Tiny Titans: Efficient Large Vision, Language and Multimodal Models through Pruning

Carolina Tavares[†], Leandro Mugnaini[†], Gustavo Henrique do Nascimento[†], Ian Pons[†], Keith Ogawa[†], Guilherme Stern[†], Lucas Libanio[†], Aline Paes[‡], Anna Helena Reali Costa[†] and Artur Jordao[†]
[‡]Universidade Federal Fluminense, [†]Universidade de Sao Paulo, Brazil

SIBGRAPI 2025

FAPESP CAPES CNPq < c²d_ centro de ciência de dados >

Preliminary Information

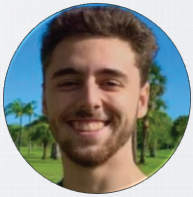
Instructors and Speakers

Preliminary Information

Carolina Tavares

Leandro Mugnaini

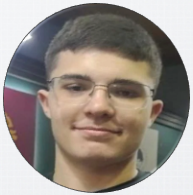
Gustavo Henrique do Nascimento



Ian Pons

Keith Ogawa

Guilherme Stern



Lucas Libanio

Aline Paes

Anna Helena Reali Costa

Artur Jordao

Agenda

Preliminary Information

- Part 1
  - Introduction and basics of Pruning
- Part 2
  - Hands-on session: pruning computer vision models
- Part 3
  - Hands-on session: pruning large language models

Tutorial Materials for Participants

Preliminary Information

- GitHub repository
  - **https://github.com/arturjordao/TinyTitans**
  - Scan the QRCode

- The materials include:
  - Paper and slides
  - Source codes (Python scripts and Jupyter notebooks)

Introduction

Advances in Deep Learning

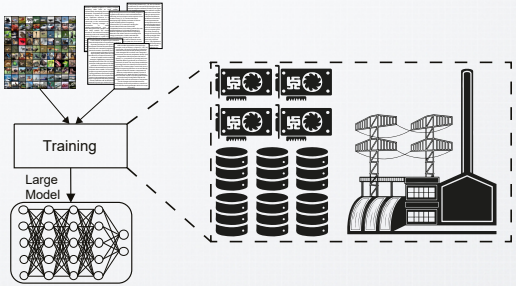
Introduction

- Visual Models
  - Video generation
- Language Models
  - Natural Language Processing
- Large Vision-Language Models (LVLMs)
  - Image and text understanding

Computational Demand

Introduction

- There is a call for efficient general-purpose AI models
  - There is a growing demand for energy to power AI workloads[1]

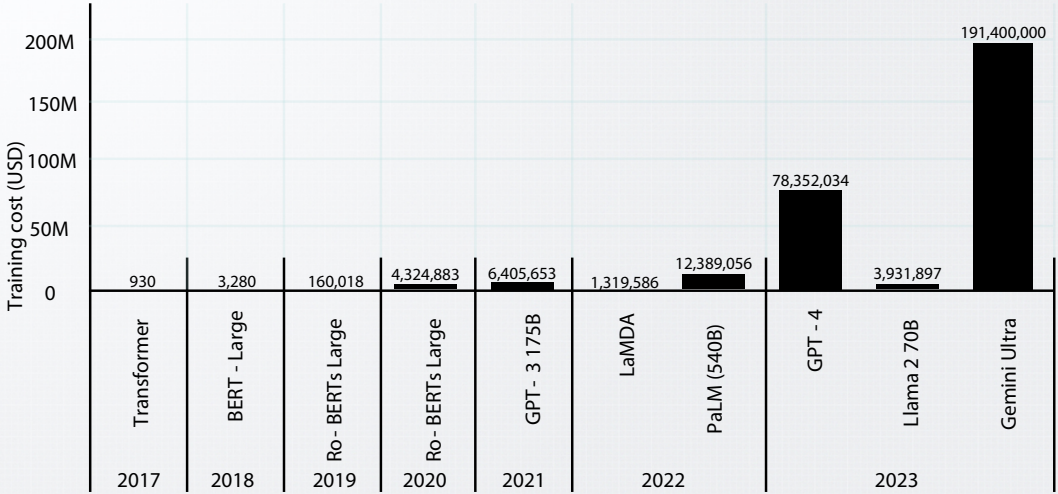


Model (#Params)	Carbon Emissions (tCO_2eq)	Equivalent to
$< 1B$	6	675 gal. of gasoline
1B	36	$40 \times$ NY $\leftrightarrow$ SF
7B	65	150 oil barrels

[1] Morrison et al. *Holistically Evaluating the Environmental Impact of Creating Language Models.* ICLR, 2025

