



(51) International Patent Classification:

G06N 3/063 (2006.01) G06N 3/08 (2006.01)
G06N 3/04 (2006.01)

(21) International Application Number:

PCT/US2020/056151

(22) International Filing Date:

16 October 2020 (16.10.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/683,217 13 November 2019 (13.11.2019) US

(71) Applicant: MICRON TECHNOLOGY, INC. [US/US];
8000 South Federal Way, Boise, Idaho 83716 (US).

(72) Inventor: LUO, Fa-Long; c/o Micron Technology, Inc.,
8000 South Federal Way, Seattle, Idaho 83716 (US).

(74) Agent: QUECAN, Andrew F. et al.; c/o Dorsey & Whitney LLP, Columbia Center, 701 Fifth Avenue, Suite 6100, Seattle, Washington 98104 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) Title: RECURRENT NEURAL NETWORKS AND SYSTEMS FOR DECODING ENCODED DATA

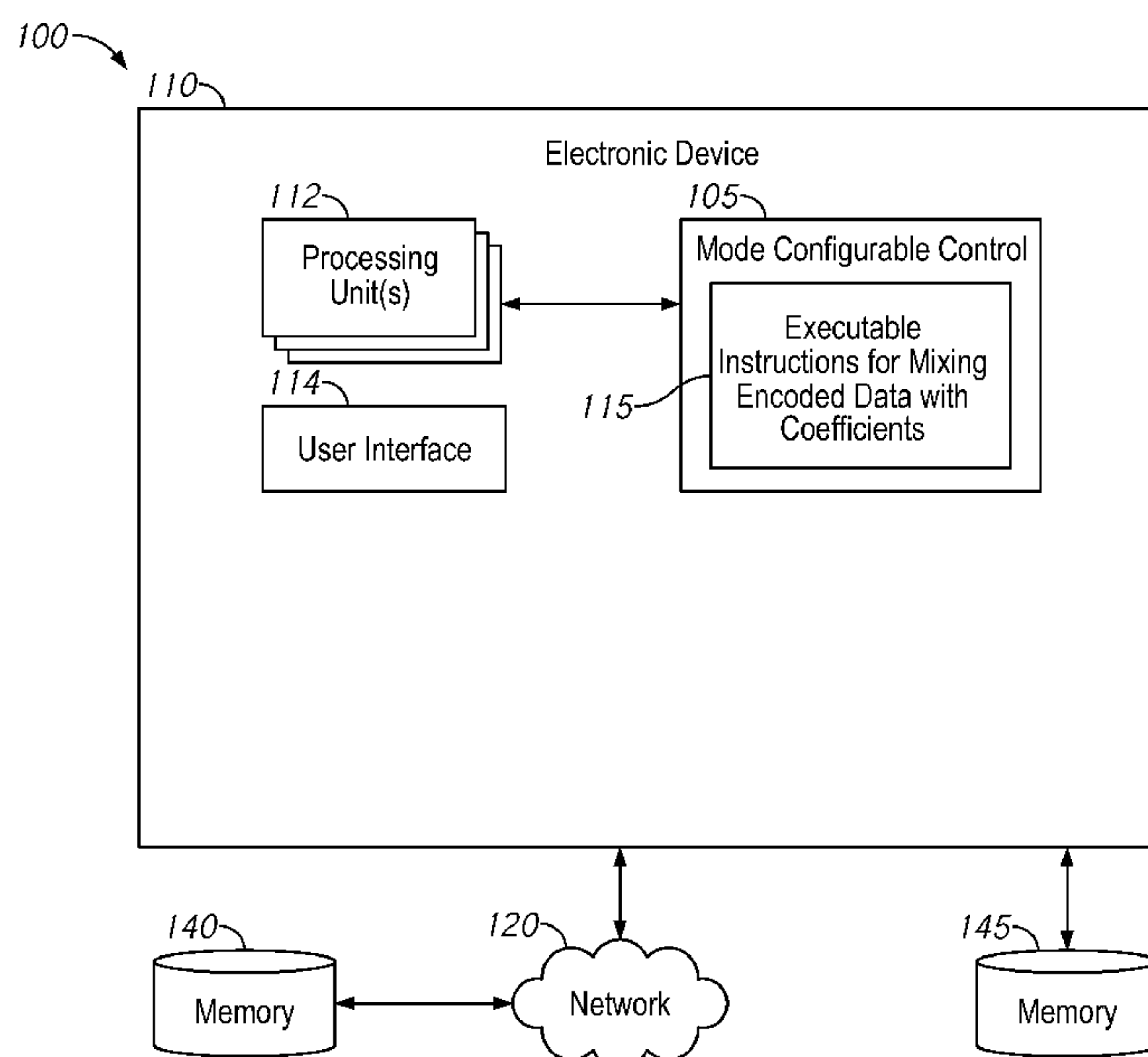


FIG. 1A

(57) Abstract: Examples described herein utilize multi-layer neural networks, such as multi-layer recurrent neural networks to decode encoded data (e.g., data encoded using one or more encoding techniques). The neural networks and/or recurrent neural networks may have nonlinear mapping and distributed processing capabilities which may be advantageous in many systems employing the neural network decoders and/or recurrent neural networks. In this manner, neural networks or recurrent neural networks described herein may be used to implement error correction coding (ECC) decoders.

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

RECURRENT NEURAL NETWORKS AND SYSTEMS FOR DECODING ENCODED DATA

TECHNICAL FIELD

[001] Examples described herein relate to neural networks for use in decoding encoded data. Examples of neural networks are described which may be used with error-correcting coding (ECC) memory, where a neural network may be used to decode encoded data.

BACKGROUND

[002] Error correction coding (ECC) may be used in a variety of applications, such as memory devices or wireless baseband circuitry. Generally, error correction coding techniques may encode original data with additional bits to describe the original bits which are intended to be stored, retrieved, and/or transmitted. The additional bits may be stored together with the original bits. Accordingly, there may be L bits of original data to be stored and/or transmitted. An encoder may provide N-L additional bits, such that the encoded data may be N bits worth of data. The original bits may be stored as the original bits, or may be changed by the encoder to form the encoded N bits of stored data. A decoder may decode the N bits to retrieve and/or estimate the original L bits, which may be corrected in some examples in accordance with the ECC technique.

[003] Bit flips (e.g., a change in charge at a memory cell) are an occurrence in non-volatile memory devices, where ECC may be applied. Thus, memory devices may operate with complex error correction techniques whose area and power needs are rising; resulting in higher cost silicon and longer firmware development times.

BRIEF DESCRIPTION OF THE DRAWINGS

[004] Figure 1A is a schematic illustration of a computing system arranged in accordance with examples described herein.

[005] Figure 1B is a schematic illustration of a neural network arranged in accordance with examples described herein.

[006] Figure 1C is a schematic illustration of a recurrent neural network arranged in accordance with examples described herein.

[007] Figures 2A-2C are schematic illustrations of processing units arranged in systems in accordance with examples described herein.

[008] Figure 3 is a schematic illustration of an apparatus arranged in accordance with examples described herein.

[009] Figures 4 and 5 are flowcharts of methods arranged in accordance with examples described herein.

SUMMARY

[010] Example apparatuses are disclosed herein. In an embodiment of the disclosure, an apparatus includes a first stage of circuitry configured to receive encoded data, to combine the encoded data, and to evaluate at least one non-linear function using combinations of the encoded data and delayed versions of the combinations of the encoded data to provide intermediate data. The apparatus further includes at least a second stage of circuitry configured to receive the intermediate data and combine the intermediate data using a set of predetermined weights to provide decoded data the set of predetermined weights based at least in part on an encoding technique associated with the encoded data.

[011] Additionally or alternatively, the set of predetermined weights are based on training of a neural network on known encoded data and delayed versions of combinations of the known encoded data.

[012] Additionally or alternatively, the neural network is configured to use the set of predetermined weights to provide an estimate of decoded data based on the received encoded data.

[013] Additionally or alternatively, the apparatus further includes a memory encoded with executable instructions to select the set of predetermined weights for the neural network in accordance with the encoding technique associated with the encoded data.

[014] Additionally or alternatively, the encoding technique comprises Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.

[015] Additionally or alternatively, the first stage of circuitry and second stage of circuitry comprises a first plurality of multiplication/accumulation units, the first plurality of multiplication/accumulation units each configured to multiply at least one bit of the encoded data with at least one of the set of predetermined weights and sum multiple weighted bits of the encoded data.

[016] Additionally or alternatively, the first stage of circuitry further comprises a first plurality of memory look-up units (MLUs), the first plurality of MLUs each configured to retrieve at least one intermediate data value corresponding to an output of a respective one of the first plurality of multiplication/accumulation units based on the at least one non-linear function.

[017] Additionally or alternatively, the first stage of circuitry further comprises a plurality of delay units configured to provide the delayed versions of the combinations of the encoded data based on the at least one intermediate data value provided by the first plurality of MLUs.

[018] Additionally or alternatively, the apparatus further includes a memory in communication with the first stage of circuitry and the second stage of MAC circuitry, the memory configured to store multiple sets of predetermined weights, including the set of predetermined weights.

[019] In another aspect of the disclosure, a method includes receiving, at a computing device that comprises a neural network, signaling indicative of a set of data pairs each comprising known encoded data and decoded data, wherein the known encoded data is encoded with a particular encoding technique; determining a set of weights for the neural network to decode data with the particular encoding technique based partly on the signaling, coefficient multiplication results associated with the signaling, and delayed versions of the coefficient multiplication results associated with the signaling; receiving signaling, from a memory of the computing device, indicative of data encoded with the particular encoding technique; decoding the data using the neural network using the weights; and writing the decoded data to or reading the decoded data from a memory or storage medium of the computing device.

[020] Additionally or alternatively, further included is determining multiple sets of weights for the neural network using additional encoded data, decoded data pairs encoded with other encoding techniques, and delayed versions of additional coefficient multiplication results based on the additional encoded data and the decoded data pairs encoded with the other encoding techniques, wherein the multiple sets of weights each correspond with a different encoding technique; selecting the set of weights associated with the particular encoding technique; and providing the set of weights to the neural network or another neural network for use in decoding the further encoded data.

[021] Additionally or alternatively, further included is receiving further input data encoded with another encoding technique; selecting a selected set of weights of the multiple sets of

weights associated with the another encoding technique; providing the selected set of weights to the neural network; and decoding the further input data using the selected set of weights.

[022] Additionally or alternatively, the particular encoding technique comprises Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.

[023] Additionally or alternatively, decoding the further encoded data using the neural network comprises weighting bits of the further encoded data using at least some of the set of weights, and summing selected weighted bits of the further encoded data to provide intermediate data, then weighting bits of the intermediate data using at least some of the weights and summing selected weighted bits of the intermediate data and selected delayed versions of the weighted bits of the intermediate data to provide decoded data.

[024] Additionally or alternatively, determining the set of weights comprises selecting weights resulting in a minimized value of an error function between an output of the neural network and known decoded data.

[025] Additionally or alternatively, determining the set of weights comprises using a k-means cluster technique to determine center vectors for the neural network

[026] In another aspect of the disclosure, a method includes mixing encoded data received at a processing unit using a plurality of coefficients and an additional plurality of coefficients, both the plurality of coefficients and the additional plurality of coefficients specific to an encoding technique associated with the encoded data. Mixing the encoded data comprises calculating, at a first layer of multiplication/accumulation processing units (MAC units) of a plurality of MAC units, the encoded data and delayed versions of respective outputs of the first layer of MAC units with the plurality of coefficients to generate first processing results; calculating, at additional layers of MAC units of the plurality of MAC units, the first processing results and delayed versions of at least a portion of the first processing results with the additional plurality of coefficients to generate second processing results; and providing output data based partly on the second processing results, the output data representative of the encoded data being decoded, such that the output data is an estimate of decoded data based on the encoded data.

[027] Additionally or alternatively, mixing the encoded data received at the processing unit using the plurality of coefficients further comprises delaying, at respective delay units associated with the first layer of MAC units, the respective outputs of the first layer of MAC units to generate the delayed versions of the respective outputs of the first layer of MAC units.

[028] Additionally or alternatively, calculating, at the first layer of multiplication/accumulation processing units (MAC units) of the plurality of MAC units, the encoded data and the delayed versions of the respective outputs of the first layer of MAC units with the plurality of coefficients comprises multiplying the encoded data and the delayed versions of the respective outputs of the first layer of MAC units with respective coefficients of the plurality of coefficients to generate the first processing results.

[029] Additionally or alternatively, further included is obtaining, at the processing unit, signaling indicative of the encoded data from a memory coupled, the signaling indicative of the encoded data including an indication from a touchscreen of a mobile communication device that the encoding technique was utilized.

[030] Additionally or alternatively, the encoding technique comprises Reed-Solomon coding, Bose-Chaudhuri-Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.

DETAILED DESCRIPTION

[031] Multi-layer neural networks and/or multi-layer recurrent neural networks may be used to decode encoded data (e.g., data encoded using one or more encoding techniques). Such neural networks may have nonlinear mapping and distributed processing capabilities which may be advantageous in many systems employing the neural network decoders. In this manner, neural networks described herein may be used to implement error correction coding (ECC) decoders.

[032] An encoder may have L bits of input data $(a_1, a_2, \dots, a_L)$. The encoder may encode the input data in accordance with an encoding technique to provide N bits of encoded data $(b_1, b_2, \dots, b_N)$. The encoded data may be stored and/or transmitted, or some other action taken with the encoded data, which may introduce noise into the data. Accordingly, a decoder may receive a version of the N bits of encoded data $(x_1, x_2, \dots, x_N)$. The decoder may decode the received encoded data into an estimate of the L bits original data $(y_1, y_2, \dots, y_L)$.

[033] Examples of wireless baseband circuitry may utilize error correction coding (such as low density parity check coding, LDPC). An encoder may add particularly selected $N-L$ bits into an original data of L bits, which may allow a decoder to decode the data and reduce and/or minimize errors introduced by noise, interferences and other practical factors in the data storage and transmission.

[034] There are a variety of particular error correction coding techniques, including low density parity check coding (LDPC), Reed-Solomon coding, Bose-Chaudhuri-Hocquenghem (BCH), and Polar coding. The use of these coding techniques, however, may come at the cost of the decrease of the frequency and/or channel and/or storage resource usage efficiency and the increase of the processing complexity. For example, the use of coding techniques may increase the amount of data which may be stored and/or transmitted. Moreover, processing resources may be necessary to implement the encoding and decoding. In some examples, the decoder may be one of the processing blocks that cost the most computational resources in wireless baseband circuitry and/or memory controllers, which may reduce the desirability of existing decoding schemes in many emerging applications such as Internet of Things (IoT) and/or tactile internet where ultra-low power consumption and ultra-low latency are highly desirable.

[035] Examples described herein utilize multi-layer neural networks (NNs), such as multi-layer recurrent neural networks (RNNs) to decode encoded data (e.g., data encoded using one or more encoding techniques). The NNs and/or RNNs have nonlinear mapping and distributed processing capabilities which may be advantageous in many systems employing the neural network decoders. For example, in some non-volatile memory devices where bit flips may degrade memory storage capabilities, NNs and/or RNNs may provide a more robust decoder that may be trained to transform noisy encoded input data to decoded data (e.g., an estimate of the decoded data). Advantageously, such noisy encoded data, may be decoded such that the NNs or RNNs reduce and/or improve errors which may be introduced by noise present in the input data. In the example, such noise may be introduced in storing the encoded data in memory that is degraded (e.g., due to bit flips).

[036] Figure 1A is a schematic illustration of an electronic device 110 arranged in a system 100 in accordance with examples described herein. The electronic device 110 is coupled to a memory 140 via network 120 or memory 145, each of which may store coefficient data. Coefficient data stored in the memory 140 or memory 145 may include coefficient data which may be mixed with noisy encoded input data received by the electronic device 110 in examples described herein. The encoded input data may also be stored in the memory 140 or 145. The electronic device 110 may obtain the encoded input data (e.g., m encoded input bits) from the memory 140 or memory 145 to decode the data to be output as decoded output data.

