(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 January 2019 (17.01.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/012527 A1

(51) International Patent Classification:
G06F 15/18 (2006.01)

(21) International Application Number:
PCT/IL2018/050746

(22) International Filing Date:
09 July 2018 (09.07.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/530,215 09 July 2017 (09.07.2017) US

(71) Applicant: **CORTICA LTD.** [IL/IL]; 103 Allenby St, 4th Floor, 6513443 Tel Aviv (IL).

(72) Inventors: **RAICHELGAUZ, Igal**; 2 HaMaccabi St., 6329302 Tel Aviv (IL). **ODINAEV, Karina**; 2 HaMaccabi St., 6329302 Tel Aviv (IL). **ZEEVI, Yehoshua, Y.**; 36 Downs St., 3434909 Haifa (IL).

(74) Agent: **SHALEV, Asaf** et al.; Shalev, Jencmen & Co., 24 Hanagar St., 4527713 Hod Hasharon (IL).

(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, 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, 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, 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,

(54) Title: DEEP LEARNING NETWORKS ORCHESTRATION

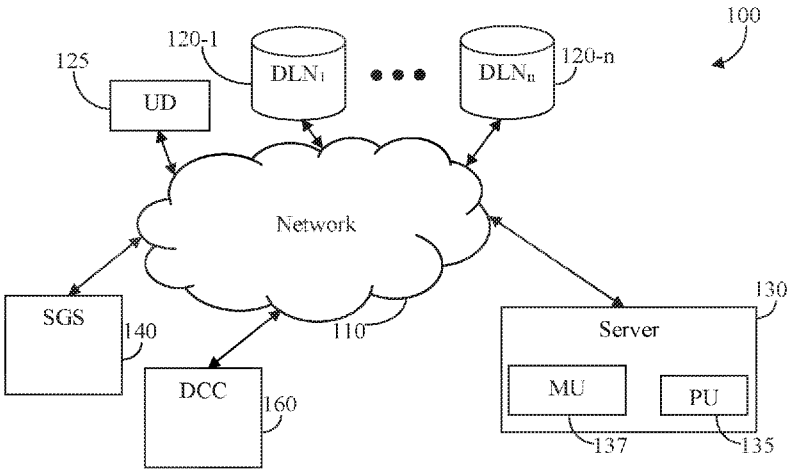

FIGURE 1

(57) Abstract: A method for responding to a query is implemented on at least one computing device and includes: receiving at least one query from a user device; determining a context for the at least one query, selecting at least one deep learning network (DLN) of a plurality of DLNs to process the at least one query, where the selecting is based at least on matching the context to the at least one DLN, sending at least a representation of the at least one query and the context to the at least one DLN, receiving at least one response to the at least one query from the at least one DLN, and sending the at least one response to the user device.

WO 2019/012527 A1

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

DEEP LEARNING NETWORKS ORCHESTRATION

RELATED APPLICATION INFORMATION

[0001] The present application claims the benefit of priority from US Provisional Patent Application, serial number 62/530215, filed on July 9, 2017 which is incorporated herein in its entirety.

TECHNICAL FIELD

[0002] The present disclosure generally relates to the orchestration of a plurality of deep learning networks coupled in an adaptively reconfigurable grid.

BACKGROUND

[0003] The reduction of multiple symbols arranged in a pattern (intentionally or seemingly randomly) to a smaller number of manageable symbols that are easily recognizable is known in the art. For example, in music, a sequence of notes may be combined into two or more notes to form a chord that is played, or otherwise heard as if being played simultaneously. Chords tend to be repetitive in nature such that the plurality of notes may be represented by a single chord symbol, thereby reducing the number of notes explicitly represented in a musical score. Accordingly, the chord, "C7", is interpreted as the root note A, the minor third C, and a perfect fifth E to be played generally simultaneously.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] The subject matter that is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the disclosure will be apparent from the following detailed description taken in conjunction with the accompanying drawings.

[0005] Figure 1 is a schematic diagram of a deep learning system, constructed and operative in accordance with embodiments described herein;

[0006] Figure 2 is a schematic diagram of a deep learning networks' reconfigurable grid architecture, constructed and operative in accordance with embodiments described herein; and

[0007] Figure 3 is a flowchart of a process for optimally responding to a query via a deep learning networks' exchange platform according to embodiments described herein.

DETAILED DESCRIPTION

Overview

[0008] A method for responding to a query is implemented on at least one computing device and includes: receiving at least one query from a user device; determining a context for the at least one query, selecting at least one deep learning network (DLN) of a plurality of DLNs to process the at least one query, where the selecting is based at least on matching the context to the at least one DLN, sending at least a representation of the at least one query and the context to the at least one DLN, receiving at least one response to the at least one query from the at least one DLN, and sending the at least one response to the user device.

