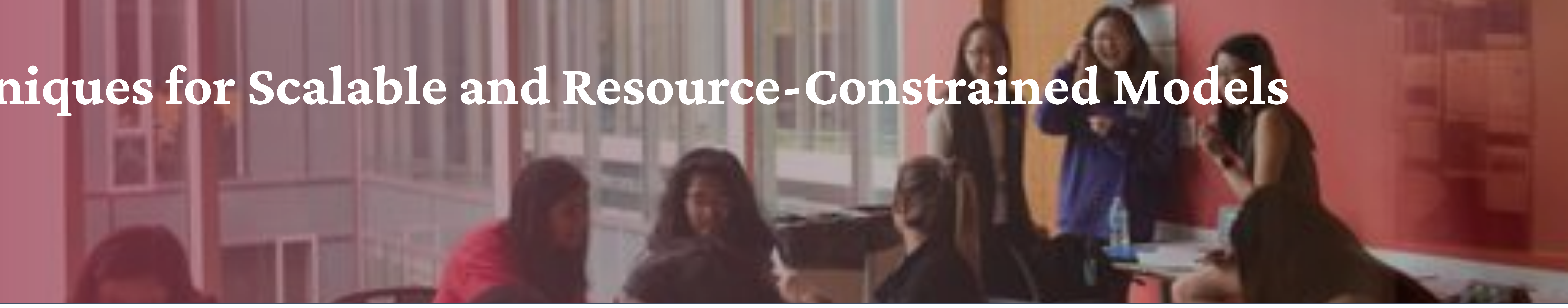
Towards Efficient AI: Techniques for Scalable and Resource-Constrained Models



Louis Sloot, Elizabeth Qiu, Maggie Chen
Mentors: Dr. Yifan (Evelyn) Gong, Joyce Li

Background

Deploying deep neural networks (DNNs) on resource-constrained devices is challenging due to high computational and storage demands, making efficiency critical.

- Energy consumption of AI has surged as model parameters grow exponentially
- Data centers account for ~3% of global carbon emissions annually

The increasing mainstream demand and popularity of large-scale AI models, including large language models (LLMs) and convolutional neural networks (CNNs), makes addressing these efficiency issues even more critical. **This study explores practical solutions to reduce model size, accelerate execution, and optimize training and inference processes.**

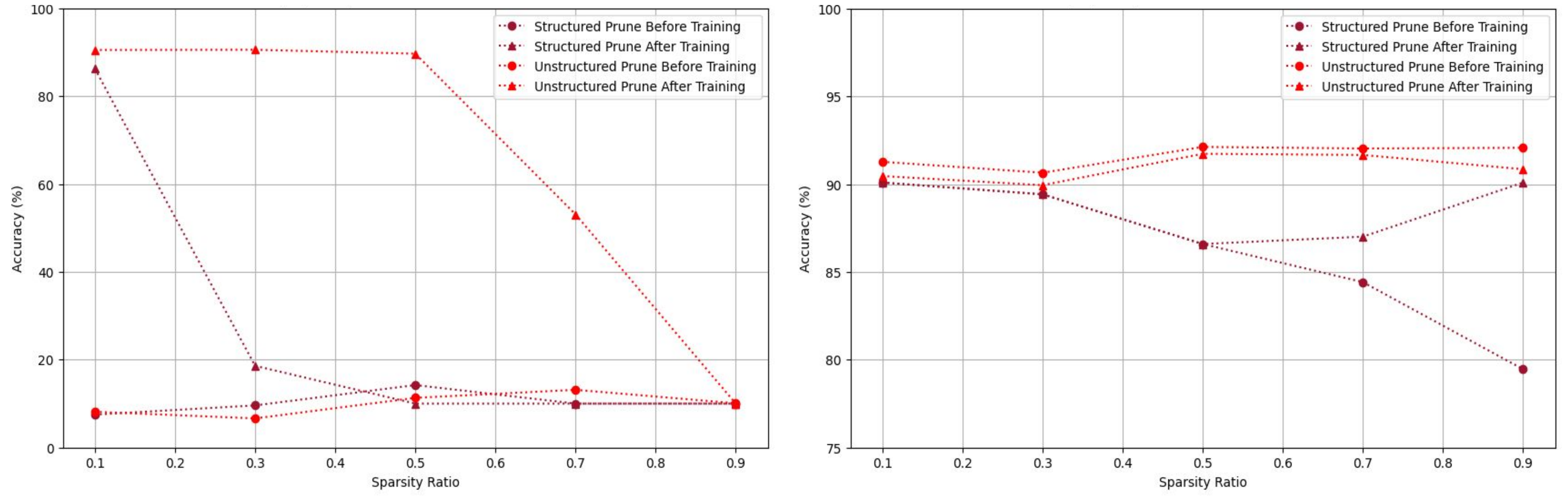


Figure 2 (*above*): Accuracy-Sparsity Ratio in Pruned Models: **Non-Retrained Models** (*Left*); **Retrained** (*Right*)

