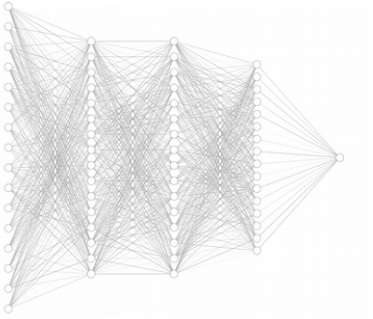
Interpretability and Deep Learning

Martin Erdmann, **Jonas Glombitza**

RWTH Aachen



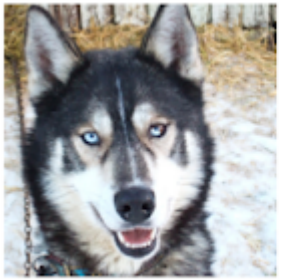
Interpretability

Deep models have thousands of parameters → open black box

- "What is the model learning?"
  - Learn from trained model
- "Can we trust the model? Does the model work as expected?"
  - Model verification
    - Systematic studies (strongly application dependent)
    - → **Interpretability**

Interpretability

ArXiv/1602.04938



(a) Husky classified as wolf

(b) Explanation

Interpretability includes

- Understanding **data**
  - “Which part of the data is most useful?”
- Understanding **predictions**
  - “Why is my model predicting a certain class / value?”
  - → Feature attribution
- Understanding the **model**
  - “How is the model working / are features formed?”
  - “How do DNNs see the world?
  - → Feature visualization

- Possible bias: we (humans) try to interpret the model

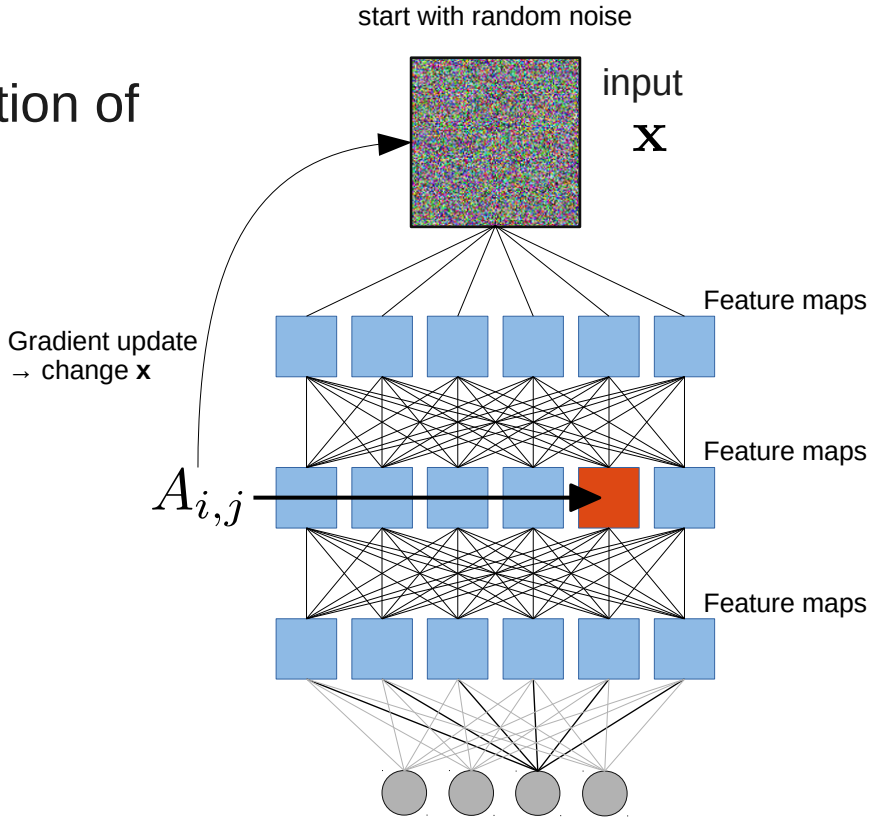
Feature Visualization

Model interpretability

Activation Maximization - CNN

Idea:

- Construct pattern which maximizes the activation of a specific feature map
- Model f_θ pre-trained, weights θ fixed
- Find $\tilde{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}}\ h(\mathbf{x}, \theta)$
- $h(\mathbf{x}, \theta) = \sum_{i,j} A_{i,j}(\mathbf{x}, \theta) + b$
- Gradient **ascent** $\mathbf{x}' \rightarrow \mathbf{x} + \alpha \frac{dh(\mathbf{x}, \theta)}{d\mathbf{x}}$

