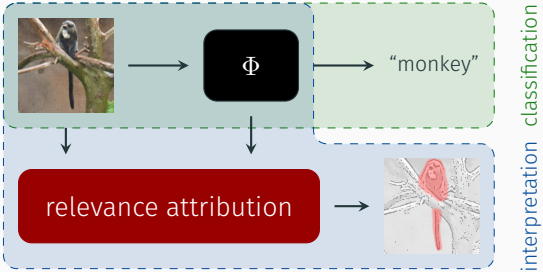
Interpretable Neural Networks with Frank-Wolfe

Sparse Relevance Maps and Relevance Orderings

Jan Macdonald — Technische Universität Berlin
Mathieu Besançon — Zuse Institute Berlin
Sebastian Pokutta — Technische Universität Berlin · Zuse Institute Berlin

39th International Conference on Machine Learning
Baltimore, July 17 – 23, 2022

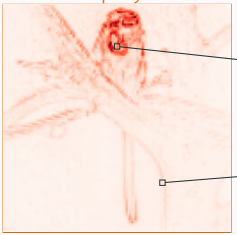
Relevance Attribution



Sensitivity [Simonyan et al. 2013], Smoothgrad [Smilkov et al. 2017], Guided Backprop [Springenberg et al. 2015], LRP-α-β [Bach et al. 2015], DeepTaylor [Montavon et al. 2018], SHAP [Lundberg and Lee 2017], LIME [Ribeiro et al. 2016], L-RDE [Macdonald et al. 2019]

What Do Relevance Scores Mean?

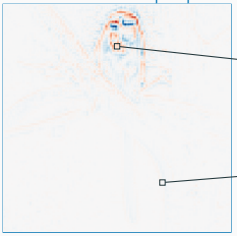
DeepTaylor



relevance score 0.75

relevance score 0.15

Guided Backprop

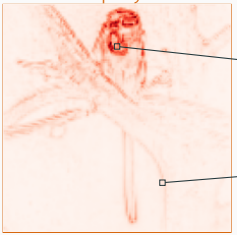


relevance score 0.35

relevance score 0.00

What Do Relevance Scores Mean?

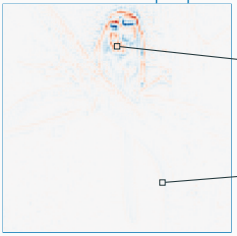
DeepTaylor



relevance score 0.75

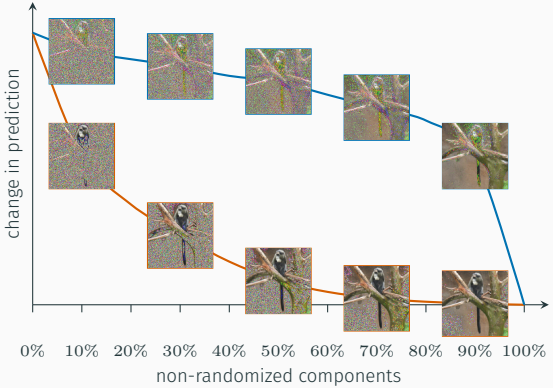
relevance score 0.15

Guided Backprop



relevance score 0.35

relevance score 0.00



Rate-Distortion Explanations (RDE):

$\|\mathbf{s}\|_1$ vs. $D(\mathbf{s}) = \mathbb{E}_{\mathbf{n}}[(\Phi(\mathbf{x}) - \Phi(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{n})^2]$

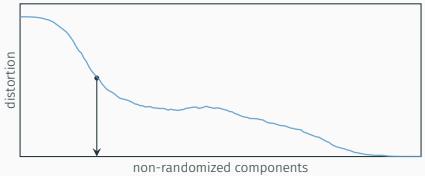
Sparse Relevance Maps and Relevance Orderings

Rate-Distortion Explanations (RDE):

$\|\mathbf{s}\|_1$ vs. $D(\mathbf{s}) = \mathbb{E}_{\mathbf{n}}[(\Phi(\mathbf{x}) - \Phi(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{n})^2]$

Rate-Constrained RDE (RC-RDE):

$\text{minimize } D(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{s}\|_1 \leq k,\ \mathbf{s} \in [0,1]^n$



Rate-Distortion Explanations (RDE):

$\|\mathbf{s}\|_1$ vs. $D(\mathbf{s}) = \mathbb{E}_{\mathbf{n}}[(\Phi(\mathbf{x}) - \Phi(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{n})^2]$

