The Forward-Forward Algorithm: Some Preliminary Investigations [1]

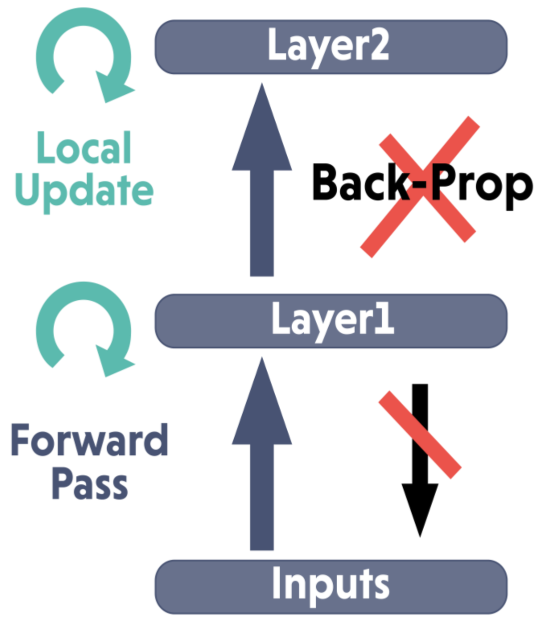
UNIVERSITY OF CAMBRIDGE

Author: Geoffrey Hinton

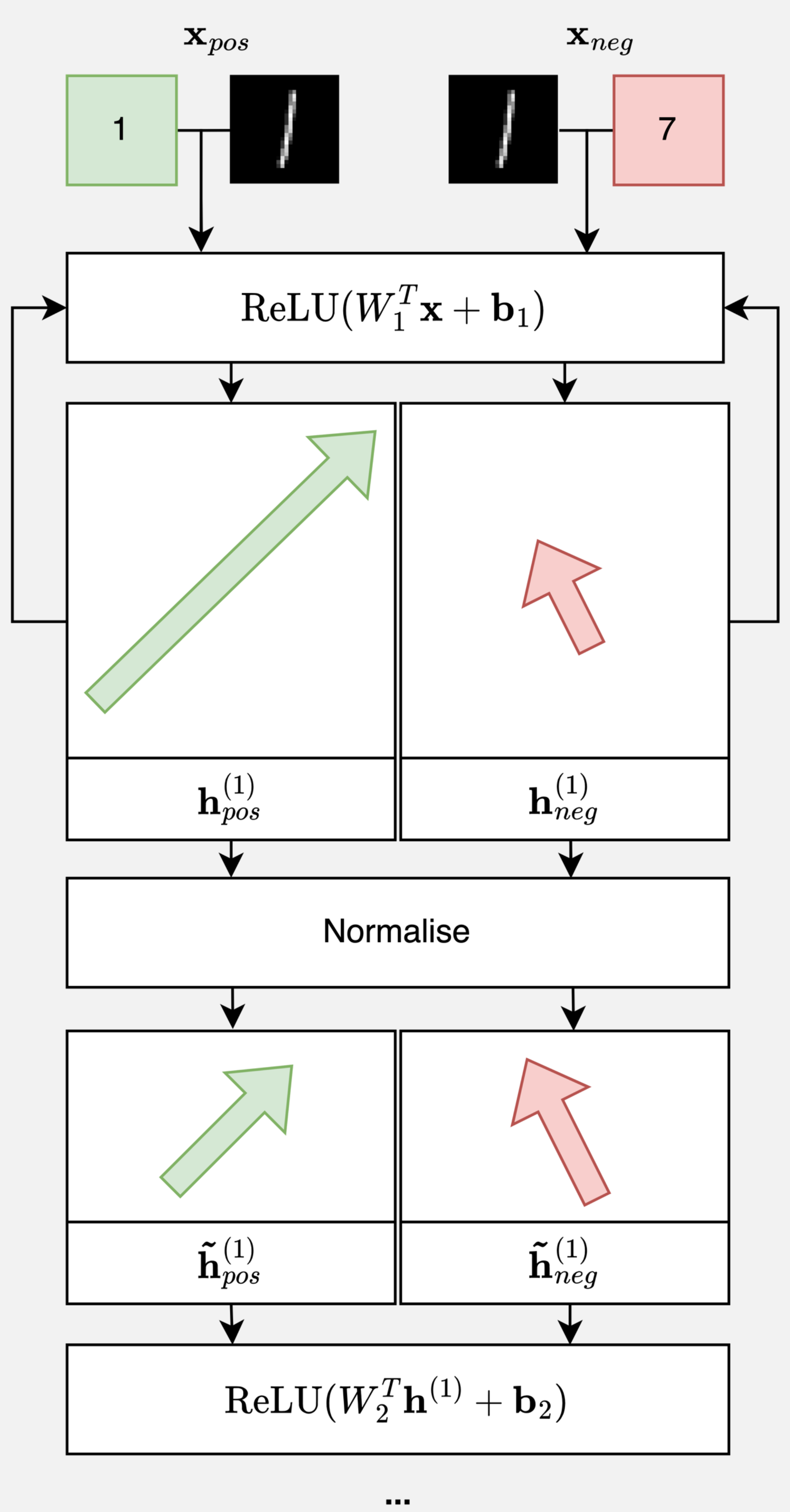
Poster: Jonas Scholz Sowmen Das Zixin Yang

