Effective Layer Pruning Through Similarity Metric Perspective

Ian Pons, Bruno Yamamoto, Anna H. Reali Costa and Artur Jordao
Escola Politécnica, Universidade de São Paulo

Deep Learning

Introduction

- Deep neural networks have been the predominant paradigm in machine learning for solving cognitive tasks

- Such models, however, are restricted by a high computational overhead, limiting their applicability on low-resource environments

- The recent Llama 3 model (Dubey et al., 2024) family comprehends a concrete example
  - Training requires up to 16K H100 GPUs globally distributed
  - The largest model (Llama 405B) requires 16 GPUs with 16-bit precision for inference

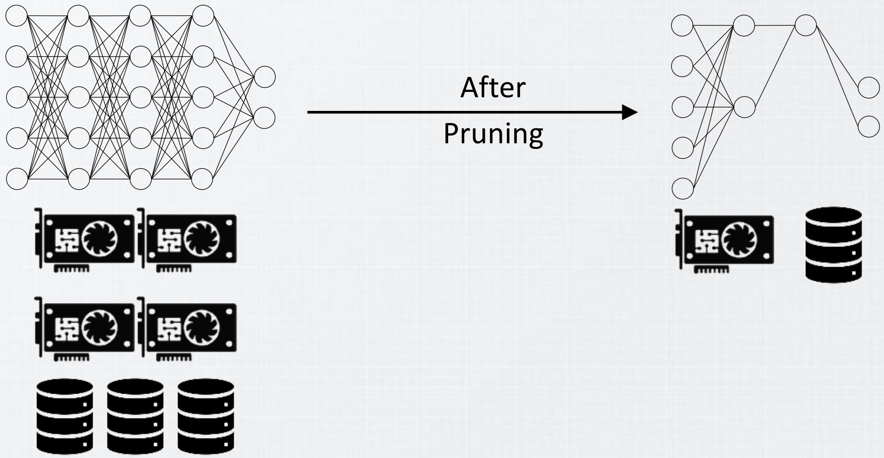
A. Dubey and et al. *The Llama 3 Herd of Models.* ArXiv, 2024

Pruning

Introduction

- Extensive research demonstrated that pruning structures from deep models is a straightforward approach to improving efficiency

- Most works in pruning focus on removing filters (He et al., 2023; Cheng et al. 2024)



He et al. *Structured Pruning for Deep Convolutional Neural Networks: A Survey*. PAMI, 2023
Cheng et al. *A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations*. PAMI, 2024

Layer Pruning

Introduction

- Layer pruning emerges as a promising alternative to standard filter pruning
  - It reduces network depth, which directly addresses model latency
- Despite positive results, most layer-pruning methods fail to preserve accuracy
  - Even when equipped with additional techniques such as knowledge distillation
- Simple criteria are unable to characterize all underlying properties exhibited by layers

Contributions

Introduction

- In this work, we propose a novel pruning criterion that leverages an effective similarity representation metric: **C**entered **K**ernel **A**lignment (CKA)

- Powered by CKA, we develop a layer-pruning method that removes entire layers from neural networks without compromising predictive ability

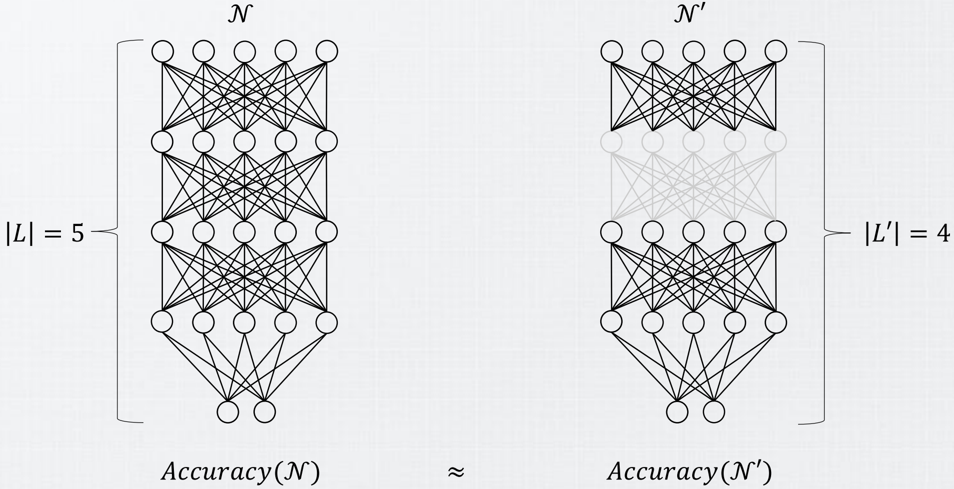
- Our criterion identifies unimportant layers: layers that, when removed, preserve similarity regarding the original model

