„Methods for interpreting and understanding deep neural networks“

Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller 2017

Presented by Philipp Wimmer

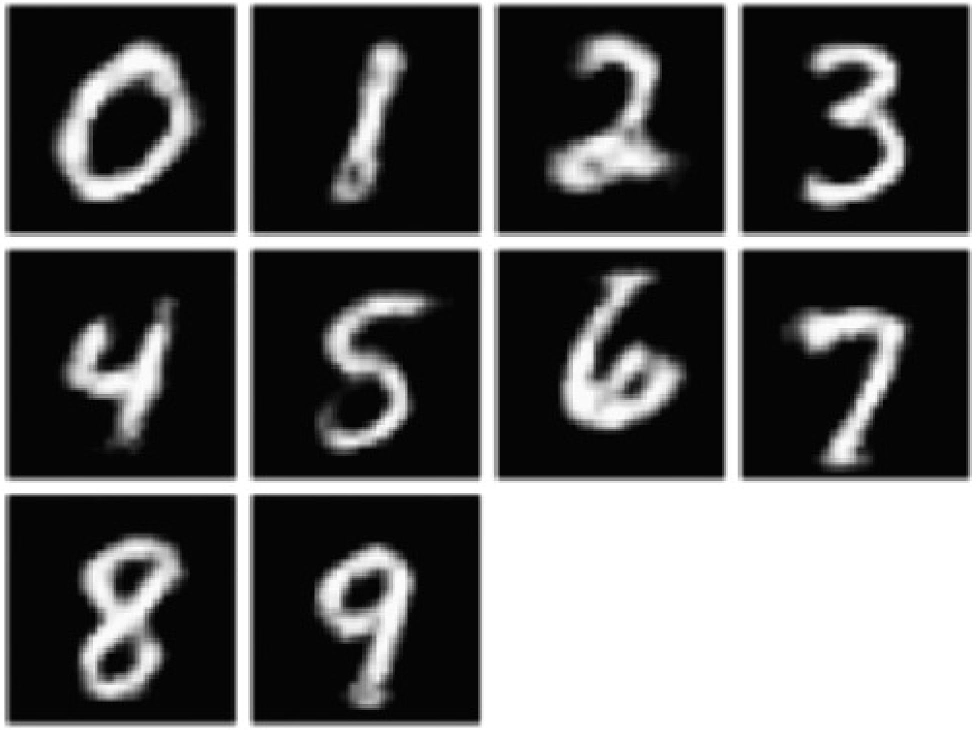
Motivation

- Understanding and validating deep neural networks is hard
  - Many parameters
  - Highly nonlinear
  - Interpratability wasn‘t a goal of DNNs
- Ability to validate is neccessary for understanding and real world applicability
- Example: Don‘t know if high prediction accuracy is due to anomaly in training data

Interpretation

*An **interpretation** is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.*

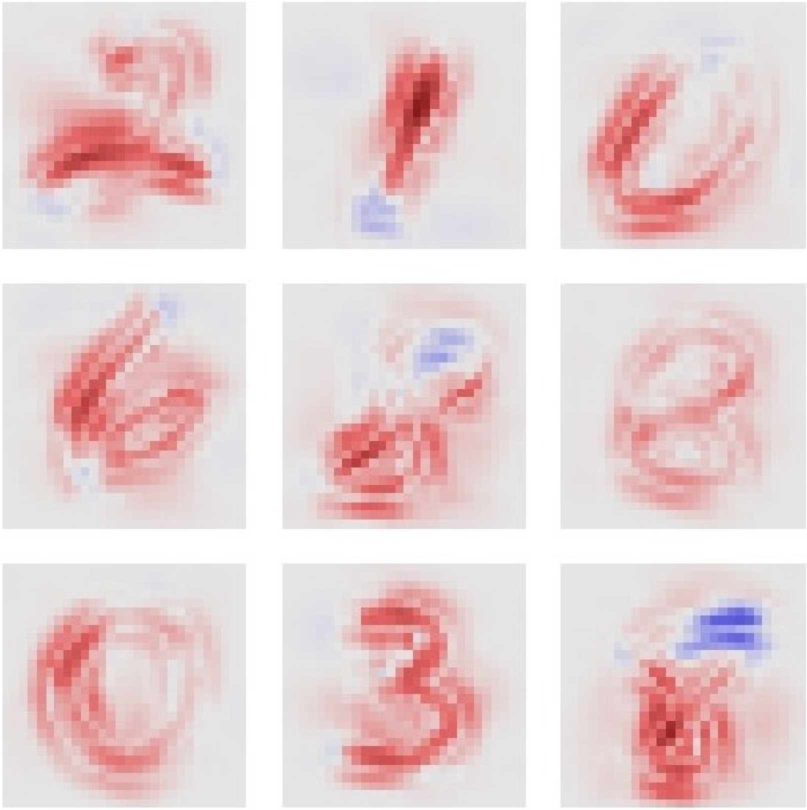
Goal: Producing a **prototype**



Explanation

*An **explanation** is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression).*

