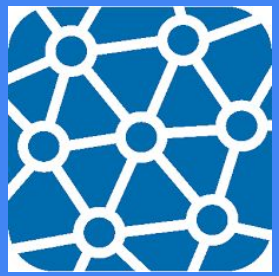
Explaining Deep Neural Network using Layer-wise Relevance Propagation and Integrated Gradients

MSc. I. Čík, MSc. A.D. Rasamoelina, doc. Ing. M. Mach CSc., prof. Ing. Peter Sinčák CSc.
Submission 74
SAMI



What does it mean “to explain”?

Oxford Dictionary definition of ‘explainable’:

A statement or account that makes something clear; a reason or justification is given for an action or belief.

Are existing AI systems explainable ?

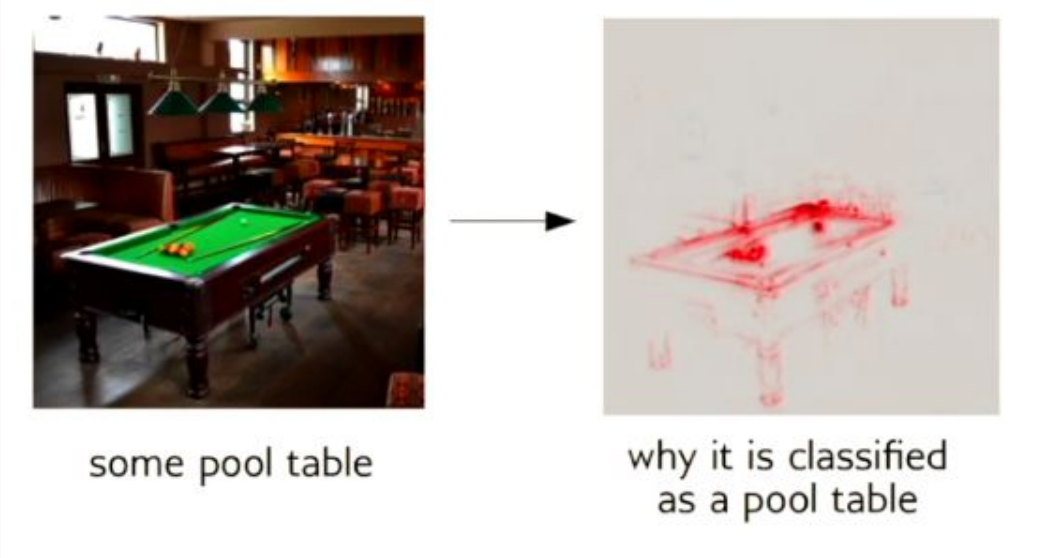
History

- Van Lentet et al. (“An explainable artificial intelligence system for small-unit tactical behavior”, 2004)
- 7±2 pieces of information at a time (Miller, 1956)
- Reasoning/explaining of expert systems (1970’s)

Explainability

"Why a given image is classified as a table ?

- Active topic
- No accepted definition
- Polarizing topic (Rudin ‘19)
- Explainability /Intepretability

