CLAIMS

1. A method, comprising:

receiving data defining a neural network;

analyzing the data to determine a depth-first cut point for a depth-first traversal portion of an overall network traversal;

performing depth-first traversal for the depth-first traversal portion of the overall network traversal;

storing weight data and output data for the depth-first traversal portion in a local memory during the depth-first portion of the overall network traversal; and

performing layer-based traversal for a layer-based portion of the overall network traversal,

wherein the depth-first cut point is determined based at least in part on a sum of a weight data size and an output data size for the depth-first traversal portion being less than a capacity of the local memory.

2. The method of claim 1, wherein the local memory is a tightly-coupled memory (TCM) associated with a processor configured to perform the depth-first traversal and the layer-based traversal.

3. The method of claim 1, wherein performing layer-based traversal for the layer-based portion of the overall network traversal comprises:

storing weight values and input values from a non-local memory in the local memory;

performing processing of a first layer of the layer-based portion of the overall network traversal using the weight values and the input values in order to generate output values;

storing the output values in the local memory; and

performing processing of a second layer of the layer-based portion of the overall network traversal using the weight values and the output values.

4. A processing system, comprising:

a memory comprising computer-executable instructions; and

a processor configured to execute the computer-executable instructions and cause the processing system to:

receive data defining a neural network;

analyze the data to determine a depth-first cut point for a depth-first traversal portion of an overall network traversal;

perform depth-first traversal for the depth-first traversal portion of the overall network traversal;

store weight data and output data for the depth-first traversal portion in a local memory during the depth-first portion of the overall network traversal; and

perform layer-based traversal for a layer-based portion of the overall network traversal,

wherein the depth-first cut point is determined based at least in part on a sum of a weight data size and an output data size for the depth-first traversal portion being less than a capacity of the local memory.

5. The processing system of claim 4, wherein the local memory is a tightly-coupled memory (TCM) associated with a processor configured to perform the depth-first traversal and the layer-based traversal.

6. The processing system of claim 4, wherein in order to perform the layer-based traversal for the layer-based portion of the overall network traversal, the processor is further configured to:

store weight values and input values from a non-local memory in the local memory;

perform processing of a first layer of the layer-based portion of the overall network traversal using the weight values and the input values in order to generate output values;

store the output values in the local memory; and

perform processing of a second layer of the layer-based portion of the overall network traversal using the weight values and the output values.

7. A method, comprising:

receiving data defining a neural network;

analyzing the data to determine a depth-first cut point for a depth-first traversal portion of an overall network traversal, wherein the depth-first cut point is determined based

at least in part on a sum of a weight data size and an output data size for the depth-first traversal portion being less than a capacity of the local memory;

performing depth-first traversal for the depth-first traversal portion of the overall network traversal; and

performing layer-based traversal for a layer-based portion of the overall network traversal, wherein performing layer-based traversal for the layer-based portion of the overall network traversal comprises:

storing weight values and input values from a non-local memory in the local memory;

performing processing of a first layer of the layer-based portion of the overall network traversal using the weight values and the input values in order to generate output values;

storing the output values in the local memory; and

performing processing of a second layer of the layer-based portion of the overall network traversal using the weight values and the output values.