Implementation of Noise-Resistant Crowd Equalisation in Optical Communication Systems with Machine Learning DSP

Karina Nurlybayeva, Diego Argüello Ron, Morteza Kamalian-Kopae, Elena Turitsyna, Sergei Turitsyn
Aston Institute of Photonic Technologies
Aston University, Birmingham, United Kingdom
d.arguelloron@aston.ac.uk

***Abstract*—We propose a solution to noisy neural networks employed in future optical communication systems. The proposed approach includes breaking down large networks into smaller ones and forming "crowds" using these elementary networks.**

***Index Terms*—Artificial Neural Networks, Equalisation, Computational Complexity, Noise Resilience.**

I. Introduction

Machine learning techniques and in particular neural networks (NNs) have been extensively used for signal equalisation in modern optical communication systems. Although the performance of some of these digital signal processing (DSP) techniques is impressive in restoring the original state of the received signal, these equalisers often require complex architecture with complicated elements such as long short-term memory nodes [1, 2]. This increases the computational complexity (CC) of such solutions beyond what is acceptable for real-time implementation, especially when compared to alternatives such as digital backpropagation, and at odds with the promise of low footprint solutions[3]. Moreover, their architecture is fundamentally different from the one that digital computers have, making this hardware not suitable for them. A solution to the challenges posed by these equalisers is an analogue implementation of them in the electrical or optical domain [4, 5, 6]. Optical platforms enjoy the intrinsic advantage of high bandwidth and low power consumption and may play an important role in the future of computing. The optical implementation of a structured NN consists of layers of matrix multiplication followed by a nonlinear activation node. This type of processors is capable of carrying out high-speed complicated ML computations with potentially low power consumption and noise floor [5]. However, this theoretical low-noise floor is yet to be realised. Nonlinear activation nodes are usually implemented using active devices that add noise to the chain of computation and lead to a cascade effect that is now the bottleneck of optical implementation of NNs [7, 6]. Thus, different techniques have been developed to deal with the non-idealities present in analogue hardware [8, 9]. In this work, we propose a simple solution to this problem by designing low-rank NNs and using several of them to equalise the signal. We demonstrate that independently trained NNs can be used to make a better decision. We draw the analogy with the "wisdom of the crowd" principle where the collective decisions of a diverse independent group of individuals outperform that of an expert. We show that various configurations can be used to take the opinion of the crowd into account and argue that this impact can be explained by the eminent presence of a random element and also the impact of additive noise at each nonlinear node [10].

This work has received funding from: EU H2020 MSCA project No. 860360 and EPSRC project TRANSNET.

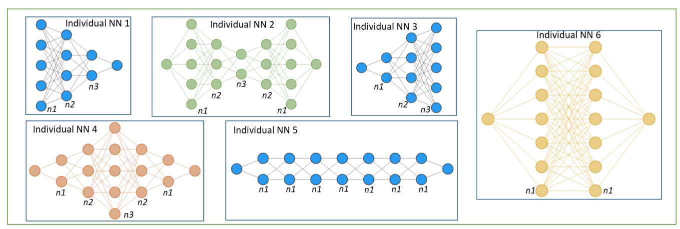

Fig. 1. Configurations of individual NNs considered in the training of the crowd. A number of neurons in each layer of individual NNs changes accordingly so that the CC equals that of the single NN divided by the crowd size.

II. Crowd equalisation

In optical communication, the task of the equaliser is to reduce the dispersive effects (defining channel memory) interfering with the nonlinearities present in the channel and/or devices [11]. Thus, the equaliser must be able to solve a regression problem where a block of consecutive symbols is used to equalise the symbol in the middle of the block[12]. The information possessed by the crowd is often more than the individuals it comprises. A famous example is when people are asked to guess the weight of a prize-winning ox, the average of all guesses is substantially more accurate than the accuracy of individual estimates [13]. This is the case when random fluctuation in independent estimates dominates bias [14]. Taking inspiration from this example, we form a crowd of individual equalisers as small NNs each capable of marginally improving

