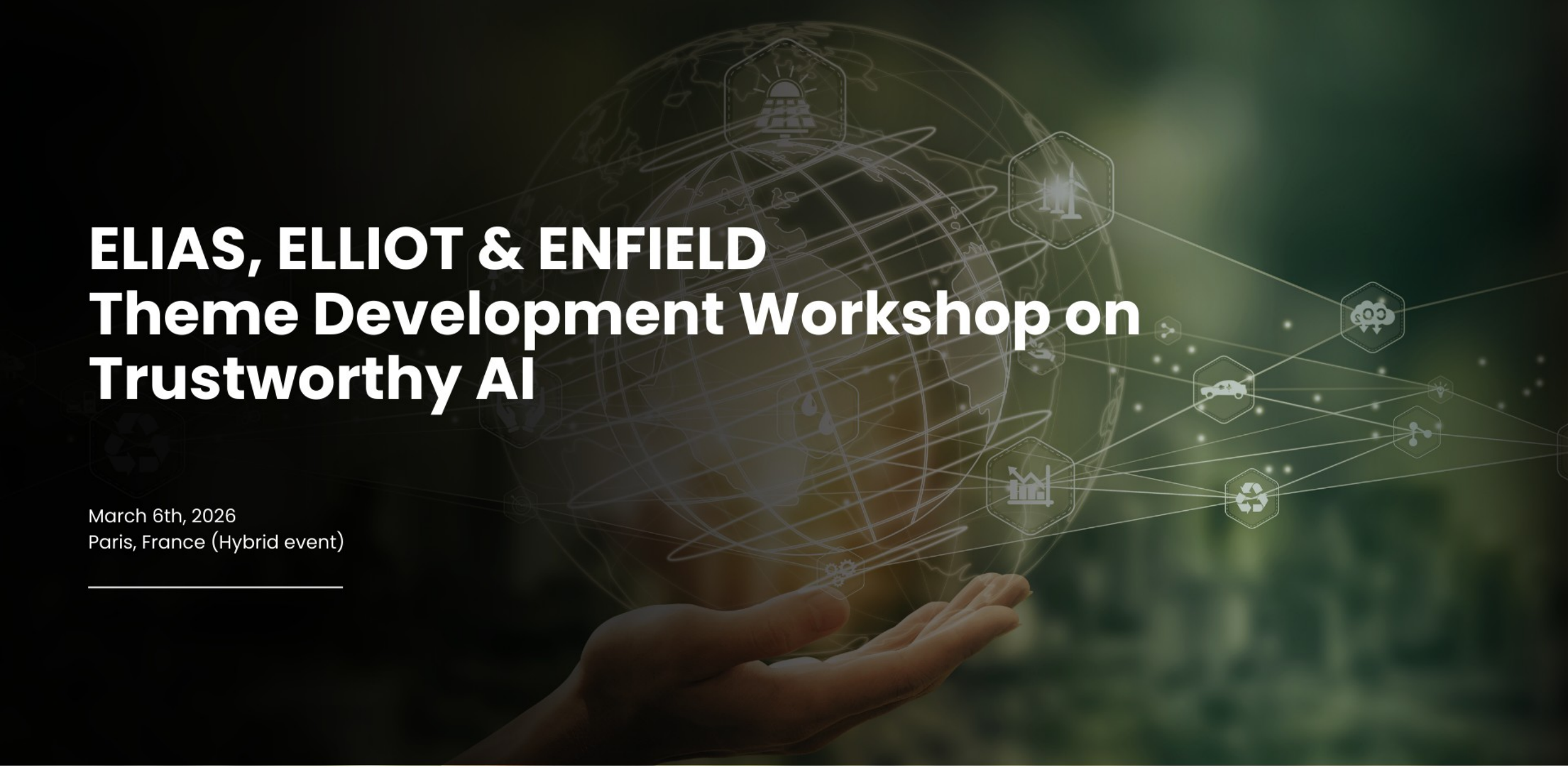


# ELIAS, ELLIOT & ENFIELD Theme Development Workshop on Trustworthy AI

March 6th, 2026
Paris, France (Hybrid event)

Funded by the European Union

**TDW on Trustworthy AI**

***Theme Development Workshops***

March 6th, 2026
Paris, France (Hybrid event)

# Beyond Traditional Pruning: Layer Collapse for Efficient Deep Learning

Enzo Tartaglione
Full Professor, Télécom Paris, Institut Polytechnique de Paris
enzo.tartaglione@telecom-paris.fr

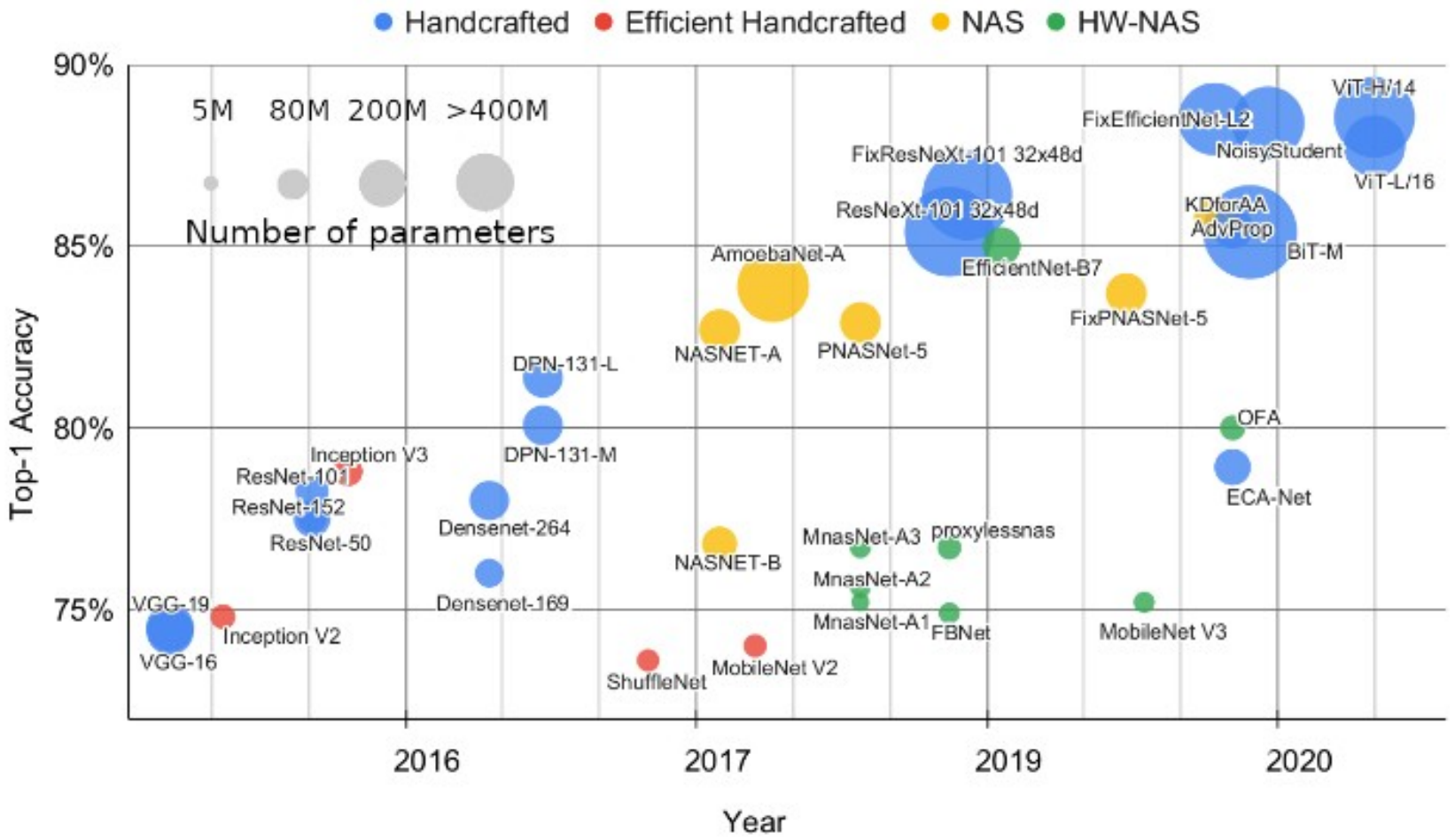
Funded by the European Union

# Outline

- Introduction to pruning and short story
- Pruning layers ??
- Parameter-efficient methods (pruning the changes in the network)
- Perspectives