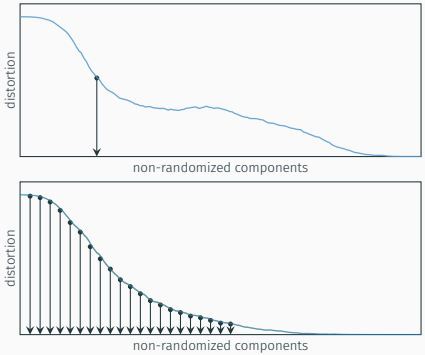
Rate-Constrained RDE (RC-RDE):

$$\text{minimize } D(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{s}\|_1 \leq k,\ \mathbf{s} \in [0,1]^n$$

Ordering RDE (Ord-RDE):

$$\text{minimize } \sum_{k=1}^{n-1} D(\mathbf{\Pi}\mathbf{p}_k) \quad \text{subject to} \quad \mathbf{\Pi} \in B_n$$

- $\mathbf{p}_k$ vector of k ones and $n - k$ zeros
- B_n Birkhoff polytope ($n \times n$ doubly stochastic matrices)



Rate-Distortion Explanations (RDE):

$\|\mathbf{s}\|_1$ vs. $D(\mathbf{s}) = \mathbb{E}_{\mathbf{n}}[(\Phi(\mathbf{x}) - \Phi(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{n})^2]$

Rate-Constrained RDE (RC-RDE):

$$\text{minimize } D(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{s}\|_1 \leq k,\ \mathbf{s} \in [0,1]^n$$

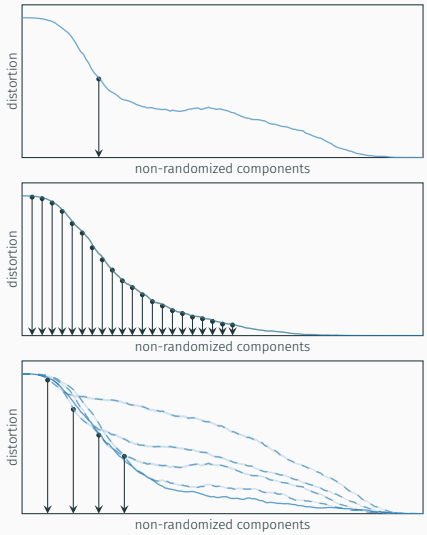
Ordering RDE (Ord-RDE):

$$\text{minimize } \sum_{k=1}^{n-1} D(\mathbf{\Pi}\mathbf{p}_k) \quad \text{subject to} \quad \mathbf{\Pi} \in B_n$$

- $\mathbf{p}_k$ vector of k ones and $n - k$ zeros
- B_n Birkhoff polytope ($n \times n$ doubly stochastic matrices)

Multi-Rate RDE (MR-RDE):

Combine multiple RC-RDE solutions at different rates k to approximate Ord-RDE.



Rate-Distortion Explanations (RDE):

$\|\mathbf{s}\|_1$ vs. $D(\mathbf{s}) = \mathbb{E}_{\mathbf{n}}[(\Phi(\mathbf{x}) - \Phi(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{n})^2]$

Rate-Constrained RDE (RC-RDE):

$$\text{minimize } D(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{s}\|_1 \leq k,\ \mathbf{s} \in [0,1]^n$$

Ordering RDE (Ord-RDE):

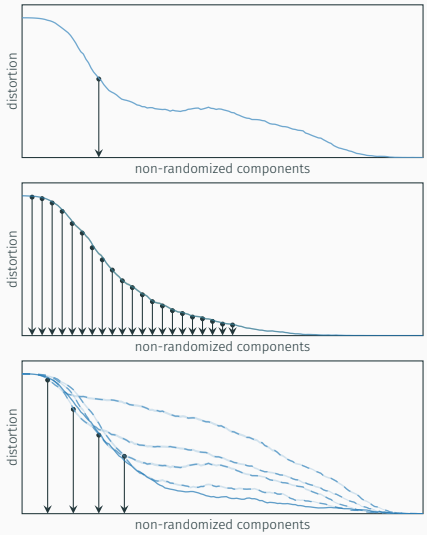
$$\text{minimize } \sum_{k=1}^{n-1} D(\mathbf{\Pi}\mathbf{p}_k) \quad \text{subject to} \quad \mathbf{\Pi} \in B_n$$

- $\mathbf{p}_k$ vector of k ones and $n - k$ zeros
- B_n Birkhoff polytope ($n \times n$ doubly stochastic matrices)

