# Adversarial Learning from Crowds

**Pengpeng Chen,**[1,3] **Hailong Sun,**[*2,3] **Yongqiang Yang,** [1,3] **Zhijun Chen** [1,3]

[1] SKLSDE Lab, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] SKLSDE Lab, School of Software, Beihang University, Beijing, China
[3] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China
{chenpp, sunhl, yangyongqiang, zhijunchen}@buaa.edu.cn

## Abstract

Learning from Crowds (LFC) seeks to induce a high-quality classifier from training instances, which are linked to a range of possible noisy annotations from crowdsourcing workers under their various levels of skills and their own preconditions. Recent studies on LFC focus on designing new methods to improve the performance of the classifier trained from crowdsourced labeled data. To this day, however, there remain under-explored security aspects of LFC systems. In this work, we seek to bridge this gap. We first show that LFC models are vulnerable to adversarial examples—small changes to input data can cause classifiers to make prediction mistakes. Second, we propose an approach, A-LFC for training a robust classifier from crowdsourced labeled data. Our empirical results on three real-world datasets show that the proposed approach can substantially improve the performance of the trained classifier even with the existence of adversarial examples. On average, A-LFC has 10.05% and 11.34% higher test robustness than the state-of-the-art in the white-box and black-box attack settings, respectively.

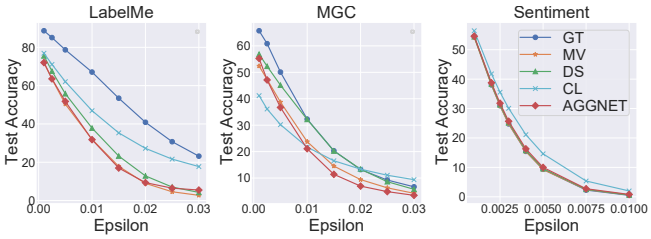



Figure 1: The distinctive influence of adversarial examples on LFC models: (i) GT, the neural networks trained in the ideal case when ground truth labels are known; (ii) MV/DS, first estimates the ground truth by MV/DS model and then trains the neural networks; (iii) CL; (iv) AggNet. We test the robustness (test accuracy (%)) of different strategies on three real-world benchmarks (LabelMe (Rodrigues and Pereira 2018), MGC (Rodrigues, Pereira, and Ribeiro 2013), and Sentiment (Rodrigues, Pereira, and Ribeiro 2013)) under the FGSM (Goodfellow, Shlens, and Szegedy 2015) with varying the perturbation scale $\epsilon$.

## Introduction

A significant prerequisite for the use of supervised learning is the availability of large, well-labeled datasets. Crowdsourcing (Zheng et al. 2017; Fang et al. 2018; Tong et al. 2020) offers an affordable method of annotating data by using freelance workers located on the internet platforms such as Amazon Mechanical Turk [1](AMT). When it comes to crowdsourced labeled data, a general rule of thumb is that the annotations may often be noisy because of the unskilled or malevolent behavior of workers. To alleviate this issue, the common practice is to ask numerous workers to provide labels for each instance. In the crowdsourcing-learning scenario, a question is raised: how to learn a decent classifier with the noisy labeled data.

The straightforward solution to the problem is to first estimate the latent true labels using the answer aggregation techniques such as majority voting (MV) (Sheng, Provost, and Ipeirotis 2008), Dawid Skene (DS) (Dawid and Skene 1979), and then train the neural networks with the aggregated labels. Alternatively, more recently proposed one-stage methods (Luo et al. 2018; Yang et al. 2018a) such as CrowdLayer (CL) (Rodrigues and Pereira 2018) and AggNet (Albarqouni and Baur 2016) simultaneously infer the true labels while learning the parameters of the deep neural network and the confusion matrices of annotators.

We note that most models in the LFC family are based on the assumption that all the examples are *benign* (Cao et al. 2019; Chen et al. 2020b) and focus on producing accurate classifiers with the estimation of ground truth labels inferred from the noisy labels of crowd workers. Unfortunately, recent studies (Goodfellow, Shlens, and Szegedy 2015; Dong et al. 2020) have found that even in the ideal case when ground truth labels are known, the classifier trained from instances could perform rather poorly in presence of adversarial examples—small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich (Carlini and Wagner 2017). In crowdsourcing settings, the noise of crowd labels aggravates the vulnerability of LFC models and existing LFC models can hardly learn a robust model, without the consideration of the ex-

*Corresponding author

[1]https://www.mturk.com

istence of adversarial examples. For instance, as shown in Figure 1, in most cases of LFC models, the final robustness drops drastically, compared with the classifier trained from instances with ground truth labels [2]. Despite such vulnerability of LFC models under adversarial attacks, LFC applications (Luo et al. 2018; Cao et al. 2019) are still prevalent in practice and even in some security-sensitive domains such as medical imaging (Raykar et al. 2010; Albarqouni and Baur 2016). Hence, there is an urgent need to investigate and understand the adversarial attacks and defense for the LFC family.

In this work, we move one step further and explore how to learn an LFC model robust to the adversarial examples. We first show that LFC models are vulnerable to adversarial attacks, in which small changes to input data can cause classifiers to make prediction mistakes. In particular, we find the adversarial attacks can significantly decrease the test accuracy of trained classifiers from crowdsourced labeled data in white-box and black-box settings. To improve the adversarial robustness of LFC, we formulate the problem of learning from crowds under adversarial attacks as a bilevel min-max problem, in which the inner maximization problem serves as a constraint of the outer minimization problem. We found this problem is highly *non-linear* and *non-convex*, which is intractable to exactly solve. To address this challenge, we propose an approach for adversarial learning from crowds (A-LFC), which solves the outer minimization problem by the expectation-maximization (EM) algorithm and the inner minimization problem by projected gradient descent. We evaluate our approach using three well-known benchmark datasets in the LFC community. For instance, in one dataset called music genres classification (MGC) dataset (Rodrigues, Pereira, and Ribeiro 2013) contains 1000 samples concerning songs involving ten music genres i.e., *country*, *disco*, etc. 700 samples are randomly selected and labeled by 44 crowd workers from AMT. Experimental results show that our approach substantially increases the test robustness.

