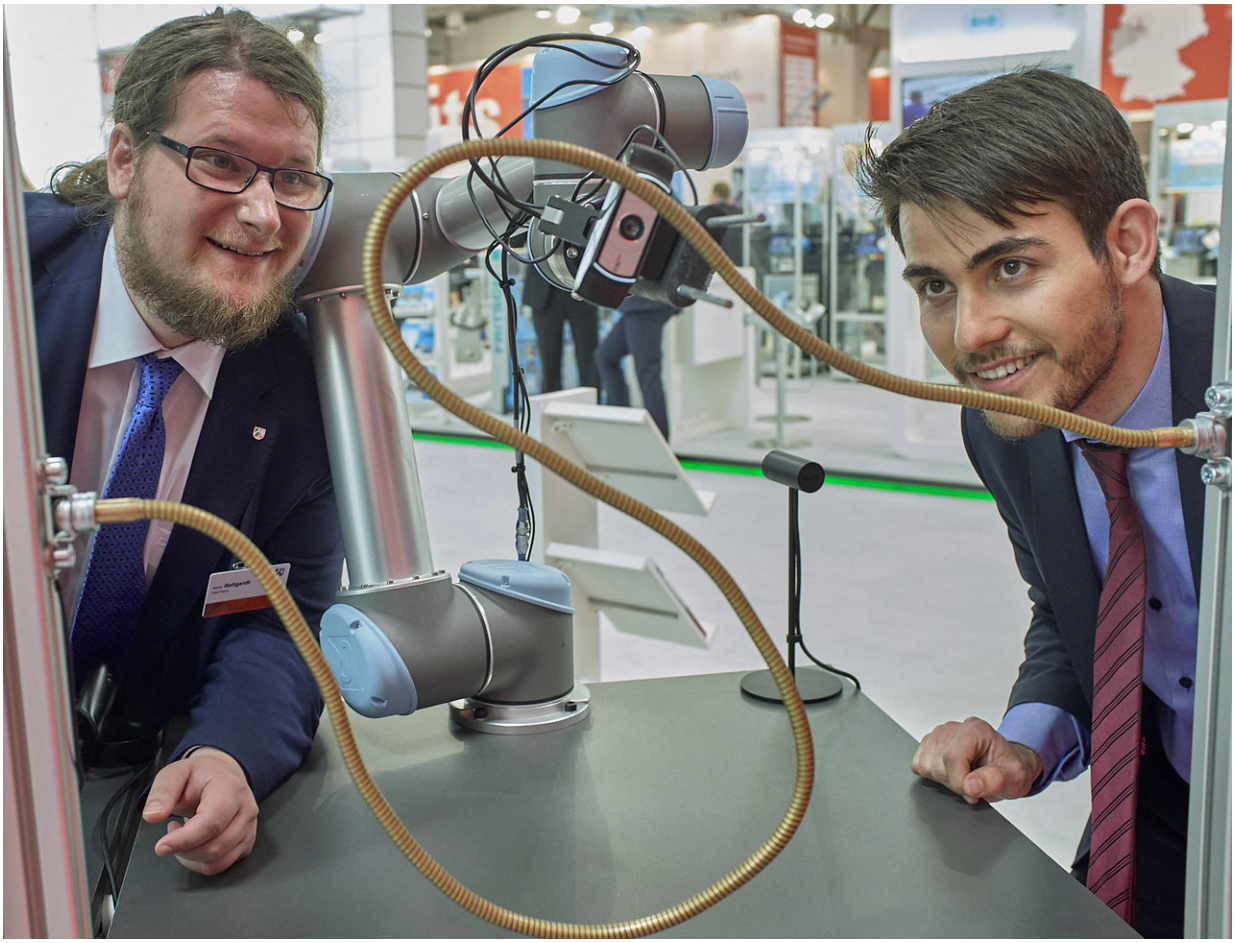
Using ablation to examine the structure of artificial neural networks

December 28 2018, by Ingrid Fadelli



Credit: Lillian, Meyes & Meisen.

