Interpretable Deep Neural Networks for Single-Trial EEG Classification

Irene Sturm[a], Sebastian Lapuschkin[b], Wojciech Samek[b], Klaus-Robert Müller[a,c]

[a] *Machine Learning Group, Berlin Institute of Technology, Berlin, Germany*
[b] *Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany*
[c] *Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea*

Abstract

Background: In cognitive neuroscience the potential of Deep Neural Networks (DNNs) for solving complex classification tasks is yet to be fully exploited. The most limiting factor is that DNNs as notorious 'black boxes' do not provide insight into neurophysiological phenomena underlying a decision. Layer-wise Relevance Propagation (LRP) has been introduced as a novel method to explain individual network decisions.
New Method: We propose the application of DNNs with LRP for the first time for EEG data analysis. Through LRP the single-trial DNN decisions are transformed into heatmaps indicating each data point's relevance for the outcome of the decision.
Results: DNN achieves classification accuracies comparable to those of CSP-LDA. In subjects with low performance subject-to-subject transfer of trained DNNs can improve the results. The single-trial LRP heatmaps reveal neurophysiologically plausible patterns, resembling CSP-derived scalp maps. Critically, while CSP patterns represent class-wise aggregated information, LRP heatmaps pinpoint neural patterns to single time points in single trials.
Comparison with Existing Method(s): We compare the classification performance of DNNs to that of linear CSP-LDA on two data sets related to motor-imaginery BCI.
Conclusion: We have demonstrated that DNN is a powerful non-linear tool for EEG analysis. With LRP a new quality of high-resolution assessment of neural activity can be reached. LRP is a potential remedy for the lack of interpretability of DNNs that has limited their utility in neuroscientific applications. The extreme specificity of the LRP-derived heatmaps opens up new avenues for investigating neural activity underlying complex perception or decision-related processes.

Keywords: Brain-Computer Interfacing, Neural Networks, Interpretability

1. Introduction

Deep Neural Networks (DNNs) are powerful methods for solving complex classification tasks in fields such as computer vision [1], natural language processing [2], video analysis [3] and physics [4]. Although researchers have recently started introducing this promising technology into the domain of cognitive neuroscience [5] and Brain-Computer Interfacing (BCI) [6, 7], most of the current techniques in these fields are still based on linear methods [8, 9]. A limiting factor for the applicability of DNN in these fields is the notion of a DNN as a *black box*. In the domain of cognitive neuroscience this is a particular drawback because obtaining neurophysiological insights is of utmost importance beyond the classification performance of a system.

Recently, the interpretability aspect of Deep Neural Networks has been addressed by the Layer-wise Relevance Propagation (LRP) [10] method. LRP explains individual classification decisions of a DNN by decomposing its output in terms of input variables. It is a principled method which has a close relation to Taylor decomposition [11] and is applicable to arbitrary DNN architectures. From a practitioner's perspective LRP adds a new dimension to the application of DNNs (e.g., in computer vision [12, 13]) by making the prediction transparent. Within the scope of cognitive neuroscience this means that DNN with LRP, may provide not only a highly effective (non-linear) classification technique that is suitable for complex high-dimensional data, but also yield detailed single-trial accounts of the distribution of decision-relevant information, a feature that is lacking in commonly applied DNN techniques and also in other state-of-the-art methods (such as those discussed below).

Here we propose using DNN with LRP for the first time for EEG analysis. For that we train a DNN to solve a classification task related to motor-imaginery BCI. On two example data sets we compare the classification performance of DNN to that of CSP-LDA, a standard technique [9]. We then apply LRP to produce *heatmaps* that indicate the relevance of each data point of a spatio-temporal EEG epoch for the classifier's decision in single trial. We present several examples of such heatmaps and demonstrate their neurophysiological plausibility. Critically, we point out that the spatio-temporal heatmaps represent a new quality of explanatory resolution that allows to explain why the classifier reaches a certain decision in a single instance. Note that such information can not be derived from CSP-LDA. Finally, we provide

a range of future applications of this technique in neuroscience. We discuss why equipping the extremely powerful non-linear technology of DNN with the diagnostic power of LRP may contribute to extending the scope of DNN techniques.

