UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

LightFF: Lightweight Inference for Forward-Forward Algorithm

LUND UNIVERSITY

Amin Aminifar*, Baichuan Huang†, Azra Abtahi†, Amir Aminifar†

* Institute of Computer Engineering, Heidelberg University, Germany

† Department of Electrical and Information Technology, Lund University, Sweden

Motivation & Research Goals

The state-of-the-art Artificial/Deep Neural Networks (ANN/DNN) consume massive amounts of energy. A prime example is GPT-3, which consumes over **1000 megawatt-hours** for training alone, equivalent to a small town's power consumption for a day [1]. The training of these ANNs/DNNs is done almost exclusively based on the back-propagation (BP) algorithm, which is known to be biologically implausible [2]. This has led to several biologically plausible alternatives, e.g., the Forward-Forward (FF) algorithm [3]. The majority of the state-of-the-art studies based on FF have mainly focused on training. However, the inference over already-trained models also consumes a massive amount of energy, e.g., accounting for around **60% of the total machine learning energy used at Google** [4]. In this work, we propose a lightweight inference algorithms for neural networks trained based on the Forward-Forward (FF) algorithm [3].

Insight

The key insight is that the local energy-based techniques provide a strong intermediate measure to decide whether the local energy or the goodness is sufficient to make a confident decision, without the need to complete the entire forward pass. This is inspired by the human nervous system. For instance, the reflexes do not pass directly into the brain, but synapse in the spinal cord. At the same time, the complex inputs that require detailed analysis are processed by the brain.

Method

In this paper, we propose a lightweight inference scheme, called LightFF, specifically designed for DNNs trained using the Forward-Forward algorithm [3]. For our lightweight inference, instead of performing the forward pass through all layers, once the operations for each layer are completed, we inspect the confidence level of the result, and based on that, we decide whether to continue the forward pass. Fig. 1 shows our lightweight inference scheme. The proposed scheme is effective because the difficulty of the classification task varies from one test sample to another. As shown, the first layer(s) is sufficient for straightforward test samples, while for other samples, more layers may be required.

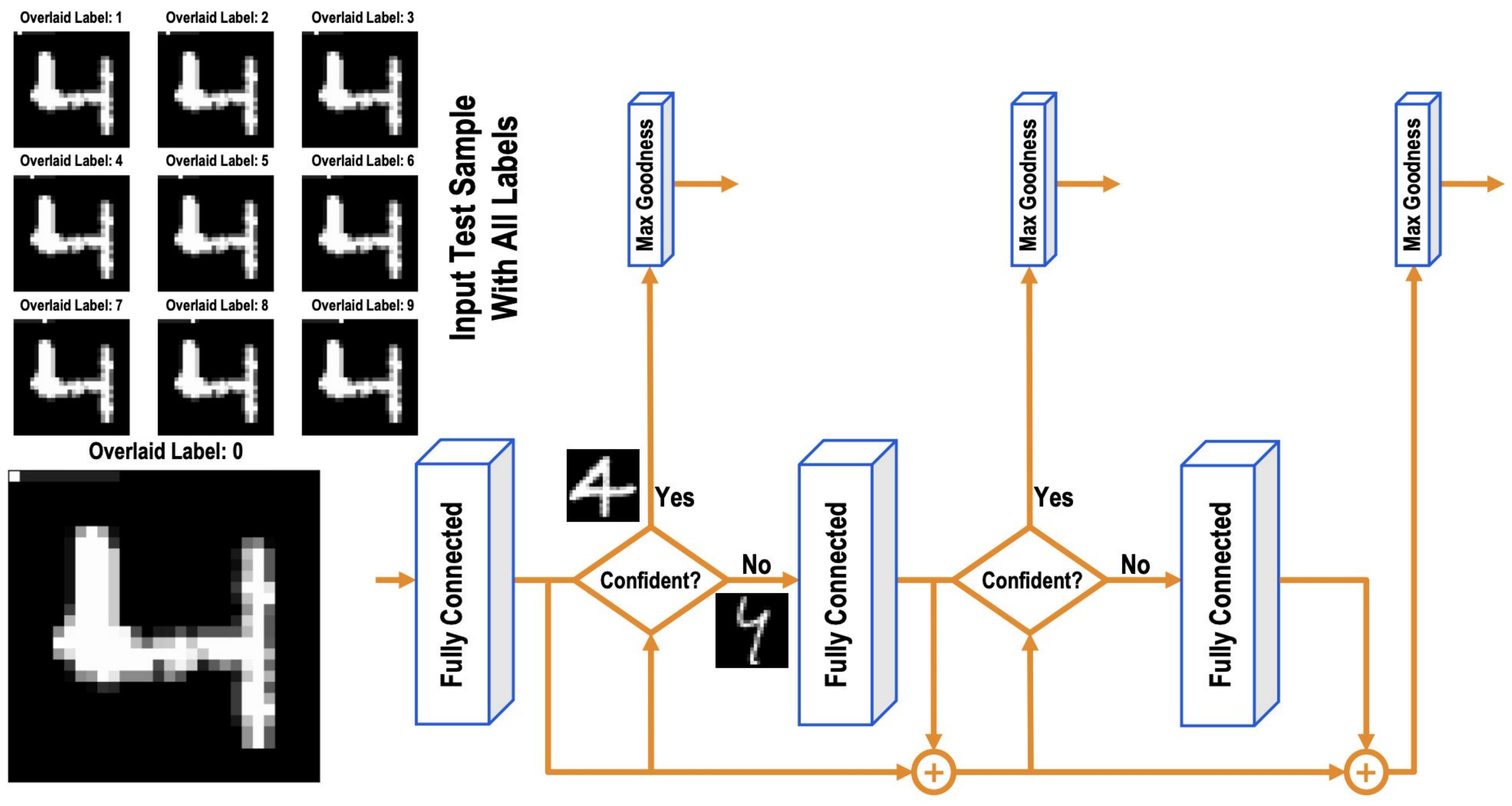

Fig. 1: Light Multi-Pass Inference

The distance between the mean values of the goodness increases for the negative and positive data as we consider more layers, as shown in Fig. 2.

