Out-of-Distribution (OOD) Learning via Generative Adversarial Networks (GANs)

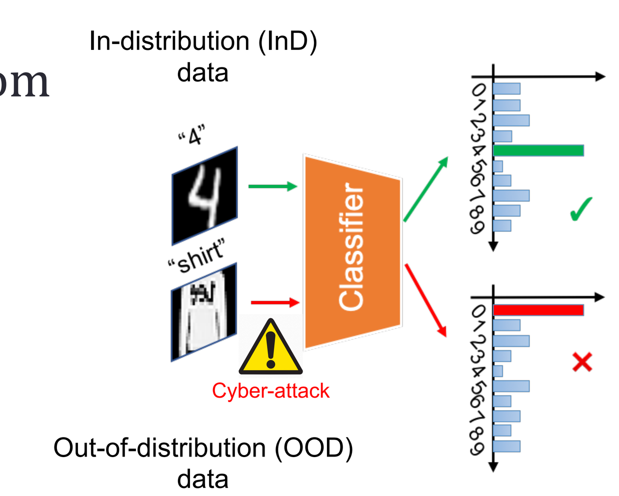
Summer 2022

Student: Xiaoyang Song Researcher: Wenbo Sun UMTRI Group: Bioscience

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Introduction

- In deep neural networks (DNNs), the training and test data are possibly from different distributions
- Classifier tends to overconfidently predict OOD data with class labels of InD data
- OOD data threatens the reliability of classifier in practice
- OOD data is rarely observed and should be augmented via generative models



Objectives

- Goal: Strengthen the classifier with the ability to detect OOD samples.
- Task 1: Train an OOD **detector** such that:

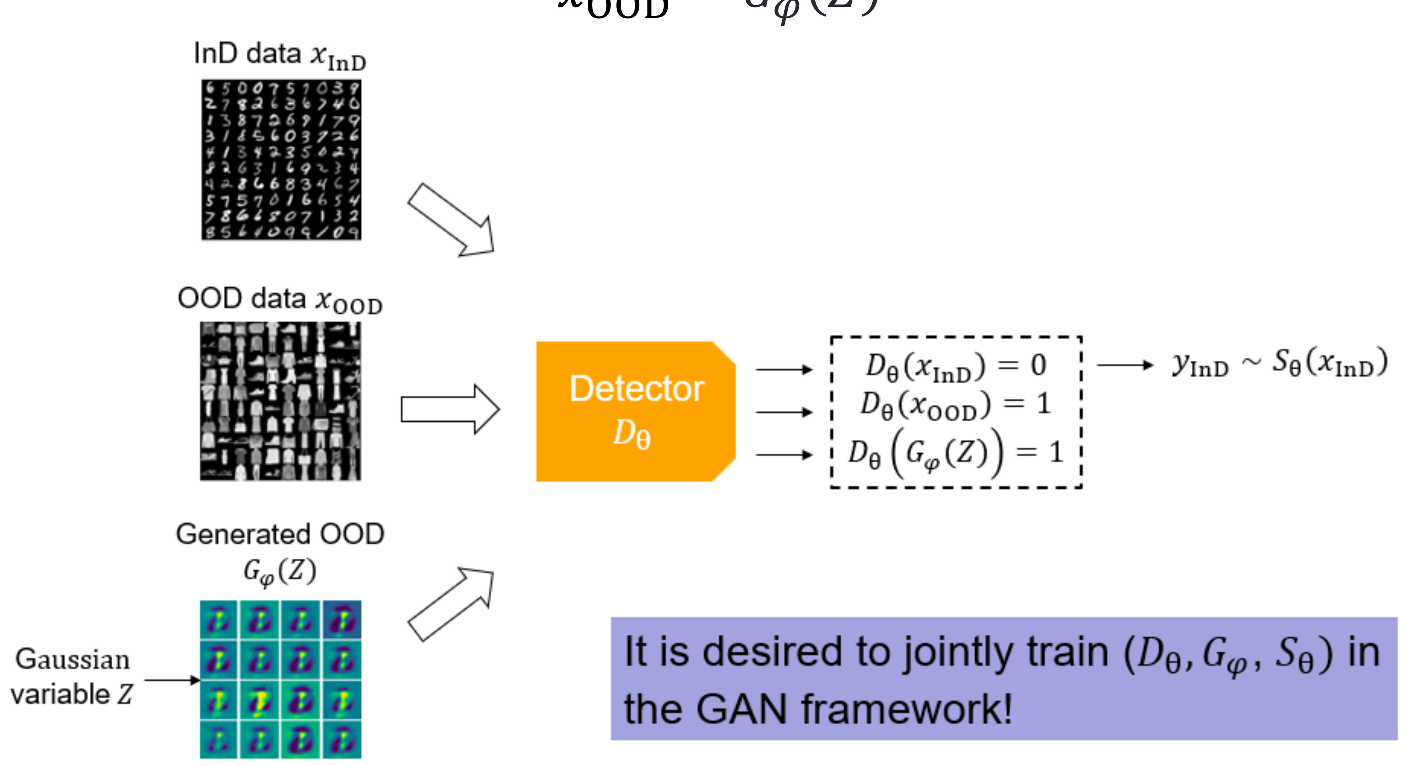
$$D_\theta(x) = \begin{cases} 1, \text{if } x \in \mathcal{X}_{\text{OOD}} \\ 0, \text{if } x \in \mathcal{X}_{\text{InD}} \end{cases}$$

