Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Hotspotting – A Probabilistic Graphical Model for Image Object Localization Through Crowdsourcing

Mahyar Salek
Microsoft Research Cambridge

Yoram Bachrach
Microsoft Research Cambridge

Peter Key
Microsoft Research Cambridge

Abstract

Object localization is an image annotation task which consists of finding the location of a target object in an image. It is common to crowdsource annotation tasks and aggregate responses to estimate the true annotation. While for other kinds of annotations consensus is simple and powerful, it cannot be applied to object localization as effectively due to the task's rich answer space and inherent noise in responses.

We propose a probabilistic graphical model to localize objects in images based on responses from the crowd. We improve upon natural aggregation methods such as the mean and the median by simultaneously estimating the difficulty level of each question and skill level of every participant.

We empirically evaluate our model on crowdsourced data and show that our method outperforms simple aggregators both in estimating the true locations and in ranking participants by their ability. We also propose a simple adaptive sourcing scheme that works well for very sparse datasets.

Introduction

Images and videos are a major part of content consumed online. However, this key data form is difficult to handle from an algorithmic perspective, making image processing challenging. Humans easily perform tasks that have proven difficult for computers, such as segmentation, object recognition and pose estimation. Tasks such as annotating an image with the best textual description or determining where a specified object is located in an image are inherently demanding, since they require a deep understanding of the image, including social, cultural or geographical knowledge. For example, finding a description for an image containing Apple's logo requires knowing the firm and its marketing material. Or consider the problem of locating the most famous person in an image of many people. This requires not only image processing capabilities, but also cultural knowledge. [1]

[1] Such a question may have a strong cultural bias. If two celebrities appear in the image, each famous in a different country, the answer likely depends on the responder's country.

People can easily answer such questions, but building an automated system for doing so is difficult. The rapid development in Internet and mobile technologies and their widespread adoption have made sharing information simpler than ever. This has triggered the the rise of new solutions to such problems. A key example is crowdsourcing, which provides a strong and cost-effective solution for carrying such tasks. Crowdsourcing marketplaces, such as Amazon's Mechanical Turk, bring together requesters interested in solving such a task at hand, and workers who are willing to perform such tasks for a payment.

Unfortunately, human judgments may vary significantly. For example: people annotating the same image may provide different descriptions; when asked whether or not a specific object appears in an image there may be wide disagreement; the ranking of multiple items may be very different between individuals. One way to overcome this is using redundancy — the requester sources many opinions, then aggregates the information obtained from the multiple workers into a single high quality solution. In some instances aggregating the information is simple: given many textual descriptions of an image, the requester can choose the most common one; or when deciding if an object appears in an image, the question only has two possible answers ("yes" or "no"), so one can use the majority vote.

One task where information aggregation is challenging is determining the location of an object in an image. We refer to this as *Hotspotting*. Hotspotting has many applications: knowing the image area corresponding to an object is vital in training vision algorithms; with popular applications in image and video sharing such as Instagram, Pinterest and YouTube, Hotspotting could enable user-profiling, non-intrusive advertising, and interactive visual experience. Hotspotting is a difficult task, that may require social or cultural knowledge, so individuals likely vary in their ability to perform it. Further, the space of possible answers is enormous, with many answers considered "acceptable".

One solution is to ask users for a rectangular bounding box for the object's location, then aggregate the information taking the mean or median of the coordinates of the corners. This clearly treats all workers as equally capable. But as discussed above, we expect a wide variance in peoples' abilities to solve Hotspotting tasks. Taking into account the ability levels of individuals, we could improve performance

by giving more weight to the opinions of high-aptitude individuals. But how could we gain such knowledge?

Our contribution: We propose a machine learning model for information aggregation in Hotspotting tasks. Our approach is based on a probabilistic graphical model that simultaneously estimates the locations of the objects in the images, the difficulty level of each image and the abilities of the people performing the task. We evaluate empirically our model's ability to infer object locations and rank participants by ability using a crowdsourced dataset. Our data was captured using Amazon's Mechanical Turk, asking 168 participants to each answer 50 Hotspotting questions (i.e. 50 different image-object pairs). We show that our method outperforms simple aggregators, both in estimating the true locations and in ranking participants. We also show that the quality of our method improves as more data (in the form of responses by participants) is fed into the model, but that the quality curve quickly saturates, so the returns from each additional participant diminish. Finally, we discuss solutions for tight budget constraints that allow only sourcing few opinions and show that in such cases a simple aggregator can perform well when sourcing the data in an adaptive way.

