

(51) International Patent Classification:
G06F 17/30 (2006.01)(21) International Application Number:
PCT/CN2017/077280(22) International Filing Date:
20 March 2017 (20.03.2017)

(25) Filing Language: English

(26) Publication Language: English

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

(72) Inventors; and

(71) Applicants (for BZ only): **SU, Zhou** [CN/CN]; Intel Lab
China, 2 Kexueyuan S Rd #9, Haidian, Beijing, 100190
(CN). **LI, Jianguo** [CN/CN]; Room 1104, Bldg. 901-Wu,
Zhong Guan Cun, Haidian District, Beijing, 100080 (CN).
YAO, Anbang [CN/CN]; 8F, Raycom Infotech Park A,
No. 2 Kexueyuan South Road, Zhong Guan Cun, HaiDian
District, Beijing, 100090 (CN). **CHEN, Yurong** [CN/CN];
Room 503, Unit 1, Building 4, Qing Feng Hua Jing Yuan,
Haidian District, Beijing, 100085 (CN).(74) Agent: **NTD PATENT AND TRADEMARK AGENCY
LIMITED**; 10th Floor, Block A, Investment Plaza, 27 Jin-
rongdajie, Xicheng District, Beijing 100033 (CN).(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, 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).

(54) Title: TOPIC-GUIDED MODEL FOR IMAGE CAPTIONING SYSTEM

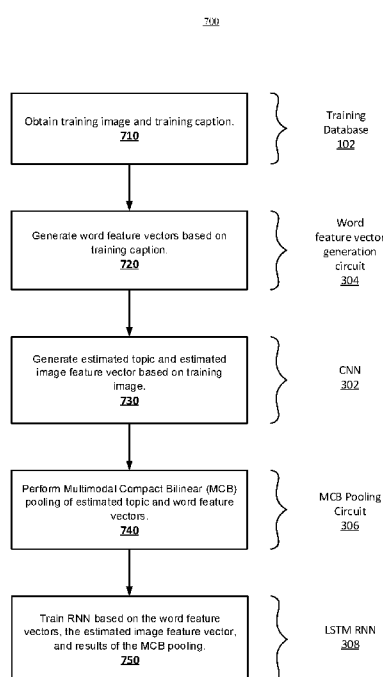


FIG. 7

(57) Abstract: Techniques are provided for training and operation of a topic-guided image captioning system. A methodology implementing the techniques includes generating image feature vectors, for an image to be captioned, based on application of a convolutional neural network (CNN) to the image. The method further includes generating the caption based on application of a recurrent neural network (RNN) to the image feature vectors. The RNN is configured as a long short-term memory (LSTM) RNN. The method further includes training the LSTM RNN with training images and associated training captions. The training is based on a combination of: feature vectors of the training image; feature vectors of the associated training caption; and a multimodal compact bilinear (MCB) pooling of the training caption feature vectors and an estimated topic of the training image. The estimated topic is generated by an application of the CNN to the training image.

Published:

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

TOPIC-GUIDED MODEL FOR IMAGE CAPTIONING SYSTEM

BACKGROUND

[0001] Image captioning systems attempt to generate descriptions (e.g., captions) for an image, based on the content of that image. Existing image captioning systems typically employ computer vision techniques to obtain details about the image and natural language processing techniques to describe the image content. These systems, however, do not always provide satisfactory results, particularly when language ambiguities arise with respect to the image content, or when relatively complex images include many components and characteristics that associate with different words.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] Features and advantages of embodiments of the claimed subject matter will become apparent as the following Detailed Description proceeds, and upon reference to the Drawings, wherein like numerals depict like parts.

[0003] Figure 1 is a top-level block diagram of an image captioning system, configured in accordance with certain embodiments of the present disclosure.

[0004] Figure 2 illustrates examples of images and associated captions, in accordance with certain embodiments of the present disclosure.

[0005] Figure 3 is a block diagram of a topic-guided model training system, configured in accordance with certain embodiments of the present disclosure.

[0006] Figure 4 is a flow diagram for training of a recurrent neural network (RNN), in accordance with certain embodiments of the present disclosure.

[0007] Figure 5 is a block diagram of a topic-guided image captioning system, configured in accordance with certain embodiments of the present disclosure.

[0008] Figure 6 is a flow diagram of the operation of an RNN system, in accordance with certain embodiments of the present disclosure.

[0009] Figure 7 is a flowchart illustrating a methodology for training of a topic-guided image captioning system, in accordance with certain embodiments of the present disclosure.

[0010] Figure 8 is a flowchart illustrating a methodology for topic-guided image captioning, in accordance with certain embodiments of the present disclosure.

[0011] Figure 9 is a block diagram schematically illustrating a platform to perform topic-guided image captioning, configured in accordance with certain embodiments of the present disclosure.

[0012] Although the following Detailed Description will proceed with reference being made to illustrative embodiments, many alternatives, modifications, and variations thereof will be apparent in light of this disclosure.

DETAILED DESCRIPTION

[0013] Generally, this disclosure provides techniques for training and operation of a topic-guided image captioning system. The topic-guided captioning system identifies and exploits connections between visual features and language features during system training using training images and associated training captions. Estimated topics are calculated from the training images and feature vectors are calculated from the training captions, which are then combined in a multimodal compact bilinear (MCB) pooling to train a recurrent neural network (RNN), as will be described in greater detail below. The combination of image topic-related guidance along with language features (e.g., the caption words) allows for improved caption generation that includes not only objects categories, but also object attributes such as shapes, colors, and relationships between objects and activities within the image. Thus, the different semantic meanings of words can be considered when captioning a given image, thereby allowing for disambiguation of complex imagery.

[0014] The disclosed techniques can be implemented, for example, in a computing system or a software product executable or otherwise controllable by such systems, although other embodiments will be apparent. The system or product is configured to perform topic-guided captioning on an image. In accordance with an embodiment, a methodology to implement these techniques includes generating image feature vectors, for an image to be captioned, based on application of a convolutional neural network (CNN) to the image. The method further includes

generating the caption based on application of a recurrent neural network (RNN) to the image feature vectors. In some example embodiments, the RNN is configured as a “long short-term memory” (LSTM) RNN, as will be explained in greater detail below. Other natural language processing techniques can be used as well, as will be appreciated in light of this disclosure. The method further includes training the LSTM RNN with training images and associated training captions. The training is based on a combination of: feature vectors calculated from the training image; feature vectors calculated from the associated training caption; and an MCB pooling of the training caption feature vectors and an estimated topic of the training image. The estimated topic is generated by an application of the CNN to the training image.

[0015] As will be appreciated, the techniques described herein may allow for improved image captioning, compared to existing methods that fail to exploit the connections between visual features and language features. The disclosed techniques can be implemented on a broad range of platforms including laptops, tablets, smart phones, workstations, and imaging devices. These techniques may further be implemented in hardware or software or a combination thereof.

[0016] Figure 1 is a top-level block diagram of an image captioning system 100, configured in accordance with certain embodiments of the present disclosure. The image captioning system 100 is shown to include a topic-guided model training system 104, a topic-guided image captioning system 106, and a training database 102. The training database 102 is configured to store training images of known content along with predetermined training captions. Each training caption is associated with a training image. For example, one training image may depict a woman playing tennis, and the associated training caption would be “a woman is playing tennis on a tennis court.” In some embodiments, other suitable mechanisms may be employed to provide pairs of training images and training captions.

[0017] The topic-guided model training system 104 is configured to employ the training images and associated captions from the training database to train the topic-guided image captioning system 106, as will be explained in greater detail below. The resulting trained topic-guided image captioning system 106 is then configured to process images 110 (e.g., of unknown content), and generate captions 120 for those images, as will also be explained below.

[0018] Figure 2 illustrates examples of images and associated captions, in accordance with certain embodiments of the present disclosure. For example, image 202 may be associated with

a caption 204 of “bridge across a river.” As another example, image 206 may be associated with a caption 208 of “a dog resting on a snow bank in front of a railing.” These example images and captions can serve to illustrate training image/caption pairings from database 102, or unknown images 110 and generated captions 120. It will be appreciated, of course, that any image can be used and that such images may include any combination of people, objects, activities, and backgrounds, etc.

[0019] Figure 3 is a block diagram of a topic-guided model training system 104, configured in accordance with certain embodiments of the present disclosure. The topic-guided model training system 104 is shown to include a convolutional neural network (CNN) 302, a word feature vector generation circuit 304, a multimodal compact bilinear pooling (MCB) circuit 306, and a long short-term memory (LSTM) recurrent neural network (RNN) 308.

[0020] The convolutional neural network (CNN) 302 is configured to generate an estimated topic T 314, associated with each training image 310. The CNN 302 is further configured to generate an estimated image feature vector V 316, associated with the training image. The topic T provides a relatively higher level abstraction of the image (e.g., semantic information) compared to the visual features of V , and helps to reduce noise and overfitting problems that may result from reliance on visual features alone. In some embodiments, the CNN may be initialized based on a pretrained Visual Geometry Group (VGG) neural network model, or based on other known techniques in light of the present disclosure.

[0021] The word feature vector generation circuit 304 is configured to generate word feature vectors $W(t)$ 318 based on the training caption 312 associated with the training image 310. In some embodiments, a training caption that comprises N words may generate N word feature vectors $W(t)$, where $t=0, 1 \dots N-1$. Word feature vectors may be generated using any known techniques, in light of the present disclosure.

