

(51) International Patent Classification:
G06K 9/62 (2006.01)(21) International Application Number:
PCT/EP2020/050083(22) International Filing Date:
03 January 2020 (03.01.2020)

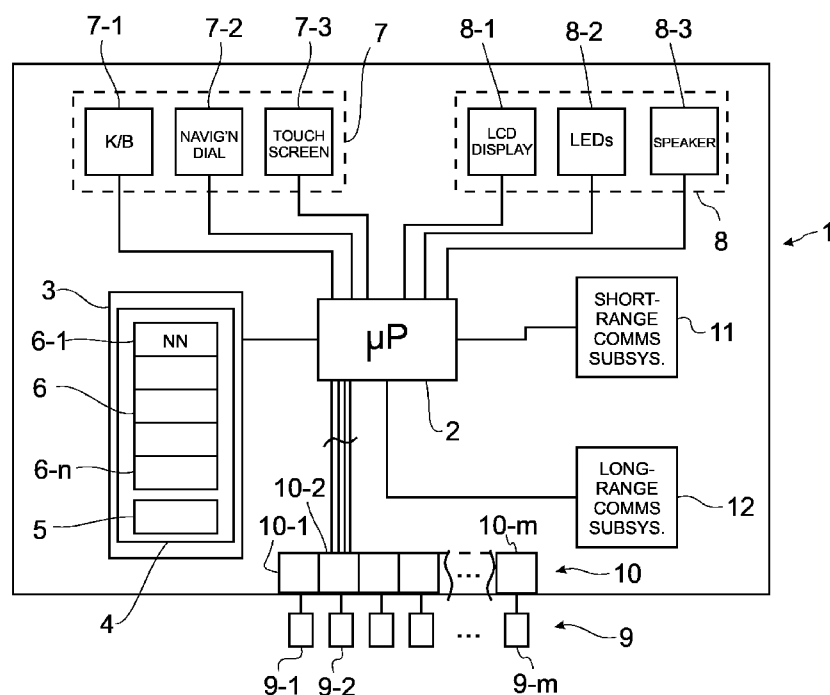
(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
LU101088 04 January 2019 (04.01.2019) LU(71) Applicant: **IEE INTERNATIONAL ELECTRONICS & ENGINEERING S.A.** [LU/LU]; Zone industrielle, 12, rue Pierre Richardot, 6468 Echternach (LU).(72) Inventors: **BEISE, Hans-Peter**; Im Kemmer 2, 66706 Perl (DE). **SCHRÖDER, Udo**; In der Rotheck 35, 54343 Föhren (DE). **DIAS DA CRUZ, Steve**; 21, Rue Widderberg, 6675 Mertert (LU). **SOKOLOWSKI, Jan**; Trierer Strasse 17, 54331 Pellingen (DE).(74) Agent: **BEISSEL, Jean** et al.; Office Freylinger S.A., 234, Route d'Arlon, BP 48, 8001 Strassen (LU).

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

(54) Title: METHOD AND SYSTEM FOR PROCESSING NEURAL NETWORK PREDICTIONS IN THE PRESENCE OF ADVERSE PERTURBATIONS



(57) **Abstract:** A method of processing predictions in the presence of adversarial perturbations in a sensing system comprising a processor and, coupled thereto, a memory. The processor may be configured to connect to one or more sensors for receiving inputs (x) therefrom. The processor may be configured to run a module in the memory for implementing a neural network. The neural network may have a network function f_{θ} , where θ are network parameters. The method may be executed by the processor and comprise generating, from the inputs (x) including at least a given input (x_0), respective outputs, the outputs being predictions of the neural network and including a given output y_0 corresponding to the given input (x_0), where $y_0 = f_{\theta}(x_0)$. The method may further comprise generating, from a plurality of outputs including the given output y , a measurement quantity (m). The measurement



SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR,
TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

quantity m may be, at or near the given input (x_0) , (i) a first measurement quantity M_1 corresponding to a gradient of the given output y , (ii) a second measurement quantity M_2 corresponding to a gradient of a predetermined objective function derived from a training process for the neural network, or (iii) a third measurement quantity M_3 derived from a combination of M_1 and M_2 . The method may further comprise determining whether the measurement quantity (m) is equal to or greater than a threshold. The method may further comprise, if the measurement quantity (m) is determined to be equal to or greater than the threshold, performing one or more remedial actions to correct for a perturbation. A method of classification based on the method is also disclosed. A corresponding sensing and/or classifying system, and a vehicle incorporating the sensing and/or classifying system, is also disclosed.

