Gradient-Based Adversarial and Out-of-Distribution Detection

Jinsol Lee [1] **Mohit Prabhushankar** [1] **Ghassan Alregib** [1]

Abstract

We propose to utilize gradients for detecting adversarial and out-of-distribution samples. We introduce *confounding labels*—labels that differ from normal labels seen during training—in gradient generation to probe the *effective expressivity* of neural networks. Gradients depict the amount of change required for a model to properly represent given inputs, providing insight into the representational power of the model established by network architectural properties as well as training data. By introducing a label of different design, we remove the dependency on ground truth labels for gradient generation during inference. We show that our gradient-based approach allows for capturing the anomaly in inputs based on the *effective expressivity* of the models with no hyperparameter tuning or additional processing, and outperforms state-of-the-art methods for adversarial and out-of-distribution detection.

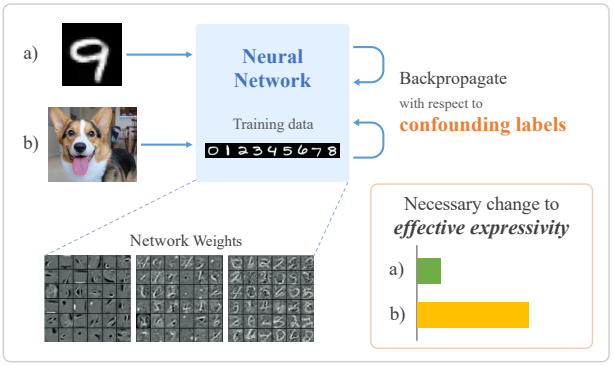


Figure 1: Overview of our approach. We argue that the necessary change observed in gradients invoked by a *confounding label* is smaller for an input similar to training data (*i.e.*, image **a**) than a more anomalous input from the model's perspective (*i.e.*, image **b**).

1. Introduction

Deep neural networks (DNNs) have achieved significant advancements over the past decade, and they are increasingly utilized in cybersecurity applications such as malware detection and intrusion detection (Mahdavifar & Ghorbani, 2019). However, many recent studies pointed out that DNNs are prone to failure when deployed in real-world environments where they often encounter data that diverge from training conditions (Temel et al., 2017; 2018; 2019). Neural networks rely heavily on the implicit assumption that any given input during inference is drawn from the same distribution as the training data (*i.e.*, in-distribution), thereby making overconfident predictions (Guo et al., 2017). In addition, the discovery of adversarial attacks by Szegedy *et al.* (Szegedy et al., 2013) highlighted the vulnerability of DNNs to inputs with small perturbations specifically designed to deceive the models. To ensure reliable performance of practical applications of a neural network, the model must be capable of distinguishing inputs that differ from training data and cannot be handled properly.

In this paper, we examine the ability of trained neural networks to handle inputs based on its *effective expressivity*. We define *effective expressivity* as the representational power of a model based on the training data. We propose to utilize backpropagated gradients to analyze the *effective expressivity* of a trained model. Gradients correspond to the amount of changes that a model requires to accurately represent a given sample. We argue that with gradients, we can characterize the anomaly in inputs based on what the model is unfamiliar with and thus incapable of representing properly. We describe the motivation of our approach in Fig. 1. A classifier is trained with handwritten digit images of 0 through 8 to capture their characteristics with its weights, visualized for some layers of the trained classifier in Fig. 1. During inference, we observe the necessary changes captured in gradients invoked by a *confounding label* with respect to an image similar to training data (*i.e.*, a handwritten digit 9) and highly dissimilar (*i.e.*, dog). Our hypothesis is that the necessary change to *effective expressivity* would be larger for the dissimilar input since the model weights learned on