Problem Statement and Proposed Method

Problem Statement

Methodology

- Given a network $\mathcal{N}$ composed of a layer set L, our goal is to remove certain layers to produce a shallower network $\mathcal{N}'$ composed by L'
  - $L' \ll L$
  - The accuracy of N' is as close as possible (ideally better) than its unpruned version $\mathcal{N}$

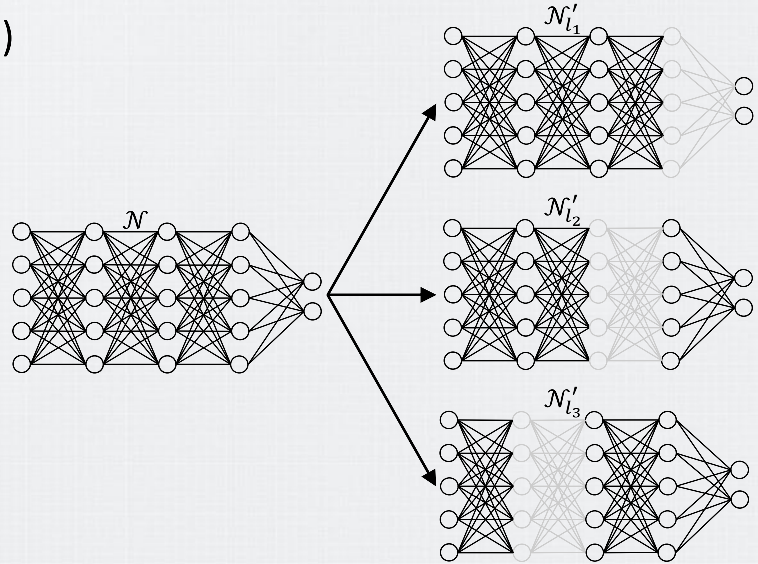


Proposed Method

Methodology

- Given a pre-trained network $\mathcal{N}$, we extract its representation (feature maps) R
  - Here, we use some training input examples

- We create a temporary **pruned** (subnetwork) model by removing a candidate layer $l_i \in L$ from $\mathcal{N}$ and, then, extract its representation R_i
  - n layers, n candidates (subnetworks)

