JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

A Joint Representation Learning and Feature Modeling Approach for One-class Recognition

Pramuditha Perera
Vishal Patel

Outline

- What is One Class Classification(OCC)?
- Two Paradigms of OCC
  - Representation Learning
  - Feature Modelling
- Limitations of Existing Methods
- Proposed Method
- Results

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

What is One Class Classification(OCC)

- An Extreme case in classification
- Knowledge of the classifier is limited to only a single class
- Given training samples from a class, the classifier is expected to reject samples from any outside class

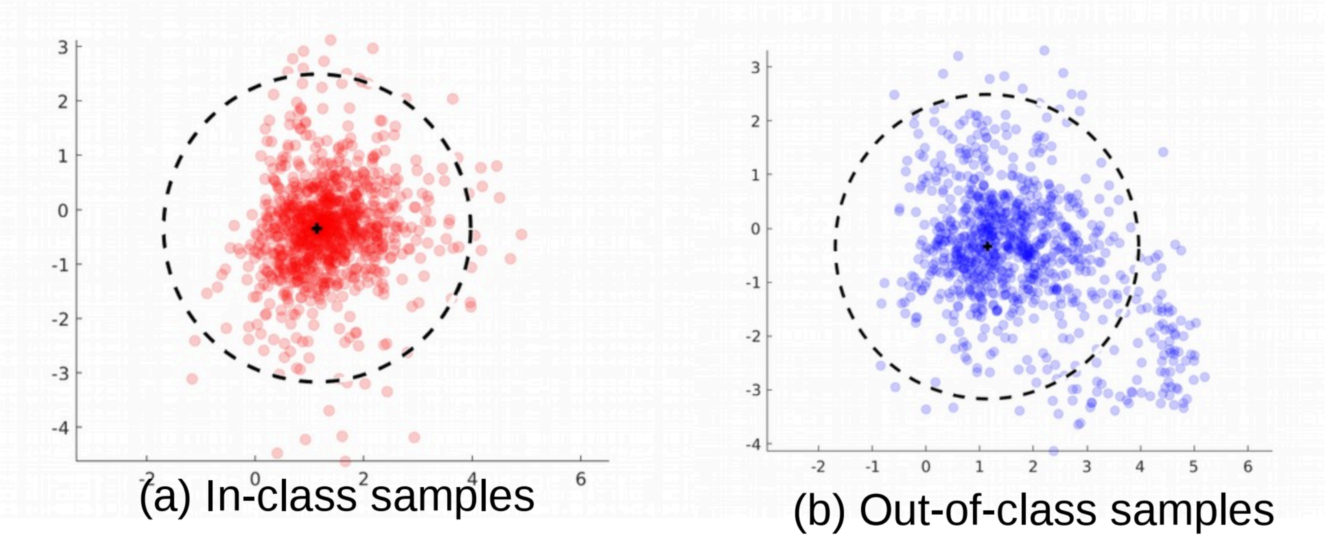
Two Paradigms of OCC

- Feature modelling
  - Use a one-class modeling method to identify the positive space in a given feature space.
  - Objects appearing outside the positive space are identified as out-of-class samples.
- Representation learning
  - An in-class representation learned during training.
  - During inference, test if the model is able to represent an input sample.

Limitations of Feature modelling

- Redundant space could be identified as a part of the positive space. Eg: redundant white-space in (a).

- Lack of guarantee that out-of-class samples will not get projected inside the identified decision boundary (b).



