University of Amsterdam

[RE] Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives

CEM #1

Mattijs Blankesteijn, Stefan Klut, Thomas van Osch & Tim Ottens

The original paper

- Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives by Dhurandhar et. al (2018)
- Proposed method
- Problem definition and space (transparency)

Targets

- Are the definitions and experimental setup given in the paper sufficiently explained for the method to be implemented?
- Are the reported results for the MNIST dataset replicable?
- Does the CEM also generalize to Fashion-MNIST?

Contrastive Explanations Method

Intuition.

- Finding pertinent positives (PP) and pertinent negatives (PN)
- PP: minimal amount of features in the input that are sufficient in themselves to yield the same classification
- PN: minimal amount of features that should be absent in the input to classify it as any other class

---

Algorithm 1 Contrastive Explanations Method (CEM)

Input: example (x_0, t_0), neural network model $\mathcal{N}$ and (optionally ($\gamma > 0$)) an autoencoder AE

1) Solve (1) and obtain,

$\boldsymbol{\delta}^{\text{neg}} \leftarrow \operatorname{argmin}_{\boldsymbol{\delta} \in \mathcal{X}/\mathbf{x}_0} \; c \cdot f_\kappa^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta \|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma \|\mathbf{x}_0 + \boldsymbol{\delta} - \text{AE}(\mathbf{x}_0 + \boldsymbol{\delta})\|_2^2.$

2) Solve (3) and obtain,

$\boldsymbol{\delta}^{\text{pos}} \leftarrow \operatorname{argmin}_{\boldsymbol{\delta} \in \mathcal{X} \cap \mathbf{x}_0} \; c \cdot f_\kappa^{\text{pos}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta \|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma \|\boldsymbol{\delta} - \text{AE}(\boldsymbol{\delta})\|_2^2.$

return $\boldsymbol{\delta}^{\text{pos}}$ and $\boldsymbol{\delta}^{\text{neg}}$. {Our Explanation: Input x_0 is classified as class t_0 because features $\boldsymbol{\delta}^{\text{pos}}$ are present and because features $\boldsymbol{\delta}^{\text{neg}}$ are absent. Code at `https://github.com/IBM/Contrastive-Explanation-Method` }

Figure: CEM pseudocode.

Contrastive Explanations Method

The Explanation Optimization.

Given sample $(\boldsymbol{x}_0, t_0)$, classifier $\mathscr{P}(\cdot)$, $\boldsymbol{I} = \boldsymbol{\delta}$ if PP and $\boldsymbol{I} = \boldsymbol{x}_0 + \boldsymbol{\delta}$ if PN, optimize:

CEM subject to $f(t_0, \boldsymbol{I}, \kappa) = 0$

$$\min_{\delta} c \cdot f(t_0, \boldsymbol{I}, \kappa) + \beta ||\delta||_1 + ||\delta||_2^2 + \gamma ||AE(\boldsymbol{I}) - \boldsymbol{I}||_2^2 \tag{1}$$

$$f^{PP}(t_0, \boldsymbol{I}, \kappa) = \max\left(\kappa + \max_{i \neq t_0} [\mathscr{P}(\boldsymbol{I})]_i - [\mathscr{P}(\boldsymbol{I})]_{t_0}, 0\right) \tag{2}$$

$$f^{PN}(t_0, \boldsymbol{I}, \kappa) = \max\left(\kappa + [\mathscr{P}(\boldsymbol{I})]_{t_0} - \max_{i \neq t_0} [\mathscr{P}(\boldsymbol{I})]_i, 0\right) \tag{3}$$

Optimization

- FISTA is used to optimize the loss of an ℓ_1 regularized function
- The perturbation δ is updated every iteration of the algorithm by a slack variable
- The slack variable is also updated with an SGD optimizer and FISTA
- The best δ^* is chosen based on $f(t_0, \boldsymbol{I}, \kappa) = 0$ and the lowest elastic net loss

Experimental Setup

- MNIST and Fashion-MNIST dataset
- Tensorflow $\longrightarrow$ PyTorch
- Convolution Neural Network Classifier (99.4% / 90.8%)
- Convolutional Autoencoder (restrict search space)

Results MNIST

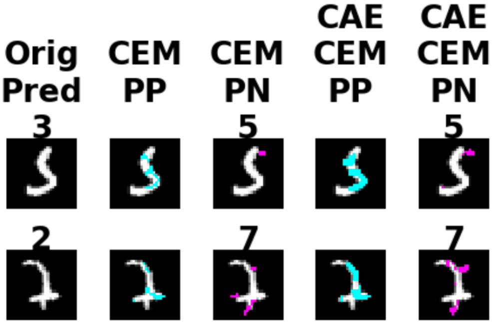

Figure: Results of Dhurandhar et al.

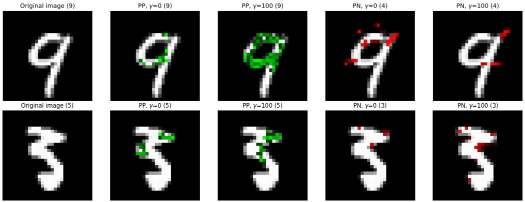

Figure: Our results.
Top row Classified: 9. PP: 9. PN 4.
Bottom row Classified: 5. PP: 5. PN: 3.

Results Fashion-MNIST

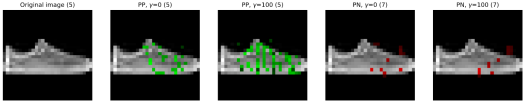

Figure: Classified: Sandal. PP: Sandal. PN: Shoe

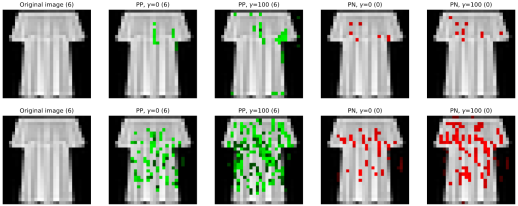

Figure: Classified: Shirt. PP: Shirt. PN: T-shirt/Top, Top: $\kappa = 10$, Bottom: $\kappa = 100$

Discussion

- MNIST comparison dependent on unreported thresholding
- Fashion-MNIST less interpretable, exploiting bias
- No guarantees for latent dimension of autoencoder
- FISTA prerequisites seem to be violated by applying CEM
- CEM needs to backpropagate through the ‘black box’

Conclusion

Targets

- The method is implementable (with original code)
- The results on MNIST are replicable, despite some inconsistencies between original paper and code and unclear parameter configurations
- Extension to Fashion-MNIST turns out to be difficult

Broader Implications:

- Discover algorithmic biases
- CEM can guide human interventions