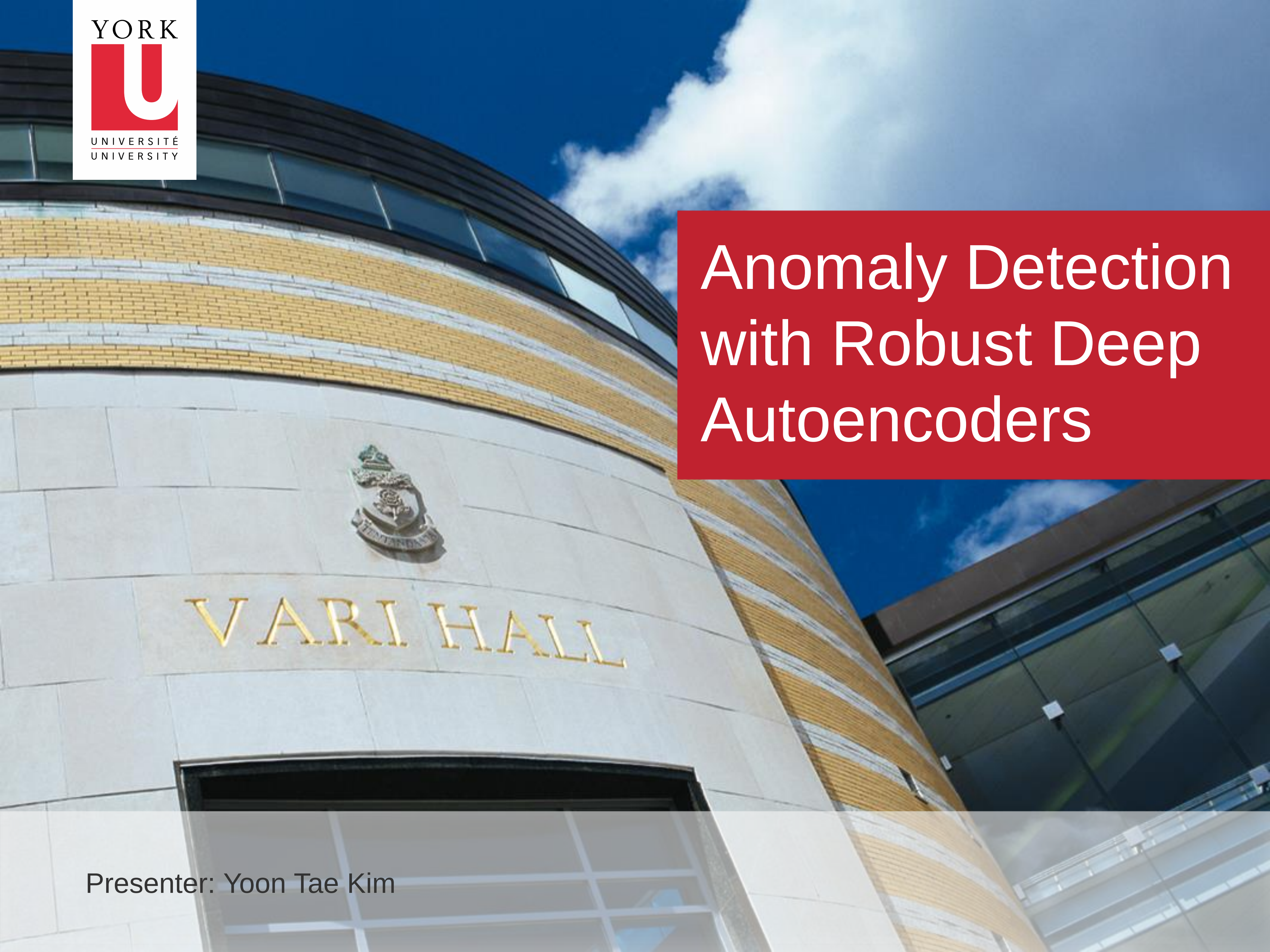


YORK U
UNIVERSITÉ
UNIVERSITY

Anomaly Detection with Robust Deep Autoencoders

Presenter: Yoon Tae Kim

Agenda

1) Main Objective

2) Related Works

3) Background

4) Methodology

5) Algorithm Training

6) Evaluation

7) Summary

YORK UNIVERSITÉ UNIVERSITY

1) Main Objective

The purpose of this paper is to introduce a novel deep autoencoder which

i) extracts high quality features and

ii) detects anomalies without any clean data

2) Related Works

i) **Denoising Autoencoders**

- A extension of standard autoencoder which is designed to detect more robust features.

- This type of autoencoders require noise-free training data.

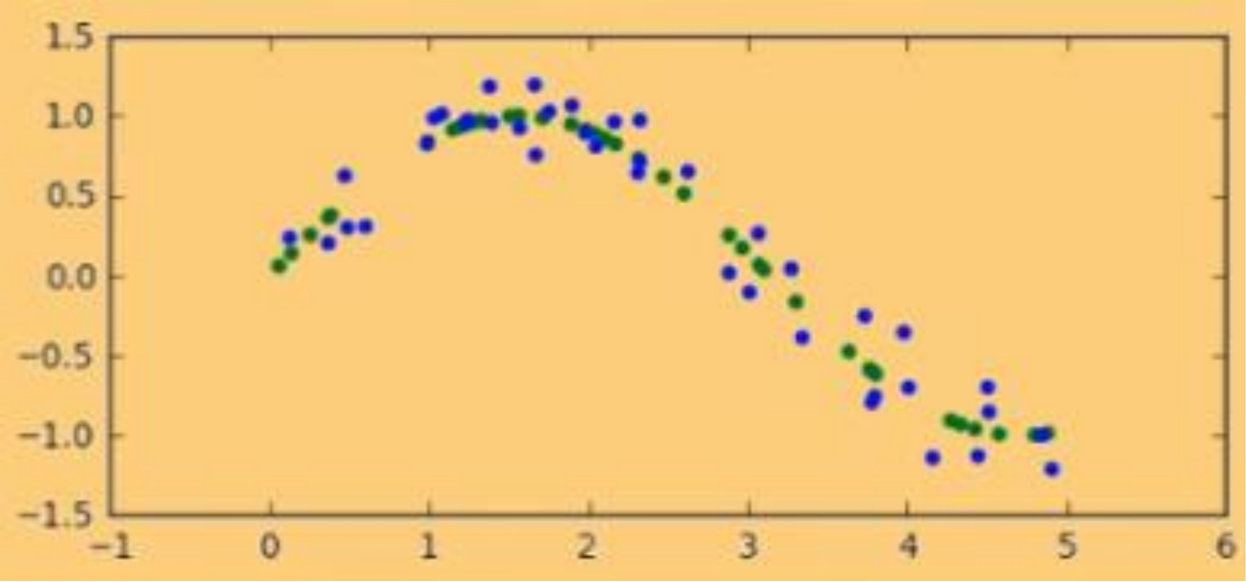
ii) **Maximum Correntropy Autoencoder**

- A deep autoencoder which uses correntropy as the reconstruction cost.