2. A Deep Neural Network for EEG Classification

2.1. Model Details

The network applied here consists of two linear sum-pooling layers with bias-inputs, followed by an activation or normalization step each. The first linear layer accepts an input of the dimensionality number of time points in epoch $\times$ number of EEG channels (for subjects aa, ..., ay 301 time point $\times$ 118 channels, for subjects od, ..., obx, recorded in a different study with a different setup, 301 time point $\times$ 58 channels) vectorized to a 33518 (od-obx: 17458) dimensional input vector and produces a 500-dimensional tanh-activated output vector. The next layer reduces the 500-dimensional space to a 2-dimensional output space followed by a softmax layer for activation in order to produce output probabilities for each class. The network was trained using a standard error back-propagation algorithm [14] using batches of 5 randomly drawn training samples. The prediction accuracies for models trained for 3000 iterations are reported in Table 1.

2.2. Explanation of classifier decisions with LRP

At prediction time, the DNN assigns a classification score $f(\mathbf{x})$ to every input data sample $\mathbf{x} = [x_1 \dots x_N]$ via a forward pass. Typically, function f consists of a sequence of layers of computation

$$z_{ij} = x_i w_{ij} \ ; \ z_j = \sum_i z_{ij} + b_j \ ; \ x_j = g(z_j) \tag{1}$$

where x_i is the layer input, x_j is its output, w_{ij} are the model parameters and $g(\cdot)$ realizes a mapping and/or pooling function. Because of this nested non-linear structure it is not obvious which input dimensions are mainly responsible for a given prediction.

Layer-wise Relevance Propagation decomposes the classifier output $f(\mathbf{x})$ in terms of relevances r_i attributing to each input component x_i its *share* with which it contributes to the classification decision

$$f(\mathbf{x}) = \sum_i r_i \tag{2}$$

By providing signed explanations it distinguishes between positive evidence ($r_i > 0$), supporting the classification decision, and negative evidence ($r_i < 0$), speaking against the prediction.

Mathematically, LRP performs a backward pass of the final score $f(\mathbf{x})$ through the neural network until the input layer. At each layer it attributes shares of upper layer relevances $r_j^{(l+1)}$ to all components i of the adjacent lower layer l, such that each component of l receives a relevance score $r_i^{(l)}$ proportional to its contribution to the output values of layer $l+1$ during the forward pass. Starting with $f(\mathbf{x})$ as the initial top level relevance, the local decomposition rule

$$r_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} r_j^{(l+1)} \tag{3}$$

is applied in a layer-by-layer manner. Note that this rule fulfills an important property, namely layer-wise relevance conservation ($\sum_i r_i^{(l)} = \sum_j r_j^{(l+1)}$), which ensures that the network's output $f(\mathbf{x})$ is fully redistributed to the input domain (see Eq. 2). In other words no relevance is lost and no additional relevance is generated.

For a more theoretical view on LRP we refer the reader to [11], where the authors show a close connection between LRP and a deep Taylor decomposition. An implementation of LRP can be found in [15] and downloaded from `www.heatmapping.org`.

3. Evaluation

3.1. Experimental Setup and Preprocessing

In order to gather a broader experience with DNN-LRP for EEG, the application of DNN with LRP on EEG data was demonstrated on two different data sets: (1) on dataset IVa from BCI competition III (cued motor imagery data with classes right hand vs. foot from 5 subjects [16]) and (2) on a subset of 5 subjects from [17] where subjects had to perform left and right hand motor imaginery while dealing with different types of distractions. Here, we only analyzed data obtained in the condition 'no distraction', a standard motor imaginery BCI setting. As in the competition, we only analyze classes right hand vs. right foot and we use the competition's partition into test and training data set for dataset IVa (subjects aa, al, av, aw, ay, for details see Table 1). For the other data set (subjects od, njy, njk, nko, obx) a leave-on-out

cross-validation was performed. In addition to the single-subject analysis, for both data sets the potential of DNN for subject-to-subject transfer was evaluated: for each subject a DNN was trained on all available data of the other four subjects and evaluated on its own test data. This was was done in a sequential fashion, so that the network was once initialized and then trained on the data of each of the four subjects successively. The entire process of training and testing was repeated five times for different orders of the four subjects and the classification performance on the test data was averaged.

