Energy-efficient Spiking Neural Network Equalization for IM/DD Systems with Optimized Neural Encoding

Alexander von Bank, Eike-Manuel Edelmann, and Laurent Schmalen
Communications Engineering Lab, Karlsruhe Institute of Technology, 76187 Karlsruhe, Germany
alexander.bank@kit.edu, edelmann@kit.edu

Abstract: We propose an energy-efficient equalizer for IM/DD systems based on spiking neural networks. We optimize a neural spike encoding that boosts the equalizer's performance while decreasing energy consumption. © 2023 The Author(s)

1. Introduction

Spiking neural networks (SNNs) enable the implementation of powerful machine-learning algorithms using energy-efficient neuromorphic hardware [1]. SNNs process information by exchanging short pulses (spikes) between their neurons. This leads to a sparse representation of information and low energy consumption since spikes are only exchanged when information is processed [2]. Recent work has shown that SNN-based equalizers are promising candidates for powerful equalizers with low-energy consumption for optical transmissions [1,3–6]. For an intensity modulation / direct detection (IM/DD) link suffering from non-linear impairments and chromatic dispersion (CD), an SNN-based equalizer and demapper, which outperforms linear as well as artificial neural network (ANN)-based equalizers was proposed in [4]. Simulation results of the equalizer proposed in [4] have been reproduced using the PyTorch-based SNN deep learning library `Norse` [7] and neuromorphic hardware [1]. In [3], the equalizer of [4] is applied to experimental IM/DD data. In [5], we proposed an SNN-based equalizer with decision feedback, which outperforms the approach of [4] for an IM/DD link [6].

The transformation of continuous data into spiking signals is called neural encoding. For instance, for the IM/DD link, the channel output is encoded and forwarded to the SNN. In [2], several neural encoding schemes and their applications are discussed. Three basic encoding schemes are [2]: Rate encoding transforms the information in the spike frequency, i.e., number of spikes per time interval; In temporal encoding, the timing of the spikes contains the information; For population encoding, the information is encoded in the interaction of different neurons. Depending on the task, encoding schemes differ in noise robustness, accuracy, energy consumption, and hardware requirements. Therefore, a crucial task when implementing SNNs is to find an efficient neural encoding.

While the work mentioned above [1,3–6] focuses on the design and learning of SNNs, the encoding is designed based on empirical knowledge and not further optimized. In [4], a log-scale encoding is proposed, which encodes the information in the relative timing of multiple spikes. In [5,6], ternary encoding is introduced, which encodes information by activating a predefined subgroup of input neurons.

This work proposes a generic neural encoding based on a learnable matrix, which determines the input pattern fed to the SNN's input layer. The learnable matrix is furthermore regularized using the ℓ_p-over-ℓ_q regularization of [8]. Using the SNN-based equalizer and simulated IM/DD link of [4], we compare the proposed encoding with log-scale encoding of [4] and ternary encoding of [5,6]. We show that the learned encoding with regularization reduces the number of SNN-generated spikes by up to 50%, resulting in reduced power consumption while enhancing the system's performance by 0.3 dB. The code is available at `https://github.com/kit-cel/OptiSpike`.

2. Spiking Neural Network-based Equalizer

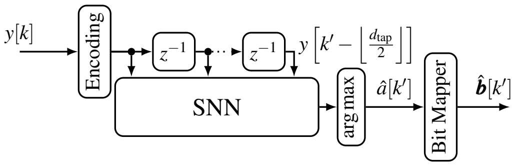

$y[k]$
Encoding
z^{-1}
z^{-1}
$y\left[k' - \left\lfloor \frac{d_{\text{tap}}}{2} \right\rfloor\right]$
SNN
arg max
$\hat{a}[k']$
Bit Mapper
$\hat{\boldsymbol{b}}[k']$

Fig. 1: Proposed equalizer structure

An SNN consists of multiple interconnected state-dependent neurons whose internal states evolve and exhibit temporal dynamics. Synapses connect neurons employing adjustable weights. A common neuron model is the leaky integrate-and-fire (LIF) model [9]: A neuron's input is scaled by the weight of the connecting synapse. The neuron's state is charged by integrating the input over time. In parallel, it loses charge over time (leakage). If the neuron is charged sufficiently, such that its neuronal threshold is exceeded, the neuron fires an output spike, which is passed to the connected upstream neurons. In this work, we simulate SNNs using `Norse` [7], which uses the backpropagation through time algorithm [9] to update the synapses' weights during training. Figure 1 outlines the structure of the SNN-based equalizer as proposed by [4]. The most recent channel output, $y[k]$, is encoded and fed to

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001899). Parts of this work were carried out in the framework of the CELTIC-NEXT project AI-NET-ANTILLAS (C2019/3-3) funded by the German Federal Ministry of Education and Research (BMBF) (grant agreement 16KIS1316).

