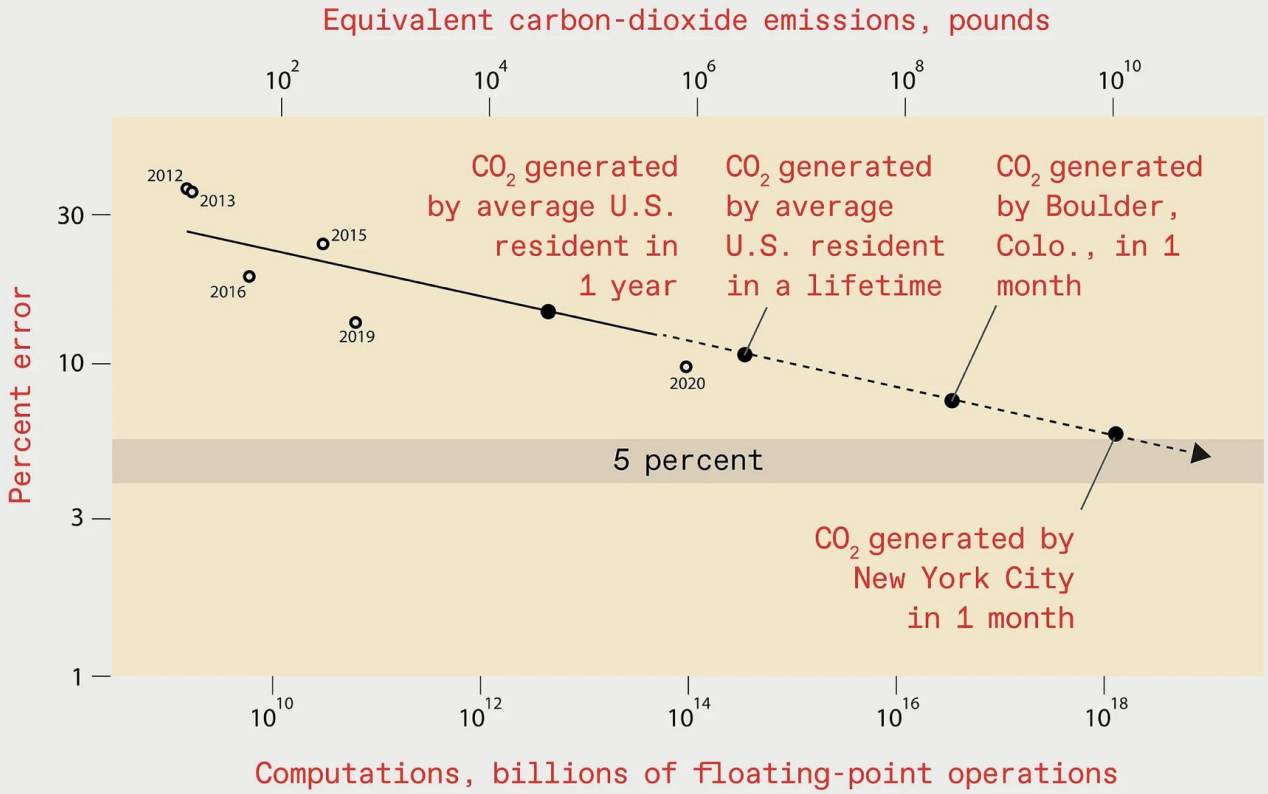
EFFICIENT NNS

28.05.2024 Computer Vision Seminar 23/24

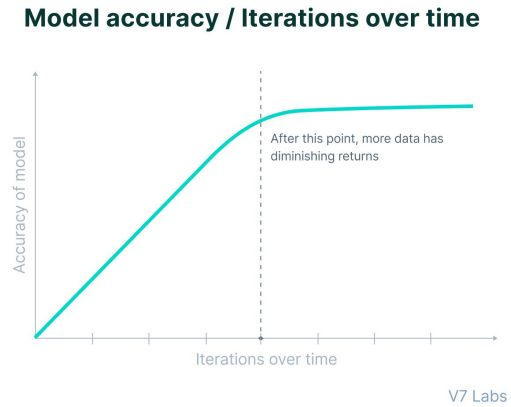
Agenda

1. Prawo malejących przychodów (diminishing returns) w DL
2. Pruning w sieciach neuronowych i jego rodzaje
3. Efektywność sieci neuronowych – praca w grupach
4. DEEP R
5. Lottery Ticket Hypothesis
6. Neptune ai