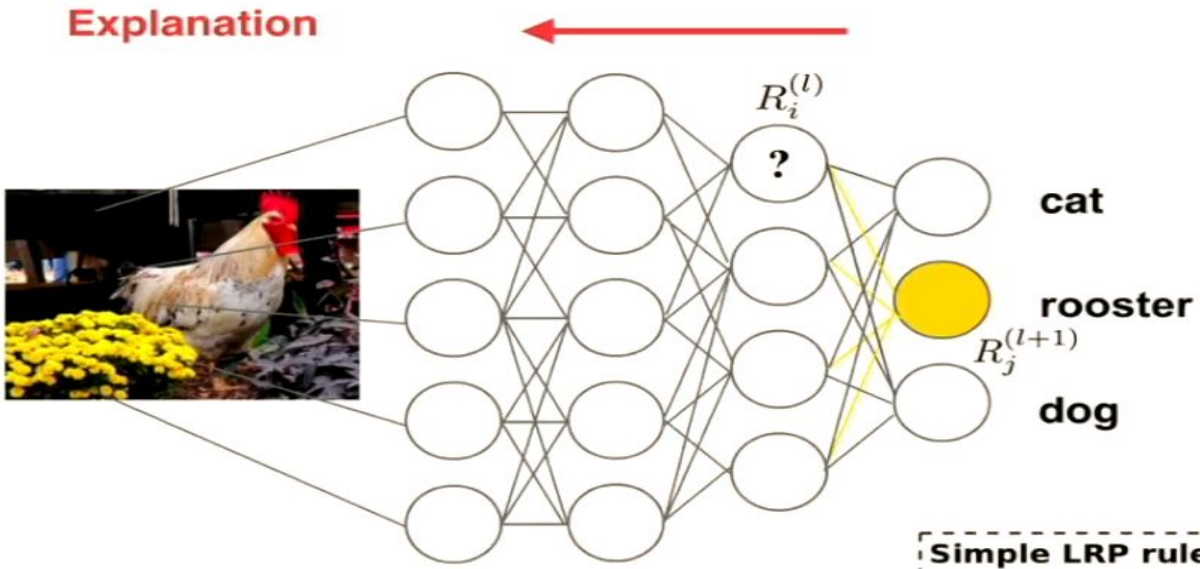


Which method to choose?
Can we learn more about the classifier?
What to do with explanations?

Explanation Methods

- Post-hoc/ Integrated (transparency based) Explainability
- Perturbation-based methods
- Function-based methods
- Surrogate-/Sampling-based methods
- Structure-based methods
- Xie ‘20, 3 types (Visualization, Distillation, Intrinsic)

Layer-wise Relevance Propagation

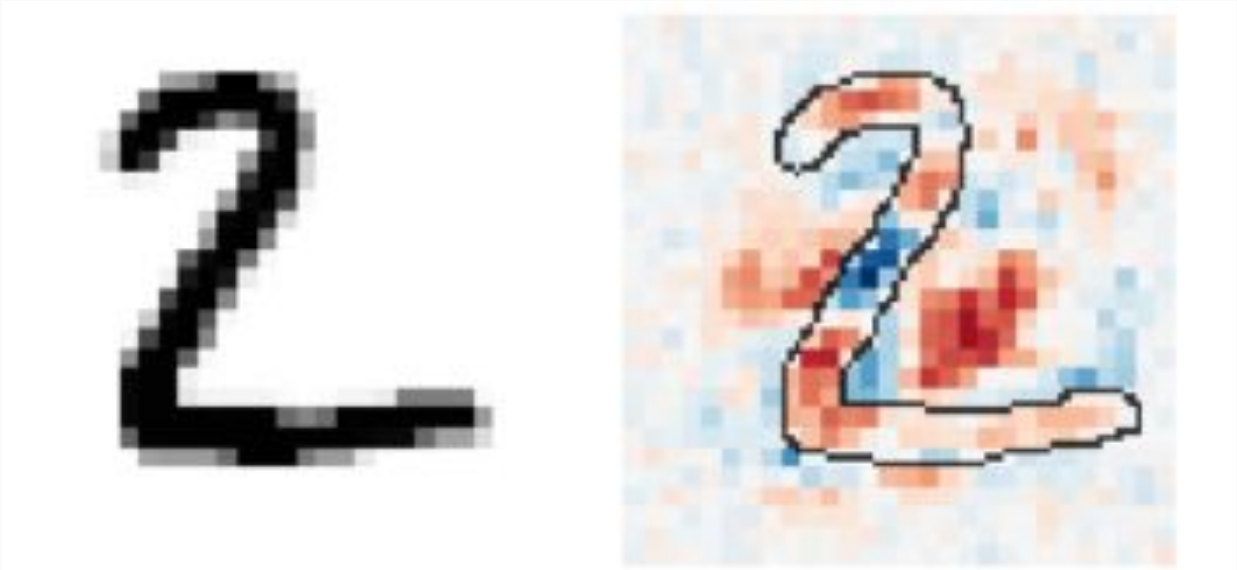


Simple LRP rule (Bach et al. 2015)