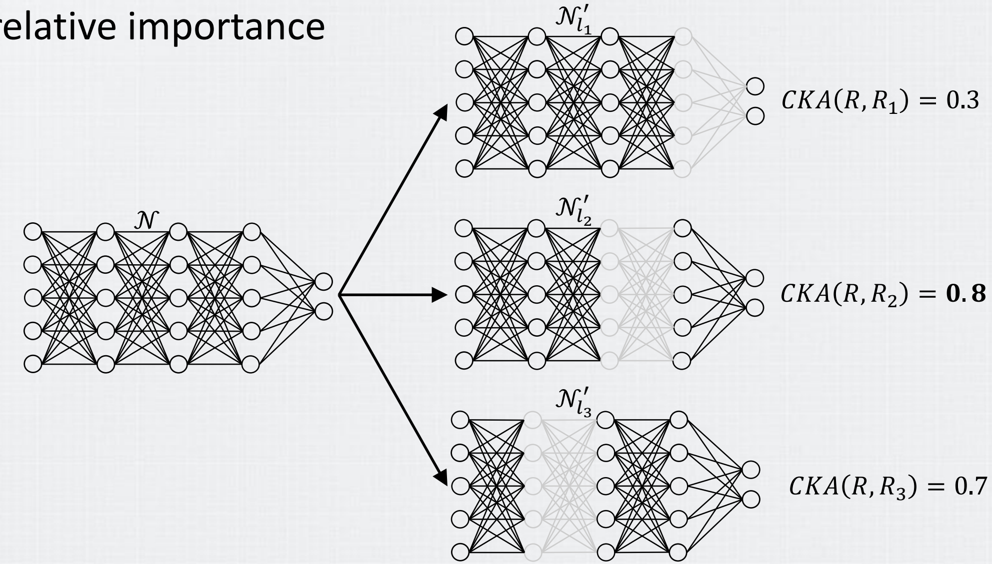


Proposed Method

Methodology

- For each subnetwork, we compare their representations with the original network using CKA

- Finally, we select the temporary network that yields the **highest similarity**
  - Similar representations between a dense (unpruned) network and its optimal sparse (pruned) candidate indicate lower relative importance



Proposed Method

Methodology

- Overview

Layer Pruning using our CKA criterion – $\boldsymbol{O}(|\boldsymbol{L}|)$

Input: Trained Network $\mathcal{N}$, Candidate Layers $l_i \in L$, Training Samples X

Output: Pruned Version of $\mathcal{N}$

$R \leftarrow M(\mathcal{N}, X) \rhd \textit{Extracts representation}$

$\boldsymbol{for}\ i \leftarrow \boldsymbol{to}\ |L|\ \boldsymbol{do}$

$\quad \mathcal{N}_{l_i} \leftarrow \mathcal{N} \backslash l_i \rhd \textit{Removes layer } l_i \textit{ from } \mathcal{N}$

$\quad R_i \leftarrow M\left(\mathcal{N}_{l_i}, X\right) \rhd \textit{Representation extraction of } \mathcal{N}_{l_i}$

$\quad S \leftarrow S\ \cup CKA(R, R_i)$

$\boldsymbol{end\ for}$

$j \leftarrow argmax(S) \rhd \textit{Index (layer) of highest similarity in } S$

$\mathcal{N} \leftarrow \mathcal{N}_{l_i} \rhd \mathcal{N} \textit{ becomes its pruned version}$

Experiments

Experimental Setup

- Datasets
  - CIFAR-10/100
  - ImageNet
  - Different tabular datasets

- Architectures
  - ResNet32/44/56/110
  - ResNet50
  - MobileV2
  - Transformer-like (for tabular data)

Effectiveness of the Proposed CKA Criterion

Experiments

- Our method outperforms existing layer pruning techniques by a large margin
  - The reason for these remarkable results is that our method carefully selects which layers to eliminate

- Additionally, our method is more cost-friendly

	Method	Δ $Acc.$	FLOPS (%)
ResNet56 on CIFAR10	PLS	(-) 0.98	30.00
	FRPP	(+) 0.26	34.80
	ESNB	(-) 0.62	52.60
	LPSR	(+) 0.19	56.29
	CKA (ours)	**(+) 0.95**	**56.29**

	Method	Δ $Acc.$	FLOPS (%)
ResNet110 on CIFAR10	ESNB	(+) 1.15	29.89
	PLS	(+) 0.06	37.73
	CKA (ours)	**(+) 1.16**	**50.33**
ResNet50 on ImageNet	LPRS	(-) 1.38	37.38
	CKA (ours)	**(-) 0.18**	39.62
	PLS	(-) 0.67	45.28
	CKA (ours)	(-) 0.90	**45.28**

Comparison with the State of the Art

Experiments

- We evaluate our method against the most recent and top-performing techniques mainly based on the survey by He et al. (2023)
  - We report the results of each method according to the original paper

