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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 August 2022 (25.08.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/178402 A1

(51) International Patent Classification:
G06N 20/00 (2019.01) *G06V 10/44* (2022.01)
G06F 21/62 (2013.01)

(21) International Application Number:
PCT/US2022/017262

(22) International Filing Date:
22 February 2022 (22.02.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/151,943 22 February 2021 (22.02.2021) US
63/184,497 05 May 2021 (05.05.2021) US

(71) Applicant: **MASSACHUSETTS INSTITUTE OF TECHNOLOGY** [US/US]; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(72) Inventors: **DEVADAS, Srinivas**; 7 Whittier Road, Lexington, MA 02420 (US). **XIAO, Hanshen**; 235 Albany St., Cambridge, MA 02139 (US).

(74) Agent: **RIMKUNAS, Zachary, J.**; Occhiuti & Rohlicek LLP, 50 Congress Street, Suite 1000, Boston, MA 02109 (US).

(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, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, 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,

(54) Title: PRIVATE LEARNING VIA TRANSFORMS AND TASK AUGMENTATION

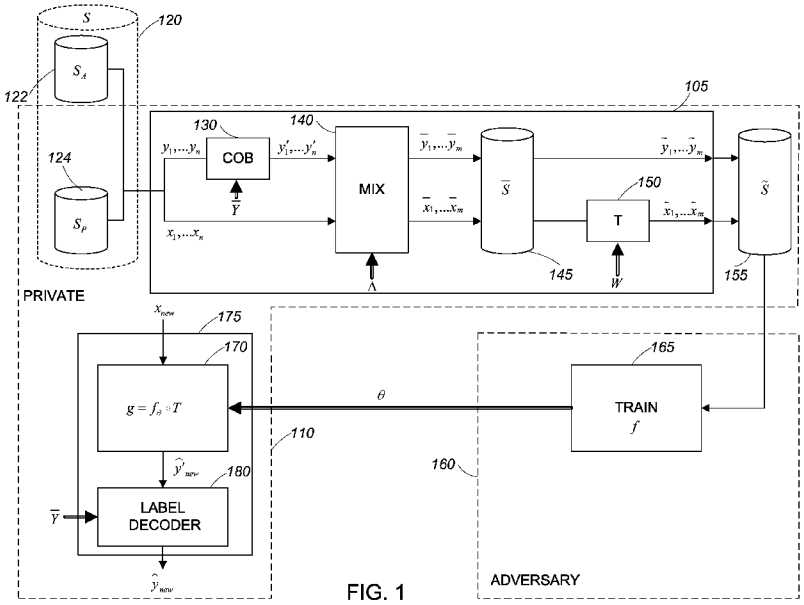

FIG. 1

(57) Abstract: A method for private learning includes, by a first party in a private computing environment, processing a first data set to form a privatized data set. The privatized data set includes a number of privatized data items, where each privatized data item is formed from multiple data items of the first data set. The privatized data set is passed to a second party for determining a public function that provides a mapping for data samples to approximations of corresponding labels.

[Continued on next page]

WO 2022/178402 A1

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))*

PRIVATE LEARNING VIA TRANSFORMS AND TASK AUGMENTATION

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 63/151,943 filed February 22, 2021 and U.S. Provisional Application No. 63/184,497 filed May 5, 2021.

BACKGROUND OF THE INVENTION

[0002] This invention relates to data privacy while having an adverse party process the data, for example, for machine learning.

[0003] Machine Learning (ML) is a generic concept used to build mathematical models based on data (samples) to make predictions or decisions. Within the broad category of data analysis, private ML is receiving growing attention. One aspect of private ML involves a party (the "private party") holding private data wishing that another party (e.g., an "adversary") process the private data to provide results of the processing. However, the private party does not want to disclose the private data to the adversary. An example of such processing includes a set of data (sample, label) items $S = \{s_i = (x_i, y_i), i = 1,...,n\}$ and wishing that the adversary provides training service that ultimately allows the user to have a function $g(x)$ such that $g(x_i) \approx y_i$, without the adversary gaining access to the data S.

SUMMARY OF THE INVENTION

[0004] Aspects described herein address a practical problem of using distributed computation without having to disclose private data to the parties performing the computation while still benefiting from the result of that computation for private functions. In particular, as is noted above, a private party does not want to disclose the data set yet wishes to have the adversary perform the bulk of the computation required to compute the function. A computation system uses a private computing environment under the control of and/or trusted by the private party, as well as an adversary computing environment under the control of the adversary and/or accessible by other parties to whom the private party does not wish to provide access to the data set. The private party

transforms the data set and passes the transformed data set to the adversary, who computes a public function. The private party uses private knowledge of the transformation applied to the data set and the public function to form a private function that corresponds to the original untransformed data set. The method can be generalized to multiple parties who wish to collaboratively have an adversary perform the bulk of the computation required to compute a combined function over all their private data sets, without any party disclosing their data set to other parties or the adversary.

[0005] In one general aspect, a method for private learning includes, by a first party in a private computing environment, processing a first data set to form a privatized data set, the privatized data set including a number of privatized data items, each privatized data item being formed from multiple data items of the first data set, and passing the privatized data set to a second party for determining a public function that provides a mapping for data samples to approximations of corresponding labels.

[0006] Aspects may include one or more of the following features.

[0007] The method may further include receiving from the second party parameter values characterizing the public function that provides a mapping for data samples to approximations of corresponding labels. The method may further include using parameter values characterizing the public function to form a private function for mapping new private sample values to corresponding estimated labels.

[0008] Processing the first data set may include transforming labels of the first data set according to a random basis transformation. Processing the first data set may include forming a mixed data set by random mixing of sets of data items from the first data set. Forming the mixed data set may include, for each data item of the mixed data set, forming a random convex combination of a random number of data items.

[0009] Processing the first data set may include transforming data samples by a transformation to form data samples of the privatized data set. The transformation may include a multiplication of a data sample by a random matrix. The transformation may include a first multi-layer artificial neural network. The transformation may preserve at least some spatial characteristics of the data sample. The transformation may include, for each subset of a number of subsets of a data sample applying at least one random nonlinear transformation to the subset.

