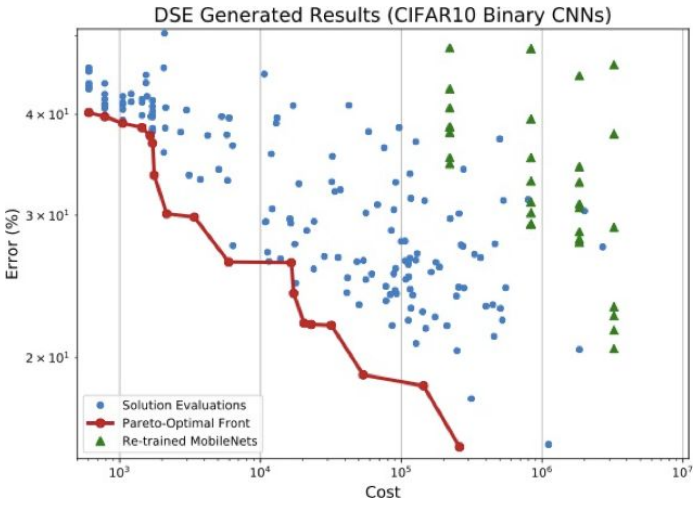
Network Architecture Search for Efficient Wake-word Detection

Warren Gross and Brett H. Meyer
McGill University and Effortless AI
Montreal, Quebec, Canada

Rapidly growing demand for energy efficient and compact neural network implementation on tiny devices has resulted in two basic approaches: (1) fine tuning and pruning off-the-shelf (OTS) networks designed and trained for one problem and target platform so they can be applied to another problem and target platform; (2) searching for and optimizing architectures specifically designed for the targeted problem and platform. The problem with (1) is that there's no guarantee that OTS networks can be sufficiently squeezed to fit onto resource-constrained architectures without sacrificing accuracy; the problem with (2) is that most approaches to network architecture search (NAS) are resource-constraint agnostic. New NAS approaches, and metrics to guide them, are needed that can quickly identify neural network architectures that find the best trade-offs between the constraints faced by *tinyML* systems (inference delay, inference energy, memory footprint, etc) and accuracy.

We have developed an automated tool (OPAL) for designing deep neural networks for resource-constrained devices. OPAL uses response surface modeling to perform multi-objective optimization of neural network hyperparameters (number of layers, per-layer parameters—including quantization—learning rate, etc), learning the combinations of hyperparameters expected to strike Pareto-optimal trade-offs. Response surface modeling has clear advantages over approaches that incrementally improve architectures (e.g., using genetic algorithms, or recurrent neural networks) since it can discover networks with radically different architectures from each other. In a small number (10's to 100's) of training runs, OPAL can outperform fine-tuned or retrained OTS networks (e.g., MobileNets), achieving equivalent accuracy at a fraction of the computational complexity. For instance, the figure illustrates that MobileNets accuracy can be achieved with 1% of the computational cost. The key insight: efficient networks are grown, not squeezed.



In the demo accompanying the poster, we will demonstrate wake-word detection on an ARM Cortex-M4 system running neural networks automatically discovered by OPAL.