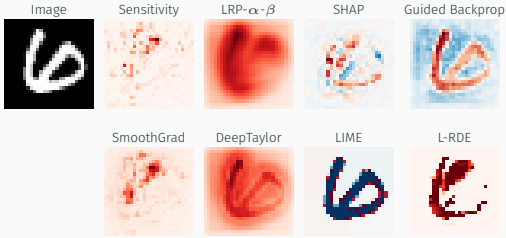
Multi-Rate RDE (MR-RDE):

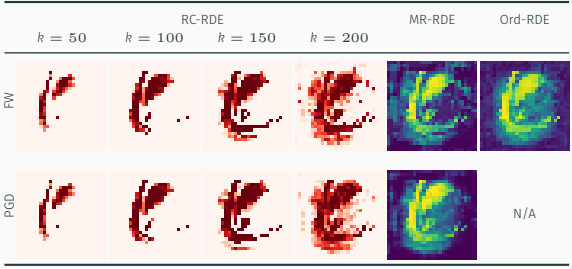
Combine multiple RC-RDE solutions at different rates k to approximate Ord-RDE.

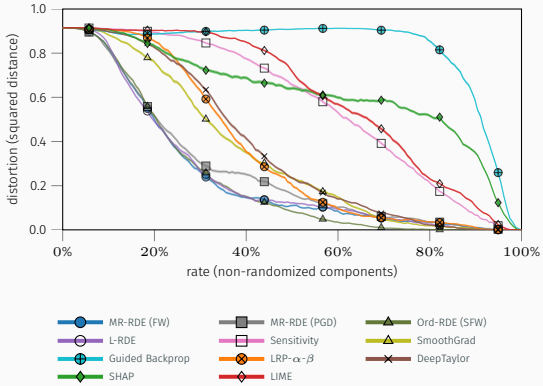
⇝ Solved with Frank-Wolfe Algorithms



Results

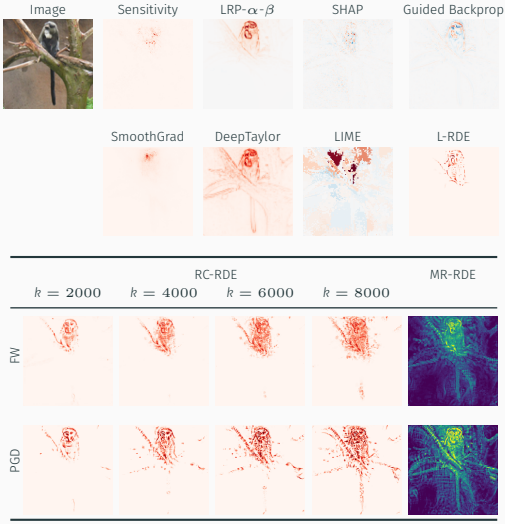


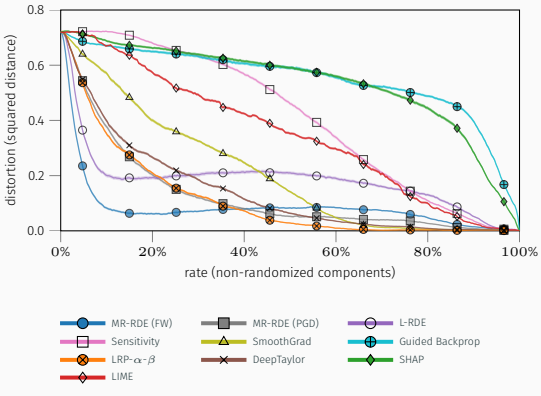




MNIST Dataset [LeCun et al. 1998]

Results





STL-10 Dataset [Coates et al. 2011]

Thank You!

macdonald@math.tu-berlin.de
@jan_maces · @matbesancon · @spokutta

arxiv.org/abs/2110.08105

github.com/ZIB-IOL/fw-rde

github.com/jmaces/rde

Bibliography I

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7 (July 2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140.

[2] A. Coates, A. Ng, and H. Lee. "An Analysis of Single-Layer Networks in Unsupervised Feature Learning". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. PMLR, Apr. 2011, pp. 215–223.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[4] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774.

[5] J. Macdonald, S. Wäldchen, S. Hauch, and G. Kutyniok. "A Rate-Distortion Framework for Explaining Neural Network Decisions". Preprint, arXiv:1905.11092. 2019.

[6] G. Montavon, W. Samek, and K.-R. Müller. "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73 (2018), pp. 1–15. DOI: 10.1016/j.dsp.2017.10.011.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[8] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". Preprint, arXiv:1312.6034. 2013.

[9] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. "SmoothGrad: removing noise by adding noise". Preprint, arXiv:1706.03825. 2017.

[10] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for Simplicity: The All Convolutional Net". Preprint, arXiv:1412.6806. 2015.