



(51) International Patent Classification:

G06N 3/08 (2006.01) G06N 99/00 (2010.01)
G06N 3/04 (2006.01)

(21) International Application Number:

PCT/US2017/014101

(22) International Filing Date:

19 January 2017 (19.01.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

15/046,894 18 February 2016 (18.02.2016) US

(71) Applicant: INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95054 (US).

(72) Inventors: ZAMORA ESQUIVEL, Julio; Floresta #6 col El briseno, Zapopan, 45236 (MX). IBARRA VON BORSTEL, Alejandro; Av. Los Pinos #520-201, Zapopan, 45127 (MX). FLORES FAJARDO, Carlos; Cascadas 82 Colonia Loreto, Tlaquepaque, 45602 (MX). LOPEZ

MEYER, Paulo; Guadalajara Design Center, Periferico Sur 7980 Edif 4-E, Tlaquepaque, 45600 (MX). WU, Xiaoyun; Building 46-202, 60 South Jin Hui Road, 31, Shanghai 201103 (CN).

(74) Agent: KELLETT, Glen M.; Barnes & Thornburg LLP, 11 South Meridian Street, Indianapolis, IN 46204 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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,

[Continued on next page]

(54) Title: TECHNOLOGIES FOR SHIFTED NEURAL NETWORKS

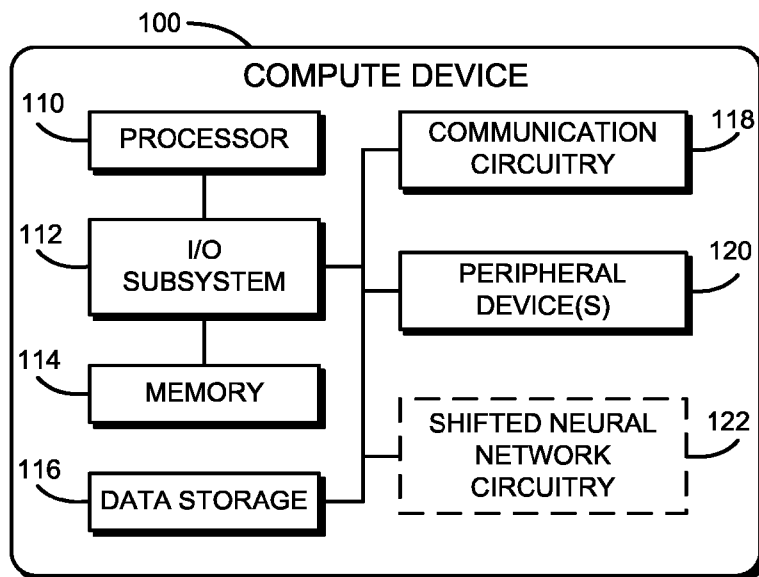


FIG. 1

(57) Abstract: Technologies for using a shifted neural network include a compute device to determine a shift-based activation function of the shifted neural network. The shift-based activation function is a piecewise linear approximation of a transcendental activation function and is defined by a plurality of line segments such that a corresponding slope of each line segment is a power of two. The compute device further trains the shifted neural network based on shift-based input weights and the determined shift-based activation function.



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, 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).

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

TECHNOLOGIES FOR SHIFTED NEURAL NETWORKS

CROSS-REFERENCE TO RELATED PATENT APPLICATION

[0001] The present application claims priority to U.S. Utility Patent Application Serial No. 15/046,894, entitled “TECHNOLOGIES FOR SHIFTED NEURAL NETWORKS,” which was filed on February 18, 2016.

BACKGROUND

[0002] The development and use of pattern recognition generally involves machine learning to recognize patterns in data by virtue, for example, of data regularities and/or irregularities. For example, a pattern recognition system may be “trained” to properly identify patterns by using training data that has pre-identified the relevant classes of the data points and/or may be used to recognize patterns in a previously unclassified set of data. One of the most common techniques for pattern recognition involves the use of Artificial Neural Networks (ANN), or simply neural networks, which were initially modeled after biological neural systems. Artificial Neural Networks are capable of learning and solving highly non-linear pattern recognition problems. In doing so, such networks typically involve the use of transcendental functions (e.g., sigmoid functions, hyperbolic tangent functions, etc.) and/or memory-intensive algorithms. Accordingly, the use of neural networks for pattern recognition is often limited or non-existent with wearable computing devices, low-power devices, and/or other hardware-limited devices (e.g., devices without a dedicated acceleration unit suitable for pattern recognition).

BRIEF DESCRIPTION OF THE DRAWINGS

[0003] The concepts described herein are illustrated by way of example and not by way of limitation in the accompanying figures. For simplicity and clarity of illustration, elements illustrated in the figures are not necessarily drawn to scale. Where considered appropriate, reference labels have been repeated among the figures to indicate corresponding or analogous elements.

[0004] FIG. 1 is a simplified block diagram of at least one embodiment of a compute device for using a shifted neural network;

[0005] FIG. 2 is a simplified block diagram of at least one embodiment of an environment of a compute device of FIG. 1;

[0006] FIG. 3 is a simplified flow diagram of at least one embodiment of a method for using a shifted neural network that may be executed by the compute device of FIG. 1;

[0007] FIG. 4 is a simplified flow diagram of at least one embodiment of a method of determining a shift-based activation function for the shifted neural network; and

[0008] FIG. 5 is a simplified graph illustrating the determination of a piecewise linear approximation of a transcendental activation function for the shifted neural network.

DETAILED DESCRIPTION OF THE DRAWINGS

[0009] While the concepts of the present disclosure are susceptible to various modifications and alternative forms, specific embodiments thereof have been shown by way of example in the drawings and will be described herein in detail. It should be understood, however, that there is no intent to limit the concepts of the present disclosure to the particular forms disclosed, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives consistent with the present disclosure and the appended claims.

[0010] References in the specification to “one embodiment,” “an embodiment,” “an illustrative embodiment,” etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may or may not necessarily include that particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to effect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described. Additionally, it should be appreciated that items included in a list in the form of “at least one A, B, and C” can mean (A); (B); (C); (A and B); (B and C); (A and C); or (A, B, and C). Similarly, items listed in the form of “at least one of A, B, or C” can mean (A); (B); (C); (A and B); (B and C); (A and C); or (A, B, and C).

[0011] The disclosed embodiments may be implemented, in some cases, in hardware, firmware, software, or any combination thereof. The disclosed embodiments may also be implemented as instructions carried by or stored on one or more transitory or non-transitory machine-readable (e.g., computer-readable) storage medium, which may be read and executed by one or more processors. A machine-readable storage medium may be embodied as any storage device, mechanism, or other physical structure for storing or transmitting information in a form readable by a machine (e.g., a volatile or non-volatile memory, a media disc, or other media device).

