Mechanistic Interpretability for Vision Models Optimization

SAPIENZA
UNIVERSITÀ DI ROMA

Course: **Computer Vision**
Candidate: Alessio Olivieri
ID: 1973323
Date: 19/06/25

Outline:

→ Analyze dataset: hand recognize labels

→ Decide sub-task for ACDC: classify animal categories

→ Use ACDC to prune the ViT

→ Train pruned ViT

→ Train Baseline: classify animal species

→ Benchmarks

Problem Statement:

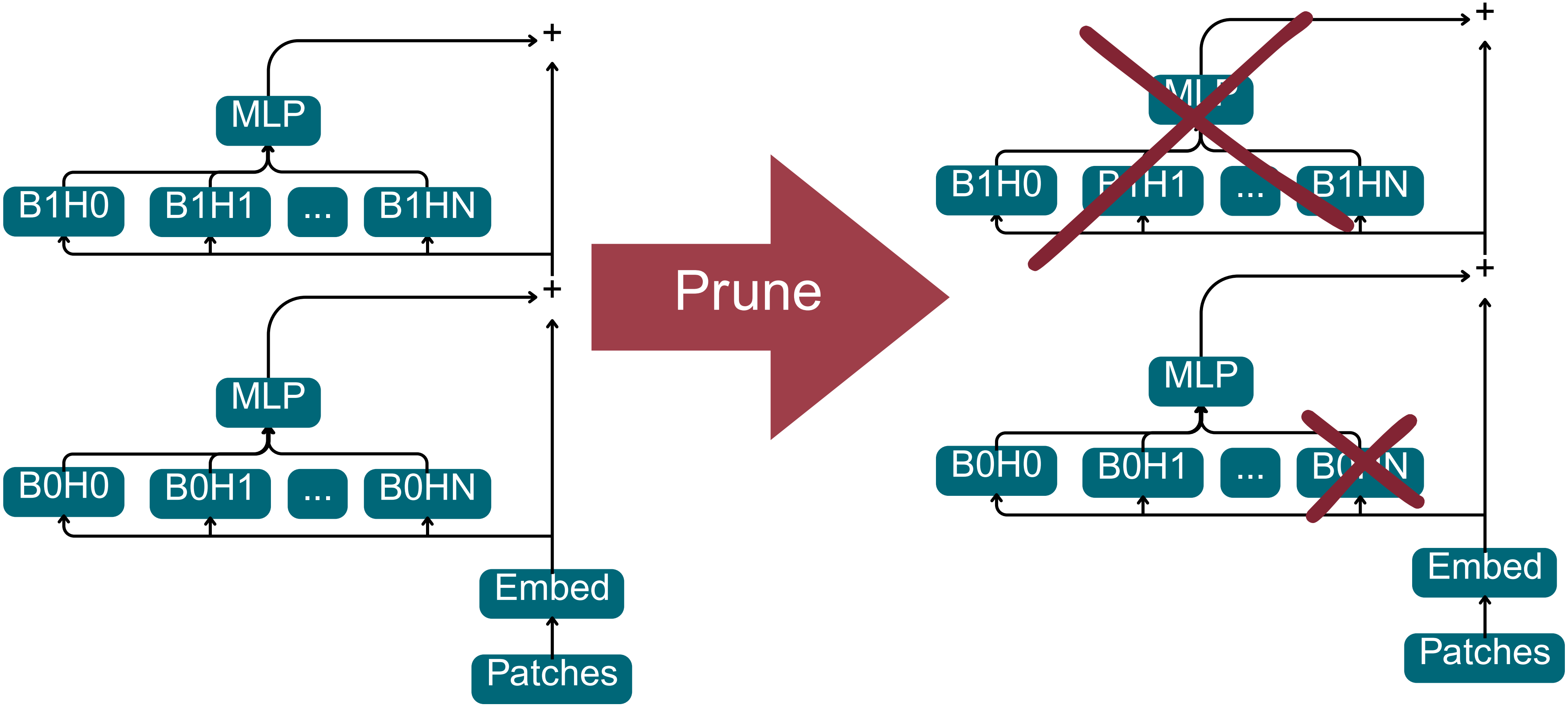
Vision Trasformers are very powerfull !

