

COGXAI ANNALYZER: COGNITIVE NEUROSCIENCE INSPIRED TECHNIQUES FOR EXPLAINABLE AI

Maral Ebrahimzadeh*, Valerie Krug* & Sebastian Stober

Artificial Intelligence Lab, Otto von Guericke University Magdeburg, Germany
maral.ebrahimzadeh@ovgu.de akrug@ovgu.de stober@ovgu.de

ABSTRACT

Over the past few years, deep Artificial Neural Networks (ANNs) have become more popular due to their great success in various tasks. However, their improvements made them more capable but less interpretable. To overcome this issue, some introspection techniques have been proposed. According to the fact that ANNs are inspired by human brains, we adapt techniques from cognitive neuroscience to easier interpret them. Our approach first computes characteristic network responses for groups of input examples, for example, relating to a specific error. We then use these to compare network responses between different groups. To this end, we compute representational similarity and we visualize the activations as topographic activation maps. In this work, we present a graphical user interface called CogXAI ANNalyzer to easily apply our techniques to trained ANNs and to interpret their results. Further, we demonstrate our tool using an audio ANN for speech recognition.

1. INTRODUCTION

Recent improvements in the field of Artificial Neural Networks (ANNs) made them popular, but hard to interpret. This is attributed to the complex architectures with more layers and neurons. ANNs are applied in a wide variety of domains, therefore, the decisions made by the network need to be trustworthy for various stakeholders. Explainable artificial intelligence (XAI) tackles this issue by developing techniques for interpreting ANN decisions [1, 2].

Considering the similarity between ANN and the brain, we follow the idea of analyzing and visualizing ANNs inspired by research that has been done in cognitive neuroscience. In this work, we introduce the CogXAI ANNalyzer toolbox, which we designed to apply neuroscience-inspired XAI techniques that we developed [3, 4] and to make it possible for interested users to analyze their own trained model. The CogXAI ANNalyzer toolbox currently supports fully-connected and convolutional neural layers

and as an example, we show the output of our tool for an exemplary speech recognition model.

2. RELATED WORK

Model introspection refers to analyzing and visualizing the internals of an ANN. Due to the fact that ANNs have a black-box nature, research in this field has proposed several methods in recent years [5, 6]. Yet, these methods are mostly suitable for computer vision tasks as the data is visually interpretable for a human [1, 2, 7].

Feature Visualization is a common technique to investigate model internals in the input space. A direct visualization of weights or activations of the network is mostly not interpretable. Therefore, feature visualization optimizes an input such that it maximally activates some neuron or feature map of the network. The obtained input pattern is then assumed to be the feature that is detected by the respective feature map or neuron [5, 8, 9].

Saliency Maps highlight the relevant input regions for a prediction made by a neural network as a heat map superimposed on the input [10]. There are various techniques for quantifying the relevance. For example, relevance is computed as the gradient of the output with respect to the input values [8], by propagating the prediction backward in a network using a decomposition approach [7] or by combining gradient and activation information [2].

Analyzing Dataset Representations can help to generally understand an ANN by investigating how the model represents different classes using many data examples. Common approaches include using linear classifiers on hidden layers [11, 12] and statistical analyses [13–15]. Further, there also are several graphical interfaces to investigate representations and learned features [16–19].

3. METHOD

Our analysis approach uses Neuron Activation Profiles (NAPs) as a characterization of network activity. The first step to obtain NAPs is to choose a dataset, a model and layers and compute the layer activations and (sensitivity-based [8]) saliency maps. We use the gradients to align the inputs or activations such that the most prediction-relevant position is located at their center position. This allows us to average the activations over groups of inputs to obtain group-characteristic activations. For better comparability of the groups, we normalize the averages by subtracting



© M. Ebrahimzadeh, V. Krug, and S. Stober. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** M. Ebrahimzadeh, V. Krug, and S. Stober, “CogXAI ANNalyzer: Cognitive Neuroscience Inspired Techniques for eXplainable AI”, in *Extended Abstracts for the Late-Breaking Demo Session of the 23rd Int. Society for Music Information Retrieval Conf.*, Bengaluru, India, 2022.

