On the Importance of Gradients for Detecting Distributional Shifts in the Wild

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting out-of-distribution (OOD) inputs has become a critical component in ensuring the safe deployment of machine learning models in the real world. Existing OOD detection approaches primarily rely on the output or feature space for deriving OOD scores, while largely overlooking information from the *gradient space*. In this paper, we present `GradNorm`, a simple and effective approach for detecting OOD inputs by utilizing information extracted from the gradient space. `GradNorm` directly employs the vector norm of gradients, backpropagated from the KL divergence between the softmax output and a uniform probability distribution. Our key idea is that the magnitude of gradients is higher for in-distribution (ID) data than for OOD data, making it informative for OOD detection. `GradNorm` establishes *state-of-the-art* performance on a comprehensive suite of benchmarks, reducing the average FPR95 by up to 14.47% compared to the previous best method.

1 Introduction

When deploying machine learning models in the real world, there is an increasingly important question to ask: *"Is the model making a faithful prediction for something it was trained on, or is the model making an unreliable prediction for something it has not been exposed to during training"?* We want models that are not only accurate on its familiar data distribution, but also aware of uncertainty outside the training distribution. This gives rise to the importance of out-of-distribution (OOD) detection, which determines whether an input is in-distribution (ID) or OOD. As of recently a plethora of literature has emerged to address the problem of OOD detection [1, 14, 18, 27, 30-32, 34].

The main challenge in OOD detection stems from the fact that modern deep neural networks can easily produce overconfident predictions on OOD inputs [37]. This phenomenon makes the separation between ID and OOD data a non-trivial task. Much of the prior work focused on deriving OOD uncertainty measurements from the activation space of the neural network, e.g., using model output [14, 18, 27, 31, 32] or feature representations [30]. Yet, this leaves an alternative space—model parameter and its *gradient space*—largely unexplored. Will a model react to ID and OOD inputs differently in its gradient space, and if so, can we discover distinctive signatures to separate ID and OOD data from gradients?

In this paper, we tackle this key question by exploring and exploiting the richness of the gradient space, ultimately showing that gradients carry surprisingly useful signals for OOD detection. Formally, we present `GradNorm`, a simple and effective approach for detecting OOD inputs by utilizing gradient extracted from a pre-trained neural network. Specifically, `GradNorm` employs the vector norm of gradients directly as an OOD scoring function. Gradients are backpropagated from the Kullback-Leibler (KL) divergence [26] between the softmax output and uniform distribution. ID data is expected to have larger KL divergence because the prediction tends to concentrate on one of the ground-truth classes and therefore is distributed less uniformly. As depicted in Figure 1, our key idea

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

is that the gradient norm is higher for ID data than that for OOD data, making it informative for OOD detection.

We provide both empirical and theoretical insights, demonstrating the superiority of `GradNorm` over both output-based and feature-based methods that are competitive in the literature. Empirically, we establish state-of-the-art performance on a suite of common OOD detection benchmarks, as well as a large-scale ImageNet benchmark. `GradNorm` outperforms the previous best method by a large margin, with up to 14.47% reduction in false-positive rate (FPR95). Theoretically, we show that `GradNorm` captures the *joint information* between the feature and the output space. The joint information results in an overall stronger separability than using either feature or output space alone.

To the best of our knowledge, there is very limited prior work studying how to use gradients for OOD detection. Closest to our work, Lee and AlRegib [28] proposed to train an auxiliary binary classifier using gradient information from ID and OOD data. However, this method can unfairly overfit the OOD test datasets and does not suit OOD-agnostic settings in the real world. In contrast, `GradNorm` is convenient without requiring training a separate classifier, and is completely hyperparameter-free without the need for tuning on validation data.

Our **key results and contributions** are summarized as follows. Code and data will be released to facilitate reproducible research.

- We propose `GradNorm`, a simple and effective gradient-based OOD detection method, which is both label-agnostic (no label required for backpropagation) and OOD-agnostic (no outlier data required). On a challenging ImageNet benchmark, our method reduces the average FPR95 by 10.89% compared to the current best method. On common CIFAR benchmark, `GradNorm` reduces the FPR95 by 14.47%.
- We perform comprehensive analysis that improves understandings of the gradient-based method under (1) different network architectures, (2) gradient norms extracted at varying depths, (3) different loss functions for backpropagation, and (4) different vector norms for aggregating gradients.
- We perform mathematical analysis of `GradNorm` and show that it can be decomposed into two terms, jointly characterizing information from both feature and output space, which demonstrates superiority.

