2005). The introduction of the Amazon Mechanical Turk crowdsourcing platform in 2006 in particular quickly led to its adoption in various image recognition tasks (Barr and Cabrera 2006; Douglis 2007; Sorokin and Forsyth 2008; Spain and Perona 2008; Deng et al. 2009).

The most common approach seeks labels for each item from multiple workers and assigns as the item's label the "majority vote" among those provided by the workers (Sheng, Provost, and Ipeirotis 2008; Snow et al. 2008; Sorokin and Forsyth 2008). Even as increasingly sophisticated approaches have been developed for aggregating the labels of independent workers there can still be significant variability in the quality of such data. Many samples receive very low agreement when labeled by multiple MTurk workers. For example, (Bell et al. 2013) collected approximately five labels per sample, and for those with 60% agreement (3 out of 5 agreement, after removing votes from low-quality workers) many of these *low-agreement samples* were mislabeled, making them unsuitable for training a high-accuracy model. As a result, (Bell et al. 2015) only used samples with at least 80% agreement (*high-agreement samples*).

Relying solely on high-agreement data can bias the data to easy-to-classify cases, which may not mirror the diversity of cases to which the trained model will be applied, negatively impacting the quality of the learned model. Further, low-agreement samples can represent cases that fall near decision boundaries, reflecting data that can be particularly valuable for improving model accuracy (Lewis and Gale 1994).

Ultimately, the problem is that MTurk workers have a high error rate on low-agreement data. If the *MTurk error rate* is the fraction of mislabed samples (compared to, for example, expert labelers or some other appropriate notion of ground truth), our goal is to reduce it so that low-agreement samples become more accurate, and thereby more useful for training computer vision models.

Reducing the MTurk error rate is not easy. The key characteristic of low-agreement data is that *independent* workers cannot agree on the label. Getting more answers from independent workers or encouraging them with agreement incentives does not get us better answers (as shown in the Experiments section). Instead, we take an approach where labels are assigned through a collaborative process involving multiple workers. We find our inspiration in two previous works. First, the graph consensus-finding work of (Judd, Kearns, and Vorobeychik 2010; Kearns 2012) showed that a network of workers can collectively reach consensus even when interactions are highly constrained. Next, the ESP Game (Von Ahn and Dabbish 2004) showed how to obtain labels from non-communicative workers by seeking agreement around a shared image. In this paper, we show how to label images by casting it as a graph consensus problem which seeks agreement between multiple, independent consensus-finding cliques of workers. We find this pattern to be effective on difficult-to-label images.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Our approach achieves greater accuracy at a greater cost than majority voting, and thus the approach is intended for use in the context of creating large-scale databases of labeled images that are not biased towards easy-to-classify samples. This approach should be used *after* using majority voting to gather labels, *only* to refine the labels of low-agreement samples.

Related Work

Others have considered different ways in which the confidence in an item's annotation may differ across items, and its implications. For example, Galaxy Zoo created "clean" and "superclean" datasets by constraining data to those in which at least 80% or 95% of workers agree on an item's label (Lintott et al. 2008), and (Hsueh, Melville, and Sindhwani 2009) use worker disagreement so as to remove ambiguous data and improve the classification of sentiment in political blog excerpts. (Dekel, Gentile, and Sridharan 2012; Deng et al. 2009; Parameswaran et al. 2014) consider how disagreement among obtained labels can be used to signal that more labels should be obtained for such items and (Ipeirotis et al. 2014) uses the uncertainty of a trained model on a dataset to target items for which additional labels can improve learning. (Wallace et al. 2011) show that learning can be improved if workers are allowed to specify their own low confidence in labeling an item so that it can be routed to more expert workers, while (Shah and Zhou 2015) proposes a payment mechanism to incentivize workers to only respond to items they have confidence about. In settings where workers provide labels for multiple items it is possible to learn models of worker performance based on factors that include the extent of ambiguity of an item's annotation (Bachrach et al. 2012; Mao, Kamar, and Horvitz 2013; Wauthier and Jordan 2011). This work is complementary to such efforts, in that rather than persist with methods in which workers assign labels in isolation, we seek to improve annotation accuracy by employing consensus agreement games in which workers act collaboratively to assign labels to items.

This work seeks more accurate annotations by engaging workers in a gaming setting similar to the ESP Game (Von Ahn and Dabbish 2004). The ESP Game gives images to pairs of players then annotates the image and rewards the players if they enter identical tags. A number of variants to the ESP Game have also been proposed. The Make a Difference game (Thogersen 2013) is similar to the ESP Game, but requires three workers to agree on a tag for it to be accepted. Further generalization to number of players and thresholds was made by (Lin et al. 2008). KissKiss-Ban (Ho et al. 2009) is a three-person game in which two players attempt to enter matching tags after a third player enters taboo words that, if entered by both other players, gives the third player points. The ESP Game and its variants bear the most similarity to consensus agreement games in that they look for the same label produced by multiple players in an interactive setting. The difference between these games and consensus agreement games is that the latter allows players to see and respond to the labels provided by the others in their clique. This allows players to guide each other to the correct answer in groups whose social dynamics avoid many of the frailties found in real-world decision-making groups (Levine and Moreland 2008; Sunstein 2006).

Our work is inspired by that of (Judd, Kearns, and Vorobeychik 2010; Kearns 2012), who explored the ability for a group of individuals to solve graph consensus problems through limited local interaction as a function of the topology of the network connecting them. Our approach is different because we require a non-collaborative agreement between disjoint subgraphs and design financial incentives which drive players toward the correct consensus rather than just any consensus.