Funded by the European Union

# Less recent trends in ANNs




Funded by the European Union

# More recent trends in ANNs

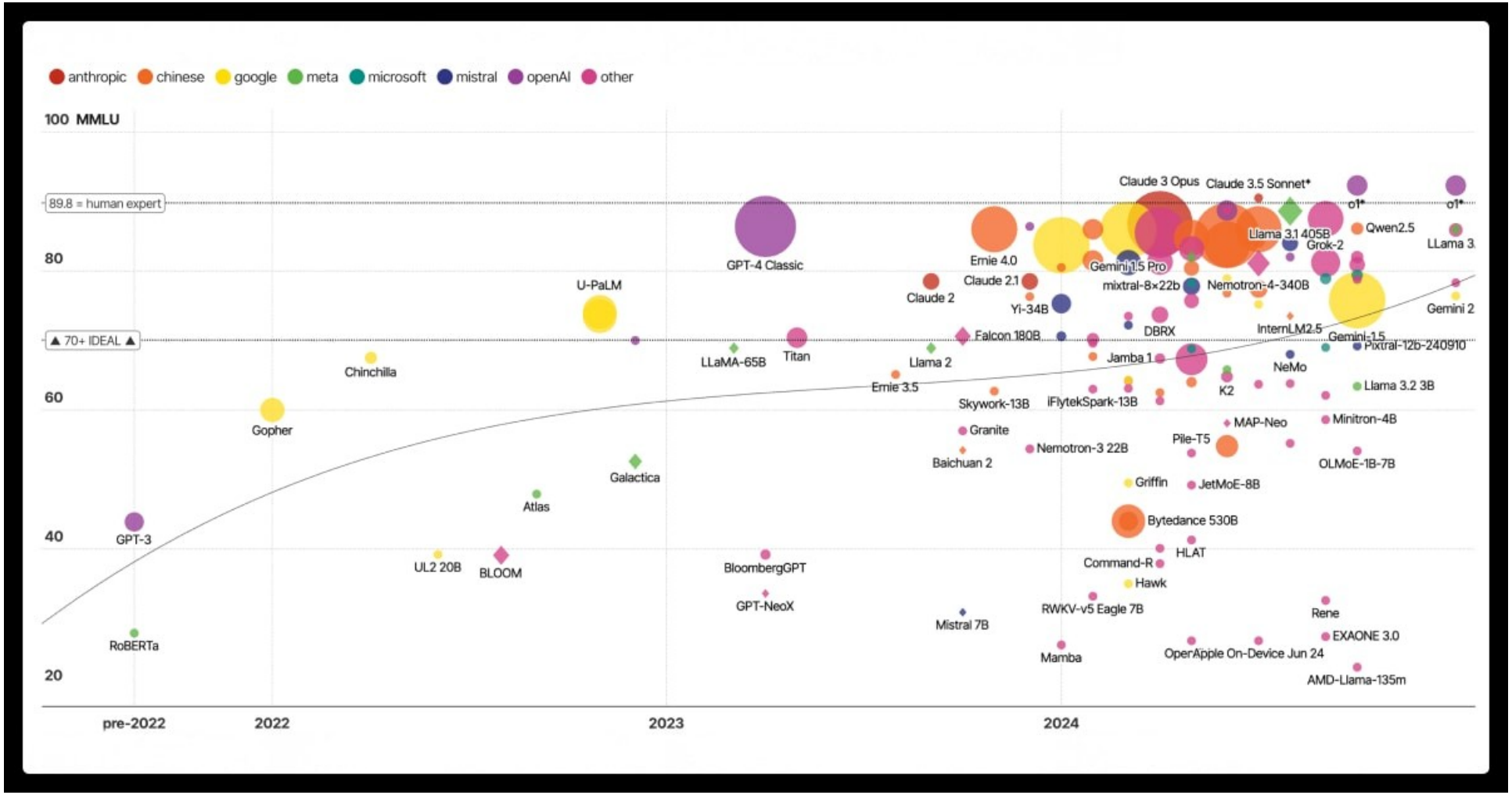



Image from https://labelyourdata.com/articles/llm-model-size

Funded by the European Union

# Pruning: a definition

- With pruning we refer to the process of removing parameters (synapses) or entire units from the deep learning model.

# Pruning: a definition

- With pruning we refer to the process of removing parameters (synapses) or entire units from the deep learning model.

- Pruning relates with **sparsification**: the weight matrix (representing a layer) becomes, indeed, sparse!

# Pruning: a definition

- With pruning we refer to the process of removing parameters (synapses) or entire units from the deep learning model.
- Pruning relates with **sparsification**: the weight matrix (representing a layer) becomes, indeed, sparse!
- The removed parameters (if they are removed in an “unstructured” way) still need to be encoded, with a “0”.
- Hence, in general, the position with a “missing” connection still needs to be encoded in some way (while using general frameworks), producing some representation overhead.
- Do we have advantages with a pruned architecture?

# Pruning: a definition

- With pruning we refer to the process of removing parameters (synapses) or entire units from the deep learning model.
- Pruning relates with **sparsification**: the weight matrix (representing a layer) becomes, indeed, sparse!
- The removed parameters (if they are removed in an “unstructured” way) still need to be encoded, with a “0”.
- Hence, in general, the position with a “missing” connection still needs to be encoded in some way (while using general frameworks), producing some representation overhead.
- Do we have advantages with a pruned architecture?
  - The number of parameters is reduced. For special designs (like ASICs) the gain is real.

# Pruning: a definition

- With pruning we refer to the process of removing parameters (synapses) or entire units from the deep learning model.
- Pruning relates with **sparsification**: the weight matrix (representing a layer) becomes, indeed, sparse!
- The removed parameters (if they are removed in an “unstructured” way) still need to be encoded, with a “0”.
- Hence, in general, the position with a “missing” connection still needs to be encoded in some way (while using general frameworks), producing some representation overhead.
- Do we have advantages with a pruned architecture?
  - The number of parameters is reduced. For special designs (like ASICs) the gain is real.
  - If they are removed in a “structured” way (entire blocks), there is no representation overhead, and the gain is real even in general frameworks!

# Pruning in deep learning: why?

- Reducing the storage memory (for a compressed model).
- Reducing the memory footprint.
- Reducing the FLOPs at inference time.
- Reducing energy consumption?
- Enhancing generalization?

Countless approaches to achieve sparse models…

Funded by the European Union

# There is a very rich (and long) story behind pruning…

There is a very rich (and long) story behind pruning…

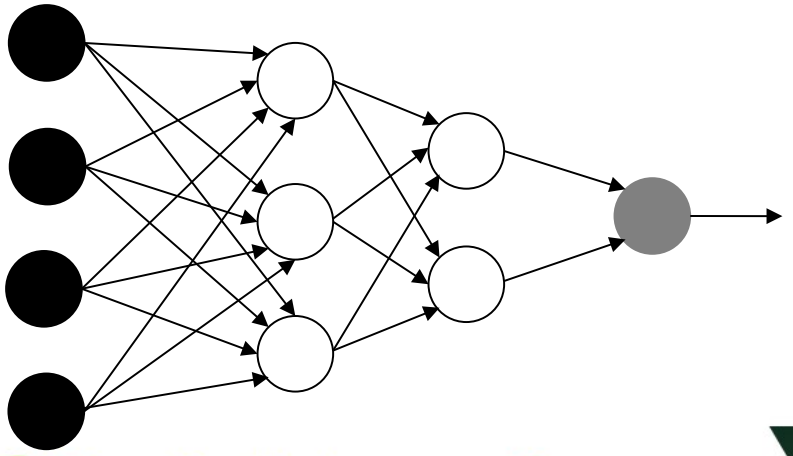
that I won’t mention here!

I will present just a simple (strong) baseline.