2 Preliminaries

We start by recalling the general setting of the supervised learning problem. We denote by $\mathcal{X} = \mathbb{R}^d$ the input space and $\mathcal{Y} = \{1, 2, ..., C\}$ the output space. A learner is given access to a set of training data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ drawn from an unknown joint data distribution P defined on $\mathcal{X} \times \mathcal{Y}$. A neural network $f(\mathbf{x}; \theta) : \mathcal{X} \to \mathbb{R}^C$ minimizes the empirical risk:

$$R_{\mathcal{L}}(f) = \mathbb{E}_D(\mathcal{L}_{\text{CE}}(f(\mathbf{x}; \theta), y)),$$

where θ is the parameters of the network, and $\mathcal{L}_{\text{CE}}$ is the commonly used cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(f(\mathbf{x}), y) = -\log \frac{e^{f_y(\mathbf{x})/T}}{\sum_{c=1}^{C} e^{f_c(\mathbf{x})/T}}. \tag{1}$$

Specifically, $f_y(\mathbf{x})$ denotes the y-th element of $f(\mathbf{x})$ corresponding to the ground-truth label y, and T is the parameter for temperature scaling [16].

Problem statement Out-of-distribution (OOD) detection can be formulated as a binary classification problem. Given a classifier f learned on training samples from in-distribution P, the goal is to design a binary function estimator,

$$g(\mathbf{x}) = \begin{cases} \text{in}, & \text{if } S(\mathbf{x}) \geq \gamma \\ \text{out}, & \text{if } S(\mathbf{x}) < \gamma, \end{cases}$$

that classifies whether a sample $\mathbf{x} \in \mathcal{X}$ is from P or not. The key challenge is to derive an OOD scoring function $S(\mathbf{x})$ for effective detection. Previous approaches have primarily relied on models' output or features for deriving OOD scores. Instead, our approach seeks to compute OOD scores based on the information extracted from the *gradient space*, which we describe in the next section.

3 Gradient-based OOD Detection

In this section, we describe our proposed method `GradNorm`. We start by introducing the loss function for backpropagation and then describe how to leverage the gradient norm as an OOD score.

We calculate gradients for each parameter by backpropagating the Kullback-Leibler (KL) divergence [26] between the softmax output and a uniform distribution. Formally, KL divergence quantifies how close a model-predicted distribution $q = \{q_i\}$ is to a reference probability distribution $p = \{p_i\}$,

$$D_{\text{KL}}(p \,||\, q) = \sum_i p_i \log \frac{p_i}{q_i} = -\sum_i p_i \log q_i + \sum_i p_i \log p_i = H(p, q) - H(p). \tag{2}$$

In particular, we set the reference distribution to be uniform $\mathbf{u} = [1/C, 1/C, ..., 1/C] \in \mathbb{R}^C$. The predictive probability distribution is the softmax output. Our KL divergence for backpropagation can be written as:

$$D_{\text{KL}}(\mathbf{u} \,||\, \text{softmax}(f(\mathbf{x})) = -\frac{1}{C}\sum_{c=1}^{C} \log \frac{e^{f_c(\mathbf{x})/T}}{\sum_{j=1}^{C} e^{f_j(\mathbf{x})/T}} - H(\mathbf{u}), \tag{3}$$

where the first term is the cross-entropy loss between the softmax output and a uniform vector $\mathbf{u}$, and the second term $H(\mathbf{u})$ is a constant. The KL divergence measures how much the predictive distribution is away from the uniform distribution. Intuitively, ID data is expected to have larger KL divergence because the prediction tends to concentrate on the ground-truth class and is thus distributed less uniformly.

GradNorm as OOD score For a given parameter w, the gradient of the above KL divergence is given by:

$$\frac{\partial D_{\text{KL}}(\mathbf{u} \,||\, \text{softmax}(f(\mathbf{x}))}{\partial w} = \frac{1}{C}\sum_{i=1}^{C} \frac{\partial \mathcal{L}_{\text{CE}}(f(\mathbf{x}), i)}{\partial w}, \tag{4}$$

where w is a component of network parameter θ. Notice that the gradient of the entropy term is 0, *i.e.* $\partial H(\mathbf{u})/\partial w = 0$. In other words, the gradient of KL divergence is equivalent to averaging the derivative of the categorical cross-entropy loss for *all* labels.

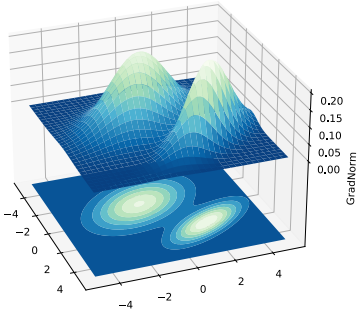


Figure 1: An example of two-dimensional input space. Input data is depicted in the xy-plane, while gradient norm for each input is depicted in the z-dimension. The magnitude of gradients is higher for ID data (light green) than that for OOD data (deep blue).

We now define the OOD score via a vector norm of gradients of the selected parameters:

$$S(\mathbf{x}) = \|\frac{\partial D_{\text{KL}}(\mathbf{u} \,\|\, \text{softmax}(f(\mathbf{x}))}{\partial \mathbf{w}}\|_p, \tag{5}$$

where $\|\cdot\|_p$ denotes L_p-norm and $\mathbf{w}$ is the set of parameters in vector form[1]. We term our method `GradNorm`, short for gradient norm. In practice, `GradNorm` can be conveniently implemented by calculating the cross-entropy loss between the predicted softmax probability and a uniform vector as the target. We will discuss choices and impacts of the selected parameter set $\mathbf{w}$ in Section 4.2.