[0010] The public function may include a multi-layer artificial neural network. Using the parameter values to form a private function may include forming a function

composition of the transformation and the public function. The first party may be a composite party consisting of mutually distrusting parties who each have private data sets and apply their own private transform to produce respective privatized data sets sent to the second party for determining the private function from the multiple privatized data sets.

[0011] In another general aspect, a method for learning includes augmenting a first data set representing a number of private (data, label) pairs known to a first party, forming a task augmented data set by combining a second data set with the first data set, wherein labels of the augmented data set represent combinations of labels from the data sets, and processing the augmented data set to determine processing parameters.

[0012] In another general aspect, a method for private learning includes accessing a first data set representing a number of private (data, label) pairs known to a first party, forming a task augmented data set by combining a second data set with the first data set, wherein labels of the augmented data set represent combinations of labels from the data sets, and processing the augmented data set to determine processing parameters without causing disclosure of the private pairs.

[0013] Aspects may include one or more of the following features.

[0014] The second data set may represent private data of a second party. The second data set may represent public data. The method may include processing a new data item by combining the new data item with multiple data items from the second data set to form one or more augmented new data items, processing the augmented new data items using the processing parameters to yield respective results, and aggregating the results to determine a processed form of the new data item. The data set may include data samples that are images and corresponding categorical labels.

[0015] Other features and advantages of the invention are apparent from the following description, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] FIG. 1 is a block diagram of a first configuration of a private learning system.

[0017] FIG. 2 is a block diagram of a second configuration of a private learning system that uses task augmentation.

[0018] FIG. 3 is a block diagram of a third configuration of a private learning system that uses federated data from multiple parties.

DETAILED DESCRIPTION

1 OVERVIEW

[0019] Referring to FIG. 1, a private computing environment 110 and an adversary computing environment 160 are connected by and communicate over a secure or publicly accessible communication link, for example, over the public Internet. The private computing environment 110 (e.g., a computer or set of interconnected computers) is under the control of and/or trusted by the private party and the adversary computing environment 160 is under the control of the adversary and/or accessible by other parties to whom the private party does not wish to provide access to the data set.

[0020] The private party does not wish to disclose its private data set S 120, yet wishes to have the adversary (i.e., an untrusted party), such as a "cloud computation service provider," perform the bulk of the computation required to determine parameters for a function such as a neural network using the private data set S 120. For example, the private party may be a hospital that wants to use private health information such as labelled grayscale mammogram images from their patients to train a neural network to detect signs of illnesses such as cancer. While the hospital wants to leverage the computational power of the adversary (e.g., an untrusted cloud computation service provider) for training their neural network, they may be legally prevented from disclosing the confidential patient information to others, including the adversary.

[0021] To address this problem, the private party (the hospital in our example) configures the private computing environment 110 to apply a transformation process 105 to the private data set S 120 (e.g., labelled mammogram images) to generate a transformed data set $\tilde{S}$ 155. The processing obfuscates the private data set S 120 such that it cannot be recovered (e.g., without substantial effort) from the transformed data set $\tilde{S}$ 155 without knowing the transformations performed by the processing. Furthermore, as is described in greater detail below, the transformations performed when generating the transformed data set $\tilde{S}$ 155 preserve certain spatial characteristics of data (e.g., image data) in the private data set S 120. The preservation of spatial characteristics may maintain amenability of the data to machine learning intended for processing image data,

for example, based on spatially-local processing as might be found in convolutional neural network approaches.

[0022] The transformed data set $\widetilde{S}$ 155 is sent over the communications link to the adversary computing environment 160, which processes the transformed data set $\widetilde{S}$ 155 in a training algorithm 165 to generate parameters, θ. The parameters, θ characterize a function, f that maps the transformed data of the transformed data set $\widetilde{S}$ 155 to approximations of transformed label data of the transformed data set $\widetilde{S}$ 155. In some examples, the function, f is convolutional (e.g., a multi-layer convolutional neural network (CNN)) and the preservation of certain spatial characteristics in the data (e.g., image data) of the transformed data set $\widetilde{S}$ 155 ensures that the training algorithm 165 is able to generate useful parameters, θ from the transformed data set $\widetilde{S}$ 155.

[0023] The parameters, θ are sent back to the private computing environment 110 over the communications link and used in an inference process 175 to infer a label, $\hat{y}_{new}$ from a new, previously unknown sample, x_{new}. For example, the hospital could use the interference process 175, parameterized by θ to process a newly acquired mammogram image to categorize the image into one of several categories (e.g., benign, suspicious, malignant, etc.).

2 TRANSFORMATION PROCESS

[0024] The transformation process 105 receives the private data set S 120 as input, where S includes n (data, label) pairs such that $S = \left((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\right)$. In some examples, the private data set S 120 is a combination of a publicly known part of the data set, S_A 122, and a privately known part of the data set, S_P 124, where the large publicly known data set S_A 122 is used to augment the relatively smaller privately known data set S_P 124. The transformation process 105 processes the private data set S to generate the transformed data set $\tilde{S}$ 155 as output, where $\tilde{S}$ includes m transformed data, label pairs such that $\widetilde{S} = \left((\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), ..., (\tilde{x}_m, \tilde{y}_m)\right)$.

[0025] In some examples, the transformation process 105 includes a change of basis (COB) module 130, a data mixing module (MIX) 140, and a data transformation module (T) 150.

2.1 Change of Basis

[0026] First, in a case where the labels come from a discrete set, the labels ($y_1,...,y_n$) of the private data set S 120 are provided to the change of basis module 130 which changes the basis of the labels such that they are unpredictable or difficult to predict by the adversary using information available to the adversary (i.e., no using the private party's information). For example, labels indicating a category associated with a particular mammogram image would be obfuscated by the change of basis module 130. The output of the change of basis module 130 is denoted by as modified label data $(y'_1,...,y'_n)$.

