



(51) International Patent Classification:

G06N 3/08 (2006.01)

G06N 7/00 (2006.01)

(21) International Application Number:

PCT/CN2021/097267

(22) International Filing Date:

31 May 2021 (31.05.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicants: **ROBERT BOSCH GMBH** [DE/DE]; Postfach 30 02 20, 70442 Stuttgart (DE). **TSINGHUA UNIVERSITY** [CN/CN]; Qinghua Yuan 1, Haidian District, Beijing 100084 (CN).

(72) Inventors: **DONG, Yinpeng**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **YANG, Xiao**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **DENG, Zhijie**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **PANG, Tianyu**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **SU, Hang**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **ZHU, Jun**; Tsinghua University, Qinghua Yuan 1, Haidian District, Beijing 100084 (CN). **CHENG, Ze**; Bosch (China) Investment Ltd., 333

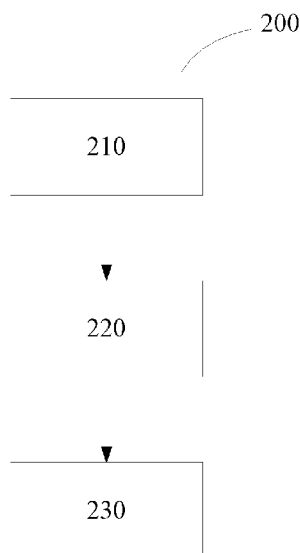
North Fuquan Road, Changning District, Shanghai 200335 (CN).

(74) Agent: **NTD PATENT AND TRADEMARK AGENCY LTD.**; 10th Floor, Tower C, Beijing Global Trade Center, 36 North Third Ring Road East, Dongcheng District, Beijing 100013 (CN).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, 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,

(54) Title: METHOD AND APPARATUS FOR DEEP LEARNING



(57) Abstract: The present disclosure provides a method for deep learning. The method comprises: receiving, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples; adversarially augmenting, by the deep learning model, the plurality of samples based on a threat model; and assigning, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

FIG. 2

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

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

METHOD AND APPARATUS FOR DEEP LEARNING

FIELD

[0001] Aspects of the present disclosure relate generally to artificial intelligence or machine learning, and more particularly, to a method and an apparatus for deep learning.

BACKGROUND

[0002] Deep neural networks (DNNs) have proved their effectiveness on a variety of domains and tasks. For example, DNNs are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. However, it has been found that DNNs are highly vulnerable to adversarial examples: adversarially perturbed examples that cause mis-classification while being nearly “imperceptible”, i.e., close to the original example. To enhance the robustness of DNNs against adversarial examples, adversarially training has proven to be among the most effective defense techniques, in which the network is trained on the adversarially augmented samples instead of on the natural or original ones.

[0003] However, a key problem of adversarially training is that, unlike in traditional deep learning, overfitting is a dominant phenomenon in adversarially training of deep networks. That is, adversarially training has the property that, after a certain point, further training will continue to substantially decrease the robust training loss, while increasing the robust test loss, witnessing a significant generalization gap. This failure on robust generalization may be surprising particularly since properly tuned DNNs (e.g., convolutional networks) rarely overfit much on standard vision datasets.

[0004] Many efforts have been devoted to studying robust generalization of adversarially training theoretically or empirically. In an aspect, some study results show that significantly increased sample complexity may be necessary for robust generalization, since robust generalization may require a more nuanced understanding of the data distribution than standard generalization. In another aspect, when the number of parameters is large, some theory suggests that some form of regularization is needed to ensure small generalization error.

[0005] It may be desirable to provide a method to mitigate robust overfitting and achieve better robustness.

SUMMARY

[0006] The following presents a simplified summary of one or more aspects according to the present disclosure in order to provide a basic understanding of such aspects. This summary is not an extensive overview of all contemplated aspects, and is intended to neither identify key or critical elements of all aspects nor delineate the scope of any or all aspects. Its sole purpose is to present some concepts of one or more aspects in a simplified form as a prelude to the more detailed description that is presented later.

[0007] In an aspect of the disclosure, a method for deep learning, comprising: receiving, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples; adversarially augmenting, by the deep learning model, the plurality of samples based on a threat model; and assigning, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

[0008] In another aspect of the disclosure, apparatus for deep learning comprises a memory; and at least one processor coupled to the memory. The at least one processor is configured to receive, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples; adversarially augment, by the deep learning model, the plurality of samples based on a threat model; and assign, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

[0009] In another aspect of the disclosure, a computer program product for deep learning comprises processor executable computer code for receiving, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples; adversarially augmenting, by the deep learning model, the plurality of samples based on a threat model; and assigning, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

[0010] In another aspect of the disclosure, a computer readable medium stores computer code for deep learning. The computer code when executed by a processor causes the processor to receive, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples; adversarially augment, by the deep learning model, the plurality of samples based on a threat model; and assign, by the deep learning model, a low predictive confidence to one or more adversarially

augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

[0011] To assign a relatively low predictive confidence for adversarially augmented samples having noisy labels in adversarially training, the presented method may leverage all training samples to suffice the need of sample complexity for adversarially training and hinder deep learning models from excessive memorization of one-hot labels in adversarially training.