Goal: Producing a **heatmap**



Part A: Interpreting

- Activation Maximazation (AM)
- AM with an expert
- AM in code space (using Generative Adverserial Networks)

Activation Maximization

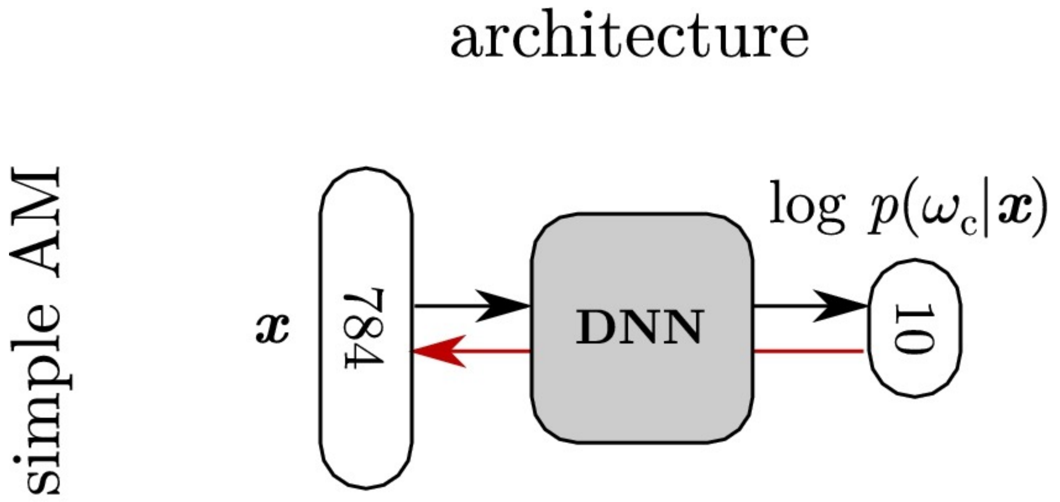
- Producing a prototype via maximizing

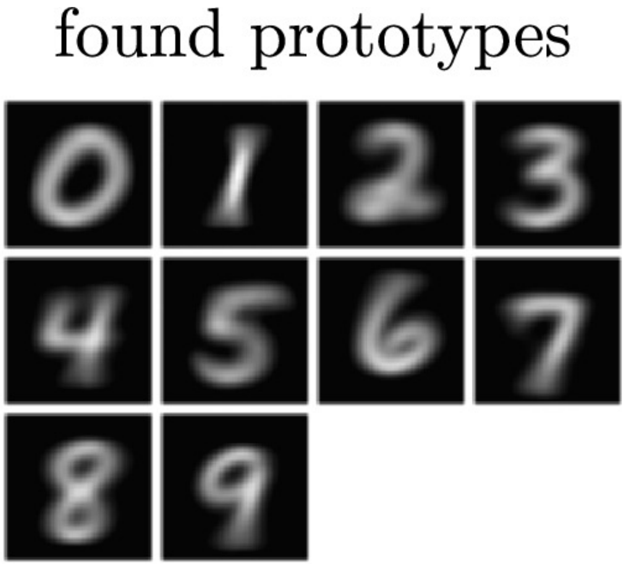
$$\underbrace{\max_{\mathbf{x}} \log p(w_c|\mathbf{x})}_{\text{Class probabilities}} - \underbrace{\lambda||\mathbf{x}||^2}_{l^2\text{ Regularizer}}$$

- Class probabilites modeled by the DNN are functions with a gradient
- Use gradient descent to maximize (just like training a DNN in reverse

Architecture of AM

- Simple to compute
- Regularizer preferes inputs close to the origin (mean of data)
- Unnatural looking protoype





