Improving Model Robustness by Adaptively Correcting Perturbation Levels with Active Queries

Kun-Peng Ning*, Lue Tao*, Songcan Chen, Sheng-Jun Huang†
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 211106

Abstract

In addition to high accuracy, robustness is becoming increasingly important for machine learning models in various applications. Recently, much research has been devoted to improving the model robustness by training with noise perturbations. Most existing studies assume a fixed perturbation level for all training examples, which however hardly holds in real tasks. In fact, excessive perturbations may destroy the discriminative content of an example, while deficient perturbations may fail to provide helpful information for improving the robustness. Motivated by this observation, we propose to adaptively adjust the perturbation levels for each example in the training process. Specifically, a novel active learning framework is proposed to allow the model to interactively query the correct perturbation level from human experts. By designing a cost-effective sampling strategy along with a new query type, the robustness can be significantly improved with a few queries. Both theoretical analysis and experimental studies validate the effectiveness of the proposed approach.

Introduction

Deep Neural Networks (DNNs) have achieved great success in many tasks with high accuracy (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2015; Sutskever, Vinyals, and Le 2014; Silver et al. 2017). On the other hand, deep models are less robust when applied to datasets with noise perturbations (Szegedy et al. 2013; Alzantot et al. 2018; Hendrycks and Dietterich 2019). Much research has been devoted to mitigating this issue in recent years (Geirhos et al. 2018; Rusak et al. 2020; Tramer et al. 2020; Hendrycks et al. 2019a; Mao et al. 2019; Hendrycks et al. 2019b; Zhang 2019). Roughly speaking, existing studies are trying to improve the model robustness by handling two different categories of perturbations. One is *adversarial perturbations*, which are maliciously designed to fool the models under some distance constraint (e.g. ℓ_∞ distance (Madry et al. 2017) or Wasserstein distance (Wong, Schmidt, and Kolter 2019)), while the appearance contents are preserved. The other one is *corruption perturbations*, which are usually incidentally generated during the process of data collection and editing (e.g. Gaussian noise (Chapelle et al. 2001), motion blur (Hendrycks and Dietterich 2019)).

In this paper, we focus on the latter case and try to handle corruption perturbations for improving the model robustness. Corruption perturbation problem is becoming a ubiquitous challenge in various applications (Hendrycks and Dietterich 2019; Michaelis et al. 2019). More and more systems based on deep learning have been deployed to real-world applications. They are typically trained and evaluated in laboratory environments. However, extensive and unexpected noises exist in real environments, which may cause serious failures if the models are not robust enough. For example, autonomous vehicles need to be able to cope with wildly varying outdoor conditions such as fog, frost, snow, sand storms, or falling leaves (Michaelis et al. 2019). Likewise, speech recognition systems should perform well regardless of the additive noise or convolutional distortions (Qian et al. 2016).

Training DNNs on perturbed examples is the primal approach to improve the model robustness (Carlini et al. 2019; Hendrycks et al. 2019b). Representative methods include noise injection (Grandvalet, Canu, and Boucheron 1997) and PGD-based robust training (Madry et al. 2017). However, most of the existing methods assign a fixed level of perturbations (e.g. fixed radius in ℓ_p norm-bounded perturbations or bandwidth in Gaussian noise) to all examples, ignoring the fact that each example has its own intrinsic tolerance to noises. In fact, excessive perturbations would destroy the class-distinguishing feature of an example, while deficient perturbations may fail to provide helpful information for improving the robustness. Intuitively, some examples are closer to the decision boundary, where tiny perturbations could change their labels, while some others are far away from the decision boundary and may tolerate higher levels of perturbation. As shown in Figure 1, under the same perturbation, the discriminability of the corrupted images is significantly different, if the original images have different intrinsic robustness. A higher-quality image is likely to be able to tolerate heavier perturbations.

Several recent works in the literature seek to adjust the perturbation levels for different examples according to prediction loss (Cheng et al. 2020; Sitawarin, Chakraborty, and Wagner 2020; Zhang et al. 2020). While it is intuitively reasonable to assign higher perturbations to examples with smaller losses,

*Equal contribution.

†Correspondence to: Sheng-Jun Huang (huangsj@nuaa.edu.cn).

Figure 1: The influence of the same perturbation ($\mathcal{N}(0, \sigma^2 I)$, where $\sigma = 0.23$) on images with different intrinsic robustness. The three perturbed images are generated by corrupting the three original images respectively with the same perturbation.

these methods may suffer from model bias and lack a correction strategy to seek help from ground-truth supervision. In the case that the model is unreliable, the perturbation adjustment strategy may be seriously misled to hurt model performance. It is thus rather important to allow the learning system to query the ground-truth information about the perturbation levels for robustness. Such an idea has been widely used in other machine learning tasks. For example, in the active learning literature, learning algorithms are allowed to query class labels from human experts to improve model accuracy (Settles 2009). Actually, the human recognition system is remarkably robust against a wide range of noises and corruptions (Rusak et al. 2020), and thus can identify the proper tolerance level of perturbations for a specific image. This motivates us to query the perturbation levels from human experts to train a robust model.

