CLAIMS

We claim:

1. A method of accessing an external memory repository to a neural network processor while performing a neural network operation for a multilayer neural network, said method comprising the steps of:

loading, from an external interface, a first subset of tensor data for said neural network operation;

loading, from said neural network processor, intermediate result tensor data for said neural network operation;

processing said first subset of tensor data and said intermediate result tensor data by a first partial layer computation by a first layer of said neural network operation in a matrix processor array to generate a subset of temporary tensor data to be used in a first subsequent layer or at least one second subsequent layer of said multilayer neural network;

loading, from said external interface, in parallel with and prior to completion of said processing of said first subset of tensor data and said intermediate result tensor data, a second subset of tensor data from said external memory repository into a first buffer within said neural network processor, said first buffer comprising a vector scalar processor coupled by an operand bus to a plurality of matrix processor circuits within said matrix processor array;

saving a first subset of said subset of temporary tensor data as said intermediate result tensor data in a second buffer having a local storage memory and discarding a second subset of said subset of temporary tensor data not required for subsequent computation of said multilayer neural network;

executing a data move operation to load the second subset of tensor data from said first buffer and said intermediate result tensor data from said second buffer into said matrix processor array; and

processing said second subset of tensor data and said intermediate result tensor data by a second partial layer computation by said first layer of said neural network operation in said matrix processor array to generate an updated subset

of temporary tensor data to be used in said first subsequent layer or said at least one second subsequent layer of said multilayer neural network.

2. The method of accessing an external memory repository to a neural network processor while performing a multilayer neural network operation as set forth in claim 1, wherein said first subset of tensor data from said external memory repository, required for additional computation of said multilayer neural network operation, comprises a subset of tensor input data for a new neural network operation.

3. The method of accessing an external memory repository to a neural network processor while performing a multilayer neural network operation as set forth in claim 1, wherein said first subset of tensor data from said external memory repository comprises a set of matrix weights required for additional computation of said multilayer neural network operation.

4. The method of accessing an external memory repository to a neural network processor while performing a multilayer neural network operation as set forth in claim 1, wherein said first subset of tensor data from said external memory repository comprises a subset of data previously stored in said external memory repository.

5. The method of accessing an external memory repository to a neural network processor while performing a multilayer neural network operation as set forth in claim 1, said method further comprising:

reloading said intermediate result tensor data from said second buffer.

6. The method of accessing an external memory repository to a neural network processor while performing a multilayer neural network operation as set forth in claim 1, said method further comprising:

storing said set of temporary data as a final result of said multilayer neural network operation in said external memory repository.

7. The method of claim 1, further comprising:

loading, from said external interface, in parallel with and prior to completion of said processing of said second subset of tensor data and said intermediate result tensor data, a third subset of tensor data from said external memory repository into a buffer within said neural network processor;

executing a further data move operation to load the third subset of tensor data and said updated subset of temporary tensor data from said buffer into said matrix processor array; and

processing said third subset of tensor data and said updated intermediate result tensor data by a third partial layer computation by said first layer of said neural network operation in said matrix processor array to generate a further updated subset of temporary tensor data to be used in said first subsequent layer or said at least one second subsequent layer of said multilayer neural network.

8. A neural network processor apparatus for performing a multilayer neural network computation, said neural network processor apparatus comprising:

an external data interface; said external data interface coupled to memory storage outside of said neural network processor apparatus;

a vector scalar processor coupled to said external data interface, said vector scalar processor for receiving incoming data;

an array of matrix processor circuits coupled to said vector scalar processor by an operand bus, said array of matrix processor circuits for performing matrix operations;

an accumulation buffer coupled to said array of matrix processor circuits, said accumulation buffer having a local storage memory;

a scheduler and sequence processor coupled to said vector scalar processor and further coupled to said array of matrix processor circuits

said scheduler and sequence processor issuing cycle commands to said vector scalar processor and to said array of matrix processor circuits, and cycle commands causing said array of matrix processor circuits to:

load, from said external data interface, a first subset of tensor data for said multilayer neural network computation;

load, from said neural network processor apparatus, intermediate result tensor data for said multilayer neural network computation;

process said first subset of tensor data and said intermediate result tensor data by a first partial layer computation by a first layer of said multilayer neural network computation to generate a subset of temporary tensor data to be used in a first subsequent layer or at least one second subsequent layer of said neural network processor apparatus;

load, from said external data interface, in parallel with and prior to completion of said processing of said first subset of tensor data and said intermediate result tensor data, a second subset of tensor data from said external memory repository into a first buffer within said neural network processor apparatus, said first buffer comprising said vector scalar processor;

save a first subset of said subset of temporary tensor data as said intermediate result tensor data in a second buffer, said second buffer comprising said accumulation buffer, and discard a second subset of said subset of temporary tensor data not required for subsequent computation of said multilayer neural network computation;

execute a data move operation to load the second subset of tensor data from said first buffer and said intermediate result tensor data from said second buffer into said array of matrix processor circuits; and

process said second subset of tensor data and said intermediate result tensor data by a second partial layer computation by said first layer of said multilayer neural network computation in said array of matrix processor circuits to generate an updated subset of temporary tensor data to be used in said first subsequent layer or said at least one second subsequent layer of said neural network processor apparatus.