Rationale of GradNorm Our operating hypothesis is that using the KL divergence for backpropagation, gradient norm is higher for ID data than that for OOD data. As we show in Section 4.2, using the gradient norm of the KL divergence is more effective than using the KL divergence directly. Moreover, `GradNorm` derived from the KL divergence with a uniform target offers two advantages over gradient norms derived from the standard cross-entropy loss.

- First, our method is *label-agnostic* and does not require any ground-truth label. It can be flexibly used during inference time when the label is unavailable for either ID or OOD data.
- Second, it captures the *uncertainty across all categories*, providing more information for OOD detection. We will provide empirical evidence to show the importance of utilizing all labels in Section 4.2.

[1] We concatenate all selected parameters into a single vector regardless of the original shapes of the parameters.

Method Space	Method	iNaturalist		SUN		Places		Textures		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Output	MSP [14]	63.69	87.59	79.98	78.34	81.44	76.76	82.73	74.45	76.96	79.29
	ODIN [31]	62.69	89.36	71.67	83.92	76.27	80.67	81.31	76.30	72.99	82.56
	Generalized ODIN [18]	66.36	84.68	62.17	85.98	66.83	83.38	67.00	81.23	65.59	83.82
	Energy [32]	64.91	88.48	65.33	85.32	73.02	81.37	80.87	75.79	71.03	82.74
Feature	Mahalanobis [30] ♣	96.34	46.33	88.43	65.20	89.75	64.46	52.23	72.10	81.69	62.02
Gradient	Lee and AlRegib [28] ♣	75.49	72.30	73.82	82.61	87.90	74.00	**43.39**	**84.16**	70.15	78.27
	GradNorm (ours)	**50.03**	**90.33**	**46.48**	**89.03**	**60.86**	**84.82**	61.42	81.07	**54.70**	**86.31**

Table 1: **Main Results.** OOD detection performance comparison between `GradNorm` and baselines. All methods utilize the standard ResNetv2-101 model pre-trained on ImageNet [5]. The classification model is trained on **ID data only**. ↑ indicates larger values are better, while ↓ indicates smaller values are better. All values are percentages. ♣ indicates a separate binary classifier needs to be trained explicitly for OOD detection.

4 Experiments

In this section, we evaluate `GradNorm` on a comprehensive suite of OOD detection benchmarks. In Section 4.1, we begin with evaluating `GradNorm` on a large-scale ImageNet benchmark [20], followed by extensive ablations and analyses that improve the understanding of our approach (Section 4.2). We further evaluate on CIFAR benchmarks that are routinely used in literature in Section 4.3.

4.1 Evaluation on ImageNet

Dataset We first use ImageNet-1k [5] as in-distribution dataset. Compared to the CIFAR benchmarks that are routinely used in literature, the ImageNet setting is not only more realistic (with higher resolution images) but also more challenging (with a larger label space of 1,000 categories). We evaluate on four OOD test datasets (see Appendix A.1 for detail). Specifically, the four test OOD datasets are from subsets of `iNaturalist` [46], `SUN` [48], `Places` [51], and `Textures` [4], with non-overlapping categories w.r.t. ImageNet-1k. The evaluations span a diverse range of domains including fine-grained images, scene images, and textural images.

Model and hyperparameters We use Google BiT-S models[2] [24] pre-trained on ImageNet-1k with a ResNetv2-101 architecture [13]. Additionally, we use L_1-norm-based OOD scores by default and explore the effect of other L_p-norms in Section 4.2. The temperature parameter T is set to be 1 unless specified otherwise, and we explore the effect of different temperatures in Section 5. At test time, all images are resized to 480×480.

Results The results for ImageNet evaluations are shown in Table 1, where `GradNorm` establishes the state-of-the-art performance. We report OOD detection performance for each OOD test dataset, as well as the average over the four datasets. For a fair comparison, all the methods use the same pre-trained backbone, without regularizing with auxiliary outlier data. In particular, we compare with MSP [14], ODIN [31], Generalized ODIN [18], Mahalanobis [30], as well as Energy [32]. *Experimental details and hyperparameters of baseline methods can be found in Appendix A.2.*

`GradNorm` outperforms the best output-based baseline, Generalized ODIN [18], by **10.89**% in FPR95. `GradNorm` also outperforms a competitive feature-based method, Mahalanobis [30], by **26.99**% in FPR95. We hypothesize that the increased size of label space makes the class-conditional Gaussian density estimation less viable. Moreover, methods such as [30] and [28] require training a separate binary classifier, whereas `GradNorm` can be conveniently used through a simple threshold comparison.

