Label-Assisted Memory Autoencoder for Unsupervised Out-of-Distribution Detection

Shuyi Zhang[1,2,3], Chao Pan[1,2], Liyan Song[1,2], Xiaoyu Wu[4], Zheng Hu[4], Ke Pei[5], Peter Tino[3], and Xin Yao✉[1,2,3]

[1] Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology (SUSTech), Shenzhen, China.
[2] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China.
[3] CERCIA, School of Computer Science, University of Birmingham, Birmingham, UK.
[4] RAMS Reliability Technology Lab, Huawei Technology Co., Ltd., Shenzhen, China.
[5] TTE-DE RAMS Lab, Huawei Technology Co., Ltd., Munich, Germany.

Abstract. Out-of-Distribution (OoD) detectors based on AutoEncoder (AE) rely on an underlying assumption that an AE network cannot reconstruct OoD data as good as in-distribution (ID) data when it is constructed based on ID data only. However, this assumption may be violated in practice, resulting in a degradation in detection performance. Therefore, alleviating the factors violating this assumption can potentially improve the robustness of OoD performance. Our empirical studies also show that image complexity can be another factor hindering detection performance for AE-based detectors. To cater for these issues, we propose two OoD detectors LAMAE and LAMAE+. Both can be trained without the availability of any OoD-related data. The key idea is to regularize the AE network architecture with a classifier and a label-assisted memory to confine the reconstruction of OoD data while retaining the reconstruction ability for ID data. We also adjust the reconstruction error by taking image complexity into consideration. Experimental studies show that the proposed OoD detectors can perform well on a wider range of OoD scenarios.

1 Introduction

Deep neural networks have been playing an increasingly important role in many applications, such as autonomous driving [37] and surveillance tracking [38]. When deploying neural networks in real-world applications, there is often very little control over the distribution of test data. The existence of test examples belonging to a different distribution from the training one, also known as Out-of-Distribution (OoD), may cause conventional classification models no longer suitable to be used [10]. Therefore, it is of crucial importance to identify OoD examples in order to maintain the reliance of classification models.

Many OoD detectors have been developed lately [10, 17–19, 25]. But a large body of them requires the availability of OoD data to tune the hyperparameters of the deep networks, being less applicable as OoD data are not typically accessible in reality. Generative models such as deep AutoEncoder (AE) [27] are exempted from this problem when used for OoD detection since they rely on the assumption that when trained with ID data only, an AE network produces higher reconstruction error for unseen OoD data than ID data.

Many AE-based OoD detectors have been proposed based on this assumption [9,33]. However, this assumption may be violated in some scenarios. Observations have demonstrated that its validity depends on the specific characteristics of OoD examples. Sometimes AE-based OoD detector can "generalize" so well that it can also reconstruct OoD data with low reconstruction error, causing unsatisfactory detection performance [7,9]. When the training dataset contains multiple classes instead of only one, which is of more practical use in the real-world, empirical studies of state-of-the-art (SOTA) AE-based OoD detectors reveal an even larger deterioration of detection performance, showing a need for dealing with such learning scenario.

In addition, there has been no systematic study characterizing different types of OoD aiming for analysing the cause of performance degradation of AE-based detectors. The only categorization of OoD is discussed in [11] based on the semantics of ID and OoD examples. The study concluded that detection difficulty would increase when OoD examples possess the same semantic meaning as the ID examples, which coincides with our findings. Nonetheless, based on our preliminary investigation on the detection performance of various types of OoD examples, we noticed that inherent image complexity may be another factor causing OoD performance degradation. As an extreme example, a constant image (i.e., with same-valued pixels) that is of low complexity can always be reconstructed very well. We further noticed that most AE-based methods suffer in such a scenario. Therefore, a more thorough OoD characterization is preferred, which can not only allow us to scrutinize the reason behind the performance variation, but also help researchers to provide more targeted solutions.

To address the above issues, we propose two OoD detectors, namely LAMAE (Label-Assisted Memory AutoEncoder) and LAMAE+, as well as a new criterion to characterize OoD scenarios. Both detectors can be trained without the availability of any information from OoD data. The key idea is to leverage the information of the class-labels of ID data so that the reconstruction of OoD data is constrained while the reconstruction capability for ID data can be retained. Hence, differentiation between ID and OoD examples can be promoted. Furthermore, we provide a finer characterization based on image complexity to investigate the reason for performance degradation of some particular types of OoD. To mitigate the bias induced by inherent image complexity, we propose an entropy-based metric, namely Complexity Normalizer (CN), to adjust the reconstruction error, and incorporate CN metric in the OoD model, forming LAMAE+.

The contributions of this paper are as follows:

1. We propose a new unsupervised OoD detector (LAMAE) that does not require OoD examples for training, neither do we make any assumptions.
2. We provide a finer characterization of OoD scenarios and discuss their relationship to detection performance.
3. Based on the proposed OoD characterization, we further propose a new metric to adjust the reconstruction error so that the refined OoD detector (LAMAE+) performs well on a wider range of different OoD types.

The rest of this paper is organised as follows. Section 2 discusses related work for OoD detection in the literature. Section 3 explains when and why existing OoD detectors may fail, based on which we propose two detectors LAMAE and LAMAE+. The effectiveness of the proposed OoD detectors are evaluated experimentally in Section 4. The paper is concluded in Section 5.