But they are incredibly computationally expensive...

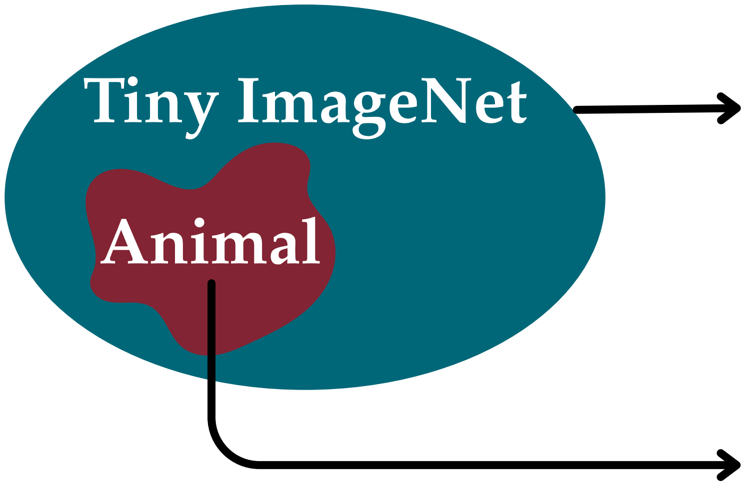
Especially for edge devices

Proposed method:

By pruning a ViT trained on a more complicated task
We can obtain a more efficient ViT



Dataset:



- 100.000 64x64 images
- 200 classes

- 58 animal classes

ACDC TASK: 6 coarse animal classes

Acquatic | Reptiles | Arthropods

Birds | Mammals | Marine Life

Experimental Setup (1)

Train the ViT to recognize the 58 fine classes

ViT architecture:

- Patch size: 8
- Hidden size: 64
- Blocks: 6
- Attention heads: 8

Data transformations:

- RandAugment
- Horizontal flip
- Random erasing: 25%

Trainer parameters:

- Dropout: 0.2
- Criterion: Soft CrossEntropy
- Optimizer: AdamW
- Mixup + CutMix
- LR: 3e-4
- WarmUp: 20 epochs
- Cosine annealing
- Patience: 50 epochs

Experimental Setup (2)

