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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 December 2022 (22.12.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/265755 A1

(51) International Patent Classification:
G06N 20/00 (2019.01)

(21) International Application Number:
PCT/US2022/028884

(22) International Filing Date:
12 May 2022 (12.05.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
202110670656.4 17 June 2021 (17.06.2021) CN

(71) Applicant: **MICROSOFT TECHNOLOGY LICENSING, LLC** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: **LI, Yanpu**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **CHEN, Yueyang**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **ZHANG, Xucheng**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: **CHATTERJEE, Aaron C.** et al.; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (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, IQ, 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.

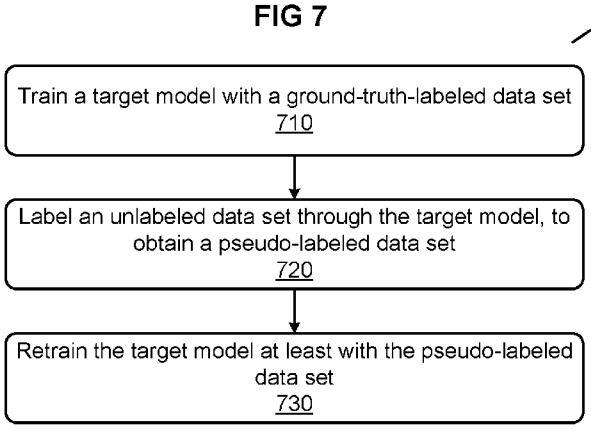
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, 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).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

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

(54) Title: MODEL SELF-TRAINING FOR NATURAL LANGUAGE PROCESSING

FIG 7



(57) Abstract: The present disclosure proposes a method, apparatus and computer program product for model self-training. A target model may be trained with a ground-truth-labeled data set. An unlabeled data set may be labeled through the target model, to obtain a pseudo-labeled data set. The target model may be retrained at least with the pseudo-labeled data set.

WO 2022/265755 A1

MODEL SELF-TRAINING FOR NATURAL LANGUAGE PROCESSING

BACKGROUND

Natural Language Processing (NLP) is a technology that uses natural language to communicate with computers, which aims to enable the computers to understand and use the natural language to achieve communications between humans and computers, thereby replacing the humans to perform various tasks related to the natural language, e.g., a classification task, a Question Answering task, a Machine Reading Comprehension task, etc. For a specific NLP task, a machine learning model may be trained with training data corresponding to the NLP task. The trained machine learning model may be deployed to perform the NLP task.

SUMMARY

This Summary is provided to introduce a selection of concepts that are further described below in the Detailed Description. It is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

Embodiments of the present disclosure propose a method, apparatus and computer program product for model self-training. A target model may be trained with a ground-truth-labeled data set. An unlabeled data set may be labeled through the target model, to obtain a pseudo-labeled data set. The target model may be retrained at least with the pseudo-labeled data set.

It should be noted that the above one or more aspects comprise the features hereinafter fully described and particularly pointed out in the claims. The following description and the drawings set forth in detail certain illustrative features of the one or more aspects. These features are only indicative of the various ways in which the principles of various aspects may be employed, and this disclosure is intended to include all such aspects and their equivalents.

BRIEF DESCRIPTION OF THE DRAWINGS

The disclosed aspects will hereinafter be described in connection with the appended drawings that are provided to illustrate and not to limit the disclosed aspects.

FIG.1 illustrates an exemplary process for model self-training according to an embodiment of the present disclosure.

FIG. 2 illustrates an exemplary process for selecting a pseudo-labeled data subset from a pseudo-labeled data set according to an embodiment of the present disclosure.

FIG.3 illustrates another exemplary process for model self-training including iterative operations according to an embodiment of the present disclosure.

FIG.4 illustrates an exemplary process for model self-training including iterative operations and a self-ensemble strategy according to an embodiment of the present disclosure.

FIG.5 illustrates an example of a process for model self-training including iterative operations and a self-ensemble strategy.

FIG.6 illustrates an exemplary process for performing a predetermined task through a retrained target model according to an embodiment of the present disclosure.

FIG.7 is a flowchart of an exemplary method for model self-training according to an embodiment of the present disclosure.

FIG.8 illustrates an exemplary apparatus for model self-training according to an embodiment of the present disclosure.

FIG.9 illustrates an exemplary apparatus for model self-training according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

The present disclosure will now be discussed with reference to several example implementations. It is to be understood that these implementations are discussed only for enabling those skilled in the art to better understand and thus implement the embodiments of the present disclosure, rather than suggesting any limitations on the scope of the present disclosure.

The performance of a machine learning model to perform a NLP task depends on a large amount of reliable training samples. Herein, a training sample may refer to a single training instance included in training data or a training data set. For some NLP tasks, there is little or no reliable training samples, which restricts the performance of machine learning models when performing these NLP tasks. For example, for a specific NLP task, there may only be a small amount of manually labeled training samples. The manually labeled training samples have high accuracy and authenticity, thus they may also be referred to as ground-truth-labeled samples herein. Accordingly, a data set formed by the ground-truth-labeled samples may be referred to as a ground-truth-labeled data set. If a machine learning model is trained with only a small amount of ground-truth-labeled samples, the performance of the trained machine learning model will be poor when it is actually deployed to perform the corresponding task.

There may be a large amount of samples that are not labeled. Herein, a sample that is not labeled may be referred to as an unlabeled sample. Accordingly, a data set formed by unlabeled samples may be referred to as an unlabeled data set. In addition to the ground-truth-labeled data set, a machine learning model may be trained with an unlabeled data set. For example, for a specific NLP task, each sample in the unlabeled data set may be labeled manually to obtain a training sample for the NLP task. However, this method requires a lot of manpower and is extremely time-consuming. In addition, the unlabeled data set may include invisible samples. Herein, an invisible sample may refer to a sample that cannot be viewed or accessed by an operator during operation on it, and it may also be referred to as invisible data. For example, some samples may

contain information related to user identity, user behavior, etc. Such samples are not allowed to be viewed or accessed during being operated, thus they cannot be labeled manually. In the case where the unlabeled data set includes invisible samples, manually labeling the unlabeled data set will cause information in the invisible samples to be leaked.

Embodiments of the present disclosure propose an improved method for training a machine learning model with an unlabeled data set. For example, a machine learning model may be trained through model self-training with an unlabeled data set. Herein, the trained machine learning model may be referred to as a target model. A target model may be trained with a ground-truth-labeled data set. The ground-truth-labeled data set may include a small amount of ground-truth-labeled samples. The target model may then label an unlabeled data set, to obtain a pseudo-labeled data set. The unlabeled data set may include a large amount of unlabeled samples, e.g., a large amount of invisible samples. Herein, a data set obtained through labeling an unannotated data set by a target model may be referred to as a pseudo-labeled data set. Subsequently, the target model may be retrained at least with the pseudo-labeled data set. This approach intends to train the target model with the training data generated through the target model, thus it may be referred to as a self-training approach. The model self-training according to the embodiments of the present disclosure may label, through the target model, a large amount of unlabeled samples, such as a large amount of invisible samples, to generate a large amount of pseudo-labeled samples with only a small amount of ground-truth-labeled samples, thereby the number of training samples may be augmented, which facilitates to train a target model with better performance. The model self-training according to the embodiments of the present disclosure may be widely applied to various NLP tasks, such as a classification task, a Question Answering task, a Machine Reading Comprehension task, etc.

In an aspect, the embodiments of the present disclosure propose to combine corpus obtained through a plurality of approaches or from a plurality of sources into an unlabeled data set for training a target model. The corpus obtained through different approaches or from different sources usually includes samples involving different domains. Herein, a domain of a sample may broadly refer to a field, a region and a type, etc. related to the sample. The corpora of different domains usually have different ways of expression and language features, and therefore have different data distribution and characteristics, etc. Training a target model with an unlabeled data set including a plurality of corpora involving a plurality of domains may guarantee the fairness of the trained target model.

In another aspect, the embodiments of the present disclosure propose to obtain a pseudo-labeled data set for training a target model through labeling an unlabeled data set by the target model. The unlabeled data set may include, e.g., a plurality of invisible samples. The unlabeled data set

may be stored in a specific data center, and operations related to the unlabeled data set may be performed through special tools. In this way, the unlabeled data set may not be viewed or accessed by operators.

