Evaluation of Out-of-Distribution Detection Performance of Self-Supervised Learning in a Controllable Environment

Jeonghoon Park[1][*], **Kyungmin Jo**[1][*], **Daehoon Gwak**[2][*], **Jimin Hong**[3],
Jaegul Choo[1], **Edward Choi**[1]
KAIST[1], Korea University[2], Humelo[3]
{jeonghoon_park, bttkm, jchoo, edwardchoi}@kaist.ac.kr
hune282@korea.ac.kr, jimin9401@gmail.com

Abstract

We evaluate the out-of-distribution (OOD) detection performance of self-supervised learning (SSL) techniques with a new evaluation framework. Unlike the previous evaluation methods, the proposed framework adjusts the distance of OOD samples from the in-distribution samples. We evaluate an extensive combination of OOD detection algorithms on three different implementations of the proposed framework using simulated samples, images, and text. SSL methods consistently demonstrated the improved OOD detection performance in all evaluation settings.

1 Introduction

An out-of-distribution (OOD) detection task is to recognize outliers or anomalies, which do not follow the distribution of the training data. To address this problem, numerous methods have been proposed in classification tasks in several modalities, including computer vision [8, 16, 17, 9, 23], natural language processing [12, 21], and two-dimensional real-number dataset (*e.g.,* Gaussian noise distribution) [19, 16, 18, 24]. Especially, [8, 16, 17] are simple yet efficient algorithms since they use existing trained models for OOD detection without additional fine-tuning with OOD samples.

Recently, the rapid progress of the Self-Supervised Learning (SSL) achieved a significant performance improvement in various tasks of computer vision [7, 3] and natural language processing [5, 1]. Also, self-supervised learning has shown its potentials in OOD detection by using contrastive learning [20, 24], masked language modeling [12] and rotation prediction [10].

However, the OOD detection performance of a model, including SSL-based ones, is typically evaluated by the AUROC score of the binary classification between in-distribution (ID) and OOD, which do not appropriately consider the characteristics between ID and OOD samples. For example, we can reasonably assume that the distance between the pixel distributions of CIFAR10 and CIFAR100 is closer than the distance between CIFAR10 and MNIST. Therefore, if the distance between ID and OOD correlates with the difficulty of OOD detection [15], then, given CIFAR10 as the ID, CIFAR100 can be considered a harder OOD dataset than MNIST.

The degree of difficulty for OOD detection was considered in some previous studies using maximum mean discrepancy between two datasets or confusion log probability [17, 24]. Additionally, [11] proposed an OOD dataset called ImageNet-O consisting of hard OOD samples with low-level statistics similar to ID (i.e., ImageNet [4]). And recently, [24, 20] argued that contrastive loss is effective for detecting harder OOD samples based on the performance using near classes or confusing examples. Although such existing methods estimate the difficulty of OOD detection to some extent,

---

[*]Equal contribution

or introduce a novel dataset to provide a more challenging setup, they cannot adjust the difficulty of OOD detection. Therefore, a principled evaluation framework which can consider the distance of two distributions (*i.e.,* the difficulty of OOD detection) is required to observe and compare the behavior of OOD detection methods to truly evaluate their performance.

In this paper, we propose a preliminary approach for a controllable OOD detection evaluation framework using datasets with three different modalities (simulated samples, images, and text). In this framework, we evaluate existing OOD detection algorithms, including SSL. Finally, we show that experimental results support the validity of our assumptions for controllable settings, and reveal interesting insights. Additionally, the results in three modalities consistently validate the effectiveness of SSL for OOD detection.

2 Performance evaluation of SSL in OOD detection

In this section, we describe the proposed controllable evaluation framework followed by the detailed experiment settings as well as evaluation results for each modality.

2.1 Controllable Evaluation Framework

We propose a controllable evaluation framework for OOD detection tasks and implement it with three different modalities: 2D simulated samples, images, and text documents[2]. The core idea of this framework is gradually adjusting a distance between ID samples and OOD samples using a distance factor.

