Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion

Ryutaro Tanno[1] * Ardavan Saeedi[2] Swami Sankaranarayanan[2]
Daniel C. Alexander[1] Nathan Silberman[2]
[1]University College London, UK [2]Butterfly Network, New York, USA
[1] {r.tanno, d.alexander}@ucl.ac.uk [2]{asaeedi,swamiviv,nsilberman}@butterflynetinc.com

Abstract

The predictive performance of supervised learning algorithms depends on the quality of labels. In a typical label collection process, multiple annotators provide subjective noisy estimates of the "truth" under the influence of their varying skill-levels and biases. Blindly treating these noisy labels as the ground truth limits the accuracy of learning algorithms in the presence of strong disagreement. This problem is critical for applications in domains such as medical imaging where both the annotation cost and inter-observer variability are high. In this work, we present a method for simultaneously learning the individual annotator model and the underlying true label distribution, using only noisy observations. Each annotator is modeled by a confusion matrix that is jointly estimated along with the classifier predictions. We propose to add a regularization term to the loss function that encourages convergence to the true annotator confusion matrix. We provide a theoretical argument as to how the regularization is essential to our approach both for the case of single annotator and multiple annotators. Despite the simplicity of the idea, experiments on image classification tasks with both simulated and real labels show that our method either outperforms or performs on par with the state-of-the-art methods and is capable of estimating the skills of annotators even with a single label available per image.

1. Introduction

In many practical applications, supervised learning algorithms are trained on noisy labels obtained from multiple annotators of varying skill levels and biases. When there is a substantial amount of disagreement in the labels, conventional training algorithms that treat such labels as the "truth" lead to models with limited predictive performance. To mitigate such variation, practitioners typically abide by the principle of "wisdom of crowds" [1] and aggregate labels by computing the majority vote. However, this approach has limited efficacy in applications where the number of annotations is modest or the tasks are ambiguous. For example, vision applications in medical image analysis [2] require annotations from clinical experts, which incur high costs and often suffer from high inter-reader variability [3, 4, 5, 6].

However, if the exact process by which each annotator generates the labels was known, we could correct the annotations accordingly and thus train our model on a cleaner set of data. Furthermore, this additional knowledge of the annotators' skills can be utilized to decide on which examples to be labeled by which annotators [7, 8, 9]. Therefore, methods that can accurately model the label noise of annotators are useful for improving not only the accuracy of the trained model, but also the quality of labels in the future.

Previous work proposed various methods for jointly estimating the skills of the annotators and the ground truth (GT) labels. We categorize these methods into two groups: (1) *two-stage* approach and (2) *simultaneous* approach. Methods in the first category perform label aggregation and training of a supervised learning model in two separate steps. The noisy labels $\widetilde{\mathbf{Y}}$ are first aggregated by building a probabilistic model of annotators. The observable variables are the noisy labels $\widetilde{\mathbf{Y}}$, and the latent variables/parameters to be estimated are the annotator skills and GT labels $\mathbf{Y}$. Then, a machine learning model is trained on the pairs of aggregated labels $\mathbf{Y}$ and input examples $\mathbf{X}$ (e.g. images) to perform the task of interest. The initial attempt was made in [10] in the early 1970s and more recently, numerous lines of research [11, 6, 12, 13, 14] proposed extensions of this work e.g. by estimating the difficulty of each example. However, in all these cases, information about the raw inputs $\mathbf{X}$ is completely neglected in the generative model of noisy labels used in the aggregation step, and this highly limits the quality of estimated true labels in practice.

The *simultaneous* approaches [15, 16, 17, 18] address this issue by integrating the prediction of the supervised learning model (i.e. distribution $p(\mathbf{Y}|\mathbf{X})$) into the probabilistic model of noisy labels, and have been shown to improve the predictive performance. These methods employ variants of the expectation-maximization (EM) algorithm during training, and require a reasonable number of labels

*A part of the work done during internship at Butterfly Network.

for each example. However, in most real world applications, it is practically prohibitive to collect a large number of labels per example, and this requirement limits their applications. A notable exception is the Model Boostrapped EM (MBEM) algorithm presented in [19] that is capable of learning even with little label redundancy.

In this paper, we propose a more effective alternative to these EM-based approaches for jointly modeling the annotator skills and GT label distribution. Our method separates the annotation noise from true labels by (1) ensuring high fidelity with the data by minimizing the cross entropy loss and (2) encouraging the estimated annotators to be maximally unreliable by minimizing the trace of the estimated confusion matrices. Our method is also simpler to implement, only requiring an addition of a regularization term to the cross-entropy loss. Furthermore, we provide a theoretical result that such regularization is capable of recovering the annotation noise as long as the average confusion matrix (CM) over annotators is diagonally dominant.

Experiments on image classification tasks with both simulated and real noisy labels demonstrate that our method, despite being much simpler, leads to better or comparable performance with MBEM [19] and generalized EM [15, 20], and is capable of recovering CMs even when there is only one label available per example. We simulated a diverse range of annotator types on MNIST and CIFAR10 data sets while we used a ultrasound dataset for cardiac view classification to test the efficacy in a real-world application. We also show importance of modeling individual annotators by comparing against various modern noise-robust methods [21, 22, 23, 24], when the inter-annotator variability is high.

Other Related Works. More broadly, our work is related to methods for robust learning in the presence of label noise. There is a large body of literature that do not explicitly model individual annotators unlike our method.

The effects of label noise are well studied in common classifiers such as SVMs and logistic regression, and robust variants have been proposed [25, 26, 27]. More recently, various attempts have been made to train deep neural networks under label noise. Reed et al. [21] developed a robust loss to model "prediction consistency", which was later extended by [28]. In [29] and [22], label noise was parametrized in the form of a transition matrix and incorporated into neural networks for binary and multi-way classification. A more effective alternative for estimating such transition matrix was proposed in [30], and a method for capturing image dependency of label noise was shown in [31]. We will later compare our model to several of these methods to test the value of modelling individual annotators in gaining robustness to label noise.

