OPEN SET RECOGNITION BY REGULARISING CLASSIFIER WITH FAKE DATA GENERATED BY GENERATIVE ADVERSARIAL NETWORKS

ICASSP-2018

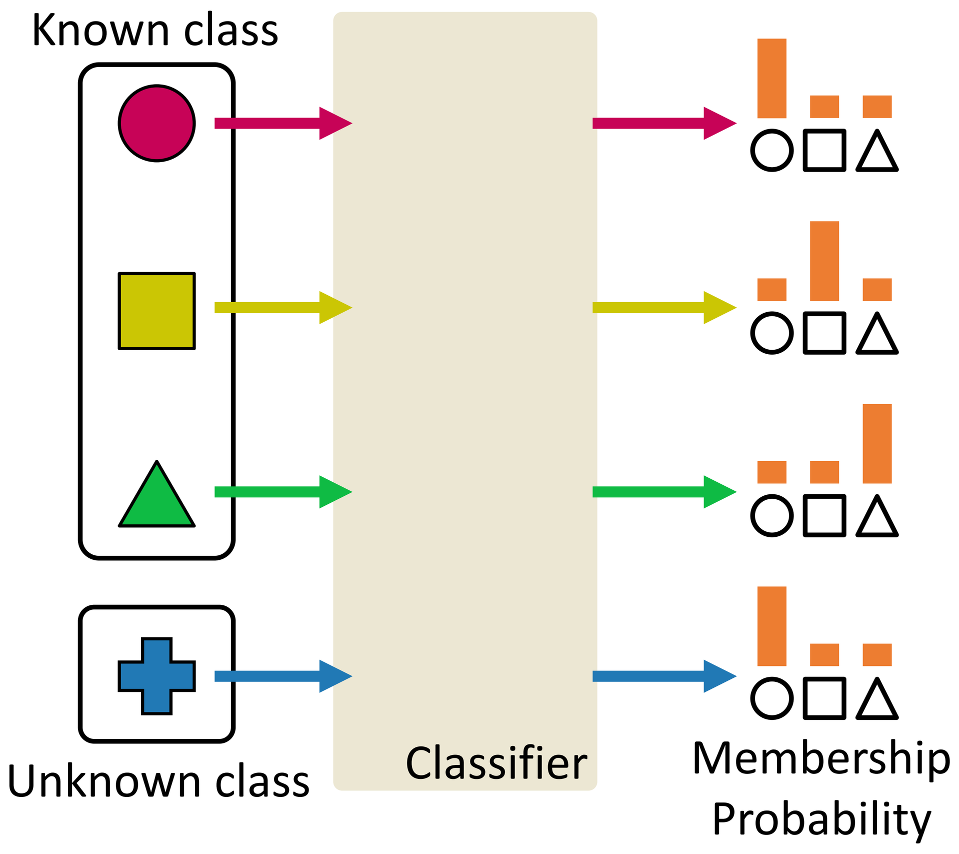
Inhyuk Jo*, Jungtaek Kim*, Hyohyeong Kang#, Yong-Deok Kim#, Seungjin Choi*