- Task 2: Train a **classifier** such that:

$$y_{\text{InD}} \sim S_\theta(x_{\text{InD}})$$

- Task 3: Train a **generator** for OOD data augmentation:

$$x_{\text{OOD}} \sim G_\varphi(Z)$$



Methods

Overall minimax objective function for GAN:

$$\min_\theta \max_\varphi \; \lambda_{CE} E_{X_{InD}}[CE(S_\theta(X_{InD}), Y_{InD})] + \lambda_d E_{Z, X_{InD}}[d(G_\varphi(Z), X_{InD})] + \lambda_W E_{X_{OOD}}[-\log W(S_\theta(X_{OOD}))] + \lambda_W E_Z[-\log W(D_\theta(G_\varphi(Z)))]$$

Methods

Objective function interpretation:

- $CE = E_{X_{InD}}[CE(S_\theta(X_{InD}), Y_{InD})]$ classification accuracy
- $W_{OOD} = E_{X_{OOD}}[-\log W(S_\theta(X_{OOD}))]$ detection of OOD samples
- $W_Z = E_Z[-\log W(D_\theta(G_\varphi(Z)))]$ detection of generated samples
- $d_{InD} = E_{Z, X_{InD}}[d(G_\varphi(Z), X_{InD})]$ forcing $G_\varphi(Z)$ away from $\mathcal{X}_{\text{InD}}$

Network architecture:

- Traditional CNN architecture for D_θ
- CNN with transposed convolutional layers for G_φ
- In each iteration, update generator twice and discriminator once for a more balanced training.

Wasserstein loss: Joint distribution Distance matrix

$$W(r, c) = \inf_{P \in \Pi(r,c)} \langle P, M \rangle \longrightarrow \text{Kronecker product}$$

$$\Pi(r, c) = \{P \in \mathbb{R}_+^{K \times K} | P\mathbf{1}_K = c, P^\top \mathbf{1}_K = r\}$$

Minimax optimization solution:

During the training stage, we use alternating gradient descent. The objective function for discriminator D_θ and classifier S_θ is the following:

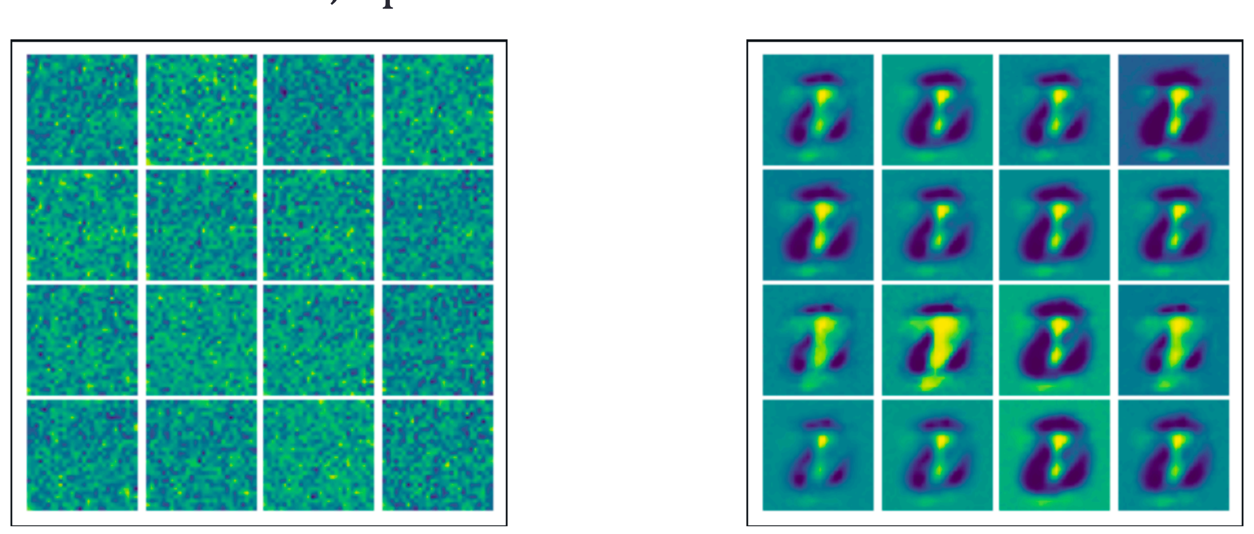
$$\min_\theta \lambda_{CE}\, CE + \lambda_W W_{OOD} + \lambda_W W_Z$$

