

ABSTRACT

The application discloses a video memory management method, apparatus, system and device. The method includes: determining the priorities of a plurality of machine learning tasks executed by a graphics processing unit; if video memory resources are to be allocated to a high-priority task, and the amount of allocatable video memory resources is smaller than the amount of video memory resources required by the high-priority task, releasing at least a part of video memory resources occupied by a low-priority task; and allocating the video memory resources to the high-priority task, so as to execute the high-priority task at least according to tensor data in a video memory space. With such processing method, when the allocated video memory resources are insufficient, the video memory resources occupied by the low-priority task are allocated to the high-priority task for use, thereby realizing dynamic scaling optimization of the GPU video memory resources occupied by the plurality of parallel machine learning tasks on a GPU. In this way, a resource utilization rate of a whole cluster can be improved on the premise of guaranteeing the performance of the high-priority task.

FIG. 1