	Method	Δ $Acc.$	FLOPS (%)
ResNet56 on CIFAR10	DECORE	(+) 0.08	26.30
	SOKS	(+) 0.16	35.91
	CKA (Ours)	**(+) 1.25**	37.52
	WhiteBox	(+) 0.28	55.60
	CLR-RNF	(+) 0.01	57.30
	CKA (Ours)	(+) 0.78	60.04
	Hrank	(-) 2.54	74.09
	GCNP	(-) 0.97	77.22
	CKA (Ours)	(-) 0.66	**78.80**

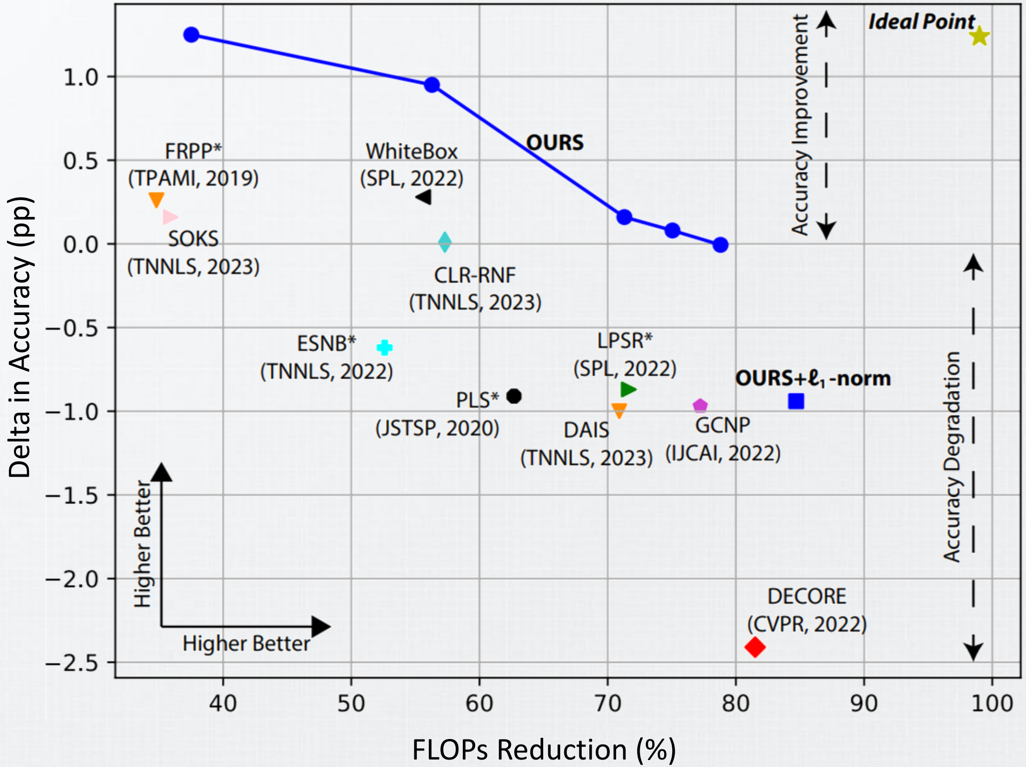
	Method	Δ $Acc.$	FLOPS (%)
ResNet110 on CIFAR10	WhiteBox	**(+) 0.62**	66.00
	CRL-RNF	(+) 0.14	66.00
	DECORE	(-) 0.79	76.83
	CKA (Ours)	(+) 0.23	**76.42**
ResNet50 on ImageNet	WhiteBox	**(-) 0.83**	**45.60**
	CLR-RNF	(-) 1.16	40.39
	SOSP	(-) 0.94	45.00
	DECORE	(-) 1.57	42.30
	CKA (Ours)	(-) 0.90	45.28

He et al. *Structured Pruning for Deep Convolutional Neural Networks: A Survey*. PAMI, 2023

