Misclassification-Aware Robust Learning from Multiple Human Labelers (Student Abstract)

Zuoyuehe Wang[1*], Chicheng Ma[2*], Pengpeng Chen[3 *], Lei Chai[3†], Yongqiang Yang[3†], Zhijun Chen[3], Jingzheng Li[4], Bing Li[5], Ting Wang[6]

[1]Nanyang Technology University
[2]University of Science and Technology Beijing
[3]Beihang University
[4]Zhongguancun Laboratory
[5]IHPC and CFAR, Agency for Science, Technology and Research (A*STAR)
[6]Xuanhang Co., Ltd.
wzuoyuehe@163.com, mcc199912@163.com, chenpp@buaa.edu.cn, chailei@buaa.edu.cn, yangyongqiang@buaa.edu.cn, zhijunchen@buaa.edu.cn, maxlijingzheng@163.com, libingsy@buaa.edu.cn, wang20250706@163.com

Abstract

Adversarial training is an effective technique for enhancing the robustness of deep neural networks (DNNs). Prior research shows that misclassified examples influence final adversarial robustness much more than correctly classified examples. Ignoring this difference during training can hurt model performance. In crowdsourcing, varying annotator expertise causes noisy, inconsistent labels. As a result, it is hard to distinguish misclassified and correctly classified examples using only provided annotations. Thus, how to use the reliability and discrepancy between these example types to improve robustness within adversarial learning remains a critical but underexplored issue. In this work, we first explore how misclassified and correctly classified examples affect learning from crowds (LFC) in adversarial environments. Then, we formulate the problem of misclassification-aware robust learning from multiple human labelers as a bilevel min-max problem. After that, we introduce MALC, a new approach to make classifiers more robust to adversarial examples via iterative adversarial example generation and parameter estimation. We conduct an extensive evaluation of the proposed MALC, showing that MALC can outperform the state-of-the-art LFC methods in both white-box and black-box settings.

Introduction

Human-labeled data played a crucial role in the supervised learning framework and should still be given heightened importance in the Large Language Model (LLM) era. Crowdsourcing offers a scalable annotation solution but introduces label noise due to annotator heterogeneity (Fang et al. 2018; Chai, Sun, and Wang 2022). While existing Learning from Crowds (LFC) methods focus on inferring true labels from noisy annotations, they overlook adversarial vulnerabilities. Recent studies show that standard classifiers fail against adversarial examples, and LFC models are further weakened by label noise (Chen et al. 2022; Chai et al. 2024). Although Adversarial learning from crowds (A-LFC) pioneers adversarial training in LFC by enhancing robustness, it treats all examples equally. This is suboptimal because misclassified examples significantly impact robustness more than correctly classified ones (Wang et al. 2020). In crowdsourcing, label noise makes distinguishing these examples inherently challenging, limiting further robustness gains (Chen et al. 2023) [1]. To address this, we propose **M**isclassification-**A**ware **L**earning from **C**rowds (**MALC**)—the first framework to explicitly model the discrepancy between misclassified and correctly classified examples during adversarial training. We formulate this as a bilevel min-max problem: Inner maximization: Generates adversarial examples. Outer minimization: Optimizes model parameters using an EM algorithm, while projecting gradients to handle non-convexity. Experiments on MGC and Labelme confirm that MALC surpasses state-of-the-art methods (e.g., **+6.43%** robustness in white-box settings). Contributions are as follows: (1) To the best of our knowledge, this is the first work that incorporates the distinction between misclassified and correctly classified examples to enhance the robustness of LFC models; (2) We investigate the distinctive influence of misclassified and correctly classified examples on LFC and formulate this problem as Misclassification-Aware robust learning from crowds; (3) We propose a novel approach, MALC for further improving classifiers robust to adversarial examples; (4) We conducted an extensive evaluation, showing that MALC outperforms the state-of-the-art in both white-box and black-box settings [2].

Method

We formulate the **MALC** problem as a bilevel min-max problem. We begin by introducing the generation of adver-

*Equal Contribution.
†Corresponding Author: Lei Chai, Yongqiang Yang

[1]We have verified this distinction in adversarial learning from crowds. Please refer to supplementary materials.
[2]Our code is available in supplementary materials

sarial examples concerning the max problem as follows:

$$\mathbf{x}_i' = \underset{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon}{\operatorname{argmax}} L\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i'\right), z_i\right), \quad (1)$$

where $L(\cdot)$ denotes the cross-entropy loss. $\mathbf{x}_i'$ denotes an adversarial example generated by solving the inner maximization problem.

$(\mathcal{X}^+)$ indicates instances successfully categorized and the other can be considered misclassified $(\mathcal{X}^-)$ in relation to the current network.

$$\begin{aligned} \mathcal{X}^+ &= \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{X}, f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \hat{z}_i\} \\ \mathcal{X}^- &= \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{X}, f_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq \hat{z}_i\}, \end{aligned} \quad (2)$$

where $\hat{z}_i$ is the label estimated from crowd labels.

