What is Claimed is:

1. An artificial neural network (ANN) quantization method for generating an output ANN implemented in a mobile computing device by quantizing an input ANN, the ANN quantization method comprising:

obtaining second parameters by quantizing first parameters of the input ANN;

obtaining a sample distribution from an intermediate ANN in which the obtained second parameters have been applied to the input ANN; and

obtaining a fractional length for the obtained sample distribution by quantizing the obtained sample distribution,

wherein the obtaining the fractional length for the obtained sample distribution comprises:

splitting the obtained sample distribution into a plurality of sample groups;

approximating the plurality of sample groups to a plurality of continuous probability distributions (CPDs), respectively;

obtaining a plurality of step sizes, based on the plurality of sample groups and the plurality of CPDs; and

selecting the fractional length, based on the obtained plurality of step sizes.

2. The ANN quantization method of claim 1, wherein the obtaining the second parameters comprises:

obtaining quantized parameters by quantizing the first parameters according to a given fractional length, and calculating quantization errors between the first parameters and the obtained quantized parameters;

calculating an evaluation value of the given fractional length, based on the calculated quantization errors; and

obtaining a final fractional length for the second parameters, based on a plurality of evaluation values corresponding to a plurality of fractional lengths.

3. The ANN quantization method of claim 2, wherein:

the calculating the evaluation value comprises calculating, as the evaluation value, a sum of squares of the calculated quantization errors; and

the obtaining the final fractional length comprises determining, as the final fractional length, a fractional length corresponding to a minimum evaluation value from among the plurality of evaluation values.

4. The ANN quantization method of any one of claims 1 to 3, wherein:

the input ANN comprises layers and channels, each having at least one parameter; and

the first parameters comprise one or more parameters from among the at least one parameter of each of the layers and the channels.

5. The ANN quantization method of claim 4, wherein the first parameters comprise at least one of weights, biases, and thresholds.

6. The ANN quantization method of any one of claims 1 to 5, wherein the splitting the obtained sample distribution into the plurality of sample groups comprises splitting the obtained sample distribution into a first sample group including negative samples and zero and a second sample group including positive samples.

7. The ANN quantization method of any one of claims 1 to 5, wherein:

the splitting of the sample distribution into the plurality of sample groups comprises splitting the sample distribution into a first sample group and a second sample group; and

the first sample group includes negative samples, and the second sample group includes zero and positive samples.

8. The ANN quantization method of any one of claims 1 to 5, wherein:

the splitting the sample distribution into the plurality of sample groups comprises splitting the sample distribution into a first sample group and a second sample group; and

the first sample group includes negative samples, and the second sample group includes positive samples.

9. The ANN quantization method of any one of claims 1 to 8, wherein the approximating the plurality of sample groups comprises approximating each of the plurality of sample groups to a generalized gamma distribution, a Gaussian distribution, or a Laplacian distribution.

10. The ANN quantization method of any one of claims 1 to 9, wherein the obtaining the plurality of step sizes comprises:

obtaining an overload distortion and a granular distortion according to a step size for each of the plurality of CPDs; and

obtaining each of the plurality of step sizes, based on the obtained overload distortion and the obtained granular distortion.

11. The ANN quantization method of any one of claims 1 to 10, wherein the selecting the fractional length comprises:

obtaining candidate fractional lengths, based on the obtained plurality of step sizes; and

selecting, as the fractional length, one from among the obtained candidate fractional lengths.

12. The ANN quantization method of claim 11, wherein the obtaining the candidate fractional lengths comprises:

obtaining fractional lengths corresponding to step sizes adjacent to the obtained plurality of step sizes; and

determining a range of fractional lengths, based on the obtained fractional lengths.

13. The ANN quantization method of claim 12, wherein:

the determining the range of the fractional lengths comprises:

determining, as a lower limit of the range, a value obtained by subtracting a first margin from a minimum fractional length among the obtained fractional lengths, and

determining, as an upper limit of the range, a value obtained by adding a second margin from a maximum fractional length among the obtained fractional lengths; and

the first margin and the second margin are determined based on a performing ability of the ANN quantization method.

14. The ANN quantization method of claim 12 or claim 13, wherein the selecting the one from among the obtained candidate fractional lengths comprises:

calculating errors corresponding to the obtained candidate fractional lengths, based on the obtained candidate fractional lengths and the plurality of CPDs; and

selecting the one candidate fractional length, based on the calculated errors.

15. The ANN quantization method of claim 12 or claim 13, wherein the selecting the one from among the obtained candidate fractional lengths comprises:

obtaining test sample distributions from test ANNs respectively depending on the obtained candidate fractional lengths;

calculating errors corresponding to the obtained test sample distributions, based on the obtained test sample distributions and the plurality of CPDs; and

selecting the one candidate fractional length, based on the calculated errors.

16. The ANN quantization method of any one of claims 1 to 15, further comprising obtaining a fixed-point ANN as the output ANN, based on the obtained second parameters and the obtained fractional length.

17. An apparatus for generating an output artificial neural network (ANN) implemented in a mobile computing device by quantizing an input ANN, the apparatus comprising:

a memory storing computer-executable instructions; and

at least one processor configured to execute the stored computer-executable instructions to implement:

a parameter quantizer configured to obtain second parameters by quantizing first parameters of the input ANN;

a neural network interface configured to obtain a sample distribution from an intermediate ANN in which the obtained second parameters have been applied to the input ANN; and

a sample quantizer configured to obtain a fractional length for the obtained sample distribution,

wherein the obtaining the fractional length for the obtained sample distribution comprises:

splitting the obtained sample distribution into a plurality of sample groups;

approximating the plurality of sample groups to a plurality of continuous probability distributions (CPDs), respectively;

obtaining a plurality of step sizes, based on the plurality of sample groups and the plurality of CPDs; and

selecting the fractional length, based on the obtained plurality of step sizes.

18. The apparatus of claim 17, wherein:

the neural network interface is further configured to provide at least one candidate fractional length for the sample distribution, received from the sample quantizer, to a test ANN, and to obtain a test sample distribution from the test ANN; and

the sample quantizer is further configured to determine one from among the at least one candidate fractional length as the fractional length for the sample distribution, based on the obtained test sample distribution.

19. A method for generating a fixed-point neural network implemented in a mobile computing device by quantizing a floating-point neural network, the method comprising:

obtaining quantized parameters by quantizing parameters in a same category in the floating-point neural network;

obtaining a sample distribution from a semifixed-point artificial neural network (ANN) in which the obtained quantized parameters have been applied to the floating-point neural network; and

obtaining a fractional length for the obtained sample distribution,

wherein the obtaining the fractional length for the obtained sample distribution comprises:

splitting the obtained sample distribution into a plurality of sample groups;

approximating the plurality of sample groups to a plurality of continuous probability distributions (CPDs), respectively;

obtaining a plurality of step sizes, based on the plurality of sample groups and the plurality of CPDs; and

selecting the fractional length, based on the obtained plurality of step sizes.