

ABSTRACT OF THE DISCLOSURE

An artificial neural network (ANN) quantization method for generating an output ANN by quantizing an input ANN includes: 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 sample distribution by quantizing the obtained sample distribution.

FIG. 3