A Case for Soft Loss Functions

Alexandra Uma,[1] **Tommaso Fornaciari,**[2] **Dirk Hovy,**[2] **Silviu Paun,**[1] **Barbara Plank,**[3] **Massimo Poesio**[1]
[1]Queen Mary University, London [2]Università Bocconi, Milano [3]IT University of Copenhagen
{a.n.uma,s.paun,m.poesio}@qmul.ac.uk,bapl@itu.dk,{tommaso.fornaciari,dirk.hovy}@unibocconi.it

Abstract

Recently, Peterson et al. provided evidence of the benefits of using probabilistic soft labels generated from crowd annotations for training a computer vision model, showing that using such labels maximizes performance of the models over unseen data. In this paper, we generalize these results by showing that training with soft labels is an effective method for using crowd annotations in several other AI tasks besides the one studied by Peterson *et al.*, and also when their performance is compared with that of state-of-the-art methods for learning from crowdsourced data.

1 Introduction

There is growing evidence that training AI models directly from the distributions of judgments produced by a crowd, thus leveraging information about disagreements as in Figure 1, not only provides a better account of the empirical data in NLP (Poesio and Artstein 2005; Recasens, Hovy, and Martí 2011; Pradhan et al. 2012; Plank, Hovy, and Søgaard 2014b; Dumitrache 2019) and computer vision (Sharmanska et al. 2016; Rodrigues and Pereira 2018), but it is also beneficial from a machine learning perspective (Peterson et al. 2019). Several methods for training directly from annotator distributions have thus been proposed (Sheng, Provost, and Ipeirotis 2008; Raykar et al. 2010; Albarqouni et al. 2016; Guan et al. 2018; Rodrigues and Pereira 2018).

Recently, Peterson et al. provided further evidence of the benefits obtained by generating probabilistic labels from the annotator distribution and training a computer vision model from these probabilistic labels. They hypothesize that the benefits are affected by the features of the dataset, and they provided an elegant demonstration that using a traditional loss function such as cross-entropy as a 'soft loss' function is optimal when the objective is to maximize performance on unseen data. However, Peterson et al. did not evaluate this proposal for other types of assessment and for other tasks. They focused on a single image classification dataset, and only compared training from human-produced probabilistic soft labels with other techniques for probabilistic label generation such as *knowledge distillation* (Hinton, Vinyals, and Dean 2015). They did not consider other methods for learning from crowd annotations, and evaluation was restricted to hard metrics and the ability to produce a distribution with minimal cross-entropy wrt the human distribution.

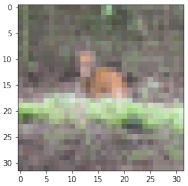

Figure 1: An example of disagreement from IC-CIFAR10H gold: $deer$, label with most votes (mass): dog, crowd counts: [dog:33, $deer$:13, $horse$:4]; see Section 5.2.

Contributions In this paper, we (i) test the hypothesis that soft loss is beneficial systematically in a variety of evaluation contexts, using crowdsourced datasets for several AI tasks and with different characteristics, and comparing the results with those obtained with state-of-the-art methods for learning from crowdsourced data; and (ii) we show that the method used to extract a probability distribution from the raw annotations matters, the choice depending on the characteristics and amount of annotators.

2 Learning from Crowdsourced Annotations

In this Section we briefly review some of the most influential approaches to learning from crowdsourced annotations, before discussing the recent proposal from Peterson et al.

2.1 Learning from Aggregated Labels

The most widely used method to learn from a crowdsourced dataset is to aggregate the crowd labels using Majority Voting (MV) or a more advanced aggregation method, see the review by Paun et al. (2018). These methods result in labels whose accuracy is sometimes comparable to that of gold labels (Paun et al. 2018). Two such approaches were considered here: MV and the best known probabilistic aggregation model (Dawid and Skene 1979).

2.2 Inducing a Classifier from Crowds

A number of methods exist to learn a model directly from the annotations (Raykar et al. 2010; Albarqouni et al. 2016; Guan et al. 2018; Rodrigues and Pereira 2018). One of the most recent such models is the *Deep Learning from Crowds* (DLC) approach, proposed by Rodrigues and Pereira (2018). DLC not only learns to combine the votes of multiple annotators, but also captures and corrects their biases while remaining computationally less complex than previous methods. Rodrigues and Pereira (2018) showed that their model outperforms several existing models when evaluated against gold truth. For these reasons, we select DLC as the representative method for this kind of approach. In particular, we used the DL-MW variant that achieved the most accurate predictions in Rodrigues and Pereira (2018).