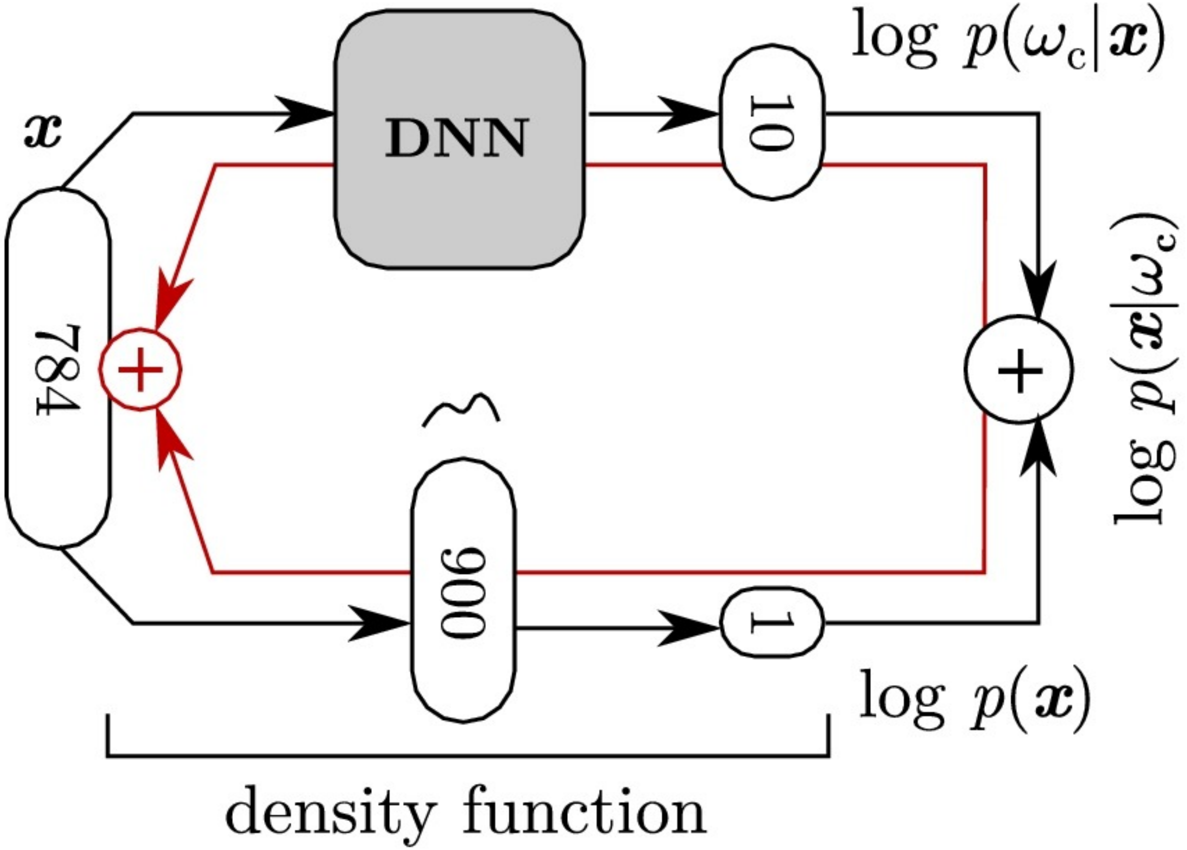
Improving AM with expert

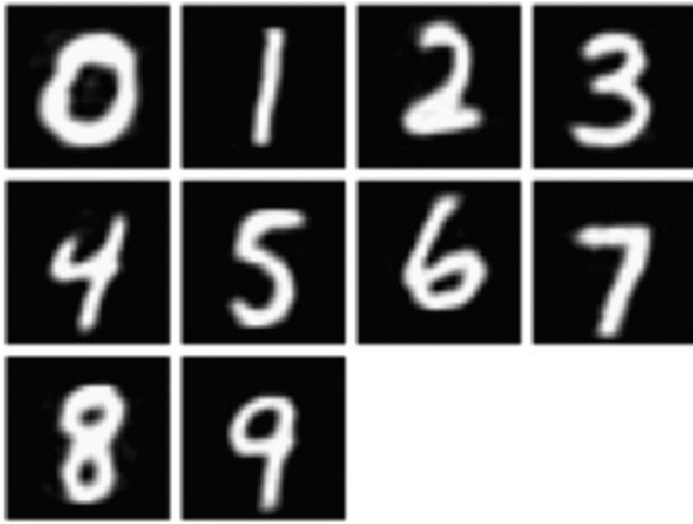
- Replace regularizer with a more sophisticated approach

$$\max_{\mathbf{x}} \underbrace{\log p(w_c|\mathbf{x})}_{\text{Class probabilities}} + \underbrace{\log p(\mathbf{x})}_{\text{Model of the data}}$$

- Expert is the data density
- For example obtained by training an Gaussian RBM
- Often more complex density models are needed

