EEE4423 Deep Learning Lab
Out of distribution Detection method for a MNIST classification:

Exploring the generative model for OOD Detection, with Hierarchical Self-Conditioned AutoEncoder

Jiho Park, 2019142056
Yonsei University
qkrwlgh0314@yonsei.ac.kr

Abstract

For the anomaly detection tasks, generative model based methods have been showing weaker performance than classifier based methods(as far as I know). This empirical result contradicts to our intuition that generative model learns the distribution of the data more directly. In this report, I mainly compare these two methods and explore the usage of the distribution learning model, on the given MNIST OOD(Out-Of-Distribution) detection task.

First, I experimented the classifier-based ***Maximum Softmax Probabilty*** *method with two techniques, perturbation and temperature scaling. Second, I experimented the Reconstruction Error based method. I present* ***HSCAE (Hierarchical Self-Conditioned AutoEncoder)****, which efficiently maps the OOD data to In-distribution data. Third, I compared these two methods and designed the* ***ensemble model****, which performed the best.*

As a results of the experiments, unlike HSCAE, the classifier based method was very sensitive to optimization and the hyperparameters (e.g. threshold, temperature scaling). HSCAE learned it's manifold in more stable manner, and showed stable performance. Lastly, from the ensemble model, HSCAE has detected the OOD data which classifier couldn't. This improvement led to the final accuracy, ***zero-acc: 99.07%, else-acc: 97.32%****.*
https://github.com/jiho314/OOD_mnist

1. Introduciton

Learning the underlying representation of the in-distribution data, is the main task of OOD(Out-Of-Distribution) detection. For example in the figure [2], if the manifold of the in-distribution is torus, learning the torus manifold will help to distinguish the OOD red dot. The researches of OOD or anomaly detection, were actively applied in various fields like industry, medical images, signal processing, and social networks.

The given task for this report is the unsupervised anomaly detection of MNIST digit data. 1 to 9 digits are defined as in-distribution and the 0 digit is defined as OOD.

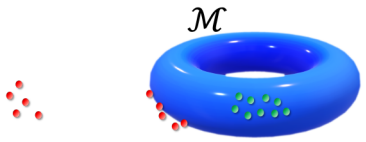

Figure 1. llustration of the manifold M (represented by torus as an example), In-distribution examples (green dots), and OOD examples close to and far from M , figure from [2]

2. Related Work

There exists two main methods for unsupervised OOD detection. One is classifier-based method and the other is reconstruction based method.

2.1. Classifier Based OOD Detection

Dan *et al*. [1] have proposed the baseline and the evaluation metric(e.g. AUPR, AUROC) for the OOD detection tasks. The **Maximum Softmax Probability** method was mainly used, which classifies the data as OOD if the maximum probability is lower than the threshold.

Shiyu *et al*. [8] showed the application of two techniques to the MSP(Maximum Softmax Probability) method, and this is called ***ODIN***. ODIN is a pre-processing method, which can be applied to any classifier and no needs training. First technique is the ***Temperature Scaling***(Eq 1),

