xAI with Layer-wise Relevance Propagation

Anna Saranti & Andreas Holzinger

human-centered.ai

11.05.2021

Outline

Introduction

LRP vs. Sensitivity Analysis (SA)

Whole dataset analysis with LRP

LRP on LSTMs and Perturbation Analysis

LRP for Pruning

LRP for NNs

LRP task

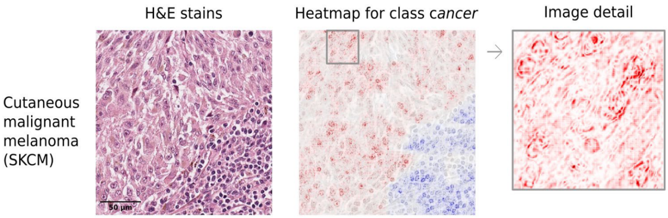
Benefits of LRP

Literature

Questions

Heatmaps

- Binary classification task
- Cancer or healthy?



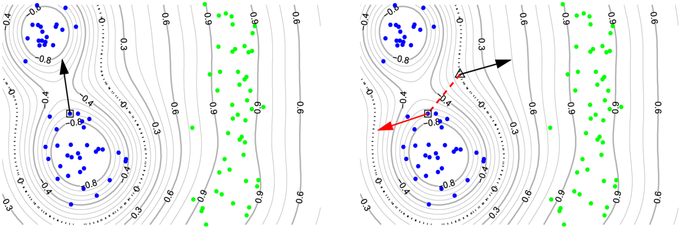
Hägele, Miriam, et al. "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods." Scientific reports 10.1 (2020): 1-12.

LRP vs. SA (1/4)

- What is a good heatmap?
- Sensitivity of a pixel p is the norm over all partial derivatives:
  $h_p = ||\frac{\partial}{\partial x_p} f(x)||$
- How much a small change in the pixel p affects the prediction (output) of the NN
- The direction of change is lost because of the norm
- Needs (locally) differentiable neurons

LRP vs. SA (2/4)

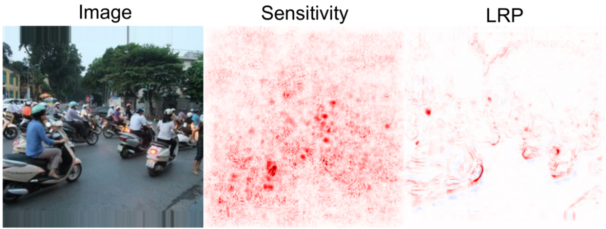
- Left: Local gradient at prediction point
- Right: Taylor approximation w.r.t. root point on decision boundary



Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

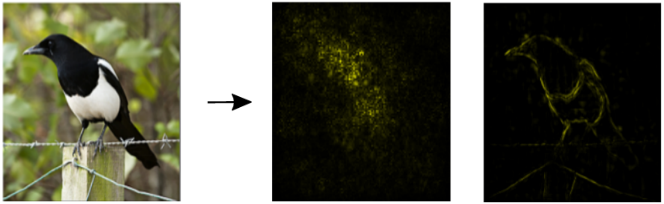
LRP vs. SA (3/4)

- Blue color denotes negative relevane
  Evidence **against** the predicted class



Samek, Wojciech, et al. "Interpreting the predictions of complex ml models by layer-wise relevance propagation." arXiv preprint arXiv:1611.08191 (2016).

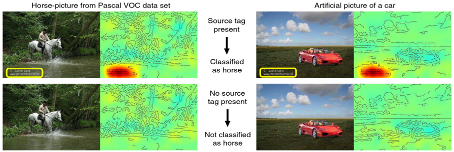
LRP vs. SA (4/4)



Samek, Wojciech, et al. "Evaluating the visualization of what a deep neural network has learned." IEEE transactions on neural networks and learning systems 28.11 (2016): 2660-2673.

Whole dataset analysis (1/2)

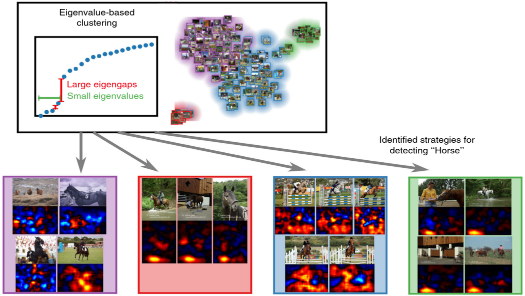
- PASCAL VOC2007 data set: horse images have a tag
- Classification by high-performing NN
- Use LRP and detect Clever Hans predictions



Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

Whole dataset analysis (2/2)

- Semi-automated Spectral Relevance Analysis
- Improve the model and the dataset



Lapuschkin, Sebastian, et al. “Unmasking clever hans predictors and assessing what machines really learn.” Nature communications 10.1 (2019): 1-8.

