Claims

1. A method of using a foundation neural network trained to perform a first computational task, to generate an adapted neural network configured to perform a second computational task which is different from the first computational task, the foundation neural network comprising a sequence of layers, each layer being configured to generate a corresponding output from a corresponding input to the layer by performing at least one corresponding function on the input, the function being based on a respective set of numerical parameters, the input to each processing layer of the sequence except a~~the~~ first layer of the sequence being based on the output of a corresponding preceding layer of the sequence, the method comprising:

forming the adapted neural network by adding one or more adapter modules to the foundation neural network;

training the adapted ~~updated~~ neural network, based on a database of training examples of the second computational task, by training the ~~update~~adaptor modules, the numerical parameters of the foundation neural network being preserved;

updating the corresponding functions of the layers to incorporate an~~the~~ effect of the ~~update~~adaptor modules into the corresponding functions; and

removing the ~~update~~adaptor modules from the ~~updated~~ adapted neural network;

wherein the first and second computational tasks are either:

tasks performed on a data input encoding at least one image and/or at least one sound signal; or

tasks of generating data encoding at least one image and/or at least one sound signal.

2. A method according to claim 1 in which each adapter module corresponds to one of the functions defined by one of the layers of the foundation model, and is configured to apply a transformation to the input or the result of the corresponding function.

3. A method according to claim 1 in which the transform is defined by a corresponding adapter matrix.

4. A method according to claim 3 in which each ~~update~~adaptor module is configured to apply a linear transformation based on the corresponding adapter matrix.

5. A method according to any of claims 1-4 in which each training example comprises input data and corresponding output data, the training including, in each of a plurality of iterations:

presenting the input data of at least one of the training examples to the input layer of the adapted neural network and modifying the adapter matrices to make an output of the adapted neural network closer to the corresponding output data of the at least one training example, the numerical parameters of the foundation neural network being preserved.

~~6. A method of using a foundation neural network trained to perform a first computational task, to generate an adapted neural network configured to perform a second computational task which is different from the first computational task, the foundation neural network comprising a sequence of layers, each layer being configured to generate a corresponding output from a corresponding input to the layer by performing at least one function on the input, the function being based on a respective set of numerical parameters, the input to each processing layer of the sequence except the first layer of the sequence being based on the output of a corresponding preceding layer of the sequence, the method comprising:~~

~~forming the adapted neural network by supplementing the foundation neural network by adding one or more adapter modules, each adapter module corresponding to one of the functions defined by one of the layers of the foundation model, and being configured to apply a linear transformation based on a corresponding adapter matrix to an input or a result of the corresponding function; and~~

~~training the adapted neural network based on a database of training examples of the second computational task, each training example comprising input data and corresponding output data, the training including, in each of a plurality of iterations:~~

~~presenting the input data of at least one of the training examples to the input layer of the adapted neural network and modifying the adapter matrices to make an output of the adapted neural network closer to the corresponding output data of the at least one training example, the numerical parameters of the foundation neural network being preserved.~~

~~7. A method according to claim 6 further comprising, following the training, a re-parameterization step of using each of the adapter matrices to update the corresponding function, thereby incorporating the effect of the linear transformation into the corresponding function, the update modules being removed from the updated neural network.~~

6~~8~~. A method according to claim 4 ~~or claim 6~~ in which each linear ~~transformation~~ projection includes adding a corresponding bias vector to a result of multiplying the corresponding adapter matrix and the input or result of the corresponding function, each iteration comprising modifying the bias vector to make the output of the adapted neural network closer to the corresponding output data of the at least one training example.

79. A method according to claim 6 ~~or claim 8~~ in which each function of each of the layers of the foundation model includes a residual connection which adds the output of the function to the input to the function, and in the adapted neural network the residual connection adds an output of the corresponding adapter module to the input of the corresponding function.