Finally, it has been shown that crowdsourcing outcomes can be improved if worker compensation depends on matching answers with those of one or more other workers. (Huang and Fu 2013b; Faltings et al. 2014) show improved outcomes if bonuses are given for a worker matching that of another single worker. (Rao, Huang, and Fu 2013) showed similar improvements when bonuses are based on a worker matching the majority vote of a set of other workers, whereas (Kamar and Horvitz 2012; Dasgupta and Ghosh 2013; Radanovic, Faltings, and Jurca 2016) provide reward schemes that match a worker's answers to more complex functions of the answers of other workers. Unlike this previous work, we use agreement to determine if all the workers (within multiple collaborative consensus decision-making cliques) have converged to the same answer. Nonetheless, we seek agreement and could benefit from the forms of agreement explored in previous works with the caveat that since our goal is to label difficult-to-label samples, lessons learned about agreement from experiments on entire datasets may not apply.

Method

Our goal is to reduce MTurk error rate by having multiple workers interactively find a consensus for low-agreement samples. In the manner of (Judd, Kearns, and Vorobeychik 2010) we formulate a consensus graph problem of $2N$ nodes organized into two disjoint cliques of N. The graph is solved when all $2N$ nodes are assigned the same label. The graph problem is made tractable by showing both cliques the same image. We hypothesize that this is sufficient information for the $2N$ players to solve the graph (based on the success of the ESP Game).

We explicate this pattern and describe how to implement it in the subsections below.

Consensus Agreement Games

Consensus agreement games (CAG) is an instantiation of the pattern described above. Namely, we take each clique as an N-way collaborative labeling game where the potential labels are constrained to the labels from a previous majority-vote labeling process plus enough random labels to make a set of K labels. K should be small so that we make full use of the information we gathered in the previous labeling process yet large enough so that cliques have a low probability of agreeing if the players collaboratively guess randomly.

During the game a player can see the selections of the other $N-1$ players and can freely change their own selection. No other interaction is allowed between players.

A game ends after a fixed time limit whereupon if all players have selected the same label then the game has reached a *consensus*. Two games, operated independently, make one CAG in which we look for *agreement* in the consensus outcomes.

A pair of games has four possible outcomes:

- Consensus agreement: both games achieve a consensus and they agree.
- Consensus disagreement: both games achieve a consensus but they disagree.
- Solitary consensus: one game achieves a consensus.
- No consensus: both games fail to achieve a consensus.

A label which achieves consensus agreement is deemed to be confident enough to be taken as an annotation for the sample.

Player labeling strategies will be determined by the game payoffs, P_i. We define three outcomes for a game (which is paired with a second game):

- Consensus agreement (P_1): a consensus is reached and it matches the consensus reached by the second game.
- Clique consensus (P_2): all players select the same label but it is not a consensus agreement.
- Discord (P_3): not all players select the same label.

We want to choose payoffs which support our goals. First, the payoff for a clique consensus must be higher than the payoff for discord. This incentivises the players to adopt a labeling strategy which is different from independent voting. However, players may adopt simple strategies to get a clique consensus (e.g., always follow the first person that votes). Therefore, the payoff for a consensus agreement must be higher than the payoff for a clique consensus. This incentivises the players to vote for the truth since they have no other way to interact with the second group of players. Thus, the payoffs must satisfy $P_1 > P_2 > P_3$.

Games have a fixed duration but not all images need the same amount of time to label. We use a 120 second timer and averaged out the needed time by packing 8 images into a single game. Accordingly, we created a payment schedule in quanta of 1/8 cents where $P_1 = \$0.02125$, $P_2 = \$0.00625$, and $P_3 = \$0.00125$ so that the maximum payout per game is $\$0.17$ and the minimum payout is $\$0.01$. These values were selected based on the results of preliminary experiments.

Worker Experience

In this section we describe one of our experimental games from the perspective of an MTurk worker.

1. The HIT reward is \$0.01 but the title advertises "(with bonuses)". The HIT description informs the worker that they will work "with other people". We require that workers have a 95% approval rate and at least 100 approved HITs.

2. A worker previewing the HIT is told that they will "play a 120 second game with other people" and the earnings are described as "If you and your group play well, you are able to each make up to \$0.17 for 120 seconds of your time. This works out to \$5.10 per hour. The base rate for your time is \$0.01 for up to 4 minutes of your time. If your group agrees on the same label, then you will receive a bonus of \$0.005 (each game consists of 8 labels, so up to an additional \$0.04 per game). If your group agrees on the correct label, then you receive an additional \$0.015 per label (up to \$0.12 per game)."

3. After accepting a HIT the worker is presented with instructions on how to play the game, definitions of the categories, and information on common mistakes. In particular, they are told "You are allowed and encouraged to change your vote as you change your opinion of what the material is.", "You will be able to see, in real time, the choice of the other players in their own rows." and "At the end of the game, you want your votes to all be the same, if a consensus is reached then you are given a bonus. Remember you get a bigger bonus if you all choose the same label and the label is correct."

4. When the worker presses a button to indicate they are ready to play then they are placed on a game queue. The worker is told that they are waiting for people to join and that "If you wait for 3 minutes and your game doesn't start, then you will get money if you stay on the webpage and submit the HIT when you are instructed to submit it."