Comparison with the State of the Art

Experiments

- Our method dominates existing pruning methods by a remarkable margin



Effectiveness in Shallow Architectures

Experiments

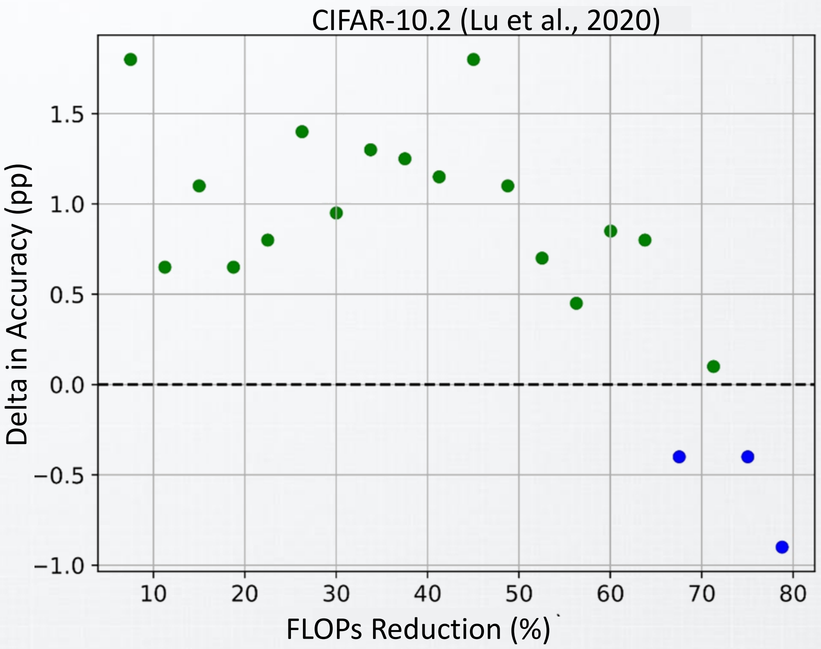
- The keys to the success of pruning is the overparameterized regime of neural networks, particularly evident in deep models

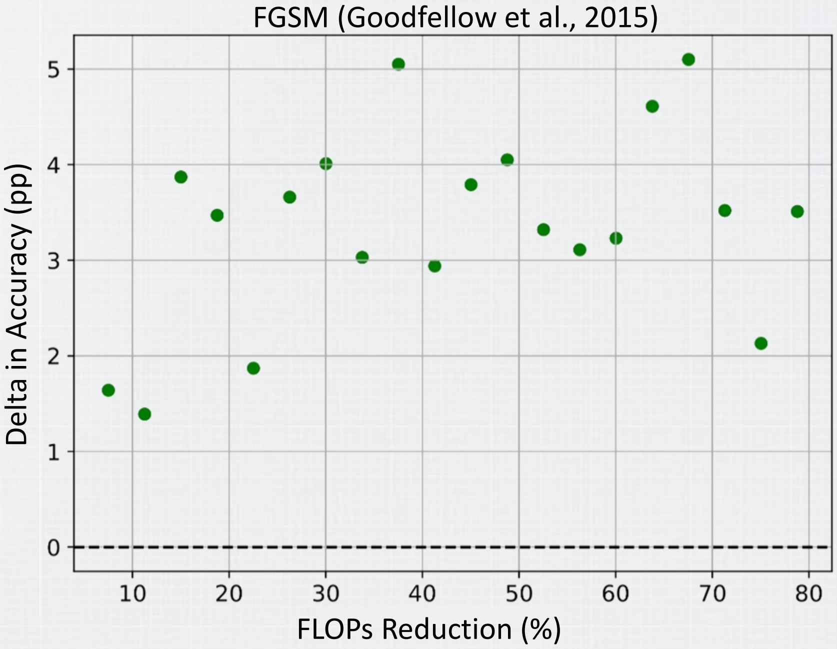
- This experiment verifies the applicability of our method to shallow models
  - ResNet32/44 and MobileV2 (please check our paper)

	Method	$\Delta\ Acc.$	FLOPS (%)
ResNet32	DAIS	**(+) 0.57**	53.90
	SOKS	(-) 0.80	54.58
	CKA (Ours)	(+) 0.05	**54.61**
ResNet44	DCP-CAC	(-) 0.03	50.04
	AGMC	(-) 0.82	50.00
	CKA (Ours)	**(+) 0.47**	**53.27**

