Learning to Predict Trustworthiness with Steep Slope Loss

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding the trustworthiness of a prediction yielded by a classifier is critical for the safe and effective use of AI models. Prior efforts have been proven to be reliable on small-scale datasets. In this work, we study the problem of predicting trustworthiness on real-world large-scale datasets, where the task is more challenging due to high-dimensional features, diverse visual concepts, and large-scale samples. In such a setting, we observe that the trustworthiness predictors trained with prior-art loss functions, i.e., the cross entropy loss, focal loss, and true class probability confidence loss, are prone to view both correct predictions and incorrect predictions to be trustworthy. The reasons are twofold. Firstly, correct predictions are generally dominant over incorrect predictions. Secondly, due to the data complexity, it is challenging to differentiate the incorrect predictions from the correct ones on real-world large-scale datasets. To improve the generalizability of trustworthiness predictors, we propose a novel *steep slope loss* to separate the features w.r.t. correct predictions from the ones w.r.t. incorrect predictions by two slide-like curves that oppose each other. The proposed loss is evaluated with two representative deep learning models, i.e., Vision Transformer and ResNet, as trustworthiness predictors. We conduct comprehensive experiments and analyses on ImageNet, which show that the proposed loss effectively improves the generalizability of trustworthiness predictors. The code and pre-trained trustworthiness predictors for reproducibility are available at `https://gitlab.com/anonymous_science/steep_slope.git`.

1 Introduction

Classification is a ubiquitous learning problem that categorizes objects according to input features. It is widely used in a range of applications, such as robotics [1], environment exploration [2], medical diagnosis [3], etc. In spite of the successful development of deep learning methods in recent decades, high-performance classifiers would still have a chance to make mistakes due to the improvability of models and the complexity in the nature of real-world data [4, 5, 6, 7].

To assess whether the prediction yielded by a classifier can be trusted or not, there are growing efforts towards learning to predict trustworthiness [8, 9]. These methods are evaluated on small-scale datasets, e.g., MNIST [10], where the data is relatively simple and existing classifiers have achieved high accuracy ($> 99\%$). As a result, there are a dominant proportion of correct predictions and the trustworthiness predictors are prone to classify incorrect predictions as trustworthy predictions. The characteristics that the simple data is easy-to-classify aggravate the situation. To further understand the prowess of predicting trustworthiness, we study this problem on the real-world large-scale datasets, e.g., ImageNet [11]. This is a challenging theme for classification in terms of boundary complexity, class ambiguity, and feature dimensionality [12]. As a result, failed predictions are inevitable.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

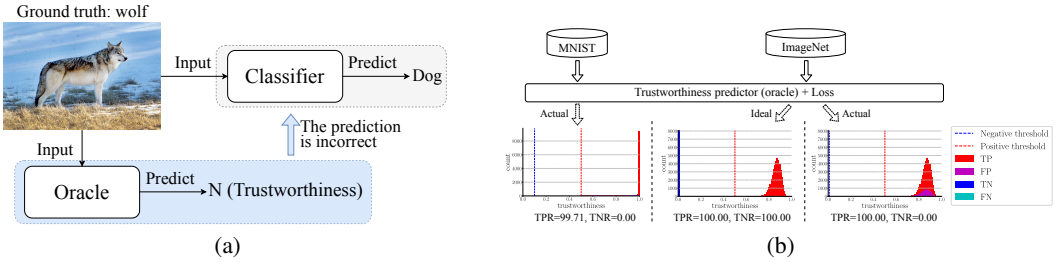


Figure 1: Conceptual illustrations of trustworthiness prediction. (a) shows the process of predicting trustworthiness where the oracle is the trustworthiness predictor. The illustration in (b) shows that the task is challenging on ImageNet, where TCP's confidence loss [9] is used in this example. The confidence that is greater (lower) than the positive (negative) threshold would be classified as a trustworthy (untrustworthy) prediction. Usually, both the positive threshold and the negative threshold are 0.5, but the negative threshold is $\frac{1}{\text{\# of classes}}$ in the case of TCP.

A general illustration of predicting trustworthiness [9, 13] is shown in Fig. 1a. The trustworthiness predictor [13] that is based on the maximum confidence have been proven to be unreliable [8, 14, 15, 16]. Instead, Corbiere et al. [9] propose the true class probability (TCP) that uses the confidence w.r.t. the ground-truth class to determine whether to trust the classifier's prediction or not. Nevertheless, the classification confidence is sensitive to the data. As shown in Fig. 1b, TCP predicts that all the incorrect predictions (0.9% in predictions) are trustworthy on MNIST [10] but predicts that all the incorrect predictions ($\sim$16% in predictions) are trustworthy on ImageNet [11].

To comprehensively understand this problem, we follow the learning scheme used in [9] and use two sate-of-the-art backbones, i.e., ResNet [5] and ViT [7], as the trustworthiness predictors. For simplicity, we call the "trustworthiness predictor" an *oracle*. We find that the oracles trained with cross entropy loss, focal loss [17], and TCP confidence loss [9] on ImageNet are prone to overfit the training samples, i.e., the true positive rate (TPR) is close to 100% while the true negative rate (TNR) is close to 0%. To improve the generalizability of oracles, we propose a novel loss function named the steep slope loss. The proposed steep slope loss consists of two slide-like curves that cross with each other and face in the opposite direction to separate the features w.r.t. trustworthy and untrustworthy predictions. It is tractable to control the slopes by indicating the heights of slides. In this way, the proposed loss is able to be flexible and effective to push the features w.r.t. correct and incorrect predictions to the well-classified regions.