Comparison with Lee and AlRegib Lee and AlRegib [28] proposed an OOD detection framework using the L_2-norm of gradients. Importantly, they do not directly use gradient norms for OOD detection, but instead, use them as the input for training a separate binary classifier. Furthermore, the binary classifier is trained on the OOD datasets, which can unfairly overfit to the test data and does not suit OOD-agnostic settings in the real world. In contrast, our methodology mitigates the shortcomings in that `GradNorm` (1) does not require any new model training, (2) is hyperparameter-free, and (3) is suitable for OOD-agnostic settings. As shown in Table 1, `GradNorm` reduces the FPR95 by **15.45**% compared with their approach. Additionally, our work comprehensively investigates and explains various important design choices in using gradient-based methods (see Section 4.2), which is otherwise not considered in [28].

[2] https://github.com/google-research/big_transfer

GradNorm is effective on alternative neural network architecture We evaluate `GradNorm` on a different architecture DenseNet-121 [19], and report performance in Table 2. `GradNorm` is consistently effective, outperforming the best baseline, Energy [32], by **10.29**% in FPR95.

Method Space	Method	iNaturalist		SUN		Places		Textures		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Output	MSP [14]	48.55	89.16	69.39	80.46	71.42	80.11	68.51	78.69	64.47	82.11
	ODIN [31]	37.00	93.29	57.30	86.12	61.91	84.14	56.49	84.62	53.18	87.04
	Generalized ODIN [18]	63.41	85.38	78.15	76.45	76.11	77.41	68.78	78.82	71.61	79.52
	Energy [32]	36.39	93.29	54.91	86.53	59.98	**84.29**	53.87	85.07	51.29	87.30
Feature	Mahalanobis [30]	97.36	42.24	98.24	41.17	97.32	47.27	62.78	56.53	88.93	46.80
Gradient	Lee and AlRegib [28]	74.97	62.39	43.59	**88.53**	55.50	82.98	60.29	70.35	58.59	76.06
	GradNorm (ours)	**23.87**	**93.97**	**43.04**	87.79	**53.92**	83.04	**43.16**	**87.48**	**41.00**	**88.07**

Table 2: OOD detection performance comparison on a different architecture, **DenseNet-121** [19]. Model is trained on ImageNet-1k [5] as the ID dataset.

4.2 Ablation Studies

Gradients from the last layer is sufficiently informative In this ablation, we investigate several variants of `GradNorm` where the gradients are extracted from different network depths. Specifically, we consider gradients of (1) **block n**: all trainable parameters in the n-th block, (2) **all parameters**: all trainable parameters from all layers of the model, and (3) **last layer parameters**: weight parameters from the last FC layer.

Table 3 contrasts the OOD detection performance using different *gradient space*. For each setting, we report the FPR95 and AUROC averaged across four OOD datasets. We observe that gradients from deeper layers tend to yield significantly better performance than shallower layers. This is desirable since gradients w.r.t. deeper layers are computationally more efficient than shallower layers. Interestingly, OOD scores obtained from the last linear layer always yield the best results among all gradient selections. Practically, one only needs to perform backpropagation w.r.t. the last linear layer, which incurs negligible computations. Therefore, our main results are based on the norm of gradients extracted from weight parameters in the last FC layer of the model.

Gradient Space	FPR95 ↓	AUROC ↑
Block 1	73.52	76.41
Block 2	74.34	76.63
Block 3	71.73	78.11
Block 4	65.07	85.11
All params	69.35	81.14
Last layer params	**54.70**	**86.31**

Table 3: Effect of `GradNorm` using different subset of gradients. Gradient norm derived from deeper layers yield better OOD detection performance.

GradNorm with one-hot v.s. uniform targets In this ablation, we contrast `GradNorm` derived using uniform targets (ours) v.s. one-hot targets. Specifically, our scoring function Equation 5 is equivalent to

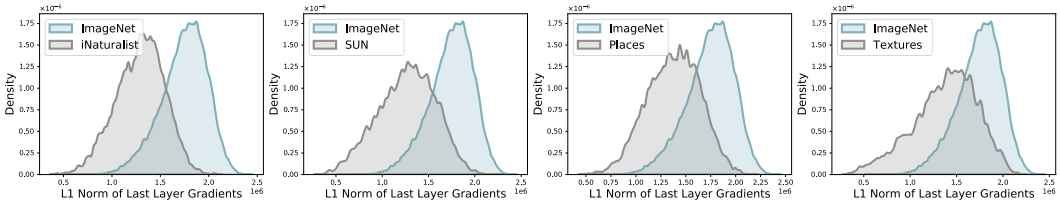
$$S(\mathbf{x}) = \|\frac{1}{C}\sum_{i=1}^{C}\frac{\partial \mathcal{L}_{\text{CE}}(f(\mathbf{x}), i)}{\partial \mathbf{w}}\|, \tag{6}$$

which captures the gradient of cross-entropy loss across *all* labels. In contrast, we compare against an alternative scoring function that utilizes *only one* dominant class label:

