ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Entropy Based Regularization Improves Performance in the Forward-Forward Algorithm

Matteo Pardi[1], Domenico Tortorella[2] and Alessio Micheli[2]

1- University of Pisa - Department of Physics 'E. Fermi'
Largo B. Pontecorvo, 3 56127 Pisa - Italy

2- University of Pisa - Department of Computer Science
Largo B. Pontecorvo, 3 56127 Pisa - Italy

Abstract. The forward-forward algorithm (FFA) is a recently proposed alternative to end-to-end backpropagation in deep neural networks. FFA builds networks greedily layer by layer, thus being of particular interest in applications where memory and computational constraints are important. In order to boost layers' ability to transfer useful information to subsequent layers, in this paper we propose a novel regularization term for the layer-wise loss function that is based on Renyi's quadratic entropy. Preliminary experiments show accuracy is generally significantly improved across all network architectures. In particular, smaller architectures become more effective in addressing our classification tasks compared to the original FFA.

1 Introduction

In the last decade, deep learning has encountered unquestionable success in a vast range of learning tasks, with increasingly larger deep neural networks trained on big datasets via end-to-end (E2E) backpropagation. However, there are application settings where particular requirements such as limited memory or energy resources make deep learning unfeasible [1, 2]. Additionally, the difficulty to train deep neural networks E2E due to problems such as exploding/vanishing gradient or slow convergence motivates the demand for alternative approaches. Hinton [3] recently proposed the forward-forward algorithm (FFA), a method that incrementally constructs the network by greedily training a layer and freezing it before moving to add the next one, thus avoiding deep backpropagation and the issues it entails. FFA is inspired by Boltzmann machines [4] and contrastive learning [5], as each layer is trained to discriminate real ('good') data samples from fake ('bad') ones. This learning method is also claimed to be more biologically plausible than backpropagation [6, 7]. However, the advantages of reduced computational demands provided by FFA may come at the expense of expressive hidden representations, since hidden layers are not optimally trained to transfer information from input layers to deeper layers as in E2E learning. In this paper, we propose to address this drawback by forcing the hidden representations learned by FFA to also improve the mutual information with the layer inputs. This is achieved by introducing a regularization term in the original FFA loss function based on Renyi's quadratic entropy (sec. 3). The preliminary experimental investigation carried on three binary and multi-class classification

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

tasks show significant and consistent improvements in accuracy across different neural network architectures with respect to the original FFA method (sec. 4).

2 Forward-Forward Algorithm

The Forward-Forward Algorithm (FFA) constructs a deep neural network by greedily training one layer at a time to maximize the response to *good* (or positive) data and minimize it on *bad* (or negative) data [3], in an approach similar to contrastive learning [5]. For a classification task on a set of labeled data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $y_i \in \{1, ..., C\}$, positive samples as generated by concatenating $\mathbf{x}_i$ with the one-hot encoding of the correct class y_i, while negative samples are generated as incorrect pairings. The first layer $l = 1$ learns a feature transformation $\mathbf{h}_i^{(1)} = f_1([\mathbf{x}_i \parallel \mathbf{y}_i])$ that maximizes the *goodness* value

$$g_i^{(1)} \propto \|\mathbf{h}_i^{(1)}\|^2 \tag{1}$$

for correct input pairings. In practice, this is done by treating the goodness value $g_i^{(1)}$ akin to the logits of a binary classifier $\hat{p} = \text{Sigmoid}(g_i^{(l)} - \theta)$ trained to discriminate positive samples from negative ones by minimizing the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$. This classifier is discarded after each layer training. Each subsequent layer $l = 2, ..., L$ is trained in a similar manner to learn features $\mathbf{h}^{(l)} = f_l(\bar{\mathbf{h}}^{(l-1)})$ from the previous layer representations normalized as $\bar{\mathbf{h}}^{(l-1)} \propto \mathbf{h}^{(l-1)}/\|\mathbf{h}^{(l-1)}\|$. The purpose of this normalization is to avoid reliance on the goodness learned by the layer predecessor and thus to exploit information not used by layer $l - 1$. The total goodness of the network $\bar{g}_i = \sum_{l=1}^L g_i^{(l)}$ is used to predict the class $\hat{y}_i$ of an unlabelled sample $\mathbf{x}_i$ as