Method

The Forward-Forward algorithm is an **alternative to backpropagation** that computes **local** gradient updates without the need for a backward pass..



Training



$$\mathcal{L} = \sigma(||\mathbf{h}^{(1)}_{neg}||_2 - \theta) - \sigma(||\mathbf{h}^{(1)}_{pos}||_2 - \theta)$$

$$g = ||\mathbf{h}||_2$$

Figure 1: During Training, maximize goodness for positive data and minimize goodness for negative data.

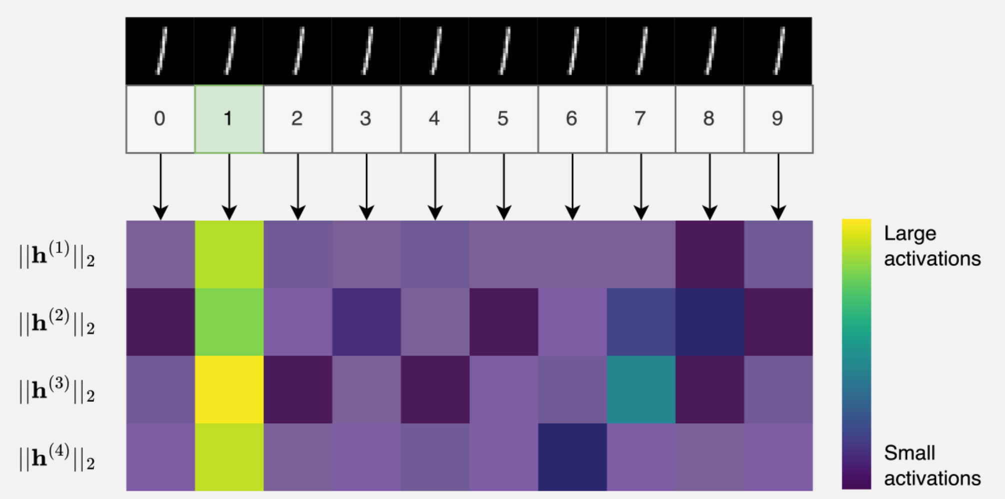
Motivations

- Local updates allow for asynchronous layer updates
- Biologically plausible
- Can be run on analog computers
- Can be easily adapted for unsupervised learning

Limitations

- Replace backpropagation outside of low-power environments
- Learns slower than backpropagation
- Lower layers do not receive higher-layer feedback

Inference



$$\hat{c} = \operatorname{argmax}_{c} \sum_{l=1}^{N_l} ||\mathbf{h}^{(l)}(c)||_2$$

Figure 2: During inference, the label leading to largest activations is chosen.

