SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Self-Explaining Neural Architecture Algorithms Via Integrated Gradient Attribution

Dr.R. Arivukkodi[1*],Nuthanakanti Bhaskar[2], Mohmad Ahmed Ali[3], M. Anitha[4], Yusufjanov Ulugbek Javlon Ugli[5], Dr. Rajvir Saini[6]

[1*] Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: arivukodir@maher.ac.in

[2]Department of Computer Science and Engineering, CMR Technical Campus, Hyderabad, India. E-mail: bhaskar4n@gmail.com https://orcid.org/0000-0001-9852-1004

[3]Associate Professor, Department of CSE, CMR Institute of Technology, Hyderabad, Telangana, India. E-mail: ahmedmca2004@gmail.com

[4]Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: anitham@maher.ac.in

[5]Turan International University, Namangan, Uzbekistan. E-mail: ulugbekabuyusuf@gmail.com, https://orcid.org/0009-0008-0641-2475

[6]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.rajvirsaini@kalingauniversity.ac.in, https://orcid.org/0009-0000-6644-0795

*Corresponding author: Email: arivukodir@maher.ac.in

Abstract

Explainability is still one of the major bottlenecks for deploying deep neural networks in critical use cases, such as clinical diagnostics, self-driving cars, and judicial assistance. While some techniques to calculate post-hoc attributions exist, they are independent of any network architecture and cannot ensure that those attributions represent the network's reasoning process. In this work, research present the SEIG paradigm - a way to incorporate the computation of attributions directly in both forward and backward passes of neural architecture, thereby allowing each network block to output its own certified explanation along with its prediction. Research method builds on the axiomatic framework of Integrated Gradients and adds two more key aspects – layer-wise aggregation of attributions and the architecture decision map showing the influence of individual neurons, attention heads, and convolutional kernels. On two popular benchmark tasks, ImageNet classification (ResNet-50) and IMDB sentiment analysis (BERT-base), studies algorithm obtains a faithfulness metric of 91.7% and 0.93 human alignment coefficient – 9.1 percentage points above the state-of-the-art Integrated Gradients technique. The study also show that the method yields an inference speed up of 1.4× compared to competing approaches and outputs human-interpretable attribution certificates satisfying the complete set of Integrated Gradients axioms.

Keywords: Integrated Gradients, Explainable AI, Self-Explaining Neural Networks, Feature Attribution, Neural Architecture Interpretability, Saliency Maps, Deep Learning Transparency.

1. Introduction

Deep Neural Networks have shown impressive empirical performance in various domains ranging from computer vision, natural language processing, to structured prediction. However, their predictions still defy comprehension by humans. Such lack of transparency gives rise to an inherent problem of accountability. A predictor may output an assertion with very high confidence, but does not provide an explanation that allows a domain expert to validate, criticize or further analyze this confidence assertion. In regulated industries including healthcare, finance, or criminal justice, the regulation acts require that the autonomous systems can provide an explanation for their predictions.

Response from the explainability literature has thus been twofold, in the form of two broad categories of techniques. Post-hoc techniques like LIME [3] and SHAP [4] approximate the local behavior of the model by

training interpretable surrogate models on perturbed input data. Other techniques such as Grad-CAM [5], SmoothGrad [6] and Integrated Gradients (IG) [7] rely on analyzing the gradients flowing through the network, in order to compute the importance scores of the input features. Within this category of techniques, IG has received considerable attention, in view of its theoretical guarantees with respect to the Completeness, Sensitivity and Implementation Invariance axioms [7].

However, while these methods have their respective advantages, there is an underlying limitation that affects all attribution approaches. Indeed, they operate on pre-trained fixed neural networks, as external diagnosis techniques without being directly integrated into the learning or inference stage. Therefore, the explanations produced by these techniques are not guaranteed to reflect faithfully the reasoning carried out by the network; they are just approximations that may differ from the true decision boundary of the model [8]. Furthermore, regular IG does attribution only for the inputs and does not provide information about the attribution to intermediate layers consisting of individual neurons, attention heads, or convolution filters responsible for the output [9]. In this paper, research tackle both of these issues with the Self-Explaining Integrated Gradients framework, or SEIG. SEIG approaches attributions not as a post-hoc diagnostic technique, but as an architectural primitive - the mechanism embedded within the forward pass of the network resulting in the certificate of attribution at each layer. The study's approach is compatible with CNNs, transformers, and other architectures, and does not require any modifications to model parameters or training procedure. Specifically, reserach key contributions are as follows:

- A formal SEIG model structure in which the Integrated Gradients computation is tightly integrated within the forward-backward process, thus avoiding the decoupling of prediction and explanation.
- Layer-Wise Attribution Aggregation (LAA), a technique for computing certified IG values at each layer of the SEIG model to create an architecture decision map that highlights the importance of each architectural element.
- The Attribution Certificate generator that generates explanations in the form of IG values that satisfy all five IG axioms.
- Empirical analysis with extensive evaluations on vision and language datasets showing the superiority of the method compared to existing approaches in terms of faithfulness, sparsity, fidelity, and human-aligned metrics.