In this paper, we propose to adaptively adjust the perturbation levels for each training example, along with a querying strategy to get ground-truth information from the human experts to correct the perturbation levels. It is worth to note that the human annotation could be costly, and thus it is less practical to query the perturbation levels for all examples. To overcome this challenge, we propose a novel active learning framework to **A**ctively **Q**uery **P**erturbation **L**evels (**AQPL** for short), aiming to train a robust model with least queries. Specifically, at each iteration of active learning, we first estimate the conformity of the current perturbation level for each example based on the prediction consistency over multiple generated noises, and then actively select the examples with the least conformity for querying. In this way, the examples with over large or over small perturbations will be corrected with the queried ground-truth information. To further reduce the annotation cost, a cost-effective query type is designed to allow human experts to easily decide the perturbation level for an image.

Experiments are conducted on multiple datasets with variant noise perturbations. Our results validate the effectiveness of the proposed AQPL method with adaptive correction of perturbation levels. Model robustness, as well as accuracy, are significantly improved by actively querying a very few times with low cost.

The main contributions of this work are summarized as follows.

- A novel framework AQPL is proposed to improve model robustness via querying the perturbation level of examples. It is a new attempt to improve the model robustness by interacting with human experts.
- An effective strategy is proposed to actively select the most useful example for perturbation level correction, which significantly reduces the query numbers for robust training.
- A cost-effective query type is designed to allow human experts to easily decide the proper perturbation level of an image with low annotation cost.

The rest of the paper is organized as follows. We review related work in Section 2 and introduce the proposed method in Section 3. Section 4 reports the experiments, followed by the conclusion in Section 5.

Related Work

Model robustness. Improving model robustness refers to the goal of ensuring the performance of machine learning models under a variety of imperfect testing conditions, which has a long history. Double backpropagation algorithm (Drucker and Le Cun 1991) is one of the earliest attempt to make models resistant to local minimal perturbations by regularizing input gradients. It reaches a consensus with Tikhonov regularization (Tikhonov and Arsenin 1977), which is equivalent to training with noise under the assumption that the noise amplitude is small enough (Bishop 1995). Nevertheless, a more practical and ubiquitous situation in the real-world is some nonnegligible noises that exist in input data that we can easily tell, such as fog, rain, or falling leaves in camera data under varying outdoor conditions. Such a situation considering perturbations in input data is known as vicinal risk minimization (Chapelle et al. 2001), along with adversarial risk minimization (Uesato et al. 2018), sparks a surge of interest in model robustness recently. To this end, a number of methods have been proposed to mitigate the problem and improve the robustness of DNNs, among which finally training with perturbations remains the most effective one (Carlini et al. 2019; Hendrycks et al. 2019b; Rusak et al. 2020). To achieve this goal, current methods assume a fixed perturbation level for all training examples (Madry et al. 2017; Wang et al. 2020; Zhang et al. 2019; Rusak et al. 2020; Cemgil et al. 2020), which hardly holds in real tasks. Most recently, (Cheng et al. 2020; Sitawarin, Chakraborty, and Wagner 2020; Zhang et al. 2020) propose to adaptively adjust the perturbation level for each example according to the capacity of the model-on-training. They adjust the perturbation level of each example to cater to the model-on-training, while we choose to define the intrinsic robustness of examples by quering the oracle. Since model vulnerability can be view as a purely human-centric phenomenon, and achieving models

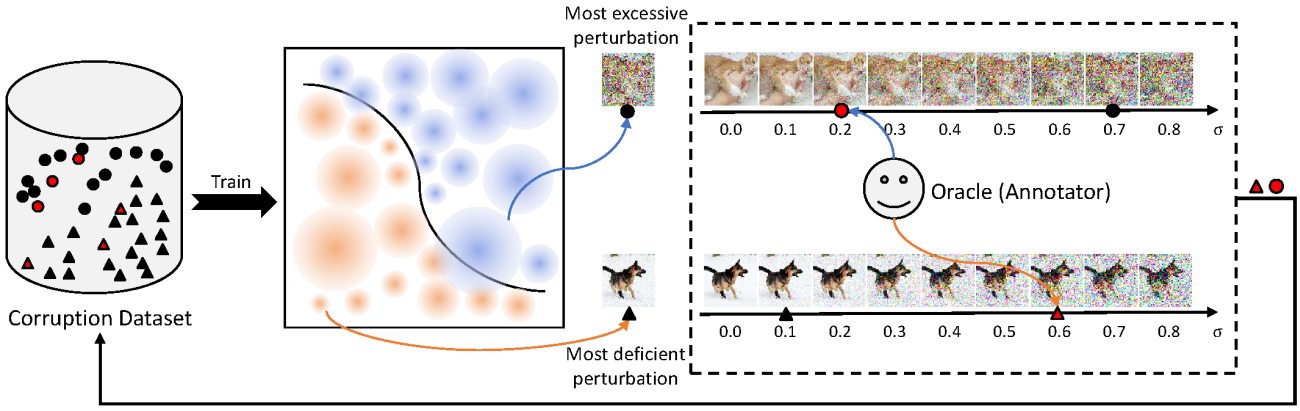


Figure 2: The proposed AQPL framework. Based on the model trained on the corruption dataset, two examples with most excessive and most deficient perturbations are selected for querying. Then, the perturbation levels (indicated by black marker) are corrected to the proper perturbation levels (indicated by red marker) by the annotator. After that, the corruption dataset and the model are updated.

that are robust and interpretable will require explicitly encoding human priors into the training process (Ilyas et al. 2019). In this paper, while we focus on improving model robustness to vicinal risk by adaptively correcting perturbation levels with active queries, our active learning framework can also be extended to tackle the adversarial risk minimization problem.