In another aspect, the embodiments of the present disclosure propose that when a pseudo-labeled data set is used to retrain a target model, a pseudo-labeled data subset that meets a quality requirement is selected from the pseudo-labeled data set, and the target model is trained with the pseudo-labeled data subset. The pseudo-labeled data subset may be formed, e.g., through selecting, from the pseudo-labeled data set, samples having task prediction values and/or domain prediction values that meet the quality requirement. A task prediction value of a sample may indicate a probability for a result of a text included in the sample for a corresponding task. The higher the task prediction value, the higher the probability that the text is the result. Taking a classification task as an example, a task prediction value of a sample may indicate a probability that a text included in the sample is classified into a corresponding class. The higher the task prediction value, the higher the probability that the text is in this class. A task prediction value of each sample in a pseudo-labeled data set may be predicted through a target model. A pseudo-labeled data subset may be formed through selecting, from the pseudo-labeled data set, samples having task prediction values higher than a predetermined threshold. Retraining the target model with such pseudo-labeled data subset may guarantee the accuracy of the retrained target model. A domain prediction value of a sample may indicate a probability that a text included in the sample is classified into a corresponding domain. The closer the domain prediction value is to an intermediate value, such as "0.5", the more ambiguous the possibility that the text is in this domain. This may indicate that the text has a general expression, rather than an expression related to a specific domain. The domain prediction value may be predicted through a domain classifier. A pseudo-labeled data subset may be formed through selecting, from the pseudo-labeled data set, samples having domain prediction values within a predetermined range. Retraining the target model with such pseudo-labeled data subset may guarantee the fairness of the retrained target model. Preferably, a pseudo-labeled data subset may be formed through selecting, from the pseudo-labeled data set, samples having task prediction values higher than a predetermined threshold and domain prediction values within a predetermined range, thereby guaranteeing the accuracy and the fairness of the retrained target model.

In another aspect, the embodiments of the present disclosure propose to iteratively perform the step of labeling an unlabeled data set to obtain a pseudo-labeled data set and the step of retraining the target model at least with the pseudo-labeled data set, thus the performance of the target model may be continuously improved to obtain a more accurate and robust target model.

In another aspect, the embodiments of the present disclosure propose that in each iteration, a

target model is configured with a comprehensive parameter set. The comprehensive parameter set may be determined through employing a self-ensemble strategy. For example, the comprehensive parameter set may be determined based on a current parameter set of the target model obtained in a current iteration and one or more previous parameter sets of the target model obtained in one or more previous iterations. In each iteration, the target model may label an unlabeled data set. The target model configured through employing the comprehensive parameter set determined through the self-ensemble strategy may be more robust and can more accurately label the unlabeled data set.

FIG.1 illustrates an exemplary process 100 for model self-training according to an embodiment of the present disclosure. In the process 100, a target model may be trained with a ground-truth-labeled data set. The ground-truth-labeled data set may include a small amount of ground-truth-labeled samples. The target model may then label an unlabeled data set, to obtain a pseudo-labeled data set. The unlabeled data set may include a large amount of unlabeled samples, e.g., a large amount of invisible samples. Subsequently, the target model may be retrained at least with the pseudo-labeled data set.

At 102, optionally, an unlabeled data set may be filtered through a predefined dictionary to obtain a filtered unlabeled data set. The predefined dictionary may include, e.g., common words that are allowed to be used, which may be words that can be searched on the Internet. When the unlabeled data set is filtered through the predefined dictionary, a word in the unlabeled data set that does not appear in the predefined dictionary may be masked, e.g., the word is deleted or replaced with a predetermined character, etc. Alternatively, when the unlabeled data set is filtered through the predefined dictionary, a sentence in the unlabeled data set that includes a word that does not appear in the predefined dictionary may be masked, e.g., the sentence is deleted or replaced with a predetermined character set, etc. In this way, words that do not appear in the predefined dictionary are not used in subsequent pretraining steps and training steps, etc., so as to prevent information related to these words from being leaked. The words that do not appear in the predefined dictionary may be terms that have not been widely used, custom words, etc.

At 104, optionally, a target model may be pretrained with the filtered unlabeled data set. The target model may be pretrained through various pretraining tasks. In an implementation, for a specific sample in the filtered unlabeled data set, some tokens in a text of the sample may be masked, and for each token of the masked tokens, the token is predicted with this token's context. Such a pretraining task may be referred to as a Masked Language Model (MLM). In another implementation, a plurality of sentence pairs may be randomly selected from the filtered unlabeled data set, and each sentence pair includes two consecutive sentences. Among the

plurality of sentence pairs, a predetermined number of sentence pairs are retained. For each sentence pair in the remaining sentence pairs, the second sentence in the sentence pair may be replaced with a sentence randomly extracted from the filtered unlabeled data set. It may then be predicted, through the target model, whether the randomly extracted sentence is the next sentence of the first sentence in the sentence pair. Such a pretraining task may be referred to as Next Sentence Prediction (NSP). Pretraining the target model with the filtered unlabeled data set may enable the target model to adapt to a sentence expression, a sentence-to-sentence relationship, or a word-to-word relationship of each sample in the filtered unlabeled data set, therefore, when subsequently labeling the filtered unlabeled data set, such as the labeling at 110, a more accurate label for each sample may be predicted. It should be appreciated that the approaches for pretraining the target model described above are only exemplary, and the target model may also be pretrained through other approaches. In addition, it should be appreciated that without performing the filtering step at 102, the target model may be pretrained directly with the unlabeled data set through the approach described above.

At 106, optionally, data augmentation may be performed on the ground-truth-labeled data set to obtain a data-augmented ground-truth-labeled data set. The ground-truth-labeled data set may be data augmented through various data augmentation methods. In an implementation, the ground-truth-labeled data set may be data-augmented through employing a back translation method. For example, for a specific sample in the ground-truth-labeled data set, an original text in the sample may be first translated into an intermediate text in another language, and then the intermediate text may be reverse-translated back to a translated text in the same language as the original text. The translated text may have the same label as the original text, but the translated text may have a different expression from the original text, and thus different training samples may be generated. In this way, the number of training samples may be effectively increased. In another implementation, the ground-truth-labeled data set may be data-augmented through employing a word replacement method. For example, for a specific sample in the ground-truth-labeled data set, some tokens in a text of the sample may be randomly selected and the selected tokens may be replaced. For example, a predetermined proportion of tokens may be selected from tokens of the text of the sample. In the selected tokens, 80% of the tokens are replaced with special tokens, such as "[PAD]", 10% of the tokens remain unchanged, and 10% of the tokens are replaces with tokens randomly selected from a vocabulary. Training the target model with such samples having replaced tokens may improve the robustness of the target model.

At 108, the target model may be trained with the data-augmented ground-truth-labeled data set. It should be appreciated that without performing the data-augmenting step at 106, the target model may be trained directly with the ground-truth-labeled data set. In addition, it should be

appreciated that having performed the pretraining step at 104, at 108, the pretrained target model may be trained with the data-augmented ground-truth-labeled data set or the ground-truth-labeled data set.

At 110, the filtered unlabeled data set may be labeled through the target model, to obtain a pseudo-labeled data set. The filtered unlabeled data set may include a plurality of unlabeled samples, e.g., a plurality of invisible samples. The target model may label each sample of the plurality of samples to obtain a pseudo-label corresponding to the sample, and obtain a pseudo-labeled sample through adding the obtained pseudo-label to the sample. A plurality of pseudo-labeled samples corresponding to the plurality of unlabeled samples may be combined into the pseudo-labeled data set.

At 112, optionally, a pseudo-labeled data subset that meets a quality requirement may be selected from the pseudo-labeled data set. The selected pseudo-labeled data subset will be used to retrain the target model later. Using the pseudo-labeled data subset that meets the quality requirement may facilitate to guarantee the accuracy and the fairness of the trained target model. An exemplary process for selecting a pseudo-labeled data subset from a pseudo-labeled data set will be illustrated later in conjunction with FIG. 2.

At 114, the target model may be retrained at least with the pseudo-labeled data subset. It should be appreciated that without performing the pseudo-labeled data subset selection at 112, the target model may be retrained at least with the pseudo-labeled data set. Preferably, the target model may be retrained with both the pseudo-labeled data subset and the data-augmented ground-truth-labeled data set.

The process 100 involves operations on the unlabeled data set, e.g., filtering the unlabeled data set at 102, pretraining the target model with the filtered unlabeled data set at 104, labeling the filtered unlabeled data set through the target model at 110, selecting the pseudo-labeled data subset at 112, retraining the target model at least with the pseudo-labeled data subset at 114, etc. In the case where the unlabeled data set includes invisible samples, the unlabeled data set may be stored in a specific data center, and operations related to the unlabeled data set may be performed through special tools, e.g., the operations involving unlabeled data set described above in the process 100 or other operations in the process 100, therefore the unlabeled data set cannot be viewed or accessed by operators.

It should be appreciated that the process for model self-training described above in conjunction FIG. 1 is merely exemplary. According to actual application requirements, the steps in the process for model self-training may be replaced or modified in any manner, and the process may include more or fewer steps. For example, one or more of the steps of filtering the unlabeled data set, pretraining the target model, data-augmenting the ground-truth-labeled data set,

selecting a pseudo-labeled data subset, etc. may be omitted from the process 100. In addition, the specific order or hierarchy of the steps in the process 100 is only exemplary, and the process for model self-training may be performed in an order different from the described one.