2 Background

This paper considers detecting OoD samples in the context of image classification. When training a classification model, we have a training dataset $D_{in} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{T}$ where $\mathbf{x}_i \in \mathcal{X} = \mathcal{R}^d$ is a d-dimensional feature vector representation of an image data and $y_i = 1, \cdots, S$ is the class label. All training samples are in-distributed as $p_{in}(\mathbf{x}, y)$. The purpose of OoD detection is to identify input examples $\mathbf{x} \sim p_{out}$ where $p_{out} \neq p_{in}$. According to [11], OoDs can be categorized into two types: semantic and non-semantic. Semantic OoDs include data from a distribution $p_{out}(\mathbf{x}, \overline{y})$ with $\{\overline{y}\} \cap \{y\} = \emptyset$. Non-semantic OoDs include data from $p_{out}(\mathbf{x}, y)$, that is, data from the same object class but presented with different styles. It was also concluded that OoD datasets with both types of distribution shifts are the easiest to detect, followed by non-semantic OoD. Semantic OoD turns out to be the hardest one to detect [11].

Many algorithms have been proposed to detect OoD examples [1, 10, 17–19, 21, 25, 29, 36, 39]. However, most of them require the aid of some kind of genuine or synthetic OoD examples in the training stage. This is an unrealistic requirement since in reality, it is often hard, if not impossible, to gain any information regarding OoD a priori. Therefore, we review only OoD detectors that do not require OoD examples for training in this section.

2.1 AE-based Detectors

OoD detectors based on generative models such as AEs naturally possess a characteristic of being able to detect OoD in an unsupervised manner [7]. There are two ways of using AEs for OoD detection [28]. Firstly, an AE can be used to learn a low-dimensional representation of the input data, then distance-based metrics can be applied to assess the discrepancy between newly arrived test examples and the ID dataset [3, 28, 35, 40]. Secondly, the reconstruction error or probability of the test example is calculated directly and used for detection. This work follows the latter strategy.

The reconstruction of AEs has been used extensively for OoD detection to tackle various issues that may exist. For instance, Zhou et al. proposed a robust AE that is capable of detecting anomalies when no clean, noise-free data is available during training [39]. Chen et al. addresses the same issue with AE ensemble by randomly varying the connectivity architecture of the base AE [6]. In contrast, this work focuses on the setting where only clean ID data are available for training.

Ana and Cho adopts a variational autoencoder (VAE) [14] as the base model and utilizes reconstruction probability in a similar manner as reconstruction error to detect OoD [2]. OoDs are expected to have low probability density. SSVAE [4] is a more advanced VAE for semi-supervised learning. The authors supplement the classification loss with the VAE loss so that the performance for both classification and OoD detection can be improved. However, VAEs have their own limitations such as the Gaussian prior assumption on the latent space. Furthermore, many recent work challenge the use of reconstruction likelihood of flow-based generative models such as VAE for OoD detection because extensive experiments have shown that it is not a reliable metric as expected [12, 23].

On the other hand, the reconstruction error of vanilla AEs is a more straightforward metric to use. However, there are also other issues with them. Denouden et al. [7] noticed that AEs can sometimes reconstruct the semantic OoD examples with less error than ID examples. To solve this issue, they adjusted the reconstruction-based detection criterion by adding the Mahalanobis distance between the test sample and the training set mean within the AE latent space. MemAE [9] is another recently proposed method aiming to improve detection for this type of OoD. It incorporates within the training stage a memory to store prototypical elements of the ID data. Hence, the reconstruction of any test examples will be forced to be more similar to the most representative ID examples. Thus, the reconstruction error will be strengthened for OoD examples. This particular issue raised by the methods above is in accordance with the findings in [11]. That is, semantic OoDs are more difficult to detect. The above-mentioned attempts are only tested on one-class ID training data only. However, the difficulty in identifying semantic OoD examples does not only exists in the one-class setting. In fact, the challenge becomes more problematic when there are multiple classes within the ID data, which is a more practical scenario. In this case, characterizing OoD examples based on their semantic meaning may be inadequate. This is explained in more detail in Section 3.2, where we also provide a novel methodology to effectively solve this problem.

Our OoD detectors are built upon MemAE [9], which is explained in more details as follows. MemAE endorses a memory component into the traditional AE architecture as shown in Fig. 1. The encoder $f_e(\cdot)$ maps an input image $\mathbf{x} \in \mathcal{X}$ to a latent space $\mathcal{Z} = \mathcal{R}^{\mathcal{C}}$ via $\mathbf{z} = f_e(\mathbf{x}; \theta_e)$, where θ_e represents the encoder-specific model parameter. Before the latent vector $\mathbf{z}$ is forwarded to the decoder, the memory module $\mathbf{M} \in \mathcal{R}^{N \times C}$ containing N prototypical vectors $\mathbf{m}_i$, each of dimension $1 \times C$, is put in place, where N is a predefined parameter for the memory size. $\mathbf{M}$ is designed to record the prototypical normal patterns of

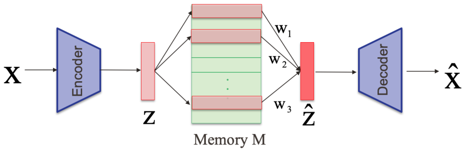


Fig. 1: Framework of MemAE [9]

ID data D_{in}, which is updated at each epoch in the training phase. Once a new training example is received, cosine similarity between the encoded vector $\mathbf{z}$ and each memory item $\mathbf{m}_i$ is calculated as $d(\mathbf{z}, \mathbf{m_i}) = \frac{\mathbf{z}\mathbf{m_i}^T}{||\mathbf{z}||\cdot||\mathbf{m_i}||}$ for $\forall\ i = \{1, \cdots, N\}$. The weight vector $\mathbf{w} = [w_i, \cdots, w_N]$ is calculated via a softmax operation $w_i = \frac{exp(d(\mathbf{z},\mathbf{m_i}))}{\sum_{j=1}^{N} exp(d(\mathbf{z},\mathbf{m_j}))}$ with $\sum_{i=1}^{N} w_i = 1$. To further limit the reconstruction ability for OoD examples, MemAE applied a hard shrinkage technique on $\mathbf{w}$, promoting the sparsity of model parameters.

