(12)Indian Patent Application

---

(21) Application Number: 201841050033

(22) Filing Date: 31/12/2018 (43) Publication Date: 03/07/2020

(71) Applicant(s): L&T TECHNOLOGY SERVICES LIMITED

(72) Inventor(s): Balaraman, Mridul
Singh, Madhusudan
Kumar, Amit
Gupta, Mrinal

(51) International Classifications: G06N 20/00 G06K 9/62 H04L 27/26 G06N 3/12 G06T 7/00

(86) International Application Number and Date: PCT/IB2019/058264 28/09/2019

(87) International Publication Number: WO/2020/065627

(54) Title: A SYSTEM AND METHOD FOR VALIDATING OPTICAL CHARACTER RECOGNITION OUTPUT

(57) Abstract: A method and a device (102) for creating and training machine learning models is disclosed. In an embodiment, a method for training a machine learning model for identifying entities from data includes creating (302) a first plurality of clusters from a first plurality of data samples in a first dataset (204) and a second plurality of clusters from a second plurality of data samples in a second dataset (206). The method further includes determining (304) a rank for each of the first plurality of clusters and a rank for each of the second plurality of clusters (306). The method includes retraining (308) the machine learning model using at least one of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

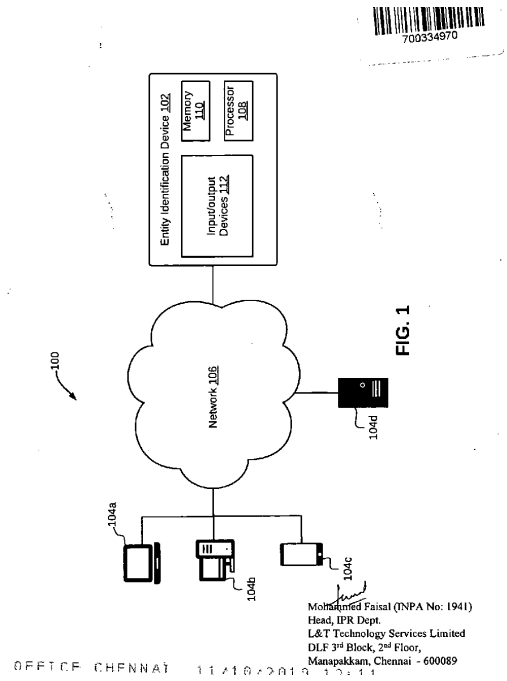

FIG. 1

ABSTRACT

A method and a device **(102)** for creating and training machine learning models is disclosed. In an embodiment, a method for training a machine learning model for identifying entities from data includes creating **(302)** a first plurality of clusters from a first plurality of data samples in a first dataset **(204)** and a second plurality of clusters from a second plurality of data samples in a second dataset **(206)**. The method further includes determining **(304)** a rank for each of the first plurality of clusters and a rank for each of the second plurality of clusters **(306)**. The method includes retraining **(308)** the machine learning model using at least one of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

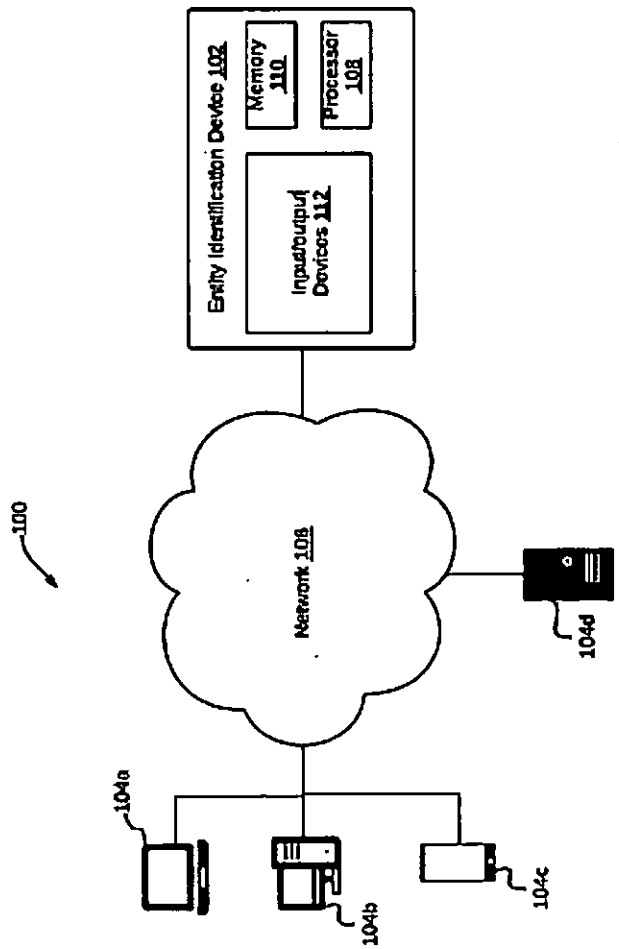

FIG. 1

We Claim:

1. A method for training a machine learning model for identifying entities from data, the method comprising:

creating **(302)** a first plurality of clusters from a first plurality of data samples in a first dataset **(204)**, based on a first set of entity attributes associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset **(206)**, based on a second set of entity attributes associated with the second plurality of data samples, wherein each of the first dataset **(204)** and the second dataset **(206)** are used to train a machine learning model to identify an entity from data, and wherein the first plurality of data samples comprise the entity and the second plurality of data samples do not comprise the entity;

determining **(304)** a rank for each of the first plurality of clusters based on a probability of identification, as determined by the machine learning model, of an associated set of data samples within the first plurality of data samples as the entity;

determining **(306)** a rank for each of the second plurality of clusters based on a probability of mismatch, as determined by the machine learning model, of an associated set of data samples within the second plurality of samples, with the entity;

retraining **(308)** the machine learning model using at least one of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters; and

creating **(702)** the first dataset **(204)** comprising the first plurality of data samples and the second dataset **(206)** comprising the second plurality of data samples.

2. The method of claim 1, wherein creating the first dataset **(204)** and the second dataset **(206)** comprises:

receiving **(704)**, by the machine learning model, a first data sample comprising the entity and a second data sample not comprising the entity;

identifying **(706)**, by the machine learning model, a first set of results matching with the first data sample and a second set of results matching with the second data sample;

determining **(708)** accuracy in identification of each result in a subset of the first set of results, based on a user input corresponding to the first data sample, wherein the subset is selected based on a predefined criterion; and

determining **(716)** error in misidentification of each result in a subset of the second set of results, based on a user input corresponding to the second data sample, wherein the subset is selected based on the predefined criterion.

3. The method of claim 2, further comprising:

adding **(712)** the first data sample to the first dataset **(204)**, when each result in the subset of the first set of results does not match with the user input; and

adding **(720)** the second data sample to the second dataset **(206)**, when each result in the subset of the second set of results does not match with the user input.

4. A method for creating a machine learning model for identifying entities from data, the method comprising:

creating **(802)** a first plurality of clusters from a first plurality of data samples in a first dataset **(204)**, based on a first set of entity attributes associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset **(206)**, based on a second set of entity attributes associated with the second plurality of data samples, wherein the first plurality of data samples comprise an entity and the second plurality of data samples do not comprise the entity;

assigning **(804)** the same rank to each of the first plurality of clusters based on the total number of the first plurality of clusters;

determining **(806)** a rank for each of the second plurality of clusters based on similarity with at least one cluster from the first plurality of clusters, wherein a cluster from the second plurality of clusters having highest similarity with the at least one cluster is assigned the highest rank and a cluster from the second plurality of clusters having lowest similarity with the at least one cluster is assigned the lowest rank; and

creating **(808)** a machine learning model to identify the entity, wherein the machine learning model is created using at least one of the first plurality of clusters weighted based on the rank assigned to each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

5. An entity identification device **(102)** for creating a machine learning model for identifying entities from data, the entity identification device comprising:

a processor **(108)**; and

a memory **(110)** communicatively coupled to the processor, wherein the memory stores processor instructions, which, on execution, causes the processor to:

create **(802)** a first plurality of clusters from a first plurality of data samples in a first dataset **(204)**, based on a first set of entity attributes associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset **(206)**, based on a second set of entity attributes associated with the second plurality of data samples, wherein the first plurality of data samples comprise an entity and the second plurality of data samples do not comprise the entity;

assign **(804)** the same rank to each of the first plurality of clusters based on the total number of the first plurality of clusters;

determine **(806)** a rank for each of the second plurality of clusters based on similarity with at least one cluster from the first plurality of clusters, wherein a cluster from the second plurality of clusters having highest similarity with the at least one cluster is assigned the highest rank and a cluster from the second plurality of clusters having lowest similarity with the at least one cluster is assigned the lowest rank; and

create **(808)** a machine learning model to identify the entity, wherein the machine learning model is created using at least one of the first plurality of clusters weighted based on the rank assigned to each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

Dated this the 31st day of December 2018

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai – 600089

DESCRIPTION

Technical Field

[001] This disclosure relates generally to machine learning models, and more particularly to method and system for creating and training machine learning models.