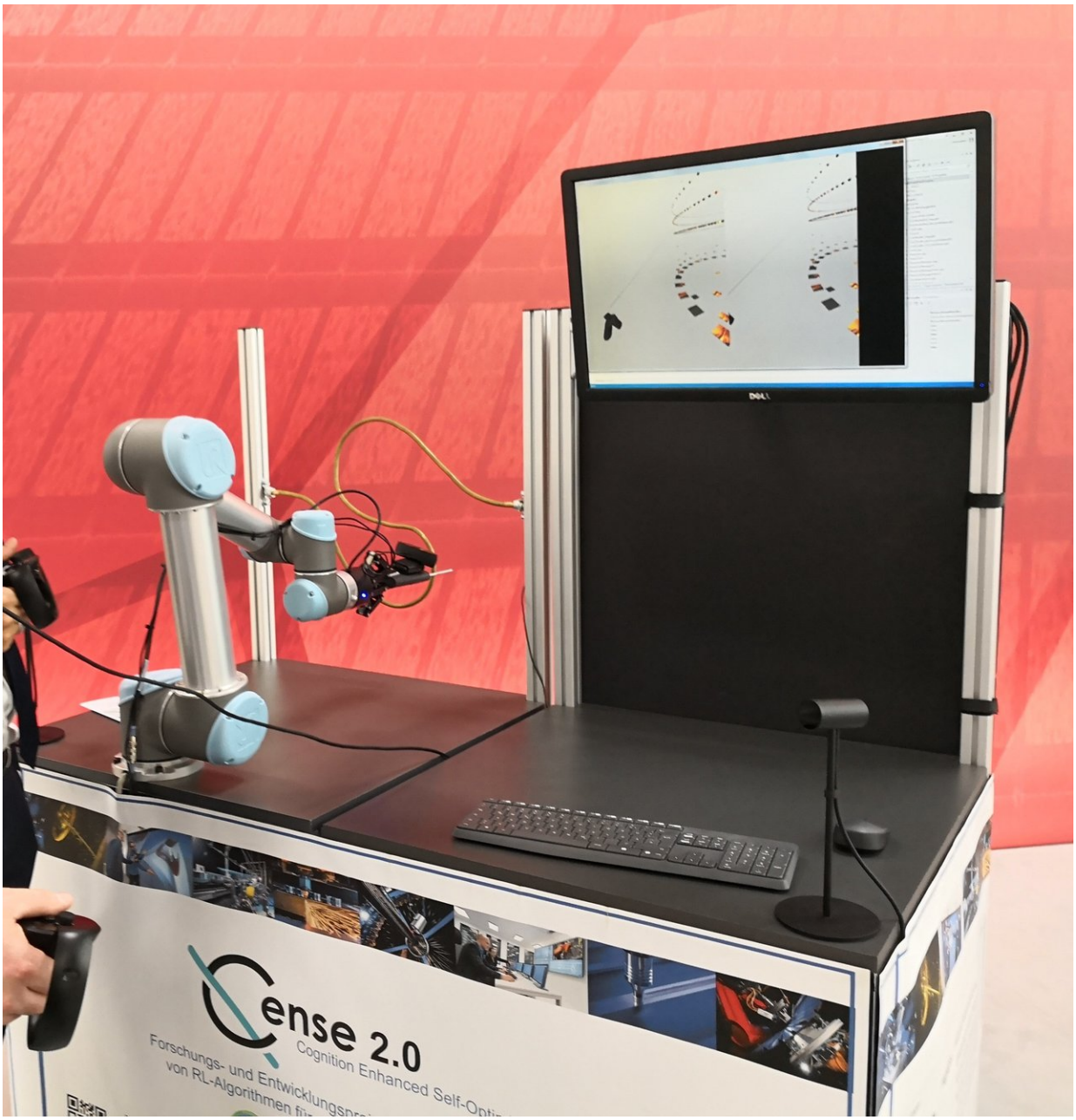
A team of researchers at RWTH Aachen University's Institute of

Information Management in Mechanical Engineering have recently explored the use of neuroscience techniques to determine how information is structured inside artificial neural networks (ANNs). In their paper, pre-published on arXiv, the researchers applied a technique called ablation, which entails cutting away parts of the brain to determine their function, on neural network architectures.

"Our idea was inspired by research in the field of neuroscience, where one of the main goals is to understand the how our brain works," Richard Meyes and Tobias Meisen, two researchers who carried out the study told TechXplore, via email. "Many insights about the brain's functionalities were discovered in ablation studies, which is an approach in which specific parts of the brain are carefully damaged in a controlled manner, affecting the brains ability to perform everyday tasks, such as generating speech, or motion."

The aim of the study carried out by Meyes, Meisen and their colleague Peter Lillian was to examine ANNs from a biological perspective, assessing their structure and the function of their different components. They decided to do this using ablation, a technique employed in neuroscience research for over two hundred years.

Essentially, ablation consists in selectively removing or destroying tissue in specific areas of the brain, with the sole purpose of observing the behavioural effects of this damage and hence better understanding the function of these areas. Ablation has already been applied to ANNs in several studies, but these studies primarily focused on tweaking the layers of the network and changing its structure, thus more closely resembling parameter searches than biological ablation.



