THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL
Signal and Image Processing Lab

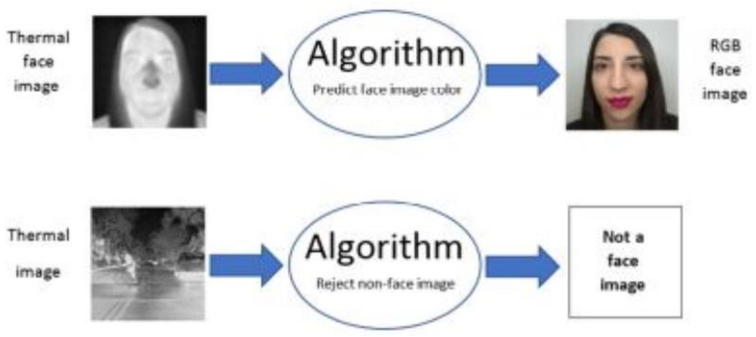
TECHNION Israel Institute of Technology

Unexpected input detection in deep neural networks

Shahar Yadin and Noam Dan, Supervised by Ev Zisselman and Ori Bryt

Introduction

- Deep neural networks (DNNs) are powerful models that achieve high performance on various tasks in machine learning, computer vision, speech and audio recognition, and language processing
- DNNs tend to behave unexpectedly when encountering input taken from an unfamiliar distribution



An example for OOD input

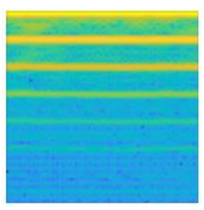
Goals

- OOD detection in regression deep neural networks
- Proposing an algorithm which solves the problem
- Examine its performances

Challenges

- No prior knowledge on OOD
- Need large data sets to calculate the distribution
- OOD and In-Distribution can be similar to each other

Motivation

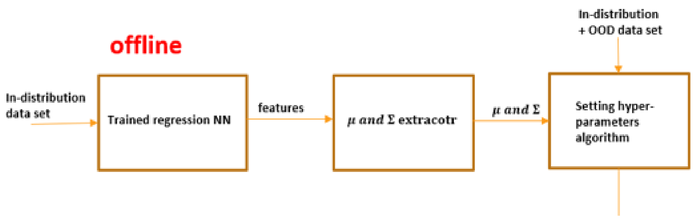


PPG signal-in distribution

X-ray hand image- OOD

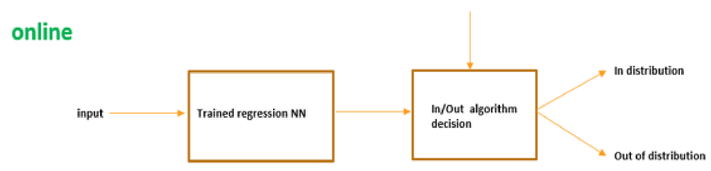
- The network predicts blood pressure according to the spectrogram image
- Accidently the X-ray image pass through the network instead of spectrogram image
- If the network doesn't detect it, it will cause a dangerous situation (a doctor can make a wrong decision)

Offline part



- Assumption: the data between the layers can be molded as a random gaussian vector.
- Calculate μ and Σ of the training data
- Calculate the weight of each layer according to its area under ROC
- Calculate the decision threshold for each layer

Online part



- Calculate mahalanobis distance for each layer

$$d_{i,j} = (X_{i,j} - \mu_i)^T \Sigma_i^{-1} (X_{i,j} - \mu_i)$$

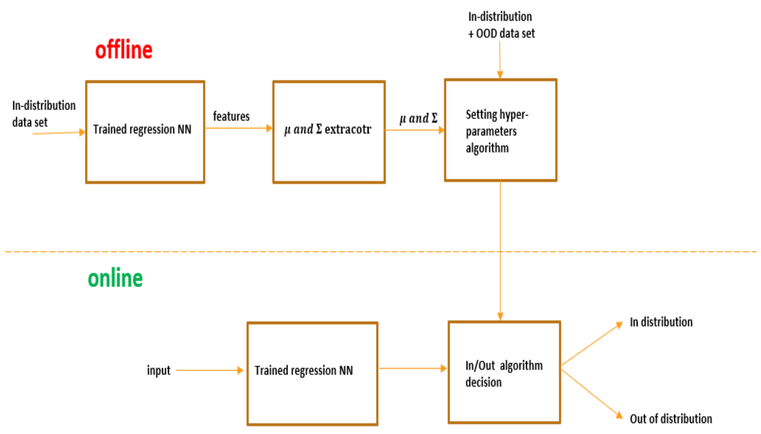
- Calculate the distance from the threshold for each layer

$$w_{i,j} = Threshold_i - d_{i,j}$$

- Decide if the input is In-Distribution or OOD according to the distances and the layers' weights.

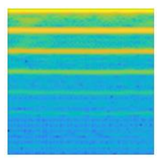
$$sign\left(\sum_i \frac{1}{1 - A_i} w_{i,j}\right)$$

Entire Block Scheme



Main Results

- In distribution: Spectrogram image of PPG signal

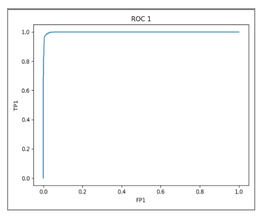


- OOD: CIFAR10 for validation,
  CIFAR10, CIFAR100, SVHN for test

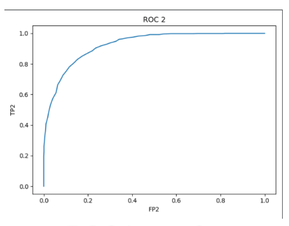
CIFAR10 example

CIFAR100 example

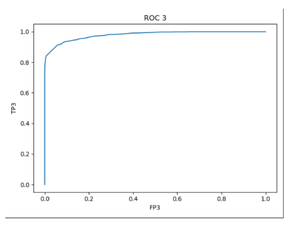
SVHN example



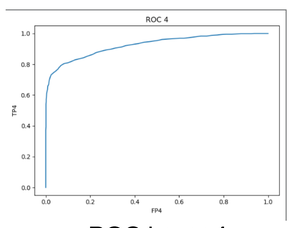
ROC Layer 1



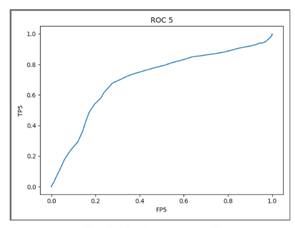
ROC Layer 2



ROC Layer 3



ROC Layer 4



ROC Layer 5

- CIFAR10: TPR=98.8%, FPR=1.4%
- CIFAR100: TPR=98.8%, FPR=1%
- SVHN: TPR=98.8%, FPR=0%

Conclusions

- OOD samples for validation achieves good generalization results.
- The network can learn on OOD from one domain and generalizes to other unseen domains.
- The algorithm doesn't work well when the OOD is too close to the in distribution.
- When the data set is small, the algorithm achieve limited results.