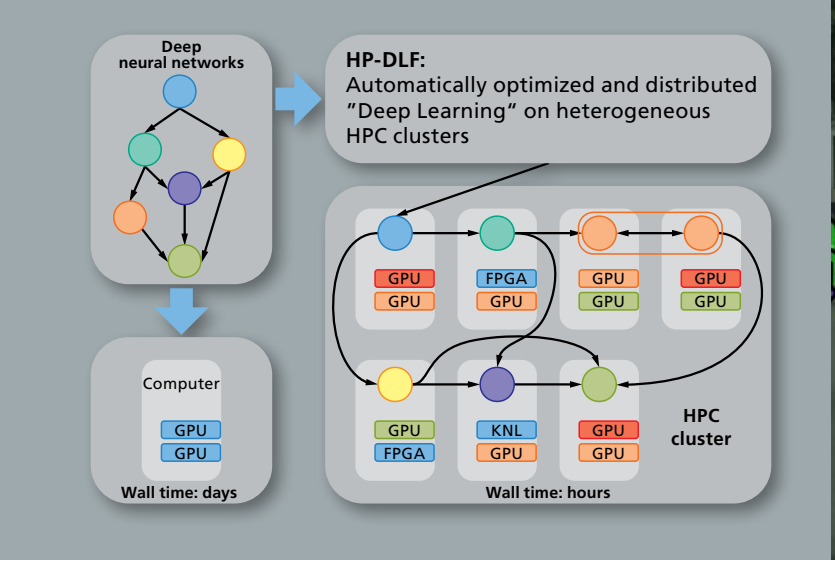

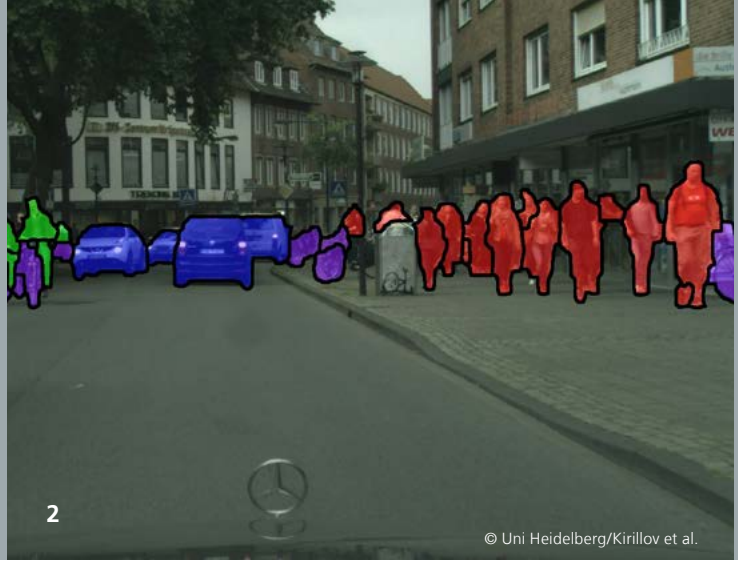

HPC FOR MACHINE LEARNING: HIGH PERFORMANCE DEEP LEARNING FRAMEWORK

Artificial neural networks have become established in many areas of machine learning in recent years. For example, they are at the leading edge of computer vision, speech and character recognition as well as machine translation. One reason for their success is the ability to create highly complex interrelationships between the raw data input and the classification (the labels) of the output data.

1 *HPC enables deep learning without storage limits.*

2 *Large amounts of data play a special role in autonomous driving.*

This often requires several million free parameters that have to be changed (i. e., learned) while training the network. Because of the large number of these so called weights, training a single neural network often takes several days or even weeks. Clearly, making these algorithms highly scalable through the use of supercomputers is highly desirable. In the ideal case, doubling the number of computers connected in parallel would halve the running time of the algorithm.

Small neural networks or fewer files?

Neural networks encounter an additional problem: they require a very large main memory. As a result, only relatively small neural networks can be trained on a single computer, or even the amount of data used for learning must be limited. Neither of these options is desirable because they reduce the capacity, i. e., the learning ability of the network. Rather, it is more desirable to train networks of twice the size with twice the number of computers. This is called "weak scalability" in the jargon of parallel computing.

High degree of scalability with GPI Space

Enabling both weak as well as strong scalability in the training of neural networks is the subject of the BMBF project "High Performance Deep Learning Framework, (HP-DLF)." A particular focus is placed on enabling the construction of neural networks of any size and ensuring easy access to existing and future high performance computing systems. No prior knowledge of parallel computing is required on the part of the user. Our in-house runtime system GPI-Space manages everything. When represented in the form of a special graph, a so-called Petri net, algorithms can be automatically and dynamically parallelized.

GPI-SPACE