2D Simulated Dataset   A simple simulated dataset was generated in a 2D space to validate the effects of self-supervised methods in OOD detection. Fig. 1 shows ID samples coming from two classes, and the controllable OOD samples that gradually moves away from the ID samples using the multiplying factor r. OOD samples follow the circle-shaped distribution whose radius is r times larger than the radius of ID samples, with $r = 1$ indicating the case that OOD equals ID. Binary classification between two ID classes is trained with 48,000 samples for training and 16,000 samples for validation. The OOD detection performance is evaluated using the 16,000 ID test samples and 16,000 OOD samples. In this experiment, a set of OOD samples with 50 different distances ($1 \leq r \leq 5.9$) was used to evaluate all OOD detection methods.

Images   In order to build a controllable dataset similar to the 2D circle dataset, Mixup[26] was used to generate data samples between ID and OOD by adjusting the mixing ratio of OOD to ID images (Fig. 5 in Appendix. A), under the assumption that different mixing ratios would correlate with the ID-OOD distance. We use each one of CIFAR10 and CIFAR100 as ID samples while using TinyImagenet, LSUN, SVHN, and CIFAR100 or CIFAR10 as OOD samples. For rigorous comparison, all experiments were conducted on the same pre-built mixup dataset consisting of 10,000 samples for each mixing ratio, using 10,000 ID and OOD samples only once for each.
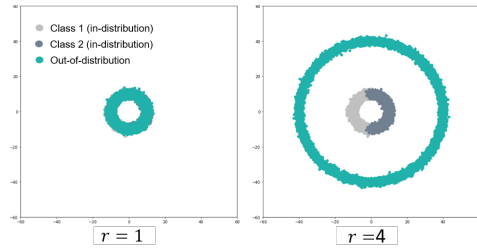

Figure 1: Examples of 2D circle dataset

2.2 Methods

Same as [10, 20, 24], we train the models using a multitask approach consisting of the supervised classification objective function L_{cls} and self-supervised objective function L_{ss}. Thus, our objective function can be written as $L_{multi} = L_{cls} + \alpha L_{ss}$, where coefficient α is a hyperparameter. In order to analyze the behavior, we conduct and evaluate the experiments in several aspects: (1) classification methods, (2) self-supervised methods, and (3) methods for OOD scoring.

Classification methods For the classification objective function L_{cls}, we use a softmax classifier with a cross-entropy loss (CE) or a One-Vs-All classifier with Distance Maximization loss (OVADM)

---

[2]All descriptions and results regarding text is provided in the appendix

Table 1: Classification accuracy(%) of in-distribution dataset

Dataset	CE	CE+SimCLR	CE+BYOL	OVADM	OVADM+SimCLR	OVADM+BYOL
2D Circle	97.44	<u>97.46</u>	97.45	**97.49**	97.45	97.44
CIFAR10	94.61	<u>94.87</u>	**95.39**	93.23	94.67	94.84
CIFAR100	76.23	76.96	**78.38**	73.8	74.38	<u>77.13</u>

[18]. We use the OVADM loss and the CE loss to analyze the effect of distance-based loss in OOD detection.

Self-supervised methods In the 2D circle dataset and the image datasets, we use the same self-supervised objective functions used in BYOL [7] and SimCLR [3]. We compare BYOL and SimCLR to compare the effect of using only positive samples when training a self-supervised task in OOD detection.

OOD scoring methods To detect OOD samples, we use three well-known OOD scoring methods, Baseline (max value of the output vector of the classifier) [8], Mahalanobis Distance(MD) [16], and ODIN [17]. The Baseline and ODIN are not distance-based, but MD is a distance-based method.

Model architecture To classify the 2D circle dataset, we used a simple network consisting of two fully-connected layers to obtain features of a penultimate layer. When training with self-supervised methods, we considered adding noise to the input to emulate the data augmentation of SimCLR and BYOL in a 2D space. The noise was sampled from a normal distribution with a mean of zero and a standard deviation of 0.01. For the experiments with image datasets, Wide ResNet [25] with the depth of 28 and the width of 10 was used. The data augmentation technique of the original SimCLR paper was applied equally to all experiments except for resizing and crop. The other training details follow the original Wide ResNet paper.

2.3 Results

We performed experiments in three different modalities and evaluated the performance of classification and OOD detection, respectively. We evaluate the classification task by its accuracy and the OOD detection performance by AUROC.