Extrapolating the gains of recent years might suggest that by 2025 the error level in the best deep-learning systems designed for recognizing objects in the ImageNet data set should be reduced to just 5 percent [top]. But the computing resources and energy required to train such a future system would be enormous, leading to the emission of as much carbon dioxide as New York City generates in one month [bottom]. SOURCE: N.C. THOMPSON, K. GREENEWALD, K. LEE, G.F. MANSO

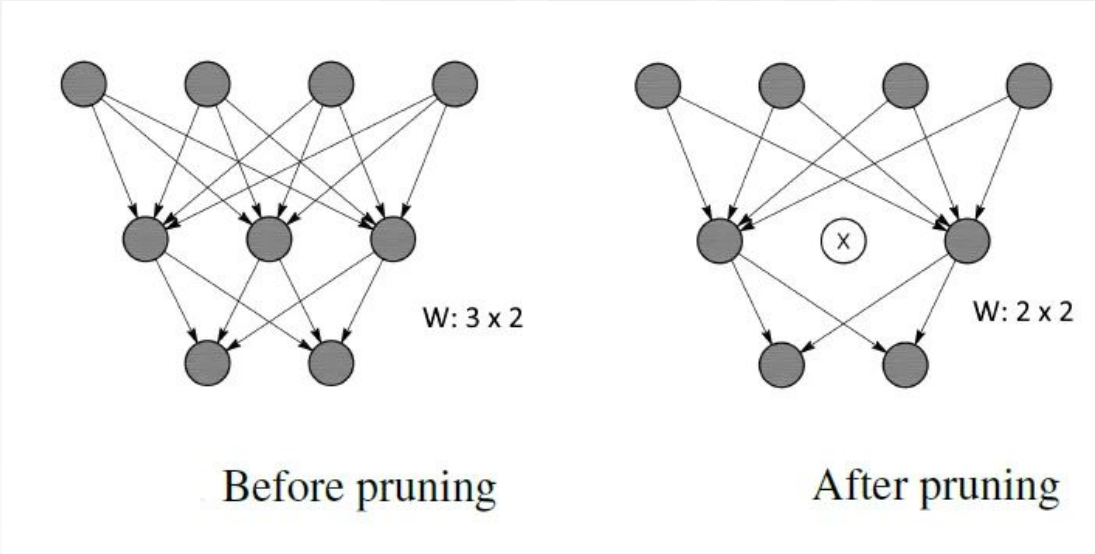
[Training Data Quality: Why It Matters in Machine Learning]



[Deep Learning's Diminishing Returns - IEEE Spectrum]

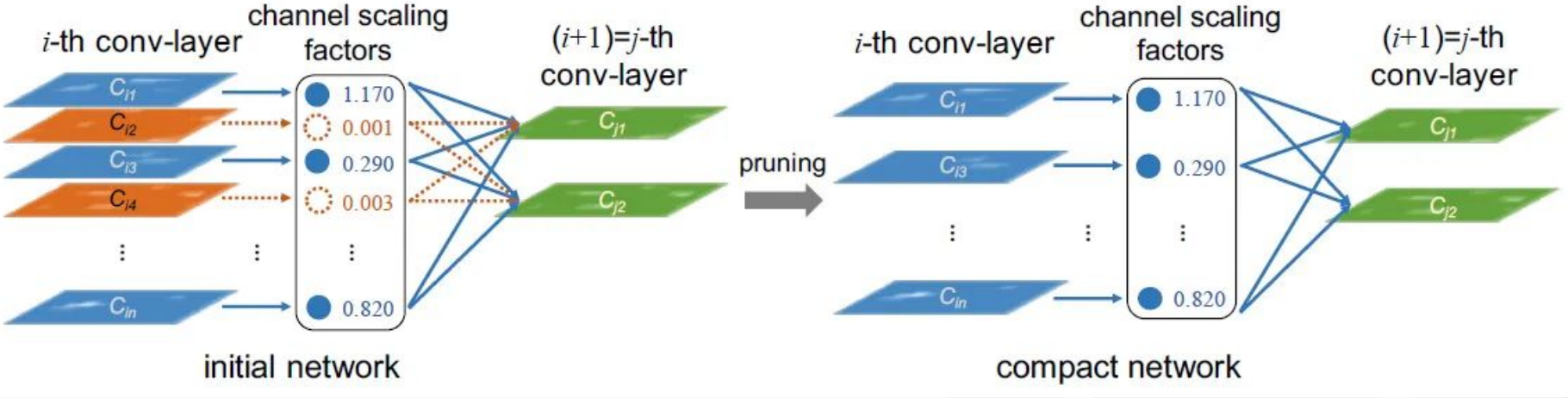
Pruning

- Usuwanie parametrów (wag) sieci
- Rodzaje:
  - Structured
  - Unstructured



[Neural Network Pruning: A Gentle Introduction | by SoonChang]