Predicting trustworthiness is similar to as well as different from conventional classification tasks. On one hand, predicting trustworthiness can be formulated as a binary classification problem. On the other hand, task-specific semantics are different between the classification task and predicting trustworthiness. The classes are referred to visual concepts, such as dog, cat, etc., in the classification task, while the ones in predicting trustworthiness are abstract concepts. The trustworthiness could work on top of the classes in the classification task. In other words, the classes in the classification task are specific and closed-form, while trustworthiness is open-form and is related to the classes in the classification task.

The contribution of this work can be summarized as follows.

- We study the problem of predicting trustworthiness with widely-used classifiers on ImageNet. Specifically, we observe that a major challenge of this learning task is that the cross entropy loss, focal loss, and TCP loss are prone to overfit the training samples, where correct predictions are dominant over incorrect predictions.
- We propose the steep slope loss function that improves the generalizability of trustworthiness predictors. We conduct comprehensive experiments and analyses, such as performance on both small-scale and large-scale datasets, analysis of distributions separability, comparison to the class-balanced loss, etc., which verify the efficacy of the proposed loss.
- To further explore the practicality of the proposed loss, we train the oracle on the ImageNet training set and evaluate it on two variants of ImageNet validation set, i.e., the stylized validation set and the adversarial validation set. The two variants' domains are quite different

from the domain of the training set. We find that the learned oracle is able to consistently differentiate the trustworthy predictions from the untrustworthy predictions.

2 Preliminaries

In this section, we first recap how a deep learning model learns in the image classification task. Then, we show how the task of predicting trustworthiness connects to the classification task.

Supervised Learning for Classification. In classification tasks, given a training sample, i.e., image $\boldsymbol{x} \in \mathbb{R}^m$ and corresponding ground-truth label $y \in \mathcal{Y} = \{1, \ldots, K\}$, we assume that samples are drawn i.i.d. from an underlying distribution. The goal of the learning task is to learn to find a classifier $f^{(cls)}(\cdot; \theta')$ with training samples for classification. θ' is the set of parameters of the classifier. Let $f_{\theta}^{(cls)}(\cdot) = f^{(cls)}(\cdot; \theta')$. The optimization problem is defined as

$$f_{\theta'}^{*(cls)} = \arg\min_{f_{\theta'}^{(cls)}} \hat{\mathcal{R}}(f_{\theta'}^{(cls)}, \ell^{(cls)}, D_{tr}), \tag{1}$$

where $f_{\theta'}^{*(cls)}$ is the learned classifier, $\hat{\mathcal{R}}$ is the empirical risk, $\ell^{(cls)}$ is a loss function for classification, and D_{tr} is the set of training samples.

Supervised Learning for Predicting Trustworthiness. In contrast to the learning task for classification, which is usually a multi-class single-label classification task [4, 5, 6, 7], learning to predict trustworthiness is a binary classification problem, where the two classes are positive (i.e., trustworthy) or negative (i.e., untrustworthy). Similar to [9], given a pair $(\boldsymbol{x}, y)$ and a classifier $f_{\theta'}^{(cls)}$, we define the ground-truth label o for predicting trustworthiness as

$$o = \begin{cases} 1, & \text{if } \arg\max f_{\theta'}^{(cls)}(\boldsymbol{x}) = y \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In other words, the classifier correctly predicts the image's label so the prediction is trustworthy in hindsight, otherwise the prediction is untrustworthy.

The learning task for predicting trustworthiness follows a similar learning framework in the classification task. Let $f_{\theta}(\cdot)$ be an oracle (i.e., a trustworthiness predictor). A generic loss function $\ell : \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$, where $\mathbb{R}_{\geq 0}$ is a non-negative space. Given training samples $(\boldsymbol{x}, y) \in D_{tr}$, the optimization problem for predicting trustworthiness is defined as

$$f_{\theta}^{*} = \arg\min_{f_{\theta}} \frac{1}{|D_{tr}|} \sum_{i=1}^{|D_{tr}|} \ell(f_{\theta}(\boldsymbol{x}_i), o_i), \tag{3}$$

where $|D_{tr}|$ is the cardinality of D_{tr}.

Particularly, we consider two widely-used loss functions for classification and the loss function used for training trustworthiness predictors as baselines. They are the cross entropy loss [18], focal loss [17], and TCP confidence loss [9]. Let $p(o = 1|\theta, \boldsymbol{x}) = 1/(1 + \exp(-z))$ be the trustworthiness confidence, where $z \in \mathbb{R}$ is the descriminative feature produced by the oracle, i.e., $z = f_{\theta}(\boldsymbol{x})$. The three loss functions can be written into