Background

[002] Digital data (for example, digital images) may be highly inconsistent and may depend on various factors. For digital images, these factors may include, but are not limited to image resolution, noise effect, or font size. Training a machine learning model to identify entities from digital data with high level of accuracy (for example, by performing perform Optical Character Recognition (OCR)) is challenging. Once the machine learning model has been trained to identify entities from a set of pre-defined data (for example, images, symbols, numbers or text), it is difficult to apply the machine learning model to identify entities from a new set of data. The new set of data may be similar to the previous set, but may have new variations that the machine learning model may not recognize.

[003] Some conventional machine learning models that are trained to identify entities from digital data (for example, by using OCR), are trained on similar set of data and are therefore not able to accurately identify data with good accuracy. Some other conventional machine learning models require adding multiple variants of data in a database in order to be trained for entity identification. However, such machine learning models need to be continuously updated in order to handle new variants of data. Pushing all such variants of data (for example, multiple variations of character images) to the database may overload it. Moreover, manually selecting and adding variants of data to the database may require people expertise and is also a time consuming process.

SUMMARY

[004] In one embodiment, a method for training a machine learning model for identifying entities from data is disclosed. The method may include creating a first plurality of clusters from a first plurality of data samples in a first dataset, based on a first set of entity attributes associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset, based on a second set of entity attributes associated with the second plurality of data samples. Each of the first dataset and the second dataset are used to train a machine learning model to identify an entity from data, and wherein the first plurality of data samples

comprise the entity and the second plurality of data samples do not comprise the entity. The method further includes determining a rank for each of the first plurality of clusters based on a probability of identification, as determined by the machine learning model, of an associated set of data samples within the first plurality of data samples as the entity. The method includes determining a rank for each of the second plurality of clusters based on a probability of mismatch, as determined by the machine learning model, of an associated set of data samples within the second plurality of samples, with the entity. The method further includes retraining the machine learning model using at least one of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

[005] In another embodiment, a method creating a machine learning model for identifying entities from data is disclosed. The methods may include creating a first plurality of clusters from a first plurality of data samples in a first dataset, based on a first set of entity attributes associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset, based on a second set of entity attributes associated with the second plurality of data samples. The first plurality of data samples comprise an entity and the second plurality of data samples do not comprise the entity. The method further includes assigning the same rank to each of the first plurality of clusters based on the total number of the first plurality of clusters. The method includes determining a rank for each of the second plurality of clusters based on similarity with at least one cluster from the first plurality of clusters, wherein a cluster from the second plurality of clusters having highest similarity with the at least one cluster is assigned the lowest rank and a cluster from the second plurality of clusters having lowest similarity with the at least one cluster is assigned the highest rank. The method further includes creating a machine learning model to identify the entity, wherein the machine learning model is created using at least one of the first plurality of clusters weighted based on the rank assigned to each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

[006] In another embodiment, an entity identification device for creating a machine learning model for identifying entities from data is disclosed. The entity identification device includes a processor and a memory communicatively coupled to the processor, wherein the memory stores processor instructions, which, on execution, causes the processor to create a first plurality of clusters from a first plurality of data samples in a first dataset, based on a first set of entity attributes

associated with the first plurality of data samples and a second plurality of clusters from a second plurality of data samples in a second dataset, based on a second set of entity attributes associated with the second plurality of data samples, wherein the first plurality of data samples comprise an entity and the second plurality of data samples do not comprise the entity; assign the same rank to each of the first plurality of clusters based on the total number of the first plurality of clusters; determine a rank for each of the second plurality of clusters based on similarity with at least one cluster from the first plurality of clusters, wherein a cluster from the second plurality of clusters having highest similarity with the at least one cluster is assigned the highest rank and a cluster from the second plurality of clusters having lowest similarity with the at least one cluster is assigned the lowest rank; and create a machine learning model to identify the entity, wherein the machine learning model is created using at least one of the first plurality of clusters weighted based on the rank assigned to each of the first plurality of clusters and at least one of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters.

[007] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[008] The accompanying drawings, which are incorporated in and constitute a part of this disclosure, illustrate exemplary embodiments and, together with the description, serve to explain the disclosed principles.

[009] **FIG. 1** illustrates a system for creating and training machine learning models, in accordance with an embodiment.

[010] **FIG. 2** is a functional block diagram of various modules within a memory of an entity identification device configured to create and train machine learning models, in accordance with an embodiment.

[011] **FIG. 3** illustrates a flowchart of a method for training a machine learning model to identify an entity from data, in accordance with an embodiment.

[012] **FIG. 4** illustrates a flowchart of a method for determining accuracy of a retrained machine learning model and accordingly retaining or retraining the retrained machine learning model, in accordance with an embodiment.

[013] **FIG. 5** illustrates a flowchart of a method for determining ranks for a first plurality of clusters created from a first dataset, in accordance with an embodiment.

[014] **FIG. 6** illustrates a flowchart of a method for determining ranks for a second plurality of clusters created from a second dataset, in accordance with an embodiment.

[015] **FIG. 7** illustrates a flowchart of a method for creating a first dataset that includes a first plurality of data samples and a second dataset that includes a second plurality of data samples, in accordance with an embodiment.

[016] **FIG. 8** illustrates a flowchart of a method for creating a machine learning model to identify an entity from data, in accordance with an embodiment.

[017] **FIG. 9** illustrates a flowchart of a method for determining accuracy of a machine learning model and accordingly retaining or retraining the machine learning model, in accordance with an embodiment.

[018] **FIG. 10** illustrates a flowchart of a method for retraining a machine learning model, in accordance with an embodiment.

DETAILED DESCRIPTION

[019] Exemplary embodiments are described with reference to the accompanying drawings. Wherever convenient, the same reference numbers are used throughout the drawings to refer to the same or like parts. While examples and features of disclosed principles are described herein, modifications, adaptations, and other implementations are possible without departing from the spirit and scope of the disclosed embodiments. It is intended that the following detailed description be considered as exemplary only, with the true scope and spirit being indicated by the following claims. Additional illustrative embodiments are listed below.

[020] In one embodiment, a system 100 for creating and training machine learning models is illustrated in the FIG. 1, in accordance with an embodiment. The system 100 may include an entity identification device 102 and a plurality of computing devices 104. The entity identification device 102 may be a computing device having data processing capability. Examples of the entity identification device 102 may include, but are not limited to a server, a desktop, a laptop, a notebook, a netbook, a tablet, a smartphone, a mobile phone, an application server, a sever, or the like. In particular, the entity identification device 102 may have the capability to create and train machine learning models to identify entities within data. The data, for example, may include, but is not limited to images, text, or sensor data. The entities, for example, may include one or more of, but are not limited to a character, an animal, an object, or a human.

[021] Further, examples of the plurality of computing devices 104 may include, but are not limited to a laptop 104a, a computer 104b, a smart phone 104c, and a server 104d. The entity identification

device 102 is in communication with the plurality of computing devices 104 via a network 106. The network 106 may be a wired or a wireless network and the examples may include, but are not limited to the Internet, Wireless Local Area Network (WLAN), Wi-Fi, Long Term Evolution (LTE), Worldwide Interoperability for Microwave Access (WiMAX), and General Packet Radio Service (GPRS). One or more of the plurality of computing devices 104 may provide data to the entity identification device 102, in order to identify one or more entities from within the data so provided.

[022] In order to create and train a machine learning model to identify entities from data, the entity identification device 102 may include a processor 108 and a memory 110. The memory 110 may store instructions that, when executed by the processor 108, may cause the processor 108 to identify entities from data, as discussed in greater detail in FIG. 2 to FIG. 10. The memory 110 may be a non-volatile memory or a volatile memory. Examples of non-volatile memory, may include, but are not limited to a flash memory, a Read Only Memory (ROM), a Programmable ROM (PROM), Erasable PROM (EPROM), and Electrically EPROM (EEPROM) memory. Examples of volatile memory may include, but are not limited to Dynamic Random Access Memory (DRAM), and Static Random-Access memory (SRAM). The entity identification device 102 may further include one or more input/output devices 112 through which the entity identification device 102 may interact with a user and vice versa. By way of an example, the input/output devices 114 may be used to render identified entities to a user. The input/output devices 112, for example, may include a display.

[023] Referring now to FIG. 2, a functional block diagram of various modules within the memory 100 of the entity identification device 102 configured to create and train machine learning models is illustrated, in accordance with an embodiment. The memory 110 may include a dataset creation module 202, a first dataset 204, a second dataset 206, a model creation module 208, a model training module 210, a clustering module 212, a ranking module 214, and a model accuracy testing module 216.