Active learning. Active learning has achieved a great success for learning with limited labeled data. Most researches focus on designing effective sampling strategies to make sure that the selected examples can improve the model performance most (Fu, Zhu, and Li 2013). During the past decades, many criteria have been proposed for selecting examples (Fu, Zhu, and Li 2013; Huang, Jin, and Zhou 2010; Lewis and Gale 1994; Seung, Opper, and Sompolinsky 1992; You, Wang, and Tao 2014; Geman, Bienenstock, and Doursat 1992; Roy and McCallum 2001). Among of these approaches, some of them prefer to select the most informative examples to reduce the model uncertainty (Lewis and Gale 1994; Seung, Opper, and Sompolinsky 1992; You, Wang, and Tao 2014), while some others prefer to select the most representative examples to match the data distribution (Geman, Bienenstock, and Doursat 1992; Roy and McCallum 2001). Moreover, some studies try to combine informativeness and representativeness to achieve better performance (Huang and Zhou 2013; Huang, Jin, and Zhou 2010). Standard active learning methods often ask the oracle to annotate data examples (Fu, Zhu, and Li 2013), (Huijser and van Gemert 2017) tries to improve the classification model by asking for annotations of decision boundary. Similarly, our approach attempts to improve the model robustness by querying for annotations of perturbation level.

The Proposed Approach

In this section, we first formalize the framework for improving model robustness via active querying, and then introduce the proposed AQPL approach in detail, followed by the theoretical analysis on the active selection strategy.

Problem Setting

We denote by $\mathcal{D}$ the clean dataset with n examples, i.e., $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{1, ..., K\} =: \mathcal{Y}$ is the ground-truth label. We also denote by $\mathcal{C}$ the dataset with common corruptions (e.g. ImageNet-C), i.e., $\mathcal{C} = \{(\hat{\mathbf{x}}_1, y_1), (\hat{\mathbf{x}}_2, y_2), ..., (\hat{\mathbf{x}}_n, y_n)\}$, where $\hat{\mathbf{x}}_i$ is the corrupted instance of $\mathbf{x}_i$ with perturbations.

A model $F_\theta(\mathbf{x}) : \mathbb{R}^d \to \mathcal{Y}$ parameterized by θ can be trained with the clean dataset $\mathcal{D}$, which however is usually less robust when applied to $\mathcal{C}$ due to the unseen corruptions. To address this issue, the mainstream methods try to train models with noise to improve the robustness against corruption perturbations. Formally, as done in (Chapelle et al. 2001; Gilmer et al. 2019; Rusak et al. 2020), we can improve the classifier F_θ by minimizing the cross-entropy loss $\mathcal{L}$ on clean dataset $\mathcal{D}$ with additive noise:

$$\min_\theta \sum_{i=1}^{n} \left[\mathbb{E}_{\boldsymbol{\varepsilon}\sim\mathcal{P}(\sigma)} \left[\mathcal{L}\left(F_\theta\left(\mathbf{x}_i + \boldsymbol{\varepsilon}\right), y_i\right)\right]\right], \tag{1}$$

where $\boldsymbol{\varepsilon}$ is the random noise generated according to the noise distribution $\mathcal{P}(\sigma)$, and σ is the perturbation level controlling the intensity of the noise. Here $\mathcal{P}(\sigma)$ can be any general noise distribution. Obviously, by minimizing the loss function, the classifier F_θ will be optimized to correctly recognize the examples perturbed by noise. In previous methods (Madry et al. 2017; Rusak et al. 2020; Gilmer et al. 2019), σ is either kept fixed or chosen uniformly from a fixed set of standard deviations. However, as discussion above, it is impractical to set a global constant σ for all examples, because each example has its own intrinsic robustness towards noises. Therefore, in this paper, we propose a more practical setting where each example has its own perturbation level. Formally, we introduce instance-dependent perturbation level σ_i to generate noises for each x_i, and define a new loss function as follows:

$$\min_\theta \sum_{i=1}^{n} \left[\mathbb{E}_{\boldsymbol{\varepsilon}_i\sim\mathcal{P}(\sigma_i)} \left[\mathcal{L}\left(F_\theta\left(\mathbf{x}_i + \boldsymbol{\varepsilon}_i\right), y_i\right)\right]\right]. \tag{2}$$

Obviously, image with stronger intrinsic robustness should receive higher values of σ_i, while image with weaker intrinsic robustness should receive lower values of σ_i. While the initialized perturbation levels are likely not to conform with the intrinsic robustness, we actively select the most useful examples and query their ground-truth information to adaptively correct the perturbation levels.