[0012] In the drawings, some structural or method features may be shown in specific arrangements and/or orderings. However, it should be appreciated that such specific

arrangements and/or orderings may not be required. Rather, in some embodiments, such features may be arranged in a different manner and/or order than shown in the illustrative figures. Additionally, the inclusion of a structural or method feature in a particular figure is not meant to imply that such feature is required in all embodiments and, in some embodiments, may not be included or may be combined with other features.

[0013] Referring now to FIG. 1, a compute device 100 for using a shifted neural network is shown. As described in detail below, in the illustrative embodiment, the compute device 100 utilizes a shifted neural network for machine learning, pattern recognition, and/or other suitable purposes. In some embodiments, the shifted neural network eliminates the use of double precision floating point units in neural network calculations, eliminates the use of transcendental functions, eliminates the need for multiplication operations in the feed-forward layers of a typical neural network, avoids input pattern normalization, and/or reduces the memory footprint of the neural network relative to typical neural networks. Accordingly, in some embodiments, hardware for floating-point operations and/or multiplication may not be included in the circuitry for shifted neural network computations. Further, the processor 110 of the compute device 100 may consume less power and/or process the neural network computations more quickly with the shifted neural network described herein.

[0014] As described below, the compute device 100 determines a shift-based activation function of the shifted neural network (e.g., a piecewise linear approximation of transcendental activation function) and stores the shifted neural network weights as bitwise shifts (e.g., as four bit values) of the input values. The compute device 100 may train the shifted neural network based on the shift-based input weights and the shift-based activation function and/or apply the shifted neural network (e.g., for pattern recognition). Although the shifted neural network techniques are described herein as being performed by the same compute device 100, it should be appreciated that one or more of the techniques may be performed by various compute devices 100 in other embodiments. For example, in some embodiments, one compute device 100 may determine the shift-based activation function and train the shifted neural network, whereas another compute device 100 may apply the shifted neural network (e.g., a wearable computing device).

[0015] The compute device 100 may be embodied as any type of computing device capable of performing the functions described herein. For example, the compute device 100 may be embodied as a wearable computing device, smartphone, cellular phone, personal digital assistant, mobile Internet device, tablet computer, netbook, notebook, Ultrabook™, laptop computer, desktop computer, server, and/or any other mobile computing/communication

device. In the illustrative embodiment, the compute device 100 is embodied as a wearable computing device or other mobile computing device having relatively limited hardware resources. As shown in FIG. 1, the illustrative compute device 100 includes a processor 110, an input/output (“I/O”) subsystem 112, a memory 114, a data storage 116, a communication circuitry 118, and one or more peripheral devices 120. Additionally, in some embodiments, the compute device 100 may include a shifted neural network circuitry 122. Of course, the compute device 100 may include other or additional components, such as those commonly found in a typical computing device (e.g., various input/output devices and/or other components), in other embodiments. Additionally, in some embodiments, one or more of the illustrative components may be incorporated in, or otherwise form a portion of, another component. For example, the memory 114, or portions thereof, may be incorporated in the processor 110 in some embodiments.

[0016] The processor 110 may be embodied as any type of processor capable of performing the functions described herein. For example, the processor 110 may be embodied as a single or multi-core processor(s), digital signal processor, microcontroller, or other processor or processing/controlling circuit. Similarly, the memory 114 may be embodied as any type of volatile or non-volatile memory or data storage capable of performing the functions described herein. In operation, the memory 114 may store various data and software used during operation of the compute device 100 such as operating systems, applications, programs, libraries, and drivers. The memory 114 is communicatively coupled to the processor 110 via the I/O subsystem 112, which may be embodied as circuitry and/or components to facilitate input/output operations with the processor 110, the memory 114, and other components of the compute device 100. For example, the I/O subsystem 112 may be embodied as, or otherwise include, memory controller hubs, input/output control hubs, firmware devices, communication links (i.e., point-to-point links, bus links, wires, cables, light guides, printed circuit board traces, etc.) and/or other components and subsystems to facilitate the input/output operations. In some embodiments, the I/O subsystem 112 may form a portion of a system-on-a-chip (SoC) and be incorporated, along with the processor 110, the memory 114, and other components of the compute device 100, on a single integrated circuit chip.

[0017] The data storage 116 may be embodied as any type of device or devices configured for short-term or long-term storage of data such as, for example, memory devices and circuits, memory cards, hard disk drives, solid-state drives, or other data storage devices. The data storage 116 and/or the memory 114 may store various data during operation of the compute device 100 as described herein.

[0018] The communication circuitry 118 may be embodied as any communication circuit, device, or collection thereof, capable of enabling communications between the compute device 100 and other devices over a network. The communication circuitry 118 may be configured to use any one or more communication technologies (e.g., wireless or wired communications) and associated protocols (e.g., Ethernet, Bluetooth[®], Wi-Fi[®], WiMAX, LTE, 5G, etc.) to effect such communication.

[0019] The peripheral devices 120 may include any number of additional peripheral or interface devices, such as speakers, microphones, additional storage devices, and so forth. The particular devices included in the peripheral devices 120 may depend on, for example, the type and/or intended use of the compute device 100.

[0020] As indicated above, in some embodiments, the compute device 100 may include a shifted neural network circuitry 122 configured to perform one or more functions related to the shifted neural network (e.g., pattern recognition). In some embodiments, the shifted neural network circuitry 122 may include various circuit elements/components to perform operations for the shifted neural network (e.g., multiplexers, summers, bit shifters, etc.). Such components may, for example, be discrete and/or embedded on an integrated circuit depending on the particular embodiment. Further, in some embodiments, floating-point circuitry and/or multiplication circuit components may be omitted from the shifted neural network circuitry 122.

[0021] Referring now to FIG. 2, in use, the compute device 100 establishes an environment for using a shifted neural network. The illustrative environment 200 includes a data handling module 202, a linear approximation module 204, a training module 206, a pattern recognition module 208, and a communication module 210. The various modules of the environment 200 may be embodied as hardware, software, firmware, or a combination thereof. For example, the various modules, logic, and other components of the environment 200 may form a portion of, or otherwise be established by, the processor 110 or other hardware components of the compute device 100. As such, in some embodiments, one or more of the modules of the environment 200 may be embodied as circuitry or collection of electrical devices (e.g., a data handling circuitry, a linear approximation circuitry, a training circuitry, a pattern recognition circuitry, and/or a communication circuitry). Additionally, in some embodiments, one or more of the illustrative modules may form a portion of another module and/or one or more of the illustrative modules may be independent of one another.