Method and System for Processing Neural Network Predictions in the Presence of Adverse Perturbations

Technical field

[0001] The present invention generally relates to the detection in sensing systems based on neural networks. More particularly, the present invention relates to sensing and/or classifying method and system, for processing predictions and/or classifications in the presence of adversarial perturbations.

Background of the Invention

[0002] The present invention finds application in any sensing system, as for example used in the automotive sector, which employs a neural network (NN) for classification/prediction purposes.

[0003] As is known, neural network models can be viewed as mathematical models defining a function $f: X \rightarrow Y$. It is known in the art that, besides the great potential of (deep-)neural networks, these functions are vulnerable to adversarial perturbations (c.f. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*). That is, correctly classified samples can be slightly perturbed in a way that the classification changes tremendously and becomes wrong. Such perturbations can be the result of an adversarial attack, but they can also occur by chance. Hence, it is necessary, in particular for safety critical applications, to have mechanisms to detect such perturbed inputs in order to interpret the corresponding classification accordingly.

[0004] The role of a derivative of the network function with respect to the input has been discussed in (i) Hein, M., & Andriushchenko, M. (2017). Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems* (pp. 2266-2276, and in (ii) Simon-Gabriel, C. J., Ollivier, Y., Schölkopf, B., Bottou, L., & Lopez-Paz, D. (2018). Adversarial Vulnerability of Neural Networks Increases With Input Dimension. *arXiv preprint arXiv:1802.01421*.

Object of the invention

[0005] A problem addressed by the present invention is how to provide effective neural network-based sensing and/or classifying methods and systems that reduce or eliminate the effects of the presence of adversarial perturbations upon predictions and/or classifications.

General Description of the Invention

[0006] In order to overcome the abovementioned problems, in one aspect there is provided a method of processing predictions in the presence of adversarial perturbations in a sensing system comprising a processor and, coupled thereto, a memory. It should be noted that in the context of the invention the expressions "processor" and "memory" are not limited to specific implementations of the processing environment. The processor and memory may e.g. be standard processors used in computers or common computing devices. On the other hand, the skilled person will appreciate that a neural network may implemented in some other hardware device that might be dedicated to neural networks (devices with a network structure burned into their circuitry are expected to be available in the future). These and other possible implementations of "processor" and "memory" devices are also encompassed by the expressions.

[0007] The processor may be configured to connect to one or more sensors for receiving inputs (x) therefrom. The processor may be configured to run a module in the memory for implementing a neural network. The neural network may have a network function f_{θ} , where θ are network parameters. The method may be executed by the processor and comprise generating, from the inputs (x) including at least a given input (x_0), respective outputs, the outputs being predictions of the neural network and including a given output y corresponding to the given input (x_0), where $y = f_{\theta}(x_0)$. The method may further comprise generating, from a plurality of outputs including the given output y , a measurement quantity (m). The measurement quantity m may be, at or near the given input (x_0), (i) a first measurement quantity M_1 corresponding to a gradient of the given output y , (ii) a second measurement quantity M_2 corresponding to a gradient of a predetermined objective function derived from a training process for the neural network, or (iii) a third measurement quantity M_3 derived from a combination of M_1 and M_2 . The

method may further comprise determining whether the measurement quantity (m) is equal to or greater than a threshold. The method may further comprise, if the measurement quantity (m) is determined to be equal to or greater than the threshold, performing one or more remedial actions to correct for a perturbation.

[0008] Preferably, the method further comprises, if the measurement quantity (m) is determined to be less than the threshold, performing a predetermined usual action resulting from y .

[0009] In an embodiment, generating the first measurement quantity M_1 comprises: computing a gradient $D_x f_\theta$ of the network function f_θ with respect to the input (x); and deriving the first measurement quantity M_1 as the value of gradient $D_x f_\theta$ corresponding to the given input (x_0). Preferably, deriving the first measurement quantity M_1 comprises determining the Euclidean norm of $D_x f_\theta$ corresponding to the given input (x_0).

[0010] In an embodiment, generating the second measurement quantity M_2 comprises: computing a gradient $D_\theta J(X, Y, f_\theta)$ of the objective function by $J(X, Y, f_\theta)$ with respect to the network parameters θ , whereby $J(X, Y, f_\theta)$ has been previously obtained by calibrating the network function f_θ in an offline training process based on given training data; and deriving the second measurement quantity M_2 as the value of gradient $D_\theta J(X, Y, f_\theta)$ corresponding to the given input (x_0). Preferably, deriving the second measurement quantity M_2 comprises determining the Euclidean norm of $D_\theta J(X, Y, f_\theta)$ corresponding to the given input (x_0).

