CLASS RELEVANCE LEARNING FOR OUT-OF-DISTRIBUTION DETECTION

Butian Xiong †, *Liguang Zhou* †, *Tin Lun Lam* *, *Yangsheng Xu*

Chinese University of Hong Kong, Shenzhen, China,
Shenzhen Institute of Artificial Intelligence and Robotics for Society

ABSTRACT

Image classification plays a pivotal role across diverse applications, yet challenges persist when models are deployed in real-world scenarios. Notably, these models falter in detecting unfamiliar classes that were not incorporated during classifier training, a formidable hurdle for safe and effective real-world model deployment, commonly known as out-of-distribution (OOD) detection. While existing techniques, like max logits, aim to leverage logits for OOD identification, they often disregard the intricate interclass relationships that underlie effective detection. This paper presents an innovative class relevance learning method tailored for OOD detection. Our method establishes a comprehensive class relevance learning framework, strategically harnessing interclass relationships within the OOD pipeline. This framework significantly augments OOD detection capabilities. Extensive experimentation on diverse datasets, encompassing generic image classification datasets (Near OOD and Far OOD datasets), demonstrates the superiority of our method over state-of-the-art alternatives for OOD detection. The code is open source and can be found on GitHub at CRL.

***Index Terms*—** out-of-distribution, class relevance learning, image classification

1. INTRODUCTION

Image classification is a well-studied task in computer vision and robotics. It aims to recognize the image with a trained classifier. Various methods have been developed to better represent image representations, with varying levels of success. These methods can be divided into different categories, including the development of stronger network architectures [1], semantic-based enhancer [2][3], and multi-modality learning [4].

Despite these successes in image classification tasks, many classifiers fail to generalize to an open set setting, wherein an image from an unknown class is mistakenly classified as a known class. For instance, the parking lot is erroneously classified as a garage due to the presence of a car in the space; the difference is that the car is parked on the road as opposed to a garage. Similarly, the restaurant was mistakenly classified as a dining room due to the presence of many dining tables in the room - a feature that is more commonly seen in restaurants compared to a home dining room.

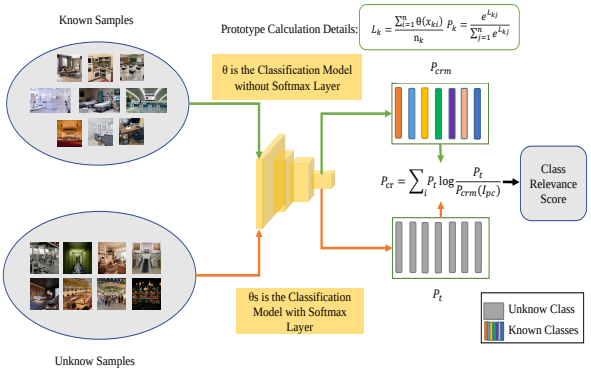

Fig. 1. Proposed class relevance learning framework for measuring the class relevance score of a test sample to the constructed class relevance matrix of training dataset.

There are various methods emerged for OOD detection [5, 6, 7, 8, 9]. Recent studies have sought to address the challenge of out-of-distribution detection by introducing an intra-class splitting method [10]. This technique aims to create atypical subsets of the known classes that can be used to model the unknown abnormal classes [11]. However, this method tends to increase the risk of falsely rejecting known classes as unknown classes.

In out-of-distribution (OOD) detection, many methods have been proposed to identify samples from a distribution different from the training dataset. Common strategies include using the maximum softmax probability [12] or max logits [13]. It is assumed that if a sample is correctly classified, its maximum value could be exploited for OOD detection. However, these methods neglect the relevance of class relationships, which is also important for analyzing OOD

* Corresponding Author: Tin Lun Lam (tllam@cuhk.edu.cn)
† Authors contributed equally to this work.

This work was partly supported by the National Natural Science Foundation of China under Grant 62073274, the Shenzhen Science and Technology Program under the Grant JCYJ20220818103000001, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202101103.

samples. Standardized max logits [14], show their findings that max logits on the range of max logits to the predicated classes. This phenomenon causes unexpected semantic classes predicated as a certain class. Therefore, standardization technology is applied to address this problem. These methods show the great potential for using the output of the model for OOD detection.