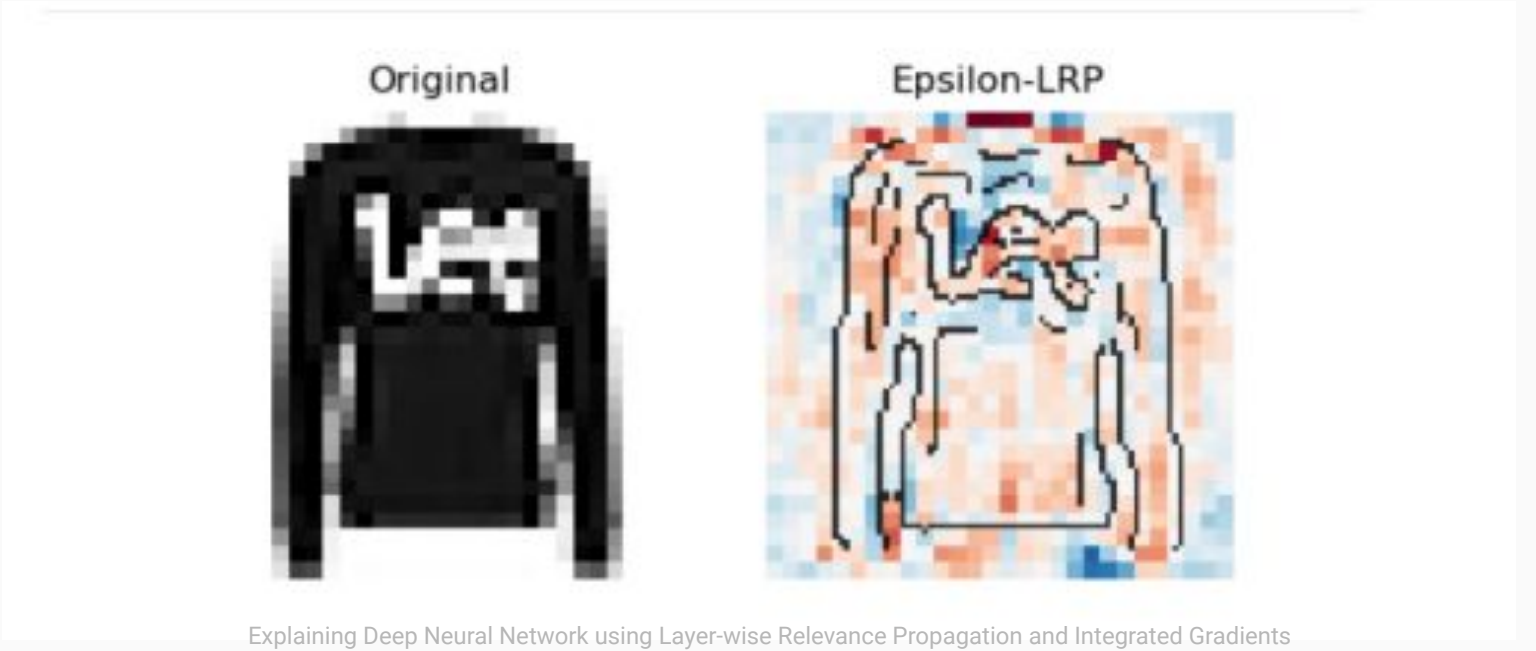
$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Every neuron gets its "share" of the redistributed relevance

