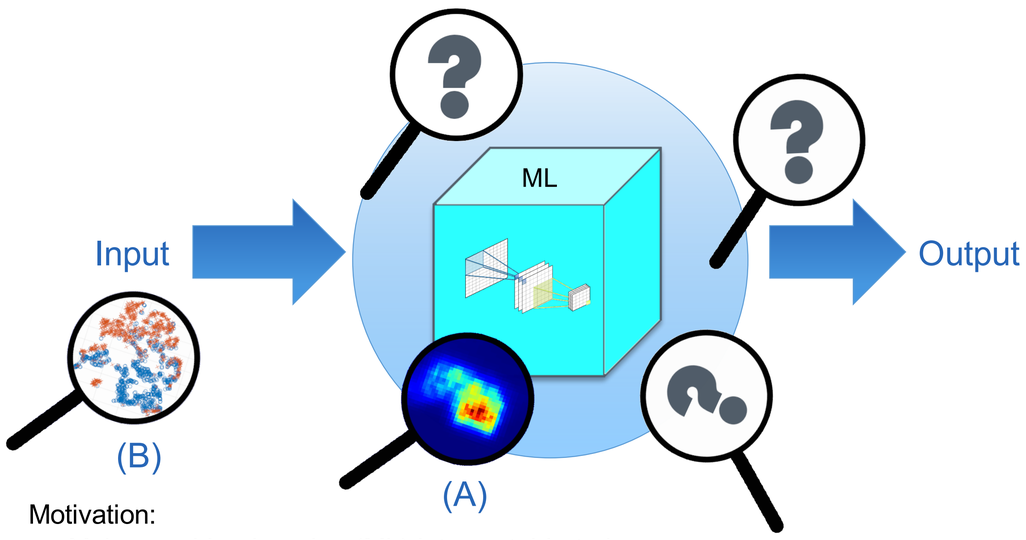
Towards Validation of Machine Learning Algorithms using Visualization Techniques

Oliver Gallitz[1,3], Oliver De Candido[2,3], Michael Botsch[1], Wolfgang Utschick[2]

1: Technische Hochschule Ingolstadt, 2. Technische Universität München, 3. AUDI AG

Introduction and Motivation



Motivation:
- Make machine learning (ML) interpretable to humans
- Crack open the black-box of ML algorithms

Approaches:
- Responsibility-Sensitive Safety[1] (RSS) or similar
- Create a safety envelope around the ML modules

Visualization techniques
- Understand the input-output relationship of the ML algorithm
- Investigate clustering of the input data

(A) Heatmapping Techniques

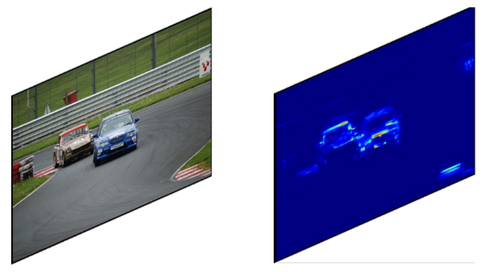
Heatmapping:
Detecting salient areas in the input image.

Representation:

Activation matrix in input dimension

$H(\boldsymbol{x}^{(i)}) \in \mathbb{R}^{H \times W}$

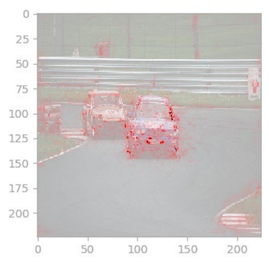
with $\boldsymbol{x}^{(i)} = \mathrm{vec}(\boldsymbol{X}^{(i)} \in \mathbb{R}^{H \times W \times C})$



Sensitivity Analysis[2]**:**

Calculate gradient of the classification score w.r.t. each input pixel

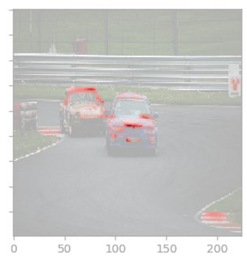
$$H(\boldsymbol{x}^{(i)}) = \frac{\partial f(\boldsymbol{x}^{(i)})}{\partial \boldsymbol{x}^{(i)}}$$



Activation-based methods:

Redistribution of prediction score backwards through all layers of the network. E.g.: Layer-Wise Relevance Propagation (LRP)[3]

