Robust and Adaptive Deep Learning via Bayesian Principles

Yingzhen Li

Department of Computing, Imperial College London
South Kensington Campus, London, SW7 2AZ, UK
yingzhen.li@imperial.ac.uk

Talk Summary

Deep learning models have achieved tremendous successes in accurate predictions for computer vision, natural language processing and speech recognition applications. However, to succeed in high-risk and safety-critical domains such as healthcare and finance, these deep learning models need to be made reliable and trustworthy. Specifically, they need to be robust and adaptive to real-world environments which can be drastically different from the training settings. In this talk, I will advocate for Bayesian principles to achieve the goal of building robust and adaptive deep learning models. I will introduce a suite of uncertainty quantification methods for Bayesian deep learning, and demonstrate applications enabled by accurate uncertainty estimates, e.g., robust prediction, continual learning and repairing model failures. I will conclude by discussing the research challenges and potential impact for robust and adaptive deep learning models.

Foundations of Bayesian Deep Learning

Bayesian neural networks (BNNs) perform posterior inference to obtain model uncertainty given data. This posterior inference procedure is computationally intractable, thus approximations are obtained by e.g., variational inference (VI) that minimises the KL-divergence from the approximation to the exact posterior. One major issue of VI is the overconfidence of obtained uncertainty estimates. To address this, Li and Turner (2016) introduced Rényi's α-divergence minimisation for posterior approximations, where the users preference for optimistic/conservative uncertainty estimates can be controlled by the parameter. With better control of uncertainty estimate via this framework, we can improve BNN's robustness and detection abilities against adversarial attacks (Li and Gal 2017).

Continual Learning for Adaptations

In continual learning, data continuously arrive, with tasks changing over time, and entirely new tasks can emerge. Continual learning systems must adapt to perform well on the entire set of tasks seen in history without revisiting all previous data. In Nguyen et al. (2018), we proposed a generic Bayesian framework for continual learning. We use the posterior distribution to enable transfer of knowledge learned in history for learning the current task, and to provide strong regularisation effect in preventing drastic model changes (thus preventing catastrophic forgetting). For practical implementations, we developed a VI-based framework that combines online VI and Monte Carlo estimation techniques for continual learning. Experiments on both prediction tasks and image generation tasks demonstrated the success of the proposed framework by avoiding catastrophic forgetting in a fully automatic way.

Repairing Model Failures

Machine learning models often exhibit unexpected failures once deployed in real-world scenarios. As one cannot anticipate beforehand all plausible failure scenarios, on-demand repair of models is needed. In Tanno et al. (2022) we tackled this challenge in the setting where data deficiency is mainly responsible for model failures, and the goal is to identify detrimental training data and remove their contributions to the trained model without re-training from scratch. We achieve this goal by introducing (Bayesian) continual learning to simultaneously address both identification and removal of detrimental data. Our framework subsumes existing work on influence function and data deletion as specific examples, and we demonstrated its benefits by extending Elastic Weight Consolidation, a continual learning algorithm, to achieve better repair of the model failures.

References

Li, Y.; and Gal, Y. 2017. Dropout Inference in Bayesian Neural Networks with Alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2052–2061. PMLR.

Li, Y.; and Turner, R. E. 2016. Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems*.

Nguyen, C. V.; Li, Y.; Bui, T. D.; and Turner, R. E. 2018. Variational Continual Learning. In *International Conference on Learning Representations (ICLR)*.

Tanno, R.; Pradier, M. F.; Nori, A.; and Li, Y. 2022. Repairing Neural Networks by Leaving the Right Past Behind. In *Advances in Neural Information Processing Systems*.