Hotspotting Tasks: A Hotspotting problem consists of a set of images, each containing a target object. The goal is to find the location of the object in each image. We make the simplifying assumption that the location is captured by a bounding box, so it can be expressed by the corners of a rectangle. The data comprises the responses of multiple participants to each of the images. We call each image-object pair an *item*, and each person providing responses a *participant*. Each participant examines every item, and chooses the location for the target object to the best of their ability. Given a sequence Q of items, and a set P of participants, we have $|Q| \cdot |P|$ locations, each of which is a tuple $R_{pq} := (x^{tl}, y^{tl}, x^{br}, y^{br})$ where (x^{tl}, y^{tl}) are the x and y coordinates of the top left corner, and (x^{br}, y^{br}) are those of the bottom right corner. The output is a set $Y = (y_1, \ldots, y_{|Q|})$ of estimated object locations, where each y_q is a 4-tuple of Cartesian coordinates for the object in item q.

DALE: Difficulty, Ability and Location

We propose a probabilistic graphical model for Hotspotting tasks, which we refer to as the Difficulty-Ability-Location-Estimation model, or DALE for short. In addition to the participants' responses regarding the location of each object, our model may also have access to "ground truth" information, in the form of the correct location of some of the objects. This allows the DALE model to improve its estimation regarding the location of the *other* objects (i.e. the ones not in the "ground truth" set), as this information is useful for better estimating the participants' ability levels, which in turn is valuable for estimating the locations of the other objects. The output of the model includes the estimated location of each object $Y = (y_1, \ldots, y_{|Q|})$, and additional information including the difficulty level of each item and the ability levels of each of the participant. The correct object locations, item difficulty levels and abilities of participants are modeled as *unobserved random variables*, whereas the responses of the participants are *observed variables*.

The model's structure is determined by conditional independence assumptions regarding the variables. In landmark work (Pearl 1988) introduced Bayesian Networks (also called directed graphical models), encoding assumptions of conditional independence through a graph where each vertex represents a variable and where edges represent dependencies between the variables. Our model is based on the extension of Bayesian Networks called a factor graph (see (Koller and Friedman 2009)), which describes the factorial structure of the joint probability distribution among the variables. Once we define the model's structure as a factor graph and set the observed variables to the observed values, namely the responses of the participants, approximate message passing (Koller and Friedman 2009) allows inferring marginal probability distributions of the target unknown variables: the true locations of the objects in the images, the ability of each participant, and the difficulty of each item.

The Graphical Model: Each participant expresses her opinion regarding an object's location by choosing a bounding box for each item. Even a participant fully aware of an object's location is unlikely to draw the exact bounding box for the "true location" (if such a notion even exists). However, such informed participants tend to choose a bounding box whose corners are close to the corners of the "true" bounding box. In other words, if a participant correctly recognizes the object, the closer a location is to the true corner of the object, the higher the probability of the participant to select it. Given the bounding box representing the object's true location and the one provided by the participant, we measure the distance between the two using *Jaccard Similarity* (described below). Using this notion we can phrase our assumption as follows: a participant who is aware of the object's true location is likely to select a bounding box that is "close" to the true bounding box (i.e. the bounding boxes are likely to have a high Jaccard similarity).

We model the process by which a participant $p \in P$ chooses the locations for each item $q \in Q$. The location participant p assigns to item q, denoted by r_{pq}, is comprised of the bounding box corners $(x^{tl}_{pq}, y^{tl}_{pq}, x^{br}_{pq}, y^{br}_{pq})$. We assume every participant has an underlying ability $a_p \in \mathbb{R}$ which determines her ability to recognize the correct location of each item $q \in Q$, and that each item q has an inherent difficulty $d_q \in \mathbb{R}$ which determines how likely it is that a participant $p \in P$ would correctly recognize the object's location. Our modelling is based on a simple generative process: the participant's ability is sampled from a Gaussian distribution reflecting the distribution of ability levels in the population of participants; the item's difficulty is sampled from a Gaussian distribution reflecting the distribution of item difficulty levels; a "performance noise" for each participant-item pair, which may be positive or negative, is added to the participant's ability. If this ability plus the random noise exceeds the sampled difficulty of the item, the participant recognizes the object's location and selects the true corners perturbed by small Gaussian noise (reflecting judgement errors or inaccuracy); if they do not know the location, they give a random answer (based on the middles of grid quadrants with a large perturbing noise). Hence, DALE is defined as a joint