[0027] In some examples, where the $y_i \in [1:c]$ (i.e., the labels y_i are in c categories), the change of basis module 130 randomly generates (or receives as input) a set of c orthogonal (or almost orthogonal) c' -dimensional (row) vectors $\overline{Y}_1,...,\overline{Y}_c$ (where $c' \geq c$). To form the $(y'_1,...,y'_n)$ output, each y_i is replaced with $y'_i = \overline{Y}_{y_i}$. In examples where the y_i are represented as "one-hot" (c -dimensional row) vectors, the transformation of the labels can be represented as $y_i\overline{Y}$ where $\overline{Y}$ is a $c \times c'$ matrix. Note that reference to "random" in this document should be understood broadly to include pseudo-random, or otherwise unpredictable or difficult to predict by the adversary using information available to adversary (i.e., not using private information of the private party).

2.2 Data Mixing

[0028] The labels output from the change of basis module 130 $(y'_1,...,y'_n)$ and the original data samples $(x_1,...,x_n)$ (e.g., image data) are then provided as inputs to the data mixing module 140, which processes the inputs to generate a mixed data set $\overline{S} = \{\overline{s}_i = (\overline{x}_j, \overline{y}_j), j = 1,...,m\}$ 145. In some examples, each mixed data item $\overline{s}_i$ includes a randomly weighted combination (e.g., element-wise random mixture, e.g., a random convex combination) of two (or more) (e.g., randomly selected) original data samples from $(x_1,...,x_n)$ and a randomly weighted combination of two or more of the modified labels $(y'_1,...,y'_n)$. In some examples, the data mixing module 140 is parameterized by random parameters denoted Λ, which includes the convex weights and the selected data and modified labels to be mixed to form the samples $\overline{s}_1,...,\overline{s}_m$ of the mixed data set $\overline{S}$ 145. When original data samples $(x_1,...,x_n)$ are image data such as mammogram images,

the mixed images $\overline{x}_1,...,\overline{x}_m$ include a randomly weighted combination of multiple original mammogram images, which may still be identifiable as comprising mammogram images. In the discussion below, the images are assumed to be of size $\sqrt{d}$ pixels square (i.e., a total of d pixels per image, and the pixels of the image may be manipulated as a d-element vector of a $\sqrt{d}\times\sqrt{d}$ pixel image as appropriate and evident from the context.)

2.3 Data Transformation

[0029] The mixed data samples $\overline{x}_1,...,\overline{x}_m$ of the mixed data set $\overline{S}$ 145 are then provided to the data transformation module 150, which applies a transform $T(x)$ to each mixed data sample $\overline{x}_j$ to generate a transformed data sample $\tilde{x}_i = T(\overline{x}_i)$. The output of the data transformation module 150 is the transformed mixed data samples, $\tilde{x}_1,...,\tilde{x}_m$.

[0030] In some examples, the transformation $T(x)$ is parameterized by one or more random values, W. In some examples, the random values include L weight matrices $W_1 \in \mathbb{R}^{d\times d}, W_2 \in \mathbb{R}^{d\times d}, ..., W_L \in \mathbb{R}^{d\times d}$. In at least some embodiments, the values in these matrices are chosen at random from the set $\{-1,0,1\}$. The L weight matrices are used to construct a neural network structure of L hidden layers with an activation function σ (e.g., a sigmoid or a hyperbolic tangent function) and weights $W[1:L]$: $\mathcal{N}_W(x) = \sigma\left(...\sigma\left(\sigma\left(x\cdot W_1\right)\cdot W_2\right)...W_L\right)$. Furthermore, the neural network structure configures each layer as a spatially local and spatially varying nonlinear transformation of the data. In particular, when the data is an image the layers of the neural network implement a separate set of different nonlinear transformations of contiguous blocks (e.g., 3x3 pixel blocks) of the image (which can be considered equivalently to requiring that elements of the W matrices are zero other than for elements that are in the same block). As is mentioned above, operating on spatially local blocks of the input data preserves spatial information that is needed by convolutional neural network training and inference algorithms to learn and use spatial hierarchies of features in the data. In the case where $\overline{x}_1,...,\overline{x}_m$ are mixed mammogram images, the transformation $T(x)$ further obfuscates the image data, making it difficult or impossible for the adversary to recover the original image data from the transformed images without knowledge of the transform.

[0031] The transformed mixed data samples $\tilde{x}_1,...,\tilde{x}_m$ output from the data transformation module 150 are combined with the mixed labels $\overline{y}_1,...,\overline{y}_n$ to form the transformed data set $\tilde{S}$ 155, which is output from the transformation process 105.

3 TRAINING

[0032] The private party then transmits a transformed data set $\tilde{S} = \left\{\tilde{s}_j = \left(\tilde{x}_j, \tilde{y}_j\right)\right\}$ to the adversary computing environment 160 where it is used to generate parameter values θ. In particular, the adversary computing environment 160 includes a training module 165 the processes the transformed data set $\tilde{S}$ 155 to determine parameter values θ that characterize a function $f\left(\tilde{x}\right)$ such that $f\left(\tilde{x}_j\right) \approx \tilde{y}_j$ for the transformed data set $\tilde{S}$ 155.

4 INFERENCE

[0033] The adversary then transmits the parameter values θ of the function f back to the private computing environment 110 where the parameters are used to configure the inference process 175. The inference process 175 includes an inference module 170 and a label decoder 180. To configure the inference module 170, the private computing environment 110 computes parameters of a function $g\left(x_i\right) \approx y_i'$ (i.e., a function that fits the original data set S) by forming a composition $g = f \circ T$ such that $y_i' \approx g(x_i)$. The label decoder 180 is configured to undo the change of basis defined by $\overline{Y}$.

[0034] The inference process 175 carried out by the private party uses a new data sample x_{new} (e.g., a previously unseen mammogram image) and to provide x_{new} to the inference module 170. The inference module 170 processes x_{new} to yield a predicted output $\hat{y}'_{new} = g(x_{new})$. The predicted output $\hat{y}'_{new}$ is provided to the label decoder, which maps $\hat{y}'_{new}$ to the original label form $\hat{y}_{new}$, which may for example in the case of one-hot encoding be implemented as $\hat{y}_{new} = \hat{y}'_{new}\overline{Y}^{-1}$. The output of the label decoder 180 $\hat{y}_{new}$ represents a label associated with the new data sample x_{new} (e.g., indicating whether and to what extent the previously unseen mammogram image indicates disease).