TABLE I

COMPLEXITY COMPARISON BETWEEN A SINGLE NN AND SUM OF THE CCS OF INDIVIDUAL NNS IN THE CROWD

Equaliser	Configurations of individual NN equalisers	total CC
Single NN	$[600, 600, 600, 600]$	4.3 mil
individuals in crowd of 2	$[900, 476, 223]$, $[732, 732]$	4.3 mil
individuals in crowd of 3	$[746, 386, 174]$, $[597, 597]$, $[175, 389, 750]$	4.3 mil
individuals in crowd of 4	$[678, 335, 114]$, $[517, 517]$, $[119, 336, 684]$, $[128, 236, 445, 236, 128]$	4.3 mil

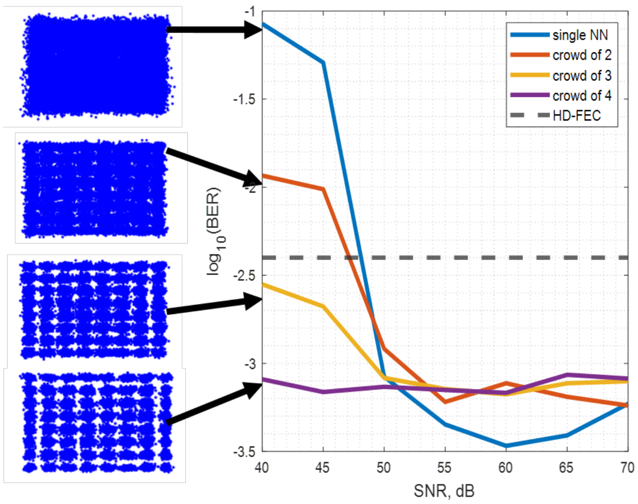

Fig. 2. Resulted BER for the cases of the single NN vs combined (crowd) NNs, when SNR is between 40 and 70 dB.

the performance of the optical communication system through equalising the received signal, see Fig. 1. This ”crowd of equalizers” is supposed to tackle the ”randomness” factor of the output of each of these individual equalisers. The weighted average of the outputs of the members of the crowd will form the final equalised symbol. Therefore, individuals are trained to individually equalise the input symbols with no knowledge about the further step of averaging their output. There are two questions to answer: i) how can we make individual equalisers linearly independent, and ii) how can we select the members of the crowd? To answer i) we need to consider that the output of a multilayer perceptron as our individual NN of choice, is the result of a series of nonlinear transformations. This means that to make two equalisers linearly independent, they need to differ in their weights, architecture, or activation functions somewhere before the last linear layer. The last activation function of all these NNs is a linear function due to the regression nature of the problem. This means that two NNs with a different number of layers or neurons are probably independent. Moreover, we guess that even a slight (but large enough) random fluctuation in the values of the link weights especially in the first layers (closer to the input layer) is enough to make two identical NNs linearly independent. However, in this work, we select NNs with different architectures as will be explained below. Regarding ii) NNs with different structures capture different features of the input data. This is handy when dealing with complicated interactions between temporal samples of the input such as in our case of an optical communication signal. Therefore, here as an exercise, we train a set of six NNs of various sizes and shapes. We do not know which one will be suitable to solve the problem but will train them and will pick the best ones to form a crowd of various sizes, see Fig. 1. This is however only one possibility to do so. An interesting alternative is to use different instances of the same small NN during the process of training as individual NNs. This speeds up the training and simplifies the architecture. The goal of this work is to show that dividing a large NN into smaller ones and using them as members of a crowd where their output is used alongside others makes the equaliser more resistant to noise. For this we consider NNs with noisy implementation, i.e. when matrix multiplication or activation function operation is noisy and reduces the SNR. In this work, we only consider noisy activation functions and assume that the matrix multiplication is ideal. The SNR is defined as the ratio of the average of the output of the activation function divided by the variance [10]. Considering noise in the process of training has been shown to improve resilience towards the ubiquitous noise and performance [8, 15, 16]. Furthermore, in any implementation of the NN noise is present, therefore a more realistic assumption is to include it in the training stage as well. We do so for training the members of the crowd and the single large NN whose performance is used as the benchmark. In all simulations, we limited the size of each member of the crowd to keep the total CC (in terms of the number of complex-valued multiplications) of the equaliser fixed, see Table I.

