

processing unit for training an NN; and an encoding unit for generating encoded data in which model header information for identifying a model of the NN, layer header information for identifying one or more layers of the NN, and pieces of weight information of respective edges belonging to each of the one or more layers identified by the layer header information are encoded, and the encoding unit encodes layer structure information indicating a layer structure of the neural network.

ADVANTAGEOUS EFFECTS OF INVENTION

[0009] According to the present invention, the encoding unit encodes layer structure information indicating a layer structure of an NN, and a new layer flag indicating whether each layer to be encoded is a layer to be updated from a corresponding layer of a reference model, or a new layer. Of pieces of data representing a structure of the NN, only information about an updated layer is encoded and transmitted, and thus, a data size of the pieces of data representing the structure of the NN can be reduced.

BRIEF DESCRIPTION OF DRAWINGS

[0010] FIG. 1 is a block diagram showing a configuration of a data processing system according to a first embodiment.

FIG. 2 is a diagram showing an exemplary configuration of an NN.

FIG. 3 is a block diagram showing a configuration of a data processing device (encoder) according to the first embodiment.

FIG. 4 is a block diagram showing a configuration of a data processing device (decoder) according to the first embodiment.

FIG. 5 is a flowchart showing operation of the data processing device (encoder)

identified from the start code.

[0046] In order to set in such a manner that 0x000001 does not occur, for example, 03 is inserted into the third byte of encoded data “0x000000 to 0x000003”, thereby resulting in 0x00000300 to 0x00000303, and upon decoding, 0x000003 is converted into 0x0000, by which the data can be brought back to the original one. Note that as long as the start code is a bit string that can be uniquely identified, a bit string other than 0x000001 may be defined as the start code. In addition, as long as a method can identify a start position of a data unit, a start code does not need to be used. For example, a bit string that can identify an end of a data unit may be added to the end of the data unit. Alternatively, a start code may be added only to the start of a non-layer data unit, and as a part of a model information header, the data size of each layer data unit may be encoded. Therefore, a separation position between any two adjacent layer data units can be identified from the above-described information.

[0047] The data unit type is data stored after the start code in the data unit to identify the type of the data unit. For the data unit type, a value is defined for each type of data unit in advance. By referring to the data unit type stored in the data unit, the decoding side can identify whether the data unit is a non-layer data unit or a layer data unit, and can further identify what kind of non-layer data unit or layer data unit the data unit is.

[0048] The model information header in the non-layer data unit includes a model identification number (`model_id`), the number of layer data units in a model (`num_layers`), and the number of encoded layer data units (`num_coded_layers`). The model identification number is a number for identifying a model of an NN. Thus, basically, individual models have numbers independent of each other, but if a data processing device (decoder) according to the first embodiment has newly received a model having the same model identification number as that of a model received in the

CLAIMS:

1. A data processing device comprising:
a data processing unit for training a neural network; and
an encoding unit for generating encoded data in which model header information for identifying a model of the neural network, layer header information for identifying one or more layers of the neural network, and pieces of weight information of respective edges belonging to each of the one or more layers identified by the layer header information are encoded, wherein
the encoding unit encodes layer structure information indicating a layer structure of the neural network.
2. The data processing device according to claim 1, wherein the encoding unit encodes the pieces of weight information of the respective edges belonging to each of the one or more layers, on a bit-plane-by-bit-plane basis from higher bits.
3. The data processing device according to claim 1 or 2, wherein the encoding unit encodes the pieces of weight information of the respective edges belonging to each of the one or more layers identified by the layer header information.
4. The data processing device according to claim 1 or 2, wherein the encoding unit encodes a difference between an edge weight value and a specific value.

5. The data processing device according to claim 1 or 2, wherein the encoding unit encodes the pieces of weight information of the respective edges as base encoded data and enhancement encoded data separately,
- the base encoded data is data obtained by quantizing weights of the edges and encoding the quantized weights, and
- the enhancement encoded data is data obtained by encoding a quantization error that is considered a corresponding one of the weights.
6. The data processing device according to claim 1 or 2, comprising a decoding unit for decoding the encoded data generated by the encoding unit, wherein
- the data processing unit trains the neural network using information decoded by the decoding unit.
7. A data processing system comprising:
- a first data processing device including:
- a first data processing unit for training a neural network; and
- an encoding unit for generating encoded data in which model header information for identifying a model of the neural network, layer header information for identifying one or more layers of the neural network, and pieces of weight information of respective edges belonging to each of the one or more layers identified by the layer header information are encoded; and
- a second data processing device including:
- a decoding unit for decoding the encoded data generated by the encoding unit; and

a second data processing unit for creating the neural network using information decoded by the decoding unit, and performing data processing using the neural network, wherein

the encoding unit encodes layer structure information indicating a layer structure of the neural network.

8. The data processing system according to claim 7, wherein the encoding unit encodes information about a portion of the neural network up to an intermediate layer, and

the second data processing device performs data processing using, as a feature, data outputted from the intermediate layer of the neural network.

9. A data processing method comprising the steps of:
training a neural network by a data processing unit; and
generating, by an encoding unit, encoded data in which model header information for identifying a model of the neural network, layer header information for identifying one or more layers of the neural network, and pieces of weight information of respective edges belonging to each of the one or more layers identified by the layer header information are encoded, wherein

the encoding unit encodes layer structure information indicating a layer structure of the neural network.