# Iterative Magnitude Pruning [Han et al., 2015]

- Parameters are randomy initialized

Train



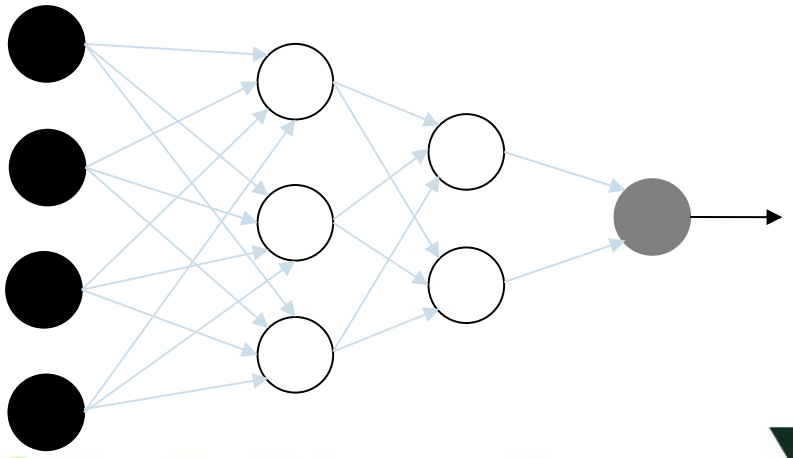
Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Funded by the European Union

# Iterative Magnitude Pruning [Han et al., 2015]

Train

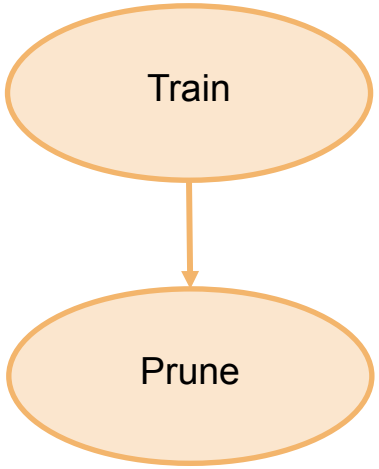
- Parameters are randomy initialized
- Parameters are updated then trained with standard gradient descent until performance is achieved (training stage)



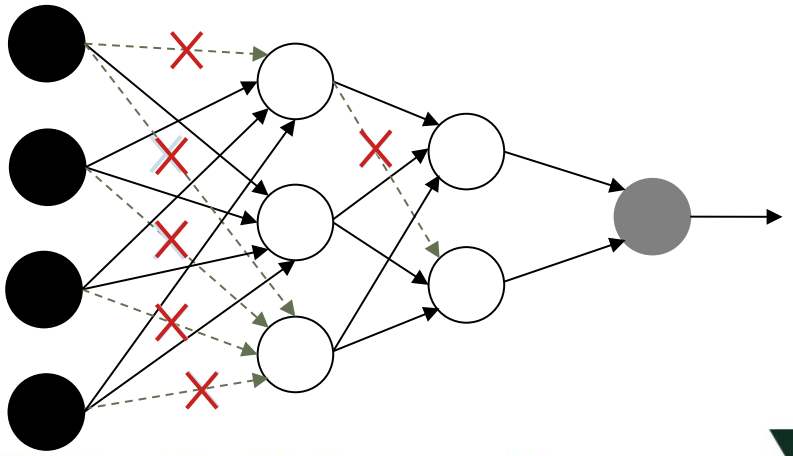
Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Funded by the European Union

# Iterative Magnitude Pruning [Han et al., 2015]



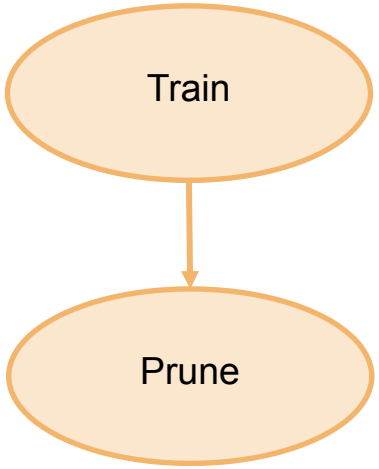

- Parameters are randomy initialized
- Parameters are updated then trained with standard gradient descent until performance is achieved (training stage)
- Parameters below threshold T are removed, pruning connections (parameter sparsification)



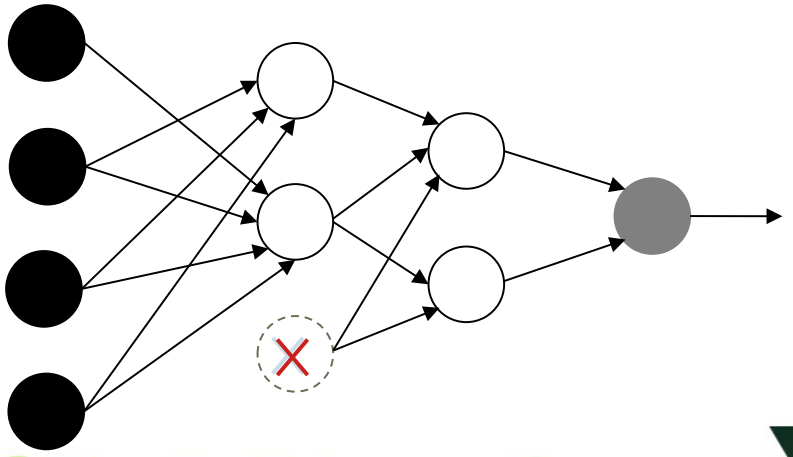
Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Funded by the European Union

# Iterative Magnitude Pruning [Han et al., 2015]



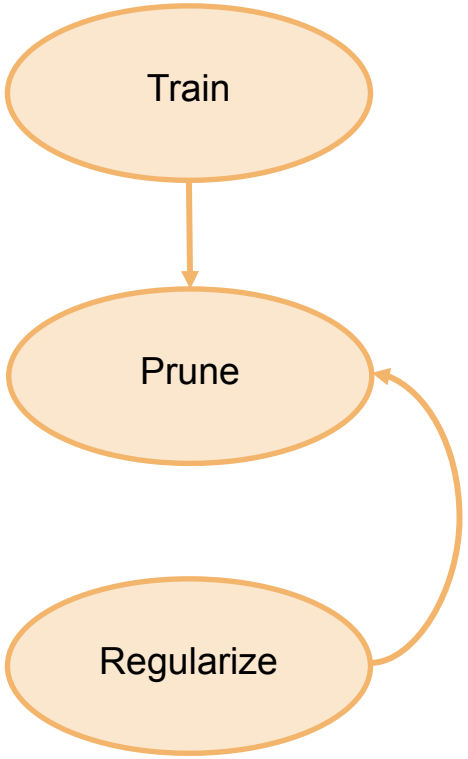

- Parameters are randomy initialized
- Parameters are updated then trained with standard gradient descent until performance is achieved (training stage)
- Parameters below threshold T are removed, pruning connections (parameter sparsification)
- Neurons without input arcs input are pruned from the network (neuron sparsification) ->Degrades network performance



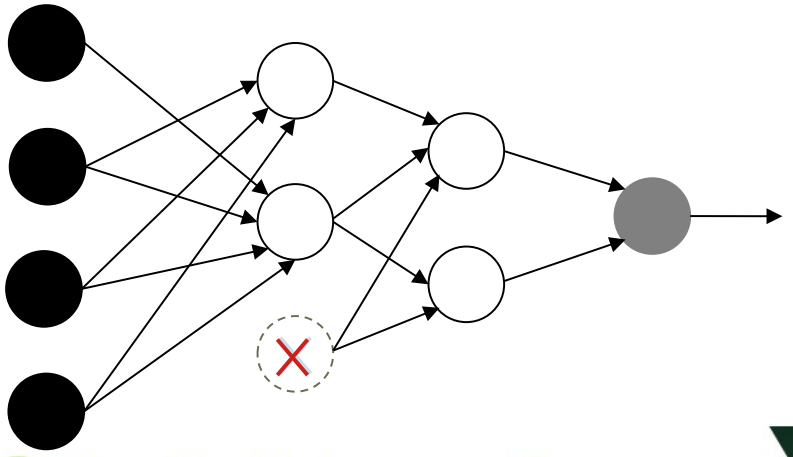
Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Funded by the European Union

