DEEPHEALTH

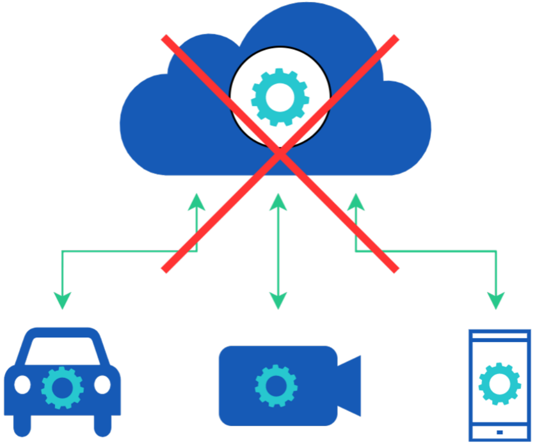
Efficiencing deep learning with pruning

Reconfigurable Architectures support in EDDL

January 27 2022

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Neural networks on embedded devices



Inferences are performed locally.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Advantages

Connectivity

Low latency

Privacy

Costs

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

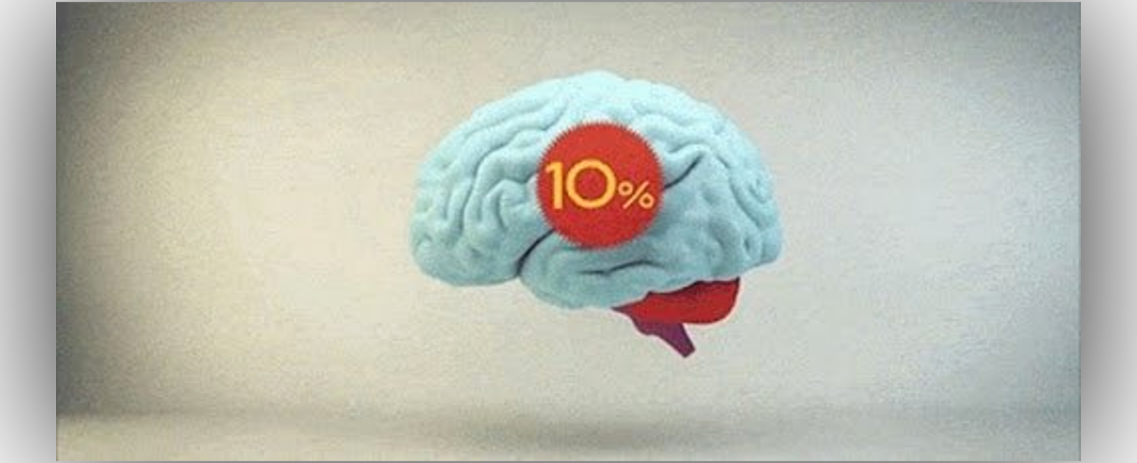
Problems

Battery

Computing power

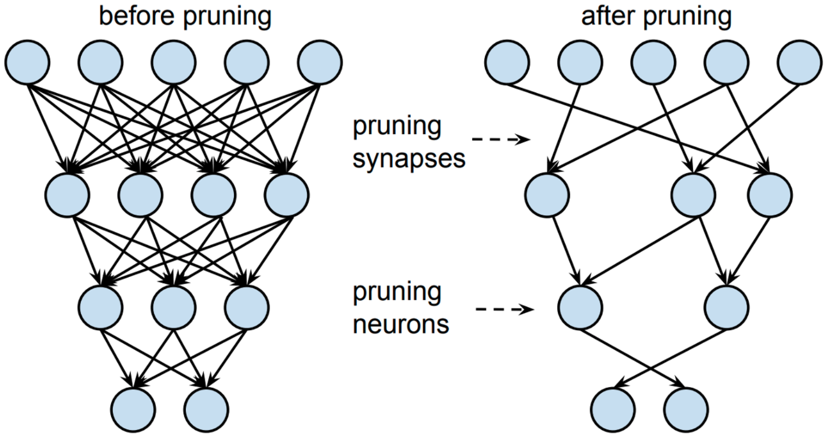
Storage

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.