FIG. 2 illustrates an exemplary process 200 for selecting a pseudo-labeled data subset from a pseudo-labeled data set according to an embodiment of the present disclosure. The process 200 may correspond to e.g., the step 112 in FIG. 1. For example, samples having task prediction values and/or domain prediction values that meet a quality requirement may be selected from a pseudo-labeled data set 202, to form a pseudo-labeled data subset 232.

The pseudo-labeled data set 202 may include a plurality of samples, and each sample may include, e.g., a text and a corresponding pseudo-label. The pseudo-label may be related to a specific task, thus it may also be referred to as a task prediction value. The task prediction values may be predicted through a target model. A task prediction value of a sample may indicate a probability for a result of a text included in the sample for a corresponding task. The higher the task prediction value, the higher the probability that the text is the result. Taking a classification task as an example, a task prediction value of a sample may indicate a probability that a text included in the sample is classified into a corresponding class. The higher the task prediction value, the higher the probability that the text is in this class. At 210, a task prediction value may be extracted from each sample in the pseudo-labeled data set 202 through task prediction value extraction, thereby obtaining a task prediction value set 212 corresponding to the pseudo-labeled data set 202.

The pseudo-labeled data set 202 may be provided to a domain classifier 220. The domain classifier 220 may be a machine learning model specially trained to predict a domain prediction value of an input sample. The domain classifier 220 may predict a domain prediction value of each sample in the pseudo-labeled data set 202, thereby obtaining a domain prediction value set 222 corresponding to the pseudo-labeled data set 202. A domain prediction value of a sample may indicate a probability that a text included in the sample is classified into a corresponding domain. The closer the domain prediction value is to an intermediate value, such as "0.5", the more ambiguous the possibility that the text is in this domain. This may indicate that the text has a general expression, rather than an expression related to a specific domain.

At 230, samples that meet a quality requirement may be selected from the pseudo-labeled data set 202 to form a pseudo-labeled data subset 232. For example, a pseudo-labeled data subset 232 may be formed through selecting, from the pseudo-labeled data set 202, samples having task prediction values higher than a predetermined threshold. For example, the pseudo-labeled data subset 232 may be formed through selecting, from the pseudo-labeled data set 202, samples having the task prediction values higher than "0.8". Retraining the target model with such

pseudo-labeled data subset may guarantee the accuracy of the retrained target model. Additionally or alternatively, the pseudo-labeled data subset 232 may be formed through selecting, from the pseudo-labeled data set 202, samples having domain prediction values within a predetermined range. For example, the pseudo-labeled data subset 232 may be formed through selecting, from the pseudo-labeled data set 202, samples having the domain prediction values between "0.4" and "0.6". Retraining the target model with such pseudo-labeled data subset may guarantee the fairness of the retrained target model. Preferably, the pseudo-labeled data subset 232 may be formed through selecting, from the pseudo-labeled data set 202, samples having the task prediction values higher than the predetermined threshold and the domain prediction values within the predetermined range, thereby guaranteeing the accuracy and the fairness of the retrained target model.

It should be appreciated that the process for selecting the pseudo-labeled data subset from the pseudo-labeled data set described above in conjunction with FIG. 2 is merely exemplary. According to actual application requirements, the steps in the process for selecting the pseudo-labeled data subset may be replaced or modified in any manner, and the process may include more or fewer steps. For example, although in the process 200, the pseudo-labeled data set 202 is provided to the domain classifier 220 to obtain the domain prediction value set 222, the unlabeled data set, such as the unlabeled data set mentioned in step 102 in FIG. 1, may also be provided to the domain classifier 220. In addition, the specific order or hierarchy of the steps in the process 200 is only exemplary, and the process for selecting the pseudo-labeled data subset may be performed in an order different from the described one.

FIG.3 illustrates an exemplary process 300 for model self-training including iterative operations according to an embodiment of the present disclosure. In the process 300, a target model may be trained with a ground-truth-labeled data set. The ground-truth-labeled data set may include a small amount of ground-truth-labeled samples. The target model may then label an unlabeled data set, to obtain a pseudo-labeled data set. The unlabeled data set may include a large amount of unlabeled samples, e.g., a large amount of invisible samples. Subsequently, the target model may be retrained at least with the pseudo-labeled data set. The labeling step and the retraining step may be performed iteratively.

Steps 302 to 314 may correspond to the steps 102 to 114 in FIG. 1. Through the steps 302 to 304, a target model may be pretrained with an unlabeled data set or a filtered unlabeled data set. Through the steps 306 to 308, the target model may be trained with a ground-truth-labeled data set or a data-augmented ground-truth-labeled data set. Through the steps 310 to 314, the target model may be retrained at least with a pseudo-labeled data set or a pseudo-labeled data subset.

At 316, it may be determined whether the target model meets a performance requirement. For

example, a test data set may be provided to the target model. The test data set may include a plurality of test samples. The target model may obtain a plurality of prediction results through predicting a plurality of test samples included in the test data set. The plurality of prediction results may be evaluated through a known way to determine whether the target model meets the performance requirement.

If it is determined at 316 that the target model does not meet the performance requirement, the process 300 may return to 310, and iteratively perform the labeling step and the retraining step through performing the steps 310 to 314 in sequence.

If it is determined at 316 that the target model meets the performance requirement, the process 300 may proceed to 318, i.e., the iteratively performing the labeling step and the retraining step may be stopped, and the process 300 may end.

In the process 300, the performance of the target model may be continuously improved through iteratively performing the step of labeling the unlabeled data set to obtain the pseudo-labeled data set and the step of retraining the target model at least with the pseudo-labeled data set, thereby a more accurate and robust target model is obtained.

It should be appreciated that the process for model self-training including the iterative operations described above in conjunction FIG. 3 is merely exemplary. According to actual application requirements, the steps in the process for model self-training including the iterative operations may be replaced or modified in any manner, and the process may include more or fewer steps. For example, one or more of the steps of filtering the unlabeled data set, pretraining the target model, data-augmenting the ground-truth-labeled data set, selecting the pseudo-labeled data subset, etc. may be omitted from the process 300. In addition, the specific order or hierarchy of the steps in the process 300 is only exemplary, and the process for model self-training including the iterative operations may be performed in an order different from the described one.

FIG.4 illustrates an exemplary process 400 for model self-training including iterative operations and a self-ensemble strategy according to an embodiment of the present disclosure. In the process 400, a target model may be trained with a ground-truth-labeled data set. The ground-truth-labeled data set may include a small amount of ground-truth-labeled samples. The target model may then label an unlabeled data set, to obtain a pseudo-labeled data set. The unlabeled data set may include a large amount of unlabeled samples, e.g., a large amount of invisible samples. Subsequently, the target model may be retrained at least with the pseudo-labeled data set. The labeling step and the retraining step may be performed iteratively. In each iteration, after performing the retraining step, the target model may be configured with a comprehensive parameter set determined through employing a self-ensemble strategy.

Steps 402 to 414 may correspond to the steps 102 to 114 in FIG. 1. Through the steps 402 to

404, a target model may be pretrained with an unlabeled data set or a filtered unlabeled data set. Through the steps 406 to 408, the target model may be trained with a ground-truth-labeled data set or a data-augmented ground-truth-labeled data set. Through the steps 410 to 414, the target model may be retrained at least with a pseudo-labeled data set or a pseudo-labeled data subset.

At 416, the target model may be configured with a comprehensive parameter set. The comprehensive parameter set may be determined through employing a self-ensemble strategy. For example, the comprehensive parameter set may be determined based on a current parameter set of the target model obtained in a current iteration and one or more previous parameter sets of the target model obtained in one or more previous iterations. In an implementation, the one or more previous iterations may include all the iterations before the current iteration. Accordingly, the one or more previous parameter sets may be all the previous parameter sets of the target model obtained in all the iterations before the current iteration. That is, the comprehensive parameter set may be determined based on the current parameter set of the target model obtained in the current iteration and all the previous parameter sets of the target model obtained in all the previous iterations. In another implementation, since the parameter set of the target model is continuously optimized, only the predetermined number of iterations before the current iteration may be considered, and some initial iterations may not be considered. That is, the comprehensive parameter set may be determined based on the current parameter set of the target model obtained in the current iteration and the predetermined number of previous parameter sets of the target model obtained in the predetermined number of previous iterations before the current iteration. As an example, the one or more previous iterations may include three iterations before the current iteration. Accordingly, the one or more previous parameter sets may be three previous parameter sets of the target model obtained in the three iterations before the current iteration.