[0022] The multimodal compact bilinear pooling (MCB) circuit 306 is configured to combine the estimated topic T 314 with the word feature vectors $W(t)$ 318. In some embodiments, the MCB circuit 306 further comprises a Fourier transform circuit and an inverse Fourier transform circuit. The Fourier transform circuit is configured to perform a Fourier transform on the estimated topic T and a Fourier transform on the word feature vectors $W(t)$. The inverse Fourier transform circuit is configured to perform an inverse Fourier transform on the multiplicative

product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors, as illustrated, for example, by the equation:

$$M(t) = FFT^{-1}(FFT(T) \cdot FFT(W(t)))$$

where $M(t)$ 320 is the resulting MCB combination vector of topic T and feature vector of word t . This multiplication in the frequency domain, which is equivalent to convolution in the time domain, provides a joint representation of image topics and language words that is superior to a simpler element-wise addition or multiplication operation. The combination vector $M(t)$, thus provides an improved representation of the potentially diverse meanings of a word when used to describe different images, compared to traditional techniques that represent each word as a single feature vector.

[0023] The long short-term memory (LSTM) recurrent neural network (RNN) 308 is configured to generate image captions, once operational (e.g., after training). The LSTM RNN 308 is trained based on the word feature vectors $W(t)$ 318, the estimated image feature vector V 316, and the MCB pooling combination of the estimated topic and the word feature vectors $M(t)$ 320, as will be explained in greater detail below in connection with Figure 4.

[0024] Figure 4 is a flow diagram 400 for training of the LSTM RNN 308, in accordance with certain embodiments of the present disclosure. The LSTM RNN 308 is shown to include a number of LSTM neural network units or nodes 308a, 308b, 308c ... 308n. The first LSTM unit 308a is configured to process the estimated image feature vector V 316 and generate initial state information $h(0)$ 412, to be propagated to the second LSTM unit 308b. The second LSTM unit 308b is configured to process a concatenation of $M(0)$ 320 and $W(0)$ 318 in conjunction with the state information $h(0)$ 412 to generate an estimated first word 406 (and associated probability) of the caption for the training image. This process continues with each subsequent LSTM unit 308c, ...308n, applied to each following pair of $M(t)$ and $W(t)$, along with propagated state information $h(t)$, to generate estimates and probabilities 408, 410 for subsequent words of the caption. The state information $h(t)$ serves as a memory of previous words of the caption sentence. For example, if the caption is “A woman is playing baseball,” it is preferable to generate “playing” based on “a woman is” rather than just the word “is” in isolation.

[0025] An error measurement 402 is performed between the estimated caption words 406, 408, 410 and the actual words of the training caption 312 (e.g., truth data). The error measurement is

fed back 404 to the LSTM units for training adaptation using known techniques, in light of the present disclosure. The training process 400 is repeated for each pair of training image and training caption from the training database 102.

[0026] Figure 5 is a block diagram of a topic-guided image captioning system 106, configured in accordance with certain embodiments of the present disclosure. The topic-guided image captioning system 106 is shown to include a trained convolutional neural network (CNN) 502 and a trained LSTM RNN 508. In some embodiments, the training has been accomplished as previously described in connection with Figures 3 and 4.

[0027] The trained CNN 502 is configured to generate estimated image feature vectors V 516, for an image to be captioned 110, based on application of a convolutional neural network to the image, using known techniques in light of the present disclosure.

[0028] The trained LSTM RNN 508 configured to generate an estimated caption 120, based on application of a recurrent neural network to the estimated image feature vectors V 516, as described below in connection with Figure 6.

[0029] In some embodiments, the generated caption 120 may be provided to one or more applications or systems, such as, for example, a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, or a system configured to aid the visually impaired.

[0030] Figure 6 is a flow diagram 600 of the operation of an LSTM RNN system, in accordance with certain embodiments of the present disclosure. The trained LSTM RNN 508 is shown to include a number of LSTM neural network units or nodes 508a, 508b, 508c ... 508m, 508n, which may operate based on known techniques in light of the present disclosure. The first LSTM unit 508a is configured to process the estimated image feature vector V 516 and generate initial state information $h(0)$ 612, to be propagated to the second LSTM unit 508b. The second LSTM unit 508b is configured to process the state information $h(0)$ 612 to generate an estimated first word 604 of the caption for the unknown image 110. This process continues with each subsequent LSTM unit 508c, ... 508n, applied to generate additional words 606, ... 608, for the caption based on propagated state information $h(t)$ 614, ... 616 until the end of the caption

sentence 610. The estimated words 604, ... 608, are combined to form the generated caption estimate 120.

Methodology

[0031] Figures 7 and 8 are flowcharts illustrating example methods for training 700 and operation of 800, a topic-guided image captioning system, in accordance with certain embodiments of the present disclosure. As can be seen, the example methods include a number of phases and sub-processes, the sequence of which may vary from one embodiment to another. However, when considered in the aggregate, these phases and sub-processes form a process for image captioning in accordance with certain of the embodiments disclosed herein. These embodiments can be implemented, for example using the system architecture illustrated in Figures 3-6 as described above. However other system architectures can be used in other embodiments, as will be apparent in light of this disclosure. To this end, the correlation of the various functions shown in Figures 7 and 8 to the specific components illustrated in the other figures is not intended to imply any structural and/or use limitations. Rather, other embodiments may include, for example, varying degrees of integration wherein multiple functionalities are effectively performed by one system. For example, in an alternative embodiment a single module can be used to perform all of the functions of methods 700 and 800. Thus, other embodiments may have fewer or more modules and/or sub-modules depending on the granularity of implementation. In still other embodiments, the methodology depicted can be implemented as a computer program product including one or more non-transitory machine readable mediums that when executed by one or more processors cause the methodology to be carried out. Numerous variations and alternative configurations will be apparent in light of this disclosure.

[0032] As illustrated in Figure 7, in an embodiment, method 700 for training of a topic-guided image captioning system commences by obtaining, at operation 710, one or more training images and associated training captions for each image. In some embodiments, the training images and associated training captions may be obtained from a training database or other suitable source.

[0033] Next, at operation 720, word feature vectors are generated based on the training caption.

[0034] At operation 730, a convolutional neural network (CNN) is applied to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image.

[0035] At operation 740, a multimodal compact bilinear (MCB) pooling is performed on the estimated topic and the word feature vectors. In some embodiments, the MCB pooling includes performing a Fourier transform on the estimated topic and on the word feature vectors; and then performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors, to generate the pooling results.

[0036] At operation 750, a recurrent neural network (RNN) is trained based on the word feature vectors, the estimated image feature vector, and the results of the MCB pooling. In some embodiments, the RNN is configured as a long short-term memory (LSTM) RNN.

[0037] As illustrated in Figure 8, in an embodiment, method 800 for topic-guided image captioning commences by generating, at operation 810, image feature vectors for an image to be captioned. In some embodiments, the image feature vectors are generated by application of a convolutional neural network (CNN) to the image.

[0038] Next, at operation 820, a caption is generated for the image, based on application of a recurrent neural network (RNN) to the image feature vectors. In some embodiments, the RNN is configured as a long short-term memory (LSTM) RNN, and the RNN is trained as described previously in connection with the description of Figure 7.

[0039] Of course, in some embodiments, additional operations may be performed, as previously described in connection with the system. For example, the generated caption may be provided to one or more applications or systems, such as, for example, a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, or a system configured to aid the visually impaired.

Example System

[0040] Figure 9 illustrates an example system 900 to perform topic-guided image captioning, configured in accordance with certain embodiments of the present disclosure. In some

embodiments, system 900 comprises an image captioning platform 910 which may host, or otherwise be incorporated into a personal computer, workstation, server system, laptop computer, ultra-laptop computer, tablet, touchpad, portable computer, handheld computer, palmtop computer, personal digital assistant (PDA), cellular telephone, combination cellular telephone and PDA, smart device (for example, smartphone or smart tablet), mobile internet device (MID), messaging device, data communication device, imaging device, and so forth. Any combination of different devices may be used in certain embodiments.

[0041] In some embodiments, platform 910 may comprise any combination of a processor 920, a memory 930, topic-guided image captioning system 106, topic-guided model training system 104, a network interface 940, an input/output (I/O) system 950, a user interface 960, an imaging source 962, and a storage system 970. As can be further seen, a bus and/or interconnect 992 is also provided to allow for communication between the various components listed above and/or other components not shown. Platform 910 can be coupled to a network 994 through network interface 940 to allow for communications with other computing devices, platforms, or resources. Other componentry and functionality not reflected in the block diagram of Figure 9 will be apparent in light of this disclosure, and it will be appreciated that other embodiments are not limited to any particular hardware configuration.

[0042] Processor 920 can be any suitable processor, and may include one or more coprocessors or controllers, such as an audio processor or a graphics processing unit to assist in control and processing operations associated with system 900. In some embodiments, the processor 920 may be implemented as any number of processor cores. The processor (or processor cores) may be any type of processor, such as, for example, a micro-processor, an embedded processor, a digital signal processor (DSP), a graphics processor (GPU), a network processor, a field programmable gate array or other device configured to execute code. The processors may be multithreaded cores in that they may include more than one hardware thread context (or “logical processor”) per core. Processor 920 may be implemented as a complex instruction set computer (CISC) or a reduced instruction set computer (RISC) processor. In some embodiments, processor 920 may be configured as an x86 instruction set compatible processor.

[0043] Memory 930 can be implemented using any suitable type of digital storage including, for example, flash memory and/or random access memory (RAM). In some embodiments, the memory 930 may include various layers of memory hierarchy and/or memory caches as are

known to those of skill in the art. Memory 930 may be implemented as a volatile memory device such as, but not limited to, a RAM, dynamic RAM (DRAM), or static RAM (SRAM) device. Storage system 970 may be implemented as a non-volatile storage device such as, but not limited to, one or more of a hard disk drive (HDD), a solid-state drive (SSD), a universal serial bus (USB) drive, an optical disk drive, tape drive, an internal storage device, an attached storage device, flash memory, battery backed-up synchronous DRAM (SDRAM), and/or a network accessible storage device. In some embodiments, storage 970 may comprise technology to increase the storage performance enhanced protection for valuable digital media when multiple hard drives are included.

[0044] Processor 920 may be configured to execute an Operating System (OS) 980 which may comprise any suitable operating system, such as Google Android (Google Inc., Mountain View, CA), Microsoft Windows (Microsoft Corp., Redmond, WA), Apple OS X (Apple Inc., Cupertino, CA), Linux, or a real-time operating system (RTOS). As will be appreciated in light of this disclosure, the techniques provided herein can be implemented without regard to the particular operating system provided in conjunction with system 900, and therefore may also be implemented using any suitable existing or subsequently-developed platform.

[0045] Network interface circuit 940 can be any appropriate network chip or chipset which allows for wired and/or wireless connection between other components of computer system 900 and/or network 994, thereby enabling system 900 to communicate with other local and/or remote computing systems, servers, cloud-based servers, and/or other resources. Wired communication may conform to existing (or yet to be developed) standards, such as, for example, Ethernet. Wireless communication may conform to existing (or yet to be developed) standards, such as, for example, cellular communications including LTE (Long Term Evolution), Wireless Fidelity (Wi-Fi), Bluetooth, and/or Near Field Communication (NFC). Exemplary wireless networks include, but are not limited to, wireless local area networks, wireless personal area networks, wireless metropolitan area networks, cellular networks, and satellite networks.

[0046] I/O system 950 may be configured to interface between various I/O devices and other components of computer system 900. I/O devices may include, but not be limited to, user interface 960, and an imaging source 962. User interface 960 may include devices (not shown) such as a display element, touchpad, keyboard, mouse, microphone, and speaker, etc. Imaging source 962 may be a camera, a scanner, a database of images, or any other suitable source. I/O

system 950 may include a graphics subsystem configured to perform processing of images for rendering on a display element. Graphics subsystem may be a graphics processing unit or a visual processing unit (VPU), for example. An analog or digital interface may be used to communicatively couple graphics subsystem and the display element. For example, the interface may be any of a high definition multimedia interface (HDMI), DisplayPort, wireless HDMI, and/or any other suitable interface using wireless high definition compliant techniques. In some embodiments, the graphics subsystem could be integrated into processor 920 or any chipset of platform 910.

[0047] It will be appreciated that in some embodiments, the various components of the system 900 may be combined or integrated in a system-on-a-chip (SoC) architecture. In some embodiments, the components may be hardware components, firmware components, software components or any suitable combination of hardware, firmware or software.

[0048] Topic-guided image captioning system 106 is configured to automatically generate captions for images, based on application of a convolutional neural network and a recurrent neural network (configured as a long short-term memory RNN), as described previously. Topic-guided model training system 104 is configured to train the image captioning system 106 based on a combination of feature vectors of training images, feature vectors of associated training captions, and a multimodal compact bilinear pooling of the training caption feature vectors and an estimated topic of the training image, as described previously. Systems 104 and 106 may include any or all of the circuits/components illustrated in Figures 3-6, as described above. These components can be implemented or otherwise used in conjunction with a variety of suitable software and/or hardware that is coupled to or that otherwise forms a part of platform 910. These components can additionally or alternatively be implemented or otherwise used in conjunction with user I/O devices that are capable of providing information to, and receiving information and commands from, a user.

[0049] In some embodiments, these circuits may be installed local to system 900, as shown in the example embodiment of Figure 9. Alternatively, system 900 can be implemented in a client-server arrangement wherein at least some functionality associated with these circuits is provided to system 900 using an applet, such as a JavaScript applet, or other downloadable module. Such a remotely accessible module or sub-module can be provisioned in real-time, in response to a request from a client computing system for access to a given server having resources that are of

interest to the user of the client computing system. In such embodiments, the server can be local to network 994 or remotely coupled to network 994 by one or more other networks and/or communication channels. In some cases, access to resources on a given network or computing system may require credentials such as usernames, passwords, and/or compliance with any other suitable security mechanism.

[0050] In various embodiments, system 900 may be implemented as a wireless system, a wired system, or a combination of both. When implemented as a wireless system, system 900 may include components and interfaces suitable for communicating over a wireless shared media, such as one or more antennae, transmitters, receivers, transceivers, amplifiers, filters, control logic, and so forth. An example of wireless shared media may include portions of a wireless spectrum, such as the radio frequency spectrum and so forth. When implemented as a wired system, system 900 may include components and interfaces suitable for communicating over wired communications media, such as input/output adapters, physical connectors to connect the input/output adaptor with a corresponding wired communications medium, a network interface card (NIC), disc controller, video controller, audio controller, and so forth. Examples of wired communications media may include a wire, cable metal leads, printed circuit board (PCB), backplane, switch fabric, semiconductor material, twisted pair wire, coaxial cable, fiber optics, and so forth.

[0051] Various embodiments may be implemented using hardware elements, software elements, or a combination of both. Examples of hardware elements may include processors, microprocessors, circuits, circuit elements (for example, transistors, resistors, capacitors, inductors, and so forth), integrated circuits, ASICs, programmable logic devices, digital signal processors, FPGAs, logic gates, registers, semiconductor devices, chips, microchips, chipsets, and so forth. Examples of software may include software components, programs, applications, computer programs, application programs, system programs, machine programs, operating system software, middleware, firmware, software modules, routines, subroutines, functions, methods, procedures, software interfaces, application program interfaces, instruction sets, computing code, computer code, code segments, computer code segments, words, values, symbols, or any combination thereof. Determining whether an embodiment is implemented using hardware elements and/or software elements may vary in accordance with any number of factors, such as desired computational rate, power level, heat tolerances, processing cycle

budget, input data rates, output data rates, memory resources, data bus speeds, and other design or performance constraints.

[0052] Some embodiments may be described using the expression “coupled” and “connected” along with their derivatives. These terms are not intended as synonyms for each other. For example, some embodiments may be described using the terms “connected” and/or “coupled” to indicate that two or more elements are in direct physical or electrical contact with each other. The term “coupled,” however, may also mean that two or more elements are not in direct contact with each other, but yet still cooperate or interact with each other.

[0053] The various embodiments disclosed herein can be implemented in various forms of hardware, software, firmware, and/or special purpose processors. For example, in one embodiment at least one non-transitory computer readable storage medium has instructions encoded thereon that, when executed by one or more processors, cause one or more of the image captioning methodologies disclosed herein to be implemented. The instructions can be encoded using a suitable programming language, such as C, C++, object oriented C, Java, JavaScript, Visual Basic .NET, Beginner’s All-Purpose Symbolic Instruction Code (BASIC), or alternatively, using custom or proprietary instruction sets. The instructions can be provided in the form of one or more computer software applications and/or applets that are tangibly embodied on a memory device, and that can be executed by a computer having any suitable architecture. In one embodiment, the system can be hosted on a given website and implemented, for example, using JavaScript or another suitable browser-based technology. For instance, in certain embodiments, the system may leverage processing resources provided by a remote computer system accessible via network 994. In other embodiments, the functionalities disclosed herein can be incorporated into other software applications, such as image perception systems, robotics, and virtual reality applications. The computer software applications disclosed herein may include any number of different modules, sub-modules, or other components of distinct functionality, and can provide information to, or receive information from, still other components. These modules can be used, for example, to communicate with input and/or output devices such as a display screen, a touch sensitive surface, a printer, and/or any other suitable device. Other componentry and functionality not reflected in the illustrations will be apparent in light of this disclosure, and it will be appreciated that other embodiments are not limited to any particular hardware or software configuration. Thus, in other embodiments system 900 may

comprise additional, fewer, or alternative subcomponents as compared to those included in the example embodiment of Figure 9.

[0054] The aforementioned non-transitory computer readable medium may be any suitable medium for storing digital information, such as a hard drive, a server, a flash memory, and/or random access memory (RAM), or a combination of memories. In alternative embodiments, the components and/or modules disclosed herein can be implemented with hardware, including gate level logic such as a field-programmable gate array (FPGA), or alternatively, a purpose-built semiconductor such as an application-specific integrated circuit (ASIC). Still other embodiments may be implemented with a microcontroller having a number of input/output ports for receiving and outputting data, and a number of embedded routines for carrying out the various functionalities disclosed herein. It will be apparent that any suitable combination of hardware, software, and firmware can be used, and that other embodiments are not limited to any particular system architecture.

[0055] Some embodiments may be implemented, for example, using a machine readable medium or article which may store an instruction or a set of instructions that, if executed by a machine, may cause the machine to perform a method and/or operations in accordance with the embodiments. Such a machine may include, for example, any suitable processing platform, computing platform, computing device, processing device, computing system, processing system, computer, process, or the like, and may be implemented using any suitable combination of hardware and/or software. The machine readable medium or article may include, for example, any suitable type of memory unit, memory device, memory article, memory medium, storage device, storage article, storage medium, and/or storage unit, such as memory, removable or non-removable media, erasable or non-erasable media, writeable or rewriteable media, digital or analog media, hard disk, floppy disk, compact disk read only memory (CD-ROM), compact disk recordable (CD-R) memory, compact disk rewriteable (CR-RW) memory, optical disk, magnetic media, magneto-optical media, removable memory cards or disks, various types of digital versatile disk (DVD), a tape, a cassette, or the like. The instructions may include any suitable type of code, such as source code, compiled code, interpreted code, executable code, static code, dynamic code, encrypted code, and the like, implemented using any suitable high level, low level, object oriented, visual, compiled, and/or interpreted programming language.

[0056] Unless specifically stated otherwise, it may be appreciated that terms such as “processing,” “computing,” “calculating,” “determining,” or the like refer to the action and/or process of a computer or computing system, or similar electronic computing device, that manipulates and/or transforms data represented as physical quantities (for example, electronic) within the registers and/or memory units of the computer system into other data similarly represented as physical quantities within the registers, memory units, or other such information storage transmission or displays of the computer system. The embodiments are not limited in this context.

[0057] The terms “circuit” or “circuitry,” as used in any embodiment herein, are functional and may comprise, for example, singly or in any combination, hardwired circuitry, programmable circuitry such as computer processors comprising one or more individual instruction processing cores, state machine circuitry, and/or firmware that stores instructions executed by programmable circuitry. The circuitry may include a processor and/or controller configured to execute one or more instructions to perform one or more operations described herein. The instructions may be embodied as, for example, an application, software, firmware, etc. configured to cause the circuitry to perform any of the aforementioned operations. Software may be embodied as a software package, code, instructions, instruction sets and/or data recorded on a computer-readable storage device. Software may be embodied or implemented to include any number of processes, and processes, in turn, may be embodied or implemented to include any number of threads, etc., in a hierarchical fashion. Firmware may be embodied as code, instructions or instruction sets and/or data that are hard-coded (e.g., nonvolatile) in memory devices. The circuitry may, collectively or individually, be embodied as circuitry that forms part of a larger system, for example, an integrated circuit (IC), an application-specific integrated circuit (ASIC), a system-on-a-chip (SoC), desktop computers, laptop computers, tablet computers, servers, smart phones, etc. Other embodiments may be implemented as software executed by a programmable control device. In such cases, the terms “circuit” or “circuitry” are intended to include a combination of software and hardware such as a programmable control device or a processor capable of executing the software. As described herein, various embodiments may be implemented using hardware elements, software elements, or any combination thereof. Examples of hardware elements may include processors, microprocessors, circuits, circuit elements (e.g., transistors, resistors, capacitors, inductors, and so forth),

integrated circuits, application specific integrated circuits (ASIC), programmable logic devices (PLD), digital signal processors (DSP), field programmable gate array (FPGA), logic gates, registers, semiconductor device, chips, microchips, chip sets, and so forth.

[0058] Numerous specific details have been set forth herein to provide a thorough understanding of the embodiments. It will be understood by an ordinarily-skilled artisan, however, that the embodiments may be practiced without these specific details. In other instances, well known operations, components and circuits have not been described in detail so as not to obscure the embodiments. It can be appreciated that the specific structural and functional details disclosed herein may be representative and do not necessarily limit the scope of the embodiments. In addition, although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described herein. Rather, the specific features and acts described herein are disclosed as example forms of implementing the claims.

Further Example Embodiments

[0059] The following examples pertain to further embodiments, from which numerous permutations and configurations will be apparent.

[0060] Example 1 is a processor-implemented method for training an image captioning system. The method comprises: obtaining, by a processor-based system, a training image and a training caption, the training caption associated with the training image; generating, by the processor-based system, word feature vectors based on the training caption; applying, by the processor-based system, a convolutional neural network (CNN) to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image; performing, by the processor-based system, a multimodal compact bilinear (MCB) pooling of the estimated topic and the word feature vectors; and training, by the processor-based system, a recurrent neural network (RNN) based on the word feature vectors, the estimated image feature vector, and results of the MCB pooling.

[0061] Example 2 includes the subject matter of Example 1, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0062] Example 3 includes the subject matter of Examples 1 or 2, wherein the MCB pooling further comprises: performing a Fourier transform on the estimated topic; performing a Fourier transform on the word feature vectors; and performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

[0063] Example 4 includes the subject matter of any of Examples 1-3, further comprising initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

[0064] Example 5 includes the subject matter of any of Examples 1-4, wherein the training image and the associated training caption are obtained from a training database.

[0065] Example 6 includes the subject matter of any of Examples 1-5, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0066] Example 7 includes the subject matter of any of Examples 1-6, further comprising: receiving an input image; and generating a caption for the input image, using the RNN.

[0067] Example 8 is a system for training an image captioning system. The system comprises: a word feature vector generation circuit to generate word feature vectors based on a training caption, the training caption associated with a training image; a convolutional neural network (CNN) to generate an estimated topic associated with the training image, and further to generate an estimated image feature vector associated with the training image; a multimodal compact bilinear (MCB) pooling circuit to combine the estimated topic and the word feature vectors; and a recurrent neural network (RNN) to generate image captions, wherein the RNN is trained on the word feature vectors, the estimated image feature vector, and the combination resulting from the MCB pooling.

[0068] Example 9 includes the subject matter of Example 8, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0069] Example 10 includes the subject matter of Examples 8 or 9, wherein the MCB pooling circuit further comprises: a Fourier transform circuit to perform a Fourier transform on the

estimated topic and the word feature vectors; and an inverse Fourier transform circuit to perform an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

[0070] Example 11 includes the subject matter of any of Examples 8-10, wherein the CNN is initialized based on a pretrained Visual Geometry Group (VGG) neural network model.

[0071] Example 12 includes the subject matter of any of Examples 8-11, further comprising a training database to provide the training image and the associated training caption.

[0072] Example 13 includes the subject matter of any of Examples 8-12, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0073] Example 14 is at least one non-transitory computer readable storage medium having instructions encoded thereon that, when executed by one or more processors, result in the following operations for image captioning. The operations comprise: obtaining a training image and a training caption, the training caption associated with the training image; generating word feature vectors based on the training caption; applying a convolutional neural network (CNN) to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image; performing a multimodal compact bilinear (MCB) pooling of the estimated topic and the word feature vectors; and training a recurrent neural network (RNN) based on the word feature vectors, the estimated image feature vector, and results of the MCB pooling.

[0074] Example 15 includes the subject matter of Example 14, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0075] Example 16 includes the subject matter of Examples 14 or 15, wherein the MCB pooling further comprises the operations: performing a Fourier transform on the estimated topic; performing a Fourier transform on the word feature vectors; and performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

[0076] Example 17 includes the subject matter of any of Examples 14-16, further comprising the operations of initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

[0077] Example 18 includes the subject matter of any of Examples 14-17, wherein the training image and the associated training caption are obtained from a training database.

[0078] Example 19 includes the subject matter of any of Examples 14-18, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0079] Example 20 is a processor-implemented method for image captioning. The method comprises: generating, by a processor-based system, image feature vectors for an image to be captioned, based on application of a convolutional neural network (CNN) to the image; and generating, by the processor-based system, the caption, based on application of a recurrent neural network (RNN) to the image feature vectors, wherein the RNN is trained on a combination of: feature vectors of a training image; feature vectors of a training caption, the training caption associated with the training image; and a multimodal compact bilinear (MCB) pooling of the training caption feature vectors and an estimated topic of the training image feature vectors.

[0080] Example 21 includes the subject matter of Example 20, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0081] Example 22 includes the subject matter of Examples 20 or 21, wherein the training image feature vectors and the estimated topic of the training image feature vectors are generated based on application of the CNN to the training image.

[0082] Example 23 includes the subject matter of any of Examples 20-22, wherein the MCB pooling further comprises: performing a Fourier transform on the estimated topic of the training image feature vectors; performing a Fourier transform on the training caption feature vectors; and performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic of the training image feature vectors and the Fourier transform of the training caption feature vectors.

[0083] Example 24 includes the subject matter of any of Examples 20-23, further comprising initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

[0084] Example 25 includes the subject matter of any of Examples 20-24, further comprising providing the generated caption to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0085] Example 26 is a system for training an image captioning system. The system comprises: means for obtaining a training image and a training caption, the training caption associated with the training image; means for generating word feature vectors based on the training caption; means for applying a convolutional neural network (CNN) to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image; means for performing a multimodal compact bilinear (MCB) pooling of the estimated topic and the word feature vectors; and means for training a recurrent neural network (RNN) based on the word feature vectors, the estimated image feature vector, and results of the MCB pooling.

[0086] Example 27 includes the subject matter of Example 26, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0087] Example 28 includes the subject matter of Examples 26 or 27, wherein the MCB pooling further comprises: means for performing a Fourier transform on the estimated topic; means for performing a Fourier transform on the word feature vectors; and means for performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

[0088] Example 29 includes the subject matter of any of Examples 26-28, further comprising means for initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

[0089] Example 30 includes the subject matter of any of Examples 26-29, wherein the training image and the associated training caption are obtained from a training database.

[0090] Example 31 includes the subject matter of any of Examples 26-30, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0091] Example 32 includes the subject matter of any of Examples 26-31, further comprising: means for receiving an input image; and means for generating a caption for the input image, using the RNN.