- Even though the model use the training data including anomalies, the highly corrupted data still reduce the quality of representations.

3) Background

Deep Autoencoder



3) Background

Robust Principal Component Analysis(RPCA)

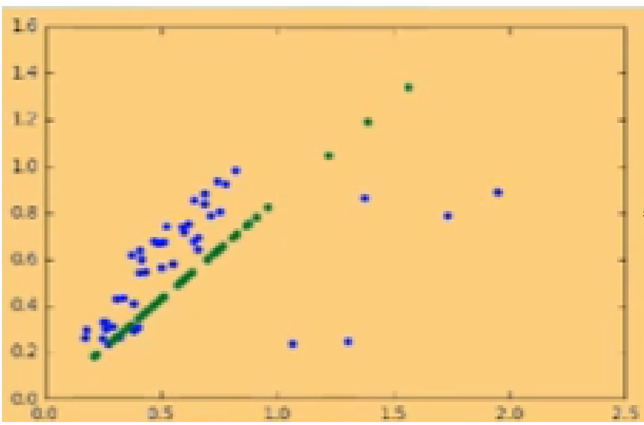
- Advanced model of Principal Component Analysis (PCA) that is more robust to outliers.
- The main idea of this model is isolating sparse noise matrix S so that the remaining low-dimensional matrix L becomes noise-free.

$$X = L + S$$

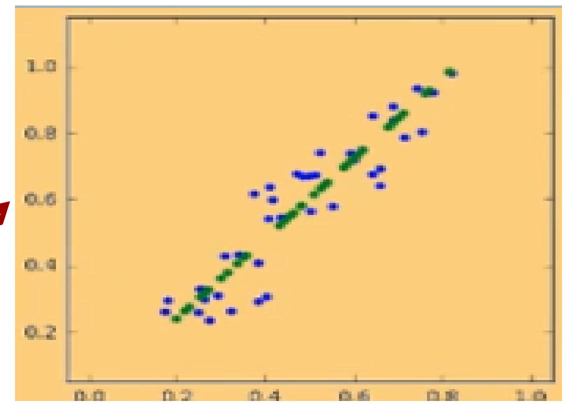
(L: Low–rank matrix, S: Sparse matrix)

3) Background

Robust Principal Component Analysis

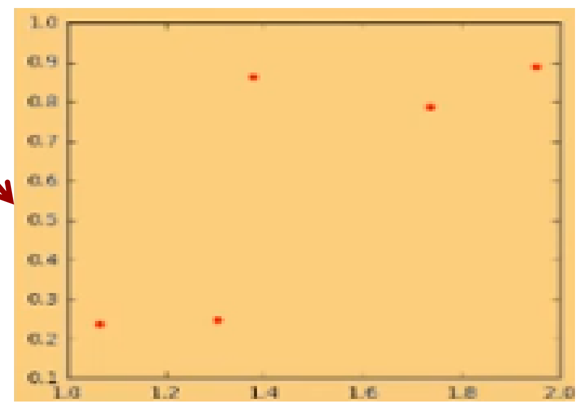


X



L

(Clean Data)



S

(Noise Data)

X = L + S

3) Background

Robust Principal Component Analysis(RPCA)

Convex Relaxations

$$\min_{L,S} \rho(L) + \lambda||S||_0 \quad \text{s.t. } \|X - L - S\|_F^2 = 0,$$

$$\longrightarrow$$

$$\min_{L,S} ||L||_* + \lambda||S||_1 \quad \text{s.t. } \|X - L - S\|_F^2 = 0,$$

Non-Convex Optimization | **Convex Optimization**

3) Background

Robust Principal Component Analysis(RPCA)

Rank of L

Zero Norm:
\# of non-zero entries in S

Convex Relaxations

Nuclear Norm:
The sum of the singular values of matrix

One Norm:
The sum of absolute values of entries

$$\min_{L,S} \rho(L) + \lambda ||S||_0$$

$$\text{s.t. } \|X - L - S\|_F^2 = 0,$$

$$\min_{L,S} ||L||_* + \lambda ||S||_1$$

$$\text{s.t. } \|X - L - S\|_F^2 = 0,$$

Non-Convex Optimization

Convex Optimization

Frobenius norm:
the square root of the sum
of the absolute squares of its elements

3) Background

<u>Advantage of Deep Autoencoder</u>

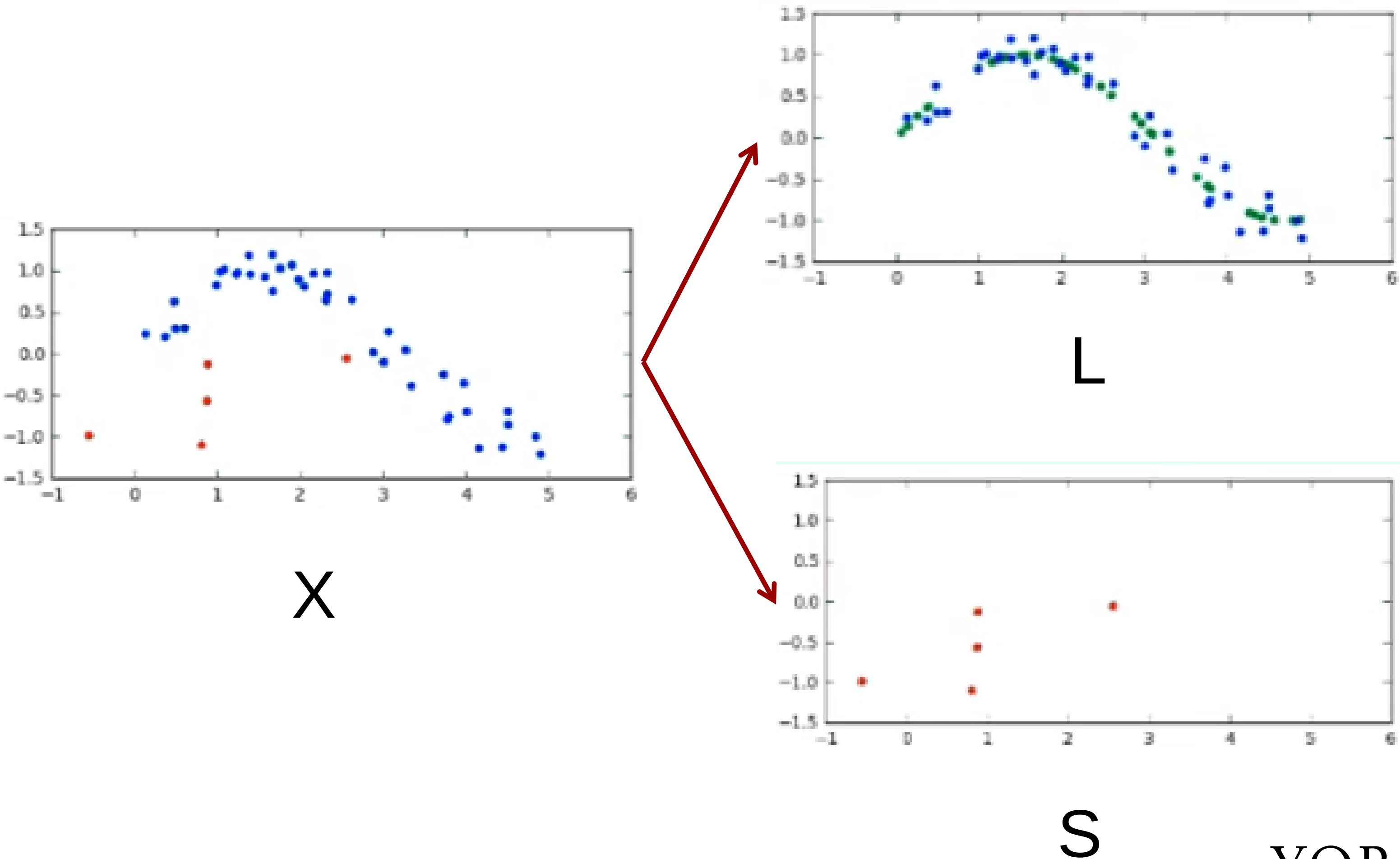
- the non-linear representation capability

<u>Advantage of RPCA</u>

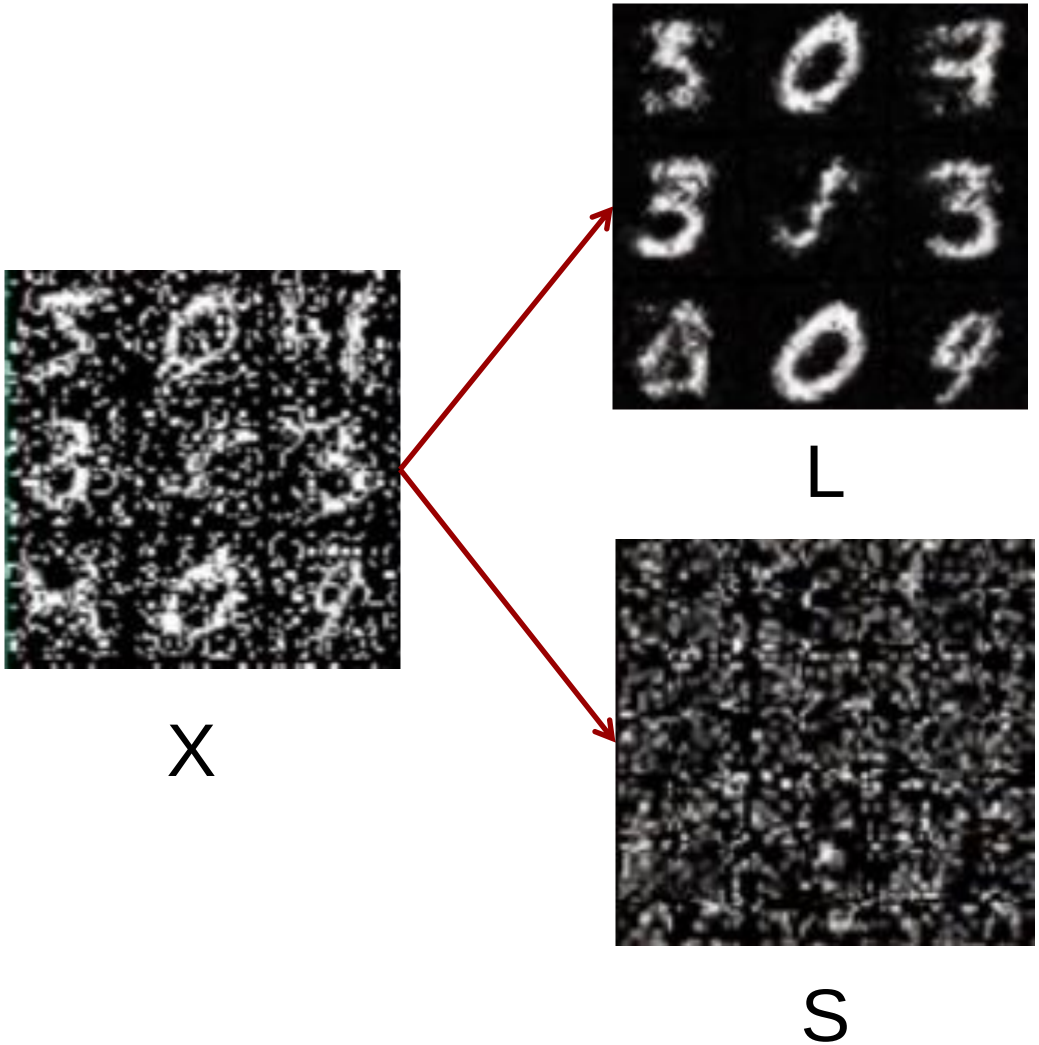
- the anomaly detection capability

=> Robust Deep Autoencoder inherits two advantages.

3) Background



3) Background



3) Methodology

Robust Deep Autoencoder

- This autoencoder is a combined model of deep autoencoder and Robust PCA.
- This autoencoder extracts robust features by isolating anomalies in training data.

Two types of Robust Deep Autoencoder

a) Robust Deep Autoencoder with L1 Regularization

b) Robust Deep Autoencoder with L2,1 Regularization

3) Methodology

I) Robust Deep Autoencoder with L1 Regularization

Convex Relaxations

$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda \|S\|_0$$
$$\text{s.t. } X - L_D - S = 0,$$

$$\longrightarrow$$

$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda \|S\|_1$$
$$\text{s.t. } X - L_D - S = 0.$$

3) Methodology

I) Robust Deep Autoencoder with L1 Regularization

Reconstruction Error of L

One Norm of S:
= The sum of absolute values of entries

Convex Relaxations

$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda\|S\|_0$$
$$\text{s.t. } X - L_D - S = 0,$$

$$\longrightarrow$$

$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda\|S\|_1$$
$$\text{s.t. } X - L_D - S = 0.$$

Zero Norm of S
= # of non-zero entries in S

3) Methodology

I) Robust Deep Autoencoder with L1 Regularization

Convex Relaxations

$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda\|S\|_0$$
$$\text{s.t. } X - L_D - S = 0,$$

