Warsztaty Badawcze

Maciej Chylak, Dawid Janus, Arkadiusz Kniaź

April 2022

1 Related works

The method presented (*5*) in their article is based on the area of knowledge distilation, which was first introduced in the (*4*) Ba and Caruana article. They proposed idea of using teacher and student network to compress bigger network to the smaller one. Many articles were later based on their researches. Interactive Knowledge Distillation, which is the name of the method used in our article, is using three different learning tactics. In the first of them, the impact of the teacher's network is constant in time. The other one is based on linear decrease of impact of the teacher's network and the last one modifies the second method by adding periodicity. The way in which the influence of a teacher's network is differentiated is not unique. It was presented previously in the article (*1*), in which supporting the learning of neural network by using hint layers decreased over time. Many networks based on the knowledge distillation method use the division of neural networks into segments, thanks to which we are able to teach individual blocks of the student's network to map the results on their teacher's network. This article is also based on this method and the idea was taken from the (*2*) Byeongho Heo, Minsik Lee, Sangdoo Yun, Jin Young Choi work, in which authors focused on determining the boundary of neuron activation. In (*6*) K. He, X. Zhang, S. Ren and J. Sun presented residual learning framework, which is used in both Student and Teacher model. This architecture allowed to easier train deeper neural networks than previously. They use references to the layers inputs to improve performance. They show that residual networks are easier to optimize, and can gain accuracy from increased depth. In (*3*) G. Hinton, O. Vinyals and J. Dean used weighted average of two different objective functions to transfer knowledge. The first is the cross entropy with the soft targets with the same high temperature in the softmax of the distilled model as used for generating the soft targets from the cumbersome model. The second is the cross entropy with the correct labels with exactly the same logits in softmax of the distilled model but at a temperature of 1. In (*7*) S. Zagoruyko, N. Komodakis focused not only on final labels and their correctness, but also on the attention of the model. They managed to define attention for convolutional neural networks and used to improve the performance of a student, by forcing it to mimic the attention maps of teacher network. To do so they use of both activation-based and gradient-based spatial attention maps.

References

(*1*) A. Romero, Y. B., N. Ballas Fitnets: hints for thin deep nets.

(*2*) B. Heo, J. Y. C., M. Lee Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons.

(*3*) G. Hinton, J. D., O. Vinyals Distilling the knowledge in a neural network, in: NeurIPS Deep Learning and Representation Learning Workshop.

(*4*) J. Ba, R. C. Do Deep Nets Really Need to be Deep?

(*5*) J. Fu, X. Y., Z. Liu Interactive Knowledge Distillation for image classification.

(*6*) K. He, J. S., X. Zhang Deep residual learning for image recognition.

(*7*) S. Zagoruyko, N. K. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer.