Causal Localization of Neural Function: The Shapley Value Method

Alon Keinan[1], Claus C. Hilgetag[2], Isaac Meilijson[1], Eytan Ruppin[1]

[1]Schools of Computer Science and Mathematics, Tel-Aviv University, Tel-Aviv, Israel
keinan@cns.tau.ac.il, *isaco@post.tau.ac.il*, *ruppin@post.tau.ac.il* (tel: +972-3-6406528)

[2]School of Engineering and Science, International University Bremen, Bremen, Germany
c.hilgetag@iu-bremen.de (tel: +49-421-2003542)

Abstract

Identifying the functional roles of elements of a neural network is one of the first challenges in understanding neural information processing. Aiming at this goal, lesion studies have been used in neuroscience, most of which employing single lesions and hence, limited in their ability to reveal the significance of interacting elements. This paper presents the *Multi-lesion Shapley value Analysis (MSA)*, an axiomatic, scalable and rigorous method, addressing the challenge of calculating the contributions of network elements from a multi-lesion data set. The successful workings of the MSA are demonstrated on artificial and biological data. MSA is a novel method for causal function localization, with a wide range of potential applications for the analysis of reversible deactivation experiments and TMS-induced "virtual lesions".

1 Introduction: The Multi-lesion Shapley Value Analysis

One of the principal challenges in understanding neural information processing is to identify the individual roles of a neural network's elements, be they single neurons, neuronal assemblies or cortical regions, depending on the scale on which the system is analyzed. Even simple nervous systems are capable of performing multiple and unrelated functions. Each function recruits some of the elements of the system, and often the same element participates in several different functions. Localization of specific functions in the nervous system is conventionally done by recording the activity during cognition and behavior and inferring the correlation between elements and different behavioral and functional observables. This correlation does not necessarily identify causality. For example, it is possible that a region does not contribute to the processing of a function, but its activity is still raised when the function is performed, because it is activated by other regions that play a role in carrying out the function. To overcome these inherent shortcomings, lesion studies have been employed in neuroscience, where the function performance is measured after lesioning different elements of the system. Lesioning enables, in principle, a correct identification of the elements that are causally responsible for a given function. Most of the lesion studies in neuroscience have been *single lesion* studies, in which only one element is lesioned at a time. Such single lesions are limited in their ability to reveal the significance of interacting elements. One obvious example is provided by two elements that exhibit a high degree of overlap in their function: Lesioning either element alone will not reveal its significance.

Acknowledging that single lesions are insufficient for localizing functions in neural systems, a Functional Contribution Analysis (FCA) was previously presented [1, 9, 5]. The FCA analyzes a data set composed of numerous multiple lesions that are afflicted upon a neural system, together with the corresponding system performance score, where in each multiple lesion experiment several elements are concurrently lesioned. The FCA uses these data to yield a prediction of the performance of the neural system when a new multiple lesion state is imposed on it. It further yields a quantification of the elements' *contributions* to each function, as a set of values minimizing that

prediction error. This contributions definition is an operative one, and hence, *there is no inherent notion of correctness of the contributions found by the method.* In particular, there are instances in which several different contribution assignments to the elements yield the same minimum prediction error. In such cases, the FCA algorithm may reach different solutions, all providing accurate predictions, but yielding different contributions.

This paper presents the *Multi-lesion Shapley value Analysis* (MSA), addressing the same challenge of defining and calculating the contributions of network elements from a data set of multiple lesions and their corresponding performance scores. In this framework, we view a set of multiple lesion experiments as a *coalitional game*, borrowing concepts and analytical approaches from the field of Game Theory. Specifically, we define the set of contributions to be the *Shapley value* [10], which stands for the unique fair division of the game's worth (the network's performance score when all elements are intact) among the different players (the network elements). The contribution of an element to a function measures its importance, i.e., the part it causally plays in the successful performance of that function. While in traditional game theory the Shapley value is more a theoretical tool that assumes full knowledge of the behavior of the game at all possible coalitions, we have developed novel methods to compute it approximately with high accuracy and efficiency from a relatively small set of multiple lesion experiments. The MSA framework further quantifies the interactions between groups of elements, allowing for higher order descriptions of the network. In contradistinction to it predecessor, the FCA, **the MSA provides a unique and axiomatically correct attribution of contributions to the system elements. It is the first method offering a fair and scalable solution to the problem of localization of function in the context of multi-lesion experiments**.