If the worker waits for 3 minutes then they are moved into the exit survey directly and will receive the HIT reward of \$0.01 (which is $8 \times P_3$) for their time.

5. During a game a worker is told "You can change your vote as many times as you want. Remember you get a bigger bonus if everyone picks the same label for each pair of images and the label is correct." Below this the worker is shown 8 pairs of images. Each pair is a crop of the sample to be labeled, a crop of the entire picture and buttons indicating the current vote of each player (Figure 1). Clicking on a crop shows a higher resolution version. They are also shown the current votes from each player for each sample.

Below the final pair the worker is told "As long as the game is running, the other players can change their votes. You may want to change your vote depending on what the others players do." At the bottom of the page the 8 pairs are repeated with much smaller images and the same vote buttons (the same as Figure 1 except the images are 85% smaller). This compact summary lets the worker view votes and vote without having to scroll the page excessively.

The time remaining (updated each second) is displayed at the top, middle and bottom of the page. When the game ends each worker is sent to an individual exit survey.

6. On the exit survey page the worker is told "You have earned at least \$0.XX and will receive more if your group agrees with the second group. You must press 'Submit HIT with bonuses' to receive all the money you have earned (may take 48 hours)." where XX is 1 cent plus 1/2 cent per clique consensus. The worker is given the opportunity to provide optional feedback. We ask: "Do you have any feedback about the image labeling task?" and "Do you have any feedback about the webpage or game?"

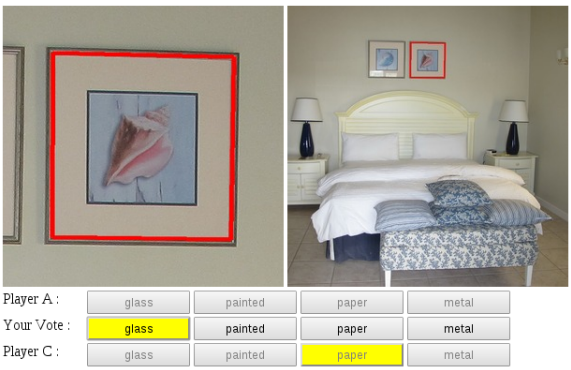

Figure 1: The collaborative labeling interface. **Top:** The material shape to be labeled is outlined in red. Clicking on either image will show a higher resolution image. The left image shows a crop of the shape, the right image shows where the shape appears in the photograph. **Bottom:** Buttons indicate the current selection of each player. Here, Player A has not yet made a selection, Player B (the current player) has selected glass and Player C has selected paper. Players may change their selection at any time, and the other players will see updates in real-time, but once a player has made a selection it is not possible for that player to return to the initial 'no-selection' state.

Figure 2: Examples of low-agreement samples in MINC. **Top-Left:** This bowl received 3 votes for ceramic and 2 votes for glass. **Top-Right:** This pig received 3 votes for ceramic, 1 vote for plastic and 1 vote for foliage. **Bottom-Left:** This bottle received 3 votes for glass and 2 votes for plastic. **Bottom-Right:** This door received 3 votes for painted and 2 votes for wood.

Experimental Datasets

We want to evaluate CAG on natural images which are difficult to label yet not ambiguous. In this section we describe how we prepared two datasets which fit these criteria.

MINC

The Materials in Context (MINC) Database[1] (Bell et al. 2015) identifies segments of uniform materials in images and annotates them with material labels (e.g., wood, plastic, etc). We choose this dataset since it has 5 votes per sample and the data is hard to classify while requiring only an understanding of everyday materials. We define low-agreement samples as those which have exactly 3 out of 5 votes in agreement (in the future this definition could be expanded to include samples with even lower agreement among workers). This definition matches the experimental settings of MINC since they defined high-agreement as four matching votes. See Figure 2 for examples.

We need ground truth to compute an error rate but ground truth is not available since the samples came from Internet photographs. Instead, we use experts to create a high-quality *expert truth* and rank different methods by comparing worker error rates against expert truth.

Three experts examined random low-agreement samples from the 10 largest categories of MINC and assigned annotations to samples which were unambiguous. In total, the experts annotated 456 samples with expert truth. The experts labeled as closely to truth as they could, and so were not limited to the 10 largest categories. Thus, we ended up with more than 10 categories.

Places

The MIT Places Database[2] (Zhou et al. 2014) is a collection of images categorized by scene type. Since we do not know a priori which samples are difficult, we first selected 12 categories in 5 mutually confusing groups and collected 5 votes for 2400 images, 200 from each category. We then looked at the samples which received 3 of 5 agreement and assigned expert truth. Two of the mutually confusing groups had sufficient samples for experimentation. In order to prevent workers from being biased we randomly subsampled the largest categories so that no category was more than twice the smallest mutually confusing category. This rule only applied to the hotel room category, which was significantly over-represented since many bedrooms were actually hotel rooms.

We ended up with a low-agreement dataset of 22 bedroom, 32 hotel room, 16 nursery (the preceding constitute one mutually confusing group), 34 coast, 44 ocean and 28 river (the second mutually confusing group) images.

Experiments

In this section we evaluate performance. We first clarify the terminology of *labels* and *annotations*. A label is assigned to a sample by a worker. An annotation is assigned to a sample by a method. Not all samples receive an annotation. For example, if 5 workers label a sample with 5 unique labels then majority vote does not assign an annotation since no label achieved a majority.