[1]Omni Lab for Intelligent Visual Engineering and Science (OLIVES), School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. Correspondence to: Jinsol Lee <jinsol.lee@gatech.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

digit 0 to 8 would not be able to capture the characteristics of a dog. One problem in utilizing gradients to observe the necessary change is that we do not have access to the labels for given inputs or any information regarding their distribution. To remove dependency on ground truth labels in gradient generation during inference, we introduce a *confounding label*—label that the network has not seen during training. While many studies proposed to manipulate the design of labels for various purposes (Tokozume et al., 2018; Zhang et al., 2017b; Yun et al., 2019; Durand et al., 2019; Duarte et al., 2021), they only analyze the effect of different label designs in training with the statistics of training dataset. In comparison, a *confounding label* does not require the knowledge of training data and can be used during inference with no access to ground truth labels. Our methodology allows for capturing anomaly in inputs from the perspective the model based on its *effective expressivity*. The contributions of this paper are three-fold:

- We propose to utilize backpropagated gradients to characterize anomaly in inputs seen during inference from the perspective of the model.
- We introduce *confounding labels* as a tool to elicit model response that can be utilized to probe the *effective expressivity* of trained neural networks.
- We validate our approach on the applications of detecting anomalous inputs including adversarial and out-of-distribution (OOD) samples, and achieve state-of-the-art performance with no hyperparameter tuning or additional processing.

2. Related Work and Preliminary

2.1. Adversarial & Out-of-Distribution Detection

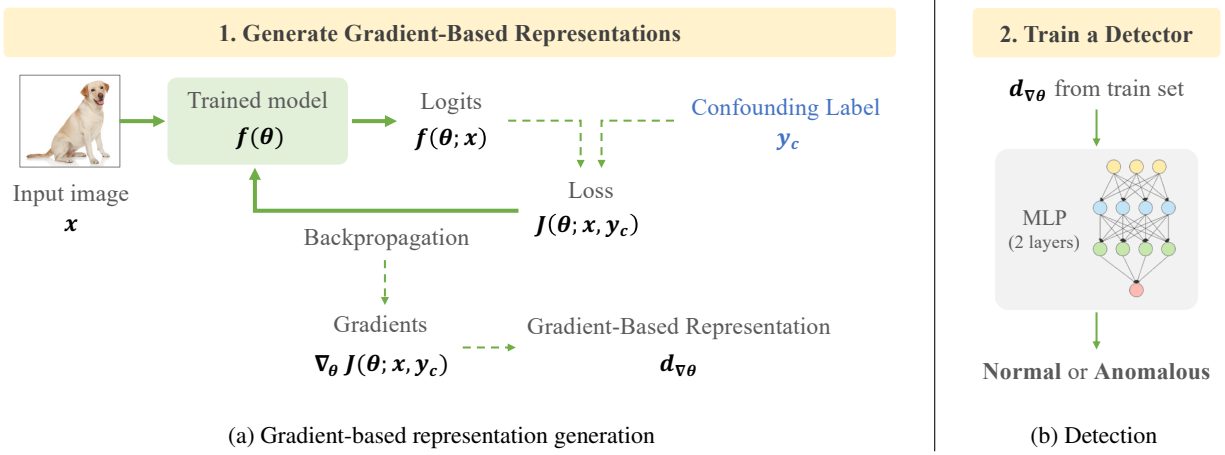
There are many recent studies of detecting anomalous samples, whether they are adversarially generated or due to statistical shift in data distribution. Hendrycks & Gimpel (2017) introduced a baseline method to threshold samples based on the predicted softmax distributions. Liang et al. (2018) proposed additional input and output processing to further improve the baseline method. DeVries & Taylor (2018) utilized prediction confidence obtained from an augmented confidence estimation branch on a pre-trained classifier. Ma et al. (2018) proposed to characterize the dimensional properties of adversarial regions with local intrinsic dimensionality. Lee et al. (2018) proposed a confidence metric using Mahalanobis distance. Liu et al. (2020) utilized energy score to capture the likelihood of occurrence during inference and training time. Most approaches utilize activation-based representations to characterize inputs via learned features, *i.e.*, what the network knows about given inputs. However, for anomalous inputs, the overconfident nature of neural networks (Guo et al., 2017) makes it counterintuitive to characterize them solely with the learned features. Rather, we argue that the anomaly in inputs should be established from the perspective of the model based on what the model is unfamiliar with and thus incapable of representing properly.