[037] Electronic device 110 also includes processing units 112 that may interact with mode configurable control 105, which may be encoded with instructions executable by the processing

unit(s) 112. In some implementations, mode configurable control 105 may be implemented as a memory. As used herein, memory may refer to computer readable media, which may include both storage media and communication media. Example computer readable media may include volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information, such as computer readable instructions. The mode configurable control 105 includes the executable instructions 115 for mixing encoded input data with coefficients. For example, the mode configuration control 105 may be implemented using circuitry (e.g., logic), one or more processor(s), microcontroller(s), controller(s), or other elements. The mode configuration control 105 may select certain weights and/or other parameters (e.g., from memory 145 which may store weights) and provide those weights and/or other parameters to one or more of the multiplication/accumulation units and/or memory look-up units of Figures 2A-2C whose details will be presented after describing Figures 1A-1C. In some examples, weights and/or other parameters stored in the memory 145 may be associated with a particular encoding technique; such as Reed-Solomon coding, Bose-Chaudhuri-Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof. During operation, the mode configuration control 105 may be used to select weights and/or other parameters in memory 140 or memory 145 based on an indicated encoding technique. For example, header information or another flag associated with encoded input data (e.g., obtained from memory 140 or 145) may indicate a particular encoding technique. In the example, the mode configuration control 105 may select weights and/or other parameters from the memory 140 or memory 145 to decode encoded input data into decoded data.

[038] The processing unit(s) 112 may be implemented using one or more processors, for example, having any number of cores. In some examples, the processing unit(s) 112 may include circuitry, including custom circuitry, and/or firmware for performing functions described herein. For example, circuitry can include multiplication unit/accumulation units for performing the described functions, as described herein. Processing unit(s) 112 can be any type including but not limited to a microprocessor or a digital signal processor (DSP), or any combination thereof. For example, processing unit(s) 112 can include levels of caching, such as a level one cache and a level two cache, a core, and registers. An example processor core can include an arithmetic logic unit (ALU), a bit manipulation unit, a multiplication unit, an accumulation unit, an adder unit, a look-up table unit, a memory look-up unit, or any

combination thereof. Examples of processing unit 112 are described herein, for example with reference to Figure 2A, 2B, or 2C.

[039] The mode configurable control 105, for example, may be encoded with executable instructions 115 for mixing encoded input data with coefficient data, e.g., to mix noisy encoded input data with coefficient data at the processing units 112. For example, in the context of decoding noisy encoded input data from memory 140 or 145, the executable instructions 115 for mixing encoded input data with coefficient data may include instructions for obtaining the noisy encoded input data from the memory 140 or 145; and to transform that obtained noisy encoded input data at the processing unit(s) 112 into decoded data (e.g., an estimate of the decoded data). For example, the executable instructions 115 for mixing encoded input data with coefficient data may further include instructions for multiplying a portion of the noisy encoded input data with coefficient data to generate a coefficient multiplication result and accumulating the coefficient multiplication result to be further multiplied and accumulated with another portion of the noisy encoded input data and coefficient data, examples of which are described herein. For example, to generate a coefficient multiplication result, a first layer of multiplication/accumulation processing units (MAC units) may calculate the noisy encoded input data with the plurality of coefficients to generate such coefficient multiplication results, or first processing results of the first layer of MAC units. Continuing in the example, to provide the output data, additional layers of MAC units may calculate the first processing results with additional pluralities of coefficients to generate additional coefficient multiplication results, or second processing results of the additional layers of MAC units. The MLUs of a last layer of the additional layers of MAC units may provide the decoded data based on the second processing results. Accordingly, the executable instructions 115 for mixing encoded input data with coefficient data may include various sets of executable instructions for different types of hardware implementations, such as those shown in Figures 2A-2C, to implement such mixing of noisy encoded input data with coefficient data.

[040] The user interface 114 may be implemented with any of a number of input devices including, but not limited to, a touchscreen, keyboard, mouse, microphone, or combinations thereof. The user interface 114 may receive input from a user, for example, regarding a processing mode selection to specify a processing mode for the processing unit(s) 112. The user interface 114 may communicate the user input to the mode configurable control 105. Example user interfaces 114 include a serial interface controller or a parallel interface

controller, which may be configured to communicate with external input devices (e.g., keyboard, mouse, pen, voice input device, touch input device, etc.).

[041] The network 120 may include a wired network or direct-wired connection, and wireless media such as acoustic, radio frequency (RF), microwave, infrared (IR) and other wireless media.

[042] The memory(s) 140, and 145 (or mode configurable control 105, if being implemented as a memory) may be implemented using any storage medium accessible to the processing unit(s) 112. For example, RAM, ROM, solid state memory, flash memory, disk drives, system memory, optical storage, or combinations thereof, may be used to implement the mode configurable control 105 or memory(s) 140, and 145. In storing encoded data in such memories, environmental or other noise may be introduced in the storing process. For example, noise may be introduced in storing the encoded data in memory 140 or 145 that is degraded (e.g., due to bit flips). Accordingly, data obtained from the memory(s) 140 or 145 may be referred to as noisy encoded input data. In some implementations, the mode configurable control 105 may store associations between coefficients and particular encoding techniques described herein.

[043] The electronic device 110 may be implemented using any of a variety of computing systems, including but not limited to one or more desktop, server, laptop, or other computers. The electronic device 110 generally includes one or more processing unit(s) 112. The computing system 100 may be implemented as a mobile communication device using any user communication device, including but not limited to, a desktop, laptop, cellular phone, tablet, appliance, automobile, or combinations thereof. The electronic device 110 may be programmed with a mobile application (e.g. processing unit(s) 112 and computer readable media encoded with instructions which, when executed, cause the electronic device 110 to perform described functions) for mixing noisy encoded input data with coefficient data. For example, the electronic device 110 may be programmed to receive an indication from a touchscreen of a mobile communication device that a particular encoding technique was utilized for the noisy encoded input data received in a 5G wireless transmission.

[044] It is to be understood that the arrangement of computing systems of the system 100 may be quite flexible, and although not shown, it is to be understood that the system 100 may include many electronic devices 110, which may be connected via the network 120 can operate in conjunction with each other to perform the systems and methods described herein. The memory 145 and/or the memory 140 may in some examples be implemented using the same media, and in other examples may be implemented using different media. For example, while

the memory 140 is shown in Figure 1A as coupled to the network 120, it can be appreciated that the memory 140 may also be implemented electronic device 110 as part of the mode configurable control 105. Additionally, while a single user interface 114 is shown in Figure 1A, it can be appreciated that the electronic device 110 may further include any number of input devices, output devices, and/or peripheral components. For example, the user interface 114 may be the interface of a mobile communication device.

[045] Figure 1B is a schematic illustration of a neural network arranged in accordance with examples described herein. The neural network 150 include three stages (e.g., layers). While three stages are shown in Figure 1B, any number of stages may be used in other examples. A first stage of neural network 150 includes node 168, node 169, node 172, and node 174. A second stage of neural network 150 includes combiner 152, combiner 154, combiner 156, and combiner 158. A third stage of neural network 150 includes combiner 160, combiner 162, combiner 164, and combiner 166. Additional, fewer, and/or different components may be used in other examples.

[046] Generally, a neural network may be used including multiple stages of nodes. The nodes may be implemented using processing units (e.g., processing units 112) which may execute one or more functions on inputs received from a previous stage and provide the output of the functions to the next stage of the neural network. The processing units may be implemented using, for example, one or more processors, controllers, and/or custom circuitry, such as an application specific integrated circuit (ASIC) and/or a field programmable gate array (FPGA). In some examples, the processing units may be implemented using any combination of one or more processing units described with respect to Figures 2A-2C. The processing units may be implemented as combiners and/or summers and/or any other structure for performing functions allocated to the processing unit. In some examples, certain of the elements of neural networks described herein perform weighted sums, e.g., may be implemented using one or more multiplication/accumulation units, which may be implemented using processor(s) and/or other circuitry. In an example, the neural network 150 may be implemented by the electronic device 110.

[047] In the example, of Figure 1B, the neural network 150 may have an input layer, which may be a first stage of the neural network including node 168, node 169, node 172, and node 174. The nodes node 168, node 169, node 172, and node 174 may implement a linear function which may provide the input signals (*e.g.*, $x_1(n), x_2(n), \dots, x_m(n)$) to another stage of the neural network (e.g., a 'hidden stage' or 'hidden layer'). Accordingly, in the example of Figure

1B, m bits of noisy encoded input data may be provided to an input stage (e.g., an input layer) of a neural network during operation. In some examples, the input data may be data encoded in accordance with an encoding technique (e.g., low density parity check coding (LDPC), Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH), and/or Polar coding). The m bits of input data may be output by the first stage of the neural network 150 to a next stage of the neural network 150. In some examples, the connection between the first stage and the second stage of the neural network 150 may not be weighted - e.g., processing units in the second stage may receive unaltered bits from the first stage in some examples. Each of the input bit may be provided to multiple ones of the processing units in the next stage. While an input layer is shown, in some examples, the input layer may not be present.

[048] The neural network 150 may have a next layer, which may be referred to as a 'hidden layer' in some examples. The next layer may include combiner 152, combiner 154, combiner 156, and combiner 158, although any number of elements may be used. While the processing elements in the second stage of the neural network 150 are referred to as combiners, generally the processing elements in the second stage may perform a nonlinear activation function using the input data bits received at the processing element. Any number of nonlinear activation functions may be used. Examples of functions which may be used include Gaussian functions, such as $f(r) = \exp\left(-\frac{r^2}{\sigma^2}\right)$. Examples of functions which may be used include multi-quadratic functions, such as $f(r) = (r^2 + \sigma^2)^{1/2}$. Examples of functions which may be used include inverse multi-quadratic functions, such as $f(r) = (r^2 + \sigma^2)^{-1/2}$. Examples of functions which may be used include thin-plate-spline functions, such as $f(r) = r^2 \log(r)$. Examples of functions which may be used include piece-wise linear functions, such as $f(r) = \frac{1}{2}(|r + 1| - |r - 1|)$. Examples of functions which may be used include cubic approximation functions, such as $f(r) = \frac{1}{2}(|r^3 + 1| - |r^3 - 1|)$. In these example functions, σ represents a real parameter (e.g., a scaling parameter) and r is the distance between the input vector and the current vector. The distance may be measured using any of a variety of metrics, including the Euclidean norm.

[049] Each element in the 'hidden layer' may receive as inputs selected bits (e.g., some or all) of the input data. For example, each element in the 'hidden layer' may receive as inputs from the output of multiple selected elements (e.g., some or all elements) in the input layer. For example, each element in the 'hidden layer' may receive as inputs from the output of multiple

selected units (e.g., some or all units) in the input layer. For example, the combiner 152 may receive as inputs the output of node 168, node 169, node 172, and node 174. While a single 'hidden layer' is shown by way of example in Figure 1B, any number of 'hidden layers' may be present and may be connected in series. While four elements are shown in the 'hidden layer', any number may be used, and they may be the same or different in number than the number of nodes in the input layer and/or the number of nodes in any other hidden layer. The nodes in the hidden layer may evaluate at least one non-linear function using combinations of the data received at the hidden layer node (e.g., element). In this manner, the hidden layer may provide intermediate data at an output of one or more hidden layers.

[050] The neural network 150 may have an output layer. The output layer in the example of Figure 1B may include combiner 160, combiner 162, combiner 164, and combiner 166, although any number of elements may be used. While the processing element in the output stage of the neural network 150 are referred to as combiners, generally the processing elements in the output may perform any combination or other operation using data bits received from a last 'hidden layer' in the neural network. Each element in the output layer may receive as inputs selected bits (e.g., some or all) of the data provided by a last 'hidden layer'. For example, the combiner 160 may receive as inputs from the outputs of combiner 152, combiner 154, combiner 156, and combiner 158. The connections between the hidden layer and the output layer may be weighted. For example, a set of weights W may be specified. There may be one weight for each connection between a hidden layer node and an output layer node in some examples. In some examples, there may be one weight for each hidden layer node that may be applied to the data provided by that node to each connected output node. Other distributions of weights may also be used. The weights may be multiplied with the output of the hidden layer before the output is provided to the output layer. In this manner, the output layer may perform a sum of weighted inputs. Accordingly, an output of the neural network 150 (e.g., the outputs of the output layer) may be referred to as a weighted sum. The output layer may accordingly combine intermediate data received from one or more hidden layers using weights to provide output data.

[051] In some examples, the neural network 150 may be used to provide L output bits which represent processed data corresponding to m input bits. For example, in the example of Figure 1B, m input bits are shown $(x_1(n), x_2(n), \dots, x_m(n))$ and L output bits are provided $(y_1(n), y_2(n), \dots, y_L(n))$. The neural network 150 may be trained such that the weights W used and/or the functions provided by the elements of the hidden layers cause the neural

network 150 to provide output bits which represent the processed data corresponding to the m encoded input bits. The input bits may have been encoded with an encoding technique, and the weights and/or functions provided by the elements of the hidden layers may be selected in accordance with the encoding technique. Accordingly, the neural network 150 may be trained multiple times - once for each encoding technique that may be used to provide the neural network 150 with input data.

[052] Examples of neural networks may be trained. Training generally refers to the process of determining weights, functions, and/or other attributes to be utilized by a neural network to create a desired transformation of input data to output data. In some examples, neural networks described herein may be trained to transform encoded input data to decoded data (e.g., an estimate of the decoded data). In some examples, neural networks described herein may be trained to transform noisy encoded input data to decoded data (e.g., an estimate of the decoded data). In this manner, neural networks may be used to reduce and/or improve errors which may be introduced by noise present in the input data. In some examples, neural networks described herein may be trained to transform noisy encoded input data to encoded data with reduced noise. The encoded data with reduced noise may then be provided to any decoder (e.g., a neural network and/or other decoder) for decoding of the encoded data. In this manner, neural networks may be used to reduce and/or improve errors which may be introduced by noise.

[053] Training as described herein may be supervised or un-supervised in various examples. In some examples, training may occur using known pairs of anticipated input and desired output data. For example, training may utilize known encoded data and decoded data pairs to train a neural network to decode subsequent encoded data into decoded data. In some examples, training may utilize known noisy encoded data and decoded data pairs to train a neural network to decode subsequent noisy encoded data into decoded data. In some examples, training may utilize known noisy encoded data and encoded data pairs to train a neural network to provide encoded data having reduced noise than input noisy encoded data. Examples of training may include determining weights to be used by a neural network, such as neural network 150 of Figure 1B. In some examples, the same neural network hardware is used during training as will be used during operation. In some examples, however, different neural network hardware may be used during training, and the weights, functions, or other attributes determined during training may be stored for use by other neural network hardware during operation.

[054] Examples of training can be described mathematically. For example, consider input data at a time instant (n), given as: $X(n) = [x_1(n), x_2(n), \dots, x_m(n)]^T$. The center vector for each element in hidden layer(s) of the neural network 150 (e.g., combiner 152, combiner 154, combiner 156, and combiner 158) may be denoted as C_i (for $i = 1, 2, \dots, H$, where H is the element number in the hidden layer).

[055] The output of each element in a hidden layer may then be given as:

$$h_i(n) = f_i(\|X(n) - C_i\|) \text{ for } (i = 1, 2, \dots, H) \quad (1)$$

[056] The connections between a last hidden layer and the output layer may be weighted. Each element in the output layer may have a linear input-output relationship such that it may perform a summation (e.g., a weighted summation). Accordingly, an output of the i 'th element in the output layer at time n may be written as:

$$y_i(n) = \sum_{j=1}^H W_{ij} h_j(n) = \sum_{j=1}^H W_{ij} f_j(\|X(n) - C_j\|) \quad (2)$$

for $(i = 1, 2, \dots, L)$ and where L is the element number of the output of the output layer and W_{ij} is the connection weight between the j 'th element in the hidden layer and the i 'th element in the output layer.

[057] Generally, a neural network architecture (e.g., the neural network 150 of Figure 1B) may include a number of elements and may have center vectors which are distributed in the input domain such that the neural network may approximate nonlinear multidimensional functions and therefore may approximate forward mapping and inverse mapping between two code words (e.g., from an N -bit input to an L -bit output). Generally, the choice of transfer function used by elements in the hidden layer may not affect the mapping performance of the neural network, and accordingly, a function may be used which may be implemented conveniently in hardware in some examples. For example, a thin-plate-spline function and/or a Gaussian function may be used in various examples and may both provide adequate approximation capabilities. Other functions may also be used.