AM + expert


x
DNN
10
log p(ω_c|x)
784
900
1
log p(x|ω_c)
log p(x)
density function

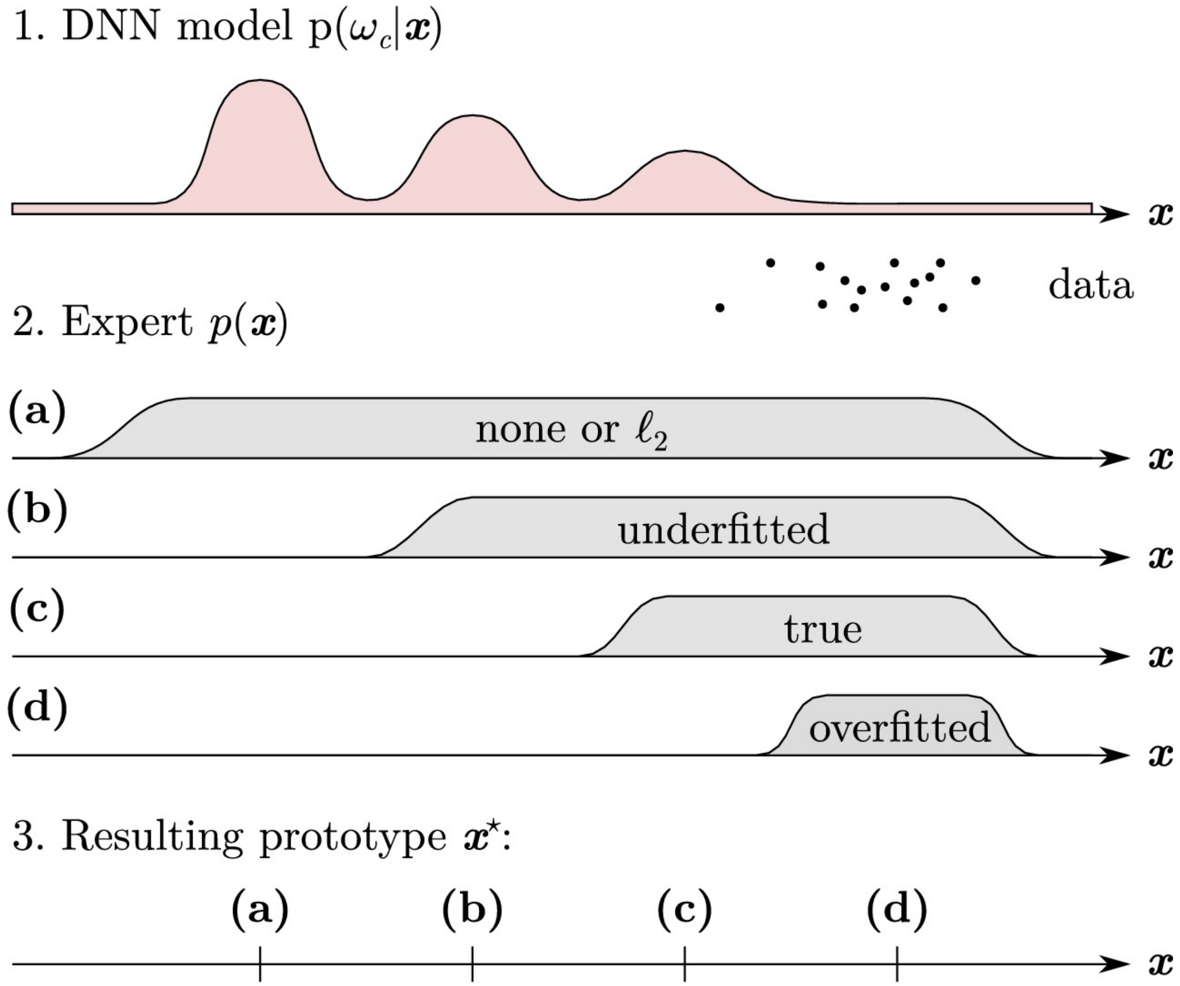

0 1 2 3
4 5 6 7
8 9

(a) maximation of class probability function

(b) favoring natural images – often sufficient.

