

CLAIMS

1. A data processing method, comprising:

inputting input data into a neural network model to obtain feature data currently output by a network layer in the neural network model;

obtaining multiple corresponding sub-matrices based on learnable gating parameters set in the neural network model;

performing inner product operation on the multiple sub-matrices to obtain transformation parameters of the neural network model;

determining, according to the transformation parameters, a normalization mode matched with the feature data, wherein the transformation parameters are used for adjusting a statistical range of statistics of the feature data, and the statistical range is used for representing the normalization mode; and

performing normalization processing on the feature data according to the determined normalization mode to obtain normalized feature data.

2. The method according to claim 1, wherein obtaining the multiple corresponding sub-matrices based on the learnable gating parameters set in the neural network model comprises:

using a sign function to process the gating parameters to obtain a binarization vector;

using a permutation matrix to permute elements in the binarization vector to generate a binarization gating vector; and

obtaining the multiple sub-matrices based on the binarization gating vector, a first fundamental matrix, and a second fundamental matrix.

3. The method according to claim 1 or 2, wherein the transformation parameters comprise a first transformation parameter, a second transformation parameter, a third transformation parameter, and a fourth transformation parameter; and

a dimension of the first transformation parameter and a dimension of the third transformation parameter are based on a batch size dimension of the feature data, and a dimension of the second transformation parameter and a dimension of the fourth transformation parameter are based on a channel dimension of the feature data;

wherein the batch size dimension is the number of pieces of data in a data batch where the feature data is located, and the channel dimension is the number of channels of the feature data.

4. The method according to claim 3, wherein determining, according to the transformation parameters, the normalization mode matched with the feature data comprises:

determining the statistical range of the statistics of the feature data as a first range, wherein the statistics comprise a mean value and a standard deviation;

adjusting the statistical range of the mean value from the first range to a second range according to the first transformation parameter and the second transformation parameter;

adjusting the statistical range of the standard deviation from the first range to a third range according to the third transformation parameter and the fourth transformation parameter; and

determining the normalization mode based on the second range and the third range.

5. The method according to claim 3, wherein the first range is each channel range of each piece of sample feature data of the feature data.

6. The method according to claim 4, wherein performing normalization processing on the feature data according to the determined normalization mode to obtain the normalized feature data comprises:

obtaining the statistics of the feature data in accordance with the first range; and

performing normalization processing on the feature data based on the statistics, the first transformation parameter, the second transformation parameter, the third transformation parameter, and the fourth transformation parameter so as to obtain the normalized feature data.

7. The method according to claim 6, wherein performing normalization processing on the feature data based on the statistics, the first transformation parameter, the second transformation parameter, the third transformation parameter, and the fourth transformation parameter so as to obtain the normalized feature data comprises:

obtaining a first normalization parameter based on the mean value, the first transformation parameter, and the second transformation parameter;

obtaining a second normalization parameter based on the standard deviation, the third transformation parameter, and the fourth transformation parameter; and

performing normalization processing on the feature data according to the feature data, the first normalization parameter, and the second normalization parameter so as to obtain the normalized feature data.

8. The method according to claim 1, wherein the transformation parameters comprise binarization matrices, and the value of each element in the binarization matrices is 0 or 1;

and/or,

wherein the gating parameters are vectors having continuous values, and the number of values in the gating parameters is consistent with the number of the sub-matrices.

9. The method according to claim 2, wherein the first fundamental matrix is an all-ones matrix, and the second fundamental matrix is a unit matrix.

10. A data processing apparatus, comprising:

a data inputting module, configured to input input data into a neural network model to obtain feature data currently output by a network layer in the neural network model;

a sub-matrix obtaining module, configured to obtain multiple corresponding sub-matrices based on learnable gating parameters set in the neural network model;

a transformation parameter obtaining module, configured to perform inner product operation on the multiple sub-matrices to obtain transformation parameters of the neural network model;

a mode determining module, configured to determine, according to transformation parameters, a normalization mode matched with the feature data, wherein the transformation parameters are used for adjusting a statistical range of statistics of the feature data, and the statistical range is used for representing the normalization mode; and

a normalization processing module, configured to perform normalization processing on the feature data according to the determined normalization mode to obtain normalized feature data.

11. The apparatus according to claim 10, wherein the sub-matrix obtaining module comprises:

a parameter processing sub-module, configured to use a sign function to process the gating parameters to obtain a binarization vector;

an element permuting sub-module, configured to use a permutation matrix to permute elements in the binarization vector to generate a binarization gating vector; and

a sub-matrix obtaining sub-module, configured to obtain the multiple sub-matrices based on the binarization gating vector, a first fundamental matrix, and a second fundamental matrix.

12. The apparatus according to claim 10 or 11, wherein the transformation parameters comprise a first transformation parameter, a second transformation parameter, a third transformation parameter, and a fourth transformation parameter; and

a dimension of the first transformation parameter and a dimension of the third transformation parameter are based on a batch size dimension of the feature data, and a dimension of the second transformation parameter and a dimension of the fourth transformation parameter are based on a channel dimension of the feature data;

wherein the batch size dimension is the number of pieces of data in a data batch where the feature data is located, and the channel dimension is the number of channels of the feature data.

13. The apparatus according to claim 12, wherein the mode determining module comprises:

a first determining sub-module, configured to determine the statistical range of the statistics of the feature data as a first range, wherein the statistics comprise a mean value and a standard deviation;

a first adjusting sub-module, configured to adjust the statistical range of the mean value from the first range to a second range according to the first transformation parameter and the second transformation parameter;

a second adjusting sub-module, configured to adjust the statistical range of the standard deviation from the first range to a third range according to the third transformation parameter and the fourth transformation parameter; and

a mode determining sub-module, configured to determine the normalization mode based on the second range and the third range.

14. The apparatus according to claim 13, wherein the first range is each channel range of each piece of sample feature data of the feature data.

15. The apparatus according to claim 13, wherein the normalization processing module comprises:

a statistics obtaining sub-module, configured to obtain the statistics of the feature data in accordance with the first range; and

a normalization processing sub-module, configured to perform normalization processing on the feature data based on the statistics, the first transformation parameter, the second transformation parameter, the third transformation parameter, and the fourth transformation parameter so as to obtain the normalized feature data.

16. The apparatus according to claim 15, wherein the normalization processing sub-module comprises:

a first parameter obtaining unit, configured to obtain a first normalization parameter based on the mean value, the first transformation parameter, and the second transformation parameter;

a second parameter obtaining unit, configured to obtain a second normalization parameter based on the

standard deviation, the third transformation parameter, and the fourth transformation parameter; and

a data processing unit, configured to perform normalization processing on the feature data according to the feature data, the first normalization parameter, and the second normalization parameter so as to obtain the normalized feature data.

17. The apparatus according to claim 10, wherein the transformation parameters comprise binarization matrices, and the value of each element in the binarization matrices is 0 or 1;

and/or,

wherein the gating parameters are vectors having continuous values, and the number of values in the gating parameters is consistent with the number of the sub-matrices.

18. The apparatus according to claim 11, wherein the first fundamental matrix is an all-ones matrix, and the second fundamental matrix is a unit matrix.

19. An electronic device, comprising:

a processor; and

a memory configured to store processor-executable instructions;

wherein the processor is configured to execute the method according to any one of claims 1 to 9.

20. A computer-readable storage medium, having computer program instructions stored thereon, wherein when the computer program instructions are executed by a processor, the method according to any one of claims 1 to 9 is implemented.