# Iterative Magnitude Pruning [Han et al., 2015]



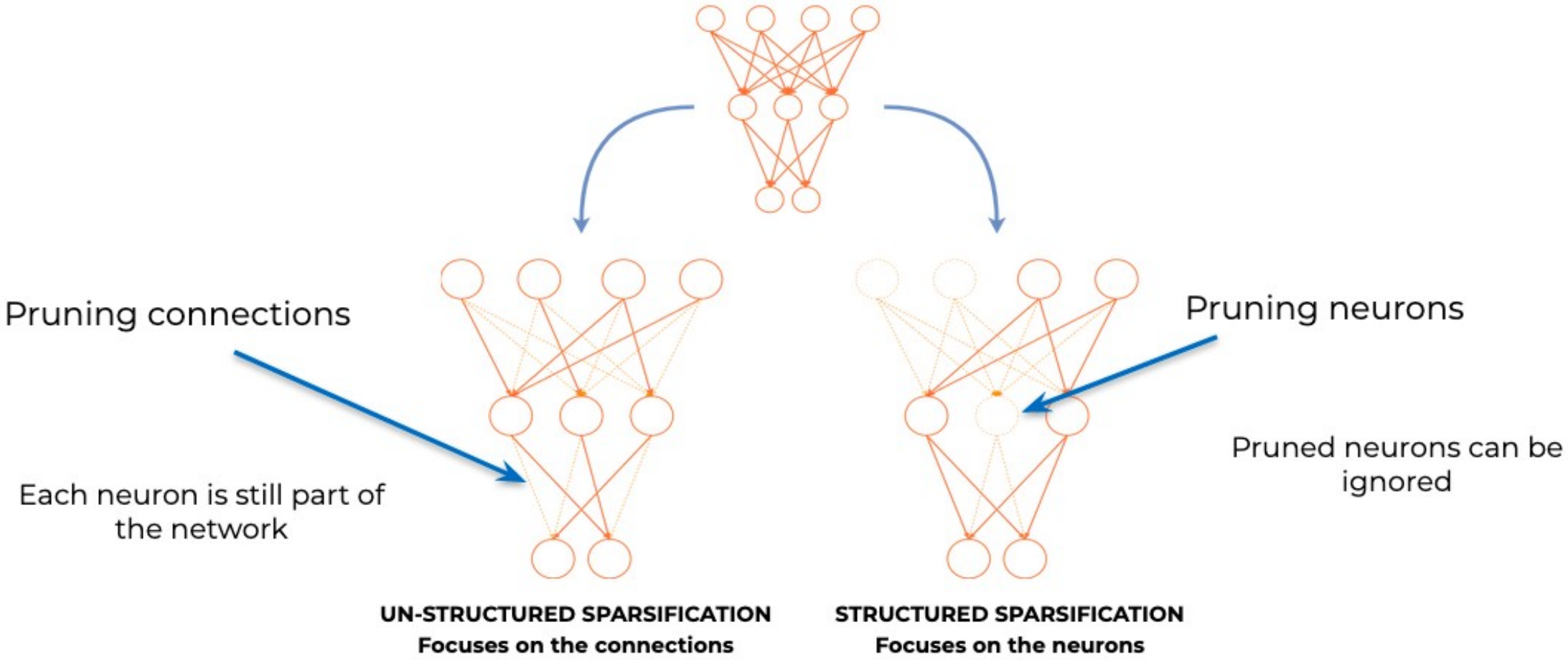

- Parameters are randomy initialized
- Parameters are updated then trained with standard gradient descent until performance is achieved (training stage)
- Parameters below threshold T are removed, pruning connections (parameter sparsification)
- Neurons without input arcs input are pruned from the network (neuron sparsification) ->Degrades network performance
- Fine-tune the model, recovering the performance and iteratively prune again.



Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Funded by the European Union

# Structured vs Unstructured sparsity




Funded by the European Union

# Structured vs Unstructured sparsity

| Dataset | Architecture | Pruning | Pruning ratio [%] | Simplified topology [MB] | Compressed bitstream [MB] | Inference time [ms] RPi 3B | P20 | MI9 | S6L | |
|---|---|---|---|---|---|---|---|---|---|---|
| CIFAR-10 | VGG-16 | No pruning | - | 60.0 | 3.6 | 647 | 204 | 153 | 251 | |
| | | LOBSTER | **92.44** | 58.61 | 1.61 | 610 | 191 | 146 | 242 | Unstructured |
| | | SeReNe | 47.16 | **31.02** | **0.34** | **594** | **99** | **85** | **106** | Structured |
| | ResNet-32 | No pruning | - | 2.0 | 0.30 | 580 | 32 | 30 | 31 | |
| | | LOBSTER | **81.19** | 1.96 | 0.12 | 545 | 32 | 26 | 30 | Unstructured |
| | | SeReNe | 52.80 | **1.0** | **0.09** | **536** | **25** | **17** | **25** | Structured |
| CIFAR-100 | AlexNet | No pruning | - | 94.6 | 10.1 | 246 | 131 | 84 | 168 | |
| | | LOBSTER | **98.90** | 48.84 | 0.40 | 224 | 95 | 67 | 120 | Unstructured |
| | | SeReNe | 59.87 | **37.07** | **0.20** | **186** | **75** | **53** | **96** | Structured |

Bragagnolo, A., Tartaglione, E., Fiandrotti, A., & Grangetto, M. (2021, September). On the Role of Structured Pruning for Neural Network Compression. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3527-3531). IEEE.

Funded by the European Union

…but GPUs can perform parallel computation!

Reducing layer’s size does not reduce the critical path!

…but GPUs can perform parallel computation!

Reducing layer’s size does not reduce the critical path!

What we want is to remove layers…

# Can we prune entire layers?

# Layer collapse

In deep learning, **layer collapse** refers to a phenomenon where a specific layer in a neural network fails to effectively differentiate the input features, causing all outputs from the layer to converge to similar or identical values regardless of input.

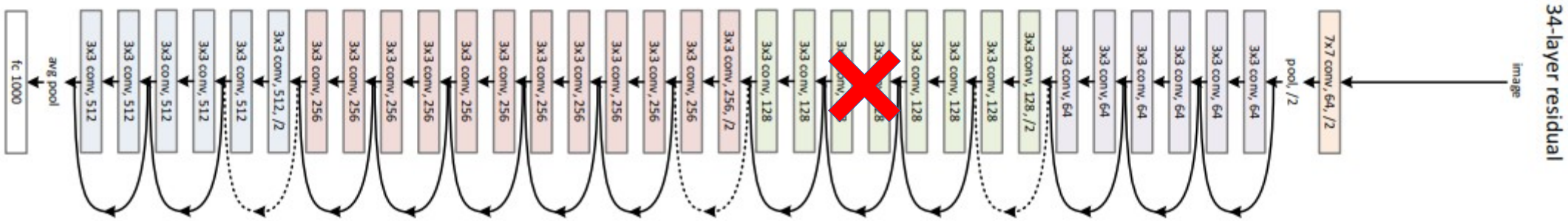
Funded by the European Union

# Layer collapse

In deep learning, **layer collapse** refers to a phenomenon where a specific layer in a neural network fails to effectively differentiate the input features, causing all outputs from the layer to converge to similar or identical values regardless of input.

Although this effect is traditionally seen as a downside of poor initialization or hyper-parameter fine-tuning (ie. the model does not learn), what if *the model learns, while some layers are collapsed?*

- If that happens, we can completely skip computation of such layer!



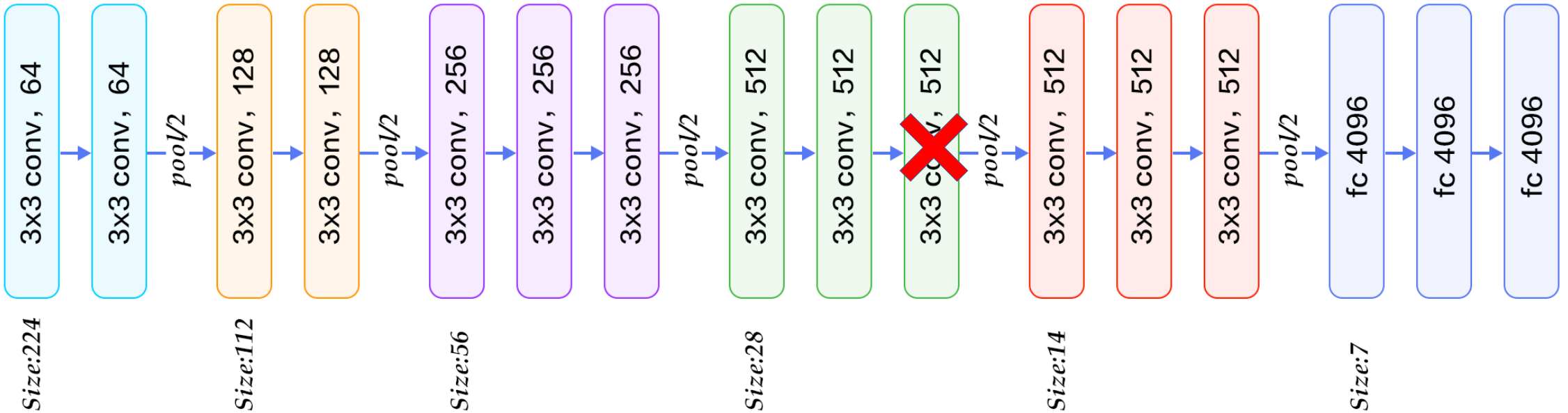

Funded by the European Union

# Layer collapse- without skip connections?


Funded by the European Union

# Layer collapse- without skip connections?

Is it really possible?