Algorithm Detail

The proposed framework AQPL is demonstrated in Figure 2. Firstly, all examples are assigned an initial perturbation level. Then at each iteration, based on the proposed conformity criterion, the two most useful examples (one with most excessive perturbation, and one with most deficient perturbation) are selected for perturbation level correction. After that, the oracle is asked to annotate a proper perturbation level that conform with the intrinsic robustness for the selected examples. Based on the queried information, the classification model will be updated, which can improve the robustness of the model as much as possible at a lower cost.

Formally, we define a triplet $(\mathbf{x}, y, \sigma)$ for each training example, which consists of the feature instance, the label and the instance-dependent perturbation level. Then the triplet dataset T with n examples is defined as follows:

$$T \coloneqq \{(\mathbf{x}_1, y_1, \sigma_1), (\mathbf{x}_2, y_2, \sigma_2), ..., (\mathbf{x}_n, y_n, \sigma_n)\}. \quad (3)$$

Next, we will discuss how to select the most useful examples from T to query the perturbation level. As discussed before, neither excessive nor deficient perturbation is helpful to improve the model robustness. If an example falls into these cases, then its perturbation should be corrected to a proper level to conform with its intrinsic robustness. Therefore, given a triplet $(\mathbf{x}, y, \sigma)$, the conformity $s(\sigma)$ of the perturbation to an example $\mathbf{x}$ can be defined as the perturbation level change before and after querying:

$$s(\sigma) \coloneqq \sigma - \sigma_o, \quad (4)$$

where σ_o is the optimal perturbation level of $\mathbf{x}$, which corresponds to the maximum perturbation that the oracle can bear to identify the semantic contents of an image. Intuitively, the larger difference between the current perturbation level and the optimal perturbation level, the more helpful information it may gain with the correction. This motivates us to select the examples that are least conform with its intrinsic robustness.

However, we cannot get the optimal level σ_o before active queries. That is why we have to find a surrogate of the conformity $s(\sigma)$. Inspired by randomized smoothing (Cohen, Rosenfeld, and Kolter 2019), we define the classification entropy to estimate the conformity of the perturbation level for an example. Specifically, for an example $\mathbf{x}$, we firstly generate M noise instances with additive Gaussian noise $\mathcal{N}(0, \sigma^2 I)$. And then, the current classifier F will predict the classes of these M noise examples. Intuitively, if the M predictions are highly consistent (with small entropy), then it implies that the example $\mathbf{x}$ has deficient perturbation. On the other hand, if the M predictions are inconsistent (with large entropy), then it is likely that $\mathbf{x}$ received excessive perturbation currently, and thus its perturbation level may need correction from the oracle.

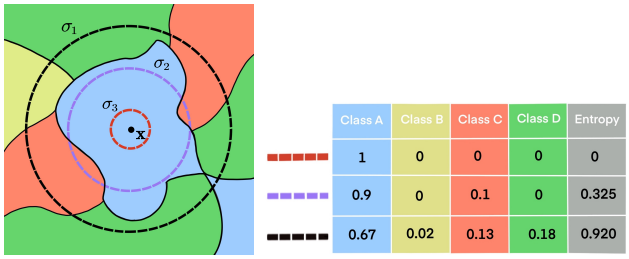


Figure 3: The decision regions of the current classifier F_θ are drawn in different color, and the three circles respectively represent different perturbation levels σ_i of the distributions $\mathcal{N}(\mathbf{x}, \sigma_i^2 I)$, where $\sigma_1 > \sigma_2 > \sigma_3$. The corresponding class distribution and the entropy are shown in the table, which indicates that a larger perturbation leads to a larger entropy.

Formally, suppose that we have a classifier F and an input $\mathbf{x}$, the probability of being classified as class k under perturbations is $p_k \coloneqq \mathbb{P}(F(\mathbf{x} + \boldsymbol{\varepsilon}) = k)$, where $\boldsymbol{\varepsilon} \sim \mathcal{P}(\sigma)$ and the noise distribution $\mathcal{P}(\sigma)$ can be any general noise distribution. Without loss of generality, we choose $\mathcal{P}(\sigma) = \mathcal{N}(0, \sigma^2 I)$ as an example in this paper. Then the classification entropy can be defined as follows:

$$H \coloneqq -\sum_{k=1}^{K} p_k \ln(p_k), \quad (5)$$

where the probability p_k is estimated using Monte Carlo sampling as discussed above.

Figure 3 presents an example to show the relation between the perturbation level and classification entropy. It can be observed that with an excessive perturbation, the classifier (corresponding to the black circle) will produce uncertain predictions with a large entropy, while with a deficient perturbation, the classifier (corresponding to the red circle) will produce consistent predictions with a small entropy. Based on the classification entropy, we then select the two examples with least conformities (one with most excessive perturbation and one with most deficient perturbation) to query their correct perturbation levels from the oracle.