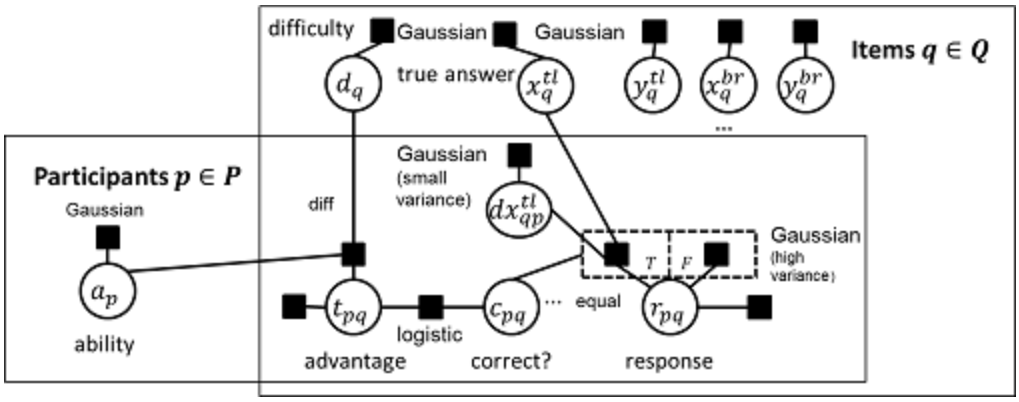

Figure 1: Factor graph for the difficulty-Ability-Location-Estimation (DALE) model.

probabilistic model. The factor graph is given in Figure 1. It has two parts. The first part models the ability of a participant p to correctly recognize the object's location in the image q (left of Figure 1), and the other ties the location of the bounding box given by the participant p for item q to whether they recognized the object or not, as represented by the variable c_{pq}. c_{pq} takes on Boolean values: 1 when the participant p recognizes the object's location in item q, and 0 when they don't. Recognizing the correct location is modeled as an interaction of the ability $a_p \in \mathbb{R}$ of participant p, and the difficulty $d_q \in \mathbb{R}$ of item q. We assume the variable depends on the difference $t_{pq} := a_p - d_q$ via the following equation:

$$P(c_{pq} = T|t_{pq}, \tau_q) := \int_{-\infty}^{\infty} \phi(\sqrt{\tau_q}(x - t_{pq}))\theta(x)\, dx = \Phi\left(\sqrt{\tau_q} t_{pq}\right). \quad (1)$$

Here ϕ is the standard Gaussian density $\phi(x) := \sqrt{2\pi}^{-1} \exp(-x^2/2)$, Φ denotes the sigmoidal cumulative Gaussian distribution $\Phi(t) := \int_{-\infty}^{t} \phi(x)\, dx$, and $\theta(\cdot)$ denotes the Heaviside step function. This integral representation indicates an alternative way to view this probability: a binary process resulting from evaluating the step function θ on a variable t with an added Gaussian noise of variance τ^{-1}.

The response r_{pq} is modeled as a mixture of two distributions. If participant p recognizes the object's location, we constrain her response to be the correct location for object with random Gaussian noise added to each coordinate in the bounding box. Namely, constrained to be $(x^{tl*} + \epsilon_x^{tl}, y^{tl*} + \epsilon_y^{tl}, x^{br*} + \epsilon_x^{br}, y^{br*} + \epsilon_y^{br})$ where the correct object location is $y_q = (x^{tl*}, y^{tl*}, x^{br*}, y^{br*})$, and $\epsilon_x^{tl}, \epsilon_y^{tl}, \epsilon_x^{br}, \epsilon_y^{br}$ are sampled from a Gaussian distribution (with a mean of zero and a low variance σ_l^2). The random noise for the x coordinate of the top left corner is shown in Figure 1 as dx_{qp}^{tl}. A similar variable occurs for the bottom right corner, and two additional variables occur for the y coordinate noise (these are omitted in the Figure to avoid a cumbersome presentation).

