Machine Learning Robustness, Fairness, and their Convergence

Jae-Gil Lee[1], Yuji Roh[1], Hwanjun Song[2], Steven Euijong Whang[1*]
[1]KAIST, [2]NAVER AI Lab
Republic of Korea
{jaegil,yuji.roh,swhang}@kaist.ac.kr,hwanjun.song@navercorp.com

ABSTRACT

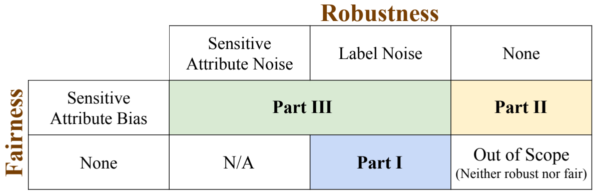
Responsible AI becomes critical where *robustness* and *fairness* must be satisfied together. Traditionally, the two topics have been studied by different communities for different applications. *Robust training* is designed for noisy or poisoned data where image data is typically considered. In comparison, *fair training* primarily deals with biased data where structured data is typically considered. Nevertheless, robust training and fair training are fundamentally similar in considering that both of them aim at fixing the inherent flaws of real-world data. In this tutorial, we first cover state-of-the-art robust training techniques where most of the research is on combating various label noises. In particular, we cover label noise modeling, robust training approaches, and real-world noisy data sets. Then, proceeding to the related fairness literature, we discuss pre-processing, in-processing, and post-processing unfairness mitigation techniques, depending on whether the mitigation occurs before, during, or after the model training. Finally, we cover the recent trend emerged to combine robust and fair training in two flavors: the former is to make the fair training more robust (i.e., robust fair training), and the latter is to consider robustness and fairness as two equals to incorporate them into a holistic framework. This tutorial is indeed timely and novel because the convergence of the two topics is increasingly common, but yet to be addressed in tutorials. The tutors have extensive experience publishing papers in top-tier machine learning and data mining venues and developing machine learning platforms.

Website: https://kdd21tutorial-robust-fair-learning.github.io

ACM Reference Format:
Jae-Gil Lee, Yuji Roh, Hwanjun Song, Steven Euijong Whang. 2021. Machine Learning Robustness, Fairness, and their Convergence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470799

1 TARGET AUDIENCE AND BACKGROUND

We target machine learning practitioners who are interested in Responsible AI issues where model training must be robust against data noise and fair against data bias. It is helpful to have working knowledge in machine learning and data mining.



		Robustness		
		Sensitive Attribute Noise	Label Noise	None
Fairness	Sensitive Attribute Bias	**Part III**		**Part II**
	None	N/A	**Part I**	Out of Scope (Neither robust nor fair)

Figure 1: Landscape of robust and fair training research and the focus on this tutorial.

2 TUTORS

Jae-Gil Lee is an associate professor at the School of Computing, KAIST. Before joining KAIST in 2010, he worked at the IBM Almaden Research Center and the University of Illinois Urbana-Champaign. His research interests encompass spatio-temporal data mining and scalable machine learning, and he is recently working on the data quality issues for deep learning. He is a senior program committee member of KDD 2021 and has served as an associate editor of IEEE TKDE since 2019.

Yuji Roh is a Ph.D. student at the School of Electrical Engineering, KAIST. Her research interests are responsible/trustworthy AI, human-centered AI, and big data - AI integration. She won the Qualcomm Innovation Fellowship Korea in 2020. She received her B.S. degree in Electrical Engineering from KAIST in 2018.

Hwanjun Song is a research scientist of NAVER AI Lab. He is particularly interested in designing advanced approaches to handle large-scale and noisy data, which are two main real-world challenges for the practical use of AI approaches. He worked as a research intern at Google Research in 2020. He earned his Ph.D. in knowledge service engineering from KAIST in 2021.

Steven Euijong Whang is an associate professor at the School of Electrical Engineering and Graduate School of AI, KAIST. His research interests are responsible AI and big data - AI integration. Previously he was a Research Scientist at Google Research and co-developed the data infrastructure of the TensorFlow Extended (TFX) end-to-end machine learning platform. He received his Ph.D. in computer science in 2012 from Stanford University. He is a recipient of the Google AI Focused Research Award in 2018, the first in Asia.

3 TUTORIAL OUTLINE

Traditionally, robust training and fair training have been studied by separate communities. Figure 1 shows the research landscape for the two topics and the scope of this tutorial. Robust training (Section 4) has mostly focused on combating label noise without regarding bias. On the other hand, fair training (Section 5) has focused on handling bias, but not necessarily noise. More recently, we are observing a convergence of robust and fair training techniques (Section 6) for (1) handling noisy or missing values in sensitive attributes (e.g., gender and race) and (2) combating label poisoning.