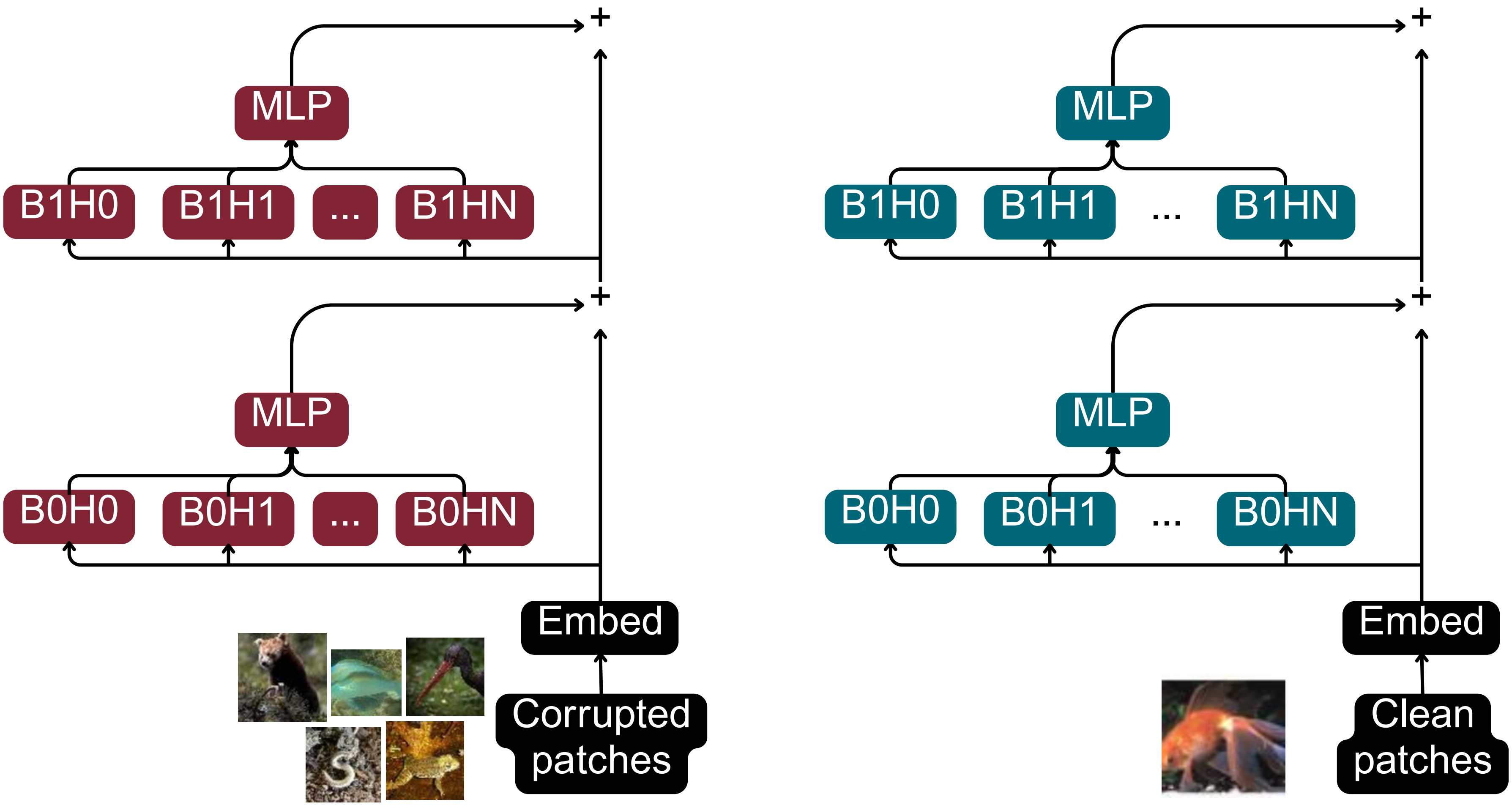
ACDC (1):

For each sample Get one "bad" sample (negative) from each of the other classes

Experimental Setup (2)

ACDC (2):

