As a target network, layers (Fully Connected Layers, Convolutional Layers, Recurrent Layers) before considering Pl@ntNet help to redesign the network itself to cope with bottlenecks. We will first concentrate on classical scheduling. At last, another opportunity is to develop high level simulation techniques, that could be used in particular to detect bottlenecks with respect to a DL network and to a parallel architecture.

Recently, several frameworks such as TensorFlow [1] and PyTorch [2] emerged and represent the DL network as a directed graph whose nodes represent convolution operations and edges represent data dependences between them. The goal of this PhD thesis is to work on how to allocate the convolution operations and how to schedule them to achieve a better efficiency, typically in the context of platforms consisting of heterogeneous resources such as GPUs and multicore nodes.

The goal of this PhD Thesis is to improve the scheduling and resource allocation strategies along several directions. First, the resource allocation algorithm does not take into account the specificities of the application. Indeed, it is for instance close to the default StarPU scheduling algorithm [3] used for general task graphs.

Second, it has been proved that for specific applications such as linear algebra kernels, injecting some static knowledge based on a more sophisticated scheduling algorithm can strongly improve the performance of greedy algorithm [4]. Third, in the context of DL, the same graph of convolution layers is used many times on different input data along the execution of the DL algorithm, what is close to the context of steady state scheduling [5], that has been proved to be more tractable than general scheduling. At last, another opportunity is to develop high level simulation techniques, that could be used in particular to detect bottlenecks with respect to a DL network and to a parallel architecture. This possibility could more speculatively be especially interesting in the context of DL, since it may help to redesign the network itself to cope with bottlenecks. We will first concentrate on classical layers (Fully Connected Layers, Convolutional Layers, Recurrent Layers) before considering Pl@ntNet [6] as a target network.
Main activities

These research directions require the joint knowledge of experts in deep learning algorithms, dynamic runtime scheduling and scheduling theory and will benefit in particular to Pl@ntNet application.

The PhD student will be localized in Bordeaux and will be co-supervised by Olivier Beaumont (RealOpt) and Alexis Joly (Zenith), in close collaboration with Guillaume Charpiat (Tau) and Samuel Thibault (Storm). Several stays (1 week) in Saclay and Montpellier will be scheduled during the PhD Thesis.

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport

Remuneration

1982€ / month (before taxes) during the first 2 years, 2085€ / month (before taxes) during the third year.