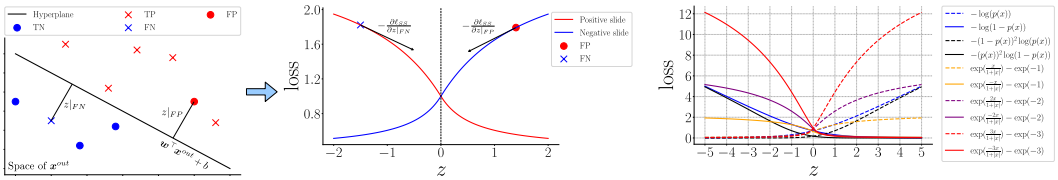
$$\ell_{CE}(f_{\theta}(\boldsymbol{x}), o) = -o \cdot \log p(o = 1|\theta, \boldsymbol{x}) - (1 - o) \cdot \log(1 - p(o = 1|\theta, \boldsymbol{x})), \tag{4}$$

$$\begin{aligned} \ell_{Focal}(f_{\theta}(\boldsymbol{x}), o) = & -o \cdot (1 - p(o = 1|\theta, \boldsymbol{x}))^{\gamma} \log p(o = 1|\theta, \boldsymbol{x}) - \\ & (1 - o) \cdot (p(o = 1|\theta, \boldsymbol{x}))^{\gamma} \log(1 - p(o = 1|\theta, \boldsymbol{x})), \end{aligned} \tag{5}$$

$$\ell_{TCP}(f_{\theta}(\boldsymbol{x}), y) = (f_{\theta}(\boldsymbol{x}) - p(\hat{y} = y|\theta', \boldsymbol{x}))^2. \tag{6}$$

In the focal loss, γ is a hyperparameter. In the TCP confidence loss, $\hat{y}$ is the predicted label and $p(\hat{y} = y|\theta', \boldsymbol{x})$ is the classification probability w.r.t. the ground-truth class.

Consequently, the learned oracle would yield z to generate the trustworthiness confidence. In the cases of ℓ_{CE} and ℓ_{Focal}, the oracle considers a prediction is trustworthy if the corresponding trustworthiness confidence is greater than the positive threshold 0.5, i.e., $p(o = 1|\theta, \boldsymbol{x}) > 0.5$. The predictions whose trustworthiness confidences are equal to or lower than the negative threshold 0.5 are viewed to be untrustworthy. In the case of ℓ_{TCP}, the positive threshold is also 0.5, but the negative threshold correlates to the number of classes in the classification task. It is defined as $1/K$ in [9].



(a) (b)

Figure 2: Conceptual workflow of the proposed steep slope loss (a) and graph comparison between the proposed loss and the conventional losses (b). In (b), the cross entropy loss and focal loss are plotted in blue and black, respectively. The TCP confidence loss is a square error and varies with the classification confidence. Therefore, it is not plotted here.

3 Methodology

In this section, we first introduce the overall learning framework for predicting trustworthiness. Then, we narrow down to the proposed steep slope loss function. At last, we provide the generalization bound that is related to the proposed steep slope loss function.

3.1 Overall Design

Corbière et al. [9] provide a good learning scheme for predicting trustworthiness. Briefly, it first trains a classifier with the training samples. Then, the classifier is frozen and the confidence network (i.e., trustworthiness predictor) is trained (or fine-tuned) with the training samples. In general, we follows this learning scheme.

This work focuses on the trustworthiness on the predictions yielded by the publicly available pre-trained classifiers, i.e., ViT [7] and ResNet [5]. We use the pre-trained backbones as the backbones of the oracles for general purposes. In this sense, the set of the oracle's parameters can be split into two parts, one is related to the backbone and the other one is related to the head, i.e., $\theta = \{\theta_{backbone}, \theta_{head}\}$. $\theta_{backbone}$ is used to generated the intermediate feature $\boldsymbol{x}^{out}$ and $\theta_{head} = \{\boldsymbol{w}, b\}$ are usually the weights of a linear function to generate the discriminative feature $z = \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$. With the classifier, the oracle, a pre-defined loss, and the training samples, we can optimize the problem (3) to find the optimal parameters for the oracle.

3.2 Steep Slope Loss

The conceptual workflow of the proposed steep slope loss is shown in Fig. 2a. The core idea is that we exploit the graph characteristics of the exponential function and the softsign function to establish two slides such that the features z w.r.t. the positive class ride down the positive slide to the right bottom and the features z w.r.t. the negative class ride down the negative slide to the left bottom. z is defined as the signed distance to the hyperplane (i.e., oracle head). Given an image $\boldsymbol{x}$, $z = f_\theta(\boldsymbol{x})$ can be broken down into

$$z = \frac{\boldsymbol{w}^\top \boldsymbol{x}^{out} + b}{\|\boldsymbol{w}\|}, \quad \boldsymbol{x}^{out} = f_{\theta_{backbone}}(\boldsymbol{x}). \tag{7}$$

The signed distance to the hyperplane has a geometric interpretation: its sign indicates in which half-space $\boldsymbol{x}^{out}$ is and its absolute value indicates how far $\boldsymbol{x}^{in}$ is away from the hyperplane.

