Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 58–60 (2004) 215–222
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Causal localization of neural function: the Shapley value method

Alon Keinan[a,*], Claus C. Hilgetag[c], Isaac Meilijson[b], Eytan Ruppin[a,d]

[a] *Schools of Computer Science, Tel-Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel*
[b] *School of Mathematical Sciences, Tel-Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel*
[c] *School of Engineering and Science, International University Bremen, Bremen, Germany*
[d] *School of Medicine, Tel-Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel*

Abstract

Identifying the functional roles of elements of a neural network is one of the fundamental challenges in understanding neural information processing. Aiming at this goal, lesion studies have been used extensively in neuroscience. Most of these employ single lesions and hence, limited ability in revealing the significance of interacting elements. This paper presents the *multi-perturbation Shapley value analysis* (*MSA*), an axiomatic, scalable and rigorous method, addressing the challenge of determining the contributions of network elements from a data set of multi-lesions or other perturbations. 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".

Keywords: Localization of function; Multi-lesions; Shapley value; Contributions analysis; Interactions; Multi-perturbations

1. Introduction: The multi-perturbation 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

* Corresponding author. Tel.: +972-3-6407864.
E-mail addresses: keinan@cns.tau.ac.il (A. Keinan), c.hilgetag@iu-bremen.de (C.C. Hilgetag), isaco@post.tau.ac.il (I. Meilijson), ruppin@post.tau.ac.il (E. Ruppin).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.01.046

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 the elements' activity 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 do play a role in carrying out the function. To overcome these inherent shortcomings, lesion studies have been employed in neuroscience, where functional 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 approaches 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 presented [1,11,6]. The FCA analyzes a data set composed of numerous multiple lesions that are afflicted upon a neural system, together with the corresponding system performance scores. 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 contribution 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-perturbation Shapley value analysis* (MSA), addressing the same challenge of defining and calculating the contributions of network elements from a data set of multiple lesions or other type of perturbations 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* [12], 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, that is, 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 methods to compute it approximately with high accuracy and efficiency from a relatively small set of multiple lesion experiments (see [5] for a more detailed account of the MSA). The MSA framework further quantifies the interactions between groups of elements, allowing for higher-order descriptions of the network. In contrast

to its 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 in 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, i.e., 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

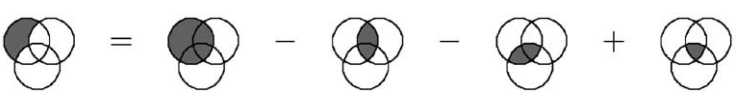
$$\gamma_1 = \frac{1}{\lambda_1} - \frac{1}{2}\cdot\frac{1}{\lambda_1+\lambda_2} - \frac{1}{2}\cdot\frac{1}{\lambda_1+\lambda_3} + \frac{1}{3}\cdot\frac{1}{\lambda_1+\lambda_2+\lambda_3}, \tag{1}$$

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

$$\sigma_1 = \frac{1}{\lambda_1} - \frac{1}{\lambda_1+\lambda_2} - \frac{1}{\lambda_1+\lambda_3} + \frac{1}{\lambda_1+\lambda_2+\lambda_3}. \tag{2}$$

Illustrating σ_1 using Venn diagrams, we obtain



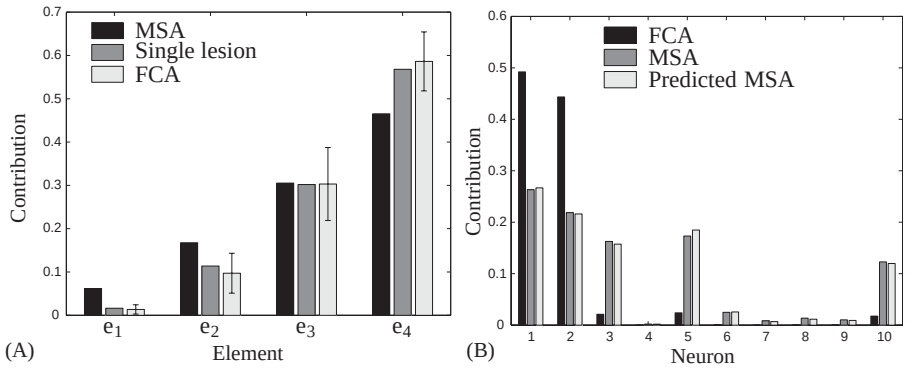


Fig. 1. (A) The MSA contributions (Shapley value) for a test case are compared with the contributions yielded by single lesion analysis and with the FCA contributions (mean and standard deviations across 10 FCA runs using the full set of all 2^4 multi-lesions). All three values are normalized such that 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.

in the same order as the terms in Eq. (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.

Fig. 1A compare 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, too, in capturing 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 its motors. EAAs are developed via genetic algorithms that apply some of the essential ingredients of inheritance and selection to a population of agents that undergo evolution, which make them a very promising model for studying neural processing and developing methods for its analysis [9]. Fig. 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 behavioral 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. Fig. 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. The MSA may be further utilized to determine the level of localization in the network and to determine the contribution of a specific feature of a neuron's dynamics [10].

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. Fig. 2 depicts the performance of the agent as a function of the number of pruned synapses, starting from the intact network. As evident, the MSA tends to keep the performance higher than the FCA throughout the incremental pruning, testifying that it better captures the inherent importance of the synapses.

Fig. 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. The figure focuses on the first 80 synapses pruned, where the agent has still viable performance.