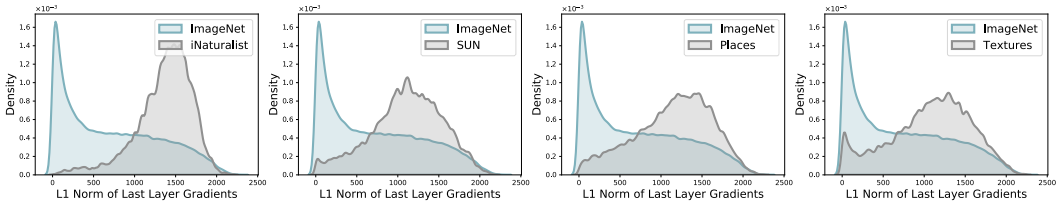
$$S_{\text{one-hot}}(\mathbf{x}) = \|\frac{\partial \mathcal{L}_{\text{CE}}(f(\mathbf{x}), \hat{y})}{\partial \mathbf{w}}\|, \tag{7}$$

where $\hat{y}$ is the predicted class with the largest output.

We first analyze the score distributions using uniform targets (**top**) and one-hot targets (**bottom**) for ID and OOD data in Figure 2. There are two salient observations we can draw: (1) using uniform target (ours), gradients of ID data indeed have larger magnitudes than those of OOD data, as the softmax prediction tends to be less uniformly distributed (and therefore results in larger KL divergence). In contrast, the gradient norm using one-hot targets shows the opposite trend, with ID data having lower magnitudes. This is also expected since the training objective explicitly minimizes the cross-entropy loss, which results in smaller gradients for the majority of ID data. (2) The score distribution using



(a) Gradient norms using KL divergence between the softmax prediction and the **uniform** target.



(b) Gradient norms using KL divergence between the softmax prediction and the **one-hot** target.

Figure 2: Comparison of L_1-norm distributions of last layer gradients between KL divergence with *uniform* target and KL divergence with *one-hot* target. We show in-distribution data in green and OOD data in gray.

one-hot targets displays a strong overlapping between ID (green) and OOD (gray) data, with large variances. In contrast, our method `GradNorm` can significantly improve the separability between ID and OOD data, resulting in better OOD detection performance.

Figure 3 reports the OOD detection performance using uniform targets (ours) v.s. one-hot targets. We use L_1-norm in both cases. For one-hot targets, we use the negative norm, *i.e.* $-S_{\text{one-hot}}(\mathbf{x})$, to align with the convention that ID data has higher scores. `GradNorm` outperforms its counterpart with one-hot targets by a large margin. For instance, `GradNorm` reduces FPR95 by **48.35%** when evaluated on the SUN dataset. Our analysis signifies the importance of measuring OOD uncertainty using all label information.

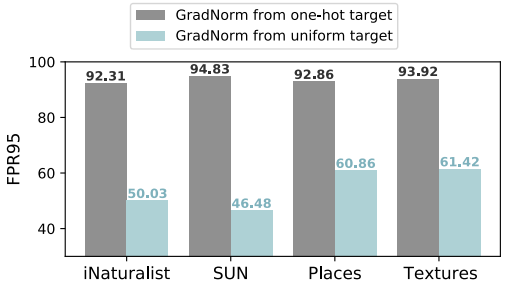


Figure 3: OOD detection performance (FPR95) comparison between uniform (ours) v.s. one-hot target.

L_1-norm is the most effective How does the choice of L_p-norm in Equation 5 affect the OOD detection performance? To understand this, we show in Figure 4 the comparison using $L_{1\sim4}$-norm, L_∞-norm, as well as a fraction norm (with $p = 0.3$). Compared with higher-order norms, L_1-norm achieves the best OOD detection performance on all four datasets. We hypothesize that L_1-norm is better suited since it captures information equally from all dimensions in the gradient space, whereas higher-order norms will unfairly highlight larger elements rather than smaller elements (due to the effect of the power of p). In the extreme case, L_∞-norm only takes the largest element (in absolute value) into account and results in the worst OOD detection performance among all norms. On the other hand, the fraction norm overall does not outperform L_1-norm. *We additionally provide results for more L_p-norms in Appendix D, with* $p = \{0.3, 0.5, 0.8, 1, 2, 3, 4, 5, 6, \infty\}$.

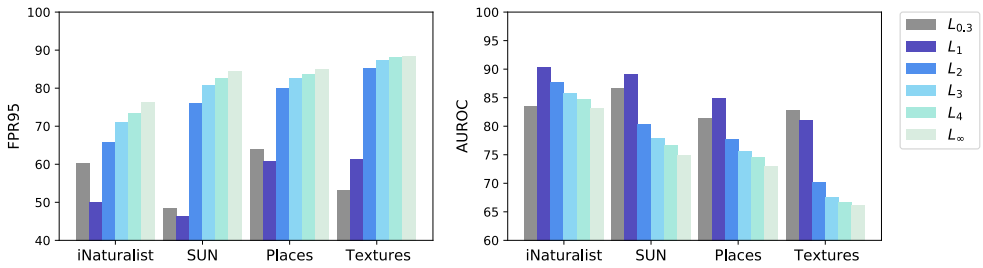


Figure 4: OOD detection performance comparison under different L_p-norms. We show FPR95 (*left*) and AUROC (*right*).

