TITLE OF THE INVENTION: MODEL COMPILING METHOD, APPARATUS, DEVICE, AND STORAGE MEDIUM

ABSTRACT

The present disclosure provides a model compiling method. The method includes that: a target computation sub-graph in a computation graph corresponding to a model to be compiled is obtained and the target computation sub-graph includes multiple memory-access-intensive operators and data corresponding to the multiple memory-access-intensive operators; multiple groups of computation tasks to be serially executed by a target thread are determined based on multiple threads set in an executor of the model and a computation task parallelism degree of each thread sequential rearrangement is performed on multiple computation tasks contained in the any group of computation tasks, and an execution instruction sequence both corresponding to the target thread and the any group of computation tasks is generated; and an executable program file for executing the model by the executor is determined based on the execution instruction sequence corresponding to the any group of computation tasks.

FIG. 1