Probabilistic Weakness Recognition for Black-Box Validation

Robert J. Moss mossr@cs.stanford.edu
Bernard Lange blange@stanford.edu
Mykel J. Kochenderfer mykel@stanford.edu
Aeronautics and Astronautics, Stanford University, Stanford, CA 94305, USA

Effective validation is key in developing complex black-box systems. When validating minor changes to a black-box system, exhaustively evaluating over the entire validation dataset may be computationally intractable. The challenge then becomes to intelligently automate selective validation given knowledge of the previous and current versions of the system failures experienced so far. Removing explicit human input in fine tuning this process—especially with an abundance of data—is the focus of this proposed work. If effective, this work could impact how black-box systems are designed and tested.

We propose an adaptive black-box validation framework that will learn system weaknesses over time and exploit this knowledge to propose validation samples that will likely result in a failure. Our proposed approach consists of two stages. First, we learn a low-dimensional representation of each dataset sample to identify features causing the system to fail. The proposed architecture includes an encoder-decoder and a failure predictor together forming a single feed-forward neural network, inspired by the domain-adversarial architecture proposed by Ganin *et al.* [1]. The encoding architecture will vary depending on the task, e.g. image classification (using an autoencoder), video prediction tasks (using a video autoencoder [2]), and decision-making tasks (using a spatio-temporal graph-structured recurrent mechanism [3]). Training would occur every system iteration to highlight changes in the system and potentially expose features that caused failures. In the second stage, the low-dimensional representation is used to select candidates which will likely result in failures. The search over the dataset is guided by the failure likelihood of the sample given the current and previous versions of the system. This probability will be encoded in a separate neural network updated after every new sample is evaluated, and hence adjusting our understanding of the system's weaknesses. The task is framed as a continual learning problem and the probability representation will be used as a prior in the next system iteration. After an evaluation budget is reached, the subset of selected samples using the current black-box system will be used to update our low-dimensional dataset representation.

References

[1] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, 2016.

[2] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *arXiv:1511.06309*, 2015.

[3] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," *arXiv:2001.03093*, 2020.