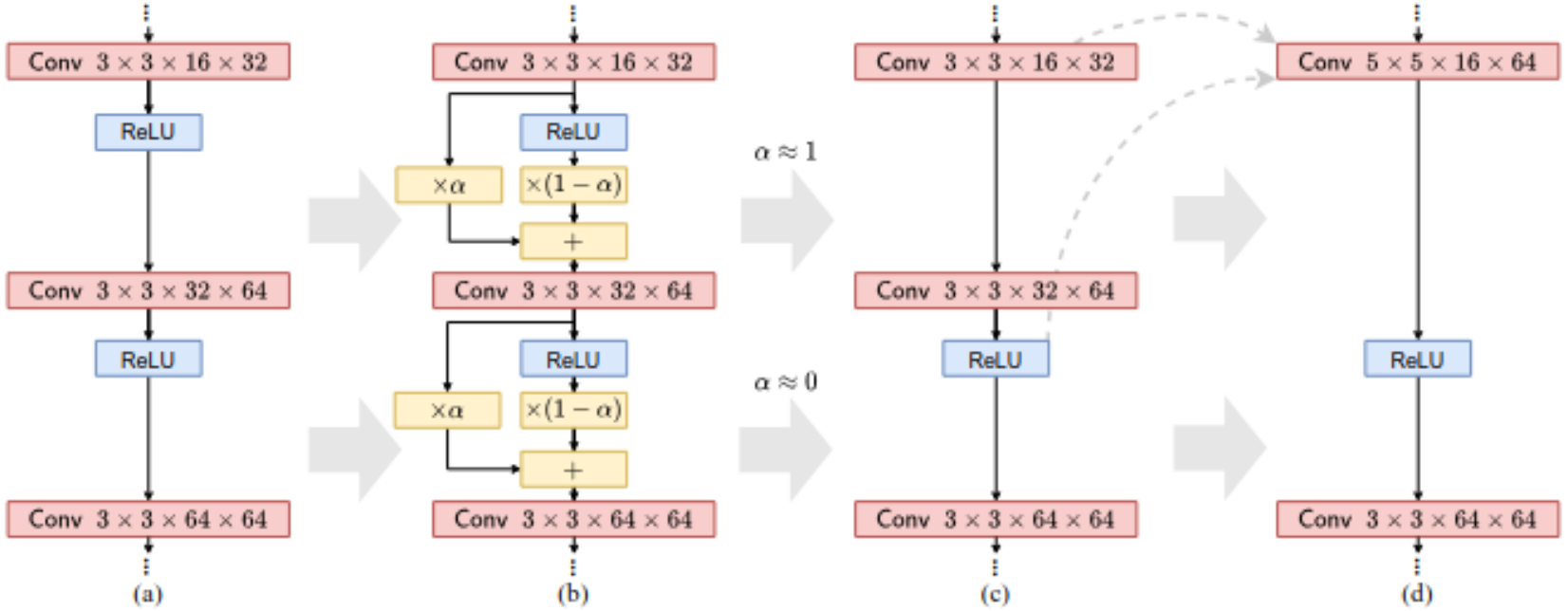



Funded by the European Union

# Layer collapse- without skip connections? [Dror et al. 2021]

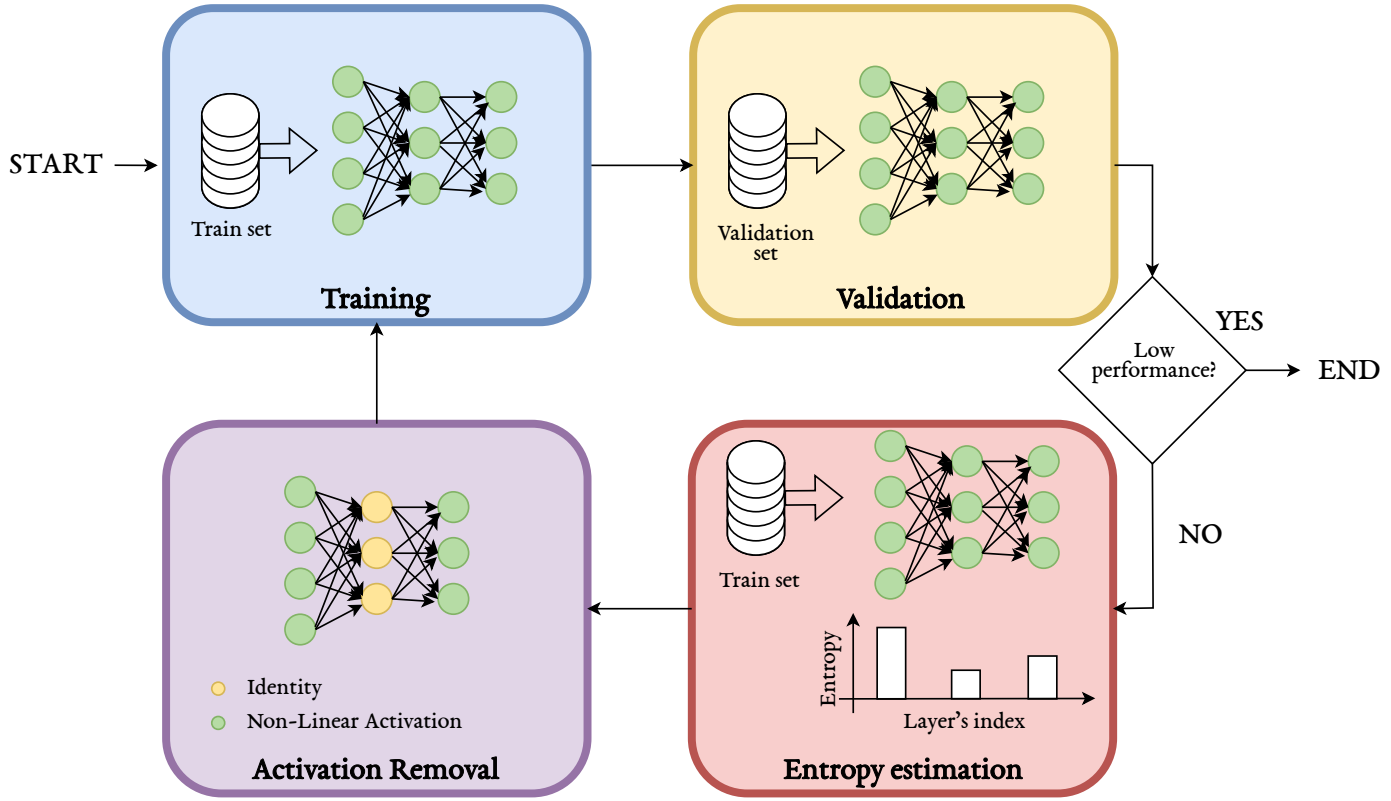
Is it really possible?

We can attack the problem differently. One way to reduce the model's depth is to *remove non-linearities!* Layer fold is one approach parametrizing the negative slope of a PReLU.




Funded by the European Union

# The naive way: EASIER [Quetu et al. 2024]




Quétu, V., Liao, Z., & Tartaglione, E. (2024, August). The simpler the better: An entropy-based importance metric to reduce neural networks' depth. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 92-108). Cham: Springer Nature Switzerland.

Funded by the European Union

# The naive way: EASIER [Quetu et al. 2024]

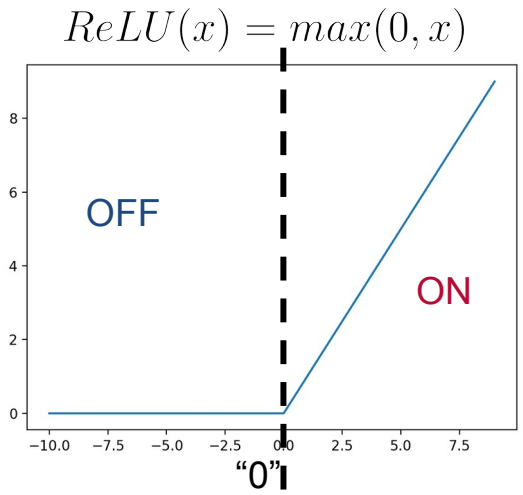
Three “states” for the neuron can be identified:

$$s^{\boldsymbol{x}}_{l,i} = sign(z^{\boldsymbol{x}}_{l,i}) = \begin{cases} +1 & \text{if } y^{\boldsymbol{x}}_{l,i} > 0 \\ -1 & \text{if } y^{\boldsymbol{x}}_{l,i} < 0 \\ 0 & \text{if } y^{\boldsymbol{x}}_{l,i} = 0 \end{cases}$$

The probability of the *i*-th neuron belonging to ON state can be calculated from the average over a batch of outputs for this neuron, and from this we define an estimator of neuron’s degeneration as

$$\mathcal{H}_{l,i} = -\sum_{s_{l,i}=\pm 1} p(s_{l,i}) \log_2 \left[p(s_{l,i})\right]$$

We can estimate layer’s degeneration averaging over all the neurons.