GradNorm is more effective than directly using KL divergence We provide an ablation, contrasting the performance of using `GradNorm` v.s. using the KL divergence derived from Equation 3 as OOD scoring function. The results are show in Figure 5, where `GradNorm` yields significantly better performance than the KL divergence directly extracted from the *output space*, demonstrating the superiority of *gradient space* for OOD detection.

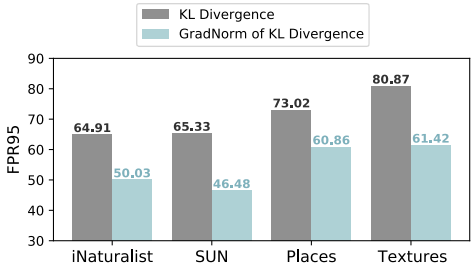


Figure 5: Comparison between `GradNorm` v.s. directly using the KL divergence as scoring function.

4.3 Evaluation on CIFAR Benchmarks

Setup We additionally evaluate `GradNorm` on a common benchmark with CIFAR-10 and CIFAR-100 [25] as ID datasets, which is routinely used in literature [14, 31, 18, 32, 30]. We use the standard split with 50,000 training images and 10,000 test images. We pre-train a ResNet-20 [12] network for 100 epochs. The learning rate is initially 0.1, and decays by a factor of 10 at epochs 50, 75 and 90 respectively. We evaluate on four common OOD benchmark datasets: `SVHN` [36], `LSUN` (crop) [50], `Places365` [51], and `Textures` [3].

Results We summarize the results in Table 4, where `GradNorm` remains competitive. In particular, `GradNorm` reduces the average FPR95 by **8.77%** on CIFAR-10 compared to the best baseline. On CIFAR-100, `GradNorm` outperforms the best baseline energy score [32] by **14.47**% in FPR95 (*full results on CIFAR-100 are available in Appendix B*). Together with our large-scale evaluation in Section 4.1, `GradNorm` overall demonstrates superior performance compared to competitive methods in literature. While some baselines such as [31, 30] may require validation datasets, `GradNorm` is hyperparameter-free and can be used in OOD-agnostic setting. Experimental details and hyperparameters of baseline methods can be found in Appendix A.2. *Additional performance comparison with generative-based methods is provided in Appendix C.*

Method	SVHN		LSUN (crop)		Places365		Textures		Average	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [14]	66.09	89.86	37.73	94.87	70.05	85.99	68.23	87.62	60.53	89.59
ODIN [31]	55.52	89.63	2.32	99.39	45.86	90.81	52.78	89.99	39.12	92.46
Generalized ODIN [18]	34.71	90.49	35.91	88.28	77.11	72.83	59.54	73.94	51.82	81.39
Energy [32]	49.80	91.97	3.86	99.03	46.48	90.55	58.67	88.79	39.70	92.59
Mahalanobis [30]	20.91	95.99	9.66	97.90	89.24	61.15	28.83	92.31	37.16	86.84
Lee and AlRegib [28]	76.09	60.98	44.26	87.00	72.11	72.04	59.79	71.84	63.06	72.96
GradNorm (ours)	17.76	96.66	0.23	99.87	57.85	85.20	37.71	90.76	**28.39**	**93.12**

Table 4: OOD detection performance on common benchmark CIFAR-10 [25].

5 Analysis of Gradient-based Method

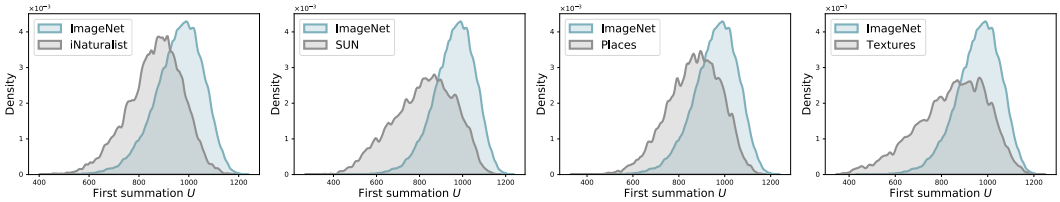
In this section, we analyze the best variant of `GradNorm`, L_1-norm of the last layer gradients (*cf.* Section 4.2), and provide insights on the mathematical insights. Specifically, we denote the last FC layer in a neural network by:

$$f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x} + \mathbf{b}, \tag{8}$$

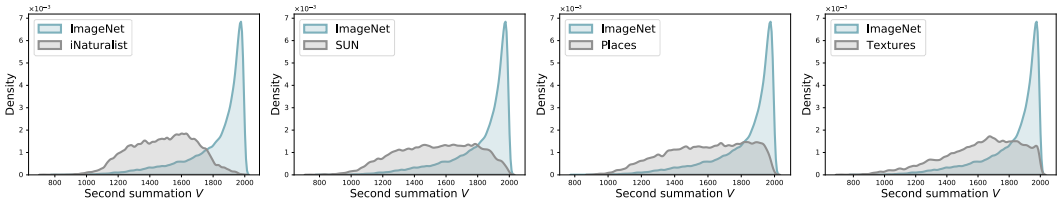
where $f = [f_1, f_2, \dots, f_C]^\top \in \mathbb{R}^C$ is the logit output, $\mathbf{x} = [x_1, x_2, \dots, x_m]^\top \in \mathbb{R}^m$ is the input feature vector, $\mathbf{W} \in \mathbb{R}^{m \times C}$ is the weight matrix, and $\mathbf{b} \in \mathbb{R}^C$ is the bias vector.

