Ethan Harvey's Prospectus

Title: Probabilistic Methods for Building Medical Imaging Classifiers from Limited Labeled Data

Advisor: Michael C. Hughes

Advisor Signature: _____ **Date:** 2024-11-12

Additional dissertation committee members: Liping Liu and Robert J. K. Jacob

Dataset size is often the limiting factor defining the ceiling of performance in a deep learning application. Without a large dataset of labeled examples available for training, deep learning models face challenges like overfitting and poor generalization. We develop probabilistic methods to overcome challenges posed by limited labeled data while working on a multi-year funded project to build models that can diagnose stroke/dementia from computed tomography (CT) and magnetic resonance imaging (MRI) images of the brain. These challenges include extrapolating classifier accuracy to larger datasets and improving model performance at the current dataset size.

When a large dataset of labeled images is not available, research projects often have a common trajectory: (1) gather a small "pilot" dataset of images and corresponding class labels, (2) train classifiers using this available data, and then (3) plan to collect an even larger dataset to further improve performance. When gathering more labeled data is expensive, practitioners face a key decision in step 3: given that the classifier's accuracy is Y% at the current size x, how much better might the model do at 2x, 10x, or 50x images? Recent approaches have focused almost entirely on estimating one single "best-fit" curve, using power laws (Hestness et al., 2017; Rosenfeld et al., 2020), piecewise power laws (Jain et al., 2023), or other functional forms (Mahmood et al., 2022). In practice, however, no single curve fit to a limited set of size-accuracy pairs can extrapolate perfectly. We offer a probabilistic approach based on Gaussian processes to extrapolate accuracy or similar performance metrics as dataset size increases (Harvey et al., 2023). Our mean functions are carefully chosen to adapt past work on point estimation to accommodate a priori common sense (accuracy or AUROC can never be larger than 1) and domain knowledge (some tasks may not reach perfect accuracy, only 1-ε).

When collecting an even larger dataset is not possible, improving model performance at the current dataset size often requires alternative strategies. We evaluate several transfer learning strategies from a Bayesian perspective (Harvey et al., 2024a). We find that standard transfer learning informed by an initialization only performs far better than reported in previous comparisons. The relative gains of methods using informative priors over standard transfer learning vary in magnitude across datasets. Among methods using informative priors, we find that an isotropic covariance (Chelba & Acero, 2006; Xuhong et al., 2018) appears competitive with a learned low-rank covariance matrix (Shwartz-Ziv et al., 2022) while being substantially simpler to understand and tune.

Finally, we propose an alternative to grid search to select regularization hyperparameters that control over-fitting: directly learning regularization hyperparameters on the full training set via model selection techniques based on the *evidence lower bound* ("ELBo") objective from variational methods (Harvey et al. 2024b). For deep neural networks with millions of parameters, we specifically recommend a modified ELBo that upweights the influence of the data likelihood relative to the prior while remaining a valid bound on the evidence for Bayesian model selection. Our proposed technique overcomes several key disadvantages of grid search: the search is

computationally expensive, requires carving out a validation set that reduces the size of available data for model training, and requires practitioners to specify candidate values. We demonstrate effectiveness on image classification tasks on several datasets, yielding heldout accuracy comparable to existing approaches with far less compute time.

References

Ciprian Chelba and Alex Acero. Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot. *Computer Speech & Language*, 20(4):382–399, 2006.

Ethan Harvey, Wansu Chen, David M. Kent, and Michael C. Hughes. A Probabilistic Method to Predict Classifier Accuracy on Larger Datasets given Small Pilot Data. In *Machine Learning for Health (ML4H)*, 2023.

Ethan Harvey, Mikhail Petrov, and Michael C. Hughes. Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported. *Transactions on Machine Learning Research (TMLR)*, 2024a. ISSN 2835-8856.

Ethan Harvey, Mikhail Petrov, and Michael C. Hughes. Learning the Regularization Strength for Deep Fine-Tuning via a Data-Emphasized Variational Objective. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024b.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Achin Jain, Gurumurthy Swaminathan, Paolo Favaro, Hao Yang, Avinash Ravichandran, Hrayr Harutyunyan, Alessandro Achille, Onkar Dabeer, Bernt Schiele, Ashwin Swaminathan, Stefano Soatto. A Meta-Learning Approach to Predicting Performance and Data Requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, and Marc T. Law. How Much More Data Do I Need? Estimating Requirements for Downstream Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations (ICLR)*, 2020.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew G. Wilson. Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Li Xuhong, Yves Grandvalet, and Franck Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. In *International Conference on Machine Learning (ICML)*, 2018.