2.3 Using Soft Loss Functions

Recently, Peterson et al. (2019) proposed to train models on crowd annotated data using 'soft labels' derived from the annotations as target distributions in a cross-entropy loss function.

Given some observed data $\{x_i, y_i\}_{i=1}^n$ at training time we want to minimize its expected loss:

$$\sum_{i=1}^{n} \sum_{c} L(f_\theta, x_i, y_i = c) p(y_i = c|x_i). \quad (1)$$

Considering the second term in (1), using hard labels only yields the optimal classifier if $p(y|x)$ is 1 for a single category and 0 for all other categories, but this has been shown to be an idealization (Poesio et al. 2019; Pradhan et al. 2012; Sharmanska et al. 2016). A more natural label categorization would be the human label distribution $p_{hum}(y|x)$ rather than a hard label (adjudicated 'gold' label or aggregated 'silver' label), and to learn from it directly. Using a negative log-likelihood, this loss reduces to the cross entropy loss function CE:

$$-\sum_{i=1}^{n} \sum_{c} p_{hum}(y_i|x_i) \ \log p_\theta(y_i = c|x_i), \quad (2)$$

where $p_\theta(x|y)$ is obtained by applying a probability function (softmax) over the logits produced by the classifier. This combination of probabilistic soft labels with a probability-comparing loss function is what we call the **soft loss function** approach. In this study, we tested this hypothesis across several tasks, not only in computer vision but also in NLP.

We further generalize the hypothesis of Peterson et al. in another direction. Peterson et al. generate $p_{hum}(y|x)$ by applying a standard normalization function over the crowd annotations for each item. Given C classes, let $d_i = [d_i^1, d_i^2, ...d_i^C]$ be a vector where some d_i^j entry stores the number of times the coders chose the j-th class for the i-th training example, using normalization,

$$p_{hum}(y_i = j|x_i) = \frac{d_i^j}{\sum_a (d_i^a)} \quad (3)$$

This implies that any class j for which the annotators provide no annotations will have a probability of 0. For datasets with numerous annotations this is a desirable effect, but for datasets with fewer annotations where some valid classes were not selected by any annotators, we hypothesize that using a softmax for normalization would be more appropriate, since $\exp(d_i^j) = 1$ when $d_i^j = 0$:

$$p_{hum}(y_i = j|x_i) = \frac{\exp(d_i^j)}{\sum_a \exp(d_i^a)} \quad (4)$$

We hypothesize that although this transformation might introduce some noise, it is a more representative distribution for datasets with fewer and/or lower quality annotations. Thus, we compared soft labels generated using the standard normalization function used by (Peterson et al. 2019) with soft labels generated using the softmax function.

3 Tasks and Models

Possibly the key claim of Peterson et al. (2019) is that properly capturing crowd judgments requires a sufficiently large dataset with a substantial number of annotator judgments. One of the key contributions of their paper is IC-CIFAR10H, a collection of over 50k judgements for 10k images in CIFAR-10. In this study, we evaluated methods for training from soft labels using not only their IC-CIFAR10H dataset, but also two datasets that are extensively used in related research (Plank, Hovy, and Søgaard 2014a; Jamison and Gurevych 2015; Rodrigues and Pereira 2018) yet differ in a number of respects from IC-CIFAR10H. The three datasets are briefly described below and summarized in Table 1. We developed near-state-of-the-art models for each of these tasks, and trained them using each of the methods for learning from crowds discussed in Section 2. We also trained these models on gold labels for comparison.

3.1 Part-of-Speech Tagging (POS)

- *Description* POS tagging is the task of assigning a part-of-speech tag (e.g., noun, verb) to every token in a text.
- *Dataset* The dataset we used–henceforth, POS–is the (Gimpel et al. 2011) dataset containing POS labels for Twitter posts, previously used in (Plank, Hovy, and Søgaard 2014a; Jamison and Gurevych 2015) and consisting of over 14k examples. Plank, Hovy, and Søgaard mapped the Gimpel tags to the universal tag set (Petrov, Das, and McDonald 2012), and crowdsourced labels for each token. We used the dataset released by (Plank, Hovy, and Søgaard 2014a) as a development set.
- *Model* We implement our own neural POS tagger inspired by (Plank, Søgaard, and Goldberg 2016) extended with attention over character and the word level representations. The model was always trained for 20 epochs using Adam (Kingma and Ba 2015) at a learning rate of 0.001 with the the model with best result saved at each epoch. This best model was used for evaluation on the test data.

3.2 Image Classification: LabelMe

- *Description* Image classification is a very general term for the task of assigning an image to the category that best describes it among a fixed set of categories.