$$\hat{y}_i = \text{argmax}_{1 \leq k \leq C}\ \bar{g}([\mathbf{x}_i \parallel \mathbf{1}_k]), \tag{2}$$

where $\mathbf{1}_k$ is the one-hot encoding of class k and $\bar{g}(\cdot)$ is the total goodness as a function of neural network input. It is actually suggested by [3] to exclude the first layer from the total goodness and to use instead $\bar{g}_i^\star = \sum_{l=2}^L g_i^{(l)}$. While the layer trained by FFA can have any form [8], in our experiments we adopt simple fully-connected layers $\mathbf{x} \mapsto \text{ReLU}(\mathbf{W}^{(l)}\,\mathbf{x} + \mathbf{b}^{(l)})$ for tasks on vector data.

3 Entropy-based regularization

To obtain an effective hierarchy of hidden representations in the neural network, the FFA greedy layer training should accomplish several objectives: (i) ensure that useful information is transferred from the network input data to all subsequent layer representations; (ii) avoid redundancy in each layer representation due to high correlations among unit activations; (iii) differentiate the representations learned by each layer. While the latter objective is addressed by the normalization of input features for layers $l > 1$, objectives (i) and (ii) cannot be accomplished simply by minimizing $\mathcal{L}_{\text{BCE}}$ and by relying on different initializations of unit weights.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Therefore, we propose to maximize the mutual information $I(\mathbf{h}^{(l)}, \bar{\mathbf{h}}^{(l-1)})$ between the layer representation $\mathbf{h}^{(l)}$ and the layer input features $\bar{\mathbf{h}}^{(l-1)}$ (for $l = 1$ we assume $\bar{\mathbf{h}}^{(0)} = [\mathbf{x} \parallel \mathbf{y}]$). For the mutual information defined from Shannon entropy $H(\cdot)$, it holds that $I(\mathbf{h}^{(l)}, \bar{\mathbf{h}}^{(l-1)}) = H(\mathbf{h}^{(l)}) - H(\mathbf{h}^{(l)}|\bar{\mathbf{h}}^{(l-1)})$, where $H(\cdot|\cdot)$ is the conditional entropy [9]. Since $\mathbf{h}^{(l)}$ is computed by a non-stochastic map, $H(\mathbf{h}^{(l)}|\bar{\mathbf{h}}^{(l-1)}) = 0$, and it is thus sufficient to maximize the entropy $H(\mathbf{h}^{(l)})$. Therefore, we change the FFA layer loss to be minimized in

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} - T \cdot H(\mathbf{h}^{(l)}), \tag{3}$$

were we call the parameter $T > 0$ *temperature*, since from a thermodynamic point of view $\mathcal{L}_{\text{BCE}}$ can be considered the energy and $\mathcal{L}$ the Helmholtz free energy. For computational efficiency we actually adopt the Renyi quadratic entropy $H_2(\mathbf{h}^{(l)})$ instead of Shannon entropy in equation (3), being $H(\mathbf{h}^{(l)}) \geq H_2(\mathbf{h}^{(l)})$. Given a batch of M samples $\{\mathbf{h}_i^{(l)}\}_{i=1}^M$ with $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$, Renyi quadratic entropy is estimated via the Parzen–Rosenblatt kernel density estimator [10] as

$$H_2 \approx -\log \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{G}_{\mathbf{S}}(\mathbf{h}_i^{(l)} - \mathbf{h}_j^{(l)}), \tag{4}$$

where $\mathcal{G}_{\mathbf{S}}(\cdot)$ is the Gaussian multi-variate kernel with bandwidth $\mathbf{S}$. We make the simplifying assumptions of normally distributed activations and diagonal $\mathbf{S}$, thus choosing $\mathbf{S} \approx (4/M(d+2))^{2/(d+4)} s^2 \mathbf{I} \approx (1/M)^{2/(d+4)} s^2 \mathbf{I}$ in accordance with Silverman's rule. The parameter $s > 0$ is called the *kernel scale*, and controls the smoothness of the density estimation in (4). Indeed, for $s \to \infty$ we obtain an uniform distribution, while for $s \to 0$ we obtain the empirical sample distribution; the choice of s has thus also an effect on mini-batch variance during stochastic gradient descent. Both temperature T and kernel scale s will be tuned via model selection in our experiments.