Next we discuss how to let the oracle decide the proper perturbation level for the selected examples. Intuitively, if the corrupted image is difficult for a human annotator to identify its semantic content, then it is likely that the image is suffering from excessive perturbation. For the selected example $\mathbf{x}^*$, we generate a series of noise images from the clean instance by varying the perturbation level from the minimum σ_{min} to the maximum σ_{max} with interval of α. Among which, the oracle is asked to choose the image that is at the threshold of identifying its semantic content. Then the corresponding perturbation level of this image is annotated as the optimal perturbation level for the queried example $\mathbf{x}^*$. This annotation process is illustrated in the dashed rectangle in Figure 2. Among the noise images generated from the selected example, the black marker indicates the one generated with the current perturbation level, while the red marker indicates the optimal perturbation level annotated by the oracle.

Algorithm 1 The AQPL algorithm

Input:
  Query batch size B;
  Triplet dataset T;
  Pretrained model F_θ;
Repeat:
  Generate M noise examples with additive Gaussian noise $\mathcal{N}(0, \sigma_i^2 I)$ for each example $\mathbf{x}_i$.
  Caculate the classification entropy $H(\mathbf{x}_i)$ for each example $\mathbf{x}_i$.
  Select two batches of examples with maximum and minimum entropy.
  Query the most acceptable perturbation level σ^* for selected examples from oracle.
  Update the triplet dataset T and update θ by minimizing Eq 2.
until query budget or expected performance reached.
Output the learned model F_θ.

After the querying, the triplet with corrected perturbation level is added into the training set for updating the model. Moreover, to improve the efficiency of learning, we also query the most acceptable perturbation level for two mini batches of examples with maximum and minimum entropy. At last, we update θ by minimizing Eq 2 until query budget or expected performance reached. Note that to maintain high accuracy on clean data, we only perturb 50% of the training data with Gaussian noise to train the model within each batch, which follows the same settings in (Rusak et al. 2020).

The process of the approach is summarized in Algorithm 1. Firstly, triplet dataset T, query batch size B and pretrained model F_θ are given. Then for each example $\mathbf{x}_i$, we generate K noise examples with additive Gaussian noise $\mathcal{N}(0, \sigma_i^2 I)$ and caculate its classification entropy $H(\mathbf{x}_i)$. We select two batches of examples with maximum and minimum entropy and ask for annotations of perturbation level σ^*. After that, we update the triplet dataset T and update θ by minimizing Eq 2 until query budget or expected performance reached.

Theoretical Analysis

In this subsection, we present theoretical analysis to show the rationality of the proposed classification entropy as a surrogate of the conformity for a simple linear case. Although it doesn't extend easily to the non-linear deep learning based classification, this analysis gives some insights into the behavior of the proposed surrogate, and how this strategy successfully reduces the query number for robust learning.

For the definition of conformity in Eq 4, if we know the optimal perturbation level σ_o, the triplet with largest or smallest value of conformity $s(\sigma)$ will be selected by the proposed algorithm. In other words, the perturbation level that changes the most before and after querying is of our interest. With the proposed method to estimate $s(\sigma)$ by the classification entropy H in Eq 5, we can get the following theorems.

Theorem 1. *Consider the case of one-layer feed-forward network for binary classification $F(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$ and $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. For any $\mathbf{x} \in \{\mathbf{x} : f(\mathbf{x}) \neq 0\}$, suppose that $\mathcal{P}(\sigma) = \mathcal{N}(0, \sigma^2 I)$ and its current perturbation level $\sigma \in (0, \infty)$, then we have*

$$\sigma \propto H. \tag{6}$$

The proof can be found in the supplementary material. Further, if there is an oracle classifier $F_o(\mathbf{x})$, then for the optimal perturbation level σ_o we have

$$\mathbb{P}(F_o(\mathbf{x} + \boldsymbol{\varepsilon}_o) = c) = \tau, \tag{7}$$

where $\boldsymbol{\varepsilon}_o \sim \mathcal{P}(\sigma_o)$ and τ is some sufficient large value (e.g. 99.73% for the empirical rule (Pukelsheim 1994)). Then for any fixed σ_o and τ, we have following theorem.

Theorem 2. *Consider the case of one-layer feed-forward network for binary classification $F(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$ and $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. For every $\mathbf{x} \in \{\mathbf{x} : f(\mathbf{x}) \neq 0\}$ and its corresponding optimal perturbation level σ_o for the oracle classifier $F_o(\mathbf{x}) = \text{sign}(\mathbf{w}_o^T\mathbf{x} + b_o)$, suppose that $\mathcal{P}(\sigma) = \mathcal{N}(0, \sigma^2 I)$. If $\mathbf{w}^T\mathbf{w}_o \neq 0$, then we have*

$$\sigma_o \propto -H. \tag{8}$$

The proof can be found in the supplementary material. Then we can further get the relation between the conformity $s(\sigma) \coloneqq \sigma - \sigma_o$ and the classification entropy H with the following corollary.

