LEVERAGING UNLABELED DATA TO PREDICT OUT-OF-DISTRIBUTION PERFORMANCE

Saurabh Garg*
Carnegie Mellon University
sgarg2@andrew.cmu.edu

Sivaraman Balakrishnan
Carnegie Mellon University
sbalakri@andrew.cmu.edu

Zachary C. Lipton
Carnegie Mellon University
zlipton@andrew.cmu.edu

Behnam Neyshabur
Google Research, Blueshift team
neyshabur@google.com

Hanie Sedghi
Google Research, Brain team
hsedghi@google.com

ABSTRACT

Real-world machine learning deployments are characterized by mismatches between the source (training) and target (test) distributions that may cause performance drops. In this work, we investigate methods for predicting the target domain accuracy using only labeled source data and unlabeled target data. We propose Average Thresholded Confidence (ATC), a practical method that learns a *threshold* on the model's confidence, predicting accuracy as the fraction of unlabeled examples for which model confidence exceeds that threshold. ATC outperforms previous methods across several model architectures, types of distribution shifts (e.g., due to synthetic corruptions, dataset reproduction, or novel subpopulations), and datasets (WILDS, ImageNet, BREEDS, CIFAR, and MNIST). In our experiments, ATC estimates target performance 2–4× more accurately than prior methods. We also explore the theoretical foundations of the problem, proving that, in general, identifying the accuracy is just as hard as identifying the optimal predictor and thus, the efficacy of any method rests upon (perhaps unstated) assumptions on the nature of the shift. Finally, analyzing our method on some toy distributions, we provide insights concerning when it works[1].

1 INTRODUCTION

Machine learning models deployed in the real world typically encounter examples from previously unseen distributions. While the IID assumption enables us to evaluate models using held-out data from the *source* distribution (from which training data is sampled), this estimate is no longer valid in presence of a distribution shift. Moreover, under such shifts, model accuracy tends to degrade (Szegedy et al., 2014; Recht et al., 2019; Koh et al., 2021). Commonly, the only data available to the practitioner are a labeled training set (source) and unlabeled deployment-time data which makes the problem more difficult. In this setting, detecting shifts in the distribution of covariates is known to be possible (but difficult) in theory (Ramdas et al., 2015), and in practice (Rabanser et al., 2018). However, producing an optimal predictor using only labeled source and unlabeled target data is well-known to be impossible absent further assumptions (Ben-David et al., 2010; Lipton et al., 2018).

Two vital questions that remain are: (i) the precise conditions under which we can estimate a classifier's target-domain accuracy; and (ii) which methods are most practically useful. To begin, the straightforward way to assess the performance of a model under distribution shift would be to collect labeled (target domain) examples and then to evaluate the model on that data. However, collecting fresh labeled data from the target distribution is prohibitively expensive and time-consuming, especially if the target distribution is non-stationary. Hence, instead of using labeled data, we aim to use unlabeled data from the target distribution, that is comparatively abundant, to predict model performance. Note that in this work, our focus is *not* to improve performance on the target but, rather, to estimate the accuracy on the target for a given classifier.

[*] Work done in part while Saurabh Garg was interning at Google

[1] Code is available at `https://github.com/saurabhgarg1996/ATC_code`.