[0011] In embodiments, the third measurement quantity M_3 is computed as a weighted sum of the first measurement quantity M_1 and the second measurement quantity M_2 .

[0012] The first measurement quantity M_1 , the second measurement quantity M_2 and/or the third measurement quantity M_3 may be generated based on a predetermined neighborhood of inputs (x) including the given input (x_0). Preferably, the predetermined neighborhood of inputs includes a first plurality of inputs prior to the the given input (x_0) and/or a second plurality of inputs after the the given input (x_0). Preferably, the number in the first plurality and/or the second plurality is 2-10, more preferably 2-5, more preferably 2-3.

[0013] In an embodiment, the one or more remedial actions comprise saving the value of $f_{\theta}(x_0)$ and wait for a next output $f_{\theta}(x_1)$ in order to verify $f_{\theta}(x_0)$ or to determine that it was a false output.

[0014] In an embodiment, the sensing system includes one or more output devices, and the one or more remedial actions comprise stopping the sensing system and issuing a corresponding warning notice via an output device.

[0015] In an embodiment, the one or more remedial actions comprise rejecting the prediction $f_{\theta}(x_0)$ and stopping any predetermined further actions that would result from that prediction.

[0016] According to another aspect, there is provided a method of classifying outputs of a sensing system employing a neural network, the method comprising, if the measurement quantity (m) is determined to be less than the threshold, performing a predetermined usual action resulting from y , wherein the predetermined usual action or the predetermined further actions comprise determining a classification or a regression based on the prediction y .

[0017] Preferably, the sensing system includes one or more output devices and one or more input devices, and wherein the method further comprises: outputting via an output device a request for a user to approve or disapprove a determined classification, and receiving a user input via an input device, the user input indicating whether the determined classification is approved or disapproved.

[0018] According to another aspect, there is provided a sensing and/or classifying system, for processing predictions and/or classifications in the presence of adversarial perturbations, the sensing and/or classifying system comprising: a processor and, coupled thereto, a memory, wherein the processor is configured to connect to one or more sensors for receiving inputs (x) therefrom, wherein the processor is configured to run a module in the memory for implementing a neural network, the neural network having a network function f_{θ} , where θ are network parameters, and wherein the processor, is configured to execute one or more embodiments of the method as described above.

[0019] According to another aspect of the invention there is provided a vehicle comprising a sensing and/or classifying system as described above.

[0020] The invention, at least in embodiments, provides a method that supports the robustness and safety of systems that implement a neural network for classification purposes. To this end, a method is formulated to measure whether a sample at hand (x_0) might be located in a region of the input space where the neural network does not perform in a reliable manner. Beneficially, the disclosed techniques exploit the analytical properties of the neural network. More precisely, the disclosed techniques implement the gradients of the neural network which then deliver sensitivity information about the decision at a given sample.

[0021] An advantage of the invention, at least in embodiments, is to reduce or eliminate the effects of the presence of adversarial perturbations upon predictions and/or classifications.

[0022] A further advantage of the invention, at least in embodiments, is that by deriving analytical characteristics from the neural network, determination of whether the neural network might have had difficulties in performing a reliable prediction is enabled.

[0023] Yet further advantages of the invention, at least in embodiments, include the following: (i) analytical properties of neural network-function may be used to measure reliability; (ii) two measures based on gradients of the neural network and on the underlying objective function used during training, are employed and can be combined to a common criterion for reliability; (iii) robustness measures are tailored to the actual neural network (directly based on the actual neural network); and (iv) the technique is applicable to any domain where neural networks are employed.

Brief Description of the Drawings

[0024] Further details and advantages of the present invention will be apparent from the following detailed description of not limiting embodiments with reference to the attached drawing, wherein:

Fig.1 is a schematic block diagram of a neural network-based sensing and/or classifying system according to an embodiment of the invention; and

Fig. 2 schematically represents the operation of the neural network-based sensing and/or classifying system of Fig. 1.

Description of Preferred Embodiments

[0025] In the drawings, like reference numerals have been used to denote like elements. Any features, components, operations, steps or other elements of one embodiment may be used in combination with those of any other embodiment disclosed herein, unless indicated otherwise hereinbelow.

[0026] Fig. 1 is a schematic block diagram of a neural network-based sensing and/or classifying system 1 (hereafter also “system”) according to an embodiment of the invention.

[0027] The system 1 includes a processor 2 and, coupled thereto, one or memories including non-volatile memory (NVM) 3. In the NVM 3 may be stored various software 4 including operating system software 5 and/or one or more software modules 6-1 to 6-n (collectively modules 6). The modules 6 may include a neural network module 6-1 implementing a neural network, as discussed further hereinbelow.

