Towards Reliable Machine Learning: Evaluating and Robustifying Deep Neural Networks

Saurabh Garg | Machine Learning Department, Carnegie Mellon University

In recent years, deep learning has gained prominence, buoyed by claims of super-human performance on a diverse range of tasks, including image classification [6], and natural language understanding [7], and cancer detection [1]. 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, whether medical diagnosis, satellite image classification, or the scene understanding models powering self-driving cars, 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, climate changes or shifts in the prevalence of various diseases (e.g., the wild shifts in the prevalence of COVID across time). Emerging markets in developing countries face this problem where models that work in developed countries might be far from optimal or even unsuitable for other parts of the world. For example, in medical diagnosis, the prevalence of diseases can change from region to region and also can evolve with passage of time presenting distribution shifts issues to the healthcare/diagnostic systems. Since the 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.

Research Overview To make deep learning robust and reliable, the primary focus of my research is to improve and evaluate deep learning models in the face of distribution shifts. My ultimate goal is to deploy robust algorithms in real-world applications, including but not limited to autonomous driving and healthcare. While, in principle, one may obtain fresh labeled data from the target distribution to re-train or evaluate the model every the distribution shifts, this can be prohibitively time-consuming and 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 primarily involves understanding the behavior of deep learning models, both theoretically and empirically, and using those insights to develop robust procedures.

Robustifying Models under Distribution Shift

To deploy machine learning models in the real world, we need to trust that the systems that have performed well in the past will continue to do so in the future. Hence, one natural question arises, *how to make these models robust in the face of distributions shifts?*

Label Shift Absent assumptions on the nature of the distribution shifts, this task is impossible. However, some assumptions can render shift detection, estimation, and on-the-fly updates to our classifiers possible with access to unlabeled data from target. In my recent work [13], I addressed the label shift setting, where we observe a shift in the label distribution (i.e., $p(y)$ changes), but the conditional distribution $p(x|y)$ does not change. Label shift arises in medical diagnostic problems, e.g., during a flu outbreak, the prevalence of flu is likely to change much faster than the probability of symptoms given disease. I developed a common framework for estimating shifts in the label distribution and contributed the first theoretical characterization of a state-of-the-art shift estimation method.

Open Set Classification Currently, I am working to extend the label-shift approach to a setting where we are confronted with data from a previously unseen class. This framework will enable machine learning models to not only handle shifts in distribution among the existing classes but also gracefully handle outlier data when it arises. Note that humans exhibit this capability routinely. Faced with new surprising symptoms, doctors can recognize the presence of a previously unseen ailment and attempt to estimate its prevalence. Motivated by this, I aim to solve the following two tasks: (i) estimate the fraction of previously unseen examples; and (ii) learn to discriminate between previously seen classes and the new class. In the recent work [15], we focused on the base case, where only one class has been seen previously. Next, I plan to extend the proposed techniques to multiclass problems, bridging my work on label shift and PU learning.

Continuous Distribution Shift Till now, I discussed the two-phase domain adaptation framework. However, real world has no fixed target distribution. Instead, distributions shift in real-time, and modeling them as being stationary can be over-simplistic, especially if the model is deployed over a long period. I wish to build on top of insight afforded by the source-target domain adaptation framework, to guide practical systems that can cope with gradual changes over time. I plan to begin with the label shift scenario, where $p(x|y)$ remains fixed but the label distribution $p(y)$ drifts over time. Building on top of insights from my work on two-phase label shift adaptation, I would like to develop algorithms for estimation and adaptation in an online setting.