We focus the rest of this paper on the application of the MSA to three different cases. In section 2 we demonstrate the workings of the MSA on a toy problem, comparing it to single lesion analysis and to the FCA. Section 3 describes results of applying the MSA to an artificial evolved neurocontroller, analyzing it both on neuronal and synaptic levels. Section 4 introduces the high-dimensional MSA and presents the results of applying it for the analysis of biological reversible deactivation experiments. Our results, their implications and further implementations of the MSA framework are discussed in section 5.

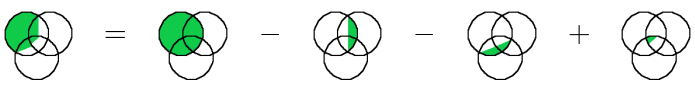
2 A Test Case

Let us define a system of elements $\{e_1, \ldots, e_n\}$, where the lifetime of element e_i is exponentially distributed with parameter λ_i (expectancy of $1/\lambda_i$), and such that the elements are independent. We define the performance of the system as the expected time when at least one of the elements is still functioning. That is, the expectancy of the maximum of the distributions. For simplicity, we focus on the case $n = 3$.

Based on the performance scores of all multi-lesions afflicted upon the network, the Shapley value is obtained, yielding a contribution

$$\gamma_1(N, v) = \frac{1}{\lambda_1} - \frac{1}{2} \cdot \frac{1}{\lambda_1 + \lambda_2} - \frac{1}{2} \cdot \frac{1}{\lambda_1 + \lambda_3} + \frac{1}{3} \cdot \frac{1}{\lambda_1 + \lambda_2 + \lambda_3} \tag{1}$$

for e_1, and similarly for the other elements. Illustrating the meaning of the resulted contribution using Venn diagrams, we get



,

in the same order as the terms in equation (1). That is, e_1 is accredited for a third of the expected time when it is functioning with both e_2 and e_3 (the rest is divided equally between the contributions of e_2 and e_3), for half of the time when it is functioning with either e_2 or e_3 (the other half is

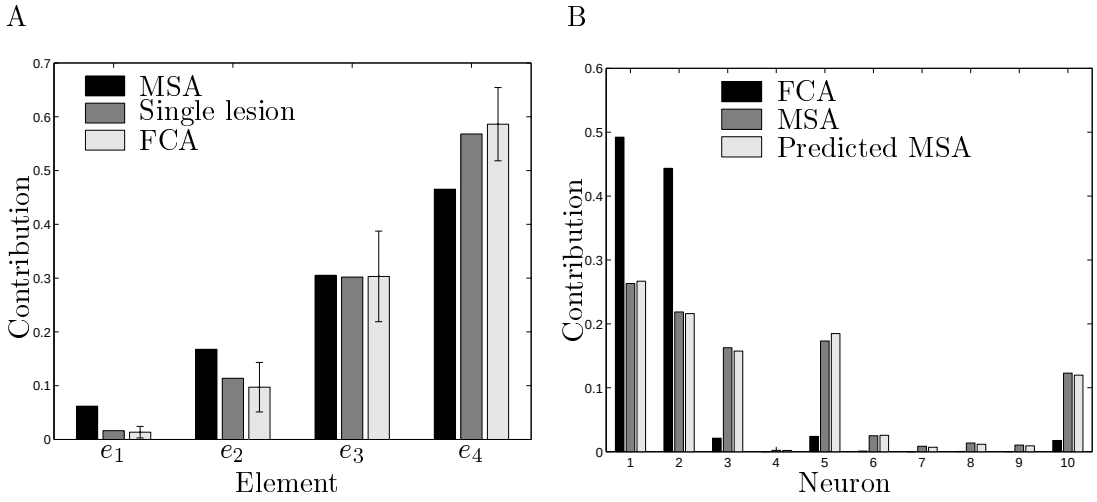
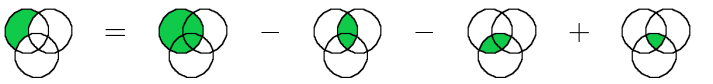