Corollary 2.1. *Consider the case of linear binary classifier $F(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$ and $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. For every $\mathbf{x} \in \{\mathbf{x} : f(\mathbf{x}) \neq 0\}$ and its corresponding optimal perturbation level σ_o for the oracle classifier $F_o(\mathbf{x}) = \text{sign}(\mathbf{w}_o^T\mathbf{x} + b_o)$, suppose that $\mathcal{P}(\sigma) = \mathcal{N}(0, \sigma^2 I)$ and its current perturbation level $\sigma \in (0, \infty)$. If $\mathbf{w}^T\mathbf{w}_o \neq 0$, then we have*

$$\sigma - \sigma_o \propto H. \tag{9}$$

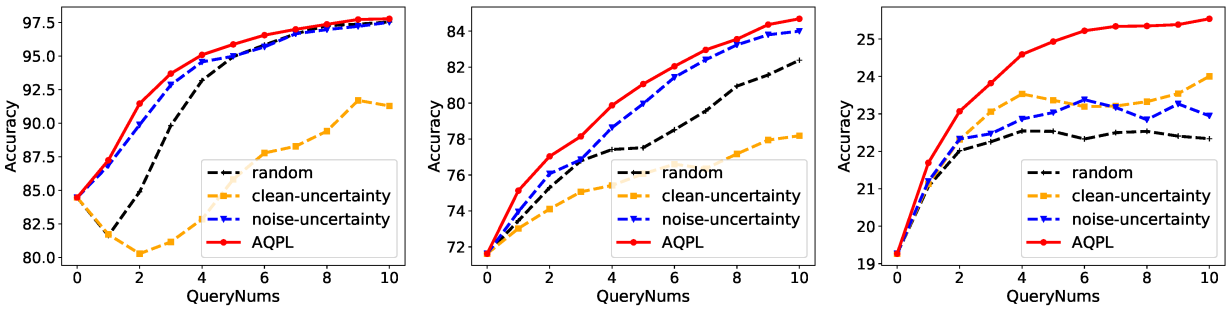
With Corollary 2.1, it can be observed that the conformity of a perturbation level to an example is directly proportional to the classification entropy. In other words, based on the classification entropy H of example $\mathbf{x}$, we can effectively select the examples which are most helpful for improving the robustness after correcting their perturbation levels.

Experiments

Settings

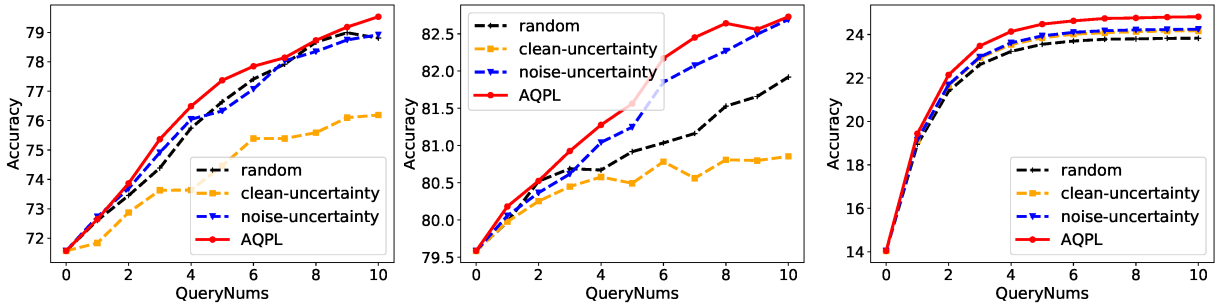
To validate the effectiveness of the proposed approach, we perform experiments on six datasets. Specifically, we train the model on MNIST (LeCun et al. 1998), CIFAR10 (Krizhevsky, Hinton et al. 2009) and Tiny-Imagenet (Yao and Miller 2015), and test on MNIST-C (Mu and Gilmer 2019), CIFAR10-C (Hendrycks and Dietterich 2019), Tiny-Imagenet-C (Hendrycks and Dietterich 2019). We employ ResNet18 model (He et al. 2016) as the base model to implement our approach as well as other compared methods.

We respectively examine the effectiveness of our approach with regard to the active sampling strategy and the querying method. To validate the effectiveness of the sampling



(a) MNIST-C (b) CIFAR10-C (c) Tiny-Imagenet-C

Figure 4: Performance comparison of different methods towards Gaussian noise in corruption datasets.



(a) MNIST-C (b) CIFAR10-C (c) Tiny-Imagenet-C

Figure 5: Average performance comparison of different methods towards 15 types of noise in corruption datasets.

strategy, we compare the following methods in the experiments: *i)* **Random**: it selects examples at random. *ii)* **Clean-uncertainty**: it selects the examples with largest uncertainty of clean example predictions (Lewis and Gale 1994). *iii)* **Noise-uncertainty**: a reasonable extension of the last strategy. It selects the examples with largest expected uncertainty of noise example preditions. Specifically, it uses the current perturbation level and clean example to generate M noise examples and selects the examples with the largest average uncertainty of these noise examples predictions. *iv)* **AQPL (ours)**: the propose approach. It selects the examples with most unsuitability of noise to examples.

Moreover, to validate the effectiveness of the querying method, the following methods are compared: *i)* **Standard**: the model are trained only on clean datasets. *ii)* **GNT** (Rusak et al. 2020): it uses a fixed perturbation level to perturb 50% of the training data with Gaussian noise within each batch, and trains the model with clean data and noise data. *iii)* **CAT** (Cheng et al. 2020): it adaptively customizes the perturbation level according to whether the model has capacity to robustly classify the example. *iv)* **AQPL-GNT (ours)**: the proposed approach. It corrects perturbation levels by interacting human experts, which based on the training model of GNT. *v)* **AQPL-CAT (ours)**: the proposed approach. It corrects perturbation levels by interacting human experts, which based on the training model of CAT.