4 Experiments and discussion

We compare the accuracy of neural networks trained by FFA with the original loss $\mathcal{L}_{\text{BCE}}$ against training with our proposed entropy regularization. We perform experiments on two binary classification tasks and one multi-class classification task for different network architectures. Apart from the accuracy of predictions computed by equation (2) with total goodness from all layers $\bar{g}$ and from all layers except the first $\bar{g}^\star$, we also report the accuracy of each individual layer; model selection is performed on $\bar{g}^\star$ following [3]. Average and standard deviation are over 5 trials. Model selection is performed for weight decay for FFA, and additionally on temperature T and kernel scaling s for FFA+ENTROPY. Code and details to reproduce our experiments are publicly available online.[1]

The double moon task [11] consists in classifying the points in the 2D plane belonging to two sets shaped like two intertwined moons (see Fig. 1). This

[1] `https://github.com/MatteoPardi/Entropy-FFA`

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

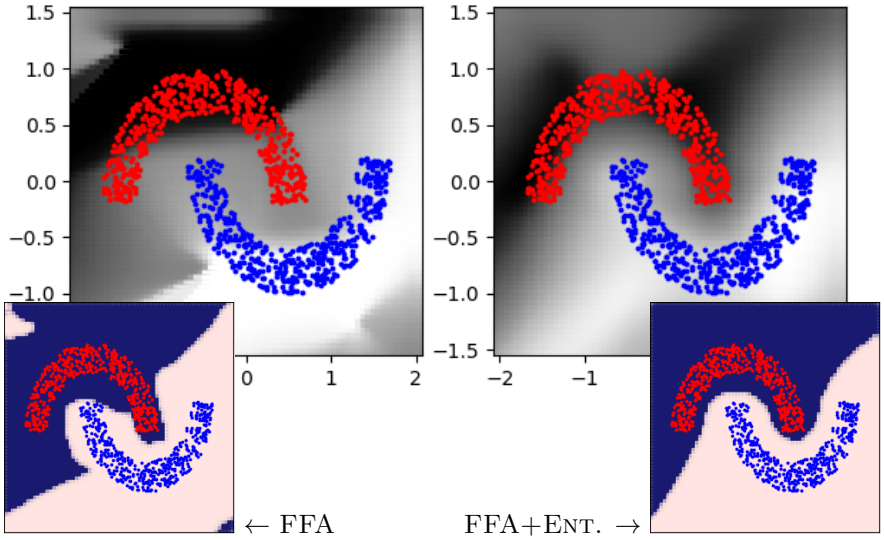

Fig. 1: Double moon task: goodness (top) and decision regions (bottom).

Layer	FFA 10	+Ent. 10	FFA 100	+Ent. 100	FFA 1000	+Ent. 1000
all	$90.0_{\pm 0.1}$	$\mathbf{99.2}_{\pm 0.7}$	$97.5_{\pm 1.8}$	$\mathbf{100.0}_{\pm 0.0}$	$97.2_{\pm 0.4}$	$\mathbf{100.0}_{\pm 0.0}$
1	$90.2_{\pm 0.6}$	$\mathbf{91.0}_{\pm 0.1}$	$\mathbf{90.5}_{\pm 0.1}$	$\mathbf{90.5}_{\pm 0.1}$	$\mathbf{86.5}_{\pm 0.1}$	$\mathbf{86.5}_{\pm 0.1}$
2	$90.2_{\pm 0.3}$	$\mathbf{99.6}_{\pm 0.2}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$
3	$63.3_{\pm 18.2}$	$\mathbf{99.5}_{\pm 0.6}$	$98.2_{\pm 2.2}$	$\mathbf{100.0}_{\pm 0.0}$	$94.0_{\pm 2.3}$	$\mathbf{100.0}_{\pm 0.0}$
$\star$ (2+3)	$89.2_{\pm 0.5}$	$\mathbf{99.7}_{\pm 0.4}$	$99.9_{\pm 0.2}$	$\mathbf{100.0}_{\pm 0.0}$	$98.2_{\pm 1.1}$	$\mathbf{100.0}_{\pm 0.0}$

Table 1: Test accuracy for the double moon binary classification task.

binary classification task showcases the ability of neural networks to learn decision boundaries in non-linearly separable tasks. We adopt fixed scaffold splits 640/160/200 of training/validation/test, training up to 3 layers of $10, 100, 1000$ hidden units. The results reported in Tab. 1 show that adding our entropy regularization leads to significant improvements in accuracy especially on smaller and deeper networks. Notice also how the accuracy of goodness in the third layers of FFA drops significantly, while FFA+Entropy is able to preserve it. In Fig. 1 we can also appreciate the shape of the decision boundary learned by FFA with entropy regularization, leading to more robust classification particularly in the concave regions of the two moons.