Structured pruning w CNN



https://arxiv.org/pdf/2106.08962

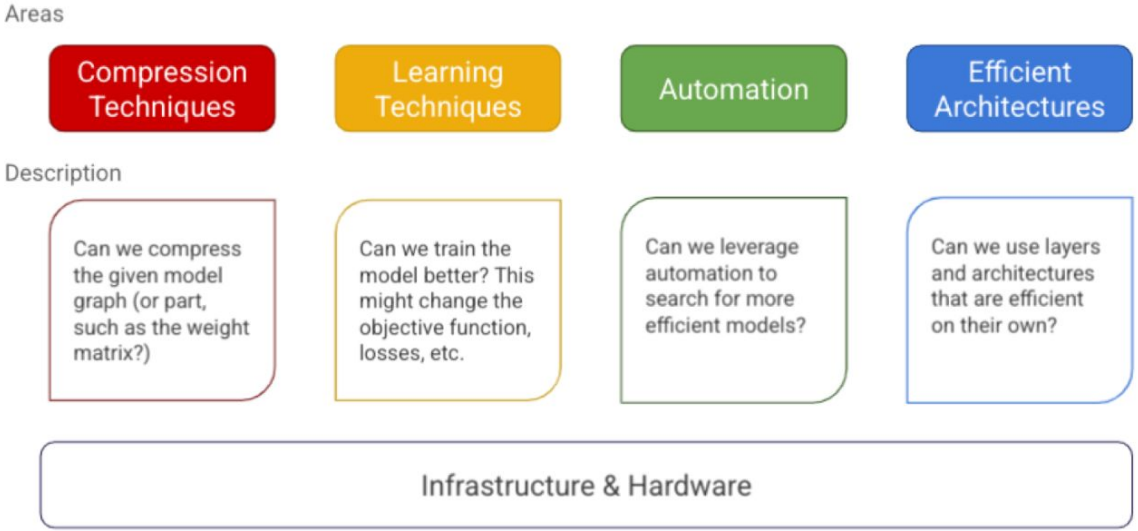


Fig. 3. A mental model for thinking about algorithms, techniques, and tools related to efficiency in Deep Learning.

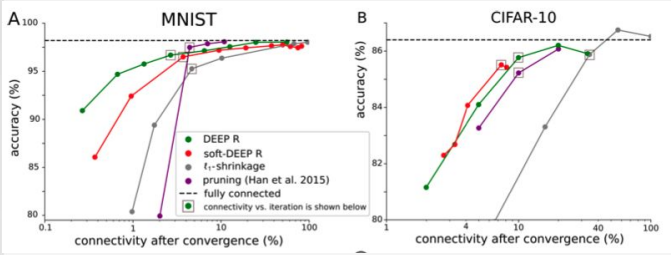
Praca w grupach

1. Grupa 1 – *3.1.1 Pruning*
2. Grupa 2 – *3.1.2 Quantization*
3. Grupa 3 – 3.2.1 Distilation
4. Grupa 4 – 3.3.1 Hyper-Parameter Optimization
5. Grupa 5 - 3.4.1 Vision – efficient architectures