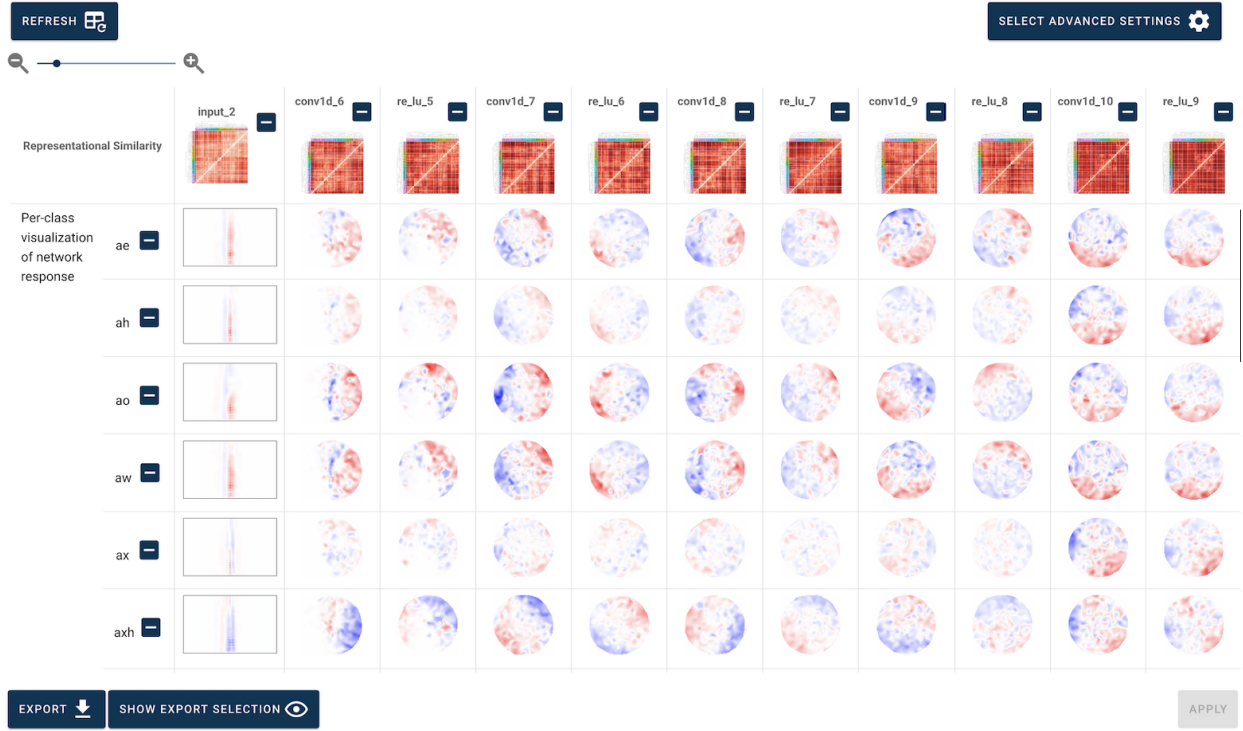


Figure 1. An overview of the visualization result in the CogXAI ANNalyzer toolbox

the average activation across all examples. Optionally, at the end, we mask prediction-irrelevant information by multiplying each group-average with the $[0,1]$ -scaled averaged saliency maps. This way, we obtain NAPs for all groups and selected layers. Details on how to compute NAPs can be found in our previous publication [3].

For visualizing NAPs, we adapt how brain activity is commonly shown as topographic activation maps in neuroscience. To this end, we perform a dimensionality reduction of the NAPs, such that the neurons or feature maps in a layer are layouted by their activation similarity. Using this layout, we color the respective position in the two-dimensional projection by the NAP-value and interpolate empty spaces. This approach of visualizing NAPs (or activations) as topographic maps is described in more detail in our previous work [4].

The CogXAI ANNalyzer tool provides an interface for performing NAP analysis and visualizing the results as topographic activation maps.

4. USING COGXAI ANNALYZER

We developed a web-based application to compute and visualize NAPs for any dataset and trained model that are compatible. Currently, it supports models with convolutional and fully-connected layers that are implemented in the TensorFlow¹ framework. We implemented a REST API using the Flask framework in Python, used Vue.js for user interface and MongoDB for data storage.