At 418, it may be determined whether the target model meets a performance requirement. The step 418 may correspond to the step 316 in FIG. 3. If it is determined at 418 that the target model does not meet the performance requirement, the process 400 may return to 410, and iteratively perform the labeling step, the retraining step and the target model configuring step through performing the steps 410 to 416 in sequence. If it is determined at 418 that the target model meets the performance requirement, the process 400 may proceed to 420, i.e., the iteratively performing the labeling step and the retraining step may be stopped, and the process 400 may end.

In the process 400, in each iteration, after the retraining step is performed, the self-ensemble strategy may be employed to determine the comprehensive parameter set, and the target model may be configured with the determined comprehensive parameter set. The target model

configured through employing the comprehensive parameter set determined by the self-ensemble strategy may be more robust and can more accurately label an unlabeled data set.

It should be appreciated that the process for model self-training including the iterative operations and the self-ensemble strategy described above in conjunction FIG. 4 is merely exemplary. According to actual application requirements, the steps in the process for model self-training including the iterative operations and the self-ensemble strategy may be replaced or modified in any manner, and the process may include more or fewer steps. For example, one or more of the steps of filtering the unlabeled data set, pretraining the target model, data-augmenting the ground-truth-labeled data set, selecting the pseudo-labeled data subset, etc. may be omitted from the process 400. In addition, the specific order or hierarchy of the steps in the process 400 is only exemplary, and the process for model self-training including the iterative operations and the self-ensemble strategy may be performed in an order different from the described one.

FIG.5 illustrates an example 500 of a process for model self-training including iterative operations and a self-ensemble strategy. The process 500 shows exemplary processes of the first iteration and the second iteration. In the first iteration, a target model 502-1 may be a target model trained e.g., through the steps 402 to 408 in FIG. 4. At 504-1, the target model 502-1 may label a filtered unlabeled data set 502-1, to obtain a pseudo-labeled data set. This step may correspond to the step 410 in FIG. 4. At 506-1, a pseudo-labeled data subset that meets a quality requirement may be selected from the pseudo-labeled data set. This step may correspond to the step 412 in FIG. 4. At 508-1, the target model may be retrained at least with the pseudo-labeled data subset, to obtain the target model 510-1. This step may correspond to the step 414 in FIG. 4. The target model 510-1 may have a parameter set 512-1.

The target model 510-1 may then be used to perform the second iteration. In the second iteration, the target model 502-2 may correspond to target model 510-1. At 504-2, the target model 502-2 may label the filtered unlabeled data set, to obtain a pseudo-labeled data set. This step may correspond to the step 410 in FIG. 4. At 506-2, a pseudo-labeled data subset that meets a quality requirement may be selected from the pseudo-labeled data set. This step may correspond to the step 412 in FIG. 4. At 508-2, the target model may be retrained at least with the pseudo-labeled data subset, to obtain the target model 510-2. This step may correspond to the step 414 in FIG. 4. The target model 510-2 may have a parameter set 512-2. At 514- 2, a comprehensive parameter set may be determined. In an implementation, the comprehensive parameter set may be determined based on the parameter sets 512-1 and 512-2. For example, for each parameter, values in the parameter sets 512-1 and 512-2 may be weighted averaged to obtain a comprehensive value of the parameter, so that the comprehensive parameter set corresponding to all the parameters may be obtained. At 516-2, the target model may be

configured with the determined comprehensive parameter set, thus the target model 518-2 may be obtained. The steps 514-2 and 516-2 may correspond to the step 416 in FIG. 4.

The target model 518-2 may then be used to perform the third iteration (not shown). In the third iteration, after a current target model is obtained through the labeling step, the selecting step, the retraining step, etc., the comprehensive parameter set may be determined based on, e.g., a current parameter set of the current target model and the parameter set 512-2, or both the parameter set 512-1 and the parameter set 512-2, and the target model may be configured with the comprehensive parameter set and the configured target model may be used in the next iteration. Through repeating the above process in one or more subsequent iterations, the target model for the next iteration may be configured through continuously using new comprehensive parameter set.

A retrained target model may be obtained through any of the process 100, the process 300, and the process 400 shown in FIG. 1, FIG. 3, and FIG. 4, respectively. A predetermined task may be performed through a retrained target model. The predetermined task may be a task that the target model is for. FIG. 6 illustrates an exemplary process 600 for performing a predetermined task through a retrained target model according to an embodiment of the present disclosure.

At 602, a target model may be obtained. The target model may be obtained through any of the process 100, the process 300, and the process 400 shown in FIG. 1, FIG. 3, and FIG. 4, respectively.

At 604, it may be determined whether the target model meets a complexity requirement. In an implementation, whether the target model meets the complexity requirement may be determined based on whether the number of parameters included in the target model exceeds a predetermined threshold. For example, when the number of parameters included in the target model does not exceed the predetermined threshold, it may be considered that the target model meets the complexity requirement. Such a target model will have a faster speed when inferring or predicting, and thus may be deployed to perform a predetermined task. When the number of parameters included in the target model exceeds the predetermined threshold, it may be considered that the target model does not meet the complexity requirement. Such a target model will be very time-consuming when inferring or predicting, and therefore it is inconvenient to be deployed to perform a predetermined task.

If it is determined at 604 that the target model meets the complexity requirement, the process may proceed to 606. At 606, the target model may be deployed to perform a predetermined task.

If it is determined at 604 that the target model does not meet the complexity requirement, an unlabeled data set may be labeled through a retrained target model to obtain a pseudo-labeled data set for training a second model. The second model may be e.g., a simple model with fewer

parameters than the target model. For example, the process 600 may proceed to 608. At 608, an unlabeled data set may be labeled through the target model, to obtain a pseudo-labeled data set.

At 610, the second model is trained at least with the pseudo-labeled data set. Preferably, the second model may be trained with both the pseudo-labeled data set and a data-augmented ground-truth-labeled data set.

At 612, the second model may be deployed to perform a predetermined task.

It should be appreciated that the process for performing the predetermined task through the retrained target model described above in conjunction with FIG. 6 is merely exemplary. According to actual application requirements, the steps in the process for performing the predetermined task through the retrained target model may be replaced or modified in any manner, and the process may include more or fewer steps. For example, the process for determining whether the target model meets the complexity requirement may be omitted from the process 600, thus the unlabeled data set may be directly labeled through the retrained target model, to obtain the pseudo-labeled data set for training the second model. In addition, the specific order or hierarchy of the steps in the process 600 is only exemplary, and the process for performing the predetermined task through the retrained target model may be performed in an order different from the described one.

FIG.7 is a flowchart of an exemplary method 700 for model self-training according to an embodiment of the present disclosure.

At 710, a target model may be trained with a ground-truth-labeled data set.

At 720, an unlabeled data set may be labeled through the target model, to obtain a pseudo-labeled data set.

At 730, the target model may be retrained at least with the pseudo-labeled data set.

In an implementation, the unlabeled data set may include a plurality of invisible samples.

In an implementation, the method 700 may further comprise: pretraining the target model with the unlabeled data set. The training a target model may comprise: training the pretrained target model with the ground-truth-labeled data set.

The method 700 may further comprise: filtering the unlabeled data set through a predefined dictionary. The pretraining the target model may comprise: pretraining the target model with the filtered unlabeled data set.

In an implementation, the method 700 may further comprise: data-augmenting the ground-truth-labeled data set. The training a target model may comprise: training the target model with the data-augmented ground-truth-labeled data set.

In an implementation, the retraining the target model may comprise: retraining the target model with the pseudo-labeled data set and the ground-truth-labeled data set.

In an implementation, the method 700 may further comprise: selecting, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement. The retraining the target model may comprise: retraining the target model at least with the pseudo-labeled data subset.

The selecting a pseudo-labeled data subset comprises: selecting, from the pseudo-labeled data set, samples having task prediction values and/or domain prediction values that meet the quality requirement, to form the pseudo-labeled data subset.

A task prediction value of each sample in the pseudo-labeled data set may be predicted through the target model.

A domain prediction value of each sample in the pseudo-labeled data set may be predicted through a domain classifier.

In an implementation, the method 700 may further comprise: iteratively performing the labeling step and the retraining step.

The method 700 may further comprise: in each iteration, after performing the retraining step, configuring the target model with a comprehensive parameter set.

The comprehensive parameter set may be determined based on a current parameter set of the target model obtained in a current iteration and one or more previous parameter sets of the target model obtained in one or more previous iterations.

The method 700 may further comprise: determining whether the target model meets a performance requirement; and in response to determining that the target model meets the performance requirement, stopping iteratively performing the labeling step and the retraining step.

In an implementation, the method 700 may further comprise: labeling the unlabeled data set through the retrained target model, to obtain a pseudo-labeled data set for training a second model.

It should be appreciated that the method 700 may further comprise any step/process for model self-training according to the embodiments of the present disclosure as mentioned above.