[1] http://opensurfaces.cs.cornell.edu/publications/minc/

[2] http://places.csail.mit.edu/

Baseline method. We want to compare the MTurk error rate of CAG against a baseline. We selected majority vote of 7 unique workers (Vote7) as a baseline since majority vote is commonly used in practice and 7 voters prevents ties and is nearly the same number of people as in CAG when $N = 3$. We cannot use the original votes for the baseline since those workers chose from more than 4 possible choices (34 choices for MINC and 12 for Places). Instead, we collected new votes using the same protocol as CAG (4 choices, the original votes plus random labels). We set the per-label cost to \$0.004 and this decision was guided by the cost of MINC annotations (reported as \$0.0033 in Table 1 of (Bell et al. 2013)).

CAG settings. We took $K = 4$ so that the chance of random agreement is low. The clique size is a free parameter. We experimented with $N = 2$ and $N = 3$ since the smallest possible clique will be the most cost effective and an odd-sized clique can prevent stalemates due to ties. We did not experiment with larger N since the cost per annotation would be too high. We report performance for both values of N on both datasets in Table 1 but for brevity we report results only for the best settings ($N = 2$ for MINC and $N = 3$ for Places) in the remainder of this section.

Comparison statistics. If one were labeling a dataset then two cliques would label each sample. However, this natural experiment would give us very little data for computing error rate. Instead we showed each sample to 3 cliques and formed all possible pairings. This gave us 3 times as much data for only 50% more experimental cost. We then used bootstrap sampling (1000 trials) to estimate performance statistics and standard errors (SE). Bootstrap sampling was not needed for the baseline method since Vote7 annotated a high-fraction of samples.

The MTurk error rate is $(1-\text{Precision})$ as defined in Equation 1. For each method the true positives are annotations which match the expert truth and the false positives are those which do not.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \qquad (1)$$

We also report cost per annotation and *throughput* (the fraction of samples which receive an annotation). CAG throughput is reduced for three reasons: a player does not cast a vote (because they abandon the game or run out of time), players disagree or cliques disagree. We remind the reader that our goal is to augment an already large dataset with correctly annotated hard-to-label samples. Thus, CAG prioritizes accuracy over throughput which leads to higher cost per annotation.

MTurk error rate. We find that CAG has a lower error rate (by at least 3 SE) than Vote7 on both datasets. See Table 1 for a summary. For MINC we had 456 low-agreement samples of which 427 received a Vote7 annotation for a throughput of 0.9364. With 344 correct annotations the MTurk error rate is 0.1944. CAG has an estimated MTurk error rate of 0.1186 with standard error 0.01257 and throughput of 0.4997 with SE 0.01335. For Places we had 176 low-agreement samples of which 170 received a Vote7 annotation for a throughput of 0.9659. With 130 correct annotations the MTurk error rate is 0.2353. CAG has an estimated MTurk error rate of 0.1490 with SE 0.02493 and throughput of 0.3842 with SE 0.02216.

Cost per annotation. The cost of a Vote7 annotation is determined by throughput and the reward of \$0.20 per HIT of 50 images which gives a cost of $7 \times \$0.004/0.9364 = \$0.0299$ per MINC annotation and $\$0.0290$ per Places annotation. For CAG the costs are variable. We pay workers for their time (\$0.01 per game), a clique for achieving a consensus (\$0.005 each) and a pair of cliques for achieving a consensus agreement (\$0.015 each). Our estimated cost per MINC annotation is \$0.104, SE \$0.00105. The breakdown is 9% for worker time, 33% for clique consensus bonuses, and 58% for consensus agreement bonuses. Our estimated cost per Places annotation is \$0.168, SE \$0.00392. The breakdown is 12% for worker time, 35% for clique consensus bonuses, and 54% for consensus agreement bonuses. Since the costs for each dataset are similar we report the combined cost in Table 1.

Acquiring more votes. We hypothesized that since we specifically selected samples which are known to be labeled with low confidence (and empirically found to have low accuracy) by an independent worker method then acquiring even more independent labels would not ultimately converge to the correct annotation. To test this hypothesis we used the baseline method to gather 21 votes per sample for the MINC dataset and the measured error rates at 7, 11 and 21 votes are 0.1944, 0.1900 and 0.1962, respectively. This confirms that for low-agreement samples additional independent votes does not converge to lower error rates.

High-agreement samples. We hypothesized that CAG would work best on low-agreement samples. For a comparison, we ran CAG and Vote7 on 352 MINC samples with 4 out of 5 agreement. The Vote7 MTurk error rate is 0.0698 with a throughput of 0.9773. The estimated CAG error rate is 0.05603 with SE 0.009223 and throughput of 0.5901 with SE 0.01496. The CAG mean is lower but within 1.5 SE so we find little difference between the two methods.

Consensus game. How important is agreement for CAG? We can conduct a single consensus game (CG) without looking for agreement between two cliques. This is attractive since the cost per annotation can be halved. The obvious downside is that there is no longer a financial incentive for the clique to label the image correctly.

For this method there are only two payoffs and they must satisfy $P_2 > P_3$. In our experiments we use $P_2 = \$0.02125$ and $P_3 = \$0.00125$ so that the pay per worker is the same as the consensus agreement games but the cost per annotation is approximately half.

Method	MINC	Places	Cost
Vote7	0.1944	0.2353	\$0.0296
CAG N=2	**0.1186**	0.2407	\$0.1043
CAG N=3	0.1410	**0.1490**	\$0.1685

Table 1: MTurk error rate for low-agreement samples and cost per annotation. Consensus agreement games (CAG) reduces the error rate. The clique size (N) which gives the best performance depends on the dataset although $N = 3$ outperforms the baseline on both datasets.

