Statistical relevance critique in machine learning explanation systems

We will give an overview of the latest development in explanation techniques in machine learning from the perspective of the Salmons theory of statistical relevance [1].While the demonstrated approaches find success in applicationsin machine learning practice, we found the critique that wastheorised by Nency Cartwright [2] could be applied in the formof adversarial attacks on neural networks [3]. We believe thatby providing this transfer of critique we could help for a better conceptualisation of explanation systems in machine learning.

REFERENCES:

[1] N. Cammarata, S. Carter, G. Goh, C. Olah, M. Petrov, L. Schubert,C. Voss, B. Egan, and S. K. Lim, Thread: circuits, Distill, vol. 5, no. 3,p. e24, 2020.
[2] N. Cartwright, “Causal laws and effective strategies,”Noˆus, pp. 419–437,1979.
[3] J. Mu and J. Andreas, Compositional explanations of neurons, arXiv:2006.14032, 2020