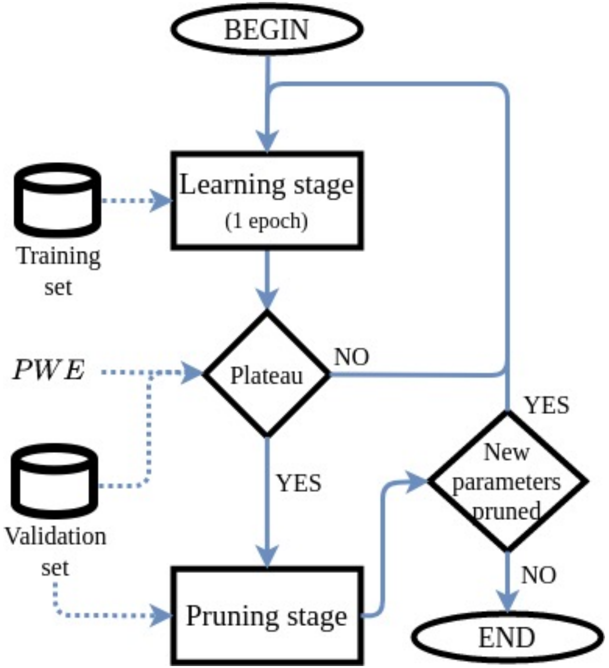
The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Neural network pruning



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

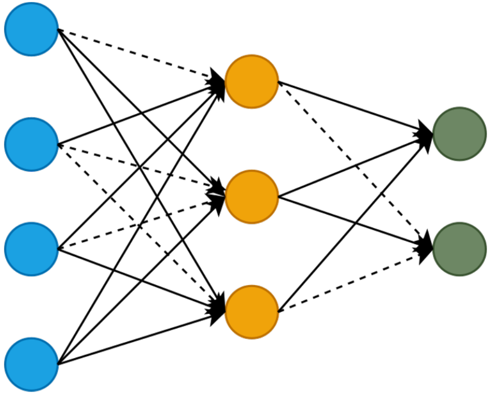
Iterative pruning strategy

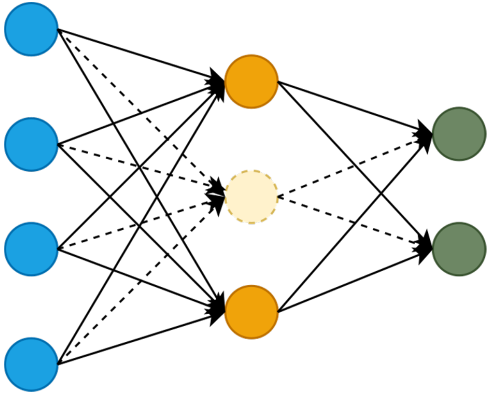


PWE= **patience on the epochs before pruning**

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Unstructured vs structured pruning





Unstructured

Removes many parameters from the network.

Can highly reduce the compressed model size.

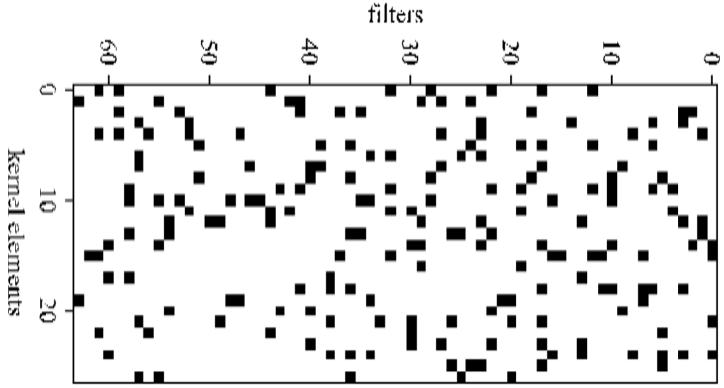
Structured

Removes entire neurons in the network.

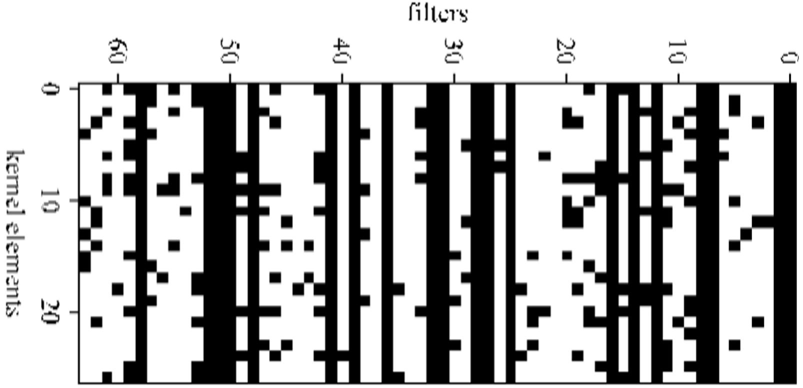
Reduces the number of operations.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Unstructured vs structured pruning