Our contributions are summarized as follows:

- We investigate the influence of adversarial examples on the performance of representative LFC models. We find that these models are very vulnerable to adversarial examples;
- We formulate the problem of LFC in the environment as a bilevel min-max problem and propose a novel approach, A-LFC for training classifiers robust to the adversarial examples;
- We conducted an extensive evaluation of the proposed A-LFC, showing that A-LFC is able to outperform the state-of-the-art in both white-box and black-box settings.

To the best of our knowledge, this is the first work on exploring the impact of adversarial examples on learning from crowdsourced labeled data and developing the LFC model robust to adversarial example attacks.

[2]Details of parameter setting are provided in the section of the experimental setup.

## Related Work

**Answer Aggregation in Crowdsourcing.** In two-stage approaches, the true labels are inferred from the crowd labels using answer aggregation methods (Chen et al. 2018, 2020a). Then, it applies the general supervised learning methods along with the inferred labels (Wang and Zhou 2016). In the stage of answer aggregation, the simplest model, MV (Sheng, Provost, and Ipeirotis 2008), derives the majority labels by counting the workers' votes for each alternative label. Due to the ignorance of varying reliability among workers, MV is error-prone. In contrast, WMV (Li and Yu 2014) and CATD (Li et al. 2014) assigns different weights to workers' votes considering workers' reliability. Besides, a major type of answer aggregation model leverages probabilistic models to estimate worker reliability. DS (Dawid and Skene 1979) models each worker's reliability with a confusion matrix and uses the EM algorithm to iteratively update the true label of each instance and the workers' confusion matrices. ZC (Demartini et al. 2012) is a simplified version of DS: it does not consider the priors and models each worker's reliability with a single probability of correct labeling. There also exist some other models that can be viewed as extensions of ZC, e.g., GLAD (Whitehill et al. 2009) and SEEK (Han et al. 2016).

In this approach, answer aggregation of labels and model training are isolated processes. After inferring the true labels, several types of additional information about each instance in the training data, such as its features, are lost. However, the additional information may be good for further improving the quality of inferred labels as well as re-training a more robust model (Zhang, Wu, and Sheng 2019).

**One-Stage Approaches.** Raykar et al. (2010) come up with the one-stage approach, LFC which jointly estimates the instances' true labels and trains the logistic regression classifier. Rodrigues et al. (2018) propose an end-to-end method named Crowd Layer which directly applies back-propagation to train deep neural networks from the crowdsourced labeled data. Considering the lack of Interpretability of Crowd Layer, Chen et al. (2020b) propose a structured end-to-end model which endows Crowd Layer the probabilistic interpretability. Chu et al. (2021) divide the confusion matrix into two components: namely frequently-shared confusion matrix and the individually-specific confusion matrix. Zhong et al. (2017) propose an approach whose objective involves the label reliability learned with the discrepancy between crowdsourced annotation from crowds and the predictions of the model. Yang et al. (2018b) describe a newly proposed model based on adversarial neural networks that relies on crowdsourced annotation data. They learn a classifier as well as the common and private features of crowd workers using data labeled by these workers.

However, the impact of adversarial attacks on the quality of learning from noisy crowdsourced labeled data has not been considered in these works. To this end, some researchers (Miao et al. 2018a,b; Fang et al. 2021) access and analyze the vulnerability of answer aggregation step of LFC models under data poisoning attacks by designing label attack strategies.

Different from these works, we are concerned with feature

attacks induced by adversarial examples. Besides accessing the vulnerability of LFC models under adversarial examples, we propose a new approach for training classifiers that is robust to the adversarial examples.

## Problem Formulation

We formulate the studied problems concerning how to use crowdsourced data to train a classifier that is robust to adversarial examples.

### LFC Problem in Adversarial Environment

We use capital letters (e.g. $\mathcal{A}$) in calligraphic math font to denote sets. We use boldface uppercase letters to denote matrices, e.g., $\mathbf{M}$, in which the entry $(i, j)$ is denoted by the corresponding lowercase letters $m_{ij}$ and the entries of $i$-th row is denoted by $\mathbf{M}_{i*}$. We use boldface lowercase letters to denote vectors (e.g., $\mathbf{v}$). Formally, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ denote the set of instances. The $i$-th instance $\mathbf{x}_i$'s unobserved ground truth is denoted by $t_i$ which takes on $K \geq 2$ possible values. Let $\mathcal{U} = \{u_j\}_{j=1}^M$ be the worker set. The workers' labels are represented as a matrix $\mathbf{Y} = (y_{ij})_{N\times M}$, in which $y_{ij}$ is the label from the worker $u_j$ to instance $\mathbf{x}_i$. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{Y}_{i*})\}_{i=1}^N$ denote the *i.i.d.* data set whose labels are from crowd workers $\mathcal{U}$, and $\mathbf{Y}_{i*} = (y_{i1}, y_{i2}, \cdots, y_{iM})$ denotes the labels from $\mathcal{U}$ to the $i$-th instance. In this work, we are concerned with the problem of how to use $\mathcal{D}$ to train a high-quality classifier $h_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta}$ for predicting $t$ given new data which may be the adversarial example.

