Maximizing Model Usability in Industry through Pruning: An Essential Optimization Technique

JESSICA LÓPEZ ESPEJEL

Novelis

Summary

I. Introduction
II. Theoretical Foundations of Pruning
III. Pruning in Practice: Real-world Examples
IV. Benefits for Industry
V. Future Trends and Directions

Introduction

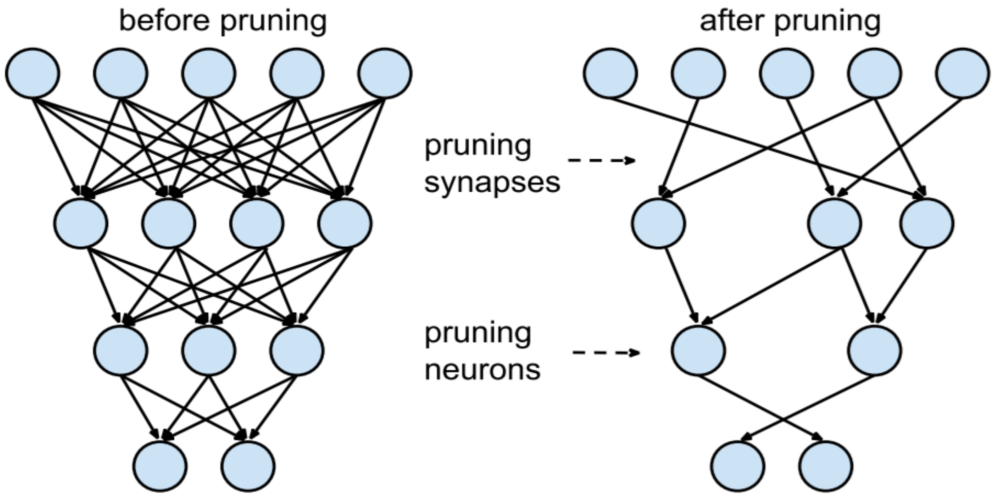
Introduction

What is Pruning?

Remove model weight values that are near or equal to zero.

It reduces the model size and achieves competitive or even better results than the original model.

It can reduce the parameters count by more than 90%.



Optimal Brain damage [LeCun et al., NeurIPS 1989]
Learning both Weights and Connections for Efficient Neural Nertwork [Han et al., NeurIPS]

Introduction

Pruning can be done in different stages:

1. Full model training
2. PEFT / LoRA
3. Post- training

Introduction

Neural Network	# Parameters		
	Before Pruning	After Pruning	Reduction
AlexNet	61 M	6,7 M	9 X
Google Net	7 M	2,0 M	3,5 X
ResNet50	26 M	7,47 M	3,4 X

Efficient Methods and Hardware for Deep Learning [Han S., Standford University]

Motivation

More Memory → More Energy

Theoretical Foundations of Pruning

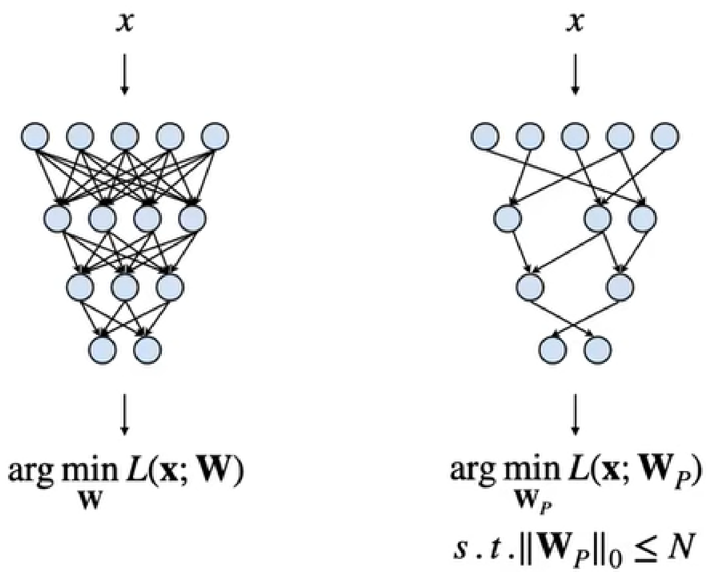
Mathematical Understanding

We can formulate the pruning as follows:

$\arg\min_{W_p} L(x; W_P)$

Subject to $\|W_p\|_0 > N$

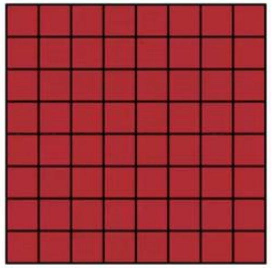
- **L** represents the objective function for neural network training
- **x** in an input, **W** is original weights, $\boldsymbol{W_P}$ is pruned weights
- $\|\boldsymbol{W_p}\|_{\boldsymbol{0}}$ calculates the #nonzeros in W_P, and **N** is the target #nonzeros



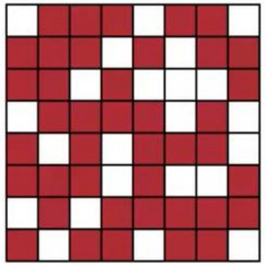
Slide by MIT 6.5940, Fall 2023

Pruning Granularity

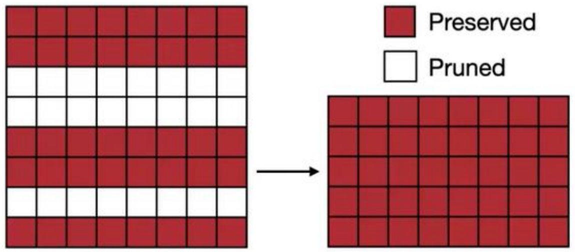
Pruning can be performed at different granularities, from structured to non-structured.