2.2. Gradients as Features

At the core of training neural networks lies gradient-based optimization techniques (Ruder, 2016). Apart from their utility as a tool to search for a converged solution, backpropagated gradients have been utilized for various purposes. We discuss the recent studies based on how the relevant gradients are generated. First category is by backpropagating model outputs (*i.e.*, logits) directly. This is widely observed in visualization techniques (Selvaraju et al., 2017; Chattopadhay et al., 2018) to highlight pixels based on classification scores. Zinkevich et al. (2017) also utilizes logits to extract holographic representation of the network. These approaches assume that the test data is drawn from the same distribution as the training data, and they observe the model response by further enforcing the output of the model. Another category of approaches is by backpropagating loss between logits and some target labels. Gradient-based adversarial attack generation fits under this category. Goodfellow et al. (2015) was first to demonstrate that a small but intentional perturbation can affect models to completely change their predictions with high confidence. Others proposed adversarial attacks that also utilize gradients in an iterative manner (Kurakin et al., 2017; Madry et al., 2017) with different initialization constraints. Another relevant work include contrastive visual explanation (Prabhushankar et al., 2020) where model classification is contrasted with other possible classes. Oberdiek et al. (2018); Lee & AlRegib (2020; 2021) also utilize loss-based approaches to generate gradients to quantify the uncertainty of neural networks. Kwon et al. (2019; 2020) employ the directional information of gradients for representation learning. These approaches utilize gradients in a similar way to model training by incorporating some forms of class labels seen during training. It is natural to utilize these in-distribution class labels when the considered test data is safe to be assumed in-distribution and it allows for inciting the model response relevant for the classes of interest. However, this becomes irrelevant when the data during inference is not drawn from the same distribution as the training data. We argue that to characterize anomaly in inputs, *unseen* labels should be introduced for gradient generation instead of seen class labels.

2.3. Manipulating Label Designs

One-hot encoding is the most widely used approach in formulating labels to handle nominal data (*i.e.*, categorical data without quantitative relationships between categories) for many applications including single-label image classifica-



(a) Gradient-based representation generation

(b) Detection

Figure 2: Overall framework for gradient-based anomalous input detection.

tion. For multi-label image classification, a combination of multiple one-hot encoded vectors is used to indicate information about multiple classes in each image. Based on these two, many recent studies analyze different formulations of labels for various purposes. Some work (Tokozume et al., 2018; Zhang et al., 2017b; Yun et al., 2019) proposed to mix two input images and their labels with pre-computed ratio as a data augmentation technique to improve generalizability as well as robustness of the models. Prabhushankar & AlRegib (2021) explored combinations of binary classification labels in extracting causal visual features for interpretability. Another type of manipulated label design is partial labels. Durand et al. (2019) tackled the problem of missing labels per image in multi-label classification and proposed to exploit the proportion of known labels per sample to weigh the contribution of the images and labels. Duarte et al. (2021) proposed to mask parts of labels based on the ratio of positive samples to negative samples to handle imbalanced dataset. These approaches explore different designs of labels only in model training and require the knowledge of training data. We propose to manipulate label designs to characterize inputs from the perspective of models with gradients during inference. Contrary to existing studies, our approach does not depend on the availability of training data statistics or ground truth labels of inference data.