Note that we are concerned with the feature attack induced by adversarial examples in LFC settings, which is different from the label attacks in existing crowdsourcing research (Miao et al. 2018a,b; Fang et al. 2021).

### Adversarial Learning from Crowds

Several defense approaches (Wang et al. 2020) *e.g.* model compression (Wang et al. 2020) and activation pruning (Das et al. 2018) have been proposed to train Deep Neural Networks (DNNs) intrinsically robust against adversarial examples. Among these, the most effective one is adversarial training. Thus, we choose the adversarial training technique to cope with the LFC problem in an adversarial environment.

We begin by introducing adversarial training which trains the classifier on adversarial examples and can be viewed the bilevel *min-max* problem as follows.

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \max_{\|\mathbf{x}'_i - \mathbf{x}_i\|_p \leq \epsilon} L\left(h_{\boldsymbol{\theta}}\left(\mathbf{x}'_i\right), t_i\right), \tag{1}$$

where $L(\cdot)$ denotes the objective *e.g.* the cross-entropy loss. $\mathbf{x}'_i$ denotes an adversarial example generated by solving the inner maximization problem, i.e., the natural example is correctly classified before perturbation, but misclassified after the perturbation. In general, the perturbation is of small size bounded with the $L_p$-norm which makes $\mathbf{x}'$ within the $\epsilon$-ball whose center is $\mathbf{x}$.

Traditional adversarial learning entails the prior knowledge of the ground truth label $t_i$; whereas $t_i$ is difficult to know beforehand in the setting of LFC. What we can obtain is the noisy labels provided by crowd workers. Thus, we reformulate the problem of adversarial learning from crowdsourced labeled data as follows.

$$\begin{aligned} &\min_{\Theta} -\alpha \log p(\mathbf{Y} \mid \mathcal{X}, \Theta) - (1-\alpha) \log p(\mathbf{Y} \mid \mathcal{X}', \Theta) \\ &s.t.\ \ \mathcal{X}' = \operatorname*{argmax}_{\mathcal{X}'} - \log p(\mathbf{Y} \mid \mathcal{X}', \Theta), \\ &and\ \ \mathcal{X}' = \{\mathbf{x}'_i \mid \|\mathbf{x}'_i - \mathbf{x}_i\|_p \leq \epsilon\}, \end{aligned} \tag{2}$$

where $\log p(\mathbf{Y}|\mathcal{X}, \Theta)$ and $\log p(\mathbf{Y}|\mathcal{X}', \Theta)$ denote the log conditional likelihood of the observed crowd labels given the natural examples and adversarial examples. $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(M)}\}$ and $\boldsymbol{\Pi}^{(j)}$ denotes the confusion matrix of worker $j$. $\alpha$ is the imitation parameter for balancing the two parts. In this bilevel min-max optimization problem, the outer subproblem is to minimize the objective function by finding optimal parameters $\Theta$ of the classifier and workers; the inner problem is to find the adversarial examples $\mathcal{X}'$ by maximizing the negative log-likelihood. The bilevel problem is general NP-hard (Fang et al. 2021). Since the outer problem is subject to the inner problem, directly implementing the common algorithms such as backpropagation cannot effectively resolve this optimization problem.

## Method

In this section, we provide the solution to the problem formalized in Equation 2. First, we give the approach to resolving its inner subproblem. Second, we aim at solving its outer subproblem. Finally, we introduce the proposed algorithm, A-LFC.

### Solving the Inner Problem

We reformulate the optimization objective of the inner problem as follows.

$$-\sum\nolimits_i \mathbb{E}_{\rho(t_i)} \log\left[p\left(t_i \mid \mathbf{x'}_i; \boldsymbol{\theta}\right)\right], \tag{3}$$

where $\rho\left(t_i\right)$ is estimation of $t_i$ which is obtained by solving the outer problem and $t_i \mid \mathbf{x'}_i; \boldsymbol{\theta} \sim \mathrm{Cat}\left(t_i; f_{\boldsymbol{\theta}}\left(\mathrm{x'}_i\right)\right)$ is a conditional categorical distribution. $f_{\boldsymbol{\theta}}$ is a flexible neural network model parametrized by $\boldsymbol{\theta}$.

Recently, adversarial training with adversarial examples generated by Projected Gradient Descent (PGD) (Madry et al. 2018) has been proved to be the sole way to prevent the trained DNNs from being fully attacked. To this end, we propose to use PGD to maximize the optimization objective of the inner problem.

### Solving the Outer Problem

In the outer problem, the optimization objective is minimizing the negative log conditional likelihoods of the observed crowd labels given features of natural and adversarial examples, w.r.t. the parameters $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(M)}\}$. These parameters can be estimated with the expectation-maximization (EM) algorithm that iteratively learns the parameters in the M-step and at the same time infers the latent ground truth $\rho\left(t_i\right)$ in the E-step.

**M-Step.** We do not directly optimize the objective in Equation 3. In our objective, we not only consider the performance of the trained classifier on adversarial examples but also take into account its performance on natural examples as follows.

$$-\alpha \log p(\mathbf{Y} \mid \mathcal{X}, \Theta) - (1-\alpha) \log p(\mathbf{Y} \mid \mathcal{X}', \Theta), \quad (4)$$

where

$$\log p(\mathbf{Y} \mid \mathcal{X}, \Theta) = \sum\nolimits_i \mathbb{E}_{\rho(t_i)} \log \left[ p\left(t_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right) \right]. \quad (5)$$

Our approach forces the neural network to update the parameters $\boldsymbol{\theta}$ for minimizing the objective in Equation 4.