If the participant does not recognize the object's location, we assume the top left corner is randomly chosen from a Gaussian distribution with wide variance σ_h^2, whose mean location is the middle of the top left quadrant, and the bottom right corner is randomly chosen from a wide variance Gaussian distribution with mean location in the middle of the bottom right quadrant. The mixture is expressed in Figure 1 using a *gate*, marked by a dashed pair of boxes, which switches the factor connecting to r_{pq}, depending on the state variable c_{pq} is in. Gates were proposed in (Minka and Winn 2008) as a flexible expression for mixture models in factor graphs. Gates represent conditional independence relations that are context-dependent, and are suited for approximate message passing inference. Note the same gate in Figure 1 controls all the coordinate responses at the same time: if c_{pq} is true then *all* responses are constrained to be the true answer plus a low variance Gaussian noise.

We discuss how DALE infers the required information. The data is given as triplets of the form participant-item-location $\mathbf{L} := \{l_{p_1 q_1}, \ldots, l_{p_n q_m}\}$ where n is the number of participants, m the number of items, and $l_{p_i q_j}$ is the location participant i provides for the object in item j (consisting of the top left x and y coordinates and the bottom right x and y coordinates). We may also obtain a set of "ground truth" locations holding the true location for some items, of the form $\mathbf{Y} = \{y_{q_{i1}, \ldots, y_{q_{ir}}}\}$, though DALE can operate without these. Given the data $\mathbf{L}$ and $\mathbf{Y}$, our goal is to obtain approximate posterior (marginal) distributions: the Gaussian distribution $p(y_q|\mathbf{R}, \mathbf{Y})$ over the correct locations of the objects in all the images; [2] the Gaussian density $p(a_p|\mathbf{R}, \mathbf{Y})$ over the ability level a_p of participants p, with means $\tilde{\mu}_p$ and variances $\tilde{\sigma}_p^2$; the Gaussian density $p(d_q|\mathbf{R}, \mathbf{Y})$ over the difficulties d_q of items q with means $\tilde{\mu}_q$ and variances $\tilde{\sigma}_q^2$; and the Bernoulli distribution $p(c_{pq}|\mathbf{R}, \mathbf{y})$ of "correctness" c_{pq} indicating whether participant p recognized the correct location of the object in item q, given by success probabilities π_{pq}.

To perform the inference using our DALE model we use approximate message passing (see (Koller and Friedman 2009) for a survey of such approaches). Our implementa-

[2] By "location" we refer to four Gaussian variables — top left x, top left y, bottom right x and bottom right y.

tion used Infer.NET (Minka et al. 2010), a toolkit for probabilistic inference. We used the expectation-propagation (EP) algorithm (Minka 2001). EP is an approach for calculating posterior distributions of target variables on an input factor graph that iteratively computes messages passed along the edges, propagating information in the graph. The underlying factor graph in our model is a loopy graph, as we have multiple participants who respond to the same set of multiple questions. In addition, the messages in some of the nodes, such as those connected to the gate, are approximations. Therefore, to obtain the resulting posterior distributions the EP procedure runs iteratively until it converges.

Empirical Analysis

We tested our DALE model using 168 participants using Amazon's Mechanical Turk. Our Hotspotting set consisted of 50 images, each with a target object. Figure 2, illustrates two example items (finding a partially hidden logo, or finding a pharmacy based on its common sign in Europe, reflecting cultural knowledge). For each of the items, we have carefully marked the true location of the object manually, yielding a set of ground truth answers.

Figure 2: Top: Hotspotting the Lenovo logo in a mall. Bottom: Hotspotting a pharmacy. Participants are expected to have (or acquire) cultural knowledge, exerting effort to locate and mark the target object in a busy scene.

Participants were awarded a performance bonus based on comparing their responses to a "ground truth" set. We used the *Jaccard Similarity* to measure score participants. For sets A, B, the Jaccard similarity is defined as $\frac{|A \cap B|}{|A \cup B|}$. In our task the Jaccard similarity is the area in the intersection of the two bounding boxes (zero if non-intersecting) divided by the area of the union of the bounding boxes. We consider an object location to be correct if its bounding box has a Jaccard similarity of at least $\alpha = 0.5$ with the ground truth. We use the *mean* and the *median* as two natural benchmark aggregators to compare our model with. We measure the quality of each aggregator against the ground truth according to two metrics. First, with **Location Quality Metric:** Given a threshold parameter $0 < \alpha < 1$, the *location quality* is the number of questions for which the aggregator scored above α in terms of Jaccard similarity to the ground truth. Secondly, using an **Ability Quality Metric:** Every participant has a Jaccard similarity for each of her items compared with the aggregator's response for that item. Summing these gives a total score per participant, reflecting her overall performance, measured according to the aggregators' inferred object locations. We *rank* participants by total score to get the aggregator's ranking. We then measure the distance between this ranking and the participant ranking based on the ground truth to obtain the aggregator's *ability quality*. We measure distances between two participant rankings (the aggregator's ranking and the ground truth ranking) using *Kendall Tau rank correlation coefficient* (Kendall 1938; Sen 1968). [3] Figure 3 is a histogram of the performance levels of participants, measured by total score (sum of Jaccard similarities to the ground truth). Recall success in an item is defined as having a Jaccard similarity of at least $\alpha = 0.5$ with the ground truth. The shape roughly follows a Gaussian distribution.