[058] Examples of neural networks may accordingly be specified by attributes (e.g., parameters). In some examples, two sets of parameters may be used to specify a neural network: connection weights and center vectors (e.g., thresholds). The parameters may be determined from selected input data (e.g., encoded input data) by solving an optimization function. An example optimization function may be given as:

$$E = \sum_{n=1}^M \|Y(n) - \widehat{Y(n)}\|^2 \quad (3)$$

where M is a number of trained input vector (e.g., trained encoded data inputs) and $Y(n)$ is an output vector computed from the sample input vector using Equations (1) and (2) above, and $\widehat{Y}(n)$ is the corresponding desired (e.g., known) output vector. The output vector $Y(n)$ may be written as:

$$Y(n) = [y_1(n), y_2(n), \dots \dots y_L(n)]^T$$

[059] Various methods (e.g., gradient descent procedures) may be used to solve the optimization function. However, in some examples, another approach may be used to determine the parameters of a neural network, which may generally include two steps – (1) determining center vectors C_i ($i = 1, 2, \dots, H$) and (2) determining the weights.

[060] In some examples, the center vectors may be chosen from a subset of available sample vectors. In such examples, the number of elements in the hidden layer(s) may be relatively large to cover the entire input domain. Accordingly, in some examples, it may be desirable to apply k-means cluster algorithms. Generally, k-means cluster algorithms distribute the center vectors according to the natural measure of the attractor (e.g., if the density of the data points is high, so is the density of the centers). k-means cluster algorithms may find a set of cluster centers and partition the training samples into subsets. Each cluster center may be associated with one of the H hidden layer elements in this network. The data may be partitioned in such a way that the training points are assigned to the cluster with the nearest center. The cluster center corresponding to one of the minima of an optimization function. An example optimization function for use with a k-means cluster algorithm may be given as:

$$E_{k_means} = \sum_{j=1}^H \sum_{n=1}^M B_{jn} \|X(n) - C_j\|^2 \quad (4)$$

where B_{jn} is the cluster partition or membership function forming an $H \times M$ matrix. Each column may represent an available sample vector (e.g., known input data) and each row may represent a cluster. Each column may include a single '1' in the row corresponding to the cluster nearest to that training point, and zeros elsewhere.

[061] The center of each cluster may be initialized to a different randomly chosen training point. Then each training example may be assigned to the element nearest to it. When all training points have been assigned, the average position of the training point for each cluster may be found and the cluster center is moved to that point. The clusters may become the desired centers of the hidden layer elements.

[062] In some examples, for some transfer functions (e.g., the Gaussian function), the scaling factor σ may be determined, and may be determined before determining the connection

weights. The scaling factor may be selected to cover the training points to allow a smooth fit of the desired network outputs. Generally, this refers to any point within the convex hull of the processing element centers may significantly activate more than one element. To achieve this goal, each hidden layer element may activate at least one other hidden layer element to a significant degree. An appropriate method to determine the scaling parameter σ may be based on the P-nearest neighbor heuristic, which may be given as,

$$\sigma_i = \frac{1}{p} \sum_{j=1}^p \|C_j - C_i\|^2 \quad (i = 1, 2, \dots, H) \text{ where } C_j \text{ (for } i = 1, 2, \dots, H) \text{ are the } P\text{-nearest neighbors of } C_i.$$

[063] The connection weights may additionally or instead be determined during training. In an example of a neural network, such as neural network 150 of Figure 1B, having one hidden layer of weighted connections an output elements which are summation units, the optimization function of Equation (3) may become a linear least-squares problem once the center vectors and the scaling parameter have been determined. The linear least-squares problem may be written as

$$\min_W \sum_{n=1}^M \|Y(n) - \widehat{Y}(n)\|^2 = \min_W \sum_{n=1}^M \|WF - \widehat{Y}\|^2 \quad (5)$$

where $W = \{W_{ij}\}$ is the $L \times H$ matrix of the connection weights, F is an $H \times M$ matrix of the outputs of the hidden layer processing elements and whose matrix elements are computed using $F_{in} = f_i(\|X(n) - C_i\|) \quad (i = 1, 2, \dots, H; n = 1, 2, \dots, M)$ and $\widehat{Y} = [\widehat{Y}(1), \widehat{Y}(2), \dots, \widehat{Y}(M)]$ is the $L \times M$ matrix of the desired (e.g., known) outputs. The connection weight matrix W may be found from Equation (5) and may be written as follows:

$$\widehat{W} = \widehat{Y} F^+ = \widehat{Y} \lim_{\alpha \rightarrow 0} F^T (FF^T + \alpha I)^{-1} \quad (6)$$

where F^+ is the pseudo-inverse of F . In this manner, the above may provide a batch-processing method for determining the connection weights of a neural network. It may be applied, for example, where all input sample sets are available at one time. In some examples, each new sample set may become available recursively, such as in the recursive-least-squares algorithms (RLS). In such cases, the connection weights may be determined as follows.

[064] First, connection weights may be initialized to any value (e.g., random values may be used). The output vector $Y(n)$ may be computed using Equation (2). The error term $e_i(n)$ of each output element in the output layer may be computed as follows:

$$e_i(n) = y_i(n) - \widehat{y}_i(n) \quad (i = 1, 2, \dots, L)$$

The connection weights may then be adjusted based on the error term, for example as follows:

$$W_{ij}(n+1) = W_{ij}(n) + \gamma e_i(n) f_j(\|X(n) - C_i\|) \quad (7)$$

$$(i = 1, 2, \dots, L; j = 1, 2, \dots, M)$$

where γ is the learning-rate parameter which may be fixed or time-varying.

[065] The total error may be computed according to the output from the output layer and the desired (known) data:

$$\epsilon = \|Y(n) - \widehat{Y(n)}\|^2 \quad (8)$$

[066] The process may be iterated by again calculating a new output vector, error term, and again adjusting the connection weights. The process may continue until weights are identified which reduce the error to equal to or less than a threshold error.

[067] Accordingly, the neural network 150 of Figure 1B may be trained to determine parameters (e.g., weights) for use by the neural network 150 to perform a particular mapping between input and output data. For example, training the neural network 150 may provide one set of parameters to use when decoding encoded data that had been encoded with a particular encoding technique (e.g., low density parity check coding (LDPC), Reed-Solomon coding, Bose-Chaudhuri-Hocquenghem (BCH), and/or Polar coding). The neural network 150 (and/or another neural network) may be trained multiple times, using different known input/output data pairs, for example. Multiple trainings may result in multiple sets of connection weights. For example, a different set of weights may be determined for each of multiple encoding techniques – e.g., one set of weights may be determined for use with decoding LDPC encoded data and another set of weights may be determined for use with decoding BCH encoded data.

[068] Recall that the structure of neural network 150 of Figure 1B is provided by way of example only. Other multilayer neural network structures may be used in other examples. For example, as described below with respect to Figure 1C, a multilayer recurrent neural network structure may be utilized as disclosed herein. Moreover, the training procedures described herein are also provided by way of example. Other training techniques (e.g., learning algorithms) may be used, for example, to solve the local minimum problem and/or vanishing gradient problem. Determined weights and/or vectors for each decoder may be obtained by an off-line learning mode of the neural network, which may advantageously provide more resources and data. In some implementations, for example as described with reference to Figures 2A-2C, the hidden layers of combiners (e.g., combiner 152, combiner 154, combiner 156, and combiner 158) may include multiplication/accumulation (MAC) units, with each layer having additional MAC units. Such implementations, having accumulated the intermediate

processing results in a respective processing elements (e.g., the respective MAC unit), may also include memory look-up (MLU) units that are configured to retrieve a plurality of coefficients and provide the plurality of coefficients as the connection weights (e.g., determined weights) for that respective layer of MAC units to be mixed with the input data.

[069] In examples of supervised learning, the input training samples: $[x_1(n), x_2(n), \dots, x_m(n)]$ may be generated by passing the encoded samples $[b_1(n), b_2(n), \dots, b_m(n)]$ through some noisy channels and/or adding noise. The supervised output samples may be the corresponding original code $[a_1(n), a_2(n), \dots, a_L(n)]$ which may be used to generate $[b_1(n), b_2(n), \dots, b_m(n)]$ by the encoder. Once these parameters are determined in offline mode, the desired decoded code-word can be obtained from input data utilizing the neural network (e.g., computing Equation (2)), which may avoid complex iterations and feedback decisions used in traditional error-correcting decoding algorithms. In this manner, neural networks described herein may provide a reduction in processing complexity and/or latency, because some complexity has been transferred to an off-line training process which is used to determine the weights and/or functions which will be used. Further, the same neural network (e.g., the neural network 150 of Figure 1B) can be used to decode an input code-word encoded from any of multiple error correction encoder by selecting different weights that were obtained by the training for the particular error correction technique employed. In this manner, neural networks may serve as a universal decoder for multiple encoder types.

[070] Figure 1C is a schematic illustration of a recurrent neural network arranged in accordance with examples described herein. The neural network 170 include three stages (e.g., layers): an inputs node 171; a combiner stage 173 and 175, and an outputs node 177. While three stages are shown in Figure 1C, any number of stages may be used in other examples, e.g., as described with reference to Figures 2A-2C. In some implementations, the neural network 170 may have multiple combiner stages such that outputs from one combiner stage is provided to another combiners stage, until being providing to an outputs node 177. As described with reference to Figure 2A, for example, there may be multiple combiner stages in a neural network 170. As depicted in Figure 1C, the delay units 175a, 175b, and 175c may be optional components of the neural network 170. When such delay units 175a, 175b, and 175c are utilized as described herein, the neural network 170 may be referred to as a recurrent neural network.

[071] The first stage of the neural network 170 includes inputs node 171. The inputs node 171 may receive input data at various inputs of the recurrent neural network. In some examples,

the inputs node 171 may include multiple input nodes, such as input node 168, node 169, node 172, and node 174 of Figure 1B. The second stage of the neural network 170 is a combiner stage including combiner units 173a, 173b, 173c; and delay units 175a, 175b, 175c. Accordingly, the combiner units 173 and delay units 175 may be collectively referred to as a stage of combiners. In some implementations, the combiner units 173a, 173b, and 173c may correspond to combiner 152, combiner 154, and combiner 156 of Figure 1B, for example. Accordingly, as described with respect to Figure 1B with processing units 112 implementing such combiners, generally processing units 112 that implement the combiner units 173a-c and delay units 175a-c in the second stage may perform a nonlinear activation function using the input data from the inputs node 171 (e.g., input signals $X_1(n)$, $X_2(n)$, and $X_3(n)$). The third stage of neural network 170 includes the outputs node 177. In some examples, the outputs node 177 may include combiner 160, combiner 162, and combiner 164 of Figure 1B. Accordingly, in some examples, the outputs nodes 177 may be referred to as a stage of combiners. Additional, fewer, and/or different components may be used in other examples.

[072] The recurrent neural network 170 includes delay units 175a, 175b, and 175c, which generate delayed versions of the output from the respective combiner units 173a-c based on receiving such output data from the respective combiner units 173a-c. In the example, the output data of combiner units 173a-c may be represented as $h(n)$; and, accordingly, each of the delay units 175a-c delay the output data of the combiner units 173a-c to generate delayed versions of the output data from the combiner units 173a-c, which may be represented as $h(n-t)$. In various implementations, the amount of the delay, t , may also vary, e.g., one clock cycle, two clock cycles, or one hundred clock cycles. That is, the delay unit 175 may receive a clock signal and utilize the clock signal to identify the amount of the delay. In the example of Figure 1C, the delayed versions are delayed by one time period, where '1' represents a time period. A time period may correspond to any number of units of time, such as a time period defined by a clock signal or a time period defined by another element of the neural network 170.

[073] Continuing in the example of Figure 1C, each delay unit 175a-c provides the delayed versions of the output data from the combiner units 173a-c as input to the combiner units 173a-c, to operate, optionally, as a recurrent neural network. Such delay units 175a-c may provide respective delayed versions of the output data from nodes of the combiner units 173a-c to respective input units/nodes of the combiner units 173a-c. In utilizing delayed versions of output data from combiner units 173a-c, the recurrent neural network 170 may train weights at the combiner units 173a-c that incorporate time-varying aspects of input data to be processed

by such a recurrent neural network 170. Once trained, in some examples, the inputs node 171 receives encoded input data that is to be decoded in the recurrent neural network 170. For example, each stream of input data may correspond to a different obtained set of encoded data that is representative of a temporal signal. Because an RNN 170 incorporates the delayed versions of output data from combiner units 173a-c, the delayed versions of output data from the combiner units 173a-c provide feedback information representative of the temporal nature, with the RNN 170 providing output decoded faster with the RNN 170 incorporating that temporal nature into calculating the output decoded data. In the example, the output decoded data may be representative of an estimate of the encoded data decoded into the original data that was encoded according to the associated encoding technique.

[074] Generally, a recurrent neural network may include multiple stages of nodes. The nodes may be implemented using processing units (e.g., processing units 112) which may execute one or more functions on inputs received from a previous stage and provide the output of the functions to the next stage of the recurrent neural network. The processing units may be implemented using, for example, one or more processors, controllers, and/or custom circuitry, such as an application specific integrated circuit (ASIC) and/or a field programmable gate array (FPGA). In some examples, the processing units may be implemented using any combination of one or more processing units 112 described with respect to Figures 2A-2C. The processing units may be implemented as combiners and/or summers and/or any other structure for performing functions allocated to the processing unit. In some examples, certain of the elements of neural networks described herein perform weighted sums, e.g., may be implemented using one or more multiplication/accumulation units, which may be implemented using processor(s) and/or other circuitry. In an example, the neural network 170 may be implemented by the electronic device 110 utilizing any combination of one or more processing units described with respect to Figures 2A-2C.

[075] Examples of recurrent neural network training and inference can be described mathematically. Again, as an example, consider input data at a time instant (n), given as: $X(n) = [x_1(n), x_2(n), \dots, x_m(n)]^T$. The center vector for each element in hidden layer(s) of the recurrent neural network 170 (e.g., combiner units 173 including combiner 152, combiner 154, combiner 156, and combiner 158 of Figure 1B) may be denoted as C_i (for $i = 1, 2, \dots, H$, where H is the element number in the hidden layer).

[076] The output of each element in a hidden layer may then be given as:

$$h_i(n) = f_i(\|X(n) + h_i(n-t) - C_i\|) \text{ for } (i = 1, 2, \dots, H) \quad (9)$$

t may be the delay at the delay unit 175 such that the output of the combiner units 173 includes a delayed version of the output of the combiner units 173. In some examples, this may be referred to as feedback of the combiner units 173. Accordingly, each of the connections between a last hidden layer and the output layer may be weighted. Each element in the output layer may have a linear input-output relationship such that it may perform a summation (e.g., a weighted summation). Accordingly, an output of the i 'th element in the output layer at time n may be written as:

$$\begin{aligned} y_i(n) &= \sum_{j=1}^H W_{ij} h_j(n) + W_{ij} h_j(n-t) \\ &= \sum_{j=1}^H W_{ij} f_j(\|X(n) + h_i(n-t) - C_j\|) \end{aligned} \quad (10)$$

for $(i = 1, 2, \dots, L)$ and where L is the element number of the output of the output layer and W_{ij} is the connection weight between the j 'th element in the hidden layer and the i 'th element in the output layer.

[077] Additionally or alternatively, while Figure 1C has been described with respect to a single stage of combiners (e.g., second stage) including the combiner units 173a-c and delay units 175a-c, it can be appreciated that multiple stages of similar combiner stages may be included in the neural network 170 with varying types of combiner units and varying types of delay units with varying delays, for example, as will now be described with reference to Figures 2A-2C. Further, the same neural network (e.g., the neural network 170 of Figure 1C) can be used to decode an input code-word encoded from any of multiple error correction encoder by selecting different weights that were obtained by the training for the particular error correction technique employed. In this manner, recurrent neural networks may serve as a universal decoder for multiple encoder types, like the neural network 150. Advantageously, the recurrent neural network 170 utilizes delayed versions of output data from combiner units 173a-c, to provide feedback information representative of a temporal nature. As such, the RNN 170 may provide output data decoded faster with the RNN 170, e.g., as compared to the neural network 150. For example, if a particular encoding technique is susceptible to noise during storing in a memory 140 or memory 145 (e.g., bit flips), obtained encoded data to be decoded by the RNN 170 may include noise that produces a time-varying effect on the obtained encoded data (e.g., temporal nature). Accordingly, the feedback information included in the delayed versions of output data may improve the efficiency of the RNN 170 (e.g., as compared to a neural network 150), to decode the data into decode data, e.g., an estimate of the encoded data as having been decoded in accordance with a decoding technique complementary to the encoding technique.