Now, we proceed to updating the parameters of the crowd annotators $\{\mathbf{\Pi}^{(1)}, \ldots, \mathbf{\Pi}^{(M)}\}$. Similar to the optimization of the parameters of the neural network, we seek the derivative of the optimization objective w.r.t. the worker's parameter and set the gradient equal to 0. Finally, we derive the closed-form solution as follows:

$$\pi_{kk'}^{(j)} = \frac{\sum_i \rho\left(t_i = k\right) \mathbb{I}\left(y_{ij} = k'\right)}{\sum_i \rho\left(t_i = k\right) \mathbb{I}\left(y_{ij} \neq \perp\right)}, \quad (6)$$

where $y_{ij} \neq \perp$ denotes the $j$-th worker provide label to instance $i$; and $\mathbb{I}$ is an indicator function. When its internal declaration is true, it takes a value of 1, or otherwise. Then, the estimation of latent ground truth $\rho\left(t_i\right)$ will be updated in the E-step presented next.

**E-Step.** Formally, given the new parameters of the classifier and the workers provided by M-step, the EM algorithm dictates that we adhere to the recipe while additionally making use of the Bayesian formula to infer the latent variable's posterior distribution.

$$\begin{aligned} &\rho\left(t_i = k\right) \propto \prod\nolimits_j p\left(y_{ij} \mid t_i = k; \mathbf{\Pi}^{(1)}, \ldots \mathbf{\Pi}^{(N)}\right) \\ &\cdot \left(\alpha p\left(t_i = k \mid \mathbf{x}_i; \boldsymbol{\theta}\right) + (1-\alpha) p\left(t_i = k \mid \mathbf{x'}_i; \boldsymbol{\theta}\right)\right), \end{aligned} \quad (7)$$

where $p\left(y_{i,j} \mid t_i = k; \mathbf{\Pi}^{(1)}, \ldots \mathbf{\Pi}^{(N)}\right)$ is the distribution of label $y_{ij}$ from worker $j$ to instance $i$ when the ground truth label of instance $i$ is $k$.

Note that we infer the latent ground truth label by considering three components *i.e.* the worker annotations, the adversarial examples, and the natural examples, which is different from the answer aggregation methods that only infer the true labels from the worker annotations.

## A-LFC Algorithm

We introduce our mechanism named A-LFC for learning a robust model from crowds. A-LFC mainly includes two steps. The former serves to resolve the inner subproblem. And the latter is responsible for tackling the outer subproblem. These two steps are iteratively conducted to derive the optimum of the bilevel min-max problem.

**Step 1**. With the estimate $\Theta$, namely the parameters of the classifier and the workers, A-LFC applies the PGD algorithm to generate the adversarial examples $\mathcal{X}'$ that maximizes the objective in Equation 3.

**Algorithm 1:** A-LFC

**Input:** Training set $\mathcal{D}$, imitation parameter $\alpha$
**Output:** Optimal $\boldsymbol{\theta}$

```
Initialize ρ_t with MV and X' with X;
while It does not achieve convergence and the limit of iteration do
    for each minibatch of the epoch do
        Update θ by the backprop. with Equation 4
    Update the worker parameters with Equation 6;
    Update the estimation of ρ_t with Equation 7;
    Update X' using PGD algorithm with Equation 3;
return The parameters of classifier θ;
```

**Step 2**. On the basis of the generated adversarial examples $\mathcal{X}'$ in step 1, A-LFC applies EM algorithm to update the parameters $\Theta$ in M-step following Equation 4 with the backpropagation and the parameters of the workers $\{\mathbf{\Pi}^{(1)}, \ldots, \mathbf{\Pi}^{(M)}\}$ with Equation 6, and update the estimation of the ground truth in E-step following Equation 7.

The procedure of A-LFC is summarized in Algorithm 1. $\rho_t$ is initialized with the aggregated labels using MV and $\mathcal{X}'$ is initialized with $\mathcal{X}$.

# Experiments

This section presents the experimental results for evaluating the effectiveness of A-LFC [3]. Specifically, we answer the following questions:

- **Q1**: Is the proposed approach sensitive to the imitation parameter $\alpha$ and how to properly set this parameter?
- **Q2**: How well does the proposed method perform under white-box attacks?
- **Q3**: How well does the proposed method perform under black-box attacks?
- **Q4**: Is the proposed method effective for learning the confusion matrices for representing the workers?

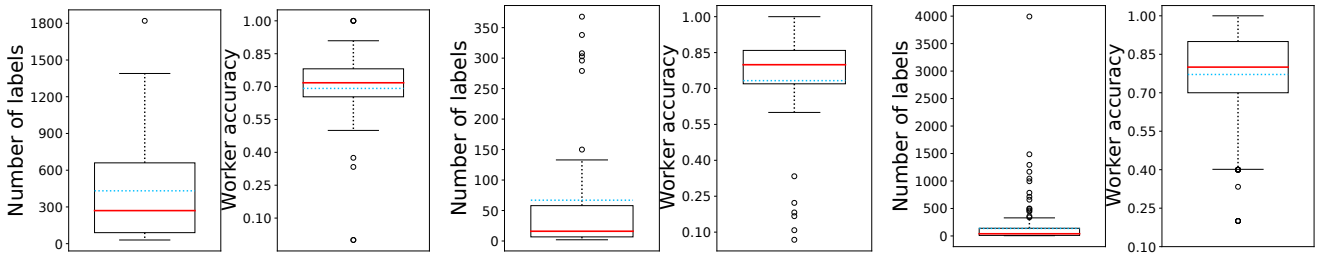
## Real-World Datasets

We use three publicly available, widely used benchmark datasets with real annotations from AMT.

