Towards Reliable Machine Learning: Evaluating and Robustifying Deep Neural Networks

Saurabh Garg

In recent years, deep learning has gained prominence, buoyed by claims of super-human performance on a diverse range of tasks, including image classification and natural language understanding. However, one of the main challenges in the successful deployment of many deep learning services is the need to deal with distribution shifts. In most real-world settings, the target distribution (from which deployment-time data arrives) can deviate markedly from the source distribution (from which we sampled our training data). Reasons for the discrepancy can include gradual changes in human behavior or differences in the demographics of the environment where the service is being used. Since real-world deployment seldom respects the i.i.d. assumption, we are left with distribution shift problems, whereby the reported and actual accuracy of deployed models can diverge.

Moreover, deep learning models lack the ability to flag uncertain decisions. While the performance of a model can be validated using held-out source data before deployment, its accuracy can deteriorate silently in presence of a distribution shift after deployment. This silent failure exacerbates the brittleness issues and limits the reliability of machine learning in practice. As a result, reliable estimates of models' performance become crucial for stakeholders to trust predictions and in order to avoid silent failure.

The primary focus of my research is to robustify and evaluate deep learning models in the face of distribution shifts. While, in principle, one may obtain fresh labeled data from the target distribution to re-train or evaluate the model every time the distribution shifts, this can be prohibitively expensive. To this end, I leverage (comparatively abundant) unlabeled data from the target domain to identify structural relationships between the target and source domains, and then use them to adapt and evaluate models. My research involves understanding the behavior of deep models, both theoretically and empirically, and using those insights to develop robust methods. In particular, my research focuses on the following questions:

Q1: How to adapt ML models in the face of distribution shifts? Absent assumptions on the nature of the distribution shifts, this task is impossible. My research in this direction is focused on formulating assumptions on distribution shift scenarios appearning in the wild and developing procedures that improve and adapt ML models under those shifts by leveraging unlabeled data [1; 4; 3; 11; 12; 5; 7].

Q2: How can we evaluating models' performance without access to labeled data? Deep learning models fail silently, i.e., they cannot flag uncertain decisions. To build reliable ML systems, obtaining certificates for accuracy is as important as robustifying these systems. My goal in this direction is to develop techniques that can leverage unlabeled data to predict model accuracy [2; 6].

Q3: How can we characterize datasets from the lens of deep models? Characterizing the properties of training data can help us: (i) derive insights for how to reconcile the ability of deep neural networks to generalize with their ability to memorize noise; (ii) identifying potentially mislabeled examples; and (iii) identifying notably challenging or rare sub-populations of examples. To this end, some of my past work focused on characterizing learning and unlearning dynamics of data points [10; 8; 9].

References

[1] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Saurabh Garg, Sivaraman Balakrishnan, Zico Kolter, and Zachary Lipton. RATT: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning (ICML)*, 2021.

[3] Saurabh Garg, Yifan Wu, Alex Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture proportion estimation and PU learning: A modern approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J Zico Kolter, Sivaraman Balakrishnan, Zachary C Lipton, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning (ICML)*, 2021.

[5] Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Saurabh Garg, Sivaraman Balakrishnan, Zachary Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

[7] Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Siva Balakrishnan, and Zachary Lipton. Rlsbench: A large-scale empirical study of domain adaptation under relaxed label shift. *Under submission*, 2022.

[8] Gal Kaplun, Nikhil Ghosh, Saurabh Garg, Boaz Barak, and Preetum Nakkiran. Deconstructing distributions: A pointwise framework of learning. *arXiv preprint arXiv:2202.09931*, 2022.

[9] Kundan Krishna, Saurabh Garg, Jeffrey Bigham, and Zachary Lipton. Downstream datasets make surprisingly good pretraining corpora. *arXiv preprint arXiv:2209.14389*, 2022.

[10] Pratyush Maini, Saurabh Garg, Zachary Lipton, and Zico Kolter. Characterizing datapoints via second-split forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Zachary Novack, Saurabh Garg, and Zachary Lipton. Chils: Zero-shot image classification with hierarchical label sets. *Under submission*, 2022.

[12] Manley Roberts, Pranav Mani, Saurabh Garg, and Zachary Lipton. Unsupervised learning under latent label shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.