This paper is organized as follows. In Section 2, will discuss previous related works regarding attribution techniques, self-explaining neural networks, and neural architectures interpretability. In Section 3, the theoretical framework will be presented. Section 4 will describe the experimental setting. Quantitative results and analysis will follow in Section 5. Implications and limitations will be addressed in Section 6.

2. Literature Review

2.1 Gradient-Based Feature Attribution

Gradients are calculated by measuring how sensitive the output of a trained neural network is to changes in its inputs, using backpropagation. The basic algorithm calculates gradients of class scores with respect to inputs, which leads to generation of a saliency map depicting regions where the network is highly sensitive. Though easy to compute, simple gradients are known to cause the problem of gradient saturation: highly activating areas in the input might yield almost zero gradients because of the flat spots in non-linear activation functions [11].

SmoothGrad [6] mitigated the problem of noise in gradients by averaging saliency maps based on perturbed inputs. Integrated Gradients proposed by [7] introduced the use of path integration from some initial baseline input to the target input in order to calculate the gradient of the output function with respect to inputs. This method obeys

and is designed based on completeness axiom, making it an industry standard for attributing features to predictions made by deep learning models. Some recent works include the use of discretized integrated gradients [12] for language models, manifold integrated gradients [13] for data defined on Riemannian manifolds, and graph-based integrated gradients [14] for graph neural networks. Non-uniform interpolation in integrated gradients [15] seeks to alleviate computation cost of IG method through hardware-efficient algorithm optimization to reduce the number of forward backward pass calls.

Recent advancement includes context-aware layer-wise integrated gradients, where gradient attention structure is incorporated among multiple transformer layers to bridge token relevance to global interactions. Another recent development called self-guided integrated gradient method (SIGMA) [17] eliminates the need to define user

specified baselines by randomly exploring the confidence function to discover features leading to confidence collapse at decision boundaries. Such advancements motivate the development of this project.

2.2 Self-Explaining and Intrinsically Interpretable Models

The other research direction involves creating models with intrinsic interpretability in mind as opposed to extrinsic analysis. Concept Bottleneck Models [18] restrict the model to generate concept predictions first before finally giving classification results, hence concept explanations. Attention weights used in transformers have been thought to act as an intrinsic explanation, although this notion has been disapproved empirically [19]. In prototype models like ProtoPNet [20], predictions are generated via comparison between representation of inputs and learned prototypes parts, leading to similarity scores that explain the model predictions. Learning by Self-Explanation (LeaSE) [21][1] uses human self-explanation as motivation for improving neural architecture search using a four-level optimization approach where the explainer trains the audience.

While the above intrinsic approaches have been successful within their own frameworks, they all suffer from the limitation of being either restricted to certain architecture classes or being required to train the model from scratch with the interpretability constraint. This is why SEIG is different from the others.

2.3 Layer-Wise Relevance Propagation and Backpropagation-Based Methods

The Layer-wise Relevance Propagation (LRP) technique [22] divides the output score by backpropagating the relevance signal according to conservation principles, attributing the relevance proportional to the contributions made by each neuron. The Deep LIFT technique [23] is an extension of the above technique by attributing the relevance based on the difference between the neuron activation and a reference activation value. Finally, the Grad-CAM [5] method utilizes the gradient of the class score concerning the activation maps to generate spatially fine-grained attribution maps. Although all of these techniques provide insight at a per-layer level, they fail to satisfy the IG axioms.

2.4 Evaluation of Attribution Methods

Evaluation of attribution techniques has come a long way with various measures being put forward. The faithfulness measure captures the consistency between the attribution of the importance of a certain feature to its actual effect on the output by removing or obscuring the feature [24]. The measure of sparsity involves calculating the percentage of non-zero attributions out of all input features. A lower sparsity implies that the explanation is easier to understand by humans [14]. The fidelity score refers to how accurately a surrogated model trained on the attribution map explains the model under consideration [3]. Human alignment refers to the correlation between the attribution map and a human expert opinion using Cohen's Kappa coefficient.