(a) In-class samples (b) Out-of-class samples

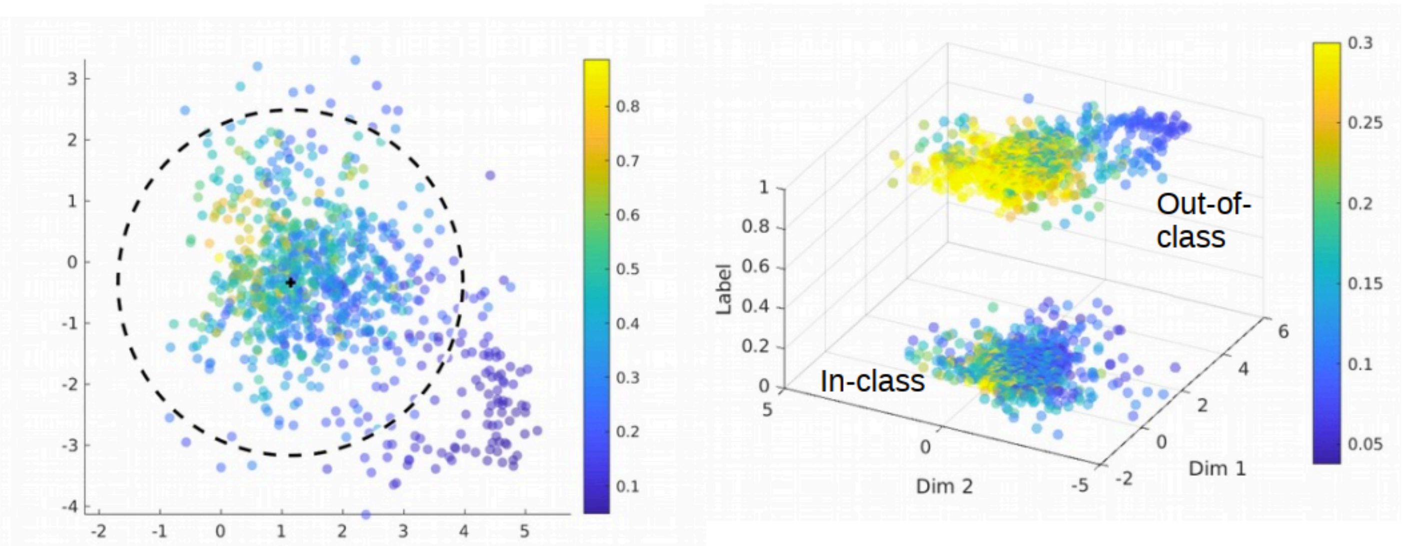
Two Paradigms of OCC

- Feature modelling
  - Use a one-class modeling method to identify the positive space in a given feature space.
  - Objects appearing outside the positive space are identified as out-of-class samples.
- Representation learning
  - An in-class representation learned during training.
  - During inference, test if the model is able to represent an input sample.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Limitations of Representation learning

- In-class samples are well represented.
- No guarantee that out of-class samples will not be represented well in the learned space.
- Specially when the representation is generic.



(c) Out-of-class samples (MSE)

(d) All samples (MSE)

Proposed Method

- Feature modelling fails only when out-of-class samples get projected inside the identified positive space.
- Provided that,
  - latent space is smooth
  - each latent code inside the positive space corresponds to an in-class sample

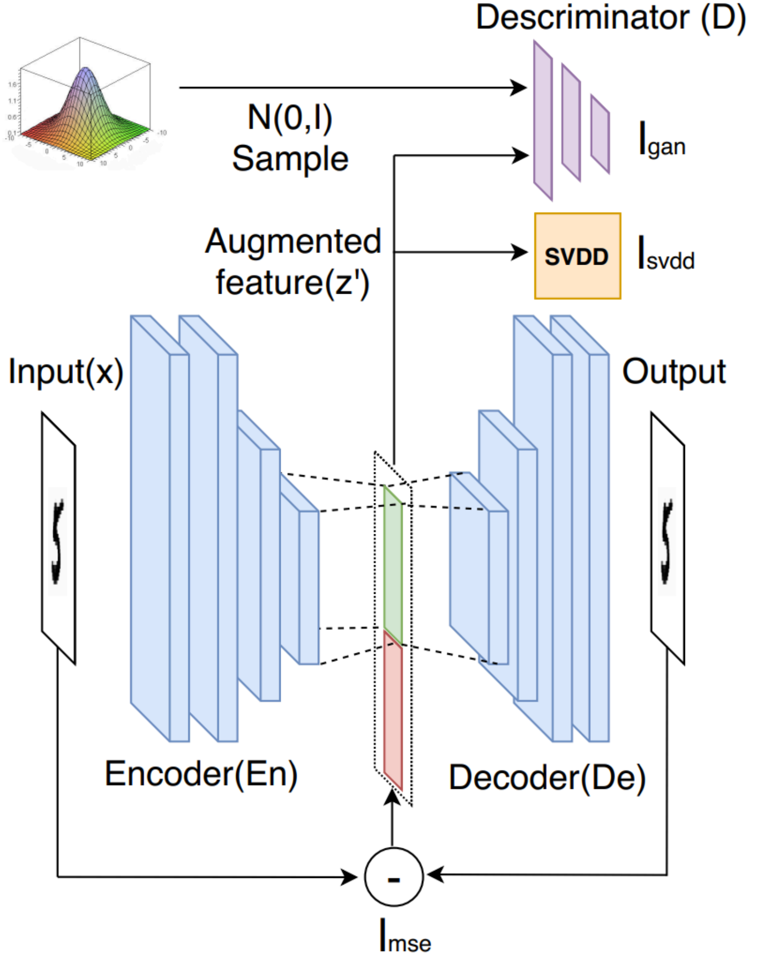
failed cases can be identified considering the reconstruction error.