Multiple lines of work have shown that a small portion of clean labels improves robustness. [32] proposed to learn from clean labels to correct the labels of noisy examples. [33] proposed a method for learning to weigh examples during each training iteration by using the validation loss on clean labels as the meta-objective. [34] employs a similar approach, but trains a separate network that proposes weighting. However, curating a set of clean labels of sufficient size is expensive for many applications, and this work focuses on the scenario of learning from purely noisy labels.

2. Methods

We assume that a set of images $\{\mathbf{x}_i\}_{i=1}^N$ are assigned with noisy labels $\{\tilde{y}_i^{(r)}\}_{i=1,...,N}^{r=1,...,R}$ from multiple annotators where $\tilde{y}_i^{(r)}$ denotes the label from annotator r given to example $\mathbf{x}_i$, but no ground truth (GT) labels $\{y_i\}_{i=1,...,N}$ are available. In this work, we present a new procedure for multiclass classification problem that can simultaneously estimate the annotator noise and GT label distribution $p(y|\mathbf{x})$ from such noisy set of data $\mathcal{D} = \{\mathbf{x}_i, \tilde{y}_i^{(1)}, ..., \tilde{y}_i^{(R)}\}_{i=1,...,N}$. The method only requires adding a regularization term, that is the average accuracy of all annotator models, to the cross-entropy loss function. Intuitively, the method biases ours models of each annotator to be as inaccurate as possible while having the model still explain the data. We will show that this is capable of decoupling the annotation noise from the true label distribution, as long as the average labels of the real annotators are "sufficiently" correct (which we formalize in Sec. 2.3). For simplicity, we first describe the method in the *dense label* scenario in which each image has labels from all annotators, and then extend to scenarios with *missing* labels where only a subset of annotators label each image. As we shall see later, the method works even when each image is only labelled by a single annotator.

2.1. Noisy Observation Model

We first describe our probabilistic model of the observed noisy labels from multiple annotators. In particular, we make two key assumptions: (1) annotators are statistically independent, (2) annotation noise is independent of the input image. By assumption (1), the probability of observing noisy labels $\{\tilde{y}^{(1)}, ..., \tilde{y}^{(R)}\}$ on image $\mathbf{x}$ can be written as:

$$p(\tilde{y}^{(1)}, ..., \tilde{y}^{(R)}|\mathbf{x}) = \prod_{r=1}^{R} \int_{y \in \mathcal{Y}} p(\tilde{y}^{(r)}|y, \mathbf{x}) \cdot p(y|\mathbf{x}) dy \quad (1)$$

where $p(y|\mathbf{x})$ denotes the true label distribution of the image, and $p(\tilde{y}^{(r)}|y, \mathbf{x})$ describes the noise model by which annotator r corrupts the ground truth label y. For classification problems, the label y takes a discrete value in $\mathcal{Y} = \{1, ..., L\}$. From assumption (2), the probability that annotator r corrupts the GT label $y = i$ to $\tilde{y}^{(r)} = j$ is independent of the image $\mathbf{x}$ i.e. $p(\tilde{y}^{(r)} = j|y = i, \mathbf{x}) = p(\tilde{y}^{(r)} = j|y = i) =: a_{ji}^{(r)}$. Here we refer to the associ-

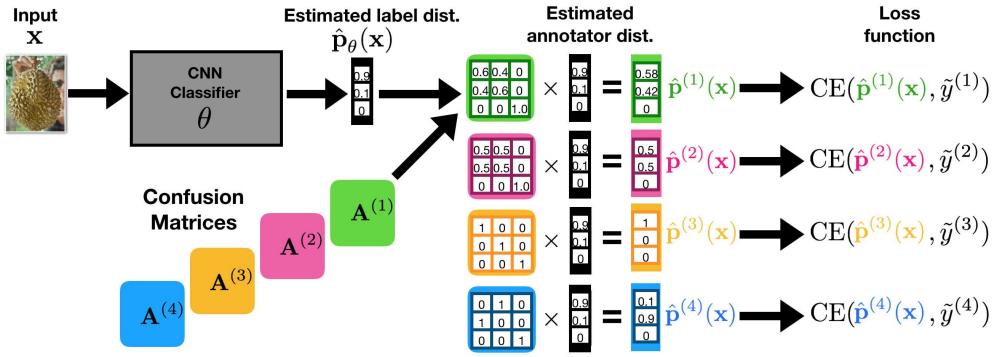


Figure 1: General schematic of the model (eq. 2) in the presence of 4 annotators. Given input image $\mathbf{x}$, the classifier parametrised by θ generates an estimate of the ground truth class probabilities, $\mathbf{p}_\theta(\mathbf{x})$. Then, the class probabilities of respective annotators $\mathbf{p}^{(r)}(\mathbf{x}) := \mathbf{A}^{(r)}\mathbf{p}_\theta(\mathbf{x})$ for $r \in \{1, 2, 3, 4\}$ are computed. The model parameters $\{\theta, \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}, \mathbf{A}^{(4)}\}$ are optimized to minimize the sum of four cross-entropy losses between each estimated annotator distribution $\mathbf{p}^{(r)}(\mathbf{x})$ and the noisy labels $\tilde{y}^{(r)}$ observed from each annotator. The probability that each annotator provides accurate labels can be estimated by taking the average diagonal elements of the associated confusion matrix (CM), which we refer to as the "skill level" of the annotator.

ated $L \times L$ transition matrix $\mathbf{A}^{(r)} = (a_{ji}^{(r)})$ as the *confusion matrix* (CM) of annotator r. The joint probability over the noisy labels is simplified to:

$$p(\tilde{y}^{(1)}, ..., \tilde{y}^{(R)}|\mathbf{x}) = \prod_{r=1}^{R} \sum_{y=1}^{L} a_{\tilde{y}^{(r)},y}^{(r)} \cdot p(y|\mathbf{x}) \qquad (2)$$

Fig. 1 provides a schematic of our overall architecture, which models the different constituents in the above joint probability distribution. In particular, the model consists of two components: the *base classifier* which estimates the ground truth class probability vector $\hat{\mathbf{p}}_\theta(\mathbf{x})$ whose i^{th} element approximates $p(y = i|\mathbf{x})$, and the set of the CM estimators $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^{R}$ which approximate $\{\mathbf{A}^{(r)}\}_{r=1}^{R}$. Each product $\hat{\mathbf{p}}^{(r)}(\mathbf{x}) := \hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x})$ represents the estimated class probability vector of the corresponding annotator. At inference time, we use the most confident class in $\hat{\mathbf{p}}_\theta(\mathbf{x})$ as the final classification output. Next, we describe our optimization algorithm for jointly learning the parameters of the base classifier, θ and the CMs, $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^{R}$.

