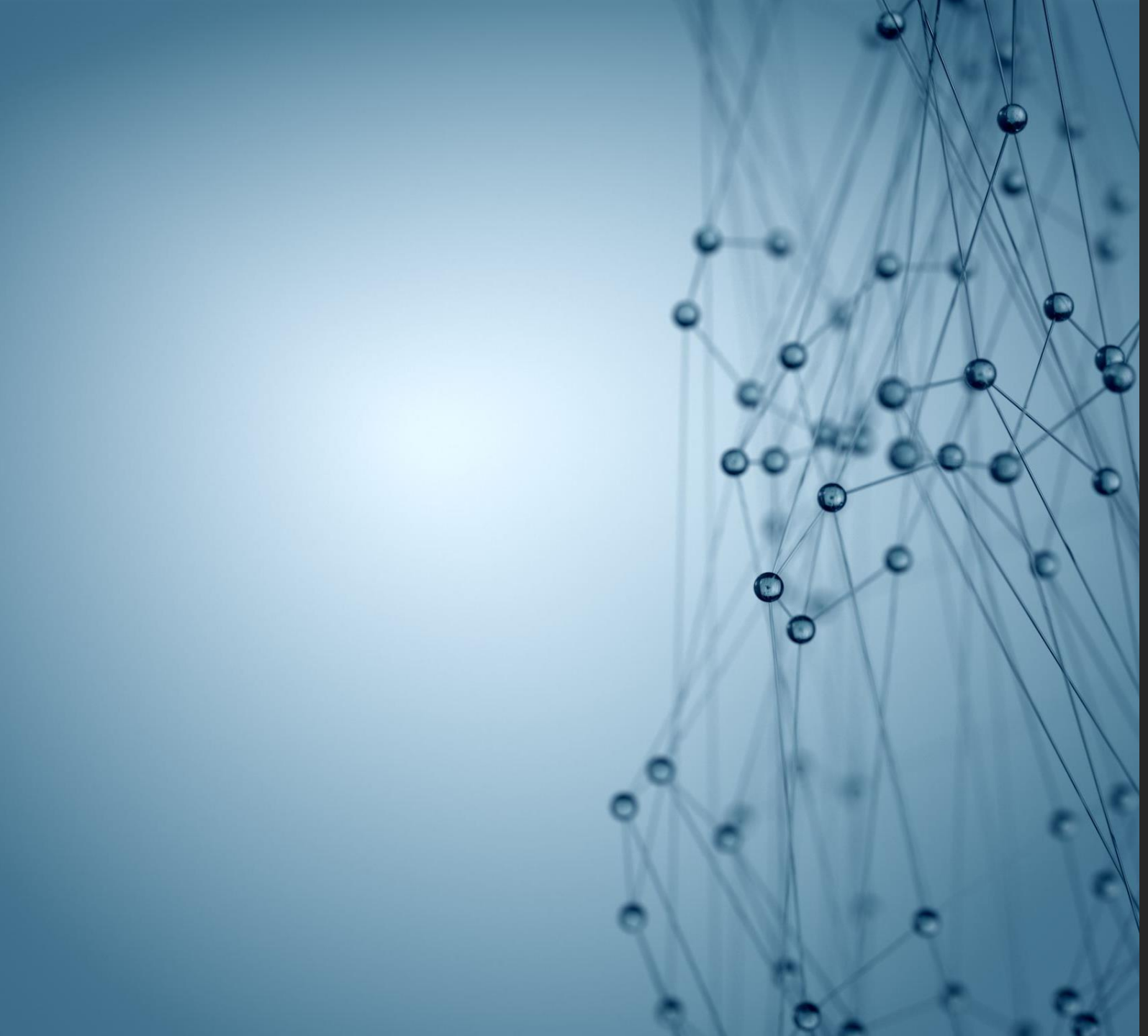


UNIVERSITÀ
DEGLI STUDI DI TRIESTE

AREA
SCIENCE PARK

SPEEDING-UP PRUNING FOR ARTIFICIAL NEURAL NETWORKS:

PRESENTING *ACCELERATED ITERATIVE MAGNITUDE PRUNING*

Marco Zullich

Eric Medvet

Felice Andrea Pellegrino

University of Trieste (Italy)