[024] The dataset creation module 202 creates the first dataset 204 that includes a first plurality of data samples and the second dataset 206 that includes a second plurality of data samples. The first plurality of data samples include an entity. Thus, the first dataset and the first plurality of data samples may correspond to a true set for the machine learning model, such that, the true set represents data that includes the entity. In contrast to the first plurality of data samples, the second plurality of data samples do not include the entity. Thus, the second dataset and the second plurality of data samples may correspond to a false set for the machine learning model, such that, the false

set represents data that includes one or more other entities that are different from the entity. Each of the first dataset and the second dataset are used to train a machine learning model to identify the entity from data fed into the machine learning model. The functionality of the dataset creation module 202 is further explained in detail in conjunction with FIG. 7.

[025] The first and second plurality of data samples may include, but are not limited to images, text, or sensor data. An entity, for example, may include, but is not limited to a character, an animal, an object, or a human. Further, the character, for example, may include but is not limited to a digit, a symbol, or an alphabet.

[026] The clustering module 212 creates a first plurality of clusters from the first plurality of data samples in the first dataset 204 and a second plurality of clusters from the second plurality of data samples in the second dataset 204. The clustering module 212 creates the first plurality of clusters based on a first set of entity attributes associated with the first plurality of data samples. Similarly, the clustering module 212 creates the second plurality of clusters based on a second set of entity attributes associated with the second plurality of data samples. This is further explained in detail in conjunction with FIG. 3.

[027] Once the clustering module 212 creates the first plurality of clusters and the second plurality of clusters, the ranking module 214 determines a rank for each of the first plurality of clusters based on a probability of identification of an associated set of data samples within the first plurality of data samples as the entity. The probability of identification of an associated set of data samples may be determined by the machine learning model. In an similar manner, the ranking module 214 determines a rank for each of the second plurality of clusters based on a probability of mismatch of an associated set of data samples within the second plurality of samples, with the entity. The probability of mismatch may be determined by the machine learning model. This is further explained in detail in conjunction with FIG. 5, FIG. 6, and FIG. 8.

[028] Once the first plurality of clusters and the second plurality of clusters have been ranked, the model training module 210 may train the machine learning model using one or more of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters. Additionally, the model training module 210 may retrain the machine learning model based on one or more of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters. This is further explained in detail in conjunction with FIG. 3.

[029] In an embodiment, the model creation module 208 may create a machine learning model to identify the entity using one or more of the first plurality of clusters weighted based on the rank

assigned to each of the first plurality of clusters. The model creation module 208 may additionally create the machine learning model based on one or more of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters. This is further explained in detail in conjunction with FIG. 8.

[030] Once the machine learning model has been retrained, the model accuracy testing module 216 tests accuracy of the retrained machine learning model in identifying the entity from each of the first plurality of data samples and discarding each of the second plurality of data samples. This is further explained in detail in conjunction with FIG. 4. Additionally, once the machine learning model has been created, the model accuracy testing module 216 may test accuracy of the newly created machine learning model in identifying the entity from each of the first plurality of data samples and discarding each of the second plurality of data samples. This is further explained in detail in conjunction with FIG. 9.

[031] Referring now to FIG. 3, a flowchart of a method for training a machine learning model to identify an entity from data is illustrated, in accordance with an embodiment. At step 302, a first plurality of clusters are created from a first plurality of data samples in a first dataset (for example, the first dataset 204). The first plurality of data samples include the entity. Thus, the first dataset and the first plurality of data samples may correspond to a true set for the machine learning model, such that, the true set represents data that includes the entity. Additionally, at step 302, a second plurality of clusters are created from a second plurality of data samples in a second dataset (for example, the second dataset 204). In contrast to the first plurality of data samples, the second plurality of data samples do not include the entity. Thus, the second dataset and the second plurality of data samples may correspond to a false set for the machine learning model, such that, the false set represents data that includes one or more other entities that are different from the entity.

[032] Each of the first and second plurality of data samples may include, but are not limited to images, text, or sensor data. An entity, for example, may include, but is not limited to a character, an animal, an object, or a human. The character, for example, may include but is not limited to a digit, a symbol, or an alphabet. Each of the first and second plurality of data samples in their respective datasets may be labelled. When data samples are image samples that include characters, each image sample may indicate the character that it refers to.

[033] By way of an example, with regards to the first plurality of data samples being the true set and the second plurality of data samples being the false set, the data may be an image and the first plurality of data samples may include images of a dog (true set), while the second plurality of data

samples may include images of a cat, a wolf, or a fox (false set). Each such image may be labelled based on the animal included therein. By way of another example, the data may be an image and the first plurality of data samples may include images of the digit (character) "3" (true set), while the second plurality of data samples may include images of all other digits (characters) other than "3" (false set). Each such image may be labelled based on the character included therein.

[034] Each of the first dataset and the second dataset are used to train a machine learning model to identify the entity from data fed into the machine learning model. It will be apparent to a person skilled in the art that multiple such machine learning models may be trained, such that, each machine learning model is trained for a given entity. By way of an example, a machine learning model may be trained to identify "dogs," while another machine learning model may be trained to identify "cats."

[035] The machine learning model may be trained on the first dataset and the second dataset to provide a probability of occurrence of the in a data sample. The machine learning model may output a probability close to 1, when the entity is identified and 0 if the identity is not identified. In an embodiment, when the first dataset and the second dataset include image samples. The data within image samples may be obtained by performing Optical Character Recognition (OCR) on the image samples in order to identify individual characters from the image samples. Each of the individual characters may later be recognized using an associated machine learning model trained for the same. A machine learning model trained to identify a character, may output a probability close to 1, when the character is identified and 0 if the character is not identified.

[036] The first plurality of clusters are created based on a first set of entity attributes associated with the first plurality of data samples. Similarly, the second plurality of clusters are created based on a second set of entity attributes associated with the second plurality of data samples. Each of the first and second set of entity attributes include one or more features descriptive of the entity. In an embodiment, when the entity is a character, the first and second set of entity attributes may include one or more of, but is not limited to a size of the character, a font of the character, a style associated with the character, a thickness of the character, or a color of the character. In an embodiment, when the first and second plurality of data samples are images, entity attributes (the first and the second set) may be extracted from each image using image processing and feature extracting algorithms. Examples of such algorithm may include, but are not limited to SIFT™ or SURF™.

[037] The first plurality of clusters may be created based on a first set of entity attributes associated with the first plurality of data samples. One cluster in the first plurality of clusters may include data

samples that have similar attributes. By way of an example, when data samples includes images of digits, each cluster may include images that are similar to each other with respect to entity attributes that may include one or more of, but is not limited to size, font, style, or thickness of the digit. In a similar manner, the second plurality of clusters are created based on the second set of entity attributes associated with the second plurality of data samples.

[038] At step 304, a rank is determined for each of the first plurality of clusters based on a probability of identification of an associated set of data samples within the first plurality of data samples, as the entity. The probability of identification of an associated set of data samples may be determined by the machine learning model. In other words, for a given cluster, the machine learning model determines a probability of identification of each data sample within the given cluster. Thereafter, based on the determined probability of identification of each data sample, the rank for the given cluster is determined. The first plurality of clusters are generated from the true set (the first plurality of data samples). Thus, when the probability of identification of data samples within a cluster is farther from 1, the cluster is assigned a higher rank so that the machine learning algorithm may be retrained for that cluster. Such retraining would increase the accuracy of the machine learning model in identifying the entity from the data samples within the cluster in future. This is further explained in detail in conjunction with FIG. 5.

[039] In a similar manner, at step 306, a rank is determined for each of the second plurality of clusters based on a probability of mismatch of an associated set of data samples within the second plurality of samples, with the entity. The probability of mismatch may be determined by the machine learning model. The second plurality of clusters are generated from the false set (the second plurality of data samples). Thus, when the probability of mismatch of data samples within a cluster is closer to 0, this would mean that the machine learning model is misidentifying the data samples within the cluster. In an alternate embodiment, when the probability of identification of data samples within a cluster is closer to 1, this would mean that the machine learning model is misidentifying the data samples within the cluster. Thus, in both the cases, the cluster is assigned a higher rank so that the machine learning algorithm may be retrained for that cluster. Such retraining would increase the accuracy of the machine learning model in identifying that the data samples within the cluster do not include the entity. This is further explained in detail in conjunction with FIG. 6.

[040] At step 308, the machine learning model may be retrained using one or more of the first plurality of clusters weighted based on the rank determined for each of the first plurality of clusters

(which are the cluster for the true set). In an embodiment, for the first plurality of clusters, one or more clusters are selected, such that, ranks assigned to each of the one or more clusters are high. For these one or more clusters, the probability of identification determined for a set of data samples within the one or more clusters may be farther from 1. Weights may be assigned to the one or more clusters in descending order of assigned ranks, such that, highest weight may be assigned to a cluster that has the lowest rank amongst the one or more clusters. By way of an example, if there are three clusters, i.e., a cluster 1 with rank 3, a cluster 2 with rank 2, and a cluster 3 with rank 1, the cluster 1 may be assigned the highest weight, while the cluster 1 may be assigned the lowest weight.