Description of embodiments

[0009] It will be appreciated that patterns of data are typically distributed unevenly in a given population of data. Some patterns may be more prominent than others and are therefore likely to have a larger number of occurrences, while other patterns may be comparatively rare. In addition, some patterns may be correlated to each other, and together form pattern-combinations which may also be very common. This may be problematic for pattern recognition systems. For example, to retrieve a similarity measurement between two content-segments, it may not be enough to consider the number of corresponding patterns; the probability of occurrence for each pattern may be of importance as well. Furthermore, correlations between patterns may also be of importance. For example, if two patterns always appear together, it may be more efficient to consider them to be a single pattern.

[0010] It will be appreciated that the issues discussed hereinabove may negatively impact on the scalability and the accuracy of pattern-recognition systems. For example, in a large system where the handling of different patterns is typically spread among multiple resources (e.g., "machines") of the pattern-recognition system, machines configured to process "less-popular" patterns may remain largely inactive, whereas machines processing "popular" patterns, may be overloaded. It also may not be possible

to distribute the handling of patterns according to their a-priory probability without knowledge of the correlations between the patterns. Furthermore, to scale up a pattern-recognition system in an efficient manner it may be beneficial to avoid duplication of the pattern-space and the need to store copies of each of the known patterns in each machine.

[0011] Reference is now made to Fig. 1 which is an exemplary and non-limiting schematic diagram of a deep learning system 100 in accordance with embodiments described herein. System 100 may be configured to provide responses to queries. The queries may include sensory inputs, such as, for example, audio elements, visual elements, etc. The audio and visual elements may be provided, for example, as multimedia content elements (MMCEs), e.g., images, graphics, video streams, video clips, audio streams, audio clips, video frames, photographs, images of signals (e.g., spectrograms, phasograms, scalograms, etc.), and/or combinations thereof and portions thereof.

[0012] System 100 comprises a network 110, a plurality of deep learning networks (DLNs) 120, and a query server 130. Network 110 may be used to communicate between different parts of system 100, and may be implemented using the Internet, the world-wide-web (WWW), a local area network (LAN), a wide area network (WAN), a metro area network (MAN), and/or any othe network(s) capable of enabling communication between the elements of the system 100.

[0013] Deep learning networks (DLNs) 120-1 through 120-n may represent a plurality of networks providing deep learning services to query server 130. It will be appreciated that deep learning as referred to herein is an application of learning tasks of artificial neural networks (ANNs) that contain a plurality of layers by a computing device. To date, deep learning has been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation and bioinformatics, producing results comparable to, and in some cases superior to, human experts. Deep learning typically uses a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from a previous layer as input. The algorithms

may be supervised or unsupervised and applications may include both unsupervised pattern analysis and supervised classification. Deep learning further enables learning of multiple levels of representations that correspond to different levels of abstraction. The levels identified in such manner may represent a hierarchy of concepts.

[0014] Query server 130 comprises a processing unit (PU) 135 and a memory unit (MU) 137. Processing unit 135 may be instantiated as processing circuitry comprising one or more hardware logic components and circuits. For example, and without limitation, illustrative types of hardware logic components that may be used include field programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), application-specific standard products (ASSPs), system-on-a-chip systems (SOCs), general-purpose microprocessors, microcontrollers, digital signal processors (DSPs), and the like, and/or any other hardware logic components capable of performing calculations and/or other manipulations of information. In accordance with embodiments described herein, processing unit 135 may be implemented as an array of at least partially statistically independent computational cores. The properties of each computational core may be set independently of those of each other core, as described further hereinbelow. Memory unit 137 contains therein a query application instantiated as instructions that when executed by processing unit 135 configures query server 130 to perform as further described herein.

[0015] Query server 130 also includes a network interface (not shown) for connectivity to DLNs 120 via network 110. In accordance with embodiments described herein, query server 130 is configured to manage DLNs 120 and to optimize their use in providing responses to queries. For example, query server 130 may generate queries and send them to one or more DLNs 120. Query server 130 may then analyze responses and associated metadata received from the DLNs 120 to improve the use and configuration of DLNs 120 with respect to specific types of queries.

[0016] As depicted in Fig. 1, user device (UD) 125 is further coupled to the network 110. User device 125 may be, for example, a personal computer (PC), a personal digital assistant (PDA), a mobile phone, a smart phone, a tablet computer, an electronic wearable device (e.g., glasses, a watch, etc.), a smart television, or another wired or

mobile appliance equipped with browsing, viewing, capturing, storing, listening, filtering, and managing capabilities enabled as further discussed herein below. It will be appreciated that for the sake of simplicity only one user device 125 is depicted in Fig. 1. However, the embodiments described herein may also support a plurality of user devices 125 that may communicate with query server 130 via network 110.