GradNorm captures joint information between feature and output First we can rewrite the KL divergence between the softmax prediction and the uniform target as:

$$\begin{aligned} D_{\text{KL}}(\mathbf{u} \| \text{softmax}(f(\mathbf{x}))) &= -\frac{1}{C} \sum_{c=1}^{C} \log \frac{e^{f_c/T}}{\sum_{j=1}^{C} e^{f_j/T}} - H(u) \\ &= -\frac{1}{C} \left(\frac{1}{T} \sum_{c=1}^{C} f_c - C \cdot \log \sum_{j=1}^{C} e^{f_j/T} \right) - H(u). \end{aligned}$$



(a) Distribution of U



(b) Distribution of V

Figure 6: We show the distributions of the two summations decomposed from the L_1-norm of the last layer gradient, for both in-distribution data (blue) and out-of-distribution data (gray).

Then we consider the derivative of D_{KL} w.r.t. each output logit f_c:

$$\begin{aligned}\frac{\partial D_{\text{KL}}}{\partial f_c} &= -\frac{1}{CT}\left(1 - CT \cdot \frac{\partial\left(\log \sum_{j=1}^{C} e^{f_j/T}\right)}{\partial f_c}\right) \\ &= -\frac{1}{CT}\left(1 - C \cdot \frac{e^{f_c/T}}{\sum_{j=1}^{C} e^{f_j/T}}\right).\end{aligned}$$

Next the derivative of D_{KL} w.r.t. the weight matrix can be written as:

$$\frac{\partial D_{\text{KL}}}{\partial \mathbf{W}} = \mathbf{x}\frac{\partial D_{\text{KL}}}{\partial f} = -\frac{1}{CT} \cdot [x_1, x_2, \dots, x_m]^\top [1 - C \cdot \frac{e^{f_1/T}}{\sum_{j=1}^{C} e^{f_j/T}}, \dots, 1 - C \cdot \frac{e^{f_C/T}}{\sum_{j=1}^{C} e^{f_j/T}}].$$

Finally, the L_1-norm of gradients of the weight matrix is simply the sum of absolute values of all elements in the gradient matrix:

$$\begin{aligned}S(\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{C}\left|\left(\frac{\partial D_{\text{KL}}}{\partial \mathbf{W}}\right)_{ij}\right| &= \frac{1}{CT}\sum_{i=1}^{m}\left(|x_i|\left(\sum_{j=1}^{C}\left|1 - C \cdot \frac{e^{f_j/T}}{\sum_{j=1}^{C} e^{f_j/T}}\right|\right)\right) \\ &= \frac{1}{CT}\left(\sum_{i=1}^{m}|x_i|\right)\left(\sum_{j=1}^{C}\left|1 - C \cdot \frac{e^{f_j/T}}{\sum_{j=1}^{C} e^{f_j/T}}\right|\right) \\ &\triangleq \frac{1}{CT} U \cdot V,\end{aligned} \tag{9}$$

where the first multiplicative term $U = \sum_{i=1}^{m} |x_i|$ is the L_1-norm of the feature vector $\mathbf{x}$, and the second term V characterizes information in the output space.

Ablation on U and V In Figure 6 we plot distribution densities of U and V, for both ID and OOD data. It's important to note that U and V measure statistical distributions in the feature space and the output space, respectively. Therefore, `GradNorm` captures the joint information between the feature and the output space. The multiplication of both U and V results in an overall stronger separability between ID and OOD, as seen in Figure 2a. We report the OOD detection performance using U and V individually as scoring functions in Table 5, both of which are less competitive than `GradNorm`.

Effect of temperature scaling We evaluate our method `GradNorm` with different temperatures T from $T = 0.5$ to $T = 1024$. As shown in Figure 7, $T = 1$ is optimal, while either increasing or decreasing the temperature will degrade the performance. This can be explained mathematically via

Method	iNaturalist		SUN		Places		Textures		Average	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
U (feature space)	77.84	74.33	61.90	78.74	76.42	72.75	67.84	72.77	71.00	74.65
V (output space)	66.14	88.45	69.49	83.13	75.95	78.98	81.13	76.06	73.18	81.66
U · V (Joint space)	**50.05**	**90.33**	**46.48**	**89.03**	**60.86**	**84.82**	**61.42**	**81.07**	**54.70**	**86.31**

Table 5: OOD detection performance using the decomposed U (feature space) and V (output space) as scoring functions. Model is trained on ImageNet-1k [5] as the ID dataset.

the V term in Equation 9. Specifically, using a large temperature will result in a smoother softmax distribution, with $C \cdot \frac{e^{f_j/T}}{\sum_{j=1}^{C} e^{f_j/T}}$ closer to 1 (and $V \to 0$). This leads to a less distinguishable distribution between ID and OOD. Our method can be hyperparameter-free by setting $T = 1$.

