(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 4 266 216 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
25.10.2023 Bulletin 2023/43

(21) Application number: **22169013.4**

(22) Date of filing: **20.04.2022**

(51) International Patent Classification (IPC):
G06N 3/04 $^{(2023.01)}$ ***G06N 3/08*** $^{(2023.01)}$

(52) Cooperative Patent Classification (CPC):
G06N 3/084; G06F 21/6254; G06N 3/045;
G06N 3/048

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(71) Applicant: **Deeping Source Inc.**
Seoul 06152 (KR)

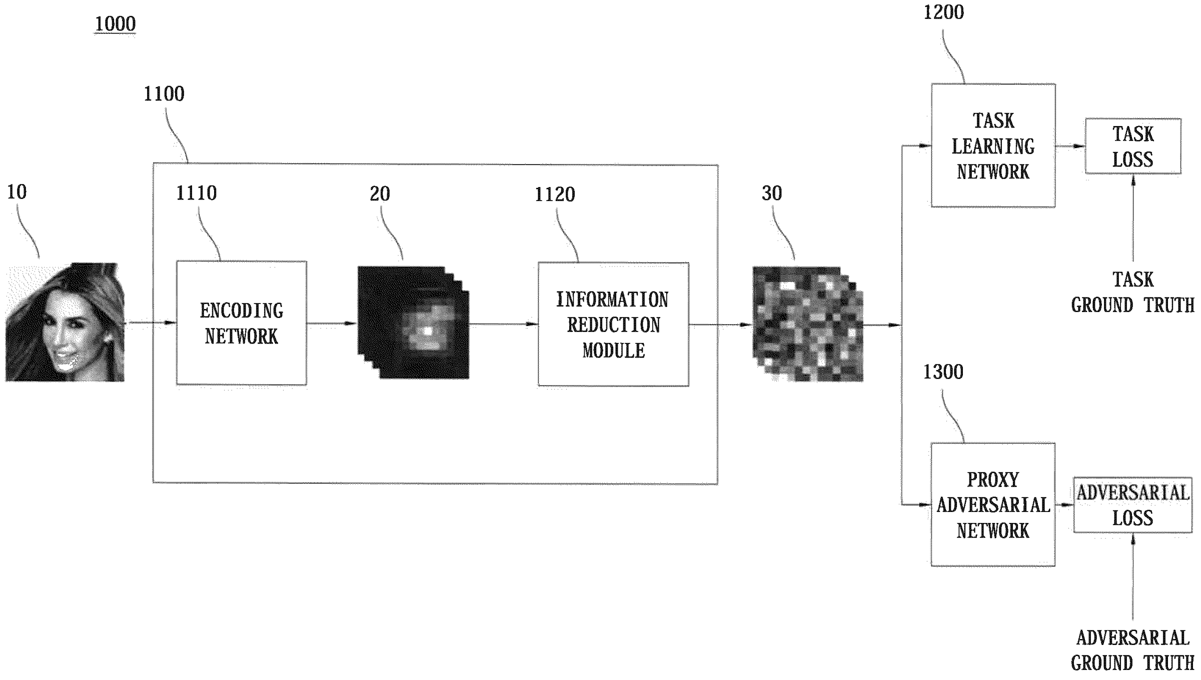
(72) Inventor: **JEONG, Jong Hu**
08748 Seoul (KR)

(74) Representative: **AWA Sweden AB**
Box 5117
200 71 Malmö (SE)

(54) LEARNING METHOD AND LEARNING DEVICE FOR TRAINING OBFUSCATION NETWORK CAPABLE OF OBFUSCATING ORIGINAL DATA FOR PRIVACY TO ACHIEVE INFORMATION RESTRICTION OBFUSCATION AND TESTING METHOD AND TESTING DEVICE USING THE SAME

(57) A learning method for training an obfuscation network, including steps of: (a) inputting a training data into the obfuscation network to (i) extract features and thus generate a data representation by performing a learning operation on the training data and (ii) transform the data representation and thus generate an anonymized data representation, and (b) inputting the anonymized data representation into a task learning network to (i) perform a task by using the anonymized data representation and thus output a task result, (ii) generate a task loss by referring to the task result and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

FIG. 3



Processed by Luminess, 75001 PARIS (FR)

Description

[0001] The present disclosure relates to a method for training an obfuscation network capable of obfuscating, e.g., anonymizing or concealing, an original data for privacy and a learning device using the same, and to a method for testing the trained obfuscation network capable of obfuscating the original data and a testing device using the same.

[0002] In recent years, machine learning as a service (MLaaS) has gained great popularity mainly due to the advances in cloud computing and deep learning. Oftentimes raw data generated on an edge device is sent to the cloud, where it is then processed and machine learning algorithms are applied to learn patterns from the data.

[0003] However, sending the raw data has the drawback that privacy-related information is directly leaked to the cloud server, which might constitute a violation of user privacy.

[0004] For example, we can consider the edge device transmitting images to the cloud, to perform person identification. While a picture of a person can be used for identification, the image can further reveal the person's gender, emotional state, race, or location. Ideally, privacy-related information should be removed from the image, while preserving task utility.

[0005] Additionally, such a private data representation should be secure against attacks from adversarial actors, who attempt to breach a user's privacy by retrieving private attributes from the private data representation. It is important to note that the service provider might be considered as a possible adversarial actor, so the client would want a model to fully remove utility unrelated information, since the representation transmitted from the client is out of their control.

[0006] To mitigate the leakage of sensitive attributes, many works have focused on the training framework of adversarial representation learning (ARL).

[0007] ARL has found its application in practical scenarios, such as information censoring, learning fair representations, the mitigation of information leakage, or collaborative inference. Commonly, the ARL framework consists of three entities, (1) an obfuscator, which transforms input data to representation that retains utility, while resolving the correlation of image features to sensitive attributes, (2) a task model, performing the utility task on the data representation, and (3) a proxy-adversary, attempting to extract sensitive attributes.

[0008] With the above scenario of MLaaS, the service provider trains the obfuscator and the task model and deploys the obfuscator to the user's client device. For the sake of the user's privacy, the obfuscator should effectively remove all information unrelated to the utility task and retain high utility with obfuscated representation.

[0009] However, previous ARL methods have focused on the task utility, i.e., privacy-utility trade-off.

[0010] In the previous ARL methods, the obfuscator is trained to generate obfuscated representation having a high task utility from the inputted data by removing all pieces of the information unrelated to the utility task.

[0011] Therefore, the trained obfuscator requires significant computations and computing resources to generate the obfuscated representation from the inputted data.

[0012] When such trained obfuscator requiring the significant computations and computing resources is installed on the edge device such as a mobile device with a limited computing resource, there is a problem in the obfuscator not functioning properly and consuming a lot of time to process the inputted data.

[0013] It is an object of the present disclosure to solve all the aforementioned problems.

[0014] It is another object of the present disclosure to provide an obfuscation network that is robust to attacks by adversarial actors while task utility is preserved.

[0015] It is still another object of the present disclosure to provide the obfuscation network that minimizes the use of computing resources in an edge device.

[0016] It is still yet another object of the present disclosure to provide the obfuscation network that can easily be applied to an off-the-shelf deep neural network (DNN).

[0017] It is still yet another object of the present disclosure to provide the obfuscation network that can easily be applied to a commonly used DNN.

[0018] In order to accomplish objects above, representative structures of the present disclosure are described as follows:

In accordance with one aspect of the present disclosure, there is provided a learning method for training an obfuscation network capable of obfuscating an original data for privacy, including steps of: (a) a learning device inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features to be used for performing a task of a task learning network and thus generate a data representation including the extracted features by performing a learning operation on the training data and (ii) transform the data representation and thus generate an anonymized data representation as an obfuscated data in which privacy-related information of the training data is protected and task utility is preserved; and (b) the learning device inputting the anonymized data representation into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation and thus output a task result, (ii) generate a task loss by referring to the task result and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

[0019] As one example, at the step of (a), the learning device inputs the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation by encoding the training data through an encoding network, and (ii) generate the anonymized data representation by reducing the features included in the data representation through an information reduction module.

[0020] As one example, the learning device instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation, (ii) a noise addition process by adding noise to the data representation, (iii) a random value replacement process by replacing parts of pixels of the data representation with a random value, (iv) a random shuffling process by shuffling position information of the data representation, and (v) a resizing process by resizing a cardinal number of pixels in the data representation to be smaller than a cardinal number of pixels in the training data, through the information reduction module.

[0021] As one example, the learning device instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation arbitrarily and thus generate an arbitrarily-resized data representation and (ii) in response to detecting that a size of the arbitrarily-resized data representation is bigger than a size of the training data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation, through the information reduction module.