(c) desired

(d) optimization of the expert itself, hides failure modes



Performing AM in code Space

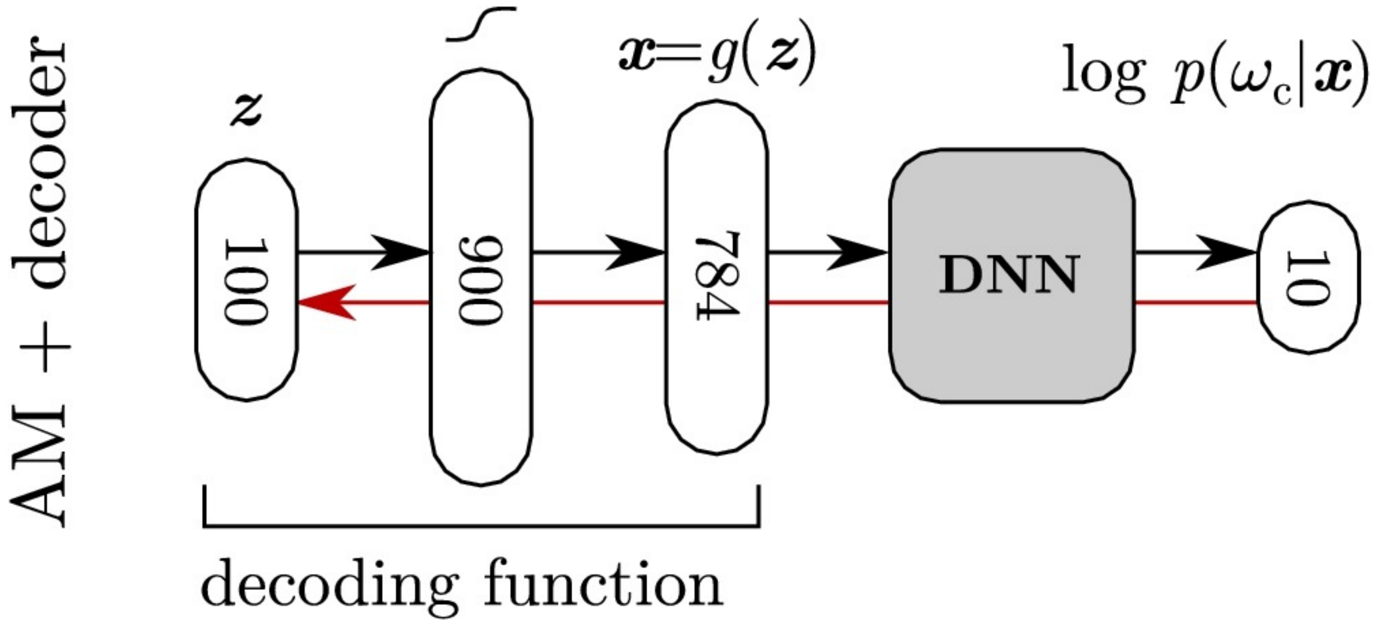
- Often learning the expert to a high accuracy is hard
- Expert often very complex such that maximizing is difficult
- A solution is to not explicitely learn p(x)
- Instead sample from an code space with known distribution which was obtained by training an GAN

$$\max_{\mathbf{z} \in Z} \log p(w_c | \underbrace{g(\mathbf{z})}_{\text{Decoded point in Code space}}) - \lambda ||\mathbf{z}||^2$$

- Then apply decoding function to get a protoype

Performing AM in code space

- Distribution in code space is by construction Gaussian
- Regularizer favors points with a high probability