5 TASK AUGMENTATION

[0035] Referring to FIG. 2, in another example, the private party does not want to disclose a private data set S_P 224, which includes (or otherwise represents) n (data,

label) pairs $\{(x_i, y_i)\}$ yet wishes to have the adversary (i.e., any untrusted party), such as a "cloud computation service provider," perform the bulk of the computation required to train a mapping from a new data item x_{new} to an estimate of the label $\hat{y}_{new}$. Very generally, a computation system uses a private computing environment 210 (e.g., a computer or set of interconnected computers) under the control of and/or trusted by the private party, as well as an adversary computing environment 260 under the control of the adversary and/or accessible by other parties to whom the private party does not wish to provide access to the data set. In some implementations, the private computing environment and the adversary computing environment are connected by a secure or publicly accessible communication link, for example, over the public Internet.

[0036] The private party uses an auxiliary data set S_A 220, which also includes (data, label) pairs, yet the private party is not necessarily concerned with disclosure of this auxiliary data set. For example, the auxiliary data set may be a publicly available data set. Rather than sending the private data set S_P 224 to the adversary, the private party forms an augmented data set $\tilde{S}$ 245. This data set is formed by combinations of randomly selected and transformed (data, label) pairs from the private data set and randomly selected and transformed (data, label) pairs from auxiliary data set. For example, as illustrated in FIG. 2, a random number generator (RNG) 232 generates a series of index pairs $(j_1, l_1), \ldots, (j_m, l_m)$. Each index pair is used to access samples from the auxiliary and private data sets, respectively (i.e., $\left(x'_{l_i}, y'_{l_i}\right)$ and $\left(x_{j_i}, y_{j_i}\right)$).

[0037] The accessed samples are transformed in data transformation modules 105a, 105b, respectively (e.g., a data transformation module described in FIG. 1), to form transformed samples $\left(\tilde{x}'_{l_i}, \tilde{y}'_{l_i}\right)$ and $\left(\tilde{x}_{j_i}, \tilde{y}_{j_i}\right)$. Then, transformed samples associated with the index pairs are combined to form a task augmented (data, label) pair $(\tilde{x}, \tilde{y}) = \left(\tilde{x}_{j_i} \mid \tilde{x}'_{l_i}, \tilde{y}_{j_i} \mid \tilde{y}'_{l_i}\right)$, where the augmentation is denoted with a "|" operator 234. For example, if the labels come from a set of K possible labels, the augmented labels may come from a set of K^2 combinations of labels. For example, if each label is encoded as a one-hot representation in a K-dimensional vector, or a representation using a random K-dimensional basis as introduced above, the result after augmentation is a $2K$-dimensional vector. Similarly, if the data items are images with d pixels or d-dimensional embeddings, the augmented data item may have $2d$ pixels or embedding elements formed by concatenation (e.g., by stacking channels or appending in pixel space). Note that in some examples, the data elements and/or the labels may not be

transformed by private transformations, and it is possible that the private data and the auxiliary data use different transformations.

[0038] The augmented data set $\tilde{S}$ 245 is passed to the adversary computing environment 260, where it is used by a training component 265 to determine values of parameters θ that are used to configure a runtime classifier 280 (also located on the adversary computing environment 260).

[0039] When the private party wishes to classify a new data item x_{new}, it retrieves a set of auxiliary data items x_i', transforms x_{new} and x_i' using data transformation modules 105c, 105d, and then combines the transformed data items to form a set of augmented data items $\{\tilde{x}\} = \{(\tilde{x}_{new} \mid \tilde{x}_i')\}$. Each of the augmented data items is passed to the adversary, where the runtime classifier 280 classifies the data items and returns a classification result $\{\hat{\tilde{y}}_{new}\} = \{(\hat{y}_{new} \mid y_i')\}$ to the private party. In the case where there are L labels, the returned classification result for each augmented data item may represent a selection from the L^2 possibilities, or alternatively a probability distribution over those entries. The private party then uses an aggregation component 290 to combine the returned classification results to yield the classification $\hat{y}_{new}$ for the private new data item.

[0040] Preferably, the auxiliary data items have similar character to the private data items. For example, if the private data items are images of animals, then preferably the auxiliary data items are also images of animals. It is not necessary that the labels be chosen from the same set, for example, a cat may be labeled as animal 3 in the private data and animal 15 in the auxiliary data.

[0041] Note that the combination operator "|" may be more complex than mere concatenation. For example, there may be further mixing and/or stacking of the data items for example, with a random linear transformation, and the labels may be processed by a random invertible transformation prior to or after forming the cross-products.

[0042] Not illustrated in FIG. 2 is an option where the parameter values θ are transferred to the private computing environment 210, and the private party executes the runtime classification component 280.

6 FEDERATED TRAINING

[0043] In some examples, multiple parties can benefit by combining their data for the purpose of training parameters θ. Referring to FIG. 3, a federated training approach facilitates a combined training scheme while maintaining privacy of the individual parties' data.

[0044] In FIG. 3, a number of private parties (referred to as parties A through Z, for example separate hospitals with private sets of mammograms) have respective private data sets $S_P^{(A)}$ through $S_P^{(Z)}$. Data items from the private data sets are processed in respective party-specific data transformation modules 105A through 105Z (e.g., the data transform module of FIG. 1 configured with random parameters that are specific to each private party). In a manner somewhat similar to the embodiment illustrated in FIG. 2, each private party provides its transformed data from its private computing environment to a combiner 234 in the adversary's computing system. The combiner 234 combines (e.g., concatenates or stacks) the transformed data items from the private parties, for example, in random combinations to form a combined dataset $\tilde{S}$ 345. The adversary processes the combined dataset $\tilde{S}$ 345 using a training module 265 to train the parameters θ for use in a runtime classification task.