$\rightarrow$

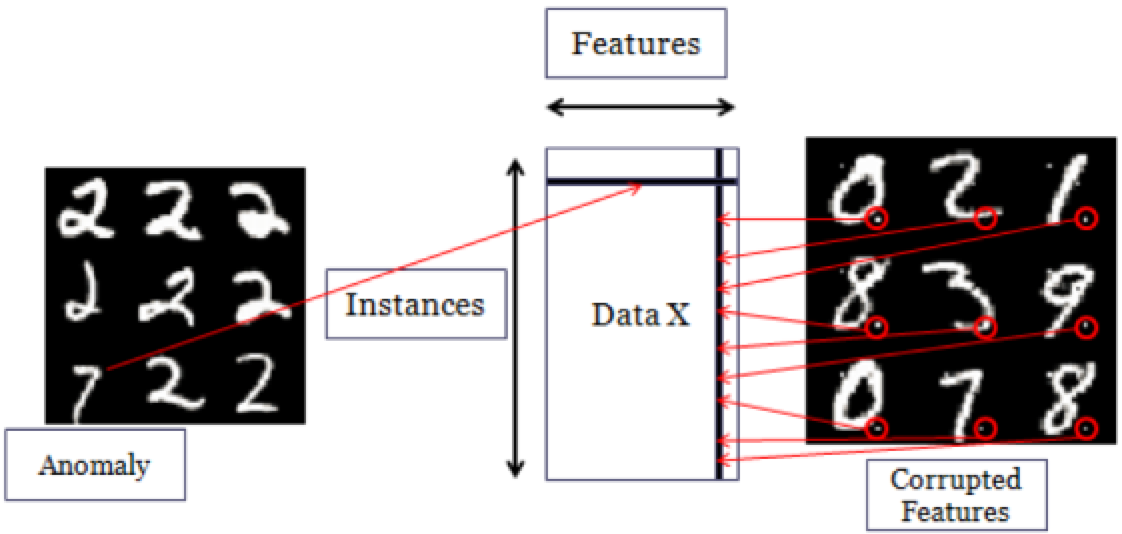
$$\min_{\theta} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda\|S\|_1$$
$$\text{s.t. } X - L_D - S = 0.$$

a) The smaller Lambda λ, The lower level of sparsity in S
b) The larger Lambda λ, The higher level of sparsity in S

Lambda λ = a parameter that controls the level of sparsity in S

3) Methodology

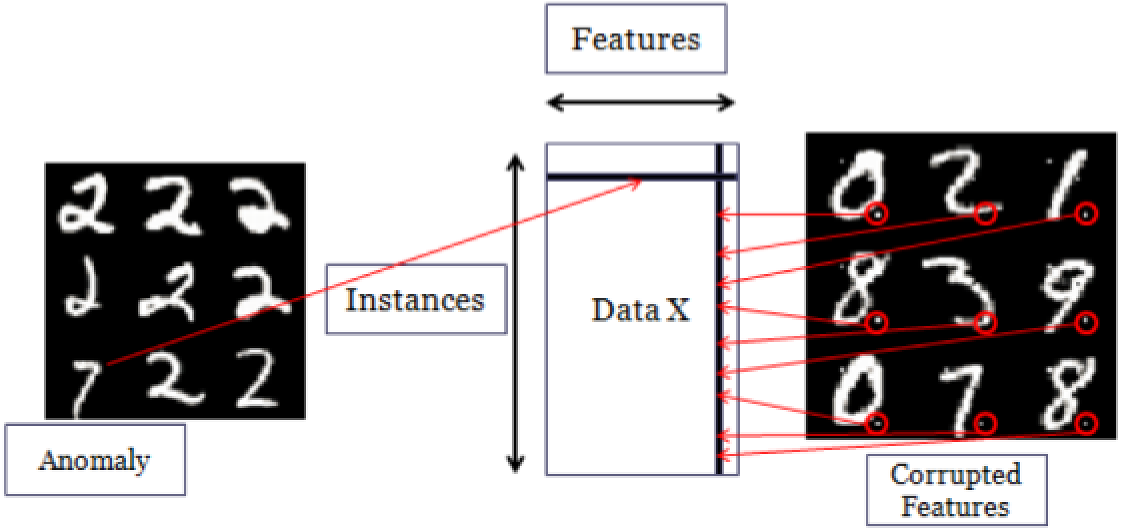
II) Robust Deep Autoencoder with L2,1 Regularization



3) Methodology

II) Robust Deep Autoencoder with L2,1 Regularization

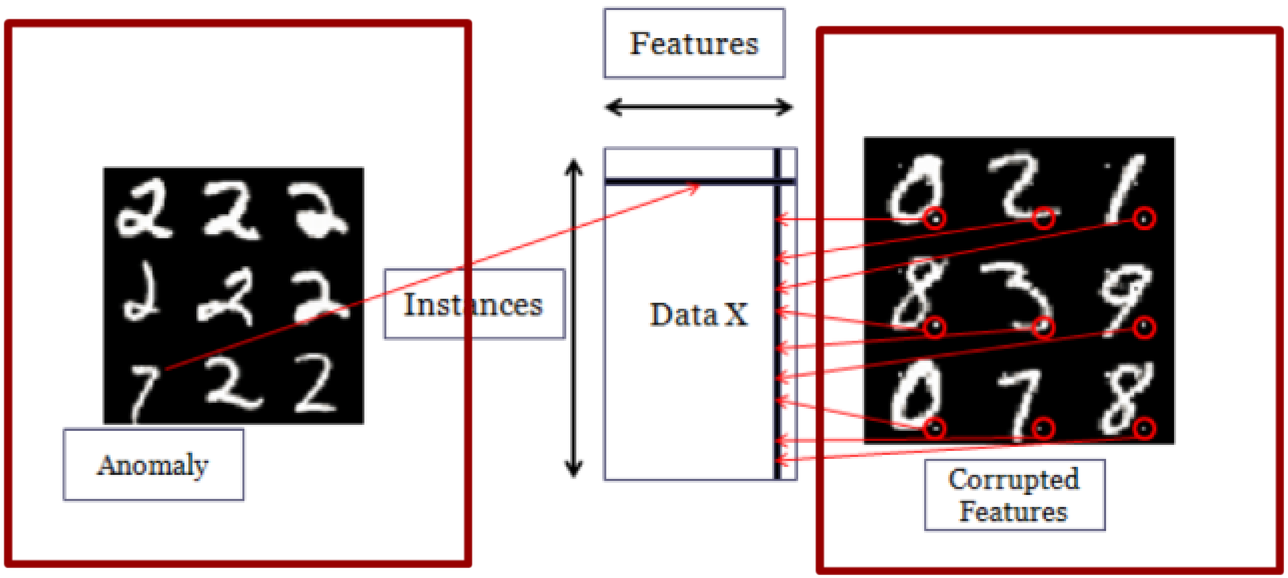
Group Anomalies



3) Methodology

II) Robust Deep Autoencoder with L2,1 Regularization

Group Anomalies



a) Particular instance is corrupted

b) Particular feature is corrupted

3) Methodology

II) Robust Deep Autoencoder with L2,1 Regularization

L2 norm of each group

$$||X||_{2,1} = \sum_{j=1}^{n} ||x_j||_2 = \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} |x_{ij}|^2\Big)^{1/2}$$

L1 norm between groups

3) Methodology

II) Robust Deep Autoencoder with L2,1 Regularization

$$\min_{\theta, S} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda \|S\|_{2,1}$$
$$\text{s.t. } X - L_D - S = 0,$$

a) Column-wise Anomaly Detection (Feature)

$$\min_{\theta, S} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda \|S^T\|_{2,1}$$
$$\text{s.t. } X - L_D - S = 0.$$

b) Row-wise Anomaly Detection (Data Instance)

5) Algorithm Training

Alternating Optimization for L1 and L2,1 RDA

- In training process, the cost function is iteratively minimized.