810. A model according to any preceding claim in which, for at least one of the layers, one of the corresponding functions is a self-attention function performed by a self-attention unit, the one or more adapter modules including an adapter module corresponding to the self-attention unit.

911. A model according to claim 810 in which the self-attention unit is a multi-head self-attention unit.

1012. A model according to any preceding claim in which, for at least one of the layers, one of the corresponding functions is a multilayer perceptron function implemented by a multilayer perceptron unit, the one or more adapter modules including an adapter module corresponding to the multilayer perceptron unit.

113. A method according to claim 102 when dependent upon claim 810 or 911, in which, for at least one of the layers of the foundation neural network, one of the functions of the layer is a self-attention function performed by a self-attention unit and another of the functions of the layer is a multilayer perceptron function implemented by a multilayer perceptron unit arranged to receive an input based on the output of the self-attention unit,

the corresponding layer of the adapted neural network including a first adapter unit configured to apply a linear projection~~transformation~~ based on a corresponding first adapter matrix to the input or result of the self-attention function, and a second adapter unit configured to apply a linear ~~transformation~~ projection based on a corresponding second adapter matrix to the input or result of the multilayer perceptron function, the input to the multilayer perceptron of the corresponding layer of the adapted neural network being based on the output of the first adapter unit.

124. A method according to any preceding claim in which the foundation neural network further includes an embedder layer arranged to receive input data, and from it to generate an input to a first of the sequence of layers

135. A method according to claim 124 in which, in the adapted neural network, only one layer of the sequence of layers comprises said adapter modules.

146. A method according to any of claims 1 to 124 in which, in the adapted neural network, a plurality of the layers of the foundation network include said adapter modules.

157. A method according to claim 146 in which, during said training, adapter matrices for multiple different ones of the layers are held the same.

168. A method according to claim 157 in which, during said training, adapter matrices for multiple different ones of the layers are trained independently.

179. A method according to any preceding claim in which the training of the adapter matrices is performed by back-propagation.

~~20. A method according to any preceding claim in which the first and second computational tasks are tasks performed on a data input encoding at least one image and/or at least one sound signal.~~

~~21. A method according to any preceding claim in which the first and/or second computational tasks are tasks of generating data encoding at least one image and/or at least one sound signal.~~

1822. A method according to any preceding claim in which the first computation task is a task of classifying input data to the foundation neural network into one of a first set of categories, and/or the second computation task is a task of classifying input data to the adapted neural network into one of a second set of categories.

1923. A method for generating a neural network model for solving a computational task, the method comprising training a foundation neural network to a perform a first computation task, and then forming an adapted neural network work by a method according to any preceding claim using the computational task as the second computational task.

204. A method of performing a computational task, the method comprising presenting input data to an adapted neural network generated by a method according to any preceding claim, using the computational task as the second computational task.

215. A method of using a foundation neural network trained to perform a first computational task, to generate an adapted neural network configured to perform a second computational task which is different from the first computational task, the foundation neural network comprising a sequence of layers, each layer being configured to generate a corresponding output from a corresponding input to the layer by performing at least one corresponding function on the input, each function receiving an input vector and generating a corresponding output vector based on a respective set of numerical parameters, the input to each processing layer of the sequence except a~~the~~ first layer of the sequence being based on the output of a corresponding preceding layer of the sequence,

the method comprising forming the adapted neural network by iteratively modifying only corresponding scale factors and mean values of the output vector of at least one of the functions based on a database of training examples of the second computational task, each training example comprising input data and corresponding output data;

wherein the first and second computational tasks are either:

tasks performed on a data input encoding at least one image and/or at least one sound signal; or

tasks of generating data encoding at least one image and/or at least one sound signal.

226. A computer system comprising at least one processor and at least one memory device, the at least one memory device storing program instructions which, when implemented by the processor, cause the processor to perform a method according to any preceding claim.

237. A computer program product storing program instructions which, when implemented by a processor, cause the processor to perform a method according to any of claims 1 to 215.