Figure 1: A. The MSA contributions (Shapley value) for a test case are compared with the contributions yielded by single lesion analysis and with the FCA contributions (mean and standard deviations across 10 FCA runs using the full set of all 2^4 multi-lesions). All three values are normalized such that the sum over all elements equals 1. B. MSA contributions, FCA contributions and predicted MSA contributions of the EAA's neurons. All are based on the full set of 2^{10} multi-lesions and are normalized such that the sum of the contributions of all the neurons equals 1.

accredited to the other element) and for the whole time when it is functioning alone, denoting a fair division of the system performance to the different elements. The contribution of e_1 according to the single lesion approach (the decrease in performance when it is lesioned) equals

$$\sigma_1(N,v) = \frac{1}{\lambda_1} - \frac{1}{\lambda_1+\lambda_2} - \frac{1}{\lambda_1+\lambda_3} + \frac{1}{\lambda_1+\lambda_2+\lambda_3}. \tag{2}$$

Illustrating σ_1 using Venn diagrams, we obtain



in the same order as the terms in equation (2). That is, when using single lesioning each element is only accredited for the expected time when it is functioning alone, without considering its previous contribution, while other elements are still functioning. Thus, **the Shapley value is much more informative in capturing the true contributions**.

Figure 1A compares the MSA contributions (the Shapley value), the single lesion contributions and the contributions yielded by the FCA for a concrete example of the test case, where $n = 4$ and $\lambda_i = 1/i$, for $i = 1, \ldots, 4$. Evidently, even in this very simple system, the contributions yielded by the MSA differ greatly from the ones yielded by the single lesion analysis. The FCA contributions resemble the contributions assigned by the single lesion analysis, testifying that the FCA fails to capture a fair attribution of contributions in this case.

3 Analysis of Evolved Autonomous Agents

A Neurally-driven *Evolved Autonomous Agent (EAA)* is a software program embedded in a simulated virtual environment, performing typical animat tasks. An agent is controlled by an artificial neural network "brain", receiving and processing sensory inputs from the surrounding environment and governing the agent's behavior via the activation of the motors. EAAs are developed via genetic algorithms that apply some of the essential ingredients of inheritance and selection to a

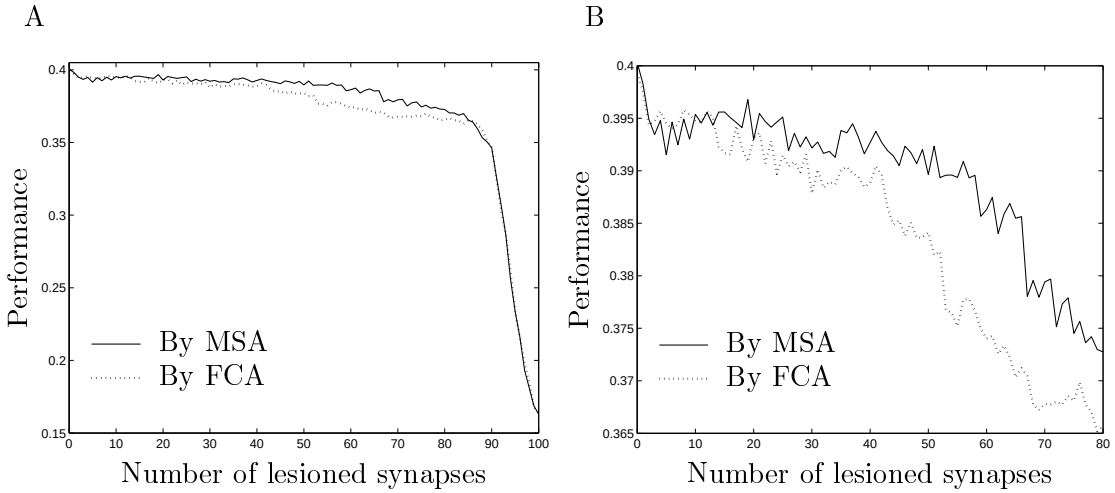


Figure 2: *Agent performance as a function of pruning level, by MSA and by FCA.* In both methods the synapses are incrementally lesioned by ascending order of their contribution. A. Incrementally pruning all synapses. B. Focusing on the interesting region of the first 80 synapses pruned, where the agent still has a viable performance.