All data sets were downsampled to 100Hz and bandpass filtered in the range of 9-13 Hz. The CSP algorithm, the first classification technique we applied, was performed on a [1000 4000] ms epoch after the cue and 3 pairs of spatial filters were selected. On the extracted features a regularized LDA classifier with analytically determined shrinkage parameter [18] was trained. For training and evaluating the second classification techique, the DNN, the envelope of each epoch (9-13 Hz bandpass-filtered data, epochs of [1000 4000] ms after cue) was calculated and an epochwise baseline of [0 300] ms before the cue was subtracted. Each epoch's spatio-temporal features (301 time points $\times$ 118 channels for aa-ay, 301 time point $\times$ 58 channels for subject od-obx) were vectorized into one vector with 33518 (17458) dimensions. Relevance maps were calculated for each trial from the two-valued DNN output according to Equation 3, yielding one relevance vector of the same dimensionality as the spatio-temporal features of the epoch.

3.2. Results

Classification results for the different methods are summarized in Table 1. Overall, classification performance of DNN is lower than that of CSP-LDA. Subjects ay and njy, the subjects with the lowest performance, represent an exception: here DNN effects an increase in classification accuracy. The performance of inter-subject DNN is inferior to that of single-subject DNN in 6/10 subjects. In the remaining four subjects inter-subject DNN effects a substantial increase in classification accuracy.

Fig. 1 (a) gives an example of relevance maps obtained with LRP for two single trials of subject od. The matrices depict examples of the output of the LRP analysis step that transformes a classifier's decision into a relevance value for each EEG channel at each time point of the epoch. The scalp maps at the bottom show the information in one column (representing a single time point in a single trial) and reveal typical lateralized motor activation patterns. The average of the spatio-temporal relevance matrix across the

Table 1: Classification accuracies for CSP-LDA, DNN and inter-subject DNN. Dataset BCI competition III IVa: aa, al, av, aw, ay. Dataset from [17]: od, njy, njz, nko, obx

subject	number of samples		class. accuray in %		
	train	test	CSP/LDA	DNN	inter-subj. DNN
aa	168	112	66	62	56
al	224	56	100	93	83
av	84	196	70	66	64
aw	56	124	99	77	71
ay	28	252	55	60	73
od	71	1	96	94	86
njy	71	1	65	69	62
njz	71	1	93	86	91
nko	71	1	81	57	68
obx	71	1	97	85	100

entire epoch (top) reveals similar scalp patterns. The average of all time-averaged relevance maps of one class (Fig. 1 (b), bottom) is highly similar in topographical distribution to the patterns of the first pair of filters obtained in the CSP analysis. Note that the LRP-based relevance maps differ from CSP patterns where the *absolute* magnitude of a weight determines its relevance and its sign the polarity. In LRP-derived heatmaps positive and negative values refer to the relevance and non-relevance with respect to the specific decision of the DNN. For instance, in a trial assigned to class 'right hand' with high confidence positive values may be understood as speaking *for* class 'right hand' membership and negative values as speaking *against* class 'right hand' membership. Fig. 1 (c) shows examples of time-averaged relevance maps for a selection of correctly/incorrectly classified trials. In those trials that were classified correctly and with high confidence (classifier output 0 or 1), relevant information is confined to small regions with neurophysiologically highly plausible distribution. In incorrectly or with less confidence classified trials influences outside the sensorimotor areas seem to have influenced the network's decision. These are located in occipital and frontal regions and may indicate the influence of visual activity and of eye movements.

4. Discussion

We have provided the first application of DNN with LRP on EEG data. In terms of classification performance, our relatively simple DNN network does not outperform the benchmark methodology of CSP-LDA. However, we provide some examples that training a network successively on several other subjects is advantageous. For instance, this substantially increased classification accuracy in a subject with particularly low accuracy. However, further investigations are required for reliable subject-to-subject transfer of learned neural representations, and, ultimately, for the advancement of subject-independent zero training strategies in BCI (e.g. [19]) using DNN technology.

The most important and novel contribution in this work is the application of LRP. We have demonstrated that LRP produces neurophysiologically highly plausible explanations of how a DNN reaches a decision. More specifically, LRP produced textbook-like motor imaginery patterns in single instants of single trials. These represent accounts of neural activity at an unprecedented level of specificity and detail. In contrast, CSP-LDA (and also other methods) only allow to examine discriminating information at the level of the whole ensemble of samples of one class. In a direct application in the BCI context LRP helped to diagnose influences that led to low-confidence or erroneous decisions of the network.