[0017] Each such user device 125 may comprise a software application (not shown) installed thereon to be executed by processing circuitry (not shown). The software application (app) may be downloaded from an application repository, such as the AppStore®, Google Play®, or any other similar repositories hosting software applications. Alternatively, the application may be pre-installed in user device 125. In accordance with some embodiments described herein, the application may be located on a remote server (e.g., in a cloud, or otherwise accessible via local or wide area network) and accessed by a web-browser application on user device 125. User device 125 may employ the application to send queries via network 110 to be analyzed by one or more of DLNs 120.

[0018] In accordance with some embodiments described herein, system 100 may also comprise a data warehouse (not shown) that is configured to store metadata associated with DLNs 120. The data warehouse may also be further configured to store queries received from user device(s) 125 and associated responses received thereof from DLNs 120. Per the exemplary embodiment of Fig. 1, query server 130 may communicate with the data warehouse through network 110. Such communication may be subject to an approval to be received from the user device 125.

[0019] In the exemplary embodiment of Fig. 1, system 100 may also comprise a signature generator system (SGS) 140 and a deep-content classification (DCC) system 160 which may be utilized by query server 130 to perform various functions as described herein. SGS 140 and DCC system 160 may be connected to query server 130 either directly or through network 110. Alternatively, DCC system 160 and/or SGS 140 may be embedded in query server 130. In accordance with one embodiment, query server 130 may also be configured with, or at least in communication with, an array of computational cores configured as discussed in more detail hereinbelow.

[0020] According to an embodiment described herein, query server 130 may be configured to receive at least one query from user device 125 over network 110. Query server 130 may analyze a query received from user device 125 to determine an optimal, or at least a preferred DLN 120 to process the query. In accordance with some embodiments described herein, query server 130 may invoke SGS 140 to generate at least one signature to be associated with the query. In accordance with some embodiments described herein, the process employed by SGS 140 to generate the signature(s) may employ lossless compression of at least part of the element(s) of the query, thereby rendering the generated signature(s) robust to noise and distortions.

[0021] DCC system 160 may comprise a database of query identifiers and query classifications. Query server 130 may use the signature as a query identifier to search for a similar query in DCC system 160. Based on this search, DCC system 160 may return a query classification to query server 130. For example, a signature representing a man with a basketball may be classified as representing "basketball."

[0022] Alternatively, or in addition, the signature may be used to search for a concept structure (or concept) in DCC system 160. A concept is a collection of signatures representing elements of the unstructured data and metadata describing the concept. As a non-limiting example, a 'Superman concept' is a signature-reduced cluster of signatures describing elements (such as multimedia elements) related to, e.g., a Superman cartoon: a set of metadata representing proving textual representation of the Superman concept. Techniques for generating concept structures are also described in US patent 8,266,185 (hereinafter '185) to Raichelgauz et al., which is assigned to common assignee, and is incorporated hereby by reference for all that it contains.

[0023] For example, user device 125 may provide one or more MMCEs as a query to query server 130, e.g., an image of a basketball player. Query server 130 may invoke SGS 140 to generate at least one signature for each of the MMCEs in the query. The signature(s) may then be used to search DCC system 160 to identify an associated concept. The metadata in the identified concept may be used to identify a context for each of the plurality of MMCEs using each of the generated signatures. For example, for the image of basketball player, the concept may be "basketball", and the keyword

"basketball" may be included in the metadata for the concept. If the query also includes a second image, e.g., the logo for the National Basketball Association (NBA), the second concept may be "NBA", and the keyword "NBA" may be included in the concept's metadata. In such an example, the context may be derived from the two concepts, yielding "NBA basketball." An exemplary technique for determining a context of multimedia elements based on signatures is described in detail in US Patent Application No. 13/770,603, filed on February 19, 2013, assigned to common assignee, which is hereby incorporated by reference for all the useful information it contains.

[0024] It will be appreciated that the embodiments described herein are not necessarily limited to the use of signatures to determine a query's context. In accordance with some embodiments, the query may include text in addition to, or instead of, one or more images. For example, the query may include the names of basketball players from the NBA. The names may be used to search DCC system 160 to identify the associated concept, e.g., "NBA basketball."