We replicate the same experimental setting for the noisy double moon task, where noise sampled from a gaussian distribution of std 0.16 was added to the original points. The results of Tab. 2 show that FFA+Entropy is more robust to noise, consistently and significantly improving accuracy with respect to FFA. This improvement in robustness can also be observed particularly in the concave region of the red moon in Fig. 2.

Our final task is the 10 digit classification task MNIST, where original 28×28 gray-scale images have been flattened into vectors. We follow the setting of [3],

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

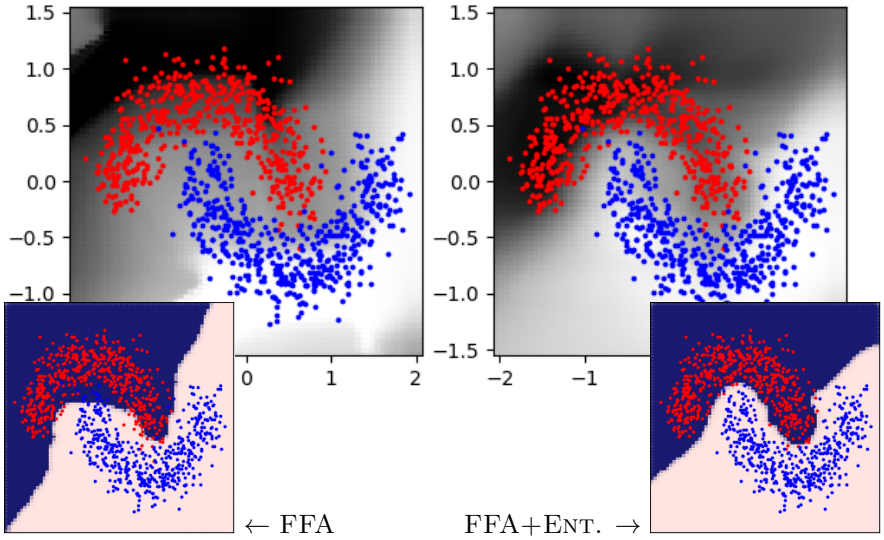

Fig. 2: Noisy double moon task: goodness (top) and decision regions (bottom).

Layer	FFA 10	+Ent. 10	FFA 100	+Ent. 100	FFA 1000	+Ent. 1000
all	$89.3_{\pm 0.3}$	$\mathbf{95.2}_{\pm 0.6}$	$95.9_{\pm 0.2}$	$\mathbf{96.8}_{\pm 0.4}$	$96.3_{\pm 0.3}$	$\mathbf{97.5}_{\pm 0.1}$
1	$\mathbf{89.6}_{\pm 1.0}$	$\mathbf{89.9}_{\pm 0.9}$	$\mathbf{90.2}_{\pm 0.4}$	$\mathbf{90.7}_{\pm 0.3}$	$\mathbf{87.5}_{\pm 0.1}$	$\mathbf{87.7}_{\pm 0.3}$
2	$90.0_{\pm 0.6}$	$\mathbf{97.6}_{\pm 0.5}$	$96.1_{\pm 0.2}$	$\mathbf{98.2}_{\pm 0.3}$	$96.6_{\pm 0.2}$	$\mathbf{98.0}_{\pm 0.1}$
3	$58.3_{\pm 11.9}$	$\mathbf{97.8}_{\pm 0.4}$	$96.1_{\pm 0.2}$	$\mathbf{98.2}_{\pm 0.3}$	$95.7_{\pm 1.2}$	$\mathbf{97.8}_{\pm 0.3}$
⋆ (2+3)	$89.4_{\pm 0.4}$	$\mathbf{98.0}_{\pm 0.4}$	$96.5_{\pm 0.1}$	$\mathbf{98.1}_{\pm 0.2}$	$96.6_{\pm 0.4}$	$\mathbf{98.0}_{\pm 0.1}$

Table 2: Test accuracy for the noisy double moon binary classification task.

adopting fixed scaffold training/validation/test splits of 50,000/10,000/10,000 samples, and training up to 3 layers of $20, 200, 2000$ hidden units. The results reported in Tab. 3 for this multi-class task confirm the trends we have observed so far. Our entropy-based regularization consistently boosts the accuracy of FFA, offering significant improvements in small- and medium-sized architectures, while preserving the ability to learn effective hidden features also in the deepest layer ($l = 3$), compared to FFA. By enhancing the ability of smaller networks to learn better hidden representations, the proposed method offers a better trade-off between accuracy and computational resources required by neural networks.

