Explanations based on the Missing

Towards Contrastive Explanations
with Pertinent Negatives

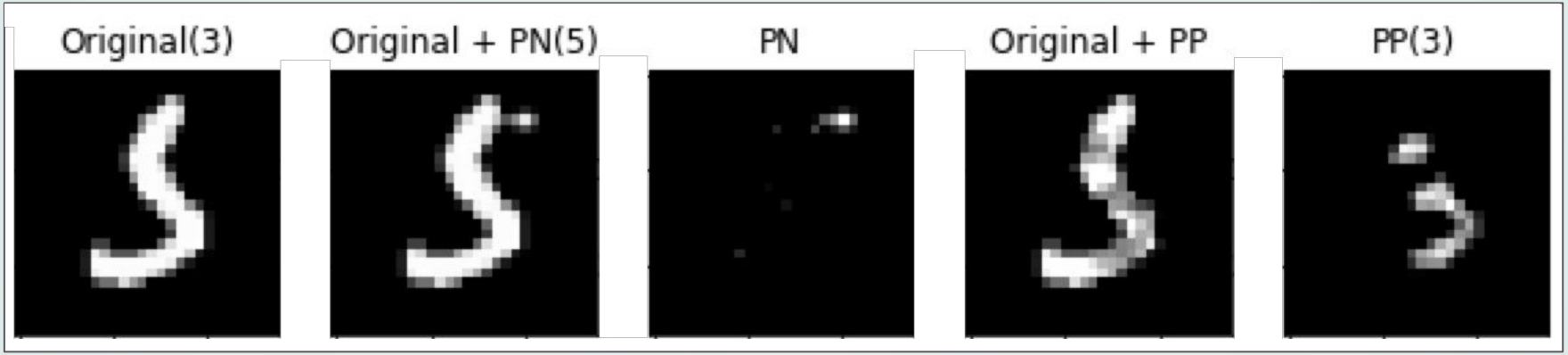
Reproducibility Paper Presentation

Introduction

- Contrastive Explanation Method (CEM)
- Intuitive explanation for ‘black box models’
- Datasets:
  - MNIST,
  - Procurement Fraud (not available)
  - fMRI (evaluated by experts)

For an intuitive explanation CEM tries to find:

Pertinent Negatives (PN) and Pertinent Positives (PP)



MNIST dataset

Pertinent Negatives are the novel idea introduced by CEM

Target questions

- Given the instructions in the paper, can we write an implementation of CEM?
  - Combine information provided by paper and publicly available code (in TensorFlow)
- How do the results of our CEM implementation compare to the original paper's results on the MNIST dataset?
  - Evaluation is done by subjective assessment of the explanations
- Does CEM generalise well to other data sets?
  - Perform CEM on FashionMNIST and assess intuitive interpretability

Method for finding pertinent negatives

- Minimize optimization objective with several terms
  - Loss function
    - Encourage probability of different class than the original one
  - Elastic net regularizer
    - For efficient feature selection in large data space
    - L1 & L2 norm
  - Autoencoder
    - Should encourage the result to be close to original data manifold (up for debate)
  - Parameters in front of every term that have to be specified beforehand

Method for finding pertinent positives

- Similar to finding the pertinent negatives
  - Loss function has different objective
    - Encourage probability of found pertinent positive to have the same class as the original sample
  - Elastic net regularizer is the same
    - For efficient feature selection in large data space
  - Autoencoder is used slightly different, with the same effect
    - Should encourage the result to be close to original data manifold (up for debate)

How do we optimize the specified objectives?

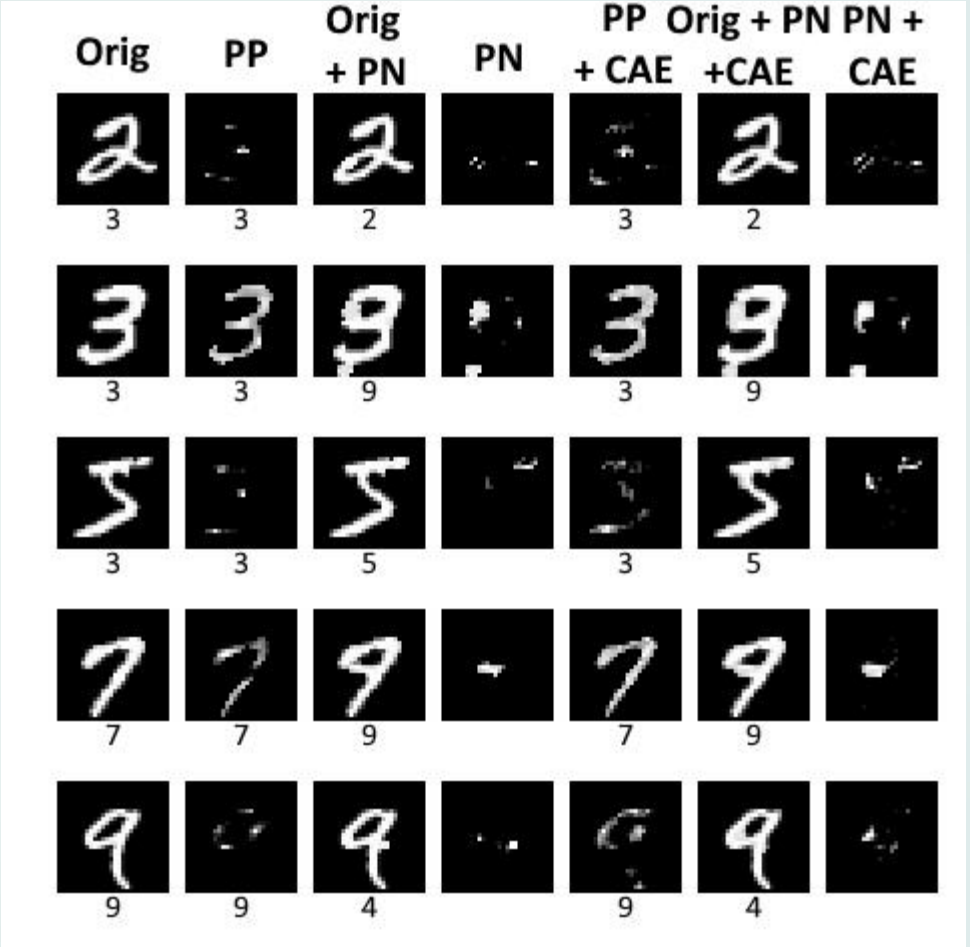
- Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)
  - Optimization algorithm for problems with L1-regularization
  - Usage of FISTA is not justified and highly debatable
  - Iteratively update value of pertinent positives / negatives
  - Uses gradient of objective function
    - Performs SGD on classifier and autoencoder
    - Classifier and autoencoder models should be differentiable (truly black box?)
  - Projection on space where pertinent positives and pertinent negatives can be found