Outside BCI, DNN with LRP may add a new dimension of explanation in any setting where detailed single-trial information is valued. In clinical applications it may represent a sensitive tool for neurophysiological interpretation of anomalies or differences between populations. Here, the opportunity to integrate prior knowledge about clinical populations through inter-subject DNN analyses may be a further advantage. In contexts where the trial-to-trial variability of EEG is not viewed as a notorious obstacle for analysis, but as a source of information, LRP can contribute high-resolving spatio-temporal representations of underlying neurophysiological phenomena. In particular, this might be interesting for linking brain indices to single instances of behavioral measures [20], for understanding subtle aspects of complex perceptual processes, such as perception of video or audio quality [21, 22], and of dynamic cognitive processes, such as decision making [23]. Finally, a trained network produces relevance maps for any (even artificially generated) DNN decision. This means that LRP can derive a representation of what a network has learned, e.g., by performing LRP on a 'ideal' speci-

men of a given class or even by systematically exploring the space of possible decisions. This might be an interesting alternative to network visualization techniques [24].

5. Conclusion

In summary, we have provided a showcase of how LRP can add an explanatory layer to the highly effective technique of DNN in the EEG/BCI domain. Our results show that LRP provides highly detailed accounts of relevant information in high-dimensional EEG data that may be useful in analysis scenarios where single trials need to be considered individually.

Acknowledgement
This work was supported by the Brain Korea 21 Plus Program and by the Deutsche Forschungsgemeinschaft (DFG). This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein. Correspondence to WS and KRM.

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Adv. in NIPS, 2012, pp. 1106–1114.

[2] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proc. of EMNLP, 2013, pp. 1631–1642.

[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proc. of CVPR, 2011, pp. 3361–3368.

[4] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, New Journal of Physics 15 (9) (2013) 095003.

[5] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, V. D. Calhoun, Deep learning for neuroimaging: a validation study, Front. Neurosci. 8 (229) (2014) doi:10.3389/fnins.2014.00229.

[6] A. Yuksel, T. Olmez, A neural network-based optimal spatial filter design method for motor imagery classification, PLOS ONE 10 (5) (2015) e0125039.

[7] H. Yang, S. Sakhavi, K. K. Ang, C. Guan, On the use of convolutional neural networks and augmented csp features for multi-class motor imagery of eeg signals classification, in: Proc. of IEEE EMBC, 2015, pp. 2620–2623.

[8] L. C. Parra, C. D. Spence, A. D. Gerson, P. Sajda, Recipes for the linear analysis of eeg, NeuroImage 28 (2) (2005) 326–341.

[9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.-R. Müller, Optimizing spatial filters for robust EEG single-trial analysis, IEEE Signal Proc. Mag. 25 (1) (2008) 41–56.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (7) (2015) e0130140.

[11] G. Montavon, S. Bach, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, arXiv preprint (2015) CoRR abs/1512.02479.

[12] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek, Analyzing classifiers: Fisher vectors and deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2912–2920.

[13] W. Samek, A. Binder, G. Montavon, S. Bach, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, arXiv preprint (2015) CoRR abs/1509.06321.

[14] Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, Efficient backprop, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 9–48.

[15] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek, The layer-wise relevance propagation toolbox for artificial neural networks, Journal of Machine Learning Research 17 (114) (2016) 1–5.

[16] B. Blankertz, K.-R. Müller, D. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, N. Birbaumer, The BCI competition III: Validating alternative approachs to actual BCI problems, IEEE Trans. Neur. Sys. Reh. 14 (2) (2006) 153–159.

[17] S. Brandl, J. Höhne, K.-R. Müller, W. Samek, Bringing BCI into everyday life: Motor imagery in a pseudo realistic environment, in: Proc. IEEE Conf. on Neural Eng. (NER), 2015, pp. 224–227.

[18] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, K.-R. Müller, Single-trial analysis and classification of ERP components – a tutorial, NeuroImage 56 (2011) 814–825.

[19] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, C. Grozea, Subject-independent mental state classification in single trials, Neural Networks 22 (9) (2009) 1305–1312.

[20] A. Delorme, M. Miyakoshi, T.-P. Jung, S. Makeig, Grand average erp-image plotting and statistics: A method for comparing variability in event-related single-trial eeg activities across subjects and conditions, J. Neurosci. Meth. 250 (2014) 3–6.

[21] A. K. Porbadnigk, M. S. Treder, B. Blankertz, J.-N. Antons, R. Schleicher, S. Möller, G. Curio, K.-R. Müller, Single-trial analysis of the neural correlates of speech quality perception, J. Neural Eg. 10 (5) (2013) 056003.

