

INFERENCE COMPUTATION METHODS AND APPARATUSES FOR LARGE MODELS

ABSTRACT

Embodiments of this specification provide inference computation methods and apparatuses
5 for large models. In a method, a large model includes N computation layers. When performing
inference computation, a graphics processing unit (GPU) requests first video memory space for
partial computation layers of the N computation layers, performs inference computation on the
partial computation layers, and stores obtained inference result data in the first video memory
space. Then the inference result data are dumped from the first video memory space to CPU
10 memory, where the released video memory space can be provided for other computation layers to
perform inference computation. When the inference result data are needed, the inference result
data in the CPU memory are transmitted to GPU video memory. When data processed by the large
model include privacy data, privacy protection needs to be implemented for the data during the
processing.

15

FIG. 3