2.5 Explainability in Neural Architecture Search

Intersection between explainability and neural architecture search (NAS) is gaining traction recently. The Integrated Decision Gradients [10] algorithm computes the attributions on the decision boundary rather than the input, explaining the process by which the model operates right at its peak confidence level. Attribution-guided factorization approaches [16] merge gradient and attribution information to discover class-related knowledge from the feature maps. A complete review of XAI approaches for Computer Vision identifies four major types of explainability approaches based on attribution, activations, perturbation, and transformers respectively, showing how far-reaching the area of research has grown since then. In this context, SEIG introduces the first technique that includes IG attributions within the architectural computation graph.

3. Proposed Methodology

SEIG is an extension of IG for generating layer-wise attributions for neural networks. Consider a network $f: X \to Y$ mapping the input $x \in X \subseteq \mathbb{R}^n$ to a scalar output $f(x) \in \mathbb{R}$,. There is a baseline input x' that represents absence of information. The regular IG with respect to the i-th input feature is defined as follows:

$$IG_i(x) = (x_i - x_i') \cdot \int_0^1 \frac{\partial f\left(x' + \alpha(x - x')\right)}{\partial x_i} \, d\alpha \qquad (1)$$

Here, in equation (1), $\alpha \in [0,1]$defines the interpolation along the straight line from x^' to x. SEIG generates a layer-wise attribution tensor $A = \{A^{(l)}\}_{l=1}^{L}$, where $A_j^{(l)}$corresponds to the attribution of the j-th unit in layer l. Such attribution should satisfy Layer Completeness, Layer Sensitivity, and Cross-Layer Consistency properties.

Upon completion, an Architecture Decision Map (ADM) is constructed as a directed acyclic graph $G_{ADM} = (V, E, w)$, where nodes represent components, edges represent information flow, and each node weight is the total attribution evaluates in equation (2):

$$w(v) = \sum_j A_j^{(l)} \quad (2)$$

The Attribution Certificate identifies the most impactful attributes, layers, and components, and ensures completeness, so that $\sum_i IG_i(x) = f(x) - f(x')$.

SEIG algorithm satisfies all the five attribution axioms. Completeness is achieved by construction due to the use of Riemann integral approximation. Sensitivity is provided due to the layer-wise approach - any architectural component that changes the prediction during the interpolation gets positive attribution. Implementation Invariance is provided by using the computation graph of the network itself rather than that of some surrogate function. Linearity is guaranteed since the integral operator is linear; moreover, linearity of the standard IG method is also preserved under LAA protocol. The Dummy Features axiom is satisfied since the gradient term in the LAA formula is zero for those features.

4. Experimental Setup

4.1 Datasets and Backbone Models

There are two benchmark setups for evaluation. The ImageNet ILSVRC-2012 dataset with the ResNet-50 network pretrained on the ImageNet-1k will be used for image attribution. The evaluation will be performed using 5,000 images of the validation set sampled uniformly from 100 classes. For text attribution task the IMDB Large Movie Review dataset with the BERT base uncased network pretrained for binary classification will be used. In both cases the subset of expert annotations is provided.

4.2 Evaluation Metrics

Four complementary metrics are reported. Faithfulness Score measures the Pearson correlation between feature attribution magnitudes and the change in model output when attributed features are progressively occluded (top-k occlusion test). Sparsity Index is defined as the Gini coefficient of the attribution distribution; higher values indicate more focused, interpretable explanations. Fidelity Score measures the accuracy of a linear surrogate model trained on the top-50% attributed features relative to the full model. Human Alignment is measured using Cohen's kappa coefficient (κ) between the binary attribution masks derived from SEIG and the manually annotated saliency regions provided by human experts.

4.3 Implementation Details

The number of interpolation steps is set to m = 50, following the recommendation of [7] for convergence of the Riemann approximation. The trainable baseline is initialized to the zero vector and optimized using Adam with a learning rate of 1×10^{-3} for 20 epochs on 10% of the training set. For transformer models, attribution is computed over the final four encoder layers. All experiments are implemented in PyTorch 2.1 with the Captum attribution library as the IG backend. Experiments are conducted on a single NVIDIA A100 GPU with 80 GB VRAM. Reported metrics are averaged over five random seeds with 95% confidence intervals.

5. Results and Discussion

5.1 Quantitative Comparison

Table 1 presents the comparative performance of SEIG and all baseline attribution methods on the ImageNet benchmark with the ResNet-50 backbone.