*Department of Computer Science and Engineering, POSTECH,

#Software R&D Center, Device Solution, Samsung Electronics, Korea

POSTECH POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

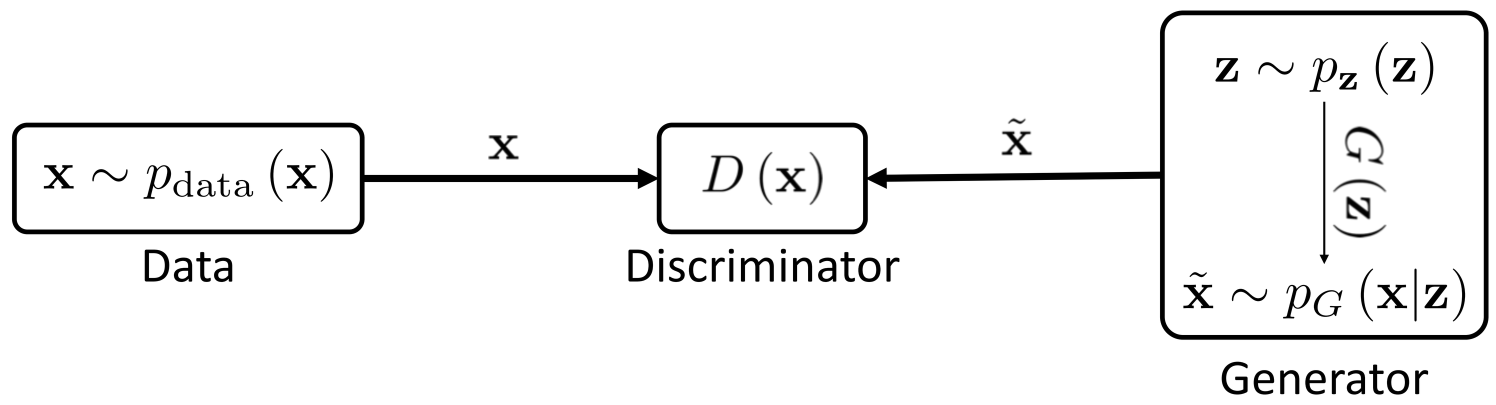
Introduction and Motivation



- Classifier based on deep learning has arrived at human-level.
- However, for unknown classes (not seen during training), the model has no way but predict them as one of known classes with high probability.

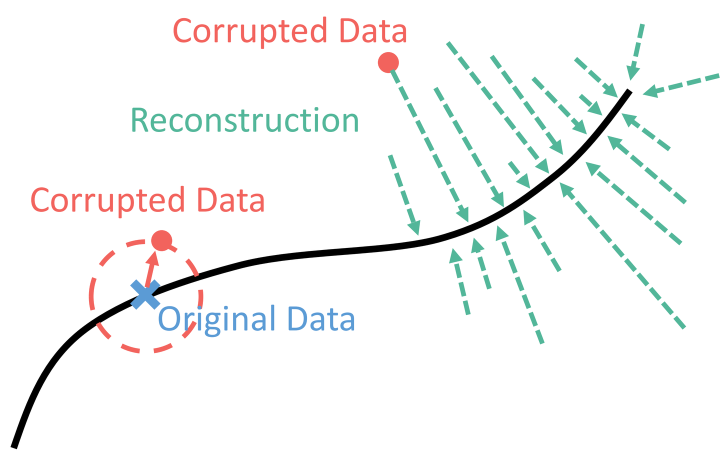
Background

❖ Generative Adversarial Networks (GAN)[1]



$$\min_G \max_D \quad \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

❖ GAN-DFM[2]



- Denoising autoencoder (DAE) models distribution of training data on feature space of discriminator
- Generator is trained to match distribution of training data on feature space of discriminator (DFM)

$$\min_G \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\|\Phi_D(G(\mathbf{z})) - \underbrace{r(\Phi_D(G(\mathbf{z})))}_{\text{fixed}}\|^2 - \log D(G(\mathbf{z}))\right]$$