It is desired that the signed distance of $\boldsymbol{x}^{out}$ with ground-truth label $o = 1$ ($o = 0$) tends towards $+\infty$ ($-\infty$) as much as possible. To this end, we define the steep slope (SS) loss function as follows

$$\ell_{SS}(f_\theta(\boldsymbol{x}), o) = o \cdot \underbrace{\left(\exp\left(\frac{\alpha^+ z}{1+|z|}\right) - \exp(-\alpha^+)\right)}_{\text{Positive slide}} + (1-o) \cdot \underbrace{\left(\exp\left(\frac{-\alpha^- z}{1+|z|}\right) - \exp(-\alpha^-)\right)}_{\text{Negative slide}} \tag{8}$$

where $\alpha^+, \alpha^- \in \mathbb{R}^+$ control the slope of the positive slide and the negative slide, respectively. If z w.r.t. the positive class is on the left-hand side of $z = 0$, minimizing the loss would push the point on the hill down to the bottom, i.e., the long tail region indicating the well-classified region. Similarly, z

w.r.t. the negative class would undergo a similar process. $\exp(-\alpha^+)$ and $\exp(-\alpha^-)$ are vertical shifts for the positive and negative slides such that ℓ_{ss} has a minimum value 0. Note that the proposed steep slope loss is in the range $[0, \text{maximum}\{\exp(\alpha^+) - \exp(-\alpha^+), \exp(\alpha^-) - \exp(-\alpha^-)\}]$, whereas the cross entropy loss and the focal loss are in the range $[0, +\infty)$. The proposed steep slope loss can work with the output of the linear function as well. Because the signed distance and the output of the linear function have a proportional relationship with each other, i.e., $\frac{\boldsymbol{w}^\top \boldsymbol{x}^{out} + b}{\|\boldsymbol{w}\|} \propto \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$.

Essentially, as shown in Fig. 2b, the cross entropy loss, focal loss, and steep slope loss work in a similar manner to encourage z w.r.t. the positive class to move the right-hand side of the positive threshold 0 and encourage z w.r.t. the negative class to move the left-hand side of the negative threshold 0, which is analogous to the sliding motion. The proposed steep slope loss is more tractable to control the steepness of slopes than the cross entropy loss and focal loss. This leads to an effective learning to yield discriminative feature for predicting trustworthiness.

3.3 Generalization Bound

With the proposed steep slope loss, we are interested in the generalization bound of trustworthiness predictors. For simplicity, we simplify a trustworthiness predictor as $f \in \mathcal{F}$, where $\mathcal{F}$ is a finite hypothesis set. The risk of predicting trustworthiness is defined as $\mathcal{R}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}[\ell_{SS}(f(\boldsymbol{x}), o)]$, where P is the underlying joint distribution of $(\boldsymbol{x}, o)$. As P is inaccessible, a common practice is to use empirical risk minimization (ERM) to approximate the risk [19], i.e.,$\hat{\mathcal{R}}_D(f) = \frac{1}{|D|}\sum_{i=1}^{|D|} \ell_{SS}(f(\boldsymbol{x}_i), o_i)$.

The following theorem provides an insight into the correlation between the generalization bound and the loss function in the learning task for predicting trustworthiness.

Theorem 3.1. *Denote maximum*$\{\exp(\alpha^+) - \exp(-\alpha^+), \exp(\alpha^-) - \exp(-\alpha^-)\}$ *as* ℓ_{SS}^{max}. $\ell_{SS} \in [0, \ell_{SS}^{max}]$. *Assume* $\mathcal{F}$ *is a finite hypothesis set, for any* $\delta > 0$*, with probability at least* $1 - \delta$*, the following inequality holds for all* $f \in \mathcal{F}$*:*

$$|\mathcal{R}(f) - \hat{\mathcal{R}}_D(f)| \leq \ell_{SS}^{max} \sqrt{\frac{\log|\mathcal{F}| + \log\frac{2}{\delta}}{2|D|}}$$

The proof sketch is similar to the generalization bound provided in [20] and the detailed proof is provided in the appendix A.

A desired characteristic of the proposed steep slope loss is that it is in a certain range determined by α^+ and α^-, as discussed in Section 3.2. This leads to the generalization bound shown in Theorem 3.1. The theorem implies that given a generic classifier, as the number of training samples increases, the empirical risk would be close to the true risk with a certain probability.

4 Related Work

Loss Function. Loss function is the key to search for optimal parameters and has been extensively studied in a range of learning tasks [9, 17, 21, 22]. Specifically, the cross entropy loss may be the most widely-used loss for classification [21], however, it is not optimal for all cases. Lin et al. [17] propose the focal loss to down-weight the loss w.r.t. well-classified examples and focus on minimizing the loss w.r.t. misclassified examples by reshaping the cross entropy loss function. Liu et al. [22] introduce the large margin softmax loss that is based on cosine similarity between the weight vectors and the feature vectors. Nevertheless, it requires more hyperparameters and entangles with the linear function in the last layer of the network. Similar to the idea of focal loss, the proposed steep slope loss is flexible to reshape the function graphs to improve class imbalance problem. TCP's confidence loss is used to train the confidence network for predicting trustworthiness [9]. We adopt TCP as a baseline.

Trustworthiness Prediction. Understanding trustworthiness of a classifier has been studied in the past decade [8, 9, 13, 23, 24]. Hendrycks and Gimpel [13] aim to detect the misclassified examples according to maximum class probability. This method depends on the ranking of confidence scores and is proved to be unreliable [8]. Monte Carlo dropout (MCDropout) intends to understand trustworthiness through the lens of uncertainty estimation [23]. However, it is difficult to distinguish the trustworthy prediction from the incorrect but overconfident predictions [9]. Jiang et al. [8] propose a confidence measure named trust score, which is based on the distance between the classifier and

a modified nearest-neighbor classifier on the test examples. The shortcomings of the method are the poor scalability and computationally expensive. Recently, instead of implementing a standalone trustworthiness predictor, Moon et al. [24] use the classifier's weights to compute the correctness ranking. The correctness ranking is taken into account in the loss function such that the classification accuracy is improved. This method relies on pairs of training samples for computing the ranking. Moreover, it is unclear if the classifier improved by the correctness ranking outperforms the pre-trained classifier when applying the method on large-scale datasets, e.g., ImageNet. In contrast, the learning scheme to train the trustworthiness predictor (i.e., confidence network) in [9] is standard and does not affect the classifiers. Its efficacy has been verified on small-scale datasets. In this work, we follows the learning scheme used in [9] and focus on predicting trustworthiness on complex dataset.

