Achieving Pareto Optimality using Efficient Parameter Reduction for DNNs in Resource-Constrained Edge Environment

Authors: **Atah Nuh Mih**, Alireza Rahimi, Asfia Kawnine, Francis Palma, Monica Wachowicz, Rickey Dubay, Hung Cao

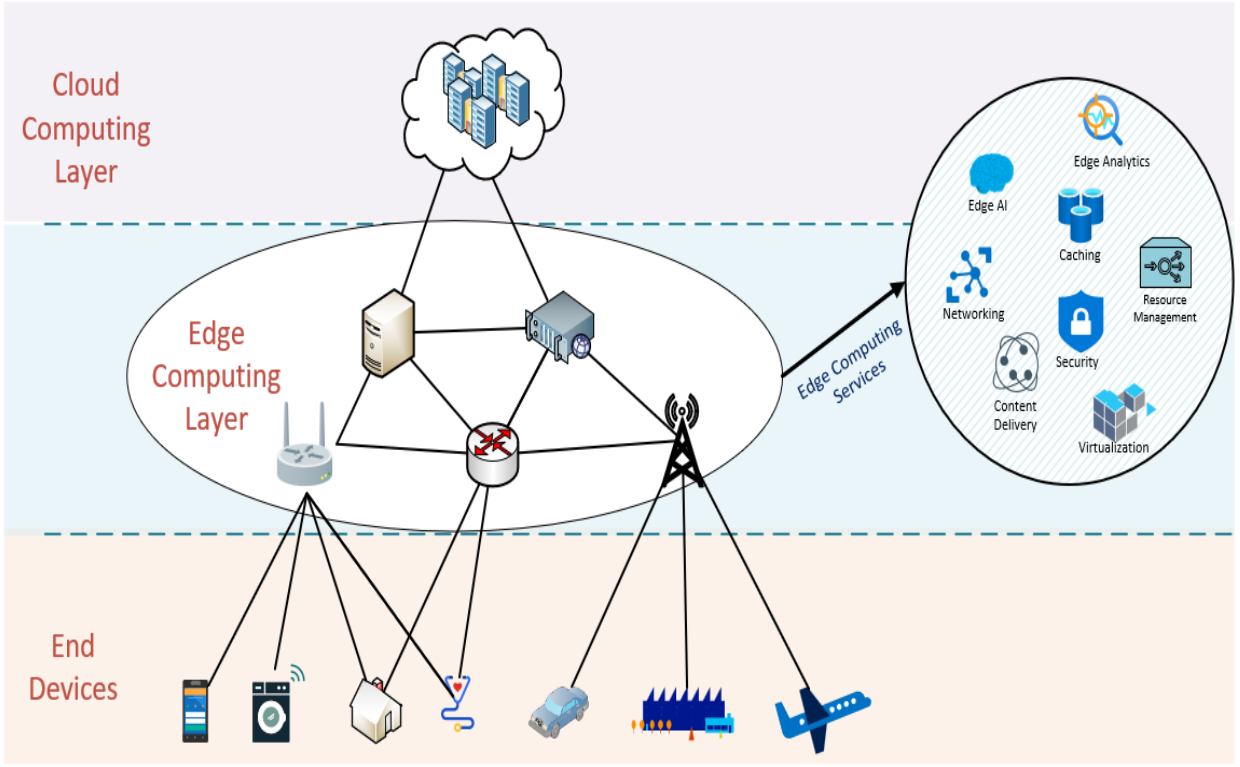
Presentation Outline

- Background
- Problem Statement
- Related Works
- Proposed Method
- Implementation
- Experiments and Results
- Discussion
- Conclusion
- Future Works

Introduction

Background

- Cloud-based machine learning
  - 'Unlimited' computation
  - High bandwidth and latency, privacy
- Edge-based machine learning
  - Closer to the user (privacy), flexible to specific needs (adaptability)
  - Low computation