L.	FFA 20	+Ent. 20	FFA 200	+Ent. 200	FFA 2000	+Ent. 2000
all	$92.86_{\pm 0.15}$	$\mathbf{93.36}_{\pm 0.21}$	$97.93_{\pm 0.10}$	$\mathbf{98.22}_{\pm 0.08}$	$\mathbf{98.52}_{\pm 0.07}$	$\mathbf{98.61}_{\pm 0.06}$
1	$\mathbf{92.42}_{\pm 0.12}$	$\mathbf{92.50}_{\pm 0.17}$	$\mathbf{97.35}_{\pm 0.12}$	$\mathbf{97.35}_{\pm 0.07}$	$96.93_{\pm 0.03}$	$\mathbf{97.36}_{\pm 0.04}$
2	$\mathbf{92.56}_{\pm 0.32}$	$\mathbf{92.92}_{\pm 0.10}$	$97.91_{\pm 0.10}$	$\mathbf{98.07}_{\pm 0.11}$	$\mathbf{98.55}_{\pm 0.06}$	$\mathbf{98.56}_{\pm 0.06}$
3	$87.47_{\pm 1.76}$	$\mathbf{92.39}_{\pm 0.24}$	$95.59_{\pm 0.19}$	$\mathbf{97.95}_{\pm 0.08}$	$97.74_{\pm 0.17}$	$\mathbf{98.39}_{\pm 0.11}$
⋆	$92.42_{\pm 0.26}$	$\mathbf{93.21}_{\pm 0.12}$	$97.79_{\pm 0.14}$	$\mathbf{98.23}_{\pm 0.09}$	$\mathbf{98.53}_{\pm 0.07}$	$\mathbf{98.58}_{\pm 0.08}$

Table 3: Test accuracy for MNIST 10 digits classification task.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

5 Conclusion

In this paper we have proposed a novel method to enhance the training of deep neural network layers for the Forward-Forward Algorithm. Our method is based on the introduction of a regularizing term in the loss function based on Renyi's quadratic entropy, with the objectives of ensuring the transfer of information through the neural network layers and of improving the quality of hidden representations. In our experiments on binary and multi-class classification tasks on vector data, our method has achieved significant and consistent improvements in accuracy across different network architectures. Particularly valuable are the performance improvements on smaller networks, thus presenting a better trade-off between accuracy and computational resources required by the models. This is significant for resource-constrained applications, such as edge computing or Internet-of-Things [1, 2]. In future works we will explore entropy-based regularization on other types of tasks such as node and graph classification [8], and on other incremental network construction algorithms beyond FFA [12, 13].

Acknowledgments Research partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] B. Sliwa, N. Piatkowski, and C. Wietfeld. LIMITS: Lightweight machine learning for iot systems with resource limitations. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020.

[2] T. Park, N. Abuzainab, and W. Saad. Learning how to communicate in the Internet of Things: Finite resources and heterogeneity. *IEEE Access*, 4:7063–7073, 2016.

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

[4] G. E Hinton, T. J Sejnowski, et al. Learning and relearning in Boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986.

[5] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th international Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[6] W. Illing, B.and Gerstner and J. Brea. Biologically plausible deep learning-but how far can we go with shallow networks? *Neural Networks*, 118:90–101, 2019.

[7] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

[8] D. Paliotta, M. Alain, B. Máté, and F. Fleuret. Graph neural networks go forward-forward. *arXiv preprint arXiv:2302.05282*, 2023.

[9] T. Cover and J. Thomas. *Elements of Information Theory, 2nd Ed.* Wiley, 2006.

[10] J. Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives.* Springer Science & Business Media, 2010.

[11] S. Haykin. *Neural networks and learning machines, 3rd Ed.* Pearson Education, 2009.

[12] A. Micheli. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 20(3):498–511, 2009.

[13] E. S. Marquez, J. S. Hare, and M. Niranjan. Deep cascade learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5475–5485, 2018.