[*] Corresponding Tutor

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470799

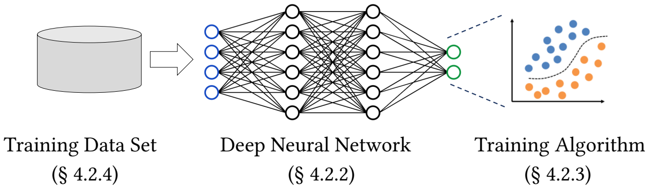


Figure 2: Categorization of algorithms robust to label noises.

4 PART I: ROBUSTNESS TO "LABEL NOISE"

4.1 Motivation and Issues

When the labels in a training data set are corrupted for several reasons, the generalizability of a resulting deep neural network (DNN) can be severely damaged, because a huge number of model parameters of a DNN make it *overfit* to the training data set even with corrupted labels. Zhang et al. [6] have shown that a DNN can easily fit to an entire training data set with any ratio of corrupted labels, which eventually induces poor performance for a test data set. Therefore, we formulate that the goal of this field is to achieve a good generalization capability very close to the ideal performance that would be obtained without any label noises.

4.2 Robust Training Approaches

4.2.1 Overview and Classification. A deep learning framework is typically involved with three components: (i) a deep neural network (DNN), (ii) a training algorithm, and (iii) a training data set. Thus, it is natural to improve one of these components to realize robust training, as in Figure 2.

4.2.2 Robust Architecture. Robustness can be achieved by inserting a *noise adaptation layer* at the top of an underlying DNN to learn label corruption process or developing a *dedicated architecture* to reliably support more diverse types of label noises.

4.2.3 Loss Adjustment. Robustness can be achieved by adjusting the loss, according to the confidence of a given label, in the middle of updating parameters. The relevant studies are further categorized into those correcting the softmax output; those reweighting the loss value; and those correcting the target label [3].

4.2.4 Sample Selection. Intuitively, robustness can be achieved by selecting only the samples with correct labels. Most studies resort to the *small-loss* trick that regards the samples with a small loss value as clean (i.e., having correct labels). Recently, MORPH [4] is proposed to be free from the small-loss trick.

5 PART II: FAIRNESS AGAINST "DATA BIAS"

5.1 Motivation and Issues

Data can easily be biased where there is too much data for certain demographics, but not for other demographics. The performance of a model against bias is measured using various fairness measures [5]. Addressing data bias can be categorized into *pre-processing*, *in-processing*, or *post-processing* approaches where the bias is mitigated before, during, or after model training, respectively, as in Figure 3.

5.2 Fair Training Approaches

5.2.1 Preparing Unbiased Data. Before model training, one can acquire, repair, reweight, or generate data to reduce bias.

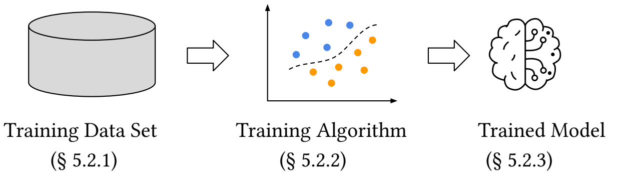


Figure 3: Categorization of algorithms fair against bias.

5.2.2 Training on Biased Data. In the presence of bias, model training can also cope with it by balancing fairness with accuracy. We introduce several representative techniques including a fairness constraints approach that adds a fairness penalty term in the loss function; and an adversarial learning-based approach that reduces the correlation between sensitive attributes and the model output.

5.2.3 Debiasing a Trained Model. After model training, one may still want to mitigate the model's bias. However, it may be impossible to go back to the data and remove its bias or re-train the model using a different algorithm. Hence, the only solution is to adjust the trained model itself to be fairer [1].

6 PART III: CONVERGENCE OF ROBUSTNESS AND FAIRNESS

Currently, there are two research directions for combining robustness and fairness: to make fairness more robust and to integrate robust and fair training in equal terms, as shown in Figure 4. Fairness techniques need to be more robust because the sensitive attributes may be noisy, poisoned, or even missing. An interesting observation is that fairness techniques usually assume structured data, so the robustness techniques, which are usually designed for image data, now have to be redesigned for structured data. The other direction of robust and fair training is to obtain both of robustness and fairness objectives at the same time [2].

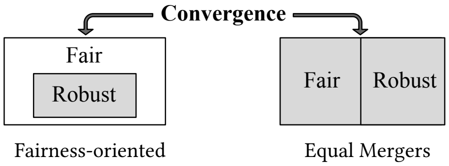


Figure 4: Categorization of algorithms with both robustness and fairness.

ACKNOWLEDGEMENTS

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00862). This work was partly supported by a Google AI Focused Research Award.

REFERENCES

[1] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*. 3315–3323.

[2] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2020. FR-Train: A Mutual Information-Based Approach to Fair and Robust Training. In *ICML*. 8147–8157.

[3] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. 2019. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. In *ICML*. 5907–5915.

[4] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. 2021. Robust Learning by Self-Transition for Handling Noisy Labels. In *KDD*.

[5] Suresh Venkatasubramanian. 2019. Algorithmic Fairness: Measures, Methods and Representations. In *PODS*. 481.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding Deep Learning Requires Rethinking Generalization. In *ICLR*.