[0022] As one example, the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the training data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

[0023] As one example, at the step of (b), the learning device inputs the anonymized data representation into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation and thus output an adversarial result in which the privacy-related information of a privacy-related region is estimated from the anonymized data representation, (ii) generate an adversarial loss by referring to the adversarial result and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0024] In accordance with another aspect of the present disclosure, there is provided a testing method for testing an obfuscation network capable of obfuscating an original data for privacy, including steps of: (a) on condition that a learning device has performed processes of (I) inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for training to be used for performing a task of a task learning network and thus generate a data representation for training including the extracted features for training by performing a learning operation on the training data and (ii) transform the data representation for training and thus generate an anonymized data representation for training as an obfuscated data in which privacy-related information for training of the training data is protected and task utility is preserved; and (II) inputting the anonymized data representation for training into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation for training and thus output a task result for training, (ii) generate a task loss by referring to the task result for training and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized, a testing device acquiring a test data; and (b) the testing device inputting the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for testing to be used for performing the task of the task learning network and thus generate a data representation for testing by performing the learning operation on the testing data and (ii) transform the data representation for testing and thus generate an anonymized data representation for testing as an obfuscated data in which privacy-related information for testing of the testing data is protected and task utility is preserved.

[0025] As one example, the testing method further includes a step of: (c) the testing device transmitting the anonymized data representation for testing to a server in which the task learning network is installed, to thereby instruct the server to acquire a task result for testing, wherein the task learning network performs the task by using the anonymized data representation for testing and thus generates the task result for testing.

[0026] As one example, at the step of (b), the testing device inputs the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation for testing by encoding the testing data through an encoding network, and (ii) generate the anonymized data representation for testing by reducing the features included in the data representation for testing through an information reduction module.

[0027] As one example, the testing device instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation for testing, (ii) a noise addition process by adding noise to the data representation for testing, (iii) a random value replacement process by replacing parts of pixels of the data representation for testing with a random value, (iv) a random shuffling process by shuffling position information of the data representation for testing, and (v) a resizing process by resizing a cardinal number of pixels in the data representation for testing to be smaller than a cardinal number of pixels in the testing data, through the information reduction module.

[0028] As one example, the testing device instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation for testing arbitrarily and thus generate an arbitrarily-resized data representation for testing and (ii) in response to detecting that a size of the arbitrarily-resized data representation for testing is bigger than a size of the testing data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation for testing, through the information reduction module.

[0029] As one example, the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the testing data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

[0030] As one example, at the step of (a), the learning device has inputted the anonymized data representation for training into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation for training and thus output an adversarial result for training in which the privacy-related information for training of a privacy-related region for training is estimated from the anonymized data representation for training, (ii) generate an adversarial loss by referring to the adversarial result for training and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0031] In accordance with still another aspect of the present disclosure, there is provided a learning device for training an obfuscation network capable of obfuscating an original data for privacy, including: at least one memory that stores instructions; and at least one processor configured to execute the instructions to perform processes of (I) inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features to be used for performing a task of a task learning network and thus generate a data representation including the extracted features by performing a learning operation on the training data and (ii) transform the data representation and thus generate an anonymized data representation as an obfuscated data in which privacy-related information of the training data is protected and task utility is preserved; and (II) inputting the anonymized data representation into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation and thus output a task result, (ii) generate a task loss by referring to the task result and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

[0032] As one example, at the process of (I), the processor inputs the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation by encoding the training data through an encoding network, and (ii) generate the anonymized data representation by reducing the features included in the data representation through an information reduction module.

[0033] As one example, the processor instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation, (ii) a noise addition process by adding noise to the data representation, (iii) a random value replacement process by replacing parts of pixels of the data representation with a random value, (iv) a random shuffling process by shuffling position information of the data representation, and (v) a resizing process by resizing a cardinal number of pixels in the data representation to be smaller than a cardinal number of pixels in the training data, through the information reduction module.

[0034] As one example, the processor instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation arbitrarily and thus generate an arbitrarily-resized data representation and (ii) in response to detecting that a size of the arbitrarily-resized data representation is bigger than a size of the training data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation, through the information reduction module.

[0035] As one example, the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the training data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

[0036] As one example, at the process of (II), the processor inputs the anonymized data representation into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation and thus output an adversarial result in which the privacy-related information of a privacy-related region is estimated from the anonymized data representation, (ii) generate an adversarial loss by referring to the adversarial result and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0037] In accordance with still yet another aspect of the present disclosure, there is provided a testing device for testing an obfuscation network capable of obfuscating an original data for privacy, including: at least one memory that stores instructions; and at least one processor configured to execute the instructions to perform processes of: (I) on condition that a learning device has performed processes of inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for training to be used for performing a task of a task learning network and thus generate a data representation for training including the extracted features for training by performing a learning operation on the training data and (ii) transform the data representation for training and thus generate an anonymized data representation for training as an obfuscated data in which privacy-related information for training of the training data is protected and task utility is preserved; and inputting the anonymized data representation for training into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation for training and thus output a task result for training, (ii) generate a task loss by referring to the task result for training and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized, acquiring a test data; and (II) inputting the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for testing to be used for performing the task of the task learning network and thus generate a data representation for testing by performing the learning operation on the testing data and (ii) transform the data representation for testing and thus generate an anonymized data representation for testing as an obfuscated data in which privacy-related information for testing of the testing data is protected and task utility is preserved.

[0038] As one example, the processor further performs a process of: (III) transmitting the anonymized data representation for testing to a server in which the task learning network is installed, to thereby instruct the server to acquire a task result for testing, wherein the task learning network performs the task by using the anonymized data representation for testing and thus generates the task result for testing.

[0039] As one example, at the process of (II), the processor inputs the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation for testing by encoding the testing data through an encoding network, and (ii) generate the anonymized data representation for testing by reducing the features included in the data representation for testing through an information reduction module.

[0040] As one example, the processor instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation for testing, (ii) a noise addition process by adding noise to the data representation for testing, (iii) a random value replacement process by replacing parts of pixels of the data representation for testing with a random value, (iv) a random shuffling process by shuffling position information of the data representation for testing, and (v) a resizing process by resizing a cardinal number of pixels in the data representation for testing to be smaller than a cardinal number of pixels in the testing data, through the information reduction module.

[0041] As one example, the processor instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation for testing arbitrarily and thus generate an arbitrarily-resized data representation for testing and (ii) in response to detecting that a size of the arbitrarily-resized data representation for testing is bigger than a size of the testing data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation for testing, through the information reduction module.

[0042] As one example, the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the testing data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

[0043] As one example, at the process of (I), the learning device has inputted the anonymized data representation for training into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation for training and thus output an adversarial result for training in which the privacy-related information for training of a privacy-related region for training is estimated from the anonymized data representation for training, (ii) generate an adversarial loss by referring to the adversarial result for training and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0044] In addition, recordable media that are readable by a computer for storing a computer program to execute the method of the present disclosure is further provided.

[0045] The following drawings to be used for explaining example embodiments of the present disclosure are only part of example embodiments of the present disclosure and other drawings can be acquired based on the drawings by those skilled in the art of the present disclosure without inventive work.

Fig. 1 is a drawing schematically illustrating a learning device for training an obfuscation network capable of obfuscating, e.g., anonymizing or concealing, an original data in accordance with one example embodiment of the present disclosure.

Fig. 2 is a drawing schematically illustrating a learning method for training the obfuscation network capable of obfuscating the original data in accordance with one example embodiment of the present disclosure.

Fig. 3 is a drawing schematically illustrating the learning method for training the obfuscation network capable of obfuscating the original data in accordance with another example embodiment of the present disclosure.

Fig. 4 is a drawing schematically illustrating a testing device for testing the obfuscation network capable of obfuscating the original data in accordance with one example embodiment of the present disclosure.

Fig. 5 is a drawing schematically illustrating a testing method for testing the obfuscation network capable of obfuscating the original data in accordance with one example embodiment of the present disclosure.

[0046] In the following detailed description, reference is made to the accompanying drawings that show, by way of illustration, specific embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention. It is to be understood that the various embodiments of the present invention, although different, are not necessarily mutually exclusive. For example, a particular feature, structure, or characteristic described herein in connection with one embodiment may be implemented within other embodiments without departing from the spirit and scope of the present invention. In addition, it is to be understood that the position or arrangement of individual elements within each disclosed embodiment may be modified without departing from the spirit and scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims, appropriately interpreted, along with the full range of equivalents to which the claims are entitled. In the drawings, like numerals refer to the same or similar functionality throughout the several views.

[0047] To allow those skilled in the art to carry out the present disclosure easily, the example embodiments of the present disclosure will be explained in detail as shown below by referring to attached drawings.

[0048] Fig. 1 is a drawing schematically illustrating a learning device for training an obfuscation network capable of obfuscating, e.g., anonymizing or concealing, an original data in accordance with one example embodiment of the present disclosure.