2.2. Joint Estimation of Confusion and True labels

Given training inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ and noisy labels $\widetilde{\mathbf{Y}}^{(r)} = \{\tilde{y}_i^{(r)}\}_{i=1}^{N}$ for $r = 1, ..., R$, we optimize the parameters $\{\theta, \hat{\mathbf{A}}^{(r)}\}$ by minimizing the negative log-likelihood (NLL), $-\log p(\widetilde{\mathbf{Y}}^{(1)}, ..., \widetilde{\mathbf{Y}}^{(R)}|\mathbf{X})$. From eq. 2, this optimization objective equates to the sum of cross-entropy losses between the observed labels and the estimated annotator label distributions:

$$-\log p(\widetilde{\mathbf{Y}}^{(1)}, ..., \widetilde{\mathbf{Y}}^{(R)}|\mathbf{X}) = \sum_{i=1}^{N}\sum_{r=1}^{R} \mathrm{CE}(\mathbf{A}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}). \qquad (3)$$

Minimizing above encourages each annotator-specific prediction $\hat{\mathbf{p}}^{(r)}(\mathbf{x}) := \hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x})$ to be as close as possible to the noisy label distribution of the corresponding annotator $\mathbf{p}^{(r)}(\mathbf{x})$. However, this loss function alone is not capable of separating the annotation noise from the true label distribution; there are infinite combinations of $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^{R}$ and classification model $\hat{\mathbf{p}}_\theta$ such that $\hat{\mathbf{p}}^{(r)}$ perfectly matches the annotator's label distribution $\mathbf{p}^{(r)}$ for any input $\mathbf{x}$.

To formalize this problem, we denote the CM of the estimated true label distribution[1] $\hat{\mathbf{p}}_\theta$ by $\mathbf{P}$. The CM of the estimated annotator's label distribution $\hat{\mathbf{p}}^{(r)}$ is then given by the product $\hat{\mathbf{A}}^{(r)}\mathbf{P}$. Minimizing the cross-entropy loss (eq. 3) encourages $\hat{\mathbf{A}}^{(r)}\mathbf{P}$ to converge to the true CM of the corresponding annotator $\mathbf{A}^{(r)}$ i.e. $\hat{\mathbf{A}}^{(r)}\mathbf{P} \to \mathbf{A}^{(r)}$. However, there are infinitely many solutions pairs $(\hat{\mathbf{A}}^{(r)}, \mathbf{P})$ that satisfy the equality $\hat{\mathbf{A}}^{(r)}\mathbf{P} = \mathbf{A}^{(r)}$. This means that we need to regularize the optimization to encourage convergence to the desired solutions i.e. $\hat{\mathbf{A}}^{(r)} \to \mathbf{A}^{(r)}$ and $\mathbf{P} \to \mathbf{I}$.

To combat this problem, we propose to add the trace of the estimated CMs to the loss in eq. 3. Extending to the "missing labels" regime in which only a subset of annotators label each example, we derive the combined loss:

$$\sum_{i=1}^{N}\sum_{r=1}^{R} \mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i)) \cdot \mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda \sum_{r=1}^{R} \mathrm{tr}(\hat{\mathbf{A}}^{(r)}) \qquad (4)$$

where $\mathcal{S}(\mathbf{x})$ denotes the set of all labels available for image $\mathbf{x}$, and $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. We simply perform gradient descent on this loss to learn $\{\theta, \hat{\mathbf{A}}^{(1)}, ..., \hat{\mathbf{A}}^{(R)}\}$.

Numerous previous work have considered the same observation model, but proposed various optimization schemes. The original work [15, 20] employed the generalized EM algorithm to estimate $\{\theta, \hat{\mathbf{A}}^{(1)}, ..., \hat{\mathbf{A}}^{(R)}\}$, and more recent work [17, 18] employed variants of hard-EM to optimize the same model. Khetan et al.,[19] proposed a

[1] $\mathbf{P}_{ji} = \int_{\mathbf{x}\in\mathcal{X}} p(\mathrm{argmax}_k[\hat{\mathbf{p}}_\theta(\mathbf{x})]_k = j|y = i)p(\mathbf{x})d\mathbf{x}$

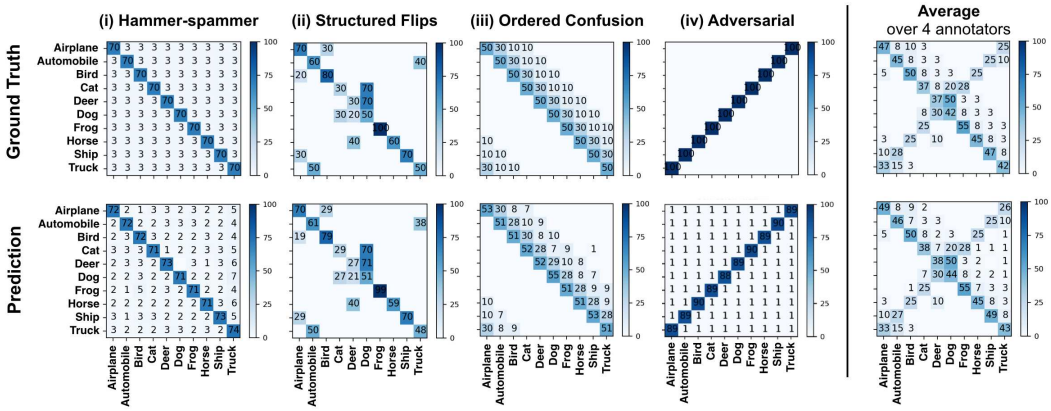


Figure 2: A diverse set of 4 simulated annotators on CIFAR-10. The top row shows the ground truths while the bottom row are the estimation from our method, trained with only one label per image.