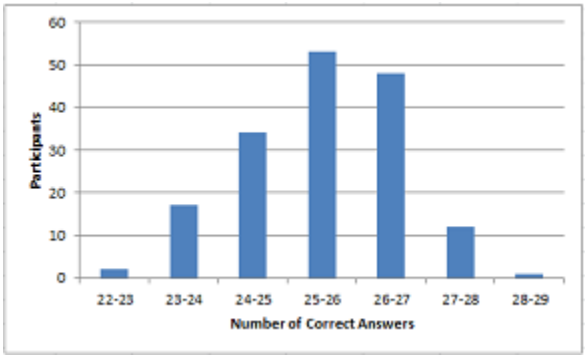

Figure 3: Participants' performance Histogram

We compared DALE's quality with simple aggregators using the two quality metrics, when all aggregators used all 168 participant opinions over all items. The mean aggregator has a very poor performance due to its sensitivity to outliers. To address this, we filtered out $\beta = 25\%$ of the participants with most extreme locations, in the spirit of robust aggregation (this was done for all aggregators for consistency [4]). DALE outperformed all benchmark aggregators, for both the location quality and ability quality metrics (under our choice of parameters), as shown in Figure 4 depicting the aggregator quality metrics on the entire input data. A paired t-test shows that the results for the location quality metric are significant at the $p = 0.08$ level. A similar comparison of the participant ability ranking shows higher significance ($p < 0.05$), but this should be taken with a grain of salt, as ability rankings are not independent. [5] Figure 4 thus shows that DALE is better than simple heuristics both in determin-

[3] Kendall Tau is defined as the difference between the number of concordant pairs and the number of discordant pairs normalized by $\frac{1}{2}n(n-1)$ where n is the number of the items in the ranking. It results in a coefficient $-1 \leq \tau \leq 1$ where agreement in rankings is captures by increasing values of τ (-1 showing perfect disagreement and 1 showing perfect agreement).

[4] As an equal number of participants from either side is removed, this has no influence on the median.

[5] We observed that the ability quality metric of DALE is highest when feeding the entire dataset for low α values (i.e. $\alpha < 0.5$).

ing object locations and in deciding who the stronger participants are. The choice of α influences the evaluation against other aggregators, but not the learning itself. For different choices of both α and β we achieve similar results.

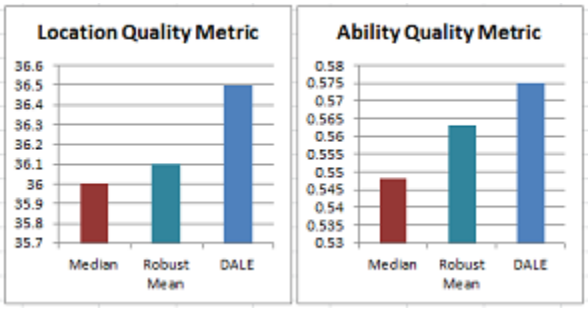

Figure 4: Aggregators' quality in the location quality metric and ability quality metric.

DALE's output includes not only the posterior distributions for the object locations, but also those of the participant ability parameters a_p and item difficulty parameters. These can be used to rank the participants by ability or items by difficulty. Figure 5 shows the ability levels a_p of the participants, sorted from lowest to highest. It shows that the model infers big differences in ability levels. These inferred abilities cause DALE to infer object locations that depend more on high aptitude participants than low aptitude ones.