Evaluating Models with Unlabeled data

In the quest for building reliable machine learning systems, obtaining certificates for models' accuracy is as important as robustifying these systems. While labeled hold-out data can be used to obtain post-hoc error estimates, this method has a few drawbacks. Provisioning a holdout dataset shrinks the training set and estimates based on holdout sets lose their validity with successive re-use of the holdout set due to adaptive overfitting [11; 2]. Furthermore, obtaining unseen fresh labeled data can be prohibitively expensive and time-consuming if the target distribution is non-stationary. To this end, we ask the following question: *Can we predict models' performance without access to held-out data?*

Guaranteeing In-Distribution Generalization To assess in-distribution generalization, classical learning theoretic tools yield vacuous guarantees for overparameterized models [10; 12]. Taking a departure from complexity-based approaches, in my recent paper [16], I introduce RATT, a method that leverages unlabeled data to produce post-training generalization bounds. Here, we assign random labels to a batch of unlabeled data, augment them with a clean training dataset, and finally, we train on this mixture following standard practices. We track the error on the randomly labeled portion of training data to upper bound the population error. Whenever classifiers achieve low error on clean data and high error on noisy data, RATT bound provides a tight upper bound on the true risk. Intuitively, our bound measures the empirical Rademacher complexity of the actual procedure.

One limitation of RATT bound is that it yields vacuous guarantees when the model memorizes all randomly labeled data. As next steps, I wish to operationalize RATT bound in the interpolation regime by discarding the randomly labeled data after the initial stages of training. I have preliminary results where we provably transfer non-vacuous guarantees of RATT bound to the interpolation regime if the model doesn't change much its output space. In the future, I also wish to make RATT training more practical and combine advantages due to implicit regularization induced by label noise training (e.g., flat solutions [4] and improved calibration [8]) simultaneously obtaining post-hoc generalization bounds.

Predicting Out-Of-Distribution (OOD) Generalization Even after robust training, the performance of machine learning models can vary substantially when models are evaluated on data samples coming from a distribution that is close to but different from the training distribution. To estimates of models' performance on OOD data sets, I am developing a method that can utilize unlabeled target data to predict models' performance. Given a classifier, held-out labeled source data (i.e., from the distribution on which the classifier was trained), and unlabeled target data, we observe that we can use shift in the distribution of models' softmax probabilities to predict models' accuracy on target data. Across more than 100 natural and synthetic distribution shifts on ImageNet and CIFAR, I proposed a method that can accurately predict estimates for models' error (mean prediction error $< 2 - 4\%$) substantially improving on top of some concurrent works [5; 3]. This work is under submission [14] and is my ongoing project in collaboration with Google.

Currently, I am working to understand the properties of distribution shifts that enable using models' softmax predictions to predict models' error. As formalized in our work [14], no method of estimating accuracy will work generally without assumptions on the source classifier or the nature of the shift. However, real-world distribution shifts seem to satisfy some astonishing simple structures that render estimation from unlabeled data. Similar surprising findings have also been highlighted in a recent line of work where it is observed that across different pairs of natural and synthetic distribution shifts performance of different models linearly correlate [9]. I hope to make connections with this literature to identify structures that enable accurate prediction on OOD data with models' softmax probabilities.

Conclusion As we enter the era of big data, where anomalies are frequent, I believe my research advances core areas critical to next-generation machine learning for reliable deployment in the real world. The research I describe here is just a small portion of my planned research agenda, and I am continuously working to make progress towards reliable machine learning models.

References

[1] A.-B. et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 2019.

[2] D. et al. Preserving statistical validity in adaptive data analysis. In *Proceedings of ACM symposium on Theory of computing*, 2015.

[3] D. et al. Are labels necessary for classifier accuracy evaluation? *arXiv preprint arXiv:2007.02915*, 2020.

[4] D. et al. Label noise sgd provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

[5] G. et al. Predicting with confidence on unseen distributions. *arXiv preprint arXiv:2107.03315*, 2021.

[6] H. et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[7] H. et al. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[8] M. et al. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.

[9] M. et al. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

[10] Z. et al. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[11] K. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[12] V. Nagarajan and Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2019.

[13] **Garg et al.** A unified view of label shift estimation. In *Advanced in Neural Information Processing (NeurIPS)*, 2020.

[14] **Garg et al.** Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shift Connecting Methods and Applications*, 2021. under review.

[15] **Garg et al.** Mixture proportion estimation and PU learning: A modern approach. In *Advanced in Neural Information Processing (NeurIPS)*, 2021.

[16] **Garg et al.** RATT: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, 2021.