We tested CG on MINC and found it much worse than Vote7. The MTurk error rate is 0.2730 and the throughput is 0.6667. Clearly workers found the HITs very lucrative since they enthusiastically snapped up CG HITs as soon as we posted them. We find that consensus does not perform well but it becomes a useful mechanism when paired with agreement.

Agreement incentives. How important is the clique consensus compared to the incentives for agreement? We conducted experiments on MINC which included the agreement incentives but excluded the collaborative clique game. This experiment's HITs are similar to a Vote7 HIT but the payment schedule is \$0.01 for the worker's time and \$0.02 per annotation (two workers agree). HITs contain 20 samples and have a time limit of 15 minutes. We found that the estimated MTurk error rate is 0.2054 with SE 0.01401 and throughput is 0.5904 with SE 0.01353 which makes this method on par with Vote7 (within 0.8 SE). Although peer incentives are effective, they become even more effective when combined with consensus games in our setting of difficult-to-label samples.

Analysis

We want to look more deeply into how the workers played the game so we instrumented some MINC CAG games on low-agreement samples and recorded all moves made by the players. We used $N = 3$ since larger cliques may have more interesting behaviors and we used MINC since it has more categories. There are two instrumented games per sample so the game pairs were used directly without the data augmentation described in the previous section.

Based on preliminary experiments we gave players 120 seconds to label 8 images. Yet, we found that the mean time of the last player activity was 77 seconds with deviation 20. This indicates players were under some time pressure and increasing the time allotted may reduce error further.

Players are allowed to change their labels so we looked for changes in the 2568 labelings. We found that in 186 cases the worker changed their label and in 162 of those cases the final label was different from the initial label.

How many different labels get considered as potential annotations for a sample? In Figure 3 we look at a histogram of the number of different labels a worker considers for a sample. In 2382 cases workers selected one label, a worker

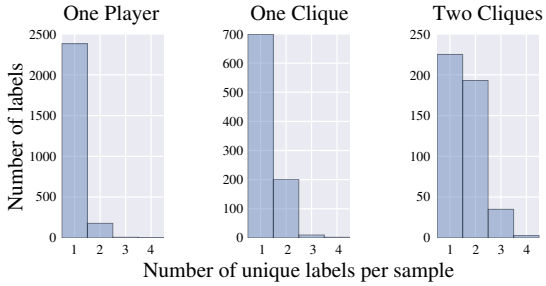

Figure 3: **Left:** The number of different choices made for a sample by a single player. **Middle:** The number of different choices made for a sample by a single clique. **Right:** The number of different choices made for a sample by both cliques. Single workers consider a small number of choices whereas the cliques increase the variety of choices considered for annotation.

selected two different labels 178 times, three different labels 7 times and all four possible labels once. We see that most players do not change their label (i.e., their first label is final) and when they do change their label they very rarely pick a third or fourth choice. What about a clique? As we can see, a sample receives more different votes from a clique than it does from just one worker. This implies that a clique does increase the amount of knowledge which is applied to a sample. How about across both cliques? Although the cliques cannot communicate directly we can still see that the label space is explored even more by 6 workers. These observations indicate the not all workers have the same knowledge. Each person is bringing their own opinions and sharing them with their clique.

Does this extra knowledge from fellow clique members help or hinder? We looked at how often an initially incorrect label was corrected and vice versa. We found that 100 initially incorrect labels were corrected and 44 initially correct labels were misled. This indicates that the cliques help guide members to the correct answer more often than they drive them away.

We know that players changed their labels, but how often do they do it? For example, do they change from A to B then back to A? If each change was decisive then there would have been 195 changes — the minimum number of changes needed (which was computed from the number of unique labels per sample made by players). We found that players changed their labels 230 times, therefore there were instances where players switched their labels more often than necessary. This could have been attempts to signal other players and/or it could be evidence of indecisiveness.

Observations About Workers

MTurk workers seemed to enjoy the game. We got lots of positive feedback: "Thanks, this was pretty fun and different way of doing it!", "It was my first time playing game like this. Much better than simple labeling", "Fun game! Well done HIT!", "That was fun!", "enjoyed a lot", and "very cute".

Workers commented on the collaborative aspects of the game: "Wish we could post a message to other players in real time. I didn't change one of my labels because I thought I had a good reason for doing so.", "I feel like the second to last was hard, it was the one my group did not agree on. It was painted, yes, but painted wood.", "It's an interesting concept. I'm amazed my group mostly agreed.", "I'm not sure what to do if I think we labeled something wrong, but we marked it so we would all agree?", and "Would have been more interesting if you could play once alone, and then with the other and see if the players do any changes."

At least one worker had the disappointing experience of the other players abandoning the game: "I don't think it's fair that I only get a bonus if other people are paying attention. Neither of the other people in my group even did the task. It should be based on how many I got correct, not how many the other people bothered to do.". One worker appears to have learned something about the other workers, namely that people often mislabel painted doors as wood: "make the directions about painted surfaces bold".

We had no problem getting workers to play our games. We could usually start one game per minute. It was observed that many workers would take another HIT and play subsequent games. This caused a phenomenon where small groups of workers may play several games together. In general, there was enough activity that workers played games rather than getting paid $\$0.01$ for waiting 3 minutes.