FIG.8 illustrates an exemplary apparatus 800 for model self-training according to an embodiment of the present disclosure.

The apparatus 800 may comprise: a target model training module 810, for training a target model with a ground-truth-labeled data set; a data set labeling module 820, for labeling an unlabeled data set through the target model, to obtain a pseudo-labeled data set; and a target model retraining module 830, for retraining the target model at least with the pseudo-labeled data set. Moreover, the apparatus 800 may further comprise any other modules configured for model self-training according to the embodiments of the present disclosure as mentioned above.

FIG.9 illustrates an exemplary apparatus 900 for model self-training according to an

embodiment of the present disclosure.

The apparatus 900 may comprise at least one processor 910 and a memory 920 storing computer-executable instructions. The computer-executable instructions, when executed, may cause the at least one processor 910 to: train a target model with a ground-truth-labeled data set; label an unlabeled data set through the target model, to obtain a pseudo-labeled data set, and retrain the target model at least with the pseudo-labeled data set.

In an implementation, the computer-executable instructions, when executed, may further cause the at least one processor 910 to: select, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement. The retraining the target model may comprise: retraining the target model at least with the pseudo-labeled data subset.

The selecting a pseudo-labeled data subset comprises: selecting, from the pseudo-labeled data set, samples having task prediction values and/or domain prediction values that meet the quality requirement, to form the pseudo-labeled data subset.

A domain prediction value of each sample in the pseudo-labeled data set may be predicted through a domain classifier.

It should be appreciated that the processor 910 may further perform any other steps/processes of methods for model self-training according to the embodiments of the present disclosure as mentioned above.

The embodiments of the present disclosure propose a computer program product for model self-training, comprising a computer program that is executed by at least one processor for: training a target model with a ground-truth-labeled data set; labeling an unlabeled data set through the target model, to obtain a pseudo-labeled data set; and retraining the target model at least with the pseudo-labeled data set. In addition, the computer programs may further be performed for implementing any other step/process of methods for model self-training according to the embodiments of the present disclosure as mentioned above.

The embodiments of the present disclosure may be embodied in a non-transitory computer-readable medium. The non-transitory computer readable medium may comprise instructions that, when executed, cause one or more processors to perform any operation of methods for model self-training according to the embodiments of the present disclosure as mentioned above.

It should be appreciated that all the operations in the methods described above are merely exemplary, and the present disclosure is not limited to any operations in the methods or sequence orders of these operations, and should cover all other equivalents under the same or similar concepts. In addition, the articles “a” and “an” as used in this specification and the appended claims should generally be construed to mean “one” or “one or more” unless specified otherwise or clear from the context to be directed to a singular form.

It should also be appreciated that all the modules in the apparatuses described above may be implemented in various approaches. These modules may be implemented as hardware, software, or a combination thereof. Moreover, any of these modules may be further functionally divided into sub-modules or combined together.

Processors have been described in connection with various apparatuses and methods. These processors may be implemented using electronic hardware, computer software, or any combination thereof. Whether such processors are implemented as hardware or software will depend upon the particular application and overall design constraints imposed on the system. By way of example, a processor, any portion of a processor, or any combination of processors presented in the present disclosure may be implemented with a microprocessor, microcontroller, digital signal processor (DSP), a field-programmable gate array (FPGA), a programmable logic device (PLD), a state machine, gated logic, discrete hardware circuits, and other suitable processing components configured for performing the various functions described throughout the present disclosure. The functionality of a processor, any portion of a processor, or any combination of processors presented in the present disclosure may be implemented with software being executed by a microprocessor, microcontroller, DSP, or other suitable platform.

Software shall be construed broadly to mean instructions, instruction sets, code, code segments, program code, programs, subprograms, software modules, applications, software applications, software packages, routines, subroutines, objects, threads of execution, procedures, functions, etc. The software may reside on a computer-readable medium. A computer-readable medium may include, by way of example, memory such as a magnetic storage device (e.g., hard disk, floppy disk, magnetic strip), an optical disk, a smart card, a flash memory device, random access memory (RAM), read only memory (ROM), programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), a register, or a removable disk. Although memory is shown separate from the processors in the various aspects presented throughout the present disclosure, the memory may be internal to the processors, e.g., cache or register.

The previous description is provided to enable any person skilled in the art to practice the various aspects described herein. Various modifications to these aspects will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other aspects. Thus, the claims are not intended to be limited to the aspects shown herein. All structural and functional equivalents to the elements of the various aspects described throughout the present disclosure that are known or later come to be known to those of ordinary skilled in the art are expressly incorporated herein and intended to be encompassed by the claims.

CLAIMS

1. A method for model self-training, comprising:

training a target model with a ground-truth-labeled data set;

labeling an unlabeled data set through the target model, to obtain a pseudo-labeled data set; and

retraining the target model at least with the pseudo-labeled data set.

2. The method of claim 1, wherein the unlabeled data set includes a plurality of invisible samples.

3. The method according to claim 1, further comprising:

pretraining the target model with the unlabeled data set, and

wherein the training a target model comprises: training the pretrained target model with the ground-truth-labeled data set.

4. The method according to claim 3, further comprising:

filtering the unlabeled data set through a predefined dictionary, and

wherein the pretraining the target model comprises: pretraining the target model with the filtered unlabeled data set.

5. The method according to claim 1, further comprising:

data-augmenting the ground-truth-labeled data set, and

wherein the training a target model comprises: training the target model with the data-augmented ground-truth-labeled data set.

6. The method of claim 1, further comprising:

selecting, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement, and

wherein the retraining the target model comprises: retraining the target model at least with the pseudo-labeled data subset.

7. The method of claim 6, wherein the selecting a pseudo-labeled data subset comprises:

selecting, from the pseudo-labeled data set, samples having task prediction values and/or domain prediction values that meet the quality requirement, to form the pseudo-labeled data subset.

8. The method according to claim 7, wherein a domain prediction value of each sample in the pseudo-labeled data set is predicted through a domain classifier.

9. The method of claim 1, further comprising:

iteratively performing the labeling step and the retraining step.

10. The method according to claim 9, further comprising, in each iteration:

after performing the retraining step, configuring the target model with a comprehensive

parameter set.

11. The method of claim 10, wherein the comprehensive parameter set is determined based on a current parameter set of the target model obtained in a current iteration and one or more previous parameter sets of the target model obtained in one or more previous iterations.

12. The method of claim 9, further comprising:

determining whether the target model meets a performance requirement; and

in response to determining that the target model meets the performance requirement, stopping iteratively performing the labeling step and the retraining step.

13. The method of claim 1, further comprising:

labeling the unlabeled data set through the retrained target model, to obtain a pseudo-labeled data set for training a second model.

14. An apparatus for model self-training, comprising:

at least one processor; and

a memory storing computer-executable instructions that, when executed, cause the at least one processor to:

train a target model with a ground-truth-labeled data set,

label an unlabeled data set through the target model, to obtain a pseudo-labeled data set, and

retrain the target model at least with the pseudo-labeled data set.

15. A computer program product for model self-training, comprising a computer program that is executed by at least one processor for:

training a target model with a ground-truth-labeled data set;

labeling an unlabeled data set through the target model, to obtain a pseudo-labeled data set; and

retraining the target model at least with the pseudo-labeled data set.