List of training algorithms

a) Alternating Direction Method of Multipliers(ADMM)

b) Dykstra’s alternating projection method

c) Back-propagation

d) Proximal gradient methods

5) Algorithm Training

a) <u>**Alternating Direction Method of Multipliers(ADMM)**</u>

- A training algorithm that solves optimization problem by breaking it into smaller pieces

b) <u>**Dykstra’s alternating projection method**</u>

- An alternating projection method that find a point in the intersection of convex sets

c) <u>**Back-propagation**</u>

- A training algorithm for deep autoencoder

d) <u>**Proximal gradient methods**</u>

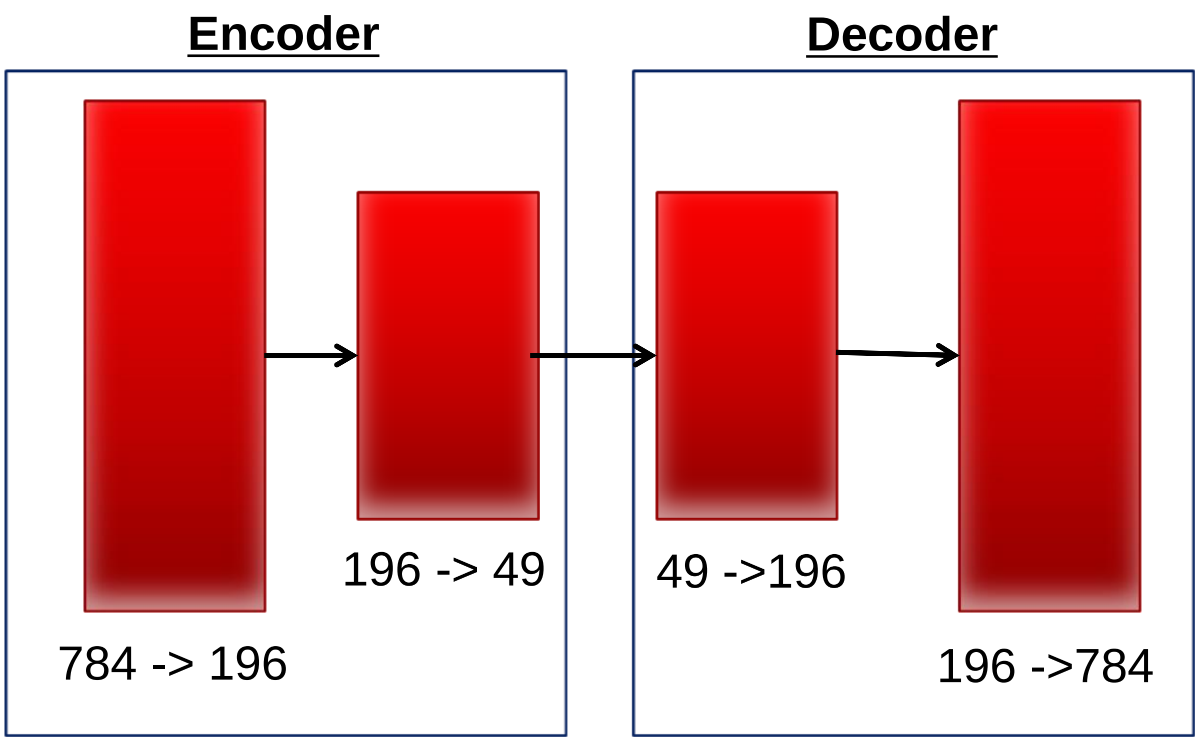
- A training algorithm for L1 and L2,1 norm of S

6) Evaluation

I) Normal Autoencoder vs L1-RDA

L1-RDA and Normal Autoencoder

- The same neural architecture (Two hidden layers)
- Both autoencoders are trained on the noise data