	Inside + Space		Outside + Space	
	Low MSE	High MSE	Low MSE	High MSE
Feature Modeling	FP	FP	TN	TN
Representation Learning	FP	TN	FP	TN
Proposed	FP	TN	TN	TN

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Proposed Method

- Autoencoder network that is trained on reconstruction loss
  $l_{mse} = \|x - \hat{x}\|^2$ where, $\hat{x} = De(En(x))$
- Extend the latent space by appending MSE to the latent feature
- Force extended latent features to follow a pre-determined distribution
  $l_{gan} = \mathbb{E}_{s \sim N(0,I) \in \mathbb{R}^{2k}}[logD(s)] + \mathbb{E}_{x \sim p_{z'}}[log(1 - D(z'))]$
- Fit a one-class classifier on extended feature space

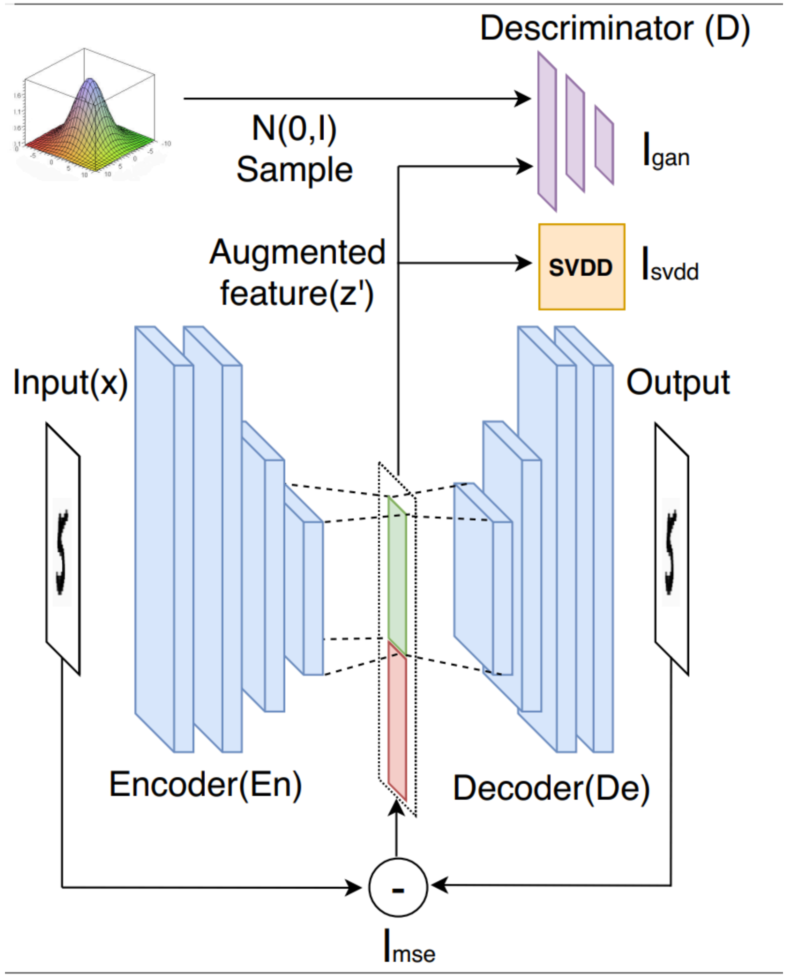


Prevent out-of-class samples from entering positive space

Proposed Method

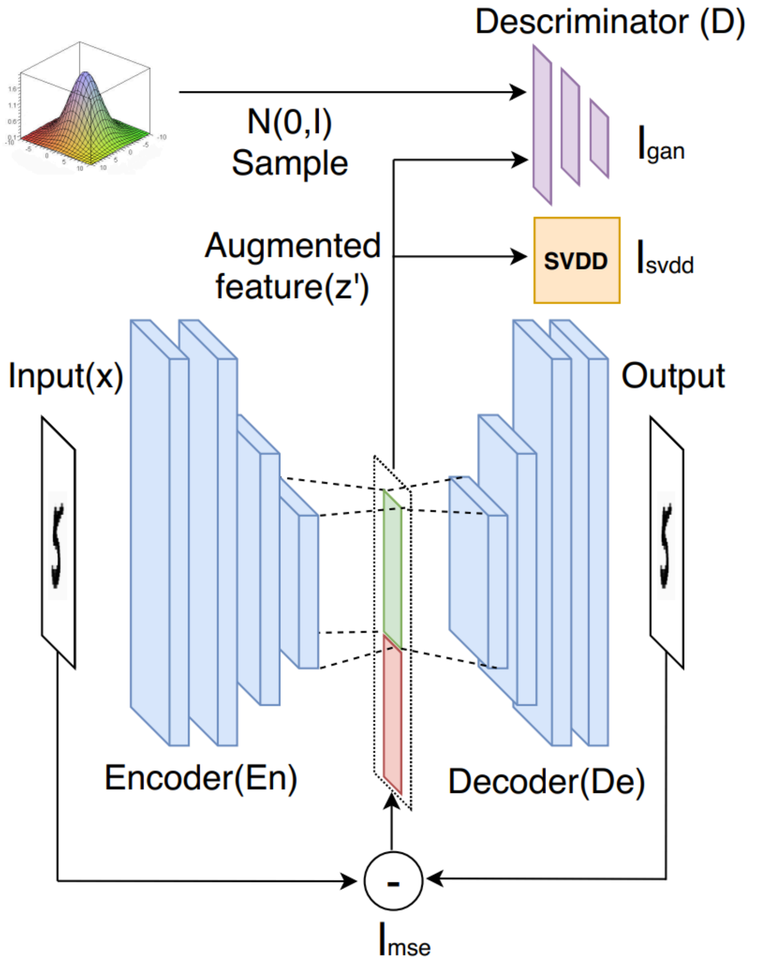
- Autoencoder network that is trained on reconstruction loss
  $l_{mse} = \|x - \hat{x}\|^2$ where, $\hat{x} = De(En(x))$
- Extend the latent space by appending MSE to the latent feature
- Force extended latent features to follow a pre-determined distribution
  $$l_{gan} = \mathbb{E}_{s \sim N(0,I) \in \mathbb{R}^{2k}}[log D(s)] + \mathbb{E}_{x \sim p_{z'}}[log(1 - D(z'))].$$
- Fit a one-class classifier on extended feature space

Reducing redundant positive space in the OCC



Proposed Method

- Autoencoder network that is trained on reconstruction loss
  $l_{mse} = \|x - \hat{x}\|^2$ where, $\hat{x} = De(En(x))$
- Extend the latent space by appending MSE to the latent feature
- Force extended latent features to follow a pre-determined distribution
  $l_{gan} = \mathbb{E}_{s \sim N(0,I) \in \mathbb{R}^{2k}}[log D(s)] + \mathbb{E}_{x \sim p_{z'}}[log(1 - D(z'))].$
- Fit a one-class classifier on extended feature space



