REDUCED CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE RECOGNITION AND OBJECT DETECTION

LAURA MENEGHETTI

Computer Vision is a thriving field increasingly exploited in several scientific and engineering contexts in order to solve complex tasks such as the recognition and detection of objects inside pictures. A possible approach to deal with these problems is represented by Convolutional Neural Networks (CNNs), a deep learning algorithm which is able to distinguish the different features of objects thanks to its deep architecture. If on the one hand this deep and complex structure manages to solve difficult tasks, on the other it leads to a highly number of parameters that needs to be calibrated during the training phase. When it comes to practical applications, especially in industries, several issues may arise, since the aforementioned networks may need to operate in embedded systems with limited hardware. The resulting memory constraints have led to the creation of a reduction strategy for the development of a reduced version of an Artificial Neural Network and in particular of a Convolutional Neural Network.
In our work [2] we propose an extension of [1] for dimensionality reduction of CNNs by developing a reduction technique based on methods widely used in the context of Reduced Order Modeling, as Proper Orthogonal Decomposition (POD) and Active Subspaces (AS). The reduced network is obtained by splitting the original one in two different nets connected by the reduction technique: the first one obtained by retaining a certain number of layers of the original CNN and a second one that deals with the classification of the features extracted by the previous part.
We finally provide the numerical results obtained by applying such method to VGG-16 for the problem of image recognition and to SSD-300 for the problem of object detection. In particular we compare the final outcome of the original net with that of its reduced version in terms of final accuracy, memory allocation, speed of the procedure.

This is a joint work with Nicola Demo and Gianluigi Rozza.

References

[1] Cui, Chunfeng and Zhang, Kaiqi and Daulbaev, Talgat and Gusak, Julia and Oseledets, Ivan and Zhang, Zheng 2020 *Active Subspace of Neural Networks: Structural Analysis and Universal Attacks.* In SIAM Journal on Mathematics of Data Science, vol. 2, pp. 1096–1122.

[2] Laura Meneghetti and Nicola Demo and Gianluigi Rozza, 2021 *A Dimensionality Reduction Approach for Convolutional Neural Networks*, arXiv: https://arxiv.org/abs/2110.09163, submitted

Mathematics Area, mathLab, SISSA, via Bonomea 265, I-34136 Trieste, Italy
Email address: laura.meneghetti@sissa.it