[0028] In embodiments, for the purpose of interaction with a user, the system 1 may include one or more input devices 7 and one or more output devices 8. The input devices 7 may include a keyboard or keypad 7-1, a navigation dial or knob/button 7-2 and/or a touchscreen 7-3. The output devices 8 may include a display (e.g. LCD) 8-1, one or more illuminable indicators (e.g. LEDs) 8-2 and/or an audio output device (e.g. speaker) 8-3.

[0029] During operation of the neural network module 6-1, the processor 2 may receive input from one or more sensors 9-1, 9-2, ..., 9-m (collectively sensors 9), for example via respective interfaces 10-1, 10-2, ..., 10-m (collectively interfaces 10), which are thereafter further processed as discussed in more detail below.

[0030] Optionally, the system 1 includes a short-range (e.g. Bluetooth, ZigBee) communications subsystem 11 and/or a long-range (e.g. cellular, such as 4G, 5G) communications subsystem 12, each interface being for receipt and/or transmission of sensor or other data, control parameters, training data, or other system-related data, or for transmission of neural network predictions and/or classifications.

[0031] Fig. 2 schematically represents the operation of the neural network-based sensing and/or classifying system of Fig. 1.

[0032] Received at the neural network module 6-1 are successive inputs or samples x , received from sensors 9 via interfaces 10. In embodiments, the neural network module 6-1 may receive the inputs x as raw data, or as pre-processed sensor data through an appropriate pre-processing technique, such as amplification, filtering or other signal conditioning. While denoted simply as x , it will be appreciated that the inputs x may be in the form of signals disposed in an array or matrix corresponding to the configuration of sensors 9.

[0033] The underlying principles of the disclosed techniques will be discussed in the following.

[0034] For the purpose of illustration, under consideration is a general sensing system that receives data from one or several sensors 9. The system employs a neural network (NN) module 6-1 to make a prediction or classification regarding the environment or some physical quantity.

[0035] As an example, the following automotive and other scenarios are envisaged:

- Interior RADAR system (for vital signs);
- LIDAR, Camera and RADAR for exterior object detection;
- Camera based gesture recognition;
- Driver monitoring system; and
- Ultrasonic based systems.

[0036] It is further assumed that the system (NN module 6-1) uses a NN denoted by f_{θ} (with θ being the network parameters) that receives the raw or pre-processed sensor data (from one or several sensors 9), denoted by x , upon which it performs a prediction or classification.

[0037] Returning to the abovementioned example scenarios, the classification / prediction might be as follows:

- Interior RADAR system (for vital signs) -> small baby is present in the car;
- LIDAR, Camera and RADAR for exterior object detection -> cyclist detected;
- Camera based gesture recognition -> gesture with intension to start a phone call detected;
- Driver monitoring system -> driver is under influence of drugs; and/or
- Ultrasonic based systems -> environment recognition.

[0038] It is assumed that f_θ has been calibrated in an offline training process (based on given training data). This training process is (as it is usually done) performed by solving an optimization problem (fit training data to desired output) that is formulated by means of a certain objective function denoted by $J(X, Y, f_\theta)$. Here X denotes the set of training data and Y are the corresponding labels (desired output).

[0039] In use, the NN module 6-1 is operative, for each input by x , to generate or determine a corresponding output, so for a given input x_0 a given output y is determined as $y = f_\theta(x_0)$.

[0040] Returning to Fig. 2, in accordance with embodiments, further processing and/or evasive/remedial action is carried out by prediction processing module 6-a (from modules 6 in Fig. 1), based on the given output y and making use of one or more measurement quantities, as discussed further below. As seen in Fig. 2, depending on further determinations/operation based on the given output y and the one or more measurement quantities, a classification stage 6-b (e.g. from modules 6 in Fig. 1) may be operable to perform a classification based on the output from the NN module 6-1. The various embodiments and actions are discussed in the following.

[0041] In embodiments of the present invention, two characteristics of f_θ and $J(X, Y, f_\theta)$ that can be used in parallel or separately are defined and employed.

[0042] In a first embodiment, the gradient of the network function f_θ with respect to the input x , which is denoted by $D_x f_\theta$, is used.