method called model-bootstrapped EM (MBEM) in which the predictions of the base neural network classifier are used in the M-step update of CMs to learn from singly labelled data, which was not viable with the prior work. However, in all of the above EM-based methods, each M-step for the parameters of NN is not available in closed form and thus performed via gradient descent. This means that every M-step requires a training of the CNN classifier, rendering each iteration of EM expensive. A naive solution to this is to perform only few iterations of gradient descent in each E-step, however, this could limit the performance if sufficient convergence is not achieved. Our approach directly maximizes the likelihood with the trace regularizer and does not suffer from these issues. In Sec. 4, we show empirically this approach leads to an improvement both in terms of accuracy and convergence rate over the previous methods on noisy labels with high inter-annotator variability.

2.3. Motivation for Trace Regularization

Here we intend to motivate the addition of the trace regularizer in eq. 4. In the last section, we saw that minimizing cross-entropy loss alone encourages $\hat{\mathbf{A}}^{(r)}\mathbf{P} \rightarrow \mathbf{A}^{(r)}$. Therefore, if we could devise a regularizer which, when minimized, uniquely ensures the convergence $\hat{\mathbf{A}}^{(r)} \rightarrow \mathbf{A}^{(r)}$, then this would make $\mathbf{P}$ tend to the identity matrix, implying that the base model fully captures the true label distribution i.e. $\text{argmax}_k[\mathbf{p}(\hat{\mathbf{x}})_\theta]_k = y \ \forall \mathbf{x}$. We describe below the trace regularizer is indeed a such regularizer when both $\hat{\mathbf{A}}^{(r)}$ and $\mathbf{A}^{(r)}$ satisfy some conditions. We first show this result assuming that there is a single annotator, and then extend to the scenario with multiple annotators.

Lemma 1 (Single Annotator)**.** *Let $\boldsymbol{P}$ be the CM of the estimated true labels $\hat{\boldsymbol{p}}_\theta$ and $\hat{\boldsymbol{A}}$ be the estimated CM of the annotator. If the model matches the noisy label distribution of the annotator i.e. $\hat{\boldsymbol{A}}\boldsymbol{P} = \boldsymbol{A}$, and both $\hat{\boldsymbol{A}}$ and $\boldsymbol{A}$ are diagonally dominant ($a_{ii} > a_{ij}$, $\hat{a}_{ii} > \hat{a}_{ij}$) for all $i \neq j$, then $\hat{\boldsymbol{A}}$ with the minimal trace uniquely coincides with the true $\boldsymbol{A}$.*

Proof. We show that each diagonal element in the true CM $\mathbf{A}$ forms a lower bound to the corresponding element in its estimation.

$$a_{ii} = \sum_j \hat{a}_{ij} p_{ji} \leq \sum_j \hat{a}_{ii} p_{ji} = \hat{a_{ii}} (\sum_j p_{ji}) = \hat{a}_{ii} \quad (5)$$

for all $i \in \{1, ..., L\}$. It therefore follows that $\text{tr}(\mathbf{A}) \leq \text{tr}(\hat{\mathbf{A}})$. We now show that the equality $\hat{\mathbf{A}} = \mathbf{A}$ is uniquely achieved when the trace is the smallest i.e. $\text{tr}(\mathbf{A}) = \text{tr}(\hat{\mathbf{A}}) \Rightarrow \mathbf{A} = \hat{\mathbf{A}}$. From (5), if the trace of $\mathbf{A}$ and $\hat{\mathbf{A}}$ are the same, we see that their diagonal elements also match i.e. $a_{ii} = \hat{a}_{ii} \forall i \in \{1, ..., L\}$. Now, the non-negativity of all elements in CMs $\mathbf{P}$ and $\hat{\mathbf{A}}$, and the equality $a_{ii} = \sum_j \hat{a}_{ij} p_{ji}$ imply that $p_{ji} = \mathbb{1}[i = j]$ i.e. $\mathbf{P}$ is the identity matrix.

□

We note that the above result was also mentioned in [22] in a more general context of label noise modelling (that neglects annotator information). Here we further augment their proof by showing the uniqueness of solutions (i.e. $\text{tr}(\mathbf{A}) = \text{tr}(\hat{\mathbf{A}}) \Rightarrow \mathbf{A} = \hat{\mathbf{A}}$). In addition, the trace regularization was never used in practice in [22] — for implementation reason, the Frobenius norm was used in all their experiments. We now extend this to the multiple annotator regime. We will show later that minimizing the mean trace of all annotators indeed enhances the estimation quality of both CM and true label distributions, particularly in the presence of high annotator disagreement.

Theorem 1 (Multiple Annotators)**.** *Let $\hat{\boldsymbol{A}}^{(r)}$ be the estimated CM of annotator r. If $\hat{\boldsymbol{A}}^{(r)}\boldsymbol{P} = \boldsymbol{A}^{(r)}$ for $r = 1, ..., R$, and the average true and estimated CMs $\boldsymbol{A}^* := R^{-1}\sum_{r=1}^R \boldsymbol{A}^{(r)}$ and $\hat{\boldsymbol{A}}^* := R^{-1}\sum_{r=1}^R \hat{\boldsymbol{A}}^{(r)}$ are diagonally dominant, then $\boldsymbol{A}^{(1)}, ..., \boldsymbol{A}^{(R)} = \text{argmin}_{\hat{\boldsymbol{A}}^{(1)}, ..., \hat{\boldsymbol{A}}^{(R)}} \left[tr(\hat{\boldsymbol{A}}^*)\right]$ and such solutions are unique. In other words, when the trace of the mean CM is minimized, the estimation of respective annotator's CMs match the true values.*