[041] At step 308, the machine learning model may additionally be retrained based on one or more of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters (which are the cluster for the false set). In an embodiment, for the second plurality of clusters, one or more clusters are selected, such that, ranks assigned to each of the one or more clusters are high. For these one or more clusters, the probability of mismatch determined for a set of data samples within the one or more clusters may be closer to 0. Weights may be assigned to one or more clusters in ascending order of assigned ranks, such that, highest weight may be assigned to a cluster that has the highest rank. By way of an example, if there are three clusters, i.e., a cluster 1 with rank 3, a cluster 2 with rank 2, and a cluster 3 with rank 1, the cluster 3 may be assigned the highest weight, while the cluster 3 may be assigned the lowest weight.

[042] n order to retrain the machine learning model, only a subset of data samples may be used from the one or more clusters identified at step 308. In an embodiment, the quantum of data samples to be extracted from a cluster may be determined based on the weight assigned to the cluster. In other words, for a cluster that is assigned a higher weight, more data samples from the cluster may be used to retrain the machine learning model. In contrast, for a cluster that is assigned a lower weight, less data samples from the cluster may be used to retrain the machine learning model. At step 310, accuracy of the retrained machine learning model in identifying the entity from each of the first plurality of data samples and discarding each of the second plurality of data samples is tested. This is further explained in detail in conjunction with FIG. 4.

[043] Referring now to FIG. 4, a flowchart of a method for testing accuracy of a retrained machine learning model and accordingly retaining or retraining the retrained machine learning model is illustrated, in accordance with an embodiment. Once the machine learning mode has been trained, at step 402, a check is performed to determine if the accuracy of the retained machine learning model is greater than a predefine accuracy threshold. The predefined accuracy threshold may be the

accuracy of the machine learning model, prior to retraining, in identifying the entity from each of the first plurality of data samples and discarding each of the second plurality of data samples. Thus, the accuracy of the retrained machine learning model may be compared with the accuracy of the machine learning model.

[044] Referring back to step 402, when the accuracy of the retrained machine learning model is greater than the predefined accuracy threshold, the retrained machine learning model is retained for further use at step 404. In this case, the machine learning model, prior to retraining, may be discarded. However, when the accuracy of the retrained machine learning model is less than or equal to the predefined accuracy threshold, the retrained machine learning model is retrained at step 406. To this end, steps 302 to 310 may be performed for the retrained machine learning model.

[045] Referring now to FIG. 5, a flowchart of a method for determining rank for a cluster from the first plurality of clusters is illustrated, in accordance with an embodiment. At step 502, the machine learning model determines a probability of data samples within the cluster. The probability of data samples within the cluster corresponds to the probability of identification of the entity from the data samples by the machine learning model. At step 504, the machine learning model determines the average probability of data samples within each of the first plurality of clusters.

[046] At step 506, a summation of the average probability of each of the first plurality of clusters is determined based on the average probability of the data samples within each of the first plurality of clusters. At step 508, a score is computed for the cluster based on the division of the average probability of the cluster by the summation of the average probability of each of the first plurality of clusters. Thereafter, at step 510, a rank of the cluster is computed as subtraction of the score of the cluster from one.

[047] By way of an example, there may be a total of three clusters, i.e., a cluster 1 with an average probability of 0.7, a cluster 2 with an average probability of 0.9, and a cluster 3 with an average probability of 0.8 respectively. A rank for each of the cluster 1, the cluster 2, and the cluster 3, may be determined using the equation 1 given below:

Cluster Rank = 1 – (Average probability of a Cluster/$\sum$ (Average probability of each of the first plurality of clusters)) ... (1)

[048] Thus, based on the equation 1 above, the cluster 1 has the highest rank, followed by the cluster 3 and the cluster 2. In other words, the cluster with the lowest average probability is assigned the highest rank, while the cluster with the highest average probability is assigned the lowest rank.

[049] Referring now to FIG. 6, a flowchart of a method for determining a rank for a cluster from the second plurality of clusters is illustrated, in accordance with an embodiment. At step 602, the machine learning model determined a probability of data samples within the cluster. The probability of data samples within the cluster corresponds to the probability of mismatch of the entity from the data samples. At step 604, the machine learning model determines the average probability of the data samples within each of the second plurality of clusters.

[050] At step 606, a summation of the average probability of each of the second plurality of clusters is determined based on the average probability of the data samples within each of the second plurality of clusters. At step 608, a score is computed for the cluster based on the division of the average probability of the cluster by the summation of the average probability of each of the second plurality of clusters. At step 610, a rank of the cluster is computed as subtraction of the score of the cluster from one.

[051] By way of an example, there may be a total of three clusters, i.e., a cluster 1 with an average probability of 0.2, a cluster 2 with an average probability of 0.5, and a cluster 3 with an average probability of 0.1 respectively. A rank for each of the cluster 1, the cluster 2, and the cluster 3, may be determined using the equation 2 given below:

Cluster Rank = 1 – (Average probability of a Cluster/$\sum$ (Average probability of each of the second plurality of clusters)) ... (2)

[052] Thus, based on the equation 2 above, the cluster 3 has the highest rank, followed by the cluster 1, and the cluster 2. In other words, the cluster with the lowest average probability is assigned the highest rank, while the cluster with the highest average probability is assigned the lowest rank.

[053] Referring now to FIG. 7, a flowchart of a method 702 for creating a first dataset that includes a first plurality of data samples and a second dataset that includes a second plurality of data samples is illustrated, in accordance with an embodiment. At step 704, a machine learning model receives a first data sample that includes the entity and a second data sample that does not include the entity. The first and second data sample, for example, may be images, such that the first data sample is an image that includes a dog, while the second data sample is an image that includes a cat. Additionally, the machine learning model may be trained to identify dogs from within images.

[054] At step 706, the machine learning model identifies a first set of results that match with the first data sample and a second set of results that match with the second data sample. After step 706, step 708 and step 716 may be executed parallelly.

[055] At step 708, the accuracy in identification of each result in a subset of the first set of results, by the machine learning model, is determined. In other words, a subset may be first selected from the first set of results. Thereafter, accuracy in identification of each result in the subset is determined. The subset may be selected based on a predefined criterion. The predefined criterion, for example, may be selecting top N results. By way of an example, top 3 results may be selected and accuracy in identification of each of these 3 results, by the machine learning model, may be determined. In an embodiment, the accuracy in identification is based on a user input corresponding to the first data sample. Each result in the subset (or the top N results) may be presented to the user via a user interface. The user may either select one of the result in the subset or may provide a correct result, that is not included in the subset.

[056] Thereafter, at step 710, a check is performed to determine whether at least one result in the subset matches with the user input. If each result in the subset does not match with the user input, the first data sample is added to the first dataset at step 712. However, if at least one result in the subset matches with the user input, no further action is taken at step 714. By way of an example, when an image of a dog is provided to the machine learning model, the output may be: Dog, Wolf, or Fox. In this case, the user may validate the accuracy in identification by the machine learning algorithm. However, when the output may be: Cat, Wolf, or Fox, the user may provide "Dog" as the correct option and may ignore each of the result provided as the output. Thus, the user validates inaccuracy in the machine learning model to identify "Dog" from the provided image. In this case, the image of dog is added to the first dataset (true set) in order to retrain the machine learning algorithm using this image.

[057] At step 716, an error in misidentification of each of a subset of the second set of results is determined. The subset is selected based on the predefined criterion. The error is determined based on a user input corresponding to the second data sample. At step 718, a check is performed to determine whether at least one result in the subset matches with the user input. If each result in the subset does not match with the user input, the second data sample is added to the second dataset at step 720. However, if at least one result in the subset matches with the user input, control goes to step 714, where no further action is taken. By way of an example, when an image of a cat is provided to the machine learning model, the output may be: Dog, Cheetah, or Fox. The user may provide "Cat" as the correct option and may ignore each of the result provided as the output. Thus, the user validates error in misidentification of "Cat" as "Dog," by the machine learning model. In this case, the image of cat is added to the second dataset (false set) in order to retrain the machine

learning algorithm using this image. However, when the output may be: Cat, Wolf, or Fox, the user may validate that there is no error in misidentification by the machine learning algorithm, for the given image.

[058] Referring now to FIG. 8, a flowchart of a method for creating a machine learning model is illustrated, in accordance with an embodiment. At step 802, based on a first set of entity attributes associated with the first plurality of data samples, a first plurality of clusters are created from the first plurality of data samples in the first dataset. The first plurality of data samples include an entity. Additionally, at step 802, based on a second set of entity attributes associated with the second plurality of data samples, a second plurality of clusters are created from the second plurality of data samples in the second dataset. In contrast to the first plurality of data samples, the second plurality of data samples do not include the entity. This has been explained in detail in conjunction with FIG. 3.

[059] Based on the total number of the first plurality of clusters, the same rank is assigned to each of the first plurality of clusters at step 804. At step 806, a rank is determined for each of the second plurality of clusters based on similarity with one or more clusters from the first plurality of clusters. A cluster from the second plurality of clusters that has the highest similarity (or closest distance) with the one or more clusters is assigned the highest rank. Similarly, a cluster from the second plurality of clusters that has lowest similarity (or highest distance) with the one or more clusters is assigned the lowest rank.