[078] Figure 2A is a schematic illustration of a processing unit 112 arranged in a system 200 in accordance with examples described herein. Such a hardware implementation (e.g., system 200) may be used, for example, to implement one or more neural networks, such as the recurrent neural network 170 of Figure 1C. Additionally or alternatively, in some implementations, the processing unit 112 may receive input data 210a, 210b, and 210c from such a computing system. The input data 210a, 210b, and 210c may be encoded data obtained from a sensor or data stored in the memory 145. In some examples, encoded data stored in the memory 140 or memory 145 may be obtained as input data to be provided to the electronic device 110 in which the processing unit 112 is implemented. Accordingly, the input data obtained may be m bits of noisy encoded input data that is provided to an input stage (e.g., an input layer) of the processing unit, with the input data 210a $X_1(i)$ corresponding to the first bit; the input data 210b $X_2(i)$ corresponding to the second bit; and the input data 210c $X_m(i)$ corresponding to the m 'th bit. In some examples, the obtained input data may be data encoded in accordance with an encoding technique (e.g., low density parity check coding (LDPC), Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH), and/or Polar coding). Additionally or alternatively, in some implementations, the input data may be noisy due to noise experienced during storing of the encoded data in the memory 140 or memory 145 (e.g., bit flips). Thus, one or more processing units 112 may implement a decoding functionality or decoding operation for the received noisy encoded input data at an electronic device 110.

[079] As denoted in the representation of the input data signals, the input data 210a $X_1(i, i-1)$ includes a current portion of the input data, at time i , and a previous portion of the input data, at time $i-1$. For example, a current portion of the input data may be a sample obtained at a certain time period (e.g., at time i), while a previous portion of the input data may be a sample obtained at a time period previous to the certain time period (e.g., at time $i-1$). Accordingly, the previous portion of the input data may be referred to as a time-delayed version of the current portion of the input data. The portions of the input data at each time period may be obtained in a vector or matrix format, for example. In an example, a current portion of the input data, at time i , may be a single value; and a previous portion of the input data, at time $i-1$, may be a single value. Thus, the input data 210a $X_1(i, i-1)$ may be a vector. In some examples, the current portion of the input data, at time i , may be a vector value; and a previous portion of the input data, at time $i-1$, may be a vector value. Thus, the input data 210a $X_1(i, i-1)$ may be a matrix.

[080] The processing unit 112 may include multiplication unit/accumulation (MAC) units 212a-c, 216a-b, and 220; delay units 213a-c, 217a-b, and 221; and memory lookup units

(MLUs) 214a-c, 218a-b, and 222 that, when mixed with input data obtained from the memory 145, may generate output data (e.g. B (1)) 230. Each set of MAC units and MLU units having different element numbers may be referred to as a respective stage of combiners for the processing unit 112. For example, a first stage of combiners includes MAC units 212a-c and MLUs 214a-c, operating in conjunction with delay units 213a-c, to form a first stage or “layer,” as referenced with respect to Figure 1B having “hidden” layers as various combiner stages. Continuing in the example, the second stage of combiners includes MAC units 216a-b and MLUs 218a-b, operating in conjunction with delay units 217a-b, to form a second stage or second layer of hidden layers. And the third stage of combiners may be a single combiner including the MAC unit 220 and MLU 222, operating in conjunction with delay unit 221, to form a third stage or third layer of hidden layers.

[081] In an example of input data being encoded in accordance with an encoding technique, the output data 230 B(1) may be the decoded data (e.g., an estimate of the decoded data) corresponding to the encoded input data in some examples. For example, the output data may be the data corresponding to the encoded input data, but having reduced and/or modified noise. In operation, the processing unit 112, may provide instructions 115, stored at the mode configurable control 105, to cause the processing unit 112 to configure the multiplication units 212a-c, 216a-c, and 220 to multiply and/or accumulate input data 210a, 210b, and 210c and delayed versions of processing results from the delay units 213a-c, 217a-b, and 221 (e.g., respective outputs of the respective layers of MAC units) with coefficient data to generate the output data 230 B(1). For example, the mode configurable control 105 may execute instructions that cause the memory 145 to provide weights and/or other parameters stored in the memory 145, which may be associated with a certain encoding technique, to the MLUs 214a-c, 218a-b, and 222 as weights for the MAC units 212a-c, 216a-b, and 220 and delay units 213a-c, 217a-b, and 221. During operation, the mode configuration control 105 may be used to select weights and/or other parameters in memory 145 based on an indicated encoding technique for the processing unit 112.

[082] As denoted in the representation of the respective outputs of the respective layers of MAC units (e.g, the outputs of the MLUs 214a-c, 218a-b, and 222), the input data to each MAC unit 212a-c, 216a-b, and 220 includes a current portion of input data, at time i , and a delayed version of a processing result, at time $i-1$. For example, a current portion of the input data may be a sample obtained at a certain time period (e.g., at time i), while a delayed version of a processing result (e.g., at time $i-1$) may be obtained from the output of the delay units

213a-c, 217a-b, and 221, which is representative of a time period previous to the certain time period (e.g., as a result of the introduced delay). Accordingly, in using such input data, obtained from both a current period and at least one previous period, output data 230 B(1) may be representative of a Markov process, such that a causal relationship between at least data from a current time period and a previous time period may improve the accuracy of weight estimation for training of coefficient data to be utilized by the MAC units and MLUs of the processing unit 112 or inference of encoded input data or noisy encoded input data in utilizing the processing unit 112. Accordingly, in utilizing delayed versions of output data from 213a-c, 217a-b, and 221 the recurrent neural network 170 provides decoded data (e.g., an estimate of the decoded data), which may have reduced and/or modified noise. Therefore, a recurrent neural network 170 may operate more efficiently to decode noisy encoded input data, e.g., data that may have been stored in a memory cells of a degraded memory device where bit flips have occurred.

[083] In an example of executing such instructions 115 for mixing encoded input data with coefficients, at a first layer of the MAC units 212a-c and MLUs 214a-c, the multiplication unit/accumulation units 212a-c are configured to multiply and accumulate at least two operands from corresponding input data 210a, 210b, or 210c and an operand from a respective delay unit 213a-c to generate a multiplication processing result that is provided to the MLUs 214a-c. For example, the multiplication unit/accumulation units 212a-c may perform a multiply-accumulate operation such that three operands, M, N, and T are multiplied and then added with P to generate a new version of P that is stored in its respective MLU 214a-c. Accordingly, the MLU 214a latches the multiplication processing result, until such time that the stored multiplication processing result is be provided to a next layer of MAC units. The MLUs 214a-c, 218a-b, and 222 may be implemented by any number of processing elements that operate as a memory look-up unit such as a D, T, SR, and/or JK latches.

[084] The MLUs 214a-c, 218a-b, and 222 shown in Figure 2A may generally perform a predetermined nonlinear mapping from input to output. For example, the MLUs 214a-c, 218a-b, and 222 may be used to evaluate at least one non-linear function. In some examples, the contents and size of the various MLUs 214a-c, 218a-b, and 222 depicted may be different and may be predetermined. In some examples, one or more of the MLUs 214a-c, 218a-b, and 222 shown in Figure 2A may be replaced by a single consolidated MLU (e.g., a table look-up). Examples of nonlinear mappings (e.g., functions) which may be performed by the MLUs 214a-c, 218a-b, and 222 include Gaussian functions, piece-wise linear functions, sigmoid functions,

thin-plate-spline functions, multiquadratic functions, cubic approximations, and inverse multiquadratic functions. Examples of functions have been described with reference also to Figure 1B. In some examples, selected MLUs 214a-c, 218a-b, and 222 may be by-passed and/or may be de-activated, which may allow an MLU and its associated MAC unit to be considered a unity gain element.

[085] Additionally in the example, the MLU 214a provides the processing result to the delay unit 213a. The delay unit 213a delays the processing result (e.g., $h_1(i)$) to generate a delayed version of the processing result (e.g., $h_1(i-1)$) to output to the MAC unit 212a as operand T. While the delay units 213a-c, 217a-b, and 221 of Figure 2A are depicted introducing a delay of '1', it can be appreciated that varying amounts of delay may be introduced to the outputs of first layer of MAC units. For example, a clock signal that introduced a sample delay of '1' (e.g., $h_1(i-1)$) may instead introduce a sample delay of '2', '4', or '100'. In various implementations, the delay units 213a-c, 217a-b, and 221 may correspond to any number of processing units that can introduce a delay into processing circuitry using a clock signal or other time-oriented signal, such as flops (e.g., D-flops) and/or one or more various logic gates (e.g., AND, OR, NOR, etc...) that may operate as a delay unit.

[086] In the example of a first hidden layer of a recurrent neural network, the MLUs 214a-c may retrieve coefficient data stored in the memory 145, which may be weights associated with weights to be applied to the first layer of MAC units to both the data from the current period and data from a previous period (e.g., the delayed versions of first layer processing results). For example, the MLU 214a can be a table look-up that retrieves one or more coefficients (e.g., specific coefficients associated with a first frequency) to be applied to both operands M and N, as well as an additional coefficient to be applied to operand T. The MLUs 214a-c also provide the generated multiplication processing results to the next layer of the MAC units 216a-b and MLUs 218a-b. The additional layers of the MAC units 216a, 216b and MAC unit 220 working in conjunction with the MLUs 218a, 218b and MLU 222, respectively, may continue to process the multiplication results to generate the output data 230 B(1). Using such a circuitry arrangement, the output data 230 B(1) may be generated from the input data 210a, 210b, and 210c.

[087] Advantageously, the processing unit 112 of system 200 may utilize a reduced number of MAC units and/or MLUs, e.g., as compared to the processing unit 112 of Figure 2B. Each subsequent layer may have a reduced portion of MAC units, delay units, and MLUs. As depicted, in Figure 2A for example, a second layer of MAC units 216a-b, delay unit 217a-b,

and MLUs 218a-b may include $m-1$ MAC units and MLUs, when $m = 3$. Accordingly, the last layer in the processing unit 112, including the MAC unit 220, delay unit 221, and MLU 222, includes only one MAC, one delay unit, and one MLU.

[088] The coefficient data, for example from memory 145, can be mixed with the input data 210a-210c and delayed version of processing results to generate the output data 230 B(1). For example, the relationship of the coefficient data to the output data 230 B(1) based on the input data 210a-c and the delayed versions of processing results may be expressed as:

$$B(1) = a^1 * f\left(\sum_{j=1}^{m-1} a^{(m-1)} f_j\left(\sum_{k=1}^m a^{(m)} X_k(i)\right)\right) \quad (11)$$

where $a^{(m)}$, $a^{(m-1)}$, a^1 are coefficients for the first layer of multiplication/accumulation units 212a-c and outputs of delay units 213a-c; the second layer of multiplication/accumulation units 216a-b and outputs of delay units 217a-b; and last layer with the multiplication/accumulation unit 220 and output of delay unit 221, respectively; and where $f(\bullet)$ is the mapping relationship which may be performed by the memory look-up units 214a-c and 218a-b. As described above, the memory look-up units 214a-c and 218a-b retrieve coefficients to mix with the input data and respective delayed versions of each layer of MAC units. Accordingly, the output data may be provided by manipulating the input data and delayed versions of the MAC units with the respective multiplication/accumulation units using a set of coefficients stored in the memory. The set of coefficients may be associated with a desired encoding technique. The resulting mapped data may be manipulated by additional multiplication/accumulation units and additional delay units using additional sets of coefficients stored in the memory associated with the desired encoding technique. The sets of coefficients multiplied at each stage of the processing unit 112 may represent or provide an estimation of the processing of the input data in specifically-designed hardware (e.g., an FPGA).

[089] Further, it can be shown that the system 200, as represented by Equation (11), may approximate any nonlinear mapping with arbitrarily small error in some examples and the mapping of system 200 may be determined by the coefficients $a^{(m)}$, $a^{(m-1)}$, a^1 . For example, if such coefficient data is specified, any mapping and processing between the input data 210a-210c and the output data 230 may be accomplished by the system 200. For example, the coefficient data may represent non-linear mappings of the input data 210a-c to the output data B(1) 230. In some examples, the non-linear mappings of the coefficient data may represent a Gaussian function, a piece-wise linear function, a sigmoid function, a thin-plate-spline function, a multi-quadratic function, a cubic approximation, an inverse multi-quadratic

function, or combinations thereof. In some examples, some or all of the memory look-up units 214a-c, 218a-b may be deactivated. For example, one or more of the memory look-up units 214a-c, 218a-b may operate as a gain unit with the unity gain. Such a relationship, as derived from the circuitry arrangement depicted in system 200, may be used to train an entity of the computing system 200 to generate coefficient data. For example, using Equation (11), an entity of the computing system 200 may compare input data to the output data to generate the coefficient data.

[090] Each of the multiplication unit/accumulation units 212a-c, 216a-b, and 220 may include multiple multipliers, multiple accumulation unit, or and/or multiple adders. Any one of the multiplication unit/accumulation units 212a-c, 216a-b, and 220 may be implemented using an ALU. In some examples, any one of the multiplication unit/accumulation units 212a-c, 216a-b, and 220 can include one multiplier and one adder that each perform, respectively, multiple multiplications and multiple additions. The input-output relationship of a multiplication/accumulation unit 212a-c, 216a-b, and 220 may be represented as:

$$B_{out} = \sum_{i=1}^I C_i * B_{in}(i) \quad (12)$$

where “ I ” represents a number to perform the multiplications in that unit, C_i the coefficients which may be accessed from a memory, such as memory 145, and $B_{in}(i)$ represents a factor from either the input data 210a-c or an output from multiplication unit/accumulation units 212a-c, 216a-b, and 220. In an example, the output of a set of multiplication unit/accumulation units, B_{out} , equals the sum of coefficient data, C_i multiplied by the output of another set of multiplication unit/accumulation units, $B_{in}(i)$. $B_{in}(i)$ may also be the input data such that the output of a set of multiplication unit/accumulation units, B_{out} , equals the sum of coefficient data, C_i multiplied by input data.

[091] Accordingly, the system 200 (e.g., a hardware implementation of recurrent neural network 170) may be used to convert an input code word (e.g., $x_1(n), x_2(n), \dots, x_m(n)$) to an output code word (e.g., $B(1)$). Examples of the conversion have been described herein with reference to Figure 1C. For example, the input code word may correspond to noisy encoded input data. A hardware implementation of recurrent neural network 170 may utilize multiplication with corresponding weights (e.g., weights obtained during training) and MLUs

to provide the output code word. The output code word may correspond to the decoded data and/or to a version of the encoded input data having reduced and/or changed noise.

[092] The mode configuration control 105 may be implemented using circuitry (e.g., logic), one or more processor(s), microcontroller(s), controller(s), or other elements. The mode configuration control 105 may select certain weights and/or other parameters from memory 145 and provide those weights and/or other parameters to one or more of the MAC units and/or MLUs of Figure 2A. In some examples, weights and/or other parameters stored in memory 145 may be associated with particular encoding techniques. During operation, the mode configuration control 105 may be used to select weights and/or other parameters in memory 145 associated with a particular encoding technique (e.g., Reed-Solomon coding, BCH coding, LDPC coding, and/or Polar coding). A hardware implementation of recurrent neural network 170 may then utilize the selected weights and/or other parameters to function as a decoder for data encoded with that encoding technique. The mode configuration control 105 may select different weights and/or other parameters stored in memory 145 which are associated with a different encoding technique to alter the operation of the hardware implementation of recurrent neural network 170 to serve as a decoder for the different encoding technique. In this manner, the hardware implementation of recurrent neural network 170 may flexibly function as a decoder for multiple encoding techniques.