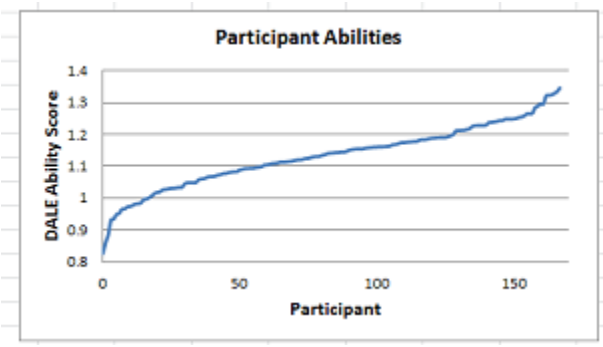

Figure 5: Inferred participant abilities (worst to best).

Figure 6 is a scatter plot showing how DALE's participants' ability parameters correlate with their total scores based on the ground truth (i.e. number of items to which the participant responded correctly, based on a threshold Jaccard similarity of $\alpha = 0.5$). The plot shows strong positive correlation, with a Pearson correlation factor of 0.71. Note that the relation need not be linear, as the ability parameter should not be interpreted as the number of items a participant would answer correctly. We expect the difficulty level of an item to be negatively correlated with the number of participants who responded correctly to the question, and indeed these have a Pearson correlation of -0.38. Unsurprisingly, these plots show that participants who responded correctly to more items tend to be inferred to have higher ability levels, and that questions who were answered correctly infrequently tend to be inferred to have high difficulty.

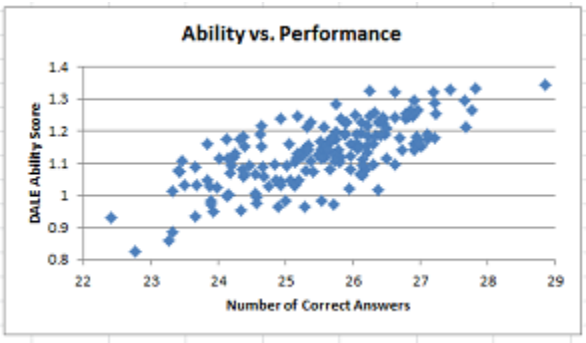

Figure 6: The correlation between participant's abilities and number of items answered correctly (using $\alpha = 0.5$).

Figure 7 shows how the quality of DALE's output improves as more data is used. The x-axis shows the number of participants whose data was fed in as input, and the y-axis shows the resulting location quality metric of the output. For any given number of participants on the x axis, we have sampled 5000 player subsets of that given size. The plot shows the average location quality under these subsets. The plot shows that the DALE model's quality improves as more data is fed in, but that the quality tends to saturate, showing that the gain of adding another participant diminishes.

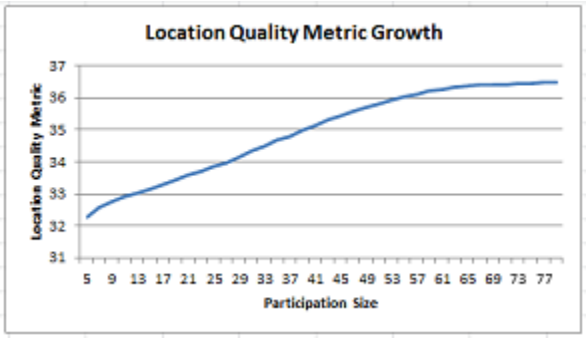

Figure 7: Location Quality Metric by participation size (number of samples) for $\alpha = 0.5$ and $\beta = 0.48$

Adaptive Sourcing

The discussion regarding Figure 7 indicates that achieving a good performance using DALE requires a large dataset of responses. In many scenarios collecting many responses is

However, for higher α values (which mean the accuracy has to be very high in order for a response to be considered "correct"), the ability quality metric is improved by screening more outliers. Further, for extremely high or low values of α ($\alpha > 0.75$ or $\alpha < 0.25$), the location quality metric of DALE improves when removing more outliers (i.e. increasing β), making DALE more similar to the median. This may be as a result of having a relatively low number of participants.