The classification accuracy is described in Table 1. The bold font indicates the highest accuracy, and the underline indicates the second highest accuracy. The classification accuracy of synthetic data is about 97%. The OVADM had the highest accuracy and secondly higher by CE+SimCLR. In the image modality, the classification accuracy is improved with self-supervised learning, but OVADM showed lower classification accuracy than CE in all combinations.

Even if the models show the same classification accuracy, the OOD detection performance could differ. The OOD detection performance for the simulated dataset is shown in Fig. 2. Naturally, distance-based methods (either using OVADM loss or MD OOD score) demonstrate a correct behavior where a larger distance between ID and OOD leads to higher AUROC. We could also see the AUROCs increase most of the time when SSL is used. The performance of OOD detection for images is shown in Fig. 3. The AUROC gradually increases as the OOD mixing ratio increases, although the slope and inflection point was different for each OOD dataset. When self-supervised learning was applied, the inflection point was reached more quickly with a steeper slope than without applying it. Comparing the three OOD scoring methods, the AUROC difference between models with and without SSL is not large for the baseline. On the other hand, ODIN shows a considerable difference in AUROC with SSL, and MD shows a large gap between the model with BYOL and the others.

3 Discussion

In 2D simulated sample, note that the models using OVADM loss and MD scoring method have the same AUROC performance at $r = 5.9$. But when r is small, the difference increases between models that use SSL and those that do not. This shows that using a fixed set of OOD data can be ineffective when comparing the OOD detection performances of different methods, which corroborates the need for a controllable OOD evaluation setup.

Models that use a distance-based method, such as OVADM or MD, increase the value of AUROC as the r increases. However, the AUROC value of models trained without a distance-based method
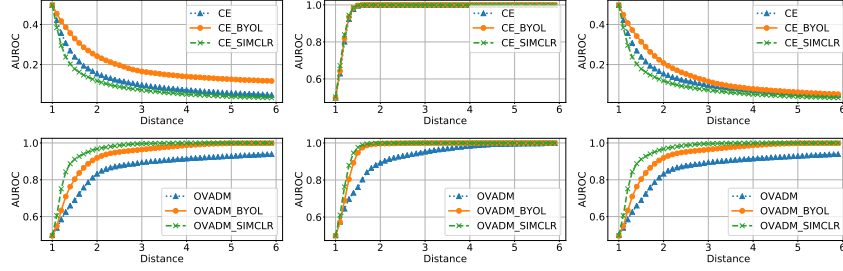
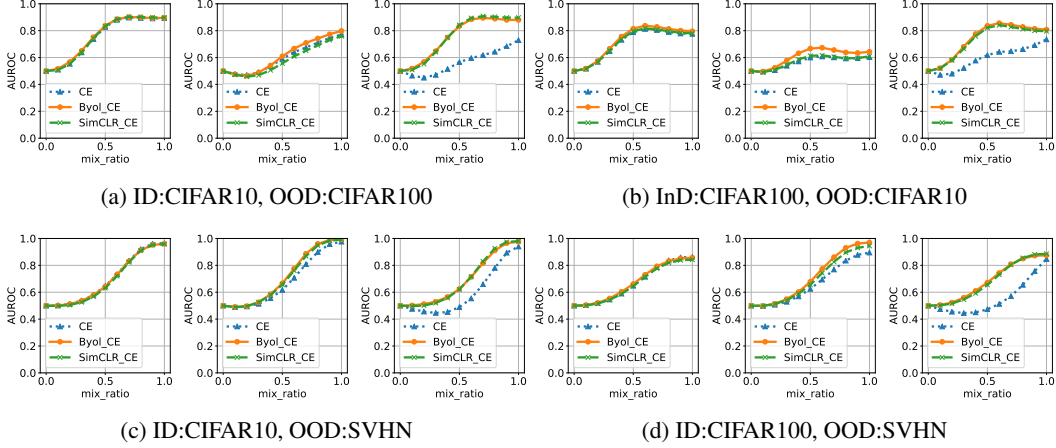Figure 2: AUROC scores in 2D dataset (Left : Baseline, Middle : MD, Right : ODIN)

(a) ID:CIFAR10, OOD:CIFAR100   (b) InD:CIFAR100, OOD:CIFAR10

(c) ID:CIFAR10, OOD:SVHN   (d) ID:CIFAR100, OOD:SVHN

Figure 3: OOD performance for images in 4 ID-OOD cases. A mixed ratio of zero means the in-distribution image. (In each case, Left: Baseline, Middle: Mahalanobis distance, Right: ODIN)

did not increase as the r value increased. As shown in Fig. 4, without the distance-based method (OVADM or MD), a model assigns a high confidence for OOD samples than models using the distance-based methods.