[0045] At runtime, when party A wishes to classify a new data item x_{new} it accesses transformed data items for the other parties (e.g., via the adversary or directly) and forms a set of augmented data items $\tilde{x}_{new}$ using the same combination approach (236) as used in training to combine the data items 234. In this example, at no point does Party A have to see the original, private data items of other parties, nor does it need to know the private transform applied by other parties. Party A then aggregates the received classifications to form the classification for its private new data item.

[0046] In some examples, rather than the training being done in the adversary environment, each party shares its transformed data with the other parties. Then these other parties can benefit from multiple sets of data to train their instance of the runtime inference component 280.

[0047] Note that in some examples, it is not feasible to concatenate or otherwise combine for task augmentation data from all the parties. In such a case, training using data from different subsets of parties may be used such that each subset is used to train a

different inference system. Then at runtime, a party's data may be combined in different subsets to yield multiple inference outputs that are then combined.

7 ALTERNATIVES

[0048] Embodiments can include one or more of the following separate features (i.e., unless otherwise indicated, none of these features is essential and the features are mutually compatible).

[0049] In some examples, the transformation T includes multiplication by a random matrix, W_1 (e.g., $\tilde{x} = x_i W_1$). In some examples, as is described above, the random values include L weight matrices $W_1 \in \mathbb{R}^{d \times d}, W_2 \in \mathbb{R}^{d \times d}, ..., W_L \in \mathbb{R}^{d \times d}$. The L weight matrices are used to construct a neural network structure of $L \geq 1$ hidden layers with an activation function σ and weights $W[1:L]$: $\mathcal{N}_W(x) = \sigma\left(..., \sigma\left(\sigma\left(x \cdot W_1\right) \cdot W_2\right), ..., W_L\right)$. Furthermore, the neural network structure configures each layer as a spatially local and spatially varying nonlinear transformation of the data.

[0050] In some examples, the weight matrices W_k are all chosen at random and σ is an elementwise non-linear function. For example, the non-linear functions may be ReLU functions (i.e., rectifier functions), which results in T being piecewise linear.

[0051] In some examples, the transformation f computed by the adversary includes a multiplication by a matrix F_1 (e.g., $\tilde{x}_i F_1$). In some examples, the transformation f is a $L' \geq 1$ layer artificial neural network, which computes $f(\tilde{x}) = \sigma\left(...\sigma\left(\sigma\left(\tilde{x}F_1\right)F_2\right)...F_{L'}\right)$, where the weight matrices F_k are estimated (i.e., optimized) by the adversary and σ is an elementwise non-linear function.

[0052] In some examples, the data set S is made up of a publicly-known (or at least known to the adversary) subset S_A and a private subset S_P, the mixing includes at least one member of the private subset in each convex combination, thereby making it more difficult for the adversary to learn the transformation T.

[0053] As is mentioned above, the transformation T maintains locality, for example, if the data samples x_i are images (e.g., arrays of pixel values), the transformation T may yield output pixels that depend only on neighborhoods of input pixel values. The functions f and f can therefore be convolutional.

[0054] Aspects described above mention to the processing of labeled grayscale images such as mammogram data. However, it is noted that the techniques (e.g., transformations) described above can be extended for use on other types of data such as RGB image data.

8 IMPLEMENTATIONS

[0055] The approaches described above can be implemented, for example, using a programmable computing system executing suitable software instructions or it can be implemented in suitable hardware such as a field-programmable gate array (FPGA) or in some hybrid form. For example, in a programmed approach the software may include procedures in one or more computer programs that execute on one or more programmed or programmable computing system (which may be of various architectures such as distributed, client/server, or grid) each including at least one processor, at least one data storage system (including volatile and/or non-volatile memory and/or storage elements), at least one user interface (for receiving input using at least one input device or port, and for providing output using at least one output device or port). The software may include one or more modules of a larger program, for example, that provides services related to the design, configuration, and execution of dataflow graphs. The modules of the program (e.g., elements of a dataflow graph) can be implemented as data structures or other organized data conforming to a data model stored in a data repository.

[0056] The software may be stored in non-transitory form, such as being embodied in a volatile or non-volatile storage medium, or any other non-transitory medium, using a physical property of the medium (e.g., surface pits and lands, magnetic domains, or electrical charge) for a period of time (e.g., the time between refresh periods of a dynamic memory device such as a dynamic RAM). In preparation for loading the instructions, the software may be provided on a tangible, non-transitory medium, such as a CD-ROM or other computer-readable medium (e.g., readable by a general or special purpose computing system or device), or may be delivered (e.g., encoded in a propagated signal) over a communication medium of a network to a tangible, non-transitory medium of a computing system where it is executed. Some or all of the processing may be performed on a special purpose computer, or using special-purpose hardware, such as coprocessors or field-programmable gate arrays (FPGAs) or dedicated, application-specific integrated circuits (ASICs). The processing may be implemented in a distributed manner in which different parts of the computation specified by the software are performed by different

computing elements. Each such computer program is preferably stored on or downloaded to a computer-readable storage medium (e.g., solid state memory or media, or magnetic or optical media) of a storage device accessible by a general or special purpose programmable computer, for configuring and operating the computer when the storage device medium is read by the computer to perform the processing described herein. The inventive system may also be considered to be implemented as a tangible, non-transitory medium, configured with a computer program, where the medium so configured causes a computer to operate in a specific and predefined manner to perform one or more of the processing steps described herein.

[0057] A number of embodiments of the invention have been described. Nevertheless, it is to be understood that the foregoing description is intended to illustrate and not to limit the scope of the invention, which is defined by the scope of the following claims. Accordingly, other embodiments are also within the scope of the following claims. For example, various modifications may be made without departing from the scope of the invention. Additionally, some of the steps described above may be order independent, and thus can be performed in an order different from that described.

WHAT IS CLAIMED IS:

1. A method for private learning comprising, by a first party in a private computing environment:

processing a first data set (S) to form a privatized data set ($\tilde{S}$), the privatized data set including a plurality of privatized data items ($\tilde{s}_j = \left(\tilde{x}_j, \tilde{y}_j\right)$), each privatized data item being formed from multiple data items of the first data set; and

passing the privatized data set to a second party for determining a public function (f) that provides a mapping for data samples ($\tilde{x}_j$) to approximations of corresponding labels ($\tilde{y}_j$).

