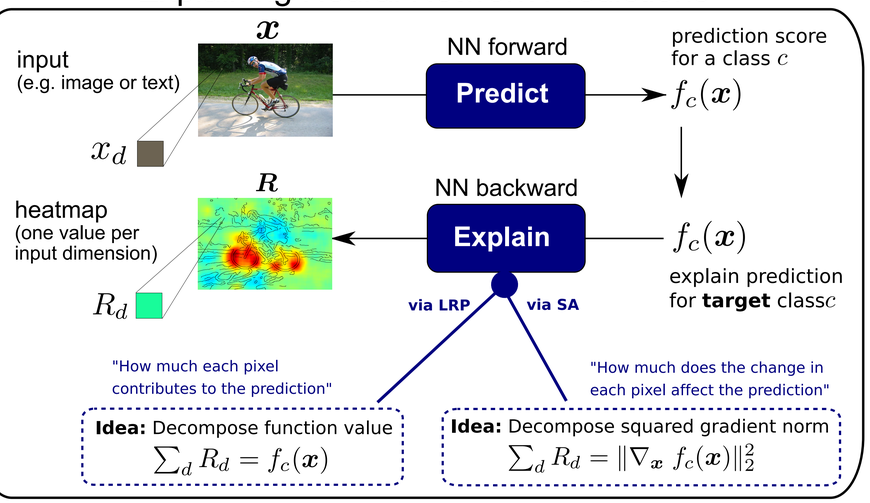
Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek

Fraunhofer Heinrich Hertz Institute

Technische Universität Berlin

Explaining Neural Network Predictions

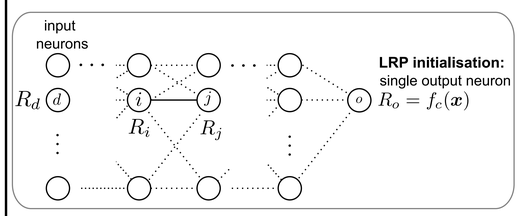


Explaining with LRP

How does LRP work?

Layer-wise Relevance Propagation (LRP)
[Bach et al. 2015, Arras et al. 2017]

LRP: backward relevance redistribution



- Element-Wise Activation $z_j = g(z_i)$

Idea: Redistribute as "Identity" $R_i = R_j$

The activated neuron value z_j is used to compute R_j.

- Weighted Linear Connection $z_j = \sum_i z_i \cdot w_{ij} + b_j$

Idea: Redistribute relevance proportionally to contribution in forward pass

Step 1: compute relevance "messages"

$$R_{i \leftarrow j} = \frac{z_i \cdot w_{ij} + \frac{\epsilon \cdot sign(z_j) + \delta \cdot b_j}{N}}{z_j + \epsilon \cdot sign(z_j)} \cdot R_j$$

with ϵ : stabilizer (small positive number)
N : number of incoming neurons
δ : bias redistribution (1.0 exact relevance conservation, 0.0 approximate conservation)

Step 2: sum up incoming messages

$R_i = \sum_j R_{i \leftarrow j}$

- Multiplicative Interaction $z_j = z_g \cdot z_s$

Idea: Redistribute all the relevance to the source

$R_s = R_j \quad R_g = 0$

gate neuron (sigmoid activated $\in [0,1]$) — "decides how much information to retain"

source neuron (tanh activated $\in [-1,1]$) — "contains the information"

The neuron values z_g and z_s are used to compute R_j.

Advantages over SA?

Sensitivity Analysis (SA)
[Dimopoulos et al. 1995, Li et al. 2016]

SA: squared partial derivative

$$R_d = \left(\frac{\partial f_c}{\partial x_d}(\boldsymbol{x})\right)^2$$

obtained by standard gradient backpropagation

LRP vs. SA:

- LRP relevances are signed, while SA relevances are positive (i.e., SA does not distinguish between positive and negative evidence).
- LRP resolves the classification decision on the current input, while SA reveals sensitivity of classifier to small changes in the input values (i.e., SA does not explain the prediction $f_c(\boldsymbol{x})$).