Normal Matrix



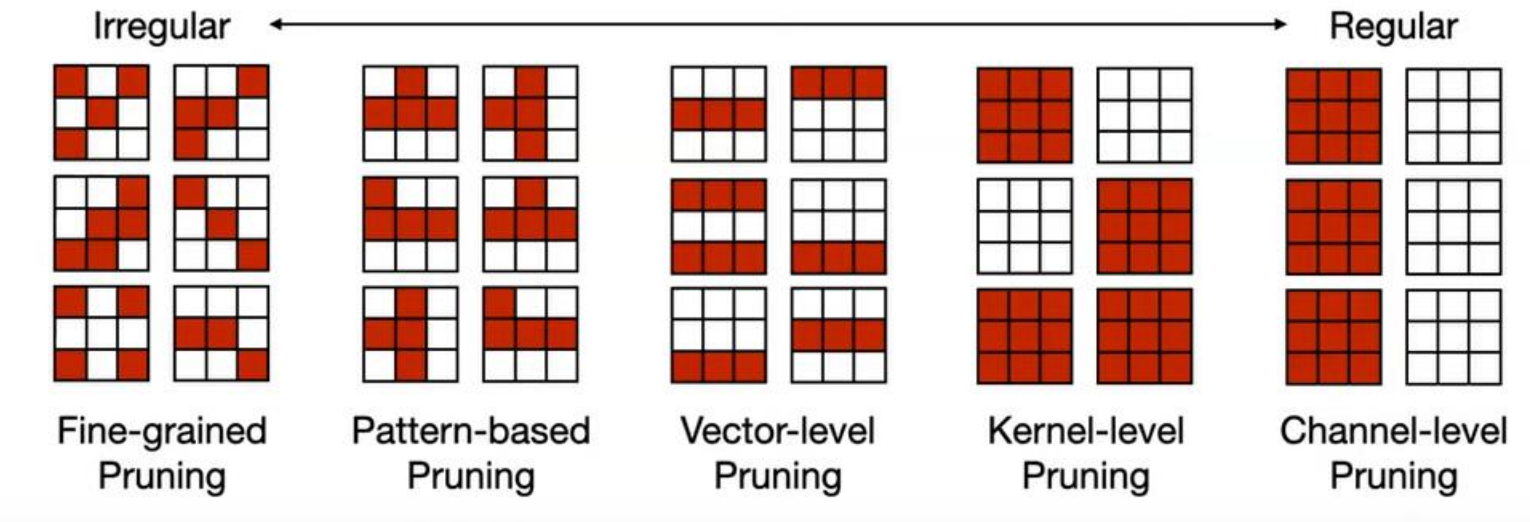
Fine-grained/
Unstructured



Coarse-grained/ Structured

Pruning Granularity

The case of convolutional layers

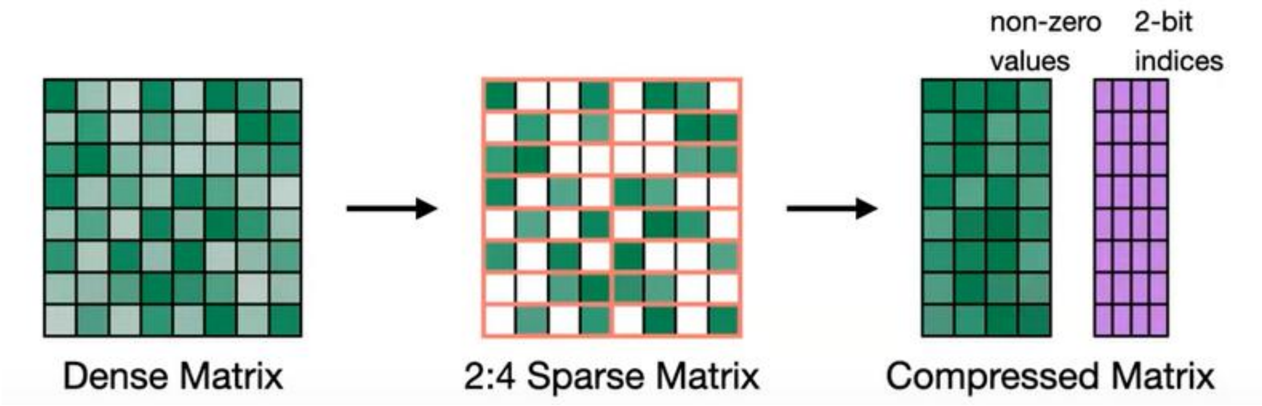


Exploring the granularity of sparcity in convolutional neural networks [Mao et al., CVPR-W]
Slide by MIT 6.5940, Fall 2023

Pruning Granularity

Pattern-based Pruning: N:M

- N:M sparsity means that in each contiguous M elements, N of them are pruned

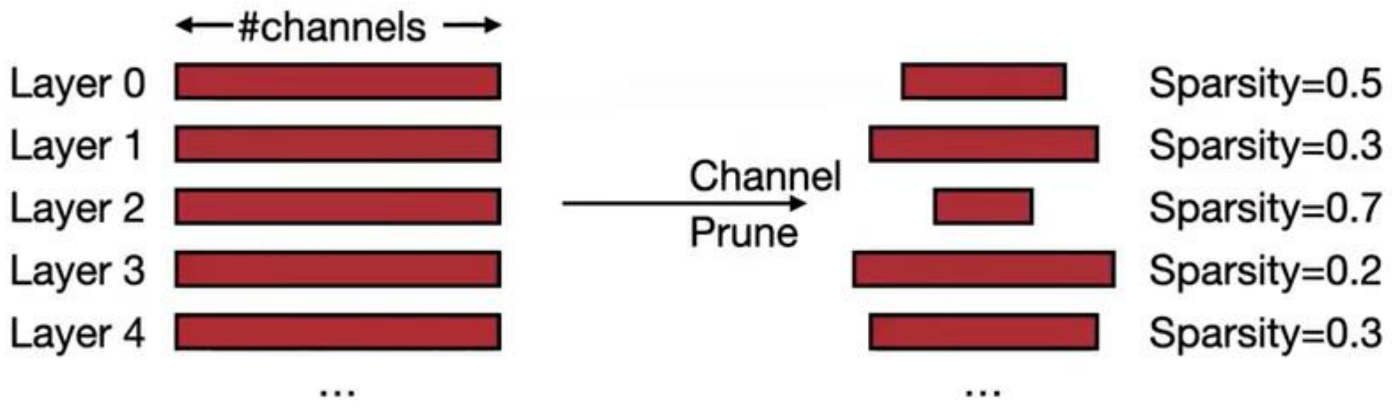


Accelerating Inference with Sparsity Using NVIDIA Ampere Architecture and NVIDIA TEnsorRT
Slide by MIT 6.5940, Fall 2023