We use the TIMIT dataset [20], which is a corpus of speech recordings of 630 speakers with 8 dialects of Amer-

ican English. Each of the 630 speakers recorded 10 out of 2345 unique sentences with a non-uniform distribution of sentences. All data are preprocessed to mel-scaled log power spectrograms using a FFT window size of 512 (32 ms) and hop size of 128 (8 ms) at 16 kHz and projecting the FFT bins to 128 mel-frequency bins.

As model, we use a variation of Wav2Letter (W2L). W2L is a 1D-convolutional architecture which includes 11 layers and is trained by using the Connectionist Temporal Classification (CTC) loss for letter prediction. Our W2L-based model predicts phonemes and uses 62 output units in the output layer. The phoneme prediction task can require fewer layer as it is an easier task than letter prediction. Therefore, we use only five convolutional layers and the output layer for phoneme prediction.

With our toolbox, we compute NAPs using the phoneme targets as groups, in all convolutional layers before and after applying the activation function, respectively. Then, we visualize the result with similarity representation and as topographic maps. Figure 1 shows the results in the visualization view of the CogXAI ANNalyzer. For each phoneme (row) and each layer (column), this view shows the NAP values as topographic maps and a clustermap as a visual representation of the phoneme similarities according to the NAPs. In the input layer (first column), NAPs indicate important characteristic of input data for each phoneme. As NAPs values are normalized, they shows whether the activation is higher (red) or lower (blue) in comparison to the global average. Furthermore, according to the topographic maps visualization of NAP values, we can see which phonemes share the same active regions.

¹ <https://www.tensorflow.org/>

5. ACKNOWLEDGMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF) as part of the project “CogXAI – Cognitive Neuroscience inspired techniques for eXplainable AI”.

The authors would like to thank Akshaya Bindu Gowri for her support in user interface implementation.

6. REFERENCES

- [1] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.
- [2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [3] A. Krug, M. Ebrahimzadeh, J. Alemann, J. Johannsmeier, and S. Stober, “Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles,” vol. 10, no. 11. Multidisciplinary Digital Publishing Institute, 2021, p. 1350.
- [4] A. Krug, R. K. Ratul, and S. Stober, “Visualizing deep neural networks with topographic activation maps,” *arXiv preprint arXiv:2204.03528*, 2022.
- [5] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, “Understanding neural networks through deep visualization,” *arXiv preprint arXiv:1506.06579*, 2015.
- [6] S. Gautam, M. M.-C. Höhne, S. Hansen, R. Jenssen, and M. Kampffmeyer, “This looks more like that: Enhancing self-explaining models by prototypical relevance propagation,” *Pattern Recognition*, p. 109172, 2022.
- [7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
- [8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [9] A. Mordvintsev, C. Olah, and M. Tyka, “Inceptionism: Going deeper into neural networks,” *Google Research Blog. Retrieved June*, vol. 20, no. 14, p. 5, 2015.
- [10] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [11] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” in *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2017.
- [12] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *International Conference on Machine Learning (ICML)*, 2018, pp. 2668–2677.
- [13] J. Fiocco, S. Choudhary, and C. Rose, “Deep neural model inspection and comparison via functional neuron pathways,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5754–5764.
- [14] A. S. Morcos, M. Raghu, and S. Bengio, “Insights on representational similarity in neural networks with canonical correlation,” *arXiv preprint arXiv:1806.05759*, 2018.
- [15] T. Nagamine, M. L. Seltzer, and N. Mesgarani, “Exploring how deep neural networks form phonemic categories,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [16] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah, “Activation atlas,” *Distill*, vol. 4, no. 3, p. e15, 2019.
- [17] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, “S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations,” *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1096–1106, 2019.
- [18] H. Park, N. Das, R. Duggal, A. P. Wright, O. Shaikh, F. Hohman, and D. H. P. Chau, “Neurocartography: Scalable automatic visual summarization of concepts in deep neural networks,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 813–823, 2021.
- [19] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, “The what-if tool: Interactive probing of machine learning models,” *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.
- [20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.