Imbalanced Classification. In classification task, the ratio of correct predictions to incorrect predictions is expected to be large due to the advance in deep learning methods. This is aligned with the nature of imbalanced classification [25, 26, 27, 28, 29], which generally employ re-sampling and re-weighting strategies to solve the problem. However, predicting trustworthiness is different from imbalanced classification as it is not aware of the visual concepts but the correctness of the predictions, whereas the classification task relies on the visual concepts. Specifically, the re-sampling based methods [25, 26, 29] may not be suitable for the problem of predicting trustworthiness. It is difficult to determine if a sample is under-represented (over-represented) and should be over-sampled (under-sampled). The re-weighting based methods [26, 28] can be applied to any generic loss functions, but they hinge on some hypothesis related to imbalanced data characteristics. For instance, the class-balanced loss [28] is based on the effective number w.r.t. a class, which presumes number of samples is known. However, the number of samples w.r.t. a class is not always assessible, e.g., in the online learning scheme [30, 31]. Instead of assuming some hypothesis, we follow the standard learning scheme for predicting trustworthiness [8, 9].

5 Experiment & Analysis

In this section, we first introduce the experimental set-up. Then, we report the performances of baselines and the proposed steep slope loss on ImageNet, followed by comprehensive analyses.

Experimental Set-Up. We use ViT B/16 [7] and ResNet-50 [5] as the classifiers, and the respective backbones are used as the oracles' backbones. We denote the combination of oracles and classifiers as $\langle$O, C$\rangle$. There are four combinations in total, i.e., $\langle$ViT, ViT$\rangle$, $\langle$ViT, RSN$\rangle$, $\langle$RSN, ViT$\rangle$, and $\langle$RSN, RSN$\rangle$, where RSN stands for ResNet. In this work, we adopt three baselines, i.e., the cross entropy loss [21], focal loss [17], and TCP confidence loss [9] for comparison purposes.

The experiment is conducted on ImageNet [11], which consists of 1.2 million labeled training images and 50000 labeled validation images. It has 1000 visual concepts. Similar to the learning scheme in [9], the oracle is trained with training samples and evaluated on the validation set. During the training process of the oracle, the classifier works in the evaluation mode so training the oracle would not affect the parameters of the classifier. Moreover, we conduct the analyses about how well the learned oracle generalizes to the images in the unseen domains. To this end, we apply the widely-used style transfer method [32] and the functional adversarial attack method [33] to generate two variants of the validation set, i.e., stylized validation set and adversarial validation set.

The oracle's backbone is initialized by the pre-trained classifier's backbone and trained by fine-tuning using training samples and the trained classifier. Training the oracles with all the loss functions uses the same hyperparameters, such as learning rate, weight decay, momentum, batch size, etc. The details for the training process and the implementation are provided in Appendix B.

For the focal loss, we follow [17] to use $\gamma = 2$, which leads to the best performance for object detection. For the proposed loss, we use $\alpha^+ = 1$ and $\alpha^- = 3$ for the oracle that is based on ViT's backbone, while we use $\alpha^+ = 2$ and $\alpha^- = 5$ for the oracle that is based on ResNet's backbone.

Following [9], we use FPR-95%-TPR, AUPR-Error, AUPR-Success, and AUROC as the metrics. FPR-95%-TPR is the false positive rate (FPR) when true positive rate (TPR) is equal to 95%. AUPR is the area under the precision-recall curve. Specifically, AUPR-Success considers the correct prediction as the positive class, whereas AUPR-Error considers the incorrect prediction as the positive class. AUROC is the area under the receiver operating characteristic curve, which is the plot of TPR versus FPR. Moreover, we use TPR and true negative rate (TNR) as additional metrics because they assess

Table 1: Performance on the ImageNet validation set. The averaged scores are computed over three runs. The oracles are trained with the ImageNet training samples. The classifier is used in the evaluation mode in the experiment. Acc is the classification accuracy (%) and is helpful to understand the proportion of correct predictions. *SS* stands for the proposed steep slope loss.