2.4. Capacity and Expressivity

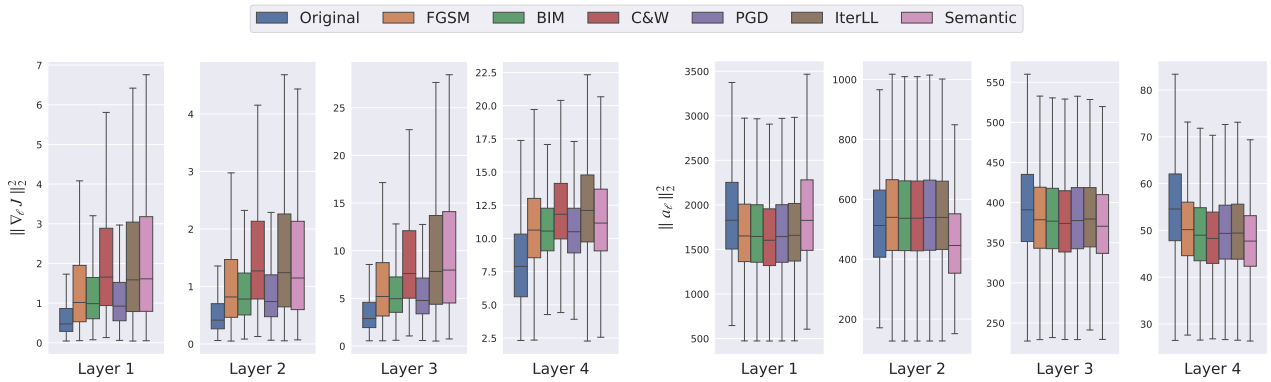
We discuss the ability of a trained neural network to handle given inputs based on two main concepts: *capacity* and *expressivity*. *Capacity* of a neural network is mainly expressed by its size or the number of parameters (Neyshabur et al., 2017; Sun et al., 2017). On the other hand, *expressivity* describes how different architectural properties of a network (depth, width, layer type) affects its ability to approximate functions (Raghu et al., 2017; Zhang et al., 2017a). Many studies discuss *expressivity* by analyzing different classes of neural networks with the same number of parameters on the functions which the networks are capable or incapable of representing. In essence, *capacity* can be seen as the maximum utilization of the network's representational power whereas *expressivity* is the observed utilization based on the connection of network parameters.

3. Method

In this section, we discuss our framework in two stages as shown in Fig. 2: generating gradient-based representations with *confounding labels* and detecting anomalous inputs. We also demonstrate the effectiveness of gradients generated with *confounding labels* in comparison with activations.

3.1. Gradients with Confounding Labels

Based on the concepts of *capacity* and *expressivity*, we define *effective expressivity* as the *expressivity* that is effectively utilized for learning given training data, which establishes the knowledge base of the model. Consider two models of the same architecture but with different training datasets of MNIST (LeCun et al., 1998) and ImageNet (Deng et al., 2009). The *capacity* and *expressivity* would be the same for both models since the network architecture remains fixed and these characteristics are independent of data. However, the *effective expressivity* would be less for the model trained with a simpler dataset of MNIST than more complex dataset of ImageNet. We propose to utilize gradients generated with a *confounding label* to examine the *effective expressivity* of a classifier during inference. Our approach focuses on the lack of knowledge that the model should offset in order to fully grasp the anomaly in the data. A *confounding label* provides an unsupervised way to elicit gradient responses relevant for the *effective ex-*



(a) Gradients (b) Activations

Figure 3: Comparison between layer-wise gradients $\nabla_\ell J$ and activations in $L2$-norms. The magnitude values are obtained from a convolutional layer from each residual block of a ResNet-18 model trained on CIFAR-10 dataset.

pressivity of a trained network that is necessary to improve its representational power and get closer to its *capacity*.