Proof. As the average CMs $\mathbf{A}^*$ and $\hat{\mathbf{A}}^*$ are diagonally dominant and we have $\mathbf{A}^* = \hat{\mathbf{A}}^*\mathbf{P}$, Lemma 1 yields that $\text{tr}(\mathbf{A}^*) \leq \text{tr}(\hat{\mathbf{A}}^*)$ with equality if and only if $\mathbf{A}^* = \hat{\mathbf{A}}^*$. Therefore, when the trace of the average CM of annotators is minimized i.e. $\text{tr}(\hat{\mathbf{A}}^*) = \text{tr}(\mathbf{A}^*)$, the estimated CM of the true label distribution $\mathbf{P}$ reduces to identity, giving $\hat{\mathbf{A}}^{(r)} = \mathbf{A}^{(r)}$ for all $r \in \{1, ..., R\}$. □

The above result shows that if each estimated annotator's distribution $\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x})$ is very close to the true noisy distribution $\mathbf{p}^{(r)}(\mathbf{x})$ (which is encouraged by minimizing the cross-entropy loss), and on average for each class c, the number of correctly labelled examples exceeds the number of examples of every other class c' that are mislabelled as c (the mean CM is diagonally dominant), then minimizing its trace will drive the estimates of CMs towards the true values. To encourage $\{\hat{\mathbf{A}}^{(1)}, ..., \hat{\mathbf{A}}^{(R)}\}$ to be also diagonally dominant, we initialize them with identity matrices. Intuitively, the combination of the trace term and cross-entropy separates the true distribution from the annotation noise by finding the maximal amount of confusion which can explain the noisy observations well.

3. Experiments

We now aim to verify the proposed method on various image recognition tasks. Particularly, we demonstrate (1) advantage of our simpler optimization scheme compared to EM-based approaches (Sec. 3.2), (2) importance of modeling multiple annotators (Sec. 3.3) and (3) the applicability of the model in a challenging real world application (Sec. 3.2). We address the first two questions by testing the proposed method on MNIST and CIFAR-10 datasets with a diverse set of simulated annotators. To answer the final question, we evaluate our approach on the task of cardiac view classification using ultrasound images where the labels are noisy and sparse, and are acquired from multiple annotators of varying levels of expertise.

3.1. Set-Up

We focus on a regime in which models have only access to noisy labels from multiple annotators. For MNIST and CIFAR-10 data sets, we simulate noisy labels from a range of annotators with different skill levels and biases.

MNIST Experiments. We consider two different models of annotator types: (i) *pairwise-flipper*: each annotator is correct with probability p or flips the label of each class to another label (the flipping target is chosen uniformly at random for each class), (ii) *hammer-spammer*: each annotator is always correct with probability p or otherwise chooses labels uniformly at random [19]. For each annotator type and skill level p, we create a group of 5 annotators by generating CMs from the associated distribution (illustration of CMs are given in the supplementary material). Given the GT labels, we generate noisy labels as defined by the CM per annotator. These noisy labels are used during training.

CIFAR-10 Experiments. We consider a diverse group of 4 annotators with different patterns of CMs as shown in Fig. 2: (i) is a "hammer-spammer" as defined above, (ii) tends to mix up semantically similar categories of images e.g. cats and dogs, and automobiles and trucks, (iii) is likely to confuse "neighbouring" classes and (iv) is an adversarial annotator who has a wrong association of class names to object categories. On average, labels generated by these annotators are correct only 45% of the time.

In synthetic experiments, we assume that equal number of labels are generated by each annotator on average. We also note that all models are trained on noisy labels and do not have access to the ground truth. Unless otherwise stated, we hold out 10% of training images as a validation set, on which the best performing model is selected. We also perform no data augmentation during training. Full details of training and model architectures are provided in the supplementary material. In Sec. 3.2 and Sec. 3.3 below, we compare our model against two separate sets of baselines to address different questions.

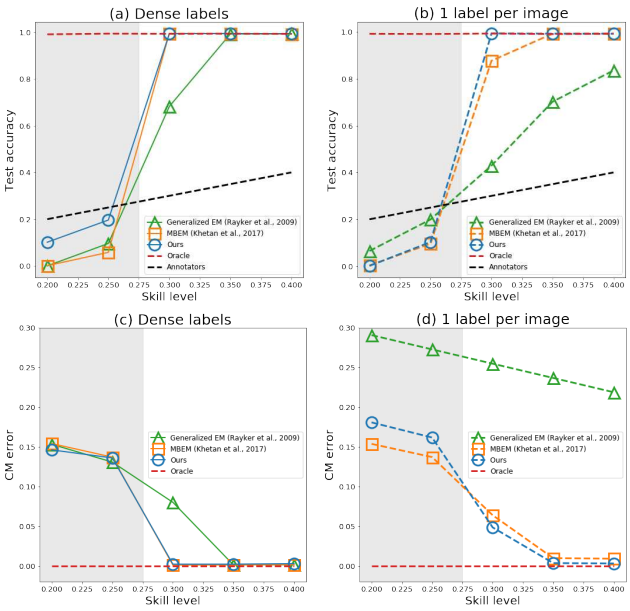


Figure 3: Comparison between our method, generalized EM, MBEM trained on noisy labels on MNIST from "pairwise flippers" for a range of mean skill level p. (a), (b) show classification accuracy in two cases, one where all annotators label each example and the other where only one label is available per example. (c), (d) quantify the CM recovery error as the annotator-wise average of the normalized Frobenius norm between each ground truth CM and its estimate. The shaded areas represent the cases where the average CM over the annotators are not diagonally dominant.