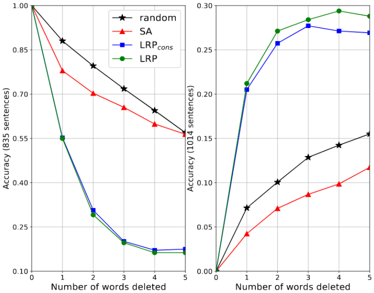
LRP on LSTMs and Perturbation Analysis (1/2)

- Sentiment classification task

true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
		1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
--	--	6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
		1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
--	--	6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .

Arras, Leila, et al. "Explaining recurrent neural network predictions in sentiment analysis." arXiv preprint arXiv:1706.07206 (2017)

LRP on LSTMs and Perturbation Analysis (2/2)

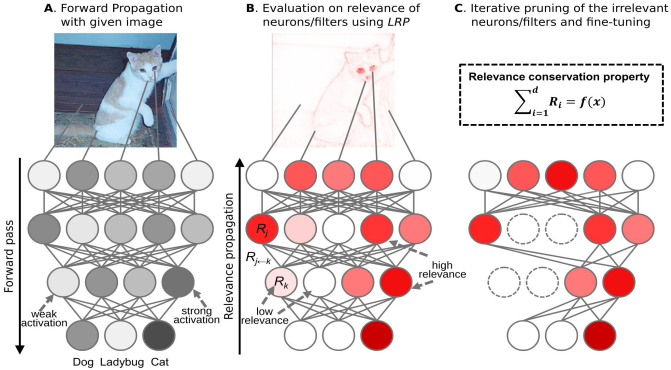


Arras, Leila, et al. "Explaining recurrent neural network predictions in sentiment analysis." arXiv preprint arXiv:1706.07206 (2017)

- How does word deleting affect performance?
- Left: Correct classification, decreasing relevance
- Right: Misclassification, increasing relevance

LRP for Pruning CNNs

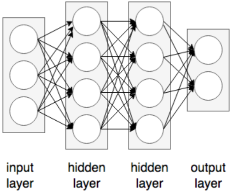
- Compress the model but keep performance
- Fight overparameterization: More parameters than training samples
- Use xAI to find out most relevant units (weights, filters) automatically



Yeom, Seul-Ki, et al. "Pruning by explaining: A novel criterion for deep neural network pruning." Pattern Recognition 115 (2021)

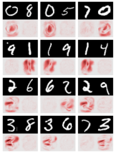
Fully Connected Neural Network

- Input image x processed by neural network (NN) f.e. for a classification task
- The neural network computes the function $f(x)$
- Function $f(x) = 0$: No object in image
  $f(x) > 0$: Object in image with a degree of certainty



Result of LRP when applied in a Convolutional Neural Network (CNN)

- Decompose the decision of the NN into the contributions of individual pixels $\mathrm{x} = \{x_p\}$
- To what extent the pixel p contributes to explaining the classification decision $f(\mathrm{x})$
- Heatmaps for correctly classified and misclassified

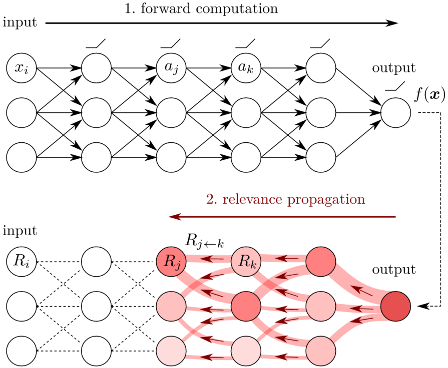


Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

NN training procedure

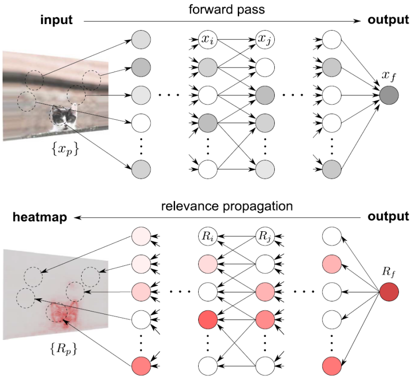
- Feedforward, although the network is trained by backpropagation of the error of its prediction with the training data
- LRP is applied after the end of the training procedure
- Training must have good performance; this will influence the quality of the explanations
- The computations of LRP use one backward pass in an already trained NN.
- For the GNN case, multiple backpropagation passes are needed - not to be confused with NN training with backpropagation.

Computational flow of Deep Taylor Decomposition (1/2)



Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." Digital Signal Processing 73 (2018): 1-15.