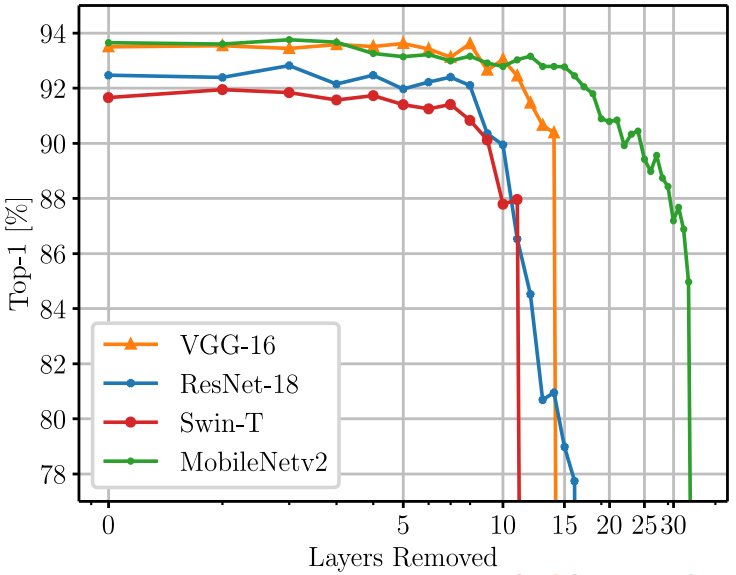

Quétu, V., Liao, Z., & Tartaglione, E. (2024, August). The simpler the better: An entropy-based importance metric to reduce neural networks’ depth. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 92-108). Cham: Springer Nature Switzerland.

Funded by the European Union

# The naive way: EASIER [Quetu et al. 2024]

EASIER applied on ResNet-18, VGG-16, Swin-T and MobileNetv2 on CIFAR-10. For each model, we gradually remove non-linear layers.

| Rem. | MFLOPs | Inference on CPU [ms] | | Inference on GPU [ms] | | | |
|---|---|---|---|---|---|---|---|
| | | Xeon E5-2640 | Raspi 4 | Jetson Orin | P2000 | RTX 2080 | A4500 |
| 0/17 | 725,47 | 13,50 | 135 | 8,52 | 4,45 | 4,43 | 3,32 |
| 1/17 | 258,24 | 9,33 | 111 | 8,31 | 4,53 | 4,43 | 3,27 |
| 2/17 | 243,46 | 9,69 | 106 | 7,83 | 4,28 | 4,21 | 3,10 |
| 3/17 | 231,79 | 9,43 | 139 | 7,38 | 4,02 | 3,93 | 2,96 |
| 4/17 | 197,85 | 10,10 | 117 | 6,91 | 3,79 | 3,68 | 2,78 |
| 5/17 | 159,05 | 11,30 | 144 | 6,44 | 3,60 | 3,46 | 2,60 |
| 6/17 | 159,99 | 8,39 | 225 | 6,13 | 4,11 | 3,18 | 1,79 |
| 7/17 | 152,36 | 9,18 | 144 | 6,06 | 4,16 | 3,10 | 1,71 |
| 8/17 | 149,84 | 9,14 | 149 | 6,14 | 3,67 | 3,21 | 1,55 |




Quétu, V., Liao, Z., & Tartaglione, E. (2024, August). The simpler the better: An entropy-based importance metric to reduce neural networks' depth. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 92-108). Cham: Springer Nature Switzerland.

Funded by the European Union

# The naive way: EASIER [Quetu et al. 2024]

EASIER applied on ResNet-18, VGG-16, Swin-T and MobileNetv2 on CIFAR-10. For each model, we gradually remove non-linear layers.

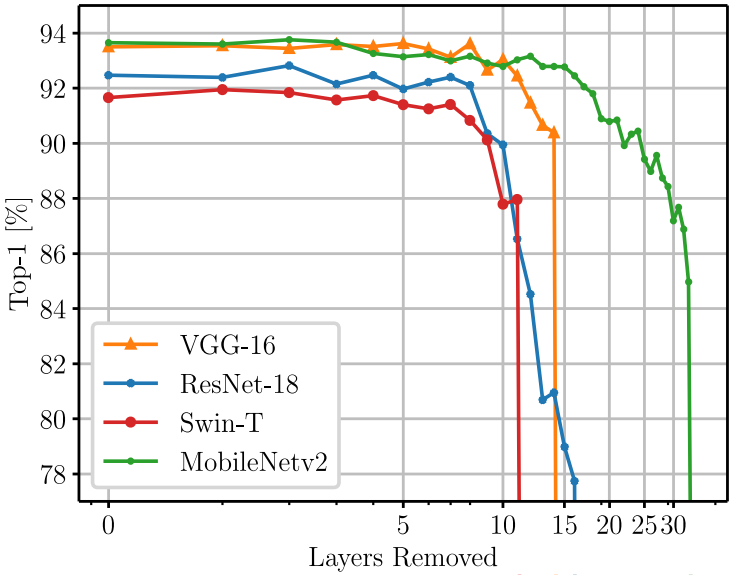
Limitations:

Performance degradation

→ Difficult to compress parameter-efficient architectures that are not overfitting without compromising performance.
For example: Swin-T on Tiny-ImageNet-200.

Training efficiency

→ Longer training time due to the iterative nature of EASIER.

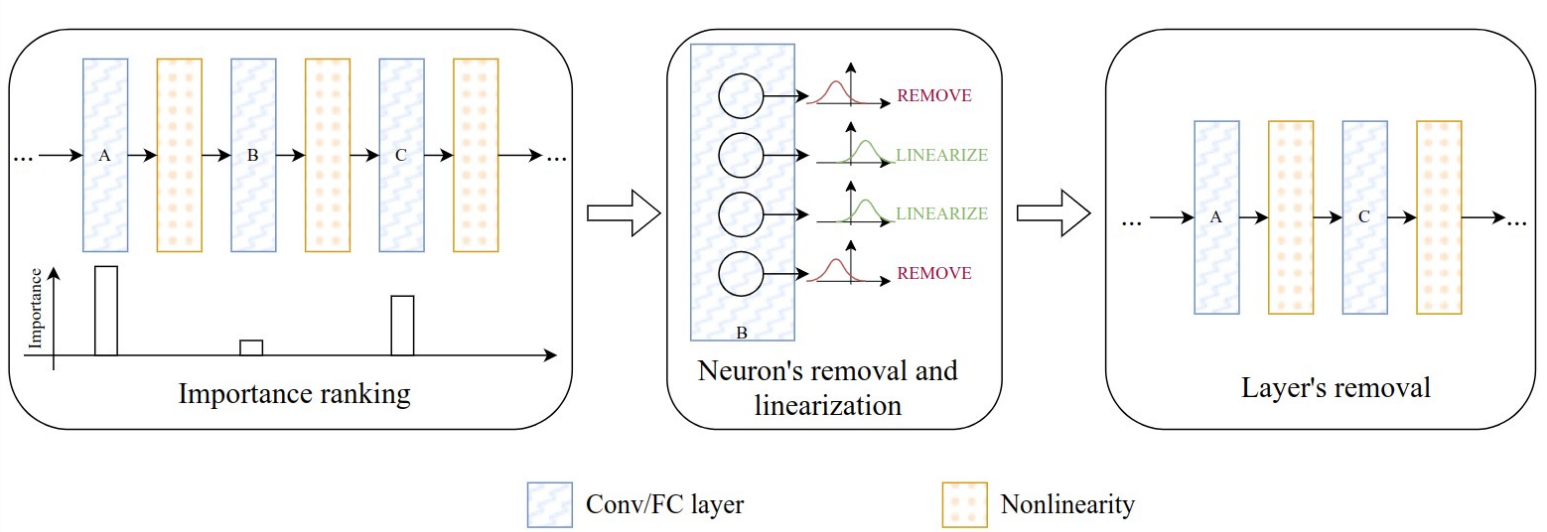



Models trained on CIFAR-10.

Quétu, V., Liao, Z., & Tartaglione, E. (2024, August). The simpler the better: An entropy-based importance metric to reduce neural networks' depth. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 92-108). Cham: Springer Nature Switzerland.

Funded by the European Union

# A better way: TLC [Liao et al. 2025]




Works by simply observing distributions in batch norms and in layer norms.

$$\hat{x}_{l,i} = \frac{x_{l,i} - \mu^B_{l,i}}{\sqrt{(\sigma^B_{l,i})^2 + \epsilon}}; \qquad z_{l,i} = \gamma_{l,i}\hat{x}_{l,i} + \beta_{l,i};$$

Still, sharing iterative fine-tuning nature, so computationally expensive….