We discuss the first stage of our framework, which is the generation of gradient-based representations with *confounding labels*, shown in Fig. 2a. An input image is passed through a trained image classifier, and the model yields logits. We utilize binary cross entropy loss $J(\theta)$ between the logits and a *confounding label*, as shown in Eq. (1), where $\hat{y}_i$ is the predicted probability for class i, $y_{c,i}$ is the true probability represented by the *confounding label*, and N is the number of classes. For this paper, we utilize the *confounding label* that combines one-hot encodings of all classes (*i.e.*, a vector of all 1's).

$$J(\theta) = \frac{1}{N}\sum_{i=0}^{N-1}\left(y_{c,i}\cdot\log\left(\hat{y}_i\right) + \left(1-y_{c,i}\right)\cdot\log\left(1-\hat{y}_i\right)\right), \tag{1}$$

The loss is then backpropagated through the model, generating gradients for each network layer. While any form of the gradients that preserve their magnitude would be valid, we measure the squared L_2-norm of the gradients for each model parameter set (*i.e.*, weight and bias parameters of network layers) and concatenate them to construct a gradient-based representation for each input. The obtained representation is of the following form,

$$d_{\nabla\theta} = \left[\, \|\nabla J_{\theta_0}(\theta)\|_2^2\,,\ \cdots\,,\ \|\nabla J_{\theta_{L-1}}(\theta)\|_2^2\,\right], \tag{2}$$

where L is the number of layers in a given network.

Validation of gradients with *confounding labels* We demonstrate the effectiveness of gradients obtained with *confounding labels* in comparison with activations. In an adversarial detection setup, we show the disparity between *effective expressivity* established with training data and the necessary adjustments to represent original image samples and adversarial samples. The key idea is that the model will need larger updates to handle adversarial samples than original images, captured by the magnitudes of the generated gradients. For this experiment, we first train a ResNet-18 (He et al., 2016) classifier trained with original CIFAR-10 (Krizhevsky & Hinton, 2009) train set. The test set of CIFAR-10 is utilized to generate adversarial attacks of the following: fast gradient sign method (FGSM) (Goodfellow et al., 2015), basic iterative method (BIM) (Kurakin et al., 2017), Carlini & Wagner attack (C&W) (Carlini & Wagner, 2017), projected gradient descent (PGD) (Madry et al., 2017), iterative least-likely class method (IterLL) (Kurakin et al., 2017), and semantic attack (Hosseini et al., 2017). We collect the L2 norm of layer-wise gradients and activations and visualize their distribution in Fig. 3. Due to the space limitation, we select a convolutional layer from each residual block of the ResNet architecture for visualization. The separation in the ranges of gradient magnitudes, shown in Fig. 3a, between original images of CIFAR-10 test set and their adversarial versions is more evident in some parts of the network than others because each layer captures information about different aspects of given inputs. But overall, we observe the smaller magnitude ranges for original images and significantly larger magnitude ranges for adversarial attack images. On the other hand, the activations throughout the network, shown in Fig. 3b, do not show as much variation in the magnitude ranges for different image sets. It supports our intuition that gradients can characterize the anomaly better than activations based on the unfamiliar aspects of given inputs from the model's perspective.

3.2. Anomalous Input Detection

We discuss the second stage of our framework in Fig. 2b, which is the detection of anomalous inputs with gradient-based representations. The gradient-based representations

Table 1: Adversarial detection results for CIFAR-10 in AUROC. For Mahalanobis method (denoted as M), we report vanilla results (V), *i.e.*, without the input pre-processing or feature ensemble, input pre-processing only (P), feature ensemble only (FE), and with both (P+FE). All values are in percentages and the best results are highlighted in bold.

MODEL	ATTACKS	BASELINE	LID	M(V)	M(P)	M(FE)	M(P+FE)	OURS
RESNET	FGSM	51.20	90.06	81.69	84.25	**99.95**	**99.95**	93.45
	BIM	49.94	99.21	87.09	89.20	**100.0**	**100.0**	96.19
	C&W	53.40	76.47	74.51	75.71	92.78	92.79	**97.07**
	PGD	50.03	67.48	56.27	57.57	65.23	75.98	**95.82**
	ITERLL	60.40	85.17	62.32	64.10	85.10	92.10	**98.17**
	SEMANTIC	52.29	86.25	64.18	65.79	83.95	84.38	**90.15**
DENSENET	FGSM	52.76	98.23	86.88	87.24	**99.98**	99.97	96.83
	BIM	49.67	**100.0**	89.19	89.17	**100.0**	**100.0**	96.85
	C&W	54.53	80.58	75.77	76.16	90.83	90.76	**97.05**
	PGD	49.87	83.01	70.39	66.52	86.94	83.61	**96.77**
	ITERLL	55.43	83.16	70.17	66.61	83.20	77.84	**98.53**
	SEMANTIC	53.54	81.41	62.16	62.15	67.98	67.29	**89.55**