2. The method of claim 1, further comprising:

receiving from the second party parameter values (θ) characterizing the public function (f) that provides a mapping for data samples ($\tilde{x}_j$) to approximations of corresponding labels ($\tilde{y}_j$).

3. The method of claim 1 or claim 2, further comprising:

using parameter values characterizing the public function (f) to form a private function (g) for mapping new private sample values (x_{new}) to corresponding estimated labels ($\hat{y}'_{new}; \hat{y}_{new}$).

4. The method of any of the preceding claims, wherein processing the first data set (S) comprises:

transforming labels y_k of the first data set according to a random basis transformation.

5. The method of any of the preceding claims, wherein processing the first data set (S) comprises:

forming a mixed data set ($\overline{S}$) by random mixing of sets of data items from the first data set.

6. The method of claim 5, wherein forming the mixed data set comprises, for each data item ($\overline{s}_j$) of the mixed data set, forming a random convex combination of a random plurality of data items.

7. The method of any of the preceding claims, wherein processing the first data set (S) comprises:

transforming data samples ($\overline{x}_j$) by a transformation (T) to form data samples ($\tilde{x}_j$) of the privatized data set.

8. The method of claim 7, wherein the transformation(T) comprises a multiplication of a data sample ($\overline{x}_j$) by a random matrix (W_1).

9. The method of claim 8, wherein the transformation (T) comprises a first multi-layer artificial neural network.

10. The method of claim 7, wherein the transformation (T) preserves at least some spatial characteristics of the data sample ($\overline{x}_j$).

11. The method of claim 8 wherein the transformation (T) includes, for each subset of a plurality of subsets of a data sample ($\overline{x}_j$) applying at least one random nonlinear transformation to the subset.

12. The method of any of the preceding claims, wherein the public function (f) comprises a multi-layer artificial neural network.

13. The method of any of the preceding claims, wherein using the parameter values to form a private function comprises forming a function composition ($f \circ T$) of the transformation (T) and the public function (f).

14. The method of any one of claims 1 through 13, wherein the first party is a composite party consisting of mutually distrusting parties (A, B) who each have private data sets and apply their own private transform (T_A, T_B) to produce respective privatized data sets sent to the second party for determining the private function from the multiple privatized data sets.

15. A method for learning comprising:

augmenting a first data set ($S_P; S_P^{(A)}$) representing a plurality of private (data, label) pairs known to a first party;

forming a task augmented data set ($\tilde{S}$) by combining a second data set ($S_A; S_P^{(Z)}$) with the first data set, wherein labels of the augmented data set represent combinations of labels from the data sets; and

processing the augmented data set to determine processing parameters (θ).

16. A method for private learning comprising:

accessing a first data set ($S_P; S_P^{(A)}$) representing a plurality of private (data, label) pairs known to a first party;

forming a task augmented data set ($\tilde{S}$) by combining a second data set ($S_A; S_P^{(Z)}$) with the first data set, wherein labels of the augmented data set represent combinations of labels from the data sets; and

processing the augmented data set to determine processing parameters (θ) without causing disclosure of the private pairs.

17. The method of claim 15 or claim 16, wherein the second data set represents private data ($S_P^{(Z)}$) of a second party.

18. The method of claim 15 or claim 16, wherein the second data set represents public data (S_A).

19. The method of any of claims 15 through 18, further comprising processing a new data item (x_{new}) by:

combining the new data item with multiple data items from the second data set to form one or more augmented new data items ($\tilde{x}_{new}$);

processing the augmented new data items using the processing parameters to yield respective results; and

aggregating the results to determine a processed form of the new data item.

20. The method of any of the preceding claims, wherein the data set comprises data samples that are images and corresponding categorical labels.

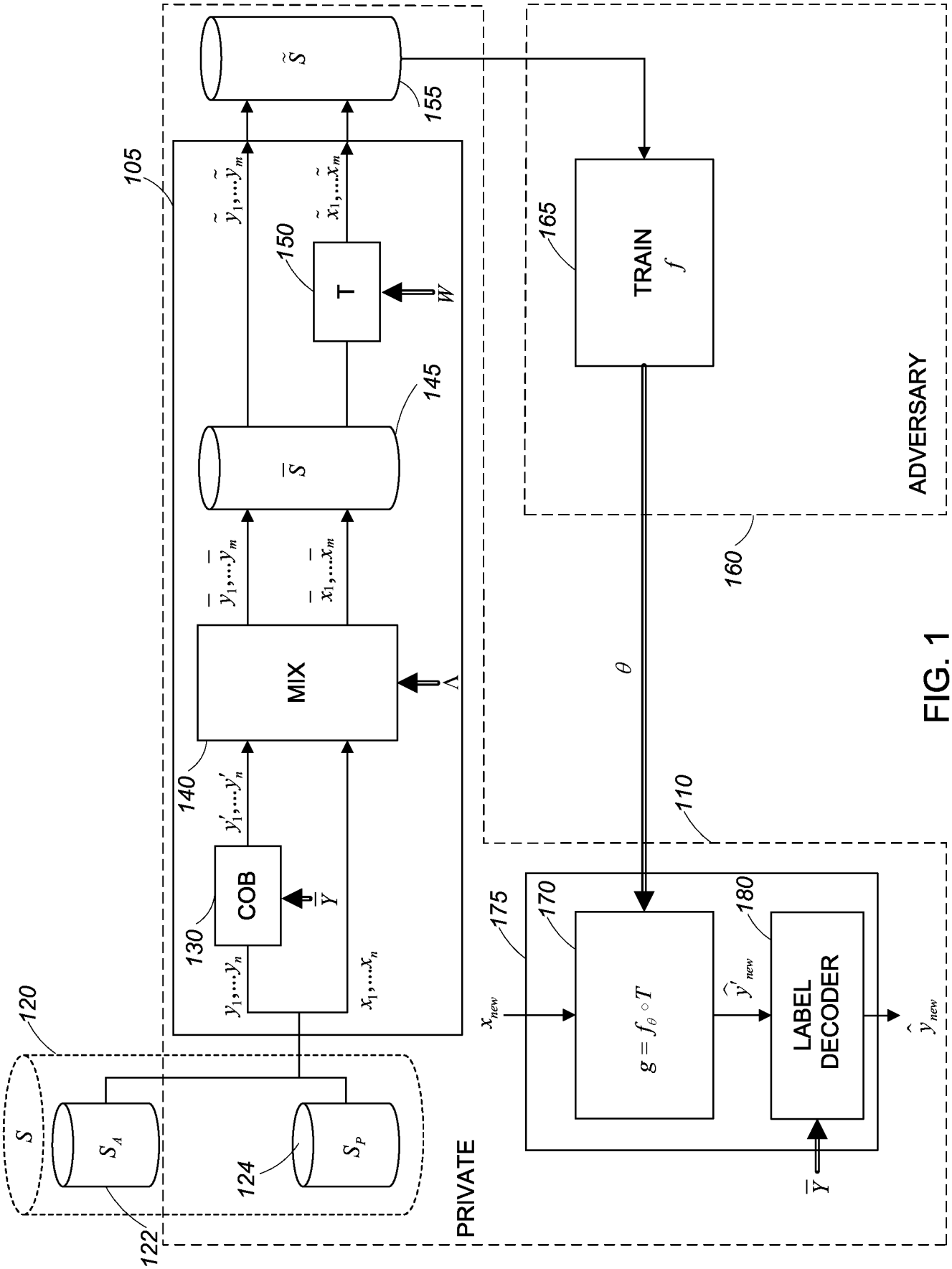
PRIVATE
ADVERSARY
MIX
COB
T
TRAIN
LABEL DECODER
105
110
120
122
124
130
140
145
150
155
160
165
170
175
180
$\tilde{S}$
$\bar{S}$
S
S_A
S_P
Λ
$\bar{Y}$
W
θ
$g = f_\theta \circ T$
x_{new}
$\hat{y}'_{new}$
$\hat{y}_{new}$
$y_1,...y_n$
$x_1,...x_n$
$y'_1,...y'_n$
$\bar{y}_1,...\bar{y}_m$
$\bar{x}_1,...\bar{x}_m$
$\tilde{y}_1,...\tilde{y}_m$
$\tilde{x}_1,...\tilde{x}_m$