Similarly, the objective function for generator G_φ is given below:

$$\max_\varphi \lambda_W W_Z + \lambda_d d_{InD}$$

Results

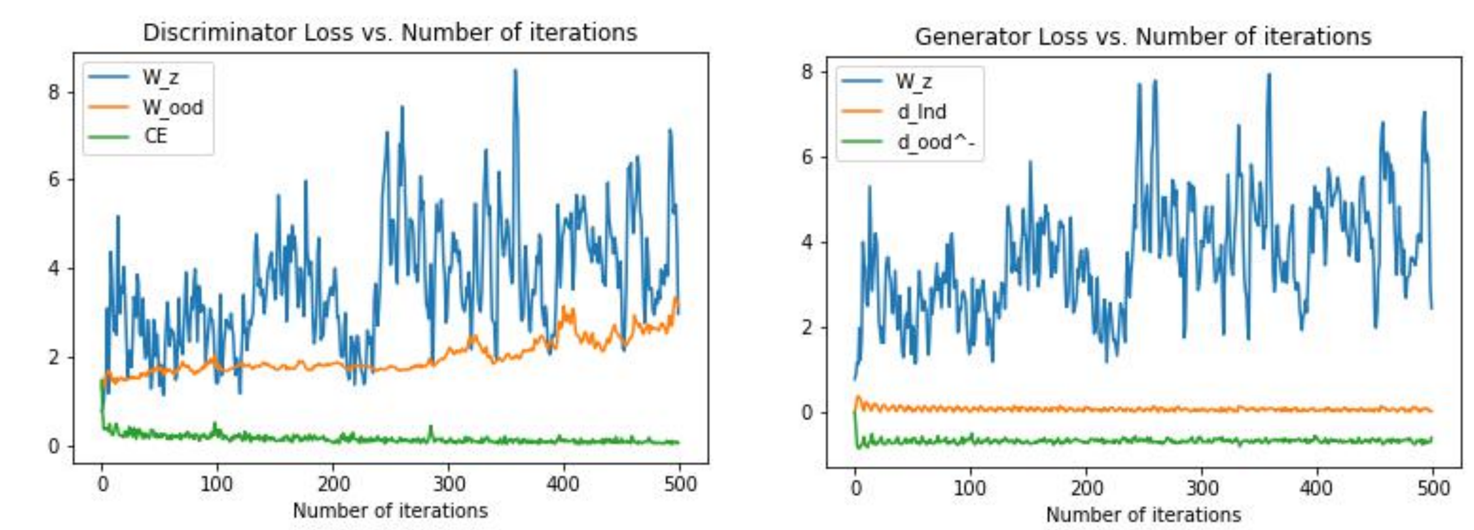
- Dataset: MNIST, InD labels: {2, 3, 6, 8, 9}, OOD labels: {1, 7}
- Distance metric in d_{InD}: image correlation coefficient
- Batch size: 256, Epochs: 15



- Left: generated OOD images before training
- Right: generated OOD images after training for 500 iterations
- The generator generated the horizontal and vertical structures in the OOD samples, while staying away from InD samples

Results

The training loss curve for discriminator (D_θ) and generator (G_φ) are shown below:



Analysis of loss curve.

- CE: Decreasing CE → Better InD sample classification accuracy.
- d_{InD}: Oscillating between [0, 0.1] → Generated images are not similar to InD images.
- d_{OoD}^-: Oscillating between [−1, −0.8] → Generated images are highly correlated (i.e. close to) OoD distributions.
- W_Z: Oscillating with decreasing magnitude → Oscillation illustrates the adversarial process of GANs.
- W_{OoD}: Reached a value of 2.5 after 15 epochs → Based on the definition of Wasserstein distance, a value of 2.5 implies that the model is likely to output uniform Softmax for OoD data.

- The discriminator achieved a prediction accuracy of 98% for InD data classification.
- The discriminator returned Softmax close to uniform for OoD data.

Conclusions

- The proposed OOD GAN discriminator returns low confidence scores for OOD samples and classifies InD data correctly
- The trained generative model recovers the OOD spaces
- With proper tuning and selection of distance metrics, the joint training scheme for OOD GAN shows its effectiveness

Acknowledgement

This is an ongoing UMTRI summer research project supervised by Dr. Wenbo Sun. I would also like to thank Prof. Judy Jin and Prof. Raed AI Kontar for their helpful suggestions.

References

1. Wang, Y., Sun, W., Jin, J., Kong, Z., & Yue, X. (2021). WOOD: Wasserstein-based Out-of-Distribution Detection. *arXiv preprint arXiv:2112.06384*.
2. Lee, K., Lee, H., Lee, K., & Shin, J. (2017). Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.
3. Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., ... & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems, 32*.
4. Choi, J., Yoon, C., Bae, J., & Kang, M. (2021). Robust Out-of-Distribution Detection on Deep Probabilistic Generative Models. *arXiv preprint arXiv:2106.07903*.
5. Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems, 31*.