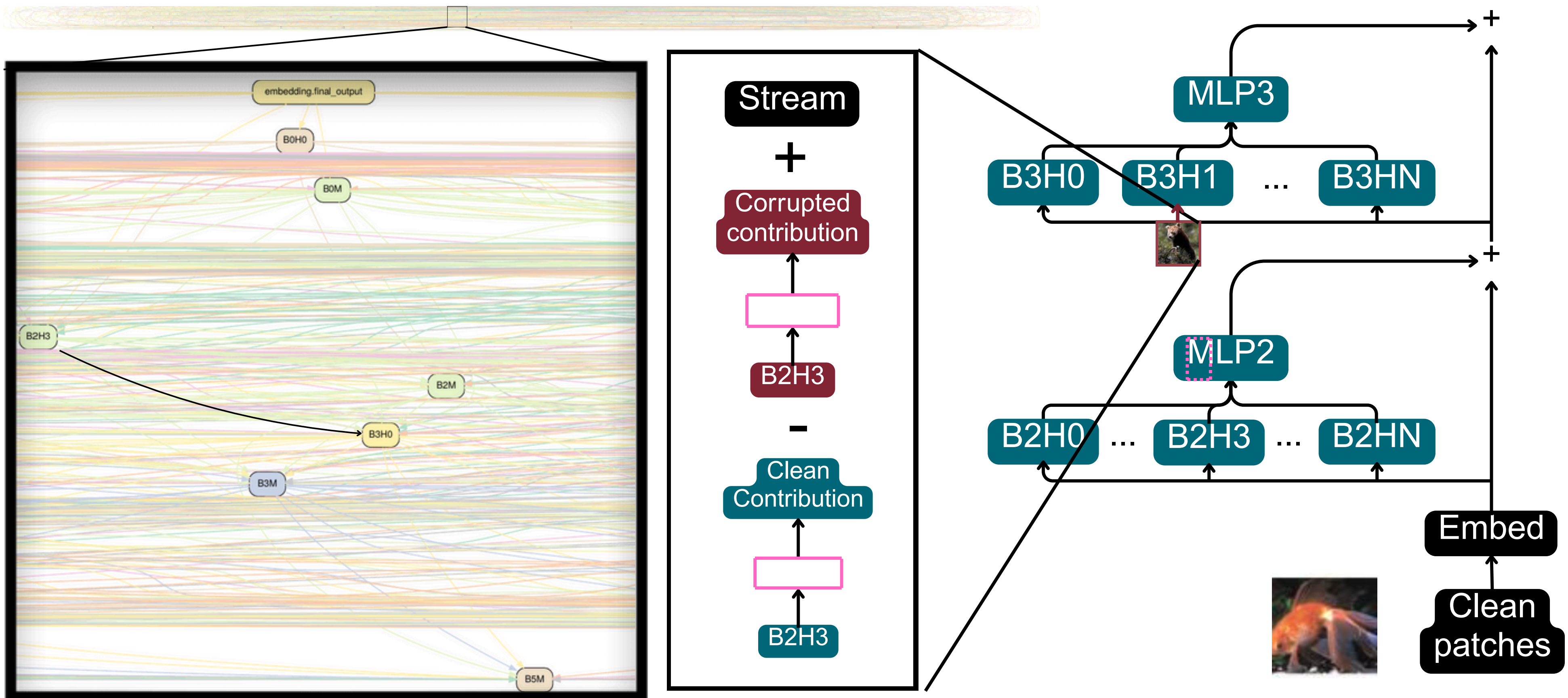
Cache all activations



Experimental Setup (2)

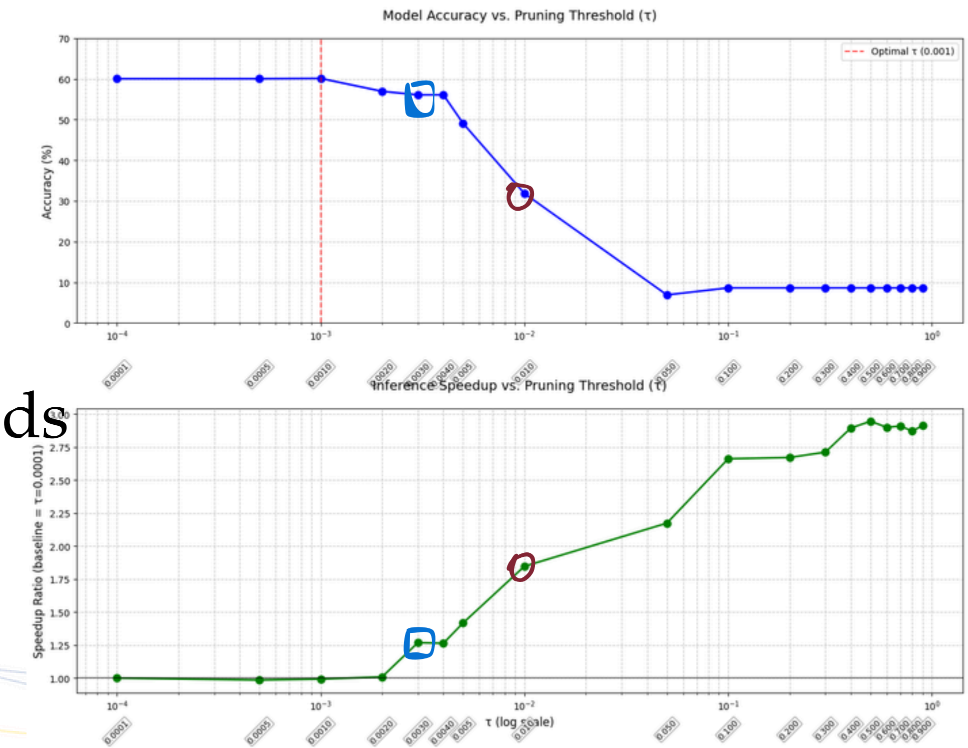
ACDC (3):