For all active learning methods, M is set to 50, and we fix the query batch size B to 100 on CIFAR10 and MNIST, and 500 on Tiny-Imagenet at each active querying iteration. In annotation process, the parameters σ_{min}, σ_{max} and α are respectively set to 0, 0.9 and 0.01. More hyper-parameters and experimental details can be found in the supplementary material.

Performance Comparison

We plot the accuracy curves of the proposed AQPL approach and compared methods with the number of queries increasing. The results with Gaussian noise are shown in Figure 4. Because of the space limitation, we present the detailed results with other types of noises in the supplementary material, and show the average results of 15 types of noise in Figure 5. The term "QueryNums" in all figures refers to the epoch of interactions with the oracle, and two batches of examples are queried from the oracle at each epoch. It is worthy to note that when comparing with other methods, we use the same base model and query batches of the same size to update the base model for fair comparison. In addition, when the query number is 0, all perturbation levels have not been updated, and the initial value is the performance of GNT.

From Figure 4 and 5, we can observe that the proposed AQPL approach outperforms the other methods in all cases. AQPL can achieve higher accuracy with fewer queries on corruption datasets. The random method, which selects examples at random, can improve the model robustness by querying perturbation levels. This phenomenon implies that it is a reasonable way to improve model robustness by cor-

Dataset	Type	Method				
		Standard	GNT	CAT	AQPL-GNT	AQPL-CAT
MNIST-C	Clean	**99.29**%	97.32%	98.43%	99.21%	99.23%
	Gaussian	16.06%	84.46%	**98.14**%	96.31%	97.90%
	All	65.34%	71.57%	80.11%	78.78%	**80.42**%
CIFAR10-C	Clean	**95.05**%	94.87%	86.42%	94.83%	94.75%
	Gaussian	43.23%	71.62%	82.78%	82.19%	**86.69**%
	All	74.24%	79.59%	71.15%	82.02%	**83.33**%
Tiny-Imagenet-C	Clean	**57.84**%	56.14%	48.62%	56.60%	55.51%
	Gaussian	19.27%	21.90%	27.98%	25.12%	**31.72**%
	All	9.99%	14.04%	23.77%	24.82%	**27.19**%

Table 1: The Top-1 accuracy of different methods on different corruption datasets.

recting perturbation levels with queried information. It is observed that the clean-uncertainty method performs poorly in most cases. One possible reason is that if the model is much uncertain about the clean example itself, then changing the perturbation level will not improve the model robustness. The noise-uncertainty method can always achieve suboptimal performance because noise examples with high uncertainty often need to adjust the perturbation level. The results in Figure 4 and 5 are consistent in general, validating that the proposed AQPL can effectively improve the model robustness with fewer queries against different types of noises.

To further validate the effectiveness of the querying method, we also show the Top-1 accuracy achieved by different methods on different corruption datasets in Table 1. It is worthy to note that the proposed AQPL-GNT and AQPL-CAT methods respectively use GNT (Rusak et al. 2020) and CAT (Cheng et al. 2020) as the based model, and the mean results over 10 queries are recorded. First of all, the standard method can always achieve the best performance on the clean test set, while performs poorly on corruption datasets. When comparing with the method that only trains on the clean datasets, the GNT method, which trains with Gaussian noise with a fixed perturbation level, significantly improves the model robustness against various noises. The CAT method has higher performance on corruption datasets than GNT by adaptively customizing the perturbation levels of examples, which implies that it is important to adaptively adjust the perturbation levels for different examples in the training process. Moreover, by allowing to query the ground-truth information on the perturbation level, the proposed approaches AQPL-GNT and AQPL-CAT can further improve the performances of GNT and CAT respectively. Most importantly, it can be observed that the proposed approach AQPL-CAT outperforms the other methods in most cases with regard to both Gaussian noise and the other 15 types of noise. Note that, when comparing with the method CAT that also adjusts perturbation level according to whether the current model has the capacity to robustly classify the examples, the AQPL-CAT can still achieve better performance. On one hand, the supervised information provided by the oracle is more reliable. On the other hand, human experts correct perturbation levels more efficiently and directly.

In summary, these results consistently demonstrate that the proposed AQPL approach can effectively improve the model robustness by actively querying the correct perturbation level from the oracle, while the sampling strategy can efficiently select the most useful examples to reduce the querying cost.

Discussion

Similar to many existing studies, the experiments are performed on image datasets in this paper. The results show that, by actively querying the supervised information about the perturbation level, model robustness against corruption perturbations on image classification tasks can be improved efficiently. In principle, the proposed method can be applied to any type of data. One challenge is that it could be difficult for human annotators to select a proper perturbation level for non-visual data. If the non-visual data can be easily visualized, such as VisArtico (Ouni, Mangeonjean, and Steiner 2012) for articulatory data, the method is still applicable. It would be an interesting future work to design feasible interfaces for annotators to decide the perturbation level for non-visual data.