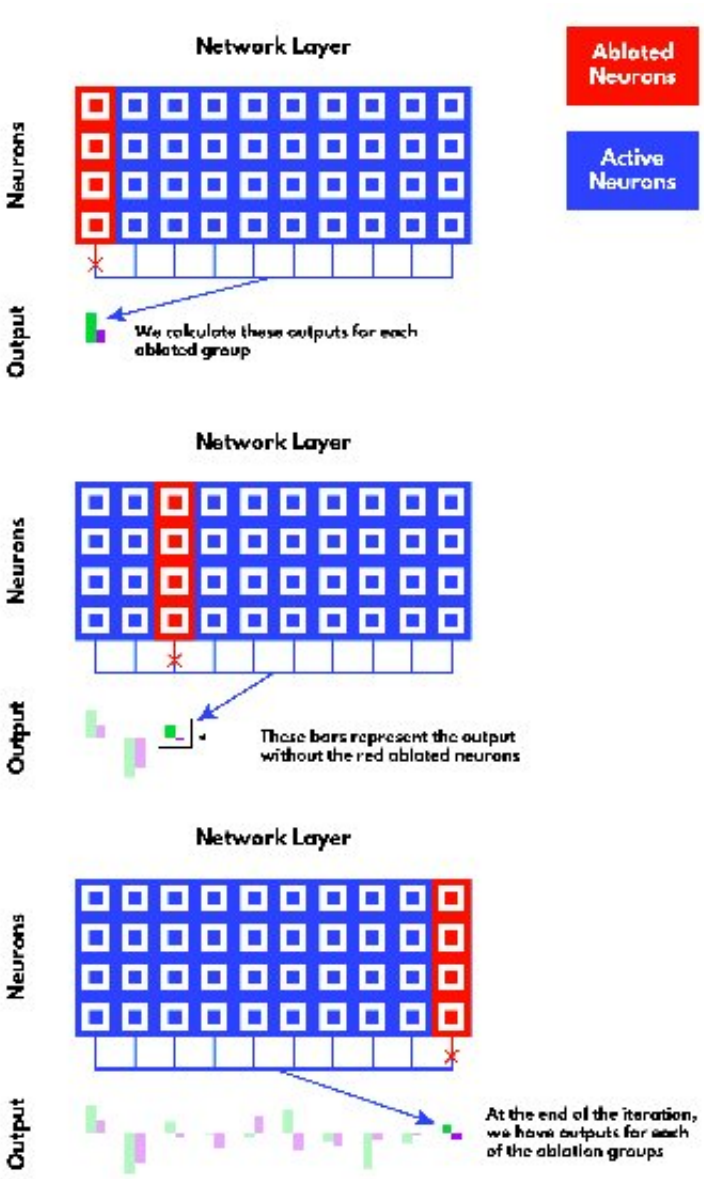
Credit: Lillian, Meyes & Meisen.

In their study, the researchers wished to damage areas of ANNs and observe how this affected their performance. Ultimately, they wished to use these observations to compare the organization of artificial neural

networks with that of biological ones.

"The idea behind ablations for artificial neural networks (ANNs) is simple," Meyes and Meisen explained. "First, we train a network to perform a specific task, e.g. to recognize handwritten digits. Second, we cut off a small part of the network and evaluate how the networks performance changes due to the damage caused. Third, we determine whether there is a relationship between the location of the damaged part and the effect it had on the network's performance. This way, we found that specific abilities of the network, e.g. to perform forward motions of the controlled robot, are locally represented and can be destroyed purposefully."

By ablating ANNs trained to navigate a wire-loop and examining how these interventions affected their output, the researchers gathered a number of interesting findings, suggesting that there are indeed links and similarities between artificial and biological networks. These similarities are related both to how the networks arrange themselves and how they store information.

