What is Claimed is:

1. A method of optimizing a neural network model, the method comprising:

receiving original model information about a first neural network model that is pre-trained, the first neural network model including a plurality of original layers;

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, the second neural network model including a plurality of compressed layers; and

outputting, on a screen, at least a part of the original model information and at least a part of the compressed model information,

wherein the outputting comprises:

displaying a first graphical representation on a graphical user interface (GUI), the first graphical representation including a comparison of a first characteristic associated with the plurality of original layers and a second characteristic associated with the plurality of compressed layers;

scoring the plurality of compressed layers for an operating efficiency, including generating a plurality of score values for the plurality of compressed layers;

outputting a result of the scoring, including displaying, on the GUI, a second graphical representation, in which the plurality of compressed layers are displayed according to different schemes based on the plurality of score values;

receiving, through the GUI, a selection of one or more of the plurality of compressed layers to be modified;

displaying, on the GUI, a menu providing a recommendation list detailing one or more candidate layers, one or more candidate structures and/or one or more candidate settings for modifying or replacing the selected one or more of the plurality of compressed layers;

receiving, through the GUI, a selection from the recommendation list; and

modifying or replacing the selected one or more of the plurality of compressed layers according to the selection from the recommendation list to produce a modified second neural network model.

2. The method of claim 1, wherein the outputting comprises:

displaying, via the graphical user interface (GUI), the at least the part of the original model information and the at least the part of the compressed model information.

3. The method of claim 1 or claim 2, wherein the first characteristic and the second characteristic are displayed by at least one of a unit of a layer or a unit of a channel.

4. The method of any one of claims 1 to 3, wherein the outputting further includes:

displaying a third graphical representation on the GUI, the third graphical representation including a network structure of the plurality of compressed layers; and

receiving, through the GUI, a first user input with respect to a first compressed layer among the plurality of compressed layers, and

wherein, based on the first user input, the first graphical representation is displayed such that a comparison of a characteristic of a first original layer among the plurality of original layers and a characteristic of the first compressed layer is displayed, the first original layer corresponding to the first compressed layer.

5. The method of claim 4, wherein:

the third graphical representation includes a plurality of layer boxes corresponding

to the plurality of compressed layers,

a first layer box that corresponds to a compressed layer satisfying a predetermined reference value is displayed according to a first scheme, and

a second layer box that corresponds to a compressed layer not satisfying the predetermined reference value is displayed according to a second scheme different from the first scheme.

6. The method of claim 5, wherein the predetermined reference value is associated with at least one of a signal-to-quantization-noise power ratio (SQNR), a latency, a power consumption, or a utilization.

7. The method of any one of claims 4 to 6, wherein the outputting further includes:

receiving, through the GUI, a second user input for grouping the plurality of compressed layers; and

displaying a fourth graphical representation on the GUI based on the second user input, the fourth graphical representation including a plurality of compressed layer groups each of which includes at least one of the plurality of compressed layers.

8. The method of claim 7, wherein the outputting further includes:

receiving, through the GUI, a third user input with respect to a first compressed layer group among the plurality of compressed layer groups; and

displaying a fifth graphical representation on the GUI based on the third user input, the fifth graphical representation including at least one compressed layer included in the first compressed layer group.

9. The method of any one of claims 4 to 8, wherein the outputting further includes:

receiving, through the GUI, a fourth user input for selecting at least one target device to execute the plurality of compressed layers; and

displaying a sixth graphical representation on the GUI based on the fourth user input, the sixth graphical representation indicating whether the plurality of compressed layers are suitable for the selected at least one target device.

10. The method of claim 9, wherein the at least one target device includes at least one of a central processing unit (CPU), a neural processing unit (NPU), a graphic processing unit (GPU), a digital signal processor (DSP), or an image signal processor (ISP).

11. The method of any one of claims 1 to 3, further comprising:

changing a setting of at least one compressed layer of the plurality of compressed layers, and outputting a result of the changing the setting.

12. The method of claim 11, further comprising:

receiving, through the GUI, a fifth user input for changing the setting of the at least one compressed layer of the plurality of compressed layers; and

updating the second characteristic based on the fifth user input,

wherein the outputting the result of the changing the setting includes displaying a seventh graphical representation on the GUI, the seventh graphical representation including a comparison of the first characteristic and the updated second characteristic.