[0025] It will be appreciated that DLNs 120 may not be configured identically. Some deep learning models have comparative advantages vis-à-vis other deep learning models for given subjects of interest. For example, some deep learning models may be observed to provide better results for face detection, whereas other deep learning models may be observed to provide better results for facial recognition (given a detected face). And even among deep learning models that provide better results for facial recognition, there may be differences in the quality of results based on ethnicity and/or other factors. It will similarly be appreciated that for practical reasons (e.g., cost/efficiency, resource availability, etc.) the DLNs 120 may be implemented with different levels of computing resources, e.g., RAM, CPU, bandwidth, etc. DLNs 120 may use large amounts of reference data; it may not be practical or efficient to store the entire universe of relevant data on each machine. Furthermore, at any given time, based on previous assignments of queries to perform, the different DLNs 120 may have different levels of resources available to perform additional queries. It will be appreciated that in operation there may be other factors differentiating between the suitability of the DLNs 120 to perform a given query.

[0026] In accordance with embodiments described herein, query server 130 may comprise a list of the various DLNs 120 that may include ratings for different contexts, tasks, and processing capabilities. The list and ratings may be based on a pre-processing analysis of actual performance and/or manual input. The ratings may represent the suitability of a given DLN 120 to perform a query in a given context in terms of processing speed and/or accuracy. Query server 130 may also track current workloads for each DLN 120 based on, for example, queries assigned to a given DLN 120 for which a response has not yet been received. Query server 130 may calculate a current processing load for each DLN 120 as a function of current workload and computing resources.

[0027] Query server 130 may be configured to use at least the identified context to select at least one DLN of the plurality of DLNs 120 that optimally serves the query in light of the relevant comparative advantages (e.g. per the rating) and current workload as described hereinabove. For example, for a query based on an image of an unknown girl, query server 130 may select a DLN 120 based on its rating for facial recognition. For a crowd scene, query server 130 may split the query into two stages: for the first stage a DLN 120 may be selected based on its rating for face detection; for the second stage a DLN 120 may be selected based on its rating for facial recognition, where the faces detected by the first DLN 120 may be provided (either directly or via query server 130) for further analysis to the second DLN 120. For a crowd scene with an identified context of "Hong Kong," the DLN 120 selected for facial recognition may be selected based on a higher rating for facial recognition among people of Asian ancestry.

[0028] Depending upon the configuration of system 100, the selected DLN 120 may return a response for the query to Query server 130 which may in turn forward the response to user device 125. Alternatively, or in addition, the selected DLN 120 may return the response directly to user device 125. In a case where there may be a plurality of outputs (e.g., from more than one DLN 120), query server 130 may cluster the outputs to a single complex output in order to optimally serve the query response. Alternatively, the plurality of outputs may be prioritized and formatted individually by query server 130 before providing the results to user device 125.

[0029] Fig. 2 depicts an exemplary and non-limiting schematic diagram of deep learning networks' reconfigurable grid architecture 200 according to an embodiment. An interface 210 is operative to receive requests from query server 130 (Fig. 1) for analyzing at least one query. The requests may include metadata associated with the query generated by query server 130 based on an analysis of the query. The metadata may include, for example, a selection of one or more DLNs (labelled herein as DLNs 230), a selection of type of DLNs 230, signatures associated with the query, concepts and/or contexts associated with the query, etc.

[0030] Thereafter, the request may be forwarded to a management unit (MU) 220 that is configured to navigate the request throughout the plurality of DLNs 230 (only one labeled in Fig. 2 for the sake of simplicity). Each DLN 230 comprises a plurality of layers (L_1 etc.) therein. Thereafter, an output 240 (only one labeled in Fig. 2 for the sake of simplicity) may be generated by the one or more selected DLNs 230 and sent to MU 220 It will be appreciated that MU 220 may be depicted twice in Fig. 2 for the sake of simplicity in representation of the flow architecture 200). MU 220 is operative to generate response 250 based on the one or more outputs. Interface 210 is operative to return response 250 to query server 130 and/or directly to user device 125.

[0031] Fig. 3 depicts an exemplary and non-limiting flowchart 300 describing an operation of an exemplary method performed by query server 130 (Fig. 1) for optimally orchestrating deep learning system 100 (Fig. 1) in response to a query. Query server 130 (Fig. 1) may receive (step 310) at least one query as input from user device 125 (Fig. 1). A query may include at least one sensory input, such as, for example, an audio and/or visual multimedia content element. The multimedia content element may be, for example, an image, a graphic, a video stream, a video clip, an audio stream, an audio clip, a video frame, a photograph, and an image of signals (e.g., spectrograms, phasograms, scalograms, etc.), and/or combinations thereof and portions thereof.

