Over the past decade, deep learning has demonstrated high accuracy on challenges in fields like computer vision and natural language processing, revolutionizing these fields in the process. Deep learning models are now a fundamental building block for applications such as autonomous driving, medical imaging, and machine translation. However, many challenges remain when deploying these models in production. Researchers and practitioners must address a diversity of questions, including how to efficiently design, train, and deploy resource-intensive deep learning models and how to automate these approaches while ensuring robustness to changing conditions.

Our work aims to improve the efficiency of deep learning training and inference, as well as the underlying systems’ robustness to changes in the environment. We address these issues by focusing on the many hyperparameters that are tuned to optimize the model’s accuracy and resource usage. These hyperparameters include the choice of model architecture, the training dataset, the optimization algorithm, the hyperparameters of the optimization algorithm (e.g., the learning rate and momentum) and the training time budget. Currently, in practice, almost all hyperparameters are tuned once before training and held static. This is sub-optimal as the conditions that dictate the best hyperparameter value change over time (e.g., training progress, inference hardware). We apply dynamic tuning to hyperparameters that have traditionally been considered static. Using three case studies, we show that using runtime information to dynamically adapt hyperparameters that are traditionally static, such as the emphasis on individual training examples and the weights updated during transfer learning, can increase the efficiency of deep learning training and inference.

Thesis Committee:
Gregory R. Ganger, Chair
David G. Andersen
Michael Kaminsky
Michael Kozuch, Intel Labs
Padmanabhan S Pillai, Intel Labs
Rahul Sukhankar, Google