Artykuł: https://arxiv.org/pdf/2106.08962

GHOST
Group of Horribly Optimistic Statisticians

DEEP R



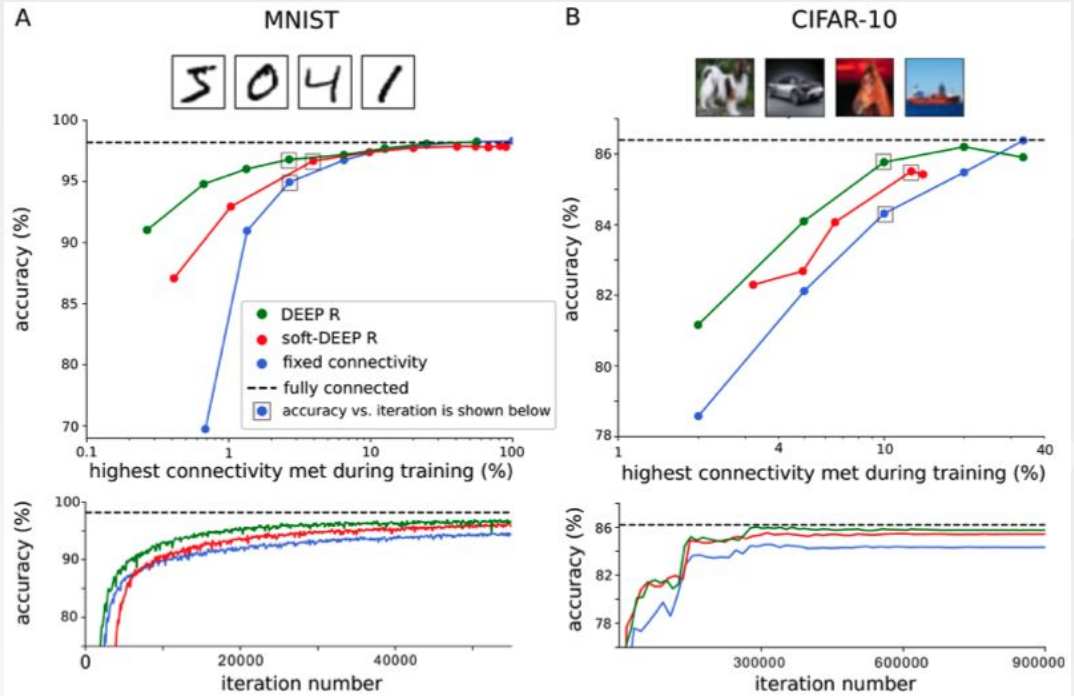


Figure 1: **Visual pattern recognition with sparse networks during training.** Sample training images (top), test classification accuracy after training for various connectivity levels (middle) and example test accuracy evolution during training (bottom) for a standard feed forward network trained on MNIST (**A**) and a CNN trained on CIFAR-10 (**B**). Accuracies are shown for various algorithms. Green: DEEP R; red: soft-DEEP R; blue: SGD with initially fixed sparse connectivity; dashed gray: SGD, fully connected. Since soft-DEEP R does not guarantee a strict upper bound on the connectivity, accuracies are plotted against the highest connectivity ever met during training (middle panels). Iteration number refers to the number of parameter updates during training.

https://openreview.net/pdf?id=BJ_wN01C-

The Lottery Ticket Hypothesis

*A randomly-initialized, dense neural network contains a **subnetwork** that is **initialized** such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

Zwycięskie bilety

Identifying winning tickets. We identify a winning ticket by training a network and pruning its smallest-magnitude weights. The remaining, unpruned connections constitute the architecture of the winning ticket. Unique to our work, each unpruned connection's value is then reset to its initialization from original network *before* it was trained. This forms our central experiment:

1. Randomly initialize a neural network $f(x; \theta_0)$ (where $\theta_0 \sim \mathcal{D}_\theta$).
2. Train the network for j iterations, arriving at parameters θ_j.
3. Prune $p\%$ of the parameters in θ_j, creating a mask m.
4. Reset the remaining parameters to their values in θ_0, creating the winning ticket $f(x; m \odot \theta_0)$.