[093] **Figure 2B** is a schematic illustration of a processing unit 112 arranged in a system 250 in accordance with examples described herein. Such a hardware implementation (e.g., system 250) may be used, for example, to implement one or more neural networks, such as the recurrent neural network 170 of Figure 1C. Additionally or alternatively, in some implementations, the processing unit 112 may receive input data 210a, 210b, and 210c from a computing system. Similarly described elements of Figure 2B may operate as described with respect to Figure 2A, but may also include additional features as described with respect to Figure 2B. For example, Figure 2B depicts MAC units 262a-c and delay units 263a-c that may operate as described with respect to MAC units 212a-c and delay units 213a-c of Figure 2A. Accordingly, elements of Figure 2B, whose numerical indicator is offset by 50 with respect to Figure 2A, include similarly elements of the processing element 112; e.g., MAC unit 266a operates similarly with respect to MAC unit 216a. The system 250, including processing element 112, also includes additional features not highlighted in the processing element 112 of Figure 2A. For example, the processing unit 112 of Figure 2B additionally includes MAC units 266c and 270b-c; delay units 267c and 271b-c; and MLUs 268c and 272b-c, such that the output

data is provided as 275a-c, rather than as singularly in Figure 2A as B(1) 230. Advantageously, the system 250 including a processing element 112 may process the input data 260a-c to generate the output data 275a-c with greater precision. For example, the output data 275a-c may process the input data 260a-260c with additional coefficient retrieved at MLU 268c and multiplied and/or accumulated by additional MAC units 266c and 270b-c and additional delay units 267c and 271b-c. For example, such additional processing may result in output data that is more precise with respect providing output data that is an estimate of decoded data (e.g., as compared to output data obtained from the processing unit 112 of Figure 2A). In implementations where board space (e.g., a printed circuit board) is not a primary factor in design, implementations of the processing unit 112 of Figure 2B may be desirable as compared to that of processing unit 112 of Figure 2A; which, in some implementations may occupy less board space as a result of having fewer elements than the processing unit 112 of Figure 2B.

[094] While processing element 112 is described in the context of Figures 2A and 2B as a single processing element 112, the features may also be implemented in the processing elements 112 of Figure 1A, such that the description of the single processing element 112 in Figure 2A or 2B is interchangeable as a processing element as implemented throughout the examples described herein. Therefore, it can be appreciated that, while referring to a single processing element 112 in Figures 2A or 2B, the electronic device 110 may have multiple processing elements 112, like Figure 1A, to also have the same features as described with respect a single processing element 112 of Figures 2A or 2B.

[095] Figure 2C is a schematic illustration of a processing unit 112 arranged in a system 280 in accordance with examples described herein. Such a hardware implementation (e.g., system 280) may be used, for example, to implement one or more neural networks, such as the recurrent neural network 170 of Figure 1C. Additionally or alternatively, in some implementations, the processor unit 112 may be implemented as a processing unit 112 in the example of Figure 1A. Similarly described elements of Figure 2C may operate as described with respect to Figure 2B, except for the delay units 263a-c, 267a-c, and 271a-c of Figure 2B. For example, Figure 2C depicts MAC units 282a-c and delay units 283a-c that may operate as described with respect to MAC units 262a-c and delay units 263a-c of Figure 2B. Accordingly, elements of Figure 2C, whose numerical indicator is offset by 20 with respect to Figure 2B, include similarly elements of the processing element 112; e.g., MAC unit 286a operates similarly with respect to MAC unit 266a.

[096] The system 280, including processing element 112, also includes additional features not highlighted in the processing element 112 of Figure 2B. Different than Figure 2B, Figure 2C depicts delay units 283a, 283b, and 283c. Accordingly, the processing unit of Figure 2C illustrate that processing units 112 may include varying arrangements to the placement of the inputs and outputs of delay units, as illustrated with delay units 283a, 283b, and 283c. For example, the output of MLUs 288b may be provided to delay unit 283b, to generate a delayed version of that processing result from the second layer of MAC units, as an input to the first layer of MAC units, e.g., as an input to MAC unit 282b. Accordingly, the processing unit 112 of system 280 is illustrative that delayed versions of processing results may be provided as inputs to other hidden layers, different than the processing unit 112 of system 250 in Figure 2B showing respective delayed versions being provided as inputs to the same layer in which those delayed versions were generated (e.g., the output of MLU 268b is provided to delay unit 267b, to generate a delayed version for the MAC unit 266b in the same layer from which the processing result was outputted). Therefore, in the example, even the output B(n) 295c may be provided, from the last hidden layer, to the first hidden layer (e.g., as an input to MAC unit 282c).

[097] Advantageously, such delayed versions of processing results, which may be provided as inputs to different or additional hidden layers, may better compensate “higher-order” memory effects in a recurrent neural network 170 that implements one or more processing units 112 of Figure 2C, e.g., as compared to the processing unit(s) 112 of Figures 2A or 2B. For example, higher-order memory effects model the effects of leading and lagging envelope signals used during training of the recurrent neural network 170, to transform obtained noisy encoded input data at the processing unit(s) 112 into decoded data (e.g., an estimate of the decoded data). The obtained encoded input data, in being stored in a memory 140 or memory 145 with noise (e.g., bit flips) or other noises experienced during storage (e.g., shot or voltage noise) may be noisy encoded input data to be decoded by the recurrent neural network 170. In the example, a recurrent neural network 170 that estimates the output data may include varying delayed versions of processing results that corresponds to such leading and lagging envelopes (e.g., of various envelopes encapsulating the noise(s)). Accordingly, implementing the processing unit 112 incorporates such higher-order memory effects, e.g., for an inference of a recurrent neural network 170, to provide output data 295a-c based on input data 281a-c.

[098] While processing element 112 is described in the context of Figures 2A, 2B, and 2C as a single processing element 112, the features may also be implemented in the processing

elements 112 of Figure 1A, such that the description of the single processing element 112 in Figure 2A, 2B, 2C is interchangeable as a processing element as implemented throughout the examples described herein. Therefore, it can be appreciated that, while referring to a single processing element 112 in Figures 2A, 2B, 2C the electronic device 110 may have multiple processing elements 112, like Figure 1A, to also have the same features as described with respect a single processing element 112 of Figures 2A, 2B, or 2C.

[0099] Figure 3 is a schematic illustration of apparatus 300 (e.g., an integrated circuit, a memory device, a memory system, an electronic device or system, a smart phone, a tablet, a computer, a server, an appliance, a vehicle, etc.) according to an embodiment of the disclosure. The apparatus 300 may generally include a host 302 and a memory system 304.

[0100] The host 302 may be a host system such as a personal laptop computer, a desktop computer, a digital camera, a mobile telephone, or a memory card reader, among various other types of hosts. The host 302 may include a number of memory access devices (e.g., a number of processors). The host 302 may also be a memory controller, such as where memory system 304 is a memory device (e.g., a memory device having an on-die controller).

[0101] The memory system 304 may be a solid state drive (SSD) or other type of memory and may include a host interface 306, a controller 308 (e.g., a processor and/or other control circuitry), and a number of memory device(s) 314. The memory system 304, the controller 308, and/or the memory device(s) 314 may also be separately considered an "apparatus." The memory device(s) 314 may include a number of solid state memory devices such as NAND flash devices, which may provide a storage volume for the memory system 304. Other types of memory may also be used.

[0102] The controller 308 may be coupled to the host interface 306 and to the memory device(s) 314 via a plurality of channels to transfer data between the memory system 304 and the host 302. The interface 306 may be in the form of a standardized interface. For example, when the memory system 304 is used for data storage in the apparatus 300, the interface 306 may be a serial advanced technology attachment (SATA), peripheral component interconnect express (PCIe), or a universal serial bus (USB), among other connectors and interfaces. In general, interface 306 provides an interface for passing control, address, data, and other signals between the memory system 304 and the host 302 having compatible receptors for the interface 306.

[0103] The controller 308 may communicate with the memory device(s) 314 (which in some embodiments can include a number of memory arrays on a single die) to control data read,

write, and erase operations, among other operations. The controller 308 may include a discrete memory channel controller for each channel (not shown in Figure 3) coupling the controller 308 to the memory device(s) 314. The controller 308 may include a number of components in the form of hardware and/or firmware (e.g., one or more integrated circuits) and/or software for controlling access to the memory device(s) 314 and/or for facilitating data transfer between the host 302 and memory device(s) 314.

[0104] The controller 308 may include an ECC encoder 310 for encoding data bits written to the memory device(s) 314 using one or more encoding techniques. The ECC encoder 310 may include a single parity check (SPC) encoder, and/or an algebraic error correction circuit such as one of the group including a Bose-Chaudhuri-Hocquenghem (BCH) ECC encoder and/or a Reed Solomon ECC encoder, among other types of error correction circuits. The controller 308 may further include an ECC decoder 312 for decoding encoded data, which may include identifying erroneous cells, converting erroneous cells to erasures, and/or correcting the erasures. The memory device(s) 314 may, for example, include one or more output buffers which may read selected data from memory cells of the memory device(s) 314. The output buffers may provide output data, which may be provided as encoded input data to the ECC decoder 312. The neural network 150 of Figure 1B and/or neural network 170 of Figure 1C (e.g., as implemented by one or more processing units of Figures 2A-2C) may be used to implement the ECC decoder 312 of Figure 3, for example. In various embodiments, the ECC decoder 312 may be capable of decoding data for each type of encoder in the ECC encoder 310. For example, the memory 145 of Figures 2A, 2B, or 2C may store parameters associated with multiple encoding techniques which may be used by the ECC encoder 310, such that the hardware implementation of neural network 150 or recurrent neural network 170 may be used as a 'universal decoder' to decode input data encoded by the ECC encoder 310, using any of multiple encoding techniques available to the ECC encoder.

[0105] The ECC encoder 310 and the ECC decoder 312 may each be implemented using discrete components such as an application specific integrated circuit (ASIC) or other circuitry, or the components may reflect functionality provided by circuitry within the controller 308 that does not necessarily have a discrete physical form separate from other portions of the controller 308. Although illustrated as components within the controller 308 in FIG. 3, each of the ECC encoder 310 and ECC decoder 312 may be external to the controller 308 or have a number of components located within the controller 308 and a number of components located external to the controller 308.

[0106] The memory device(s) 314 may include a number of arrays of memory cells (e.g., non-volatile memory cells). The arrays can be flash arrays with a NAND architecture, for example. However, embodiments are not limited to a particular type of memory array or array architecture. Floating-gate type flash memory cells in a NAND architecture may be used, but embodiments are not so limited. The cells may be multi-level cells (MLC) such as triple level cells (TLC) which store three data bits per cell. The memory cells can be grouped, for instance, into a number of blocks including a number of physical pages. A number of blocks can be included in a plane of memory cells and an array can include a number of planes. As one example, a memory device may be configured to store 8 KB (kilobytes) of user data per page, 128 pages of user data per block, 2048 blocks per plane, and 16 planes per device.

[0107] According to a number of embodiments, controller 308 may control encoding of a number of received data bits according to the ECC encoder 310 that allows for later identification of erroneous bits and the conversion of those erroneous bits to erasures. The controller 308 may also control programming the encoded number of received data bits to a group of memory cells in memory device(s) 314.

[0108] The apparatus shown in Figure 3 may be implemented in any of a variety of products employing processors and memory including for example cameras, phones, wireless devices, displays, chip sets, set top boxes, gaming systems, vehicles, and appliances. Resulting devices employing the memory system may benefit from examples of neural networks described herein to perform their ultimate user function.

[0109] From the foregoing it will be appreciated that, although specific embodiments have been described herein for purposes of illustration, various modifications may be made while remaining within the scope of the claimed technology. Certain details are set forth herein to provide an understanding of described embodiments of technology. However, other examples may be practiced without various of these particular details. In some instances, well-known circuits, control signals, timing protocols, neural network structures, algorithms, and/or software operations have not been shown in detail in order to avoid unnecessarily obscuring the described embodiments. Other embodiments may be utilized, and other changes may be made, without departing from the spirit or scope of the subject matter presented here.

[0110] Figure 4 is a flowchart of a method 400 arranged in accordance with examples described herein. The example method may include block 402, which may be followed by block 404, which may be followed by block 406, which may be followed by block 408, which

may be followed by block 410. Additional, fewer, and/or different blocks may be used in other examples, and the order of the blocks may be different in other examples.

[0111] Block 402 recites “receive known encoded and decoded data pairs, the encoded data encoded with a particular encoding technique.” The known encoded and decoded data pairs may be received by a computing device (e.g., electronic device 110) that includes a neural network, such as the neural network 150 of Figure 1B, neural network 170 of Figure 1C (e.g., as implemented by one or more processing units of Figures 2A-2C), and/or the ECC decoder 312 of Figure 3. Signaling indicative of the set of data pairs may be provided to the computing device.

[0112] Block 404 may follow block 402. Block 404 recites “determine a set of weights for a neural network to decode data encoded with the particular encoding technique.” For example, a neural network (e.g., any of the neural networks described herein) may be trained using the encoded and decoded data pairs received in block 402. The weights may be numerical values, which, when used by the neural network, allow the neural network to output decoded data corresponding encoded input data encoded with a particular encoding technique. The weights may be stored, for example, in the memory 145 of Figures 2A, 2B, or 2C. In some examples, training may not be performed, and an initial set of weights may simply be provided to a neural network, e.g., based on training of another neural network.

[0113] In some examples, multiple sets of data pairs may be received (e.g., in block 402), with each set corresponding to data encoded with a different encoding technique. Accordingly, multiple sets of weights may be determined (e.g., in block 404), each set corresponding to a different encoding technique. For example, one set of weights may be determined which may be used to decode data encoded in accordance with LDPC coding while another set of weights may be determined which may be used to decode data encoded with BCH coding.

[0114] Block 406 may follow block 404. Block 406 recites “receive data encoded with the particular encoding technique.” For example, data (e.g., signaling indicative of data) encoded with the particular encoding technique may be retrieved from a memory of a computing device and/or received using a wireless communications receiver. Any of a variety of encoding techniques may have been used to encode the data.

[0115] Block 408 may follow block 406. Block 408 recites “decode the data using the set of weights.” By processing the encoded data received in block 406 using the weights, which may have been determined in block 404, the decoded data may be determined. For example, any neural network described herein may be used to decode the encoded data (e.g., the neural

network 150 of Figure 1B, neural network 170 of Figure 1C (e.g., as implemented by one or more processing units of Figures 2A-2C), and/or the ECC decoder 312 of Figure 3). In some examples, a set of weights may be selected that is associated with the particular encoding technique used to encode the data received in block 406. The set of weights may be selected from among multiple available sets of weights, each for use in decoding data encoded with a different encoding technique.

[0116] Block 410 may follow block 408. Block 410 recites “writing the decoded data to or reading the decoded data from memory.” For example, data decoded in block 408 may be written to a memory, such as the memory device 314 of Figure 3. In some examples, instead of or in addition to writing the data to memory, the decoded data may be transmitted to another device (e.g., using wireless communication techniques). While block 410 recites memory, in some examples any storage device may be used.

[0117] In some examples, blocks 406-410 may be repeated for data encoded with different encoding techniques. For example, data may be received in block 406, encoded with one particular encoding technique (e.g., LDPC coding). A set of weights may be selected that is for use with LDPC coding and provided to a neural network for decoding in block 408. The decoded data may be obtained in block 410. Data may then be received in block 406, encoded with a different encoding technique (e.g., BCH coding). Another set of weights may be selected that is for use with BCH coding and provided to a neural network for decoding in block 408. The decoded data may be obtained in block 410. In this manner, one neural network may be used to decode data that had been encoded with multiple encoding techniques.

[0118] Figure 5 is a flowchart of a method 500 in accordance with examples described herein. Example method 500 may be implemented using, for example, system 100 in Figure 1A, the neural network 150 of Figure 1B, neural network 170 of Figure 1C (e.g., being implemented by one or more processing units of Figures 2A-2C), and/or the ECC decoder 312 of Figure 3, or any system or combination of the systems depicted in the aforementioned Figures. In some examples, the blocks in example method 500 may be performed by a computing device such as an electronic device 110 of Figure 1A and/or in conjunction with a processing unit, such as processing unit 112 of Figure 2A, 2B, or 2C. The operations described in blocks 504-516 may also be stored as computer-executable instructions in a computer-readable medium such as the mode configurable control 105, storing the executable instructions 115.