Experiments with LRP

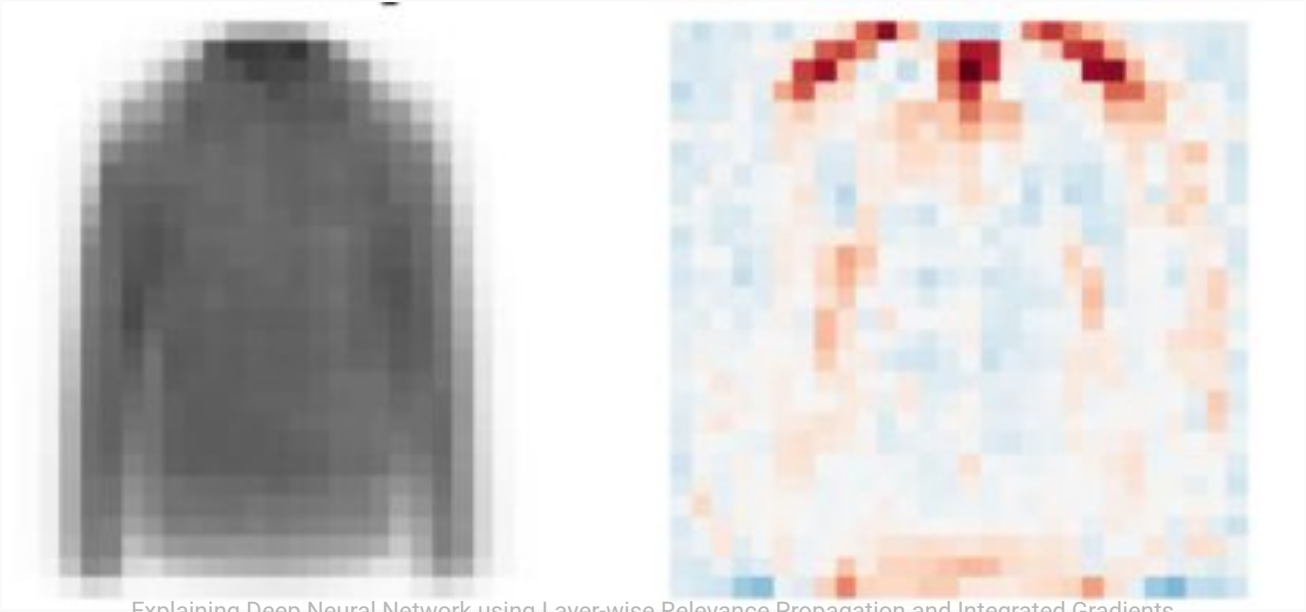


Experiments with LRP - Individual



Explaining Deep Neural Network using Layer-wise Relevance Propagation and Integrated Gradients
IEEE 19th World Symposium on Applied Machine Intelligence and Informatics

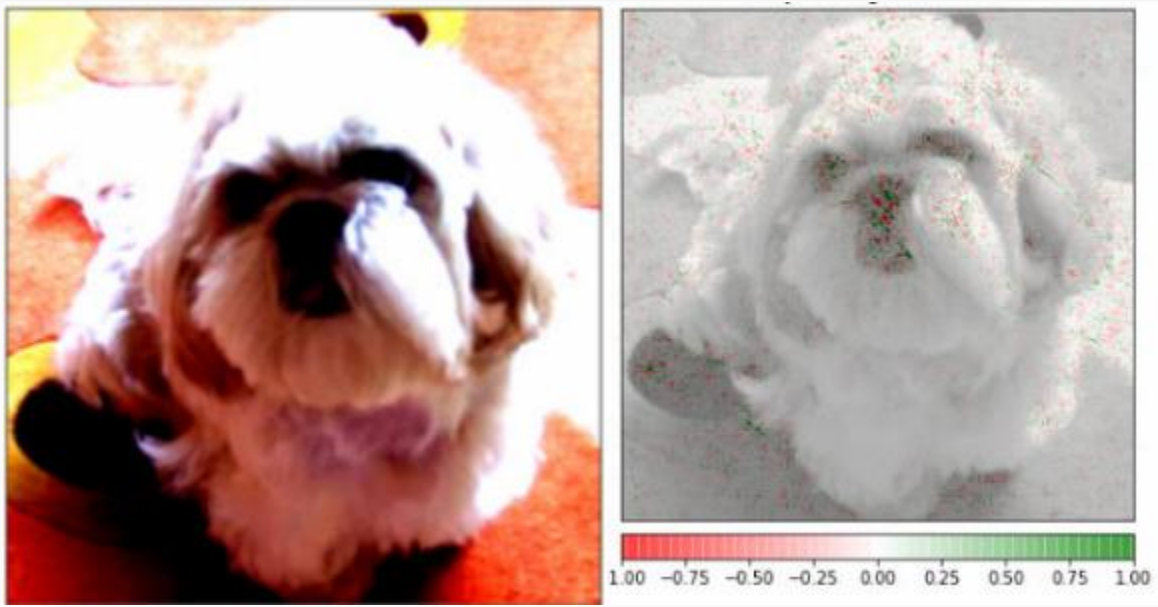
Experiments with LRP - Cumulated



Integrated Gradients

Goal: Computes the integral of the gradients of the output prediction for the class with respect to the input image pixels.