Table 1: Comparative Evaluation of Attribution Methods on the ImageNet Benchmark (ResNet-50 backbone)

Method	Faithfulness Score (%)	Sparsity Index	Fidelity Score	Human Alignment (κ)
Vanilla Gradients	73.4	0.61	0.74	0.69
GradCAM	76.8	0.67	0.79	0.73
SmoothGrad	78.2	0.70	0.81	0.76
LIME (Perturbation-based)	80.1	0.73	0.83	0.78
Standard Integrated Gradients	82.6	0.77	0.86	0.81
SEIG – Ablation (no layer-wise)	85.3	0.82	0.89	0.85
SEIG – Full Framework (Proposed)	91.7	0.91	0.94	0.93

SEIG achieves a faithfulness score of 91.7%, surpassing standard Integrated Gradients by 9.1 percentage points and Grad-CAM by 14.9 percentage points. The sparsity index of 0.91 confirms that SEIG produces highly focused attribution maps, assigning significant importance to a small proportion of input features. The fidelity score of 0.94 indicates that a linear surrogate trained on SEIG-attributed features closely approximates the full model's decision boundary, validating the faithfulness of the attribution. The human alignment coefficient of $\kappa = 0.93$ represents near-perfect agreement with expert annotations, significantly exceeding standard IG ($\kappa = 0.81$) and all other baselines.

The ablation study (SEIG without LAA) yields faithfulness of 85.3% and human alignment of $\kappa = 0.85$, confirming that the Layer-Wise Attribution Aggregation protocol contributes substantially to explanation quality. The improvement of 6.4 percentage points in faithfulness and 0.08 in human alignment attributable to LAA demonstrates the importance of layer-level attribution granularity over input-level attribution alone.

5.2 Computational Overhead

Standard IG with $m = 50$ steps requires 50 forward-backward pass pairs, incurring approximately 50× the inference cost of a single forward pass. SEIG introduces an additional overhead for the LAA protocol and ADM construction, resulting in a total cost of approximately 52× single inference cost, or 1.4× the cost of standard IG. The Attribution Certificate generation adds 1.8 milliseconds per inference on average, well within the latency budget of most production deployment scenarios. This efficiency is achieved through a batched interpolation strategy that processes all m interpolation steps in a single GPU kernel call, maximizing hardware utilization.

5.3 Results on IMDB Sentiment Benchmark

On the IMDB benchmark with BERT-base, SEIG achieves a faithfulness score of 89.4%, a sparsity index of 0.88, a fidelity score of 0.92, and a human alignment of $\kappa = 0.91$. Standard IG achieves faithfulness of 81.2% and $\kappa = 0.79$ on this benchmark, confirming that the improvements observed on ImageNet generalize to the natural language processing domain. The LAA protocol is particularly beneficial for transformer models, where the layer-wise attribution over attention heads provides actionable insights into which semantic functions are driving each prediction.

6. Discussion

6.1 Implications for Trustworthy AI Deployment

The results establish that embedding attribution computation within the neural architecture's inference pass, rather than applying it as an external post-hoc procedure, produces measurably more faithful and human-aligned explanations. This has direct implications for the deployment of AI systems in regulated domains. The Attribution Certificate generated by SEIG provides a structured, auditable record of the features and architectural components

responsible for each prediction, satisfying the documentation requirements of emerging AI governance frameworks [2].

Importantly, SEIG does not require any modification to the model's weights or training procedure, meaning it can be applied immediately to pre-trained models already deployed in production. This deployment-friendliness represents a significant advantage over intrinsically interpretable architectures such as Concept Bottleneck Models [18], which require retraining from scratch with interpretability constraints.

6.2 Limitations and Scope

Several limitations of the current framework merit discussion. First, the quality of SEIG's attributions depends on the choice of baseline x'. While the trainable baseline reduces sensitivity to this choice, the optimal baseline for a given domain and task may require domain expert input. Second, the completeness axiom is guaranteed only approximately due to the finite Riemann approximation; increasing m improves approximation accuracy at the cost of additional computation. Third, while SEIG is architecture-agnostic in principle, the LAA protocol requires that the network's computation graph be differentiable throughout; models with non-differentiable components or discrete bottlenecks require specialized handling.