[0049] Referring to Fig. 1, the learning device 1000 may include a memory 1001 for storing instructions to be used in training the obfuscation network capable of obfuscating the original data for privacy and a processor 1002 for performing processes in training the obfuscation network capable of obfuscating the original data for privacy according to the instructions stored in the memory 1001 in accordance with one example embodiment of the present disclosure.

[0050] Specifically, the learning device 1000 may typically achieve a desired system performance by using combinations of at least one computing device and at least one computer software, e.g., a computer processor, a memory, a storage, an input device, an output device, or any other conventional computing components, an electronic communication device such as a router or a switch, an electronic information storage system such as a network-attached storage (NAS) device and a storage area network (SAN) as the computing device and any instructions that allow the computing device to function in a specific manner as the computer software.

[0051] Also, the processors of such devices may include hardware configuration of MPU (Micro Processing Unit) or CPU (Central Processing Unit), cache memory, data bus, etc. Additionally, the computing device may further include operating system (OS) and software configuration of applications that achieve specific purposes.

[0052] Such description of the computing device does not exclude an integrated device including any combination of a processor, a memory, a medium, or any other computing components for implementing the present disclosure.

[0053] Meanwhile, on condition that at least one training data, e.g., at least one training image, is acquired, the processor 1002 of the learning device 1000 may (i) input the training data into the obfuscation network, to thereby instruct the obfuscation network to (i-1) extract features to be used for performing a task of a task learning network and thus generate a data representation including the extracted features by performing a learning operation on the training data and (i-2) transform the data representation and thus generate an anonymized data representation as an obfuscated data in which privacy-related information of the training data is protected and task utility is preserved and (ii) input the anonymized data representation into the task learning network, to thereby instruct the task learning network to (ii-1) perform the task by using the anonymized data representation and thus output a task result, (ii-2) generate a task loss by referring to the task result and its corresponding ground truth, (ii-3) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (ii-4) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

[0054] Herein, the task utility may be any measurements that can represent how effective a given task is performed. For reference, the meaning of the task utility is well-known to those skilled in the art and thus the detailed explanation thereon is omitted.

[0055] Further, the learning device 1000 may input the anonymized data representation into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform an adversarial task by using the anonymized data representation and thus output an adversarial result in which the privacy-related information of a privacy-related region is estimated from the anonymized data representation, (ii) generate an adversarial loss by referring to the adversarial result and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0056] A method for training the obfuscation network capable of obfuscating, e.g., anonymizing or concealing, the original data for privacy by using the learning device 1000 in accordance with one example embodiment of the present disclosure is described by referring to Figs. 2 as follows.

[0057] First, if the training data 10 is acquired, the learning device 1000 may input the training data 10 into the obfuscation network 1100, to thereby instruct the obfuscation network 1100 to (i) extract the features to be used for performing the task of the task learning network and thus generate the data representation 20 including the extracted features by performing the learning operation on the training data 10 and (ii) transform the data representation 20 and thus generate an anonymized data representation 30 as the obfuscated data.

[0058] Herein, the learning device 1000 may input the training data 10 into the obfuscation network 1100, to thereby instruct the obfuscation network 1100 to (i) generate the data representation 20 by encoding the training data 10 through an encoding network 1110, and (ii) generate the anonymized data representation 30 by reducing the features included in the data representation 20 through an information reduction module.

[0059] For example, the learning device 1000 may input a training image as the training data 10 into the obfuscation network 1100, to thereby instruct the obfuscation network 1100 to (i) generate a feature map as the data representation 20 by performing the learning operation on the training image and (ii) transform the feature map and thus generate an anonymized feature map as the anonymized data representation 30. That is, the learning device 1000 may instruct the encoding network 1110 of the obfuscation network 1100 to encode the training image and thus generate the feature map in which feature values as measurements for recognizing the training image are included and may input the feature map into the information reduction module 1120 of the obfuscation network 1100, to thereby instruct the information reduction module 1120 to transform the feature map and thus generate the anonymized feature map.

[0060] Meanwhile, the encoding network 1110 may be a deep neural network capable of performing the task by performing the learning operation on the training data 10 or may be parts of front-end layers of the deep neural network. Further, the task learning network 1200 may be the deep neural network capable of performing the task by performing learning operation on the training data 10 or may be parts of rear-end layers of the deep neural network. Specifically, the encoding network 1110 and the task learning network 1200 may be sub-networks included in the deep neural network capable of performing the task by performing the learning operation on the training data 10. Herein, the encoding network includes earlier layers of the deep neural network, and the task learning network includes remaining layers of the deep neural network.

[0061] For example, the encoding network 1110 may be configured as a convolutional neural network like U-Net or earlier layers of U-Net developed for biomedical image segmentation by the Department of Computer Science of the University of Freiburg, ResNet or earlier layers of ResNet which is Deep Residual Learning for Image Recognition developed by Microsoft, AlexNet or earlier layers of AlexNet which won the 2012 ILSVRC (ImageNet Large Scale Visual Recognition Challenge) competition, MobileNetV2 or earlier layers of MobileNetV2 developed by Google, etc. Further, the task learning network 1200 may be configured as ResNet or latter layers of ResNet, AlexNet or latter layers of AlexNet, or MobileNetV2 or latter layers of MobileNetV2, etc.

[0062] Furthermore, the task performed by the task learning network 1200 may include various tasks that may be performed by deep learning methods, such as a classification task, a regression task, and a segmentation task, etc.

[0063] Furthermore, the information reduction module 1120 may generate the anonymized data representation 30 by reducing the features included in the data representation 20 through the information reduction module 1120.

[0064] Herein, the information reduction module 1120 may perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation 20, (ii) a noise addition process by adding noise to the data representation 20, (iii) a random value replacement process by replacing parts of pixels of the data representation 20 with a random value, (iv) a random shuffling process by shuffling position information of the data representation 20, and (v) a resizing process by resizing a cardinal number of pixels in the data representation 20 to be smaller than a cardinal number of pixels in the training data 10. As a result, the information reduction module 1120 may generate the anonymized data representation.

[0065] As an example, for the training image with size 178x178x3 as its height, width, and channel, the encoding network 1110 configured as earlier five layers of ResNet18 may perform convolutional operations on the training image to generate the feature map with size 45x45x64 and then transform the feature map into the anonymized feature map with size 45x45x60 through the information reduction module 1120. Herein, although total pixels 121,500 of the anonymized feature map with the size 45x45x60 is more than the total pixels 95,052 of the training image with size 178x178x3, a lot of information may have been reduced through the information reduction module 1120.

[0066] As another example, for the training image with size 178x178x3 as its height, width, and channel, the encoding network 1110 configured as U-Net may perform the convolutional operations and deconvolutional operations on the training image to generate the feature map with size 178x178x3 and then transform the feature map into the anonymized feature map with size 178x178x3 through the information reduction module 1120. Herein, although the total pixels of the anonymized feature map and the total pixels of the training image are the same, a lot of information may have been reduced through the information reduction module 1120.

[0067] Meanwhile, the learning device 1000 may instruct the obfuscation network 1100 to (i) perform the resizing process to thereby change a size of the data representation 20 arbitrarily and thus generate an arbitrarily-resized data representation and (ii) in response to detecting that a size of the arbitrarily-resized data representation is greater than a size of the training data 10, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation, through the information reduction module 1120. Herein, the noise addition process through the information reduction module 1120 may or may not be performed if the size of the arbitrarily-resized data representation is less than the size of the training image. That is, the obfuscation network 1100 may perform the learning operation on the training image to generate the arbitrarily-resized data representation and transform the arbitrarily-resized data into the anonymized data representation such that the number of total pixels of the anonymized data representation is less than those of the training image or such that even if the resultant number of the total pixels of the anonymized data representation is greater than those of the training image, information included in the anonymized data representation is reduced.

[0068] Next, the learning device 1000 may input the anonymized data representation 30 into the task learning network 1200, to thereby instruct the task learning network 1200 to perform the task by using the anonymized data representation 30 and thus output the task result.

[0069] Herein, the task result may be characteristic information generated by performing the learning operation on the anonymized data representation 30, or a task-specific output generated by using the characteristic information.

[0070] And, the characteristic information may be features or logits corresponding to the anonymized data representation 30. Also, the characteristic information may be feature values related to certain features in the anonymized data representation 30, or the logits including values of at least one of vectors, matrices, and coordinates related to the certain features. For example, if the training data 10 is facial image data, the result above may be classes for face recognition, facial features, e.g., laughing expressions, coordinates of facial landmark points, e.g., both end points on far sides of an eye.