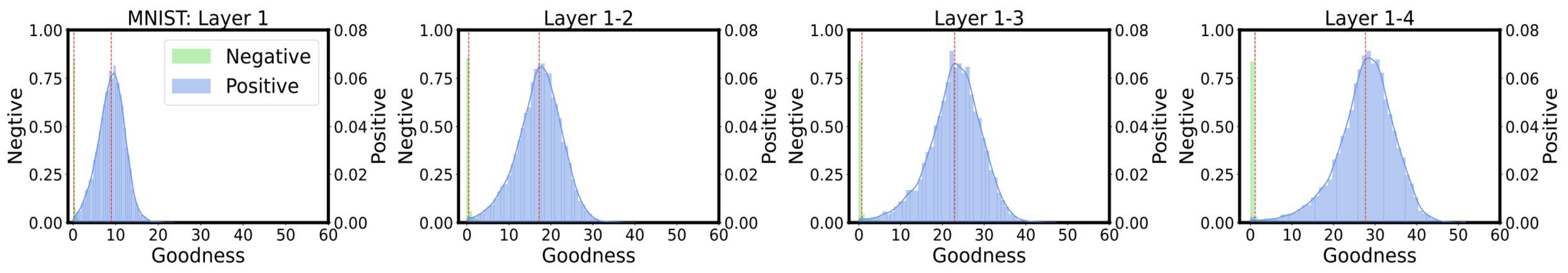

Fig. 2: Goodness Distribution

Results

We evaluate LightFF in terms of prediction performance (Error), computational complexity (MACs), and the actual execution time. LightFF improves computational efficiency with a comparable classification error, compared to the Forward-Forward algorithm.

Dataset	Error		MACs		Execution Time	
	FF	LightFF	FF	LightFF	FF	LightFF
MNIST	**1.51%**	**1.51%**	12.94M	**2.95M**	22.81ms	**3.39ms**
CIFAR-10	50.65%	**46.05%**	17.30M	**10.31M**	29.57ms	**16.38ms**
CHB-MIT	39.62%	**34.83%**	13.39M	**8.39M**	4.73ms	**2.85ms**
MIT-BIH	10.74%	**10.07%**	11.90M	**1.99M**	11.26ms	**1.83ms**

We apply the proposed lightweight inference scheme for a network trained based on BP. In this case, we generally observe a degradation in the classification performance of networks trained using BP, as shown below.

Dataset	BP	Light-BP	FF (OP)	LightFF (OP)
MNIST	**1.33%**	5.21%	1.53%	**1.02%**
CIFAR-10	**43.62%**	54.22%	47.78%	**46.25%**
CHB-MIT	**25.63%**	40.69%	37.98%	**28.23%**
MIT-BIH	**8.25%**	11.55%	10.86%	**10.54%**

LightFF decreases the mean number of layers used in inference.

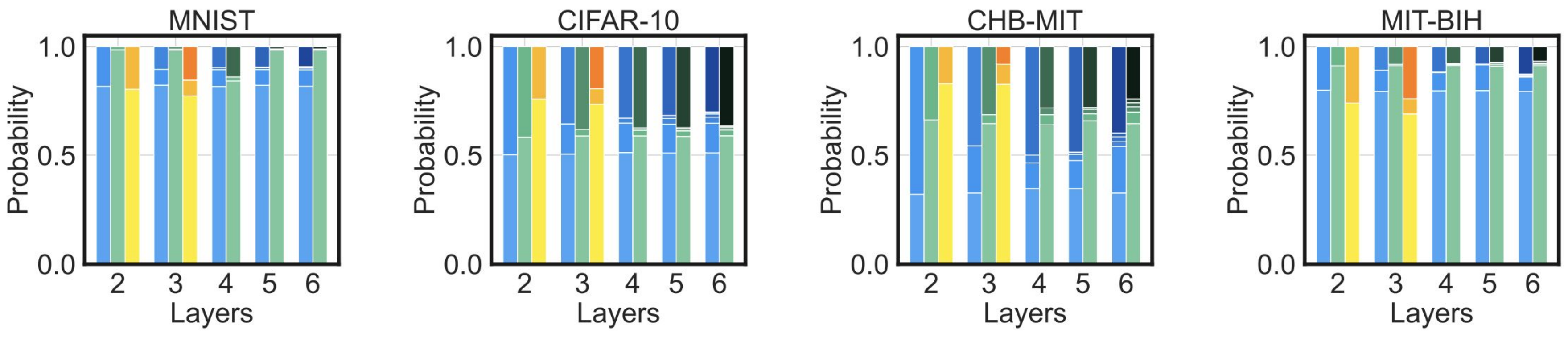

Fig. 3: Probability of Each Layer

References

[1] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon emissions and large neural network training, 2021.

[2] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. Nature communications, 7(1):13276, 2016.

[3] G. Hinton. The forward-forward algorithm: Some preliminary investigations. arXiv preprint arXiv:2212.13345, 2022.

[4] Patterson, David, et al. "The carbon footprint of machine learning training will plateau, then shrink." *Computer* 55.7 (2022): 18-28.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Swedish Research Council