

DESCRIPTION

TITLE OF INVENTION: DATA PROCESSING DEVICE, DATA PROCESSING METHOD, AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] The invention relates to a data processing device and a data processing method that encode and compress information about a configuration of a neural network, and relates to a storage medium storing compressed data.

BACKGROUND ART

[0002] As a method for solving classification problems and regression problems of input data, there is machine learning.

For the machine learning, there is a technique called a neural network that imitates a brain's neural circuit (neurons). In the neural network, classification (discrimination) or regression of input data is performed using a probabilistic model (a discriminative model or a generative model) represented by a network in which neurons are mutually connected to each other.

Furthermore, in a convolutional neural network which is a neural network having not only a fully-connected layer but also a convolution layer and a pooling layer, a network can be created that implements data processing other than classification and regression, such as a network that implements a data filtering process. For example, a convolutional neural network can implement an image or audio filtering process that achieves noise removal of an input signal, an improvement in quality, or the like, with an image or audio being an input, a high-frequency restoration process for audio with missing high frequencies such as compressed audio, inpainting for an image whose

Unsupervised Feature Learning NIPS Workshop, 2011.

SUMMARY OF INVENTION

TECHNICAL PROBLEM

[0006] However, the optimal values of weights assigned to a respective plurality of edges in a neural network vary depending on network training results and are not fixed.

Hence, variations occur in the compression size of edge weights, and the technique described in Non-Patent Literature 1 has a problem that high compression of parameter data about edges of a neural network cannot be achieved.

[0007] The invention is to solve the above-described problem, and an object of the invention is to obtain a data processing device and a data processing method that can highly compress parameter data of a neural network, and a storage medium storing compressed data.

SOLUTION TO PROBLEM

[0008] A data processing device according to the invention includes a data processing unit, a compression controlling unit, and an encoding unit. The data processing unit processes input data using a neural network. The compression controlling unit determines quantization steps and generates quantization information that defines the quantization steps, the quantization steps being used when parameter data of the neural network is quantized. The encoding unit encodes network configuration information and the quantization information to generate compressed data, the network configuration information including the parameter data quantized using the quantization steps determined by the compression controlling unit.

ADVANTAGEOUS EFFECTS OF INVENTION

controlling unit 102 determines the quantization steps shown in FIG. 10 or the quantization steps shown in FIG. 12, as quantization steps for weight information of a plurality of respective edges in the network. FIG. 14 shows the syntax of the quantization steps shown in FIG. 10 or 12.

[0061] The syntax shown in FIG. 14 will be described.

As described above, the input layer (first layer) has no edges.

Hence, quantization steps are set for a layer_id+1th layer in which the ID information "layer_id" is 1 to L-1.

First, in the layer_id+1th layer, the compression controlling unit 102 sets base_quant_step[j] (j = 0, 1, ..., E-1) indicating a quantization step for the first node shown in FIG. 10 (the first kernel in FIG. 12).

Note that E is the number of edges $N_{\text{layer_id-1}}$ or the kernel size $K_{\text{layer_id}}$.

[0062] Then, for the second and subsequent nodes (or kernels) (i ≥ 1), the compression controlling unit 102 refers, on a node-by-node (or kernel-by-kernel) basis, to the flag "prev_quant_copy_flag[i-1]" indicating whether the quantization step is the same as that for a node (or a kernel) with an immediately previous index.

When the flag "prev_quant_copy_flag[i-1]" is 1 (true), an i+1th node (or kernel) has the same quantization step as that for an ith node (or kernel).

On the other hand, when the flag "prev_quant_copy_flag[i-1]" is 0 (false), the compression controlling unit 102 sets the difference value "diff_quant_value[i-1][j]" (i = 1, 2, ..., C-1, and j = 0, 1, ..., E-1) as information for generating a quantization step for the i+1th node (or kernel).

[0063] The quantization step can be generated by adding a quantization step set for an immediately previous node (kernel) to the difference value "diff_quant_value[i-1][j]".

Namely, in the second node (or kernel) (i = 1), base_quant_step[j] +

CLAIMS

1. A data processing device comprising:
 - a data processing unit for processing input data using a neural network;
 - a compression controlling unit for determining quantization steps and generating quantization information that defines the quantization steps, the quantization steps being used when parameter data of the neural network is quantized; and
 - an encoding unit for encoding network configuration information and the quantization information to generate compressed data, the network configuration information including the parameter data quantized using the quantization steps determined by the compression controlling unit.

2. A data processing device comprising:
 - a data processing unit for processing input data using a neural network; and
 - a decoding unit for decoding compressed data obtained by encoding quantization information and network configuration information, the quantization information defining quantization steps used when parameter data of the neural network is quantized, and the network configuration information including the parameter data quantized using the quantization steps in the quantization information, wherein
 - the data processing unit inversely quantizes the parameter data using the quantization information and the network configuration information which are decoded from the compressed data by the decoding unit, and constructs the neural network using the network configuration information including the inversely quantized parameter data.

3. The data processing device according to claim 1 or 2, wherein the parameter data of the neural network is weight information assigned to edges that connect nodes in the neural network.
4. The data processing device according to claim 1, wherein
the compression controlling unit changes the quantization steps on an edge-by-edge basis, and
the encoding unit encodes the quantization information that defines the edge-by-edge quantization steps.
5. The data processing device according to claim 1, wherein
the compression controlling unit changes the quantization steps on a node-by-node or kernel-by-kernel basis, and
the encoding unit encodes the quantization information that defines the node-by-node or kernel-by-kernel quantization steps.
6. The data processing device according to claim 1, wherein
the compression controlling unit changes the quantization steps on a layer-by-layer basis of the neural network, and
the encoding unit encodes the quantization information that defines the layer-by-layer quantization steps for the neural network.
7. A data processing method comprising:
a step of, by a decoding unit, decoding compressed data obtained by encoding quantization information and network configuration information, the quantization

information defining quantization steps used when parameter data of a neural network is quantized, the network configuration information including the parameter data quantized using the quantization steps in the quantization information; and

a step of, by a data processing unit, inversely quantizing the parameter data using the quantization information and the network configuration information which are decoded from the compressed data by the decoding unit, constructing the neural network using the network configuration information including the inversely quantized parameter data, and processing input data using the neural network.

8. A storage medium storing compressed data, the compressed data obtained by encoding

quantization information that defines quantization steps used when parameter data of a neural network is quantized; and

network configuration information including the parameter data quantized using the quantization steps in the quantization information, wherein

the compressed data causes a data processing device to inversely quantize the parameter data using the quantization information and the network configuration information which are decoded from the compressed data by the data processing device, and to construct the neural network using the network configuration information including the inversely quantized parameter data.