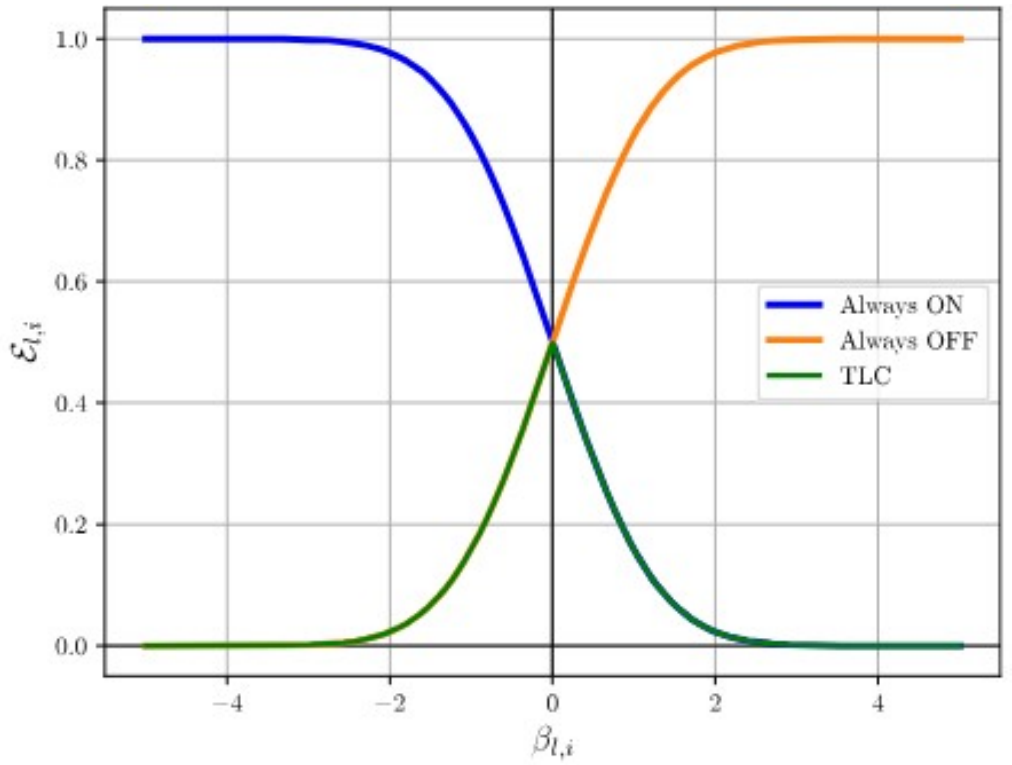
Liao, Z., Hezbri, N., Quétu, V., Nguyen, V. T., & Tartaglione, E. (2025). Till the Layers Collapse: Compressing a Deep Neural Network through the Lenses of Batch Normalization Layers. AAAI 2025 **(oral)**.

Funded by the European Union

# A better way: TLC [Liao et al. 2025]

$$\hat{x}_{l,i} = \frac{x_{l,i} - \mu^B_{l,i}}{\sqrt{(\sigma^B_{l,i})^2 + \epsilon}}; \qquad z_{l,i} = \gamma_{l,i}\hat{x}_{l,i} + \beta_{l,i};$$

$$\mathcal{E}_{l,i} = \Phi\left(-\frac{|\beta_{l,i}|}{\gamma_{l,i}}\right)$$




Liao, Z., Hezbri, N., Quétu, V., Nguyen, V. T., & Tartaglione, E. (2025). Till the Layers Collapse: Compressing a Deep Neural Network through the Lenses of Batch Normalization Layers. AAAI 2025 **(oral)**.

Funded by the European Union

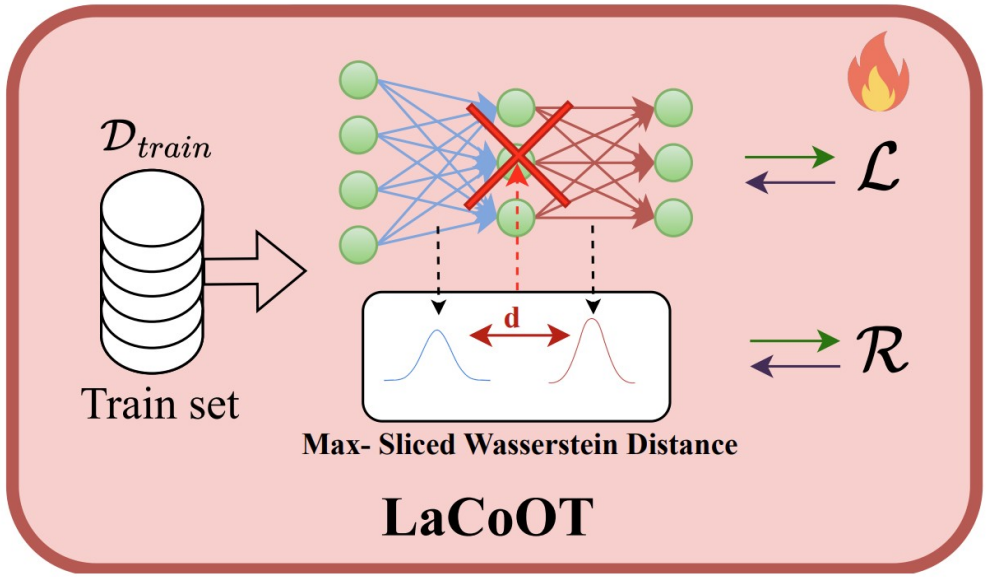
# A better way: TLC [Liao et al. 2025]

| Dataset | Approach | ResNet-18 | | Swin-T | | MobileNet-V2 | | VGG-16bn | |
|---|---|---|---|---|---|---|---|---|---|
| | | top-1 | Rem. | top-1 | Rem. | top-1 | Rem. | top-1 | Rem. |
| CIFAR-10 | Dense model | 92.00 | 0/17 | 91.63 | 0/12 | 93.64 | 0/35 | 93.09 | 0/15 |
| | Smallest weights | 88.49 | 11/17 | 86.92 | 3/12 | 10.00 | 1/35 | 90.53 | 7/15 |
| | Smallest gradients | 88.60 | 11/17 | 86.96 | 3/12 | 10.00 | 1/35 | 90.4 | 7/15 |
| | EGP | 90.64 | 5/17 | 86.04 | 6/12 | 92.22 | 6/35 | 10.00 | 1/15 |
| | LF | 90.65 | 1/17 | 85.73 | 2/12 | 89.24 | 9/35 | 86.46 | 1/15 |
| | EASIER | 86.53 | 11/17 | 91.25 | 6/12 | 92.45 | 16/35 | 93.03 | 7/15 |
| | TLC | **90.91 ± 0.57** | **12/17** | **91.98 ± 0.07** | **6/12** | **92.97 ± 0.38** | **17/35** | **93.61 ± 0.23** | **7/15** |
| Tiny-Inet | Dense model | 41.86 | 0/17 | 75.88 | 0/12 | 45.70 | 0/35 | 58.44 | 0/15 |
| | Smallest weights | 37.42 | 8/17 | 72.90 | 1/12 | 0.5 | 1/35 | 56.88 | 1/15 |
| | Smallest gradients | 37.88 | 8/17 | 72.92 | 1/12 | 0.5 | 1/35 | 57.34 | 1/15 |
| | LF | 37.86 | 4/17 | 50.54 | 1/12 | 25.88 | 12/35 | 31.22 | 1/15 |
| | EGP | 37.44 | 5/17 | 71.48 | 1/12 | 46.88 | 1/35 | — | — |
| | EASIER | 35.84 | 6/17 | 70.94 | 1/12 | 47.58 | 11/35 | 55.16 | 1/15 |
| | TLC | **38.66 ± 0.68** | **9/17** | **74.07 ± 0.02** | **1/12** | **47.84 ± 0.55** | **16/35** | **57.63 ± 0.65** | 1/15 |

Liao, Z., Hezbri, N., Quétu, V., Nguyen, V. T., & Tartaglione, E. (2025). Till the Layers Collapse: Compressing a Deep Neural Network through the Lenses of Batch Normalization Layers. AAAI 2025 **(oral)**.