⟨O, C⟩	Loss	Acc↑	FPR-95%-TPR↓	AUPR-Error↑	AUPR-Success↑	AUC↑	TPR↑	TNR↑
⟨ViT, ViT ⟩	CE	83.90	93.01	**15.80**	84.25	51.62	**99.99**	0.02
	Focal [17]	83.90	93.37	15.31	84.76	52.38	99.15	1.35
	TCP [9]	83.90	88.38	12.96	87.63	60.14	99.73	0.00
	SS	83.90	**80.48**	10.26	**93.01**	**73.68**	87.52	**38.27**
⟨ViT, RSN⟩	CE	68.72	93.43	30.90	69.13	51.24	**99.90**	0.20
	Focal [17]	68.72	93.94	**30.97**	69.07	51.26	93.66	7.71
	TCP [9]	68.72	83.55	23.56	79.04	66.23	94.25	0.00
	SS	68.72	**77.89**	20.91	**85.39**	**74.31**	68.32	**67.53**
⟨RSN, ViT⟩	CE	83.90	93.29	14.74	85.40	53.43	**100.00**	0.00
	Focal [17]	83.90	94.60	**14.98**	85.13	52.37	**100.00**	0.00
	TCP [9]	83.90	91.93	14.12	86.12	55.55	**100.00**	0.00
	SS	83.90	**88.70**	11.69	**90.01**	**64.34**	96.20	**9.00**
⟨RSN, RSN⟩	CE	68.72	94.84	29.41	70.79	52.36	**100.00**	0.00
	Focal [17]	68.72	95.16	**29.92**	70.23	51.43	99.86	0.08
	TCP [9]	68.72	88.81	24.46	77.79	62.73	99.23	0.00
	SS	68.72	**86.21**	22.53	**81.88**	**67.92**	79.20	**42.09**

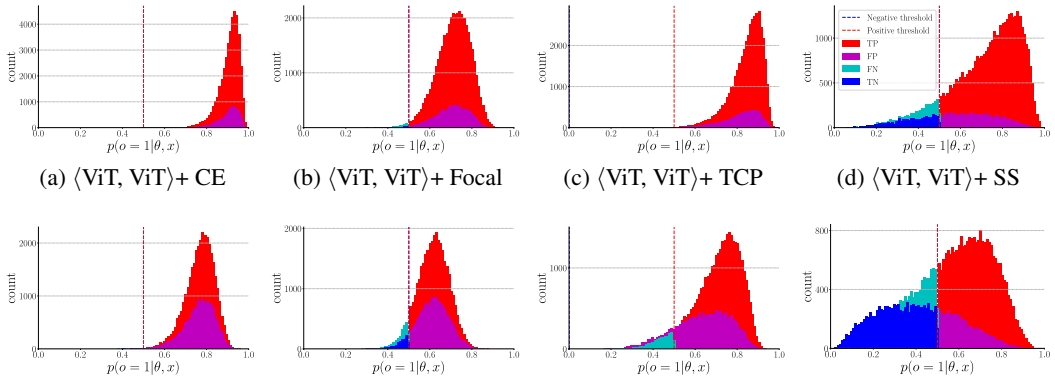
overfitting issue, e.g., TPR=100% and TNR=0% imply that the trustworthiness predictor is prone to view all the incorrect predictions to be trustworthy.

Performance on ImageNet. The result are reported in Table 1. We have two key observations. Firstly, training with the cross entropy loss, focal loss, and TCP confidence loss lead to overfitting the imbalanced training samples, i.e., the dominance of trustworthy predictions. Specifically, TPR is higher than 99% while TNR is less than 1% in all cases. Secondly, the performance of predicting trustworthiness is contingent on both the oracle and the classifier. When a high-performance model (i.e., ViT) is used as the oracle and a relatively low-performance model (i.e., ResNet) is used as the classifier, cross entropy loss and focal loss achieve higher TNRs than the loss functions with the other combinations. In contrast, the two losses with ⟨ResNet, ViT⟩ lead to the lowest TNRs (i.e., 0%).

Despite the combinations of oracles and classifiers, the proposed steep slope loss can achieve significantly higher TNRs than using the other loss functions, while it achieves desirable performance on FPR-95%-TPR, AUPR-Success, and AUC. This verifies that the proposed loss is effective to improve the generalizability for predicting trustworthiness. Note that the scores of AUPR-Error and TPR yielded by the proposed loss are lower than that of the other loss functions. Recall that AUPR-Error aims to inspect how easy to detect failures and depends on the negated trustworthiness confidences w.r.t. incorrect predictions [9]. The AUPR-Error correlates to TPR and TNR. When TPR is close to 100% and TNR is close to 0%, it indicates the oracle is prone to view all the predictions to be trustworthy. In other words, almost all the trustworthiness confidences are on the right-hand side of $p(o = 1|\theta, \boldsymbol{x}) = 0.5$. Consequently, when taking the incorrect prediction as the positive class, the negated confidences are smaller than -0.5. On the other hand, the oracle trained with the proposed loss intends to yield the ones w.r.t. incorrect predictions that are smaller than 0.5. In general, the negated confidences w.r.t. incorrect predictions are greater than the ones yielded by the other loss functions. In summary, a high TPR score and a low TNR score leads to a high AUPR-Error.

To intuitively understand the effects of all the loss functions, we plot the histograms of trustworthiness confidences w.r.t. true positive (TP), false positive (FP), true negative (TN), and false negative (FN) in Fig. 3. The result confirms that the oracles trained with the baseline loss functions are prone to predict overconfident trustworthiness for incorrect predictions, while the oracles trained with the proposed loss can properly predict trustworthiness for incorrect predictions.