**Music Genre Classification dataset (MGC).** The MGC dataset (Rodrigues, Pereira, and Ribeiro 2013) contains one thousand samples concerning songs of different genres. All songs belong to ten music genres (i.e., class 0 to class 9) and each song takes 30 seconds. The ten music genres of songs are *country*, *disco*, *rock*, *pop*, *hiphop*, *jazz*, *reggae*, *metal*, *classical*, and *blues*. 700 samples are randomly selected and labeled by 44 crowd workers from Mechanical Turk with a mean accuracy of 73.28%. Each worker provides an average of 66.93 labels. Finally, 2,946 crowd labels are obtained. The feature of each instance is extracted to 124 dimensions

**LabelMe**. Dataset LabelMe is an image classification dataset (Rodrigues and Pereira 2018) involving 8 classes

[3] Our code is available at https://github.com/yongqiangyang/A-LFC.




(a) Dataset LabelMe (b) Dataset MGC (c) Dataset Sentiment

Figure 2: Boxplots concerning the distribution of the number of labels from per worker and the distribution of the worker accuracy among the workers in dataset LabelMe, dataset MGC, and dataset Sentiment.

from class 0 to class 7, i.e., *highway*, *inside city*, *tall building*, *street*, *forest*, *coast*, *mountain*, and *open country*. Totally, it contains 2,688 instances. Among them, 1,000 instances were labeled by workers from AMT. An instance corresponds to an average of 2.547 labels elicited from the crowds. The average worker accuracy is 69.2%.

**Sentiment Polarity Classification (Sentiment)**. The Sentiment dataset (Rodrigues, Pereira, and Ribeiro 2013) contains 5000 sentences (with crowdsourced annotations) from movie reviews extracted from the website RottenTomatoes.com and their sentiment polarity was classified as positive or negative, The datasets received a total of 27747 annotations from 203 distinct annotators on AMT. For the tasks, 5429 instances are provided as test sets.

For the three real-word datasets, Figure 2 presents the boxplots concerning the distribution of the number of labels per worker and the distribution of the worker accuracy among the workers. It reveals that the distribution of the number of labels provided per worker follows a highly skewed distribution, that is a small number of workers provide the great majority of labels.

## Baselines

We compare the method with the following representative baselines.

- *MV* (Wang and Zhou 2016): MV is used to determine the true labels, after which a neural network classifier is trained on the basis of the aggregate results.
- *DS*: Like MV, but using Dawid Skene's label aggregation algorithm (Zheng et al. 2017).
- *AggNet* (Albarqouni and Baur 2016): The generic version of LFC model, in which the classifier is based on a deep neural networks.
- *Crowd Layer (CL)* (Rodrigues and Pereira 2018): We implement the MV version of CL. The deep neural network is trained directly from the crowd labels by using backpropagation.

## Experimental Setup

**Details of Parameter Setting.** For the LabelMe dataset pre-processed by the pre-trained convolutional neural networks (CNN) layers of VGG network (Rodrigues and Pereira 2018), we apply one fully connected (FC) layer with

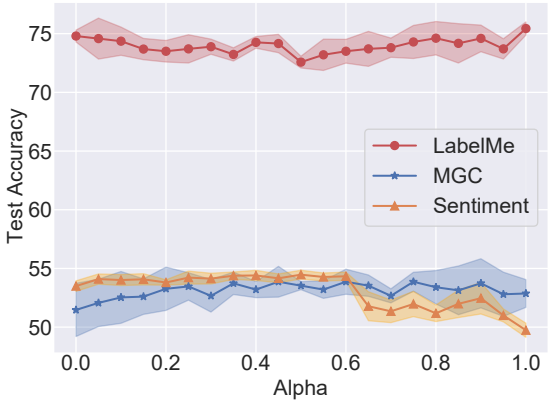
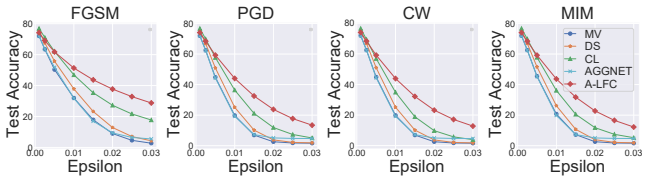



Figure 3: Sensitivity to imitation parameter $\alpha$

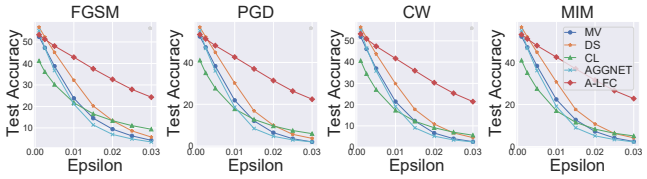
128 units. For the MGC and Sentiment dataset, we apply one fully connected (FC) as the one hidden layer with 128 units, respectively. For each dataset, we use ReLU activations, 50% dropout, and Adam stochastic optimization. The learning rate is 0.001, the batch size is 64, and the number of epoch is 200. The worker parameters were initially with the result of MV. The training attack is 10-step PGD with random start and step size $\epsilon/4$. The perturbation $\epsilon$ of training attack is 8/255 and the parameter $\alpha$ is set to 0.5. All experiments were performed 50 times on NVIDIA Tesla V100 GPUs and we report the average result.

**The Adversarial Attacks.** We evaluate the A-LFC model under the black-box and white-box attack settings, and implement the following four adversarial attack methods to generate the adversarial examples.