III. RESULTS

In this work, we use complex-valued NNs for all the simulations[17]. For the backpropagation we chose the Adam optimiser with a learning rate of 0.001, which minimises the complex-valued mean-squared error loss. The configuration of a single large NN, as well as the structure of the input feature vector to the each of NNs, is the same as in [18], where several QAM symbols before and after the symbol of interest are considered in each run. Our data set consists of experimental data containing transmitted and received QAM symbols in an optical communication system of length 400 km. The experimental setup consists of a polarisation division multiplexing 28-Gbaud 64QAM transmitter, 4 spans of SSMF ($\alpha = 0.2$ dB/km, $D = 17$ ps/nm/km, $\gamma = -1.3$/W/km) and a coherent receiver with no nonlinearity compensation DSP, see [18] for details. We divide the whole data set into four parts: training (70%), validation (10%), testing (10%) and

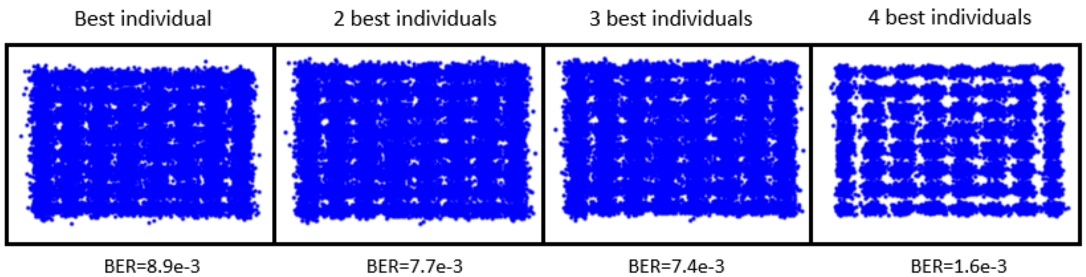

Fig. 3. Improvement in BER and constellations for a crowd of 4 NNs when averaging takes place over 1, 2, 3, 4 best individual NNs.

crowd testing (10%). The last crowd-testing set is only used to test the performance of the crowd and is not included in the training or testing stages of the NNs. After training six individuals of Fig. 1, we select the bests two, three, and four, to form crowds of sizes two, three, and four, respectively and save the full models. We used batches of size 2000 and continued training up to 300 epochs with typical provisions to prevent overfitting. Then we use these models on a separate test set (crowd test set) and combine the predicted output of them. This combination is a weighted sum of the outputs according to the competence of each individual, i.e. the more accurate the individual is in acting on the testing set the more weight its output will obtain. Therefore the output of the equaliser is as below:

$$y = \frac{1}{\sum_{i\in\mathcal{C}} BER_i\left(\mathbf{x}^{\text{test}}\right)} \sum_{i\in\mathcal{C}} \frac{f_i\left(\mathbf{x}^{\text{crowd test}}\right)}{BER_i\left(\mathbf{x}^{\text{test}}\right)} \tag{1}$$

where i is an individual in the crowd $\mathcal{C}$, $\mathbf{x}^{\text{test}}$ and $\mathbf{x}^{\text{crowd test}}$ are the input feature vector drawn from the test and crowd test set, respectively, $f_i\left(\mathbf{x}\right)$ is the output of the ith individual, and $BER_i\left(\mathbf{x}\right)$ is the BER of the ith individual. In Eq. (1), y is the QAM symbol output of the equaliser attributed to the middle symbol in $\mathbf{x}^{\text{crowd test}}$. This y is then used in a minimum distance detector to detect bits. The noise added to the output of each neuron is a Gaussian random variable with zero mean and different variance levels. Figure 2 shows the achieved BER (calculated as the number of mismatches of the transmitted and received bits) versus the SNR for a single equaliser, crowds of two, three, and four. The single equaliser performance significantly deteriorates as the noise power increases. As is shown, the sensitivity of the performance of these equalisers decreases as the size of the crowd increases. On the other hand, CC is defined as the number of complex-valued multiplications [3]:

$$CC = (1-s)\cdot\left(n_s n_i n_1 + \sum_{l=1}^{L-1} n_l n_{l+1} + n_o n_L\right) \tag{2}$$

The total CC of all these four equalisers (i.e. one single and three crowds of different sizes) is the same, see Table. I. On the left-hand side of Fig. 2, the resulting constellations after equalisation are presented showing a clear improvement in reducing noise from the received symbols. Also as can be seen in Fig. 2, at high noise powers, increasing the size of the crowd decreases the BER especially at $\mathrm{SNR} \leq 50$ dB. In Fig. 3 the improvement in the separation of clouds in the received constellation as a result of including more individuals in the crowd for the case of the crowd of four is illustrated. This yet again suggests that what is removed by means of averaging is of a noise nature.