The latent representation fed to the decoder is then $\hat{\mathbf{z}} = \mathbf{w}\mathbf{M} = \sum_{i=1}^{N} w_i \mathbf{m_i}$. The reconstructed image is $\hat{\mathbf{x}} = f_d(\hat{\mathbf{z}}; \theta_d)$ where θ_d represents the decoder-specific model parameter. The MemAE loss function considers two terms: the reconstruction loss and an entropy for promoting the sparsity of $\mathbf{w}$. The memory is fixed after the training stage. In the testing phase, all examples are forced to be constructed with prototypical components of the ID data, resulting in significant reconstruction errors for OoDs.

2.2 Non-AE-based Detectors

One-class classification are popularly used for OoD detection [24, 26]. Nonetheless, when the number of data dimensions is high, which is a typical issue of image data, these approaches can suffer from the curse of dimensionality.

Other types of OoD detectors also exist. For instance, Shalev et al. [30] utilize extra supervision by training networks to predict word embedding of class labels. It needs to combine the outputs of several similar networks to detect OoD examples. GODIN [11] is a very recently proposed OoD detector. It is an improvement of one of the benchmark detectors, ODIN [19], which utilizes class posterior probability produced by a softmax classifier for detection. Unlike ODIN, it decomposes the class posterior probability using the rule of conditional probability during training and uses only the numerator, i.e., the joint class-domain probability for detection. GODIN frees the algorithm from explicit parameter-tuning with respect to specific OoD datasets.

3 Label-Assisted Memory AutoEncoder

This section explains two OoD detectors, namely Label-Assisted Memory AutoEncoder (LAMAE) and LAMAE+ (a refined adaptation of LAMAE) in Section 3.1 and Section 3.2, respectively. Source codes of our proposed algorithms are available at `https://github.com/fzjcdt/LAMAE`.

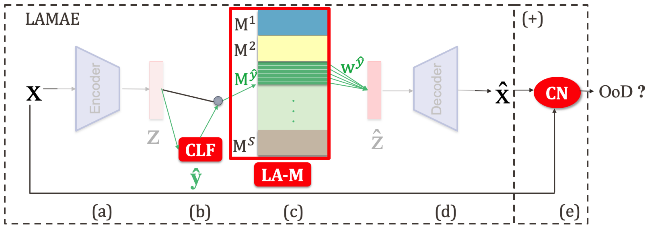


Fig. 2: Framework of our proposed LAMAE and LAMAE+. `CLF` denotes the classifier module (Section 3.1.1). `LA-M` denotes the label-assisted memory (Section 3.1.2). `CN` denotes a normalizer to refine the reconstruction (Section 3.2).

3.1 Label-Assisted Memory AutoEncoder (LAMAE)

As shown in Fig. 2, the network architecture of Label-Assisted Memory AutoEncoder (LAMAE) consists of four components: (a) an encoder (`Encoder`) to compress the intrinsic data features, (b) a classifier (`CLF`) to regularize the memory for a better targeted reconstruction, (c) a label-assisted memory (`LA-M`) reserving class-conditional memory chunks and their associated weights, and (d) a decoder (`Decoder`) to recreate the input based on the information stored in (c). Components (a) and (d) have been explained in Section 2.1, so this section focuses on the newly proposed components (b) and (c).

3.1.1 Classifier Module

Our preliminary experiment shows that when ID data consist of multiple classes, the performance of existing AE-based detectors can deteriorate significantly. Fig. 3 provides an illustrative example, where digit “7” is OoD and the rest nine digits are ID. We can see that MemAE can reconstruct both ID and OOD examples very well (Fig. 3), such that it would be difficult to identify the OoD examples based on reconstruction error. Fig. 4(a) shows the histograms of the reconstruction errors of ID and OoD data, further confirming the difficulty of achieving good OoD performance by using MemAE under this circumstance.

A potential reason is that the latent space learned from the multi-class ID dataset allows for a combination of features from various ID classes to reconstruct unseen OoD examples. This combination may not have much effect on the reconstruction of ID examples, but can be detrimental for OoD detection since the reconstruction error is no longer distinguishable. To tackle this issue, we propose to regulate the reconstruction of test images by exploiting their class labels, which is implemented by placing a classifier (`CLF`) and a label-assisted memory (`LA-M`) in the AE framework as shown in Fig. 2. The two modules are explained in this section and Section 3.1.2, respectively.