Robustness to Adversarial Samples

Experiments



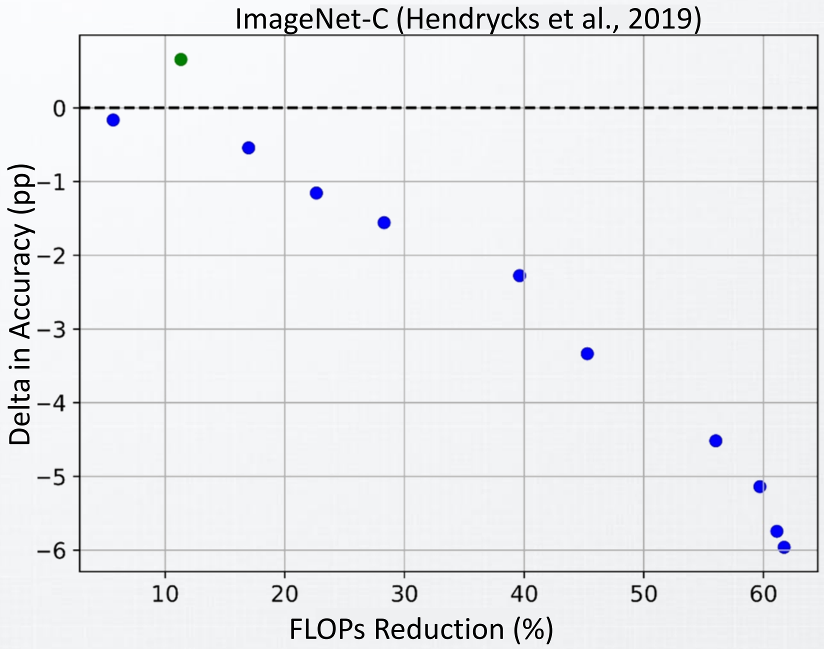


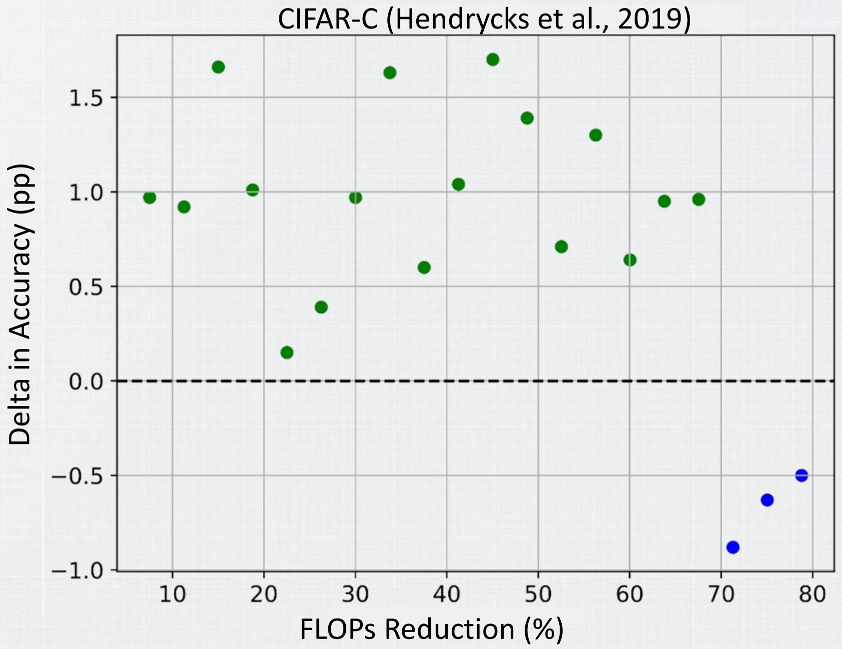
Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR, 2015

Lu et al. *Harder or Different? A Closer Look at Distribution Shift in Dataset Reproduction*. ICML, 2020