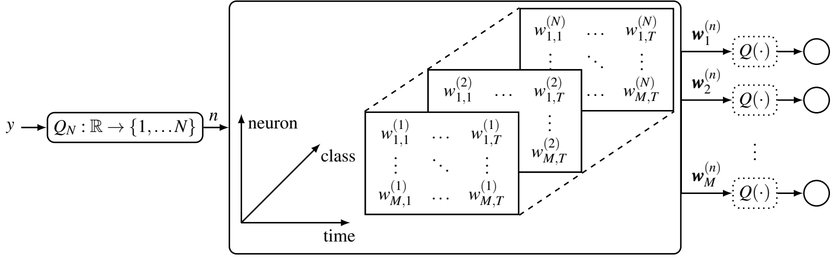

y
$Q_N : \mathbb{R} \to \{1,\dots N\}$
n
neuron
class
time
$w_{1,1}^{(1)}$ … $w_{1,T}^{(1)}$
$w_{M,1}^{(1)}$ … $w_{M,T}^{(1)}$
$w_{1,1}^{(2)}$ … $w_{1,T}^{(2)}$
$w_{M,T}^{(2)}$
$w_{1,1}^{(N)}$ … $w_{1,T}^{(N)}$
$w_{M,T}^{(N)}$
$\boldsymbol{w}_1^{(n)}$
$\boldsymbol{w}_2^{(n)}$
$\boldsymbol{w}_M^{(n)}$
$Q(\cdot)$
$Q(\cdot)$
$Q(\cdot)$

Fig. 2: Sketch of the embedding. The received value y is classified into a class n with $n \in \{1,\dots N\}$. Depending on n, the embedding matrix $\boldsymbol{W}^{(n)} \in \mathbb{R}^{M \times T}$ is chosen, where $w_{m,t}^{(n)}$ is the embedding value of class n, fed to the m-th input neuron at time step t. To simulate quantized input to neuromorphic hardware, quantizers $Q(\cdot)$ can be included.

the equalizer to estimate the k'-th transmitted bit pattern $\boldsymbol{b}[k']$ corresponding to the k'-th transmit symbol, where $k' = k - \lceil d_{\text{tap}}/2 \rceil$. The number of equalizer taps d_{tap} matches the number of significant channel taps when sampling the channel at the symbol rate. The encoded sample and previous encoded samples are fed to the SNN. At the SNN's output layer, the index $\hat{a}$ of the highest state output neuron is determined and transformed to bits via a bit mapper.

3. Encoding

In Fig. 2, we outline the structure of a parameterized encoding, whose parameters are jointly optimized with the loss function of the SNN's learning task. For encoding, $y \in \mathbb{R}$ is quantized by a uniform quantizer $Q_N(\cdot)$ with N quantization levels and mapped to a class n, indicating the quantized value. Matrices $\boldsymbol{W}^{(n)} \in \mathbb{R}^{M\times T}$ are initialized with random i.i.d. elements $w_{m,t}^{(n)} \sim \mathcal{N}(0,1)$, where $m \in \{1,\dots M\}$ denotes the SNN's input neuron, $t \in \{1,\dots T\}$ the SNN's discrete simulation time step and $n \in \{1,\dots N\}$ the input value class. Depending on n, the matrix $\boldsymbol{W}^{(n)}$ is chosen, where the m-th row of $\boldsymbol{W}^{(n)}$, denoted as $\boldsymbol{w}_m^{(n)}$, is fed to the m-th SNN's input neuron over time. During training, both the matrices $\boldsymbol{W}^{(n)}$ and the SNN parameters can be jointly optimized by minimizing the cross entropy loss $J_{\text{CE}}(a,\hat{a})$ of the transmit symbol's actual index a and the estimated index $\hat{a}$.