[0092] Example 33 is a system for image captioning. The system comprises: means for generating image feature vectors for an image to be captioned, based on application of a convolutional neural network (CNN) to the image; and means for generating the caption, based on application of a recurrent neural network (RNN) to the image feature vectors, wherein the RNN is trained on a combination of: feature vectors of a training image; feature vectors of a training caption, the training caption associated with the training image; and a multimodal compact bilinear (MCB) pooling of the training caption feature vectors and an estimated topic of the training image feature vectors.

[0093] Example 34 includes the subject matter of Example 33, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

[0094] Example 35 includes the subject matter of Examples 33 or 34, wherein the training image feature vectors and the estimated topic of the training image feature vectors are generated based on application of the CNN to the training image.

[0095] Example 36 includes the subject matter of any of Examples 33-35, wherein the MCB pooling further comprises: means for performing a Fourier transform on the estimated topic of the training image feature vectors; means for performing a Fourier transform on the training caption feature vectors; and means for performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic of the training image feature vectors and the Fourier transform of the training caption feature vectors.

[0096] Example 37 includes the subject matter of any of Examples 33-36, further comprising means for initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

[0097] Example 38 includes the subject matter of any of Examples 33-37, further comprising means for providing the generated caption to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

[0098] The terms and expressions which have been employed herein are used as terms of description and not of limitation, and there is no intention, in the use of such terms and expressions, of excluding any equivalents of the features shown and described (or portions thereof), and it is recognized that various modifications are possible within the scope of the claims. Accordingly, the claims are intended to cover all such equivalents. Various features, aspects, and embodiments have been described herein. The features, aspects, and embodiments are susceptible to combination with one another as well as to variation and modification, as will be understood by those having skill in the art. The present disclosure should, therefore, be considered to encompass such combinations, variations, and modifications. It is intended that the scope of the present disclosure be limited not be this detailed description, but rather by the claims appended hereto. Future filed applications claiming priority to this application may claim the disclosed subject matter in a different manner, and may generally include any set of one or more elements as variously disclosed or otherwise demonstrated herein.

CLAIMS

What is claimed is:

1. A processor-implemented method for training an image captioning system, the method comprising:
 - obtaining, by a processor-based system, a training image and a training caption, the training caption associated with the training image;
 - generating, by the processor-based system, word feature vectors based on the training caption;
 - applying, by the processor-based system, a convolutional neural network (CNN) to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image;
 - performing, by the processor-based system, a multimodal compact bilinear (MCB) pooling of the estimated topic and the word feature vectors; and
 - training, by the processor-based system, a recurrent neural network (RNN) based on the word feature vectors, the estimated image feature vector, and results of the MCB pooling.
2. The method of claim 1, wherein the RNN is configured as a long short-term memory (LSTM) RNN.
3. The method of claim 1, wherein the MCB pooling further comprises:
 - performing a Fourier transform on the estimated topic;
 - performing a Fourier transform on the word feature vectors; and
 - performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.
4. The method of any of claims 1-3, further comprising initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

5. The method of any of claims 1-3, wherein the training image and the associated training caption are obtained from a training database.

6. The method of any of claims 1-3, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

7. The method of any of claims 1-3, further comprising:
receiving an input image; and
generating a caption for the input image, using the RNN.

8. A system for training an image captioning system, the system comprising:
a word feature vector generation circuit to generate word feature vectors based on a training caption, the training caption associated with a training image;
a convolutional neural network (CNN) to generate an estimated topic associated with the training image, and further to generate an estimated image feature vector associated with the training image;
a multimodal compact bilinear (MCB) pooling circuit to combine the estimated topic and the word feature vectors; and
a recurrent neural network (RNN) to generate image captions, wherein the RNN is trained on the word feature vectors, the estimated image feature vector, and the combination resulting from the MCB pooling.

9. The system of claim 8, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

10. The system of claim 8, wherein the MCB pooling circuit further comprises:
a Fourier transform circuit to perform a Fourier transform on the estimated topic and the word feature vectors; and

an inverse Fourier transform circuit to perform an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

11. The system of any of claims 8-10, wherein the CNN is initialized based on a pretrained Visual Geometry Group (VGG) neural network model.

12. The system of any of claims 8-10, further comprising a training database to provide the training image and the associated training caption.

13. The system of any of claims 8-10, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

14. At least one non-transitory computer readable storage medium having instructions encoded thereon that, when executed by one or more processors, result in the following operations for image captioning, the operations comprising:

obtaining a training image and a training caption, the training caption associated with the training image;

generating word feature vectors based on the training caption;

applying a convolutional neural network (CNN) to the training image, to generate an estimated topic associated with the training image, and to generate an estimated image feature vector associated with the training image;

performing a multimodal compact bilinear (MCB) pooling of the estimated topic and the word feature vectors; and

training a recurrent neural network (RNN) based on the word feature vectors, the estimated image feature vector, and results of the MCB pooling.

15. The computer readable storage medium of claim 14, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

16. The computer readable storage medium of claim 14, wherein the MCB pooling further comprises the operations:

performing a Fourier transform on the estimated topic;
performing a Fourier transform on the word feature vectors; and
performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic and the Fourier transform of the word feature vectors.

17. The computer readable storage medium of any of claims 14-16, further comprising the operations of initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

18. The computer readable storage medium of any of claims 14-16, wherein the training image and the associated training caption are obtained from a training database.

19. The computer readable storage medium of any of claims 14-16, wherein the image captioning system is trained to generate a caption to be provided to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

20. A processor-implemented method for image captioning, the method comprising:
generating, by a processor-based system, image feature vectors for an image to be captioned, based on application of a convolutional neural network (CNN) to the image; and

generating, by the processor-based system, the caption, based on application of a recurrent neural network (RNN) to the image feature vectors,

wherein the RNN is trained on a combination of: feature vectors of a training image; feature vectors of a training caption, the training caption associated with the training image; and a multimodal compact bilinear (MCB) pooling of the training caption feature vectors and an estimated topic of the training image feature vectors.

21. The method of claim 20, wherein the RNN is configured as a long short-term memory (LSTM) RNN.

22. The method of claim 20, wherein the training image feature vectors and the estimated topic of the training image feature vectors are generated based on application of the CNN to the training image.

23. The method of claim 20, wherein the MCB pooling further comprises:
performing a Fourier transform on the estimated topic of the training image feature vectors;
performing a Fourier transform on the training caption feature vectors; and
performing an inverse Fourier transform on the multiplicative product of the Fourier transform of the estimated topic of the training image feature vectors and the Fourier transform of the training caption feature vectors.

24. The method of any of claims 20-23, further comprising initializing the CNN based on a pretrained Visual Geometry Group (VGG) neural network model.

25. The method of any of claims 20-23, further comprising providing the generated caption to at least one of a photo-captioning application of a social media platform or service, a human-robot interaction application, a content-based image retrieval application, a gaming application, an entertainment application, and a system configured to aid visually impaired people.