Our adversarial risk is defined for misclassified instances $\mathcal{X}^-$ as:

$$L\left(f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right), \hat{z}_i\right) + L\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right), f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right)\right). \quad (3)$$

For correctly classified example, we have $L\left(f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right), \hat{z}_i\right) = L\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right), f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right)\right)$. Combining the proposed two adversarial risks in an adversarial training framework, we can train a network that minimizes the following risk:

$$\begin{aligned} &\min_{\Theta} L\left(f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right), \hat{z}_i\right) + \lambda L\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right), f_{\boldsymbol{\theta}}\left(\hat{\mathbf{x}}_i'\right)\right) \cdot \mu_i \\ &s.t.\ \mathbf{x}_i' = \underset{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon}{\operatorname{argmax}} L\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right), f_{\boldsymbol{\theta}}\left(\mathbf{x}_i'\right)\right), \end{aligned} \quad (4)$$

where $\mu_i = p\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right) \neq \hat{z}_i\right)$ which denotes the probability that the instance is misclassified. $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(M)}\}$ and $\boldsymbol{\Pi}^{(j)}$ denotes the confusion matrix of worker j. λ is the imitation parameter for balancing the two parts.

MALC Algorithm

We present MALC, a crowdsourcing mechanism for learning robust models. Two main stages make up MALC, to find the best solution to the bilevel min-max problem, these two stages are repeated iteratively.

Stage 1. With the estimate Θ, namely the parameters of the classifier and the workers, MALC applies the PGD algorithm to generate the adversarial examples $\mathcal{X}'$ with the inferred ground truth in Equation 1.

Stage 2. MALC applies the EM algorithm to solve the external subproblem. MALC first estimates the ground truth based on the generated adversarial examples $\mathcal{X}'$ in E-step, and simultaneously distinguishes the misclassified examples $\mathcal{X}^-$ and correctly classified examples $\mathcal{X}^+$ within $\mathcal{X}'$. After that, the model parameters Θ and the parameters of the workers $\{\boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(M)}\}$ are updated via backpropagation with Equation 4, using the distinguished subsets $\mathcal{X}^+$ and $\mathcal{X}^-$ in M-step. Details are introduced in the Appendix.

Experiment

We conducted comprehensive evaluations of the LFC models' robustness against four types of adversarial attacks—FGSM, PGD (a 10-step variant), CW, and MIM—on two real-world datasets: MGC and LabelMe, comparing against six baseline models.

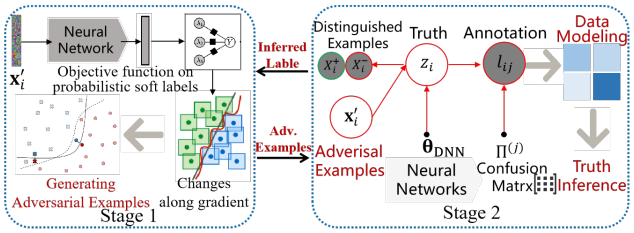

Figure 1: MALC Framework Empowered by Probabilistic Graphical Model

Method	MGC				
	FGSM	PGD	CW	MIM	Avg.
MV†	43.33	38.00	1.93	40.00	30.82
AggNet	40,80	40.87	1.73	42.93	31.58
CL (VW)§	46.40	45.20	11.20	45.87	37.17
CL (VW-B)§	46.47	44.80	12.80	45.40	37.37
CL (MW)§	32.47	31.53	13.80	31.80	27.40
A-LFC	54.13	53.60	19.83	53.53	45.27
MALC	**57.27**	**57.00**	**21.33**	**57.13**	**48.18**

[1] †: "MV" denotes the method of label aggregation. §: "MW", "VW" and "VW-B" refer to three different ways of parameterizing the annotator reliability of CL.

Table 1: White-box robustness on MGC dataset .

The LFC models' white-box robustness is shown in Table 1. First, MALC achieves the best robustness against all four types of attacks on MGC. On average, MALC has a 6.43% higher test robustness than the state-of-the-art model, A-LFC. When compared to other baselines, the suggested LFC model, MALC is more robust. When compared to competing methods, MALC proves to be more resilient. This suggests that the target model may be able to use adversarial learning from crowds in physical world scenarios where it is trying to hide from potential attackers.

The comprehensive experimental results (including the complete White-box and Black-box robustness evaluation on Labelme and MGC, sensitivity to the Imitation Parameter λ, and performance of representing workers) and the experiment settings are presented in the Appendix.

Conclusion

We propose MALC, a novel defense algorithm that explicitly differentiates between misclassified and correctly classified examples during adversarial training. MALC formulates misclassification-aware learning under adversarial attacks as a bilevel min-max optimization problem, solved via: EM algorithm for the outer minimization and Projected Gradient Descent (PGD) for the inner maximization. Experiments on real-world benchmarks demonstrate MALC's superiority, improving test robustness by 6.43% (white-box). Future work will extend MALC to defend against data poisoning attacks.

References

Chai, L.; Qi, L.; Sun, H.; and Li, J. 2024. RA 3: A Human-in-the-loop Framework for Interpreting and Improving Image Captioning with Relation-Aware Attribution Analysis. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 330–341. IEEE.

Chai, L.; Sun, H.; and Wang, Z. 2022. An error consistency based approach to answer aggregation in open-ended crowdsourcing. *Information Sciences*, 608: 1029–1044.

Chen, P.; Sun, H.; Yang, Y.; and Chen, Z. 2022. Adversarial learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5304–5312.

Chen, Z.; Sun, H.; He, H.; and Chen, P. 2023. Learning from noisy crowd labels with logics. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 41–52. IEEE.

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.

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.