ABSTRACT

Embodiments of this application provide a model training method and apparatus, and a compute device cluster. The method is applied to a training cluster. The training cluster includes a first model training group and a second model training group among a plurality of model training groups. The first model training group and the second model training group are configured to train a neural network model in parallel in a data parallel manner. The method includes: generating checkpoint data of a first training node and a second training node after one round of training ends; storing, by the first training node, the checkpoint data in a memory of the first training node, storing, by the second training node, the checkpoint data in a memory of the second training node; and if the first training node is faulty in a process of training a model, obtaining, by the first training node, the checkpoint data from the memory of the second training node. According to the model training method in embodiments of this application, consumption of communication resources can be reduced, and performance and bandwidth requirements for remote storage can be lowered, thereby significantly reducing overall system costs.

Fig. 5