[0012] Other aspects or variations of the disclosure, as well as other advantages thereof will become apparent by consideration of the following detailed description and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] 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.

[0014] **FIG. 1** shows an example of robust overfitting.

[0015] **FIG. 2** illustrates an example flow diagram of method according to one or more aspects of the present disclosure.

[0016] **FIG. 3** shows a conceptual illustration of decision boundaries.

[0017] **FIG. 4A** and **FIG. 4B** show the natural and robust training accuracies of PGD-AT and TRADES, respectively, when trained on true or random labels, with explicit regularizations turned off.

[0018] **FIG. 4C** shows the generalization gap under varying levels of label noise from 0% (true labels) to 100% (completely random labels).

[0019] **FIG. 5** illustrates an example of a hardware implementation for an apparatus according to one or more aspects of the present disclosure.

DETAILED DESCRIPTION

[0020] 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.

[0021] Recent studies identify the robust overfitting phenomenon in adversarially training, i.e., shortly after the first learning rate decay, further training will continue to decrease the robust test accuracy. It further identifies that several remedies for overfitting, including explicit ℓ_1 and ℓ_2 regularization, data augmentation, etc., cannot

gain improvements upon early stopping. Although robust overfitting has been thoroughly investigated, there may still lack an explanation of why it occurs. **FIG. 1** shows an example of robust overfitting, where four accuracy curves of TRADES (TRadeoff-inspired Adversarial DEFense via Surrogate-loss minimization) on CIFAR-10 illustrate a standard generalization gap between natural training accuracy and natural test accuracy, and a robust generalization gap between robust training accuracy and robust test accuracy respectively. As shown in **FIG. 1**, after the first learning rate decay (e.g., around 100 epochs), further training will continue to increase the natural test accuracy, however to decrease the robust test accuracy steadily, witnessing growing robust generalization gap.

[0022] The present disclosure proposes a method for mitigating robust overfitting. The present disclosure identifies that the cause of robust overfitting of adversarially training may lie in memorization of one-hot labels. Some samples may naturally lie close to a decision boundary (e.g., may be referred to as “hard” samples hereafter), and should be assigned a low prediction confidence for the corresponding worst-case adversarially augmented samples relative to the samples that are located relatively far from the decision boundary (e.g., may be referred to as “easy” samples hereafter). Because it may be difficult for a network to assign one-hot labels for all perturbed samples within a perturbation budget, and labels can be inappropriate for some adversarially augmented samples that lie close to the decision boundary. In one aspect of the present disclosure, the true labels of some training samples may be noisy for adversarially training. As an example, the true one-hot labels of some training samples may become wrong labels when being attacked by augmenting some perturbations, particularly for the “hard” samples. As another example, after a certain epoch in training, a deep learning model may start to memorize these “hard” training samples with noisy labels, leading to a degeneration of test robustness.

[0023] To address this identified problem, the presented method for mitigating robust overfitting proposes to fit relatively “easy” samples with one-hot labels and assign a low predictive confidence for “hard” samples in adversarially training. The presented method may leverage all training samples to suffice the need of sample complexity for adversarially training and hinder deep learning models from excessive memorization of one-hot labels in adversarially training.

[0024] **FIG. 2** illustrates an example flow diagram of method 200 according to one or more aspects of the present disclosure. Method 200 may be performed by a deep learning model. At block 210, a plurality of samples and a plurality of labels corresponding to the plurality of samples may be received by the deep learning model. In one aspect of the present disclosure, the plurality of samples and the plurality of

labels corresponding to the plurality of samples may comprise a training dataset.

[0025] At block 220, the plurality of samples may be adversarially augmented by the deep learning model based on a threat model. In one aspect of the present disclosure, the deep learning model may comprise an adversarial defense method or model in deep learning under black-box attacks or white-box attacks.

[0026] At block 230, a low predictive confidence may be assigned by the deep learning model to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model. In one aspect of the present disclosure, after being adversarially augmented, “hard” samples may be assigned a low predictive confidence than “easy” samples.