Findings

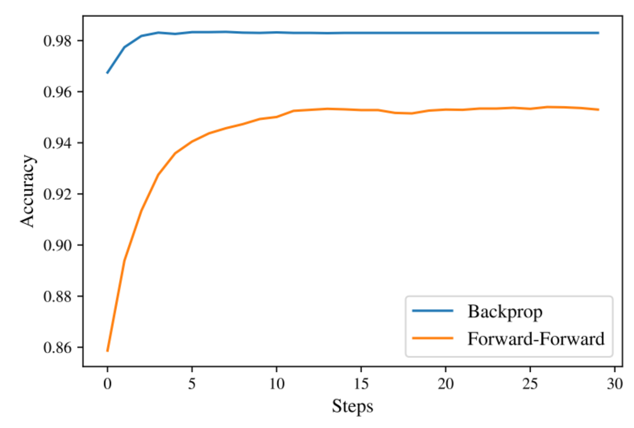


Figure 3: Performance of backpropagation vs Forward-Forward.

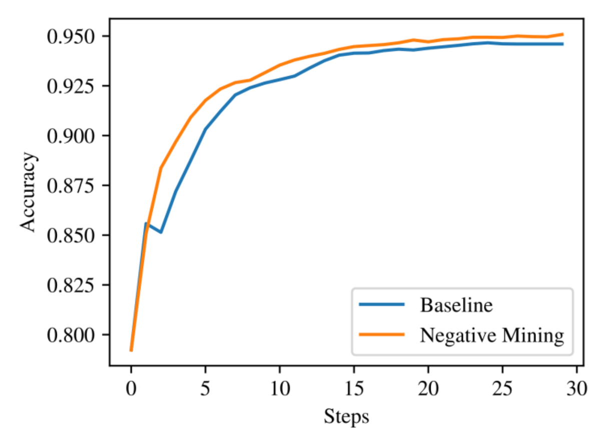


Figure 4: Hard negative mining improves accuracy.