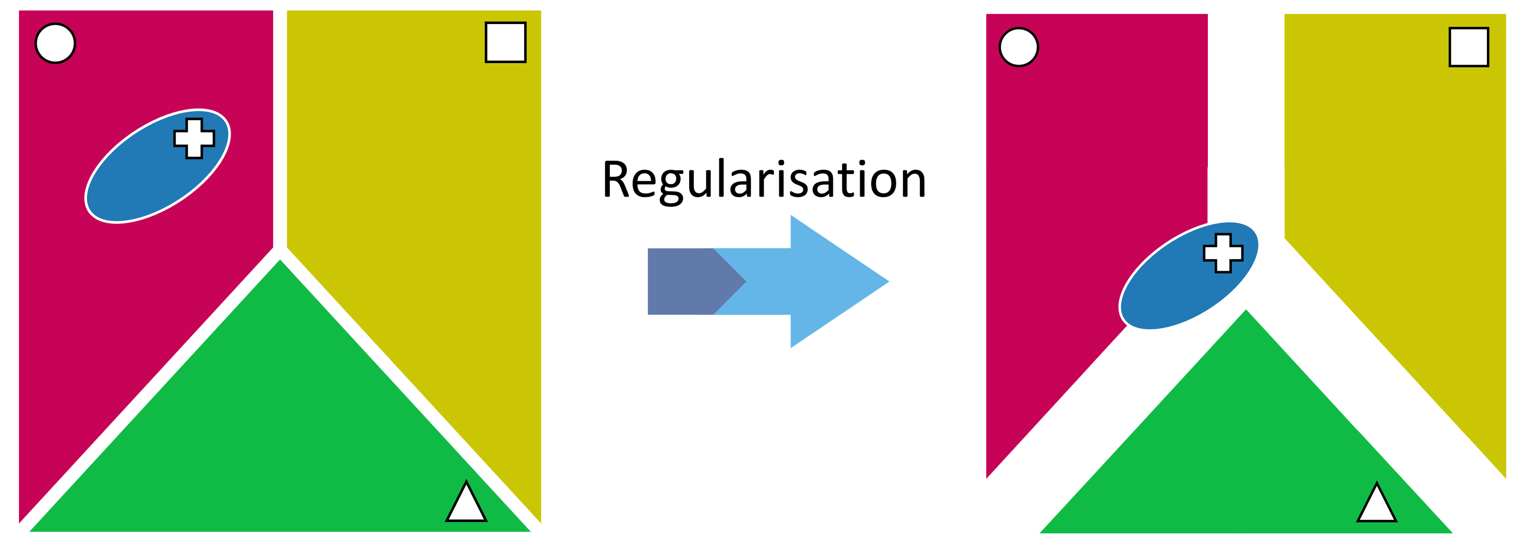
Selected References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Gen-erative adversarial nets," inAdvances in Neural InformationProcessing Systems (NIPS), 2014.

[2] D. Warde-Farley and Y. Bengio, "Improving generative adver-sarial networks with denoising feature matching," inProceed-ings of the International Conference on Learning Representa-tions (ICLR), 2017.

GAN-Marginal DFM (GAN-MDFM)

❖ Problematic behavior of classifier



Decision boundary on feature space of classifier
Known class (circle, square, triangle), Unknown class (cross)

- If we could generate fake data located on feature space nearby space of known classes, classifier would be easily trained by minimising cross entropy and miximising entropy of fake data.
- To train classifier end-to-end, fake data is necessary.
  - ➡ Tightening decision boundary.
  - ➡ Separating unknown classes from known classes on feature space.

❖ Objective

- Classifier: minimise cross entropy with positive data (known classes) maximise entropy with fake negative data (unknown classes)