One problem we encountered was that a small number of workers would take many HITs at once. Those workers learned not to enter the queue more than 2 or 3 times simultaneously because it is difficult to play multiple games at once. However, by holding the HITs they prevented other players from joining the queue so that games could not begin. We decided to allow multiple-queueing, but lowered the HIT duration so that a worker could only comfortably take 4 HITs at once before they had to begin returning HITs.

To better understand the workers the authors took the role of a player (in games not used for analysis or experimentation). We observed two interesting behaviors. In one case, the workers agreed that a sample was painted but we kept our vote as wallpaper (which was the correct answer). As we approached the end of the game one of the workers rapidly changed their vote between painted and wallpaper. They may have been trying to draw attention to the disagreement in the hopes of getting someone to change their vote. Or, they may have been trying to signal that they were indifferent to the two choices. The second interesting behavior was in a different game. In the beginning, the third player did not cast any votes. This can happen if they are absent or have returned the HIT after joining the game queue. However, near the end of the game that third player suddenly cast a vote for every consensus that the two other players had established. We hypothesize that this player was following a strategy that maximized their payout without requiring their full attention — they simply let the other two players do all the work and copied their votes.

We anticipated that workers may try to collude to force a global consensus every time. This is hard since workers do not know when the second clique will run (or even if they are in the first or second clique) and our queueing system prevents any worker from the first clique entering the second clique. Nonetheless, it was suggested to the authors that workers could use a predetermined strategy of always picking the label which comes first lexicographically. We tried to oppose such strategies by including at least one random label in the set of candidates and shuffling the button order for each game.

In one case, we observed two workers that would very often play the game together. This could simply be because they enjoyed the game and were actively seeking our HITs. Another hypothesis is that this was the same person or two people in the same room taking control of the game by controlling two votes. We inspected the game records and found that these two players did not always agree and thus concluded that they were not colluding. However, we recommend that the queueing system prevent pairs of workers from entering the same games too often.

Discussion

This paper has demonstrated that consensus agreement games can improve the annotations of images on which independent workers disagree. We have explored only two instances of this approach, and one can imagine varying design choices in this space to explore their impact on cost and effectiveness. For example, using cliques of size larger than three or varying the number of parallel cliques that must agree would impact the cost and accuracy of consensus agreement games. Indeed, while our results show that agreement is necessary, one could imagine doing so only at random on a subset of cases, to reduce cost while hopefully maintaining effectiveness. Our current approach only allows workers to change labels once the worker has provided a label, with no opportunity to "erase" the vote without picking an alternative. Allowing vote erasing, restricting the number of times an answer can be changed, or allowing workers to define arbitrary labels as in the ESP Game would impact cost and effectiveness in as yet unexplored ways.

There are also design choices relevant to the worker experience. Payment sizes, the number of images presented per game, amount of time given for each game, worker qualifications, throttling repeat players, and rules about simultaneous queueing and game playing could also all affect MTurk error rate and throughput. Results may be further improved by implementing cheating prevention mechanisms such as random clique assignment and blocking previously seen worker pairings. Worker attention may be improved by adding sentinel samples. Since workers seem to enjoy playing it may not be necessary to pay for the time workers wait on the game queue. Simply allowing the worker to return a HIT if they tire of waiting would simplify the game logic and lower costs. Worker enthusiasm can also be taken as a sign that rewards can be reduced even further. It is not possible to pay fractional cents so there is a limit to how low payoffs can go, and while we combine multiple images into each game, there are attentional limits to increasing that number too high. One way to reduce rewards is to allow workers to chain multiple games and combine their rewards into a single payout.

Whereas (Judd, Kearns, and Vorobeychik 2010; Kearns 2012) explored how a group of individuals reach consensus as a function of the topology of the network connecting them, cliques are one case in the space of networks — consensus agreement games could be based on other network structures. We followed Judd, et al's approach of minimal communication channel amonst workers, but given the value that social transparency (Huang and Fu 2013a) and larger-scale communication (Mao et al. 2016) can have on online group performance, the nature of the communication allowed between workers could be expanded, such as letting a player highlight portions of images or providing a chatbox (as suggested by one worker). Given our observation of free riding in some game instances (Latane, Williams, and Harkins 1979), mechanisms for player self-policing (Kraut et al. 2012; Horton 2010; Dow et al. 2012), such as allowing players to identify and remove idle players, might have value. Also, workers could become more effective consensus members if they are instructed to be open to different opinions, view conflict as natural, and avoid early decision making (Gentry 1982).

On the other hand, the social psychology literature has found a wide range of ways in which teams and groups do not work as well as we might expect (Kerr and Tindale 2004; Levine and Moreland 2008; Sunstein 2006), finding for example that group behavior can suffer from group polarization, in-group/out-group biases, social loafing, and amplification of individual cognitive biases. CAG avoids many of these issues, creating small group structures that violate the premises of much of such work. For example, social influence effects have limited relevance when your teammates are anonymous and only together for a short time with limited means of communications. There is minimal opportunity to have differentiated member roles, to see the effects of a team leader, to consider time-varying group composition, to have intragroup conflict, or to see impacts of suboptimal work flows when all workers are given identical tasks and incentives, remain anonymous to each other, and are teamed for timescales measured in minutes or seconds. Consensus games have found a way to take what had previously been perceived as problematic issues for small groups and has instead harnessed them as an asset. Social forces that encourage conformity are known to damage group performance, serving as a barrier to getting a diversity of knowledge and approaches. We instead present group tasks that actually seek conformity, both within each collaborative labeling clique and by using payments that target conformity to a second game's outcomes. Indeed, consensus agreement games may turn group polarization into an asset.