Funded by the European Union

# Layer Collapse by Regularization [Quetu et al. 2024]



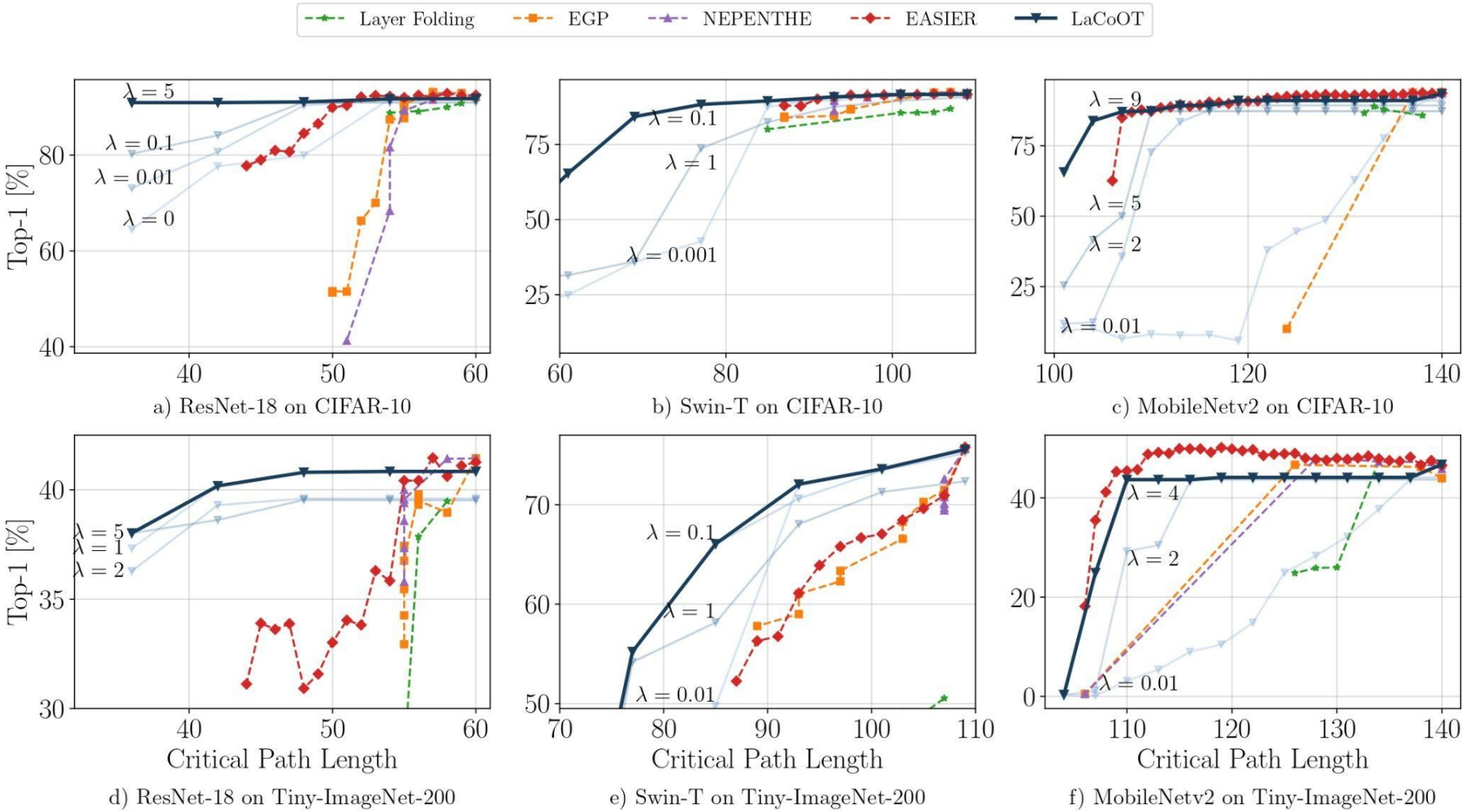

Training performed by minimizing distribution distance between adjacent layers (or group of) through a regularization term.

Layer pruning is performed in one-shot without retraining.

Quétu, V., Liao, Z., Hezbri, N., Pizzati, F., & Tartaglione, E. (2025). LaCoOT: Layer Collapse through Optimal Transport. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20497-20507).

Funded by the European Union

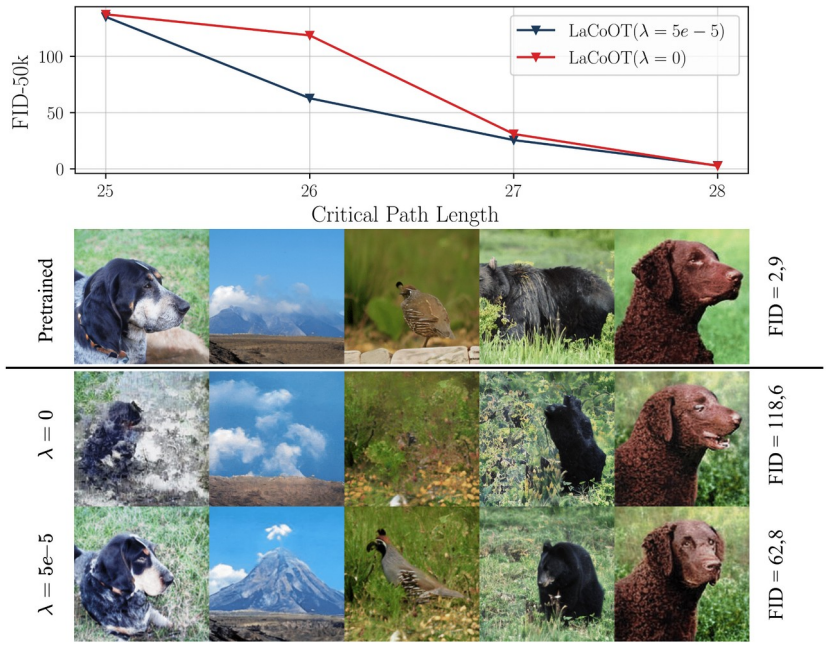
# Layer Collapse by Regularization [Quetu et al. 2024]




Quétu, V., Liao, Z., Hezbri, N., Pizzati, F., & Tartaglione, E. (2025). LaCoOT: Layer Collapse through Optimal Transport. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20497-20507).

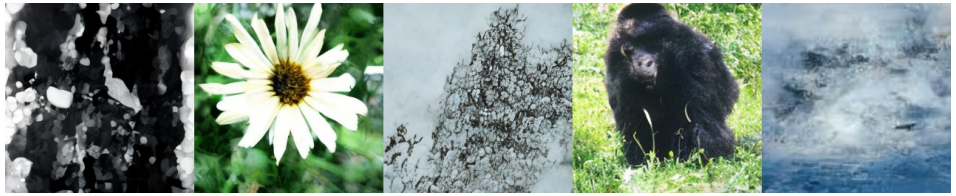
Funded by the European Union

# Layer Collapse by Regularization [Quetu et al. 2024]




FID-50k as a function of the critical path length achieved by a DiT-XL/2 finetuned on ImageNet.

Pretrained DiT-XL/2



DiT-XL/2 finetuned with λ=0 – 2 blocks removed

DiT-XL/2 finetuned with λ=5e-5 – 2 blocks removed

Quétu, V., Liao, Z., Hezbri, N., Pizzati, F., & Tartaglione, E. (2025). LaCoOT: Layer Collapse through Optimal Transport. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20497-20507).

Funded by the European Union

# Want to join the MultiMedia équipe in Paris?

WE ARE HIRING YOUNG PROFESSORS!

- Assistant/Associate Professor in Video Processing and Animation with Deep Learning
  https://institutminestelecom.recruitee.com/l/en/o/assistantassociate-professor-in-video-processing-and-animation-with-deep-learning-3-years-fixed-term-contract

- Assistant/Associate Professor in Coding for Images and Video
  https://institutminestelecom.recruitee.com/l/en/o/assistantassociate-professor-in-coding-for-images-and-video-permanent-contract

Funded by the European Union

# Follow us!

**Subscribe to the newsletter by scanning the QR code.**

ELIAS | elias-ai.eu seeks to establish Europe as a leader in AI research, advancing sustainable innovation and economic growth. By connecting academic researchers with industry professionals, ELIAS ensures that AI research promotes a sustainable future, strengthens societal cohesion, and upholds individual rights.

ELIAS

elias_project

elias-ai-project

www.elias-ai.eu

elias-coordination@unitn.it

Funded by
the European Union

# Follow us!

**Check our LinkedIn page by scanning the QR Code.**

ELLIOT | elliot-ai.eu develops the next generation of open, trustworthy multimodal foundation models built on European values. ELLIOT combines real and synthetic data, scalable infrastructure, and open science to create robust AI systems capable of generalising across diverse, dynamic data streams. It promotes reproducibility, ethical deployment, and societal uptake across key application domains.

elliot
MULTIMODAL AI MADE IN EUROPE

elliot-eu

elliot-eu

www.elliot-ai.eu

ikom@iti.gr

Funded by the European Union

# Follow us!

**Check our LinkedIn page by scanning the QR Code.**

ENFIELD | enfield-project.eu advances adaptive, green, human-centric, and trustworthy AI across Europe. Bringing together academia, industry, SMEs, and the public sector—ENFIELD develops AI solutions for healthcare, energy, manufacturing, and space. By delivering applications, publications, and strategic roadmaps, it fosters reproducibility, ethical deployment, and societal uptake of AI across Europe.

ENFIELD

HorizonEnfield

horizon-enfield

www.enfield-project.eu

Funded by
the European Union

# TDW on Trustworthy AI

***Theme Development Workshops***

March 6th, 2026
Paris, France (Hybrid event)

# Thank you!

Funded by the European Union