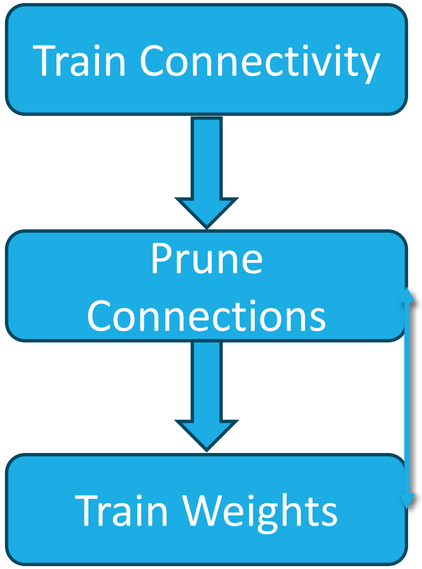
Pruning Granularity

Channel Pruning

- Pro: Direct speed up due to reduced channel numbers
- Con: smaller compression ratio

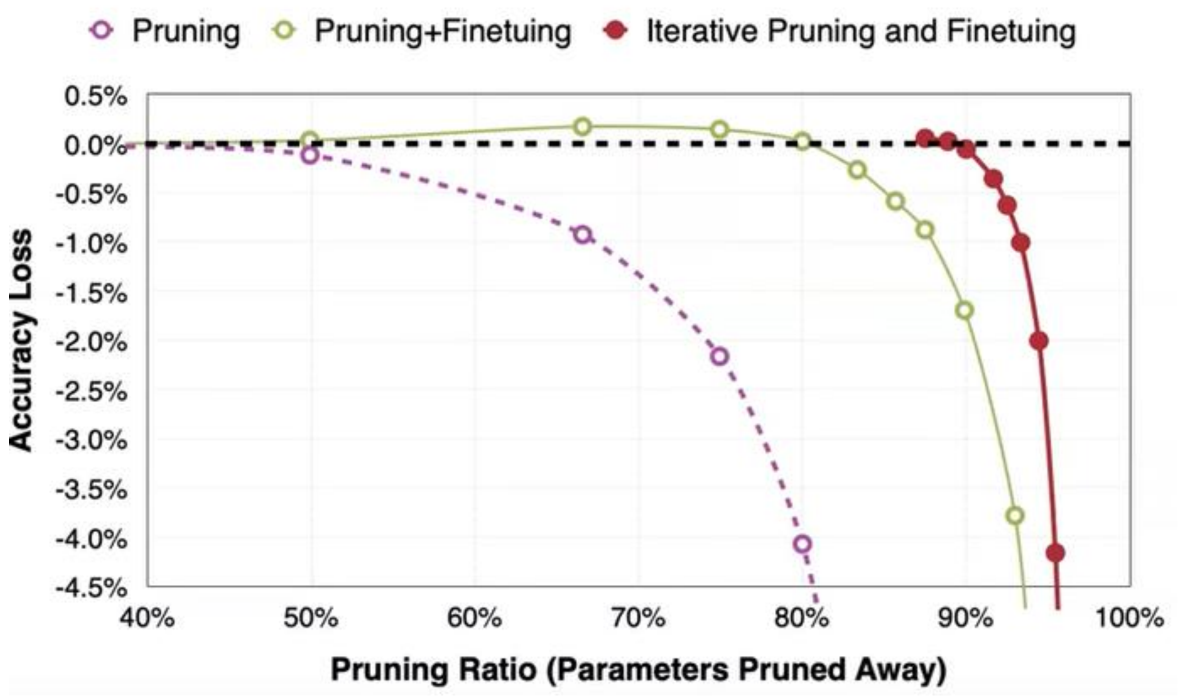


Full Model Training



Learning both Weights and Connections for Efficient Neural Nertwork [Han et al., NeurIPS]
Inspired by EfficientML.ai Lecture 3 - Pruning and Sparsity [MIT, 2023]

Pruning + Finetuning



Learning both Weights and Connections for Efficient Neural Nertwork [Han et al., NeurIPS]
Inspired by EfficientML.ai Lecture 3 - Pruning and Sparsity [MIT, 2023]

