Towards Improving Class Parallelism for Edge Environments

Main-stream serving paradigms for distributed models, such as data parallelism and model parallelism, are not suitable when it comes to inference for tasks that require low latency and have atomic input streams. A recent effort, Sensai, proposes a new generic approach called class parallelism that aims to distribute a base convolution neural network (CNN) model across several homogeneous machines.

The model distribution paradigm decomposes a CNN into disconnected subnets, each responsible for predicting specific classes or groups of classes. They claim that this approach enables fast, in-parallel inference on live data with minimal communication overhead, significantly reducing inference latency on single data items without compromising accuracy.

Class Parallelism, however, comes with its own set of challenges and limitations. For instance, since the generated models should be created in a homogeneous manner; they share similar characteristics. Further, regardless of the input, all sub-models have to be executed to get the final prediction, which directly impacts the robustness and scalability of the system.

During the first stage of the thesis, our goal is to reproduce the results from the paper. Later, we want to improve the existing method to become more robust and possibly extend it to new use cases besides image classification. Finally, if time permits, we want to evaluate the trained models in an edge environment.

Advisors

Navidreza Asadi