The assumption that the different mixing ratio correlates to the distance between ID and OOD is debatable, but Fig. 3 shows similar monotonicity as Fig. 2 (especially for more distinct dataset pairs CIFAR10-SVHN and CIFAR100-SVHN), indicating the validity of our assumption. It is notable that SSL not only consistently increases the OOD detection performance, but also guides the model to demonstrate monotonically increasing OOD detection performance as the mixing ratio increases (*i.e.* ID-OOD distance increases). Furthermore, as shown in Fig.2 and 3, the benefit of SSL on OOD detection tended to be more evident on hard samples (*i.e.* close distance between ID and OOD), which is in line with the previous claims [24, 20].
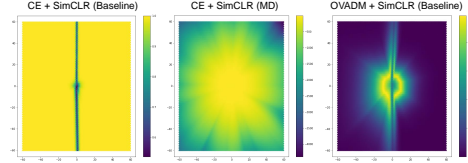

Figure 4: OOD score landscapes in 2D simulated data. (Darker means OOD)

4 Conclusions

In this paper, we proposed the controllable framework for evaluating OOD detection performance. With this framework, we compared the OOD detection performance and observed the behavior of various methods according to the distance from in-distribution in three different modalities: the 2D synthetic dataset, images, and natural language. As mentioned earlier, the setting of controllable datasets of the computer vision and natural language has limitations, and extension into a more principled and realistic manner using generative models such as VAE[13] or GAN[6] can be future work.

Broader Impact

We believe this is not applicable to us.

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. The IWSLT 2015 evaluation campaign. In *International Conference on Spoken Language*, 2015.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255, 2009.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

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

[9] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proc. the International Conference on Learning Representations (ICLR)*, 2019.

[10] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 15663–15674, 2019.

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[12] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic Rishabh Krishnan, and Dawn Song. Pre-trained transformers improve out-of-distribution robustness. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. the International Conference on Learning Representations (ICLR)*, 2014.

[14] Guilaume Lample and Alexis Conneau. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019.

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

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

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

[18] Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. *arXiv preprint arXiv:2007.05134*, 2020.

[19] 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 *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 14707–14718, 2019.

[20] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[21] Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. Out-of-domain detection for low-resource text classification tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3566–3572, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[23] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 550–564, 2018.

[24] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[25] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
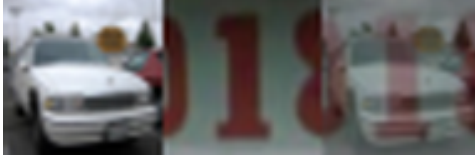
(a) CIFAR10, SVHN, Mixup	(b) CIFAR100, TinyImagenet, Mixup

Figure 5: Mixup examples in image domain

Example	Mixup rate
So what we need to do is to stand up boldly, singly and together, to push governments, to push regional fisheries management organizations, to declare our right to declare certain areas off-limits to high seas fishing, so that the freedom to fish no longer means the freedom to fish anywhere and anytime.	**0.0**
Em đã sng nhng đêm tri có ánh trăng chiu vàng. Em đã sng nhng đêm ngoài kia bin ru b cát. c gì! anh đây gi này. our right to declare certain areas off-limits to high seas fishing, so that the freedom to fish no longer means the freedom to fish anywhere and anytime.	**0.5**
Em đã sng nhng đêm tri có ánh trăng chiu vàng. Em đã sng nhng đêm ngoài kia bin ru b cát. c gì! anh đây gi này. c gì! anh cùng em chuyn trò. Cùng nhau nghe sóng xô ghnh đá ngàn câu hát yên bình.Em đã bit cô đn là th mi khi cách xa anh.	**1.0**

Figure 6: OOD examples in text domain. RED denotes Vietnamese documents used as OOD samples.