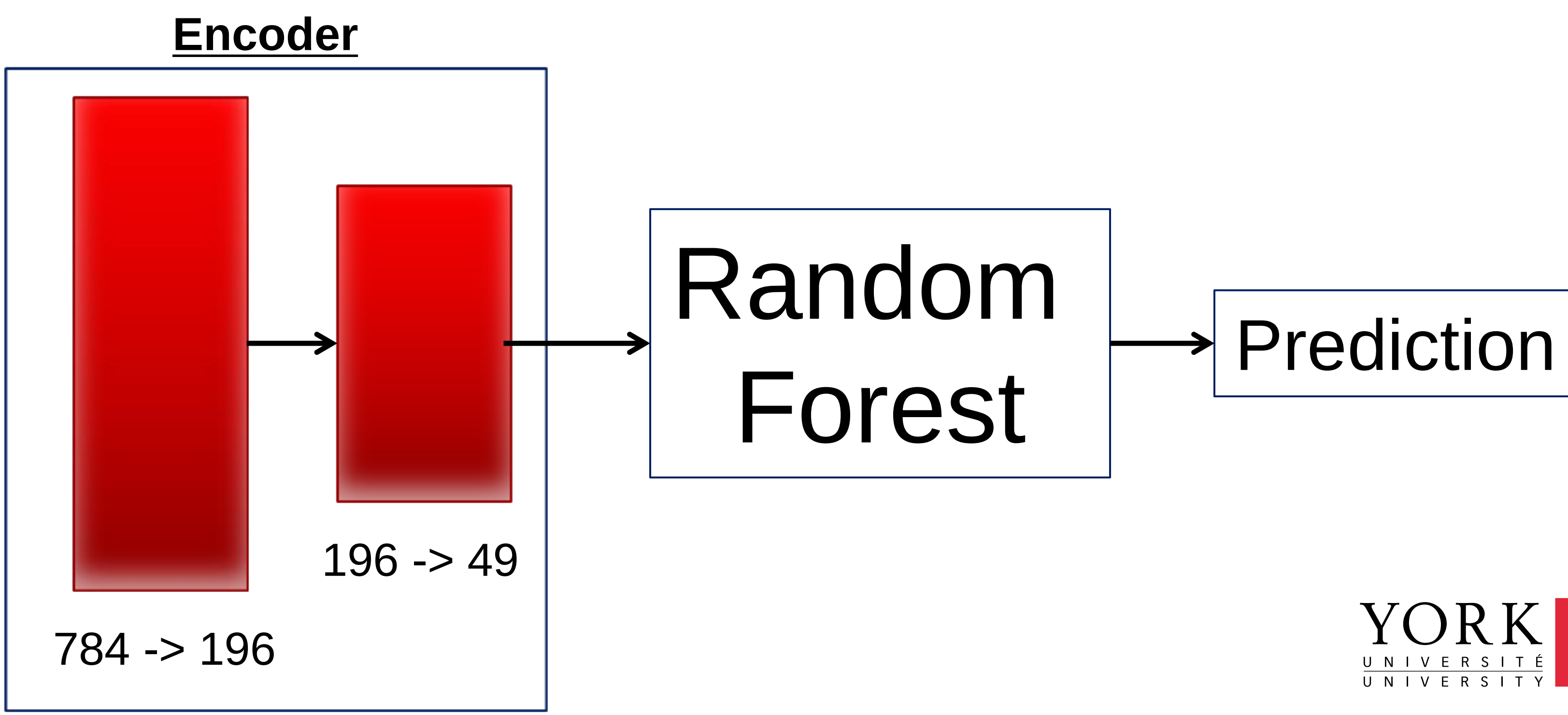
6) Evaluation

Evaluation of feature quality

6) Evaluation

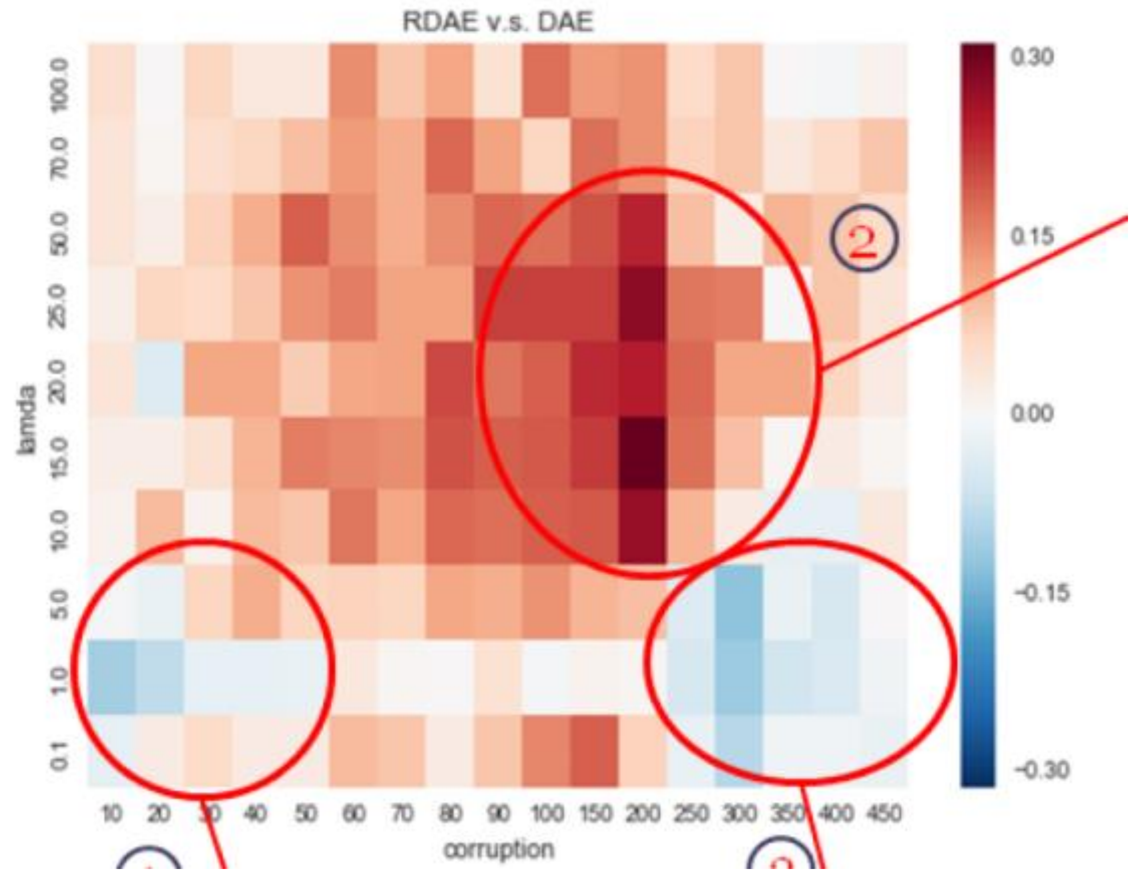
Evaluation of feature quality

- The higher test error, the lower feature quality.
- Normal autoencoder has up to 30 % higher error than RDA.
- Overall, RDA shows better performance in feature quality!

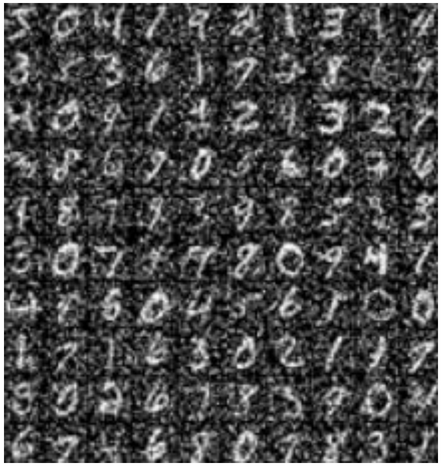


YORK UNIVERSITÉ UNIVERSITY

6) Evaluation



6) Evaluation



Corrupted Images

RDA

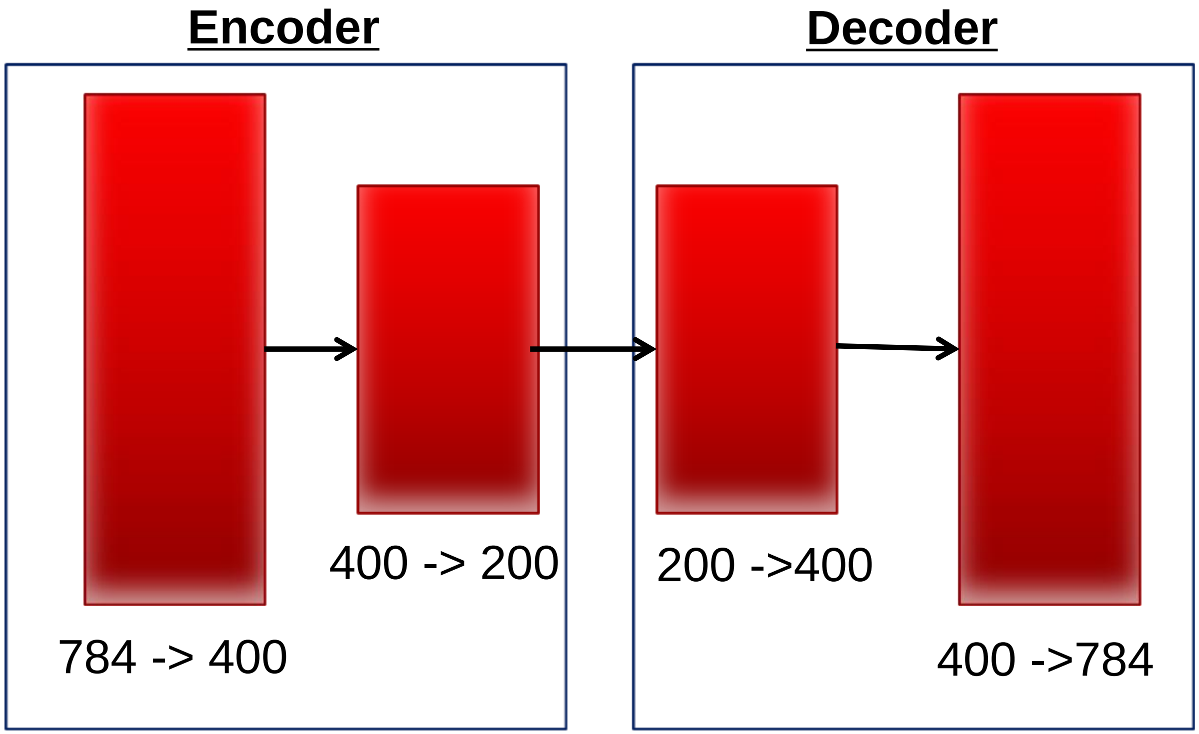
Normal Autoencoder

6) Evaluation

II) <u>L2,1-RDA vs Isolation Forest</u>

L2,1-RDA

- Two hidden layers, but different layer size



6) Evaluation

Isolation Forest

- The model discover outliers using isolation technique.