[060] At step 808, a machine learning model to identify the entity is created using one or more of the first plurality of clusters weighted based on the rank assigned to each of the first plurality of clusters. The machine learning model is additionally created based on one or more of the second plurality of clusters weighted based on the rank determined for each of the second plurality of clusters. This has already been discussed in detail in conjunction with FIG. 3. At step 810, accuracy of the machine learning model in identifying the entity from each of the first plurality of data samples and discarding each of the second plurality of data samples is tested. This is further explained in detail in conjunction with FIG. 9.

[061] FIG. 9 illustrates a flowchart of a method for testing accuracy of a machine learning model and accordingly retaining or retraining the machine learning model, in accordance with an embodiment. Once the machine learning mode has been created, at step 902, a check is performed to determine if the accuracy of the machine learning model is greater than a predefined accuracy threshold. The predefined accuracy threshold may be the accuracy of a similar machine learning

model, in identifying a similar entity from each of a true set data samples and discarding each of the false set of data samples. Thus, the accuracy of the machine learning model may be compared with the accuracy of similar existing machine learning models.

[062] Referring back to step 902, when the accuracy of the machine learning model is greater than the predefined accuracy threshold, the machine learning model is retained for further use at step 904. However, when the accuracy of the machine learning model is less than or equal to the predefined accuracy threshold, the machine learning model is retrained at step 906. This is further explained in detail in conjunction with FIG. 10.

[063] Referring now to FIG. 10, a flowchart of a method for retraining a machine learning model, in accordance with an embodiment. Once the machine learning model has been created as detailed in FIG. 8, at step 1002, a retraining rank is determined for each of the first plurality of clusters based on a probability of identification, as determined by the machine learning model, of an associated set within the first plurality of data samples as the entity. At step 1004, a retraining rank is determined for each of the second plurality of clusters based on a probability of mismatch, as determined by the machine learning model, of an associated set within the second plurality of data samples with the entity. At step 1006, the machine learning model is retrained using one or more of the first plurality of clusters weighted based on the retraining rank determined for each of the first plurality of clusters and one or more of the second plurality of clusters weighted based on the retraining rank determined for each of the second plurality of clusters. This has already been explained in detail on conjunction with FIG. 3.

[064] It will be appreciated that, for clarity purposes, the above description has described embodiments of the invention with reference to different functional units and processors. However, it will be apparent that any suitable distribution of functionality between different functional units, processors or domains may be used without detracting from the invention. For example, functionality illustrated to be performed by separate processors or controllers may be performed by the same processor or controller. Hence, references to specific functional units are only to be seen as references to suitable means for providing the described functionality, rather than indicative of a strict logical or physical structure or organization.

[065] It is intended that the disclosure and examples be considered as exemplary only, with a true scope and spirit of disclosed embodiments being indicated by the following claims.

