ABSTRACT

A communication method, apparatus, and system are provided, and relate to the field of communication technologies. The method includes: receiving first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, and N is a positive integer; and performing inference based on the N first modules according to the first active parameter. According to the foregoing solution, a first apparatus for inference does not need to store full parameters of the first module, but may obtain active parameters in the full parameters of the first module through the first information. Compared with a solution in which the full parameters are stored locally in the first apparatus and the active parameters are read from the full parameters, the foregoing solution can save storage space for the first apparatus, so that the first apparatus with limited storage space can run a large model with a large parameter quantity, thereby improving inference performance of the large model.

FIG. 4