17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Open-set Recognition with Supervised Contrastive Learning

Yuto Kodama†, Yinan Wang†, Rei Kawakami‡, Takeshi Naemura†
† The University of Tokyo, Tokyo Institute of Technology
{kodama, wangyn, naemura}@hc.ic.i.u-tokyo.ac.jp, reikawa@c.titech.ac.jp

Abstract

Open-set recognition is a problem in which classes that do not exist in the training data can be presented at test time. Existing methods mostly take a multi-task approach that integrates N-class classification and self-supervised pretext tasks, and they detect outliers by examining the distance to each class center in the feature space. Instead of relying on the learning through reconstruction, this paper explicitly uses distance learning to obtain the feature space for the open-set problem. In addition, although existing methods concatenate features from multiple tasks to measure the abnormality, we calculate it in each task-specific space independently and merge the results later. In experiments, the proposed method partially outperforms the state-of-the-art methods with significantly fewer parameters.

1 Introduction

Closed-set recognition relies on an implicit assumption that the data presented during testing is covered by the training data. However, this assumption is unrealistic for applications that require interaction with the real world, such as autonomous driving. Open-set recognition, therefore, addresses the problem that unknown data can be presented in addition to known data. It is an *N+1*-class classification problem, where *N known* classes and one *unknown* class must be identified at testing, while no information about the *unknown* class is available at training.

To solve this problem, how to obtain feature representations that are effective for known classes but also for unknown classes becomes an issue. To obtain such good features, existing studies have regularized classification networks using other tasks such as reconstruction or other self-supervised pretext tasks. However, it is still not obvious what kind regularization is the most effective. In addition, a network using reconstruction requires a decoder to generate images from features, which increases computational cost and memory usage.

In this paper, we show that it is possible to obtain useful feature representations by using supervised contrastive distance learning [8] without learning reconstruction. Supervised contrastive learning prepares two images with different data augmentations, and they are learned to become closer or farther apart in feature space depending on their labels. Since images with same labels deformed by different data augmentations are learned to have the same features, the feature space may be more robust to such deformations and may have a more adequate representation to detect anomaly than the features learned with reconstruction. We also present that calculating the anomaly score in each task-specific space and merging the score later by averaging is better than concatenating the features from multiple tasks and calculate the score in a unified space. Experiments show that the proposed method partially outperforms the state-of-the-art methods with significantly fewer parameters on standard open-set/out-of-distribution data sets such as SVHN and Tiny-Imagenet.

2 Related work and preliminaries

Open-set recognition In anomaly detection, autoencoders (AEs), variational autoencoders (VAEs), and generative adversarial nets (GANs) are trained to minimize the reconstruction error [19, 1, 7]. They assume the reconstruction error will be large for unknown classes; however, if the AEs are trained with many classes and generalized well, they may be able to reconstruct the unknowns as well. Another way to solve open-set classification is to extend Softmax and use $1 - \max(y_i)$ as the anomaly score for the class i, where y_i is the output from Softmax [2]. We will refer to this method as Softmax*. It assumes that the confidence will not be high for unknown classes, while this is often not valid [6].

OpenMax [2] utilizes the extreme value theory (EVT) and shows how to calculate anomaly scores in a feature representation learned by classification. Many followers are more focused on how to obtain good representation for being able to classify knowns and detect outliers simultaneously. CROSR [21] and C2AE [14] add AEs to classifiers in a multi-task learning manner. CGDL [20] replaces AEs in CROSR [21] to VAEs. GDFR [15] regularizes the classification space by adding a self-supervised task to estimate rotations. To enrich the input representation, an AE is attached to the front end of the classifiers and they are cascaded. OpenHybrid [22] replaces AEs with a Flow network that can measure the likelihood of data in the latent space. It is currently the best performing network because of its ability to represent long-tail distributions, but the network size is large because of the flow network.