[0119] Example method 500 may begin with a block 504 that starts execution of the mixing encoded input data with coefficient data routine. The method may include a block 508 that

recites “obtaining encoded data associated with an encoding technique.” As described herein, the processing unit may be configured to obtain a variety of types of input data that may be encoded with various encoding techniques, such as data that has been encoded with a low density parity check coding (LDPC), a Reed-Solomon coding, a Bose–Chaudhuri–Hocquenghem (BCH), and/or Polar coding. In the context of apparatus 300, the encoded data may be obtained from the memory device(s) 314, as described with respect to Figure 3.

[0120] Block 508 may be followed by block 512 that recites “retrieving a plurality of coefficients from a memory, the plurality of coefficients associated with the encoding technique.” As described herein, the processing unit may retrieve coefficients for mixing with encoded input data; for example, utilizing a memory look-up unit (MLU). For example, the memory may store (e.g., in a database) associations between coefficients and encoding techniques described herein. For example, the processing unit may request the coefficients from a memory part of the implementing computing device, from a memory part of an external computing device, or from a memory implemented in a cloud-computing device. In turn, the plurality of coefficients may be retrieved from the memory as requested by the processing unit.

[0121] Block 512 may be followed by block 516 that recites “calculating, at a first layer of multiplication/accumulation processing units (MAC units), the encoded data with the plurality of coefficients to generate first processing results.” As described herein, the processing unit utilizes the plurality of coefficients such that mixing the coefficients with encoded input data generates output data that reflects the processing of the input data with coefficients by the circuitry of Figures 2A, 2B, or 2C. In some implementations of a recurrent neural network (e.g., recurrent neural network 170), delayed versions of respective outputs of the first layer of MAC units and the encoded input data are mixed to generate the output data (e.g., an estimate of decoded data). For example, various ALUs in an integrated circuit may be configured to operate as the circuitry of Figures 2A, 2B, or 2C, thereby mixing the input data and delayed versions of respective outputs of the first layer of MAC units with the coefficients as described herein. For example, with reference to Figure 2A, the input data and delayed versions of respective outputs of the first layer of MAC units may be calculated with the plurality of coefficients to generate first processing results, at a first layer of multiplication/accumulation processing units (MAC units). In some examples, various hardware platforms may implement the circuitry of Figure 2A, 2B, or 2C, such as an ASIC, a DSP implemented as part of a FPGA, or a system-on-chip.

[0122] Block 516 may be followed by block 520 that recites “calculating, at additional layers of MAC units, the first processing results with the additional plurality of coefficients to generate second processing results.” As described herein, the processing unit utilizes additional plurality of coefficients such that mixing the coefficients with certain processing results generates output data that reflects the processing of the input data with coefficients by the circuitry of Figures 2A, 2B, or 2C. In some implementations of a recurrent neural network (e.g., recurrent neural network 170), the first processing results and delayed versions of at least a portion of the first processing results are mixed to generate the output data (e.g., an estimate of decoded data). For example, with reference to Figure 2A, the processing results of the first layer (e.g., multiplication processing results) and delayed versions of at least a portion of those processing results may be calculated with the additional plurality of coefficients to generate second processing results, at a second layer of multiplication/accumulation processing units (MAC units). The processing results of the second layer may be calculated with an additional plurality of coefficients to generate the output data B(1) 230.

[0123] Block 520 may be followed by block 524 that recites “providing output data representative of the encoded data being decoded, such that the output data is an estimate of decoded data based on the encoded data.” As described herein, the neural network 150 or recurrent neural network 170 provides output data as output bits which represent the processed data, corresponding to the encoded input data (e.g., m encoded input bits) having been decoded according to the encoding technique. Accordingly, neural networks described herein transform encoded input data to decoded data (e.g., an estimate of the decoded data). Block 524 may be followed by block 528 that ends the example method 500. In some examples, block 512 may be an optional block.

[0124] The blocks included in the described example methods 400 and 500 are for illustration purposes. In some embodiments, the blocks may be performed in a different order. In some other embodiments, various blocks may be eliminated. In still other embodiments, various blocks may be divided into additional blocks, supplemented with other blocks, or combined together into fewer blocks. Other variations of these specific blocks are contemplated, including changes in the order of the blocks, changes in the content of the blocks being split or combined into other blocks, etc.

[0125] Examples described herein may refer to various components as “coupled” or signals as being “provided to” or “received from” certain components. It is to be understood that in some examples the components are directly coupled one to another, while in other examples the

components are coupled with intervening components disposed between them. Similarly, signal may be provided directly to and/or received directly from the recited components without intervening components, but also may be provided to and/or received from the certain components through intervening components.

CLAIMS

What is claimed is:

1. An apparatus, comprising:
a first stage of circuitry configured to receive encoded data, to combine the encoded data, and to evaluate at least one non-linear function using combinations of the encoded data and delayed versions of the combinations of the encoded data to provide intermediate data; and
at least a second stage of circuitry configured to receive the intermediate data and combine the intermediate data using a set of predetermined weights to provide decoded data the set of predetermined weights based at least in part on an encoding technique associated with the encoded data.
2. The apparatus of claim 1, wherein the set of predetermined weights are based on training of a neural network on known encoded data and delayed versions of combinations of the known encoded data.
3. The apparatus of claim 2, wherein the neural network is configured to use the set of predetermined weights to provide an estimate of decoded data based on the received encoded data.
4. The apparatus of claim 2, further comprising a memory encoded with executable instructions to select the set of predetermined weights for the neural network in accordance with the encoding technique associated with the encoded data.
5. The apparatus of claim 1, wherein the encoding technique comprises Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.
6. The apparatus of claim 1, wherein the first stage of circuitry and second stage of circuitry comprises a first plurality of multiplication/accumulation units, the first plurality of multiplication/accumulation units each configured to multiply at least one bit of the encoded data with at least one of the set of predetermined weights and sum multiple weighted bits of the encoded data.

7. The apparatus of claim 6, wherein the first stage of circuitry further comprises a first plurality of memory look-up units (MLUs), the first plurality of MLUs each configured to retrieve at least one intermediate data value corresponding to an output of a respective one of the first plurality of multiplication/accumulation units based on the at least one non-linear function.

8. The apparatus of claim 1, wherein the first stage of circuitry further comprises a plurality of delay units configured to provide the delayed versions of the combinations of the encoded data based on the at least one intermediate data value provided by the first plurality of MLUs.

9. The apparatus of claim 1, further comprising a memory in communication with the first stage of circuitry and the second stage of MAC circuitry, the memory configured to store multiple sets of predetermined weights, including the set of predetermined weights.

10. A method comprising:

receiving, at a computing device that comprises a neural network, signaling indicative of a set of data pairs each comprising known encoded data and decoded data, wherein the known encoded data is encoded with a particular encoding technique;

determining a set of weights for the neural network to decode data with the particular encoding technique based partly on the signaling, coefficient multiplication results associated with the signaling, and delayed versions of the coefficient multiplication results associated with the signaling;

receiving signaling, from a memory of the computing device, indicative of data encoded with the particular encoding technique;

decoding the data using the neural network using the weights; and

writing the decoded data to or reading the decoded data from a memory or storage medium of the computing device.

11. The method of claim 10, further comprising:

determining multiple sets of weights for the neural network using additional encoded data, decoded data pairs encoded with other encoding techniques, and delayed versions of additional coefficient multiplication results based on the additional encoded

data and the decoded data pairs encoded with the other encoding techniques, wherein the multiple sets of weights each correspond with a different encoding technique;

selecting the set of weights associated with the particular encoding technique; and

providing the set of weights to the neural network or another neural network for use in decoding the further encoded data.

12. The method of claim 11, further comprising:

receiving further input data encoded with another encoding technique;

selecting a selected set of weights of the multiple sets of weights associated with the another encoding technique;

providing the selected set of weights to the neural network; and

decoding the further input data using the selected set of weights.

13. The method of claim 10, wherein the particular encoding technique comprises Reed-Solomon coding, Bose–Chaudhuri–Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.

14. The method of claim 10, wherein decoding the further encoded data using the neural network comprises weighting bits of the further encoded data using at least some of the set of weights, and summing selected weighted bits of the further encoded data to provide intermediate data, then weighting bits of the intermediate data using at least some of the weights and summing selected weighted bits of the intermediate data and selected delayed versions of the weighted bits of the intermediate data to provide decoded data.

15. The method of claim 10, wherein determining the set of weights comprises selecting weights resulting in a minimized value of an error function between an output of the neural network and known decoded data.

16. The method of claim 10, wherein determining the set of weights comprises using a k-means cluster technique to determine center vectors for the neural network.

17. A method comprising:

mixing encoded data received at a processing unit using a plurality of coefficients and an additional plurality of coefficients, both the plurality of coefficients and the additional

plurality of coefficients specific to an encoding technique associated with the encoded data, wherein mixing the encoded data comprises:

calculating, at a first layer of multiplication/accumulation processing units (MAC units) of a plurality of MAC units, the encoded data and delayed versions of respective outputs of the first layer of MAC units with the plurality of coefficients to generate first processing results;

calculating, at additional layers of MAC units of the plurality of MAC units, the first processing results and delayed versions of at least a portion of the first processing results with the additional plurality of coefficients to generate second processing results; and

providing output data based partly on the second processing results, the output data representative of the encoded data being decoded, such that the output data is an estimate of decoded data based on the encoded data.

18. The method of claim 17, wherein mixing the encoded data received at the processing unit using the plurality of coefficients further comprises:

delaying, at respective delay units associated with the first layer of MAC units, the respective outputs of the first layer of MAC units to generate the delayed versions of the respective outputs of the first layer of MAC units.

19. The method of claim 17, wherein calculating, at the first layer of multiplication/accumulation processing units (MAC units) of the plurality of MAC units, the encoded data and the delayed versions of the respective outputs of the first layer of MAC units with the plurality of coefficients comprises:

multiplying the encoded data and the delayed versions of the respective outputs of the first layer of MAC units with respective coefficients of the plurality of coefficients to generate the first processing results.

20. The method of claim 17, further comprising:

obtaining, at the processing unit, signaling indicative of the encoded data from a memory coupled, the signaling indicative of the encoded data including an indication from a touchscreen of a mobile communication device that the encoding technique was utilized.

21. The method of claim 17, wherein the encoding technique comprises Reed-Solomon coding, Bose-Chaudhuri-Hocquenghem (BCH) coding, low-density parity check (LDPC) coding, Polar coding, or combinations thereof.