[0043] Here, it is noted that, given an actual input x_0 during life-time (of operation of system 1), the magnitude of the entries in the gradient $D_x f_\theta(x_0)$ scales with the sensitivity of the classification in the neighbourhood of the sample x_0 . In other words, the higher the entries $D_x f_\theta(x_0)$ the more the output $f_\theta(x_0 + \delta)$ will change for certain perturbations δ . This in turn provides information that allows the determination of whether the input region around the sample x_0 constitutes a region of high fluctuation in the classification or not. This gives information about the reliability of the output $f_\theta(x_0)$.

[0044] In this first embodiment, therefore, a suitable quantity is derived from $D_x f_\theta(x_0)$ denoted by $M_1(D_x f_\theta(x_0))$ (with, for instance, M_1 the Euclidean norm). If

this quantity exceeds a predefined threshold, then the system can react accordingly (concrete reactions are formulated below).

[0045] In a second embodiment, there is employed $D_{\theta} J(X, Y, f_{\theta})$ – the gradient of the objective function with respect to the network parameters θ .

[0046] Here, given an actual input x_0 during life-time and the corresponding output $f_{\theta}(x_0) = y_0$, the magnitude of the entries in the gradient $D_{\theta} J(x_0, y_0, f_{\theta})$ provides information about whether the system would have learned something when the pair (x_0, y_0) would have been part of the training data. That is, the higher the entries in $D_{\theta} J(x_0, y_0, f_{\theta})$ the more the system could have learned from (x_0, y_0) . This in turn allows it to be concluded whether there has been sufficient training data in that input region and whether the system should be capable or not to classify the latter with a sufficiently high confidence. The underlying assumption is that an adversarial perturbation would have given information (high entries in $D_{\theta} J(x_0, y_0, f_{\theta})$) to the training process.

[0047] In this second embodiment, therefore, a quantity $M_2(D_{\theta} J(x_0, y_0, f_{\theta}))$ derived from $D_{\theta} J(x_0, y_0, f_{\theta})$ is used to quantify to what extent one can trust the output $f_{\theta}(x_0)$. Such quantity M_2 could for instance be the Euclidean norm or any other mathematical mapping to a size or length. If this quantity exceeds a predefined threshold, the system can react accordingly.

[0048] Both measures M_1, M_2 can also be evaluated in a reasonable neighbourhood around the sample x_0 . For example, a predetermined number of values obtained for samples (inputs) prior to and/or after input x_0 may be used.

[0049] If one or two of the proposed measures M_1, M_2 indicate that the prediction $f_{\theta}(x_0)$ is not reliable, then, in embodiments the following are remedial/evasive actions that may be executed:

Reject the prediction $f_{\theta}(x_0)$ and stop any further actions that would result from it (for instance classification);

Save the value of $f_{\theta}(x_0)$ and wait for a next output $f_{\theta}(x_1)$ in order to falsify or verify $f_{\theta}(x_0)$;

Stop the whole system and issue a corresponding warning notice; and/or

Ask a potential user to approve the classification.

[0050] For illustration, let $M(x, f_\theta)$ be one of the introduced quantities $M_1(D_x f_\theta(x_0))$, $M_2(D_\theta J(x_0, y_0, f_\theta))$, a combination (like weighted sum) of the latter, or any other useful mapping.

Then a pseudocode of the system may be as follows:

While live-time of the system

 receive sensor data x

$y \leftarrow f_\theta(x)$

$m \leftarrow M(x, f_\theta)$

 if $m < \text{confidence threshold}$ then

 perform usual action resulting from y

 else

 perform an evasive action

 end

end

[0051] While embodiments have been described by reference to embodiments of survey devices having various components in their respective implementations, it will be appreciated that other embodiments make use of other combinations and permutations of these and other components.

[0052] Reference throughout this specification to “one embodiment” or “an embodiment” means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases “in one embodiment” or “in an embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to one of ordinary skill in the art from this disclosure, in one or more embodiments.

[0053] Thus, while there has been described what are believed to be the preferred embodiments of the invention, those skilled in the art will recognize that other and further modifications may be made thereto without departing from the scope of the invention as defined by the claims, and it is intended to claim all such changes and modifications as fall within the scope of the invention. For example, any formulas given above are merely representative of procedures that may be