Out-of-distribution detection (OoDD) The out-of-distribution (OoD) detection and open-set recognition (OSR) deals with similar problems. The OSR targets to detect unknown classes in the same domain of the training dataset, while OoDD focuses on the detection of samples from a different dataset/domain. Early method is based on softmax thresholding, also showing that the OoD samples may have high softmax values. In [11], they increase the difference between the softmax value distribution of in-distribution (ID) and

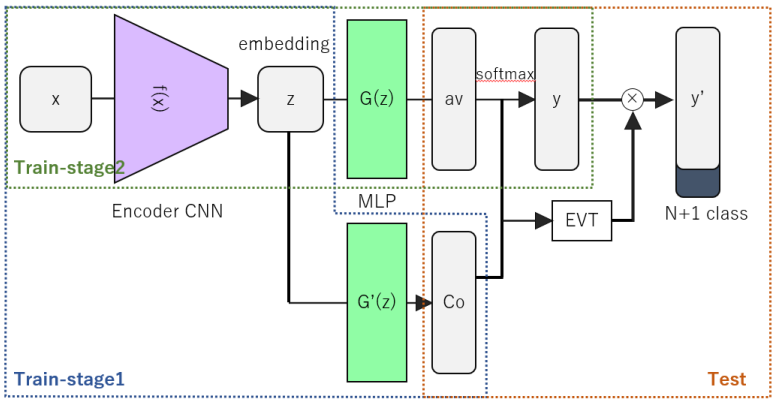

Figure 1. The proposed network. First, we pre-train an encoder by supervised contrastive learning, and then we train the MLP for classification by freezing the encoder's weights. At test time, the anomaly score for each class is calculated using the features from the two tasks and is converted to N+1-class probability.

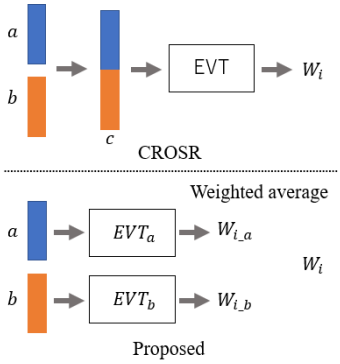

Figure 2. While CROSR calculates the anomaly score from a single vector that includes multiple features, the proposed method takes the weighted sum of the anomaly scores calculated from each feature.

OoD samples by adding perturbation to the input and adding temperature scaling to softmax. Gaussian distribution analysis is utilized to model the distribution of each layer output of DNN [10]. The mahalanobis distance between the feature maps of a test sample and the closest class-conditional Gaussian distribution are examined, and the confidence score is the weighted average of the distances. Some studies utilized generative models trained on ID data with a view that the likelihood of the OoD sample will be smaller than that of ID samples, but a study shows that such hypothesis is not valid [12]. Ren et al. [16] argue that this problem is due to the background that has an influence on the likelihood of the semantic part of the image; thus, an additional generative model focusing on the background is trained and an alternative metric is defined to reduce the influence of the background.

Contrastive learning Contrastive learning explicitly learns distance so that positive samples (e.g., samples from the same class) become closer and vice versa. It has been extensively studied in applications such as face recognition and query-based similarity search. Recently, it draws more attention since metric learning with self-supervision shows excellence in obtaining well-generalized representation. Several metric-learning loss are proposed, such as triplet, N-pair, max-margin, and anchor-based methods, depending on the sampling method of the data in the batch and the loss design. The state-of-the-art methods in contrastive learning are SimCLR [3], Moco [4], BYOL [5] etc., where two images (views) with different augmentations are generated from a single image in a batch, and learning is performed in such a way that the distance between the pair becomes closer and the others becomes farther apart. The contrastive loss is as follows:

$$\mathcal{L} = -\sum_{i \in I} \log \frac{\exp\left(z_i \cdot z_{j(i)}/\tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a/\tau\right)}, \qquad (1)$$

where z_i is the feature of i-th sample, a is a sample in A which is a set of negatives to z_i. $z_{j(i)}$ is a positive to i-th sample, and τ is a scaling factor.