A   Dataset Details

Text To test OOD detection in a text-domain, the task of our models is to classify the language of a document into two categories: English and French. We use IWSLT 2015 dataset [2] which consists of 12 languages. We randomly select 30,000 examples for training and evaluate the model on 3,000 examples for each language as in-distribution datasets. Vietnamese documents are used as OOD samples since it is written in alphabet-base characters and linguistically different to English and French. As can be seen in Fig. 6, we construct controllable 6,000 examples by replacing consecutive sentences in in-distribution documents with Vietnamese sentences according to the mixing ratio, where mixing ratio of 1.0 equals pure OOD.

B   Implementation Details

2D simulated dataset For synthetic-based experiments, we trained a fully-connected network with 2 hidden layers which have 8 units each. The network was trained for 10 epochs on a batch size of 128 using Adam optimizer. The learning rate was 0.01. When utilizing the self-supervised method, we sampled noise values from a normal distribution with a mean of zero and a standard deviation of 0.01. These values are added to the samples to simulate the augmentation of the self-supervision method.

Text In the text-based expeirment, we took a 12-layer Transformer [22] and trained it for the language classification task. Using stochastic gradient descent with Adam optimizer, we trained the network for 5 epochs with batch size of 16 following the default implementation of HuggingFace [3]. When using the self-supervised method, unlike the other two modalities, the model was not trained in a multi-task manner. We used a pre-trained model which was sufficiently trained with self-supervised training objective and we fine-tuned it on the text classification task as [12]. We chose XLM-mlm-enfr-1024 model [14] as a pre-trained language model considering the in-distribution languages.

Image domain We used Wide ResNet[25] with depth 28 and width 10 for all the vision-based experiments. Like the Wide ResNet original paper, we trained the network for 200 epochs with a dropout rate of 0.3 and a batch size of 128. We used stochastic gradient descent with Nesterov momentum for the optimizer. The learning rate started at 0.1 and was multiplied by 0.1 at 50% and 75% of the training epoch, and the momentum was 0.9. The data augmentation technique of SimCLR original paper was applied equally to all experiments except for resizing and crop.

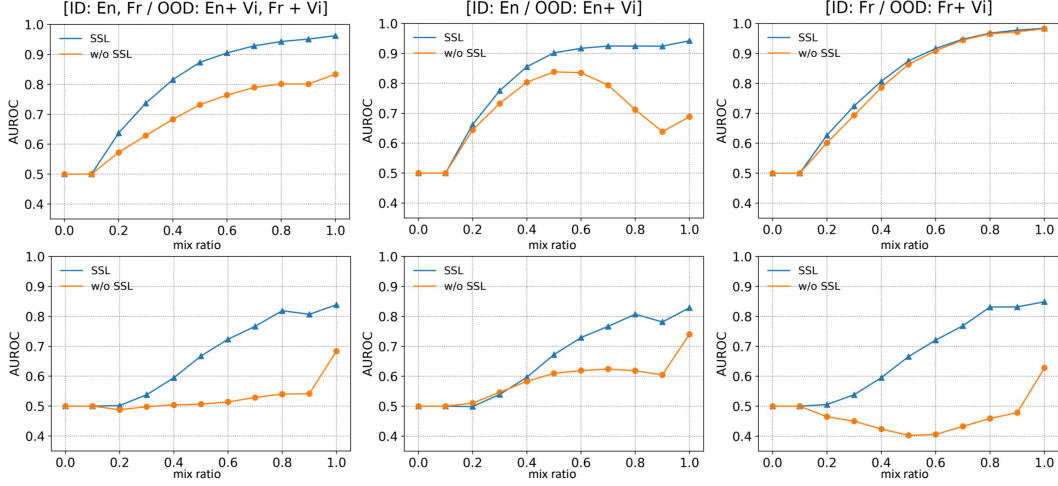[3]https://github.com/huggingface/transformers/tree/master/examples/text-classification