$$y = \sum_k R_k^{L-1} = \cdots = \sum_i R_i^1$$



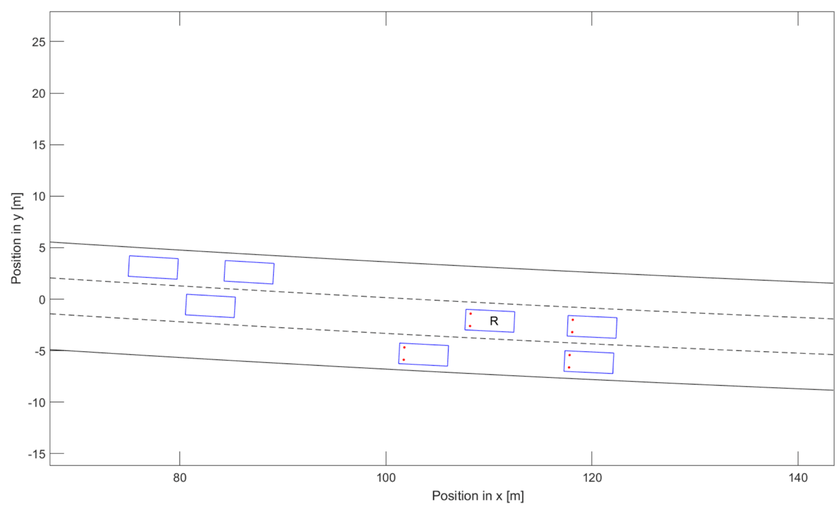
Critical Braking Scenario Simulator

Motivation:
- Create “clinically-clean” input data
- Model a highway convoy scenario with high traffic
- Capable to simulate: lane changing, lane following, braking profiles, etc.

Input data:
- Follow the relevant vehicle (R) and leader vehicle
- Take eight features from random freezeframe out of a scenario where:

$$\tau_{\text{init}} > d_V > \tau_{\text{crit}}$$

- Label random freezeframe from sequence as non-critical or critical
- Generated $M = 2000$ freezeframes with 50% non-critical/critical



(B) Dimensionality Reduction Techniques

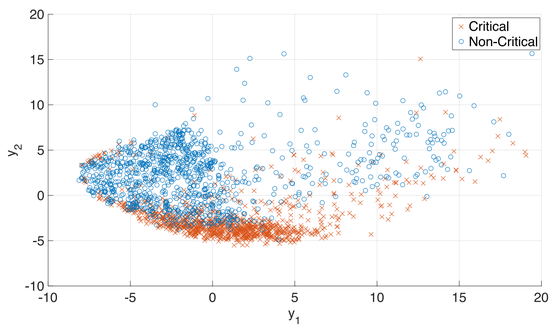
$$\phi: \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\right] = \boldsymbol{X} \in \mathbb{R}^{N \times M} \mapsto \left[\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(M)}\right] = \boldsymbol{Y} \in \mathbb{R}^{p \times M}$$

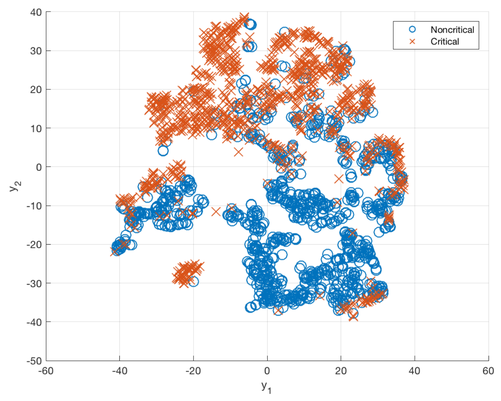
Principal Component Analysis (PCA)[4,5]
- Minimize the projection error **or** retain most of the variance

$$\min_{\mathbf{U_p}} \sum_{i=1}^{M} \left\| \boldsymbol{x}^{(i)} - \left(\boldsymbol{U}_p \boldsymbol{y}^{(i)} + \boldsymbol{\mu}_{\boldsymbol{x}}\right) \right\|_2^2$$

$$\max_{\mathbf{U_p}} \mathrm{tr}(\boldsymbol{U}_p^T \boldsymbol{C} \boldsymbol{U}_p) \quad \text{s.t.} \quad \boldsymbol{U}_p^T \boldsymbol{U}_p = \boldsymbol{I}_p$$

- Linear weighting matrix: $\boldsymbol{U}_p \in \mathbb{R}^{N \times p}$
- Sample covariance $\boldsymbol{C}$ and mean $\boldsymbol{\mu}_{\boldsymbol{x}}$





t-Distributed Stochastic Neighborhood Embedding (t-SNE)[6]
- Minimize Kullback-Leibler (KL) divergence between pairwise similarities

$$\min_{\{y^{(i)}\}} D_{\mathrm{KL}}(f_{i,j} \| g_{i,j})$$

- High-dimensional similarities:
  - Gaussian distribution
- Low-dimensional similarities:
  - Student’s t-distribution

References:

[1] S. Shalev-Shwartz, S. Shammah, and A. Shashua, “On a Formal Model of Safe and Scalable Self-driving Cars,“ Mar. 2018. [Online] Available: http://arxiv.org/pdf/1708.06374v5.

[2] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: Removing Noise by Adding Noise,” Jun. 2017. [Online] Available: http://arxiv.org/pdf/1706.03825v1.

[3] S. Bach *et al.,* “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” (eng), *PLoS ONE*, vol. 10, no. 7, 2015.

[4] K. Pearson, “LIII. On Lines and Planes of Closest Fit to Systems of Points in Space,” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11 pp. 559–572, 1901.

[5] H. Hotelling, “Analysis of a Complex of Statistical Variables into Principal Components,” Journal of Educational Psychology, vol. 24, no. 6, pp. 417–441, 1933.

[6] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

Technische Hochschule Ingolstadt
Fakultät für Elektrotechnik und Informatik
CARISSMA

Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik
Professur für Methoden der Signalverarbeitung