A classifier $f_c(\cdot)$ is incorporated into the MemAE architecture by connecting the latent space $\mathcal{Z}$ and memory $\mathbf{M}$ as shown in Fig. 2. $f_c(\cdot)$ can be a single-layered or multi-layered network, depending on the complexity of the application. The predicted label $\hat{y} = f_c(\mathbf{z}; \theta_c)$ of a training image $\mathbf{x}$, where θ_c denotes the classifier

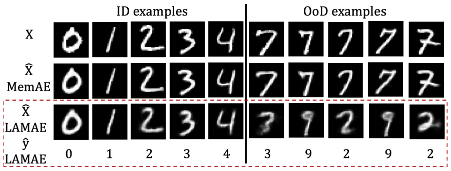
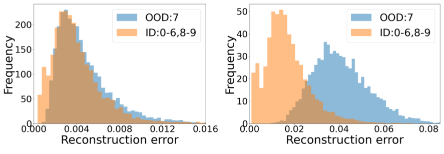


Fig. 3: Original and reconstructed images of ID and OoD for MemAE and LAMAE.



(a) MemAE (b) LAMAE

Fig. 4: Density of reconstruction error of ID and OoD for MemAE and LAMAE.

parameter, can then be used to guide the learning of the label-assisted memory. Given a test image, the latent representation induced by the encoder is forwarded to the classifier for the predicted label. We can see from the dotted box in Fig. 3 that the classifier (trained purely on ID data) always assigns the OoD example with one of the existing ID labels.

3.1.2 Label-Assisted Memory Module

The Label-Assisted Memory module (`LA-M`) aims to record the most representative prototypical patterns for each individual class. Therefore, the whole memory $\mathbf{M}$ of size N is divide into S mutually exclusive class-conditional memory chunks $\{\mathbf{M}^s | s = 1, \cdots, S\}$ where S is the number of ID classes, i.e., $\mathbf{M} = \cup_{s=1}^{S} \mathbf{M}^s$. We use N^s and $\mathbf{w}^s$ to denote the size and associated weight vector of $\mathbf{M}^s$, respectively. This study assigns the same size for $\mathbf{M}^s$, i.e., $N^1 = \cdots = N^S$.

Similar to MemAE, cosine similarity is used to calculate the weights (see Section 2.1). However, thanks to the information from $f_c(\cdot)$, the latent feature $\mathbf{z}$ is only compared with each memory item $\{\mathbf{m}_i^{\hat{y}} \in \mathbf{M}^{\hat{y}} | i = 1, \cdots, N^{\hat{y}}\}$. The associated weight $\mathbf{w}^{\hat{y}}$ is calculated based on the similarity of the memory items and $\mathbf{z}$. Only $\mathbf{M}^{\hat{y}}$ and $\mathbf{w}^{\hat{y}}$ are used to formulate $\hat{\mathbf{z}}$ as

$$\hat{\mathbf{z}} = \mathbb{1}_{s=\hat{y}} \sum_{i=1}^{N^s} w_i^s \mathbf{m}_i^s. \tag{1}$$

Note that weight vector $\mathbf{w}$ is rather sparse since $\mathbf{w}^s = 0 \ \forall \ s \neq \hat{y}$, so we do not need to apply the hard shrinkage technique as in MemAE.

In the testing phase, $\mathbf{M}$ is fixed. Given a test image, one employs the classifier to predict its label, so that only the predicted-class-conditional memory chunk is referred to construct the latent feature according to Eqn. (1), which is then decoded to obtain the reconstruction error. Modules (c) and (d) in Fig. 3 demonstrate this process.

LAMAE is expected to induce higher reconstruction error for OoD examples, as they are forced to be reconstructed with the prototypical features of a mislabelled class. In contrast, the reconstruction performance for ID examples can be maintained, given the typically good performance of CLF. Fig. 4(b) also demonstrates the effectiveness of LAMAE in separating ID and OoD data.

3.1.3 Training Objective

The loss function is formulated as the sum of reconstruction error and classification error on the training data as

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}[R(\mathbf{x}^t, \hat{\mathbf{x}}^t) + \beta L(y^t, \hat{y}^t|\mathbf{x}^t)], \tag{2}$$

where T is the training set size, β is a tuning parameter, $R(\mathbf{x}^t, \hat{\mathbf{x}}^t) = ||\mathbf{x}^t - \hat{\mathbf{x}}^t||_2^2$ is the mean-squared reconstruction error, and $L(y^t, \hat{y}^t|\mathbf{x}^t) = -\sum_{s=1}^{S} \mathbb{1}_{\hat{y}^t=s} log(p(y^t = s))$ is cross entropy classification error. In this work, softmax activation is adopted and β is set to 1 without tuning.

In the training process, ID examples, along with their true labels, are used to minimize the overall loss during which the predictive performance of the classifier module is also guaranteed.

3.2 LAMAE with Complexity Normalizer (LAMAE+)

Further exploration into our experiments shows that although LAMAE generally achieves better performance than other AE-based methods, it may still fail on some specific types of semantic OoDs. For instance, the detection performance is 26% lower when digit “1” is taken as OoD compared with the case when “0” or “2” is taken. In fact, almost all existing AE-based detectors suffer in such a scenario. This section aims to investigate the reason why the detection of some types of OoD is of greater difficulty. After that, we propose a new way to categorize OoD examples. Finally, we design an image complexity-based metric (module (e) in Fig. 2) to upgrade LAMAE, inducing LAMAE+.

3.2.1 Image Reconstruction and Complexity

According to the taxonomy in [11], our experimental setting based on handwritten digits belongs to the *semantic* OoD scenario. Nevertheless, our experimental results show that the detection performances can still vary by a large extent when different digit is treated as OoD, suggesting that it is not adequate to explain the performance variation purely from the perspective of semantics.