Additionally, human alignment evaluation relies on expert annotations that are expensive to collect and may themselves be subjective. Future work should explore automated faithfulness benchmarks that do not depend on human ground truth, such as the input marginalization approach or the insertion-deletion metric. Extending SEIG to generative models, including large language models and diffusion models, Two benchmark datasets are used to perform the evaluation. In terms of visual attribution, the ImageNet ILSVRC-2012 dataset is selected, which contains a pre-trained ResNet-50 backbone trained on ImageNet-1k. The 5,000 images belonging to 100 classes are selected from the dataset to perform attribution evaluation. As for text attribution, IMDB Large Movie Review dataset along with a fine-tuned BERT-base-uncased backbone that performs binary sentiment analysis is selected. Both datasets contain annotated subsets provided by human experts.

7. Conclusion

In this work, research have introduced SEIG, which is a new method by which certified feature attribution is embedded into the inference process of the neural network architecture itself. With the addition of the LAA protocol and the ADM onto the IG technique, SEIG creates an explanation that is faithful, sparse, fidelity-preserving, and aligned with human reasoning. It can be seen through empirical results that SEIG attains a faithfulness of 91.7% and a human alignment coefficient of $\kappa = 0.93$ in the ImageNet dataset. This is better compared with the performance of all baselines tested, including the standard Integrated Gradients technique, in which the overhead is only 1.4× that of standard IG. Furthermore, the Attribution Certificate component makes it possible to convert numerical feature attribution values into meaningful explanation certificates required in many regulatory processes. In doing so, SEIG establishes the principle of self-explanation as one of the fundamental architectural components of a neural network, laying the groundwork towards accountability in neural networks. Directions for future works include the extension of SEIG to generative architectures, automation of faithfulness benchmarks through elimination of human annotation, and attribution certificate protocol integration with model governance tools.

References

1. Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning. MIT Press.
2. Senthilkumaran, B., Singh, G., Bostani, A., Krithika, S., Khalbayeva, Z., & Nanthini, P. (2025). Design and implementation of an AI-based medical analytics framework employing deep neural networks and advanced machine learning models for precision healthcare. Archives for Technical Sciences, 3(34), 991–1005. https://doi.org/10.70102/afts.2025.1834.991
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30).
5. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74
6. Irshad Ahamed, M. (2026). Ethical AI development and ensuring transparency and fairness in algorithmic decision-making. Global Tech Management Digest, 2(1), 13–19.
7. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319–3328). PMLR.
8. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (Vol. 31).
9. Mousavi, A., & Karshenasan, A. (2017). Forecasting stock prices of banks using artificial neural networks (GMDH). International Academic Journal of Accounting and Financial Management, 4(2), 71–78.
10. Walker, C., Jha, S., Chen, K., & Ewetz, R. (2024). Integrated decision gradients: Compute your attributions where the model makes its decision. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(6), 5289–5297. https://doi.org/10.1609/aaai.v38i6.28298
11. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning (pp. 3145–3153). PMLR.
12. Sanyal, S., & Ren, X. (2021). Discretized integrated gradients for explaining language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10285–10299). https://doi.org/10.18653/v1/2021.emnlp-main.808
13. Tuama, H. S., & Abdulameer, R. Y. (2023). Using time series models and neural networks to predict the sales of motor oils in Iraq. International Academic Journal of Business Management, 10(1), 1–11. https://doi.org/10.9756/IAJBM/V10I1/IAJBM1001
14. Simpson, L., Millar, K., Cheng, A., Lim, C. C., & Chew, H. G. (2025). Graph-based integrated gradients for explaining graph neural networks. In Australasian Joint Conference on Artificial Intelligence (pp. 150–162). Springer Nature.
15. Agrab, A. S. (2022). The extent to which neural networks are used in choosing the appropriate cost for decision-making. International Academic Journal of Economics, 9(1), 20–30. https://doi.org/10.9756/IAJE/V9I1/IAJE0903
16. Gur, S., Ali, A., & Wolf, L. (2021). Visualization of supervised and self-supervised neural networks via attribution guided factorization. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(13), 11545–11554. https://doi.org/10.1609/aaai.v35i13.17343
17. Dhurgadevi, N., Nandhini, P., Pavithra, S., & Suganya, R. (2022). Pothole detection using deep learning. International Academic Journal of Innovative Research, 9(2), 1–4. https://doi.org/10.9756/IAJIR/V9I2/IAJIR0908
18. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In Proceedings of the 37th International Conference on Machine Learning (pp. 5338–5348). PMLR.
19. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3543–3556). https://doi.org/10.18653/v1/N19-1357
20. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems (Vol. 32).
21. Ashour, H., & Chandrakumar, R. (2025). Blockchain-driven cybersecurity solutions for secure and scalable Internet of Energy architectures. National Journal of Intelligent Power Systems and Technology, 55–63.
22. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

23. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820