Problem Statement

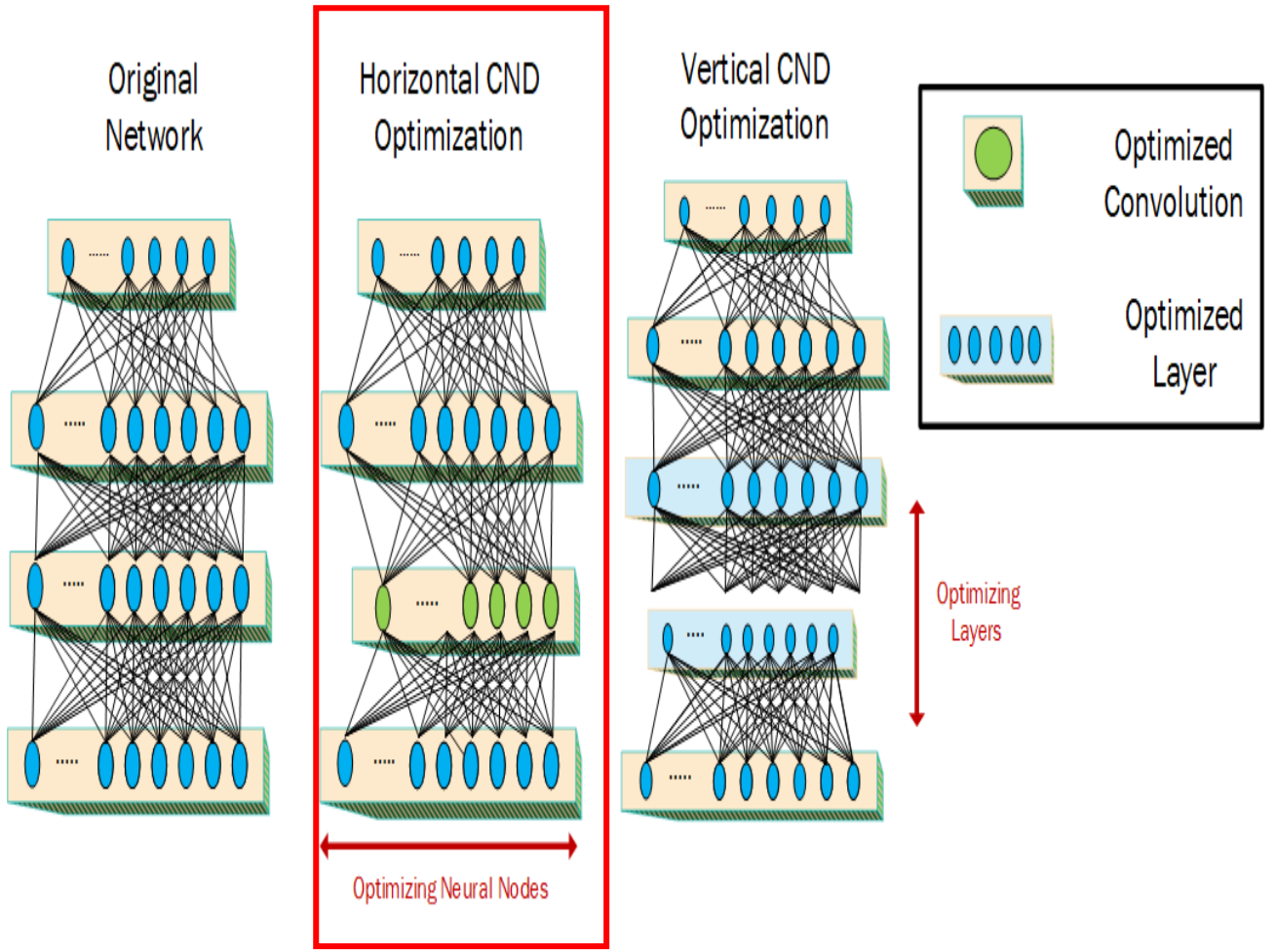
- Model training computationally intensive
- Lightweight models vs Heavyweight models
  - Accuracy and hardware resource utilization
- Can DL models be optimized for low resource utilization but maintaining high accuracy?

Related Works

- Model optimization using **post-training quantization**
  - Hardware-friendly post-training quantization - Habi et al. (2021)
  - 4-bit (Banner et al. 2019) and 8-bit (Wu et al. 2020) quantization
- Model optimization with **neural architecture search**
  - ResNet-18 backbone - Li et al. (2023)
  - MobileNetV2 backbone - Lyu et al. (2021)
- AI models at the edge
  - Student-teacher approach - Kukreja et al. (2019)
  - On-device Learning Anomaly Detector (ONLAD) - Tsukada et al. (2020)

Proposed Method

- Model optimization for DNNs
  - Use of compact network design
  - Applying the optimization on a DNN (Xception)
- Transfer learning for efficient resource usage?