[0071] Meanwhile, the task-specific output may have various results according to the task of the task learning network 1200, such as a probability of a class for classification, coordinates resulting from regression for location detection, etc. Further, an activation function of an activation unit may be applied to the characteristic information, to thereby generate the task-specific output. Herein, the activation function may include a sigmoid function, a linear function, a softmax function, an rlinear function, a square function, a sqrt function, an srlinear function, an abs function, a tanh function, a brlinear function, etc. but the scope of the present disclosure is not limited thereto.

[0072] As an example, when the task learning network 1200 performs the task for the classification, it may map the characteristic information onto each of classes, to thereby generate one or more probabilities of the anonymized data representation 30 for each of the classes. Herein, each of the probabilities for each of the classes may represent each of probabilities of the characteristic information for each of the classes being true. For example, if the training data is the facial image data, a probability of the face having a laughing expression may be outputted as 0.75, and a probability of the face not having the laughing expression may be outputted as 0.25, and the like. Herein, a softmax algorithm may be used for mapping the characteristic information onto each of the classes, but the scope of the present disclosure is not limited thereto, and various algorithms may be used for mapping the characteristic information onto each of the classes.

[0073] Meanwhile, in case the feature map is generated by performing the convolutional operations on the training image with the size 178x178x3 through the encoding network 1110 configured as the earlier five layers of ResNet18 in the obfuscation network 1100 and then the anonymized feature map with the size 45x45x60 is generated by transforming the feature map, the task learning network 1200 may be configured as the remaining layers of ResNet18 after the earlier five layers and may generate the task result by performing the learning operation on the anonymized feature map with the size 45x45x60.

[0074] Additionally, in case the feature map is generated by performing the convolutional operations and the deconvolutional operations on the training image with the size 178x178x3 through the encoding network 1110 configured as U-Net of the obfuscation network 1100 and then the anonymized feature map with size 178x178x3 is generated by transforming the feature map, the task learning network 1200 may be configured as ResNet18 and may generate the task result by performing the learning operation on the anonymized feature map with the size 178x178x3.

[0075] Next, the learning device 1000 may (i) generate a task loss by referring to the task result and its corresponding ground truth, (ii) train the task learning network 1200 through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network 1100 through a second backpropagation of the task loss such that the task loss is minimized.

[0076] The method of training the obfuscation network 1100 using one task learning network 1200 designed to perform a specific task is described above, however, contrary to above, it is also possible to train the obfuscation network 1100 using multiple task learning networks 1200 that are designed to perform different tasks.

[0077] Next, by referring to Fig. 3, another method for training the obfuscation network 1100 in accordance with another embodiment of the present disclosure will be explained as below.

[0078] Fig. 3 schematically illustrates a method for training the obfuscation network by adding an additional proxy adversarial network to the configuration of Fig. 2. Herein the proxy adversarial network may be able to (i) breach the privacy-related information by acquiring sensitive attributes related to the privacy-related information from the anonymized data representation or (ii) reconstruct the training image by using the anonymized data representation. The detailed explanations that can be easily inferred from the description with reference to Fig. 2 will be omitted.

[0079] First, if the training data 10 is acquired, the learning device 1000 may input the training data 10 into the obfuscation network 1100, to thereby instruct the obfuscation network 1100 to (i) extract the features to be used for performing the task of the task learning network 1200 and thus generate the data representation 20 including the extracted features by performing the learning operation on the training data 10 and (ii) transform the data representation 20 and thus generate the anonymized data representation 30 as the obfuscated data.

[0080] Next, the learning device 1000 may input the anonymized data representation 30 into the task learning network 1200, to thereby instruct the task learning network 1200 to perform the task by using the anonymized data representation 30 and thus generate the task result.

[0081] Additionally, the learning device 1000 may input the anonymized data representation 30 into the proxy adversarial network 1300, to thereby instruct the proxy adversarial network 1300 to perform the adversarial task by using the anonymized data representation 30 and thus output the adversarial result in which the privacy-related information of the privacy-related region is estimated from the anonymized data representation 30.

[0082] Herein, the adversarial task may be a task for detecting the privacy-related information by extracting private attributes from the anonymized data representation 30 or a task for reconstructing the training image by using the anonymized data representation. For example, the adversarial task may be an adversarial attack aiming to detect a gender, an emotional state, a race, etc., of an image or to reconstruct the privacy-related information within the image.

[0083] Meanwhile, the proxy adversarial network 1300 may be the deep neural network capable of performing the adversarial task by performing learning operation on the training data 10 or may be parts of rear-end layers of the deep neural network.

[0084] For example, the proxy adversarial network 1300 may be configured as U-Net or latter layers of U-Net, ResNet or latter layers of ResNet, AlexNet or latter layers of AlexNet, or MobileNetV2 or latter layers of MobileNetV2, etc.

[0085] As an example, in case the feature map is generated by performing the convolutional operations on the training image with the size 178x178x3 through the encoding network 1110 configured as the earlier five layers ResNet18 in obfuscation network 1100 and then the anonymized feature map with the size 45x45x60 is generated by transforming the feature map, the proxy adversarial network 1300 may be configured as the remaining layers of ResNet18 after the earlier five layers and may generate the task result by performing the adversarial task on the anonymized feature map with the size 45x45x60.

[0086] Additionally, in case the feature map is generated by performing the convolutional operations and the deconvolutional operations on the training image with the size 178x178x3 through the encoding network 1110 configured as U-Net in the obfuscation network 1100 and then the anonymized feature map with the size 178x178x3 is generated by transforming the feature map, the proxy adversarial network 1300 may be configured as U-Net and may generate the adversarial result by performing an adversarial operation on the anonymized feature map with the size 178x178x3.

[0087] Next, the learning device 1000 may (i) generate an adversarial loss by referring to the adversarial result and its corresponding ground truth, (ii) train the proxy adversarial network 1300 through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iii) train the obfuscation network 1100 through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0088] That is, the learning device 1000 may generate the task loss by referring to the task result of the task learning network 1200 and its corresponding task ground truth, and may generate the adversarial loss by referring to the adversarial result of the proxy adversarial network 1300 and its corresponding adversarial ground truth. Next, the learning device 1000 may (i) train the task learning network 1200 through the first backpropagation of the task loss such that the task loss is minimized, (ii) train the proxy adversarial network 1300 through the third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iii) train the obfuscation network 1100 through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0089] The method of training the obfuscation network 1100 additionally using the proxy adversarial network 1300 designed to perform the adversarial task is described above, however, contrary to above, it is also possible to train the obfuscation network 1100 using multiple additional proxy adversarial networks 1300 that are designed to perform different adversarial tasks.

[0090] Next, Fig. 4 is a drawing schematically illustrating a testing device for testing the trained obfuscation network in accordance with one example embodiment of the present disclosure.

[0091] For reference, in the description below, the phrase "for training" is added for terms related to the learning processes, and the phrase "for testing" is added for terms related to testing processes, to avoid possible confusion.

[0092] Referring to Fig. 4, the testing device 2000 may include a memory 2001 for storing instructions to be used in testing the trained obfuscation network capable of obfuscating the original data for privacy and a processor 2002 for performing processes in testing the trained obfuscation network capable of obfuscating the original data for privacy according to the instructions stored in the memory 2001 in accordance with one example embodiment of the present disclosure.

[0093] Specifically, the testing device 2000 may typically achieve a desired system performance by using combinations of at least one computing device and at least one computer software, e.g., a computer processor, a memory, a storage, an input device, an output device, or any other conventional computing components, an electronic communication device such as a router or a switch, an electronic information storage system such as a network-attached storage (NAS) device and a storage area network (SAN) as the computing device and any instructions that allow the computing device to function in a specific way as the computer software.

[0094] Also, the processors of such devices may include hardware configuration of MPU (Micro Processing Unit) or CPU (Central Processing Unit), cache memory, data bus, etc. Additionally, the computing device may further include operating system (OS) and software configuration of applications that achieve specific purposes.

[0095] Such description of the computing device does not exclude an integrated device including any combination of a processor, a memory, a medium, or any other computing components for implementing the present disclosure.

[0096] Meanwhile, the processor 2002 of the testing device 2000, (1) on condition that the learning device has performed processes of (I) inputting the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for training to be used for performing the task of the task learning network and thus generate a data representation for training including the extracted features for training by performing the learning operation on the training data and (ii) transform the data representation for training and thus generate an anonymized data representation for training as the obfuscated data in which privacy-related information for training of the training data is protected and the task utility is preserved and (II) inputting the anonymized data representation for training into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation for training and thus output a task result for training, (ii) generate the task loss by referring to the task result for training and its corresponding ground truth, (iii) train the task learning network through the first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss such that the task loss is minimized, may acquire at least one test data, and (2) according to the instructions stored in the memory 2001, may input the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for testing to be used for performing the task of the task learning network and thus generate a data representation for testing by performing the learning operation on the testing data and (ii) transform the data representation for testing and thus generate an anonymized data representation for testing as an obfuscated data in which privacy-related information for testing of the testing data is protected and the task utility is preserved.