- Test each edge changing the input stream in input to the dest node
- If the average KL-Divergenge over train dataset between clean logits is less than τ, Prune the edge



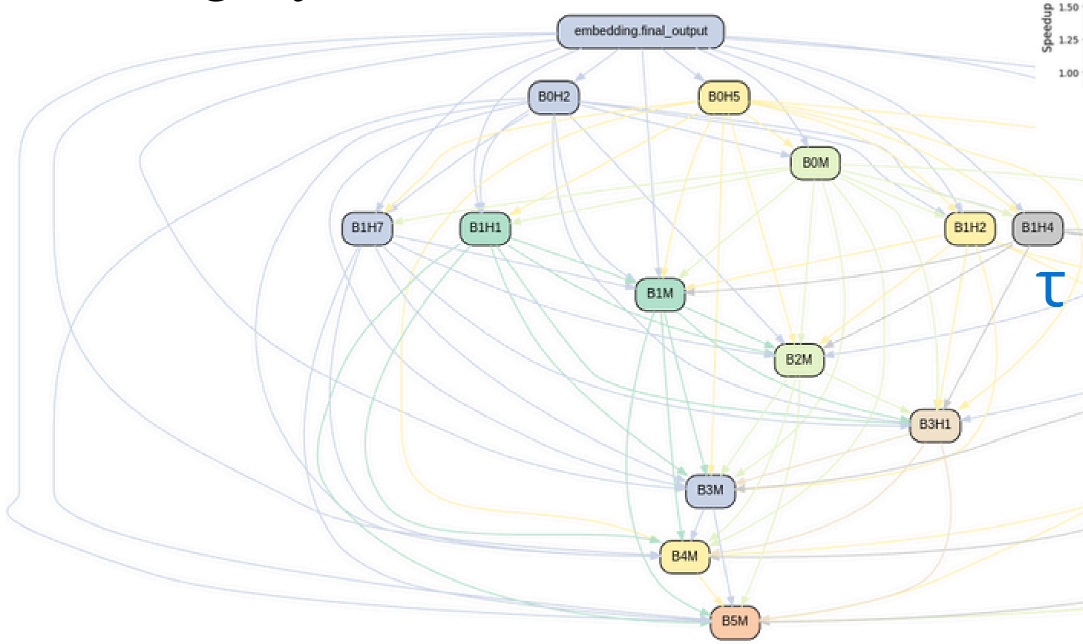
Model Evaluation (1)

Inference time and accuracy over τ



$\tau = 0.01$:

- ACDC pruned 94% of edges
- pruned 41 unused attention heads
- 50% accuracy decrease
- Roughly 2 times faster on CPU