IV. Conclusion

We have demonstrated the power of crowd equalisation in mitigating the problem of noise propagation and amplification in the implementations of artificial NNs. Using experimental samples from a 400 km 28-Gbaud 64QAM transmission, we have shown that breaking down large NN equalisers into smaller and less powerful ones and using them in a crowd rather than a single NN makes them more powerful in tackling the random factor of the NN implementations. Future works include investigating the impact of different architecture of the individual members and studying their impact on different types of noise.

Acknowledgment

The numerical simulations of this work were partially run on EPS Machine Learning Server of the Computer Science department at Aston University.

References

[1] Oleksandr Kotlyar et al. "Convolutional long short-term memory neural network equalizer for nonlinear Fourier transform-based optical transmission systems". In: *Optics Express* 29.7 (2021), pp. 11254–11267.

[2] Pedro J Freire et al. "Domain Adaptation: the Key Enabler of Neural Network Equalizers in Coherent Optical Systems". In: *2022 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE. 2022, pp. 1–3.

[3] Diego Argüello Ron et al. "Experimental Implementation of a Neural Network Optical Channel Equalizer in Restricted Hardware Using Pruning and Quantization". In: (2022).

[4] Gordon Wetzstein et al. "Inference in artificial intelligence with deep optics and photonics". In: *Nature* 588.7836 (2020), pp. 39–47.

[5] Yichen Shen et al. "Deep learning with coherent nanophotonic circuits". In: *Nature Photonics* 11.7 (2017), pp. 441–446.

[6] Ryan Hamerly et al. “Large-scale optical neural networks based on photoelectric multiplication”. In: *Physical Review X* 9.2 (2019), p. 021032.
[7] Bhavin J Shastri et al. “Principles of neuromorphic photonics”. In: *arXiv preprint arXiv:1801.00016* (2017).
[8] Chuteng Zhou et al. “Noisy machines: Understanding noisy neural networks and enhancing robustness to analog hardware errors using distillation”. In: *arXiv preprint arXiv:2001.04974* (2020).
[9] Nadezhda Semenova, Laurent Larger, and Daniel Brunner. “Understanding and mitigating noise in trained deep neural networks”. In: *Neural Networks* 146 (2022), pp. 151–160.
[10] Nadezhda Semenova et al. “Fundamental aspects of noise in analog-hardware neural networks”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.10 (2019), p. 103128.
[11] John R Barry. *Wireless infrared communications*. Vol. 280. Springer Science & Business Media, 1994.
[12] Pedro J Freire et al. “Neural Networks-based Equalizers for Coherent Optical Transmission: Caveats and Pitfalls”. In: *arXiv preprint arXiv:2109.14942* (2021).
[13] Francis Galton. “Vox populi (the wisdom of crowds)”. In: *Nature* 75.7 (1907), pp. 450–451.
[14] Edward Vul and Harold Pashler. “Measuring the Crowd Within: Probabilistic Representations Within Individuals”. In: *Psychological Science* 19.7 (2008). PMID: 18727777, pp. 645–647. DOI: 10.1111/j.1467-9280.2008.02136.x.
[15] Naresh Nagabushan, Nishank Satish, and S Raghuram. “Effect of injected noise in deep neural networks”. In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE. 2016, pp. 1–5.
[16] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. “Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 588–597.
[17] J Agustin Barrachina. *Complex-Valued Neural Networks (CVNN)*. Version v1.0.3. Jan. 2021. DOI: 10.5281/zenodo.4452131. URL: https://doi.org/10.5281/zenodo.4452131.
[18] Morteza Kamalian-Kopae et al. “Neural Network-Enhanced Optical Phase Conjugation for Nonlinearity Mitigation”. In: *2022 Optical Fiber Communications Conference and Exhibition (OFC)*. 2022, pp. 1–3.