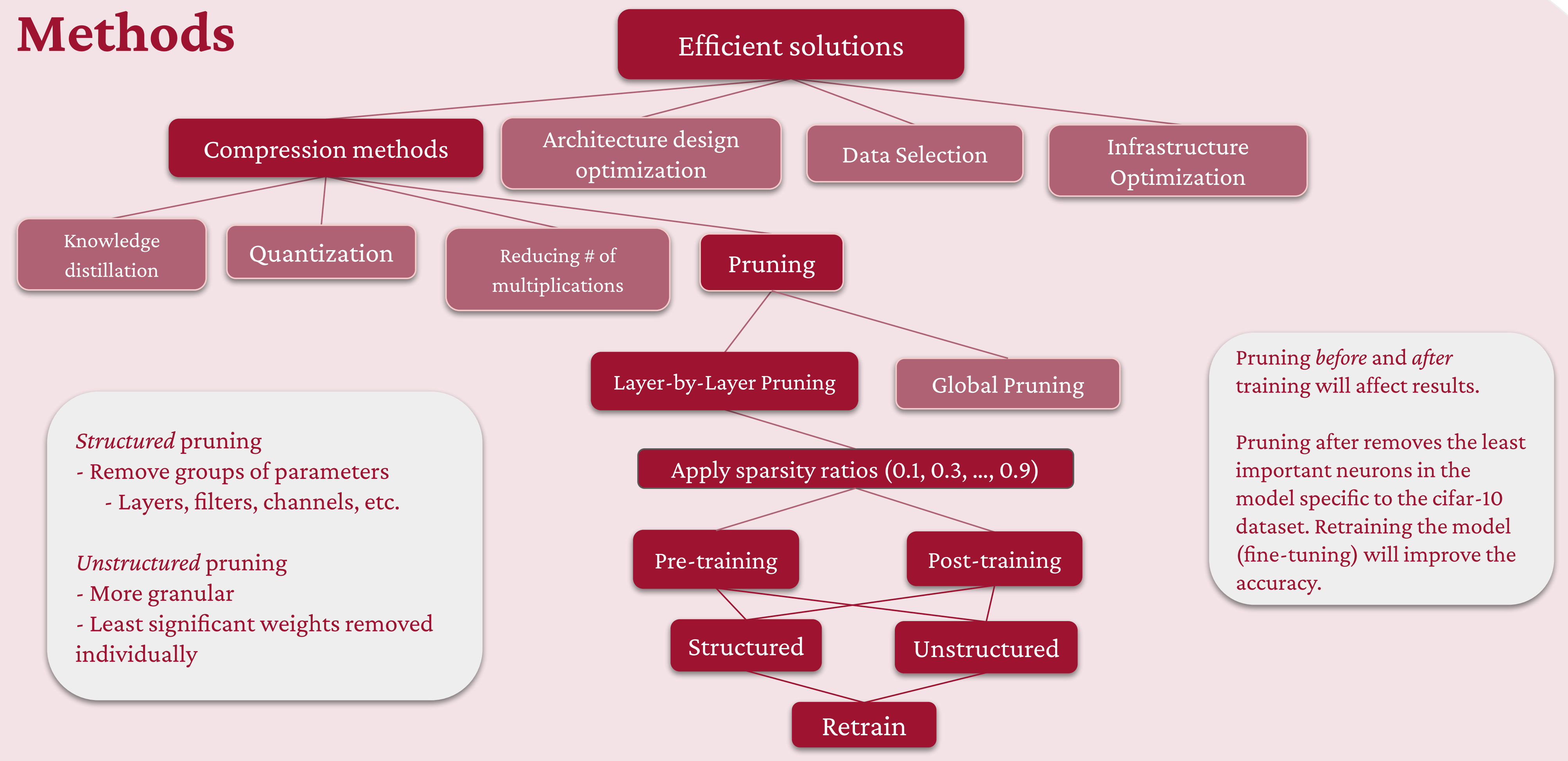
Table 1 (*below*): Multiply-Accumulate Operations (MACs) based on sparsity ratio

Sparsity Ratio	**0.0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
Predicted MACs (millions)	**72.3**	**65.1**	**50.6**	**36.2**	**21.7**	**7.24**

Takeaways

- Unstructured pruning more resistant to accuracy decrease, but theoretically harder to gain efficiency
- Structured pruning impacts accuracy; easier to leverage efficiency (smaller matrix dimensions)
- Overall, (small) portions of models can be pruned without critical loss in accuracy while making theoretical efficiency gains

Methods



Structured pruning
- Remove groups of parameters
    - Layers, filters, channels, etc.