[0022] The data handling module 202 is configured to identify, receive, and/or retrieve input values by which to train the shifted neural network and/or to which to apply the shifted

neural network. For example, in some embodiments, the data handling module 202 may identify/retrieve a set of predefined image patterns for recognition (e.g., Mixed National Institute of Standards and Technology (MNIST) database or other image patterns approved by a standards body). Further, the data handling module 202 may identify and/or retrieve the set of shift-based input weights for each neuron of the shifted neural network at a given point in time. As described herein, in the illustrative embodiment, the weights may be stored as four bits in memory (e.g., in a four-bit memory region) such that three of the bits are indicative of an amount of bitwise shift of the corresponding weight and the other bit is indicative of a sign of the weight. Accordingly, in the illustrative embodiments, each of the weights is essentially a power of two (i.e., may be expressed as 2^n for some integer n) if multiplication were to be used instead of a bitwise shift operation.

[0023] The linear approximation module 204 is configured to determine a shift-based activation function of the shifted neural network. In particular, in the illustrative embodiment, the shift-based activation function is a piecewise linear approximation of a transcendental activation function (see FIG. 5) and is defined by a plurality of line segments such that the slope of each line segment is a power of two. Accordingly, in the illustrative embodiment, the slopes may be stored and/or calculated as bitwise shifts. The transcendental functions may be embodied, for example, as a sigmoid function or a hyperbolic tangent function depending on the desired operating range. As described below, in some embodiments, the shift-based activation function may be determined based on k-means clustering.

[0024] The training module 206 is configured to train the shifted neural network based on the shift-based input weights and the determined shift-based activation function. It should be appreciated that the training module 206 may utilize any suitable techniques, algorithms, and/or mechanisms for training the shifted neural network. For example, in some embodiments, the training module 206 may utilize many neurons and variations in the shift-based weights iteratively in order to distinguish/label various classes of data. Depending on the particular embodiment, the training module 206 may utilize supervised and/or unsupervised training techniques.

[0025] The pattern recognition module 208 is configured to apply the shifted neural network to various inputs for pattern recognition. For example, in some embodiments, the pattern recognition module 208 may apply the trained shifted neural network to a set of unclassified input data points in order to determine the corresponding classes of those data points. As indicated above, in some embodiments, various techniques described herein may be employed on different devices. For example, one compute device 100 may perform the training

of the shifted neural network and another compute device 100 may apply the shifted neural network (e.g., for pattern recognition).

[0026] The communication module 210 handles the communication between the compute device 100 and other computing devices. For example, in some embodiments, the compute device 100 may transmit the shift-based activation function, shift-based weights, and/or other data associated with the shifted neural network to another compute device 100.

[0027] Referring now to FIG. 3, in use, the compute device 100 may execute a method 300 for use of a shifted neural network. The illustrative method 300 begins with block 302 in which the compute device 100 identifies, receives, and/or retrieves a set of input values by which to train the shifted neural network and/or to which to apply the shifted neural network. For example, in block 304, the compute device 100 may identify/receive one or more image patterns such as MNIST patterns. Of course, in other embodiments, it should be appreciated that the techniques described herein may be applied to non-image data.

[0028] In typical neural network implementations, the input values are often normalized (e.g., from the signed char range of (-128,127) to the floating-point range of (-1.0f, 1.0f)) to reach consistency within the data's dynamic range, which results in floating-point multiplications in the convolutional neural network engine. However, in the illustrative embodiment, the input values may be utilized without normalization in the shifted neural network, which may save memory, time, operations, and/or power.

[0029] In block 306, the compute device 100 determines a shift-based activation function of the shifted neural network. In particular, in block 308, the compute device 100 determines a piecewise linear approximation of a transcendental activation function such that the piecewise linear approximation is defined by a set of line segments, each having a slope that is a power of two (and therefore calculable by virtue of bitwise shift operations). For example, the compute device 100 may generate a linear approximation of a sigmoid function in block 310 or generate a linear approximation of a hyperbolic tangent function in block 312 (see, for example, function 504 of FIG. 5). Of course, it should be appreciated that the compute device 100 may utilize another transcendental function suitable for the performance of the functions described herein in other embodiments.

[0030] In some embodiments, in block 314, the compute device 100 may generate the linear approximation of the transcendental activation function based on k-means clustering. To do so, the compute device 100 may execute a method 400 of FIG. 4. The illustrative method 400 begins with block 402 in which the compute device 100 computes a slope value as a base-two logarithm of a tangent line (see, for example, tangent lines 502 of FIG. 5) of the

transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function (see, for example, function 504 of FIG. 5).

More formally, the compute device 100 may determine $y_j = Af\left(\frac{x_j}{B}\right)$ and

$$\sigma_j = -\log_2\left(Af\left(\frac{x_j}{B}\right)\right),$$

where j is an index, A and B are scale constants, f is the transcendental

activation function, y_j is the corresponding output value, x_j is the corresponding input value, and σ_j is a slope of the function. It should be appreciated that, in some embodiments, the tangent lines may be initialized randomly or according to some other suitable algorithm.

[0031] In block 404, the compute device 100 associates each slope with the nearest cluster. In other words, the compute device 100 determines the “distance” of each slope value to each of the centroid values and associates the slope with the nearest cluster. More formally, the compute device 100 may determine $S_j^{(t)} = \{\sigma_p : \|\sigma_p - m_i^{(t)}\| < \|\sigma_p - m_j^{(t)}\|, \forall 1 \leq j \leq k\}$, where $S_j^{(t)}$ is a particular set/cluster (of the k clusters), σ_p is a corresponding slope of the function, $m_i^{(t)}$ and $m_j^{(t)}$ are means of the corresponding clusters, and p, j, i , and k are indices.

[0032] In block 406, the compute device 100 updates the centroid (e.g., mean) associated with each cluster and computes the offset for each tangent line. More formally, the

$$\text{compute device 100 may determine } b_i = \frac{1}{|S_i^{(t)}|} \sum_{\sigma_j \in S_i^{(t)}} y_j - \frac{x_j}{2^{m_i}} \text{ and } m_i^{(t+1)} = \text{Round}\left(\frac{1}{|S_i^{(t)}|} \sum_{\sigma_j \in S_i^{(t)}} \sigma_j\right)$$

, where b_i is an offset of the corresponding tangent line relative to the origin, $S_j^{(t)}$ is a particular set/cluster, σ_j is a corresponding slope of the function, $m_i^{(t+1)}$ is a mean of the corresponding cluster, y_j is the corresponding output value, and x_j is the corresponding input value.

