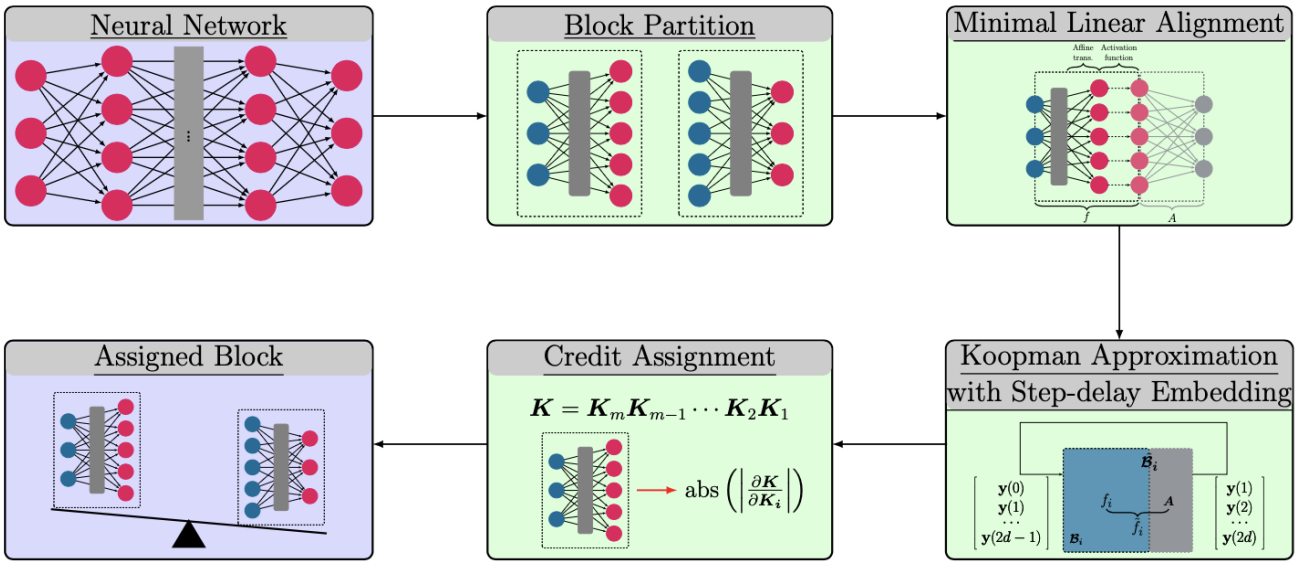
Credit Assignment for Trained Neural Networks Based on Koopman Operator Theory

Zhen LIANG, Changyuan ZHAO, Wanwei LIU, Bai XUE, Wenjing YANG, Zhengbin PANG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2629-4

Problems & Ideas

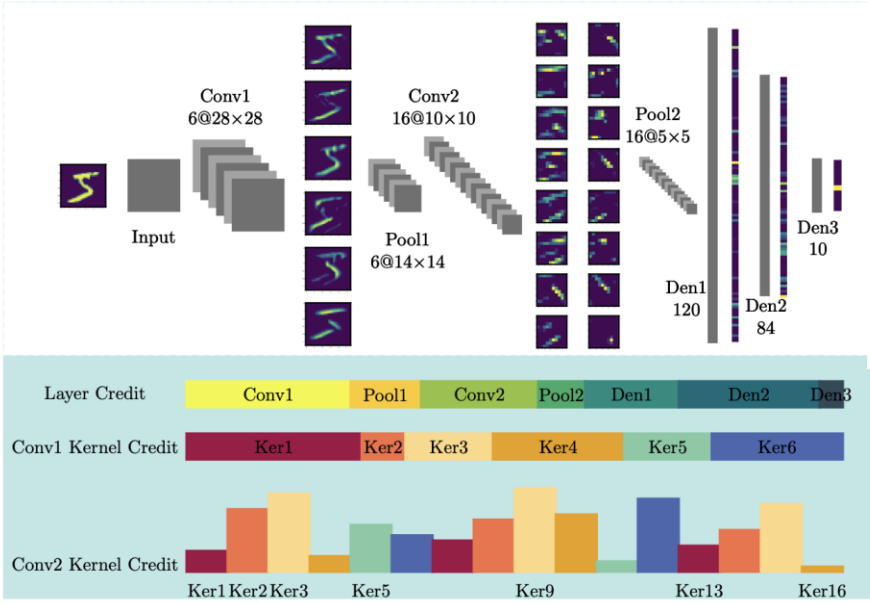
- Drawbacks of credit assginement problems (CAPs) of neural networks:
  - Most methods are utilized to solve CAPs on untrained neural networks, instead of trained neural networks.
  - Methods for trained neural networks are mainly limited in vision domains and lack of formal analysis and guarantee.
- Ideas: An alternative approach to CAPs of trained neural networks from a linear dynamical system perspective.

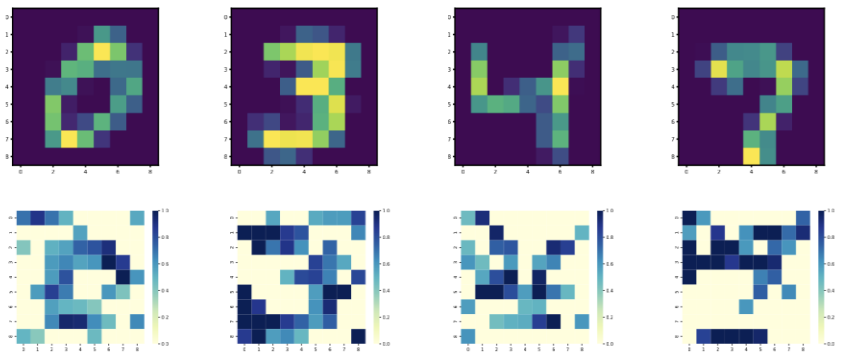


Workflows of tackling the credit assignment problems on trained neural networks.

Main Contributions & Results

- Contributions:
  - We migrate CAPs to trained neural networks and regard neural networks as dynamical systems to be linearly approxiamted with the Koopman Operator theory;
  - We present a minimal linear dimension alignment approach, which provides an insight into the dimension difference encountered in neural networks;
  - We define a credit metric with comprehensible algebraic explanation for assigning credits to network components.





Experimental results of credit assignment and learned feature maps. Left: Credit assignment of layers and convolution kernels in neural network Lenet-5; Top: Learned feature weights from a fully-connectd neural network.