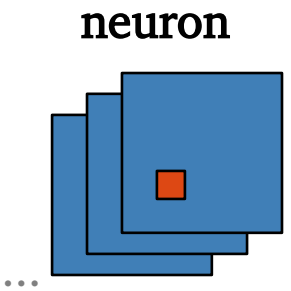


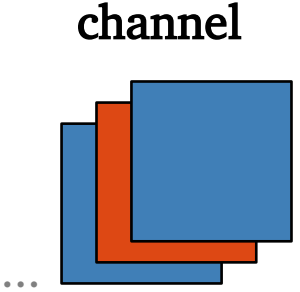
https://doi.org/10.1142/12294

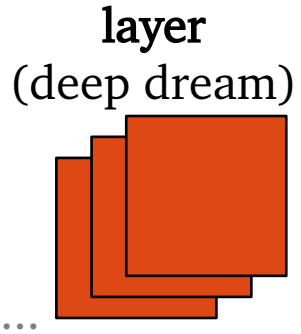
Activation Maximization

- Visualization of neurons, channels, layers

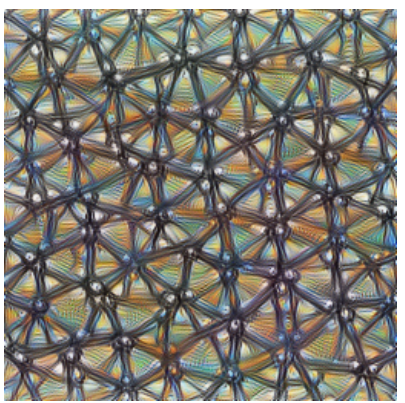
objective

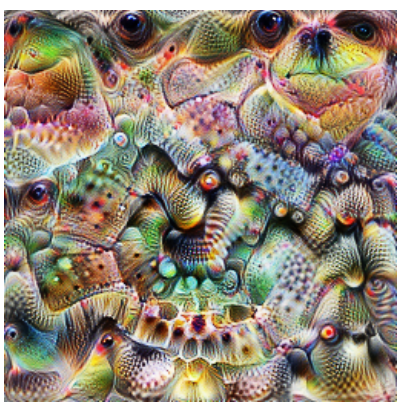






obtained visualizations

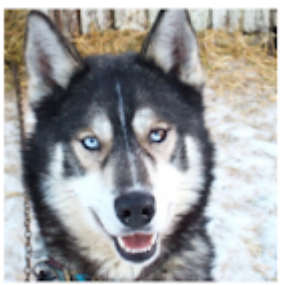




https://doi.org/10.1142/12294

Analysis of predictions

Sensitivity Analyses
(Feature) Attribution