In block 408, the compute device 100 determines whether each centroid value is stable. In other words, the compute device 100 determines whether the centroid values of the clusters are unchanged (or minimally changed) relative to the previous iteration. More formally, the compute device 100 may determine whether $m_i^{(t+1)} = m_i^{(t)}, \forall 1 \leq i \leq k-1$. If the centroid values are not stable, the method 400 returns to block 404 in which the compute device 100 again determines the nearest cluster to each of the slopes and re-associates the slopes accordingly (i.e., to the extent there is a change). However, if the compute device 100

determines that the centroid values are stable, the compute device 100 determines the piecewise linear approximation in block 410. In doing so, in the illustrative embodiment, the compute device 100 computes the intersections of the tangent lines in block 412. More formally, the compute device 100 may determine the intersections according to $r_i = m_i m_{i+1} \left(\frac{b_i - b_{i+1}}{m_i - m_{i+1}} \right), \forall 1 \leq i \leq k-1$ and may determine the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein r_i is an intersection of two tangent lines and $\gg$ is a bitwise shift operator. It should be appreciated that the techniques described herein may be employed to map a continuous neural network into a shifted neural network. In some embodiments, the shifted neural network may have significant memory compression (e.g., eight times) relative to typical neural networks. Further, in some embodiments, the shifted neural network may permit an implementation on a small SoC by avoiding floating-point and integer multiplication operations as described above.

[0033] Referring back to FIG. 3, in block 316, the compute device 100 trains the shifted neural network based on the relevant shift-based weights and the determined shift-based activation function. As indicated above, it should be appreciated that the compute device 100 may utilize any suitable techniques, algorithms, and/or mechanisms for training the shifted neural network. For example, in some embodiments, the compute device 100 may utilize many neurons and variations in the shift-based weights iteratively in order to distinguish/label various classes of data.

[0034] In typical neural networks, it should be appreciated that the neuron weights may be accumulated according to $s = \sum w_i x_i$, where w_i is a weight and x_i is the corresponding input. In other words, the input values are multiplied by various weights and summed. However, in the illustrative embodiment, each of the weights is indicative of an amount by which to perform a bitwise shift on the input value. Accordingly, in memory, the compute device 100 may utilize three bits to store the shift amount (e.g., 0, 1, 2, 3, 4, 5, 6, or 7) and one bit for the sign. As such, in the illustrative embodiment, each of the weights is a power of two (e.g., -128, -64, -32, -16, -8, -4, -2, -1, 1, 2, 4, 8, 16, 32, 64, or 128). It should be appreciated that such a data representation and operations may represent a significant memory compression relative to integer and/or floating-point multiplication associated with a convolutional engine.

[0035] In block 318, the compute device 100 applies the shifted neural network to various inputs for pattern recognition. For example, in some embodiments, the compute device 100 may apply the trained shifted neural network to a set of unclassified input data points in order to determine the corresponding classes of those data points. As indicated above, in some embodiments, it should be appreciated that one compute device 100 may perform the training of the shifted neural network and another compute device 100 may apply the shifted neural network (e.g., for pattern recognition).

EXAMPLES

[0036] Illustrative examples of the technologies disclosed herein are provided below. An embodiment of the technologies may include any one or more, and any combination of, the examples described below.

[0037] Example 1 includes a compute device for use of a shifted neural network, the compute device comprising a linear approximation module to determine a shift-based activation function of the shifted neural network, wherein the shift-based activation function is a piecewise linear approximation of a transcendental activation function and is defined by a plurality of line segments, and wherein a corresponding slope of each line segment is a power of two; and a training module to train the shifted neural network based on shift-based input weights and the determined shift-based activation function.

[0038] Example 2 includes the subject matter of Example 1, and further including a data handling module to identify a set of input values and a set of shift-based input weights for each neuron of the shifted neural network.

[0039] Example 3 includes the subject matter of any of Examples 1 and 2, and wherein the set of input values comprises a set of predefined image patterns for recognition.

[0040] Example 4 includes the subject matter of any of Examples 1-3, and wherein the set of predefined image patterns comprises a set of image patterns approved by a standards body.

[0041] Example 5 includes the subject matter of any of Examples 1-4, and wherein the transcendental activation function comprises a sigmoid function.

[0042] Example 6 includes the subject matter of any of Examples 1-5, and wherein the transcendental activation function comprises a hyperbolic tangent function.

[0043] Example 7 includes the subject matter of any of Examples 1-6, and wherein to determine the shift-based activation function comprises to determine the shift-based activation function based on k-means clustering.

[0044] Example 8 includes the subject matter of any of Examples 1-7, and wherein to determine the shift-based activation function based on k-means clustering comprises to compute a slope value as a base-two logarithm of a tangent line of the transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function; associate each slope value with a corresponding nearest cluster; and update a centroid associated with each cluster.

[0045] Example 9 includes the subject matter of any of Examples 1-8, and wherein to determine the shift-based activation function based on k-means clustering further comprises to compute an offset of the tangent line relative to an origin.

[0046] Example 10 includes the subject matter of any of Examples 1-9, and wherein to determine the shift-based activation function based on k-means clustering further comprises to compute intersections of the tangent lines in response to a determination that each centroid value is stable.

[0047] Example 11 includes the subject matter of any of Examples 1-10, and wherein to determine the shift-based activation function comprises to generate the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein x is an input value, σ_i is the slope value of the i^{th} tangent line, b_i is an offset of the i^{th} tangent line relative to the transcendental activation function, r_i is an intersection of two tangent lines, and $\gg$ is a bitwise shift operator.

[0048] Example 12 includes the subject matter of any of Examples 1-11, and further including a pattern recognition module to apply the shifted neural network to an input for pattern recognition.

[0049] Example 13 includes the subject matter of any of Examples 1-12, and wherein each corresponding slope is computed by a bitwise shift operation.

[0050] Example 14 includes the subject matter of any of Examples 1-13, and wherein each of the input weights is stored in a four-bit memory region; and wherein three bits of the four-bit memory region store a value associated with an amount of bitwise shift of the corresponding input weight and one bit is indicative of a sign of the weight.

[0051] Example 15 includes a method for use of a shifted neural network, the method comprising determining, by a compute device, a shift-based activation function of the shifted neural network, wherein the shift-based activation function is a piecewise linear approximation

of a transcendental activation function and is defined by a plurality of line segments, and wherein a corresponding slope of each line segment is a power of two; and training, by the compute device, the shifted neural network based on shift-based input weights and the determined shift-based activation function.

[0052] Example 16 includes the subject matter of Example 15, and further including identifying, by the compute device, a set of input values and a set of shift-based input weights for each neuron of the shifted neural network.

[0053] Example 17 includes the subject matter of any of Examples 15 and 16, and wherein the set of input values comprises a set of predefined image patterns for recognition.

[0054] Example 18 includes the subject matter of any of Examples 15-17, and wherein the set of predefined image patterns comprises a set of image patterns approved by a standards body.

[0055] Example 19 includes the subject matter of any of Examples 15-18, and wherein the transcendental activation function comprises a sigmoid function.

[0056] Example 20 includes the subject matter of any of Examples 15-19, and wherein the transcendental activation function comprises a hyperbolic tangent function.

[0057] Example 21 includes the subject matter of any of Examples 15-20, and wherein determining the shift-based activation function comprises determining the shift-based activation function based on k-means clustering.