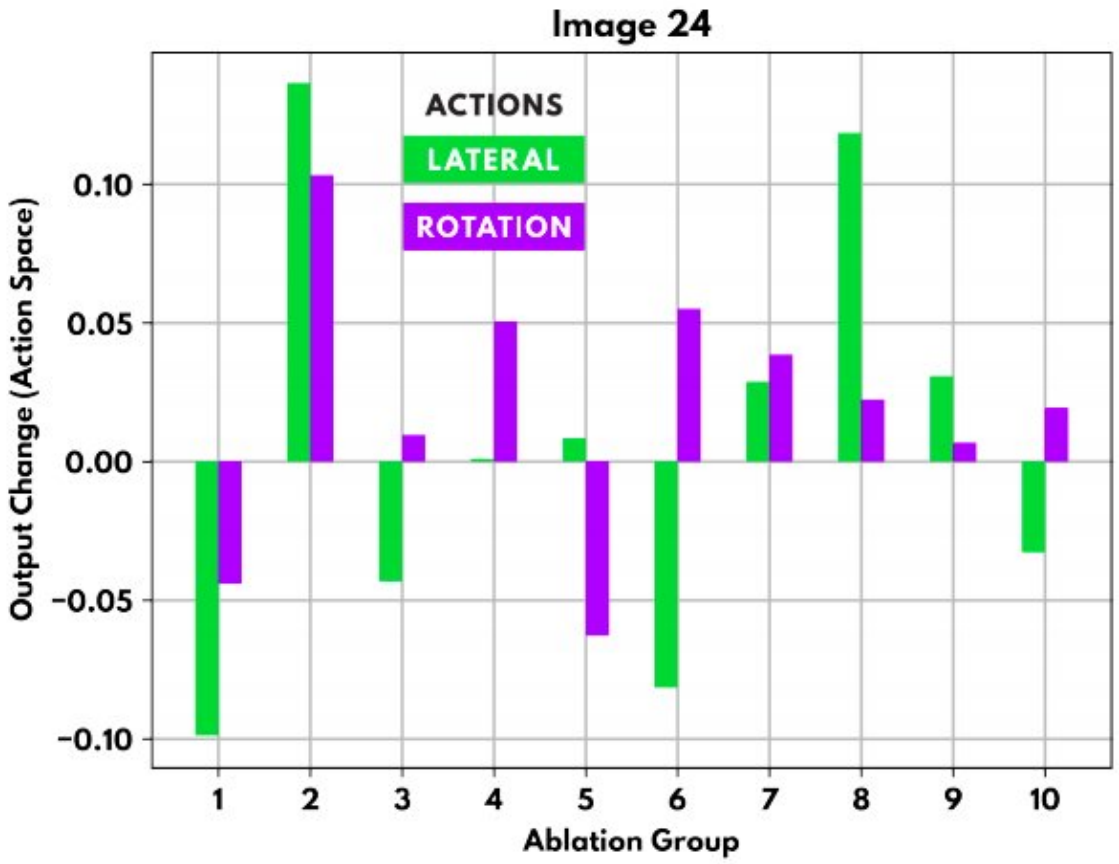


As each ablation group is removed, the output without that group is saved. After ablating each group, the researchers ended up with a list of outputs showing how the network changes when its parts are removed. Only one group is ablated at a time. Credit: Lillian, Meyes & Meisen.

"Our most interesting finding is the observation that a damaged network's performance generally decreases, while very specific abilities of the network, e.g. to recognize a specific digit, can be enhanced by

damaging specific parts," Meyes and Meisen said. "Our study suggests that a neural network's performance can be increased by carefully damaging it in the right regions. Furthermore, our study implies that the application of neuroscientific methods for ANNs may open up new perspectives on understanding artificial intelligence."

Despite the fascinating results gathered by Meyes, Meisen and Lillian, their study had several limitations and was merely a first step in examining the connections between biological and artificial neural networks. For instance, their experiments were limited by the use of reinforcement learning and relied on a model trained robotically, in real-time. Future research could examine the similarities between ANNs and brain networks in more detail and at larger scales.



A typical network's ablation results (how its output changed) for an image—the method used by the researchers matches each ablation group with its counterparts in the other trials. This data makes up part of the expanded action space. The researchers have omitted the longitudinal action due to its highly constant value. Credit: Lillian, Meyes & Meisen.

"We now plan to continue exploring our general idea of conducting neuroscience inspired research on ANNs," Meyes and Meisen said. "One of our next steps will be to visualize activity in ANNs just like brain activity can be visualized with imaging methods such as fMRI. We aim to make to make the decision-making process in ANNs more transparent and get a new perspective on AI in general."

More information: Peter Lillian et al. Ablation of a robot's brain: neural networks under a knife. arXiv:1812.05687 [cs.NE]. arxiv.org/abs/1812.05687

Richard Meyes et al. Continuous Motion Planning for Industrial Robots based on Direct Sensory Input, *Procedia CIRP* (2018). DOI: 10.1016/j.procir.2018.03.067

Richard Meyes et al. Motion Planning for Industrial Robots using Reinforcement Learning, *Procedia CIRP* (2017). DOI: 10.1016/j.procir.2017.03.095