Towards Efficient and Reliable Data Curation for Machine Learning

Naiqing Guan

Department of Computer Science
University of Toronto
naiqing.guan@mail.utoronto.ca

1 Introduction

Modern machine learning models require large training datasets to achieve good accuracy, yet manual labelling and curation of large datasets are both expensive and time-consuming. Thus, acquiring labelled datasets has become one of the main bottlenecks in applying machine learning in practical scenarios. This has motivated researchers to investigate approaches to reduce annotation costs and instigated corporate activity on labelling services.

The programmatic weak supervision (PWS) framework [Ratner et al.(2016), Ratner et al.(2017), Zhang et al.(2022)] provides an approach to automatically label large datasets without manually annotating specific instances. In the PWS framework, users represent weak supervision sources in the form of label functions (LFs), which are programs that provide noisy labels to a subset of data. Since the label functions have varying accuracy and may exhibit ad-hoc correlations, a label model is designed to aggregate noisy, weak labels into probabilistic labels. The aggregated labels are then used to train the downstream model.

While the PWS framework reduces annotation costs, it still has some limitations. First, the design of LFs requires substantial endeavours and domain expertise, while automatic LF design is still challenging. Secondly, the labels generated by the PWS framework are usually noisy, which deteriorates the performance of downstream models. How to evaluate, control and improve the quality of LFs and generated labels requires further investigation.

My research aims to improve the efficiency and reliability of data curation for machine learning, with a focus on the PWS framework. In this presentation, I will discuss two of my recent works in this direction, focusing on automatic LF design and improving label quality, respectively.

2 Presentation Outline

In the first part of the presentation, I will describe the background for efficient data curation and introduce the PWS framework.

In the second part, I will describe two of my recent works in enhancing the efficiency and reliability of the PWS framework. I will first introduce DataSculpt [Guan et al.(2023)], which automatically designs LFs by prompting large language models. We explored an expansive design landscape in DataSculpt and identified the strengths and limitations of contemporary LLMs in LF design. Then I will briefly describe ActiveDP, which combines PWS with active learning [Settles(2012)] to combine the strengths of both paradigms and improve the label quality.

In the third part, I will describe the limitations of the current PWS framework and propose some future research directions in this area.

References

[Guan et al.(2023)] Naiqing Guan, Kaiwen Chen, and Nick Koudas. 2023. Can Large Language Models Design Accurate Label Functions? arXiv:2311.00739 [cs.CL]

[Ratner et al.(2017)] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.

[Ratner et al.(2016)] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems* 29 (2016), 3567–3575.

[Settles(2012)] Burr Settles. 2012. Active learning. *Synthesis lectures on artificial intelligence and machine learning* 6, 1 (2012), 1–114.

[Zhang et al.(2022)] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433* (2022).