

ABSTRACT

This application discloses a computational graph inference method and system, and a storage medium. The method includes: monitoring a to-be-inferred original-type computational graph; in response to a failure in converting the monitored original-type computational graph into a corresponding target-type computational graph, dividing the
5 original-type computational graph into first-level original-type computational subgraphs; converting the first-level original-type computational subgraph into at least one corresponding target-type computational graph; and generating a to-be-inferred file by using the target-type computational graph corresponding to the first-level original-type
10 computational subgraph, and sending the to-be-inferred file to a remote device. A technical problem of low inference efficiency of computational graphs is resolved in this application.

Fig. 4