Consensus games, nonetheless are a form of small group activity, and need not be immune from the known inefficiencies of small groups. For example, our observations suggest the presence of social loafing. More generally, consensus games can be studied from the lens of social psychology, exploring such questions as what communication patterns (as constrained as they are) impact outcomes? How does what we know about the effects of time scarcity on small groups apply here? How do anchoring, priming, and other effects impact outcomes, for example in terms of what sequence of games (images, teammates) a player is given? What are the effects of sex, age, and personality differences on group performance? How can group outcomes be improved through individual performance feedback? Do forms of organizational learning occur, and if so are they helpful? Are there ways in which mutual presence of team members (most typically visual) can positively impact group performance as it does in other small group settings? The answers to such questions are not just academic. The selection of who should be teamed and how their efforts should be structured can be informed by such knowledge, so that rather than assembling anonymous workers, we would assemble teams of the right people given the right tasks in the right way (Gentry 1982).

Whether one uses independent workers, consensus agreement games, or other crowdsourcing approaches for annotating data, they can all be loosely seen as sampling from a large population of individuals and eliciting information from them either independently or collaboratively so as to approximate what the majority vote of the entire population would be. Thus, for example, one might ask if an image of a cat is "cute", where this judgment should reflect what a typical person might answer, yet we are attempting to answer this without sampling the entire population. There are a variety of approaches that have been taken to perform tasks of this sort (Caragiannis, Procaccia, and Shah 2013; Dasgupta and Ghosh 2013; Donmez, Carbonell, and Schneider 2009; Ertekin, Hirsh, and Rudin 2012; Goel and Lee 2012; 2014; Lee et al. 2014; Mao, Procaccia, and Chen 2012; Montanari and Saberi 2009), and they might similarly suggest methods for using interacting workers to label data. Further, while we have proposed a hard-coded pragmatic approach of starting with votes among independent workers and then turning to consensus agreement games when there is insufficient support for a label, one could consider explicitly reasoning over workflows incorporating these and other consensus-seeking approaches (Shahaf and Horvitz 2010; Zhang, Horvitz, and Parkes 2013; Weld 2015).

Conclusion

We introduce consensus agreement games — a method for refining the majority vote labels on hard-to-label samples. We demonstrated this method on two real-world image datasets and showed that error rate was on average 37.8% lower than a majority vote baseline. The cost per annotation is high, but the method does not need to be run on the entire dataset. We suggest the following procedure. First, label all data with cost-efficient independent labeling tasks. Next, divide the dataset into subsets based on estimated difficulty. Next, take a small part of each subset, annotate them with expert truth then run CAG with $N = 2$ and $N = 3$. Select the best performing N and compare the difference of error rates for CAG annotations and independent worker annotations against the expert truth. Finally, use CAG to fully annotate those subsets for which the error rate (or difference of error rates) is larger than some threshold. The threshold is determined by either estimating the value of reducing error rate or setting a tolerance for maximum error rate.

In this paper we grounded the method in two practical, unrelated image labeling tasks that demonstrate its success

in the setting for which it was conceived. Yet, the pattern is general and may prove useful for other application domains. We believe there is value in the future to explore the merit of consensus agreement games on human computation tasks outside of image labeling.

Acknowledgments

PI Bala would like to thank our funding agencies including a Google Faculty Research Grant and NSF IIS 1617861.

References

Bachrach, Y.; Graepel, T.; Minka, T.; and Guiver, J. 2012. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. *arXiv preprint arXiv:1206.6386*.

Barr, J., and Cabrera, L. F. 2006. Ai gets a brain. *Queue* 4(4):24–29.

Bell, S.; Upchurch, P.; Snavely, N.; and Bala, K. 2013. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)* 32(4).

Bell, S.; Upchurch, P.; Snavely, N.; and Bala, K. 2015. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3479–3487.

Caragiannis, I.; Procaccia, A. D.; and Shah, N. 2013. When do noisy votes reveal the truth? In *Proceedings of the fourteenth ACM conference on Electronic commerce*, 143–160. ACM.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, 319–330. ACM.

Dekel, O.; Gentile, C.; and Sridharan, K. 2012. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research* 13(1):2655–2697.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Donmez, P.; Carbonell, J. G.; and Schneider, J. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 259–268. ACM.

Douglis, F. 2007. From the editor in chief: The search for jim, and the search for altruism. *IEEE Internet Computing* 11(3):4.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1013–1022. ACM.

Ertekin, S.; Hirsh, H.; and Rudin, C. 2012. Learning to predict the wisdom of crowds. *arXiv preprint arXiv:1204.3611*.

Faltings, B.; Jurca, R.; Pu, P.; and Tran, B. D. 2014. Incentives to counter bias in human computation. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Gentry, M. E. 1982. Consensus as a form of decision making. *J. Soc. & Soc. Welfare* 9:233.

Goel, A., and Lee, D. 2012. Triadic consensus. In *Internet and Network Economics*. Springer. 434–447.

Goel, A., and Lee, D. T. 2014. Large-scale decision-making via small group interactions: the importance of triads1. In *Workshop on Computational Social Choice (COMSOC)*.

Ho, C.-J.; Chang, T.-H.; Lee, J.-C.; Hsu, J. Y.-j.; and Chen, K.-T. 2009. Kisskissban: a competitive human computation game for image annotation. In *Proceedings of the acm sigkdd workshop on human computation*, 11–14. ACM.