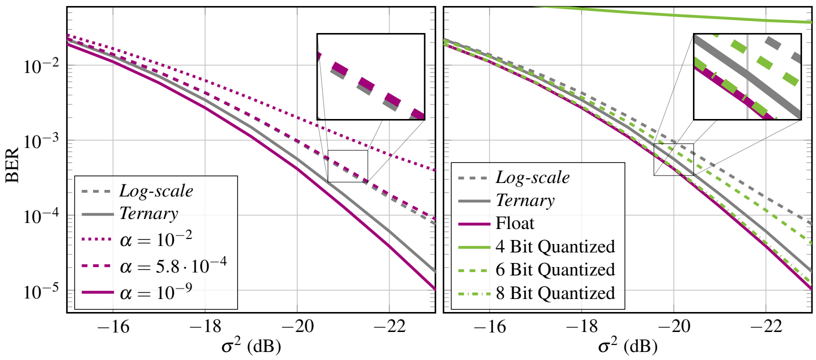
Sparsity describes how much a signal's energy is concentrated on a few samples [8]. To reduce the number of spikes and, therefore, the network's energy consumption, we propose incorporating a sparsity-inducing penalty for $\boldsymbol{W}^{(n)}$. The average ℓ_p-over-ℓ_q quasinorm-ratio over all classes provides a measure of sparseness. In particular, the normalized ℓ_1-over-ℓ_2 quasinorm-ratio $\ell_{1,2}\left(\boldsymbol{W}^{(n)}\right) = \sum_{m=1}^{M}\sum_{t=1}^{T}|w_{m,t}^{(n)}| \left(\sum_{m'=1}^{M}\sum_{t'=1}^{T}|w_{m',t'}^{(n)}|^2\right)^{-1/2}$ is frequently used [8]. The overall loss function can be defined as $\left(a,\hat{a},\boldsymbol{W}^{(n)}\right) = (1-\alpha)J_{\text{CE}}(a,\hat{a}) + \alpha\frac{1}{N}\sum_{n=1}^{N}\ell_{1,2}\left(\boldsymbol{W}^{(n)}\right)$, $\alpha \in [0,1]$. To avoid exploding parameters, the matrices $\boldsymbol{W}^{(n)}$ are normalized according to $\boldsymbol{W}^{(n)} \leftarrow \boldsymbol{W}^{(n)}\left(\max_{m,t} \quad w_{m,t}^{(n)}\right)^{-1}$, $\forall\, n \in N$, after each optimization step. Consequently, low values $w_{m,t}^{(n)}$ are pushed even further to zero.

The proposed encoding is limited by the real-valued nature of the sequences $\boldsymbol{w}_m^{(n)}$, losing the binary character of log-scale and ternary encoding. However, Intel's Loihi 2 chip, as state-of-the-art neuromorphic hardware, supports the input of quantized values, so-called *graded spikes*, with up to 32-bit resolution [10]. The setup of Fig. 2 can be extended by quantizers $Q(\cdot)$ to simulate the impact of quantization of $\boldsymbol{w}_m^{(n)}$.

4. Results

We compare the proposed encoding with ternary encoding [5] and log-scale encoding [4] for an IM/DD link and the SNN-based equalizer of Fig. 1. The IM/DD link is simulated as in [4] with parameters like channel B of [6]: A single mode fiber of 5 km, RRC pulse shaping with roll-off factor $\beta = 0.2$, 100 GBd baud rate, wavelength of 1270 nm, dispersion coefficient of $-17\,\text{ps}\,\text{nm}^{-1}\,\text{km}^{-1}$ and $d_{\text{tap}} = 41$ equalizer taps are used in the simulation. The pulse amplitude modulated transmit symbols are taken from the set $\mathcal{C} = \left\{0, 1, \sqrt{2}, \sqrt{3}\right\}$ with Gray mapping. After pulse shaping, a bias is added. For the log-scale encoding [4], an input neuron count of $M = 10$ was used with $T = 30$ discrete time steps. In contrast, ternary encoding [5] and the proposed encoding have an input neuron count of $M = 8$ and $T = 10$ discrete time steps. Furthermore, the proposed encoding uses $N = 256$ matrices $\boldsymbol{W}^{(n)}$. Each SNN consists of $N_{\text{o}} = 4$ output neurons and $N_{\text{h}} = 80$ hidden neurons. The training was carried out using a batch size of 200 000, 5 Epochs, 2000 batches per epoch, and a learning rate of 10^{-3}. In [6], we have shown that for the given link, SNNs trained at a noise power of $\sigma^2 = -17\,\text{dB}$ perform best. Hence, we fix $\sigma^2 = -17\,\text{dB}$ for training.