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 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 [7]). 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 [8]. Nineteen single and multi-lesion experiments were performed [7] and another 14 lesions were deduced by assuming mirror-symmetric effects resulting from lesions of the two hemispheres. Fig. 3A shows the predicted MSA contributions of the different regions involved in the experiments, using Projection Pursuit Regression (PPR) [3] for prediction, trained on 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 was lesioned, the superficial one was lesioned as well (regions 5 and 7 in Fig. 3A). Nevertheless, the MSA successfully revealed that only the deep SC regions are the ones of significance.

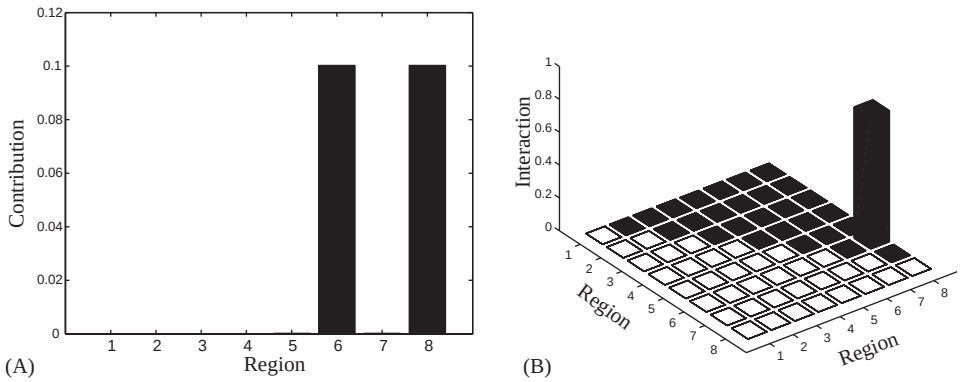


Fig. 3. *Two-dimensional MSA of reversible deactivation experiments.* (A) Predicted MSA contributions of the eight regions. Regions 6 and 8 represent SC_R-deep and SC_L-deep, respectively. (B) The symmetric interaction between each pair of regions.

We further performed a two-dimensional MSA to quantify the interactions between each pair of regions, finding only one significant interaction, between SC_R-deep and SC_L-deep (Fig. 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 effect, uncovering the type of interaction assumed to take place in this function [4,7]. *This analysis testifies to the usefulness of the MSA in deducing from lesioning data the functionally important regions as well as their significant interactions.*

5. Conclusions

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.

MSA is a novel method for assigning, based on any type of multi-perturbations inflicted upon a system, numerical values of responsibility in an objective and reasonable way. We aim to focus future work on the analysis of several neural networks, both artificial and biological. On the biological level, for instance, we now 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 are 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.

Acknowledgements

We acknowledge the valuable contributions and suggestions made by Ranit Aharonov, Shay Cohen, Zohar Ganon, Ehud Lehrer and Keren Saggie and the technical help provided by Oran Singer. This research has been supported by the Adams Super Center for Brain Studies in Tel-Aviv University and by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities.

References

[1] R. Aharonov, L. Segev, I. Meilijson, E. Ruppin, Localization of function via lesion analysis, Neural Comput. 15 (4) (2003) 885–913.
[2] R. Aharonov-Barki, T. Beker, E. Ruppin, Emergence of memory-driven command neurons in evolved artificial agents, Neural Comput. 13 (3) (2001) 691–716.
[3] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Statist. Assoc. 76 (376) (1981) 817–823.
[4] C.C. Hilgetag, S.G. Lomber, B.R. Payne, Neural mechanisms of spatial attention in the cat, Neurocomputing 38 (2000) 1281–1287.
[5] A. Keinan, B. Sandbank, C.C. Hilgetag, I. Meilijson, E. Ruppin, Fair attribution of functional contribution in artificial and biological networks, Neural Computat., to appear.
[6] A. Keinan, I. Meilijson, E. Ruppin, Controlled analysis of neurocontrollers with informational lesioning, Philos. Trans. Roy. Soc. London: Ser. A 361 (1811) (2003) 2123–2144.
[7] S.G. Lomber, B.R. Payne, 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 (2001) 44–57.
[8] S.G. Lomber, B.R. Payne, J.A. Horel, The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function, J. Neurosci. Methods 86 (1999) 179–194.
[9] E. Ruppin, Evolutionary autonomous agents: a neuroscience perspective, Nature Rev. Neurosci. 3 (2002) 132–141.
[10] K. Saggie, A. Keinan, E. Ruppin, Spikes that count: rethinking spikiness in neurally embedded systems, Twelfth Annual Computational Neuroscience Meeting (CNS*2003), 2003. Neurocomputing, this issue.
[11] L. Segev, R. Aharonov, I. Meilijson, E. Ruppin, High-dimensional analysis of evolutionary autonomous agents, Artificial Life 9 (1) (2003) 1–20.
[12] L.S. Shapley, A value for n-person games, in: H.W. Kuhn, A.W. Tucker (Eds.), Contributions to the Theory of Games, Vol. 2 of Annals of Mathematics Studies 28, Princeton University Press, Princeton, 1953, pp. 307–317.

Alon Keinan earned his B.Sc. degree *summa cum laude* in computer science, statistics and operations research from Tel Aviv University in 1997. He is currently a Ph.D. student at the school of computer science, Tel Aviv University, working in the computational neuroscience lab, under the guidance of Eytan Ruppin and Isaac Meilijson.

Claus C. Hilgetag is an Assistant Professor of Neuroscience at the International University Bremen. He studied Biophysics in Berlin and Neuroscience in Edinburgh, Oxford, Newcastle and Boston. His current research focuses on computational analyses of cortical architecture and connectivity, and on understanding the mechanisms of spatial attention and inattention in mammalian brains.