[22] L. Acqualagna, S. Bosse, A. K. Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand, B. Blankertz, EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs), J. Neural Eng. 12 (2) (2015) 026012.

[23] A. Tzovara, M. M. Murray, N. Bourdaud, R. Chavarriaga, J. d. R. Millán, M. De Lucia, The timing of exploratory decision-making revealed by single-trial topographic EEG analyses, NeuroImage 60 (4) (2012) 1959–1969.

[24] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint (2015) CoRR abs/1506.06579.

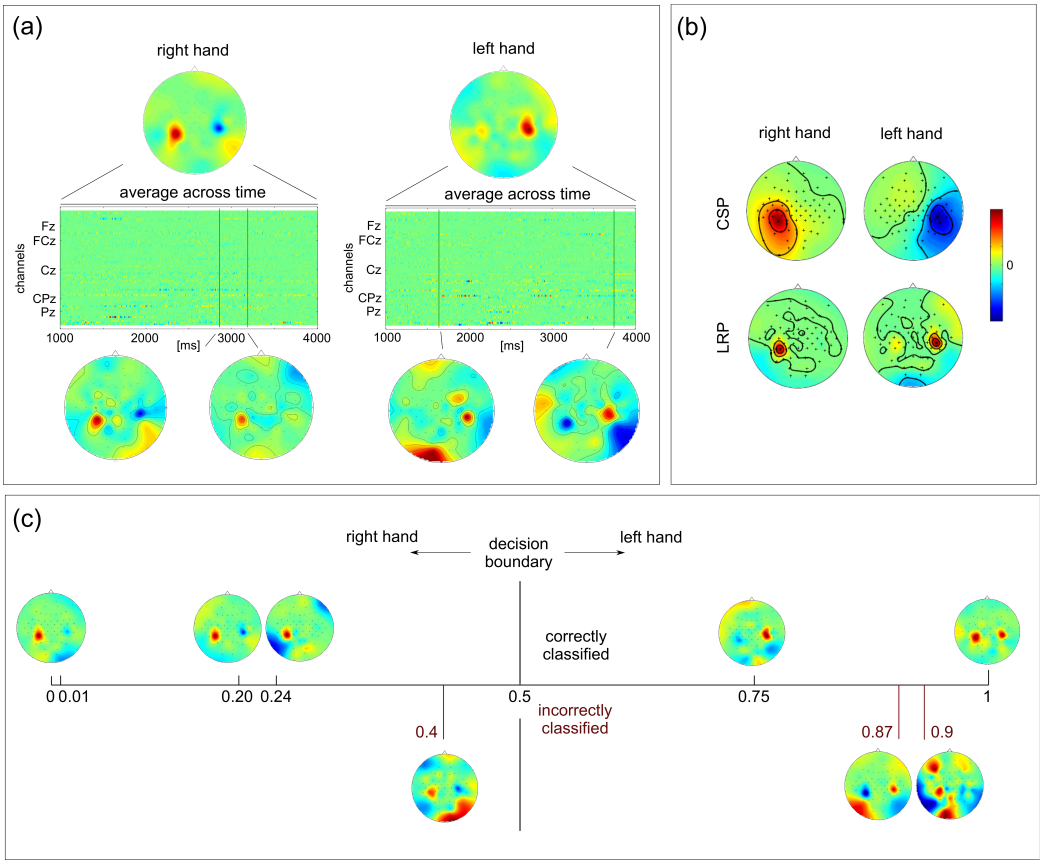

Figure 1: (a) Example of LRP relevance maps for a single trial of each class of subject od. The matrices indicate the relevance of each time point (ordinate) and EEG channel (abscissa). Below the matrix the relevance information for two single time points (indicated by the green line) is plotted as a scalp topography. The scalp plot above the matrices depict the average relevance map across the time window of the entire epoch. (b) CSP patterns (top) and relevance maps (bottom) for subject od. Here, the relevance maps represent the average of all trials of one class, additionally averaged across the time window of the entire epoch. The CSP pattern represents the whole ensemble of samples of one class. (c) Relevance maps and DNN output. Examples of (time-averaged) relevance maps for single trials with different classification outcomes. Values above 0.5 indicate a decision for class 'left hand', values below 0.5 for class 'right hand'. Values close to the extrema 0 and 1 indicate high confidence of the decision. Correctly classified samples appear above the axis, incorrectly classified samples below.