However, these methods have only considered the logits/softmax probability of the class itself when judging out-of-distribution (OOD) samples, neglecting the relevance of class relationships which is essential for OOD judgment. To address this, we propose a class relevance learning framework to learn prototypes of each class into two levels. At the logits level, the maximum class logits is utilized, while at the same time, the class relevance prototype is developed to capture the relationship between different classes. This framework takes into account the relevance of class relationships, thus allowing for a more comprehensive judgment of OOD samples.

Our main contributions are summarized as follows: 1) We propose a simple yet effective post-processing method, namely class relevance learning to statistically compute the class relevance matrix for in-distribution (ID) classes. 2) Different from previous methods, including MSP, max logits, and standardized max logits, that merely exploit the logits/softmax probability on the class level, we first take the class relevance matrix into consideration. 3) Extensive experimental results on diverse image classification datasets verify the superior performance of the proposed method for OOD detection.

2. METHODOLOGY

2.1. Problem Statement

Out-of-distribution detection is a learning problem wherein a model is trained with an ID dataset of labeled images, denoted as D_{ID}, to recognize various categories, represented as $C_i = \{1, 2, ..., n_i\}$, where n_i is the number of classes in the training set. During the test, the model is tested on both ID and OOD datasets, with some categories not present in the training set, denoted as D_{OOD}. The number of classes in the test set is denoted as $C_o = \{1, 2, ..., n_i, .., n_o\}$, where n_o is the total number of classes in the test set, and the difference between n_o and n_i indicates the number of OOD classes. The model should be able to recognize the known samples in the ID classes while detecting the unknown samples in the OOD classes.

2.2. Class Relevance Learning

Traditional OOD detection algorithms only have a fixed prototype that merely uses the network outputs, such as logits/softmax probability, which is limited and does not exploit class relationships. To address this, we propose a novel class relevance learning framework to statistically describe the class relevance among each class in the D_{ID} dataset after the training phase. Then, the class relevance matrix of the training dataset is statistically established.

As indicated in Fig.1, the model is firstly trained on the D_{ID} with known samples. Then, the parameters of the model are fixed. After that, the classification model without the softmax layer, denoted as θ, is obtained. The θ is utilized to obtain the class prototype by averaging the output logits of each class in the known samples as L_k, where n_k is the number of samples of $k-th$ class in the training dataset, and x_{ki} is i-th images in the $k-th$ classes.

$$L_k = \frac{\sum_{i=1}^{n_k} \theta(x_{ki})}{n_k} \in R^{n_i \times 1} \tag{1}$$

$$P_k = \sigma(L_k) \in R^{n_i \times 1} \tag{2}$$

$$P_{crm} = \{P_1, ..., P_{ni}\} \in R^{n_i \times n_i} \tag{3}$$

Then, the softmax version of average logits L_k is calculated as P_k with softmax function σ. In particular, the output P_k is the prototype probability of $k-th$ class. We iteratively calculate the prototype of each class following this process. Finally, the class relevance matrix, denoted as P_{crm}, is constructed through this process, where each row represents the prototype probability of a certain class.

After obtaining the class relevance matrix, we can calculate the distance between a test sample and the class relevance matrix. First, the softmax of a test sample can be obtained and defined as P_t, where the θ_s is the classification model with the softmax layer and z is a test image of size 224*224*3. I_{pc} means the pseudo class index that best matches the input sample z. The reason why it is pseudo class is that it may not be the true class of the input sample. Then, to check out where is the closest class to the test sample, index of pseudo class I_{pc} of a test sample is introduced by taking the Argmax function on P_t, where the class number of most likely ID classes is identified. Hence, we can estimate the distance between a test sample to the prototype of the most relevant class for the ID dataset by looking at the class relevance matrix P_{crm}.

$$\begin{aligned} P_t &= \theta_s(z) \in R^{n_i \times 1} \\ I_{pc} &= argmax(\theta_s(z)) \\ P_{cr} &= \sum P_t \log \frac{P_t}{P_{crm}(I_{pc})} \end{aligned} \tag{4}$$