[0097] Herein, the learning device may have inputted the anonymized data representation for training into the proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation for training and thus output an adversarial result for training in which the privacy-related information for training of a privacy-related region for training is estimated from the anonymized data representation for training, (ii) generate an adversarial loss by referring to the adversarial result for training and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0098] Additionally, the testing device 2000 may transmit the anonymized data representation for testing to a server in which the task learning network is installed, to thereby instruct the server to acquire a task result for testing. Herein the task learning network may perform the task by using the anonymized data representation for testing and thus generates the task result for testing.

[0099] The testing device 2000 for testing the obfuscation network capable of obfuscating, e.g., anonymizing or concealing, the original data for privacy in accordance with one example embodiment of the present disclosure is described by referring to Fig. 5 as follows. In the description below, parts easily deducible regarding the learning method from Figs. 2 and 3 will be omitted.

[0100] Referring to Fig. 5, the testing device 2000 may acquire the test data 11 on condition that the obfuscation network capable of obfuscating the original data for privacy has been trained.

[0101] Herein, the obfuscation network may have been trained by using the learning method illustrated in Figs. 2 and 3.

[0102] That is, as illustrated in Fig. 2, the learning device has performed processes of (I) inputting the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract the features for training to be used for performing the task of the task learning network and thus generate the data representation for training including the extracted features for training by performing the learning operation on the training data and (ii) transform the data representation for training and thus generate the anonymized data representation for training as the obfuscated data in which the privacy-related information for training of the training data is protected and the task utility is preserved and (II) inputting the anonymized data representation for training into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation for training and thus output the task result for training, (ii) generate the task loss by referring to the task result for training and its corresponding ground truth, (iii) train the task learning network through the first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss such that the task loss is minimized.

[0103] Further, as illustrated in Fig. 3, the learning device may have inputted the anonymized data representation for training into the proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation for training and thus output the adversarial result for training in which the privacy-related information for training of the privacy-related region for training is estimated from the anonymized data representation for training, (ii) generate the adversarial loss by referring to the adversarial result for training and its corresponding ground truth, (iii) train the proxy adversarial network through the third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

[0104] Next, the testing device 2000 may input the test data 11 into the obfuscation network 1100, to thereby instruct the obfuscation network 1100 to (i) extract the features for testing to be used for performing the task of the task learning network 1200 and thus generate the data representation for testing 21 by performing the learning operation on the testing data 11 and (ii) transform the data representation for testing 21 and thus generate the anonymized data representation for testing 31 as the obfuscated data in which the privacy-related information for testing of the testing data 21 is protected and the task utility is preserved.

[0105] Further, the testing device 2000 may transmit the anonymized data representation for testing 31 to the server 3000 in which the task learning network 1200 is installed, to thereby instruct the server 3000 to acquire the task result for testing, wherein the task learning network 1200 performs the task by using the anonymized data representation for testing and thus generates the task result for testing.

[0106] Meanwhile, the testing device 2000 may be an edge device such as a mobile device with limited computing resources. In this case, the trained obfuscation network 1100 may be installed on the mobile device and the trained task learning network 1200 may be installed on the server 3000.

[0107] For example, if the testing device 2000 is the mobile device, a user may use the mobile device to take photos or record videos. The frames of the photos and videos may be obfuscated through the obfuscation network 1100, to thereby generate the anonymized data representation 31 in which privacy-related information of the training data is protected and the task utility is preserved. Then, the anonymized data representation 31 may be transmitted to the server 3000, to thereby allow the server to perform the task through the task learning network 1200.

[0108] Herein, even if the anonymized data representation 31 transmitted from the mobile device is exposed to the adversarial network 4000, the adversarial network 4000 may not be able to extract the private attributes therefrom, thereby preventing the adversarial attack by the adversarial network 4000.

[0109] Meanwhile, comparisons between a performance of the obfuscation model according to one embodiment of the present disclosure and a performance of the conventional obfuscation models are as follows.

[0110] First, the applicant set up an experiment as follows.

Datasets

[0111] The applicant conducted experiments on CelebA, FairFace, and CIFAR10. Following the utility and privacy task setting from DISCO, the applicant set "smiling" as the utility attribute and "male" as the privacy attribute for CelebA, "gender" as the utility attribute and "race" as the private attribute for FairFace. For CIFAR10, following the setting from MaxEnt, the utility task is defined as classifying living objects (e.g., "bird", "cat", etc.) or non-living objects (e.g., "airplane", "automobile", etc.) and privacy task as classifying separate 10 classes.

Models

[0112] For the full task learning model F, the applicant used ResNet18. ResNet18 consists of one convolution layer, and 4 residual blocks each of which consists of four convolution layers and residual connection, and finally a fully connected layer. The applicant chose the splitting point to be right after each of the 4 residual blocks. The applicant indicated the different configurations as RN18{1, 2, 3, 4} respectively, where the subscript number indicates the block after which the network was split. For the task learning model and the proxy adversarial model the applicant used the remaining part of the split architecture, e.g., for RN184 the remaining part would consist of the fully connected layer. For the noise parameter, σ= 1920 was used for FairFace, and σ= 3840 was used for CelebA and CIFAR10. The parameter was chosen based on privacy-utility trade-off on given dataset and model. A separate Adam optimizer was used for all 3 models with learning rate 10-3, and λ= 10-2 was used for balancing the losses. The commonly used cross-entropy loss was used as the utility and information leakage attack loss. The applicant reported top-1 accuracy for the utility and privacy task, respectively.

Attacks

[0113] For the information leakage attack, the applicant used a latter part of the split architecture. After the training of the obfuscator network was done, the adversarial network was trained also with Adam optimizer and the highest privacy accuracy was reported. The reconstruction attack was performed on CelebA dataset with a decoder from DeepObfs, which was also trained with Adam optimizer with learning rate 10-3. The reconstruction loss was computed between original image and reconstructed image with the MSE loss. The applicant depicted the qualitative results, but additionally provided quantitative comparison by reporting MSE, L2, SSIM, and PSNR.

[0114] The applicant compared the present invention with various baselines for the privacy-utility trade-off.

ResNet18

[0115] To indicate the practical performance bounds, the applicant reported the utility and privacy performance for a ResNet18 model trained on the respective task.

Image Noise

[0116] As a simple baseline image privacy remover, the applicant added Gaussian noise sampled from N(o, σ2) to the input image directly, while obeying the image range of pixels in the range (0, 1). That is to say, there are no model parameters to be trained for this model. For CelebA and FairFace σ= 2 was used and σ= 0.8 for CIFAR10. The σ was chosen based on the noise that fully obfuscates the image for the human eye. The applicant used the entire ResNet18 for both utility and privacy model.

No Noise

[0117] To indicate the effectiveness of noise addition module (i.e., the noise adding process of the information restriction module) the applicant trained the split ResNet model without adding noise η to the intermediate representation. The applicant reported the performance for RN3 and RN4.

MaxEnt

[0118] The applicant compare to MaxEnt ARL method which uses full ResNet18 as a client-side obfuscator, the last output of the obfuscator is a vector which has length d. d= 128 was used for CIFAR10, which is the setting used from the paper, and d=256 was used for FairFace and CelebA. The loss consists of task loss and adversarial loss, while the adversarial loss is based on the entropy of the prediction of the adversarial model.

DISCO

[0119] The applicant reported the privacy-utility trade-off numbers as in the original work. The applicant reconfirmed the reconstruction vulnerability of DISCO as reported in their work with their parameters.

DeepObfuscator

[0120] Since the authors of DeepObfuscator did not open-source their code, the applicant re-implemented DeepObfuscator to the best of the applicant's knowledge using the provided information in their paper.

[0121] The test results according to the method of the present invention and the conventional method according to these tests are as follows.

[0122] Table 1. Efficiency of each client model

<Table 1>

Benchmark	DeepObfs.	DISCO	MaxEnt	RN18	RN18$_1$	RN18$_2$	RN18$_3$	RN18$_4$
Computational Cost (GFLOPs)	6.00	2.52	2.52	2.52	0.75	1.31	1.92	2.52
Memory (MB)	1.00	42.80	43.17	42.69	0.60	2.61	10.63	42.67

[0123] Table 1 shows a computational cost and a memory usage for each of the models, and images with the size 178x178x3 were used to measure the performance.

[0124] As shown in Table 1, all of present invention's variants have equal or lower computational costs than the others. It is also appreciated that the present invention requires smaller memory usage compared to the other modules, except for DeepObfs. However, it is noted that DeepObfs shows the largest computational cost.

