# [Review] Learning perturbations to Explain Time Series Predictions

ICML 23 Poster accepted

[Author] Joseph Enguehard

[Reviewer] Chanmoo Park
December 19, 2023

# Perturbation-based saleincy methods

- Fong and Vedaldi (2017)


- When the black box model $f : \mathcal{X} \to \mathcal{Y}$ is given, with the mask $m \in [0,1]^{\text{dim}(\mathcal{X})}$ and a perturbation $\Phi(\mathbf{x}, \mathbf{m}) : \mathcal{X} \times [0,1]^{\text{dim}(\mathcal{X})} \to \mathcal{X}$ can be trained in two ways...
  - Deletion game (minimum mask with maximum loss)

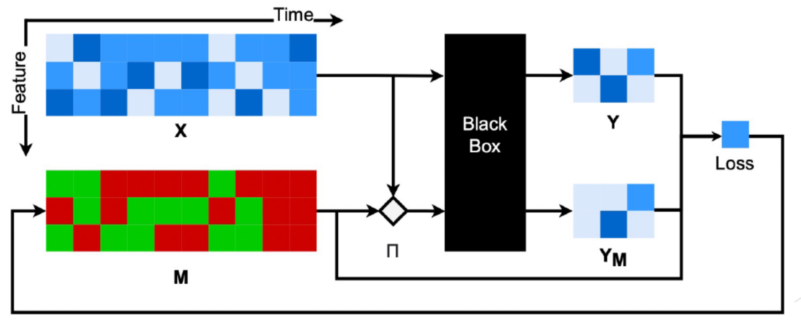
$$\underset{\mathbf{m} \in [0,1]^n}{\arg\min} \lambda \|\mathbf{1} - \mathbf{m}\|_1 - \mathcal{L}(f(\mathbf{x}), f(\Phi(\mathbf{x}, \mathbf{m})))$$

  - Preservation game (maximum mask with minimum loss)

$$\underset{\mathbf{m} \in [0,1]^n}{\arg\min} \lambda \|\mathbf{m}\|_1 + \mathcal{L}(f(\mathbf{x}), f(\Phi(\mathbf{x}, \mathbf{m})))$$

# Perturbation methods applied to the time series data

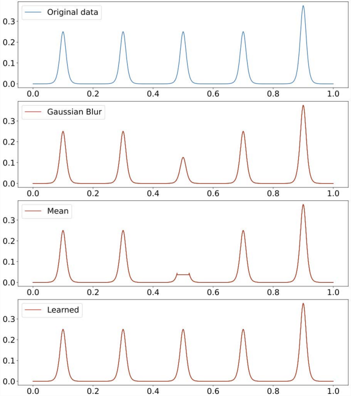
- Crabbé and Van Der Schaar (2021)




- Considered multivariate ($n$-feature) time series ($t \in [1 : T]$) as a $n \times T$ feature matrix and applied Perturbation-based methods
- The explanation on the multivariate time series data:
  - What feature is critical?
  - When it becomes to critical?

# Perturbation methods applied to the time series data

- Crabbé and Van Der Schaar (2021) used ”Learnable” mask, but ”Fixed” Perturbation




- Author pointed out that this fixed perturbation could lead an wrong explanation, since it couldn’t consider the long-term dependencies of time-series.

# "Learning Perturbation" to Explain Time Series Predictions

- Fixed Perturbation

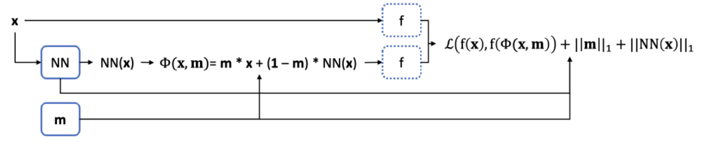
$$\Phi(\mathbf{x},\mathbf{m}) = \begin{cases} \mathbf{m}\times\mathbf{x} + (\mathbf{1}-\mathbf{m})\times\mu_0 & \text{(local mean)} \\ \mathbf{m}\times\mathbf{x} + (\mathbf{1}-\mathbf{m})\times\nu & \text{(Gaussian noise)} \\ \int g_{\sigma_0\times(1-\mathbf{m})}(\mathbf{y}-\mathbf{x})d\mathbf{y} & \text{(Gaussian blur)} \end{cases}$$

- **Learnable Perturbation** (Proposed method)

$$\Phi(\mathbf{x},\mathbf{m}) = \mathbf{m}\times\mathbf{x} + (\mathbf{1}-\mathbf{m})\times \mathrm{NN}(\mathbf{x})$$

  - Neural Net $\mathrm{NN}(x)$ : used GRU (Gated Recurrent Unit)
  - Learned through the preservation game:

$$\underset{\mathbf{m},\Theta\in\mathrm{NN}}{\arg\min}\lambda_1\|\mathbf{m}\|_1 + \lambda_2\|\mathrm{NN}(\mathbf{x})\|_1 + \mathcal{L}(f(\mathbf{x}), f(\Phi(\mathbf{x},\mathbf{m})))$$



# Experiment : MIMIC-III dataset (Johnson et al., 2016)

- MIMIC-III dataset
  - Electronic Health Records of more than 60,000 critical care patients.
  - 96 different longitudinal real-valued measurements over a period of 48 hours after patient admission.
  - Task : to predict in-hospital mortality based on 48 hours data at each hour.

| **Method** | Acc ↓ | Comp ↑ | CE ↑ | Suff ↓ |
|---|---|---|---|---|
| DeepLift | 0.988 (0.002) | -4.36E-4 (0.001) | 0.097 (0.006) | 2.86E-3 (0.001) |
| DynaMask | 0.990 (0.001) | 2.21E-4 (0.001) | 0.097 (0.005) | 2.99E-3 (0.001) |
| IG | 0.988 (0.003) | 2.24E-4 (0.002) | 0.098 (0.006) | 2.21E-3 (0.001) |
| GradientShap | 0.987 (0.004) | -2.19E-3 (0.001) | 0.095 (0.006) | 3.99E-3 (0.001) |
| Lime | 0.996 (0.001) | -7.36E-4 (0.001) | 0.094 (0.005) | 3.39E-3 (0.001) |
| Occlusion | 0.988 (0.001) | -1.93E-3 (0.001) | 0.095 (0.005) | 4.57E-3 (0.001) |
| Aug Occlusion | 0.989 (0.001) | 4.59E-4 (0.001) | 0.098 (0.005) | 1.90E-3 (0.002) |
| Retain | 0.989 (0.001) | -3.79E-3 (0.001) | 0.093 (0.005) | 7.70E-3 (0.001) |
| Ours | **0.981** (0.004) | **1.53E-2** (0.004) | **0.118** (0.008) | **-1.19E-2** (0.004) |

* Acc and CE : 'Accuracy' and 'Cross Entropy change' when salient feature is masked

** Comp(Comprehensiveness) and Suff(Sufficiency) : Softmax prob. changes when salient feature is masked or only salient feature is used

# Experiment : MIMIC-III dataset (Johnson et al., 2016)

- Feature-wise, Time-wise importance