Pruning in Practice: Real-world Examples

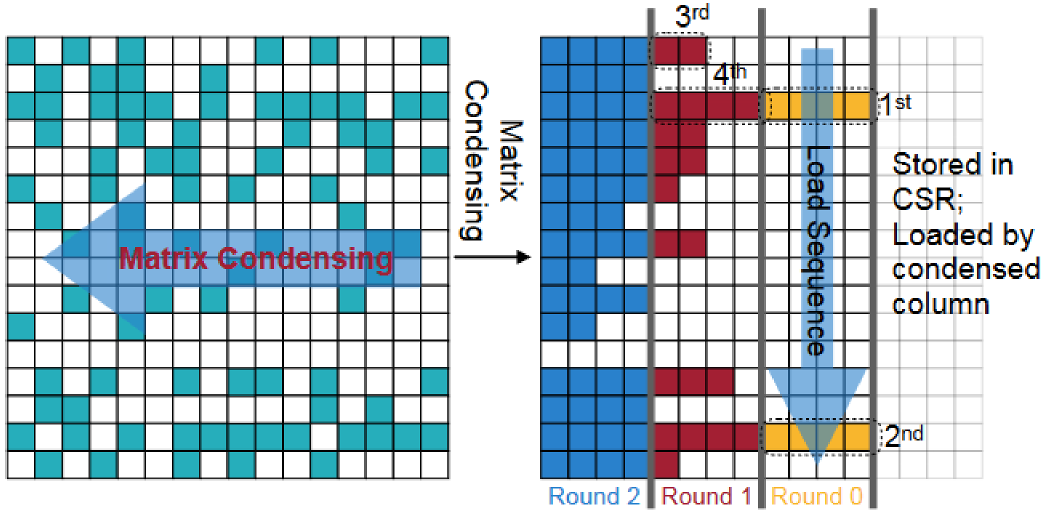
Pruning in the Industry

The 26th IEEE International Symposium on High-Performance Computer Architecture (HPCA 2020)

SpArch: Efficient Architecture for Sparse Matrix Multiplication

Zhekai Zhang*, Hanrui Wang*, Song Han
EECS
Massachusetts Institute of Technology
Cambridge, MA, US
{zhangzk, hanrui, songhan}@mit.edu

William J. Dally
Electrical Engineering
Stanford University / NVIDIA
Stanford, CA, US
dally@stanford.edu



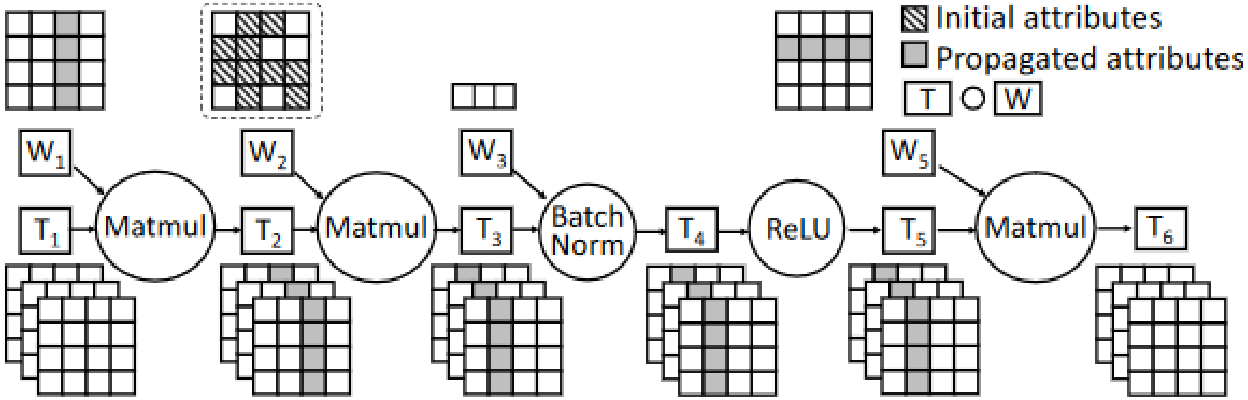
Produces 16 partial matrices

Produces 12 partial matrices

Matrix Condensing. Condense the sparse matrix to the left, reducing the number of columns, thus reducing the number of partial matrices. It can be stored naturally using the CSR format [Zhang et al., 2020]