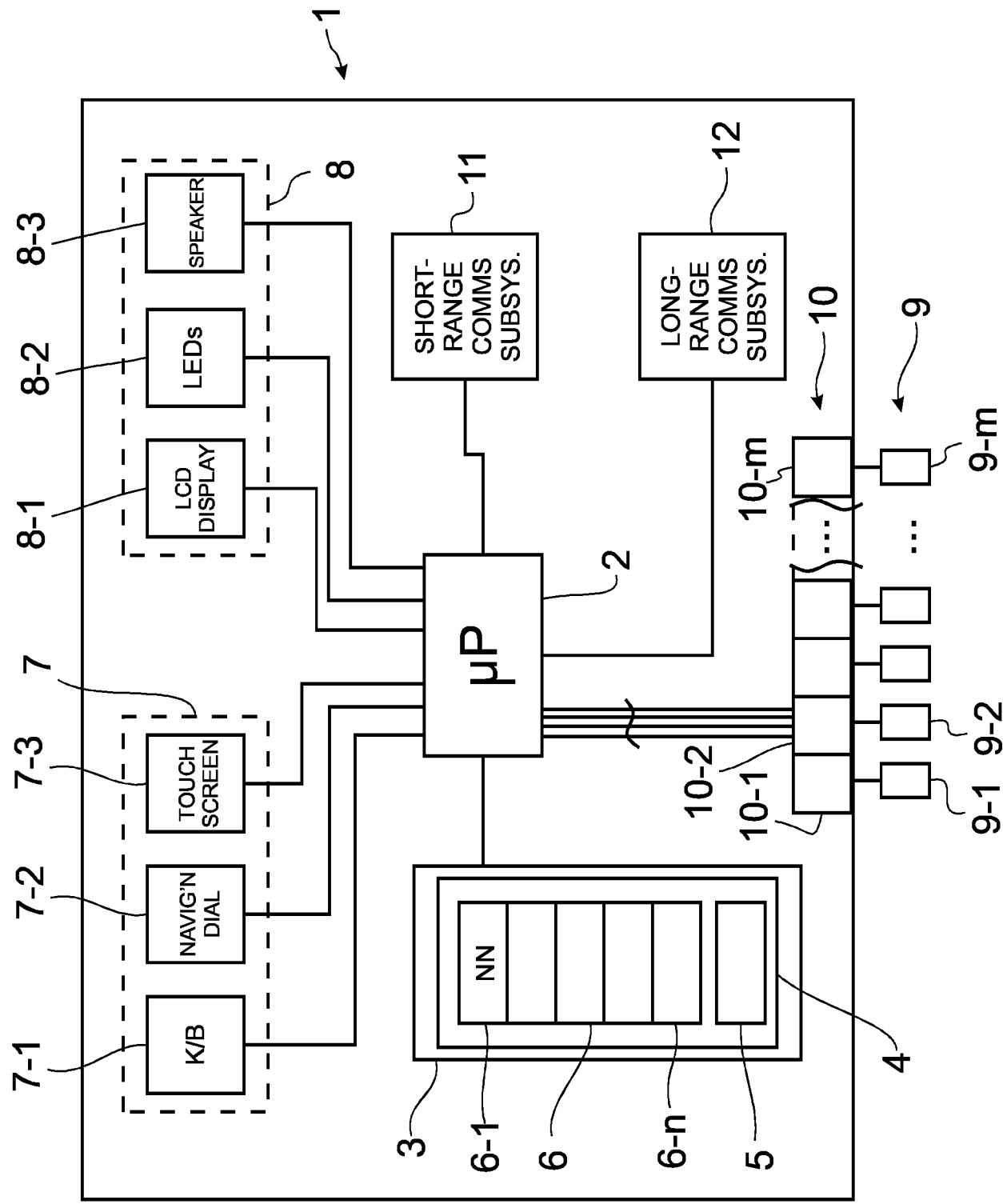
used. Functionality may be added or deleted from the block diagrams and operations may be interchanged among functional blocks. Steps may be added or deleted to methods described within the scope of the present invention.