The class relevance score, denoted as P_{cr}, quantizes the distance between a test sample to its pseudo class prototype. P_{cr} serves as a measurement for the system to choose which sample is OOD and which is ID. If P_{cr} is large, the distance between a test sample to its pseudo class prototype is large. In other words, this sample is very likely to be an OOD sample. Instead, if P_{cr} is small, it means the relevance of the test sample and its pseudo class prototype is small, which indicates the test sample is very likely to be an ID sample.

$$P_{cf} = -max(\theta(z)) * \alpha - \frac{1}{P_{cr}} * \beta \quad (5)$$

In addition to introducing the class relevance score, we also preserve the maximum logits as a complementary score, which can be regarded as the class score. The final sample OOD score, denoted as P_{cf}, quantifies the degree to which a sample may be considered an OOD sample, where α is 5, β is 0.5 by default. These parameters control the influence of max logits and class relevance score for OOD detection. Lower values of P_{cf} correspond to higher likelihoods of an ID sample, whereas higher values of P_{cf} correspond to greater likelihoods of an OOD sample.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

Datasets To examine the generalization ability of the proposed method for OOD detection, we used the generic image classification datasets, including the CIFAR10 as an ID dataset. Two benchmarks were used for OOD detection: Near-OOD and Far-OOD [15]. In the Near-OOD setting, ID samples are prevented from being wrongly introduced into OOD sets. The Near-OOD dataset contains samples that are closer to the ID dataset and are therefore more difficult to distinguish from it. The Far-OOD dataset, on the other hand, contains samples that are more easily distinguished from the ID dataset, as they are significantly different. The CIFAR100 [16] and TinyImageNet [17] datasets were used as Near-OOD datasets. The MINIST [18], SVHN [19], Texture [20], and Places365 [21] datasets are used as far OOD datasets after 1,305 images were removed due to semantic overlap. The detailed test set split for OOD can be found in [15].

Evaluation Metrics The performance of OOD detection is evaluated using the following metrics: the false positive rate at 95% true positive rate (FPR95) and area under the receiver operating characteristic curve (AUROC).

3.2. Experimental Results

In this section, we present the results of the proposed CRL method in comparison to several state-of-the-art techniques on various datasets using ResNet18 as the backbone network. The goal is to assess the performance of CRL, in the context of OOD detection. Table 1 presents a detailed comparison of CRL (Ours), with three other prominent techniques in the field: ODIN [22], DICE [23], and SHE [24]. The table showcases the performance of these methods on various datasets and provides insights into their effectiveness in OOD detection.

In the Near-OOD setting, CRL achieves an outstanding FPR95 of 42.34%, significantly outperforming ODIN (73.89%), DICE (68.81%), and SHE (77.94%). Moreover, CRL exhibits an impressive AUROC score of 89.70%, surpassing the competitors by a substantial margin. In the challenging Far-OOD scenario, CRL maintains its superiority with an FPR95 of 31.77%, substantially lower than ODIN (60.34%), DICE (55.64%), and SHE (74.85%), accompanied by a remarkable AUROC score of 91.48

In conclusion, our proposed method, CRL, demonstrates remarkable performance in OOD detection across a diverse range of datasets. It consistently outperforms existing state-of-the-art techniques, as evidenced by the substantial reduction in FPR95 and the high AUROC scores. This suggests that CRL has the potential to significantly enhance the reliability and robustness of AI systems, making it a valuable contribution to the field of OOD detection.

Ablation Study We present the results of an ablation study designed to assess the contributions of our method. We compare CRL, against the baseline method: Maxlogits [13]. The evaluation is performed on two settings of the out-of-distribution (OOD) dataset: Near-OOD and Far-OOD. The primary goal is to analyze the impact of class relevance information components on the performance of OOD detection. Table 2 presents the results of the ablation study. In the Far-OOD setting, Maxlogits exhibits slightly lower FPR95 at 48.63% and an AUROC of 89.85%. Once again, CRL stands out by achieving the lowest FPR95 of 31.77% and the highest AUROC of 91.48%. This highlights the robustness and effectiveness of CRL in identifying Far-OOD samples. In the more challenging Near-OOD setting, Maxlogits performs worse with an FPR95 of 60.02% and an AUROC of 87.84%. Notably, our proposed CRL outperforms Maxlogits with an impressive FPR95 of 42.34% and an AUROC of 89.70%. These results demonstrate the superiority of CRL in effectively detecting samples that are near the in-domain distribution.