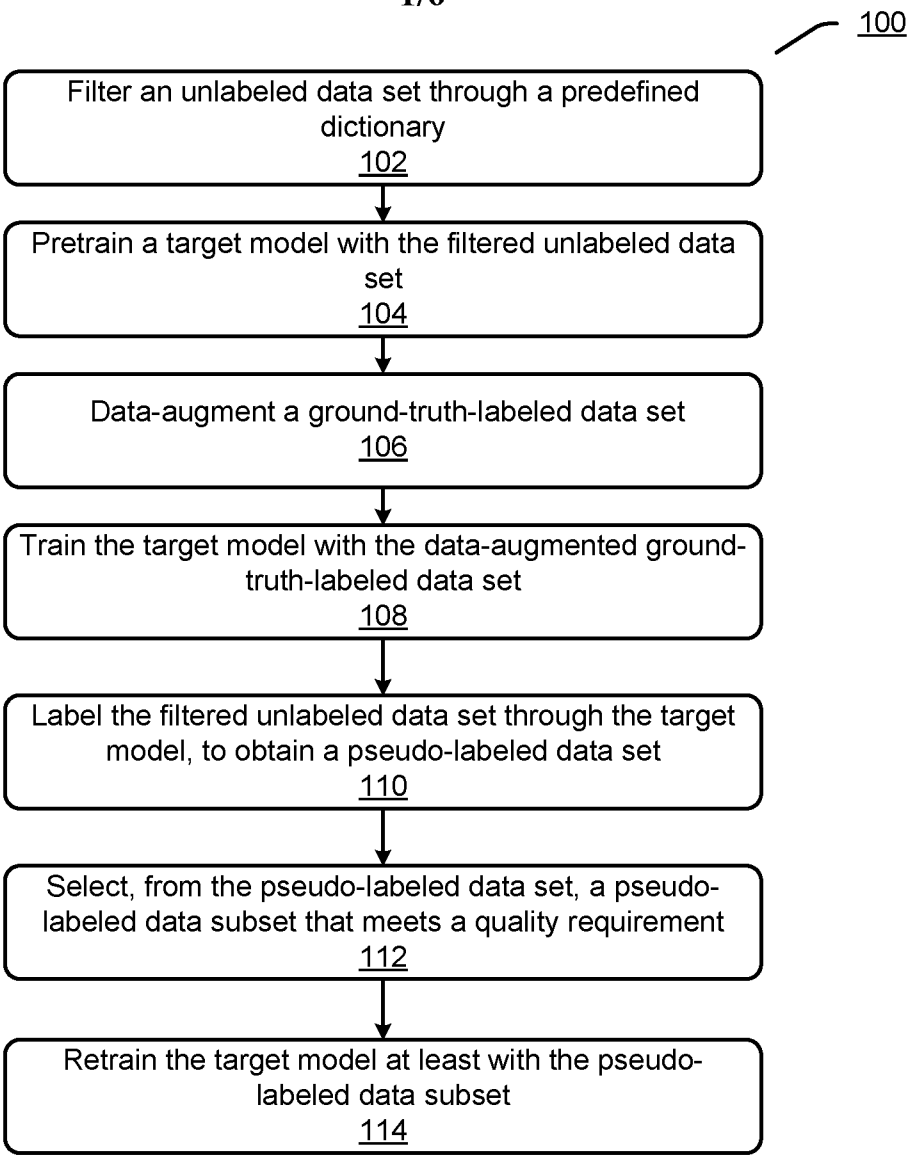

100
Filter an unlabeled data set through a predefined dictionary
102
Pretrain a target model with the filtered unlabeled data set
104
Data-augment a ground-truth-labeled data set
106
Train the target model with the data-augmented ground-truth-labeled data set
108
Label the filtered unlabeled data set through the target model, to obtain a pseudo-labeled data set
110
Select, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement
112
Retrain the target model at least with the pseudo-labeled data subset
114

FIG 1

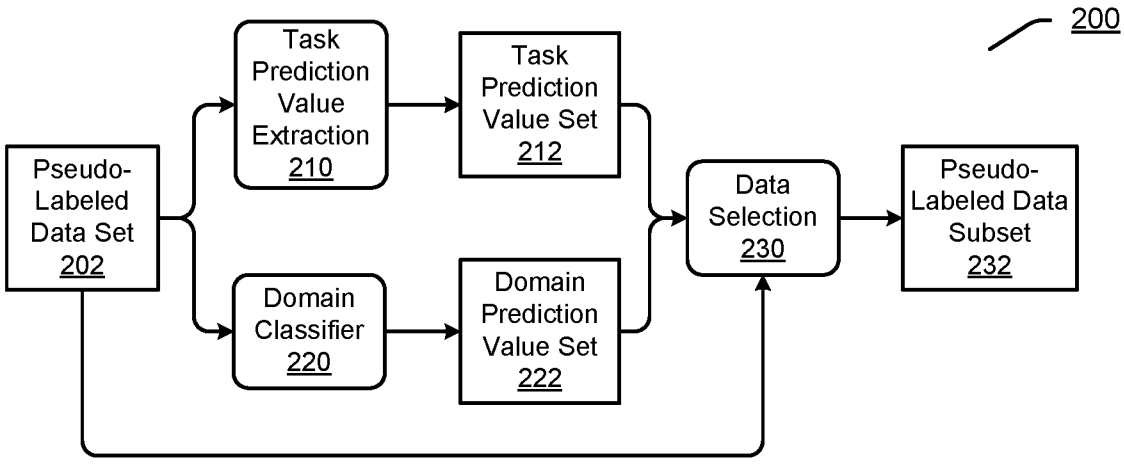

200
Pseudo-Labeled Data Set
202
Task Prediction Value Extraction
210
Task Prediction Value Set
212
Domain Classifier
220
Domain Prediction Value Set
222
Data Selection
230
Pseudo-Labeled Data Subset
232

FIG 2

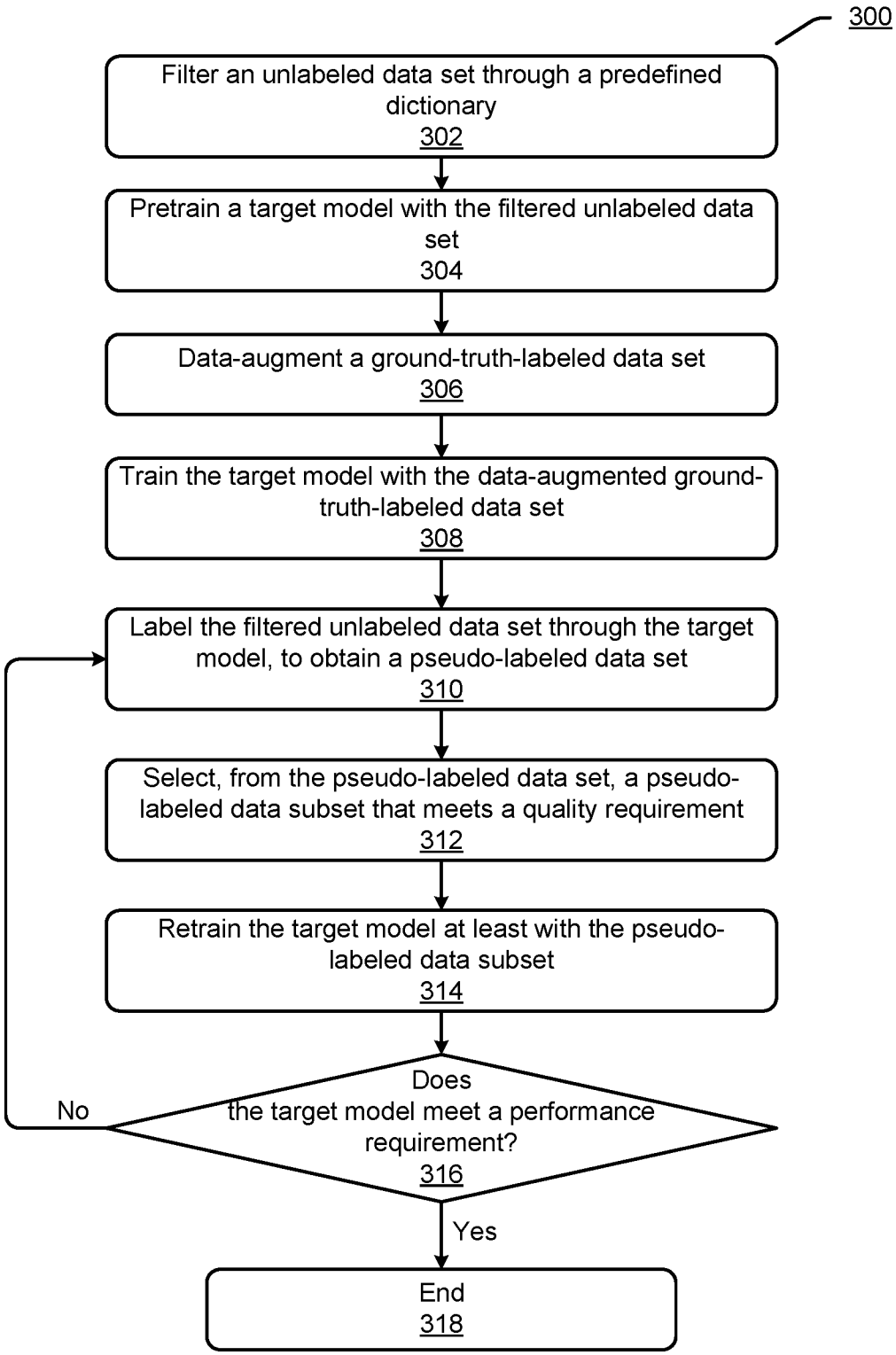

300
Filter an unlabeled data set through a predefined dictionary
302
Pretrain a target model with the filtered unlabeled data set
304
Data-augment a ground-truth-labeled data set
306
Train the target model with the data-augmented ground-truth-labeled data set
308
Label the filtered unlabeled data set through the target model, to obtain a pseudo-labeled data set
310
Select, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement
312
Retrain the target model at least with the pseudo-labeled data subset
314
Does the target model meet a performance requirement?
316
No
Yes
End
318

