Efficient and Flexible Neural Network Training Through Layer-wise Feedback Propagation

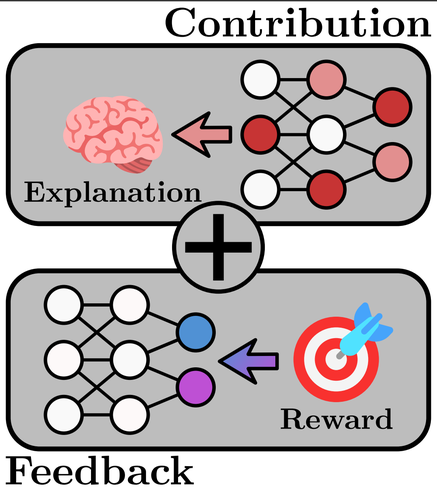
Leander Weber, Jim Berend, Moritz Weckbecker, Alexander Binder,
Thomas Wiegand, Wojciech Samek, Sebastian Lapuschkin

LFP
Layer-wise Feedback Propagation

TL;DR

Idea

Contribution

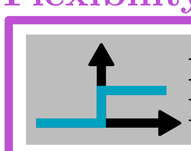


Feedback

We propose LFP, a novel **training paradigm** that leverages **Local XAI** to distribute a **Feedback Signal**

Strengths

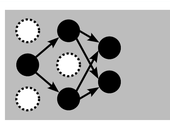
Flexibility



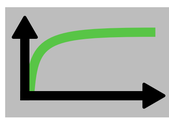
Non-Differentiable Models

Non-Differentiable Objectives

Efficiency



Sparsity



Fast Convergence

Main Results

Convergence properties similar to SGD

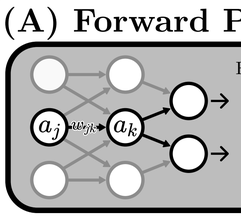
Approximation-free training of non-differentiable models

Improved model s**parsity** and **efficient** information representation

Promising for **energy-restricted applications** such as edge devices

Method

(A) Forward Pass



From input, compute **pre-activations** ...

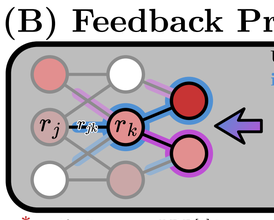
$$z_k = \sum_j w_{jk} a_j$$

... and **activations** ...

$$a_k = \phi_k(z_k)$$

... to generate an output.

(B) Feedback Propagation



Use **contributions*** to propagate **initial reward** ...

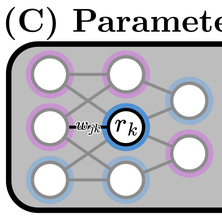
$$r_{jk} = \frac{w_{jk} a_j}{z_k} \cdot r_k$$

... and obtain **local rewards.**

$$r_j = \sum_k r_{jk}$$

*specifically, we use **LRP**[1] here, but other local XAI methods may work as well

(C) Parameter Update



Use **local reward** ...

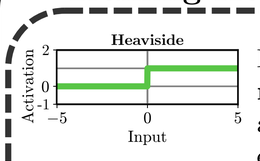
$$d = \frac{|w_{jk}| a_j}{z_k} \cdot r_k$$

... for a **Hebbian-like update**

$$w_{jk}^{\text{new}} = w_{jk} + \eta d$$

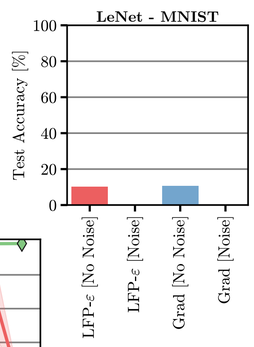
Results

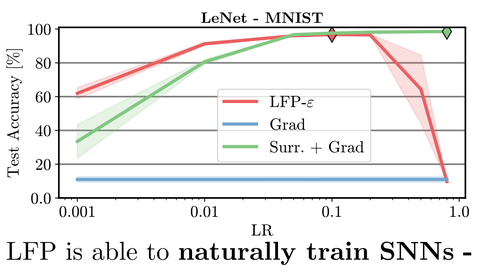
Training Heaviside-activated Models



Heaviside activations promise **memory- and energy-efficiency** and - but have **zero gradients** everywhere

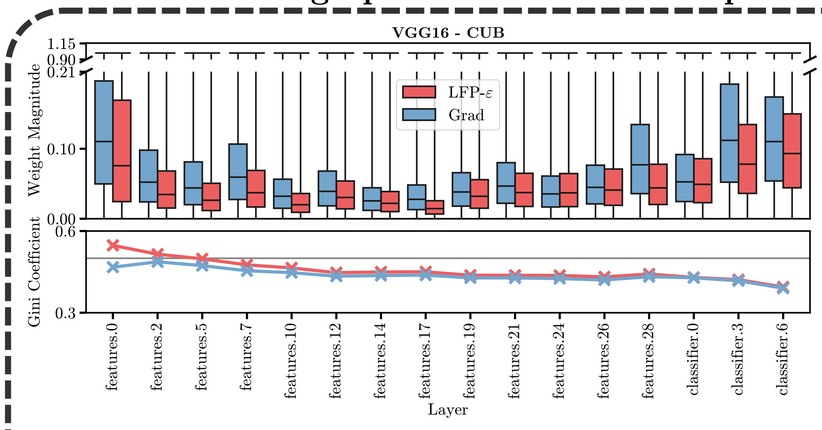
Training Heaviside-activated ANNs is **difficult** for backpropagation-based algorithms. However, adding noise to activations yields a **valid training signal** for LFP.





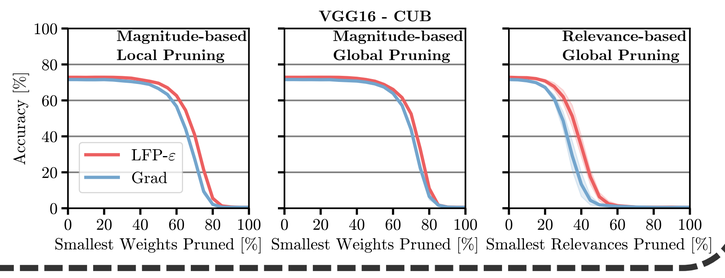
LFP is able to **naturally train SNNs - without requiring surrogates** for Heaviside in the backward pass.

Obtaining Sparse and Efficient Representations



Despite achieving **comparable accuracy**, parameters of LFP-trained models are **sparser**

Consequently, LFP-trained models can be **pruned more easily**, indicating more **efficient representation** of the same information



Check out the paper...

... and the code!

[1] Bach, Sebastian, et al. "**On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.**" PloS one 10.7 (2015): e0130140.
[2] Achtibat, Reduan, et al. "**AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers.**" Proceedings of the 41st International Conference on Machine Learning, in Proceedings of Machine Learning Research (2024): 235:135-168

ICLR