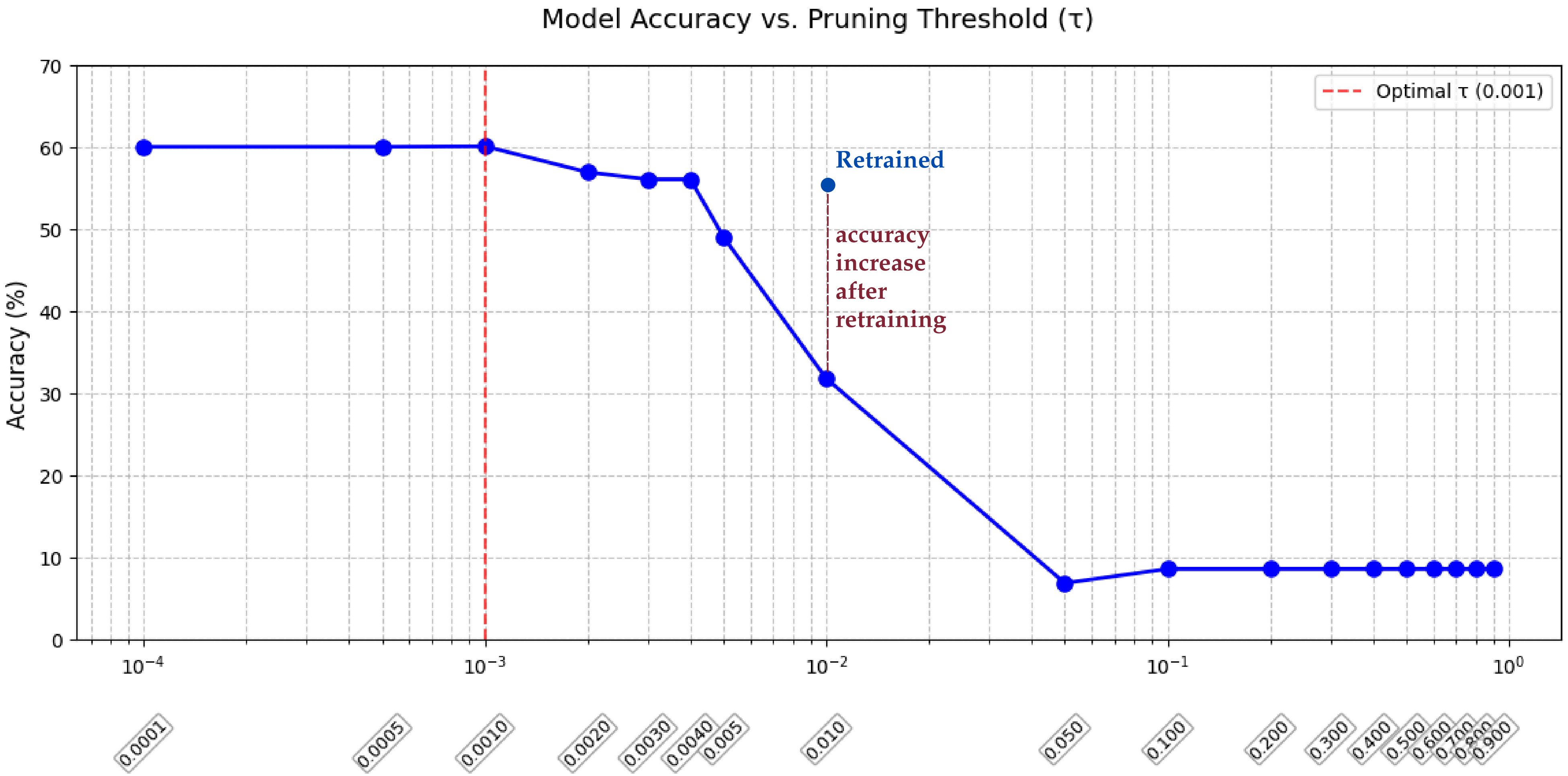


$\tau = 0.003$:

- ACDC pruned 57% of edges
- pruned 18 unused heads
- 9% accuracy decrease
- 25% speedup

Model Evaluation (2)

Accuracy on coarse classes after re initialization and training of the pruned ViT on the 58 fine labels



Conclusions:

- Mechanistic interpretability allows to squeeze performances from existing models
- 2 times speed increase and only 9% decrease of performances on CPU

Future work

- Create a prunable ViT architecture that scales well on GPU
- Benchmark performances on GPU

SAPIENZA
UNIVERSITÀ DI ROMA

References

Miller et al. (2024). **Transformer Circuit Evaluation Metrics Are Not Robust.**
Conference on Language Modeling.

Conmy et al. (2023). **Towards Automated Circuit Discovery for Mechanistic Interpretability.**
NeurIPS. arXiv:2304.14997 [cs.LG].

Elhage, et al. (2021) "**A Mathematical Framework for Transformer Circuits**", Transformer Circuits Thread.

Matt Nguyen. (2024). "**Building a Vision Transformer Model from Scratch.**" Medium.

Yuechun (Ethan) Gu (2024) "**Fine-Tuning Vision Transformer (ViT) on Tiny ImageNet Dataset**"