MODIFYING MACHINE LEARNING MODELS TO IMPROVE LOCALITY

BACKGROUND

[0001] This specification generally relates to modifying machine learning models based on characteristics of memory to improve locality.

[0002] Neural networks are machine learning models that employ one or more layers of models to generate an output, e.g., a classification, for a received input. Some neural networks include one or more hidden layers in addition to an outer layer. The output of each hidden layer is used as input to the next layer in the network, i.e., the next hidden layer or the output layer of the network. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

[0003] A neural network can be represented as a graph with nodes and edges. A node can represent an operation that can have zero or more inputs and zero or more outputs. An edge can represent data and/or control dependencies among operations. A machine learning processor can execute the graph repeatedly with different data (typically, a batch of input samples at a time). For each operation, the machine learning processor can obtain input data from memory and store the output data in memory.

SUMMARY

[0004] This specification describes technologies relating to improving locality of machine learning models and using the improved models to perform machine learning computations.

[0004a] According to a first aspect of the present invention, there is provided a method performed by data processing apparatus, the method comprising: receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations; receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by

the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory; generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and performing machine learning computations using the updated machine learning model.

[0004b] According to a second aspect of the present invention, there is provided a system, comprising: a data processing apparatus; and a memory storage apparatus in data communication with the data processing apparatus, the memory storage apparatus storing instructions executable by the data processing apparatus and that upon such execution cause the data processing apparatus to perform operations comprising: receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations; receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory; generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in

based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and performing machine learning computations using the updated machine learning model.

[0004c] According to a third aspect of the present invention, there is provided a non-transitory computer storage medium encoded with a computer program, the program comprising instructions that when executed by one or more data processing apparatus cause the data processing apparatus to perform operations comprising: receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations; receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory; generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and performing machine learning computations using the updated machine learning model.

[0005] There is disclosed a method that includes receiving data of a machine learning model. The data can represent operations of the machine learning model and data dependencies between the operations. Data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed is received. The memory hierarchy can include multiple memories at multiple memory levels for storing machine learning data used by the

machine learning processor when performing machine learning computations using the machine learning model. The characteristics can include a data storage capacity of each memory and a memory bandwidth of each memory. At least one of the memories can have a different memory bandwidth than at least one other memory. An updated machine learning model can be generated based on the data of the machine learning model and the characteristics of the memory hierarchy. The updated machine learning model can be generated by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy. Machine learning computations can be performed using the updated machine learning model.

[0006] These and other implementations can each optionally include one or more of the following features. In some examples, the data of the machine learning model can include a graph that represents the operations of the machine learning model, the control dependencies between the operations, and data dependencies between the operations.

[0007] In some examples, generating the updated machine learning model can include selecting, for at least a portion of the operations, one of the memories to store outputs of the operation based on when the outputs will be used as inputs to another operation.

[0008] In some examples, generating the updated machine learning model can include determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation. The first memory can have a lower memory bandwidth than a second memory of the multiple memories. In response to the determining, the updated machine learning model can include first control data that causes the machine learning processor to store the output data for the first operation in the first memory after the output data is generated by the first operation, and second control data that causes the machine learning processor to transfer the output data from the first memory to the second memory prior to the output data being used as input to the second operation.

[0009] In some examples, the second control data causes the machine learning processor to transfer the output data from the first memory to the second memory in response to a third operation being executed. The third operation can be different from the first and second operations. Determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation can include determining that the output data for the first operation is to be stored in the first memory based on at least one of (i) a number of operations that will be executed between the first operation and the second operation or (ii) an estimated duration of time between when the first operation will be executed and the second operation will be executed.

[0010] In some examples, generating the updated machine learning model can include determining that input data for a particular sequence of operations of the machine learning model requires more data storage capacity than a particular memory of the multiple memories. In response, the updated machine learning model can include multiple sequences of operations that include a same sequence of operations as the particular sequence of operations, first control data that causes the machine learning processor to split the input data into multiple portions of data, second control data that causes the machine learning processor to assign each portion of data to a respective sequence of operations of the multiple sequence of operations; and third control data that causes the machine learning processor to perform the multiple sequences of operations in series.

[0011] The subject matter described in this specification can be implemented in particular embodiments so as to realize one or more of the following advantages. Machine learning models can be modified based on characteristics of a memory hierarchy of a machine learning processor on which the models are to be implemented to improve locality of data access (e.g., locality in data placement and/or data movement), resulting in improved memory bandwidth (e.g., faster data accesses) and utilization of the memory. This improvement in memory bandwidth and memory utilization enables the machine learning processor to perform machine learning computations faster and more efficiently than using the original machine learning model.

CLAIMS

What is claimed is:

1. A method performed by data processing apparatus, the method comprising:

receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations;

receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory;

generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and

performing machine learning computations using the updated machine learning model.

2. The method of claim 1, wherein the data of the machine learning model comprises a graph that represents the operations of the machine learning model, the control dependencies between the operations, and data dependencies between the operations.

3. The method of claim 1, wherein generating the updated machine learning model comprises selecting, for at least a portion of the operations, one of the memories to

store outputs of the operation based on when the outputs will be used as inputs to another operation.

4. The method of claim 1, wherein generating the updated machine learning model comprises:

determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation, the first memory having a lower memory bandwidth than a second memory of the multiple memories and, in response:

including in the updated machine learning model:

first control data that causes the machine learning processor to store the output data for the first operation in the first memory after the output data is generated by the first operation; and

second control data that causes the machine learning processor to transfer the output data from the first memory to the second memory prior to the output data being used as input to the second operation.