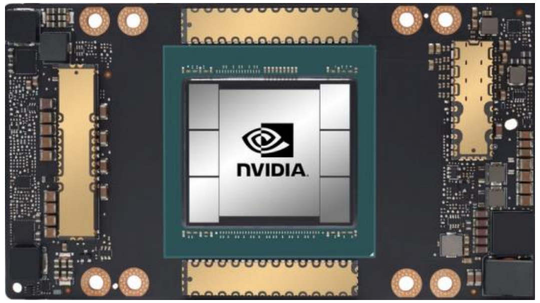
Pruning in the Industry

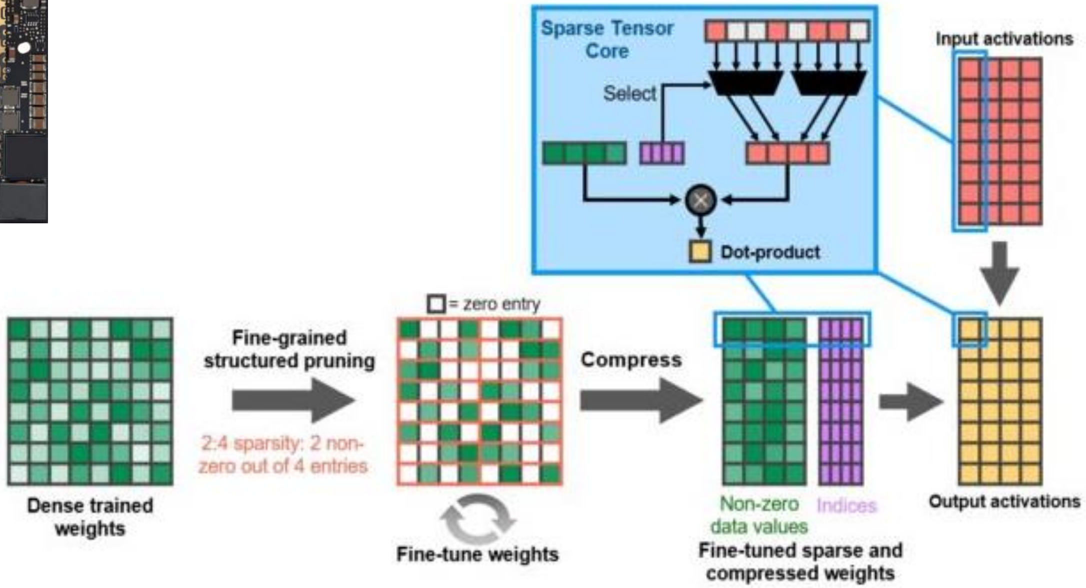
Microsoft



The sparsity attribute of one tensor can be propagated along the deep learning network [Zhen et al., 2022].

Pruning in the Industry





Structure is enforced, through a new 2:4 **sparse matrix** definition that **allows two non-zero** values in every four-entry vector. A100 supports 2:4 structured sparsity on rows.

Benefits for Industry

Benefits for Industry

✓ Improved Efficiency

- It reduces the number of parameters in a neural network, making it more computationally efficient.
- The industries can rely on real-time or resource-constrained applications.

✓ Reduced Memory Footprint

- Smaller models require less memory, making them suitable for edge devices.

✓ Faster Inference

- It allows industries to process data more quickly.

✓ Energy Savings

- Pruning can result in lower energy consumption.

Benefits for Industry

- ✓ Scalability
  - Its easier to scale neural networks for larger datasets and more complex tasks.
  - It is adaptable for growing industries and evolving applications.
- ✓ Lower Training Time
  - The training is faster
- ✓ Reduced Model Deployment Costs
  - It can lead to cost savings in industries where data transmission and storage costs are significant.
- ✓ Regulatory Compliance
  - Pruning can help industries meet regulatory requirements related to data privacy and model explainability.

Future Trends and Directions

Future Trends and Directions

✓ Structured Pruning

- Entire neurons, channels, or blocks of a neural network are pruned together.

✓ Dynamic Pruning

- They aim to adapt the network structure **during training or inference based on the input data distribution** are expected to become more prevalent, **improving model adaptability and accuracy**.

✓ Automated Pruning

- The development of automated pruning algorithms, possibly driven by reinforcement learning or evolutionary approaches, will **simplify the process of selecting and implementing the most effective pruning strategies.**

Future Trends and Directions

- ✓ Hardware-aware Pruning
  - Pruning techniques will become more tailored to specific architectures, ensuring optimal performance on various platforms, such as GPUs, TPUs, and custom AI accelerators.
- ✓ Transfer Learning with Pruning
  - Explore how pre-trained models can be pruned and fine-tuned more effectively to reduce the environmental footprint and improve task-specific performance.
- ✓ Cross-domain Pruning
  - Applying pruning techniques developed in one domain or application area to other domains, possibly using domain adaptation methods.

Thank you!

Jessica López Espejel
Linkedin: https://www.linkedin.com/in/jessicalopezespejel/