Selective Risk Analysis. Risk-coverage curve is an important technique for analyzing trustworthiness through the lens of the rejection mechanism in the classification task [9, 34]. In the context of predicting trustworthiness, selective risk is the empirical loss that takes into account the decisions, i.e., to trust or not to trust the prediction. Correspondingly, coverage is the probability mass of



(a) ⟨ViT, ViT⟩+ CE (b) ⟨ViT, ViT⟩+ Focal (c) ⟨ViT, ViT⟩+ TCP (d) ⟨ViT, ViT⟩+ SS

(e) ⟨ViT, RSN⟩+ CE (f) ⟨ViT, RSN⟩+ Focal (g) ⟨ViT, RSN⟩+ TCP (h) ⟨ViT, RSN⟩+ SS

Figure 3: Histograms of trustworthiness confidences w.r.t. all the loss functions on the ImageNet validation set. The oracles that are used to generate the confidences are the ones used in Table 1. The histograms generated with ⟨RSN, ViT⟩and ⟨RSN, RSN⟩are provided in appendix D.

the non-rejected region. As can see in Fig. 4a, the proposed loss leads to significantly lower risks, compared to the other loss functions. We present the risk-coverage curves w.r.t. all the combinations of oracles and classifiers in Appendix E. They consistently exhibit similar pattern.

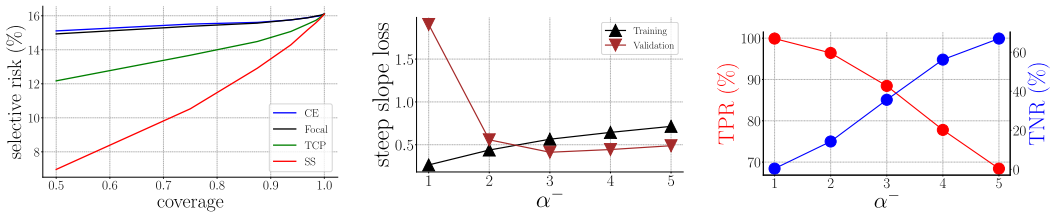
Performance on Small-Scale Datasets. We also provide comparative experimental results on small-scale datasets, i.e., MNIST [10] and CIFAR-10 [35]. The experiment details and results are reported in Appendix F. The proposed loss outperforms TCP [9] on metrics FPR-95%-TPR and AUC on MNIST, while it outperforms TCP on metrics FPR-95%-TPR, AUPR-Error, AUPR-Success, and AUC on CIFAR-10. This shows the proposed loss is able to adapt to relatively simple data.

Generalization to Unseen Domains. In practice, the oracle may run into the data in the domains that are different from the ones of training samples. Thus, it is interesting to find out how well the learned oracles generalize to the unseen domain data. Using the oracles are trained with the ImageNet training set (i.e., the ones used in Table 1), we evaluate it on the stylized ImageNet validation set and adversarial ImageNet validation set. ⟨ViT, ViT⟩ is used in the experiment.

The results are reported in Table 2. As both unseen domains are different from the one of the training set, the classification accuracies are much lower than the ones in Table 1. The adversarial validation set is also more challenging than the stylized validation set. As a result, the difficulty affects the scores across all metrics. The oracles trained with the baseline loss functions are still prone to recognize the incorrect prediction to be trustworthy. The proposed loss consistently improves the performance on FPR-95%-TPR, AUPR-Sucess, AUC, and TNR. Note that the adversarial perturbations are computed on the fly [33]. Instead of truncating the sensitive pixel values and saving into the images files, we follow the experimental settings in [33] to evaluate the oracles on the fly. Hence, the classification accuracies w.r.t. various loss function are slightly different but are stably around 6.14%.

Table 2: Performance on the stylized ImageNet validation set and the adversarial ImageNet validation set. ⟨ViT, ViT⟩ is used in the experiment and the domains of the two validation sets are different from the one of the training set that is used for training the oracle. The corresponding histograms are available in Appendix D.

Dataset	Loss	Acc↑	FPR-95%-TPR↓	AUPR-Error↑	AUPR-Success↑	AUC↑	TPR↑	TNR↑
Stylized [32]	CE	15.94	95.52	84.18	15.86	49.07	**99.99**	0.02
	Focal [17]	15.94	95.96	**85.90**	14.30	46.01	99.71	0.25
	TCP [9]	15.94	93.42	80.17	21.25	57.29	99.27	0.00
	SS	15.94	**89.38**	75.08	**34.39**	**67.68**	44.42	**81.22**
Adversarial [33]	CE	6.14	94.35	**93.70**	6.32	51.28	**99.97**	0.06
	Focal [17]	6.15	93.67	93.48	6.56	52.39	99.06	1.43
	TCP [9]	6.11	93.94	92.77	7.55	55.81	99.71	0.00
	SS	6.16	**90.07**	90.09	**13.07**	**65.36**	87.10	**24.33**



(a) (b) (c)

Figure 4: Analyses based on ⟨ViT, ViT⟩. (a) are the curves of risk vs. coverage. Selective risk represents the percentage of errors in the remaining validation set for a given coverage. (b) are the curves of loss vs. α^-. (c) are TPR and TNR against various α^-.

Ablation Study. In contrast to the compared loss functions, the proposed loss has more hyperparameters to be determined, i.e., α^+ and α^-. As the proportion of correct predictions is usually larger than that of incorrect predictions, we would prioritize α^- over α^+ such that the oracle is able to recognize a certain amount of incorrect predictions. In other words, we first search for α^- by freezing α^+, and then freeze α^- and search for α^+. Fig. 4b and 4c show how the loss, TPR, and TNR vary with various α^-. In this analysis, the combination ⟨ViT, ViT ⟩ is used and $\alpha^+ = 1$. We can see that $\alpha^- = 3$ achieves the optimal trade-off between TPR and TNR. We follow a similar search strategy to determine $\alpha^+ = 2$ and $\alpha^- = 5$ for training the oracle with ResNet backbone.