[0058] Example 22 includes the subject matter of any of Examples 15-21, and wherein determining the shift-based activation function based on k-means clustering comprises computing a slope value as a base-two logarithm of a tangent line of the transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function; associating each slope value with a corresponding nearest cluster; and updating a centroid associated with each cluster.

[0059] Example 23 includes the subject matter of any of Examples 15-22, and wherein determining the shift-based activation function based on k-means clustering further comprises computing an offset of the tangent line relative to an origin.

[0060] Example 24 includes the subject matter of any of Examples 15-23, and wherein determining the shift-based activation function based on k-means clustering further comprises computing intersections of the tangent lines in response to a determination that each centroid value is stable.

[0061] Example 25 includes the subject matter of any of Examples 15-24, and wherein determining the shift-based activation function comprises generating the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein x is an input value, σ_i is the slope value of the i^{th} tangent line, b_i is an offset of the i^{th} tangent line relative to the transcendental activation function, r_i is an intersection of two tangent lines, and $\gg$ is a bitwise shift operator.

[0062] Example 26 includes the subject matter of any of Examples 15-25, and further including applying, by the compute device, the shifted neural network to an input for pattern recognition.

[0063] Example 27 includes the subject matter of any of Examples 15-26, and wherein each corresponding slope is computed by a bitwise shift operation.

[0064] Example 28 includes the subject matter of any of Examples 15-27, and wherein each of the input weights is stored in a four-bit memory region; and wherein three bits of the four-bit memory region store a value associated with an amount of bitwise shift of the corresponding input weight and one bit is indicative of a sign of the weight.

[0065] Example 29 includes a computing device comprising a processor; and a memory having stored therein a plurality of instructions that when executed by the processor cause the computing device to perform the method of any of Examples 15-28.

[0066] Example 30 includes one or more machine-readable storage media comprising a plurality of instructions stored thereon that in response to being executed result in a computing device performing the method of any of Examples 15-28.

[0067] Example 31 includes a computing device comprising means for performing the method of any of Examples 15-28.

[0068] Example 32 includes a compute device for use of a shifted neural network, the compute device comprising means for determining a shift-based activation function of the shifted neural network, wherein the shift-based activation function is a piecewise linear approximation of a transcendental activation function and is defined by a plurality of line segments, and wherein a corresponding slope of each line segment is a power of two; and means for training the shifted neural network based on shift-based input weights and the determined shift-based activation function.

[0069] Example 33 includes the subject matter of Example 32, and further including means for identifying a set of input values and a set of shift-based input weights for each neuron of the shifted neural network.

[0070] Example 34 includes the subject matter of any of Examples 32 and 33, and wherein the set of input values comprises a set of predefined image patterns for recognition.

[0071] Example 35 includes the subject matter of any of Examples 32-34, and wherein the set of predefined image patterns comprises a set of image patterns approved by a standards body.

[0072] Example 36 includes the subject matter of any of Examples 32-35, and wherein the transcendental activation function comprises a sigmoid function.

[0073] Example 37 includes the subject matter of any of Examples 32-36, and wherein the transcendental activation function comprises a hyperbolic tangent function.

[0074] Example 38 includes the subject matter of any of Examples 32-37, and wherein the means for determining the shift-based activation function comprises means for determining the shift-based activation function based on k-means clustering.

[0075] Example 39 includes the subject matter of any of Examples 32-38, and wherein the means for determining the shift-based activation function based on k-means clustering comprises means for computing a slope value as a base-two logarithm of a tangent line of the transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function; means for associating each slope value with a corresponding nearest cluster; and means for updating a centroid associated with each cluster.

[0076] Example 40 includes the subject matter of any of Examples 32-39, and wherein the means for determining the shift-based activation function based on k-means clustering further comprises means for computing an offset of the tangent line relative to an origin.

[0077] Example 41 includes the subject matter of any of Examples 32-40, and wherein the means for determining the shift-based activation function based on k-means clustering further comprises means for computing intersections of the tangent lines in response to a determination that each centroid value is stable.

[0078] Example 42 includes the subject matter of any of Examples 32-41, and wherein the means for determining the shift-based activation function comprises means for generating the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein x is an input value, σ_i is the slope value of the i^{th} tangent line, b_i is an offset of the i^{th} tangent line relative to the transcendental activation function, r_i is an intersection of two tangent lines, and $\gg$ is a bitwise shift operator.

[0079] Example 43 includes the subject matter of any of Examples 32-42, and further including means for applying the shifted neural network to an input for pattern recognition.

[0080] Example 44 includes the subject matter of any of Examples 32-43, and wherein each corresponding slope is computed by a bitwise shift operation.

[0081] Example 45 includes the subject matter of any of Examples 32-44, and wherein each of the input weights is stored in a four-bit memory region; and wherein three bits of the four-bit memory region store a value associated with an amount of bitwise shift of the corresponding input weight and one bit is indicative of a sign of the weight.

WHAT IS CLAIMED IS:

1. A compute device for use of a shifted neural network, the compute device comprising:

a linear approximation module to determine a shift-based activation function of the shifted neural network, wherein the shift-based activation function is a piecewise linear approximation of a transcendental activation function and is defined by a plurality of line segments, and wherein a corresponding slope of each line segment is a power of two; and

a training module to train the shifted neural network based on shift-based input weights and the determined shift-based activation function.

2. The compute device of claim 1, further comprising a data handling module to identify a set of input values and a set of shift-based input weights for each neuron of the shifted neural network.

3. The compute device of claim 2, wherein the set of input values comprises a set of predefined image patterns for recognition.

4. The compute device of claim 3, wherein the set of predefined image patterns comprises a set of image patterns approved by a standards body.

5. The compute device of claim 1, wherein the transcendental activation function comprises a sigmoid function.

6. The compute device of claim 1, wherein the transcendental activation function comprises a hyperbolic tangent function.

7. The compute device of claim 1, wherein to determine the shift-based activation function comprises to determine the shift-based activation function based on k-means clustering.

8. The compute device of claim 7, wherein to determine the shift-based activation function based on k-means clustering comprises to:

compute a slope value as a base-two logarithm of a tangent line of the transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function;

associate each slope value with a corresponding nearest cluster; and

update a centroid associated with each cluster.

9. The compute device of claim 8, wherein to determine the shift-based activation function based on k-means clustering further comprises to compute an offset of the tangent line relative to an origin.

10. The compute device of claim 8, wherein to determine the shift-based activation function based on k-means clustering further comprises to compute intersections of the tangent lines in response to a determination that each centroid value is stable.

11. The compute device of claim 10, wherein to determine the shift-based activation function comprises to generate the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein x is an input value, σ_i is the slope value of the i^{th} tangent line, b_i is an offset of the i^{th} tangent line relative to the transcendental activation function, r_i is an intersection of two tangent lines, and $\gg$ is a bitwise shift operator.

