CLAIMS

1. A neural network circuit device including at least an input layer, one or more intermediate layers, and an output layer, the neural network circuit device comprising, in the one or more intermediate layers:

an input value Xi to which convolution is applied;

a convolution operation circuit configured to receive a weight Wi and to perform a convolution operation thereof;

a sum circuit configured to take a sum of operation values each subjected to the convolution operation and a bias W0; and

an activating function circuit configured to convert a signal Y generated by taking the sum, using an activating function f(u),

wherein the convolution operation circuit is configured to skip a weight Wi having a zero weight and to perform a convolution operation based on a nonzero weight and an input value Xi corresponding thereto,

wherein the convolution operation circuit includes a third storage part that stores therein a nonzero weight and an absolute address of an input value Xi to which convolution is applied, and

wherein the convolution operation circuit is configured to, in performing an operation in which a zero weight is skipped,

specify an absolute address of interest in the third storage part and retrieve

a nonzero weight and an input value Xi each corresponding thereto, and

perform a convolution operation based on the retrieved input value Xi and the nonzero weight corresponding thereto.

2. A neural network circuit device including at least an input layer, one or more intermediate layers, and an output layer, the one or more intermediate layer including a first intermediate layer and a second intermediate layer, the neural network circuit device comprising, in each of the first intermediate layer and the second intermediate layer:

a convolution operation circuit configured to receive a weight Wi and perform a convolution operation thereof;

a sum circuit configured to take a sum of operation values each subjected to the convolution operation and a bias W0; and

an activating function circuit configured to convert a signal Y generated by taking the sum, using an activating function f(u),

wherein, in the convolution operation circuit in the first intermediate layer, a tri-state {-multibit, 0, +multibit} weight Wi skips a zero weight, and a convolution operation is performed based on a nonzero weight and an input value Xi corresponding thereto, and

wherein, in the convolution operation circuit in the second intermediate layer, a

ternary {-1, 0, +1} or a binary {-1, +1} weight Wi skips a zero weight, and a convolution operation is performed based on a nonzero weight and an input value Xi corresponding to the nonzero weight,

wherein the convolution operation circuit includes a third storage part that stores therein a nonzero weight and an absolute address of an input value Xi to which convolution is applied,

wherein the convolution operation circuit is configured to, in performing an operation in which a zero weight is skipped,

specify an absolute address of interest in the third storage part and retrieve

a nonzero weight and an input value Xi each corresponding thereto, and

perform a convolution operation based on the retrieved input value Xi and the nonzero weight corresponding thereto, and

wherein the weight Wi in the convolution operation circuit is expressed by Formula as follows:

$$W_i = \begin{cases} -1 & W_{hid} < \rho \\ 0 & -\rho \leqq W_{hid} < \rho \\ +1 & \rho \leqq W_{hid} \end{cases} \quad \cdots(2)$$

wherein Whid: multibit weight Wi, and ρ: threshold, except when p>0 or except when p>0 and p≤Whid<p.

3. The neural network circuit device according to claim 1,

wherein the convolution operation circuit is configured to perform a convolution operation of a tri-state {-W (multibit), 0, +W (multibit} weight Wi and an input value Xi, and

wherein the weight Wi in the convolution operation circuit is expressed by Formula as follows:

$$W_i = \begin{cases} -1 & W_{hid} < \rho \\ 0 & -\rho \leqq W_{hid} < \rho \\ +1 & \rho \leqq W_{hid} \end{cases} \quad \cdots(2)$$

wherein Whid: multibit weight Wi, and ρ: threshold, except when p>0 or except when p>0 and p≤Whid<p.

4. The neural network circuit device according to claim 1,

wherein the weight Wi in the convolution operation circuit is expressed by Formula as follows:

$$W_i = \begin{cases} W_{hid} & W_{hid} < \rho \\ 0 & -\rho \leqq W_{hid} < \rho \\ W_{hid} & \rho \leqq W_{hid} \end{cases} \quad \cdots(3)$$

wherein, Whid: multibit weight Wi, and ρ: threshold, except when p>0 or except when p>0 and p≤Whid<p.

5. The neural network circuit device according to claim 1,

wherein the convolution operation circuit is configured to perform a convolution operation of a ternary {-1, 0, +1} weight Wi and an input value Xi

wherein the weight Wi in the convolution operation circuit is expressed by Formula as follows:

$$W_i = \begin{cases} -1 & W_{hid} < \rho \\ 0 & -\rho \leqq W_{hid} < \rho \\ +1 & \rho \leqq W_{hid} \end{cases} \quad \cdots(2)$$

wherein Whid: multibit weight Wi, and ρ: threshold, except when p>0 or except when p>0 and p≤Whid<p.