are collected for both normal and anomalous images with *confounding labels*, and they are split into train, validation, and test sets (40%-40%-20% split). Each representation is labeled with the binary information regarding its corresponding input image (positive for anomalous inputs, negative for normal inputs). A multi-layer perceptrons (MLP) of 2 layers is trained as an anomalous input detector using the train and validation sets, and the trained MLP model is applied on the test set to report the detection accuracy.

4. Experiments

In this section, we utilize gradient-based representations generated in response to *confounding labels* for detecting anomalous inputs: adversarial detection and out-of-distribution detection.

4.1. Adversarial Detection

For adversarial detection, we train ResNet and DenseNet (Huang et al., 2017) architectures for a classification task on CIFAR-10 training set. The test set of CIFAR-10 is utilized to generate adversarial attacks mentioned in the validation portion of Section 3.1. We utilize the pristine CIFAR-10 test set as negative samples and the adversarial images as positive samples. For comparison, we employ the Baseline method (Hendrycks & Gimpel, 2017), local intrinsic dimensionality (LID) scores (Ma et al., 2018), and the Mahalanobis method (Lee et al., 2018). Following the setup detailed in Section 3.2, the performance is measured in AUROC and reported in Table 1. For the Mahalanobis method, we report vanilla results (V), *i.e.*, without input pre-processing or feature ensemble, results with input pre-processing only (P), results with feature ensemble only (FE), results with both (P+FE) to demonstrate its dependency on delicate calibration of hyperparameters.

The Mahalanobis method shows the best performance for the adversarial attack types of FGSM and BIM. On the other hand, our approach outperforms all state-of-the-art methods for the C&W, PGD, IterLL and Semantic attacks. Note that the Mahalanobis method requires fine-tuning of hyperparameters. For all attacks, we utilize the range of hyperparameter values that they have explored in their paper and report the performance with the best parameter values. Interestingly enough, the Mahalanobis approach shows less saturated performance for the types of attacks that were not included in their work. These attack types also correspond to the ones for which our method outperforms all the state-of-the-art approaches. This indicates the need for further hyper parameter tuning outside the suggested parameter values, depending on the types of considered adversarial attacks. It highlights the benefit of our approach in its simplicity of obtaining the gradient-based representations with no hyperparameter tuning or any pre-/post-processing. We remark that our method can characterize the adversarial attack samples effectively, even for the attack types that are widely considered to be highly challenging to detect: C&W and PGD attacks.

4.2. Out-of-distribution Detection

We employ various image classification datasets for OOD detection: CIFAR-10 and SVHN (Netzer et al., 2011) as in-distribution, and additional TinyImageNet (Deng et al., 2009) and LSUN (Yu et al., 2015) as OOD. We demonstrate our approaches using ResNet-18 and DenseNet. For

Table 2: Out-of-distribution detection results in detection accuracy, AUROC, and AUPR. All values are in percentages and the best results are highlighted in bold.