FIG 3

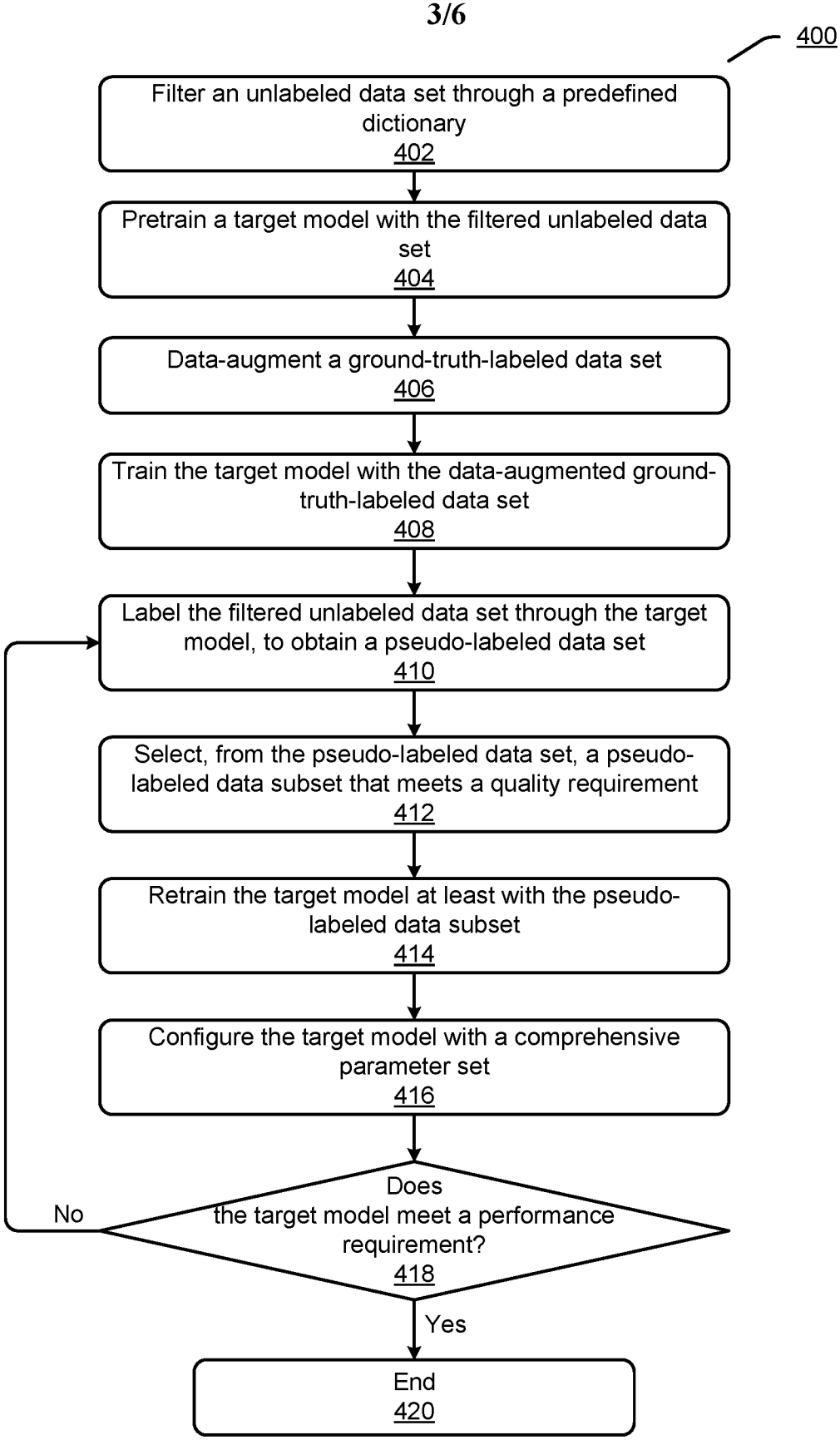

400
Filter an unlabeled data set through a predefined dictionary
402
Pretrain a target model with the filtered unlabeled data set
404
Data-augment a ground-truth-labeled data set
406
Train the target model with the data-augmented ground-truth-labeled data set
408
Label the filtered unlabeled data set through the target model, to obtain a pseudo-labeled data set
410
Select, from the pseudo-labeled data set, a pseudo-labeled data subset that meets a quality requirement
412
Retrain the target model at least with the pseudo-labeled data subset
414
Configure the target model with a comprehensive parameter set
416
Does the target model meet a performance requirement?
418
No
Yes
End
420

FIG 4

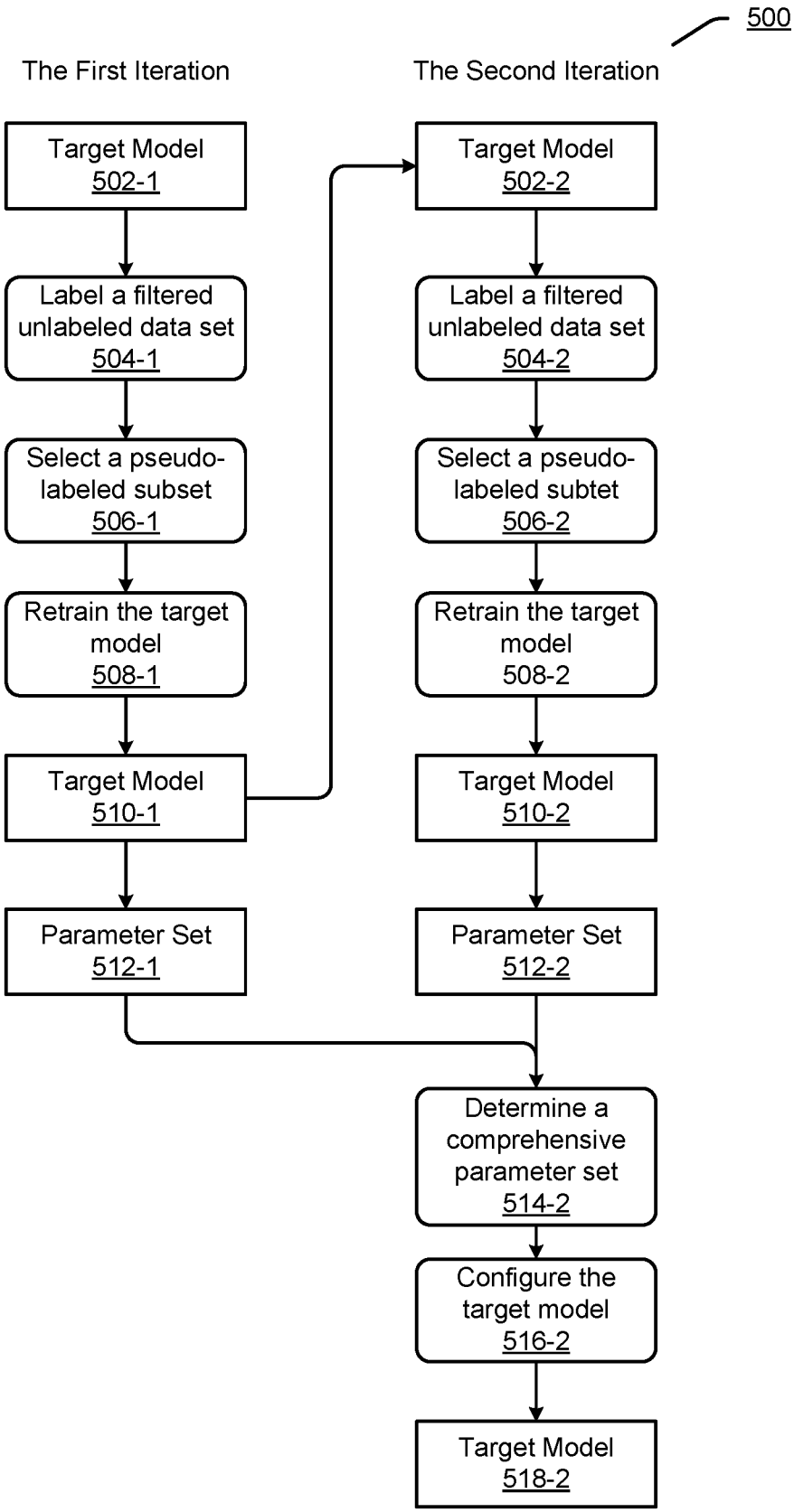

500
The First Iteration
The Second Iteration
Target Model 502-1
Label a filtered unlabeled data set 504-1
Select a pseudo-labeled subset 506-1
Retrain the target model 508-1
Target Model 510-1
Parameter Set 512-1
Target Model 502-2
Label a filtered unlabeled data set 504-2
Select a pseudo-labeled subtet 506-2
Retrain the target model 508-2
Target Model 510-2
Parameter Set 512-2
Determine a comprehensive parameter set 514-2
Configure the target model 516-2
Target Model 518-2