First Layer Features

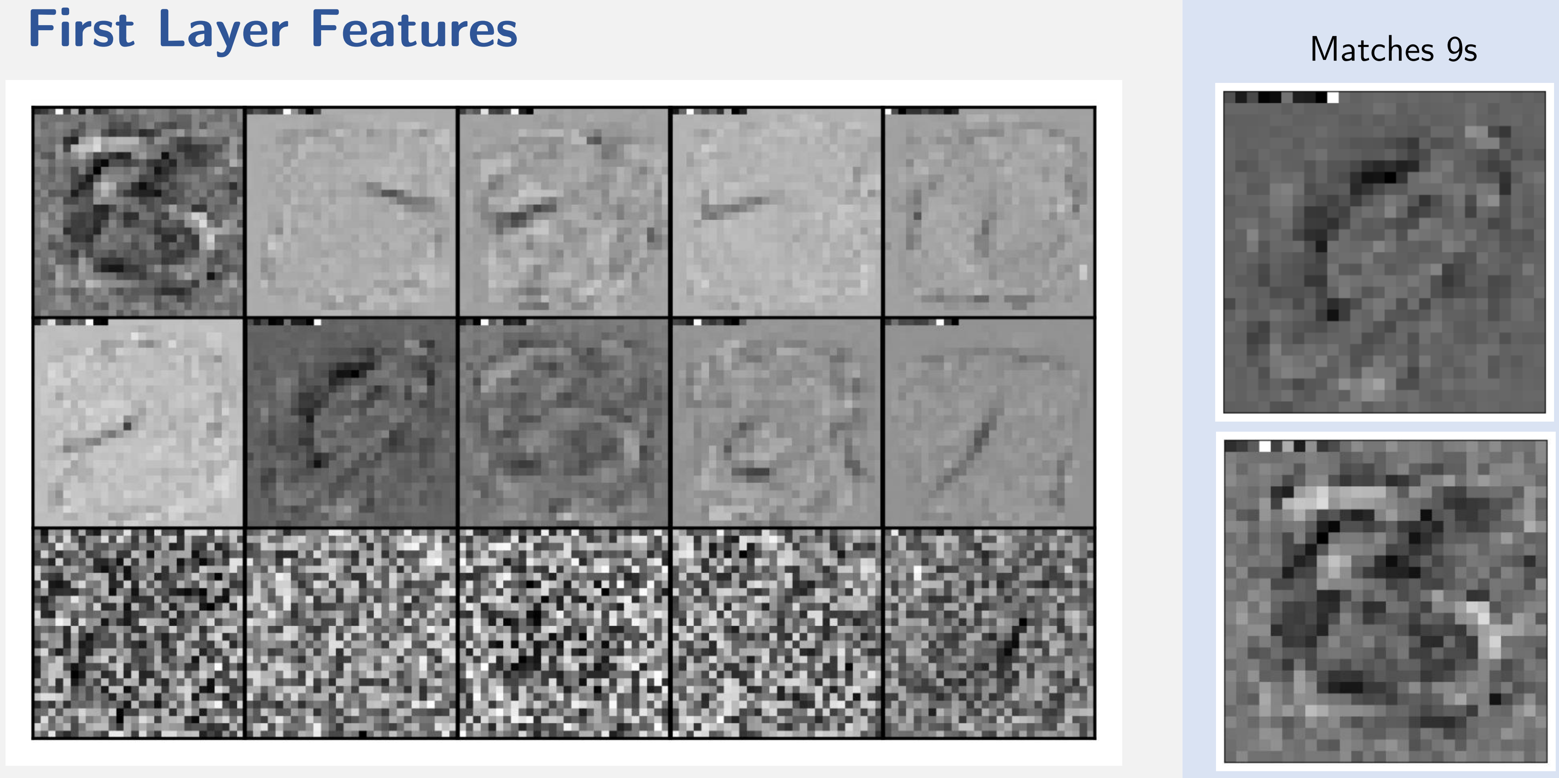


Figure 5: Some neurons learn sensible features (top), others do not (bottom).

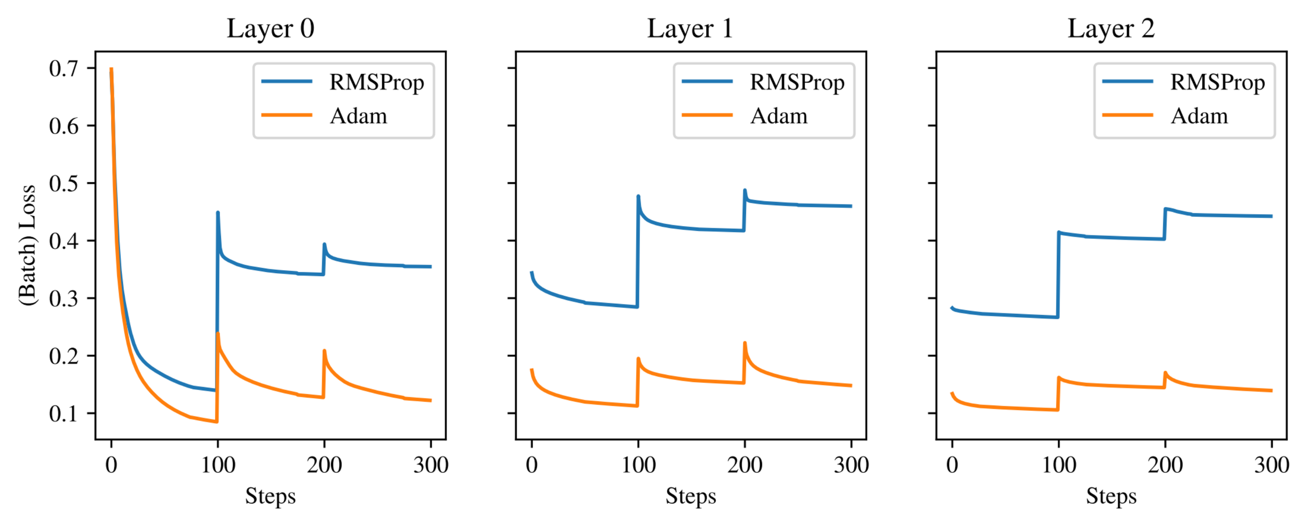


Figure 6: Adam significantly outperforms RMSProp, SGD never converges.