[0125] Table 2. Comparison for the privacy-utility trade-off

<Table 2>

	Fairface			CelebA			CIFAR10		
Method	Privacy ↓	Utility ↑	*Δ* ↑	Privacy ↓	Utility ↑	*Δ* ↑	Privacy ↓	Utility ↑	*Δ* ↑
RN18	63.57	92.11	-	98.14	93.48	-	94.51	98.79	-
MaxEnt [39]	24.56	90.52	65.96	59.28	93.43	34.15	24.61	97.74	73.13
DISCO [45]	19.00	81.50	62.50	61.20	91.00	29.80	22.30	91.98	69.68
DeepObfs.	50.83	89.64	38.81	97.63	91.92	-5.71	73.79	92.86	19.07
Image Noise	42.61	74.33	31.72	91.71	85.38	-6.33	54.37	87.77	33.40
No Noise (RN18$_3$)	45.22	89.55	44.33	94.54	93.38	-1.16	69.34	97.64	28.30
No Noise (RN18$_4$)	31.56	89.87	58.31	93.19	93.43	0.24	56.02	97.97	41.95
Ours (RN18$_3$)	19.47	89.08	**69.61**	57.77	93.07	**35.30**	21.71	96.92	**75.21**
Ours (RN18$_4$)	15.60	88.34	**72.74**	53.77	90.86	**37.09**	19.81	94.25	**74.44**

[0126] Table 2 shows a comparison between the present invention and existing ARL approaches focusing on the privacy-utility trade-off.

[0127] In terms of privacy-utility trade-off (Δ), the present invention outperforms all other methods while showing comparable utility accuracy with the performance bound. Comparison with 'No Noise' method shows the effectiveness of our noise addition module.

[0128] Table 3. Quantitative result of reconstruction attack on CelebA. Similarity scores between original images and the reconstruction images.

<Table 3>

Method	MSE ↑	L1 ↑	SSIM ↓	MS-SSIM ↓	PSNR ↓	LPIPS ↑
MaxEnt	4955.44	58.83	0.3893	0.4057	11.19	0.6619
DeepObfs.	182.63	9.47	0.7834	0.9298	25.52	0.1864
DISCO	567.17	15.94	0.5765	0.7611	20.60	0.4351
Image Noise	584.88	16.97	0.6017	0.7776	20.46	0.3710
No Noise (RN18$_3$)	1391.39	26.89	0.4666	0.6155	16.70	0.4882
No Noise (RN18$_4$)	1841.70	31.70	0.4558	0.5829	15.48	0.4857
Ours (RN18$_3$)	5437.02	63.22	0.3086	0.1682	10.78	0.8045
Ours (RN18$_4$)	**5454.12**	**63.48**	**0.3301**	**0.1571**	**10.77**	**0.8197**

[0129] According to Table 3, the present invention showed the best robustness to the reconstruction attack in terms of both qualitative and quantitative result.

[0130] The present disclosure has an effect of providing the obfuscation network that is robust to attacks by adversarial actors while the task utility is preserved.

[0131] The present disclosure has another effect of providing the obfuscation network that minimizes the use of the computing resources in the edge device.

[0132] The present disclosure has still another effect of providing the obfuscation network that can easily be applied to an off-the-shelf deep neural network (DNN).

[0133] The present disclosure has still yet another effect of providing the obfuscation network that can easily be applied to a commonly used DNN.

[0134] The embodiments of the present disclosure as explained above can be implemented in a form of executable program command through a variety of computer means recordable in computer readable media. The computer readable media may include solely or in combination, program commands, data files, and data structures. The program commands recorded to the media may be components specially designed for the present disclosure or may be usable to a skilled human in a field of computer software. Computer readable media include magnetic media such as hard disk, floppy disk, and magnetic tape, optical media such as CD-ROM and DVD, magneto-optical media such as floptical disk and hardware devices such as ROM, RAM, and flash memory specially designed to store and carry out program commands. Program commands may include not only a machine language code made by a compiler but also a high-level code that can be used by an interpreter etc., which is executed by a computer. The aforementioned hardware device can work as more than a software module to perform the action of the present disclosure and they can do the same in the opposite case.

[0135] As seen above, the present disclosure has been explained by specific matters such as detailed components, limited embodiments, and drawings. They have been provided only to help more general understanding of the present disclosure. It, however, will be understood by those skilled in the art that various changes and modification may be made from the description without departing from the spirit and scope of the disclosure as defined in the following claims.

[0136] Accordingly, the thought of the present disclosure must not be confined to the explained embodiments, and the following patent claims as well as everything including variations equal or equivalent to the patent claims pertain to the category of the thought of the present disclosure.

Claims

1. A learning method for training an obfuscation network capable of obfuscating an original data for privacy, comprising steps of:

(a) a learning device inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features to be used for performing a task of a task learning network and thus generate a data representation including the extracted features by performing a learning operation on the training data and (ii) transform the data representation and thus generate an anonymized data representation as an obfuscated data in which privacy-related information of the training data is protected and task utility is preserved; and
(b) the learning device inputting the anonymized data representation into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation and thus output a task result, (ii) generate a task loss by referring to the task result and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

2. The learning method of Claim 1, wherein, at the step of (a), the learning device inputs the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation by encoding the training data through an encoding network, and (ii) generate the anonymized data representation by reducing the features included in the data representation through an information reduction module.

3. The learning method of Claim 2, wherein the learning device instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation, (ii) a noise addition process by adding noise to the data representation, (iii) a random value replacement process by replacing parts of pixels of the data representation with a random value, (iv) a random shuffling process by shuffling position information of the data representation, and (v) a resizing process by resizing a cardinal number of pixels in the data representation to be smaller than a cardinal number of pixels in the training data, through the information reduction module.

4. The learning method of Claim 3, wherein the learning device instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation arbitrarily and thus generate an arbitrarily-resized data representation and (ii) in response to detecting that a size of the arbitrarily-resized data representation is bigger than a size of the training data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation, through the information reduction module.

5. The learning method of Claim 2, wherein the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the training data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

6. The learning method of Claim 1, wherein, at the step of (b), the learning device inputs the anonymized data representation into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation and thus output an adversarial result in which the privacy-related information of a privacy-related region is estimated from the anonymized data representation, (ii) generate an adversarial loss by referring to the adversarial result and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

7. A testing method for testing an obfuscation network capable of obfuscating an original data for privacy, comprising steps of:

(a) on condition that a learning device has performed processes of (I) inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for training to be used for performing a task of a task learning network and thus generate a data representation for training including the extracted features for training by performing a learning operation on the training data and (ii) transform the data representation for training and thus generate an anonymized data representation for training as an obfuscated data in which privacy-related information for training of the training data is protected and task utility is preserved; and (II) inputting the anonymized data representation for training into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation for training and thus output a task result for training, (ii) generate a task loss by referring to the task result for training and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized, a testing device acquiring a test data; and
(b) the testing device inputting the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features for testing to be used for performing the task of the task learning network and thus generate a data representation for testing by performing the learning operation on the testing data and (ii) transform the data representation for testing and thus generate an anonymized data representation for testing as an obfuscated data in which privacy-related information for testing of the testing data is protected and task utility is preserved.

8. The testing method of Claim 7, further comprising a step of:
(c) the testing device transmitting the anonymized data representation for testing to a server in which the task learning network is installed, to thereby instruct the server to acquire a task result for testing, wherein the task learning network performs the task by using the anonymized data representation for testing and thus generates the task result for testing.

9. The testing method of Claim 7, wherein, at the step of (b), the testing device inputs the test data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation for testing by encoding the testing data through an encoding network, and (ii) generate the anonymized data representation for testing by reducing the features included in the data representation for testing through an information reduction module.

10. The testing method of Claim 9, wherein the testing device instructs the obfuscation network to perform at least one of (i) a frequency filtering process by passing at least one preset frequency band or rejecting the preset frequency band of the data representation for testing, (ii) a noise addition process by adding noise to the data representation for testing, (iii) a random value replacement process by replacing parts of pixels of the data representation for testing with a random value, (iv) a random shuffling process by shuffling position information of the data representation for testing, and (v) a resizing process by resizing a cardinal number of pixels in the data representation for testing to be smaller than a cardinal number of pixels in the testing data, through the information reduction module.

11. The testing method of Claim 10, wherein the testing device instructs the obfuscation network to (i) perform the resizing process to thereby change a size of the data representation for testing arbitrarily and thus generate an arbitrarily-resized data representation for testing and (ii) in response to detecting that a size of the arbitrarily-resized data representation for testing is bigger than a size of the testing data, perform the noise addition process to thereby reduce information included in the arbitrarily-resized data representation for testing, through the information reduction module.

12. The testing method of Claim 9, wherein the encoding network and the task learning network are sub-networks included in a deep neural network capable of performing the task by performing the learning operation on the testing data, wherein the encoding network includes earlier layers of the deep neural network, and wherein the task learning network includes remaining layers of the deep neural network.