Figure 7: AUROC scores in the text modality using the Baseline (first row) and the Mahalanobis distance (second row). Each column represents the model performance for combined datasets, English only and French only.
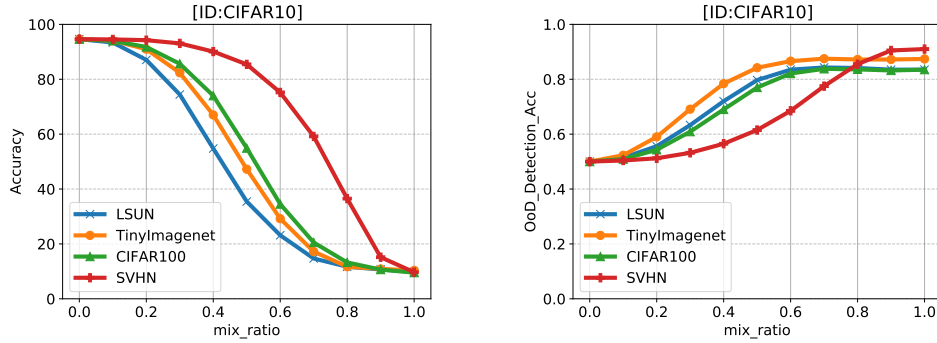

Figure 8: ID classification accuracy(left) and OOD detection accuracy(right) in image domain when CIFAR10 is utilized as ID

C   Experimental Results from Text

In the text modality, all models were trained to achieve the classification accuracy above 99%. As shown in the first column of Fig. 7, SSL-based methods consistently demonstrated better OOD detection performance (*i.e.* higher AUROC scores) compared to the ones without pre-training.

For further analysis, we separately studied the results of OOD detection for each language (English and French), where the two settings exhibited different behaviors. In the English datasets, the one without SSL demonstrated increasing OOD detection performance as the mixing ratio increased (*i.e.,* supposedly, the distance between ID and OOD becoming greater) until the midpoint ($0.6 \sim 0.7$), but after that, the performance decreased except for the high AUROC when the ratio is 1.0. Also, in the French datasets, the transformer without SSL using MD showed the opposite behaviors, decreasing OOD detection performance as the mixing ratio increased until the intermediate ratio, but the performance increased after that ratio. It is also notable that the models without SSL using MD in the English datasets demonstrated a similar behavior as (b) in Fig. 3.

On the other hand, the SSL-based model showed a consistently increasing OOD detection performance as the mixing ratio increased in all languages. The overall results from the text modality indicate that the suggested controllable framework for text (*i.e.,* mixing Vietnamese sentences to original sentences) might demonstrate the validity of the SSL in OOD detection. However, we believe the setting of natural language could be extended to a more principled manner, such as using various combinations of datasets with diverse difficulties as ID and OOD.