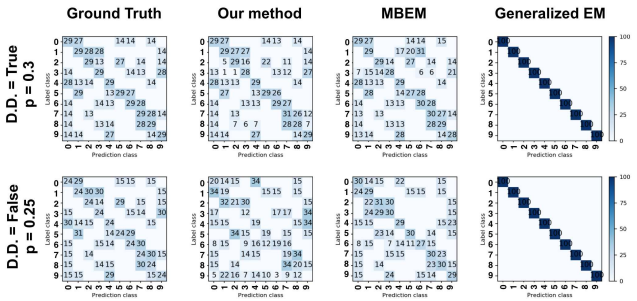


Figure 4: Visualization of the mean CM estimates when the diagonal dominance (D.D.) holds (mean skill level, $p = 0.3$) and does not hold ($p = 0.25$). In all cases, only one label is provided per image. The numbers are rounded to nearest integers. Here the respective models are trained on the noisy labels from 5 "pairwise flippers". Note that when each image receives only 1 label, the generalised EM [15] completely fails to recover the CM due to the failure of M-step for updating the confusion matrices (see Algorithm. 2 in the supplementary material).

3.2. Comparing with EM-based Approaches

This section examines the ability of our method in learning the CMs of annotators and the GT label distribution on MNIST and CIFAR-10. In particular, we compare against two prior methods: (1) generalized EM [20], the first method for end-to-end training of the CM model in the presence of multiple annotators, and (2) Model Bootstrapped EM (MBEM) [19], the present state-of-the-art method. We analyze the performance in two cases, one in which all labels from 5 annotators are available for each image ("dense labels"), and another where only one randomly selected annotator labels each example ("1 label per image"). We quantify the error of CM estimation by the average Frobenius norm between each CM and its estimate over the annotators, and this metric is normalized to be in the range $[0, 1]$ by dividing by the number of classes L i.e. $R^{-1}L^{-1}\sum_r \sum_{i,j} ||a_{ij}^{(r)} - \hat{a}_{ij}^{(r)}||^2$.

Performance Comparison. Fig. 3 compares the classification accuracy and the error of CM estimation on MNIST for a range of mean skill-levels p where labels are generated by a group of 5 "pairwise-flippers". The "oracle" model is the idealistic scenario where CMs of the annotators are a priori known to the model while "annotators" indicate the average labeling accuracy of each annotator group.

Fig. 3 shows a strong correlation between the classification accuracy and the error of CM estimation. We observe our model displays consistently better or comparable performance in terms of both classification accuracy and estimation of CMs with dense labels (Fig. 3(a) and (c)). When each example receives only one label from one of the annotators, we observe the same trend as long as the mean CMs are diagonally dominant (Fig. 3(b,d)). We also observe that when the diagonal dominance holds, all three methods perform better than the annotators. On the other hand, when the diagonal dominance does not hold (see the grey regions), all models undergo a steep drop in classification accuracy due to the inability to estimate CMs accurately as reflected in Fig. 3(c,d), which is consistent with Theorem. 1. Fig. 4 also visualizes the average of the estimated CMs at this break point. We also note that with only one label per image, the generalized EM algorithm [15, 20] is not capable of recovering CMs at all and predict identity matrices (Fig. 4), which renders the model equivalent to a vanilla classifier directly trained on noisy labels. A similar set of results in the "spammer-hammer" case are also available in the supplementary materials.

Method	Accuracy	CM error
Our method	$\mathbf{81.23 \pm 0.21}$	$\mathbf{0.72 \pm 0.01}$
Our method (no trace norm)	80.29 ± 0.65	1.37 ± 0.12
MBEM [19]	73.33 ± 0.46	2.53 ± 0.24
generalized EM [15]	70.49 ± 0.23	6.13 ± 0.28
Single CM [22]	68.82 ± 2.27	-
Weighted Doctor Net [24]	60.11 ± 1.80	-
Soft-bootstrap [21]	54.73 ± 1.33	-
Vanilla CNN [21]	52.33 ± 0.31	-

(a) Dense labels

Method	Accuracy	CM error
Our method	$\mathbf{77.65 \pm 0.31}$	$\mathbf{1.22 \pm 0.01}$
Our method (no trace norm)	76.31 ± 0.49	1.46 ± 0.27
MBEM [19]	55.97 ± 1.23	4.58 ± 0.64
generalized EM [15]	53.38 ± 0.71	4.47 ± 0.64
Single CM [22]	59.91 ± 0.98	-
Weighted Doctor Net [24]	57.98 ± 0.14	-
Soft-bootstrap [21]	42.91 ± 1.08	-
Vanilla CNN [21]	36.04 ± 1.04	-

(b) 1 label per image

Table 1: Mean classification accuracy and CM estimation errors ($\times 10^{-2}$) on CIFAR-10 with dense labels. Average annotator accuracy is 45%. Standard deviations are computed based on 3 runs with varied weight initialization.

On CIFAR-10 dataset, Tab. 1 shows that our method outperforms MBEM and the generalized EM in terms of both classification accuracy and CM estimation by a large margin. In addition, the standard deviations of these metrics are generally smaller for our method than for the baselines. Fig. 2 illustrates that our method can estimate CMs of the 4 very different annotators even when each image receives only one label. Interestingly, Tab. 1 shows that even removing the trace norm can achieve reasonably high classification accuracy and low CM estimation error. We believe this is because of the unexplained robustness of a deep CNN to label noise. Nevertheless, adding the trace norm improves the performance, and we also observe on MNIST that such improvement is pronounced in the presence of larger noise (see supplementary materials).

Sensitivity to Hyper-parameters. We next study the robustness of our method against the generalized EM and MBEM to the specification of hyper-parameters. We used the group of five pairwise-flippers with the mean skill level $p = 0.35$ to generate noisy labels on MNIST data set. For our model, we compare the effects of the scaling λ of the trace-norm in eq. 4 on the trajectory of classification accuracy on the validation set and the quality of CM estimation. For the baselines, we experiment by varying the number of EM steps (denoted by T) and the number of stochastic gradient descent for each E-step (denoted by G) while fixing the total number of training iterations at $100,000$. We observed our model presents robustness to different values of λ as long as the trace-norm loss is not larger than the cross-entropy loss (where the estimated CMs will start to diffuse too much), and Fig. 5 shows the stability of the validation curves for $\lambda \in \{0.1, 0.01, 0.001\}$. Both the MBEM and generalized EM show evident dependence on the values of T and G and by and large display slower convergence than our method. We also observe that if too few gradient descents are performed ($G = 1000$) during each E-step, the model converges to a lower accuracy in both classification and CM estimation.