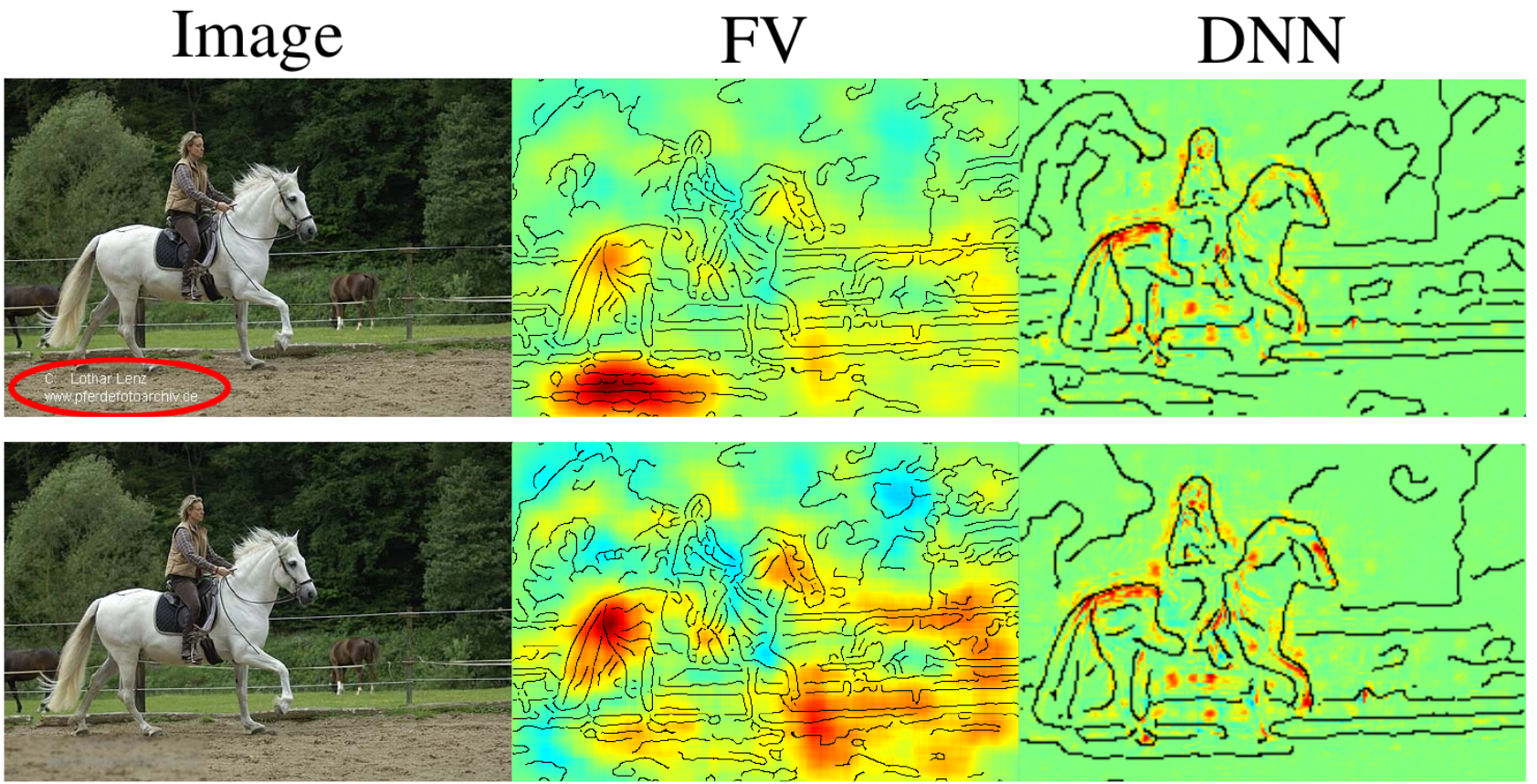


(a) Husky classified as wolf

(b) Explanation

ArXiv/1602.04938

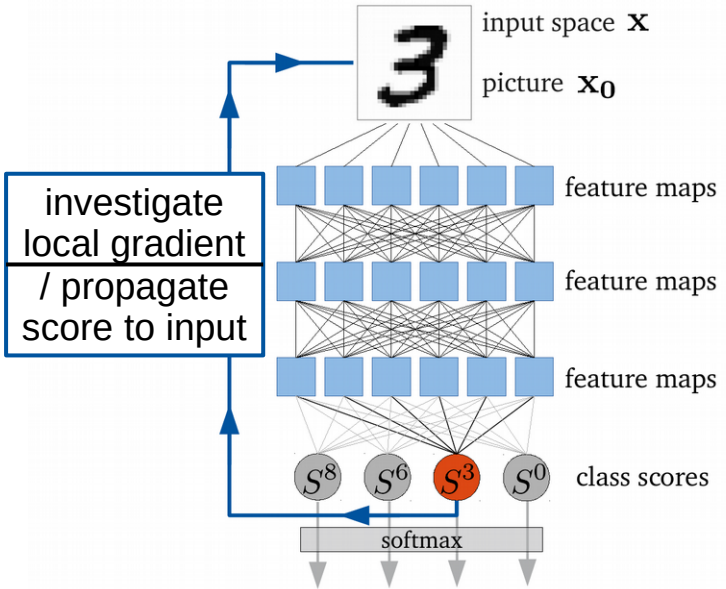
Semantic Misinterpretation



Bach et. Al. - Analyzing Classifiers: Fisher Vectors and Deep Neural Networks, arXiv:1512.00172

Prediction analysis

- **Sensitivity analyses**
  - saliency maps (ArXiv/1312.6034)
  - study to what the DNN is **sensitive** to



https://doi.org/10.1142/12294

- **Attribution analyses**
  - fulfill **completeness:** $\sum_i R_i^c = S^c(\mathbf{x_0})$