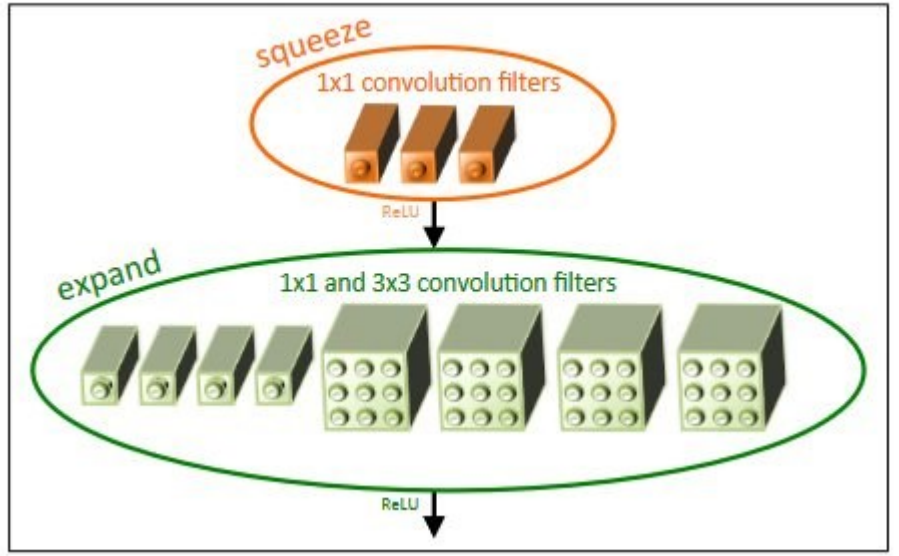
DNN = Deep Neural Network

Proposed Method: Optimizing a DNN

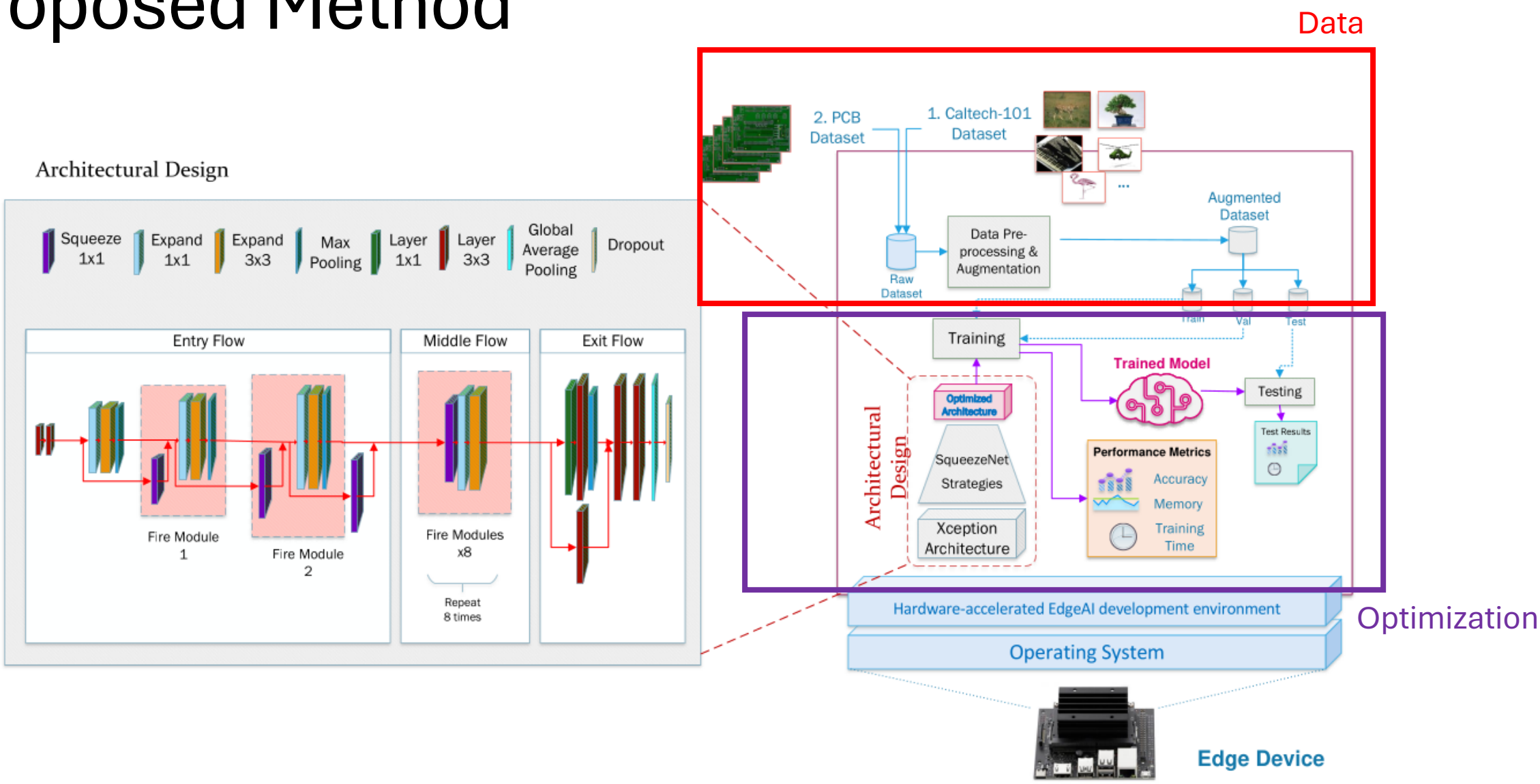
- Efficient parameter reduction strategies - SqueezeNet
  - 3x3 filters with 1x1 filters
  - Decrease number of input channels into 3x3 filters
  - Downsample late in the network