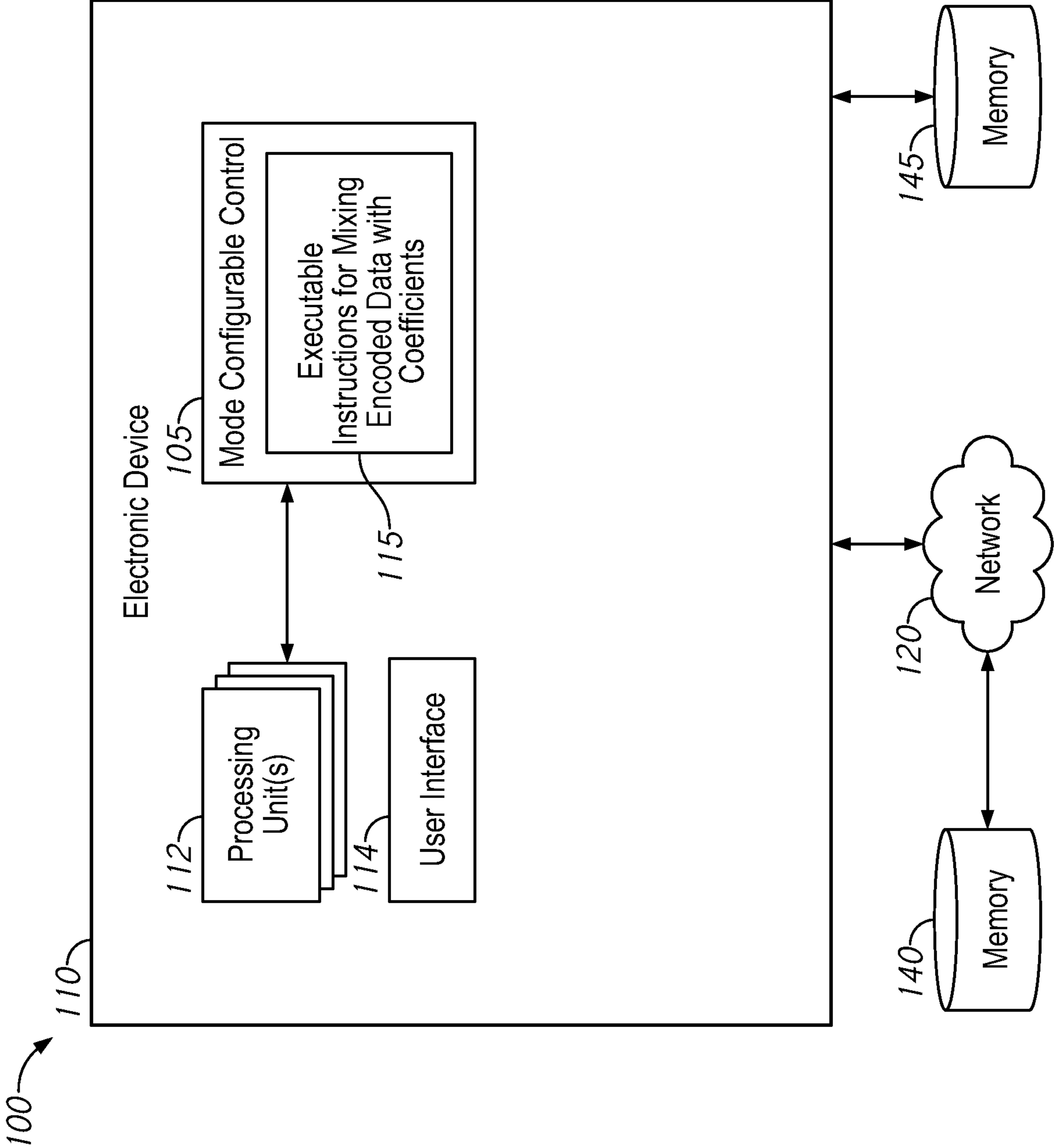


FIG. 1A

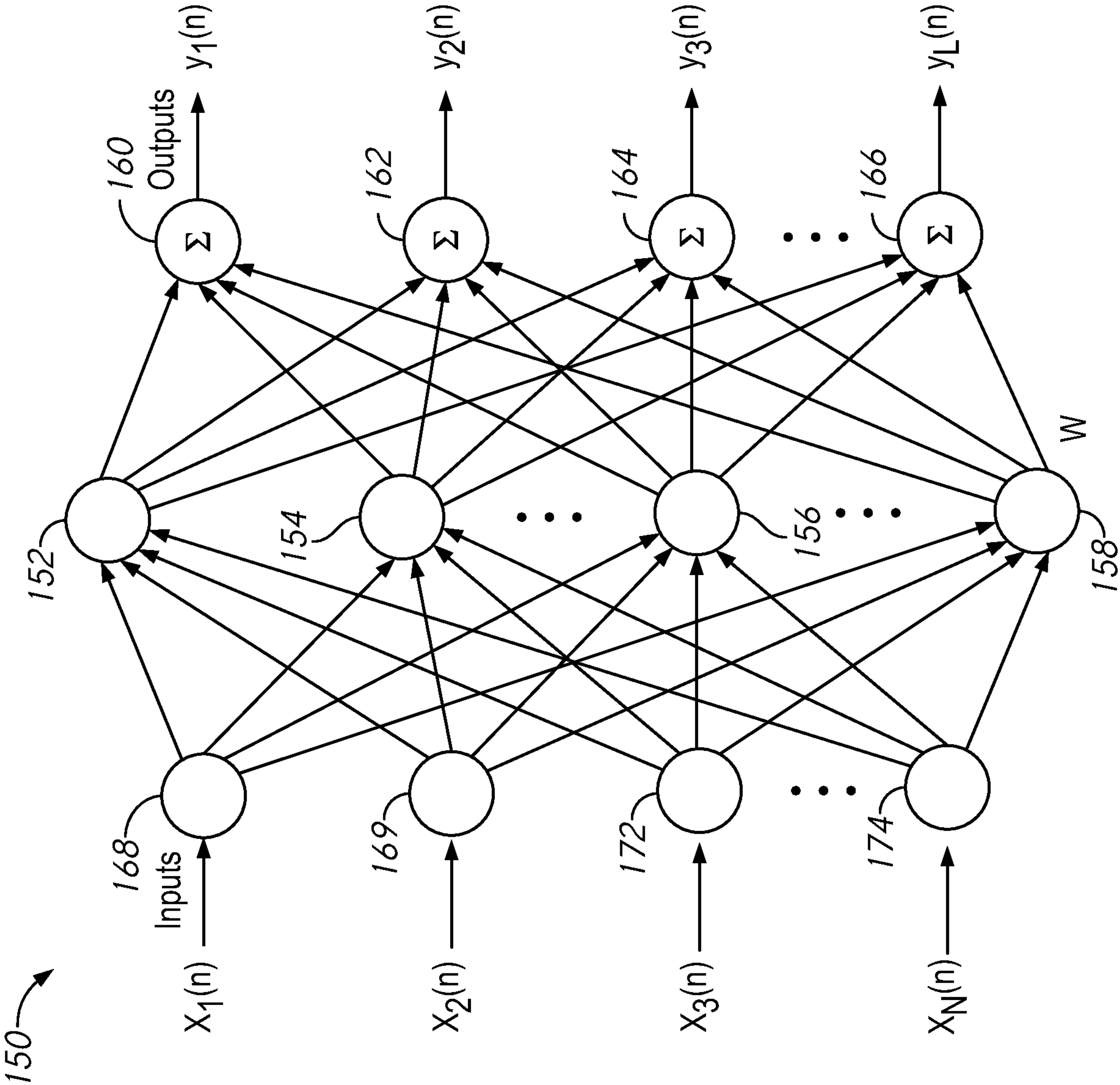


FIG. 1B

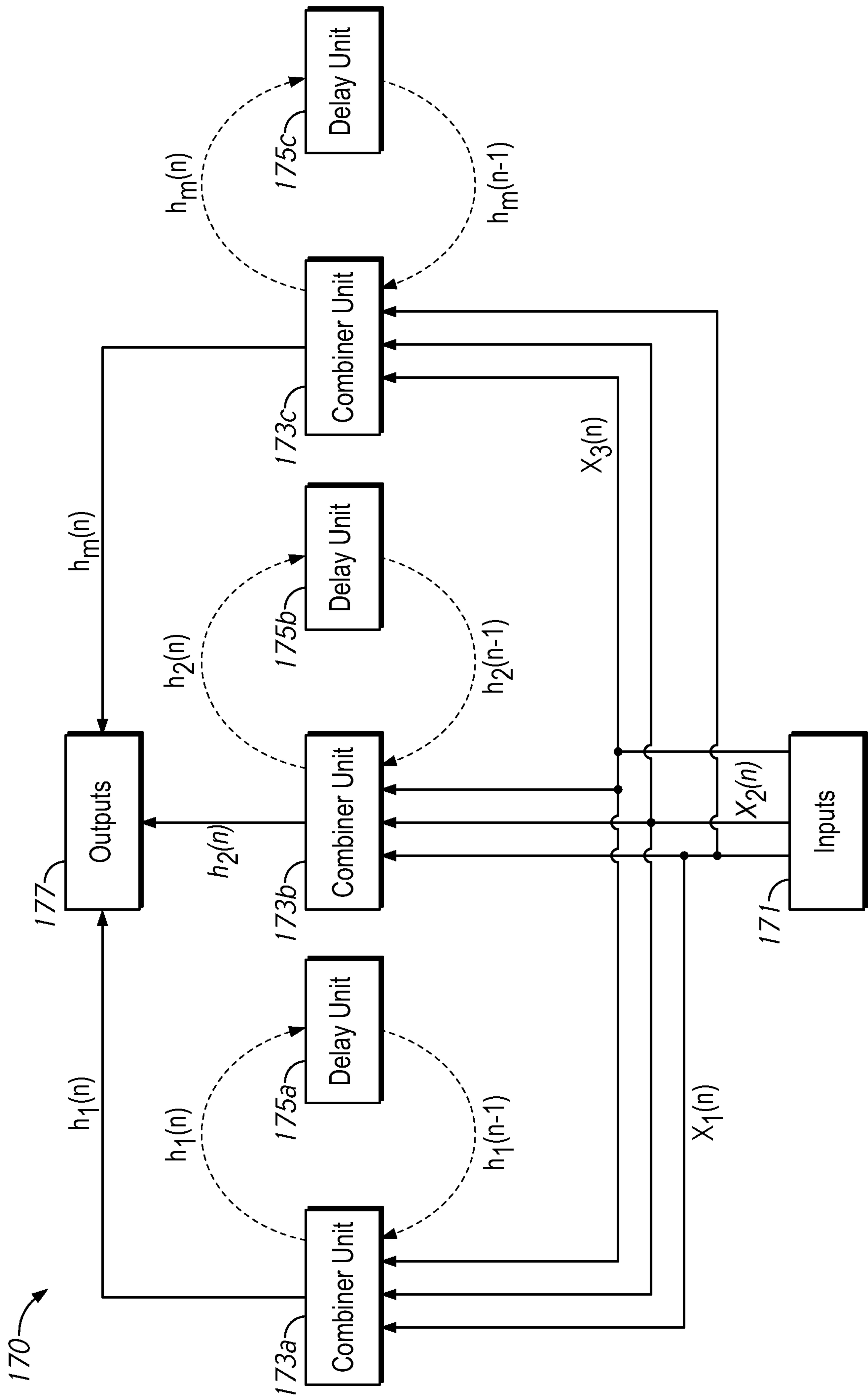


FIG. 1C

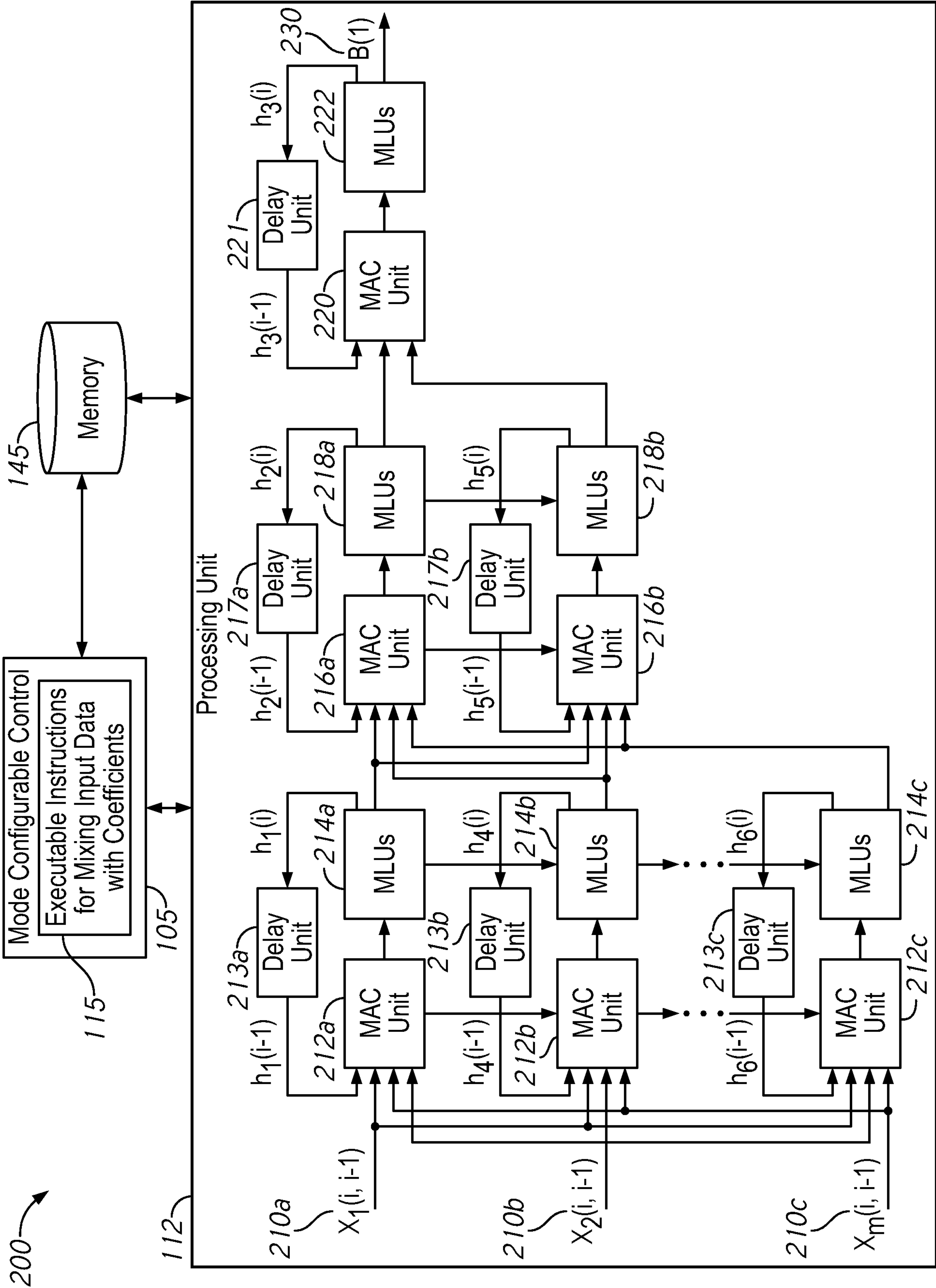


FIG. 2A

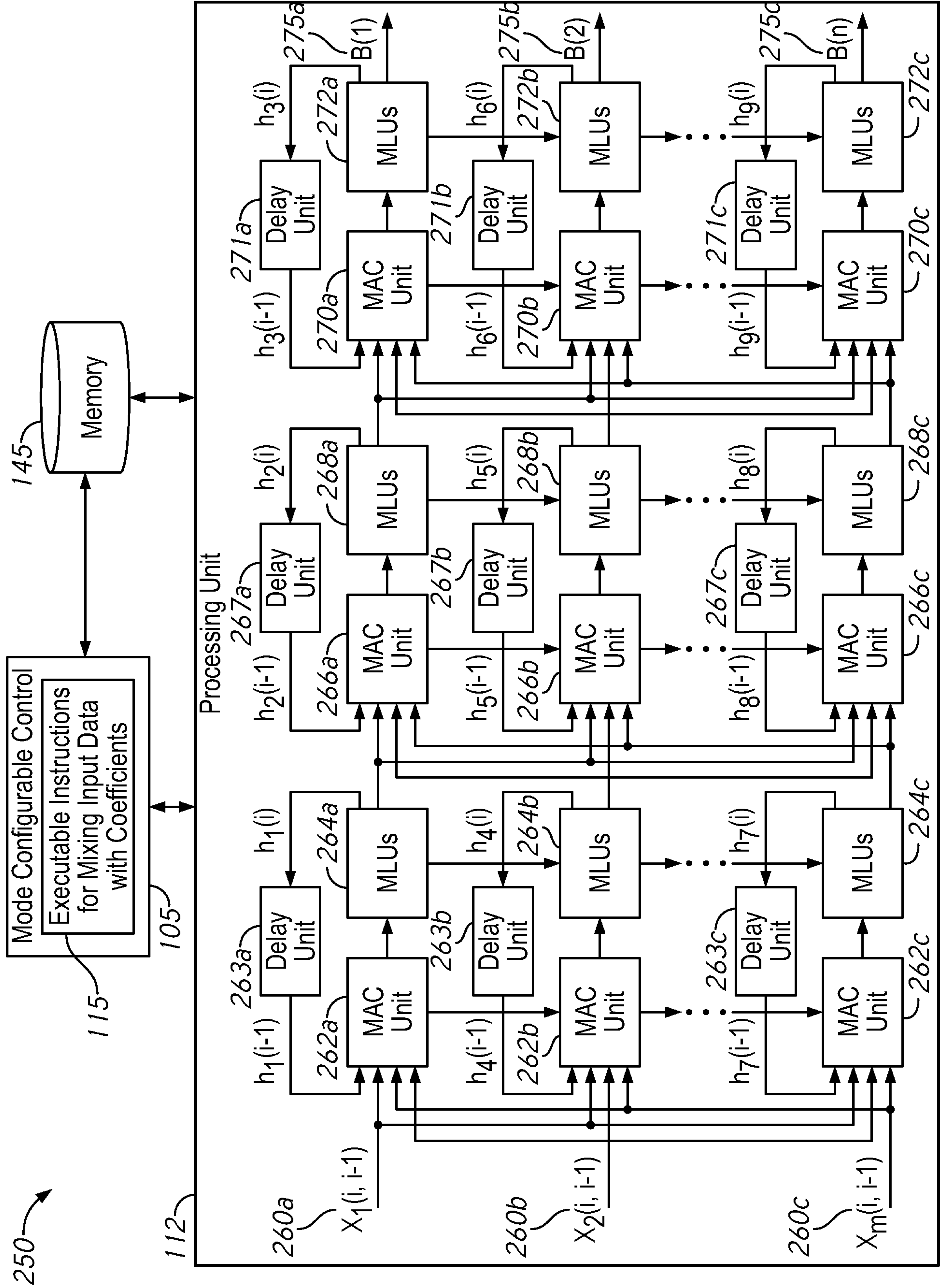


FIG. 2B

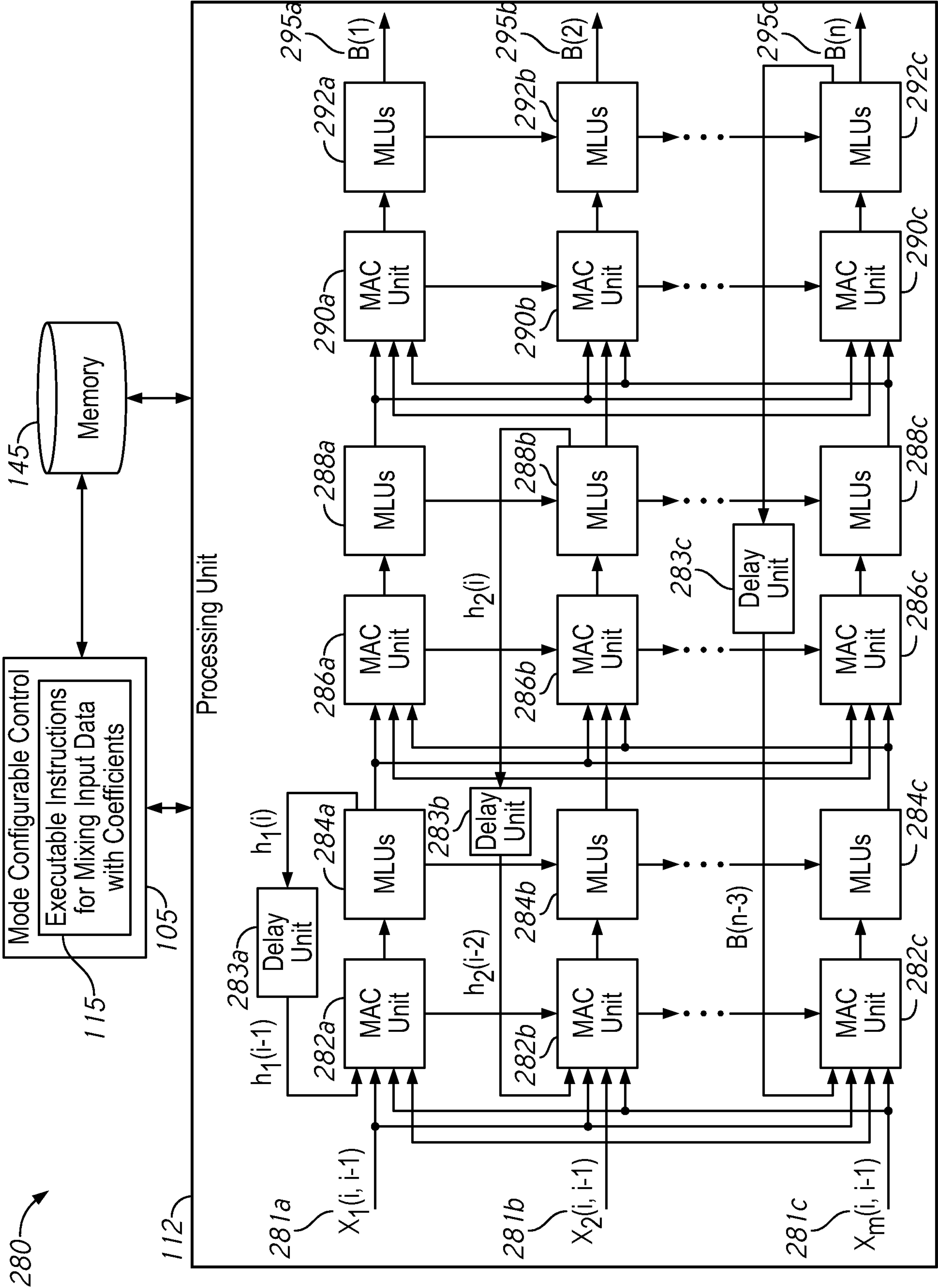


FIG. 2C

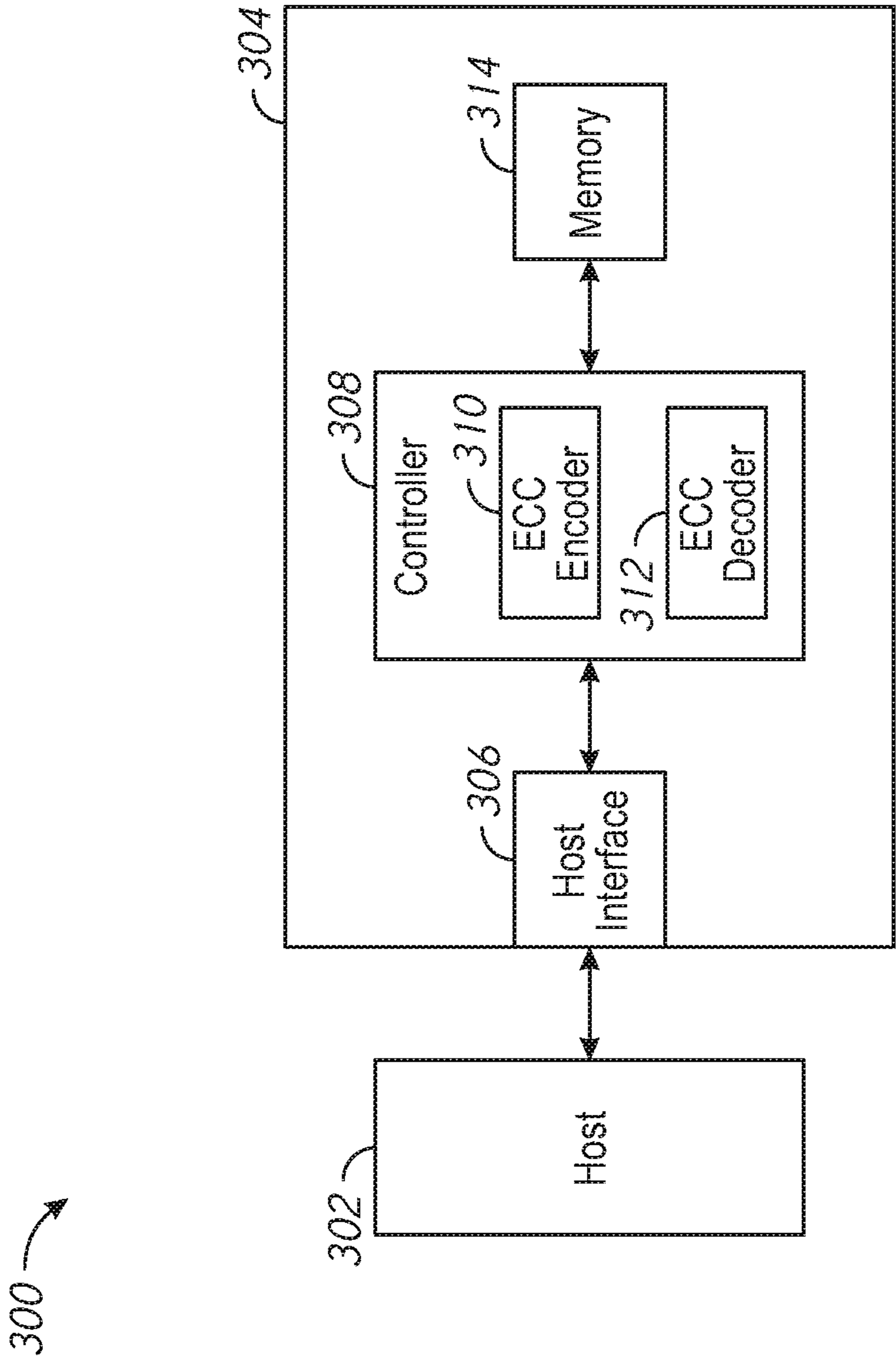


FIG. 3

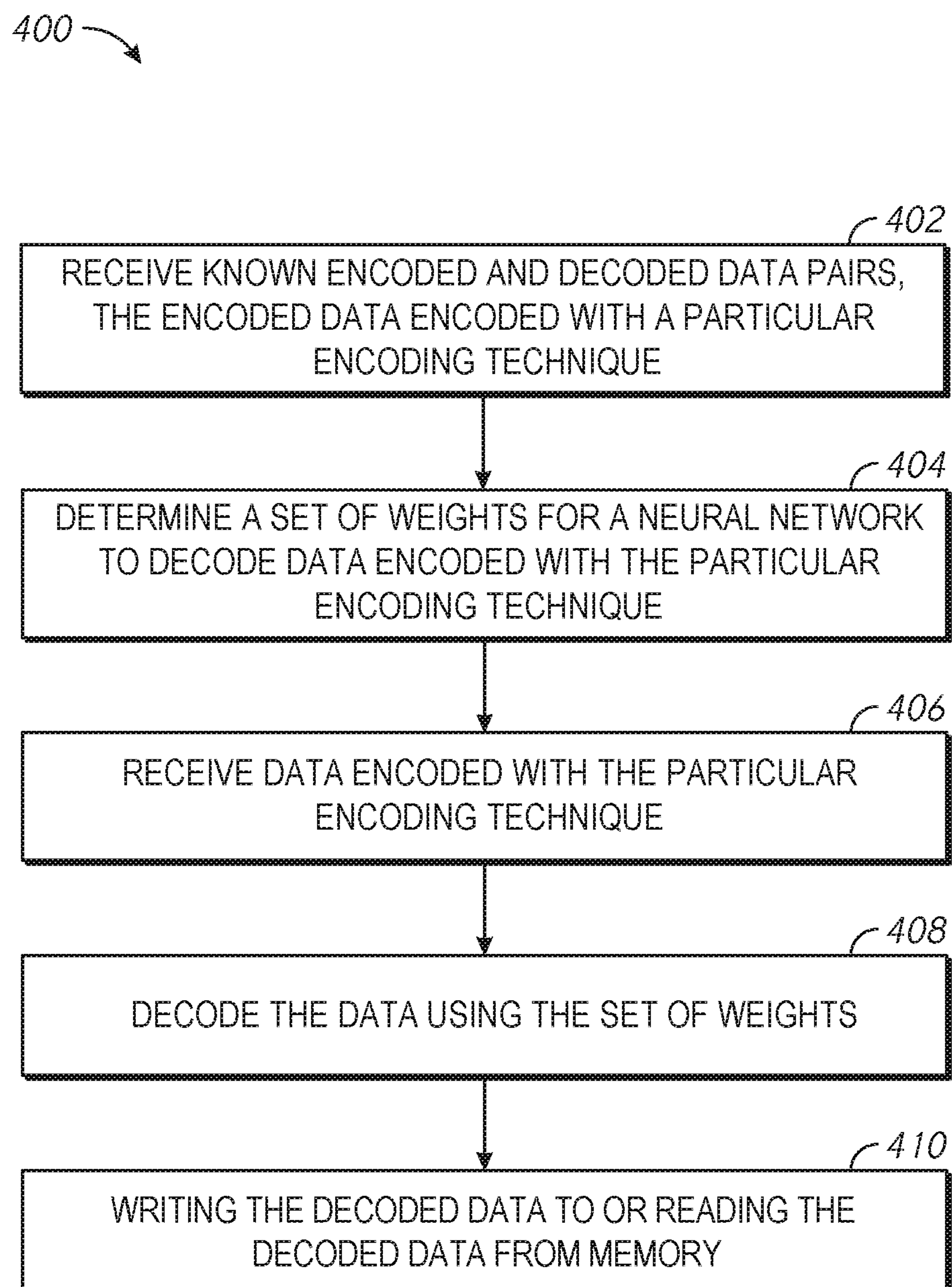
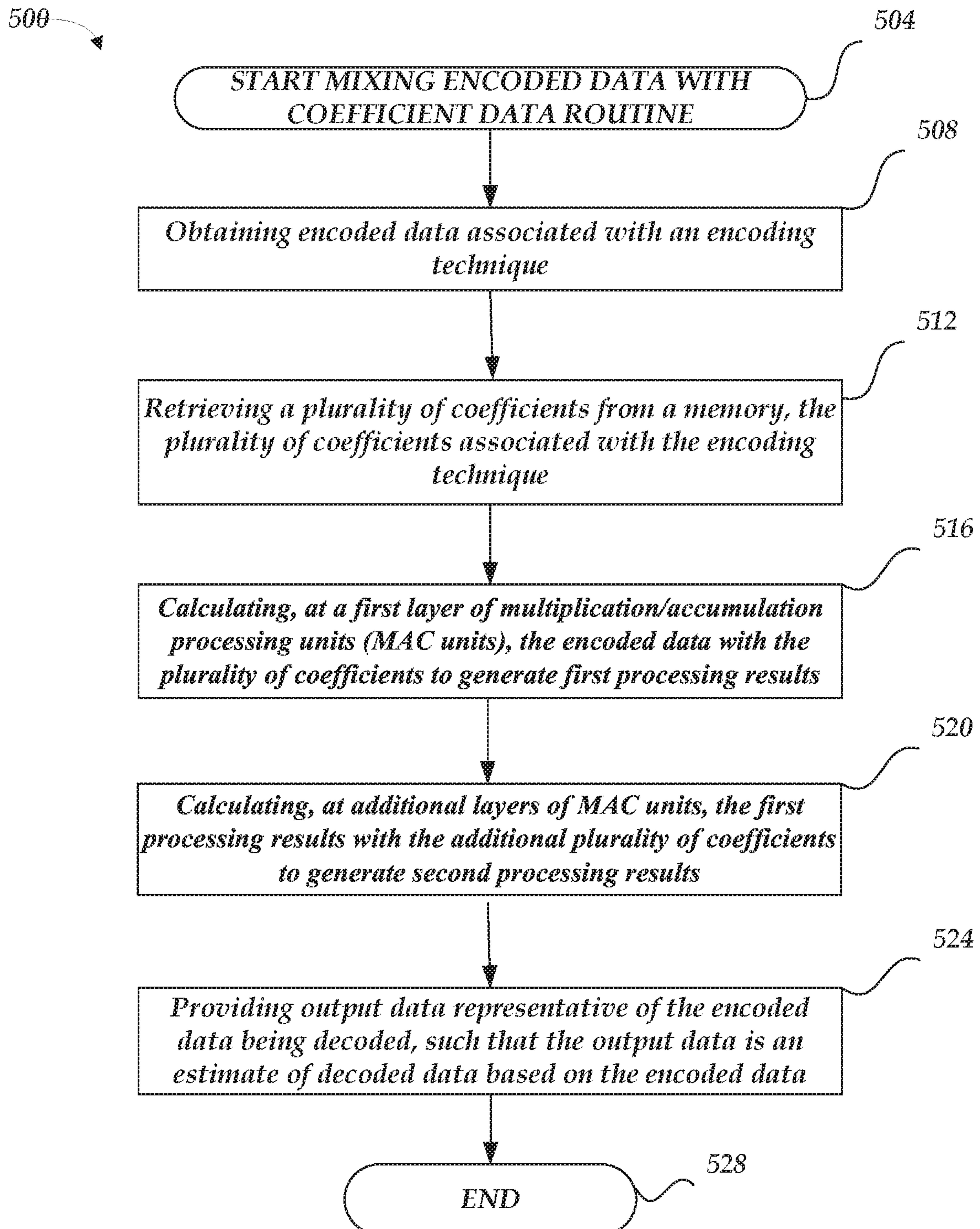


FIG. 4

**Fig. 5**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2020/056151

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063(2006.01)i; G06N 3/04(2006.01)i; G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N 3/063(2006.01); G06F 7/523(2006.01); G06N 3/04(2006.01); G06N 3/08(2006.01); G10L 15/16(2006.01); H04N 19/117(2014.01); H04N 19/50(2014.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: recurrent neural network (RNN), delay, set of weights, determining, decoding, encoding

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 10176802 B1 (AMAZON TECHNOLOGIES, INC.) 08 January 2019 (2019-01-08) column 5, line 64 - column 6, line 7; column 11, lines 53-64; column 12, lines 2-43; column 19, lines 59-63; and figures 3-4	1-13,15-16
A		14,17-21
Y	US 2018-0249158 A1 (MEDIATECK INC.) 30 August 2018 (2018-08-30) paragraphs [0009], [0014], [0032], [0034]; and claim 12	1-13,15-16
Y	US 2018-0046897 A1 (BEIJING DEEPHI INTELLIGENCE TECHNOLOGY CO., LTD.) 15 February 2018 (2018-02-15) paragraphs [0109], [0161]	6-7
A	US 2004-0015459 A1 (HERBERT JAEGER) 22 January 2004 (2004-01-22) paragraphs [0052]-[0081]	1-21

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family
--	--

Date of the actual completion of the international search

25 January 2021

Date of mailing of the international search report

26 January 2021

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsu-ro, Seo-gu, Daejeon
35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

YANG, Jeong Rok

Telephone No. +82-42-481-5709