12. The compute device of claim 1, further comprising a pattern recognition module to apply the shifted neural network to an input for pattern recognition.

13. The compute device of any of claims 1-12, wherein each corresponding slope is computed by a bitwise shift operation.

14. The compute device of any of claims 1-12, wherein each of the input weights is stored in a four-bit memory region; and

wherein three bits of the four-bit memory region store a value associated with an amount of bitwise shift of the corresponding input weight and one bit is indicative of a sign of the weight.

15. A method for use of a shifted neural network, the method comprising:
determining, by a compute device, a shift-based activation function of the shifted neural network, wherein the shift-based activation function is a piecewise linear approximation of a transcendental activation function and is defined by a plurality of line segments, and wherein a corresponding slope of each line segment is a power of two; and
training, by the compute device, the shifted neural network based on shift-based input weights and the determined shift-based activation function.

16. The method of claim 15, further comprising identifying, by the compute device, a set of input values and a set of shift-based input weights for each neuron of the shifted neural network.

17. The method of claim 16, wherein the set of input values comprises a set of predefined image patterns for recognition.

18. The method of claim 17, wherein the set of predefined image patterns comprises a set of image patterns approved by a standards body.

19. The method of claim 15, wherein the transcendental activation function comprises a sigmoid function.

20. The method of claim 15, wherein the transcendental activation function comprises a hyperbolic tangent function.

21. The method of claim 15, wherein determining the shift-based activation function comprises determining the shift-based activation function based on k-means clustering by:

computing a slope value as a base-two logarithm of a tangent line of the transcendental activation function for each function value of a plurality of function values in the range of the transcendental activation function;

associating each slope value with a corresponding nearest cluster; and

updating a centroid associated with each cluster.

22. The method of claim 21, wherein determining the shift-based activation function based on k-means clustering further comprises:

computing an offset of the tangent line relative to an origin; and

computing intersections of the tangent lines in response to a determination that each centroid value is stable.

23. The method of claim 22, wherein determining the shift-based activation function comprises generating the piecewise linear approximation according to

$$y \approx \begin{cases} x \gg \sigma_1 + b_1, & 0 < x < r_1 \\ \vdots & \\ x \gg \sigma_k + b_k, & r_{k-1} < x < r_k \end{cases}$$

wherein x is an input value, σ_i is the slope value of the i^{th} tangent line, b_i is an offset of the i^{th} tangent line relative to the transcendental activation function, r_i is an intersection of two tangent lines, and $\gg$ is a bitwise shift operator.

24. The method of claim 15, wherein each corresponding slope is computed by a bitwise shift operation.

25. One or more machine-readable storage media comprising a plurality of instructions stored thereon that in response to being executed result in a computing device performing the method of any of claims 15-24.