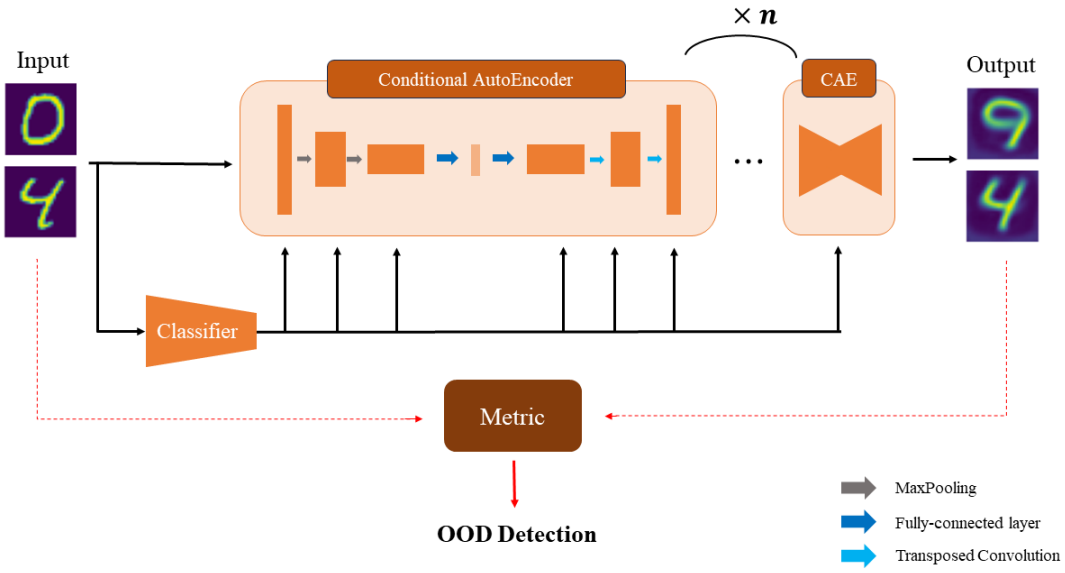

Figure 2. The Architecture of Hierarchical Self-Conditioned AutoEncoder

which was first proposed from Geoffrey *et al*. [3]. It's simple logit-division technique which efficiently addresses the high-confidence problem. Second is the ***Perturbation***, which came from FGSM(Fast Gradient Sign Method) [5]. It computes the gradient that increases the loss, and subtract it to the input image(Eq 2). The subtraction increases the softmax score of the in-distribution data.

$$S_i(x;T) = \frac{\exp(f_i(x)/T)}{\sum_{j=1}^{N}\exp(f_j(x)/T)} \tag{1}$$

$$\tilde{x} = x - \epsilon \text{sign}(-\nabla_x \log S_y(x;T)) \tag{2}$$

2.2. AutoEncoder

Stanislav *et al*. [9] have explained the mathematical relation between the encoder/decoder and it's manifold. As their interpretation, autoencoder does not just compress the input data, but also denoises the noise between the data and manifold. This means the ***autoencoder actually can map the input data to it's trained distribution***. So, autoencoder and it's variations were actively applied to anomaly detection tasks, including [6, 9].

3. Method

3.1. ODIN

The CNN(Convolution Neural Network) based classifier were used for the ODIN method. Then, I applied the temperature scaling and the perturbation to the classifier. For the perturbation, the label is necessary to compute the loss and it's gradient, but there are no labels for evaluation data. So, there needs an assumption that the trained classifier can classify the unseen 1 to 9 digit data properly. With this assumption, we can make the pseudo-label by converting the prediction into one-hot vector. This method is valid since the data domain is easy and the classifier actually achieves the high accuracy of the test mnist data. The temperature scaling factor, perturbation weight and the threshold delta were the hyperparameters and have been adjusted after several grid searches.

3.2. Hierarchical Self-Conditioned AutoEncoder

The goal of HSCAE is the mapping of OOD data to in-distribution. After this mapping, in-distribution data will be reconstructed well, while OOD data transforms to digit among 1 to 9. OOD detection is done with the reconstruction error and it's threshold. The architecture is shown in Fig 2.

AutoEncoder: Due to the domain data set, the important part of the AE(AutoEncoder) design is the ***dimension of the latent vector***(encoded vector). If the dimension is too high, AE can learn to embed sufficient information and reconstruct the image without the specific knowledge of the train data distribution. So the latent dimension is set as relatively small values(e.g. 6, 8, 10). In compensation, to remain high-quality reconstruction for in-distribution data, each encoder and decoder were designed deeper than other AEs for mnist reconstruction. Both of them consists of 14 convolution layer and 1 fully-connected layer.

Why not other AutoEncoders?: There are many developed versions of AE, but they were not applied. This is due to make the reconstruction more difficult for OOD data.

VAE(Variational-AE) directly learns the distribution by maximizing the evidence lower bound. So it is well known to have better generalization performance. I've wanted my AE to overly fit the 1 to 9 digit domain, so VAE was not applied. Also, for the same reason, residual connection was not applied. Since these are my theoretical argument, more experiments are necessary for rigorous verification.

Self-Conditioning: Classifier's prediction is used as a condition. First, to match with the channel dimension, the condition is linearly projected to layer's channel dimension, with non-linear activation. Then, it is added to the layer(equally on every pixel). This conditioning method was referred to classifier-free guidance diffusion [7]. In addition, the classifier is the one used on ODIN experiment, and is frozen while training AE.

The self-conditioning forces the mapping of OOD data to in-distribution, using the 'high-confidence' property of the softmax. Even if the given data is OOD, the data will be predicted among 1 to 9 digit in high probability. This is because of the softmax, which increases the logit exponentially. Therefore, the OOD data can be more easily mapped to the conditioned digit, which is the in-distribution. This efficiently increases the reconstruction error of the OOD data while decreasing the one of the in-distribution.

Hierarchy: After training a single self-conditioned AE, hierarchical reconstruction is applied. It strengthens the mapping to the in-distribution. Hierarchy level is set as hyperparamter, which can be adjusted after the training.

Metric: For reconstruction error, the distance metric influences the performance a lot. Softmax probability distance, maximum softmax probability distance, mean squared error, vgg-content loss and SSIM were tested. For softmax probability distance metric, the distance between the softmax probability of its input and reconstruction were measured. Temperature scaling factor 2 was applied to the probability computation. For HSCAE alone, the softmax probability distance metric performed best, and for ensemble, MSE metric performed the best.