Based on this, we hypothesize that the inherent complexity of the image is positively correlated to its reconstruction difficulty, which impacts detection performance. To test this hypothesis, we train and test two AEs on two datasets with very different complexities (handwritten digits with lower complexity and natural images with higher complexity). We adopt Shannon entropy [31, 32], which has a well-established information-theoretic basis, to measure the image complexity as

$$H(S) = -\sum_{S_i=0}^{n-1} p(S_i) log(p(S_i)), \tag{3}$$

where n is the number of grey levels, S_i are the grey level pixel values contained in image S and $p(S_i)$ is the probability of pixel having level S_i. A Pearson correlation of 0.7952 between image entropy and reconstruction error is derived from a total of 20,000 test images (10,000 from each dataset), suggesting a strong positive correlation. Our hypothesis has been verified.

3.2.2 A Characterization of OoD

Experimentally, we found that image complexity played an important role in OoD detection. Hence, we propose to further characterize OoD according to the complexity of OoD examples as compared to the ID ones. When OoD examples have a lower image complexity, such as images with constant pixels, we categorize them as "*plain*". For OoD examples with a higher image complexity, such as images with random pixels, we categorize them as "*fancy*". Altogether, OoDs can be categorized into six classes: *S+P*, *S+F*, *NS+P*, *NS+F*, *NS+S+P*, and *NS+S+F*, where *P, F, S, NS* stand for *plain*, *fancy*, *semantic* and *non-semantic*, respectively.

By nature, the reconstruction for *plain* images should be easier than that of the *fancy* images, since fewer features are required for their description, leading to lower errors. This property would probably mislead OoD detectors towards classifying *plain* images as ID even when they are actually OoD. Therefore, *semantic-and-plain (S+P)* OoD is the hardest to detect among all types of OoD and image complexity should be catered for when making OoD detection based on the criterion of reconstruction error.

3.2.3 Complexity-normalized Test Statistic

To tackle the challenge of detecting *S+P* OoDs, we propose a new metric called Complexity Normalizer (`CN`) to adjust reconstruction error for detection. Indeed, the mechanism of CN can be used in combination with any AEs. When CN is equipped with LAMAE, we form LAMAE+.

To perform LAMAE+ for each test image, we calculate an entropy-based normalizer $CN = log(H(S) + 1)$ and re-scale the reconstruction error derived from LAMAE as:

$$\widehat{Err} = \frac{||\mathbf{x}^t - \hat{\mathbf{x}}^t||_2^2}{CN + \gamma}, \tag{4}$$

where $\gamma > 0$ is a tiny value to avoid the numerical problem of zero division (fixed as 1e-9). $\widehat{Err}$ is the correction of the reconstruction error, taking image complexity under consideration for OoD detection.

4 Experimental Studies

This section carries out two sets of experiments. Experiment 1 validates the proposed LAMAE+ by comparing with SOTA OoD detectors. Experiment 2 examines the effectiveness of each component in LAMAE+. Comparisons between LAMAE and LAMAE+ can also be found in Experiment 2.

4.1 Experimental Setup

Our experiments are based on the following benchmark datasets with each image standardized to [0,1] channel-wise.
1. **MNIST** [16] contains gray-scale images of handwritten digits 0-9.
2. **Fashion MNIST** (FMNIST) [34] contains gray-scale images of Zalando's article images from 10 classes including sneakers, trousers, pullover, etc.
3. **CIFAR10** [15] contains natural color images from 10 classes including airplane, ship, dog, cat, etc.
4. **CelebA** [20] contains face images of 10,177 celebrities.
5. **notMNIST** [5] contains gray-scale images of English letters from A to J.
6. **Constant** contains images of plain color. All pixels of an image has the same value uniform-randomly drawn from the set $\{0, \cdots, 255\}$.
7. **Noise** contains images of uniform noise. Pixel values are independently drawn from the uniform distribution on the set $\{0, \cdots, 255\}$.

To show the generality and applicability of the proposed detectors, we conduct experiments on three different settings. Setting 1 is built with MNIST dataset. In each experiment, one class is used as the OoD class, and the rest are seen as ID. The procedure is repeated for all classes. Within this setting, *S+P* OoDs and *S+F* OoDs exist. In Setting 2 and 3, FMNIST and CIFAR10 are used as the ID dataset respectively. Various OoD datasets including CelebA, notMNIST, FMNIST, Constant and Noise are adopted. Within this setting, both *NS+S+P* and *NS+S+F* OoDs exist.

Due to page limit, we report the area under the receiver operating characteristic curve (AUROC) which plots the true positive rate (TPR) of ID against the false positive rate (FPR) of OoD data by a varying threshold. The average detection performance and the standard deviation of 10 repetitive experiments are reported. Performance measured by area under the precision-recall curve (AUPRC) shows similar trends.

4.2 Experiment 1: Comparative Studies with SOTA Detectors

In this section we validate the proposed methods for the detection of various types of OoD. We compare LAMAE+ with unsupervised detectors including traditional AE, VAE [2], SSVAE [4], MemAE [9] and the latest non-reconstruction-based detector GODIN [11].

On MNIST and FMNIST, we implement the encoder using three convolution layers as in MemAE [9]. For GODIN [11] where MNIST and FMNIST are not used for training, we experimented with the same structure as the setting for encoder-and-classifier component adopted for LAMAE+. On CIFAR10, with higher data complexity, deeper encoder and decoder are constructed for MemAE and LAMAE+. A skip connection from $\mathbf{z}$ to $\hat{\mathbf{z}}$ with dimension 16 is added to further assist reconstruction. Except for the last layer, each layer is followed by a batch normalization (BN) [13] and a Rectified Linear (ReLU) activation [22]. Batch size is set to 128 and we use an Adam optimization procedure.