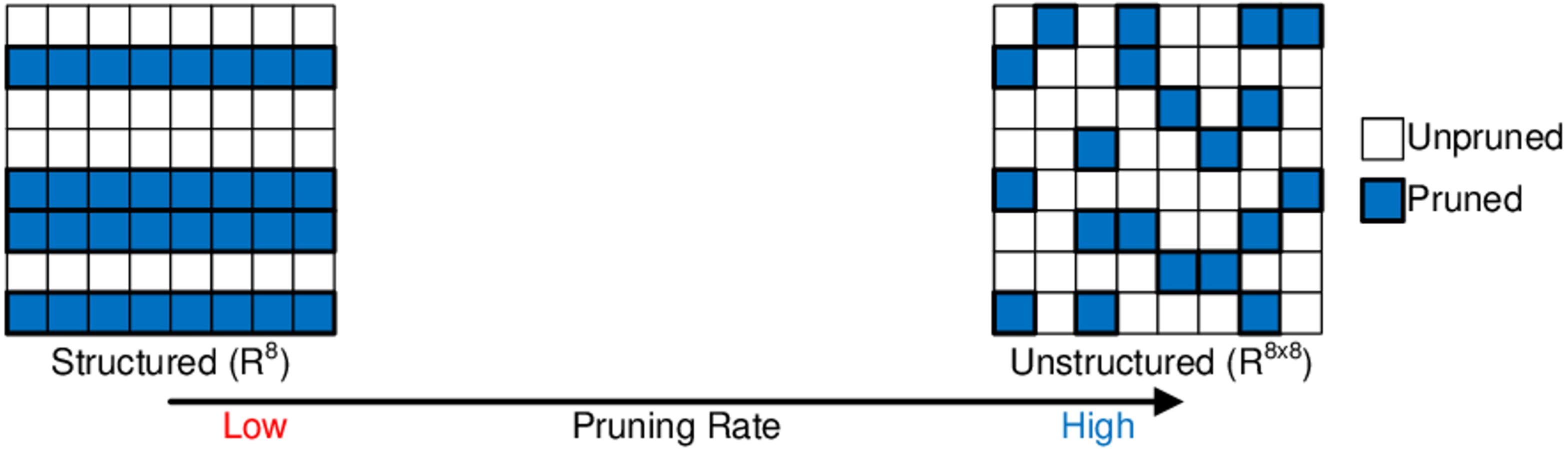
UNSTRUCTURED sparsity



STRUCTURED sparsity

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Unstructured vs structured pruning



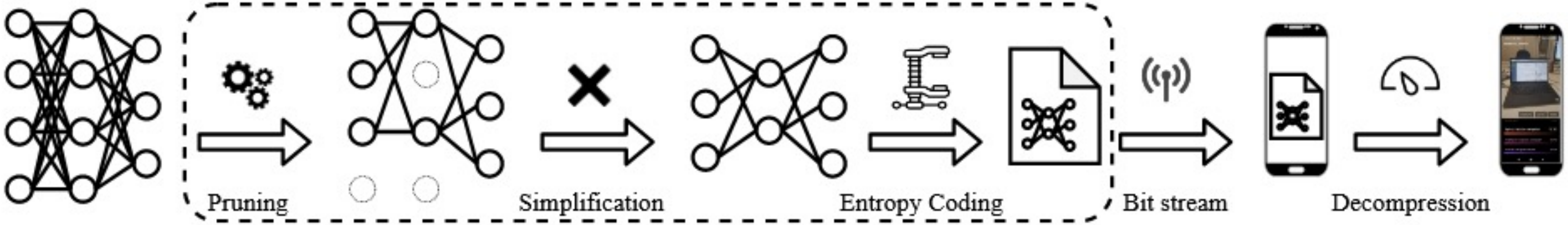
The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Pruning alone is not enough!

- Need for a real removal of the neurons/channels
- SIMPLIFY, available at https://github.com/EIDOSlab/simplify, does that!
- Paper with description in detail available at https://reader.elsevier.com/reader/sd/pii/S2352711021001576?token=8C19E9E2A04C913B545980F7567737F387925763A5A15B994AA9E2BC3D0A67D3B301B42C4791FC7F0113B420874CCF63&originRegion=eu-west-1&originCreation=20220120212108

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

The pruning pipeline for compression



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

References

- Tartaglione, E., Bragagnolo, A., Fiandrotti, A., & Grangetto, M. (2022). Loss-based sensitivity regularization: towards deep sparse neural networks. *Neural Networks*, *146*, 230-237.
- Tartaglione, E., Bragagnolo, A., Odierna, F., Fiandrotti, A., & Grangetto, M. (2021). SeReNe: Sensitivity-Based Regularization of Neurons for Structured Sparsity in Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*.
- Bragagnolo, A., & Barbano, C. A. (2022). Simplify: A Python library for optimizing pruned neural networks. *SoftwareX*, *17*, 100907.
- 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.
- Tartaglione, E., Lepsøy, S., Fiandrotti, A., & Francini, G. (2018, December). Learning sparse neural networks via sensitivity-driven regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 3882-3892).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.