Dataset Distribution		Baseline / ODIN / Mahalanobis / Energy / Ours		
In	Out	Detection Accuracy	AUROC	AUPR
CIFAR-10 (ResNet)	SVHN	83.36 / 88.81 / 91.95 / - / **98.04**	88.30 / 94.93 / 97.10 / 95.72 / **99.84**	88.26 / 95.45 / 96.12 / 97.26 / **99.98**
	ImageNet	84.01 / 85.21 / **97.45** / - / 91.28	90.06 / 91.86 / **99.68** / 84.95 / 96.92	89.26 / 91.60 / **99.60** / 84.60 / 97.21
	LSUN	87.34 / 88.42 / **98.60** / - / 98.37	92.79 / 94.48 / **99.86** / 92.39 / **99.86**	92.30 / 94.22 / 99.82 / 95.40 / **99.87**
SVHN (ResNet)	CIFAR-10	79.98 / 80.12 / 88.84 / - / **97.90**	81.50 / 81.49 / 95.05 / 98.85 / **99.79**	81.01 / 80.95 / 90.25 / **99.57** / 98.11
	ImageNet	81.70 / 81.92 / 96.17 / - / **97.74**	83.69 / 83.82 / 99.23 / 98.18 / **99.77**	82.54 / 82.60 / 98.17 / **99.30** / 97.93
	LSUN	80.96 / 81.15 / 97.50 / - / **99.04**	82.85 / 82.98 / 99.54 / 97.72 / **99.93**	81.97 / 82.01 / 98.84 / 99.09 / **99.21**
CIFAR-10 (DenseNet)	SVHN	87.71 / 86.80 / 95.75 / - / **98.92**	93.51 / 94.51 / 98.96 / 89.33 / **99.94**	94.60 / 95.20 / 97.21 / 94.01 / **99.99**
	ImageNet	84.93 / 91.16 / **96.83** / - / 90.07	91.50 / 96.93 / **99.45** / 86.17 / 96.44	91.02 / 96.85 / **99.29** / 87.20 / 96.41
	LSUN	85.58 / 91.91 / 98.08 / - / **99.20**	92.09 / 97.51 / 99.74 / 92.77 / **99.96**	91.71 / 97.51 / 99.72 / 95.57 / **99.96**
SVHN (DenseNet)	CIFAR-10	86.61 / 86.25 / 96.50 / - / **98.34**	91.90 / 91.76 / 99.02 / 90.63 / **99.82**	92.12 / 92.38 / 94.40 / 94.30 / **98.36**
	ImageNet	90.22 / 90.31 / **98.89** / - / 98.21	94.78 / 95.08 / **99.84** / 87.51 / 99.81	94.76 / 95.53 / **99.37** / 92.15 / 98.16
	LSUN	89.14 / 89.13 / 99.49 / - / **99.52**	94.12 / 94.52 / 99.92 / 83.45 / **99.96**	94.32 / 95.20 / 99.54 / 89.33 / **99.70**

evaluation, we measure detection accuracy, area under receiver operating characteristic curve (AUROC), and area under precision-recall curve (AUPR). The results of OOD detection are reported in Table 2 with other state-of-the-art methods, including Baseline (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Mahalanobis (Lee et al., 2018), and Energy (Liu et al., 2020). For Energy method, detection accuracy is omitted as they do not specifically determine the threshold values for energy scores for OOD detection.

We observe that our method outperforms other state-of-the-art methods in AUROC when CIFAR-10 is used as in-distribution and SVHN as OOD, and SVHN as in-distribution and all others as OOD for ResNet. For DenseNet, our method performs better in every experiment except with TinyImageNet as OOD. Our method is especially effective when there is a larger difference in the complexity of in-distribution and OOD datasets. The results support our intuition behind the *effective expressivity* of a trained network based on its training data. The models trained on CIFAR-10 have better *effective expressivity* than those trained on the simpler SVHN, and they are capable of extracting more diverse features. The models trained on SVHN would require larger updates to handle inputs of higher complexity, exhibiting more apparent gap in the model responses and resulting in the better performance in OOD detection.