[0027] As an example, the plurality of labels may comprise one-hot labels. As another example, the deep learning model may comprise one or more of projected gradient descent adversarial training (PGD-AT), or TRADES. As a further example, the threat model may comprise one or more of ℓ_2 -norm threat model, or ℓ_∞ -norm threat model.

[0028] It will be appreciated by those skilled that other models or labels may be possible, and one or more additional steps may be included in the method 200.

[0029] **FIG. 3** shows a conceptual illustration of decision boundaries. In order to illustrate the identified cause of robust overfitting and influences of “hard” samples and “easy” samples on adversarially training, the present disclosure proposes to consider models trained on random labels, in addition to true labels as usual. It is shown in **FIG. 3** decision boundaries learned via standard training and adversarially training with true and random labels, respectively. As shown in **FIG. 3**, there may be a significantly more complicated decision boundary when training samples with random labels.

[0030] In order to illustrate potential behaviors of defense models under extreme condition, such as random labels, adversarially training is performing by PGD-AT and TRADES with random labels sampled uniformly over all classes. **FIG. 4A**, **FIG. 4B**, and **FIG. 4C** illustrate adversarially training with random labels on CIFAR-10. **FIG. 4A** and **FIG. 4B** show the natural and robust training accuracies of PGD-AT and TRADES, respectively, when trained on true or random labels, with explicit regularizations turned off. For both methods, the network achieves almost 100% natural and robust training accuracies when trained on true labels. When the labels are random, totally different behaviors are observed between PGD-AT and TRADES — PGD-AT fails to converge while TRADES still reaches nearly 100% training accuracies. **FIG. 4C** shows the generalization gap under varying levels of label noise

from 0% (true labels) to 100% (completely random labels).

[0031] As shown in **FIG. 4A** and **FIG. 4B**, PGD-AT fails to converge and TRADES can converge under this circumstance. The latter may demonstrate that deep learning networks may have sufficient capacity to memorize adversarial samples of training dataset with completely random labels. Furthermore, the convergence issue of PGD-AT when trained on random labels may result from the intrinsic algorithm, i.e., adversarial loss function of PGD-AT.

[0032] PGD-AT formulates adversarial training as the following robust optimization problem:

$$\min_{\theta} \sum_{i=1}^n \max_{x'_i \in S(x_i)} \mathcal{L}(f_{\theta}(x'_i), y_i) \quad \text{Eq. (1)}$$

Where let $D = \{(x_i, y_i)\}_{i=1}^n$ denote a training dataset with n samples, where $x_i \in R^d$ is a natural sample and $y_i \in \{1, \dots, C\}$ is its true label often encoded as an one-hot vector 1_{y_i} with totally C classes. Where f_{θ} may be a DNN classifier with parameters θ that predicts probabilities over all classes, $\mathcal{L}$ is the classification loss (e.g., the cross-entropy loss as $\mathcal{L}(f_{\theta}(x), y) = -1_y^T \log f_{\theta}(x)$), and $S(x) = \{x': \|x' - x\|_p \leq \epsilon\}$ is an adversarial region centered at x with radius $\epsilon > 0$ under the ℓ_p -norm threat models (e.g., ℓ_2 and ℓ_{∞} norms).

[0033] TRADES formulates adversarial training by minimizing a different adversarial loss:

$$\min_{\theta} \sum_{i=1}^n \mathcal{L}(f_{\theta}(x_i), y_i) + \beta \cdot \max_{x'_i \in S(x_i)} \mathcal{D}(f_{\theta}(x_i) || f_{\theta}(x'_i)) \quad \text{Eq. (2)}$$

Where $\mathcal{L}$ is the classification loss (e.g., the cross-entropy loss), $\mathcal{D}$ is the Kullback–Leibler divergence, and β is a balancing hyperparameter.

[0034] In one aspect of the present disclosure, it is identified that TRADES in Eq. (2) may minimize a clean cross-entropy loss on natural samples, which may make DNNs memorize natural samples with random labels at first, before fitting adversarial samples. It can be seen from **FIG. 4B** that at the very early stage of TRADES training (e.g., the first 25 epochs), the natural accuracy starts to increase while the robust

accuracy does not. But differently, PGD-AT in Eq. (1) directly minimizes the cross-entropy loss on adversarial samples with random labels, which could introduce unstable gradients with large variance, making it fail to converge.

[0035] Based on the analysis, the cross-entropy loss on natural samples may be added into the PGD-AT objective (e.g., Eq. (1)) to resemble the learning of TRADES with random labels, which may be written as:

$$\min_{\theta} \sum_{i=1}^n (1 - \gamma) \cdot \mathcal{L}(f_{\theta}(x_i), y_i) + \gamma \cdot \max_{x'_i \in \mathcal{S}(x_i)} \mathcal{L}(f_{\theta}(x'_i), y_i) \quad \text{Eq. (3)}$$

Where γ is gradually increased from 0 to 1.