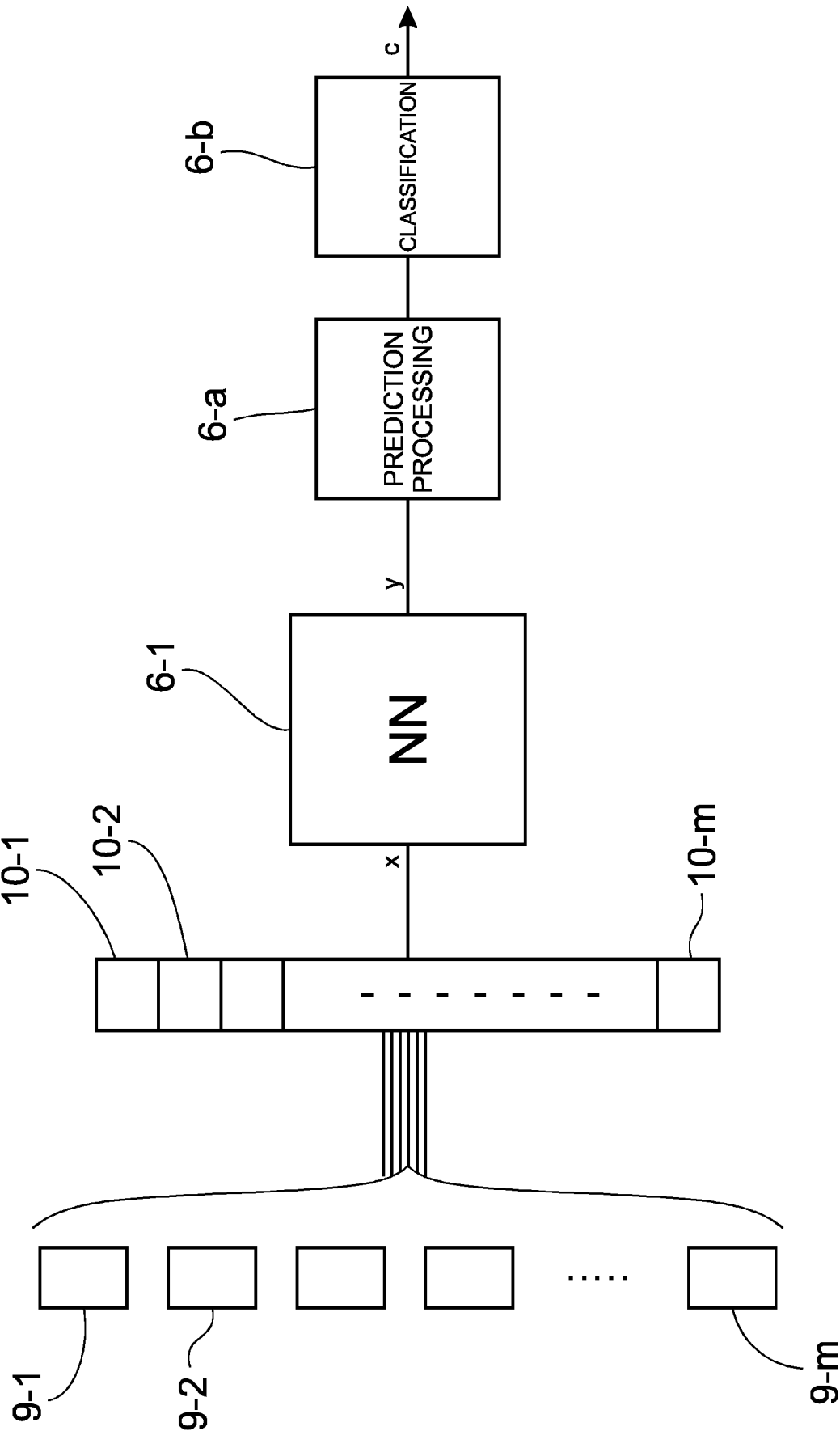
Claims

1. A method of processing predictions in the presence of adversarial perturbations in a sensing system comprising a processor and, coupled thereto, a memory, the processor being configured to connect to one or more sensors for receiving inputs (x) therefrom, the processor being configured to run a module in the memory for implementing a neural network, the neural network having a network function f_θ , where θ are network parameters, the method being executed by the processor and comprising:
 - generating, from the inputs (x) including at least a given input (x_0), respective outputs, the outputs being predictions of the neural network and including a given output y_0 corresponding to the given input (x_0), where $y_0 = f_\theta(x_0)$;
 - generating, from a plurality of outputs including the given output y_0 , a measurement quantity (m), where m is, at or near the given input (x_0), (i) a first measurement quantity M_1 corresponding to a gradient of the given output y , (ii) a second measurement quantity M_2 corresponding to a gradient of a predetermined objective function derived from a training process for the neural network, or (iii) a third measurement quantity M_3 derived from a combination of M_1 and M_2 ;
 - determining whether the measurement quantity (m) is equal to or greater than a threshold, and
 - if the measurement quantity (m) is determined to be equal to or greater than the threshold, performing one or more remedial actions to correct for a perturbation.
2. The method according to claim 1, further comprising, if the measurement quantity (m) is determined to be less than the threshold, performing a predetermined usual action resulting from y .
3. The method according to claim 1 or 2, wherein generating the first measurement quantity M_1 comprises:
 - computing a gradient $D_x f_\theta$ of the network function f_θ with respect to the input (x),