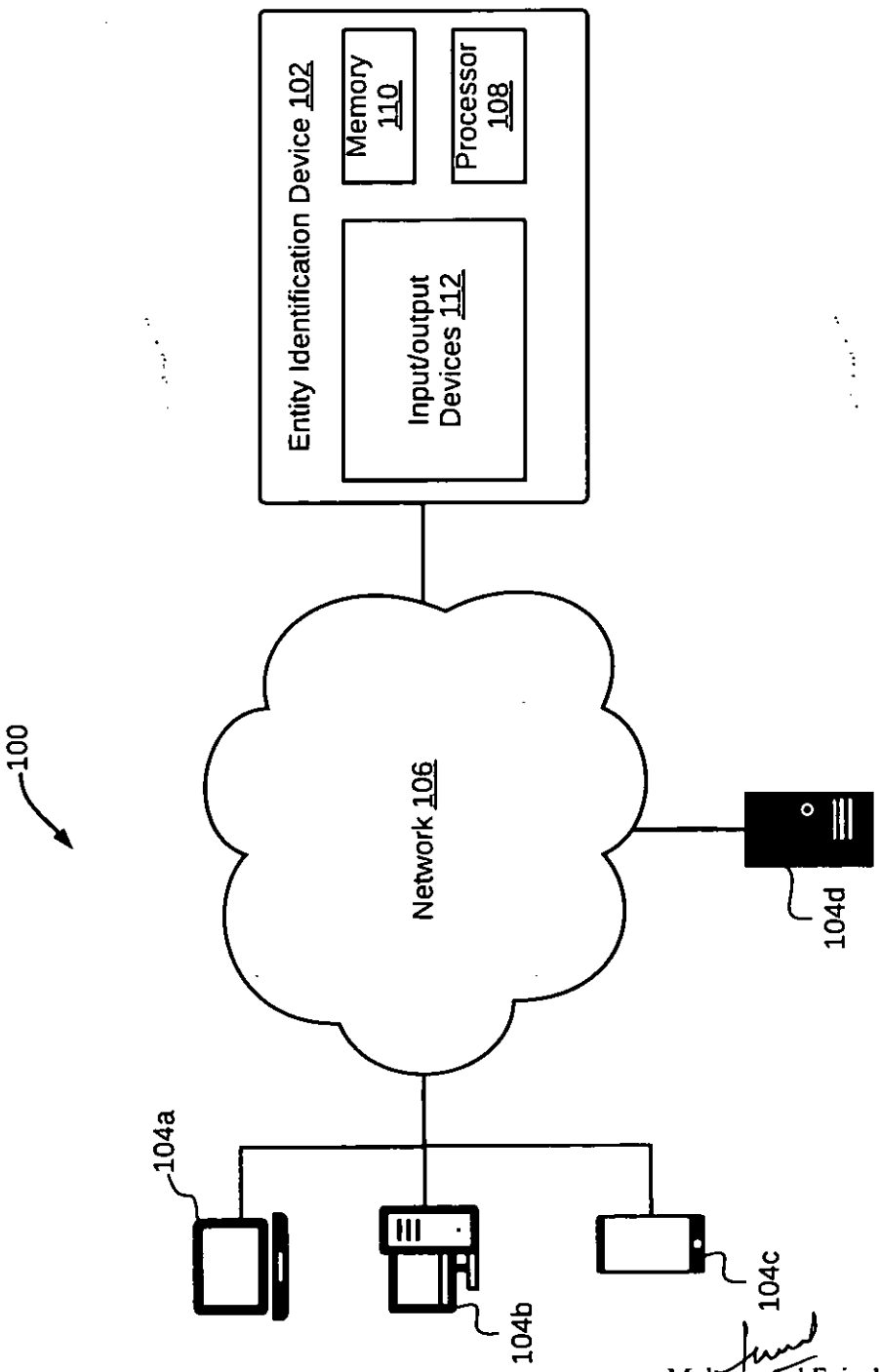

FIG. 1

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

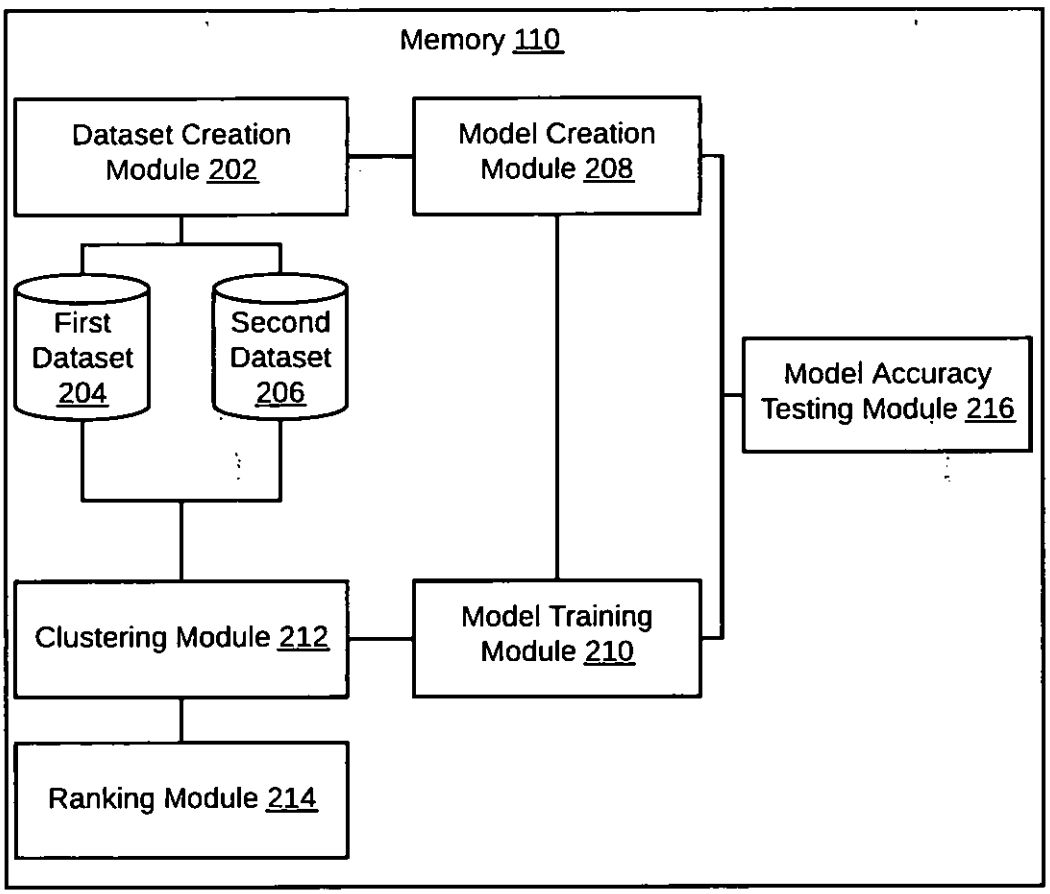

FIG. 2

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

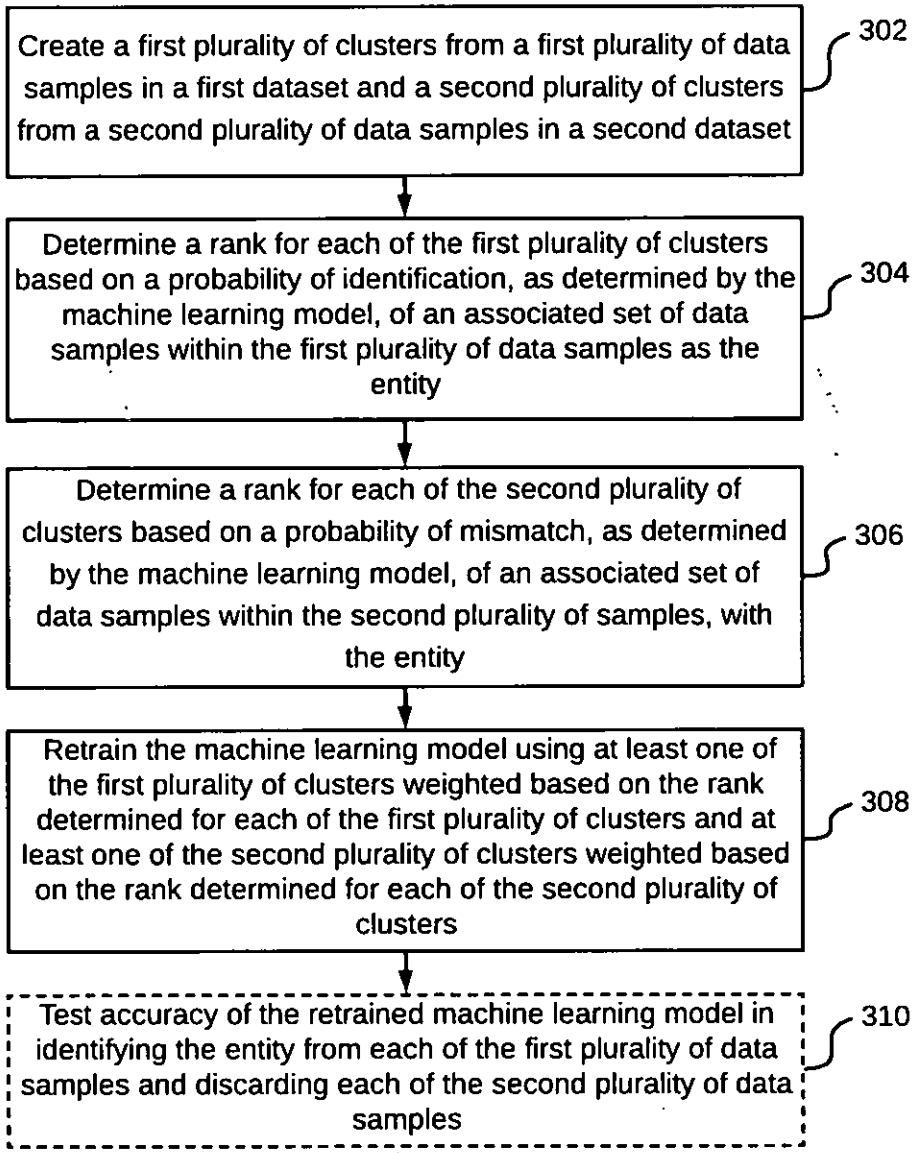

FIG. 3

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

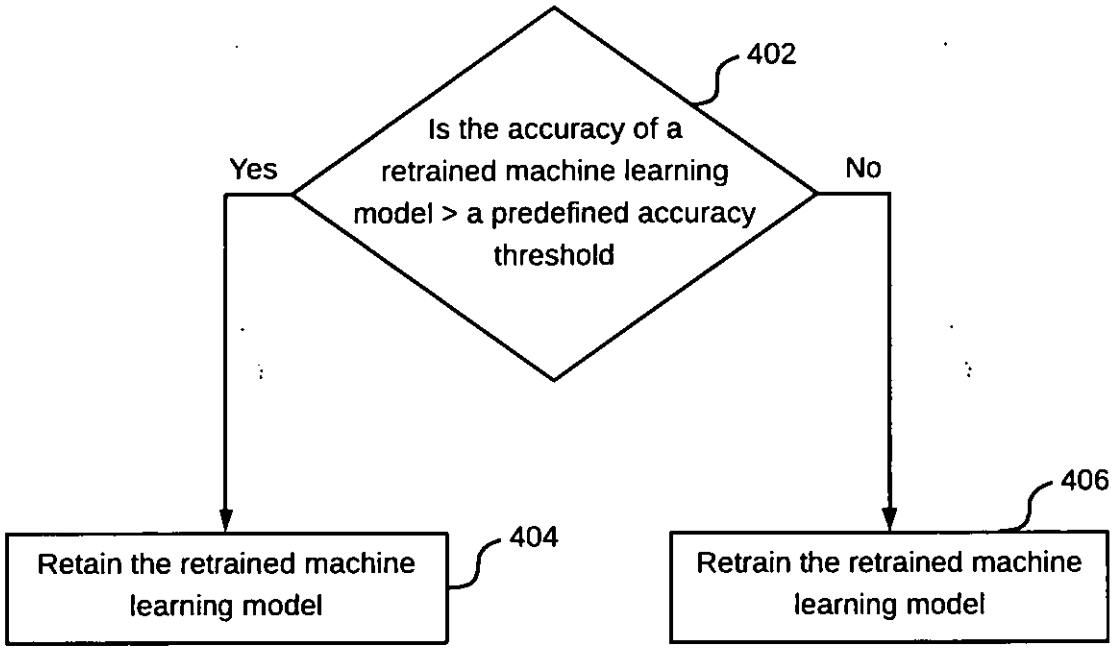

FIG. 4

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

