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.

2. The method of claim 1, further comprising: determining the depth-first cut point 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.

3. The method of claim 2, 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.

4. 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.

5. The method of claim 2, wherein the depth-first cut point is determined by a compiler based on the capacity of the local memory.

6. 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.

7. The processing system of claim 6, wherein the processor is further configured to determine the depth-first cut point 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.

8. The processing system of claim 7, 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.

9. The processing system of claim 6, 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.

10. The processing system of claim 7, wherein the depth-first cut point is determined by a compiler based on the capacity of the local memory.

11. A method, comprising: receiving data defining a neural network; receiving hardware characteristic data defining a plurality of hardware resources, including a local memory capacity; analyzing the data to determine a cut point for a depth-first traversal portion of an overall network traversal; determining a first cut point based at least in part on the data and the local memory capacity, wherein the first cut point is at a maximum depth of the neural network given the local memory capacity; determining a second cut point, wherein the second cut point is shallower than the first cut point, and wherein the second cut point increases hardware parallelism compared to the first cut point; and performing depth-first traversal for the depth-first portion of the overall network traversal using two or more hardware resources of the plurality of hardware resources and the second cut point.

12. The method of claim 11, further comprising: 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.

13. The method of claim 12, wherein the local memory is a tightly-coupled memory (TCM).

14. The method of claim 11, wherein performing depth-first traversal for the depth-first traversal portion of the overall network traversal does not require writing data to a non-local memory.

15. The method of claim 11, wherein the second cut point enables cone parallelization during the depth-first traversal for the depth-first portion of the overall network traversal.

16. 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; 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.