Effects of using $z = \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$. Using the signed distance as z, i.e., $z = \frac{\boldsymbol{w}^\top \boldsymbol{x}^{out} + b}{\|\boldsymbol{w}\|}$, has a geometric interpretation as shown in Fig. 2a. However, the main-stream models [5, 6, 7] use $z = \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$. Therefore, we provide the corresponding results in appendix G, which are generated by the proposed loss taking the output of the linear function as input. In comparison with the results of using $z = \frac{\boldsymbol{w}^\top \boldsymbol{x}^{out} + b}{\|\boldsymbol{w}\|}$, using $z = \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$ yields comparable scores of FPR-95%-TPR, AUPR-Error, AUPR-Success, and AUC. Also, TPR and TNR are moderately different between $z = \frac{\boldsymbol{w}^\top \boldsymbol{x}^{out} + b}{\|\boldsymbol{w}\|}$ and $z = \boldsymbol{w}^\top \boldsymbol{x}^{out} + b$, when α^+ and α^- are fixed. This implies that TPR and TNR are sensitive to $\|\boldsymbol{w}\|$.

Separability between Distributions of Correct Predictions and Incorrect Predictions. As observed in Fig. 3, the confidences w.r.t. correct and incorrect predictions follow Gaussian-like distributions. Hence, we can compute the separability between the distributions of correct and incorrect predictions from a probabilistic perspective. Given the distribution of correct predictions $\mathcal{N}_1(\mu_1, \sigma_1^2)$ and the distribution of correct predictions $\mathcal{N}_2(\mu_2, \sigma_2^2)$, we use the average Kullback–Leibler (KL) divergence $\bar{d}_{KL}(\mathcal{N}_1, \mathcal{N}_2)$ [36] and Bhattacharyya distance $d_B(\mathcal{N}_1, \mathcal{N}_2)$ [37] to measure the separability. More details and the quantitative results are reported in Appendix H. In short, the proposed loss leads to larger separability than the baseline loss functions. This implies that the proposed loss is more effective to differentiate incorrect predictions from correct predictions.

Steep Slope Loss vs. Class-Balanced Loss. We compare the proposed loss to the class-balanced loss [28], which is based on a re-weighting strategy. Following [28], we use the hyperparameters $\beta = 0.999$ and $\gamma = 0.5$ for the class-balanced loss. The results are reported in Appendix I. Overall, the proposed loss outperforms the class-balanced loss, which implies that the imbalance characteristics of predicting trustworthiness is different from that of imbalanced data classification.

6 Conclusion

In this work, we study the problem of predicting trustworthiness on a large-scale dataset. We observe that the oracle, i.e., trustworthiness predictor, trained with the cross entropy loss, focal loss, and TCP confidence loss lean towards viewing incorrect predictions to be trustworthy due to overfitting. To improve the generalizability of the oracles, we propose the steep slope loss that encourages the features w.r.t. correct predictions and incorrect predictions to be separated from each other. We evaluate the proposed loss on ImageNet through the lens of the trustworthiness metrics, selective classification metric, and separability of distributions, respectively. Experimental results show that the proposed loss is effective in improving the generalizability of trustworthiness predictors.

References

[1] Daniel Leidner, Christoph Borst, Alexander Dietrich, Michael Beetz, and Alin Albu-Schäffer. Classifying compliant manipulation tasks for automated planning in robotics. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2015.

[2] Antoine Caté, Lorenzo Perozzi, Erwan Gloaguen, and Martin Blouin. Machine learning as a tool for geologists. *The Leading Edge*, 36(3):215–219, 2017.

[3] José Antonio Sanz, Mikel Galar, Aranzazu Jurio, Antonio Brugos, Miguel Pagola, and Humberto Bustince. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Applied Soft Computing*, 20:103–111, 2014.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[9] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Mitra Basu and Tin Kam Ho. *Data complexity in pattern recognition*. Springer Science & Business Media, 2006.

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[14] Foster J. Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 445–453, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[16] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[18] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[19] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[20] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[21] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[22] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 507–516. JMLR.org, 2016.

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[24] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7034–7044. PMLR, 13–18 Jul 2020.

[25] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384, 2016.

[26] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5016–5025, 2019.

[27] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019.

[29] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13893–13902, 2020.

[30] Koby Crammer, Jaz Kandola, and Yoram Singer. Online classification on a budget. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

[31] Yahav Bechavod, Christopher Jung, and Steven Z. Wu. Metric-free individual fairness in online learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11214–11225. Curran Associates, Inc., 2020.

[32] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[33] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[35] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[36] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[37] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 1.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see the discussion in the ablation study in Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please refer to the abstract, the experimental set-up in Section 5, and Section B in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer to the experimental set-up and the ablation study in Section 5, and Section B in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Please refer to Table 1, Table A1, and Section B in the appendix.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to Section B in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please refer to Section 5 and Section B in the appendix.
   (b) Did you mention the license of the assets? [Yes] Please refer to Section C in the appendix.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please refer to the abstract and Section C in the appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]