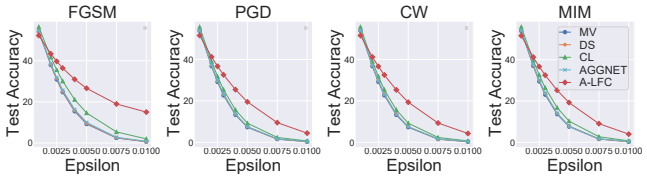
- *FGSM* (Goodfellow, Shlens, and Szegedy 2015): It is a representative method fast to execute, in which the sign of the perturbation is based on the gradients of the objective w.r.t. the examples.
- *PGD* (Madry et al. 2018): This is one of the most powerful gradient-based attacks. Given a natural example, the process of PGD starts with a random perturbation and proceeds by updating the perturbation iteratively.
- *CW* (Carlini and Wagner 2017): It minimizes a nonlinear mapped perturbation to get a considerably smaller




(a) Dataset LabelMe.

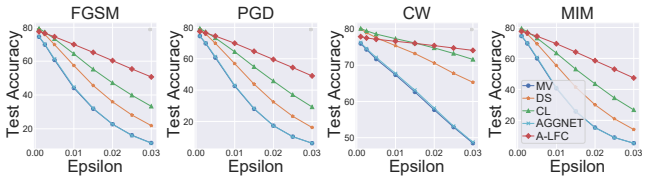



(b) Dataset MGC.

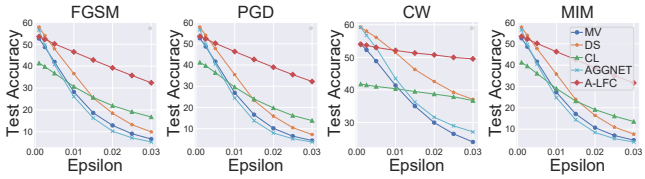



(c) Dataset Sentiment.

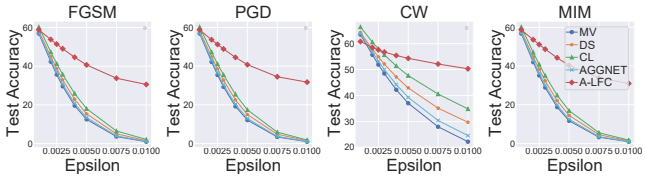
Figure 4: White-box robustness (test accuracy (%) of the classifier under white-box attacks) on real-world datasets LabelMe, MGC, and Sentiment.




(a) Dataset LabelMe.




(b) Dataset MGC.




(c) Dataset Sentiment.

Figure 5: Black-box robustness (test accuracy (%) of the classifier under black-box attacks) on real-world datasets LabelMe, MGC, and Sentiment.

perturbation size, resulting in a huge speed reduction.

– *MIM* (Dong et al. 2018): The momentum element is incorporated into the iterative process; it also helps stabilize the attack's update direction, allowing the attacker to avoid weak local maxima and therefore producing more transferrable adversarial examples.

## Exp 1: Sensitivity to Imitation Parameter $\alpha$

In this experiment, we further investigate the imitation parameter $\alpha$ in objective function defined in Equation 2. which controls the weight of optimization of the natural examples and the adversarial examples of the proposed method A-LFC. The test attack is 10-step PGD with random start, the size of $\epsilon$ is 0.001, and the step size is $\epsilon/4$. We present the results in Figure 3 concerning the three datasets, i.e., dataset LabelMe, dataset MGC, and dataset Sentiment. By exploiting adversarial learning from crowds, A-LFC excels in many settings of imitation parameter $\alpha$ and delivers excellent stability and robustness. In all of our experiments, we used $\alpha = 0.5$. We did not see the need to search for other better setting of this hyperparameter because it worked well enough.

## Exp 2: White-Box Robustness

We evaluate the robustness of all the five LFC models against four types of adversarial attacks for datasets LabelMe, MGC, and Sentiment: FGSM, PGD (10-step PGD), CW, and MIM. Every adversarial attack is fully able to utilize the model parameters and is subject to the constraints of the same perturbation scale. The LFC models' white-box robustness is shown in Figure 4. First, A-LFC achieves the best robustness against all four types of attacks on dataset LabelMe, MGC, and Sentiment. On average, A-LFC has a 10.05% higher test robustness than the state-of-the-art model, CrowdLayer. Second, we can observe that no matter what strategy is used to instigate an attack, there appears to be a direct relationship between the scale of the perturbation and the amount of damage it can do to the test accuracy, as verified in previous studies (Goodfellow, Shlens, and Szegedy 2015; Dong et al. 2018). In addition, A-LFC outperforms baseline approaches more when the value of perturbation grows. The reason is that A-LFC adopts adversarial training from crowdsourced labeled data to establish that their model is robust even in the face of significant perturbation. Additionally, we examine that the attacks (such as PGD) that have tremendous potential to have better attack success rates than weaker attacks (e.g., FGSM).

## Exp 3: Black-Box Robustness

Black-box attacks are crafted from the natural test examples by attacking a substitute model with an architecture that is a duplicate of MV (using MV to infer the labels and train the DNNs with the hidden layer containing a Fully Connected (FC) layer of 32 units). A variety of attacking techniques have been employed: FGSM, PGD (10-step PGD), CW, and MIM. We report the black-box robustness of all LFC models in Figure 5. Again, the proposed LFC model A-LFC achieves higher robustness than other baselines. A-LFC is found to achieve more robustness than those of the others. On average, A-LFC has 11.34% higher test robustness than the state-of-the-art, CL. We can also observe that