Financial Cost

Introduction

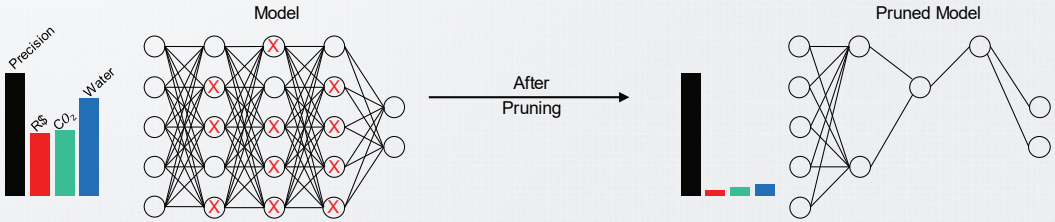


Compression through Pruning

Introduction

- Among existing solutions to overcome the previous issues, pruning emerges as an effective and hardware-agnostic technique for improving model efficiency

- Pruning involves removing structures from the architecture (model)
  - Weights, filters or even entire layers



Basics of Pruning

General Pruning Algorithm

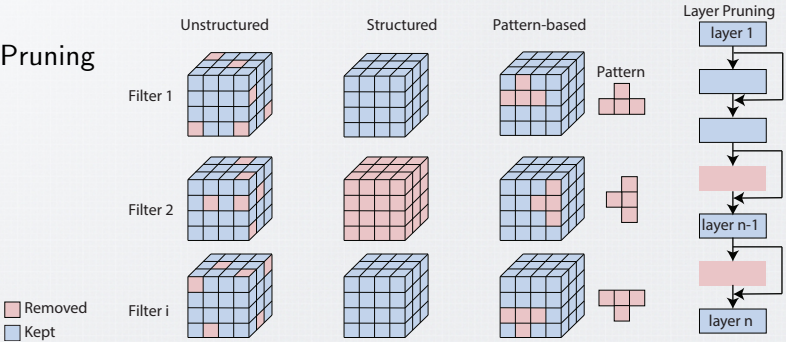
Basics of Pruning

Input: Network $\mathcal{F}$, Pruning Criterion c, Pruning Ratio p
Input: Number of Iterations K
Output: Pruned Network $\mathcal{F}'$ (a.k.a Subnetwork)
for $k \leftarrow 1$ **to** K **do**
  $S \leftarrow c(\mathcal{F})$ ▷ Set importance for each structure
  $I \leftarrow p\%$ Unimportant structures based on S
  $\mathcal{F}' \leftarrow \mathcal{F} \setminus I$ ▷ Removes the structures indexed by I
  Update $\mathcal{F}'$
  $\mathcal{F} \leftarrow \mathcal{F}'$
end for

Forms of Pruning

Basics of Pruning

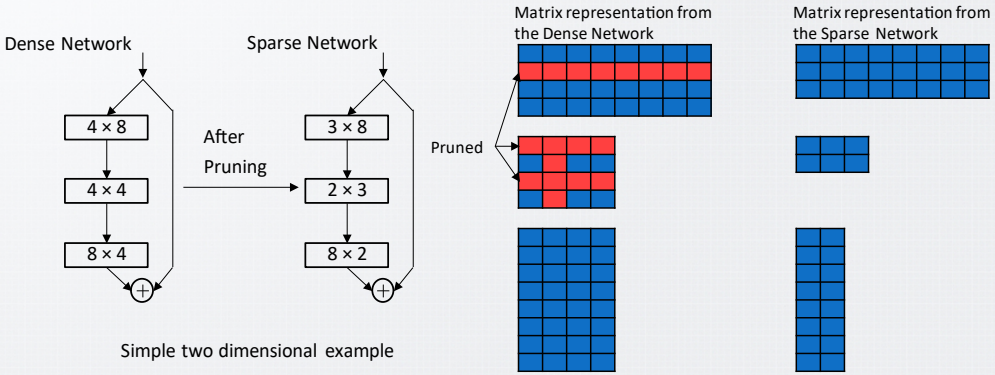
- Unstructured Pruning
- Structured Pruning
- Pattern-based Pruning



Technical Details behind Pruning

Basics of Pruning

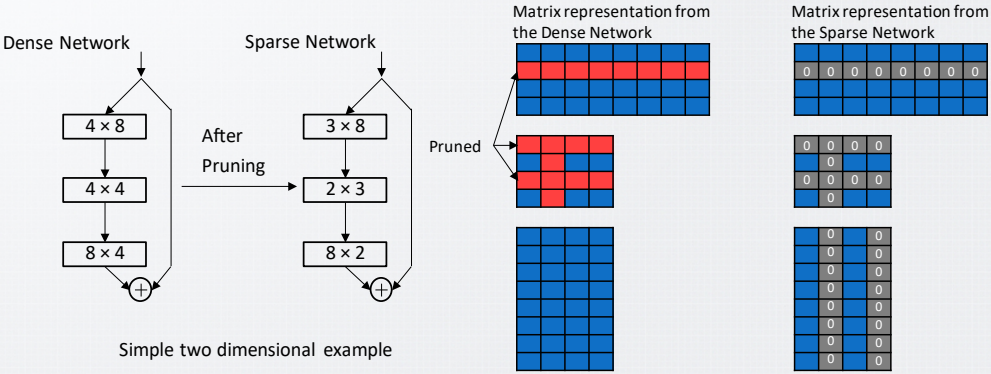
- Tensor multiplication
  - The weight tensors of the ResNet50 and Transformer architectures are four- and three-dimensional (TensorFlow)