It is interesting to note the performance of our gradient-based approach when TinyImageNet is used as OOD for both architectures. With SVHN as the in-distribution dataset and TinyImageNet as OOD, our approach achieves the best performance, if not a very close second. However, with CIFAR-10 as in-distribution, our approach shows inferior performance to the Mahalanobis method. We provide samples images of each dataset for visual analysis in Fig. 4. With CIFAR-10 as in-distibution, TinyImageNet shows the most visual similarity with CIFAR-10 in the types and scales of objects in the images among the three OOD datasets shown in Fig. 4 (b)-(d). SVHN is a street-view house number dataset, which is vastly different from CIFAR-10 of natural images. LSUN is a scene understanding dataset with scene categories as classifications. Our experiment results show that the more similar the in-distribution and OOD datasets, the smaller the gap in the spreads of gradient magnitudes, which makes it more challenging to differentiate OOD samples from in-distribution samples. Overall, we remark that our gradient-based method outperforms all activation-based methods in most cases.

(a) CIFAR-10 (b) SVHN (c) ImageNet (d) LSUN

Figure 4: Sample dataset images.

5. Conclusion

In this paper, we propose to examine the *effective expressivity* of neural networks with backpropagated gradients for adversarial and out-of-distribution detection. With gradients, we characterize the anomaly in inputs seen during inference based on the lack of knowledge of the model, rather than with its learned features. We introduce *confounding labels* as a tool to elicit gradient response to remove the dependency on ground truth labels during inference. We show that our approach outperforms state-of-the-art methods for adversarial and out-of-distribution detection with no need for hyperparameter tuning or additional processing.

References

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Duarte, K., Rawat, Y., and Shah, M. Plm: Partial label masking for imbalanced multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2739–2748, 2021.

Durand, T., Mehrasa, N., and Mori, G. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 647–657, 2019.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations*, 2017.

Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. On the limitation of convolutional neural networks in recognizing negative images. In *IEEE International Conference on Machine Learning and Applications*, pp. 352–358, 2017.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *International Conference on Learning Representations Workshop*, 2017.

Kwon, G., Prabhushankar, M., Temel, D., and AlRegib, G. Distorted representation space characterization through backpropagated gradients. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2651–2655. IEEE, 2019.

Kwon, G., Prabhushankar, M., Temel, D., and AlRegib, G. Backpropagated gradient representations for anomaly detection. *arXiv preprint arXiv:2007.09507*, 2020.

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

Lee, J. and AlRegib, G. Gradients as a measure of uncertainty in neural networks. In *IEEE International Conference on Image Processing*, 2020.

Lee, J. and AlRegib, G. Open-set recognition with gradient-based representations. In *IEEE International Conference on Image Processing*, 2021.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of International Conference on Learning Representations*, 2018.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *Proceedings of the International Conference on Learning Representations*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2017.

Mahdavifar, S. and Ghorbani, A. A. Application of deep learning to cybersecurity: A survey. *Neurocomputing*, 347:149–176, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Oberdiek, P., Rottmann, M., and Gottschalk, H. Classification uncertainty of deep neural networks based on gradient information. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pp. 113–125. Springer, 2018.

Prabhushankar, M. and AlRegib, G. Extracting causal visual features for limited label classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3697–3701. IEEE, 2021.

Prabhushankar, M., Kwon, G., Temel, D., and AlRegib, G. Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3289–3293. IEEE, 2020.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

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

Temel, D., Kwon, G., Prabhushankar, M., and AlRegib, G. Cure-tsr: Challenging unreal and real environments for traffic sign recognition. *NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems*, 2017.

Temel, D., Lee, J., and AlRegib, G. Cure-or: Challenging unreal and real environments for object recognition. In *IEEE International Conference on Machine Learning and Applications*, pp. 137–144, 2018.

Temel, D., Lee, J., and AlRegib, G. Object recognition under multifarious conditions: A reliability analysis and a feature similarity-based performance estimation. In *IEEE International Conference on Image Processing*, pp. 3033–3037, 2019.

Tokozume, Y., Ushiku, Y., and Harada, T. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5486–5494, 2018.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *Proceedings of the International Conference on Learning Representations*, 2017a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.

Zinkevich, M. A., Davies, A., and Schuurmans, D. Holographic feature representations of deep networks. In *UAI*, 2017.