[0036] By using the improved adversarial loss (i.e., Eq. (3)), the training on random labels can successfully converge. The improvement from Eq. (1) (failing to converge) to Eq. (3) (able to converge) may indicate that it may be beneficial to fit relatively “easy” samples with one-hot labels and assign a low predictive confidence for “hard” samples in adversarially training. In one aspect of the present disclosure, as shown in **FIG. 3**, in the two images of right column, relative to the natural samples (for example, as indicated by 310), the corresponding adversarially augmented samples (for example, as indicated by 320) may become closer to or even cross the decision boundary due to the attacking, which may result in noisy labels. Therefore, memorizing natural samples (that may be easier relative to corresponding adversarially augmented samples) with random labels at first, before fitting adversarial samples, as identified in Eq. (2) and utilized in Eq. (3), can improve the adversarially training even under extreme condition, such as random labels.

[0037] In one aspect of the present disclosure, the present disclosure identifies that several typical methods training models on the identified clean samples may be not suitable for adversarially training. Because these methods will neglect a portion of training data with noisy labels, which can lead to inferior results for adversarially training due to the reduction of sample complexity. In another aspect of the present disclosure, the present disclosure proposes to regularize the predictions of adversarial samples from being over-confident by integrating the temporal ensembling (TE) approach into the adversarial training frameworks. TE maintains an ensemble prediction of each data and penalizes the difference between the current prediction and the ensemble prediction. In further aspect of the present disclosure, the present disclosure identifies that TE is suitable for adversarially training since it enables to leverage all training samples and hinders the network from excessive memorization of one-hot labels with a regularization term.

[0038] In an example, by adding a regularization term of TE, the training objective of PGD-AT may be rewritten as:

$$\min_{\theta} \sum_{i=1}^n \max_{x'_i \in S(x_i)} \zeta(f_{\theta}(x'_i), y_i) + \omega \cdot \|f_{\theta}(x'_i) - \hat{P}_i\|_2^2$$

Eq. (4)

Where ensemble prediction of a training sample x_i is denoted as P_i , which is updated after each training epoch as $P_i \leftarrow \eta \cdot P_i + (1 - \eta) \cdot f_{\theta}(x_i)$, where η is the momentum term. Wherein $\hat{P}_i$ is the normalization of P_i as a probability vector and ω is a balancing weight.

[0039] In another example, TE may be similarly integrated with TRADES with the same regularization term. With the regularization term, the deep learning model would learn to fit relatively “easy” samples with one-hot labels and assign low confidence for “hard” samples at initial. After the learning rate decays, the regularization term may avoid fitting one-hot labels for the “hard” samples. Therefore, the proposed algorithm may enable to learn under label noise in adversarially training and alleviate the robust overfitting problem.

[0040] It will be appreciated by those skilled that other algorithms may be possible to perform the method 200 described above with reference to **FIG. 2**.

[0041] To evaluate the performance of the proposed algorithm, experiment results are listed in Table 1. It is reported in Table 1 the test accuracy on the best checkpoint obtaining the highest robust accuracy under PGD-10 and the final checkpoint, as well as the difference between these two checkpoints. Besides PGD-10, PGD-1000, C&W-1000, AutoAttack and SPSA also be adopted for a more rigorous robustness evaluation. It can be observed that the difference between best and final test accuracies is reduce to around 1%. Due to the mitigation of robust overfitting, the proposed method or algorithm may achieve better robustness.

Table 1. Test accuracy (%) of several methods on CIFAR-10 under the ℓ_{∞} norm with $\epsilon = 8/255$ based on the ResNet-18 architecture.

Methods	Natural Accuracy			PGD-10			PGD-1000			C&W-1000			AutoAttack			SPSA		
	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff
PGD-AT	83.75	84.82	-1.07	52.64	44.92	7.72	51.22	43.74	8.48	50.11	43.63	7.48	47.74	41.84	5.90	53.51	47.48	6.03
PGD-AT+TE	82.35	82.79	-0.44	55.79	54.83	0.96	54.65	53.30	1.35	52.30	51.73	0.57	50.59	49.62	0.97	55.11	54.87	0.24
TRADES	81.19	82.48	-1.29	53.32	50.25	3.07	52.44	48.67	3.77	49.88	48.14	1.74	49.03	46.80	2.23	52.99	51.22	1.77
TRADES+TE	83.86	83.97	-0.11	55.15	54.42	0.73	53.74	53.03	0.71	50.77	50.63	0.14	49.77	49.20	0.57	54.42	54.14	0.28