costly, so aggregators that have a high quality for very sparse or small datasets may be preferred. We investigate a simple heuristic for "budgeted" settings called *adaptive sourcing*. Suppose we have the budget to get b responses of a single participant to a single item. Adaptive sourcing is a technique in which we choose an item to be answered that we are most uncertain about. [6] Intuitively, those are the answers on which a consensus has not formed yet. For the space of bounding boxes, we propose a heuristic based on Jaccard similarity to quantify consensus. Let $\hat{R}_q$ be the aggregate response for question q using an aggregator of choice, based on the current data at hand. We define ψ_q, the *confidence level* of item q as $\sum_{p \in P} \frac{J(R_{pq}, \hat{R}_q)}{|P|}$ where $J(R_{pq}, \hat{R}_q)$ is the Jaccard index between the aggregate response and that of participant p. This is a measure of agreement, as it measures the average participants' similarity to the aggregated response. Our adaptive sourcing scheme works by iteratively collecting a single response for the current item with minimal confidence, $q_{\min} := \arg\min_{q \in Q} \psi_q$. Figure 8 shows the location quality metric of adaptive vs. non-adaptive approaches for the median aggregator, where the non-adaptive approach distributes the budget across all questions uniformly. The x axis is the number of item responses sourced. Figure 8 clearly shows that the adaptive sourcing outperforms uniform non-adaptive sourcing for any budget.

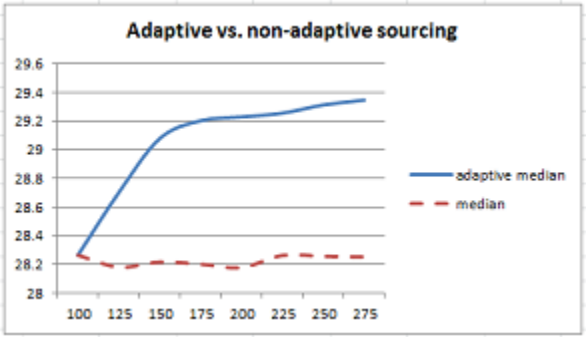

Figure 8: Performance of adaptive vs. non-adaptive sourcing (using median, and thresholded Jaccard scoring of $\alpha = 0.5$).

Related Work

Several approaches for information aggregation were proposed. One prominent method is voting, studied in social choice theory (Sen 1986). One of the earliest results in this field is Condorcet's Jury Theorem which bounds the probability that a set of agents using majority voting would reach the correct decision (Austen-Smith and Banks 1996; McLennan 1998; List and Goodin 2001). The related field of judgment aggregation (List and Puppe 2009) examines aggregating judgments on interconnected propositions given in a formal logic language into collective judgments on these propositions. However, such results are typically restricted to settings where the space of answers is very small, making voting a tractable mechanism. In our setting the answer space is to big to use such approaches. Various mechanisms were proposed to *incentive* participants to exert effort and provide opinions, including prediction markets (Pennock and Sami 2007), strategyproof learning (Everaere, Konieczny, and Marquis 2007; Dekel, Fischer, and Procaccia 2008) and crowdsourcing contests (Howe 2006; Archak 2010; Chawla, Hartline, and Sivan 2012; Gao et al. 2012). An approach to the incentive problem is *Games with a Purpose* (von Ahn 2006), where games are designed so that data collected could be used as useful inputs for algorithms. (Ahn, Liu, and Blum 2006) proposed Peekaboom, a game designed to incentivize work on Hotpostting tasks. We deal with the orthogonal topic of aggregating information once the agents have already provided their opinions.

Our model outperforms simpler aggregators such as the mean or median by making probabilitic inference on people's ability to solve the Hotspotting task. Psychologists noted that people significantly vary in their ability to perform cognitive tasks (Lubinski 2004; Schmidt and Hunter 2004) and designed tests for measuring human intelligence. The way our DALE model relates a participant's ability and her knowing where the object is located is somewhat reminiscent of "Item Response Theory" from psychology (Hambleton, Swaminathan, and Rogers 1991), which was used to develop such IQ tests. Recent work extends the concept of individual IQ to *collective intelligence* (Woolley et al. 2010). Some computational methods for aggregating opinions have recently been proposed (Lyle 2008; Bachrach et al. 2012a; 2012b; Kosinski et al. 2012; Demartini 2012), but they rely heavily on having a small answer space and thus cannot be used for Hotspotting tasks. Our approach is similar to machine learning approaches for aggregating labels given to images (Whitehill et al. 2009; Raykar et al. 2010; Welinder et al. 2010; Yan et al. 2011), but we exploit the physical / spacial nature of the Hotspotting problem.