Table 1: AUROC detection performance for MNIST in Experiment 1. (Each time the model is trained on 9 of the 10 classes and the left-out class is considered to be the OoD class. **Bold** indicates the best scores.)

OoD Class	AE	VAE(e)	VAE(p)	SSVAE(p)	MemAE	GODIN	LAMAE+
0	81.4±0.7	87.0±1.6	94.9±0.1	96.9±0.1	83.0±1.4	86.3±8.0	**98.5±0.2**
1	12.7±0.1	27.2±1.3	47.0±3.0	9.5±0.6	14.2±1.0	86.4±5.0	**90.1±2.5**
2	92.9±0.3	96.0±0.6	96.1±0.1	97.2±0.0	94.8±0.4	88.7±4.4	**99.0±0.1**
3	82.0±0.4	94.2±0.5	84.8±0.3	90.2±0.2	83.1±1.2	70.6±6.8	**97.7±0.3**
4	76.6±0.8	91.4±0.6	70.8±0.4	75.1±0.3	75.7±0.6	79.4±6.0	**95.5±0.6**
5	82.6±0.3	92.2±0.7	86.0±0.3	89.4±0.1	84.1±0.8	64.8±9.9	**97.5±0.3**
6	83.6±0.6	86.1±1.1	92.3±0.01	96.0±0.2	85.9±0.9	83.1±7.6	**97.0±0.5**
7	56.9±1.0	67.2±1.7	66.9±0.2	75.5±0.5	57.7±0.8	79.8±8.7	**94.7±1.1**
8	90.5±0.3	95.3±0.5	89.1±0.4	92.2±0.2	90.1±0.6	85.8±3.8	**97.5±0.3**
9	59.2±0.7	67.3±0.8	62.0±2.0	67.7±0.5	56.5±0.9	79.1±9.1	**89.7±2.0**

For AE-based detectors, the maximum number of training epochs is set to 200, 200 and 500 and the class-conditional memory size N^s in LAMAE and LAMAE+ is set to 10, 10 and 50 for MNIST, FMNIST and CIFAR10 respectively. Later we demonstrate experimentally that performance is insensitive to the selection of memory size. An extra fully connected layer with softmax output is taken as the classifier component. A 10% validation set is extracted from the ID training dataset. Early stopping is adopted to choose the model that achieves the lowest loss on the validation dataset. Note that this validation dataset is still ID so the models are trained without access to any information about OoD.

4.2.1 Performance on Semantic OoD Only

Table 1 reports the results of MNIST. Results of VAE(p) and SSVAE(p) based on reconstruction probability are taken from the original papers [2,4]. We also report the results based on reconstruction error (VAE(e)).

We can see that LAMAE+ achieves the best AUROC in all 10 cases. The improvement is especially substantial for digits “1”, “4” , “7” and “9”. Detailed explanations of how exactly each component in LAMAE contributed to this outcome is presented later in Section 4.3. It is also worth noting that the standard deviation of GODIN is much larger than that of AE reconstruction-based approaches, signifying that this type of approaches are more stable than softmax-based approaches.

4.2.2 Performance on Both Semantic and Non-semantic OoD

Table 2 reports the results of FMNIST and CIFAR10 where various OoD datasets are selected. VAE and SSVAE based on reconstruction likelihood have not been tested within these settings and the exact formulation of reconstruction likelihood is not provided. Hence, we report only VAE(e) and SSVAE(e) based on

Table 2: AUROC detection performance for FMNIST and CIFAR10 in Experiment 1. (**Bold** indicates the best scores.)

ID	OoD	AE	VAE(e)	SSVAE(e)	MemAE	GODIN	LAMAE+
FMNIST	MNIST	96.2±0.2	99.0±0.1	98.9±0.1	97.1±0.1	79.0±4.3	**99.9±0.0**
	notMNIST	99.6±0.0	99.8±0.0	**99.9±0.0**	99.8±0.0	64.0±5.8	**99.9±0.0**
	Constant	68.1±2.2	63.2±1.7	82.0±7.0	72.3±1.1	84.4±9.6	**100.0±0.0**
	Noise	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	86.5±6.3	99.9±0.0
CIFAR10	FMNIST	71.7±0.7	69.4±0.9	84.8±1.9	**98.5±0.0**	94.2±1.7	95.0±0.6
	CelebA	55.0±0.5	58.2±0.3	60.0±1.2	70.0±0.0	**75.7±2.2**	59.5±1.0
	Constant	0.0±0.0	0.0±0.0	0.0±0.0	51.2±0.0	92.7±2.0	**100.0±0.0**
	Noise	100.0±0.0	**100.0±0.0**	**100.0±0.0**	**79.6±0.0**	91.0±8.8	**100.0±0.0**

Table 3: AUROC detection performance for MNIST in Experiment 2. The second column lists the average image complexity of each OoD class on 1000 images measured by Eqn. (3). **Bold** indicates the best scores among each subgroup.