- Fire Module

- Integrating into a DNN (Xception)



Proposed Method



Experiments & Result

Implementation

- Edge Device: A203 Mini PC
  - Nvidia's Jetson Xavier NX 8GB module
- Two experiments
  - Caltech-101
  - PCB Defects
- Evaluation
  - Optimized Model
  - Xception
  - MobileNetV2
  - EfficientNetV2B1

Experiment 1 – Caltech 101

- Caltech-101
  - 101 object classes
  - 9,146 images
- Training setup
  - Base models

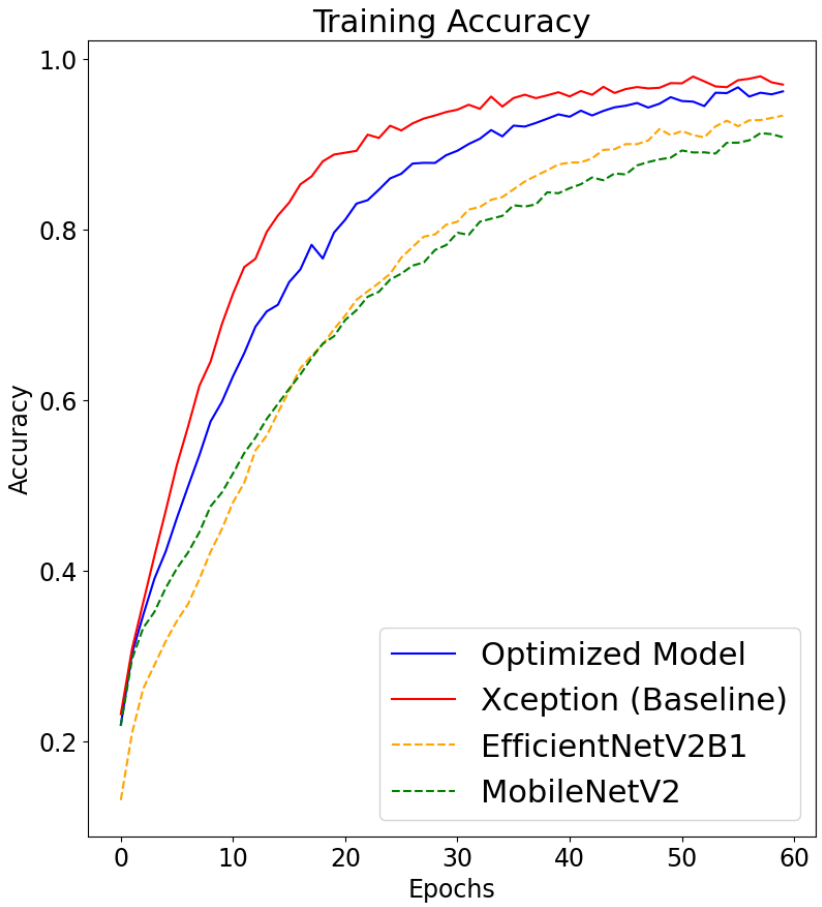
Results: Optimization Performance

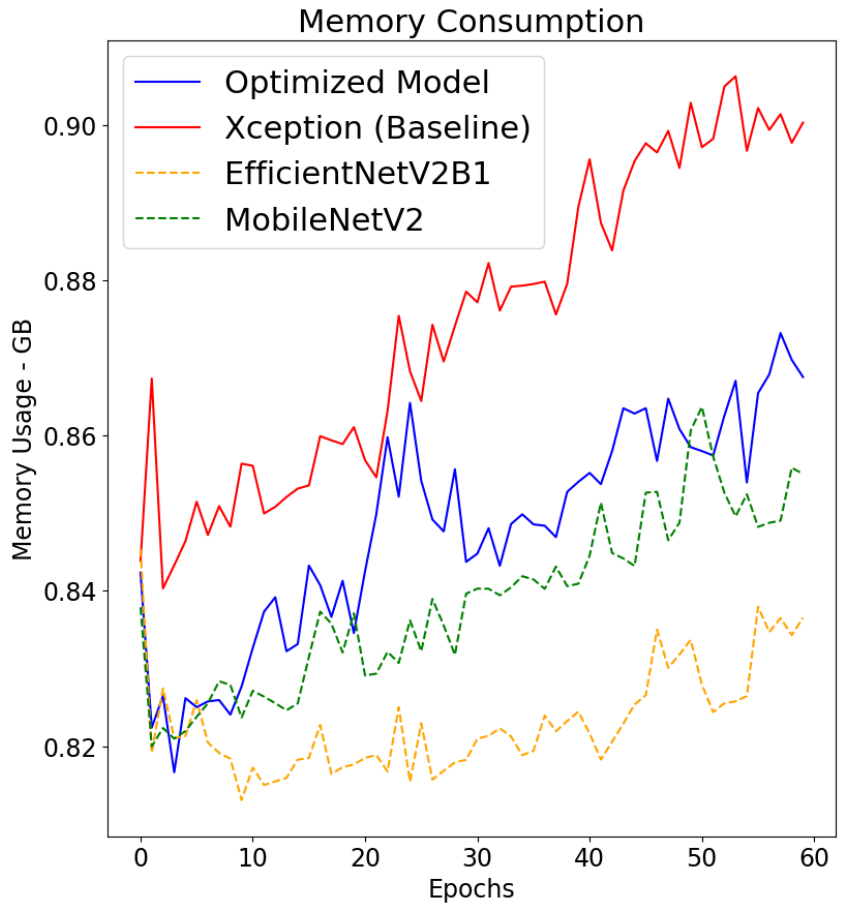
