Personal Statement

Hwanjun Song, Research Scientist @ NAVER AI Lab [Google Scholar]
ghkswns91@gmail.com https://songhwanjun.github.io/ +82-10-5308-2512

Introduction

My research mainly focuses on practical machine learning (ML) techniques, which can perform well under diverse real-world scenarios. To this end, my research topics cover a wide range of research fields, such as database, data mining, machine learning, and computer vision. More specifically, three research topics have been addressed so far: (1) robust deep learning for handling data with close- and open-set noise, (2) large-scale learning & processing to achieve high accuracy and efficiency simultaneously, and (3) applied data science to apply modern ML techniques to numerous real-world applications.

Topic 1. Robust Deep Learning (2019 – Current)

In supervised learning, the success of DNNs is conditioned on the availability of massive data with carefully annotated labels. However, data labels are often corrupted because they can be complex even for an experienced person, i.e., close-set noise. Modern DNNs are trained in an over-parameterized regime, and such DNNs easily overfit to any ratio of noisy labels, eventually resulting in poor generalization. In addition, many examples in different domains irrelevant to a target domain are inevitably included from less controlled data collection processes; they are called either open-set noise or out-of-distribution examples. Learning DNNs in the presence of open-set or close-set noise is much more practical in real-world scenarios. To cope with these challenges, I presented two robust learning methods for label noise (ICML19, KDD21) and wrote a highly influential Survey Paper with a Tutorial at KDD21.

Topic 2. Large-scale Learning & Processing (2017 – Current)

The extremely high computational cost for large-scale data and models is another real-world challenge for ML. To meet the requirement of practical use, it is necessary to balance the trade-off between accuracy and efficiency. In this aspect, I have contributed to multiple parts of modern ML and distributed processing pipelines: (i) an efficient and effective fully transformer-based object detector (ICLR22), which achieves the state-of-the-art performance trade-off between AP and FPS, (ii) two adaptive batch sampling approaches (MLJ20, CIKM20), which accelerate the training convergence of DNNs, and (iii) two distributed clustering algorithms (KDD17, SIGMOD18), which save the processing time by up to 100 times. In particular, I have strong experience in multi-node distributed training with PyTorch and processing with Hadoop and Spark.

Topic 3. Applied Data Science (2020 – Current)

Beyond robustness and large-scale learning, I have experience with multiple applied science problems including user modeling/recommender system (PAKDD20, AAAI21, AAAI22), representation learning (WWW20), and COVID-19 related prediction (KDD20, AAAI22).

In addition to the listed topics, I have studied robust federated learning, continual learning, and active learning (all of them are submitted to either CVPR22 or ICML22); and multimodal learning for object detection (BMVC21), which was done during a research internship at Google Research. Supported by these diverse experiences and achievements, I have a very fast learning curve and a unique thinking model that will be a great advantage to explore new topics for the advanced ML.

Areas to Work

I believe that ability to integrate multiple disciplines is necessarily needed towards practical approaches, considering that (1) data size is increasing explosively and (2) data noise is unavoidable. In that sense, as supported by my previous work, my wide research background would be of great help to boost the usability of deep learning for real-world scenarios, which many industrial companies and academic institutes pursue.