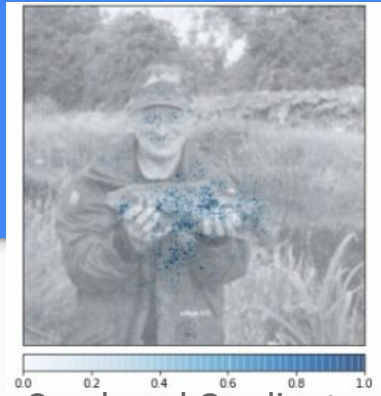
Overlayed Integrated Gradients



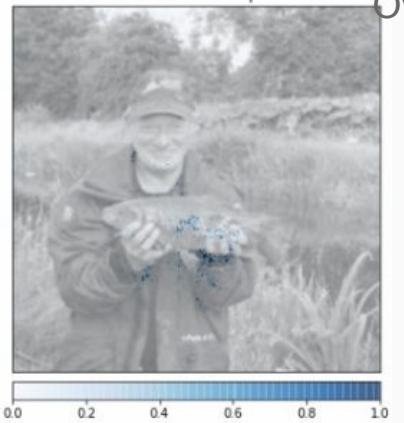
Integrated Gradients

Original image

Overlayed deeplift



Overlayed Gradient Magnitudes

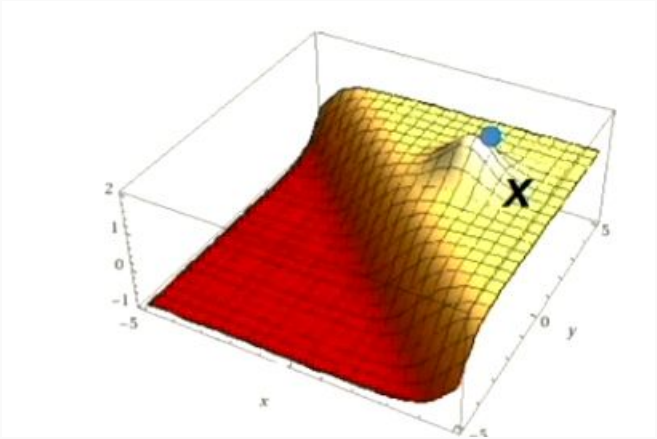


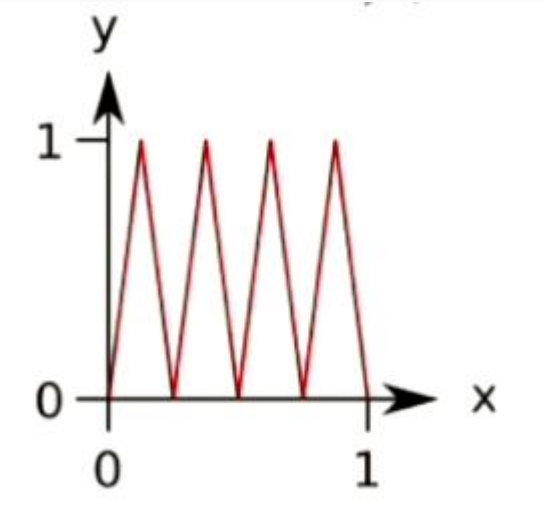
Overlayed Integrated gradients

Overlayed Integrated Gradients with SmothGrad Squared

Explaining Deep Neural Network using Layer-wise Relevance Propagation and Integrated Gradients
IEEE 19th World Symposium on Applied Machine Intelligence and Informatics

Why do we have noise explanations ?





Conclusion

- High performance <-> Low interpretability
- Do we need to understand our algorithms?
- Do we still have to develop new XAI techniques?
- What does it mean “to explain” or “to understand”?

DISCUSSION

THANKS FOR YOUR ATTENTION