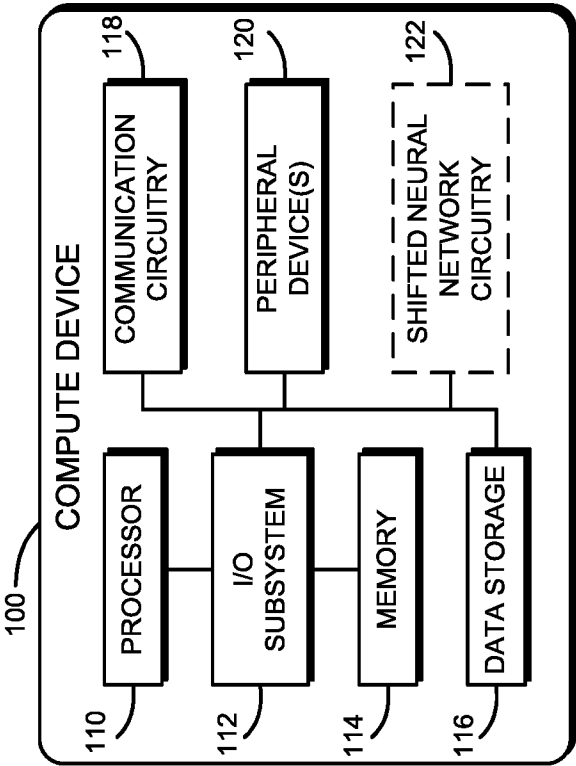


FIG. 1

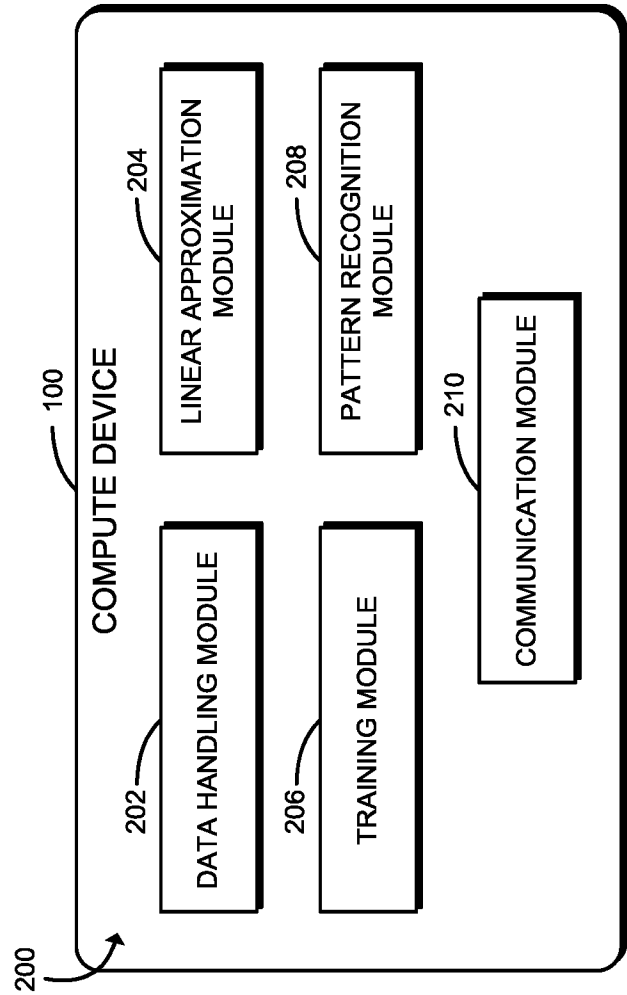


FIG. 2

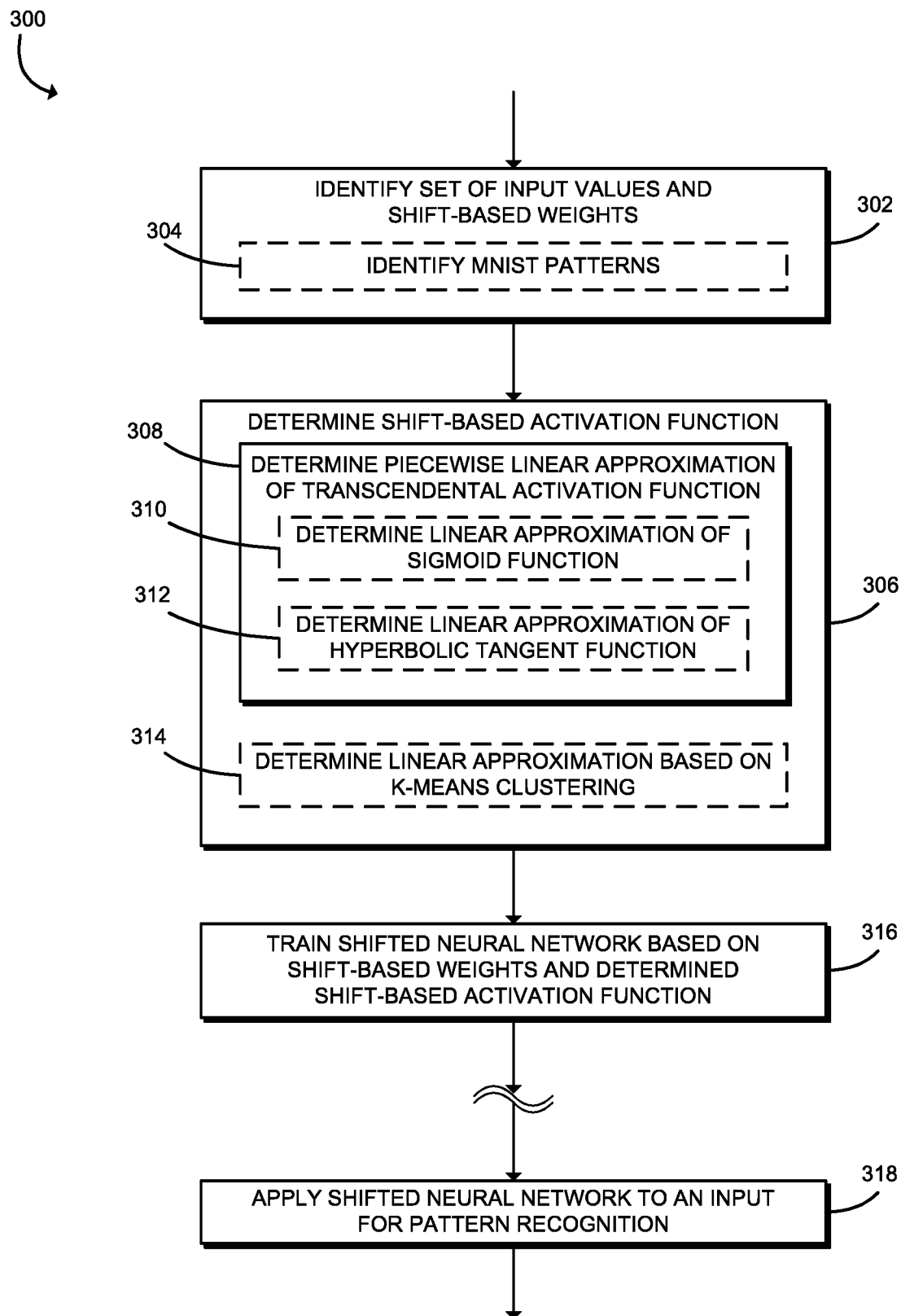


FIG. 3

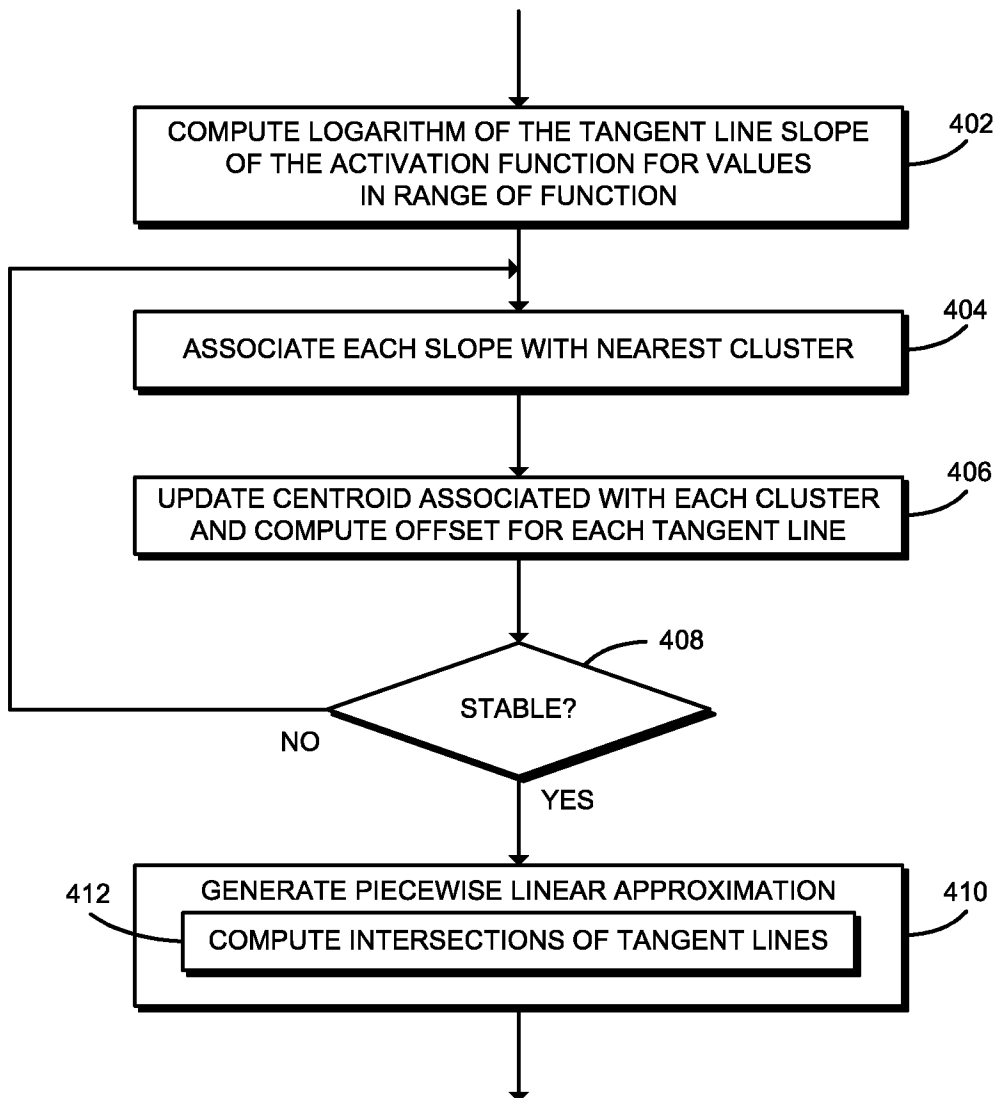
400


FIG. 4

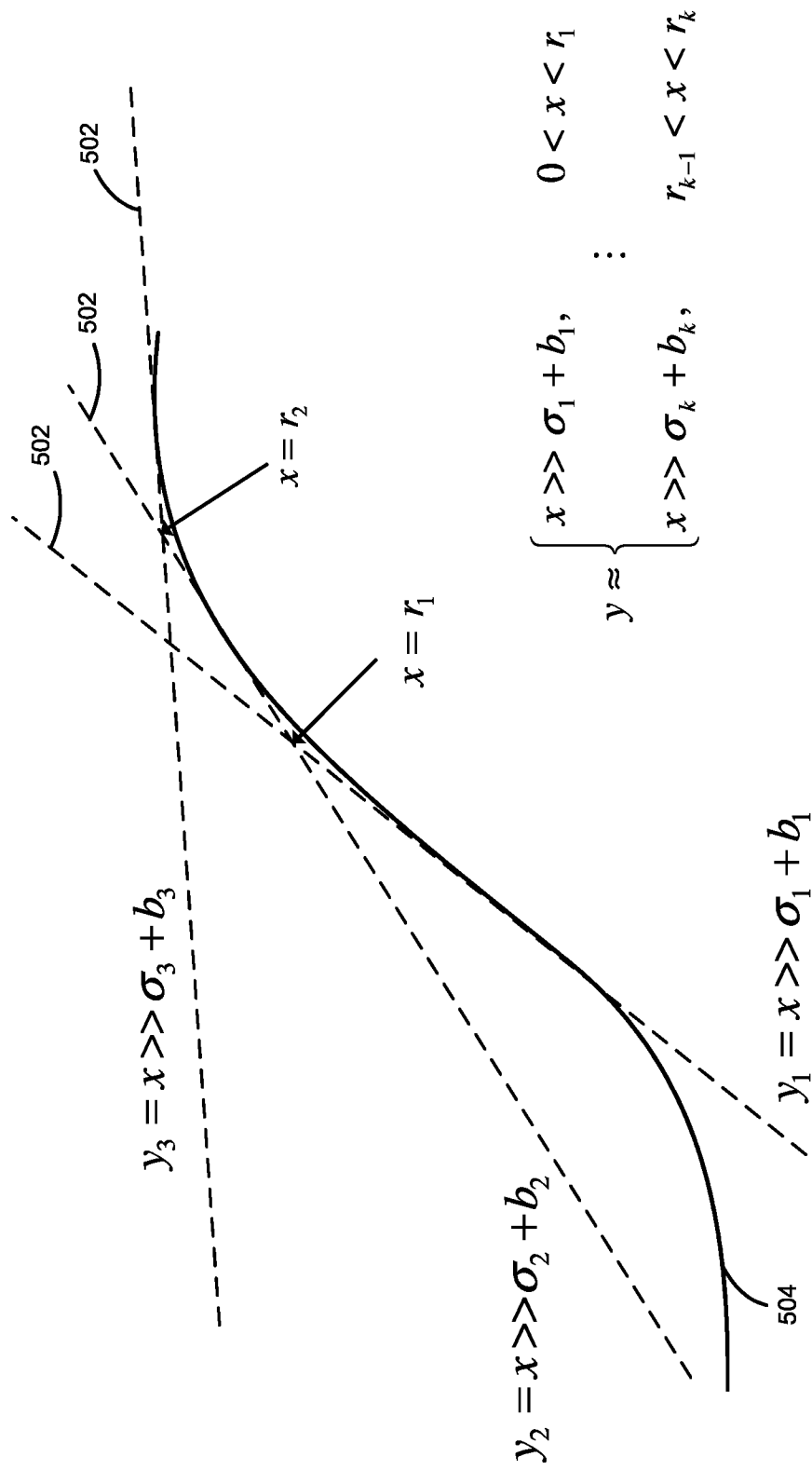


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2017/014101**A. CLASSIFICATION OF SUBJECT MATTER****G06N 3/08(2006.01)i, G06N 3/04(2006.01)i, G06N 99/00(2010.01)i**

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

B. FIELDS SEARCHEDMinimum documentation searched (classification system followed by classification symbols)
G06N 3/08; G06N 3/063; H04N 19/136; G06N 3/04; G06K 9/62; G06F 15/18; G06N 99/00Documentation 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 modelsElectronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & keywords: shifted neural network, activation function, piecewise linear approximation, training, weight, recognition, k-means clustering, and similar terms.**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCHUYLER ELDRIDGE et al., 'Towards General-Purpose Neural Network Computing', In: 2015 International Conference on Parallel Architecture and Compilation, 21 October 2015, pp. 99-112 <URL: http://ieeexplore.ieee.org/document/7429298 > See pages 99, 104, 106-107.	1-7, 12-20, 24-25
A		8-11, 21-23
Y	US 2007-0094176 A1 (JAMES HOWARD GOODNIGHT et al.) 26 April 2007 See paragraphs [0021]-[0025].	1-7, 12-20, 24-25