Computational flow of Deep Taylor Decomposition (2/2)



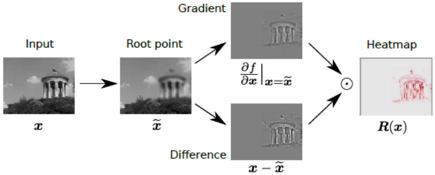
Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

Taylor Decomposition

- Taylor expansion of a function $f(x)$ at point a:
  $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$
- $\sum_j R_j = \left(\left. \frac{\partial(\sum_j R_j)}{\partial\{x_i\}} \right|_{\partial\{\tilde{x}_i\}} \right)^T (\{x_i\} - \{\tilde{x}_i\}) + \epsilon =$
  $\sum_i \sum_j \left. \frac{\partial R_j}{\partial x_i} \right|_{\partial\{\tilde{x}_i\}} (x_i - \tilde{x}_i) + \epsilon$

Pixel-wise decomposition of a function

- Goal: redistribute the neural network output onto the input variables; the relevance R_j to lower-level relevances $\{R_i\}$



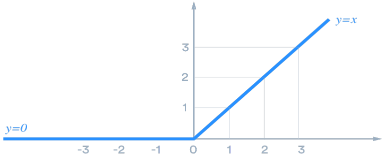
Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

How to find the neighbouring point $\tilde{x}$?

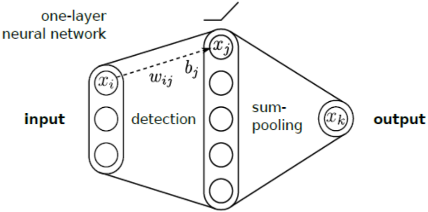
- Find a neighbouring point $\tilde{\mathrm{x}}$, for which $f(\tilde{\mathrm{x}}) = 0$ (root point)
- A good root point is the one that removes elements of the data point x that cause the $f(\mathrm{x})$ to be positive
  (object detected)
- Similar image, object not recognizable from the classifier - hence the output $f(\tilde{\mathrm{x}}) = 0$

Properties

- Conservation: $\sum_i R_i = \sum_j R_j$
  (i and j are layers)
- Rectified Linear Unit (ReLU):



Example (1/3)



Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

- $x_j = max(0, \sum_i x_i w_{ij} + b_j)$ (ReLU nonlinearity)
- $x_k = \sum_j x_j$ (Sum pooling)

Example (2/3)

R_k of output layer: Total relevance that must be backpropagated:

- $R_k = x_k = \sum_j x_j$

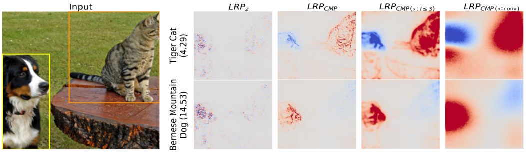
R_j of hidden layer: Taylor decomposition on $\{\tilde{x}_j\} = 0$:

- $R_j = R_k(\tilde{x}) + \frac{\partial R_k}{\partial x_j}\Big|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j) = x_j = max(0, \sum_i x_i w_{ij} + b_j)$
- For which $\tilde{x}$ is $R_k(\tilde{x}) = 0$?
  Since ReLU ensures that $\{\forall j : \tilde{x}_j \geq 0\}$ and
  $\frac{\partial R_k}{\partial x_j} = \frac{\partial \sum_j x_j}{\partial x_j} = 1$

Example (3/3)

R_i of input layer:

- $R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)})$
- $R_i = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j$: Relevance weighted proportionally



Kohlbrenner, Maximilian, et al. "Towards best practice in explaining neural network decisions with LRP." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

LRP task

- Use the equations above to compute numerically the relevance of all layers of the network depicted in the figure.
- Use your weight values (w_{ij}) but think on weighting schemes that are typically used in neural networks.
  See `https://keras.io/initializers/`
- Verify that the conservation and positivity rules properties apply.
- Provide descriptions of the interpretations
- Code: `https://github.com/albermax/innvestigate`

Benefits of LRP

1. More interpretable heatmaps
   Positive and negative relevance
2. Discover artefacts in big datasets
   Actionable insights
3. Principles applied to new NN architectures (GNN)
   Supports Quality Management (QM)

Literature (1/3)

Main LRP paper:

- Montavon, Grégoire, et al. "Explaining nonlinear classification decisions with deep taylor decomposition." Pattern Recognition 65 (2017): 211-222.

Literature (2/3)

Differences with Sensitivity Analysis (SA):

- Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." Digital Signal Processing 73 (2018): 1-15.
- Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

Literature (3/3)

Whole dataset analysis:

- Lapuschkin, Sebastian, et al. "Unmasking clever hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

LRP on LSTMs and Perturbation Analysis:

- Arras, Leila, et al. "Explaining recurrent neural network predictions in sentiment analysis." arXiv preprint arXiv:1706.07206 (2017).

LRP for Pruning:

- Yeom, Seul-Ki, et al. "Pruning by explaining: A novel criterion for deep neural network pruning." Pattern Recognition 115 (2021).

- Questions?
- Dipl. -Ing. Anna Saranti
  anna.saranti@medunigraz.at