In Table 3, we delve into the impact of two key hyperparameters, α, and β, on the performance of our model in OOD detection tasks. Focusing on the case with a fixed $\alpha = 5.0$, we observe that varying β has a discernible effect on both Near-OOD and Far-OOD AUROC scores. Notably, as β increases from 0.5 to 5.0, both Near-OOD and Far-OOD AUROC scores exhibit a consistent upward trend, culminating in peak performances of 89.70 for Near-OOD and 91.48 for Far-OOD at $\beta = 5.0$. These findings emphasize the crucial role of hyperparameter fine-tuning, particularly with respect to β, in enhancing the robustness of our model for OOD detection tasks.

Visualization of confidence scores The elucidation of distinctions between max logits and class relevance learning is explicated in Figure 2. The ResNet18 architecture is employed as the foundational neural network, undergoing training on the CIFAR10 dataset. Subsequently, an evaluation is conducted on the TinyImageNet dataset [17] and the Places365 dataset, adhering to the protocol established by Yang et al. [15]. In the leftmost column of the figure, the con-

Table 1. Comparison of our method with other state-of-the-art approaches using various datasets. The ID dataset is CIFAR10, while the OOD dataset consists of two settings: Near-OOD and Far-OOD. The backbone network employed is ResNet18.

OOD Dataset	ODIN [22]		DICE [23]		SHE [24]		CRL (Ours)	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR100	74.80	83.21	71.41	77.17	79.34	80.67	46.47	88.77
TinyImageNet	72.98	84.83	66.21	78.83	76.54	83.34	38.22	90.63
Near-OOD	73.89	84.02	68.81	78.00	77.94	82.01	**42.34**	**89.70**
MNIST	41.27	92.47	45.62	82.01	70.96	83.87	29.10	91.87
SVHN	69.01	85.56	30.79	91.06	66.70	85.33	27.72	91.96
Texture	57.08	89.22	68.34	79.28	81.56	82.73	28.40	92.07
Places365	74.02	84.58	77.79	74.02	80.18	81.34	41.84	90.00
Far-OOD	60.34	87.96	55.64	81.59	74.85	83.32	**31.77**	**91.48**

Table 2. Ablation study. The comparison between the Maxlogits and CRL is displayed.

OOD Dataset	Maxlogits [13]		CRL (Ours)	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR100	64.97	86.60	46.47	88.77
TinyImageNet	55.08	89.07	38.22	90.63
Near-OOD	60.02	87.84	**42.34**	**89.70**
MINIST	43.23	90.66	29.10	91.87
SVHN	43.68	90.37	27.72	91.96
Texture	45.67	90.30	28.40	92.07
Places365	61.93	88.08	41.84	90.00
Far-OOD	48.63	89.85	**31.77**	**91.48**

Table 3. Ablation study. The selection of hyperparameters.

Parameters		Near-OOD	Far-OOD
α	β	AUROC↑	AUROC↑
1.0	0.5	89.15	90.92
2.0	0.5	89.27	91.04
5.0	0.5	89.42	91.20
5.0	0.7	89.47	91.25
5.0	1.0	89.52	91.30
5.0	3.0	89.66	91.43
5.0	5.0	**89.70**	**91.48**

fidence score distribution of the proposed Class Relevance Learning (CRL) method is portrayed, while the rightmost column illustrates the Maxlogits approach.

Evidently, the CRL method manifests a superior confidence distribution for out-of-distribution (OOD) detection. A discernible distinction arises, wherein the majority of OOD samples are conspicuously distinguished. Furthermore, the demarcation between in-distribution (ID) samples and OOD samples is more distinctly discerned. In contrast, the Maxlogits approach exhibits a higher proportion of samples that reside in the ambiguous region between the ID and OOD boundaries, rendering them challenging to differentiate.