- deriving the first measurement quantity M_1 as the value of gradient $D_x f_\theta$ corresponding to the given input (x_0) .
4. The method according to claim 3, wherein deriving the first measurement quantity M_1 comprises determining the Euclidean norm of $D_x f_\theta$ corresponding to the given input (x_0) .
 5. The method according to claim 1 or 2, wherein generating the second measurement quantity M_2 comprises:
 - computing a gradient $D_\theta J(X, Y, f_\theta)$ of the objective function by $J(X, Y, f_\theta)$ with respect to the network parameters θ , whereby $J(X, Y, f_\theta)$ has been previously obtained by calibrating the network function f_θ in an offline training process based on given training data; and
 - deriving the second measurement quantity M_2 as the value of gradient $D_\theta J(X, Y, f_\theta)$ corresponding to the given input (x_0) .
 6. The method according to claim 5, wherein deriving the second measurement quantity M_2 comprises determining the Euclidean norm of $D_\theta J(X, Y, f_\theta)$ corresponding to the given input (x_0) .
 7. The method according to any of the preceding claims, wherein the third measurement quantity M_3 is computed as a weighted sum of the first measurement quantity M_1 and the second measurement quantity M_2 .
 8. The method according to any of the preceding claims, wherein the first measurement quantity M_1 , the second measurement quantity M_2 and/or the third measurement quantity M_3 is generated based on a predetermined neighborhood of inputs (x) including the given input (x_0) .
 9. The method according to claim 8, wherein the predetermined neighborhood of inputs includes a first plurality of inputs prior to the given input (x_0) and/or a second plurality of inputs after the given input (x_0) .
 10. The method according to claim 9, wherein the number in the first plurality and/or the second plurality is 2-10, more preferably 2-5, more preferably 2-3.
 11. The method according to any of the preceding claims, wherein the one or more remedial actions comprise saving the value of $f_\theta(x_0)$ and wait for a next output $f_\theta(x_1)$ in order to verify $f_\theta(x_0)$ or to determine that it was a false output.

12. The method according to any of the preceding claims, wherein the sensing system includes one or more output devices, and the one or more remedial actions comprise stopping the sensing system and issuing a corresponding warning notice via an output device.
- 5 13. The method according to any of the preceding claims, wherein the one or more remedial actions comprise rejecting the prediction $f_{\theta}(x_0)$ and stopping any predetermined further actions that would result from that prediction.
14. A method of classifying outputs of a sensing system employing a neural network, the method comprising the method according to claim 2, or any claim
10 dependent thereon; wherein the predetermined usual action or the predetermined further actions comprise determining a classification or a regression based on the prediction y .
15. The method according to claim 14, wherein the sensing system includes one or more output devices and one or more input devices, and wherein the
15 method further comprises:
- outputting via an output device a request for a user to approve or disapprove a determined classification, and
 - receiving a user input via an input device, the user input indicating whether the determined classification is approved or disapproved.
- 20 16. A sensing and/or classifying system, for processing predictions and/or classifications in the presence of adversarial perturbations, the sensing and/or classifying system comprising:
- a processor and, coupled thereto,
 - a memory,
 - 25 - wherein the processor is configured to connect to one or more sensors for receiving inputs (x) therefrom,
 - wherein the processor is configured to run a module in the memory for implementing a neural network, the neural network having a network function f_{θ} , where θ are network parameters, and
 - 30 - wherein the processor, is configured to execute the method of any of the preceding claims.
17. A vehicle comprising a sensing and/or classifying system according to claim 16.





INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2020/050083

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/62
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)
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)

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/249480 A1 (LEFEBVRE W CURT [US] ET AL) 9 December 2004 (2004-12-09) paragraphs [0006], [0012], [0013], [0020] - [0030]	1-17
A	----- CARL-JOHANN SIMON-GABRIEL ET AL: "Adversarial Vulnerability of Neural Networks Increases With Input Dimension", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 5 February 2018 (2018-02-05), XP081420357, cited in the application Sections 2 and 4 ----- -/--	1-17



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

31 March 2020

Date of mailing of the international search report

09/04/2020

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Miclea, Sorin

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2020/050083

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GAO JING ET AL: "Per-pixel bias-variance decomposition of continuous errors in data-driven geospatial modeling: A case study in environmental remote sensing", ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING, vol. 134, 1 December 2017 (2017-12-01), pages 110-121, XP085253934, ISSN: 0924-2716, DOI: 10.1016/J.ISPRSJPRS.2017.11.001 Sections 1, 2.1, 5 -----	1-17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2020/050083

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2004249480 A1	09-12-2004	US 2004249480 A1	09-12-2004
		US 2005216103 A1	29-09-2005
		US 2007118238 A1	24-05-2007
		WO 2004109468 A2	16-12-2004