Technical Details behind Pruning

Basics of Pruning

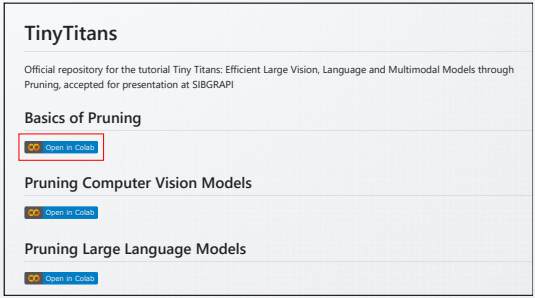
- The zeroed-out trick
  - Sets zero (0) to the pruned weights. This strategy only provides **practical** benefits on specialized frameworks for sparse computations



Technical Details behind Pruning

Basics of Pruning

- Follow along in Colab
  - https://github.com/arturjordao/TinyTitans
  - Open Colab: **Basics of Pruning**



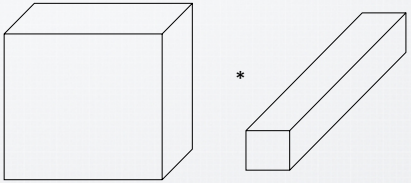
Technical Details behind Pruning

Basics of Pruning

- Metrics
  - Floating Points Operations (FLOPs)
  - Number of parameters
  - Latency (inference time)

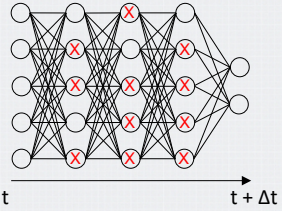
FLOPs

$224 \times 224 \times 3 \times 3 \times 3 \times 512 \Rightarrow 924{,}844{,}032$



Latency

Time for predicting a sample, generating an image, etc.



Parameter Adjustments after Pruning

- Takeaways
  - Essential for predictive performance recovery after pruning
  - Effectiveness depends on training recipe (e.g., optimizer, lr scheduler, etc.)
  - Knowledge distillation is an alternative to full fine-tuning
  - Introduces additional costs

- In this tutorial, we exclude the fine-tuning step due to time and resource constraints

GreenAI Metrics

Basics of Pruning

- GreenAI-Related metrics
  - Carbon footprint
  - Water consumption
  - USD

- How to measure the carbon footprint in Deep Learning?
  - CO_2 Emission = Energy Consumption $\times$ Carbon Energy Efficiency
  - Energy Consumption (kWh): $W * hours/1000$
  - Carbon Energy Efficiency = CO_2 emissions per kWh (Default value 0.432)

GreenAI Metrics

Basics of Pruning

- User-friendly toolkit (deep learning-specific[1])
  - https://mlco2.github.io/impact/
- Other user-friendly toolkits
  - (LLM-specific[2]) https://github.com/SotaroKaneda/MLCarbon
  - (General Purpose[3]) http://calculator.green-algorithms.org/

Deep learning-specific

LLM-specific

General Purpose

[1] Lacoste et al. *Quantifying the carbon emissions of machine learning*. NeurIPS, 2019

[2] Faiz et al. *LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models*. ICLR, 2024

[3] Lannelongu et al. *Green Algorithms: Quantifying the Carbon Footprint of Computation*. Advanced Science, 2021

Popular Benchmarks for Pruning across Domains

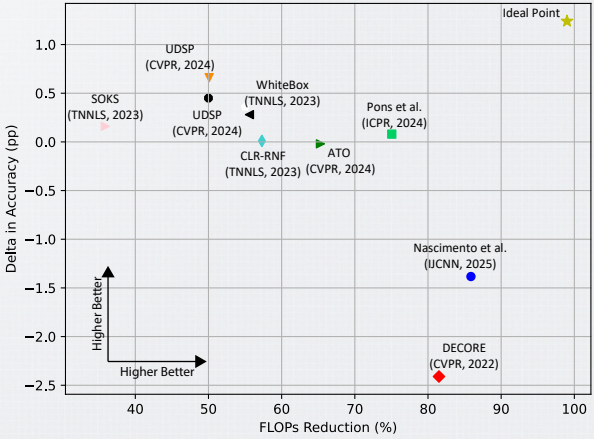
Basics of Pruning

- Computer vision benchmarks
  - CIFAR-10, ImageNet
  - PASCAL VOC, nuScenes, StreamPETR
- Natural Language benchmarks
  - WinoGrande, HellaSwag, ARC-e/ARC-c and PIQA
- Multimodal (vision and language) benchmarks
  - ScienceQA, Vizwiz, MMVet, LLaVABench