Obtaining anomaly score Rather than simple sofmax thresholding, we follow OpenMax [2] to calculate anomaly scores. OpenMax is based on EVT [17], which is one of the methods of meta-recognition. An ordinary classification network is trained on the training dataset, and the activation vectors (AV) before the softmax layer are extracted; namely, $AV = f(x)$ where x is a feature from the previous layer. The mean average vectors (MAV) for each class and the distances between MAV and AV for each class are calculated. The distances are fitted to the Weibull distribution using EVT. During testing, the distance calculated from the MAV and AV of the input image is examined and the degree of anomaly for each class is calculated; the degree of anomaly is weighted by the probabilities of the N classes, and their weighted sum becomes the anomaly score. In Softmax, each class confidence y_i is obtained as follows:

$$y_i = \text{Softmax}(AV)_i = \frac{exp(AV_i)}{\sum_{j=1}^{N} exp(AV_j)}. \qquad (2)$$

In OpenMax,

$$w_i = \text{EVT}(||AV - MAV_i||_2) \ (1 \le i \le N), \qquad (3)$$
$$\hat{y}_i = \begin{cases} y_i w_i & (i \le N), \\ \sum_{i=1}^{N} y_i(1 - w_i) & (i = N+1). \end{cases}$$

In CROSR [21], OpenMax is extended to use intermediate outputs of another network in addition to AV:

$$c = \text{concat}([AV, z_1, z_2, ...]) \qquad (4)$$
$$w_i = \text{EVT}(||c - c_i||_2) \ (1 \le i \le N)$$

3 Proposed Method

Utilizing supervised contrastive learning We propose a two-step learning process that pretrains the network using contrastive learning. Networks trained with supervised contrastive learning make the features of augmented images to be close to each other, and

thus can be more aware of semantic concepts invariant to such augmentations compared to reconstruction networks. This paper utilizes supervised contrastive learning [8], which use SimCLR with a single encoder. The learning process is as follows. In the first stage, we perform supervised contrastive learning, as shown in Fig. 1. Because of the limited space, the loss function for positive pairs are shown:

$$\mathcal{L}_{in}^{sup} = \sum_{i \in I} -\log \left(\frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \right), \tag{5}$$

where $P(i)$ is the set of images that have the same label as i. For negative pairs, the log operation is applied before taking the average over I as the number of negatives are much larger. The inner product of features is used as the distance.

In the second stage, we train the MLP for classification using the features obtained by the encoder trained in the first stage. The vectors obtained in supervised contrastive learning has higher dimension than the number of classes N; thus, the MLP finds the best mapping between those vectors and N-dimensional vector.

Anomaly scores in each task space At the time of testing, the unknown detection is performed using the features of the learned network. In CROSR, distances of a sample to each MAV are calculated using vectors that concatenate features from multiple outputs. In the proposed method, EVT is applied to the individual features to be used, and the weighted sum of the anomaly scores calculated from each feature is used as the final anomaly score. In this way, the influence of differences in the distribution of values in each feature can be removed. The inner product is used as the distance function to match the loss function of the contrastive learning. As shown in Fig. 2, suppose we have feature vectors a and b from different tasks and their MAVs. Then, the anomaly score W_i for i-th sample is calculated as follows:

$$W_a(i) = EVT_{a_i}(a_i \cdot MAV_{a_i}) \tag{6}$$

$$W_b(i) = EVT_{b_i}(b_i \cdot MAV_{b_i}) \tag{7}$$

$$W_i = mean(W_a(i), W_b(i)) \tag{8}$$

4 Experiments

Experiments were conducted on the open-set detection, which classifies known and unknown into two classes, and on the open-set classification, which classifies N known classes and one unknown class.

Datasets In the open-set detection setting, following [13], we used CIFAR10, SVHN, and Tiny-Imagenet to measure the performance. CIFAR10 and SVHN contain 10 classes each and we divided each dataset into 6 known classes and 4 unknown classes. In CIFAR+10 and CIFAR+50, we used 4 animal classes from CIFAR10 as known classes, and used 10 and 50 classes of non-animal classes from CIFAR100 [9] as unknown classes respectively. Tiny-Imagenet contains 200 classes and we randomly divided 20 classes as known classes and the other 180 classes as unknown. The image resolution was 64 times 64 pixels. The area under the ROC curve (AUROC) was used for the evaluation of open-set recognition and we took the average of five experiments to compensate stochasticity.