[0032] Query server 130 may determine (step 320) a context for the query. In accordance with some embodiments described herein, step 320 may comprise query server 130 generating (step 322) a signature for at least one multimedia content element in the query; using the generated signature(s) to search (step 324) DCC 160 (Fig. 1) for

a concept associated with the at least one multimedia content element; and determining (step 326) a context based on the concept(s) associated with each of the multimedia content elements in the query. Alternatively, or in addition, step 320 may comprise looking up one or more keywords from the query in DCC 160 to determine the context.

[0033] Based at least on the determined context, query server 130 may select (step 330) one or more DLNs 120 (Fig. 1) to perform the query. In accordance with some embodiments described herein, step 330 may comprise query server 130 selecting (step 332) one or more DLNs 120 based on their associated ratings for processing queries with the determined context(s). Query server 130 may then check for sufficient processing capacity (step 334) on the selected DLN(s), e.g., according to current processing load. If one or more DLNs 120 does not have sufficient processing capacity, control may return to step 332 where a different DLN 120 may be selected, e.g., the DLN 120 with the next highest rating for the determined context. Otherwise processing may continue to step 340.

[0034] Query server 130 may send (step 340) the query to the selected DLN(s) 120, and subsequently receive (step 350) the query response(s) from the selected DLN(s) 120 after the query is processed. It will be appreciated that depending on the configuration of system 100, the query as sent to DLN (s) 120 may not necessarily be identical to the query as received from user device 125, but rather a representation of the original query. For example, the query sent in step 340 may include one or more MMCEs from the query received from user device 125, one or more signatures derived from the MMCE(s), and/or the context(s) as determined in step 326. If necessary, query server 130 may combine (step 360) multiple responses into a single combined response. Query server 130 may then return (step 370) the (combined) response(s) to user device 125.

[0035] In accordance with some embodiments described herein, query server 130 may be configured to adjust the ratings for DLNs 120 in accordance with the results of process 300. For example, if the actual response to receive a query response from a given DLN 120 is slower/faster than anticipated as per its current rating for the associated context, query server 130 may adjust the rating accordingly. Similarly, in

some implementations, query server 130 may be configured with a feedback mechanism to receive feedback from user device 125 regarding the usefulness of the query responses received in process 300. Query server 130 may also be configured to adjust an associated rating in accordance with the feedback.

[0036] It is important to note that the embodiments disclosed herein are only examples of the many advantageous uses of the teachings herein. In general, statements made in the specification of the present application do not necessarily limit any of the various claimed inventions. Moreover, some statements may apply to some inventive features but not to others. In general, unless otherwise indicated, singular elements may be in plural and vice versa with no loss of generality. In the drawings, like numerals refer to like parts through several views.

[0037] The various embodiments disclosed herein may be implemented as hardware, firmware, software, or any combination thereof. Moreover, the software is preferably implemented as an application program tangibly embodied on a program storage unit or computer readable medium consisting of parts, or of certain devices and/or a combination of devices.

[0038] The application program may be uploaded to, and executed by, a machine comprising any suitable architecture. Preferably, the machine is implemented on a computer platform having hardware such as one or more central processing units ("CPUs"), a memory, and input/output interfaces. The computer platform may also include an operating system and microinstruction code.

[0039] The various processes and functions described herein may be either part of the microinstruction code or part of the application program, or any combination thereof, which may be executed by a CPU, whether or not such a computer or processor is explicitly shown. In addition, various other peripheral units may be connected to the computer platform such as an additional data storage unit and a printing unit. Furthermore, a non-transitory computer readable medium is any computer readable medium except for a transitory propagating signal.

[0040] All examples and conditional language recited herein are intended for pedagogical purposes to aid the reader in understanding the disclosed embodiments and

the concepts contributed by the inventor to furthering the art, and are to be construed as being without limitation to such specifically recited examples and conditions.

[0041] Moreover, all statements herein reciting principles, aspects, and embodiments of the invention, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents as well as equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure.

What is claimed is:

1. A method for responding to a query, the method implemented on at least one computing device and comprising:

receiving at least one query from a user device;

determining a context for said at least one query;

selecting at least one deep learning network (DLN) of a plurality of DLNs to process said at least one query, wherein said selecting is based at least on matching said context to said at least one DLN;

sending at least a representation of said at least one query and said context to said at least one DLN;

receiving at least one response to said at least one query from said at least one DLN; and

sending said at least one response to said user device.

2. The method according to claim 1 wherein said at least one query includes at least one multimedia content element (MMCE), and the method further comprises:

using at least one identifier associated with said at least one MMCE to search a deep content classification system for at least one concept associated with said at least one MMCE; and

determining said context according to metadata associated with said at least one concept.

3. The method according to claim 2 wherein:

said at least one identifier is at least one signature derived from said at least one MMCE; and

said concept is a collection of associated signatures and metadata.