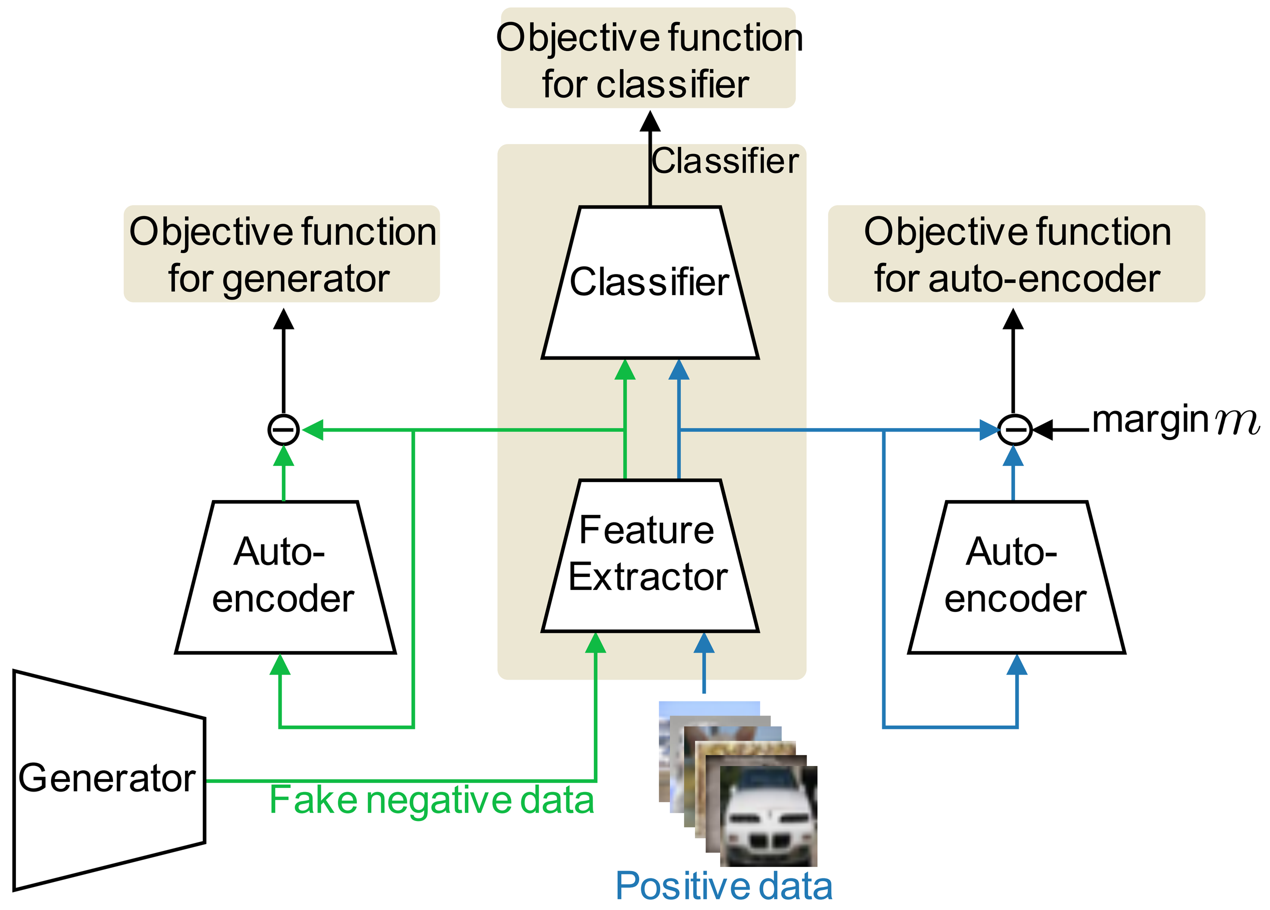
$$\min_C -\mathbb{E}_{\mathbf{x},y\sim p_{\text{data}}(\mathbf{x},y)}[\log p_C(y|\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim p_G(\mathbf{x})}[H(p_C(y|\mathbf{x}))]$$

- Marginal Denoising Autoencoder (MDAE): model m noisy feature distribution of known classes

$$\min_M \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[\left|\|\Phi(\mathbf{x}) - M(n(\Phi(\mathbf{x})))\|^2 - m\right|\right]$$

- Generator: generate fake negative data that is located on **m** away feature space from the one known classes

$$\min_G \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\|\Phi(G(\mathbf{z})) - \underbrace{M(\Phi(G(\mathbf{z})))}_{\text{fixed}}\|^2\right]$$