FIG 5

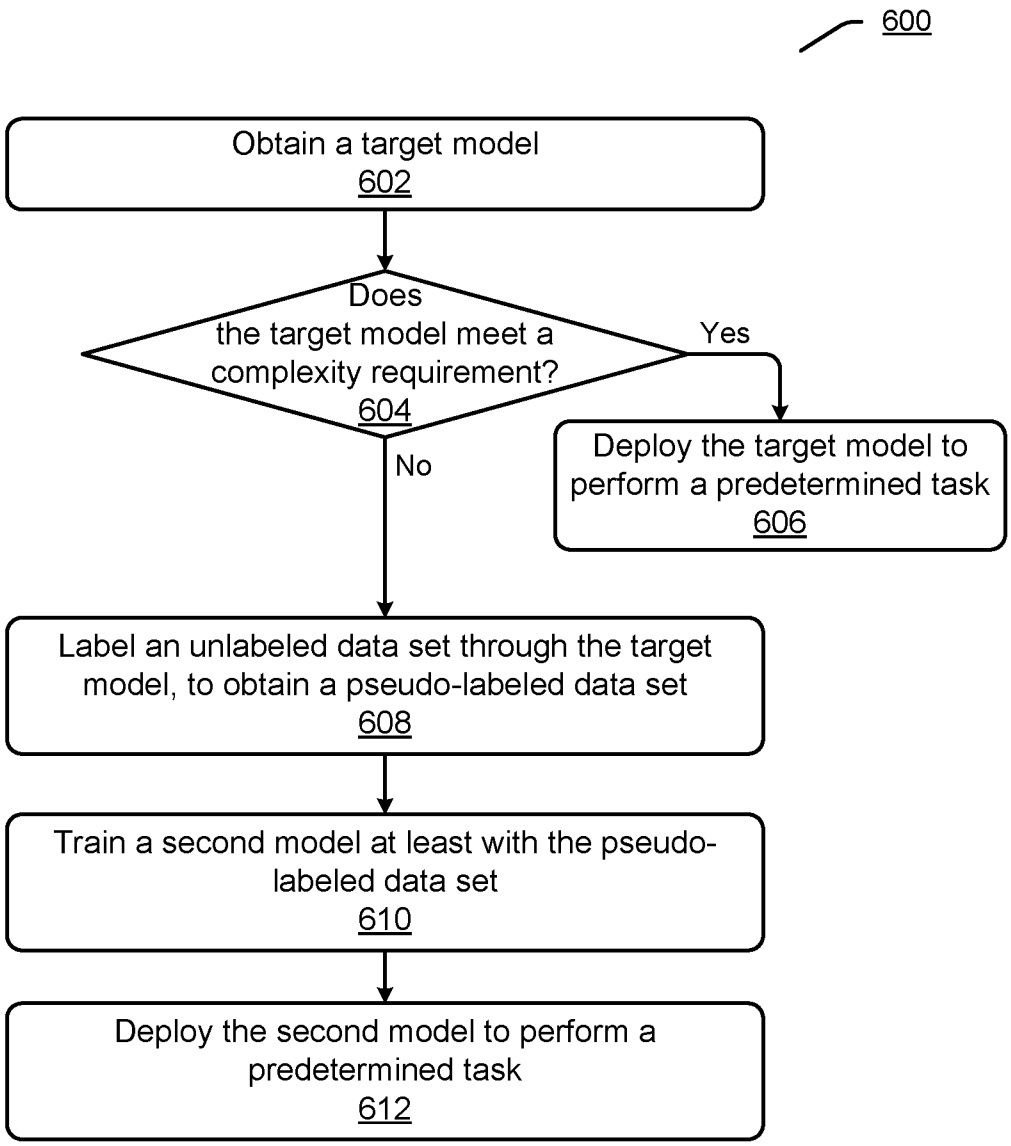

600
Obtain a target model
602
Does the target model meet a complexity requirement?
604
Yes
No
Deploy the target model to perform a predetermined task
606
Label an unlabeled data set through the target model, to obtain a pseudo-labeled data set
608
Train a second model at least with the pseudo-labeled data set
610
Deploy the second model to perform a predetermined task
612

FIG 6

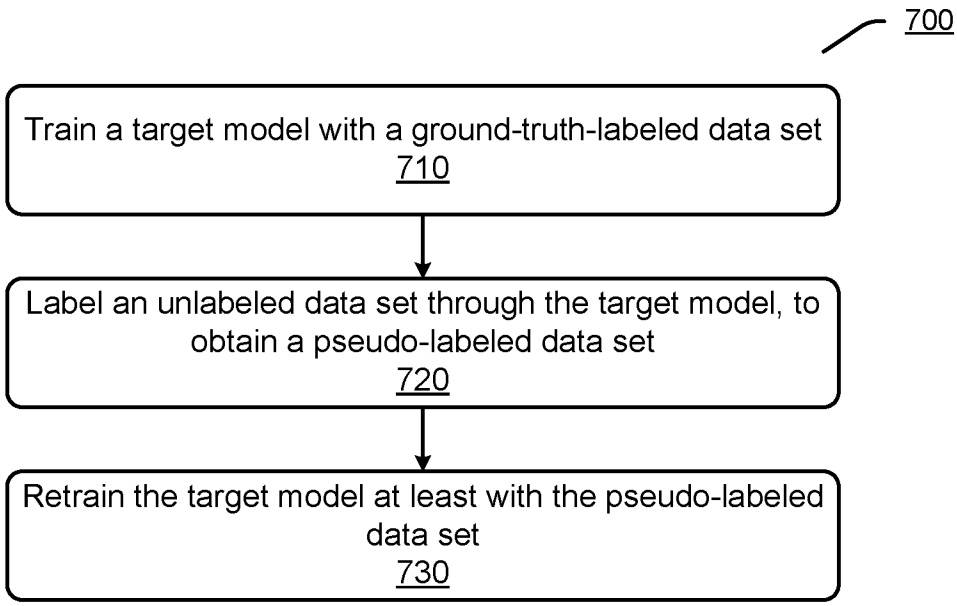


FIG 7

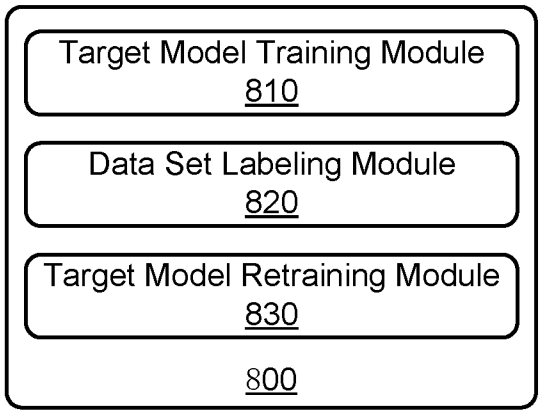


FIG 8

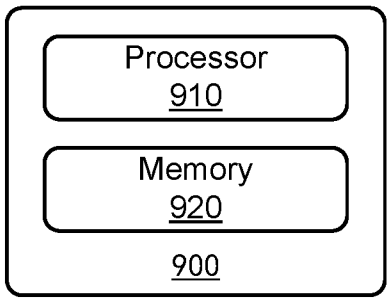


FIG 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/028884

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N20/00
ADD.

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

B. FIELDS SEARCHED

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LI ZHEN-ZHEN ET AL: "Learning to select pseudo labels: a semi-supervised method for named entity recognition", FRONTIERS OF INFORMATION TECHNOLOGY & ELECTRONIC ENGINEERING, ZHEJIANG UNIVERSITY PRESS, HEIDELBERG, vol. 21, no. 6, 27 December 2019 (2019-12-27), pages 903-916, XP037181900, ISSN: 2095-9184, DOI: 10.1631/FITEE.1800743 [retrieved on 2019-12-27] abstract; figures 1, 2 Section 1, 3 ----- -/--	1-15

[x] Further documents are listed in the continuation of Box C.

[x] 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
8 August 2022	19/08/2022

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	Houtgast, Ernst

1

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/028884

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018/137433 A1 (DEVARAKONDA MURTHY V [US] ET AL) 17 May 2018 (2018-05-17) abstract; figures 1,5,6 paragraphs [0007], [0022] - [0026], [0132] - [0134]	1-15
X	WO 2020/068784 A1 (SCHLUMBERGER TECHNOLOGY CORP [US]; SCHLUMBERGER CA LTD [CA] ET AL.) 2 April 2020 (2020-04-02) abstract; figures 1-8 paragraphs [0004] - [0012], [0050], [0051], [0064] - [0075]	1-15

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2022/028884

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018137433 A1	17-05-2018	NONE	
WO 2020068784 A1	02-04-2020	US 2021406644 A1	30-12-2021
		WO 2020068784 A1	02-04-2020