4. The method according to claim 2 wherein said at least one identifier is metadata associated with said MMCE.

5. The method according to claim 1 wherein said at least one query includes at least text, and the method further comprises:

using said at least text to search a deep content classification system for at least one concept associated with said at least text; and

determining said context according to metadata associated with said at least one concept.

6. The method according to claim 1 further comprising:

tracking a processing load for each of said DLNs, wherein said selecting is further based on said processing load.

7. The method according to claim 1 wherein said at least one DLN is at least two DLNs.

8. The method according to claim 7 wherein said receiving at least one response comprises:

receiving said at least one response from each of said at least two DLNs; and

combining said at least one response from each of said at least two DLNs into a combined response, wherein said sending said at least one response comprises sending said combined response to said user device.

9. The method according to claim 7 wherein said receiving at least one response comprises:

receiving a first response from one of said at least two DLNs;

sending at least said first response as part of said at least one query to another DLN from said at least two DLNs; and

receiving a second response from said another DLN, wherein said sending said at least one response comprises sending said second response to said user device.

10. The method according to claim 1 further comprising rating each of said plurality of DLNs for performance of queries associated with a plurality of contexts, wherein:

said context is determined from among said plurality of contexts; and

said selecting is further based at least on said rating.

11. A deep learning system comprising:

a plurality of deep learning networks (DLNS) instantiated on at least one computing device;

a query server instantiated on a computing and operative to:

receive at least one query from a user device,

determine a context for said at least one query,

select at least one DLN of said plurality of DLNs to process said at least one query, wherein said query server is operative to select said at least one based at least on matching said context to said at least one DLN,

send at least a representation of said at least one query and said context to said at least one DLN,

receiving at least one response to said at least one query from said at least one DLN, and

send said at least one response to said user device.

12. The method according to claim 11 wherein said query server is further operative to:

receive at least one multimedia content element as at least part of said query;

use at least one identifier associated with said at least one MMCE to search a deep content classification system for at least one concept associated with said at least one MMCE; and

determine said context according to metadata associated with said at least one concept.

13. The method according to claim 12 wherein:

said at least one identifier is at least one signature derived from said at least one MMCE; and

said concept is a collection of associated signatures and metadata.

14. The method according to claim 12 wherein said at least one identifier is metadata associated with said MMCE.

15. The method according to claim 11 wherein said at least one query includes at least text, and the method further comprises:

using said at least text to search a deep content classification system for at least one concept associated with said at least text; and

determining said context according to metadata associated with said at least one concept.

16. The method according to claim 11 further comprising:

tracking a processing load for each of said DLNs, wherein said selecting is further based on said processing load.

17. The method according to claim 11 wherein said at least one DLN is at least two DLNs, and said query server is further configured to:

receive said at least one response from each of said at least two DLNs; and

combine said at least one response from each of said at least two DLNs into a combined response; and

send said combined response to said user device.

18. The method according to claim 11 wherein said at least one DLN is at least two DLNs, and said query server is further configured to:

receive a first response from one of said at least two DLNs;

send at least said first response as part of said at least one query to another DLN from said at least two DLNs;

receive a second response from said another DLN; and

send said second response to said user device.

19. The method according to claim 11 wherein said query server is further operative to:

select said at least one DLN based at least on a rating, wherein said rating represents a suitability for each of said plurality of DLNs to perform queries associated with each of said plurality of contexts; and

adjust said suitability rating for an associated DLN for a given context based at least on performance of said associated DLN in providing said at least one response for said given context, wherein said given context is from said plurality of contexts.

20. A query response system instantiated on at least one computing device and comprising:

means for receiving at least one query from a use device;

means for determining a context for said at least one query;

means for selecting at least one deep learning network (DLN) of a plurality of DLNs to process said at least one query, wherein said selecting is based at least on matching said context to said at least one DLN;

means for sending said at least one query and said context to said at least one DLN;

means for receiving at least one response to said at least one query from said at least one DLN; and

means for sending said at least one response to said user device.