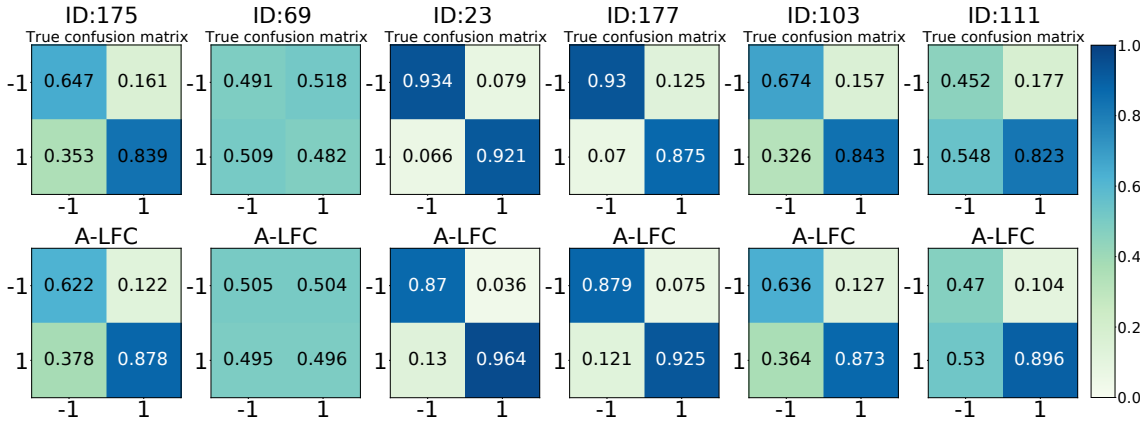



Figure 6: Comparison among true confusion matrices and confusion matrices learned with A-LFC on the real-world dataset Sentiment. The principal diagonal elements of a confusion matrix denote the reliability parameters. The color intensity of the cells increases with the relative magnitude of value.

all LFC models are more robust than those in the white-box attack setting. This hints that the potential of adversarial learning from crowds for being extended to the scenarios of the physic world where the target model conceals from possible attackers. Additionally, CL's performance is slightly larger than that of A-LFC when the perturbation is smaller than 0.02 in some cases. In this case, no matter whether the adversarial examples are included in the learning process, their effect on the model is quite small and even can be ignored. Along with the increase of perturbation scale, A-LFC achieves the best robustness among all the LFC models on the three datasets.

### Exp 4: Performance of Representing Workers

Besides the trained classifier, we also evaluate the learned reliability parameters of each worker which can be represented as a confusion matrix on three datasets LabelMe, MGC, and Sentiment. The training attack is 10-step PGD with step size $\epsilon/4$. For the six normal workers who provide the most labels in the three datasets, Figure 7, Figure 8, and Figure 9 show the comparison among the true confusion matrices, the learned weight matrices of A-LFC, where darker cells correspond to a larger value, while lighter cells correspond to a smaller value. The true confusion matrices are calculated with the workers' labels and ground-truth labels of instances in the datasets. Even we incorporate adversarial examples into the learning process, we can see that the learned confusion matrices are much close to the true confusion matrices, which indicates the good performance of A-LFC for representing the workers.

## Conclusion

In this work, we perform a systematic study on the effect of adversarial examples on LFC models from attack and defense perspectives. On the one hand, we have demonstrated that the LFC-based crowdsourced learning system is vulnerable to adversarial examples. On the other hand, we formulate the problem of LFC in the adversarial environment as a

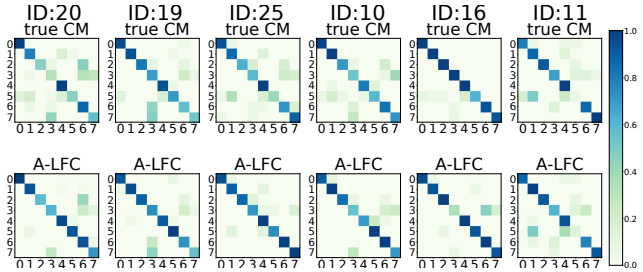



Figure 7: Comparison among true confusion matrices, Confusion Matrices (CM) learned with A-LFC on the real-world dataset LabelMe.

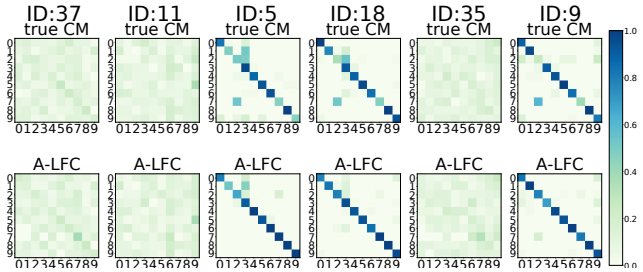



Figure 8: Comparison among the true confusion matrices, confusion matrices learned with A-LFC on the real-world dataset MGC.

bilevel min-max problem and propose a novel LFC framework robust to the adversarial examples. Extensive validation on several real-world benchmark datasets shows that A-LFC is an effective approach to learning from crowdsourced labeled data and substantially outperforms the state-of-the-art in white-box and black-box attack settings. On average, A-LFC has a 10.05% and 11.34% higher test robustness than CrowdLayer in white-box and black-box attack settings, respectively. In future work, we plan to investigate approaches to defending against other types of adversarial attacks such as data poisoning.

## Acknowledgments

This work was supported partly by National Key Research and Development Program of China under Grant No.2019YFB1705902, partly by National Natural Science Foundation under Grant Nos.(61972013, 61932007). Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

## References

Albarqouni, S.; and Baur, C. 2016. AggNet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging*, 35(5): 1313–1321.

Cao, P.; Xu, Y.; Kong, Y.; and Wang, Y. 2019. Max-MIG: an Information Theoretic Approach for Joint Learning from Crowds. In *Proceedings of 7th International Conference on Learning Representations*. New Orleans, LA, USA: OpenReview.net.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.

Chen, P.; Sun, H.; Fang, Y.; and Huai, J. 2018. Collusion-Proof Result Inference in Crowdsourcing. *Journal of Computer Science and Technology*, 33(2): 351–365.