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Proposed Method

- Pre-determined distribution should be chosen to minimize white space volume in the positive half space
- In our work we chose SVDD as our choice of OCC
- We considered following criterion when selecting a distribution:
  - Distribution should be unimodal.
  - Distribution should be isotropic
  - Distribution should not have long tails.
- Gaussian distribution, student-t distribution and Cauchy distribution are good candidates
- We experimented using the Gaussian distribution

Johns Hopkins Whiting School of Engineering

Results

Average AUC on MNIST dataset

Class	OCSVM[10]		KDE[7]		IF[7]		DCAE[19]		ANOGAN[27] (IPIM17)		SDOCC[4] (ICML18)		DOCC[4] (ICML18)		AND*[6] (CVPR19)		OCGAN*[39] (CVPR19)		AE+SVDD		Ours	
0	98.6	*0.0*	97.1	*0.0*	98.0	*0.3*	97.6	*0.0*	96.6	*1.3*	97.8	*0.7*	98.0	*0.7*	99.3	*0.0*	**99.8**	*0.0*	96.8	*0.0*	99.6	*0.1*
1	99.5	*0.0*	98.9	*0.0*	97.3	*0.4*	98.3	*0.0*	99.2	*0.6*	99.6	*0.1*	99.7	*0.1*	**99.9**	*0.0*	**99.9**	*0.0*	99.3	*0.0*	98.8	*0.7*
2	82.5	*0.1*	79.0	*0.0*	88.6	*0.5*	85.4	*0.0*	85.0	*2.9*	89.5	*1.2*	91.7	*0.8*	95.9	*0.0*	94.2	*0.0*	83.4	*0.0*	**97.2**	*0.5*
3	88.1	*0.0*	86.2	*0.0*	89.9	*0.4*	86.7	*0.0*	88.7	*2.1*	90.3	*2.1*	91.9	*1.5*	**96.6**	*0.0*	96.3	*0.0*	86.8	*0.0*	95.5	*0.3*
4	94.9	*0.0*	87.9	*0.0*	92.7	*0.6*	86.5	*0.0*	89.4	*1.3*	93.8	*1.5*	94.9	*0.8*	95.6	*0.0*	**97.5**	*0.0*	92.4	*0.0*	95.7	*0.4*
5	77.1	*0.0*	73.8	*0.0*	85.5	*0.8*	78.2	*0.0*	88.3	*2.9*	85.8	*2.5*	88.5	*0.9*	96.4	*0.0*	**98.0**	*0.0*	75.8	*0.0*	96.3	*0.5*
6	96.5	*0.0*	87.6	*0.0*	95.6	*0.3*	94.6	*0.0*	94.7	*2.7*	98.0	*0.4*	98.3	*0.5*	**99.4**	*0.0*	99.1	*0.0*	93.1	*0.0*	98.8	*0.3*
7	93.7	*0.0*	91.4	*0.0*	92.0	*0.4*	92.3	*0.0*	93.5	*1.8*	92.7	*1.4*	94.6	*0.9*	98.0	*0.0*	**98.1**	*0.0*	92.6	*0.0*	95.7	*0.3*
8	88.9	*0.0*	79.2	*0.0*	89.9	*0.4*	86.5	*0.0*	84.9	*2.1*	92.9	*1.4*	93.9	*1.6*	95.3	*0.0*	93.9	*0.0*	88.9	*0.0*	**95.4**	*0.4*
9	93.1	*0.0*	88.2	*0.0*	93.5	*0.3*	90.4	*0.0*	92.4	*1.1*	94.9	*0.6*	96.5	*0.3*	**98.1**	*0.0*	**98.1**	*0.0*	93.7	*0.0*	97.7	*0.2*
Mean	91.3	*0.0*	86.9	*0.0*	92.3	*0.4*	89.7	*0.0*	91.3	*1.9*	93.5	*1.2*	94.8	*0.8*	**97.5**	*0.0*	**97.5**	*0.0*	90.2	*0.0*	97.1	*0.4*