Robustness to Adversarial Samples

Experiments





Hendrycks et al. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. ICLR, 2019

Conclusions

Conclusions

- We proposed a novel criterion for identifying and removing unimportant layers
  - Our criterion leverages the Centered Kernel Alignment (CKA) to select unimportant layers from a set of candidates
  - Our criterion capture underlying properties exhibited by layers and preventing model collapse
- Powered by CKA, we showed that similar representations between a dense (unpruned) network and its optimal pruning candidate indicate lower relative importance

Conclusions

- Unlike most existing layer-pruning criteria that fail to capture underlying properties of layers, our method effectively assigns layer importance and thus prevents model collapse

- We believe our results open new opportunities to prune through the lens of similarities metrics and encourage further efforts on layer pruning

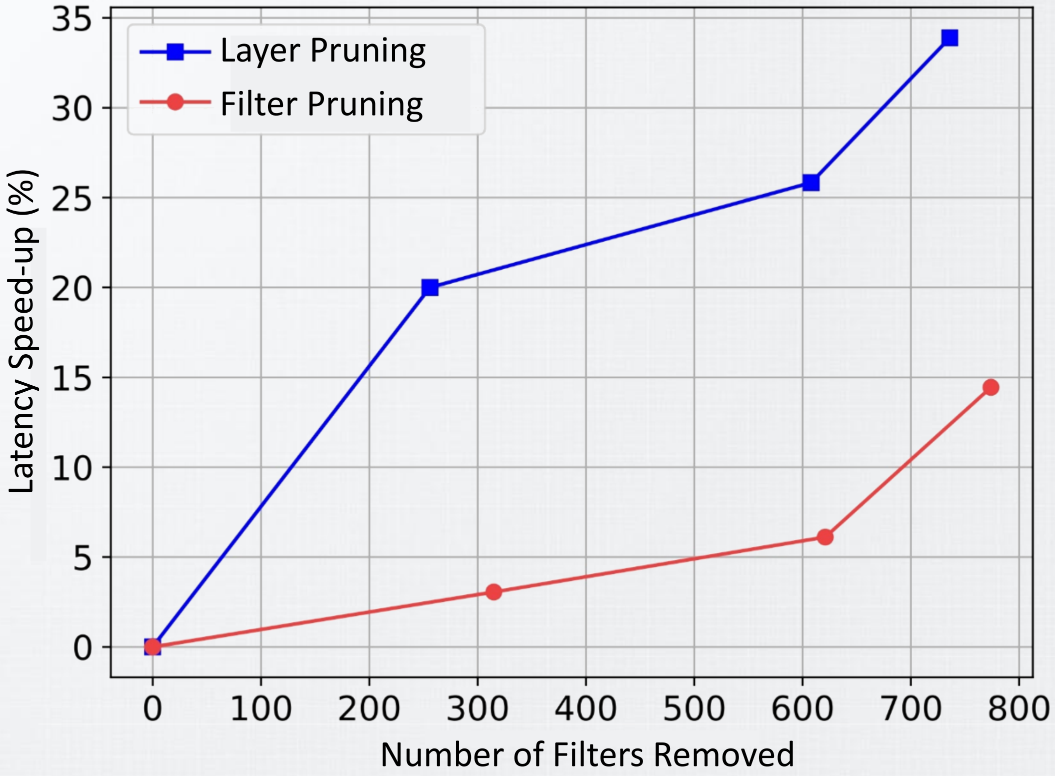
Acknowledgements

- The authors would like to thank grant #2023/11163-0, São Paulo Research Foundation (FAPESP), and grant #402734/2023-8, National Council for Scientific and Technological Development (CNPq). Artur Jordao Lima Correia would like to thank Edital Programa de Apoio a Novos Docentes 2023. Processo USP nº: 22.1.09345.01.2. Anna H. Reali Costa would like to thank grant #312360/2023-1 CNPq. This study was also partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001

Appendix

The Effect of Layer Pruning on Efficiency

Appendix



Technical Details Behind Layer Pruning

Appendix