	Detector	1	2	3	4	5	6
OOD Class	Complexity	AE	AE+	MemAE	MemAE+	LAMAE	LAMAE+
0	1.91	**81.4±0.7**	78.3±0.8	**83.0±1.4**	79.9±1.6	**98.8±0.1**	98.5±0.2
1	0.94	12.7±0.1	**27.0±0.3**	14.2±1.0	**32.6±1.8**	72.8±4.2	**90.1±2.5**
2	1.76	**92.9±0.3**	**92.9±0.3**	**94.8±0.4**	**94.8±0.4**	98.9±0.1	**99.0±0.1**
3	1.74	**82.0±0.4**	**82.0±0.4**	**83.1±1.2**	**83.1±1.3**	97.5±0.4	**97.7±0.3**
4	1.55	76.6±0.8	**79.2±0.7**	75.7±0.6	**78.7±0.7**	93.9±0.8	**95.5±0.6**
5	1.67	82.6±0.3	**83.1±0.3**	84.1±0.8	**84.6±0.9**	97.2±0.3	**97.5±0.3**
6	1.71	83.6±0.6	**84.3±0.6**	85.9±0.9	**86.6±0.9**	96.5±0.5	**97.0±0.5**
7	1.42	56.9±1.0	**62.1±1.0**	57.5±0.8	**63.4±0.8**	91.7±1.5	**94.7±1.1**
8	1.88	**90.5±0.3**	89.1±0.4	**90.1±0.6**	88.5±0.7	**98.0±0.3**	97.5±0.3
9	1.59	59.2±0.7	**60.9±0.7**	56.5±0.9	**58.2±1.0**	87.4±2.1	**89.8±2.0**

reconstruction error. We can see that LAMAE+ ranked the first in 5 out of the 8 cases. In particular, it is capable of detecting the OoD examples belonging to the Constant dataset better than the other methods, demonstrating its effectiveness in identifying the *plain* OoDs.

The performance is not as good when CIFAR10 is used as the training dataset. This may be due to the fact that the network structure is not deep enough to account for the complicated details within the CIFAR10 dataset. In addition, a single classifier layer may also be inadequate for this dataset. For instance, the backbone classifier used by GODIN is Resnet-34 [11]. Increasing the complexity of network may lead to improvements in detection performance at a cost of an increasing computational burden.

4.3 Experiment 2: Analysis of LAMAE+

In this section, we analyse the functionality of each component in LAMAE+. We experimentally demonstrate that the combination of label-assisted memory and CN-adjusted test statistic helps the detector to achieve better results on the most difficult OoD type, i.e., *S+P* OoDs, with the MNIST dataset.

4.3.1 Effect of the Classifier and the Label-Assisted Memory

To illustrate the effectiveness of the classifier and the label-assisted memory, we compare the results for AE, MemAE and LAMAE on the MNIST dataset in Table 3 (Detectors 1, 3 and 5).

We can see that our results for AE and MemAE confirmed the benefit of establishing a memory component as discussed in MemAE [9], for which MemAE achieved higher AUROC than AE in 7 out of the 10 cases. Furthermore, in all 10 cases, LAMAE achieved a significant improvement in AUROC when compared with both MemAE and AE, demonstrating the dominant advantage of using a classifier and a label-assisted memory for detecting *semantic* OoDs when the ID dataset contains multiple classes. Moreover, for OoD digits “1”, “4”, “7” and “9” whose image complexity measured with Shannon entropy (Eqn. 3) ranked the lowest four, there is still a gap when compared with the rest cases. We will address this issue with *plain* OoDs with the CN component in the following section.

4.3.2 Effect of CN-Adjustment

This section demonstrates that the CN-adjustment can further improve the detection performance, especially for the most difficult OoD type *S+P*. As discussed earlier, CN can be used with any AE reconstruction-based OoD detectors and an improvement in detection performance can be anticipated. We verify this conjecture experimentally by taking AE, MemAE and LAMAE as the base detectors and modify only the reconstruction error-based test statistic. We rename the CN-adjusted detectors by suffixing “+”. Results are presented in Table 3.

It can be noted that in 8 of the 10 cases, using a CN-adjusted test statistic indeed leads to a significant improvement in detection performance for the digits “1” , “4” , “7” and “9” , which are the hardest ones to detect among all digits [4] and can be characterized as *S+P* by us. This is true for various types of AEs. For digits “0” and “8” , CN caused a slight decrease in AUROC. This is due to the fact that these two digits are already the most complex 2 with the highest entropy values. Regarding this, we suggest that better complexity measurements may be created in the future so that the detection performance on slightly more complicated images can be maintained.

Examining the overall average performance, we conclude that the improvement in detection performance is attributed to the combination of the classifier component, the label-assisted memory and the complexity normalizer.

4.3.3 Sensitivity to Memory Size

This section provides a sensitivity analysis of the detection performance for LAMAE. We present the performance under different memory size settings for the MNIST experiment. Fig. 5 suggests that LAMAE is robust to different memory sizes and for simple datasets such as MNIST, even a small memory size can achieve satisfactory performance.

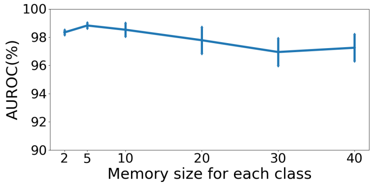


Fig. 5: Sensitivity of detection performance to class-conditional memory size on MNIST when digit "0" is held as OoD. Similar trends can be observed for others.