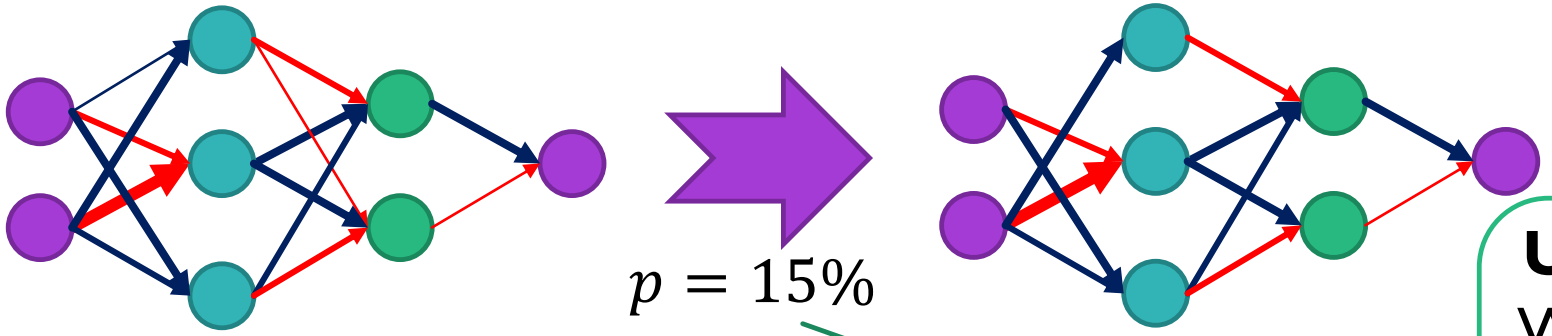
Alessio Ansuini

Area Research & Technology (Italy)

ITERATIVE MAGNITUDE PRUNING: STATE OF THE ART

PRUNING IN ANN & MAGNITUDE PRUNING

- Pruning a Neural Network → removing parameters from it
- Large number of criteria for pruning
- Magnitude pruning deletes parameters having small magnitude



Positive parameter

Negative parameter

Pruning rate

proportion of parameters to remove from the network

=

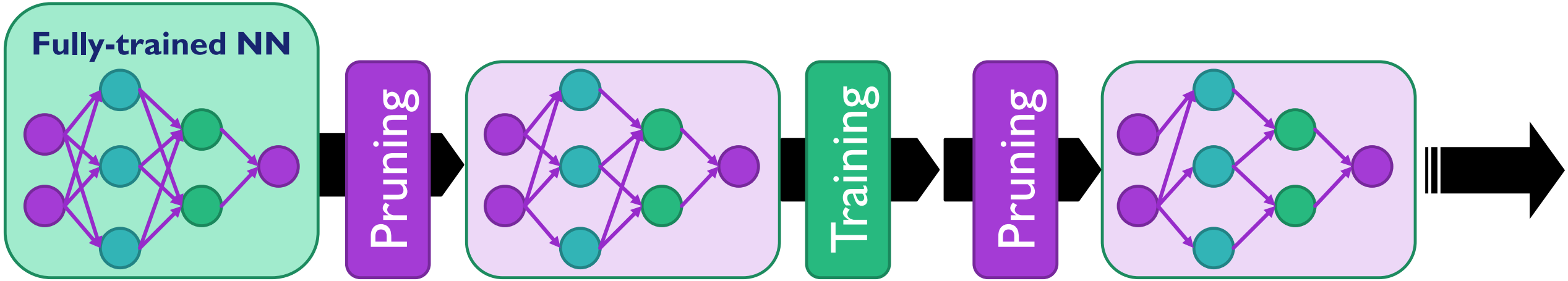
prune the lowest 15% of parameters in magnitude

Unstructured sparsity

We can take advantage of it via specific libraries (CUsparse) or HW (Nvidia A100 GPU)

PRUNING AND RE-TRAINING

- A simple application of pruning degrades the ANN performance
- After pruning, a **re-training phase** follows
- Re-training is operated only on parameters having survived the pruning



ITERATIVE MAGNITUDE PRUNING (IMP)

MAIN TECHNIQUES FOR RE-TRAINING

Weight Rewind (WR)

- Store random initialization of parameters Θ_0 of unpruned ANN
- After pruning, **reset** surviving weights to Θ_0
- Re-train for the same number of epochs keeping the same LR annealing schedule

Frankle and Carbin, 2019 «The Lottery Ticket Hypothesis»

Learning Rate Rewind (LRR)

- **Do not reset** weights after pruning
- Re-train for the same number of epochs keeping the same LR annealing schedule

Renda, Frankle and Carbin, 2020 «Comparing Rewinding and Fine-tuning in Neural Network Pruning»

PROS AND CONS OF WR & LRR

PROS

Reach very high pruning rates (>95%) with **similar or better performance** w.r.t. unpruned network