- *Dataset* LabelMe[1] (Russell et al. 2008) is a widely used, community-created image classification dataset where images are assigned to one of 8 categories (highway, inside city, tall building, street, forest, coast, mountain, open country). Rodrigues and Pereira collected crowd labels for 10k images using Amazon Mechanical Turk from 59 annotators producing at least one label for each image. In this study we used this version of LabelMe. We randomly split the 10K images into training and test data (8,882 and 1,118 images respectively) to allow for ground truth and probabilistic evaluation. 500 images from the dataset with gold labels were used as development set.
- *Model* The model from (Rodrigues and Pereira 2018) was replicated for this task. Training was carried out for 20 epochs using the Adam optimizer (Kingma and Ba 2015) at a learning rate of 0.001. The model with the best development result was saved and used for testing.

3.3 Image Classification: CIFAR-10H

- *Dataset* Peterson et al. (2019) collected human annotations for the 10k test portion of the 10-category image classification CIFAR-10[2] using Amazon Mechanical Turk, creating the CIFAR-10H dataset.[3] (The categories are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). We used the CIFAR-10H dataset for training and testing using a 70:30 random split, ensuring that the number of images per class remained balanced as in the original dataset. We also use a subset of the CIFAR-10 training dataset (3k images) as our development set.
- *Model* The model trained for this task is the ResNet-34A model (He et al. 2016), a deep residual framework which is one of the best performing systems for the CIFAR-10 image classification. We used a publicly available Pytorch implementation of this ResNet model.[4] We trained the model with for a total of 65 epochs divided into segments of 50, 5 and 10, using a learning rate of 0.1 and decaying the learning rate by 1e-4 at the end of every segment. The model used for the evaluation phase was the model with the best development performance.

4 Evaluation metrics: soft and hard

We evaluate the trained models using *hard* and *soft* metrics. As hard metric we used **accuracy** as in Peterson et al. (2019): it assesses how good a model is at learning the preferred (gold) label for an item. A hard metric however ignores the fact that human judges may assign non-null probability to different labels for that item. Peterson et al. (2019) use **cross-entropy** to capture how well the model captures human's assessment of the probability not just of the top label, but also of the other labels.

We use cross-entropy as well, but we also introduce a different way to assess the similarity of the model to human behaviour: We measure the ability of a trained model to capture human uncertainty in its prediction using **entropy correlation**. To measure the ability of model θ to predict with an uncertainty correlated to the human uncertainty, we compute on an item basis the normalized entropy of the probability distribution produced by the model $p_\theta(y_i|x_i)$, the normalized entropy of the probabilistic soft labels $p_{hum}(y_i|x_i)$, and compute the Pearson correlation between the two values. The average of this correlation coefficient across all items is what we call 'entropy correlation'.

5 Results

Table 2 compares the two methods of extracting soft labels from the crowd annotations: softmax and standard normalization. Table 3 shows the results of training models for each dataset described in Section 3 using the methods described in Section 2 using the hard and soft metrics. To account for non-deterministic model training effects, we average over 30 runs, except for IC-CIFAR10H which was run 10 times due to model complexity. The mean results and standard deviation from the mean is reported.

5.1 Extracting probabilistic labels: softmax vs standard normalization

Table 2 shows that the normalization approach used to produce soft labels from the crowd annotations does affect the results. In 2 of the 3 tasks (IC-LABELME and POS), generating soft labels using softmax yielded better results than generating them using standard normalization. The opposite was true for IC-CIFAR10H. This difference can be explained in terms of the differences between the datasets.

Part of the explanation is the property of softmax of smoothing a distribution, already mentioned in Section 2.3. The softmax assigns a probability to every possible label in a dataset even if it received no annotations (see Section 2.3). In the POS and IC-LABELME datasets, the gold interpretation received no annotations for 11% of the items; i.e. $d_{x_i}^{y^j} = 0$, where j is the gold/preferred class for 11% of instances, i. This is not unexpected, given that on average, POS has 5 annotations per item and 17 possible classes (a coder to label ratio of 0.294) and IC-LABELME has an average of 2.54 annotations per item and 10 possible classes (a coder to label ratio of 0.254). In IC-CIFAR10H, by contrast, there is no item for which the gold annotation is not produced by at least one annotator. As a consequence, using a softmax function to generate probabilistic labels might allow the model to learn that those classes are at least probable.