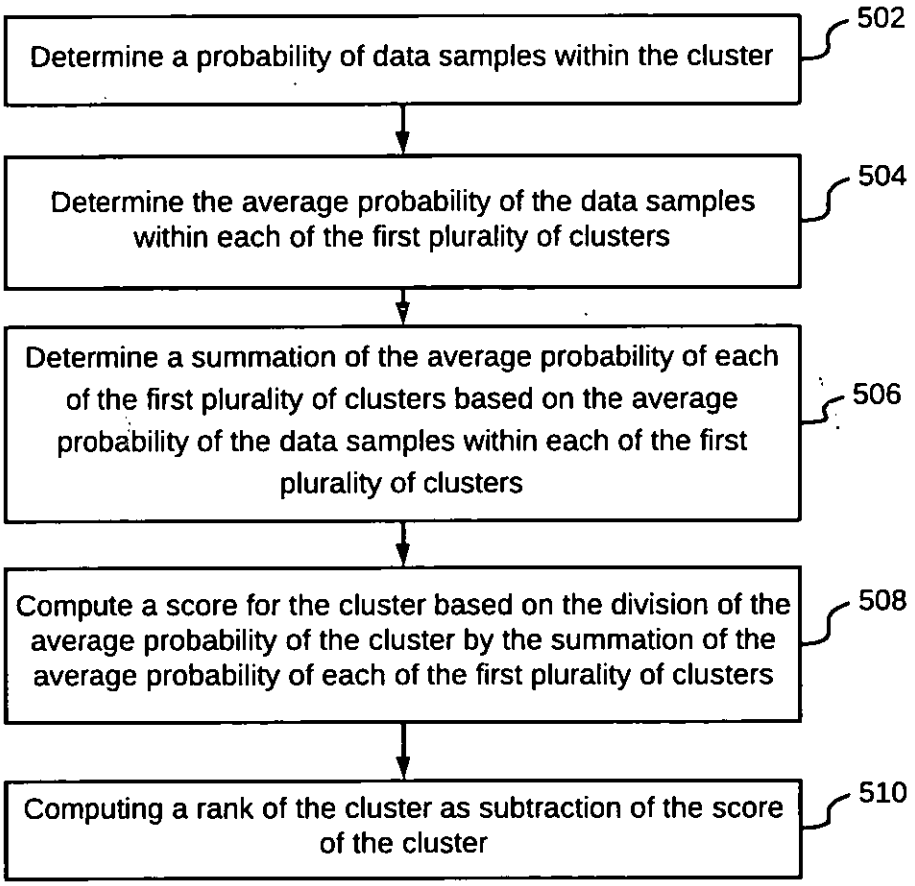

FIG. 5

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

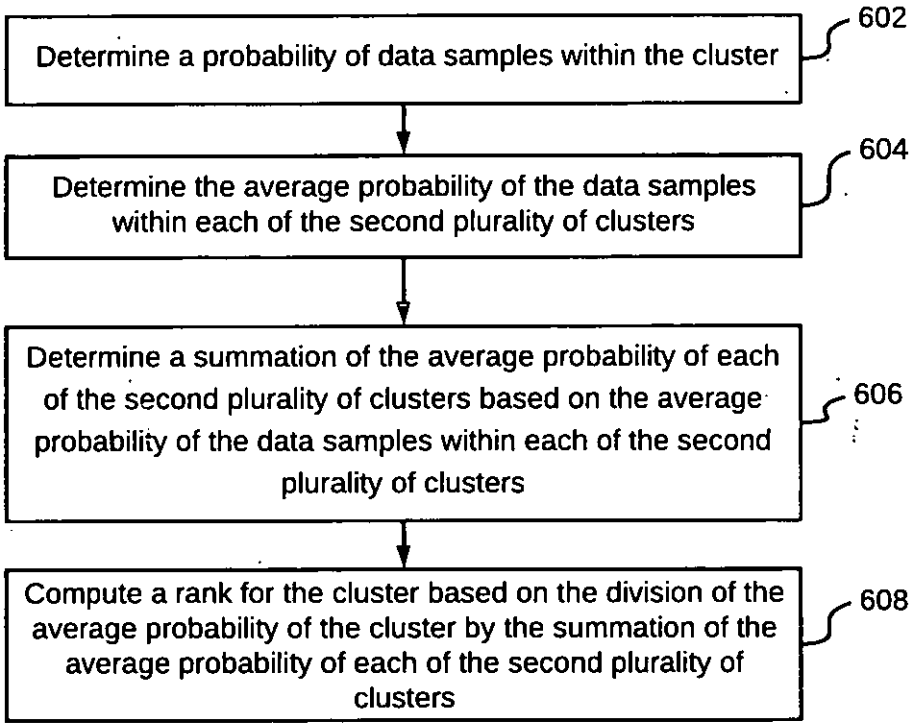

FIG. 6

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

7/11

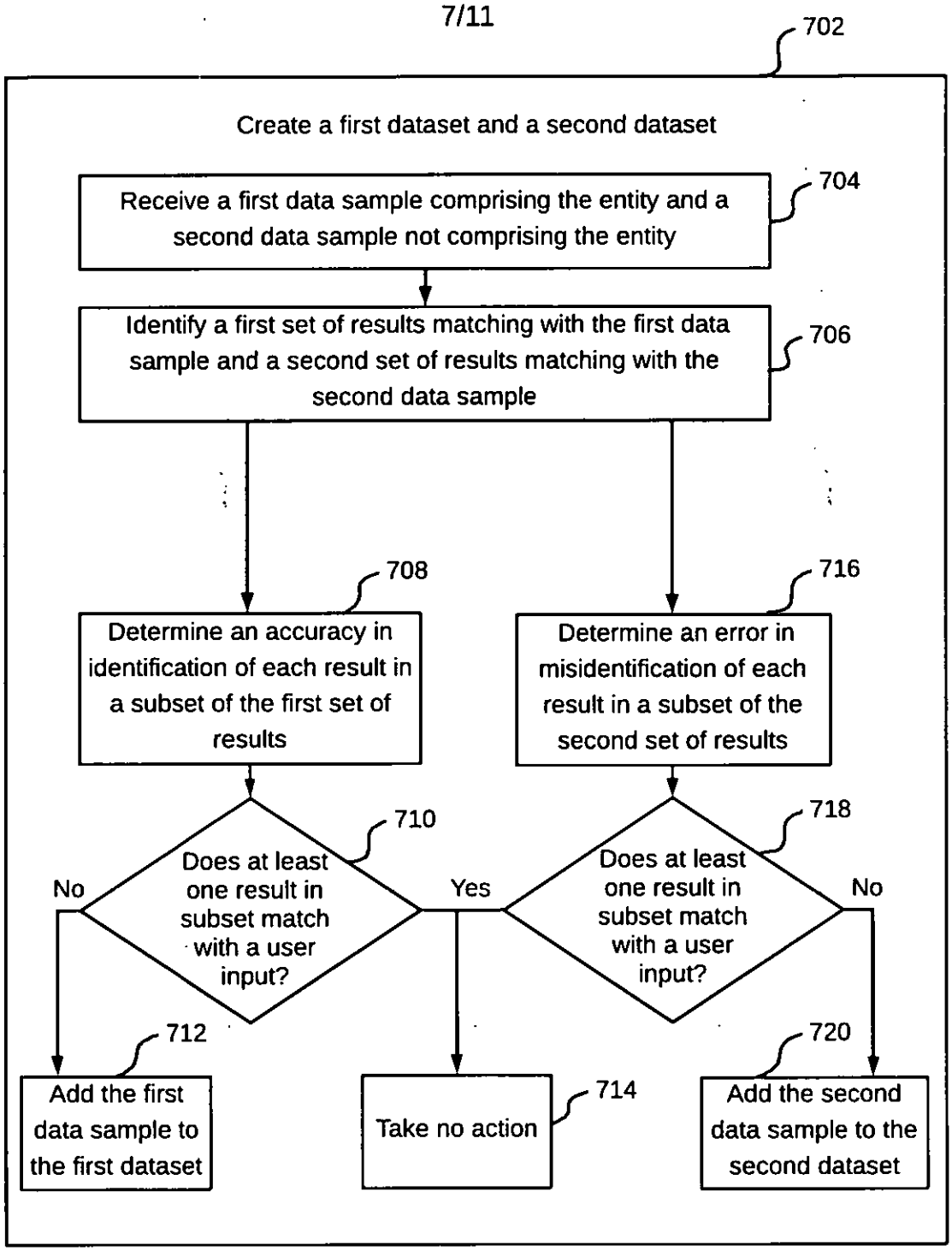

FIG. 7

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

8/11

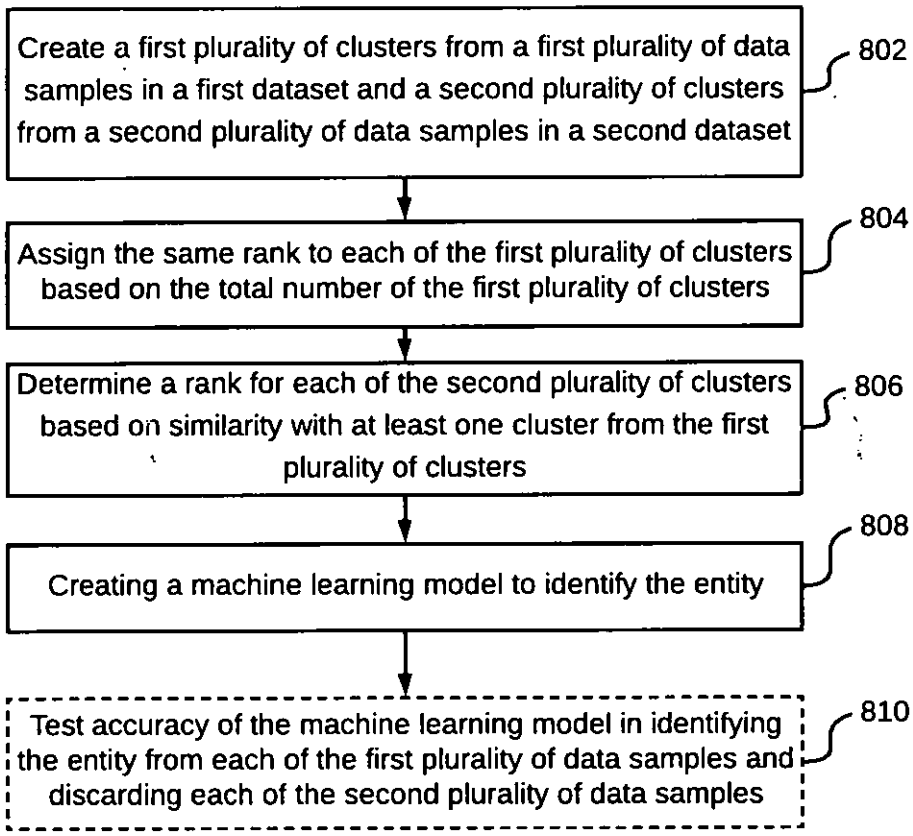

FIG. 8

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