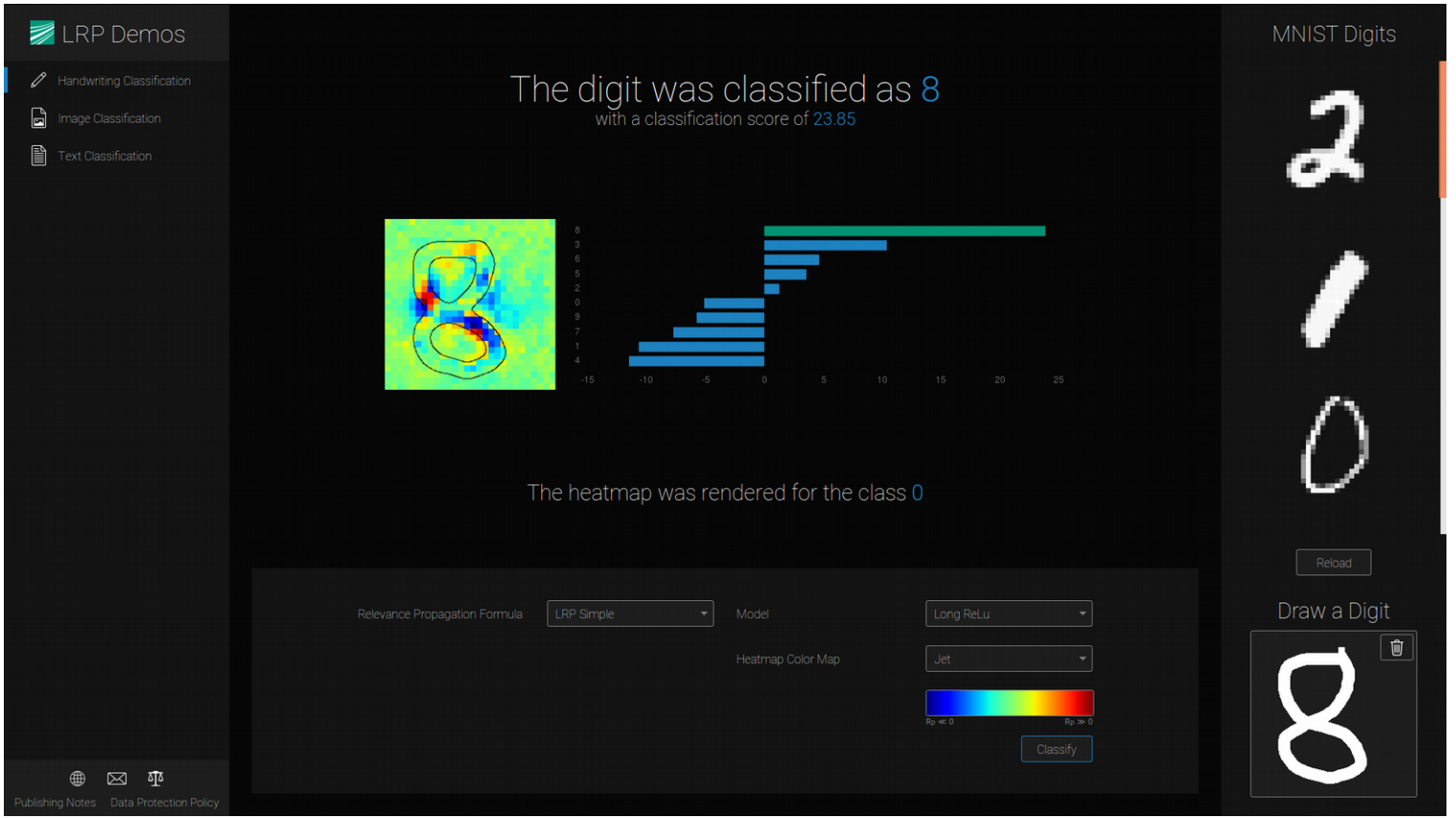
prediction

sum runs over all inputs

relevance/attribution

- sum over all input relevances = prediction
- common methods:
  - layer-wise relevance propagation, IntegratedGradients, DeepLIFT

DEMO - Handwriting



https://lrpserver.hhi.fraunhofer.de/handwriting-classification

Summary: understanding deep networks

Feature visualization

- understanding the model – “What is learned by the network?”

Prediction analysis

- interpret a prediction – Why a specific pattern caused certain reconstruction

Fast growing field of research → many methods ‘on the market’

- ➔ study your network using a collection of techniques
- understand your model, debug your architectures
- perform other tasks (segmentation), learn about the data

- ➢ Software libraries: iNNvestigate, DeepExplain, Captum