13. The testing method of Claim 7, wherein, at the step of (a), the learning device has inputted the anonymized data representation for training into a proxy adversarial network, to thereby instruct the proxy adversarial network to (i) perform the adversarial task by using the anonymized data representation for training and thus output an adversarial result for training in which the privacy-related information for training of a privacy-related region for training is estimated from the anonymized data representation for training, (ii) generate an adversarial loss by referring to the adversarial result for training and its corresponding ground truth, (iii) train the proxy adversarial network through a third backpropagation of the adversarial loss such that the adversarial loss is minimized, and (iv) train the obfuscation network through the second backpropagation of the task loss and the adversarial loss such that the task loss is minimized and such that the adversarial loss is maximized.

14. A learning device for training an obfuscation network capable of obfuscating an original data for privacy, comprising:

at least one memory that stores instructions; and
at least one processor configured to execute the instructions to perform processes of (I) inputting a training data into the obfuscation network, to thereby instruct the obfuscation network to (i) extract features to be used for performing a task of a task learning network and thus generate a data representation including the extracted features by performing a learning operation on the training data and (ii) transform the data representation and thus generate an anonymized data representation as an obfuscated data in which privacy-related information of the training data is protected and task utility is preserved; and (II) inputting the anonymized data representation into the task learning network, to thereby instruct the task learning network to (i) perform the task by using the anonymized data representation and thus output a task result, (ii) generate a task loss by referring to the task result and its corresponding ground truth, (iii) train the task learning network through a first backpropagation of the task loss such that the task loss is minimized, and (iv) train the obfuscation network through a second backpropagation of the task loss such that the task loss is minimized.

15. The learning device of Claim 14, wherein, at the process of (I), the processor inputs the training data into the obfuscation network, to thereby instruct the obfuscation network to (i) generate the data representation by encoding the training data through an encoding network, and (ii) generate the anonymized data representation by reducing the features included in the data representation through an information reduction module.

FIG. 1

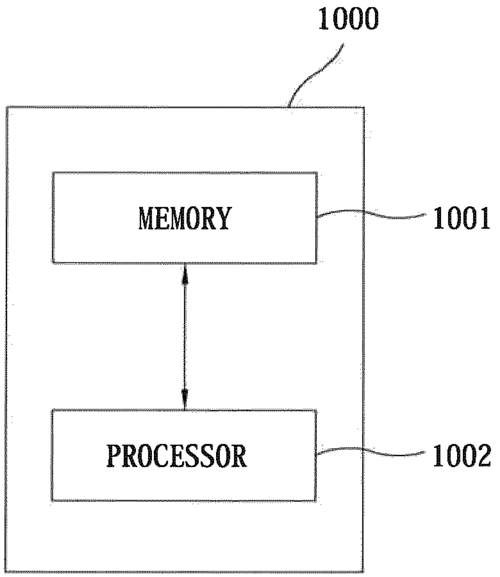


FIG. 2

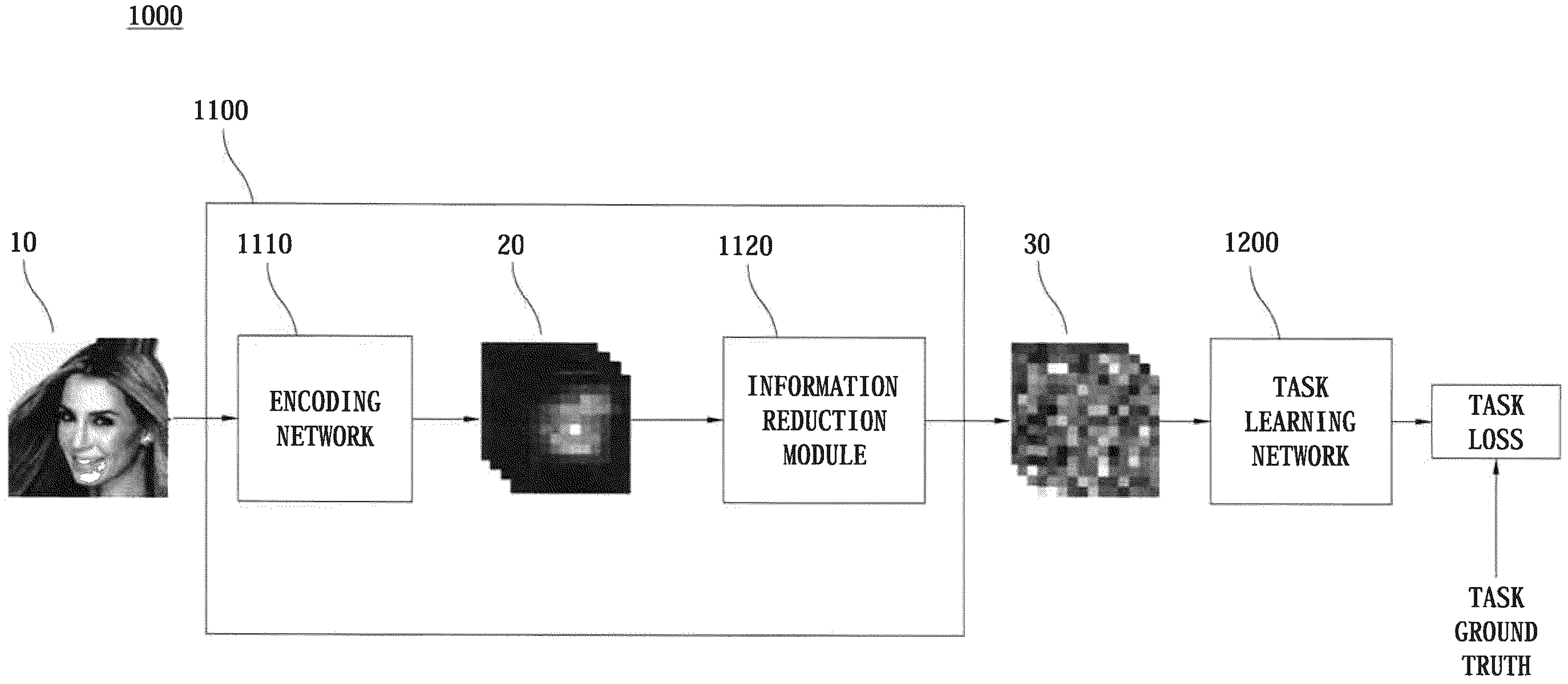

1000
1100
10
1110
ENCODING NETWORK
20
1120
INFORMATION REDUCTION MODULE
30
1200
TASK LEARNING NETWORK
TASK LOSS
TASK GROUND TRUTH