13. The method of any one of claims 1 to 3, wherein:

the second graphical representation includes a plurality of layer boxes corresponding to the plurality of compressed layers,

a first layer box that corresponds to a compressed layer having a score value greater than a reference score value is displayed according to a first scheme, and

a second layer box that corresponds to a compressed layer having a score value less than or equal to the reference score value is displayed according to a second scheme different from the first scheme.

14. The method of claim 13, wherein the scoring includes scoring the plurality of score values based on at least one of a result of estimating a compression performance of the plurality of compressed layers, a result of determining whether the plurality of compressed layers are suitable for a target device, types of the plurality of compressed layers, a result of predicting capacity of the plurality of compressed layers, or a memory footprint utilization of the plurality of compressed layers.

15. The method of any one of claims 1 to 3, 13 and 14, further comprising:

changing at least one of the plurality of compressed layers based on the result of the scoring.

16. A method of optimizing a neural network model, the method comprising:

receiving original model information about a first neural network model that is pre-trained and includes a plurality of original layers;

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, the second neural network model including a plurality of compressed layers;

displaying a first graphical representation on a graphical user interface (GUI), the first graphical representation including a network structure of the plurality of compressed layers;

receiving, through the GUI, a first user input with respect to a first compressed layer among the plurality of compressed layers;

displaying a second graphical representation on the GUI based on the first user input, the second graphical representation including a comparison of a characteristic of a first original layer among the plurality of original layers and a characteristic of the first compressed layer, the first original layer corresponding to the first compressed layer;

receiving, through the GUI, a second user input for changing a setting of a second compressed layer among the plurality of compressed layers;

updating a characteristic of the second compressed layer based on the second user input;

displaying a third graphical representation on the GUI based on the second user input, the third graphical representation including a comparison of a characteristic of a second original layer among the plurality of original layers and the updated characteristic of the second compressed layer, the second original layer corresponding to the second compressed layer;

generating a plurality of score values for the plurality of compressed layers;

displaying a fourth graphical representation on the GUI, the fourth graphical representation including the plurality of compressed layers that are displayed according to different schemes based on the plurality of score values;

displaying a fifth graphical representation on the GUI, the fifth graphical representation including the plurality of compressed layers at least one of which is

changeable based on the plurality of score values;

receiving, through the GUI, a selection of one or more of the plurality of compressed layers to be modified;

displaying, on the GUI, a menu providing a recommendation list detailing one or more candidate layers, one or more candidate structures and/or one or more candidate settings for modifying or replacing the selected one or more of the plurality of compressed layers;

receiving, through the GUI, a selection from the recommendation list; and

modifying or replacing the selected one or more of the plurality of compressed layers according to the selection from the recommendation list to produce a modified second neural network model.

17. A method of providing a graphical user interface (GUI) related to optimizing a neural network model, the method comprising:

receiving first model information about a first neural network model that is pre-trained, the first neural network model including a plurality of original layers;

generating a second neural network model and second model information about the second neural network model by performing a data processing on the first neural network model, the second neural network model including a plurality of compressed layers; and

providing a graphical user interface (GUI), which displays a graphical representation including at least a part of the first model information and at least a part of the second model information in comparison with each other,

wherein the providing comprises:

displaying a first graphical representation on the GUI, the first graphical representation including a comparison of a first characteristic associated with the plurality

of original layers and a second characteristic associated with the plurality of compressed layers;

generating a plurality of score values for the plurality of compressed layers;

displaying, on the GUI, a second graphical representation, in which the plurality of compressed layers are displayed according to different schemes based on the plurality of score values;

receiving, through the GUI, a selection of one or more of the plurality of compressed layers to be modified;

displaying, on the GUI, a menu providing a recommendation list detailing one or more candidate layers, one or more candidate structures and/or one or more candidate settings for modifying or replacing the selected one or more of the plurality of compressed layers;

receiving, through the GUI, a selection from the recommendation list; and

modifying or replacing the selected one or more of the plurality of compressed layers according to the selection from the recommendation list to produce a modified second neural network model.