The second part of the explanation is that the difference is due to the different characteristics of the two probability distributions. The distribution obtained through the softmax exacerbates the differences in probability mass between the classes compared to the distribution obtained through standard normalization–the larger the number of annotations collected for an item, the smaller the entropy of the resulted distribution. Thus, which type of normalization is best suited for a dataset depends on the characteristics of the annotations–such as those shown in Table 1. When we have a large number of annotations (and from good quality

[1] http://labelme.csail.mit.edu/

[2] https://www.cs.toronto.edu/ kriz/cifar.html

[3] https://github.com/jcpeterson/cifar-10h

[4] https://github.com/KellerJordan/ResNet-PyTorch-CIFAR10

Table 1: Annotations and Annotators

	POS	IC-LABELME	IC-CIFAR10H
Average annotations per item	5.00	2.50	51.10
Average coder to label ratio	0.30	0.26	5.11
Average observed agreement per item	0.73	0.73	0.94
Average annotator accuracy	0.68	0.69	0.95
Percentage of annotators with accuracy above 0.75	0.29	0.42	1.00

Table 2: Comparing Accuracy for Methods for Generating Soft Labels from Crowd Annotations

	POS	IC-LABELME	IC-CIFAR10H
Standard Normalization	78.99 ± 0.36	83.46 ± 0.82	$\mathbf{66.64 \pm 0.81}$
Softmax	$\mathbf{80.03 \pm 0.28}$	$\mathbf{84.85 \pm 0.50}$	65.50 ± 1.10

Table 3: Accuracy, Cross entropy and Entropy correlation results across all methods and tasks

	POS			**IC-LABELME**			**IC-CIFAR10H**		
	Accuracy	Cross Entropy	Correlation	Accuracy	Cross Entropy	Correlation	Accuracy	Cross Entropy	Correlation
Gold	89.22 ± 0.70	3.34 ± 0.40	0.41 ± 0.02	97.21 ± 0.49	4.86 ± 0.14	-0.01 ± 0.01	65.22 ± 0.76	2.61 ± 0.07	0.13 ± 0.09
Majority Voting	77.90 ± 0.84	2.58 ± 0.13	0.52 ± 0.02	80.36 ± 0.57	3.07 ± 0.09	0.15 ± 0.01	65.68 ± 1.01	2.63 ± 0.10	0.13 ± 0.03
Dawid and Skene	77.46 ± 1.75	2.52 ± 0.09	0.50 ± 0.02	83.43 ± 0.79	2.90 ± 0.23	0.11 ± 0.05	65.65 ± 1.15	2.55 ± 0.14	0.13 ± 0.02
DLFC	76.50 ± 2.57	2.16 ± 0.04	0.60 ± 0.04	83.54 ± 0.57	2.80 ± 0.15	0.28 ± 0.06	$\mathbf{68.25 \pm 1.06}$	3.51 ± 0.14	0.12 ± 0.01
Soft-loss	$\mathbf{80.03 \pm 0.28}$	$\mathbf{1.67 \pm 0.52}$	$\mathbf{0.67 \pm 0.01}$	$\mathbf{84.85 \pm 0.50}$	$\mathbf{1.64 \pm 0.02}$	$\mathbf{0.39 \pm 0.08}$	66.64 ± 0.81	$\mathbf{1.11 \pm 0.04}$	$\mathbf{0.22 \pm 0.01}$

coders), as in the case of IC-CIFAR10H, we have a good representation for the target distribution, and so a regular normalization is a natural choice as it does not change the class proportions. A softmax in this case would result in losing the richness of the original representation. On the other hand, if we only have a few annotations and of a lower quality, as in POS and IC-LABELME, then our target representation is likely to be noisy. For this reason a softmax is preferred in these circumstances, as it reduces the noise from the original votes, in contrast to standard normalization.

As a result of this analysis, the results for soft-loss training in Table 3 were computed using softmax soft labels for POS and IC-LABELME, and using standard normalized soft labels for IC-CIFAR10H.

5.2 Evaluation using hard metrics

Three observations are apparent from Table 3: (i) CE soft-loss learning achieved better results at learning to predict gold labels than learning from silver labels (MV and Dawid and Skene); (ii) CE soft-loss outperforms gold training in a single task, CIFAR-10H image classification (IC-CIFAR10H); and (iii) CE soft-loss outperforms the state-of-art method for learning directly from soft labels (DLC) in POS or IC-LABELME, but not in IC-CIFAR10H.

The first observation suggests that a soft aggregation of labels from annotators that retains the uncertainty of the crowd is beneficial over a hard consensus that aims to 'even out the noise', irrespective of the level of expertise of the annotators or their level of disagreement. IC-CIFAR10H, annotated by highly accurate ('expert') annotators and POS and IC-LABELME annotated by a mixed crowd of annotators, all benefit from probabilistic soft labelling over hard labelling from crowds (see Table 1 for dataset characteristics).

