

ABSTRACT

A method of optimizing a neural network model includes receiving original model information about a first neural network model that is pre-trained; generating a second neural network model and compressed model information about the second neural network model by performing a compression on the first neural network model; and outputting, on a screen, at least a part of the original model information and at least a part of the compressed model information.

FIG. 1