Explaining Internal Representations in Deep Networks: Adversarial Vulnerability of Image Classifiers and Learning Sequential Tasks with Sparse Reward

prof. Ing. Igor Farkaš, Dr.

Centre for Cognitive Science
Department of Applied Informatics
Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
Mlynská dolina, 842 48 Bratislava
Slovak Republic
E-mail: farkas@fmph.uniba.sk

Abstract

Modern artificial intelligence based on deep neural networks has demonstrated in the past decade great achievements in concrete tasks, sometimes even surpassing human performance. On the other hand, there exist fundamental problems in these models, be it image classification or natural language tasks. Since deep networks are inherently black boxes, it is important to design and apply techniques that help shed light on the functioning of the trained models. In the talk, we will discuss two domains. First, in the context of image classification we will illustrate the effect of adversarial examples that can easily fool trained models, hence revealing their lack of robustness. In the second part, we will deal with sequential tasks, such as computer games, with an extremely sparse reward. Introducing the concept of intrinsic motivation, we will describe the neural networks based model that can successfully use reinforcement learning and self-supervised knowledge distillation to solve these tasks thanks to optimized organization of its internal representations. Finally, we briefly mention our current work related to cognitive robotics.

Short Bio

Igor Farkaš graduated from Slovak University of Technology in Bratislava (technical cybernetics), where he also received a doctoral degree (applied informatics). He worked as a postdoc at the University of Richmond, Virginia, USA. After returning home in 2003, he became affiliated with the Faculty of mathematics, physics and informatics, Comenius University in Bratislava, and served as the the Chair of Department of Applied Informatics in 2015-2022. Since 2014 he has been a guarantor of the unique international interdisciplinary master program in cognitive science (MEi:CogSci). His research areas span the related

fields of cognitive science and artificial intelligence, with the focus on studying artificial neural networks and their use in various tasks such as natural language and robotics. He has been involved in several projects, currently serving as the coordinator of the Horizon Europe project TERAIS. In 2024, he received two national awards: Scientist of the Year 2023 in the category Personality of International Cooperation, and the ESET Science Award in the category Outstanding Academic in Slovakia.