(a) [CE] ID:CIFAR10, OOD:LSUN   (b) [OVADM] ID:CIFAR10, OOD:LSUN

(c) [CE] ID:CIFAR10, OOD:TinyImagenet   (d) [OVADM] ID:CIFAR10, OOD:TinyImagenet

(e) [CE] ID:CIFAR10, OOD:CIFAR100   (f) [OVADM] ID:CIFAR10, OOD:CIFAR100

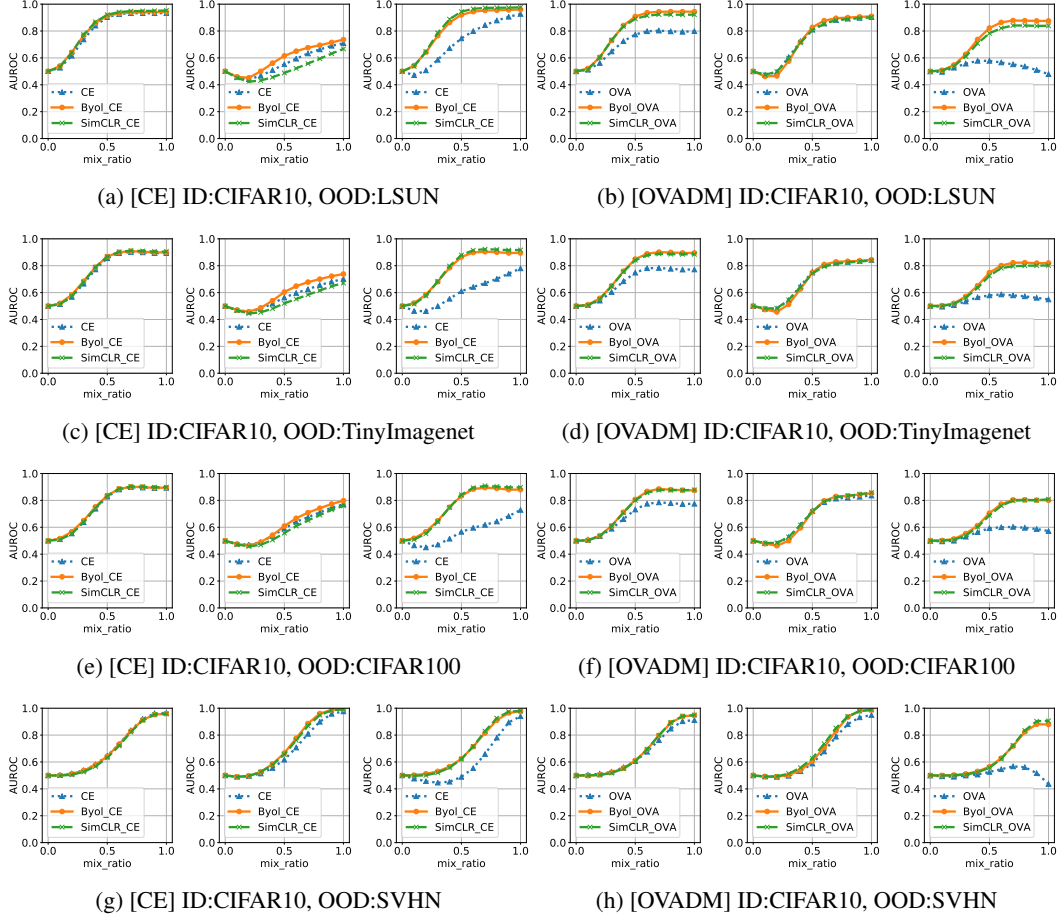(g) [CE] ID:CIFAR10, OOD:SVHN   (h) [OVADM] ID:CIFAR10, OOD:SVHN

Figure 9: Out-of-distribution performance for CIFAR10 in-distribution images. A mixed ratio of zero means the in-distribution image. (a),(c),(e),(g): CE loss is used, (b),(d),(f),(h): OVADM loss is used. (In each case, Left: Baseline, Middle: Mahalanobis distance, Right: ODIN)

D   Additional Results from Images

As mentioned earlier, it is assumed that the mixed image gradually approaches from ID to OOD in the image domain, as the mixing ratio increases. As shown in the Fig. 8, as the mixing ratio increases, the in-distribution classification accuracy of the mixed image is monotonically decreasing, while the OOD detection accuracy is mostly monotonically increasing. These results suggest that the proposed method is a valid controllable framework to simulate various difficulty levels of OOD detection.

We found it informative that the speed of change in accuracy depends on the dataset. When mixing ID with either LSUN, Tiny Imagenet, or CIFAR100 tend to yield a similar effect on the classification accuracy. On the other hand, when mixing ID with SVHN, the one with most distinct characteristics from ID compared to the other three, the classification accuracy degrades slower as the mixing ratio increases. Specifically, the classification accuracy stays relatively the same when mixing ID with SVNH, until the mixing ratio reaches 0.3. When mixing ID with LSUN, on the other hand, the classification accuracy starts to drop as soon as the mixing ratio reaches 0.1.

Fig. 9 illustrates the AUROC performance of each OOD dataset for CIFAR10 according to the change of mixing ratio. Each row represents a different OOD dataset. The order of these rows was taken from the order of its accuracy curve in Fig. 8. The three columns on the left represent the cases where CE loss is applied, and the three columns on the right are the results of applying the OVADM loss. In addition, the first and fourth columns are the results of using the Baseline, the second and fifth columns are the results of using the MD, and the third and sixth columns are the results of using ODIN.