Pruning for Computer Vision

Basics of Pruning

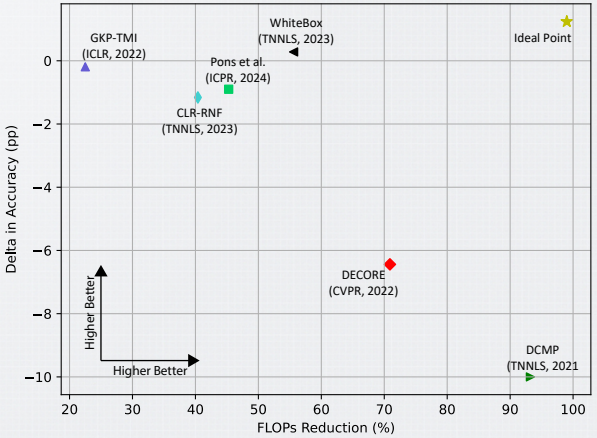
- ResNet56 on CIFAR-10



Pruning for Computer Vision

Basics of Pruning

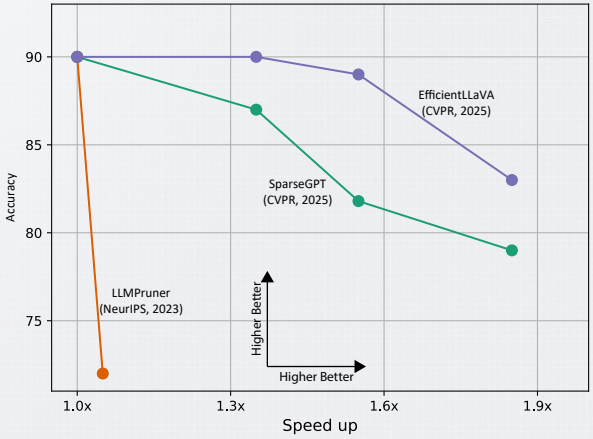
- ResNet50 on ImageNet



Pruning for Computer Vision

Basics of Pruning

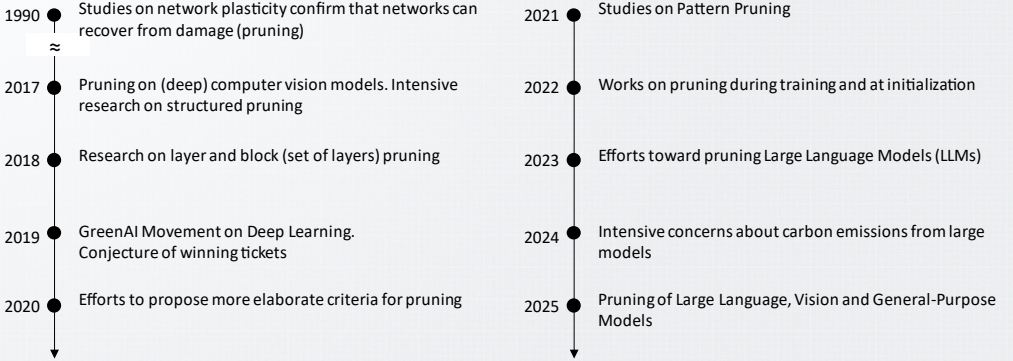
- LLaVA-1.5 on ScienceQA



Historical Trends

Basics of Pruning

- Noteworthy progress and seminal research in pruning

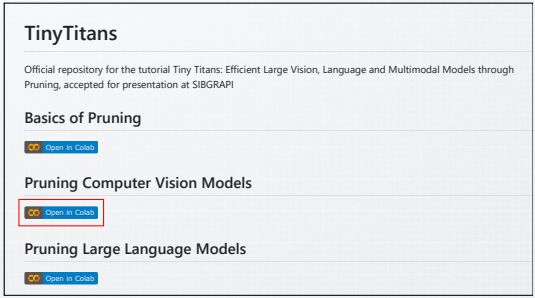


Hands-On
Pruning Computer Vision Models

Hands-On

Hands-On Pruning Computer Vision Models

- Follow along in Colab
  - https://github.com/arturjordao/TinyTitans
  - Open Colab: **Pruning Computer Vision Models**



Hands-On
Pruning Large Language Models

Hands-On

Hands-On Pruning Large Language Models

- Follow along in Colab
  - https://github.com/arturjordao/TinyTitans
  - Open Colab: **Pruning Large Language Models**