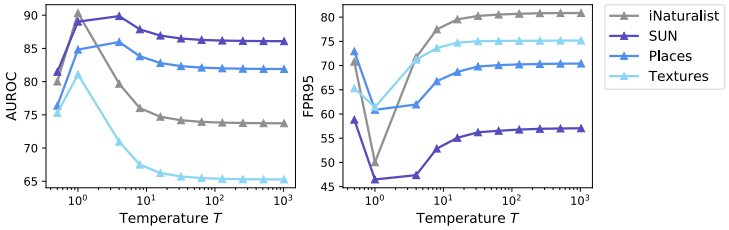


Figure 7: OOD detection performance of `GradNorm` with varying temperature parameter T. We show AUROC (*left*) and FPR95 (*right*).

6 Related Work

Discriminative-based OOD detection Hendrycks and Gimpel [14] proposed a baseline approach for OOD detection utilizing the maximum softmax probability from a pre-trained model. Several works attempt to improve the OOD detection performance by calibrating confidence scores [31, 18] or energy score [32] derived from the output space. Information from the feature space [30] was proven effective on small-scale benchmarks. Other methods seek improvements using deep ensembles [27] or adding a special branch for confidence estimation [6]. These approaches derive OOD score either from output or feature space. In contrast, we show that *gradient space* carries surprisingly useful information for OOD detection, which was underexplored in the literature.

OOD detection with auxiliary outlier data An orthogonal line of work reveals that models can produce lower confidence on anomalous samples with exposure to auxiliary outlier data [1, 9, 33, 34, 43]. Auxiliary outlier data can either be realistic images [15, 32, 34, 39] or synthetic images generated by GANs [29]. However, in many applications, it can be prohibitive to construct an auxiliary outlier dataset since the in-distribution data has a much larger coverage of concepts (e.g., ImageNet). Importantly, our method does not assume the availability of any auxiliary data, and can be *flexibly* used for any pre-trained model.

OOD detection with generative models Alternative approaches for detecting OOD inputs resort to generative models that directly estimate density [7, 21, 38, 41, 44]. An input is deemed as OOD if it lies in the low-density regions. However, Nalisnick et al. [35] showed that deep generative models can assign a high likelihood to OOD data. Several methods improve OOD detection with generative models, such as improved metrics [2], likelihood ratios [40, 42] and likelihood regret [49]. However, generative models [17] can be prohibitively challenging to train and optimize, and the performance can often lag behind discriminative counterpart [23, 47]. In contrast, our method relies on a discriminative classifier, which is much easier to optimize and achieves stronger performance.

7 Conclusion

In this paper, we propose `GradNorm`, a novel OOD detection approach utilizing information extracted from the *gradient space*. Experimental results on a suite of common benchmarks show that our gradient-based method can improve the performance of OOD detection by up to 14.47% in FPR95, establishing state-of-the-art performance. Extensive ablations provide further understandings of our approach. We hope that our research brings to light the informativeness of gradient space, and inspires future work to utilize gradient space for OOD detection.

References

[1] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

[2] Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. 2018.

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[5] 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. Ieee, 2009.

[6] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[8] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

[9] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pages 2151–2159. PMLR, 2019.

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

[11] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. In *International Conference on Learning Representations*, 2018.

[18] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] Rui Huang and Yixuan Li. Towards scaling out-of-distribution detection for large semantic space. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[22] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[23] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV 2020*, 2020.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[28] Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2416–2420. IEEE, 2020.

[29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[31] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

[33] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[34] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020.

[35] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2018.

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[37] 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.

[38] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4797–4805, 2016.

[39] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *arXiv preprint arXiv:1906.03509*, 2019.

[40] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14680–14691, 2019.

[41] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[42] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020.

[43] Akshayvarun Subramanya, Suraj Srinivas, and R Venkatesh Babu. Confidence estimation in deep neural networks via density modelling. *arXiv preprint arXiv:1707.07013*, 2017.

[44] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[45] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[46] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[47] Ziyu Wang, Bin Dai, David Wipf, and Jun Zhu. Further analysis of outlier detection with deep generative models. *Advances in Neural Information Processing Systems*, 33, 2020.

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[49] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in Neural Information Processing Systems*, 33, 2020.

[50] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[51] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 4 and 5.
   (b) Did you describe the limitations of your work? [N/A]
   (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] See Section 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 5.
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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See experimental setups in Section 4 and also Appendix A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (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]
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] See Section 4 Datasets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
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]