Final Prediction: After the detection of the OOD, the HSCAE outputs the prediction of in-distribution data with it's classifier.

3.3. Ensemble

The classifier and the autoencoder learns the underlying representation in different manner. In other words, two models can map the OOD data in different manifold. Therefore, the ensemble of two can achieve the complementary OOD detection. In the ensemble model, the data is detected as OOD if either of two models classifies it as OOD. Since there exists two chances to be detected as OOD, more strict thresholds were applied for both ODIN and HSCAE.

		ODIN OOD Detection			
		without perturbation		with perturbation	
Models	**1∼9 Eval**	zero_acc	else_acc	zero_acc	else_acc
ResNet18-ep50-1	99.47	96.475	98.049	**98.230**	**97.266**
ResNet18-ep50-2	99.60	95.095	96.343	95.874	95.791
ResNet18-ep100	99.63	93.602	94.213	94.991	93.313
SpinalVGG-ep50	**99.64**	94.814	95.375	95.634	94.793

Table 1. Experiment results of the classifiers.

	HSCAE OOD Detection					
	hierarchy 1		**hierarchy 2**		**hierarchy 3**	
latent dim	zero_acc	else_acc	zero_acc	else_acc	zero_acc	else_acc
6	96.865	96.428	96.502	96.538	-	-
8	97.112	96.020	**97.505**	**96.020**	96.024	95.966
10	96.415	95.726	96.700	95.509	97.109	95.014

Table 2. Experiment Results of HSCAE

Ensemble: ODIN(resnet18) + HSCAE(latent dim 8)									
hierarchy1		hierarchy 2		hierarchy 3		hierarchy 4		hierarchy 5	
zero	else	zero	else	zero	else	zero	else	zero	else
97.77	97.54	98.10	97.54	98.43	97.53	**98.77**	**97.52**	**99.07**	**97.32**

Table 3. Experiment Results of Ensemble

4. Experiment

Batch size of 128, random crop and rotation augmentation were applied for train dataset. Every optimization were done by Adam optimizer without extra scheduling. For evaluation of each model, best threshold values were chosen by grid search.

4.1. ODIN

Three ResNet18 models were tested. Two trials with 50 epochs of training, and one trial with 100 epochs of training. SpinalVGG [4] were tested after 50 epochs for comparison. Temperature scaling factor of 10 were applied for all. The results are in table 1. The OOD detection accuracies of ResNet18 differ a lot. Even with the same 50 epoch of training, there exists 1∼3% of accuracy difference. For SpinalVGG, since it is one of the sota mnist digit classifier, it shows the best accuracy on 1 to 9 prediction. However it shows weaker OOD detection performance. Lastly the perturbation shows the trade-off between zero-acc and else-acc.

Analysis: Through this results, we can learn few properties of MSP OOD detection. First, better classification performance doesn't ensures the better OOD detection. Second, we can see that performance is very sensitive to it's optimization. Third it's threshold is also sensitive. The best threshold values for each ResNet18-ep50-1, ResNet18-ep50-2, ResNet18-ep100 were ***0.235, 0.28, 0.34***.

As a result, we can say that the classifier based MSP OOD detection forms it's manifold in stochastic manner.

4.2. HSCAE

After several tests to find the best latent dimension, I've chose 6, 8, 10 of the dimension and tested on hierarchy level 1, 2 and 3. Each of them was trained for 200 epochs. As in table 2, latent dimension of 8 and hierarchy level 2 performed the best.

Analysis: HSCAE outperformed the ODIN classifiers except ResNet18-ep50-1. Also, unlike ODIN classifiers, it's best threshold values were stable. The best threshold values were ***1.2e-4, 1.0e-4, 1.0e-4*** for each latent dimension 6, 8, 10. As a result, we can see that the HSCAE learns the manifold in more stable manner.

4.3. Ensemble

For ensemble model, ResNet18-ep50-1 and HSACE with latent dim 8 were applied. After several tests, more strict thresholds were chosen. Threshold of 0.231 for classifier, and 1.0(with MSE metric) for HSCAE were applied. As in table 3, hierarchy level 4 and 5 performed the best.

Analysis: The ensemble model outperformed both ODIN and HSCAE detector. We can ensure that HSCAE detects the OOD data which ODIN can't.

5. Discussion

Mapping from OOD to In-distribution: To verify the mapping function of the HSCAE. I've done two qualitative evaluations, about latent dimension and hierarchy level. First, HSCAE with appropriate latent dimension can actually map the OOD data to in-distribution. Fig 3 is the reconstruction result from HSCAE with latent dimension 8. As we can see, HSCAE successfully maps the OOD to in-distribution. As I mentioned in Sec 3.2, I argue that self-conditioning eases this mapping. However, the latent dimension of 10 shows different results. By Fig 4, I've confirmed that latent dimension 10 is high enough to learn complete reconstruction without using the conditioned knowledge. Considering qualitative and quantitative results, I've chose to use the latent dimension 8. Second, the hierarchical reconstruction could strengthen the mapping. As in Fig 5, some inputs become more in-distribution digit after several hierarchies. These two results could explain the high performance of the ensemble model with HSCAE of latent dimension 8 and hierarchy level 5.