Experimental Setup

- Performed experiments on MNIST and FashionMNIST
- Trained CNN and Autoencoder ourselves
- Hyperparameters of original paper could mostly be used as starting point
    - No exact hyperparameters for CNN and AE training
    - Played around with parameters to achieve the best results
- PyTorch package written using paper and existing implementation

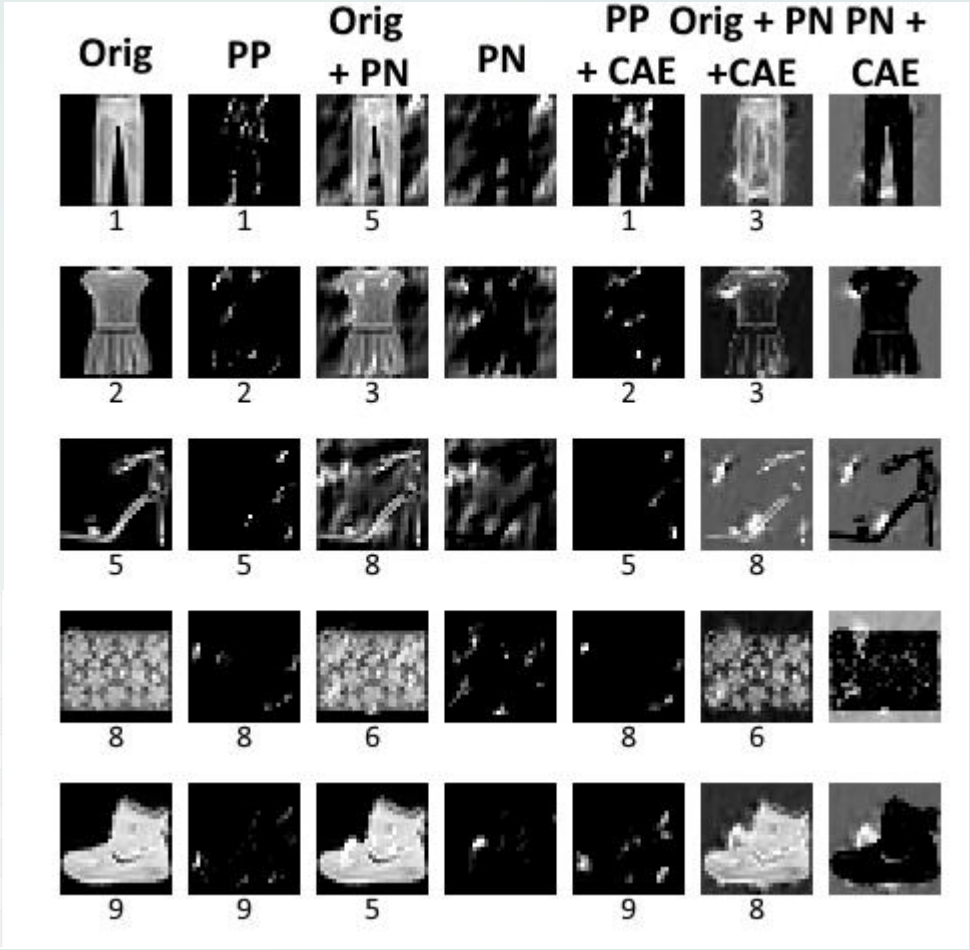
MNIST Results

- Intuitive cases
  - 7 is transformed to a 9
- Edge cases
  - 2 is as 3
  - 5 as 3



FashionMNIST Results

- High tendency to classify as a bag
- Less intuitive results than MNIST



0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Discussion

- Reimplementing CEM
  - Existing implementation was different from the paper
  - No specification of the gradient used in objective function
  - Regularisation coefficients are dependent on classification
- Evaluating the results
  - Subjective individual evaluation
  - Results are reproducible

Conclusion

- Paper does not describe process well enough
- Experiments were not all reproducible
- Paper lacks theoretical substantiation