5 Conclusion

We proposed LAMAE, a novel AE-based OoD detector with a label-assisted memory. Specifically, we injected a classifier and a class-conditional memory into the AE network architecture to avoid combination of features from different ID classes and thus, constrain the reconstruction of OoD examples while retaining the generalization on ID examples. The detection performance of semantic OoD examples improved significantly. We also proposed a new way to characterize OoD based on image complexity and a new metric, CN, to eliminates the bias associated with the reconstruction error induced by inherent image complexity. Thereby, the refined detector LAMAE+ is capable of detecting the most difficult type of OoD that previous work cannot handle well. It is also worth pointing out that both detectors are purely unsupervised.

In the current work, we only used the basic Shannon entropy to measure image complexity. More suitable measures may also exist [8]. Besides, the performance of the classifier component is of crucial importance to the results. Potential improvements can be made to further improve the detection performance on more complex datasets. Various sizes for each class-conditional memory can also be considered.

Acknowledgement This work was supported by National Natural Science Foundation of China (Grant No. 62002148), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07X386), Shenzhen Science and Technology Program (Grant No. KQTD2016112514355531), Shenzhen Fundamental Research Program (Grant No. JCYJ20190809121403553), Research Institute of Trustworthy Autonomous Systems, and Huawei.

References

1. Abdelzad, V., Czarnecki, K., Salay, R., Denounden, T., Vernekar, S., Phan, B.: Detecting out-of-distribution inputs in deep neural networks using an early-layer output. arXiv preprint arXiv:1910.10307 (2019)
2. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE **2**(1), pp. 1-18. (2015)

3. Andrews, J.T., Morton, E.J., Griffin, L.D.: Detecting anomalous data using autoencoders. International Journal of Machine Learning and Computing **6**(1), 21. (2016)
4. Berkhahn, F., Keys, R., Ouertani, W., Shetty, N., Geißler, D.: Augmenting variational autoencoders with sparse labels: A unified framework for unsupervised, semi-(un) supervised, and supervised learning. arXiv preprint arXiv:1908.03015. (2019)
5. Bulatov, Y.: Notmnist dataset, `http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html` (2020)
6. Chen, J., Sathe, S., Aggarwal, C., Turaga, D.: Outlier detection with autoencoder ensembles. In: SIAM International Conference on Data Mining. pp. 90-98. (2017)
7. Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., Vernekar, S.: Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. arXiv preprint arXiv:1812.02765. (2018)
8. Gao, P., Li, Z., Zhang, H.: Thermodynamics-based evaluation of various improved shannon entropies for configurational information of gray-level images. Entropy **20**(1), 19. (2018)
9. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: IEEE/CVF International Conference on Computer Vision. pp. 1705-1714. (2019)
10. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations. (2018)
11. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10951-10960 (2020)
12. Huang, Y., Dai, S., Nguyen, T., Baraniuk, R.G., Anandkumar, A.: Out-of-distribution detection using neural rendering generative models. arXiv preprint arXiv:1907.04572. (2019)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448-456. (2015)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations. (2014)
15. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto. (2009)
16. LeCun, Y.: The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`. (1998)
17. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations. (2018)
18. Lee, K., Lee, K., Lee, H., 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)
19. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations. (2018)
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision. pp. 3730-3738. (2015)
21. Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., Lopez, A.M.: Metric learning for novelty and anomaly detection. In: British Machine Vision Conference **64**. (2018)

22. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning pp. 807-814.(2010)
23. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: International Conference on Machine Learning. (2019)
24. Perera, P., Patel, V.M.: Learning deep features for one-class classification. IEEE Transactions on Image Processing **28**(11), pp.5450-5463 (2019)
25. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., DePristo, M.A., Dillon, J.V., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: Advances in Neural Information Processing Systems pp. 14680-14691. (2019)
26. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International Conference on Machine Learning. pp. 4393-4402. (2018)
27. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Technical report La Jolla Inst for Cognitive Science (1985)
28. Sarafijanovic-Djukic, N., Davis, J.: Fast distance-based anomaly detection in images using an inception-like autoencoder. In: International Conference on Discovery Science. pp. 493-508. (2019)
29. Shafaei, A., Schmidt, M., Little, J.: A less biased evaluation of ood sample detectors. In: British Machine Vision Conference (2019)
30. Shalev, G., Adi, Y., Keshet, J.: Out-of-distribution detection using multiple semantic label representations. In: Advances in Neural Information Processing Systems. pp. 7375-7385. (2018)
31. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), pp. 379-423. (1948)
32. Tsai, D.Y., Lee, Y., Matsuyama, E.: Information entropy measure for evaluation of image quality. Journal of Digital Imaging **21**(3), pp. 338-347. (2008)
33. Tuluptceva, N., Bakker, B., Fedulova, I., Schulz, H., Dylov, D.V.: Anomaly detection with deep perceptual autoencoders. arXiv preprint arXiv:2006.13265. (2020)
34. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747. (2017)
35. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. In: British Machine Vision Conference **8**. (2015)
36. Yu, Q., Aizawa, K.: Unsupervised out-of-distribution detection by maximum classifier discrepancy. In: IEEE/CVF International Conference on Computer Vision. pp. 9518-9526. (2019)
37. Yuan, Y., Wang, D., Wang, Q.: Anomaly detection in traffic scenes via spatial-aware motion reconstruction. IEEE Transactions on Intelligent Transportation Systems **18**(5), pp. 1198-1209. (2016)
38. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: ACM International Conference on Multimedia. pp. 1933-1941. (2017)
39. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 665-674. (2017)
40. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations. (2018)