CONS

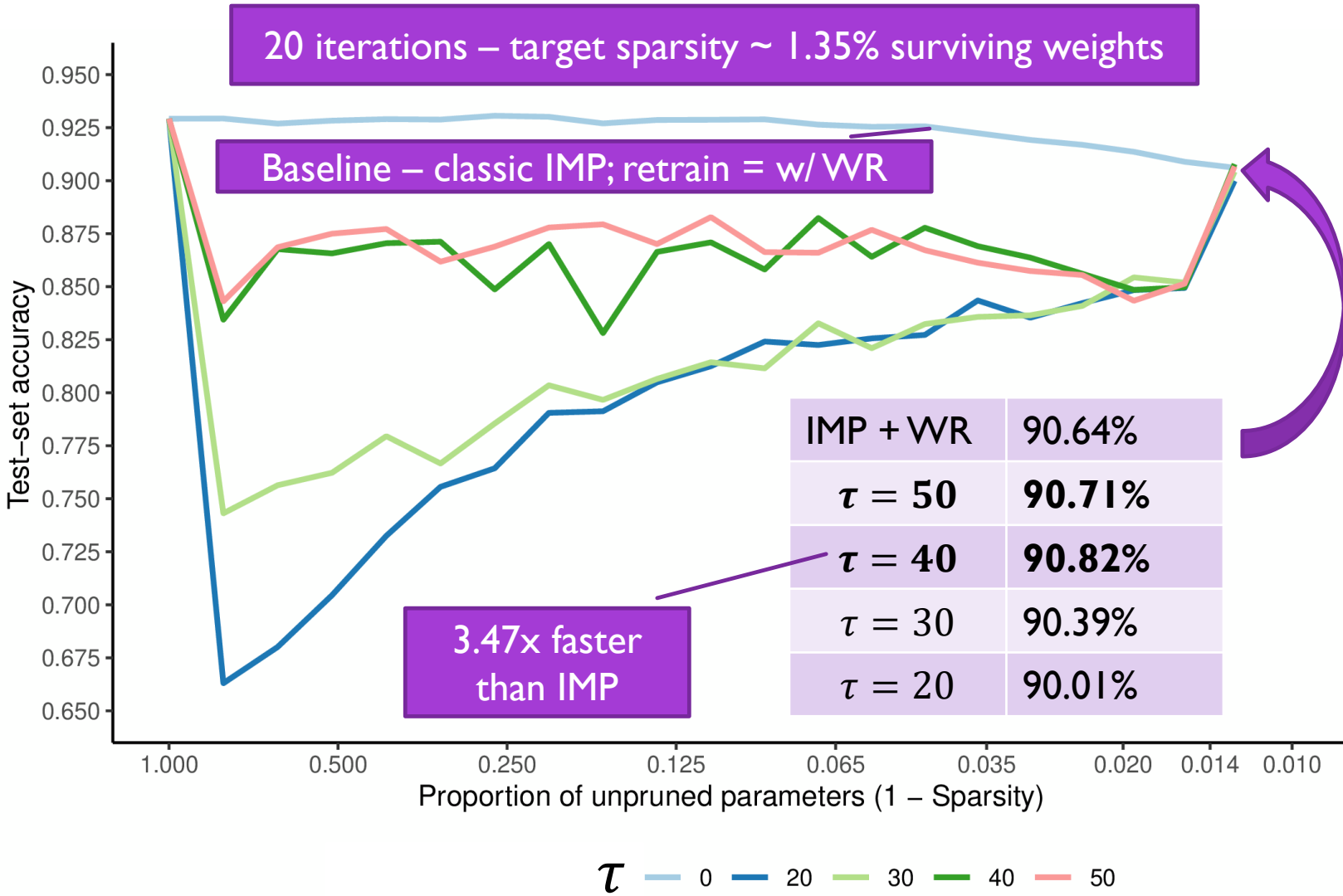
Especially if compared to other methods, requires application of many **sequential iterations**

If a target sparsity is known from the beginning, is it possible to fast-forward the execution of IMP for all the iterations but the last one?

PRESENTING ACCELERATED ITERATIVE MAGNITUDE PRUNING

ACCELERATING IMP

- Unpruned ANN trained for T epochs
- Prune for K iterations
- Iterations $1, \dots, K-1$: re-train for τ epochs, $\tau \ll T$
- **Accelerated Iterative Magnitude Pruning** (AIMP)
- Test with VGG-19 on CIFAR10 dataset
- $T = 160; K = 20; p = 0.2$



IMP + WR	90.64%
$\tau = 50$	**90.71%**
$\tau = 40$	**90.82%**
$\tau = 30$	90.39%
$\tau = 20$	90.01%

DRAWBACKS & DIRECTIONS FOR FUTURE WORK

- Trials on **IMP + LRR** were not as satisfying as IMP + WR
- Median accuracy: 93.68 % VS. 63.62 % ($\tau = 50$)
- No proper criterion to determine an optimal τ
- AIMP seems to work only when **overall pruning rate** is very high ($\geq 98\%$)

THE END

Thanks for the attention!

Contacts: **marco.zullich@phd.units.it**

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

AREA
SCIENCE PARK