Table 1. Comparison of AUROC in calculating features in a unified space and task-specific spaces

	CIFAR10	SVHN	TinyImangenet
concat, norm (CROSR)	81.16	94.64	72.72
concat, inner product	83.97	95.12	76.58
split, inner product (Ours)	84.24	95.53	77.04

In the open-set classification setting, we used 10 classes of CIFAR10 as known classes. For testing, Tiny-Imagenet and LSUN are used as unknowns. There are two choices of resize or crop to match the resolution of CIFAR10; thus, we have four datasets in total. The average of F1 scores for each classes (macro-F1 score) is used for the evaluation.

Augmentation In SimCLR [3], two augmented images need to be generated. According to the paper, powerful augmentation methods such as fast-augment and auto-augment can be used for better training, but in this study, only *RandomCrop* and *Colorjitter* are used for augmentation for fair comparison to the existing methods.

Network settings As mentioned in [21], higher classification performance tends to improve the AUC. Therefore, for a fair comparison, we use the same encoder network as the existing method [13]. This encoder CNN consists of eleven 3x3 convolutional layers, with batch normalization and LeakyReLU (0.2) between each layer. In this network, beside the final output, there are three blocks where the resolution changes, and the skip connection of the encoder-decoder model is placed between these blocks. The intermediate outputs from these blocks are denoted as $z1$, $z2$, and $z3$.

In both 1st- and 2nd-stages, SGD was used to train the models, and the learning rate was set to decay from 3e-1 to 6e-5 by the cosine function in the 1st-stage, and to decay by 0.2 every 20 epochs from 5e-1 in the 2nd-stage. The vector dimension of contrastive features was 192. In the 1st-stage, all the training data were used, and in the 2nd-stage, the training data was divided to create validation data, and this validation data is used to determine the epoch for the classifier's training.

We utilized libMR[18] for the calculation of EVT, and the hyperparameter `tail-size` was set to 20, following OpenMax.

Results We selected the features for the open-set recognition, since the proposed network can obtain six types of features including intermediate outputs, such as *av*, *co* and z in Fig 1. The details of feature selection is provided in the supplementary material. The experiment shows the intermediate outputs from the encoders are not as useful *av*, *co* and z. In addition, since the embedding z varies greatly in dimension depending on the structure of the encoder CNN, only *co*

Table 2. The AUROC for unknown and known binary classification in each data set. Numbers are in percentages and are averaged by five experiments. Scores for existing methods are taken from the respective papers.

Method	CNN parameter	CIFAR10	CIFAR+10	CIFAR+50	SVHN	Tiny-Imagenet
SoftMax	1	67.7	81.6	80.5	88.6	57.7
OpenMax [2]	1	69.5	81.7	79.6	89.4	57.6
G-OpenMax [23]		67.5	82.7	81.9	89.6	58.0
OSRCI [13]		69.9	83.8	82.7	91.0	58.6
C2AE [14]	2	71.1	81.0	80.3	89.2	58.1
CROSR [21]	1.4	88.3	91.2	90.5	89.9	58.9
GDFR [15]	3	80.7	92.8	92.6	93.5	60.8
CGDL [20]	10.8	90.3	95.9	95.0	93.5	76.2
Ours	1	84.2	95.0	94.6	**95.5**	**77.0**

Table 3. The macro F1 scores obtained by the classification for each class. Scores of existing methods are taken from each paper.

Method	CNN parameter	Tiny-Imagenet Crop	Tiny-Imagenet Resize	LSUN Crop	LSUN Resize
SoftMax	1	63.9	65.3	64.2	64.7
OpenMax [2]	1	66.0	68.4	65.7	66.8
DHRNet+Softmax*	1.07	64.5	64.9	65.0	64.9
DHRNet+OpenMax	1.07	65.5	67.5	65.6	66.4
CROSR [21]	1.07	72.1	73.5	72.0	74.9
C2AE [14]	2	83.7	82.6	78.3	80.1
GDFR + activation [15]	4	75.7	79.2	75.1	80.5
GDFR + Softmax* [15]	4	82.1	77.7	84.3	80.5
CGDL [20]	8.09	84.0	83.2	80.6	81.2
Ours	1	80.9	78.9	**85.9**	**81.3**

and av are used in the following.