Chen, P.; Sun, H.; Fang, Y.; and Liu, X. 2020a. CONAN: A framework for detecting and handling collusion in crowdsourcing. *Information Sciences*, 515: 44–63.

Chen, Z.; Wang, H.; Sun, H.; Chen, P.; Han, T.; Liu, X.; and Yang, J. 2020b. Structured Probabilistic End-to-End Learning from Crowds. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1512–1518. Online: AAAI Press.

Chu, Z.; Ma, J.; and Wang, H. 2021. Learning from Crowds by Modeling Common Confusions. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 5832–5840. AAAI Press.

Das, N.; Shanbhogue, M.; Chen, S.-T.; Hohman, F.; Li, S.; Chen, L.; Kounavis, M. E.; and Chau, D. H. 2018. Compression to the rescue: Defending from Adversarial Attacks across Modalities. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied statistics*, 20–28.

Demartini, G.; Difallah, D. E.; Cudr; and Mauroux, P. 2012. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *Proceedings of International Conference on World Wide Web*, 469–478. New York, USA: ACM.

Dong, Y.; Fu, Q.; Yang, X.; Pang, T.; Su, H.; Xiao, Z.; and Zhu, J. 2020. Benchmarking Adversarial Robustness on Image Classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 318–328. Seattle, USA: Computer Vision Foundation.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9185–9193. Salt Lake City, USA: Computer Vision Foundation.

Fang, M.; Sun, M.; Li, Q.; Gong, N. Z.; Tian, J.; and Liu, J. 2021. Data Poisoning Attacks and Defenses to Crowdsourcing Systems. In Leskovec, J.; Grobelnik, M.; Najork, M.; Tang, J.; and Zia, L., eds., *Proceedings of the Web Conference 2021*, 969–980.

Fang, Y.; Sun, H.; Chen, P.; and Huai, J. 2018. On the Cost Complexity of Crowdsourcing. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 1531–1537. Stockholm, Sweden: AAAI Press.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *Proceedings of 3rd International Conference on Learning Representations, ICLR 2015*. San Diego, USA: OpenReview.net.

Han, T.; Sun, H.; Song, Y.; Fang, Y.; and Liu, X. 2016. Incorporating External Knowledge into Crowd Intelligence for More Specific Knowledge Acquisition. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1541–1547. Menlo Park, California: AAAI Press.

Li, H.; and Yu, B. 2014. Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing. *arXiv preprint arXiv:1411.4086*.

Li, Q.; Li, Y.; Gao, J.; Su, L.; Zhao, B.; Demirbas, M.; Fan, W.; and Han, J. 2014. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *Proceedings of the VLDB Endowment*, 8(4): 425–436.

Luo, Y.; Tian, T.; Shi, J.; Zhu, J.; and Zhang, B. 2018. Semi-Crowdsourced Clustering with Deep Generative Models. In *Proceedings of Annual Conference on Neural Information Processing Systems*, 3216–3226. Montréal, Canada: Curran Associates, Inc.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of 6th International Conference on Learning Representations*. Vancouver, Canada: OpenReview.net.

Miao, C.; Li, Q.; Su, L.; Huai, M.; Jiang, W.; and Gao, J. 2018a. Attack under Disguise: An Intelligent Data Poisoning Attack Mechanism in Crowdsourcing. In *Proceedings of the 2018 World Wide Web Conference*, 13–22. Lyon, France: ACM.

Miao, C.; Li, Q.; Xiao, H.; Jiang, W.; Huai, M.; and Su, L. 2018b. Towards Data Poisoning Attacks in Crowd Sensing Systems. In *Proceedings of the Nineteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 111–120. Los Angeles, CA, USA: ACM.

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

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013. Learning from Multiple Annotators: Distinguishing Good from Random Labelers. *Pattern Recognition Letters*, 34(12): 1428–1436.

Rodrigues, F.; and Pereira, F. C. 2018. Deep Learning from Crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 1611–1618. New Orleans, Louisiana: AAAI Press.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 614–622. New York, USA: ACM.

Tong, Y.; Zhou, Z.; Zeng, Y.; Chen, L.; and Shahabi, C. 2020. Spatial crowdsourcing: a survey. *VLDB J.*, 29(1): 217–250.

Wang, L.; and Zhou, Z. 2016. Cost-Saving Effect of Crowdsourcing Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2111–2117. Menlo Park, California: AAAI Press.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *Proceedings of 8th International Conference on Learning Representations*. Addis, Ethiopia: OpenReview.net.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proceedings of Annual Conference on Neural Information Processing Systems*, 2035–2043. Vancouver, Canada: Curran Associates, Inc.

Yang, J.; Drake, T.; Damianou, A. C.; and Maarek, Y. 2018a. Leveraging Crowdsourcing Data for Deep Active Learning an Application: Learning Intents in Alexa. In *Proceedings of the 2018 World Wide Web Conference*, 23–32. Lyon, France: ACM.

Yang, Y.; Zhang, M.; Chen, W.; Zhang, W.; Wang, H.; and Zhang, M. 2018b. Adversarial learning for chinese ner from crowd annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, J.; Wu, M.; and Sheng, V. S. 2019. Ensemble Learning from Crowds. *TKDE.*, 31(8): 1506–1519.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth Inference in Crowdsourcing: Is the Problem Solved? *Proceedings of the VLDB Endowment*, 10(5): 541–552.

Zhong, J.; Yang, P.; and Tang, K. 2017. A quality-sensitive method for learning from crowds. *IEEE Transactions on Knowledge and Data Engineering*, 29(12): 2643–2654.