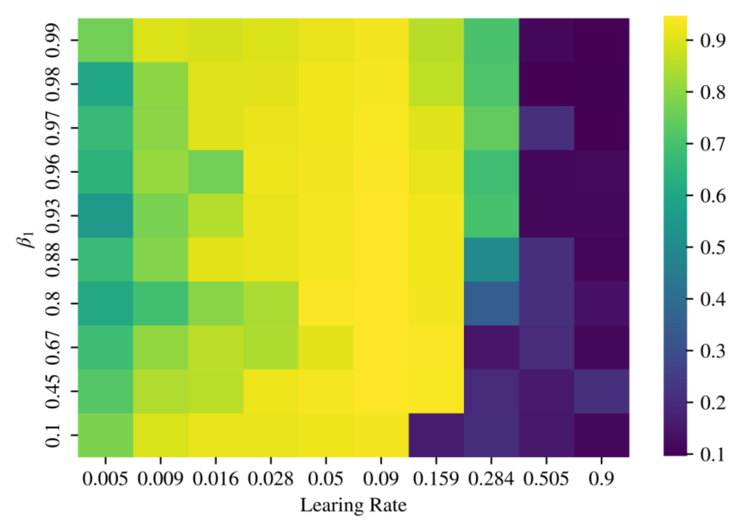


Figure 7: Model performance is sensitive to optimiser parameters, drops sharply.

Reproducibility Issues

- Original code not released
- Hyperparameters not specified
- Found high sensitivity to hyperparameters & specifically optimiser (SGD doesn't work)
- Achieved 96.8% accuracy vs Hinton's 99.2%

Extensions

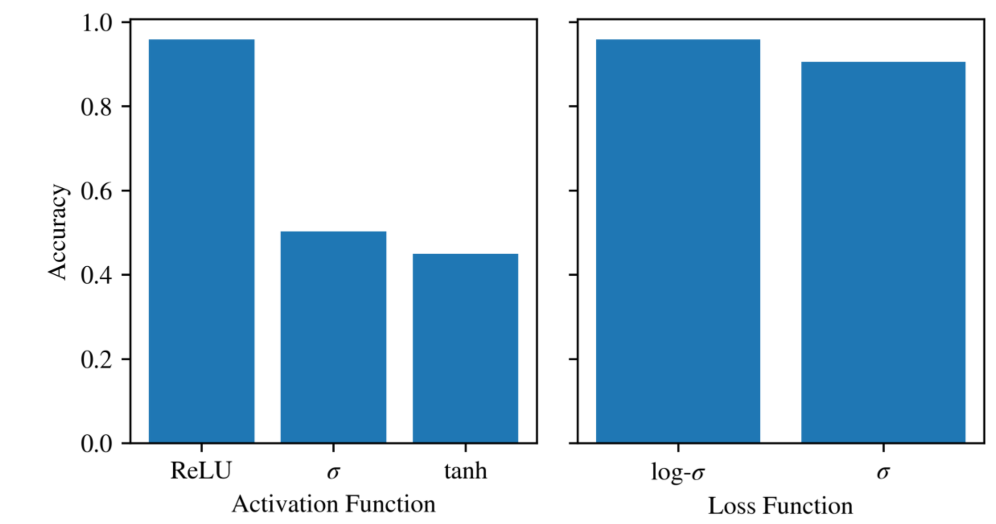


Fig. 8: Left: Activations functions that constrain activations perform worse.