- The model had showed the state-of-the-art performance in outlier detection before RDA was introduced.

More information

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

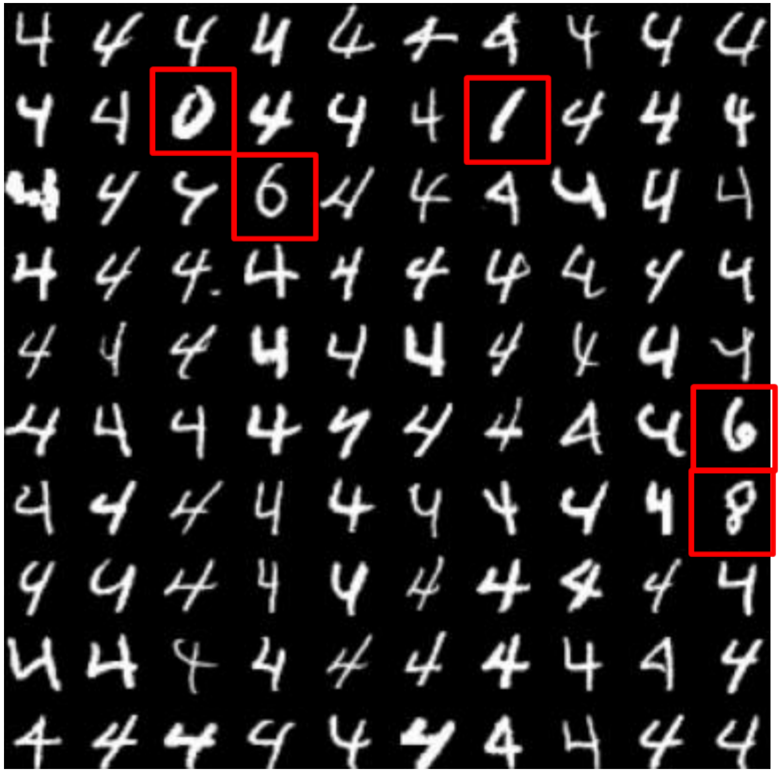
https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/tkdd11.pdf

6) Evaluation

100 examples

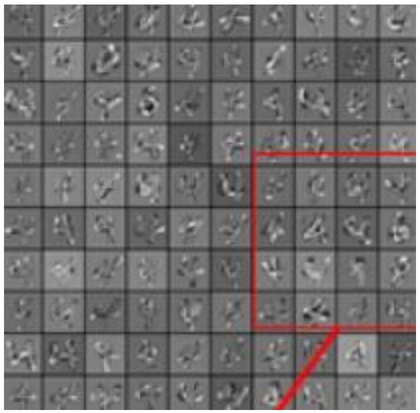
6) Evaluation



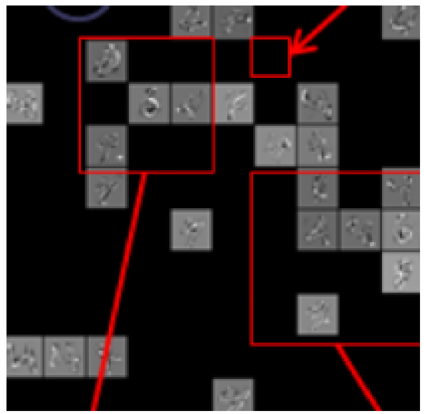
Anomalies

6) Evaluation

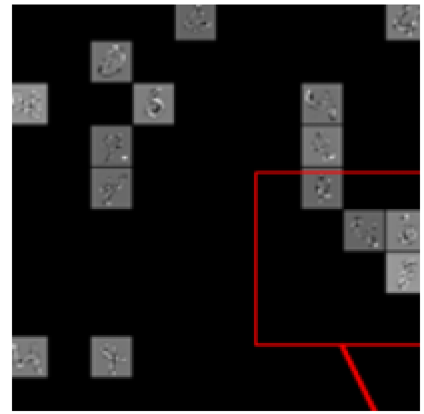
Lamda = 0.00005



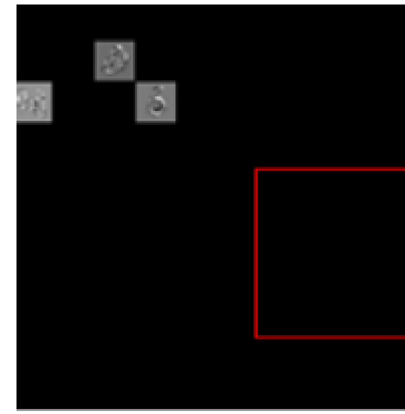
Lamda = 0.0005



Lamda = 0.00055



Lamda = 0.00065



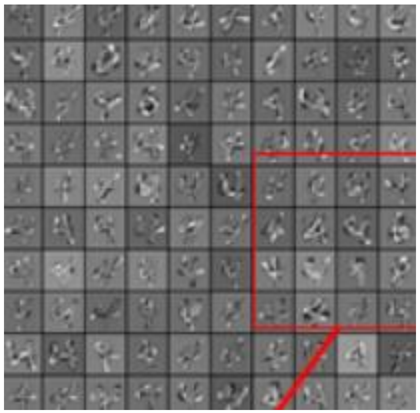
Trade Off

More False-Positives
Less False-Negatives

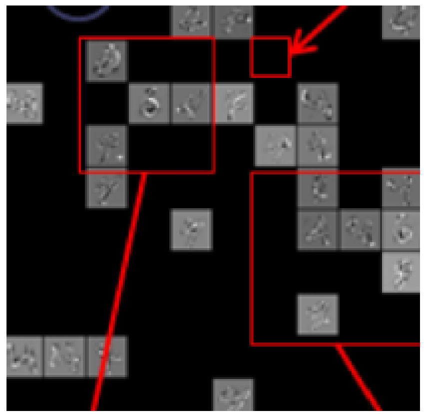
Less False-Positives
More False-Negatives