Secondly, gold labels are usually the aggregate or adjudicated consensus of expert annotators, and as such can be very useful during learning, but as noted several times in the literature, they may present an idealization of the task which may be excessive in cases when the disagreement is real (Poesio et al. 2019; Pradhan et al. 2012; Sharmanska et al. 2016). As seen in Figure 1, in a complex task like image classification disagreements in annotation can be information about the underlying difficulty of a given example. Although several annotators chose dog as the label for that image, $deer$ and $horse$ also received substantial amounts of votes, and the diverging opinions are clearly an indication of the confusing nature of the image. Probabilistic soft labels preserve label uncertainty without detracting from hard aggregated accuracy: in this case, the probabilistic soft label combines the high accuracy of majority voting with uncertainty preservation. The higher accuracy of CE soft-loss training compared to gold training for this task seems to suggest that particularly when the annotators are of expert quality, training using all expert annotations rather than a consensus gold label yields better results.

Finally, the third observation seems to indicate that although for noisy datasets, CE soft loss outperforms the state-of-art method for learning from soft labels, DLC is better suited for learning from soft labels when the annotators can be trusted. Further experimentation is needed to ascertain the extent to which these observations hold true.

5.3 Evaluation using soft metrics

From the result in Table 3 we can see how similar the probability distribution produced by a model is to the probability distribution produced by the human annotators, as measured using cross-entropy. Training using CE soft loss achieves better results for the 3 tasks, for all the methods, including DLC and gold training, making it advantageous over the two methods. CE soft loss also produces the distribution whose uncertainty best correlates with the uncertainty

demonstrated by the crowd in classifying the given item, as measured using entropy correlation–although the two soft metrics do not produce different results for these datasets.

6 Conclusions

In this study we found that training with CE as a soft loss function works well not only to train models that generalize well to unseen data, as demonstrated by Peterson et al. (2019), and not only on datasets with the characteristics of IC-CIFAR10H, but in general as a method for training models from soft labels and for a variety of tasks, subject to some conditions.

We also found that although this type of training does not in general outperform gold with respect to hard evaluation metrics, it does so with datasets with a substantial number of annotations per item and high quality annotations, such as IC-CIFAR10H). Also, soft-loss training systematically outperforms gold training when the objective is to achieve a model whose output mimics most closely the distribution of labels produced by the annotators, either in respect to relative ranking or in terms of uncertainty.

Acknowledgments

This research was supported in part by the DALI project, ERC Grant 695662.

References

Albarqouni, S.; Baur, C.; Achilles, F.; Belagiannis, V.; Demirci, S.; and Navab, N. 2016. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. on Medical Imaging* 35:1313–1321.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society* 28(1):20–28.

Dumitrache, A. 2019. *Truth in Disagreement*. Ph.D. Dissertation, Free University Amsterdam.

Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; and Smith, N. A. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th ACL*, 42–47. Portland, Oregon, USA: ACL.

Guan, M. Y.; Gulshan, V.; Dai, A. M.; and Hinton, G. E. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the 32nd AAAI*, 3109–3118. New Orleans, Louisiana: AAAI Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jamison, E., and Gurevych, I. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of EMNLP*, 291–297. Lisbon, Portugal: ACL.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics* 6:571–585.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 9616–9625.

Petrov, S.; Das, D.; and McDonald, R. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th LREC*. ELRA.

Plank, B.; Hovy, D.; and Søgaard, A. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th EACL*.

Plank, B.; Hovy, D.; and Søgaard, A. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd ACL*.

Plank, B.; Søgaard, A.; and Goldberg, Y. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th ACL*.

Poesio, M., and Artstein, R. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Meyers, A., ed., *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, 76–83.

Poesio, M.; Chamberlain, J.; Paun, S.; Yu, J.; Uma, A.; and Kruschwitz, U. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of NAACL*, 1778–1789. Minneapolis, Minnesota: ACL.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40. Jeju Island, Korea: Association for Computational Linguistics.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.

Recasens, M.; Hovy, E.; and Martí, M. A. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua* 121(6):1138–1152.

Rodrigues, F., and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the 32nd AAAI*, 1611–1618. AAAI Press.

Russell, B.; Torralba, A.; Murphy, K.; and Freeman, W. 2008. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77.

Sharmanska, V.; Hernández-Lobato, D.; Hernández-Lobato, J. M.; and Quadrianto, N. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of CVPR*, 2194–2202.

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*, KDD '08, 614–622. New York, NY, USA: Association for Computing Machinery.