Next, methods of how to calculate anomaly was evaluated. Three different experiments were conducted depending on (1) whether two features should be concatenated or calculated in task-specific manner, and (2) whether feature vectors should be normalized or not. As Table 1 shows, the performance of the direct inner product was better than normalized inner product, and calculating the anomaly score in each feature is better than the space that concatenates them.

We show the performance comparison of open-set detection experiment in Table 2. G-OpenMax [23] and OSRCI [13] are methods that use unknown images generated by a generative model as input; C2AE is a method that uses reconstruction errors by training a conditional AE [14]. Among the existing studies, only those that use the same encoder network as ours are listed. CGDL uses VGG13 as the encoder but we listed it as the scores are similar to ours. As shown in Table 2, although the proposed method have fewer parameters, the performance is comparable to heavy networks that use discriminative reconstruction such as CGDL, and even outperform it on TinyImageNet and SVHN.

We show the results of open-set classification experiment in Table 3. A threshold has to be determined to categorize whether a sample is an anomaly or not for several methods including ours. In the EVT-based methods such as OpenMax, CROSR and ours, it is possible to take the argmax as the output is the probability for N+1 classes. However, GDFR [15] describes a method for determining the threshold value from the training data, and we follow the same protocol in this experiment. The scores of the existing methods are taken from each paper. The proposed method outperforms existing methods on LSUN-Crop and LSUN-Resize dataset as shown in Table 3. The proposed method uses the smallest network in the list; thus, the encoder should have lower expressive power and there is no decoder. This indicates the effectiveness of the proposed method for the open-set classification.

5 Conclusion

In this paper, we have shown that features trained by the supervised contrastive learning is effective for open-set recognition. The number of CNN parameters can be significantly reduced compared to the existing studies that utilize reconstruction. In addition, when using multiple features for anomaly detection, we have proposed a method to calculate the anomaly score from each feature separately and take the weighted sum of them, instead of concatenating them. We have also showed that the inner product should not be normalized when detecting anomalies while training is done with normalized features. Experiments on standard datasets have shown that the performance of the proposed method was comparable to SOTA with fewer parameters.

Acknowledgement

This work was in part supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair.

References

[1] S. Akcay *et al.* Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2019.

[2] A. Bendale, T. Boult. Towards open set deep networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016.

[3] T. Chen *et al.* Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[4] X. Chen *et al.* Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] J.-B. Grill *et al.* Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020.

[6] M. Hein *et al.* Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.

[7] Y. Kawachi *et al.* Complementary set variational autoencoder for supervised anomaly detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2366–2370, 2018.

[8] P. Khosla *et al.* Supervised contrastive learning. In *NeurIPS*, 2020.

[9] A. Krizhevsky *et al.* Cifar-100 (canadian institute for advanced research).

[10] K. Lee *et al.* A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7167–7177, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] S. Liang *et al.* Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv: Learning*, 2018.

[12] E. Nalisnick *et al.* Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[13] L. Neal *et al.* Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[14] P. Oza, V. M. Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[15] P. Perera *et al.* Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] J. Ren *et al.* Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019.

[17] E. M. Rudd *et al.* The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):762–768, Mar 2018.

[18] W. J. Scheirer *et al.* Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33:1689–1695, 2011.

[19] T. Schlegl *et al.* Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *IPMI*, 2017.

[20] X. Sun *et al.* Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] R. Yoshihashi *et al.* Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] H. Zhang *et al.* Hybrid models for open set recognition. In *European Conference on Computer Vision*, 2020.

[23] S. D. Zongyuan Ge, R. Garnavi. Generative openmax for multi-class open set classification. In G. B. Tae-Kyun Kim, Stefanos Zafeiriou, K. Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 42.1–42.12. BMVA Press, September 2017.