[0042] **FIG. 5** illustrates an example of a hardware implementation for an apparatus 500 according to one or more aspects of the present disclosure. The apparatus 500 for deep learning may comprise a memory 510 and at least one processor 520. The processor 520 may be coupled to the memory 510 and configured

to perform the method 200 described above with reference to **FIG. 2** and one or more aspects of the present disclosure. The processor 520 may be a general-purpose processor, or may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, multiple microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration. The memory 510 may store the input data, output data, data generated by processor 520, and/or instructions executed by processor 520.

[0043] The various operations, models, and networks described in connection with the disclosure herein may be implemented in hardware, software executed by a processor, firmware, or any combination thereof. According one or more aspects of the disclosure, a computer program product for deep learning may comprise processor executable computer code for performing the method 200 described above with reference to **FIG. 2** and one or more aspects of the present disclosure. According to one or more aspects of the disclosure, a computer readable medium may store computer code for deep learning, the computer code when executed by a processor may cause the processor to perform the method 200 described above with reference to **FIG. 2** and one or more aspects of the present disclosure. Computer-readable media includes both non-transitory computer storage media and communication media including any medium that facilitates transfer of a computer program from one place to another. Any connection may be properly termed as a computer-readable medium. Other embodiments and implementations are within the scope of the disclosure.

[0044] The method described in the present disclosure may be implemented by software, hardware, firmware, or any combination thereof, and may provide a machine learning model that is configured to perform one or more particular machine learning tasks. In one example, the machine learning tasks may include speech and visual recognition tasks. In another example, the machine learning tasks may be an agent control task carried out in a control system for automatic driving, a control system for an industrial facility, or the like.

[0045] The preceding description of the disclosed embodiments is provided to enable any person skilled in the art to make or use the various embodiments. Various modifications to these embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments without departing from the scope of the various embodiments. Thus, the claims are not intended to be limited to the embodiments shown herein but is to be accorded the widest scope consistent with the following claims and the principles and novel features disclosed herein.

WHAT IS CLAIMED IS:

1. A method for deep learning, comprising:
receiving, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples;
adversarially augmenting, by the deep learning model, the plurality of samples based on a threat model; and
assigning, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.
2. The method of claim 1, wherein the plurality of labels comprising one-hot labels.
3. The method of claim 2, wherein the one or more adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model are located closely to a decision boundary.
4. The method of claim 1, wherein the assigning the low predictive confidence to the one or more adversarially augmented samples having noisy labels is performed by regularizing predictions of the plurality of adversarially augmented samples via temporal ensembling (TE).
5. The method of claim 4, wherein the deep learning model comprising one or more of projected gradient descent adversarial training (PGD-AT), or TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES).
6. The method of claim 1, wherein the threat model comprising one or more of ℓ_2 -norm threat model, or ℓ_∞ -norm threat model.
7. An apparatus for deep learning, comprising:
a memory; and
at least one processor coupled to the memory and configured to perform the method of one of claims 1-6.

8. A computer program product for deep learning, comprising: processor executable computer code for performing the method of one of claims 1-6.

9. A computer readable medium, storing computer code for deep learning, the computer code when executed by a processor, causing the processor to perform the method of one of claims 1-6.

10. A method for visual recognition, comprising:
receiving, by a deep learning model, a plurality of samples and a plurality of labels corresponding to the plurality of samples, wherein the deep learning model is configured to perform a visual recognition task and the plurality of samples comprising a plurality of images;
adversarially augmenting, by the deep learning model, the plurality of samples based on a threat model; and
assigning, by the deep learning model, a low predictive confidence to one or more adversarially augmented samples of the plurality of adversarially augmented samples having noisy labels due to the adversarially augmenting based on the threat model.

11. The method of claim 10, wherein the plurality of samples and the plurality of labels corresponding to the plurality of samples comprising one or more of CIFAR-10 training dataset, or MNIST training dataset.