Results

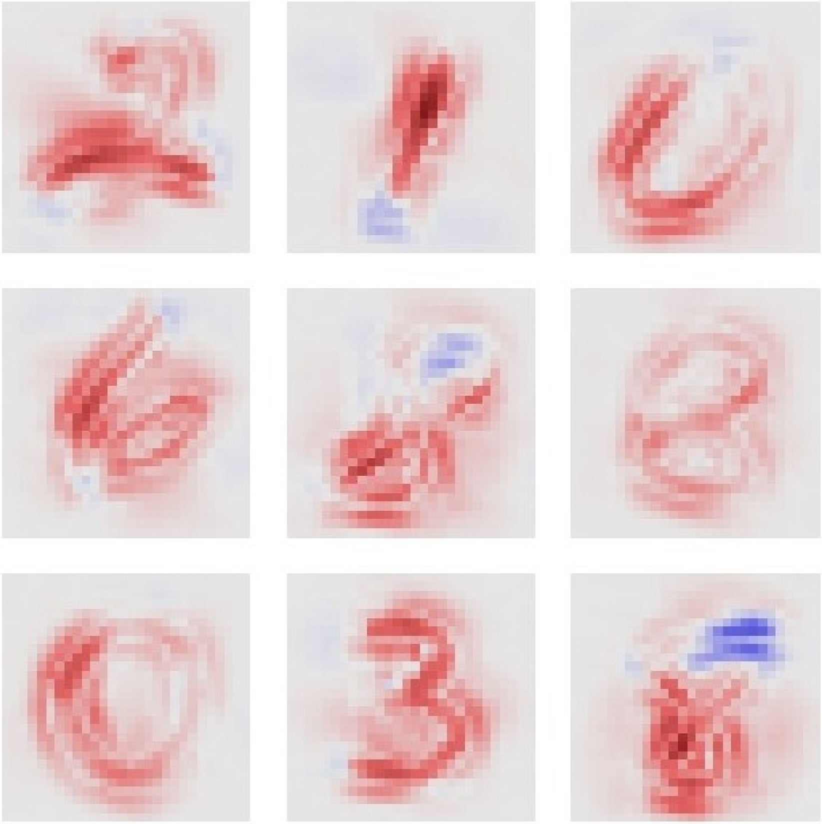
Simple AM

AM with expert

AM in code space

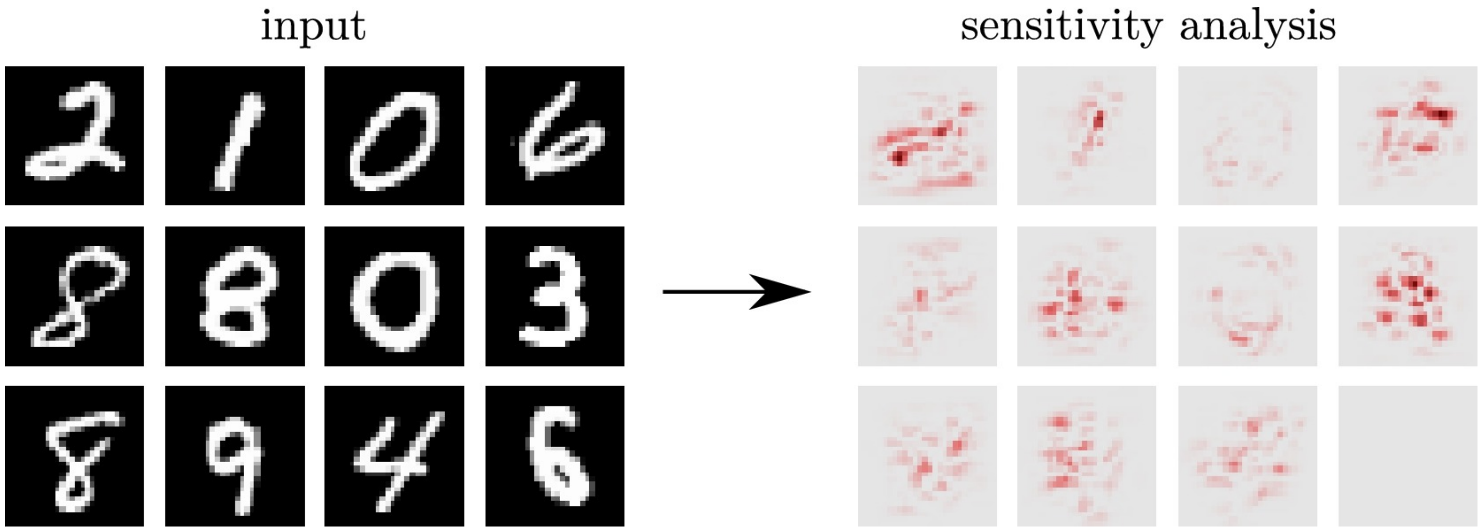
Part B: Explaining

- Sensitivity Analysis
- (Layerwise) Relevance Propagation



Sensitivity Analysis

$$R_i(\mathbf{x}) = \left(\frac{\partial f}{\partial x_i}\right)^2$$



Sensitivity Analysis

- Gradient is easily calculated via backpropagation
- The measured relevance score is not what is wanted