FIG. 3

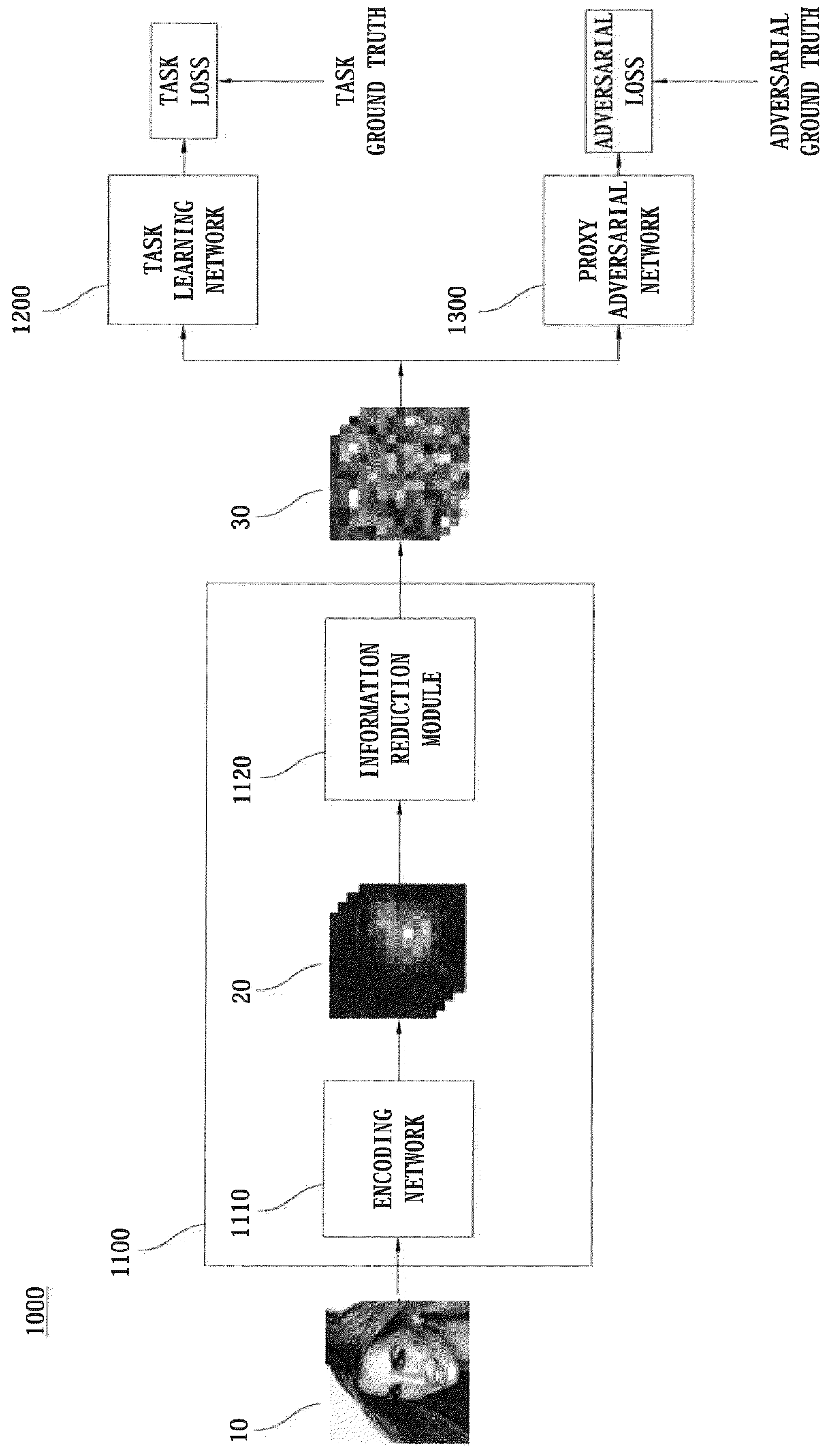

1000
10
1100
1110
ENCODING NETWORK
20
1120
INFORMATION REDUCTION MODULE
30
1200
TASK LEARNING NETWORK
TASK LOSS
TASK GROUND TRUTH
1300
PROXY ADVERSARIAL NETWORK
ADVERSARIAL LOSS
ADVERSARIAL GROUND TRUTH

FIG. 4

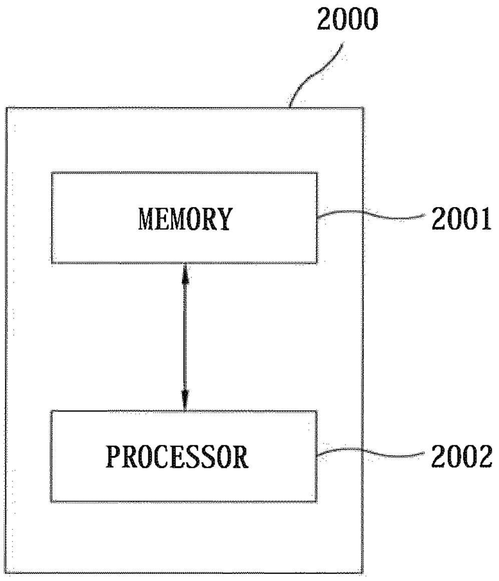
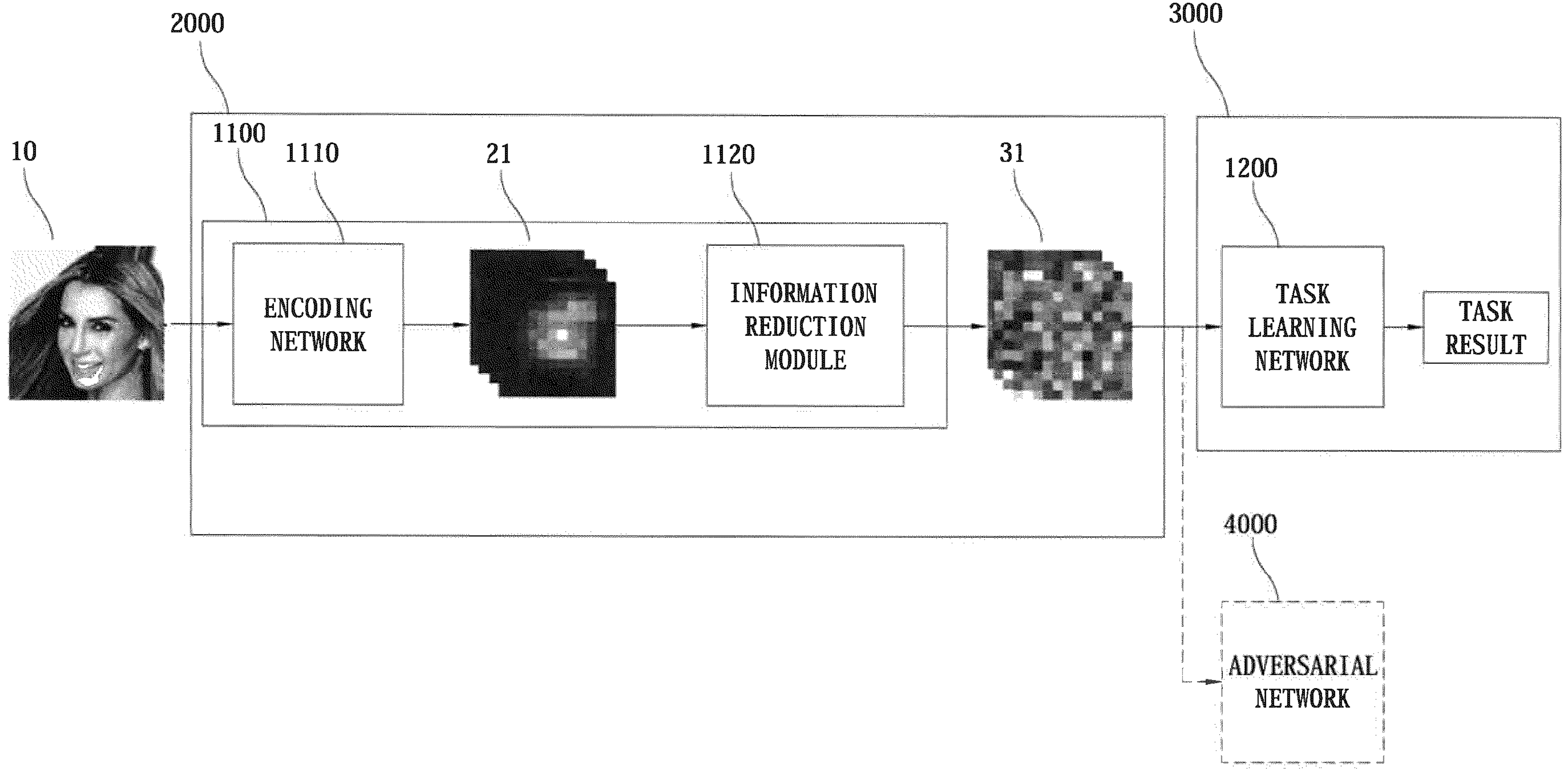


FIG. 5


2000
3000
10
1100
1110
21
1120
31
1200
ENCODING NETWORK
INFORMATION REDUCTION MODULE
TASK LEARNING NETWORK
TASK RESULT
4000
ADVERSARIAL NETWORK

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 22 16 9013

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 2021/141926 A1 (FERRER CRISTIAN CANTON [US] ET AL) 13 May 2021 (2021-05-13) * abstract; claims 1-20; figures 1,2 * * paragraphs [0005], [0020] - [0027], [0031] * -----	1-15	INV. G06N3/04 G06N3/08
A	SINGH ABHISHEK ET AL: "DISCO: Dynamic and Invariant Sensitive Channel Obfuscation for deep neural networks", 2021 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), IEEE, 20 June 2021 (2021-06-20), pages 12120-12130, XP034008778, DOI: 10.1109/CVPR46437.2021.01195 [retrieved on 2021-10-15] * the whole document * -----	1-15	
A	US 11 200 342 B1 (KIM TAE HOON [KR] ET AL) 14 December 2021 (2021-12-14) * the whole document * -----	1-15	
			TECHNICAL FIELDS SEARCHED (IPC)
A	US 11 244 248 B1 (KIM TAE HOON [KR]) 8 February 2022 (2022-02-08) * the whole document * -----	1-15	G06N

The present search report has been drawn up for all claims

1

Place of search	Date of completion of the search	Examiner
The Hague	13 October 2022	Manfrin, Max

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
....
& : member of the same patent family, corresponding document

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO.

EP 22 16 9013

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

13-10-2022

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2021141926 A1	13-05-2021	NONE	
US 11200342 B1	14-12-2021	KR 20220052837 A	28-04-2022
		US 11200342 B1	14-12-2021
		WO 2022086145 A1	28-04-2022
US 11244248 B1	08-02-2022	KR 20220052839 A	28-04-2022
		US 11244248 B1	08-02-2022
		WO 2022086147 A1	28-04-2022

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82