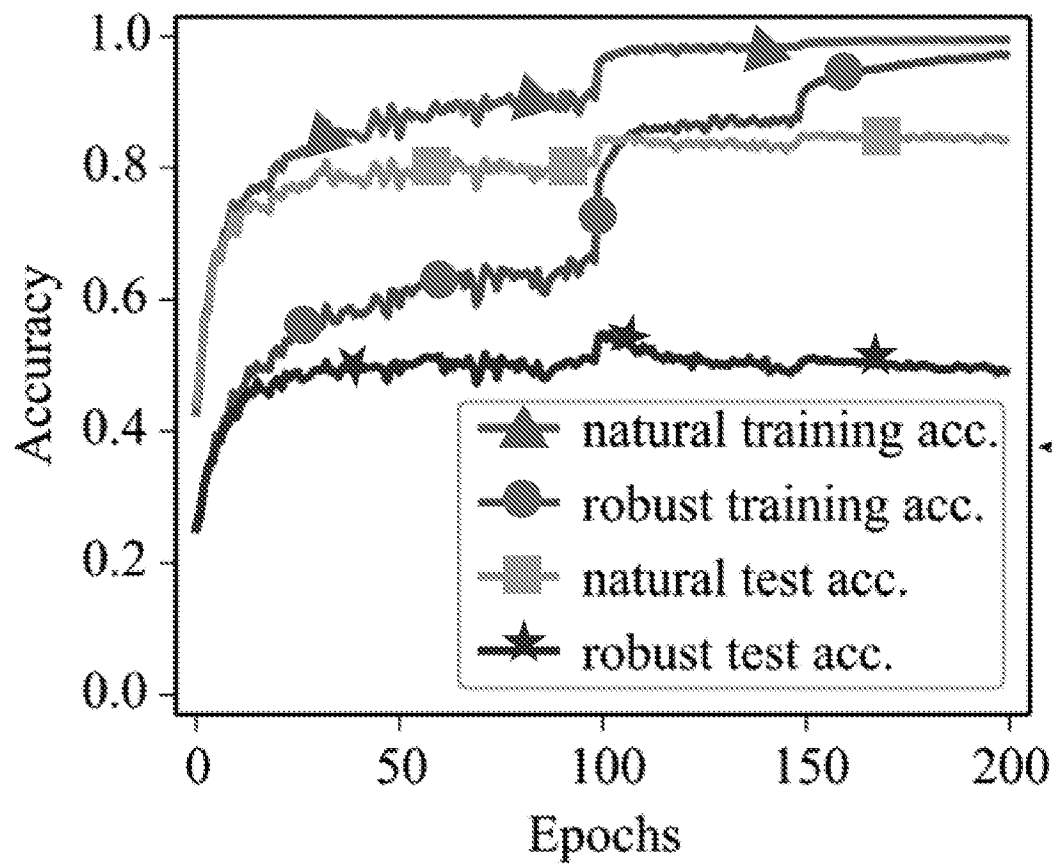


FIG. 1

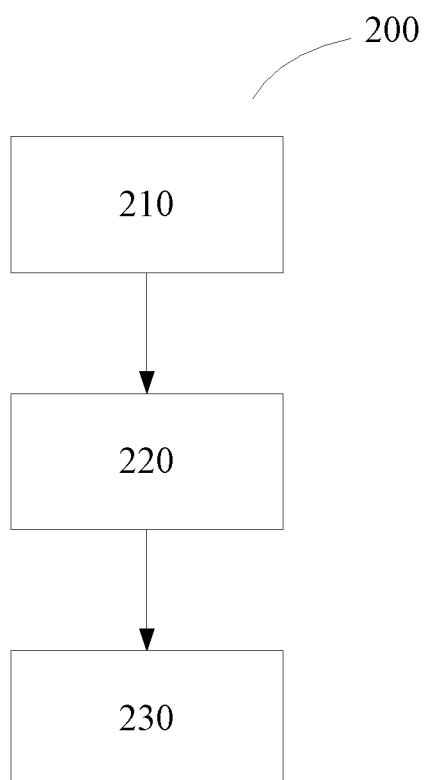


FIG. 2

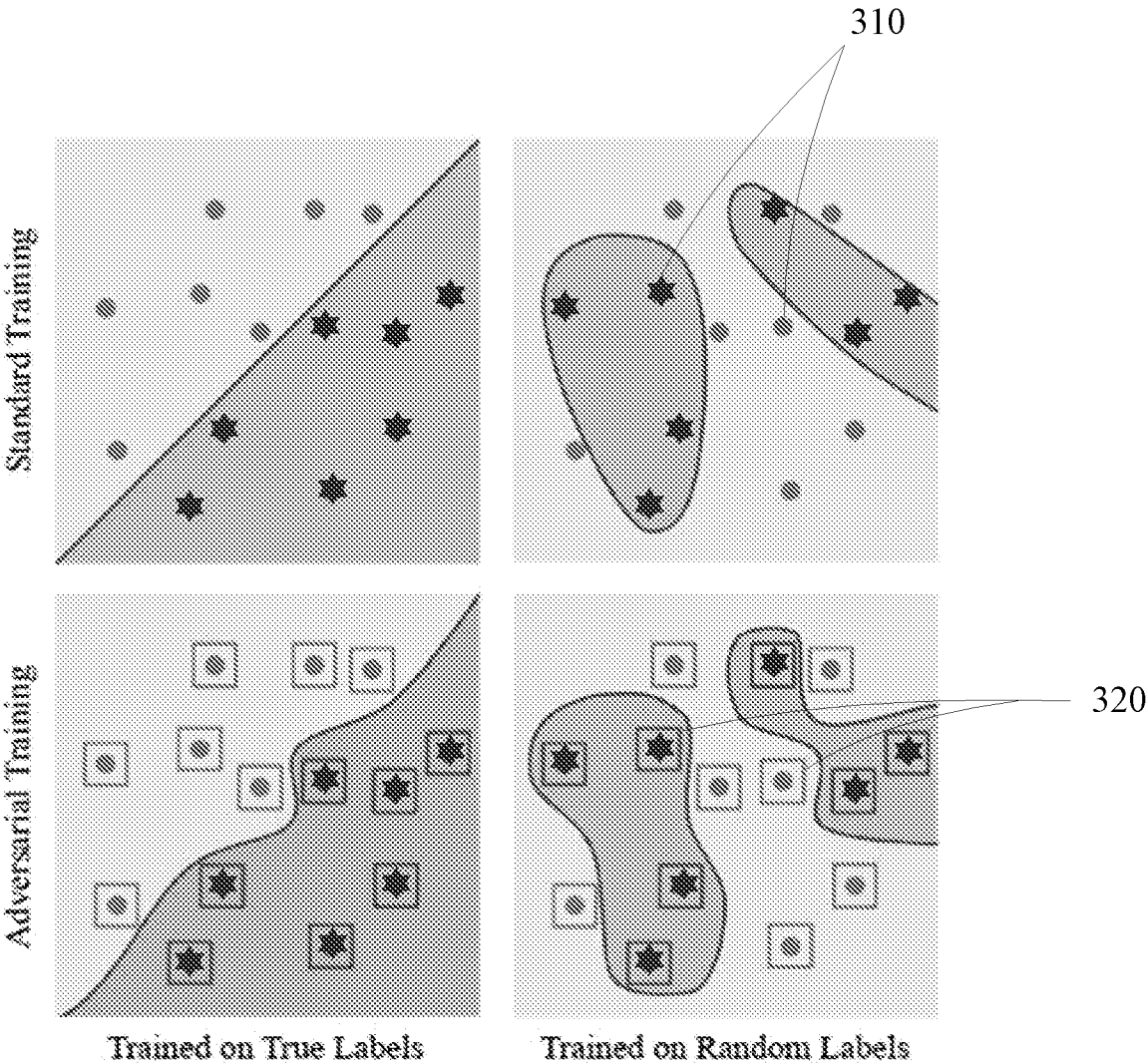


FIG. 3

4/5

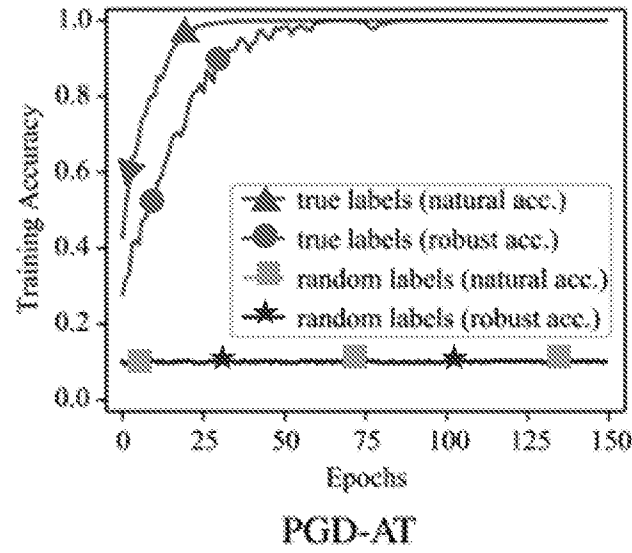


FIG. 4A

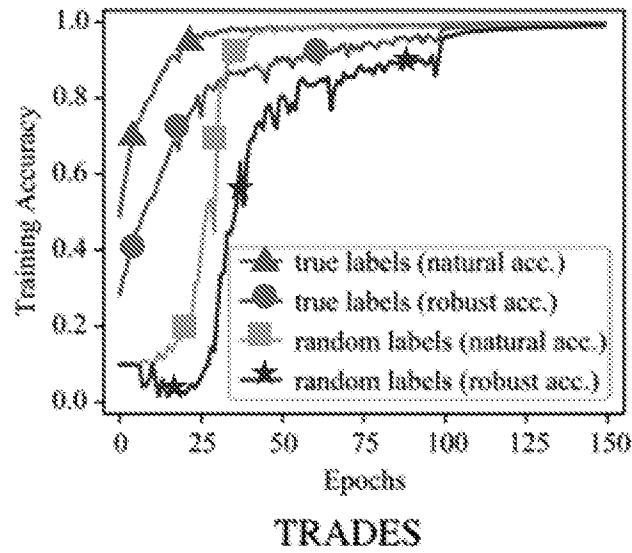


FIG. 4B

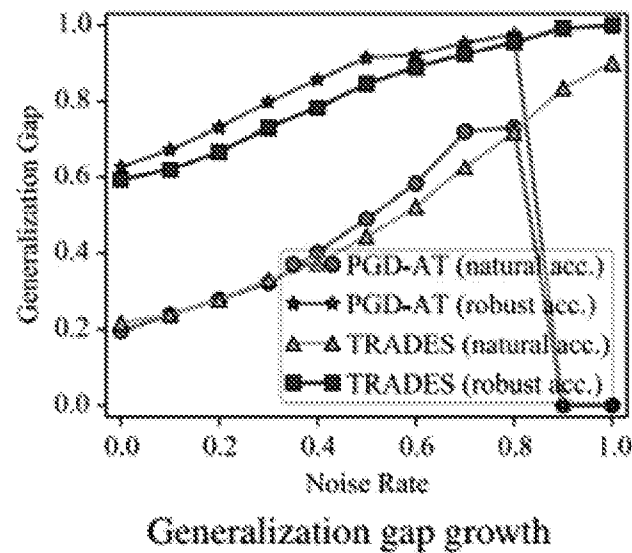


FIG. 4C

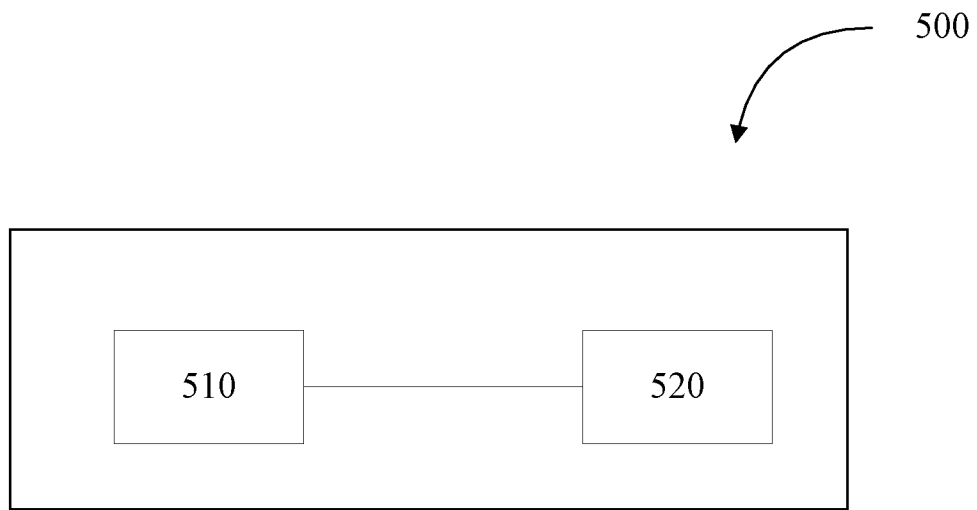


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/CN2021/097267

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N3/08 G06N7/00
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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TAHERI SHAYAN ET AL: "RazorNet: Adversarial Training and Noise Training on a Deep Neural Network Fooled by a Shallow Neural Network", BIG DATA AND COGNITIVE COMPUTING, vol. 3, no. 3, 1 September 2019 (2019-09-01), page 43, XP055776488, DOI: 10.3390/bdcc3030043 page 10 - page 13; figure 5; table 3 ----- -/--	1-11



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

3 November 2021

Date of mailing of the international search report

17/11/2021

Name and mailing address of the ISA/

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

Authorized officer

Falco, Gabriele

INTERNATIONAL SEARCH REPORT

International application No
PCT/CN2021/097267

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BO HAN ET AL: "A Survey of Label-noise Representation Learning: Past, Present and Future", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 9 November 2020 (2020-11-09), XP081809994, page 10, paragraph 5 - page 11 -----	1-11
A	LESLIE RICE ET AL: "Overfitting in adversarially robust deep learning", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 26 February 2020 (2020-02-26), XP081608597, page 1 - page 2 -----	6