Complementary of MSE Metric: For HSCAE alone, SP(Softmax Probability) metric worked best. However on ensemble, MSE metric performed better. This might have

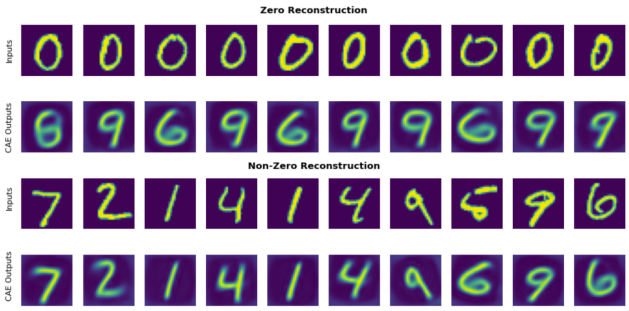

Figure 3. Comparison of the HSCAE(latent dim = 8) reconstruction between zero and non-zero

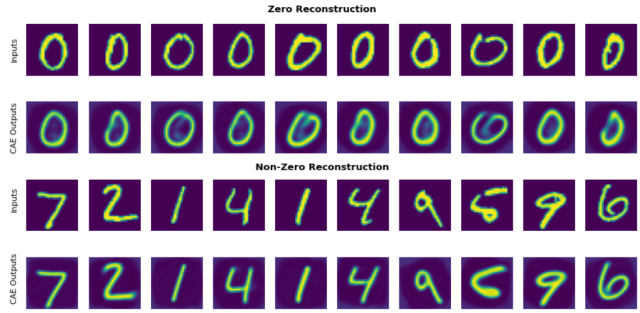

Figure 4. Comparison of the HSCAE(latent dim = 10) reconstruction between zero and non-zero

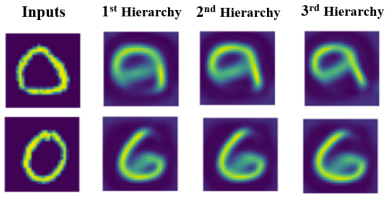

Figure 5. Hierarchical reconstruction example from the HSCAE (latent dim = 8)

happened because SP was already used in ODIN. Using MSE metric could help HSCAE to form more different OOD detection boundary, and so complement the ODIN classifier.

Test Metric: Most of all, ***test metric was incomplete***. The single accuracy value after the adjustment of threshold was used for evaluation. However, this is not pure evaluation, more close to hyper parameter tuning process. This might be the reason of the high performance of ResNet18-ep50-1. As a result, for rigorous evaluation, ***AUPR or AUROC evaluation is necessary, which is regardless of threshold values***.

6. Conclusion

For MNIST digit dataset, the presented structural improvement from AE, strengthened the mapping to in-distribution and improved the OOD detection performance.

ODIN classifier and HSCAE each learned the manifold of in-distribution, in different manner. This is because objectives of the classifier and the generative model are different. ***Therefore, even generative models generally show weaker performance on OOD detection, it's distribution learning property can complement the classifier-based methods.***

References

[1] Kevin Gimpel Dan Hendrycks. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*. 1

[2] Budrul Ahsan Takashi Nishide Genki Osada1, Tsubasa Takahashi. Out-of-distribution detection with reconstruction error and typicality-based penalty. *WACV*. 1

[3] Jeff Dean Geoffrey Hinton, Oriol Vinyals. Distilling the knowledge in a neural network. *NIPS*. 2

[4] Seyed Mohammad Jafar Jalali Abbas Khosravi Amir F Atiya Saeid Nahavandi Dipti Srinivasan H M Dipu Kabir, Moloud Abdar. Spinalnet: Deep neural network with gradual input. 3

[5] Christian Szegedy Ian J. Goodfellow, Jonathon Shlens. Explaining and harnessing adversarial examples. *ICLR*. 2

[6] Sungzoon Cho Jinwon An. Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center*. 2

[7] Tim Salimans Jonathan Ho. Classifier-free diffusion guidance. *NeurIPS*. 3

[8] R. Srikant Shiyu Liang, Yixuan Li. Enhancing the reliability of out-of-distribution image detection in neural networks. *ICLR*. 1

[9] Donald A. Adjeroh Gianfranco Doretto Stanislav Pidhorskyi, Ranya Almohsen. Generative probabilistic novelty detection with adversarial autoencoders. *CVPR*. 2