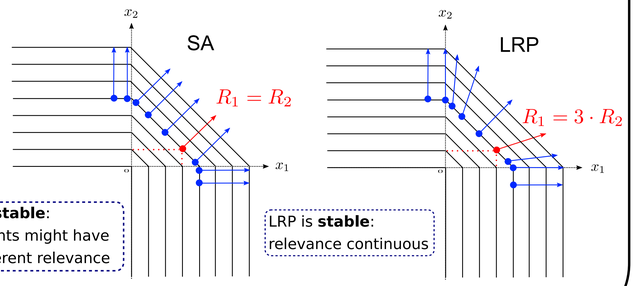
Relevance Aggregation: **for both LRP and SA by summation**

i.e. Word-Level Relevance Value:

$R(\text{word}) = \sum_{d \in \text{word embedding}} R_d$

Intuition -Toy Example

two ReLU neurons: $f(\boldsymbol{x}) = \max(0, x_1) + \max(0, x_2)$



SA is **unstable**: close points might have very different relevance

LRP is **stable**: relevance continuous

Quantitative validation of results

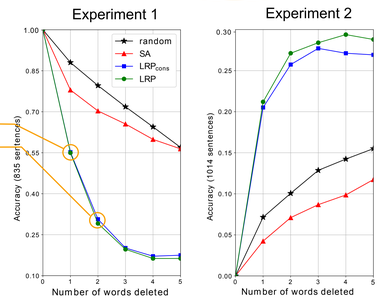
For these experiments:
- use as target class the **true** sentence class
- consider only sentences with length >= 10 words

deletion = word-embedding set to zero in the input sentence

Validation Setup: **delete up to 5 words per sentence accordingly to their relevance and track the impact on the classification performance**

e.g. word deleting by decreasing LRP relevance:

original sentence: the cat sat on the mat
1 deleted word: the ◯ sat on the mat
2 deleted words: the ◯ ◯ on the mat
3 deleted words: ◯ ◯ ◯ on the mat



Conclusion: most pertinent impact obtained by LRP

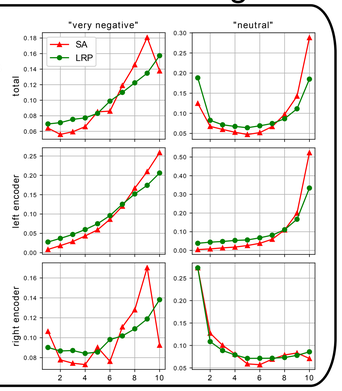
LRP is most appropriate to identify words speaking **for** or **against** the classifier's decision

start with:	correctly classified	falsely classified
deletion order:	decreasing relevance	increasing relevance
best explanation:	lowest curve	highest curve

Relevance distr. over sentence length

For these experiments:
- use as target class **one of the classes**
- consider all sentences with length >= 19 words
- use total rel. or only rel. from left/right encoder

Relevance Statistic: **divide sentence length into 10 intervals and sum up absolute word-level relevances per interval, then normalize to one**



Observation: relevance increases over sentence length for all classes besides for the class "neutral"

for longer sentences strong sentiments tend to appear at the sentence end

Use Case: Recurrent NN Model & Task

Contribution: Extend LRP to recurrent nets and compare to SA.

Model: word-based bidirectional LSTM
word embeddings of dimension 60, one hidden layer of size 60
takes as input a sequence of word embeddings $(x_1, x_2, ..., x_T)$
[we employ the model released by Li et al. 2016]

Task: five-class sentiment prediction
very negative, negative, neutral, positive, very positive
model trained on phrases and sentences from the Stanford Sentiment Treebank [dataset released by Socher et al. 2013]

recurrence of the form:

$$i_t = \text{sigm}\left(W_i\, h_{t-1} + U_i\, x_t + b_i\right)$$
$$f_t = \text{sigm}\left(W_f\, h_{t-1} + U_f\, x_t + b_f\right)$$
$$o_t = \text{sigm}\left(W_o\, h_{t-1} + U_o\, x_t + b_o\right)$$
$$g_t = \tanh\left(W_g\, h_{t-1} + U_g\, x_t + b_g\right)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$

[Hochreiter&Schmidhuber 1997, Gers et al. 2000]

Decomposing Sentiment onto Words

For these exemplary sentence visualizations:
- we use as the target class the **true** sentence class
- we visualize word-level relevance values

Notation: -- very negative, - negative, ++ very positive, + positive

Sentence Heatmap: **map positive relevance to red, negative to blue, and normalize color opacity to extremal relevance per sentence**

LRP

true	predicted	N°	
		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 .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork clichés ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
	++	19.	a worthy entry into a very difficult genre .
++		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

SA

true	predicted	N°	
		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 .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork clichés ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
	++	19.	a worthy entry into a very difficult genre .
++		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Further information and code www.heatmapping.org

Previous Work
[1] Bach et al. **On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation** PLoS ONE 2015
[2] Samek et al. **Evaluating the Visualization of what a Deep Neural Network has learned** IEEE TNNLS 2016
[3] Montavon et al. **Explaining NonLinear Classification Decisions with Deep Taylor Decomposition** Pattern Recognition 2017
[4] Arras et al. **"What is relevant in a text document?": An interpretable machine learning approach** PLoS ONE 2017
[5] Montavon et al. **Methods for Interpreting and Understanding Deep Neural Networks** arXiv 2017