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results

Average AUC on CIFAR10 dataset

Class	OCSVM[10]		KDE[7]		IF[7]		DCAE[19]		ANOGAN[27] (IPIM17)		SDOCC[4] (ICML18)		DOCC[4] (ICML18)		AND*[6] (CVPR19)		OCGAN*[39] (CVPR19)		AE+SVDD		Ours	
Plane	61.6	*0.9*	61.2	*0.0*	60.1	*0.7*	59.1	*5.1*	67.1	*2.5*	61.7	*4.2*	61.7	*4.1*	73.5	*0.0*	**75.7**	*0.0*	55.2	*0.0*	66.4	*1.5*
Car	63.8	*0.6*	64.0	*0.0*	50.8	*0.6*	57.4	*2.9*	54.7	*3.4*	64.8	*1.4*	65.9	*2.1*	58.0	*0.0*	53.1	*0.0*	73.0	*0.0*	**78.5**	*0.6*
Bird	50.0	*0.5*	50.1	*0.0*	49.2	*0.4*	48.9	*2.4*	52.9	*3.0*	49.5	*1.4*	50.8	*0.8*	**69.0**	*0.0*	64.0	*0.0*	49.1	*0.0*	54.9	*0.6*
Cat	55.9	*1.3*	56.4	*0.0*	55.1	*0.4*	58.4	*1.2*	54.5	*1.9*	56.0	*1.1*	**59.1**	*1.4*	54.2	*0.0*	62.0	*0.0*	53.6	*0.0*	57.3	*0.6*
Deer	66.0	*0.7*	66.2	*0.0*	49.8	*0.4*	54.0	*1.3*	65.1	*3.2*	59.1	*1.1*	60.9	*1.1*	**76.1**	*0.0*	72.3	*0.0*	61.1	*0.0*	73.6	*0.1*
Dog	62.4	*0.8*	62.4	*0.0*	58.5	*0.4*	62.2	*1.8*	60.3	*2.6*	62.1	*2.4*	**65.7**	*2.5*	54.6	*0.0*	62.0	*0.0*	60.4	*0.0*	63.1	*0.4*
Frog	74.7	*0.3*	74.9	*0.0*	42.9	*0.6*	51.2	*5.2*	58.5	*1.4*	67.8	*2.4*	67.7	*2.6*	75.1	*0.0*	72.3	*0.0*	62.6	*0.0*	**80.8**	*0.1*
Horse	62.6	*0.6*	62.6	*0.0*	55.1	*0.7*	58.6	*2.9*	62.5	*0.8*	65.2	*1.0*	67.3	*0.9*	53.5	*0.0*	57.5	*0.0*	69.1	*0.0*	**72.0**	*1.1*
Ship	74.9	*0.4*	75.1	*0.0*	74.2	*0.6*	76.8	*1.4*	75.8	*4.1*	75.6	*1.7*	75.9	*1.2*	71.7	*0.0*	**82.0**	*0.0*	74.7	*0.0*	80.3	*0.6*
Truck	75.9	*0.3*	76.0	*0.0*	58.9	*0.7*	67.3	*3.0*	66.5	*2.8*	71.0	*1.1*	73.1	*1.2*	54.8	*0.0*	55.4	*0.0*	77.8	*0.0*	**79.9**	*1.0*
Mean	64.8	0.6	64.9	*0.0*	55.5	*0.6*	59.4	*2.7*	61.8	*2.6*	63.3	*1.8*	64.8	*1.8*	64.1	*0.0*	65.7	*0.0*	63.6	*0.0*	**70.7**	*0.7*

Results

Average AUC on GTSRB STOP SIGN dataset (Adversarial sample detection)

OCSVM [10]	67.5	*1.2*
KDE [7]	60.5	*1.7*
IF [7]	73.8	*0.9*
DCAE [19]	79.1	*3.0*
SDOCC [4]	77.8	*4.9*
DOCC [4]	80.3	*2.8*
Ours	**85.2**	*0.7*

Results

True positives

False Negatives

False Positives

Thank You!