- Measures not the relevance, but the local slope of it

Relevance Propagation

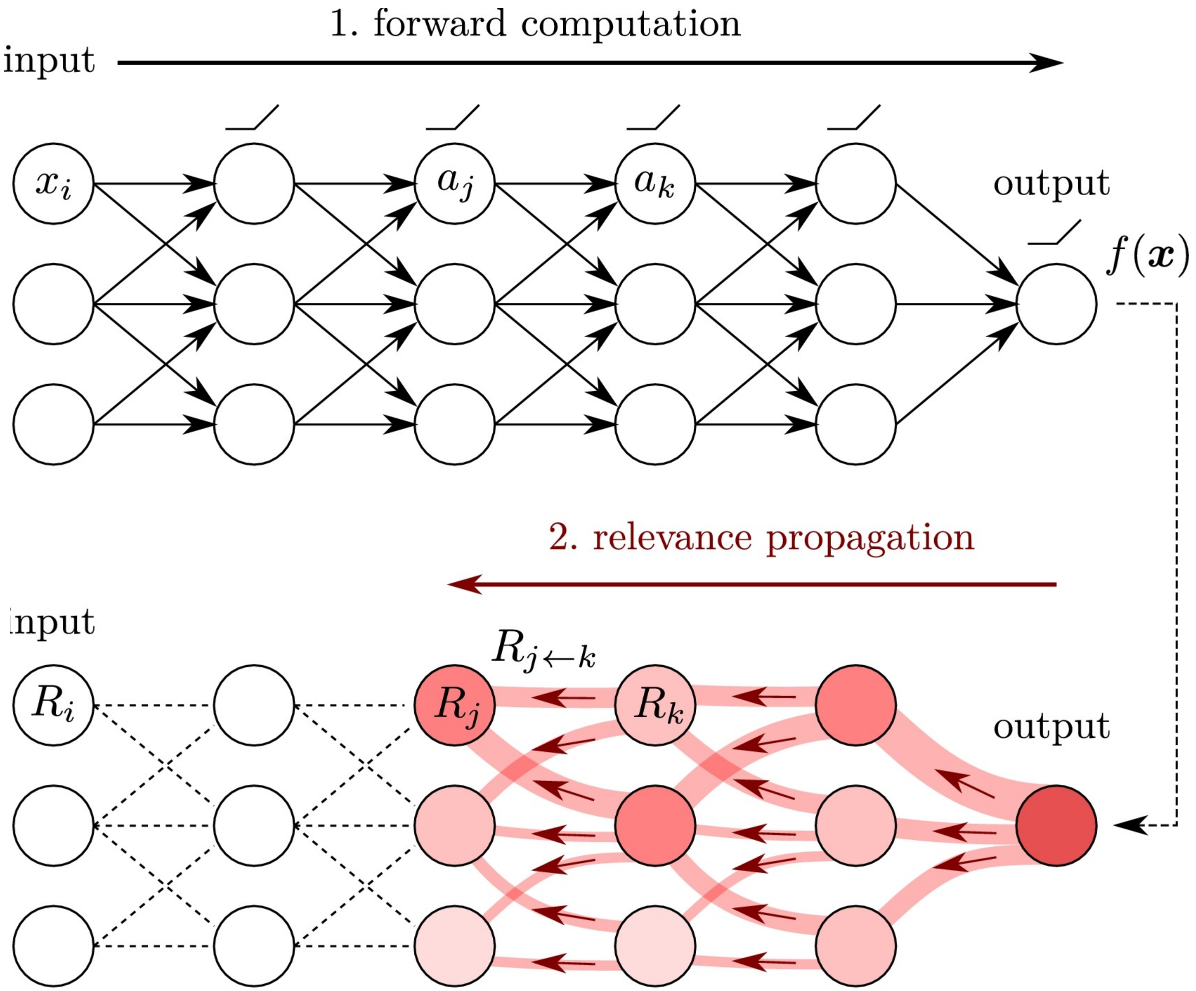
- Make use of the graph structure of DNNs
- Propagate the relevance score backwards through the network is similar to the backpropagation of the error during the training phase
- Relevance has to be conserved (similar to current in an electric circuit)
- Local conservation at each neuron
- Filtering: Able to block the flow through certain neurons