Horton, J. J. 2010. Employer expectations, peer effects and productivity: Evidence from a series of field experiments. *Peer Effects and Productivity: Evidence from a Series of Field Experiments (August 3, 2010)*.

Hsueh, P.-Y.; Melville, P.; and Sindhwani, V. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, 27–35. Association for Computational Linguistics.

Huang, S.-W., and Fu, W.-T. 2013a. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 621–630. ACM.

Huang, S.-W., and Fu, W.-T. 2013b. Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 639–648. ACM.

Ipeirotis, P. G.; Provost, F.; Sheng, V. S.; and Wang, J. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28(2):402–441.

Judd, S.; Kearns, M.; and Vorobeychik, Y. 2010. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences* 107(34):14978–14982.

Kamar, E., and Horvitz, E. 2012. Incentives and truthful reporting in consensus-centric crowdsourcing. Technical report, Technical report, MSR-TR-2012-16, Microsoft Research.

Kanefsky, B.; Barlow, N. G.; and Gulick, V. C. 2001. Can distributed volunteers accomplish massive data analysis tasks. *Lunar and Planetary Science* 1.

Kearns, M. 2012. Experiments in social computation. *Communications of the ACM* 55(10):56–67.

Kerr, N. L., and Tindale, R. S. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55:623–655.

Kraut, R. E.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y.; Kittur, N.; Konstan, J.; Ren, Y.; and Riedl, J. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Latane, B.; Williams, K.; and Harkins, S. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology* 37(6):822.

Lee, D. T.; Goel, A.; Aitamurto, T.; and Landemore, H. 2014. Crowdsourcing for participatory democracies: Efficient elicitation of social choice functions. In *HCOMP-2014*.

Levine, J. M., and Moreland, R. L. 2008. *Small groups: key readings*. Psychology Press.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 3–12. Springer-Verlag New York, Inc.

Lin, C.-W.; Chen, K.-T.; Chen, L.-J.; King, I.; and Lee, J. H. 2008. An analytical approach to optimizing the utility of esp games. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, 184–187. IEEE.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; et al. 2008. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society* 389(3):1179–1189.

Mao, A.; Mason, W.; Suri, S.; and Watts, D. J. 2016. An experimental study of team size and performance on a complex task. *PloS one* 11(4):e0153048.

Mao, A.; Kamar, E.; and Horvitz, E. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *HCOMP-2013*.

Mao, A.; Procaccia, A. D.; and Chen, Y. 2012. Social choice for human computation. In *HCOMP-12: Proc. 4th Human Computation Workshop*. Citeseer.

Montanari, A., and Saberi, A. 2009. Convergence to equilibrium in local interaction games. In *FOCS'09. 50th Annual IEEE Symposium on*, 303–312. IEEE.

Parameswaran, A.; Boyd, S.; Garcia-Molina, H.; Gupta, A.; Polyzotis, N.; and Widom, J. 2014. Optimal crowd-powered rating and filtering algorithms. *Proceedings of the VLDB Endowment* 7(9):685–696.

Radanovic, G.; Faltings, B.; and Jurca, R. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7(4):48.

Raddick, J.; Lintott, C.; Schawinski, K.; Thomas, D.; Nichol, R.; Andreescu, D.; Bamford, S.; Land, K.; Murray, P.; Slosar, A.; et al. 2007. Galaxy zoo: an experiment in public science participation. In *Bulletin of the American Astronomical Society*, volume 39, 892.

Rao, H.; Huang, S.-W.; and Fu, W.-T. 2013. What will others choose? how a majority vote reward scheme can improve human computation in a spatial location identification task. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2008. LabelMe: A database and web-based tool for image annotation. *IJCV* 77(1-3):157–173.

Shah, N. B., and Zhou, D. 2015. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Advances in Neural Information Processing Systems*, 1–9.

Shahaf, D., and Horvitz, E. 2010. Generalized task markets for human and machine computation. In *AAAI*.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622. ACM.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, 254–263. Association for Computational Linguistics.

Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk.

Spain, M., and Perona, P. 2008. Some objects are more equal than others: Measuring and predicting importance. In *Computer Vision–ECCV 2008*. Springer. 523–536.

Sunstein, C. R. 2006. *Infotopia: How many minds produce knowledge*. Oxford University Press.

Thogersen, R. 2013. Data quality in an output-agreement game: A comparison between game-generated tags and professional descriptors. In *Collaboration and Technology*. Springer. 126–142.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. ACM.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Who should label what? instance allocation in multiple expert active learning. In *SDM*, 176–187. SIAM.

Wauthier, F. L., and Jordan, M. I. 2011. Bayesian bias mitigation for crowdsourcing. In *Advances in Neural Information Processing Systems*, 1800–1808.

Weld, D. S. 2015. Intelligent control of crowdsourcing. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 1–1. ACM.

Westphal, A. J.; Butterworth, A. L.; Snead, C. J.; Craig, N.; Anderson, D.; Jones, S. M.; Brownlee, D. E.; Farnsworth, R.; and Zolensky, M. E. 2005. Stardust@ home: a massively distributed public search for interstellar dust in the stardust interstellar dust collector.

Zhang, H.; Horvitz, E.; and Parkes, D. C. 2013. Automated workflow synthesis. In *AAAI*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using Places database. In *Advances in neural information processing systems*, 487–495.