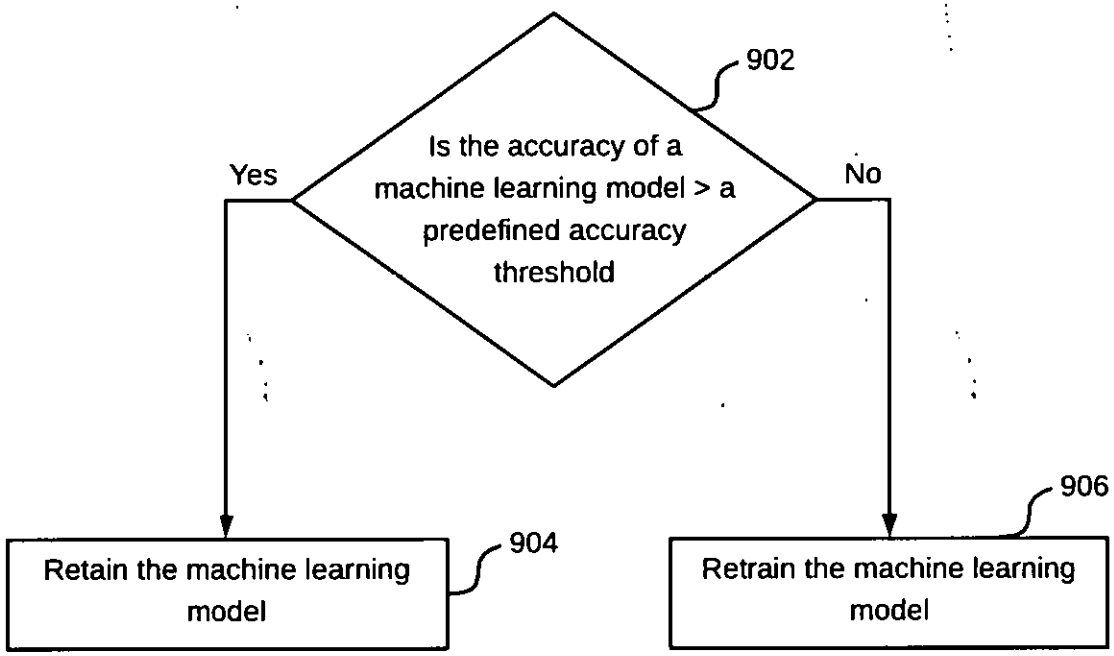

FIG. 9

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

10/11

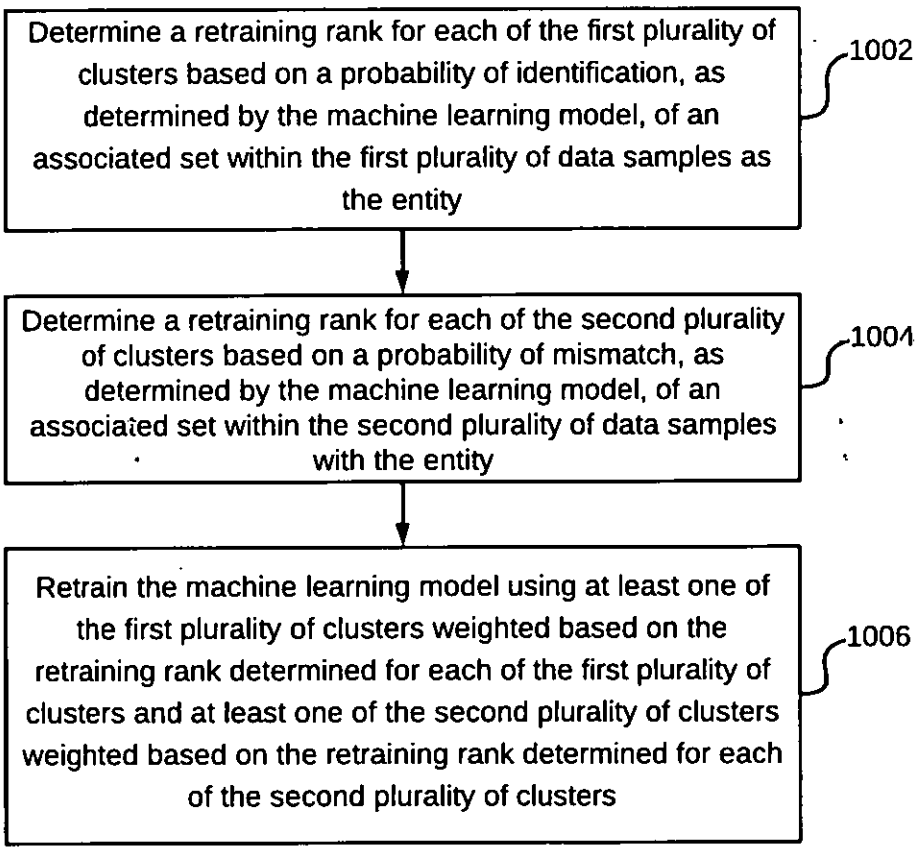

FIG. 10

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai - 600089

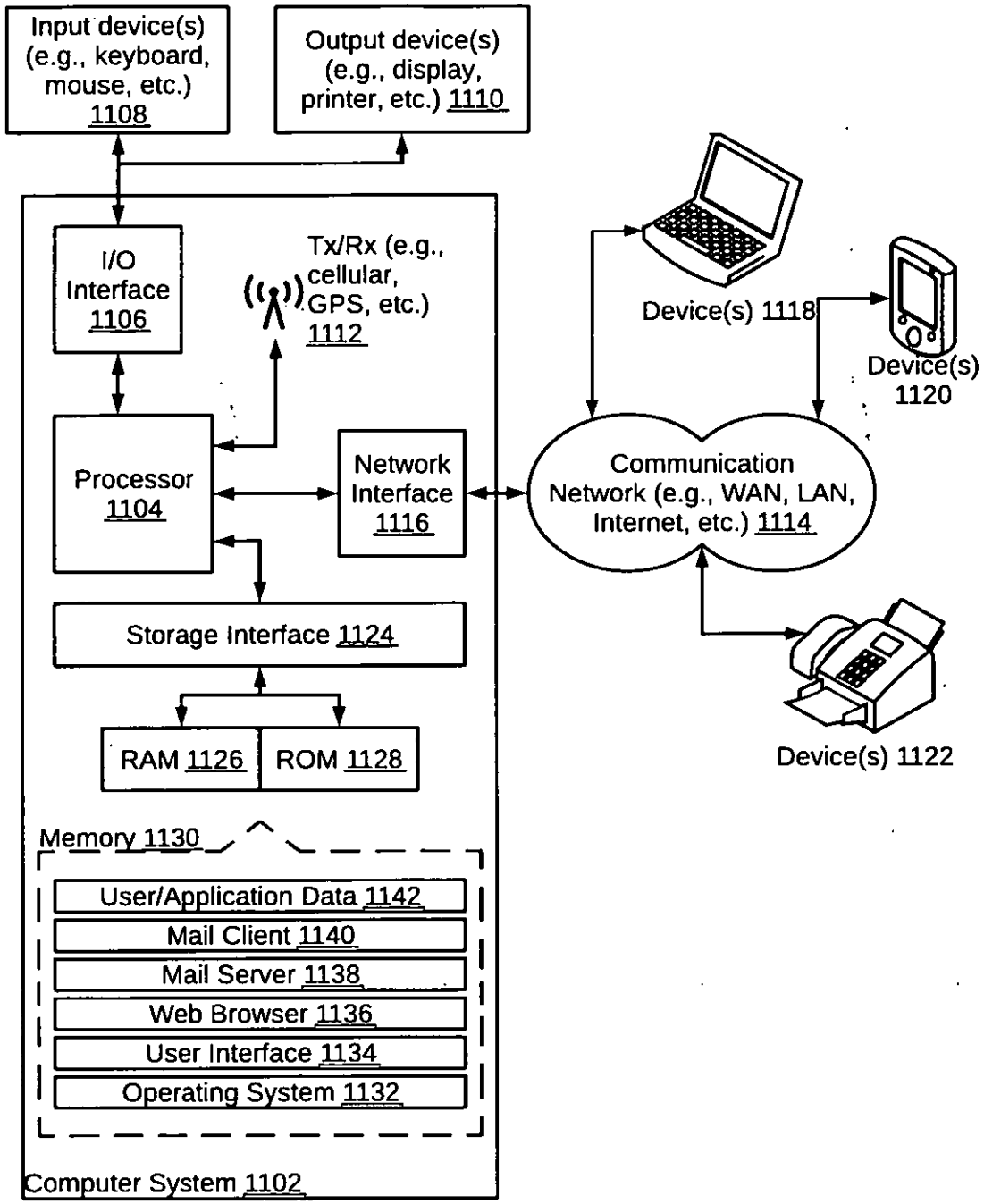

FIG. 11

Mohammed Faisal (INPA No: 1941)
Head, IPR Dept.
L&T Technology Services Limited
DLF 3rd Block, 2nd Floor,
Mānapākkam, Chennai - 600089