	Model	#Params	Train Accuracy	Test Accuracy	Avg Mem/Epoch	Avg Time/Epoch	Avg Inference Time
Baseline Comparison	Optimized Model	↓15.8M	↓96.16%	↑76.21%	↓847.9MB	↓523.88s	↓465ms
	Xception	21.1M	96.95%	75.89%	874.6MB	702.19s	520ms
Other Lightweight Models	EfficientNetV2B1	7.1M	93.32%	30.53%	823.0MB	381.44s	383ms
	MobileNetV2	2.4M	90.79%	58.11%	838.6MB	301.12s	306ms

↓ - Decrease wrt Xception

↑ - Increase wrt Xception

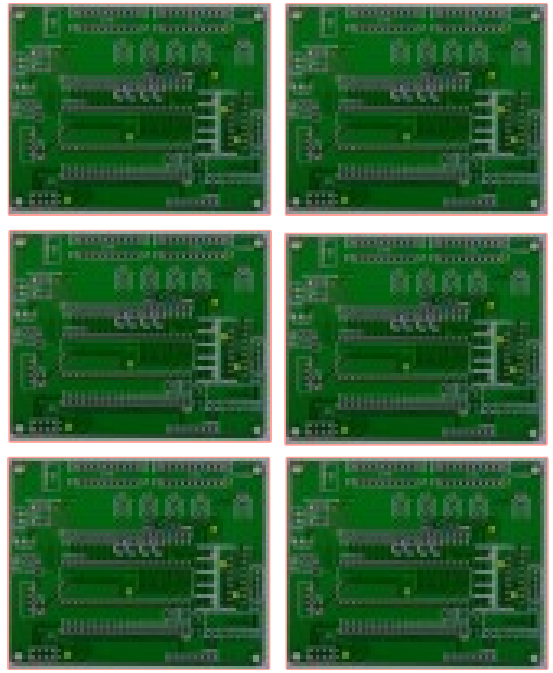
Training pattern for Caltech-101 image classification