C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2018-0084988 A (GOOGLE LLC) 25 July 2018 (2018-07-25) paragraphs [0008]-[0044]; and figure 1	1-21
A	ZACHARY C. LIPTON et al., ‘A Critical Review of Recurrent Neural Networks for Sequence Learning’, arXiv:1506.00019v4, 17 October 2015 [retrieved on 2021.01.11]. Retrieved from <https://arxiv.org/abs/1506.00019v4>. sections 3-5	1-21

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2020/056151

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	10176802	B1	08 January 2019	US	10210862	B1	19 February 2019
US	2018-0249158	A1	30 August 2018	CA	2997193	A1	09 March 2017
				CN	107925762	A	17 April 2018
				CN	107925762	B	27 November 2020
				EP	3342164	A1	04 July 2018
				EP	3342164	B1	15 April 2020
				KR	10-2018-0052651	A	18 May 2018
				KR	10-2124714	B1	19 June 2020
				WO	2017-036370	A1	09 March 2017
US	2018-0046897	A1	15 February 2018	CN	107239823	A	10 October 2017
				CN	107239829	B	11 December 2020
				CN	107657316	B	07 April 2020
				CN	107689224	B	01 September 2020
				CN	107689948	B	01 September 2020
				CN	107704916	B	03 July 2020
				US	10621486	B2	14 April 2020
				US	10643124	B2	05 May 2020
				US	10698657	B2	30 June 2020
				US	10762426	B2	01 September 2020
				US	10802992	B2	13 October 2020
				US	10810484	B2	20 October 2020
				US	10832123	B2	10 November 2020
US	2004-0015459	A1	22 January 2004	AU	2001-95591	A1	22 April 2002
				CA	2424929	A1	18 April 2002
				CA	2424929	C	03 April 2012
				EP	1356430	A2	29 October 2003
				EP	1356430	B1	14 July 2010
				JP	2004-511866	A	15 April 2004
				JP	4093858	B2	04 June 2008
				US	7321882	B2	22 January 2008
				WO	02-31764	A2	18 April 2002
				WO	02-31764	A3	21 August 2003
KR	10-2018-0084988	A	25 July 2018	CN	107038476	A	11 August 2017
				DE	102016125918	A1	03 August 2017
				EP	3374932	A1	19 September 2018
				JP	2019-509539	A	04 April 2019
				JP	6706326	B2	03 June 2020
				US	2017-0220925	A1	03 August 2017
				WO	2017-136070	A1	10 August 2017