population of agents that undergo evolution, which make them a very promising model for studying neural processing and developing methods for its analysis [8]. Figure 1B shows that the MSA contributions for the neurons of the analyzed agent differ significantly from the FCA contributions. Specifically, neuron number 5, **the command neuron determining the agent's behavioural strategy [2], is assigned a significant contribution by the MSA but a near-vanishing one by the FCA**. The MSA may also calculate the Shapley value based on the prediction of the performance in all multi-lesions, instead of the actual performance scores, resulting in *predicted MSA contributions*. This is highly important as it *enables the MSA to use only a small sample of multi lesions out of the possibly vast lesion space*. Figure 1B shows the predicted MSA contributions, based on the FCA's prediction, to be very similar to the actual MSA contributions. This demonstrates that although the FCA contributions are far from the MSA ones, the FCA's prediction (based on its contributions) allows for the good estimation of the latter.

We turn to analyzing the neurocontroller at the level of its synapses, capturing the synaptic backbone of the network. Considering that it is impossible to calculate even the predicted performance of all 2^{100} synaptic multi-lesions, the synaptic MSA contributions can be based on *an estimate of the Shapley value*. The synaptic contributions may serve as a guide for pruning a neural network, by lesioning the synapses according to the magnitude of their contributions (in ascending order). This has been done for the FCA [1], showing that pruning by the FCA contributions outperforms pruning by synaptic weights magnitude. To compare the contributions obtained by the MSA with those obtained by the FCA, we incrementally prune the full recurrent synaptic network of the agent using the two methods. Figure 2 depicts the performance of the agent as a function of the number of lesioned synapses, starting from the intact network. **As evident, the MSA tends to keep the performance higher throughout the incremental pruning, showing that the MSA is better than the FCA in quantifying the contributions of elements**.

4 MSA of Biological Data: Reversible Deactivation Experiments

The Shapley value stands for the average marginal importance of an element. For complex networks, where the importance of an element may depend on the state (lesioned or intact) of other elements, a higher order description is required to capture the characteristics of the network. Focusing here

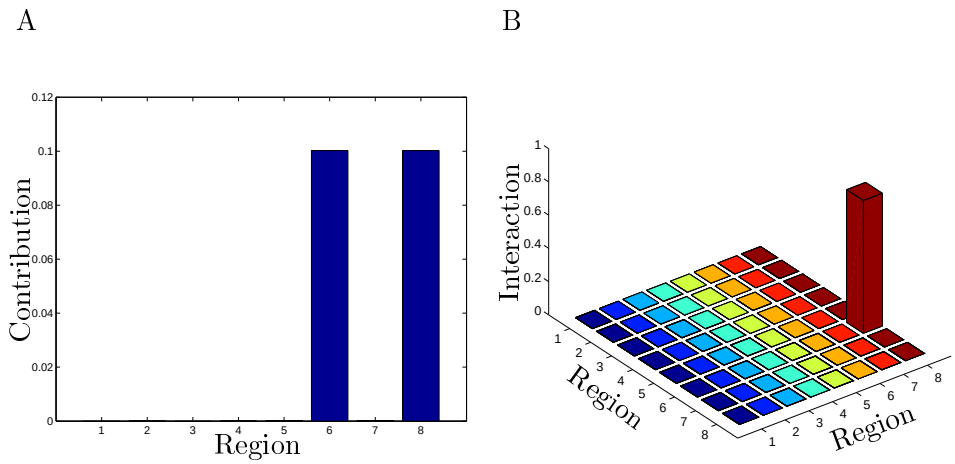


Figure 3: *Two-dimensional MSA of reversible deactivation experiments.* A. Predicted MSA contributions of the eight regions. Regions 6 and 8 represent SC_R-deep and SC_L-deep, respectively. B. The interaction between each pair of regions. The symmetric interaction is plotted only above the main diagonal.

on a two-dimensional analysis, we define the interaction between a pair of elements as *how much larger (or smaller) the average marginal importance of the two combined elements is compared with the sum of the average marginal importance of each of them separately when the other one is lesioned.* Furthermore, the MSA classifies the type of interaction based on the average marginal importance of each element when the other element is lesioned and when the other element is intact.