First of all, as the mixing ratio increases, the AUROC quickly reaches the inflection point in the order of LSUN, TinyImagenet, CIFAR100, and SVHN, which is the same behavior as the classification accuracy of each datasets.
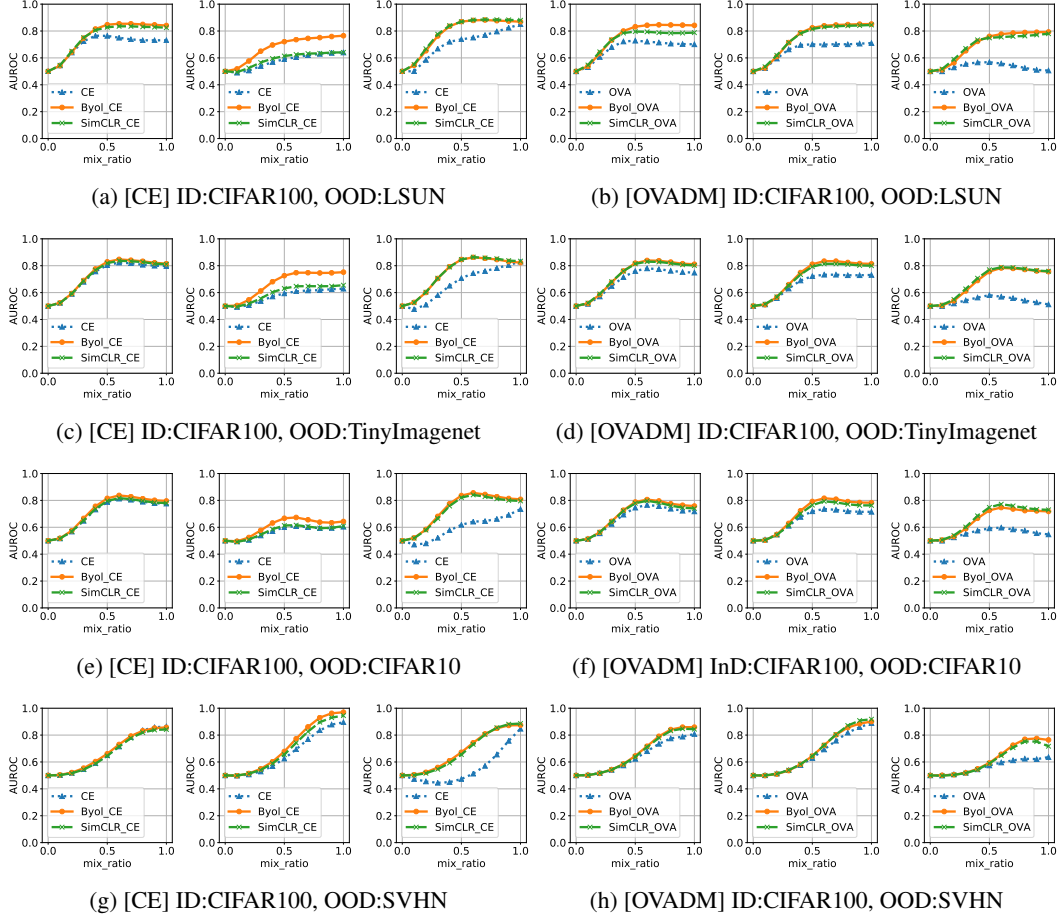
Figure 10: Out-of-distribution performance for CIFAR100 in-distribution images. A mixed ratio of zero means the in-distribution image. (a),(c),(e),(g): CE loss is used, (b),(d),(f),(h): OVADM loss is used. (In each case, Left: Baseline, Middle: Mahalanobis distance, Right: ODIN)

In other words, the faster the classification accuracy decreases as the mixing ratio increases, the longer the flat section appears in the AUROC plot. In addition, when the OOD score was measured by Baseline, the either using SSL or not made little difference when combined with the CE loss. Strangely, the same behavior was shown when using the OVADM loss, but when the OOD score was measured by MD, not Baseline. While most of the plot show monotonically increasing AUROC as the mixing ratio increases, the AUROC changes little when using the combination of the OVADM loss and the ODIN method without SSL. The similar tendency was also observed when CIFAR100 was used as the ID (Fig. 10).