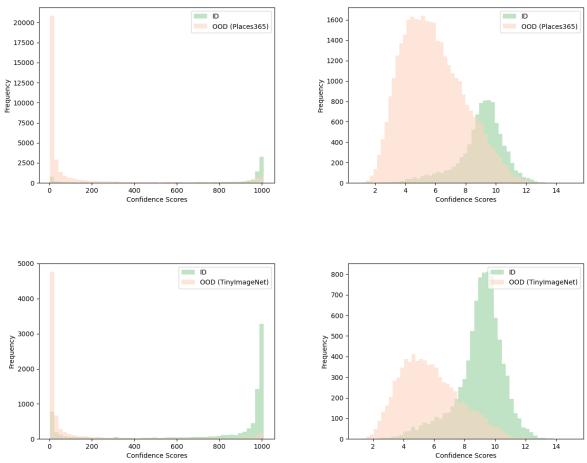

Fig. 2. This figure displays the difference between max logits and class relevance learning. We train the ResNet18 model on the CIFAR10 dataset and test it on the TinyImageNet [17] and Places365 dataset. The left column shows the confidence score distribution of CRL and the right column shows the distribution of Maxlogits.

4. CONCLUSION

In this paper, we propose a class relevance learning for OOD detection. Unlike previous methods, which only exploit the single logits/softmax probability, we first build up the class relevance concept by statistically analyzing the inter-class relationship and constructing a class relevance matrix. During the test stage, the logits and class relevance matrix are utilized for OOD score estimation. The result will serve as an OOD score to distinguish which sample is OOD. Experiment results on diverse OOD benchmarks show CRL has superior performance than previous state-of-the-art methods.

5. REFERENCES

[1] Liguang Zhou, Yuhongze Zhou, Xiaonan Qi, Junjie Hu, Tin Lun Lam, and Yangsheng Xu, "Feature pyramid attention based residual neural network for environmental sound classification," *arXiv preprint arXiv:2205.14411*, 2022.

[2] Liguang Zhou, Jun Cen, Xingchao Wang, Zhenglong Sun, Tin Lun Lam, and Yangsheng Xu, "Borm: Bayesian object relation model for indoor scene recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 39–46.

[3] Bo Miao, Liguang Zhou, Ajmal Saeed Mian, Tin Lun Lam, and Yangsheng Xu, "Object-to-scene: Learning to transfer object knowledge to indoor scene recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2069–2075.

[4] Liguang Zhou, Yuhongze Zhou, Xiaonan Qi, Junjie Hu, Tin Lun Lam, and Yangsheng Xu, "Attentional graph convolutional network for structure-aware audio-visual scene classification," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[5] Jinggang Chen, Xiaoyang Qu, Junjie Li, Jianzong Wang, Jiguang Wan, and Jing Xiao, "Detecting out-of-distribution examples via class-conditional impressions reappearing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Xiongjie Chen, Yunpeng Li, and Yongxin Yang, "Batch-ensemble stochastic neural networks for out-of-distribution detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] Sabri Mustafa Kahya, Muhammet Sami Yavuz, and Eckehard Steinbach, "Mcrood: Multi-class radar out-of-distribution detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] Zaharah Bukhsh and Aaqib Saeed, "On out-of-distribution detection for audio with deep nearest neighbors," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] Yuhongze Zhou, Issam Hadj Laradji, Liguang Zhou, and Derek Nowrouzezahrai1, "Osm: An open set matting framework with ood detection and few-shot learning," in *British Machine Vision Conference (BMVC)*, 2022.

[10] William HB Smith, Michael Milford, Klaus D McDonald-Maier, Shoaib Ehsan, and Robert B Fisher, "Openscenevlad: Appearance invariant, open set scene classification," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[11] Patrick Schlachter, Yiwen Liao, and Bin Yang, "Open-set recognition using intra-class splitting," in *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

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

[13] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song, "Scaling out-of-distribution detection for real-world settings," *International Conference on Machine Learning (ICML)*, 2022.

[14] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo, "Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15425–15434.

[15] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al., "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32598–32611, 2022.

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

[17] Ya Le and Xuan Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, pp. 3, 2015.

[18] Li Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

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

[20] Gustaf Kylberg, *Kylberg texture dataset v. 1.0*, Centre for Image Analysis, Swedish University of Agricultural Sciences and ..., 2011.

[21] 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*, vol. 40, no. 6, pp. 1452–1464, 2017.

[22] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *International Conference on Learning Representations (ICLR)*, 2018.

[23] Yiyou Sun and Yixuan Li, "Dice: Leveraging sparsification for out-of-distribution detection," in *European Conference on Computer Vision (ECCV)*, 2022.

[24] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al., "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.