

CLAIMS

1. A computer-implemented method, comprising:

obtaining, by a data processing apparatus, a trained convolutional neural network comprising at least a first convolutional layer and a second convolutional layer, each of the at least the first convolutional layer and the second convolutional layer comprising a corresponding plurality of filters with known filter parameters;

pre-computing, by the data processing apparatus, a first reusable factor for the first convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;

pre-computing, by the data processing apparatus, a second reusable factor for the second convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;

storing the first reusable factor and the second reusable factor in a data storage;

identifying, by the data processing apparatus, a first input data to the first convolutional layer of the trained convolutional neural network;

retrieving, from the data storage, the first reusable factor and the second reusable factor;

computing, by the data processing apparatus, a first output of the first convolutional layer using a Winograd convolutional operator based on the retrieved first reusable factor and the first input data;

using, by the data processing apparatus, the first output as a second input data to the second convolutional layer of the trained convolutional neural network;

computing, by the data processing apparatus, a second output of the second convolutional layer using the Winograd convolutional operator based on the retrieved second reusable factor and the second input data;

identifying a second input data to the first convolutional layer of the trained convolutional neural network;

retrieving, from the data storage, the first reusable factor and the second reusable factor;

and

computing, based on the retrieved first reusable factor and the retrieved second reusable factor, an additional output corresponding to the second input data.

2. The computer-implemented method of claim 1, further comprising saving the first reusable factor for the first convolutional layer.

3. The computer-implemented method of claim 1, wherein computing a first output of the first convolutional layer using a Winograd convolutional operator based on the first reusable factor and the first input data comprises computing the first output of the first convolutional layer according to a Winograd minimal filtering algorithm with the first reusable factor and the first input data.

4. The computer-implemented method of claim 1, wherein the first input data to the trained convolutional neural network comprises one or more images.

5. The computer-implemented method of claim 1, further comprising:

determining, by the data processing apparatus, output data of the trained convolutional neural network based on the second output of the second convolutional layer.

6. The computer-implemented method of claim 1, further comprising outputting output data of the trained convolutional neural network via a user interface.

7. A non-transitory, computer-readable medium storing one or more instructions executable by a computer system to perform operations comprising:

obtaining a trained convolutional neural network comprising at least a first convolutional layer and a second convolutional layer, each of the at least the first convolutional layer and the second convolutional layer comprising a corresponding plurality of filters with known filter parameters;

pre-computing a first reusable factor for the first convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;

pre-computing a second reusable factor for the second convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;
storing the first reusable factor and the second reusable factor in a data storage;
identifying a first input data to the first convolutional layer of the trained convolutional neural network;
retrieving, from the data storage, the first reusable factor and the second reusable factor;
computing a first output of the first convolutional layer using a Winograd convolutional operator based on the retrieved first reusable factor and the first input data;
using the first output as a second input data to the second convolutional layer of the trained convolutional neural network;
computing a second output of the second convolutional layer using the Winograd convolutional operator based on the retrieved second reusable factor and the second input data;
identifying a second input data to the first convolutional layer of the trained convolutional neural network;
retrieving, from the data storage, the first reusable factor and the second reusable factor;
and
computing, based on the retrieved first reusable factor and the retrieved second reusable factor, an additional output corresponding to the second input data.

8. The non-transitory, computer-readable medium of claim 7, the operations further comprising saving the first reusable factor for the first convolutional layer.

9. The non-transitory, computer-readable medium of claim 7, wherein computing a first output of the first convolutional layer using a Winograd convolutional operator based on the first reusable factor and the first input data comprises computing the first output of the first convolutional layer according to a Winograd minimal filtering algorithm with the first reusable factor and the first input data.

10. The non-transitory, computer-readable medium of claim 7, wherein the first input data to the trained convolutional neural network comprises one or more images.

11. The non-transitory, computer-readable medium of claim 7, the operations further comprising:
determining output data of the trained convolutional neural network based on the second output of the second convolutional layer.
12. The non-transitory, computer-readable medium of claim 7, the operations further comprising outputting output data of the trained convolutional neural network via a user interface.
13. A computer-implemented system, comprising:
one or more computers; and
one or more computer memory devices interoperably coupled with the one or more computers and having tangible, non-transitory, machine-readable media storing instructions, that when executed by the one or more computers, perform operations comprising:
obtaining a trained convolutional neural network comprising at least a first convolutional layer and a second convolutional layer, each of the at least the first convolutional layer and the second convolutional layer comprising a corresponding plurality of filters with known filter parameters;
pre-computing a first reusable factor for the first convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;
pre-computing a second reusable factor for the second convolutional layer based on the known filter parameters regardless of any input data to the trained convolutional neural network;
storing the first reusable factor and the second reusable factor in a data storage;
identifying a first input data to the first convolutional layer of the trained convolutional neural network;
retrieving, from the data storage, the first reusable factor and the second reusable factor;
computing a first output of the first convolutional layer using a Winograd convolutional operator based on the retrieved first reusable factor and the first input data;
using the first output as a second input data to the second convolutional layer of the trained convolutional neural network;

computing a second output of the second convolutional layer using the Winograd convolutional operator based on the retrieved second reusable factor and the second input data;

identifying a second input data to the first convolutional layer of the trained convolutional neural network;

retrieving, from the data storage, the first reusable factor and the second reusable factor;
and

computing, based on the retrieved first reusable factor and the retrieved second reusable factor, an additional output corresponding to the second input data.

14. The computer-implemented system of claim 13, the operations further comprising saving the first reusable factor for the first convolutional layer.

15. The computer-implemented system of claim 13, wherein computing a first output of the first convolutional layer using a Winograd convolutional operator based on the first reusable factor and the first input data comprises computing the first output of the first convolutional layer according to a Winograd minimal filtering algorithm with the first reusable factor and the first input data.

16. The computer-implemented system of claim 13, the operations further comprising:

determining output data of the trained convolutional neural network based on the second output of the second convolutional layer.

17. The computer-implemented system of claim 13, further comprising a user interface; and the operations further comprising outputting output data of the trained convolutional neural network via the user interface.

18. The computer-implemented method of claim 1, wherein computing the second output comprises:

computing a convolution result using the Winograd convolutional operator based on the second reusable factor and the second input data; and

adding a bias to the convolution result to generate the second output.

19. The computer-implemented method of claim 18, wherein the Winograd convolutional operator comprises a function of a bias addition operator, and wherein adding the bias comprises adding the bias using the Winograd convolutional operator.