5. The method of claim 4, wherein the second control data causes the machine learning processor to transfer the output data from the first memory to the second memory in response to a third operation being executed, the third operation being different from the first and second operations.

6. The method of claim 4, wherein determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation comprises:

determining that the output data for the first operation is to be stored in the first memory based on at least one of (i) a number of operations that will be executed between the first operation and the second operation or (ii) an estimated duration of time between when the first operation will be executed and the second operation will be executed.

7. The method of claim 1, wherein generating the updated machine learning model comprises:

determining that input data for a particular sequence of operations of the machine learning model requires more data storage capacity than a particular memory of the multiple memories and, in response:

including in the updated machine learning model:

multiple sequences of operations that include a same sequence of operations as the particular sequence of operations;

first control data that causes the machine learning processor to split the input data into multiple portions of data;

second control data that causes the machine learning processor to assign each portion of data to a respective sequence of operations of the multiple sequence of operations; and

third control data that causes the machine learning processor to perform the multiple sequences of operations in series.

8. The method of claim 1, wherein generating the updated machine learning model comprises using a second machine learning model to generate the updated machine learning model based on the machine learning model and the characteristics of the memory hierarchy.

9. A system, comprising:

a data processing apparatus; and

a memory storage apparatus in data communication with the data processing apparatus, the memory storage apparatus storing instructions executable by the data processing apparatus and that upon such execution cause the data processing apparatus to perform operations comprising:

receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations;

receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory;

generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and

performing machine learning computations using the updated machine learning model.

10. The system of claim 9, wherein the data of the machine learning model comprises a graph that represents the operations of the machine learning model, the control dependencies between the operations, and data dependencies between the operations.

11. The system of claim 9, wherein generating the updated machine learning model comprises selecting, for at least a portion of the operations, one of the memories to store outputs of the operation based on when the outputs will be used as inputs to another operation.

12. The system of claim 9, wherein generating the updated machine learning model comprises:

determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation, the first memory having a lower memory bandwidth than a second memory of the multiple memories and, in response:

including in the updated machine learning model:

first control data that causes the machine learning processor to store the output data for the first operation in the first memory after the output data is generated by the first operation; and

second control data that causes the machine learning processor to transfer the output data from the first memory to the second memory prior to the output data being used as input to the second operation.

13. The system of claim 12, wherein the second control data causes the machine learning processor to transfer the output data from the first memory to the second memory in response to a third operation being executed, the third operation being different from the first and second operations.

14. The system of claim 12, wherein determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation comprises:

determining that the output data for the first operation is to be stored in the first memory based on at least one of (i) a number of operations that will be executed between the first operation and the second operation or (ii) an estimated duration of time between when the first operation will be executed and the second operation will be executed.

15. The system of claim 9, wherein generating the updated machine learning model comprises:

determining that input data for a particular sequence of operations of the machine learning model requires more data storage capacity than a particular memory of the multiple memories and, in response:

including in the updated machine learning model:

multiple sequences of operations that include a same sequence of operations as the particular sequence of operations;

first control data that causes the machine learning processor to split the input data into multiple portions of data;

second control data that causes the machine learning processor to assign each portion of data to a respective sequence of operations of the multiple sequence of operations; and

third control data that causes the machine learning processor to perform the multiple sequences of operations in series.

16. The system of claim 9, wherein generating the updated machine learning model comprises using a second machine learning model to generate the updated machine learning model based on the machine learning model and the characteristics of the memory hierarchy.

17. A non-transitory computer storage medium encoded with a computer program, the program comprising instructions that when executed by one or more data processing apparatus cause the data processing apparatus to perform operations comprising:

receiving data of a machine learning model, the data representing operations of the machine learning model and data dependencies between the operations;

receiving data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed, the memory hierarchy including multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model, the characteristics including a data storage capacity of each memory and a memory bandwidth of each memory, wherein at least one of the memories has a different memory bandwidth than at least one other memory;

generating, based on the data of the machine learning model and the characteristics of the memory hierarchy, an updated machine learning model by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy and modifying, for each of one or more of the operations of the machine learning model, which memory of the multiple memories of the machine learning processor that outputs of the operation are stored in based on the characteristics of the memory hierarchy including the data storage capacity and the memory bandwidth of each memory; and

performing machine learning computations using the updated machine learning model.

18. The non-transitory computer storage medium of claim 17, wherein the data of the machine learning model comprises a graph that represents the operations of the machine learning model, the control dependencies between the operations, and data dependencies between the operations.

19. The non-transitory computer storage medium of claim 17, wherein generating the updated machine learning model comprises selecting, for at least a portion of the operations, one of the memories to store outputs of the operation based on when the outputs will be used as inputs to another operation.

20. The non-transitory computer storage medium of claim 17, wherein generating the updated machine learning model comprises:

determining that output data for a first operation is to be stored in a first memory of the multiple memories based on when the output data for the first operation will be used as input by a second operation, the first memory having a lower memory bandwidth than a second memory of the multiple memories and, in response:

including in the updated machine learning model:

first control data that causes the machine learning processor to store the output data for the first operation in the first memory after the output data is generated by the first operation; and

second control data that causes the machine learning processor to transfer the output data from the first memory to the second memory prior to the output data being used as input to the second operation.