A	US 2015-0347870 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 03 December 2015 See paragraph [0078]; claim 1; and figure 11.	1-25
A	US 2015-0106310 A1 (UNIVERSITY OF TENNESSEE RESEARCH FOUNDATION) 16 April 2015 See claims 1, 44.	1-25
A	US 2015-0036920 A1 (FUJITSU LIMITED) 05 February 2015 See paragraphs [0082]-[0093]; claim 1; and figures 13-14.	1-25



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

"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

04 May 2017 (04.05.2017)

Date of mailing of the international search report

04 May 2017 (04.05.2017)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea



Facsimile No. +82-42-481-8578

Authorized officer

BYUN, Sung Cheal

Telephone No. +82-42-481-8262



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/014101

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007-0094176 A1	26/04/2007	US 2003-0014378 A1 US 2006-0010089 A1 US 6941289 B2 US 7162461 B2 US 7340440 B2	16/01/2003 12/01/2006 06/09/2005 09/01/2007 04/03/2008
US 2015-0347870 A1	03/12/2015	US 2016-0232430 A1 US 9373058 B2 US 9536179 B2	11/08/2016 21/06/2016 03/01/2017
US 2015-0106310 A1	16/04/2015	US 2015-0106306 A1 US 2015-0106311 A1 US 2015-0106314 A1 US 2015-0106315 A1 US 2015-0106316 A1	16/04/2015 16/04/2015 16/04/2015 16/04/2015 16/04/2015
US 2015-0036920 A1	05/02/2015	CN 104346622 A EP 2833295 A2 JP 2015-032308 A	11/02/2015 04/02/2015 16/02/2015