Towards more useful explanations of deep learning-based image classification

Master Thesis Proposal in *Image Analysis and Machine Learning* at the Division Vi3, Dept. of Information Technology.

Supervisor & Contact

Prof. Joakim Lindblad, joakim.lindblad@it.uu.se

Background

With the growing prevalence of AI and convolutional neural networks (CNNs) to support decision making throughout society, there is an urgent demand to explain what their decisions are based on. This has led to the creation of the rapidly growing field of explainable AI (XAI). Several techniques have been developed to shed light on how the decisions of AI systems are made. However, the vast majority of such methods are offering explanations of the type where in the image the important factors are appearing. These heat-mapping, saliency, activation, attention, or attribution methods highlight where the network is “looking” when reaching a decision, but they do not answer what the network finds as important at these positions. Is it colour, shape, texture?

Aim

This project aims to explore methods which can deliver information about conceptual factors behind CNN decision making, enabling improved understanding – essential for confident usage of AI for critical tasks such as medical diagnostics, but also for better insight in the underlying phenomena.

Prerequisites

- Proficiency in computer programming (Python is a must)
- Image analysis course passed with grade 4 or higher
- Knowledge and own experience of deep learning (PyTorch)

Relevant courses

- Deep Learning for Image Analysis - 1MD120
- Advanced Probabilistic Machine Learning - 1RT705

References

- Tobias Hammarström. Towards explainable decision-making strategies of deep convolutional neural networks: An exploration into explainable AI and potential applications within cancer detection, 2020. http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1505847
- Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International conference on machine learning, pp 2668–2677. PMLR, 2018.
- Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. International Conference on Data Science and Advanced Analytics (DSAA), pp 80–89, 2018.
- Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin. "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.