6) Evaluation

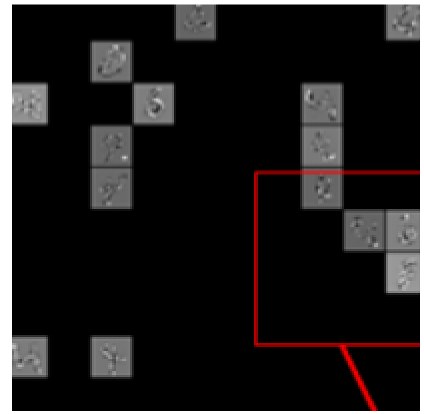
Lamda = 0.00005



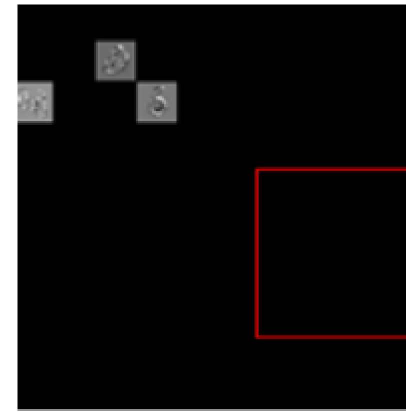
Lamda = 0.0005



Lamda = 0.00055



Lamda = 0.00065



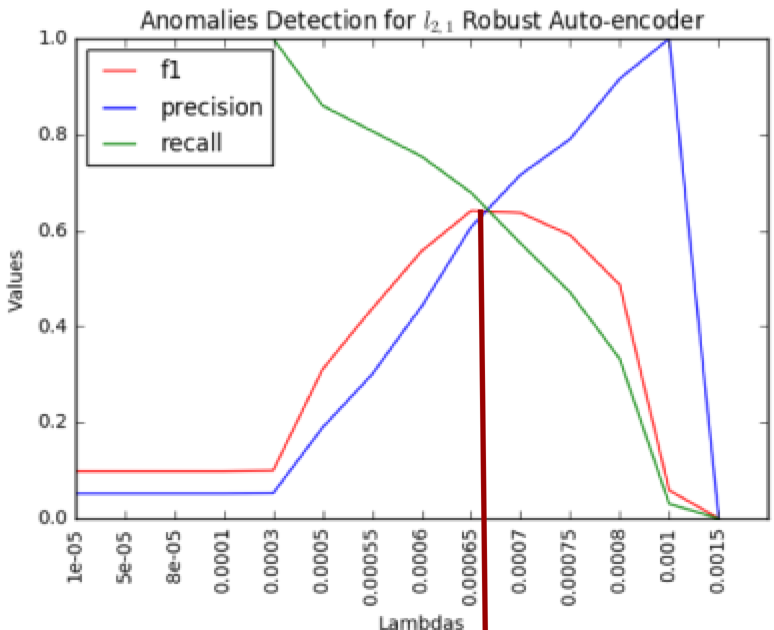
Trade Off

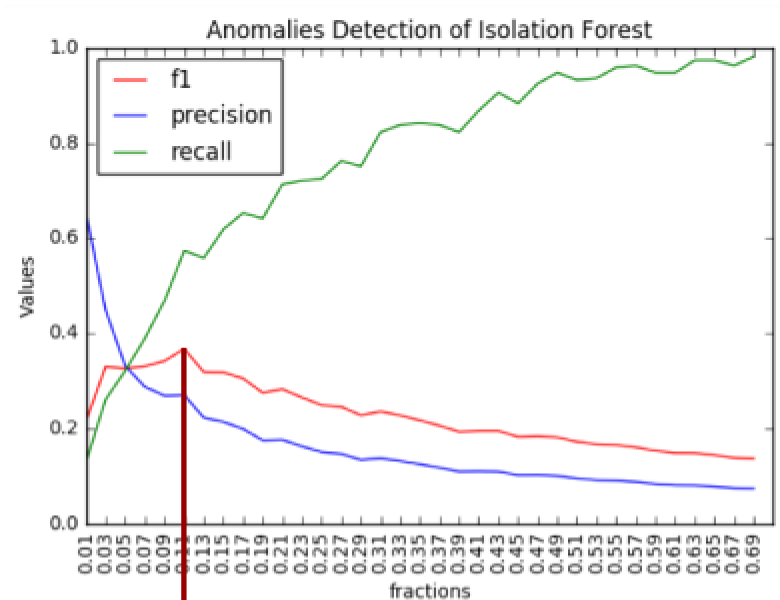
More False-Positives
Less False-Negatives

Less False-Positives
More False-Negatives

F1 Score to find the optimal lambda!

6) Evaluation





Optimal Lambda = 0.00065

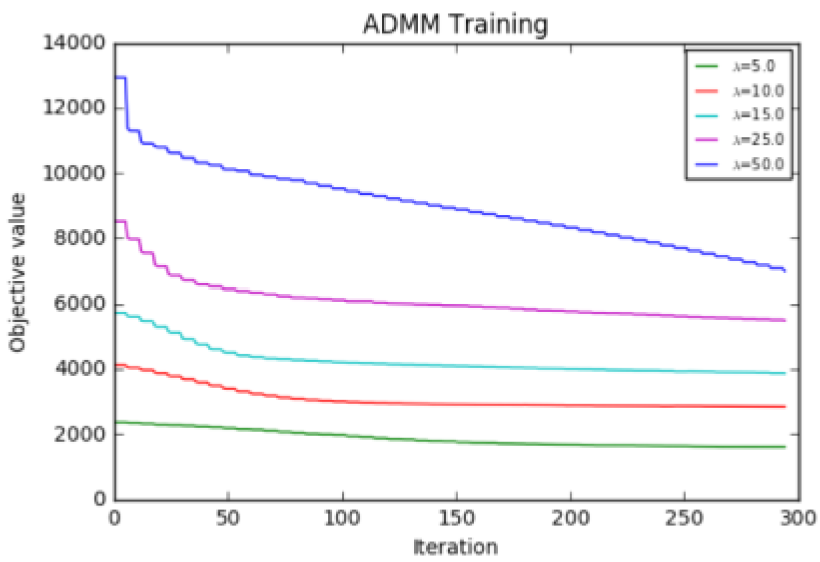
0.64 > **0.37**

RDA | Isolation Forest

6) Evaluation

Evaluation of Training Algorithm

- In most cases, the convergence of ADMM algorithm is fast.
- However, ADMM algorithm with large lambda value converges slowly.



7) Summary

i) Robust Deep Autoencoder is a combined model of Robust PCA and Deep Autoencoder. Therefore, RDA inherits advantages of two models.

ii) Robust Deep Autoencoder shows the state of art performance in anomaly detection without any clean data.

iii) Limitations

a) The convergence rate of ADMM algorithm with large lambda value is slow

b) The performance in anomaly detection largely depends on lambda value.

References

I) <u>Paper</u>

- https://www.eecs.yorku.ca/course_archive/2018-19/F/6412/reading/kdd17p665.pdf

II) <u>KDD 2017 Presentation 01</u>

- https://www.youtube.com/watch?v=npVO4RH4428

III) <u>KDD 2017 Presentation 02</u>

- https://www.youtube.com/watch?v=eFQVvFMHlC8

IV) Wikipedia – Dykstra's alternating projection method

- https://en.wikipedia.org/wiki/Dykstra%27s_projection_algorithm

Q & A