In this paper, we focus on the corruption perturbations both in our theoretical and experimental analysis. We believe that corruption perturbations commonly occur in real tasks. On the other hand, it would be interesting to extend the study for improving adversarial robustness. Actually, the average-case robustness under a specific noise distribution could bring non-negligible adversarial robustness (Wong and Kolter 2020). More importantly, the optimal perturbation level for a clean example considered in this paper, essentially, represents an adversarial (worst-case) noise distribution on the example with regard to the oracle.

Conclusion

In this work, we propose a novel active learning framework to improve the model robustness by querying the conform perturbation levels. On one hand, instead of assuming a fixed noise for the whole training set, the perturbation levels are adjusted adaptively for different examples during the training process. On the other hand, by estimating the conformity with classification entropy, the most useful examples are actively selected to achieve effective learning with lower annotation cost. Both theoretical and empirical results validate the effectiveness of the proposed approach. In the future, we plan to extend the framework to handle adversarial perturbations.

Acknowledgments

This research was supported by the National Key R&D Program of China (2020AAA0107000), NSFC (62076128) and the China University S&T Innovation Plan Guided by the Ministry of Education.

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998* .

Bishop, C. M. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural computation* 7(1): 108–116.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705* .

Cemgil, T.; Ghaisas, S.; Dvijotham, K. D.; and Kohli, P. 2020. Adversarially Robust Representations with Smooth Encoders. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1gfFaEYDS.

Chapelle, O.; Weston, J.; Bottou, L.; and Vapnik, V. 2001. Vicinal risk minimization. In *Advances in neural information processing systems*, 416–422.

Cheng, M.; Lei, Q.; Chen, P.-Y.; Dhillon, I.; and Hsieh, C.-J. 2020. Cat: Customized adversarial training for improved robustness. *arXiv preprint arXiv:2002.06789* .

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, 1310–1320.

Drucker, H.; and Le Cun, Y. 1991. Double backpropagation increasing generalization performance. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, 145–150. IEEE.

Fu, Y.; Zhu, X.; and Li, B. 2013. A survey on instance selection for active learning. *Knowledge and information systems* 35(2): 249–283.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* .

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural computation* 4(1): 1–58.

Gilmer, J.; Ford, N.; Carlini, N.; and Cubuk, E. 2019. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In *International Conference on Machine Learning*, 2280–2289.

Grandvalet, Y.; Canu, S.; and Boucheron, S. 1997. Noise injection: Theoretical prospects. *Neural Computation* 9(5): 1093–1108.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations* .

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019a. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, 15637–15648.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019b. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv preprint arXiv:1912.02781* .

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2010. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, 892–900.

Huang, S.-J.; and Zhou, Z.-H. 2013. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th International Conference on Data Mining*, 1079–1084. IEEE.

Huijser, M.; and van Gemert, J. C. 2017. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE international conference on computer vision*, 5286–5295.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 125–136.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images .

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Lewis, D. D.; and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* .

Mao, C.; Zhong, Z.; Yang, J.; Vondrick, C.; and Ray, B. 2019. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, 478–489.

Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A. S.; Bethge, M.; and Brendel, W. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484* .

Mu, N.; and Gilmer, J. 2019. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337* .

Ouni, S.; Mangeonjean, L.; and Steiner, I. 2012. VisArtico: a visualization tool for articulatory data. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Pukelsheim, F. 1994. The three sigma rule. *The American Statistician* 48(2): 88–91.

Qian, Y.; Bi, M.; Tan, T.; and Yu, K. 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(12): 2263–2276.

Roy, N.; and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. Int. Conf. on Machine Learning.

Rusak, E.; Schott, L.; Zimmermann, R.; Bitterwolf, J.; Bringmann, O.; Bethge, M.; and Brendel, W. 2020. Increasing the robustness of DNNs against image corruptions by playing the Game of Noise. *arXiv preprint arXiv:2001.06057* .

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 287–294.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature* 550(7676): 354–359.

Sitawarin, C.; Chakraborty, S.; and Wagner, D. 2020. Improving Adversarial Robustness Through Progressive Hardening. *arXiv preprint arXiv:2003.09347* .

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Tikhonov, A. N.; and Arsenin, V. Y. 1977. Solutions of ill-posed problems. *New York* 1–30.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347* .

Uesato, J.; O'Donoghue, B.; Oord, A. v. d.; and Kohli, P. 2018. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666* .

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*.

Wong, E.; and Kolter, J. Z. 2020. Learning perturbation sets for robust machine learning. *arXiv preprint arXiv:2007.08450* .

Wong, E.; Schmidt, F. R.; and Kolter, J. Z. 2019. Wasserstein adversarial examples via projected sinkhorn iterations. *arXiv preprint arXiv:1902.07906* .

Yao, L.; and Miller, J. 2015. Tiny imagenet classification with convolutional neural networks. *CS 231N* 2(5): 8.

You, X.; Wang, R.; and Tao, D. 2014. Diverse expected gradient active learning for relative attributes. *IEEE transactions on image processing* 23(7): 3203–3217.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573* .

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. 2020. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. *arXiv preprint arXiv:2002.11242* .

Zhang, R. 2019. Making Convolutional Networks Shift-Invariant Again. In *International Conference on Machine Learning*, 7324–7334.