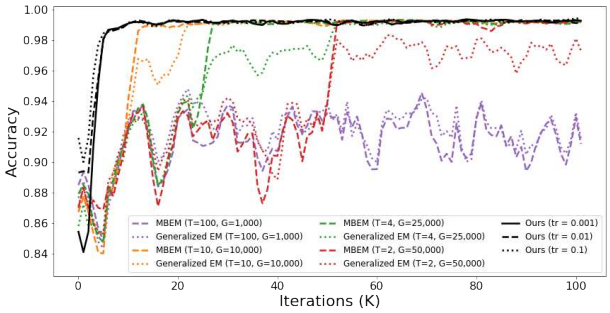


Figure 5: Curves of validation accuracy during training of our method, generalized EM and MBEM for a range of hyper-parameters. For our method, the scaling of the trace regularizer is varied in $[0.001, 0.01, 0.1]$. while, for EM and MBEM, we vary the number of EM steps (T), and the number of gradient descent steps per E-step (G) while fixing the total number of training iterations at $100,000$.

3.3. Value of Modelling Individual Annotators

Now, we compare the performance of our method against the prior work that aim to improve robustness to noisy labels without explicitly modelling the individual annotators. The first baseline is the vanilla classifier trained on the majority vote labels. We also compare against the noise robust approaches proposed in [21] and [22]. Reed *et al.* [21] adds to the cross-entropy loss a label consistency term based on the negative entropy of the softmax outputs, and we used the default hyper-parameter $\beta = 0.95$ for comparison. Sukhbaatar *et al.* [22] explicitly accounts for the label noise with a single CM, but does not model individual

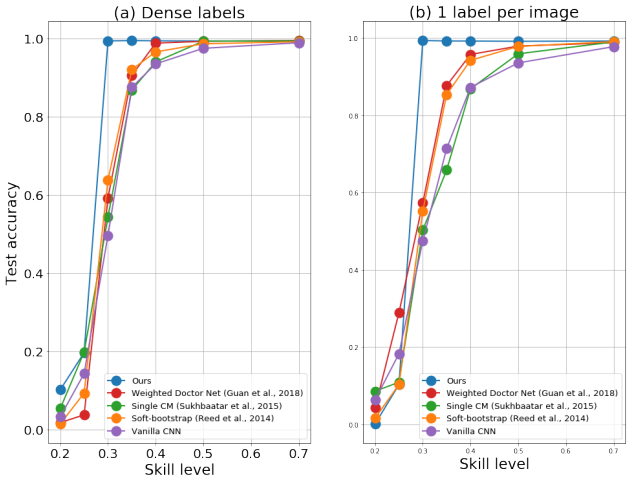
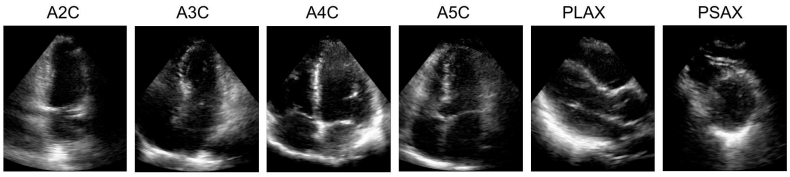


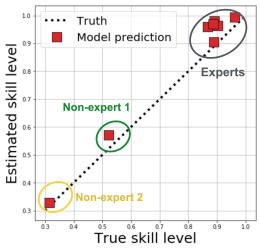
Figure 6: Classification accuracy on MNIST of different noise-robust models as a function of the mean annotator skill level p in two cases. Here, for each mean skill-level p, a group of 5 "pairwise flippers" is formed and used to generate labels. (a). each example receives labels from all the annotators. (b). each example is labelled by only 1 randomly selected annotator.

annotators. We add the trace-norm of the same scaling used in our method ($\lambda = 0.01$) to the loss function for training. We also include Weighted Doctor Net architecture (WDN) [24] in the comparison, a recent method that models the annotators individually and then learns averaging weights for combining them. It should be noted that this model considers a different observation model of the labels and does not explicitly model the true label distribution. When we have access to multiple labels per example, with the exception of WDN, we aggregated the labels by computing the majority vote and trained all models. This is because we observed a consistent improvement on validation accuracy (thus poses a tougher challenge against our method) and this would be a more realistic utilization of such data set. For both MNIST and CIFAR-10 experiments, we test on the same set of simulated labels as used in Sec. 3.2.

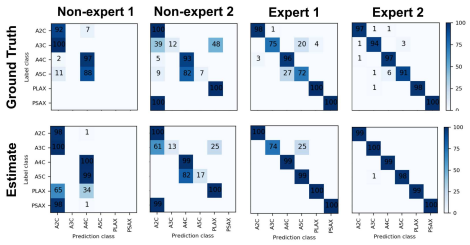
Fig. 6 shows better or comparable classification accuracy than all the baselines when the diagonal dominance of the mean CM holds. In particular, our methods show significant improvement when the mean skill level of the annotators are relatively low (e.g. $p = 0.3$ and 0.35). The results are pronounced in the case with only one label available per image for which the baseline methods undergo a steep drop in accuracy (see Fig. 6(b)). Results in the "spammer-hammer" case are available in the supplementary material. Similarly on CIFAR-10 data set, Tab. 1 shows that our method improves the classification accuracy upon the baselines. Such improvement is pronounced in the case of sparse labels. On the other hand, a vanilla CNN with only L2 weight decay overfits to the training data very quickly in the presence of such high noise.