Conclusions

We proposed DALE, a probabilistic graphical model for Hotspotting: determining the locations of objects in images. DALE also models the participant abilities and item difficulties so as to yield accurate results. Though DALE outperforms heuristic techniques given large datasets, its advantage shrinks for sparse datasets. In light of this, we proposed and evaluated an adaptive sourcing approach to Hotspotting, by obtaining opinions for images with least consensus and aggregating information using the median.

Several directions remain open for future work. Can better results be achieved using different noise models and generative processes? Can active learning techniques be used in the DALE mode to achieve an adaptive sourcing scheme that chooses the next opinion to source based on the expected reduction in uncertainty? Could the information regarding question difficulty and participant ability outputted by DALE be used to construct good incentive schemes? Is there a way to combine DALE with algorithms from computer vision to achieve a smaller search space and better results? Finally, Can DALE framework be applied to other domains such as Natural Language Processing?

[6] This heuristic mimics an active learning method where we examine the variable that maximizes the expected entropy reduction.

References

Ahn, L. V.; Liu, R.; and Blum, M. 2006. Peekaboom: A game for locating objects in images. In *In ACM CHI*, 55–64. ACM Press.

Archak, N. 2010. Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder. com. In *Proceedings of the 19th international conference on World wide web*, 21–30. ACM.

Austen-Smith, D., and Banks, J. 1996. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review* 34–45.

Bachrach, Y.; Graepel, T.; Kasneci, G.; Kosinski, M.; and Van Gael, J. 2012a. Crowd IQ - aggregating opinions to boost performance. In *AAMAS*.

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

Chawla, S.; Hartline, J.; and Sivan, B. 2012. Optimal crowdsourcing contests. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, 856–868. SIAM.

Dekel, O.; Fischer, F.; and Procaccia, A. 2008. Incentive compatible regression learning. In *SODA*.

Demartini, G., D. D. C.-M. P. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*, 469–478.

Everaere, P.; Konieczny, S.; and Marquis, P. 2007. The strategy-proofness landscape of merging. *Journal of Artificial Intelligence Research*.

Gao, X.; Bachrach, Y.; Key, P.; and Graepel, T. 2012. Quality expectation-variance tradeoffs in crowdsourcing contests. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Hambleton, R.; Swaminathan, H.; and Rogers, H. 1991. *Fundamentals of item response theory*, volume 2. Sage Publications, Inc.

Howe, J. 2006. The rise of crowdsourcing. *Wired magazine* 14(6):1–4.

Kendall, M. 1938. A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*.

Kosinski, M.; Bachrach, Y.; Kasneci, G.; Van-Gael, J.; and Graepel, T. 2012. Crowd iq: Measuring the intelligence of crowdsourcing platforms. *ACM Web Sciences*.

List, C., and Goodin, R. 2001. Epistemic democracy: generalizing the condorcet jury theorem. *Journal of Political Philosophy* 9(3):277–306.

List, C., and Puppe, C. 2009. Judgment aggregation: A survey. *Handbook of Rational and Social Choice*.

Lubinski, D. 2004. Introduction to the special section on cognitive abilities: 100 years after spearman's (1904)"general intelligence,'objectively determined and measured". *JPSP*.

Lyle, J. 2008. Collective problem solving: Are the many smarter than the few?

McLennan, A. 1998. Consequences of the condorcet jury theorem for beneficial information aggregation by rational agents. *American Political Science Review* 413–418.

Minka, T., and Winn, J. 2008. Gates. *NIPS* 21.

Minka, T.; Winn, J.; Guiver, J.; and Knowles, D. 2010. Infer.NET 2.4.

Minka, T. 2001. *A family of algorithms for approximate Bayesian inference*. Ph.D. Dissertation.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems : networks of plausible inference*.

Pennock, D., and Sami, R. 2007. Computational aspects of prediction markets.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *JMLR*.

Schmidt, F., and Hunter, J. 2004. General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology* 86(1):162.

Sen, P. 1968. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association* 63(324):1379–1389.

Sen, A. 1986. Social choice theory. *Handbook of mathematical economics* 3:1073–1181.

von Ahn, L. 2006. Games with a purpose. *Computer* 39(6):92–94.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The multidimensional wisdom of crowds. In *NIPS*, volume 6, 8.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *NIPS* 22:2035–2043.

Woolley, A.; Chabris, C.; Pentland, A.; Hashmi, N.; and Malone, T. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686.

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. 2011. Active learning from crowds. In *ICML*.