Experimental Results

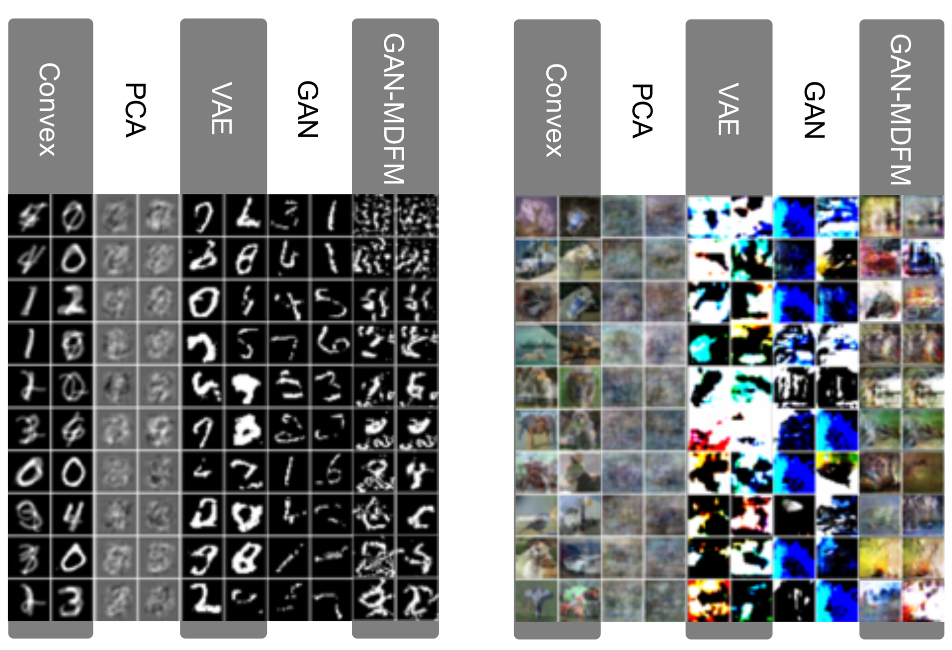
- Regularisation of GAN-MDFM did not degenerate accuracy at all but even improved on CIFAR10.
- Our model outperformed other methods in terms of Area Under the Curve (AUC).
- Generated data seemed similar to training data but not exactly as our purpose.

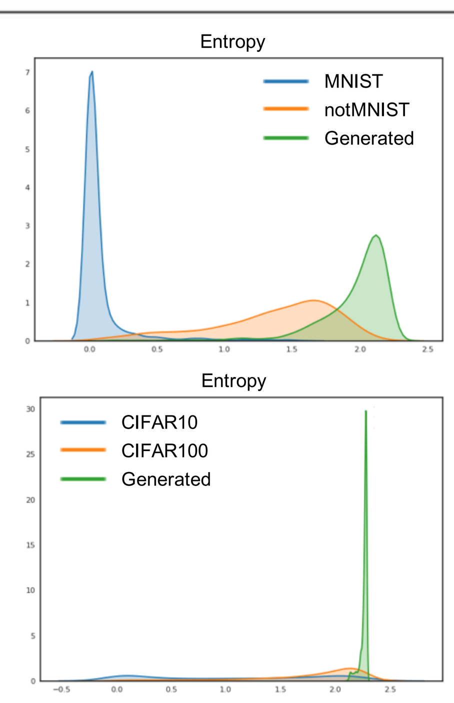
Table 1. Classification accuracy

	Baseline	Convex	PCA	VAE	GAN	GAN-MDFM
MNIST	0.987	0.986	0.987	0.988	**0.991**	0.987
CIFAR10	0.707	0.654	0.700	0.702	0.616	**0.728**

Table 2. Area under the curve

		Baseline	Convex	PCA	VAE	GAN	T-scaling	GAN-MDFM
MNIST vs. notMNIST	entropy	0.930	0.976	0.907	0.926	**0.987**	0.938	**0.987**
	max logit	0.885	0.969	0.840	0.865	0.982	0.887	**0.991**
CIFAR10 vs. CIFAR100	entropy	0.666	0.671	0.656	0.666	0.641	0.707	**0.729**
	max logit	0.696	0.664	0.687	0.691	0.641	0.696	**0.721**





Conclusions

- We have proposed a unknown class generator with assistance of MDAE.
- Our generated data is well-designed augmented data for regularising classifier.