9. The neural network processor apparatus as set forth in claim 8, wherein said first subset of tensor data from said external memory repository comprises a subset of tensor input data for a new neural network operation.

10. The neural network processor apparatus as set forth in claim 8, wherein said first subset of tensor data from said external memory repository comprises a set of matrix weights for a neural network operation.

11. The neural network processor apparatus as set forth in claim 8, wherein said cycle commands further cause said array of matrix processor circuits to reload said intermediate result tensor data from said second buffer.

12. The neural network processor apparatus as set forth in claim 8, wherein said cycle commands further cause said array of matrix processor circuits to store said subset of temporary tensor data as a final result of said multilayer neural network operation in said external memory repository.

13. The neural network processor apparatus of claim 8, each matrix processor circuit of the array of matrix processor circuits comprising a control system coupled by a control bus to a local wide static random access memory bank, the local wide static random access memory bank coupled to a bank of multiplexors, the bank of multiplexors couple to an operand register and a bank of multiply and accumulate units.

14. A method of accessing an external memory repository to a neural network processor while performing a neural network operation for a multilayer neural network, said method comprising the steps of:

loading, from an external interface, a first subset of tensor data for said neural network operation;

loading, from said neural network processor, intermediate result tensor data for said neural network operation;

processing said first subset of tensor data and said intermediate result tensor data by a first partial layer computation by a first layer of said multilayer neural network in a matrix processor array to generate a first subset of temporary data to be used in a first subsequent layer or at least one second subsequent layer of said multilayer neural network;

loading, from said external interface, in parallel with and prior to completion of said processing of said first subset of tensor data and said intermediate result tensor data, a second subset of tensor data from said external memory repository into a first buffer within said neural network processor, said first buffer comprising a vector scalar processor coupled by an operand bus to a plurality of matrix processor circuits within said matrix processor array;

processing said first subset of tensor data and said intermediate result tensor data by a second partial layer computation for the first subsequent layer of said multilayer neural network in said matrix processor unit to generate a second subset of temporary data to be used in a subsequent layer of said multilayer neural network;

saving a subset of said set of temporary data as intermediate result tensor data in a second buffer having a local storage memory;

executing a data move operation to load the second subset of tensor data from said first buffer into said matrix array;

processing said second subset of tensor data and said intermediate result tensor data by a second partial layer computation by said first layer of said neural network operation in said matrix processor array;

reloading said second subset of temporary data to process said at least one second subsequent layer of said multilayer neural network;

calculating a first final output of said neural network operation of said at least one second subsequent layer.

15. The method of accessing an external memory to a neural network processor while performing multilayer neural network operations as set forth in claim 14, wherein said first subset of tensor data from said external memory repository comprises a subset of tensor input data for a new neural network operation.

16. The method of accessing an external memory to a neural network processor while performing multilayer neural network operations as set forth in claim 14, wherein said first subset of tensor data from said external memory repository comprises a set of matrix weights for a neural network operation.

17. The method of accessing an external memory to a neural network processor while performing multilayer neural network operations as set forth in claim 14, wherein said first subset of tensor data from said external memory repository comprises a subset of data previously stored in said external memory repository.

18. The method of accessing an external memory to a neural network processor while performing multilayer neural network operations as set forth in claim 14, said method further comprising:

reloading said intermediate result tensor data from said second buffer.

19. The method of accessing an external memory to a neural network processor while performing multilayer neural network operations as set forth in claim 14, said method further comprising:

storing said set of temporary data as a final result of said multilayer neural network operation in said external memory repository.

20. The method of accessing an external memory repository to a neural network processor while performing a neural network operation for a multilayer neural network of claim 14, each matrix processor circuit of the array of matrix processor circuits comprising a control system coupled by a control bus to a local wide static random access memory bank, the local wide static random access memory bank coupled to a bank of multiplexors, the bank of multiplexors coupled to an operand register and a bank of multiply and accumulate units.