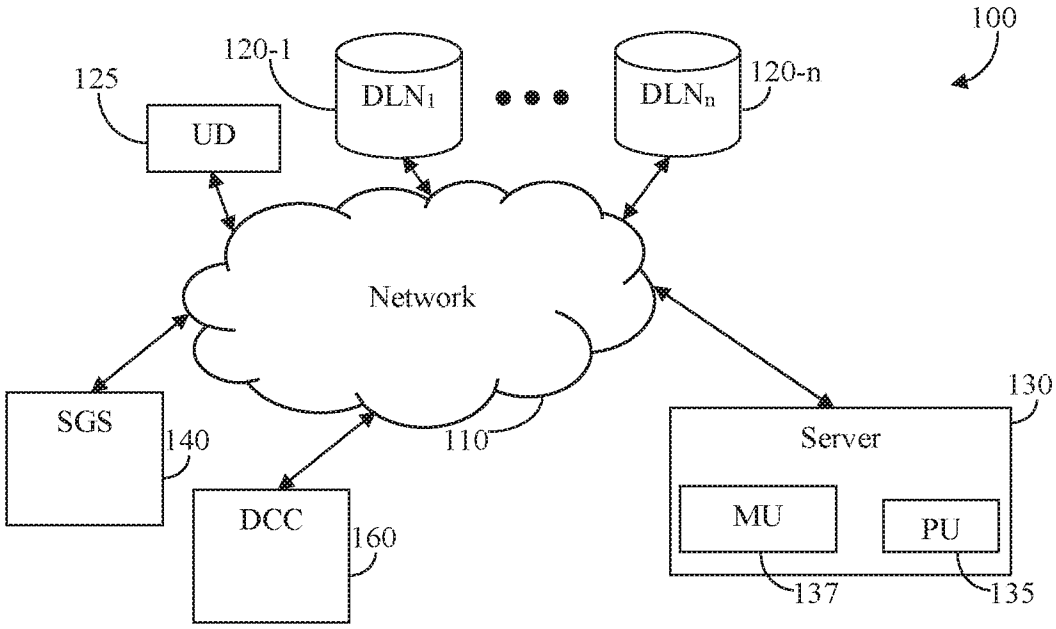
100
120-1
125
DLN1
DLNn
120-n
UD
Network
SGS
140
110
130
Server
MU
PU
DCC
160
137
135

FIGURE 1

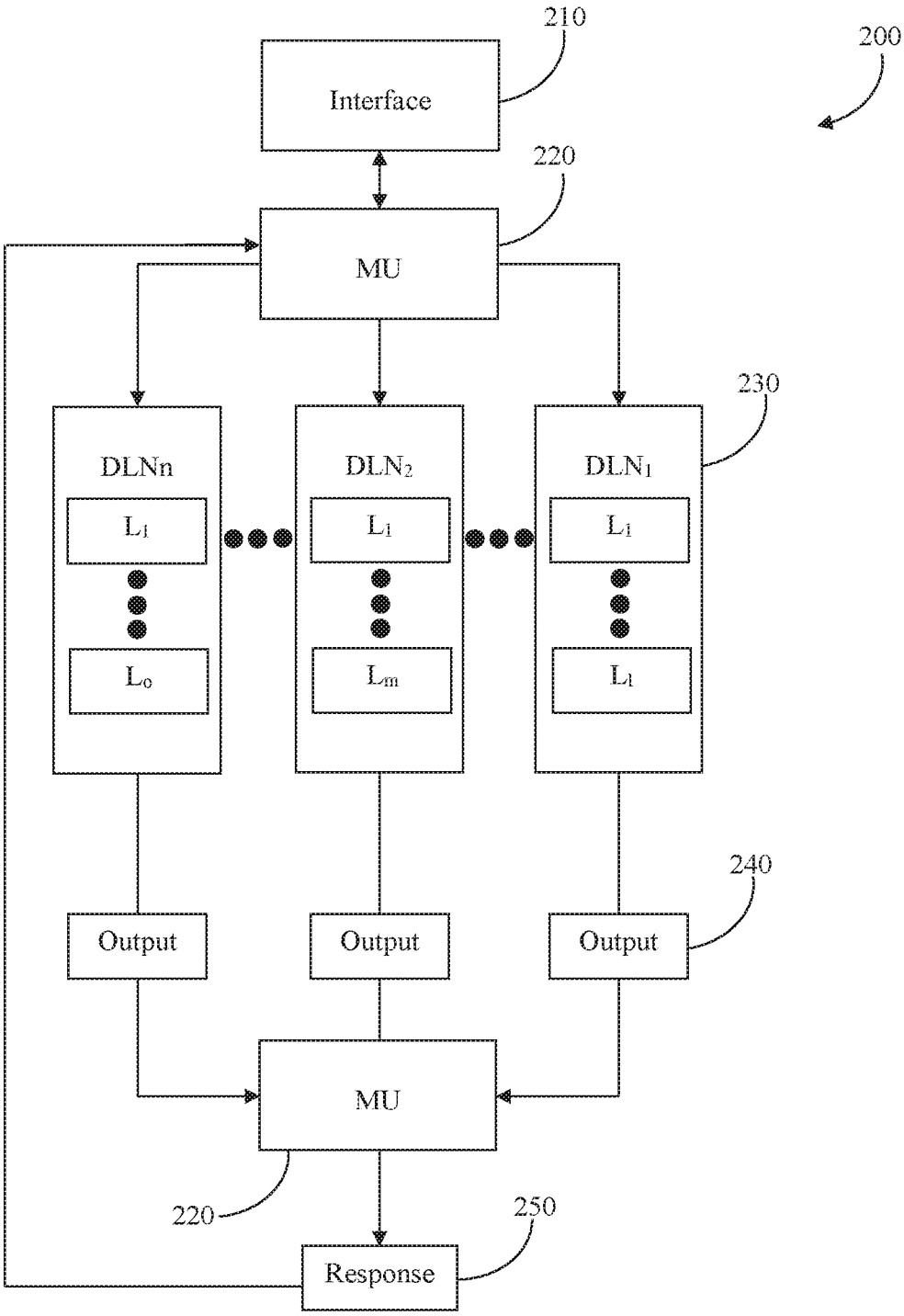
200
210
Interface
220
MU
230
DLNn
DLN2
DLN1
L1
L1
L1
Lo
Lm
L1
240
Output
Output
Output
MU
220
250
Response

