WHAT IS CLAIMED IS:

1. A video memory management method, comprising:

determining priorities of a plurality of machine learning tasks executed by a graphics processing unit;

if video memory resources are to be allocated for a high-priority task, and an amount of allocatable video memory resources is smaller than an 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 while continuing to execute the low-priority task; and

allocating the video memory resources released from the low-priority task to the high-priority task, so as to execute the high-priority task at least according to tensor data in a video memory space.

2. The method according to claim 1, further comprising:

releasing idle video memory resources occupied by the plurality of machine learning tasks.

3. The method according to claim 2, wherein releasing the idle video memory resources occupied by the plurality of machine learning tasks comprises:

determining usage status information of video memory resources of the plurality of machine learning tasks; and

if the information satisfies a video memory resource release condition, releasing the idle video memory resources.

4. The method according to claim 3, wherein

the usage status information comprises: an upper limit of video memory resources actually used by the task; and

the release condition comprises: a duration during which an amount of video memory resource allocated to the task is larger than the upper limit reaches a duration threshold.

5. The method according to claim 1, further comprising:

releasing idle video memory resources of other high-priority tasks.

6. The method according to claim 1, wherein releasing at least a part of the video memory resources occupied by the low-priority task comprises:

if an amount of idle video memory resources of the low-priority task is larger than or equal to an amount of the required video memory resources, releasing the idle video memory resources occupied by the low-priority task.

7. The method according to claim 1 or 6, wherein releasing at least a part of the video memory resources occupied by the low-priority task comprises:

if an amount of the idle video memory resources of the low-priority task is smaller than an amount of the required video memory resources,

allocating internal memory resources to the low-priority task; and

releasing at least a part of the video memory resources used by the low-priority task, so as to continue to execute the low-priority task at least according to tensor data in an internal memory space.

8. The method according to claim 7, further comprising:

if the amount of the allocatable video memory resources increases to an amount of the video memory resources required by the low-priority task, allocating video memory resources to the low-priority task, so as to continue to execute the low-priority task according to the tensor data in the video memory space.

9. The method according to claim 7, wherein the low-priority task comprises an iterative learning task; and releasing the video memory resources used by the low-priority task comprises:

after the low-priority task completes current iterative learning, releasing at least a part of the video memory resources used by the low-priority task.

10. The method according to claim 9, wherein before the releasing at least a part of the video memory resources used by the low-priority task, the method further comprises:

allocating internal memory resources to the high-priority task, so as to execute the high-priority task according to tensor data in the internal memory space and the tensor data in the video memory space.

11. The method according to claim 10, wherein after the allocating the video memory resources released from the low-priority task to the high-priority task, the method further comprises:

releasing the internal memory resources of the high-priority task.

12. The method according to claim 1, wherein the machine learning tasks comprise distributed deep learning tasks.

13. A video memory management method, comprising:

running a machine learning task by a graphics processing unit;

determining usage status information of video memory resources of the machine learning task; and

if the information satisfies a video memory resource release condition, releasing idle video memory resources occupied by the task while continuing to execute the task, so as to allocate the idle video memory resources to other machine learning tasks that are executed in parallel by the graphics processing unit.

14. A video memory management apparatus, comprising:

a priority determination unit configured to determine priorities of a plurality of machine learning tasks executed by a graphics processing unit;

a video memory release unit configured to release at least a part of video memory resources occupied by a low-priority task while continuing to execute the low-priority task if video memory resources are to be allocated for a high-priority task, and an amount of allocatable video memory resources is smaller than an amount of video memory resource required by the high-priority task; and

a video memory allocation unit configured to allocate the video memory resources released from the low-priority task to the high-priority task, so as to execute the high-priority task at least according to tensor data in a video memory space.

15. An electronic device, comprising a processor and a memory, wherein the memory is configured to store a program for implementing the video memory management method according to any one of claims 1 to 12, and the device is powered on and executes the program of the method by the processor.

16. A video memory management apparatus, comprising:

a task executing unit configured to execute a machine learning task by a graphics processing unit;

an information determining unit configured to determine usage status information of video memory resources of the machine learning task; and

a video memory release unit configured to release idle video memory resources occupied by the task if the information satisfies a video memory resource release condition while continuing to execute the task, so as to allocate the idle video memory resources to other machine learning tasks that are executed in parallel by the graphics processing unit.

17. An electronic device, comprising:

a processor and a memory, wherein

the memory is configured to store a program for implementing the video memory management method according to claim 13, and the device is powered on and executes the program of the method by the processor.

18. A video memory management system, comprising:

a storage resource coordinator configured to:

determine priorities of a plurality of machine learning tasks that are executed by a graphics processing unit; and

if video memory resources are to be allocated for a high-priority task, and an amount of allocatable video memory resources is smaller than an amount of video memory resources of the high-priority task,

send a video memory resource release instruction to a first storage resource allocator of low-priority tasks; and

send a video memory resource allocation instruction to a second storage resource allocator of high-priority tasks;

a first storage resource allocator configured to release at least a part of video memory resources occupied by a low-priority task according to the release instruction while continuing to execute the low-priority task; and

a second storage resource allocator configured to allocate the video memory resources released from the low-priority task to the high-priority task according to the allocation instruction, so as to execute the high-priority task at least according to tensor data in a video memory space.

19. The system according to claim 18, wherein the allocator is further configured to send usage status information of the video memory resources of the tasks; and

the coordinator is further configured to send the video memory resource release instruction to the allocator if the information satisfies a video memory resource release condition.

20. The system according to claim 19, wherein the allocator is further configured to send the information to the coordinator according to a preset period.

21. A machine learning system, comprising:

a client configured to send priority information of machine learning tasks to a server; and

a server configured to determine priorities of the 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 an amount of allocatable video memory resources is smaller than an amount of video memory resources required by the high-priority task, release at least a part of video memory resources occupied by a low-priority task while continuing to execute the low-priority task; and allocate the video memory resources released from the low-priority task to the high-priority task, so as to execute the high-priority task at least according to tensor data in a video memory space.