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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 April 2024 (25.04.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/084097 A1

(51) International Patent Classification:
G06N 3/045 (2023.01) *G06N 5/022* (2023.01)
G06N 3/092 (2023.01)

(21) International Application Number:
PCT/EP2023/079389

(22) International Filing Date:
21 October 2023 (21.10.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
20220100868 21 October 2022 (21.10.2022) GR

(71) Applicant: **DEEPMIND TECHNOLOGIES LIMITED** [GB/GB]; 5 New Street Square, London EC4A 3TW (GB).

(72) Inventors: **KOSSEN, Jannik Lukas**; 6 Pancras Square, London N1C 4AG (GB). **BELGRAVE, Danielle Charlotte Mary**; 6 Pancras Square, London N1C 4AG (GB). **TOMASEV, Nenad**; 6 Pancras Square, London N1C 4AG (GB). **CANGEA, Catalina-Codruta**; 6 Pancras Square, London N1C 4AG (GB). **KTENA, Sofia Ira**; 6 Pancras Square, London N1C 4AG (GB). **VÉRTES, Eszter**; 6 Pancras Square, London N1C 4AG (GB). **PATRAUCEAN, Viorica**; 6 Pancras Square, London N1C 4AG (GB). **JAEGLE, Andrew Coulter**; 6 Pancras Square, London N1C 4AG (GB).

(74) Agent: **FISH & RICHARDSON P.C.**; Highlight Business Towers, Mies-van-der-Rohe-Str. 8, 80807 Munich (DE).

(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, CV, CZ, DE, DJ, DK, DM,

(54) Title: SELECTIVE ACQUISITION FOR MULTI-MODAL TEMPORAL DATA

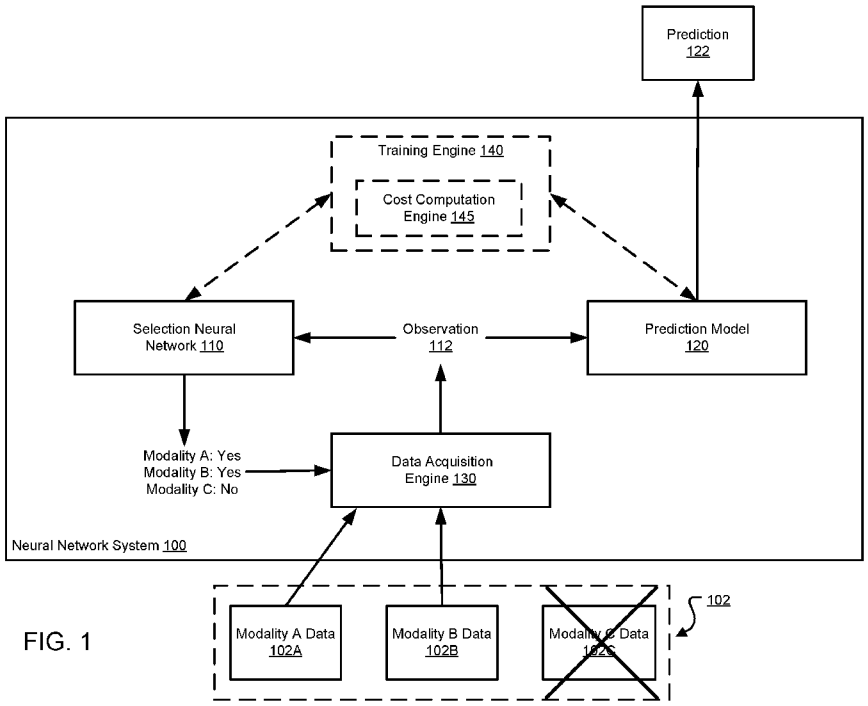

FIG. 1

(57) Abstract: Methods, systems, and apparatus, including computer programs encoded on computer storage media, for generating a prediction characterizing an environment. In one aspect, a method includes obtaining a respective observation characterizing a state of an environment for each time step in a sequence of multiple time steps, comprising, for each time step after a first time step in the sequence of time steps: processing a network input that comprises observations obtained for one or more preceding time steps to generate a plurality of acquisition decisions; obtaining an observation for the time step, wherein the observation includes data corresponding to modalities that are selected for acquisition at the time step, does not include data corresponding to modalities that are not selected for acquisition at the time step; and processing a model input that includes the observation for each time step in the sequence of time steps to generate the prediction.

WO 2024/084097 A1

[Continued on next page]

DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

SELECTIVE ACQUISITION FOR MULTI-MODAL TEMPORAL DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to GR national application No. 20220100868, filed on October 21, 2022. The disclosure of the prior application is considered part of and is incorporated by reference in the disclosure of this application.

BACKGROUND

[0002] This specification relates to processing data using machine learning models.

[0003] Machine learning models receive an input and generate an output, e.g., a predicted output, based on the received input. Some machine learning models are parametric models and generate the output based on the received input and on values of the parameters of the model.

[0004] Some machine learning models are deep models that employ multiple layers of models to generate an output for a received input. For example, a deep neural network is a deep machine learning model that includes an output layer and one or more hidden layers that each apply a non-linear transformation to a received input to generate an output.

SUMMARY

[0005] This specification generally describes a system implemented as computer programs on one or more computers in one or more locations that generates a prediction characterizing an environment.

[0006] According to one aspect, there is provided a method performed by one or more computers, the method comprising: obtaining a respective observation characterizing a state of an environment for each time step in a sequence of multiple time steps, comprising, for each time step starting from a first time step in the sequence of time steps: processing a network input that comprises observations obtained for any preceding time steps using a selection neural network to generate a plurality of acquisition decisions, wherein each acquisition decision corresponds to a respective modality from a set of multiple modalities and defines whether data corresponding to the modality is selected for acquisition at the time step; obtaining an observation for the time step, wherein the observation includes only data corresponding to modalities selected for acquisition at the time step; and processing a model input that includes the observation for each

time step in the sequence of time steps using a prediction model to generate a prediction characterizing the environment.

[0007] In some implementations, the method further comprises: determining an acquisition cost based on the respective modalities selected for acquisition at each time step in the sequence of time steps; determining a reward based at least in part on the acquisition cost; and training the selection neural network based on the reward using a reinforcement learning technique.

[0008] In some implementations, each modality in the set of modalities is associated with a respective cost factor, and wherein determining the acquisition cost comprises: determining, for each time step in the sequence of time steps, a respective acquisition cost for the time step based on the respective cost factor associated with each modality selected for acquisition at the time step; and determining the acquisition cost as a combination of the acquisition costs for the time steps.

[0009] In some implementations, for each time step in the sequence of time steps, determining the acquisition cost for the time step comprises: determining the acquisition cost for the time step as a sum of the cost factor associated with each modality selected for acquisition at the time step.

[0010] In some implementations, determining the acquisition cost as a combination of the acquisition costs for the time steps comprises: determining the acquisition costs as a sum over the acquisition costs for the time steps.

[0011] In some implementations, for one or more of the modalities, the cost factor for the modality is based at least in part on an amount of resource usage required to capture data corresponding to the modality.

[0012] In some implementations, the resource usage required to capture data corresponding to the modality characterizes at least energy usage required to capture data corresponding to the modality.

[0013] In some implementations, the resource usage required to capture data corresponding to the modality characterizes at least an amount of time required to capture data corresponding to the modality.

[0014] In some implementations, for one or more of the modalities, the cost factor for the modality is based at least in part on a risk associated with capturing data corresponding to the modality.

[0015] In some implementations, the environment comprises a patient, and the risk associated with capturing data corresponding to the modality is based at least in part on a medical risk to the patient resulting from capturing data corresponding to the modality.

[0016] In some implementations, the method further comprises: determining a prediction error that measures an error in the prediction generated by the prediction model; and determining the reward based on both: (i) the acquisition cost, and (ii) the prediction error.

[0017] In some implementations, the prediction model is a machine learning model.

[0018] In some implementations, the prediction model comprises a neural network.

[0019] In some implementations, the method further comprises: training the prediction machine learning model to optimize an objective function that depends on a prediction error of the prediction machine learning model.

[0020] In some implementations, for each time step in the sequence of time steps, the network input to the selection neural network at the time step further comprises: data identifying the acquisition decision for any modality at any preceding time step.

[0021] In some implementations, the method further comprises: for each of one or more time steps in the sequence of time steps: processing a model input that includes the observation for the time step and observations for one or more preceding time steps in the sequence of time steps using the prediction model to generate an intermediate prediction characterizing the environment; and determining an intermediate prediction error that measures an error in the intermediate prediction generated by the prediction model; and determining the reward based at least in part on the intermediate prediction errors.

[0022] In some implementations, the set of modalities includes an imaging modality, and wherein data corresponding to the imaging modality comprises image data.

[0023] In some implementations, the set of modalities includes a medical imaging modality.

[0024] In some implementations, the environment is a medical environment that comprises a patient.

[0025] In some implementations, the prediction characterizing the environment comprises a predicted medical diagnosis of the patient.

[0026] In some implementations, the prediction characterizing the environment comprises a prediction for a medical treatment to be applied to the patient.

[0027] In some implementations, the method further comprises, for each time step after the first time step in the sequence of time steps: determining that: (i) data corresponding to modalities, from the set of modalities, that are selected for acquisition at the time step will be included in the observation for the time step, and (ii) data corresponding to modalities, from the set of modalities, that are not selected for acquisition at the time step will not be included in the observation for the time step.

[0028] In some implementations, for each of one or more time steps after the first time step in the sequence of time steps: only a proper subset of the modalities in the set of modalities are selected for acquisition at the time step.

[0029] In some implementations, the method further comprises, for each time step after the first time step in the sequence of time steps: causing data to be acquired only for modalities selected for acquisition at the time step.

[0030] According to another aspect there is provided a system comprising: one or more computers; and one or more storage devices communicatively coupled to the one or more computers, wherein the one or more storage devices store instructions that, when executed by the one or more computers, cause the one or more computers to perform the operations of the methods described herein.

[0031] According to another aspect there are provided one or more non-transitory computer storage media storing instructions that when executed by one or more computers cause the one or more computers to perform the operations of the methods described herein.

[0032] The subject matter described in this specification can be implemented in particular implementations so as to realize one or more of the following advantages.

[0033] This specification describes a system for processing multi-modal data captured over a sequence of time points to generate a prediction characterizing an environment. In many real-world scenarios, capturing data corresponding to a modality can incur significant cost, e.g., in terms of resource consumption (e.g., consumption of energy or time), or in terms of risk (e.g., medical risk, e.g., resulting from exposing a patient to radiation from acquiring medical images of the patient, e.g., x-ray images or CT images). Moreover, processing data corresponding to certain modalities can also incur significant cost, e.g., in terms of computational resources (e.g., memory and computing power), e.g., for high-dimensional data such as image data, video data, or audio data. The system described in this specification can adaptively determine which data

modalities to acquire at each time point, and for certain time points, can acquire fewer than all the available modalities (or can even refrain from acquiring any modalities).

[0034] The system can be trained, using machine learning techniques, to optimize a trade-off between acquisition cost and predictive performance. In particular, the system can be trained to achieve an acceptable predictive performance while minimizing acquisition cost across the available modalities, thus enabling more efficient use of resources (e.g., energy resources or computational resources) and reduction of risk (e.g., medical risk). In some cases, the system can be trained to optimize the predictive performance while encouraging (or requiring) acquisition costs to satisfy a cost budget.

[0035] The details of one or more implementations of the subject matter of this specification are set forth in the accompanying drawings and the description below. Other features, aspects, and advantages of the subject matter will become apparent from the description, the drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0036] FIG. 1 shows an example neural network system.

[0037] FIG. 2 is a flow chart of an example process for generating a prediction characterizing an environment.

[0038] FIG. 3 is a flow chart of an example process for training a selection neural network.

[0039] FIG. 4 is a flow chart of an example process for sub-steps of one of the steps of the process of FIG. 3.

[0040] FIG. 5 is a flow chart of another example process for sub-steps of one of the steps of the process of FIG. 3.

[0041] FIG. 6 is an example illustration of generating a prediction using the selection neural network and determining one or more updates to the parameter values of the selection neural network based on the prediction.

[0042] Like reference numbers and designations in the various drawings indicate like elements.

DETAILED DESCRIPTION

[0043] FIG. 1 shows an example neural network system 100. The neural network system 100 is an example of a system implemented as computer programs on one or more computers in one or

more locations, in which the systems, components, and techniques described below can be implemented.

[0044] The neural network system 100 includes a selection neural network 110, a prediction model 120, a data acquisition engine 130, and, optionally, in some implementations, a training engine 140.

[0045] Generally, the neural network system 100 is a system that acquires input data 102 that is generated within or about an environment at each time step over a sequence of multiple time steps, and makes one or more predictions 122 that characterize one or more aspects of the environment. For example, the neural network system 100 can output a prediction 122 after having acquired input data 102 for the last time step in the sequence of multiple time steps.

[0046] At any given time step in the sequence, the input data 102 that is acquired by the system can potentially (but not necessarily) include data from a set of two or more available modalities. In this specification, a data "modality" refers to a type of data, e.g., that is generated using a specified sensor or diagnostic technique (e.g., medical diagnostic technique).

[0047] The set of modalities can include any appropriate modalities. A few examples of possible modalities are described next. In some implementations, the set of modalities comprises one or more of these examples.

[0048] In some implementations, the set of modalities includes an imaging modality, and data corresponding to the imaging modality comprises image data (e.g., one-dimensional (1D) image data, two-dimensional (2D image data, three-dimensional (3D) image data, etc.). The image data may comprise pixel value data, e.g. color or monochrome pixel value data.

[0049] For instance, the set of modalities can include one or more medical imaging modalities, e.g., a computed tomography (CT) modality, an ultrasound (US) modality, a magnetic resonance imaging (MRI) modality, an x-ray modality, a histological imaging modality, an electroencephalogram (EEG) modality, an electromyography (EMG) modality, an electrocardiogram (ECG) modality, etc.

[0050] As another example, the set of modalities can include a camera modality, e.g., where data corresponding to the camera modality is captured using a camera, e.g., a visible spectrum camera or an infrared spectrum camera.

[0051] In some implementations, the set of modalities can include a genetic data modality, and data corresponding to the genetic data modality comprises genetic data. Genetic data can include,

e.g., data defining a respective expression level (in a subject) of each gene in a set of genes. The genetic data may be obtained by a suitable diagnostic technique, such as DNA or RNA sequencing, performed on genetic material obtained from the subject.

[0052] In some implementations, the set of modalities can include a proteomic data modality, and data corresponding to the proteomic data modality comprises proteomic data. Proteomic data can include, e.g., data defining a respective expression level (in a subject) of each protein in a set of proteins.

[0053] In some implementations, the set of modalities can include a blood testing modality, and data corresponding to the blood testing modality can include data defining levels of one or more components of the blood of a subject, e.g., sodium, potassium, chloride, bicarbonate, blood urea nitrogen, magnesium, creatinine, glucose, calcium, cholesterol, etc.

[0054] In some implementations, the set of modalities can include an audio modality, and data corresponding to the audio modality can include audio data, e.g., audio data characterizing words spoken by a person, audio data characterizing sounds made by one or more body parts of a person (e.g., the heart, the digestive system, the lungs, etc.), etc. The audio data may comprise data defining an audio waveform such as a series of values in the time and/or frequency domain defining the waveform.

[0055] In some implementations, the set of modalities can include a biopsy modality, and data corresponding to the biopsy modality can characterize a sample of cells or tissue obtained from a patient by way of a biopsy. For instance, data corresponding to the biopsy modality can include a microscope image of the sample obtained from the patient.

[0056] In some implementations, the set of modalities can include modalities that measure one or more of: humidity, light, air quality, sound, temperature, wind speed, pH, etc.

[0057] In a particular implementation, the set of modalities includes at least: a medical imaging modality and a blood testing modality.

[0058] In a particular implementation, the set of modalities includes at least: a medical imaging modality, a blood testing modality, and a biopsy modality.

[0059] In a particular implementation, the set of modalities includes at least: a medical imaging modality, a blood testing modality, a biopsy modality, and a genetic data modality.

[0060] In some of these implementations, these data types differ not only in feature spaces and dimensionalities, but also in data capturing processes and costs associated with capturing data

corresponding to these data types. For example, medical imaging can be ordered at the discretion of a physician and needs to be captured at a relatively higher cost, e.g., in terms of resource consumption or in terms of risk, while blood pressure and temperature can be monitored on a regular basis and can be captured at a relatively lower cost.

[0061] The environment can be any appropriate environment, e.g., a real-world environment, e.g., a medical environment, an agriculture environment, an aquaculture environment, an industrial environment, or a scientific environment.

[0062] A medical environment can include a patient, and one or more of the modalities can be modalities that generate data characterizing the patient, as described above.

[0063] An industrial environment can include, e.g., a manufacturing facility (e.g., that includes one more industrial machines used for the production of manufactured goods), a chemical processing facility (e.g., that includes one more industrial machines used for chemical processing), a data center facility (e.g., that includes a collection of computing units, e.g., processors, used for performing computing tasks), or an energy production facility (e.g., a nuclear plant, a hydroelectric plant, a photovoltaic power station, etc.). The one or more modalities can be modalities that generate data characterizing the facility, e.g. data generated by one or more sensors located within or around the facility, e.g. sensors for measuring the states of industrial machines or computing units within the facility. A prediction characterizing the industrial environment may comprise predicted values for one or more properties that may be determined based on the sensor values, or one or more properties measured by the sensor(s), e.g. the predicted values may comprise predicted sensor values.

[0064] A scientific environment can include a collection of subjects being studied for scientific purposes, where the subjects can include, e.g., plants, animals, cells, tissues, etc.

[0065] The environment can evolve over time, and thus the input data 102 generated within or about the environment at a first time step can have different values than the input data 102 generated within or about the environment at a second time step. The collection of input data 102 that is acquired over the sequence of multiple time steps may thus be referred to as a "temporal" sequence input data, because in some implementations, the input data is arranged according to the time step at which it was captured. For example, the most recent input data is the last input data in a temporal sequence of input data and the least recent input data is the first input data in the temporal sequence.

[0066] The one or more predictions 122 that characterize one or more aspects of the environment are made by the prediction model 120 based on the input data 102. The prediction model 120 can be configured as a machine learning model that can have any appropriate machine learning model architecture. For instance, the prediction model 120 can be implemented as a neural network, or a decision tree, or a random forest, or a support vector machine, or a linear regression model, and so forth. In a particular example, the prediction model 120 can be implemented as a neural network that can include any appropriate types of neural network layers (e.g., fully connected layers, attention layers, convolutional layers, and so forth) in any appropriate number (e.g., 5 layers, or 10 layers, or 100 layers) and connected in any appropriate configuration (e.g., as a directed graph of layers).

[0067] A few examples of possible predictions 122 are described next.

[0068] In some implementations, the environment is a medical environment that includes a patient, and the prediction defines a predicted medical treatment to be applied to the patient. For instance, the prediction can include a respective score for each medical treatment in a set of medical treatments, where the score for a medical treatment defines a likelihood that the medical treatment should be applied to the patient. The set of medical treatments can include medical treatments corresponding to administering a drug to the patient, performing an intervention (e.g., surgery) on the patient, etc.

[0069] In some implementations, the environment is a medical environment that includes a patient, and the prediction defines a predicted medical diagnosis for the patient. For instance, the prediction can include a respective score for each medical diagnosis in a set of medical diagnoses, where the score for a medical diagnosis defines a likelihood that the medical diagnosis applies to the patient. The set of medical diagnoses can include, e.g., diagnoses for one more diseases, e.g., cancer, diabetes, heart failure, Alzheimer's, flu, measles, strep throat, sepsis, etc.

[0070] In some implementations, the environment is an agriculture environment (e.g., an environment where crops are cultivated) or an aquaculture environment (e.g., an environment where aquatic organisms are cultivated), and the prediction defines a predicted yield (e.g., measured in tons of crops or aquatic organisms), or a predicted amount of time until crops or aquatic organisms in the environment should be harvested (e.g., measured in days).

[0071] In some implementations, the environment is an industrial environment, and the prediction defines a predicted production level of the industrial environment over a predefined time range (e.g., 1 hour, 1 day, or 1 week), e.g., a number of units of product generated by a manufacturing facility, or a quantity of chemicals generated by a chemical processing facility, or a number of computing tasks completed by a data center facility, or an amount of energy produced by an energy production facility.

[0072] In some implementations, the environment is a scientific environment, and the prediction defines a predicted result of a scientific study, e.g., the health of subjects of the scientific study, e.g., the integrity of cell walls of a collection of cells at the conclusion of the study, or the weight of animals in a population of animals at the conclusion of the study.

[0073] In these example environments and many other real-world environments, capturing data corresponding to a modality oftentimes incur significant cost, e.g., in terms of resource consumption (e.g., consumption of energy or time), or in terms of risk (e.g., medical risk, e.g., resulting from exposing a patient to radiation from acquiring medical images of the patient, e.g., x-ray images or CT images). Moreover, processing data corresponding to certain modalities can also incur significant cost, e.g., in terms of computational resources (e.g., memory and computing power), e.g., for high-dimensional data such as image data, video data, or audio data.

[0074] Therefore, although the neural network system 100 could potentially receive input data corresponding to each and every modality in the set of two or more available modalities at each time step, the system may not actually do so, and may instead only acquire (and thereafter receive) input data corresponding to each modality in a proper subset of the set of two or more available modalities at each of one or more time steps. A proper subset includes at least one modality in the set of two or more available modalities, but less than all of the modalities in the set.

[0075] In the example of FIG. 1, the neural network system 100 could potentially receive data corresponding respectively to a set of three data modalities that might be made available to the system: data 102A corresponding to modality A, data 102B corresponding to modality B, and data 102C corresponding to modality C.

[0076] However, as illustrated, the actually acquired input data 102 is multi-modal data that only includes data corresponding respectively to two of the three available data modalities: data 102A corresponding to modality A, and data 102B corresponding to modality B. That is, data 102C

corresponding to modality C is not acquired and is therefore not received by the system, and the actually acquired input data 102 does not include data 102C corresponding to modality C.

[0077] In other examples, the input data can potentially (but not necessarily) include data corresponding to a smaller (e.g., two) or larger (e.g., ten, one hundred, or more) set of available modalities. Analogously, in those examples, the actually acquired input data 102 can include data corresponding to each modality in a proper subset of the smaller or larger set of available modalities.

[0078] In particular, for each time step after a first time step in the sequence of time steps, the neural network system 100 uses the selection neural network 110 to make acquisition decisions that define, for each modality in the set of modalities, whether data corresponding to the modality will be selected for acquisition.

[0079] At any given time step after the first time step in the sequence of multiple time steps, the selection neural network 110 processes a network input that includes (i) previous observations 112 obtained for one or more preceding time steps and, optionally, (ii) data identifying the acquisition decision for any modality at any preceding time step, to generate a plurality of acquisition decisions for the given time step. Each acquisition decision corresponds to a respective modality from the set of multiple modalities and defines whether data corresponding to the modality should be selected for acquisition at the given time step. As will be explained further below, an "observation" refers to data that is generated by the data acquisition engine 130 from the acquired input data 102 and that is provided to the selection neural network 110 and/or prediction model 120 for further processing.

[0080] The selection neural network 110 can have any appropriate neural network architecture that allows the selection neural network 110 to generate acquisition decisions from previous observations. In particular, the selection neural network can include any appropriate types of neural network layers (e.g., fully connected layers, attention layers, convolutional layers, and so forth) in any appropriate number (e.g., 5 layers, or 10 layers, or 100 layers) and connected in any appropriate configuration (e.g., as a directed graph of layers).

[0081] As a particular example, the selection neural network 110 and the prediction model 120 can each be configured as a respective neural network having one of the architectures described in Andrew Jaegle, et al. PerceiverIO: A general architecture for structured inputs & outputs. In International Conference on Representation Learning, 2022.

[0082] In the example of FIG. 1, at a particular time step, the selection neural network 110 generates a total of three acquisition decisions: decision A corresponding to modality A, decision B corresponding to modality B, and decision C corresponding to modality C. Specifically, decision A defines that data corresponding to modality A should be selected for acquisition at the time step, decision B defines that data corresponding to modality B should be selected for acquisition at the time step, and decision C defines that data corresponding to modality C should not be selected for acquisition at the time step. The selection neural network 110 can generate different acquisition decisions at other time steps.

[0083] Each acquisition decision can be generated deterministically, e.g. by an output of the selection neural network 110. For example, an output layer of the selection neural network 110 can include a respective neuron corresponding to each modality; and each modality is selected for acquisition only if the activation of the neuron exceeds a predefined threshold. Alternatively, each acquisition decision can be generated stochastically e.g. where the output of the selection neural network 110 parameterizes a distribution from which the acquisition decision is sampled. For example, each acquisition decision can be a binary decision, with 0 indicating that data corresponding to a particular modality should not be selected for acquisition, and 1 indicating that data corresponding to the particular modality should be selected for acquisition.

[0084] The neural network system 100 then uses the data acquisition engine 130 to effectuate the acquisition decisions generated by the selection neural network 110. That is, the neural network system 100 provides the acquisition decisions to the data acquisition engine 130—and the data acquisition engine 130 causes data to be acquired only for modalities selected for acquisition at the given time step.

[0085] In the example of FIG. 1, the data acquisition engine 130 acquires, in accordance with the acquisition decisions generated by the selection neural network 110, data corresponding to modality A and data corresponding to modality B. The data acquisition engine 130 refrains from acquiring data corresponding to modality C.

[0086] In some implementations, the data acquisition engine 130 can effectuate (e.g. execute or enact) the acquisition decisions by passing an electronic signal to a sensor, or another electronic device having environment sensing capabilities that is communicatively coupled to the system, to capture data corresponding to one of the selected modalities. In response to receiving the electronic signal, the sensor operates to capture the data with or about the environment.

[0087] In some implementations, the data acquisition engine 130 can effectuate the acquisition decisions by generating and outputting a prompt for presentation to a user through a user interface device. The user interface device can be any appropriate stationary or mobile computing device, such as a desktop computer, a workstation in a medical environment, a tablet, a smartphone, or a smartwatch.

[0088] The prompt can help guide a user in capturing data according to the acquisition decisions generated by the selection neural network 110. The prompt can instruct the user on what modalities of data to capture. For example, the prompt can be presented within a window with text asking that data corresponding to one of the selected modalities should be captured. A user can interact with the user interface device to view the selected modality, and upload data corresponding to the selected modality after the data is captured.

[0089] The neural network system 100 then generates an observation 112 for the given time step from the input data 102, which is acquired by the data acquisition engine 130 in accordance with the plurality of acquisition decisions generated by the selection neural network 110.

[0090] The observation 112 for the given time step: (i) includes data corresponding to modalities, from the set of modalities, that are selected for acquisition at the given time step, and that (ii) excludes, i.e., does not include, data corresponding to modalities, from the set of modalities, that are not selected for acquisition at the given time step. After being generated, the observation 112 is then provided to the prediction model 120 for further processing.

[0091] In this way, although the data that is potentially available to the system includes multi-modal data corresponding respectively to the set of modalities, only data corresponding to a proper subset of the modalities in the set of modalities may actually be selected for acquisition by the neural network system 100. For example, only data that corresponds respectively to a small number of modalities within a relatively large number of modalities may be selected for acquisition by the neural network system 100, and, thereafter, used by the prediction model 120 to generate the prediction 122.

[0092] By incorporating the selection neural network 110 and acquiring data in accordance with the acquisition decisions generated by the selection neural network 110, the neural network system 100 can reduce the amount of computational resources consumed by the prediction process because repeatedly acquiring and subsequently processing data from all of the set of modalities is no longer necessarily required. Instead, at each of at least some of the time steps,

only data from a relatively small number of selected modalities needs to be acquired and then processed.

[0093] The training engine 140, when included, can train the selection neural network 110 and, optionally, the prediction model 120 to determine trained parameter values of the selection neural network 110 and, optionally, trained parameter values of the prediction model 120 that enable the selection neural network 110 to generate acquisition decisions that can result in reduced consumption of computational resources by the system while still maintaining predictive performance, e.g., in terms of the accuracy of the predictions 122. Thus, in some implementations, the selection neural network 110 and the prediction model 120 may be jointly trained by the training engine 140.

[0094] In the example of FIG. 1, the training engine 140 includes or has access to a cost computation engine 145. The cost computation engine 145 is configured to compute acquisition costs associated with the modalities that are selected for acquisition according to the acquisition decisions generated by the selection neural network 110.

[0095] The training engine 140 can thus apply a reinforcement learning technique that uses a reward derived from the acquisition costs to train the selection neural network 110 jointly with the prediction model 120 to optimize a trade-off between acquisition cost and predictive performance.

[0096] In particular, the training engine 140 can train the selection neural network 110 and the prediction model 120 to achieve an acceptable predictive performance while minimizing acquisition cost across the available modalities, thus enabling more efficient use of resources (e.g., energy resources or computational resources) and reduction of risk (e.g., medical risk).

[0097] The cost computation engine 145 can be configured to compute the acquisition cost for a modality in a set of modalities based on any appropriate criteria. A few examples of possible criteria for setting acquisition costs for modalities are described next.

[0098] In some implementations, the acquisition cost for a modality can be based at least in part on an amount of resource usage (e.g., energy or time) required to acquire data corresponding to the modality.

[0099] In some implementations, the acquisition cost for a modality can be based at least in part on an amount of risk required to acquire data corresponding to the modality. For example, in a medical environment, acquiring data corresponding to a biopsy modality may incur a risk of

infection in the patient, and acquiring data corresponding to an x-ray modality may incur a risk of exposing the patient to unhealthy levels of radiation. An amount of risk may be determined based on statistics characterizing different outcomes (e.g. patient outcomes) when data corresponding to the modality is acquired.

[0100] In some implementations, the acquisition cost for a modality can be based at least in part on a level of disruption caused by acquiring data corresponding to the modality. For example, in an industrial environment, acquiring data corresponding to a modality can include running diagnostic tests that reduce production of the industrial facility. As another example, in a scientific environment, acquiring data corresponding to a modality can include disrupting conditions in the environment (e.g., by performing tests on one or more subjects in the environment) in a manner that could compromise the validity or accuracy of results of the experiment. Training the selection neural network 110 will be described further below with reference to FIGS. 3-6.

[0101] FIG. 2 is a flow diagram of an example process 200 for generating a prediction characterizing an environment. For convenience, the process 200 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 200.

[0102] The environment can be any appropriate environment, e.g., a real-world environment, e.g., a medical environment, an agriculture environment, an aquaculture environment, an industrial environment, or a scientific environment.

[0103] The system repeatedly performs steps 202 and 204 to obtain a respective observation characterizing a state of an environment for each time step in a sequence of multiple time steps. That is, the system performs one iteration of steps 202 and 204 for each time step in the sequence of multiple time steps.

[0104] In some implementations, the number of time steps is fixed (predefined). For example, the system can generate a sequence-level prediction after a predefined number of time steps have elapsed. In other implementations, the number of time steps is flexible, and different sequences can include varying numbers of time steps. For example, the system can repeatedly perform iterations of steps 202 and 204 until a termination signal (e.g., a flag or another indicator) is received at a given time step indicating that the given time step is the last time step in the

sequence. For example, a flag can be set to a first value if the given time step is not the last time step in a sequence and the flag can be set to a second value if the given time step is the last time step in the sequence. The termination signal may be based on the prediction(s) generated by the system.

[0105] For each time step after a first time step in the sequence of multiple time steps, the system processes, using a selection neural network, a network input that includes (i) observations obtained for one or more preceding time steps and, optionally (ii) data identifying the acquisition decision for any modality at any preceding time step, to generate a plurality of acquisition decisions for the time step (step 202). An "observation" refers to data that is generated by the data acquisition engine 130 from the acquired input data 102 and that is provided to the selection neural network 110 and/or prediction model 120 for further processing. Each acquisition decision corresponds to a respective modality from a set of multiple modalities, and defines whether data corresponding to the modality is selected for acquisition at the time step.

[0106] For the first time step, because there is no preceding time step, some implementations of the system can instead provide a predetermined network input, i.e., an input having predetermined values, for processing by the selection neural network. Some other implementations of the system can alternatively acquire a default (e.g., random or predefined) set of modalities, i.e., without using the selection neural network to generate any acquisition decisions for the first time step.

[0107] The system obtains an observation for the time step in accordance with the plurality of acquisition decisions generated by the selection neural network (step 204). For example, the system can use a data acquisition engine to acquire data corresponding to each modality that is selected for acquisition according to the acquisition decisions, and then include the acquired data in the observation.

[0108] In particular, the observation (i) includes data corresponding to modalities, from the set of modalities, that are selected for acquisition at the time step, and (ii) does not include data corresponding to modalities, from the set of modalities, that are not selected for acquisition at the time step.

[0109] The selected modalities may, and generally will, vary from one time step to another. In other words, the system may obtain data corresponding to different modalities at different time steps.

[0110] In some examples, for one or more time steps in the sequence of multiple time steps, the system can obtain an observation that includes data corresponding to all of the modalities in the set. In another example, for one or more time steps, the system can obtain an observation that includes data corresponding to a proper subset of the set of modalities (and does not include data corresponding to any remaining modality that is not in the proper subset). A "proper" subset of a set is a subset that includes one or more but not all of the elements in the set. In another example, for one or more time steps, the system can obtain a null observation that does not include data corresponding to any modality in the set.

[0111] After having performed the iteration of steps 202 and 204 for the last time step in the sequence of multiple time steps, the system processes, using a prediction model, a model input that includes the observation for each time step in the sequence of time steps to generate a prediction characterizing the environment (step 206).

[0112] An example algorithm for generating a prediction is shown below.

Algorithm 1 A2MT

Inputs: Test input $\boldsymbol{x}$, agent π, model f.

for $t = 1$ to T **do**
  Sample $a_t \sim \pi(\cdot|\tilde{\boldsymbol{x}}_{1:t-1}, \boldsymbol{a}_{1:t-1}; \boldsymbol{\theta})$.
  for $m = 1$ to M **do**
    if $a_{t,m} = 1$ **then**
      Acquire: $\tilde{\boldsymbol{x}}_{t,m} \leftarrow \boldsymbol{x}_{t,m}$.
    else
      Do not acquire: $\tilde{\boldsymbol{x}}_{t,m} \leftarrow \emptyset$.
    end if
  end for
end for
Return prediction $f(\tilde{\boldsymbol{x}}_{1:T})$.

[0113] In Algorithm 1, each input x includes a sequence of observations $x_i = (x_{i,1}, \ldots, x_{i,T})$. At each time step t, the observation $x_{i,t}$ includes data corresponding to M modalities $x_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,M})$. Each modality may be high-dimensional, $x_{i,t,m} \in R^{d_m}$. For example, $x_{i,t,m}$ could be a single frame in a video having dimensionality $d_m = H \cdot W \cdot C$, where H is height, W is width, and C is number of color channels.

[0114] At each timestep $t \in (1, \ldots, T)$, a plurality of acquisition decisions across modalities $a_t = (a_{t,1}, \ldots, a_{t,M})$ are generated by sampling from the output of the selection neural network (which is referred to in Algorithm 1 as an agent): $a_t \sim \pi(\cdot\ |\tilde{x}_{1:t-1}, a_{1:t-1}; \theta)$. Here, $a_{t,M} \in (0, 1)$ is a binary indicator of whether modality m was acquired at time step t. $\tilde{x}$ is used instead of x to

highlight that the input data may contain missing entries, and θ are the parameters of the selection neural network. At each timestep t, and for each modality m, data $x_{m,t}$ corresponding to the modality m is acquired only if $a_{m,t} = 1$.

[0115] By repeatedly performing the process 200, the system can generate different predictions that characterize the same or different aspects of the environment. That is, the process 200 can be performed as part of generating a prediction from a sequence of observations for which the desired output, i.e., the desired prediction that should be generated by the system from the sequence of observations, is not known. One or more actions may be performed based on the prediction(s). For example, an agent, such as an electromechanical agent, interacting with a real-world environment to perform a task, may select one or more action to perform in the real-world environment according to the prediction(s).

[0116] Some of all of the steps of the process 200 can also be performed as part of processing sequences of observations derived from a set of training data, i.e., sequences of observations derived from input data for which the predictions that should be generated by the system are known, in order to train the trainable components of the system to determine trained values for the parameters of these components.

[0117] FIG. 3 is a flow diagram of an example process 300 for training a selection neural network. For convenience, the process 300 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 300.

[0118] During training, process 300 can be performed subsequent to process 200 on each training input selected from a set of training data derived from a plurality of temporal sequences of input data generated within or about an environment (e.g., one of the physical environments mentioned above or a computer simulation of one of these physical environments). That is, for each training input, the system performs process 200 to generate a prediction characterizing the environment using the selection neural network and in accordance with the current values of the parameters of the selection neural network, and then performs process 300 to determine one or more updates to the parameter values of the selection neural network based on the prediction generated in process 200.

[0119] The process 300 is illustrated in FIG. 6, which shows an example of generating a prediction using the selection neural network and determining one or more updates to the parameter values of the selection neural network based on the prediction.

[0120] As illustrated in FIG. 6, at any given time step, the selection neural network ("Agent") generates a total of three acquisition decisions: a first decision corresponding to a text modality, a second decision corresponding to an image modality, and a third decision corresponding to a numeric modality. To generate these acquisition decisions for the given time step, the selection neural network processes a network input that includes observations obtained for one or more preceding time steps to generate an output π which parameterizes a distribution from which the acquisition decision is sampled.

[0121] For example, at a first time step, the first decision defines that text data should be selected for acquisition at the time step, the second decision defines that image data should be selected for acquisition at the time step, and the third decision defines that numeric data should not be selected for acquisition at the time step. As represented by the [masked] tokens, some input data may contain missing entries.

[0122] After generating the prediction based on the modalities selected for acquisition at each time step by the selection neural network, the system determines an acquisition cost based on the respective modalities selected for acquisition at each time step in the sequence of time steps (step 302). In some implementations, each modality in the set of modalities is associated with a respective cost factor. In these implementations, the system can perform sub-steps 402-404, as is explained in more detail with reference to FIG. 4, to determine the acquisition cost.

[0123] FIG. 4 is a flow diagram of sub-steps 402-404 of step 302 of the process of FIG. 3.

[0124] In implementations where each modality in the set of modalities is associated with a respective cost factor, the system can determine, for each time step in the sequence of time steps, a respective acquisition cost for the time step based on the respective cost factor associated with each modality selected for acquisition at the time step (step 402). For example, the acquisition cost for the time step can be computed as a sum, either weighted or unweighted, of the cost factor associated with each modality selected for acquisition at the time step.

[0125] The system determines the acquisition cost as a combination of the acquisition costs for the time steps in the sequence (step 404). For example, the acquisition costs can be computed as

a sum, either weighted or unweighted, over the respective acquisition costs for the time steps in the sequence.

[0126] For each of one or more of the modalities, the cost factor for the modality is based at least in part on an amount of resource usage required to capture data corresponding to the modality. For example, the resource usage required to capture data corresponding to the modality characterizes at least energy usage required to capture data corresponding to the modality. As another example, the resource usage required to capture data corresponding to the modality characterizes at least an amount of time required to capture data corresponding to the modality.

[0127] Additionally or alternatively, for each of one or more of the modalities, the cost factor for the modality is based at least in part on a risk associated with capturing data corresponding to the modality. When the environment is a medical environment that includes a patient, for example, the risk associated with capturing data corresponding to the modality is based at least in part on a medical risk to the patient resulting from capturing data corresponding to the modality.

[0128] The system determines a reward based at least in part on the acquisition cost the selected modalities (step 304). The acquisition cost can be included in the reward, which is typically a numeric value, in any appropriate manner.

[0129] For example, the system can determine the reward based at least in part on a comparison of the acquisition cost to a threshold referred to as a "cost budget." In some implementations, the system can reduce the reward by a predefined or adaptive amount if the acquisition cost exceeds the cost budget. The cost budget can indicate, e.g., an acceptable level of acquisition cost, e.g., an acceptable amount of energy usage, or an acceptable amount of medical risk (e.g., based on a tolerable amount of radiation exposure for a patient), or an acceptable amount of computational resources (e.g., memory and computing power) used for processing data from the acquired modalities.

[0130] In some implementations, the reward depends on both (i) the acquisition cost and (ii) a prediction error that measures an error in the prediction generated by the prediction model. In these implementations, as illustrated in FIG. 6, the reward can for example be computed as an expectation value:

$$R = \mathbb{E}\left[-C(\boldsymbol{a}) - \mathcal{L}(f(\tilde{\boldsymbol{x}}_{1:T}), y)\right], \quad \text{where} \quad C(\boldsymbol{a}) = \sum_{t=1}^{T}\sum_{m=1}^{M} c_m a_{t,m}. \qquad (1)$$

[0131] Here, the expectation is over a training input (x, y), where x is a sequence of observations and y is the ground truth prediction, a represents acquisition decisions generated by

the selection neural network; $C(a)$ represents the total acquisition cost of the sequence of observations; C_m is a modality-specific cost factor, and $\mathcal{L}(f(\tilde{x}_{1:T}), y)$ is log likelihood loss of the prediction generated by the prediction model with respect to the ground truth prediction (although other loss functions may of course be used, i.e. loss functions comparing the prediction generated by the prediction model with the ground truth prediction).

[0132] Optionally, in some implementations, the system adds intermediate prediction errors to the reward, e.g., the reward computed using Equation (1). The intermediate prediction errors, when used, encourage the selection neural network to decrease the prediction error. In these implementations, the system can perform sub-steps 502-506, as is explained in more detail with reference to FIG. 5, to determine the reward.

[0133] FIG. 5 is a flow diagram of sub-steps 502-506 of step 304 of the process of FIG. 3.

[0134] For each of one or more time steps in the sequence of time steps, the system processes a model input that includes the observation for the time step and observations for one or more preceding time steps in the sequence of time steps using the prediction model to generate an intermediate prediction characterizing the environment (step 502).

[0135] For each of the one or more time steps in the sequence of time steps, the system determines an intermediate prediction error that measures an error in the intermediate prediction generated by the prediction model (step 504).

[0136] The system determines the reward based at least in part on the intermediate prediction errors that have been determined for the one or more time steps (step 506). For example, the system can add the intermediate prediction errors to the reward computed using Equation (1). The intermediate prediction errors can for example be computed as:

$$I = -\alpha \sum_{t=1}^{T} (\mathcal{L}(f(\tilde{x}_{1:t}), y) - \gamma \mathcal{L}(f(\tilde{x}_{1:t-1}), y)), \quad (2)$$

where α is a hyperparameter (e.g., a predefined constant value), γ is the discount factor, x is a sequence of observations and y is the ground truth prediction, and $\mathcal{L}(f(\tilde{x}_{1:T}), y)$ is log likelihood loss of the prediction generated by the prediction model with respect to the ground truth prediction.

[0137] The system trains the selection neural network based on the reward using a reinforcement learning technique to adjust the values of the parameters of the selection neural network (step 304). In particular, the system trains the selection neural network to generate acquisition decisions that maximize the reward that is determined based at least in part on the acquisition

cost. For example, the reinforcement learning technique can be a policy gradient technique, e.g., an advantage actor critic (A2C) policy gradient technique, that applies Gumbel parameterization to the (discrete) acquisition decisions.

[0138] In some implementations, the system also trains the prediction model based on the reward, e.g., the reward computed using Equation (1), which depends on both the acquisition cost and the prediction error, to simultaneously adjust the values of the parameters of the prediction model.

[0139] For example, the system can train the prediction model and the selection neural network together to jointly update the parameter values of both the selection neural network and the prediction model, e.g., in order to allow the prediction model to adapt specifically to the combinations of modalities frequently selected by the selection neural network. In this example, as the parameter values of the prediction model are updated, the rewards received by the selection neural network may consequently change.

[0140] Alternatively, in other implementations, the system can train the prediction model separately from the training of the selection neural network (during which the parameter values of the prediction model are held fixed), e.g., based on optimizing an objective function that depends on the prediction error of the prediction machine learning model.

[0141] For example, the system can pre-train the prediction model to process masked sequences of observations to generate corresponding predictions. The system then trains the selection neural network to update the parameter values of the selection neural network, while holding the pre-trained parameter values of the prediction model fixed.

[0142] This specification uses the term "configured" in connection with systems and computer program components. For a system of one or more computers to be configured to perform particular operations or actions means that the system has installed on it software, firmware, hardware, or a combination of them that in operation cause the system to perform the operations or actions. For one or more computer programs to be configured to perform particular operations or actions means that the one or more programs include instructions that, when executed by data processing apparatus, cause the apparatus to perform the operations or actions.

[0143] Embodiments of the subject matter and the functional operations described in this specification can be implemented in digital electronic circuitry, in tangibly-embodied computer software or firmware, in computer hardware, including the structures disclosed in this

specification and their structural equivalents, or in combinations of one or more of them. Embodiments of the subject matter described in this specification can be implemented as one or more computer programs, i.e., one or more modules of computer program instructions encoded on a tangible non-transitory storage medium for execution by, or to control the operation of, data processing apparatus. The computer storage medium can be a machine-readable storage device, a machine-readable storage substrate, a random or serial access memory device, or a combination of one or more of them. Alternatively or in addition, the program instructions can be encoded on an artificially-generated propagated signal, e.g., a machine-generated electrical, optical, or electromagnetic signal, that is generated to encode information for transmission to suitable receiver apparatus for execution by a data processing apparatus.

[0144] The term "data processing apparatus" refers to data processing hardware and encompasses all kinds of apparatus, devices, and machines for processing data, including by way of example a programmable processor, a computer, or multiple processors or computers. The apparatus can also be, or further include, special purpose logic circuitry, e.g., an FPGA (field programmable gate array) or an ASIC (application-specific integrated circuit). The apparatus can optionally include, in addition to hardware, code that creates an execution environment for computer programs, e.g., code that constitutes processor firmware, a protocol stack, a database management system, an operating system, or a combination of one or more of them.

[0145] A computer program, which may also be referred to or described as a program, software, a software application, an app, a module, a software module, a script, or code, can be written in any form of programming language, including compiled or interpreted languages, or declarative or procedural languages; and it can be deployed in any form, including as a stand-alone program or as a module, component, subroutine, or other unit suitable for use in a computing environment. A program may, but need not, correspond to a file in a file system. A program can be stored in a portion of a file that holds other programs or data, e.g., one or more scripts stored in a markup language document, in a single file dedicated to the program in question, or in multiple coordinated files, e.g., files that store one or more modules, sub-programs, or portions of code. A computer program can be deployed to be executed on one computer or on multiple computers that are located at one site or distributed across multiple sites and interconnected by a data communication network.

[0146] In this specification the term "engine" is used broadly to refer to a software-based system, subsystem, or process that is programmed to perform one or more specific functions. Generally, an engine will be implemented as one or more software modules or components, installed on one or more computers in one or more locations. In some cases, one or more computers will be dedicated to a particular engine; in other cases, multiple engines can be installed and running on the same computer or computers.

[0147] The processes and logic flows described in this specification can be performed by one or more programmable computers executing one or more computer programs to perform functions by operating on input data and generating output. The processes and logic flows can also be performed by special purpose logic circuitry, e.g., an FPGA or an ASIC, or by a combination of special purpose logic circuitry and one or more programmed computers.

[0148] Computers suitable for the execution of a computer program can be based on general or special purpose microprocessors or both, or any other kind of central processing unit. Generally, a central processing unit will receive instructions and data from a read-only memory or a random access memory or both. The essential elements of a computer are a central processing unit for performing or executing instructions and one or more memory devices for storing instructions and data. The central processing unit and the memory can be supplemented by, or incorporated in, special purpose logic circuitry. Generally, a computer will also include, or be operatively coupled to receive data from or transfer data to, or both, one or more mass storage devices for storing data, e.g., magnetic, magneto-optical disks, or optical disks. However, a computer need not have such devices. Moreover, a computer can be embedded in another device, e.g., a mobile telephone, a personal digital assistant (PDA), a mobile audio or video player, a game console, a Global Positioning System (GPS) receiver, or a portable storage device, e.g., a universal serial bus (USB) flash drive, to name just a few.

[0149] Computer-readable media suitable for storing computer program instructions and data include all forms of non-volatile memory, media and memory devices, including by way of example semiconductor memory devices, e.g., EPROM, EEPROM, and flash memory devices; magnetic disks, e.g., internal hard disks or removable disks; magneto-optical disks; and CD-ROM and DVD-ROM disks.

[0150] To provide for interaction with a user, embodiments of the subject matter described in this specification can be implemented on a computer having a display device, e.g., a CRT

(cathode ray tube) or LCD (liquid crystal display) monitor, for displaying information to the user and a keyboard and a pointing device, e.g., a mouse or a trackball, by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback, e.g., visual feedback, auditory feedback, or tactile feedback; and input from the user can be received in any form, including acoustic, speech, or tactile input. In addition, a computer can interact with a user by sending documents to and receiving documents from a device that is used by the user; for example, by sending web pages to a web browser on a user's device in response to requests received from the web browser. Also, a computer can interact with a user by sending text messages or other forms of message to a personal device, e.g., a smartphone that is running a messaging application, and receiving responsive messages from the user in return.

[0151] Data processing apparatus for implementing machine learning models can also include, for example, special-purpose hardware accelerator units for processing common and compute-intensive parts of machine learning training or production, i.e., inference, workloads.

[0152] Machine learning models can be implemented and deployed using a machine learning framework, e.g., a TensorFlow framework or a JAX framework.

[0153] Embodiments of the subject matter described in this specification can be implemented in a computing system that includes a back-end component, e.g., as a data server, or that includes a middleware component, e.g., an application server, or that includes a front-end component, e.g., a client computer having a graphical user interface, a web browser, or an app through which a user can interact with an implementation of the subject matter described in this specification, or any combination of one or more such back-end, middleware, or front-end components. The components of the system can be interconnected by any form or medium of digital data communication, e.g., a communication network. Examples of communication networks include a local area network (LAN) and a wide area network (WAN), e.g., the Internet.

[0154] The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other. In some embodiments, a server transmits data, e.g., an HTML page, to a user device, e.g., for purposes of displaying data to and receiving user input from a user interacting with the device, which acts as a client. Data

generated at the user device, e.g., a result of the user interaction, can be received at the server from the device.

[0155] While this specification contains many specific implementation details, these should not be construed as limitations on the scope of any invention or on the scope of what may be claimed, but rather as descriptions of features that may be specific to particular embodiments of particular inventions. Certain features that are described in this specification in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable subcombination. Moreover, although features may be described above as acting in certain combinations and even initially be claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to a subcombination or variation of a subcombination.

[0156] Similarly, while operations are depicted in the drawings and recited in the claims in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Moreover, the separation of various system modules and components in the embodiments described above should not be understood as requiring such separation in all embodiments, and it should be understood that the described program components and systems can generally be integrated together in a single software product or packaged into multiple software products.

[0157] Particular embodiments of the subject matter have been described. Other embodiments are within the scope of the following claims. For example, the actions recited in the claims can be performed in a different order and still achieve desirable results. As one example, the processes depicted in the accompanying figures do not necessarily require the particular order shown, or sequential order, to achieve desirable results. In some cases, multitasking and parallel processing may be advantageous.

CLAIMS

1. A method performed by one or more computers, the method comprising:

obtaining a respective observation characterizing a state of an environment for each time step in a sequence of multiple time steps, comprising, for each time step after a first time step in the sequence of time steps:

processing a network input that comprises observations obtained for one or more preceding time steps using a selection neural network to generate a plurality of acquisition decisions,

wherein each acquisition decision corresponds to a respective modality from a set of multiple modalities and defines whether data corresponding to the modality is selected for acquisition at the time step;

obtaining an observation for the time step, wherein the observation: (i) includes data corresponding to modalities, from the set of modalities, that are selected for acquisition at the time step, (ii) does not include data corresponding to modalities, from the set of modalities, that are not selected for acquisition at the time step; and

processing a model input that includes the observation for each time step in the sequence of time steps using a prediction model to generate a prediction characterizing the environment.

2. The method of claim 1, further comprising:

determining an acquisition cost based on the respective modalities selected for acquisition at each time step in the sequence of time steps;

determining a reward based at least in part on the acquisition cost; and

training the selection neural network based on the reward using a reinforcement learning technique.

3. The method of claim 2, wherein each modality in the set of modalities is associated with a respective cost factor, and wherein determining the acquisition cost comprises:

determining, for each time step in the sequence of time steps, a respective acquisition cost for the time step based on the respective cost factor associated with each modality selected for acquisition at the time step; and

determining the acquisition cost as a combination of the acquisition costs for the time steps.

4. The method of claim 3, wherein for each time step in the sequence of time steps, determining the acquisition cost for the time step comprises:

determining the acquisition cost for the time step as a sum of the cost factor associated with each modality selected for acquisition at the time step.

5. The method of any one of claims 3-4, wherein determining the acquisition cost as a combination of the acquisition costs for the time steps comprises:

determining the acquisition costs as a sum over the acquisition costs for the time steps.

6. The method of any one of claims 3-5, wherein for one or more of the modalities, the cost factor for the modality is based at least in part on an amount of resource usage required to capture data corresponding to the modality.

7. The method of claim 6, wherein the resource usage required to capture data corresponding to the modality characterizes at least energy usage required to capture data corresponding to the modality.

8. The method of any one of claims 6-7, wherein the resource usage required to capture data corresponding to the modality characterizes at least an amount of time required to capture data corresponding to the modality.

9. The method of any one of claims 3-8, wherein for one or more of the modalities, the cost factor for the modality is based at least in part on a risk associated with capturing data corresponding to the modality.

10. The method of claim 9, wherein the environment comprises a patient, and the risk associated with capturing data corresponding to the modality is based at least in part on a medical risk to the patient resulting from capturing data corresponding to the modality.

11. The method of any one of claims 2-10, further comprising:

determining a prediction error that measures an error in the prediction generated by the prediction model; and

determining the reward based on both: (i) the acquisition cost, and (ii) the prediction error.

12. The method of any preceding claim, wherein the prediction model is a machine learning model.

13. The method of claim 12, wherein the prediction model comprises a neural network.

14. The method of any one of claims 12-13, further comprising:

training the prediction machine learning model to optimize an objective function that depends on a prediction error of the prediction machine learning model.

15. The method of any preceding claim, wherein for each time step in the sequence of time steps, the network input to the selection neural network at the time step further comprises:

data identifying the acquisition decision for any modality at any preceding time step.

16. The method of any one of claims 2-15, further comprising:

for each of one or more time steps in the sequence of time steps:

processing a model input that includes the observation for the time step and observations for one or more preceding time steps in the sequence of time steps using the prediction model to generate an intermediate prediction characterizing the environment; and

determining an intermediate prediction error that measures an error in the intermediate prediction generated by the prediction model; and

determining the reward based at least in part on the intermediate prediction errors.

17. The method of any preceding claim, wherein the set of modalities includes an imaging modality, and wherein data corresponding to the imaging modality comprises image data.

18. The method of claim 17, wherein the set of modalities includes a medical imaging modality.

19. The method of any preceding claim, wherein the environment is a medical environment that comprises a patient.

20. The method of claim 19, wherein the prediction characterizing the environment comprises a predicted medical diagnosis of the patient.

21. The method of any one of claims 19-20, wherein the prediction characterizing the environment comprises a prediction for a medical treatment to be applied to the patient.

22. The method of any preceding claim, further comprising, for each time step after the first time step in the sequence of time steps:

determining that: (i) data corresponding to modalities, from the set of modalities, that are selected for acquisition at the time step will be included in the observation for the time step, and (ii) data corresponding to modalities, from the set of modalities, that are not selected for acquisition at the time step will not be included in the observation for the time step.

23. The method claim 22, wherein for each of one or more time steps after the first time step in the sequence of time steps:

only a proper subset of the modalities in the set of modalities are selected for acquisition at the time step.

24. The method of any one of claims 22-23, further comprising, for each time step after the first time step in the sequence of time steps:

causing data to be acquired only for modalities selected for acquisition at the time step.

25. A system comprising:

one or more computers; and

one or more storage devices communicatively coupled to the one or more computers, wherein the one or more storage devices store instructions that, when executed by the one or more computers, cause the one or more computers to perform operations of the respective method of any one of claims 1-24.

26. One or more non-transitory computer storage media storing instructions that when executed by one or more computers cause the one or more computers to perform operations of the respective method of any one of claims 1-24.

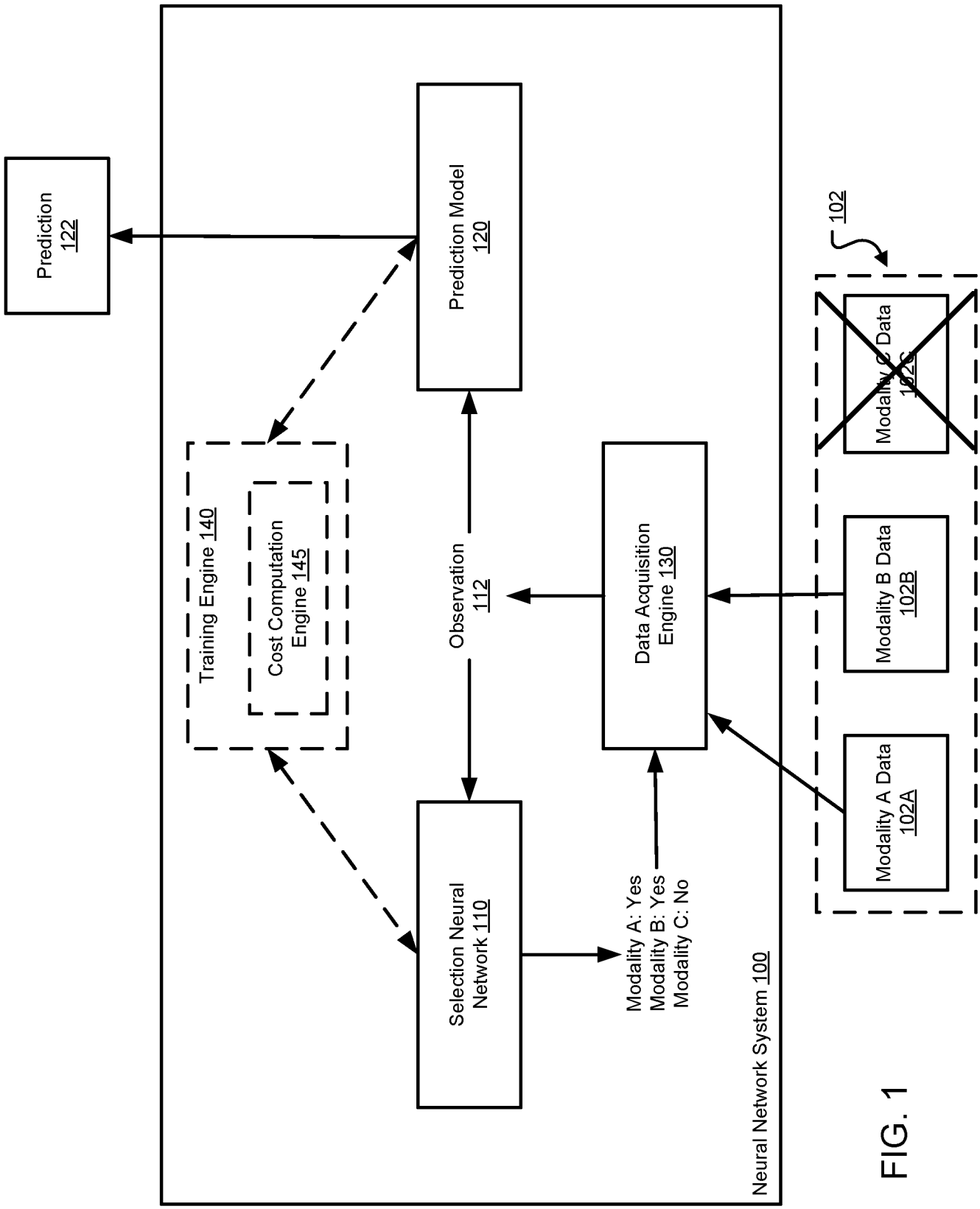
Prediction 122
Prediction Model 120
Training Engine 140
Cost Computation Engine 145
Observation 112
Data Acquisition Engine 130
Selection Neural Network 110
Modality A: Yes
Modality B: Yes
Modality C: No
Neural Network System 100
102
Modality A Data 102A
Modality B Data 102B
Modality C Data 102C

FIG. 1

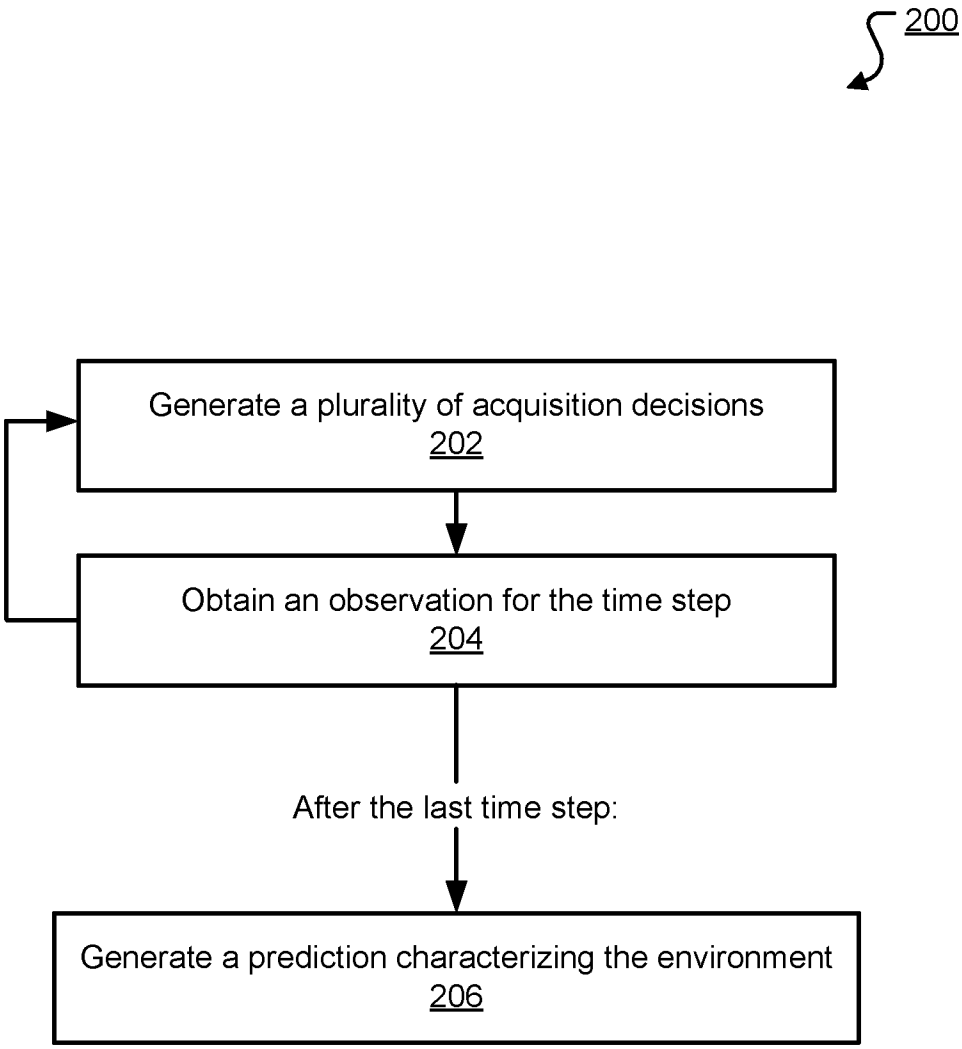
200
Generate a plurality of acquisition decisions
202
Obtain an observation for the time step
204
After the last time step:
Generate a prediction characterizing the environment
206

FIG. 2

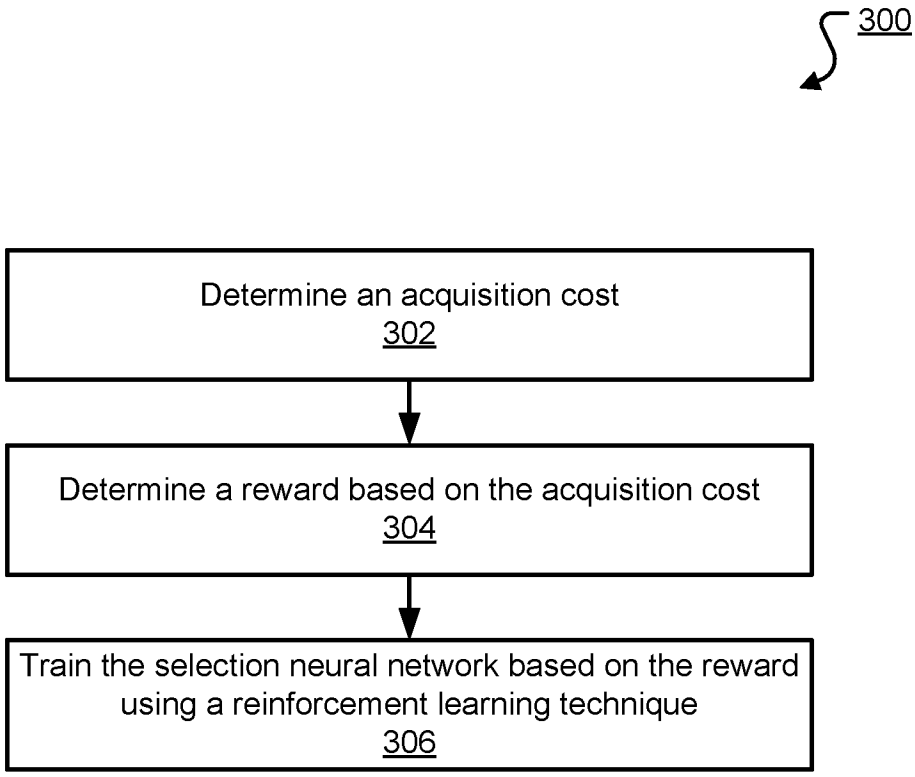
300
Determine an acquisition cost
302
Determine a reward based on the acquisition cost
304
Train the selection neural network based on the reward using a reinforcement learning technique
306

FIG. 3

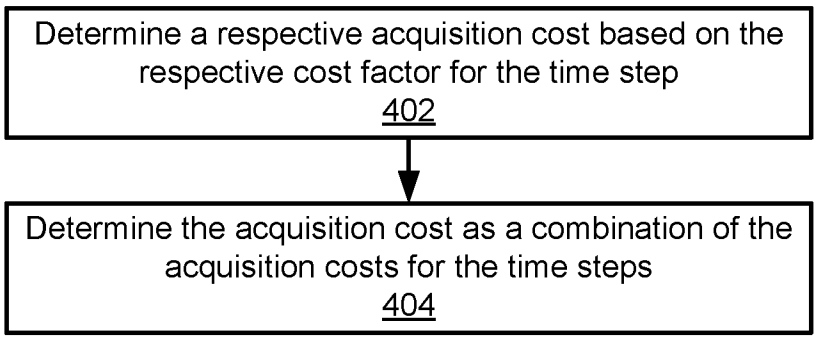
Determine a respective acquisition cost based on the respective cost factor for the time step
402
Determine the acquisition cost as a combination of the acquisition costs for the time steps
404

FIG. 4

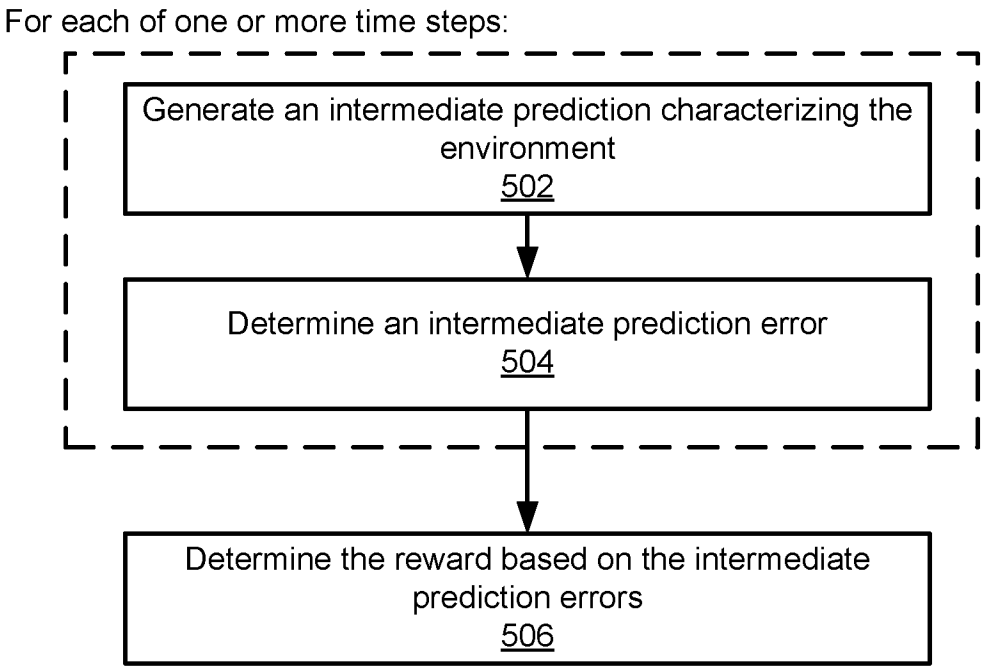
For each of one or more time steps:
Generate an intermediate prediction characterizing the environment
502
Determine an intermediate prediction error
504
Determine the reward based on the intermediate prediction errors
506

FIG. 5

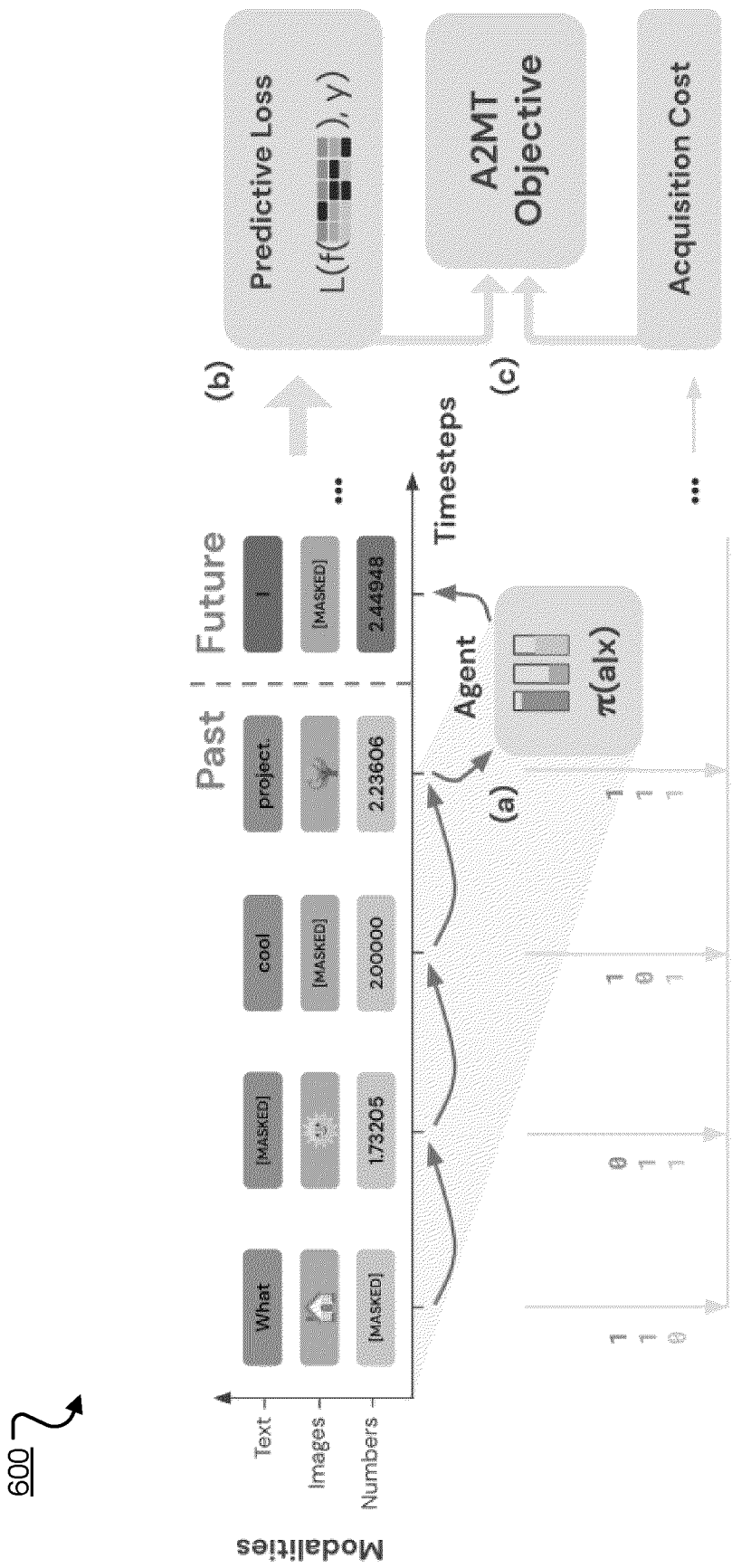
600
Modalities
Text
Images
Numbers
Past
Future
Timesteps
What
[MASKED]
1.73205
[MASKED]
cool
[MASKED]
2.00000
project.
2.23606
1
[MASKED]
2.44948
Agent
(a)
π(a|x)
(b)
Predictive Loss
L(f(
), y)
(c)
A2MT
Objective
Acquisition Cost

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2023/079389

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N3/045 G06N3/092 G06N5/022
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06N

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)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YANG LI ET AL: "Active Feature Acquisition with Generative Surrogate Models", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 11 February 2021 (2021-02-11), XP081877434, abstract Chapter 1. Introduction; page 1 - page 2; figures 1, 2 Chapter 2. Methods; page 2 - page 5; figure 3 Section Medical Diagnosis; page 7 Algorithm 2; page 12 section B.3 Medical Diagnosis; page 12 -/--	1-26

[X] 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
23 January 2024	02/02/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Hasnas, Sergiu

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2023/079389

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Section C. Hyperparameters; page 13 -----	
X	CHUN-HAO CHANG ET AL: "Dynamic Measurement Scheduling for Event Forecasting using Deep RL", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 24 January 2019 (2019-01-24), XP081009017, abstract Chapter 1. Introduction - 3. Methods; page 1 - page 5; figure 1 page 6, column 1, paragraph 3; figure 3 page 11, paragraph 2 -----	1-26

1