(a) Different classes of cardiac views



(b) Skill estimation



(c) Learned CMs

Method	Accuracy	CM error
Ours	$\mathbf{75.57 \pm 0.16}$	$\mathbf{11.48 \pm 0.48}$
w/o trace norm	70.99 ± 3.31	15.22 ± 0.94
MBEM [19]	73.91 ± 0.11	12.18 ± 0.29
WDN [24]	59.15 ± 1.60	-
Single-CM [22]	74.38 ± 0.29	-
Soft-bootstrap [21]	72.99 ± 0.17	-
Vanilla CNN	70.95 ± 0.44	-

(d) Performance comparison

Figure 7: Results on the cardiac view classification dataset: (a) illustrates examples of different cardiac view images. (b) plots the estimated skill level of each annotator (average of the diagonal elements of its estimated CM) against the ground truth (c) compares the estimated CMs of the two least skilled and two most skilled annotators according the GT labels (d) summarizes the classification accuracy and error of CM estimation for different methods.

3.4. Experiments on Cardiac View Classification

Lastly, we illustrate the results of our approach for a real data set with sparse and noisy labels from the medical domain. This data set consists of images of the cardiac region in different views, acquired using a hand-held ultrasound probe. The task is to classify a given ultrasound image into one of six different view classes (see Fig. 7(a)). The process of obtaining a cardiac view label is crucial for guiding the user to the correct locations of measurements, and affects the quality of the downstream cardiac tasks.

A committee of sonographers (with varying levels of experience) were tasked with providing the cardiac view labels to a large volume of ultra-sound images, and each example is only labelled by a subset of them. To acquire ground truth in this setting, we chose those samples where the three most experienced sonographers agreed on a given label. The resulting data set consists of noisy labels provided by the remaining less experienced 6 sonographers for a total of $240,000$ training images and $22,000$ validation images. In addition, we also acquired labels from two non-expert users and included in the training data.

We estimated the skill-level of each annotator by computing the average value of the diagonal elements in the corresponding learned CM, and Fig. 7(b) shows that the group of experts can be separated from the two non-experts with varying levels of experinces (one is less competent than the other). Fig. 7(c) shows that confusion between $A3C$ and $A5C$, even common among experts, can be detected (see the result for 'Expert 1') while clearly capturing the patterns of mistakes for the non-experts. In addition, Fig. 7(d) shows that our model outperforms MBEM [19] again in classification accuracy and the quality of CM estimation. Lastly, the higher classification accuracy of our model with respect to the other baseline models illustrates again that modelling individual annotators improves robustness to label noise.

4. Discussion and Conclusion

We introduced a new theoretically grounded algorithm for simultaneously recovering the label noise of multiple annotators and the ground truth label distribution. Our method enjoys implementation simplicity, requiring only adding a regularization term to the loss function. Experiments on both synthetic and real data sets have shown superior performance over the common EM-based methods in terms of both classification accuracy and the quality of confusion matrix estimation. Comparison against the other modern noise-robust methods demonstrates that the modelling individual annotators improves robustness to label noise. Furthermore, the method is capable of estimating annotation noise even when there is a single label per image.

Our work was primarily motivated by medical imaging applications for which the number of classes are mostly limited to below 10. However, future work shall consider imposing structures on the confusion matrices to broaden up the applicability to massively multi-class scenarios e.g. introducing taxonomy based sparsity [18] and low-rank approximation. We also assumed that there is only one ground truth for each input; this no longer holds true when the input images are truly ambiguous—recent advances in modelling multi-modality of label distributions [35, 36] potentially facilitate relaxation of such assumption. Another limiting assumption is the image independence of the annotator's label noise. The majority of disagreement between annotators arise in the difficult cases. Integrating such input dependence of label noise [16, 37] is also a valuable next step.

Acknowledgments
We would like to thank Alon Daks, Israel Malkin and Pouya Samangouei at Butterfly Network for their feedback, and Dr. Linda Moy, MD of NYU Langone Medical Center for providing references on inter-reader variability in radiology. RT is supported by Microsoft Research Scholarship.

References

[1] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[3] Takeyuki Watadani, Fumikazu Sakai, Takeshi Johkoh, Satoshi Noma, Masanori Akira, Kiminori Fujimoto, Alexander A Bankier, Kyung Soo Lee, Nestor L Müller, Jae-Woo Song, et al. Interobserver variability in the ct assessment of honeycombing in the lungs. *Radiology*, 266(3):936–944, 2013.

[4] Andrew B Rosenkrantz, Ruth P Lim, Mershad Haghighi, Molly B Somberg, James S Babb, and Samir S Taneja. Comparison of interreader reproducibility of the prostate imaging reporting and data system and likert scales for evaluation of multiparametric prostate mri. *American Journal of Roentgenology*, 201(4):W612–W618, 2013.

[5] Elizabeth Lazarus, Martha B Mainiero, Barbara Schepps, Susan L Koelliker, and Linda S Livingston. Bi-rads lexicon for us and mammography: interobserver variability and positive predictive value. *Radiology*, 239(2):385–391, 2006.

[6] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[7] Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 25–32. IEEE, 2010.

[8] Chengjiang Long, Gang Hua, and Ashish Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3000–3007, 2013.

[9] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2839–2847, 2015.

[10] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[11] Padhraic Smyth, Usama M Fayyad, Michael C Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, pages 1085–1092, 1995.

[12] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[13] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.

[14] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436, 2013.

[15] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pages 889–896. ACM, 2009.

[16] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATs*, pages 932–939, 2010.

[17] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017.

[18] Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, Cornell Tech, and Pietro Perona. Lean multiclass crowdsourcing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2723, 2018.

[19] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[20] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[21] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[22] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[23] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017.

[24] Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. *AAAI*, 2018.

[25] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

[26] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[27] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 143–158. Springer, 2012.

[28] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *International conference on machine learning*, pages 567–574, 2012.

[30] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pages 2233–2241, 2017.

[31] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[32] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge J Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pages 6575–6583, 2017.

[33] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018.

[34] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2309–2318, 2018.

[35] Ardavan Saeedi, Matthew D Hoffman, Stephen J DiVerdi, Asma Ghandeharioun, Matthew J Johnson, and Ryan P Adams. Multimodal prediction and personalization of photo edits with deep generative models. *arXiv preprint arXiv:1704.04997*, 2017.

[36] Simon AA Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus H Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *arXiv preprint arXiv:1806.05034*, 2018.

[37] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.