6. The neural network circuit device according to claim 1 or claim 2,

wherein the convolution operation circuit includes: a first storage part that stores therein a nonzero weight and a relative address of an input value X to which convolution is applied; and a second storage part that stores therein an address of the input value Xi, and

wherein the convolution operation circuit is configured to, in performing an operation in which a zero weight is skipped,

retrieve a nonzero weight and a relative address corresponding thereto, from the first storage part,

reference the second storage part and retrieve therefrom a subsequent input value X to which convolution is applied, using the retrieved relative address and a current address, and

perform a convolution operation based on the input value Xi retrieved from the second storage part and a nonzero weight corresponding thereto.

7. The neural network circuit device according to claim 1 or claim 2,

wherein, when learning is repeated in which: an error between an output

obtained by forward propagation in a neural network and a training data is calculated; and the calculated error is subjected to backward propagation therein, based on which the weight Wi is updated, then, the weight Wi is not subjected to rounding at least in the intermediate layer, until the learning is terminated.

8. The neural network circuit device according to claim 1 or claim 2,

wherein, when learning is repeated in which: an error between an output obtained by forward propagation in a neural network and a training data is calculated; and the calculated error is subjected to backward propagation therein, based on which the weight Wi is updated, then, after the learning is terminated, the weight Wi having been maintained without being subjected to rounding is made to zero, to thereby create a neural network having sparse weights.

9. A neural network processing method of a neural network circuit device including at least an input layer, one or more intermediate layers, and an output layer, the neural network circuit method comprising, in the one or more intermediate layers:

a convolution operation step of receiving an input value X to which convolution is applied, and a weight Wi, and performing a convolution operation thereof;

a step of taking a sum of operation values each subjected to the convolution operation and a bias W0; and

a step of converting a signal Y generated by taking the sum, using an activating function f(u),

wherein, in the convolution operation step, a weight Wi having a zero weight is skipped, and a convolution operation is performed based on a nonzero weight and an input value Xi corresponding thereto,

the neural network circuit method: further comprising,

a step of, using a third storage part configured to store therein a nonzero weight and an absolute address of an input value Xi to which convolution is applied, in performing an operation in which a zero weight is skipped, specifying an absolute address of interest in the third storage part and retrieving a nonzero weight and an input value Xi each corresponding thereto, and

a step of performing a convolution operation based on the retrieved input value Xi and the nonzero weight corresponding thereto.

10. The neural network processing method according to claim 9,

wherein, when learning is repeated in which: an error between an output obtained by forward propagation in a neural network and a training data is calculated; and the calculated error is subjected to backward propagation therein, based on which the weight Wi is updated, then, the weight Wi is not subjected to rounding at least in the intermediate layer, until the learning is terminated.

11. The neural network processing method according to claim 9,

wherein, when learning is repeated in which: an error between an output obtained by forward propagating in a neural network and a training data is calculated; and the calculated error is subjected to backward propagation therein, based on which the weight Wi is updated, then, after the learning is terminated, the weight Wi having been maintained without being subjected to rounding is made to zero, to thereby create a neural network having sparse weights.

12. A neural network execution program causing a computer to serve as a neural network circuit device, the neural network circuit device including at least an input layer, one or more intermediate layers, and an output layer, the neural network execution program realizing functions, in the intermediate layer, of:

a convolution operation circuit part configured to receive an input value X to which convolution is applied, and a weight Wi, and to perform a convolution operation thereof;

a sum circuit part configured to take a sum of operation values each subjected to the convolution operation and a bias W0;

an activating function circuit part configured to convert a signal Y generated by taking the sum, using an activating function f(u); and

a third storage part configured to store therein a nonzero weight and an absolute address of an input value Xi to which convolution is applied,

wherein the convolution operation circuit part is configured to skip a weight Wi having a zero weight and to perform a convolution operation based on a nonzero weight and an input value Xi corresponding thereto, and

wherein the convolution operation circuit is configured to, in performing an operation in which a zero weight is skipped,

specify an absolute address of interest in the third storage part and retrieve

a nonzero weight and an input value Xi each corresponding thereto, and

perform a convolution operation based on the retrieved input value Xi and the nonzero weight corresponding thereto.