Experiment 2 – PCB Defects

- Dataset
  - 6 object classes
  - 1386 images augmented to 22,000 images
- Training setup
  - Base models
  - Models pre-trained on Caltech-101



Results: Base Models vs Pre-trained

	Base Models				Pre-trained			
Model	Train Accuracy	Test Accuracy	Avg Mem/Epoch	Avg Inf Time	Train Accuracy	Test Accuracy	Avg Mem/ Epoch	Avg Inf Time
Optimized Model	97.81%	90.30%	865.8MB	465ms	70.71%	69.80%	833.7MB	446ms
Xception	97.66%	88.10%	893.6MB	519ms	70.84%	71.00%	831.5MB	509ms
EfficientNetV2 B1	88.87%	55.25%	874.8MB	346ms	65.96%	67.40%	828.7MB	335ms
MobileNetV2	96.13%	50.50%	849.4MB	295ms	67.89%	69.05%	818.2MB	295ms

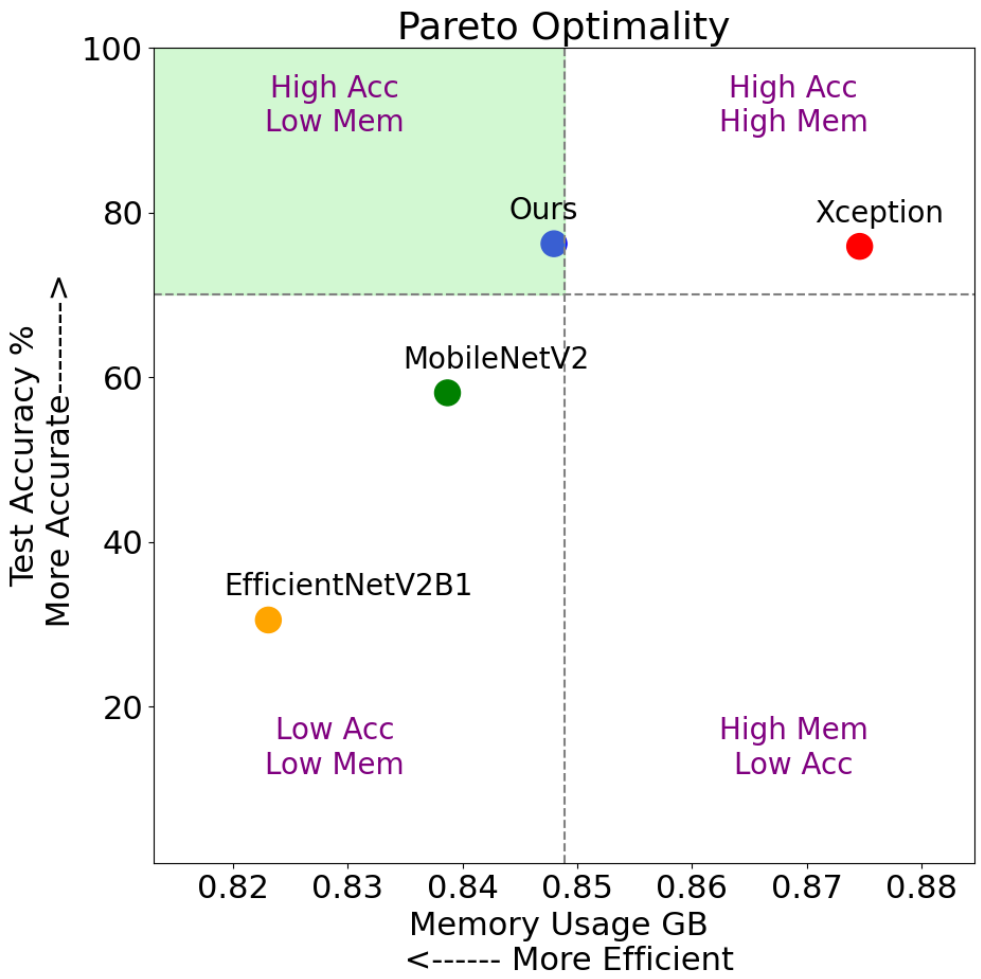
Good accuracy

Resource efficient

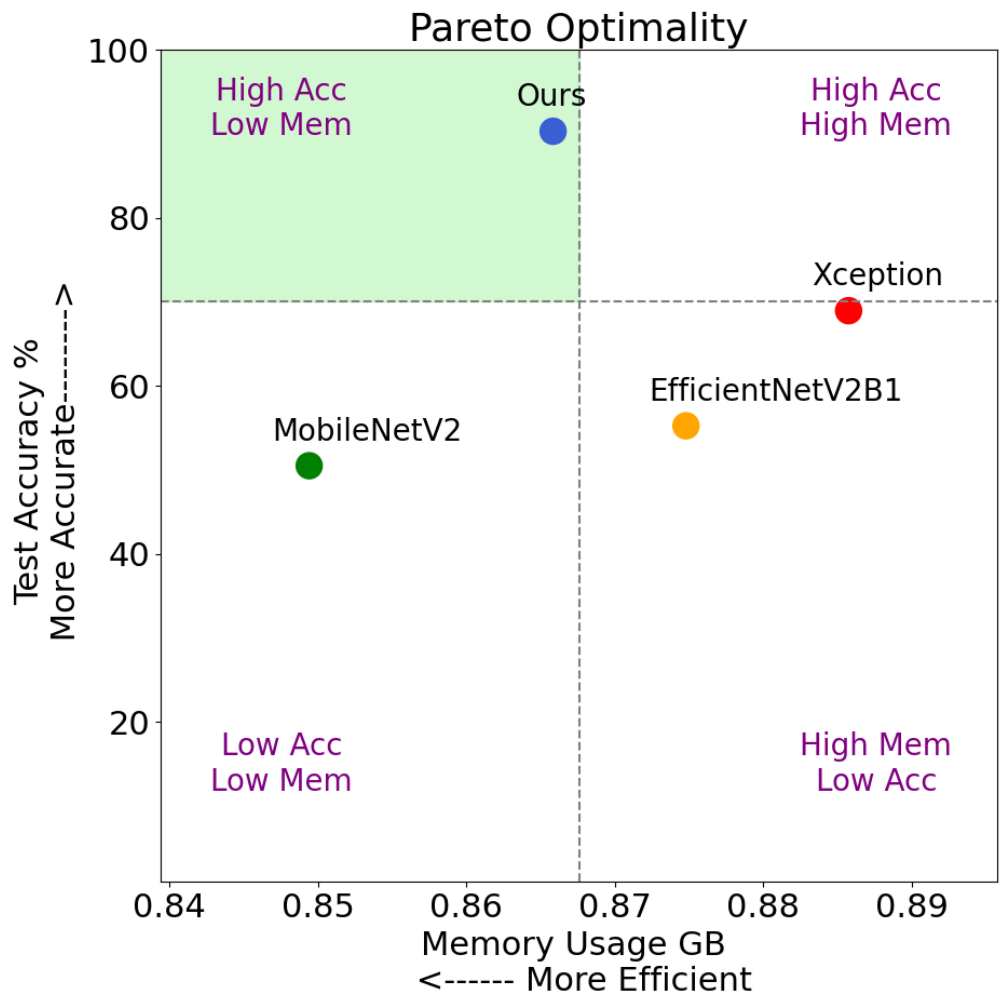
Discussion & Conclusion

Evaluating Dual Objectives

Caltech 101 Image Classification



PCB Defect Detection



Conclusion

Optimization maintains high acc but low consumption?

TL improves resource efficiency

TL = Transfer Learning

Future Works

- Diverging from manual design processes
- Neural architecture search

ANALYTICS EVERYWHERE LAB

Dr Hung Cao
Assistant Professor
Lab Director
University Of New Brunswick, Canada
hcao3@unb.ca

Dr Francis Palma
Assistant Professor
Faculty of Computer Science
University Of New Brunswick, Canada

Dr Monica Wachowicz
Adjunct Professor
Associate Dean
Geospatial Science
RMIT University, Australia

Dr Trevor Hanson
Professor
Faculty of Civil Engineering
University Of New Brunswick, Canada

Atah Nuh Mih
MSc Student

Alireza Rahimi
MSc Student

Asfia Kawnine
MSc Student

Truong Thanh Hung Nguyen
PhD Student

Simran Dadhich
MSc Student

...