FIG. 1

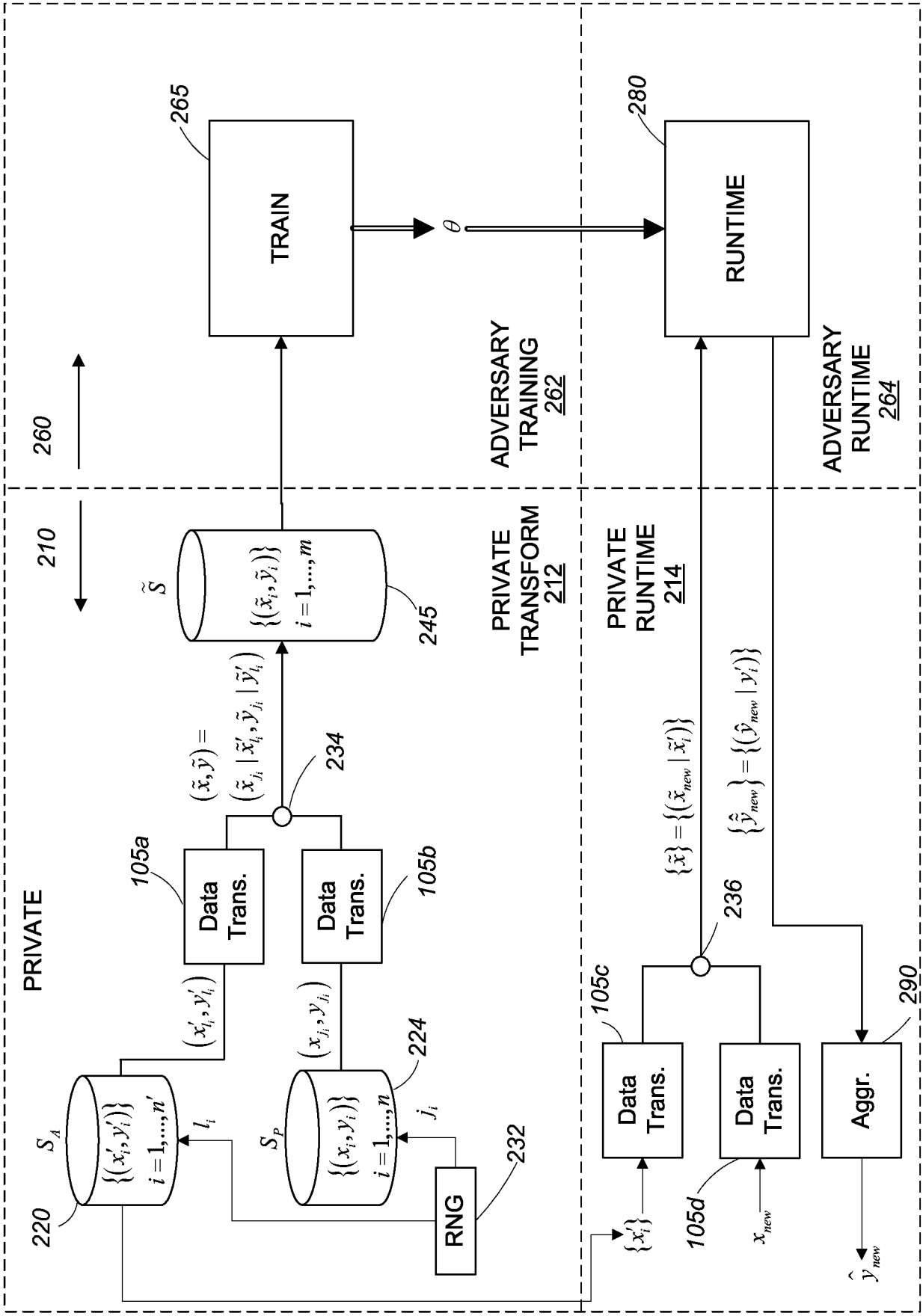
PRIVATE
210
260
ADVERSARY TRAINING 262
PRIVATE TRANSFORM 212
PRIVATE RUNTIME 214
ADVERSARY RUNTIME 264
TRAIN
RUNTIME
Data Trans.
Data Trans.
Data Trans.
Data Trans.
Aggr.
RNG
105a
105b
105c
105d
220
224
232
234
236
245
265
280
290
θ