Unstructured pruning
- More granular
- Least significant weights removed individually

Pruning *before* and *after* training will affect results.

Pruning after removes the least important neurons in the model specific to the cifar-10 dataset. Retraining the model (fine-tuning) will improve the accuracy.

Evaluation metrics:
- Accuracy (%)
- Efficiency (MACs)

Model architecture: ResNet-18

Dataset: CIFAR-10

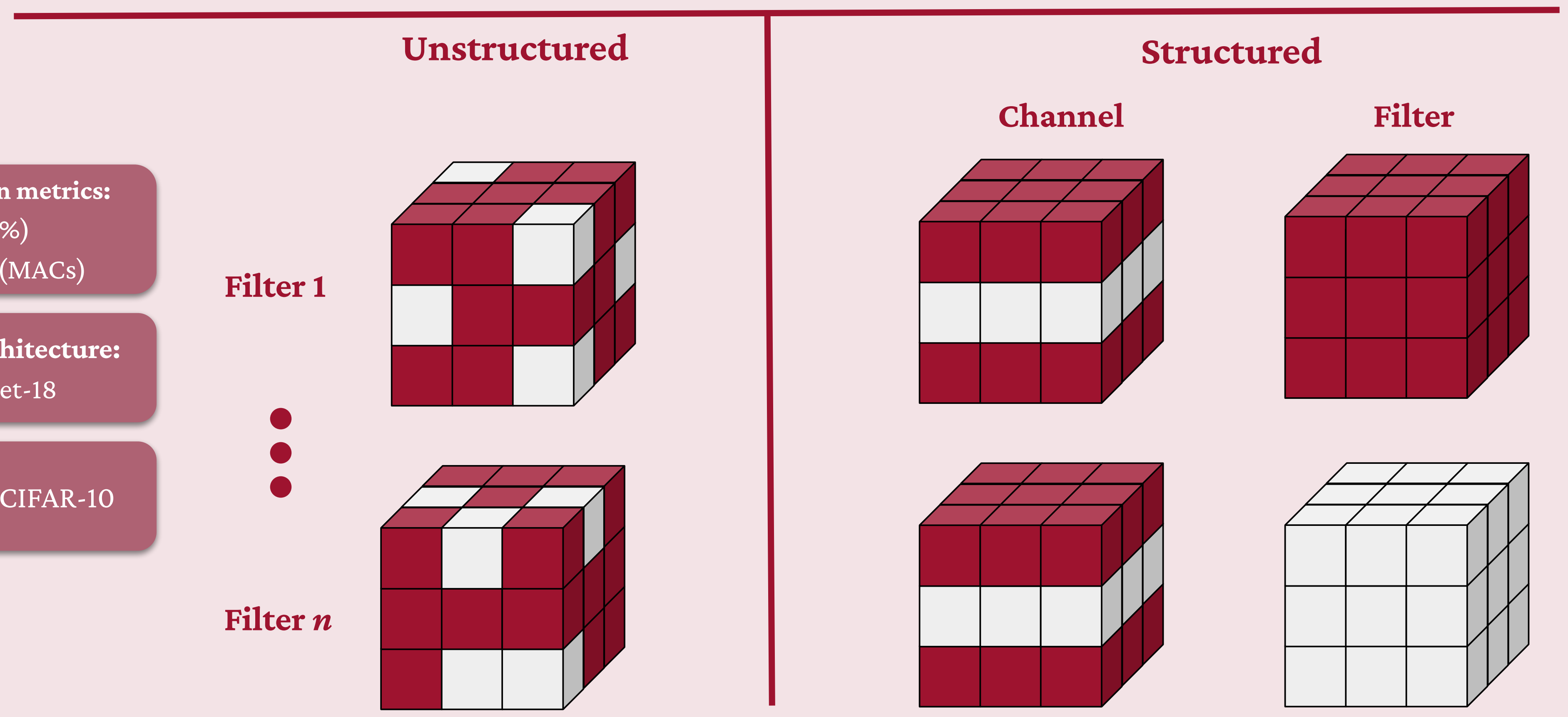


Figure 1: Pruning Methods

Future Actions

Gaining Efficiency from Zeroed Weights: Investigate, develop algorithm(s) that convert zeroed weights into lowered MACs and greater efficiency

Pruning During Training: Faster training times than pruning after training and higher accuracy than pruning before training

Sparsity Ratios: Varying sparsity ratios distributed across different layers accounts for some layers influencing results more than others

Experimenting with Different Norms: Currently using L1 norms to determine structured pruning, but other norms may yield different results

Carnegie Mellon University School of Computer Science

DEPARTMENT OF COMPUTER SCIENCE

Link to Codebase

Adobe