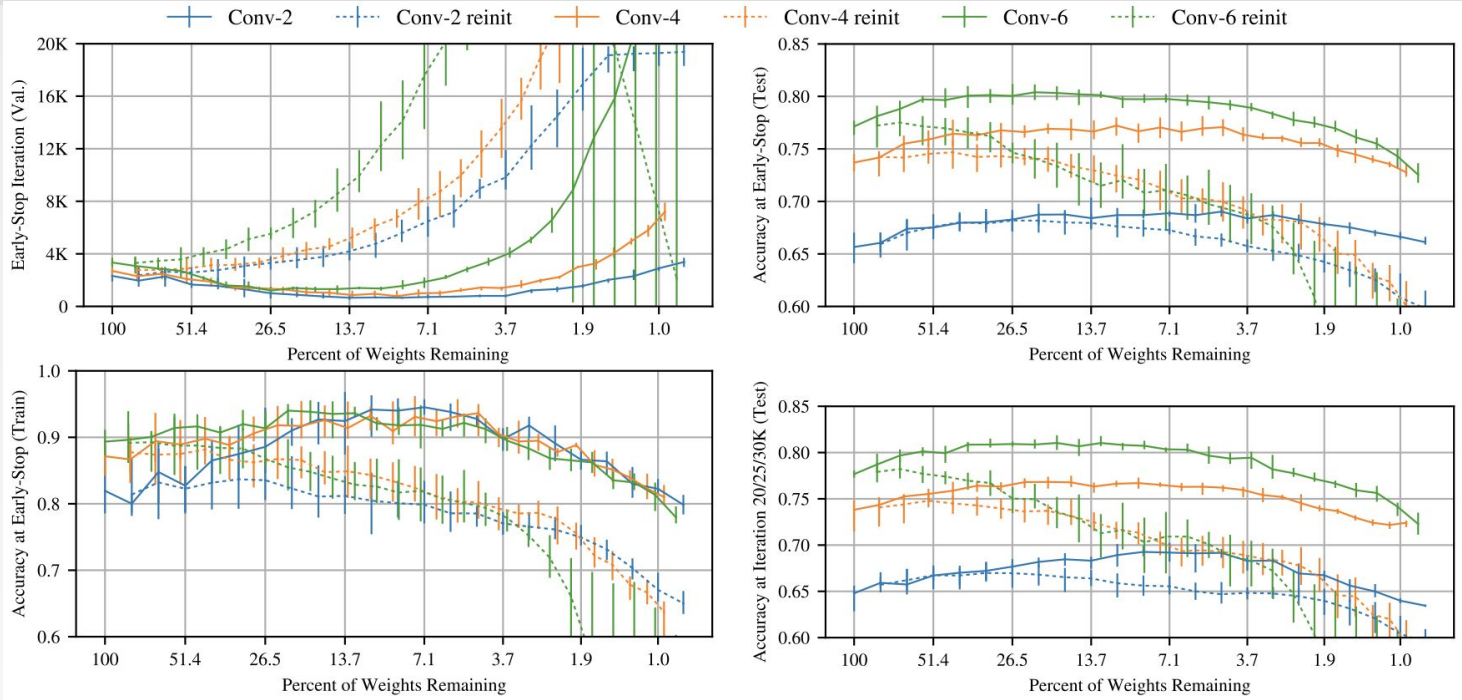


Figure 5: Early-stopping iteration and test and training accuracy of the Conv-2/4/6 architectures when iteratively pruned and when randomly reinitialized. Each solid line is the average of five trials; each dashed line is the average of fifteen reinitializations (three per trial). The bottom right graph plots test accuracy of winning tickets at iterations corresponding to the last iteration of training for the original network (20,000 for Conv-2, 25,000 for Conv-4, and 30,000 for Conv-6); at this iteration, training accuracy $\approx 100\%$ for $P_m \geq 2\%$ for winning tickets (see Appendix D).

Bilety a dropout

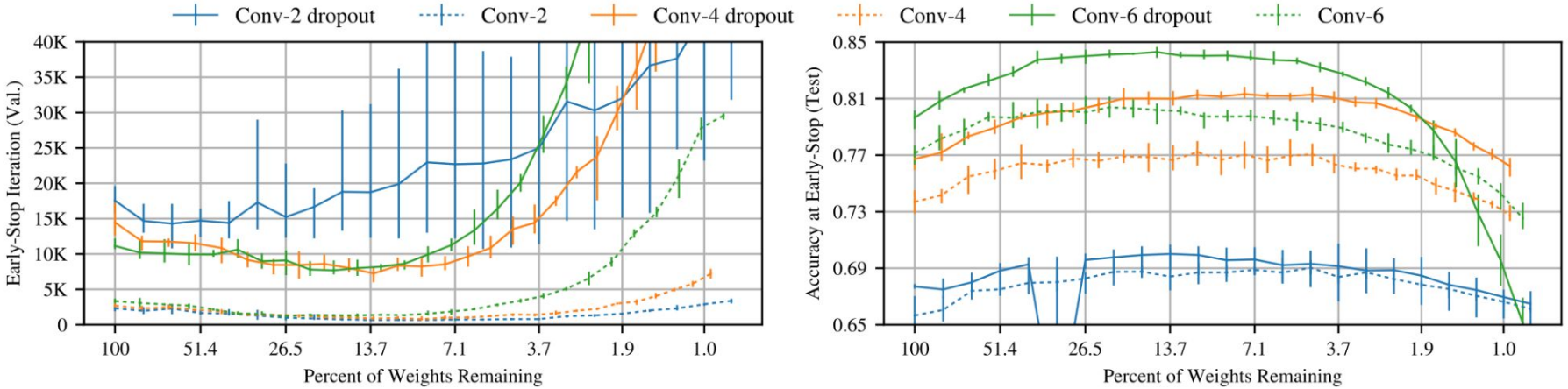


Figure 6: Early-stopping iteration and test accuracy at early-stopping of Conv-2/4/6 when iteratively pruned and trained with dropout. The dashed lines are the same networks trained without dropout (the solid lines in Figure 5). Learning rates are 0.0003 for Conv-2 and 0.0002 for Conv-4 and Conv-6.

Materiały

- https://www.datature.io/blog/a-comprehensive-guide-to-neural-network-model-pruning
- https://medium.com/@wongsirikuln/cnn-model-compression-via-pruning-461c2fd167f6
- https://www.youtube.com/watch?v=ZVVnvZdUMUk