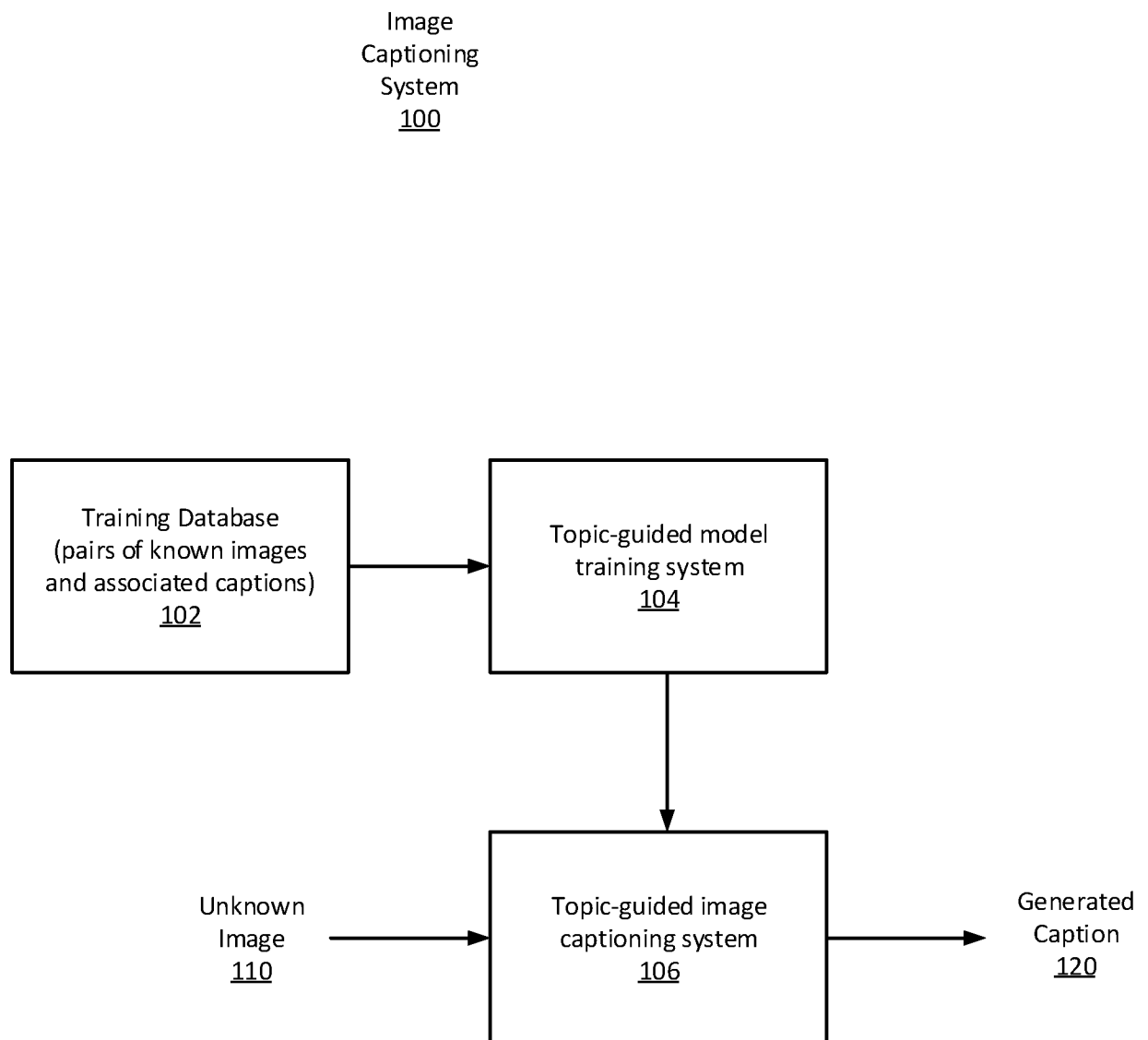


FIG. 1

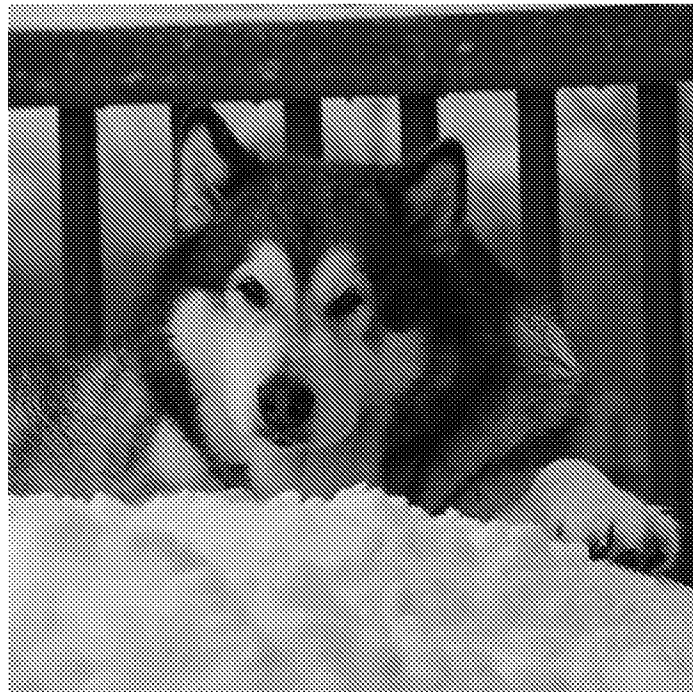
Image
202



Caption
204

Bridge across a river

Image
206



Caption
208

A dog resting on a snow bank
in front of a railing

FIG. 2

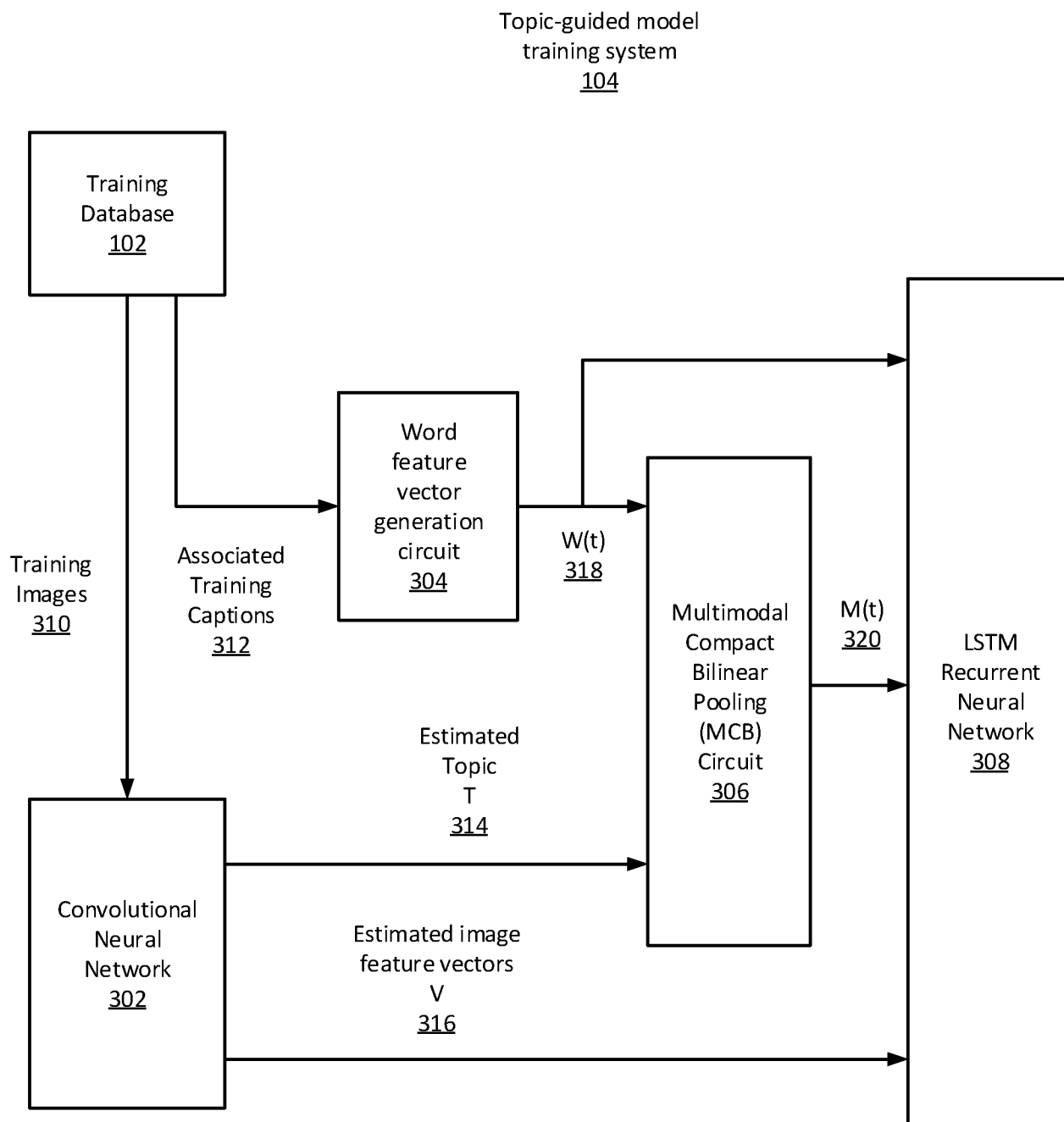


FIG. 3

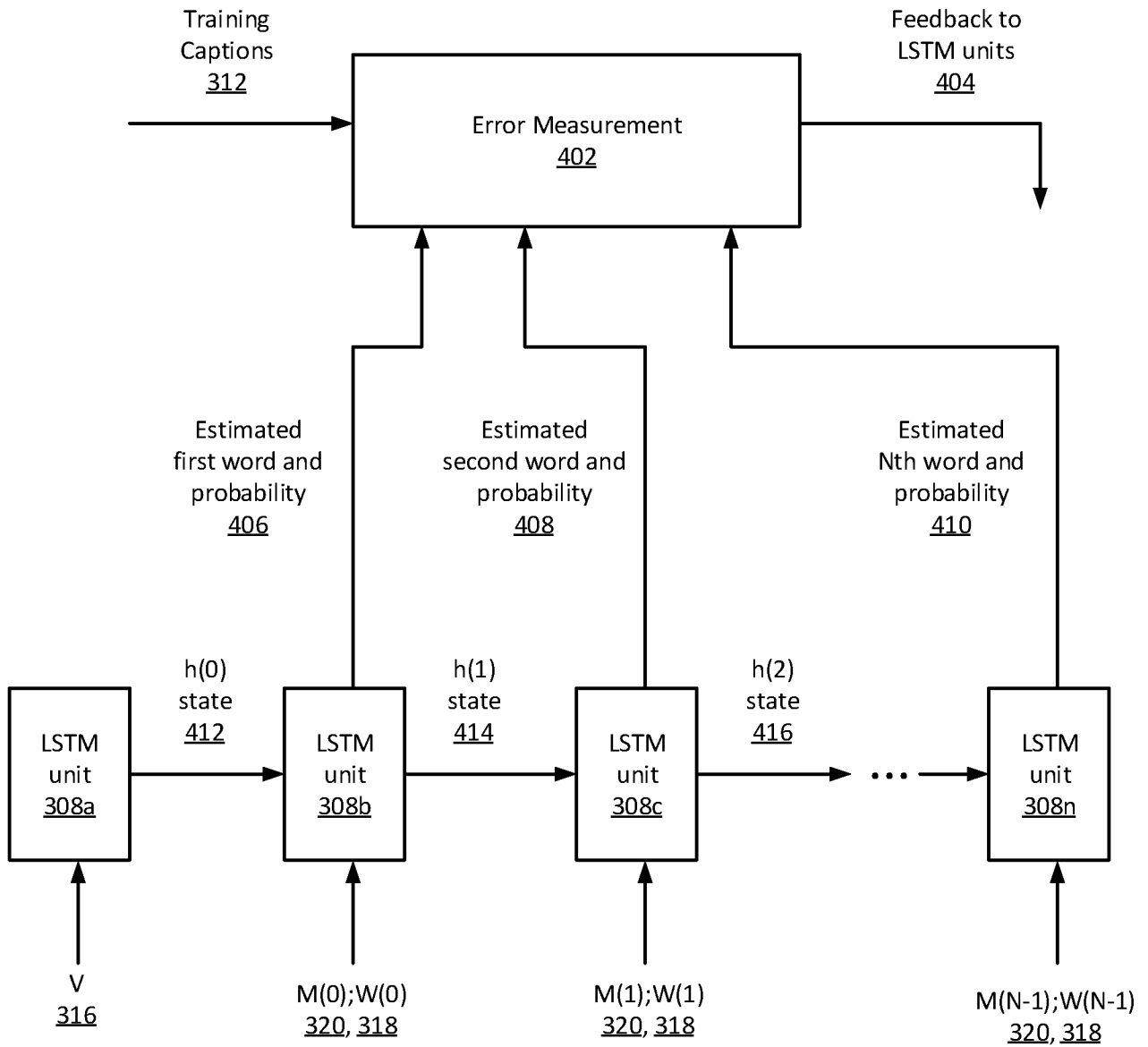
LSTM RNN training
400

FIG. 4

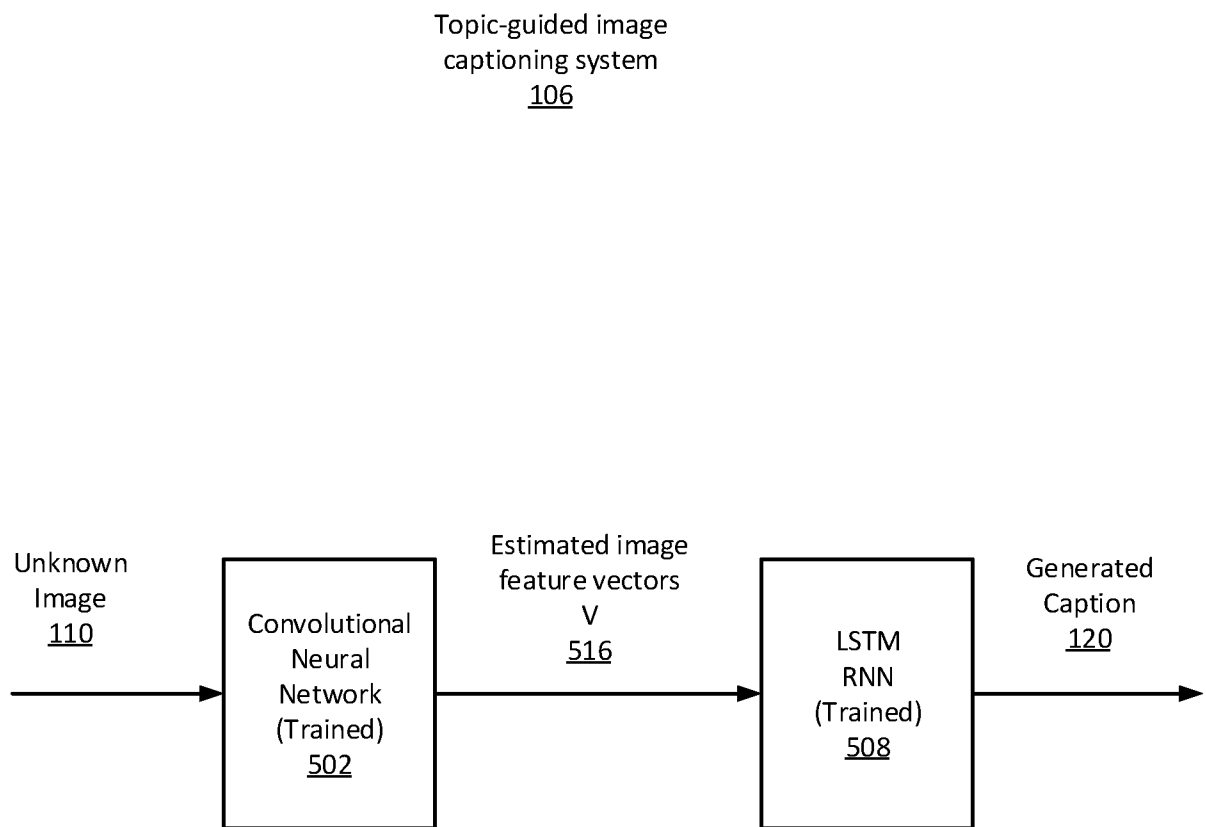


FIG. 5

Operation of Trained LSTM RNN
600

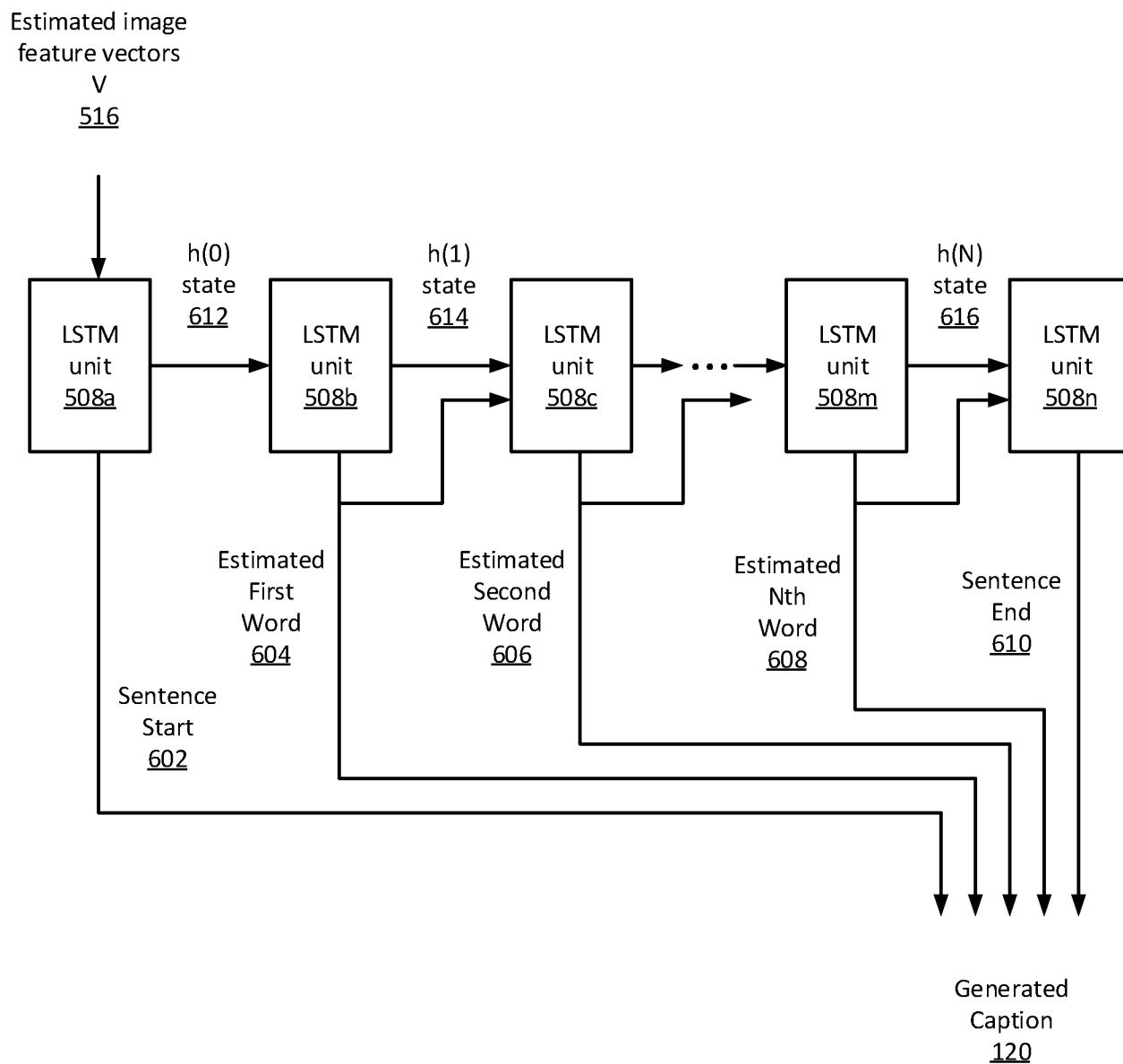


FIG. 6

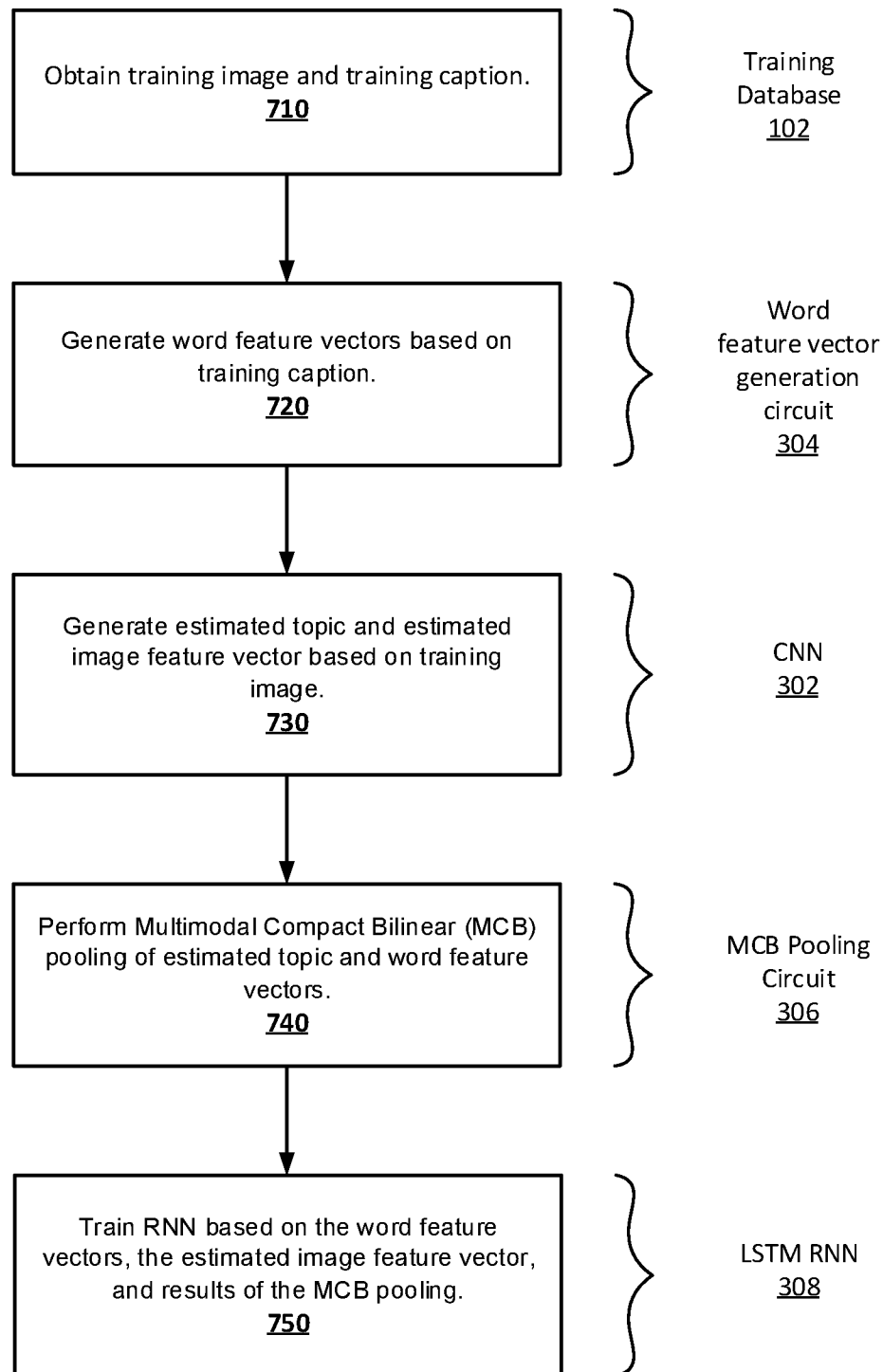
700

FIG. 7

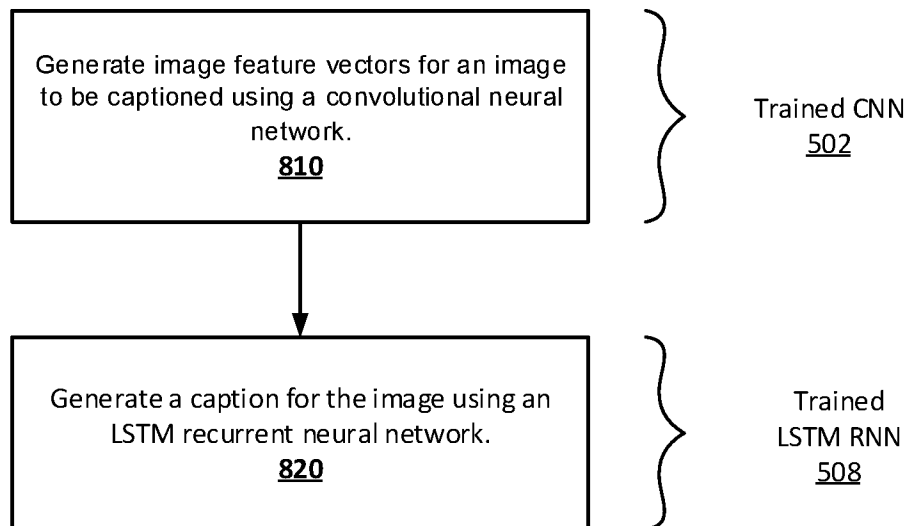
800

FIG. 8

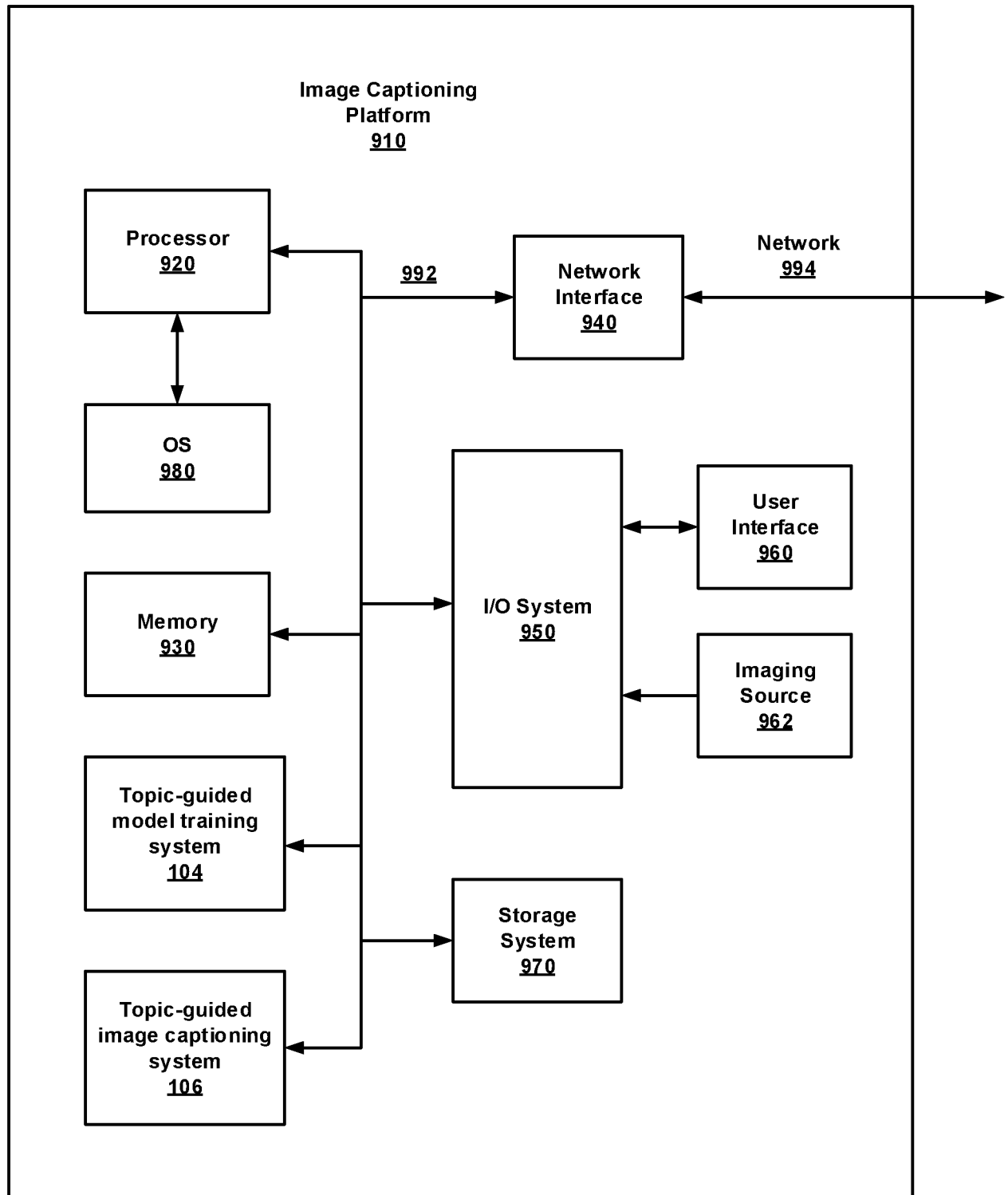
900

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/077280

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, G06K

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

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

CNKI, CNPAT, WPI, IEEE: CNN, RNN, image caption, description, annotation, MCB, train, visual feature, language feature

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106503055 A (TIANJIN UNIVERSITY) 15 March 2017 (2017-03-15) description, paragraphs [0013]-[0021] and figure 6	1-25
A	CN 105989067 A (HUAWEI TECHNOLOGIES CO., LTD.) 05 October 2016 (2016-10-05) the whole document	1-25
A	CN 105279495 A (TIANJIN UNIVERSITY) 27 January 2016 (2016-01-27) the whole document	1-25
A	CN 106446782 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 22 February 2017 (2017-02-22) the whole document	1-25
A	WO 2016197381 A1 (SENSETIME GROUP LTD.) 15 December 2016 (2016-12-15) the whole document	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

30 November 2017

Date of mailing of the international search report

27 December 2017

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

SHU, Qi

Telephone No. (86-10)62414064

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2017/077280

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	106503055	A	15 March 2017	None	
CN	105989067	A	05 October 2016	None	
CN	105279495	A	27 January 2016	None	
CN	106446782	A	22 February 2017	None	
WO	2016197381	A1	15 December 2016	None	