FIG. 2

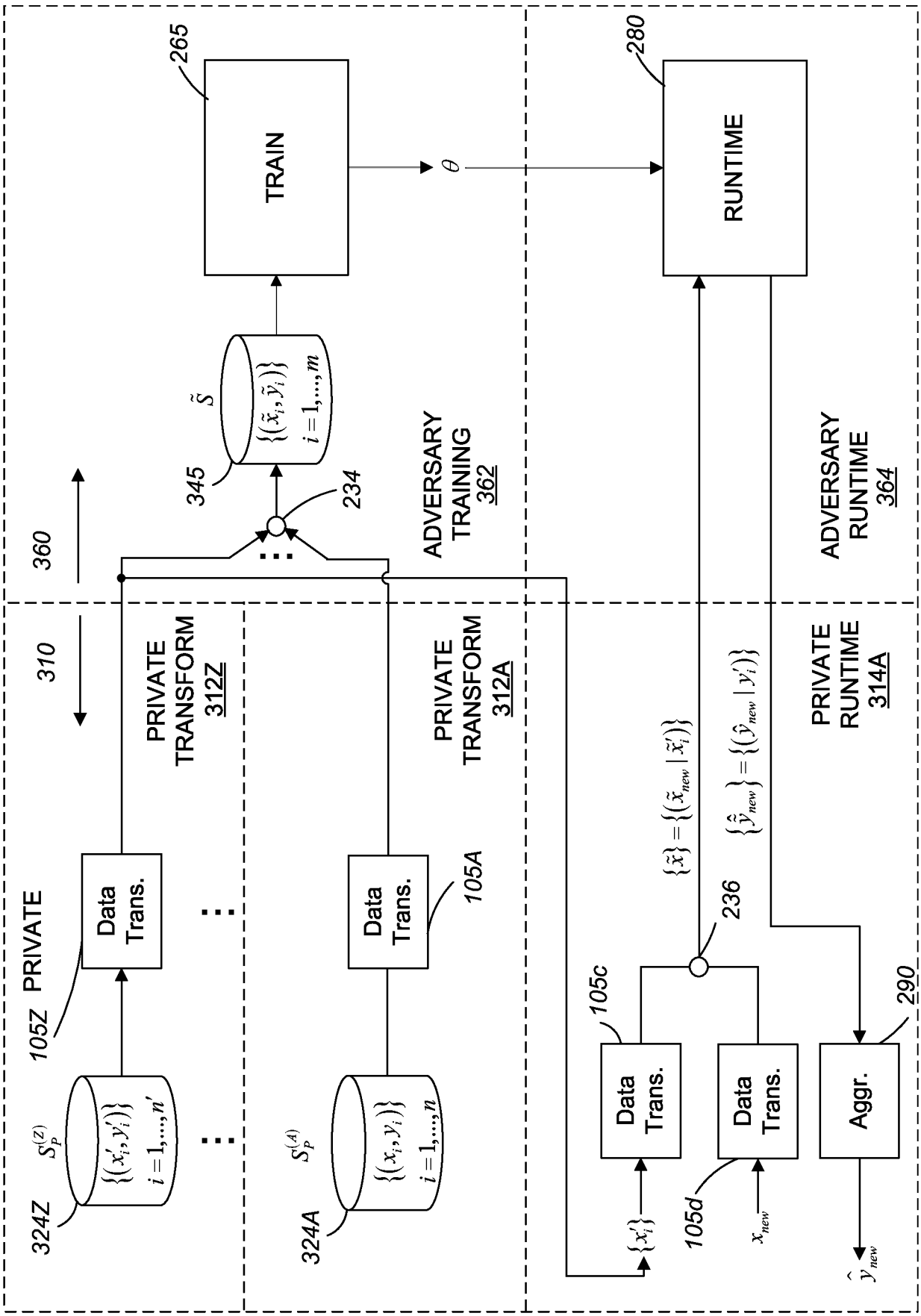
360
310
PRIVATE
105Z
PRIVATE TRANSFORM 312Z
Data Trans.
324Z
$S_P^{(Z)}$
$\{(x'_i, y'_i)\}$
$i=1,...,n'$
105A
PRIVATE TRANSFORM 312A
Data Trans.
324A
$S_P^{(A)}$
$\{(x_i, y_i)\}$
$i=1,...,n$
265
TRAIN
θ
345
$\tilde{S}$
$\{(\tilde{x}_i, \tilde{y}_i)\}$
$i=1,...,m$
234
ADVERSARY TRAINING 362
280
RUNTIME
ADVERSARY RUNTIME 364
PRIVATE RUNTIME 314A
$\{\tilde{x}\} = \{(\tilde{x}_{new} \mid \tilde{x}'_i)\}$
$\{\hat{\tilde{y}}_{new}\} = \{(\hat{\tilde{y}}_{new} \mid y'_i)\}$
236
105c
Data Trans.
105d
Data Trans.
290
Aggr.
$\{x'_i\}$
x_{new}
$\hat{y}_{new}$

FIG. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/017262

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N20/00 G06F21/62 G06V10/44
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)
H04L G06N G06F G06K G06V

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ALEX MANSBRIDGE ET AL: "Learning to Noise: Application-Agnostic Data Sharing with Local Differential Privacy", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 23 October 2020 (2020-10-23), XP081795323, sections 2, 3, 4, 5 -----	1-20
A	MATHILDE RAYNAL ET AL: "Image Obfuscation for Privacy-Preserving Machine Learning", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 20 October 2020 (2020-10-20), XP081790794, sections 2, 3, 4 -----	1-20

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 May 2022	24/05/2022

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

1