FIGURE 2

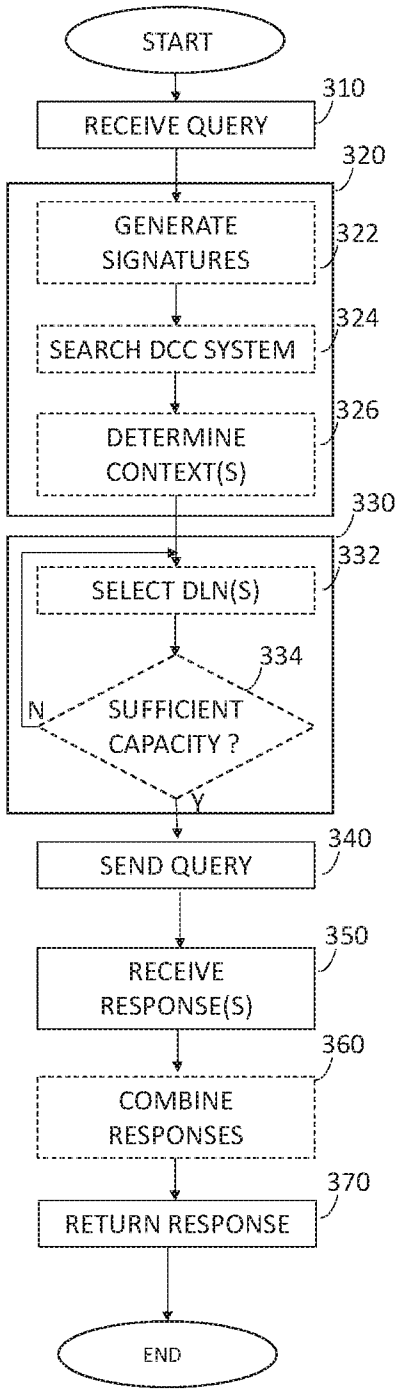
START
300
310
RECEIVE QUERY
320
GENERATE SIGNATURES
322
SEARCH DCC SYSTEM
324
DETERMINE CONTEXT(S)
326
330
SELECT DLN(S)
332
334
N
SUFFICIENT CAPACITY ?
Y
SEND QUERY
340
RECEIVE RESPONSE(S)
350
COMBINE RESPONSES
360
RETURN RESPONSE
370
END

FIGURE 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IL 18/50746

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 15/18 (2018.01)
CPC - G06N 3/08, G06N 99/005, G06N 3/0454, G06N 3/0427, G06K 9/6269

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)
See Search History Document

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2015/0293976 A1 (MICROSOFT CORPORATION) 15 October 2015 (15.10.2015), entire document, especially abstract, para [0030], [0032], [0034]-[0035], [0040]-[0042], [0048], [0053]-[0055], [0059], [0072]-[0073], [0116].	1-20
Y	US 2003/0158839 A1 (FAYBISHENKO ET AL.) 21 August 2003 (21.08.2003), entire document, especially abstract, para [0048], [0051]-[0052], [0071]-[0072], [0074], [0090]-[0091], [0099]-[0100], [0117], [0146]-[0147], [0158], [0162], [0178], [0240].	1-20
A	US 2017/0132510 A1 (FACEBOOK, INC.) 11 May 2017 (11.05.2017), entire document, especially abstract.	1-20

☐ 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	Date of mailing of the international search report
16 September 2018 (16.09.2018)	11 OCT 2018
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774