We turn to analyzing data from reversible cooling deactivation experiments studying the localization of spatial attention to auditory stimuli (paradigm described in [6]). The experiments tested auditory stimuli detection and orienting responses in intact and reversibly lesioned cats, using cryoloops implanted over cortical and superior collicular (SC) target structures following an established standard procedure [7]. Nineteen single and multi-lesion experiments were performed [6] and another 14 lesions were deduced by assuming mirror-symmetric effects resulting from lesions of the two hemispheres. Figure 3A shows the predicted MSA contributions of the different regions involved in the experiments, using Projection Pursuit Regression (PPR) [3] for prediction, trained using the 33 lesions. It is clear that only regions 6 and 8 (SC_R-deep and SC_L-deep) play a role in determining the performance. Both have a contribution equal to half the overall performance of the system (0.2). Due to the experimental approach, when a deep component of the SC is lesioned, the superficial one is lesioned as well (regions 5 and 7 in Figure 3A). Nevertheless, **the MSA successfully reveals that only the deep SC regions are the ones of significance**.

We further performed a two-dimensional MSA to quantify the interactions between each pair of regions, finding only one significant interaction, between the pair (SC_R-deep,SC_L-deep) (Figure 3B). Furthermore, observing the negative contribution of each of the two regions when the other one is lesioned (-0.3) and the positive contribution when the other one is intact (0.5), **the MSA concluded that the two regions exhibit "paradoxical lesioning"**, uncovering the type of interaction assumed to take place in this function [4, 6]. **This analysis testifies to the usefulness of the MSA in deducing from lesioning data the important regions and the important interactions among the regions.**

5 Conclusion

We describe a new framework for quantitative function localization via multi-lesion experiments, based on a rigorous definition of the elements' contributions. The Shapley value as a unique fair

solution concept has been used in many fields beyond theoretical Game Theory (including cost allocation, politics, international environmental problems and economic theory), testifying to its usefulness. The MSA accurately approximates the Shapley value, in a scalable manner, making it a more accurate and efficient method for function localization than its predecessor, the FCA. The prediction and estimation variants of the MSA are specifically geared toward neuroscience analysis applications where it is possible to perform only a limited number of multi-lesion experiments.

We aim to focus future work on the analysis of several neural networks, both artificial and biological. On the biological level, we plan to continue the application of the MSA to data from reversible cooling deactivation experiments in cats. More importantly, we plan to apply the MSA to the localization of spatial attention in the human brain. To this end, occipital, parietal, temporal, prefrontal and motor cortical regions in human subjects will be reversibly deactivated in multiple deactivation experiments using the "virtual lesion" technique of Transcranial Magnetic Stimulation (TMS). These experiments will be analyzed by the MSA to yield precise quantitative localization of processing, to study the general profile of spatial localization across subjects and to determine the important functional interactions between regions involved in spatial attention processing in the cortex.

References

[1] R. Aharonov, L. Segev, I. Meilijson, and E. Ruppin. Localization of function via lesion analysis. *Neural Computation*, 15(4), 2003.

[2] R. Aharonov-Barki, T. Beker, and E. Ruppin. Emergence of memory-driven command neurons in evolved artificial agents. *Neural Computation*, 13(3):691–716, 2001.

[3] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.

[4] C. C. Hilgetag, S. G. Lomber, and B. R. Payne. Neural mechanisms of spatial attention in the cat. *Neurocomputing*, 38:1281–1287, 2000.

[5] A. Keinan, I. Meilijson, and E. Ruppin. Controlled analysis of neurocontrollers with informational lesioning. *Phil. Trans. R. Soc. Lond. A, to appear*, 2003.

[6] S. G. Lomber, B. R. Payne, and P Cornwell. Role of the superior colliculus in analyses of space: Superficial and intermediate layer contributions to visual orienting, auditory orienting, and visuospatial discriminations during unilateral and bilateral deactivations. *J Comp Neurol*, 441:44–57, 2001.

[7] S. G. Lomber, B. R. Payne, and J. A. Horel. The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function. *J Neurosci Methods*, 86:179–194, 1999.

[8] E. Ruppin. Evolutionary autonomous agents: A neuroscience perspective. *Nature Reviews Neuroscience*, 3:132–141, 2002.

[9] L. Segev, R. Aharonov, I. Meilijson, and E. Ruppin. High-dimensional analysis of evolutionary autonomous agents. *Artificial Life*, 9(1), 2003.

[10] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume II of *Annals of Mathematics Studies 28*, pages 307–317. Princeton University Press, Princeton, 1953.