Double Backpropagation
with Applications to Robustness and Saliency Map Interpretability

Supervisor: Prof. Dr. Dr. h.c. Peter Maass

RTG 2224

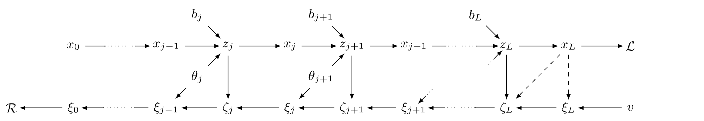
Concept

The phenomenon of double backpropagation comes into play, whenever the loss function of a neural network contains derivatives with respect to inputs.

In this work, a first in-depth study of the involved theoretical and practical phenomena is presented. In a challenging MALDI mass spectra tumor classification task (for which a specialized neural network architecture was developed), the efficacy of a specific double backpropagation loss for the increase of inter-lab robustness as well as saliency map interpretability is proven. Furthermore, the theoretical connection between the interpretability of saliency maps and the robustness of a classifier is examined. The thesis is based on the following publications:

1. Tumor Typing for Mass Spectrometry Imaging
   - Behrmann, J., **Etmann, C.**, Boskamp, T., Casadonte, R., Kriegsmann, J., & Maass, P. (2017). *Deep Learning for Tumor Classification in Imaging Mass Spectrometry*. Bioinformatics, 34(7), 1215-1223.
   - **Etmann, C.**, Schmidt, M., Behrmann, J., Boskamp, T., Casadonte, R., Hauberg-Lotte, L., Peter, A., Kriegsmann, J., & Maass, P. (2019). *Deep Relevance Regularization: Interpretable and Robust Tumor Typing of Imaging Mass Spectrometry Data*. Manuscript submitted for publication.
2. The Mathematics of Double Backpropagation
   - **Etmann C.** (2019), *A Closer Look at Double Backpropagation*. Manuscript submitted for publication.
3. Explaining the Connection Between Adversarial Robustness and Saliency Map Interpretability
   - **Etmann, C.**, Lunz, S., Maass, P. & Schönlieb, C. (2019). *On the Connection Between Adversarial Robustness and Saliency Map Interpretability*. Proceedings of the 36th International Conference on Machine Learning, in PMLR 97:1823-1832

Double Backpropagation



Neural network model:

$$\left.\begin{aligned} z_j &= K_j(\theta_j, x_{j-1}) + b_j \\ x_j &= \varphi_j(z_j) \end{aligned}\right\} \text{ for } j = 1, \ldots, L, \qquad (1)$$

where K_j is a continuous bilinear operator between real Hilbert spaces, φ_j is a activation function. This yields a *coordinate-free description!*
Main tool:

Theorem 1 (Etmann, 2019). *For a continuous, bilinear operator $K : \mathcal{P} \times \mathcal{X} \to \mathcal{Y}$ (with real Hilbert spaces $\mathcal{P}, \mathcal{X}, \mathcal{Y}$), there exist two continuous, bilinear operators $K^T : \mathcal{P} \times \mathcal{Y} \to \mathcal{X}$ and $K^\square : \mathcal{X} \times \mathcal{Y} \to \mathcal{P}$, such that $\langle K(\theta, x), y \rangle_\mathcal{Y} = \langle x, K^T(\theta, y) \rangle_\mathcal{X} = \langle K^\square(x, y), \theta \rangle_\mathcal{P}$, which are unique up to the order of arguments. These operations commute.*

Other results:

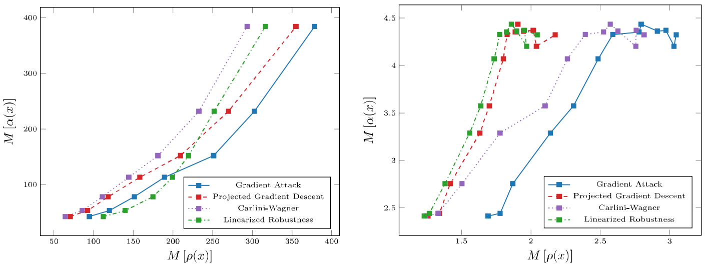
- Complexity analysis
- Description of loss surface
- Improved algorithm with up to a third of saved computations when using ReLU activations

Adversarial Robustness

Neural networks that were trained to be more robust to adversarial attacks seem to exhibit the added benefit of having more 'interpretable' saliency maps. We show that this is caused by the angle between the saliency map and the input image decreasing, as the distance to the descision boundary increases (up to linearization). We prove tight upper bounds of the linearized robustness in terms of the alignment:

$$\overbrace{\tilde{\rho}(x)}^{\text{linearized robustness}} \leq \underbrace{\alpha(x)}_{\text{image-gradient-alignment}} + \overbrace{\|x\| \cdot \|\bar{g}^\dagger - \bar{g}\|}^{\text{multi-class effects}} + \underbrace{\frac{|\beta^\dagger|}{\|g^\dagger\|}}_{\text{non-homogeneity effects}}. \qquad (2)$$

The theoretical findings are validated by training various classification models on image datasets, where the adversarial robustness is varied using double backpropagation losses.



MALDI Imaging

Mass spectra from histopathological samples of cancerous tissue were analyzed for the specific subtype of cancer. For this, the custom *IsotopeNet* neural network architecture was developed. In inter-lab scenarios, these however suffer from overadapting to dataset characteristics. By employing the custom *deep relevance regularization* loss

$$\min_\theta \ell(f_\theta(x), y) + \lambda_1 \|r(x, y)\|_1 + \lambda_2 \|r(x, y)\|_2^2,$$

sparsity on the classification explanations was enforced. Here, $r(x, y)$ is the gradient-based *layer-wise relevance propagation* proposed by Bach et. al (2015). This is intended to restrict the model to only take into account the most salient features, which in this case promotes the utilization of biologically relevant features instead of measurement artifacts.

	IsotopeNet	IsotopeNet+DRR	Linear Baseline Method
bal. Acc. (spot)	37.3%	**77.4%**	75.5%
bal. Acc. (patient)	34.9%	**80.6%**	78.8%

The proposed regularization scheme succeeds in robustifying the neural network. Furthermore, the biological plausibility of the model is increased.

DFG

Universität Bremen