$$(1) \quad a_k = \sigma\left(\sum_j a_j w_{jk} + b_k\right)$$

$$(2) \quad \sum_j R_{j\leftarrow k} = R_k$$

$$(3) \quad R_j = \sum_k R_{j\leftarrow k}$$

$$(4) \quad \sum_{i=1}^{d} R_i = f(\mathbf{x})$$



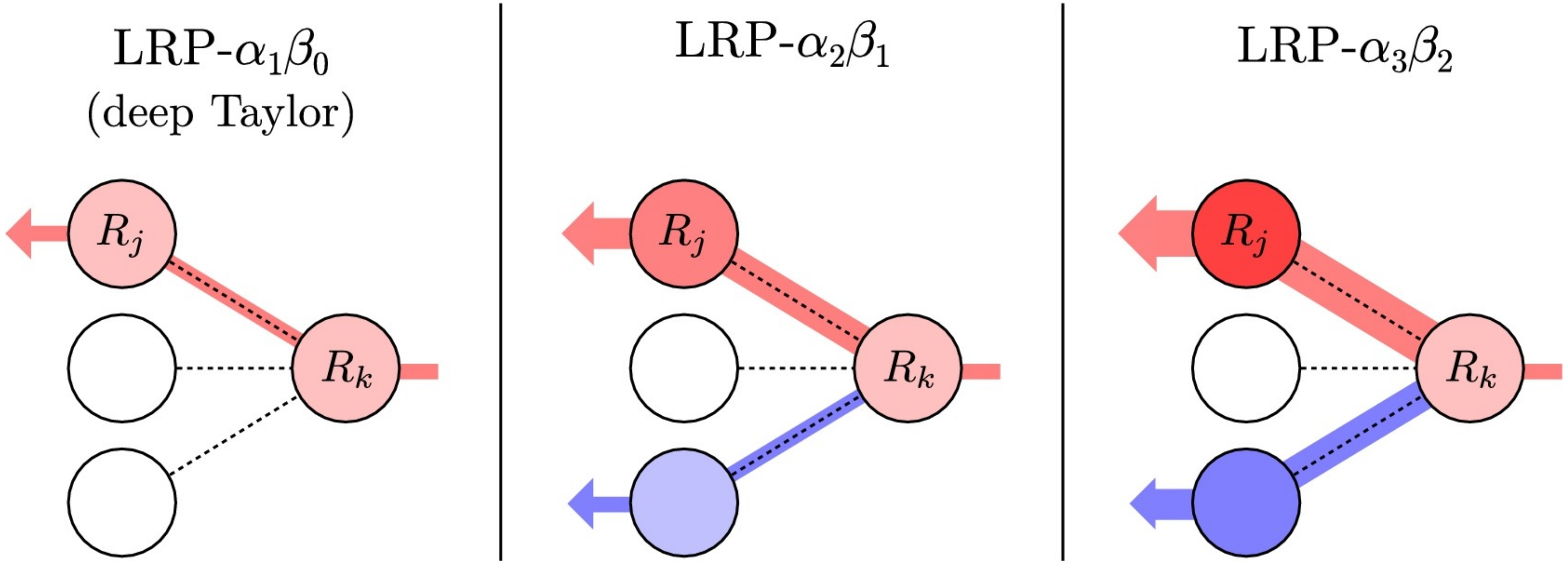
Propagation Rule

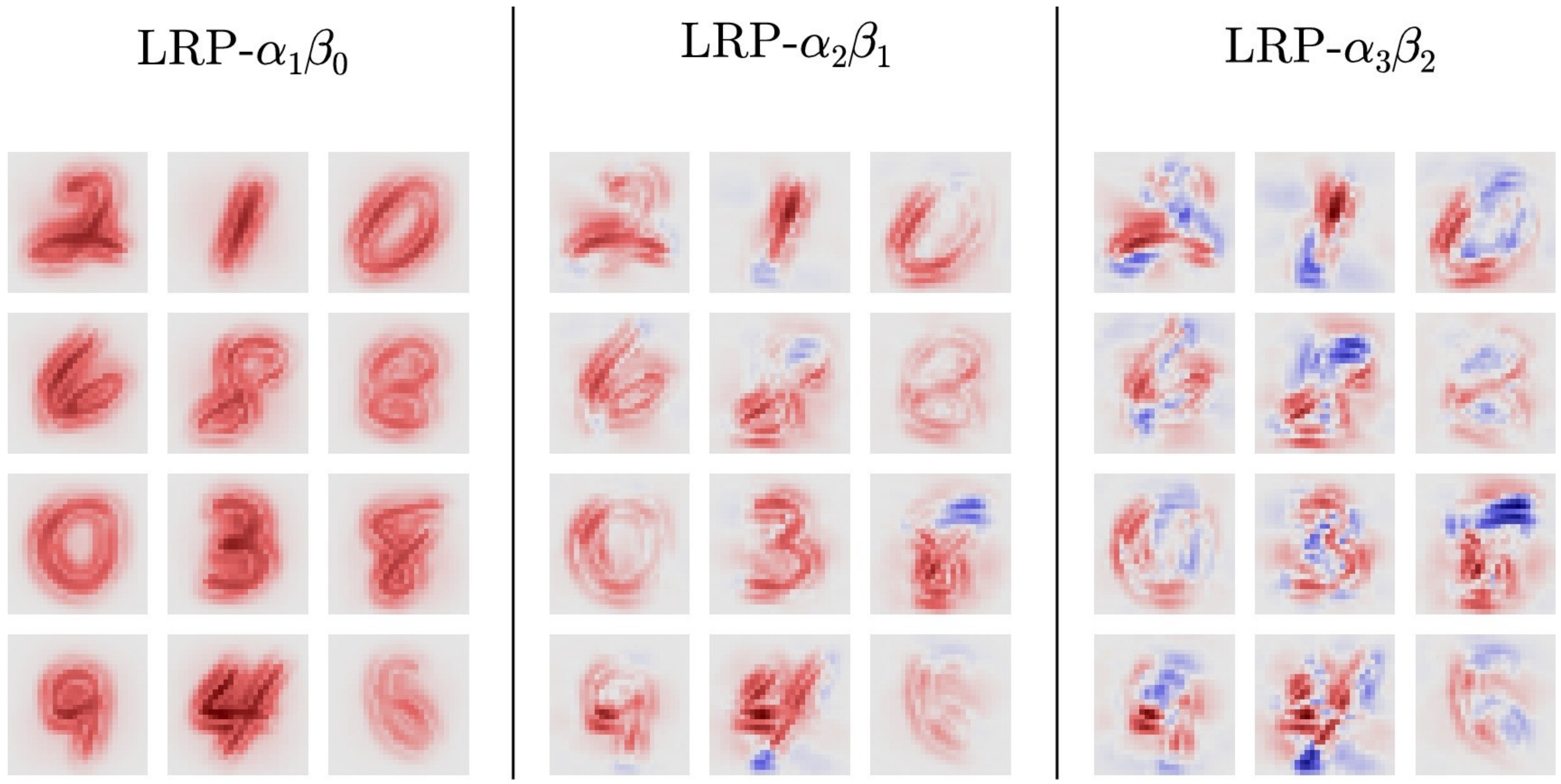
$$R_j = \underbrace{\sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} \alpha R_k}_{\text{Relevance}} - \underbrace{\sum_k \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \beta R_k}_{\text{Counter-Relevance}}$$

$$\alpha - \beta = 1, \beta \geq 0$$

Hyperparameters of LRP

- Ratio of α and β determines the influence of counter variance




LRP-$\alpha_1\beta_0$
LRP-$\alpha_2\beta_1$
LRP-$\alpha_3\beta_2$

Conclusion

- Two methods for increasing post-hoc interpretability
- No need to change existing algorithms
- Enables better understanding and validation
- Should be in the toolbox of everyone using DNNs