Right: using log-sigmoid loss **improves upon the paper's sigmoid loss.**

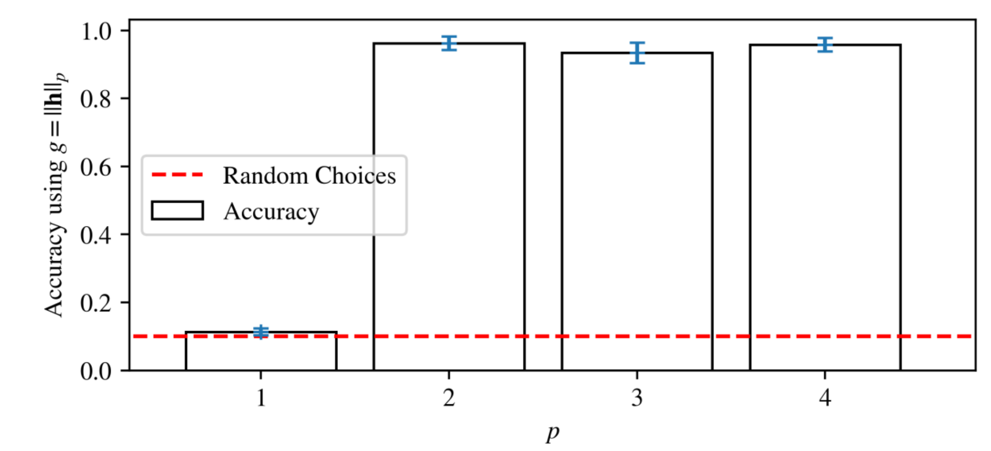


Fig. 9: The arbitrarily chosen goodness function (L2 norm) performs similarly to higher power L-norms.

Future Work

Replicate and explore the recurrent version of FF.

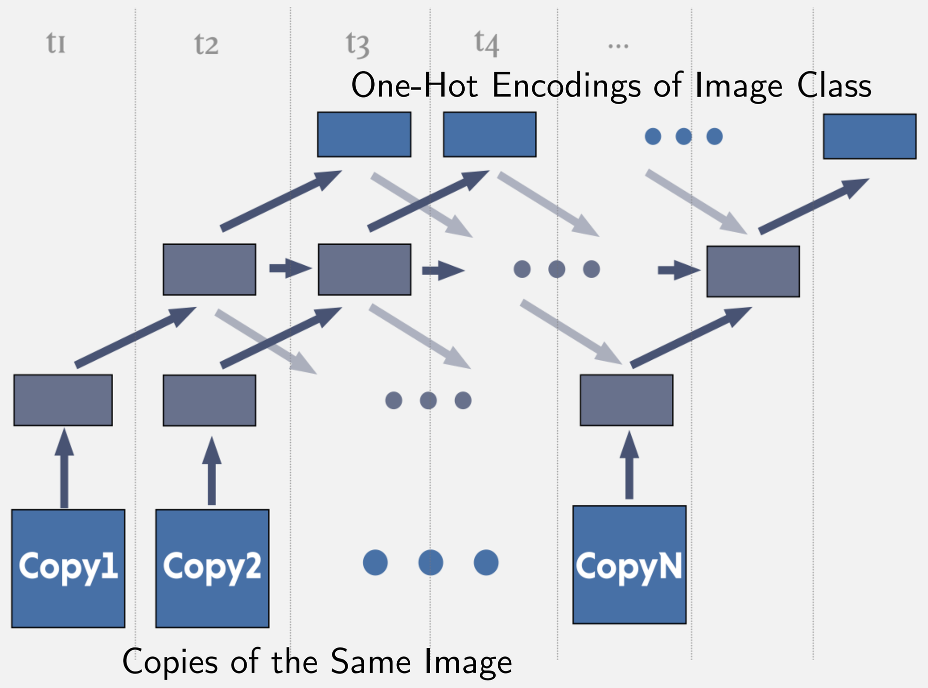


Figure 10: Recurrent architecture that resolves the major weakness of FF: copies of static image form a 'video' that enables learning of later layers from earlier layers.

References

[1] Hinton, Geoffrey. "The forward-forward algorithm: Some preliminary investigations." *arXiv preprint arXiv:2212.13345* (2022).

Scan to view original paper