Figure 3 shows the performance of the proposed encoding and benchmarks. The different approaches are compared regarding their equalization performance (BER) and energy efficiency, measured by their respective spike rate. The spike rate is defined by $S_{\text{h}}/(N_{\text{h}}T)$, where S_{h} is the number of spikes generated by the hidden layer, and $N_{\text{h}}T$ is the number of possible spikes of the discrete-time SNN simulation. Figure 3(a) compares the proposed encoding with the benchmarks for different α. If α is sufficiently low, our encoding is superior to log-scale and ternary encoding and enhances system performance.



(a) (b)

Encoding	spike rate (%)
Log-scale	14.5
Ternary	12.5
$\alpha = 10^{-2}$	5.2
$\alpha = 5.8 \cdot 10^{-4}$	6.0
$\alpha = 10^{-9}$	7.3
4 Bit Quant.	7.6
6 Bit Quant.	7.3
8 Bit Quant.	7.3

(c)

Fig. 3: Comparison of the proposed encoding against the benchmark encodings: (a) The impact of α, (b) the impact of the quantization of $\boldsymbol{W}(n)$ for fixed $\alpha = 10^{-9}$ and (c) the resulting spike rate measured at fixed $\sigma^2 = 19\,$dB.

Furthermore, the spike rate is significantly reduced. For $\alpha = 10^{-9}$, it is reduced by roughly 50% compared to the log-scale and roughly 42% compared to the ternary encoding, see Fig. 3(c). Over the range of simulated σ^2, the spike rate is near constant. The spike rate can be further decreased by increasing α and thus the impact of the ℓ_1-over-ℓ_2 penalty on the loss function. However, this rate reduction comes at the cost of decreasing system performance. For a given target BER, a suitable choice of α enables the flexible reduction of the spike rate. Figure 4 shows the impact of α on the distribution of $w_{m,t}^{(n)}$. Prior to learning, the parameters are initialized by sampling independently from $\mathcal{N}(0,1)$. For both $\alpha = 10^{-2}$ and $\alpha = 10^{-9}$, approximately 1.6% of the SNN's input has the maximal amplitude of one. The number of graded spikes reduces by increasing α, which in turn increases the BER. The histograms and the performance of $\alpha = 10^{-2}$ and $\alpha = 10^{-9}$ indicate that most input information is encoded in the graded spikes of $\boldsymbol{W}^{(n)}$. Hence, quantizing the encoding values decreases performance, as shown in Fig. 3(b). If the number of quantization bits is sufficiently high, the quantization has only a minor effect on the system's performance. However, if the number of quantization bits is set too low, the system performance will deteriorate considerably. As mentioned above, Intel's Loihi 2 supports up to 32-bit quantization. Hence, the quantization effect can be neglected. Notably, the spike rate is only negligibly affected by quantization.

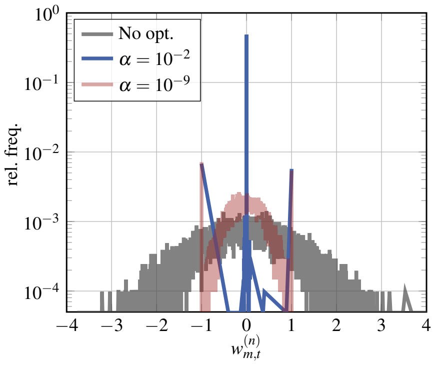


Fig. 4: Histogram of the elements of $\boldsymbol{W}$ before optimization (No opt) and with optimization using α.

5. Conclusion

We proposed a novel neural encoding based on learnable parameter matrices. We have shown that the encoding enhances system performance while decreasing the spike rate significantly.

References

1. E. Arnold *et al.*, "Spiking neural network nonlinear demapping on neuromorphic hardware for IM/DD optical communication," in *J. Lightw. Technol.*, vol. 41, No. 11, pp. 3424-3431, June 2023
2. D. Auge *et al.*, "A survey of encoding techniques for signal processing in spiking neural network," in *Neural Process. Lett.*, vol. 53, pp. 4693-4710, 2021.
3. G. Böcherer *et al.*, "Spiking neural network linear equalization: Experimental demonstration of 2km 100Gb/s IM/DD PAM4 optical transmission," in *Optical Fiber Communications Conference (OFC)*, San Diego, CA, USA, 2023
4. E. Arnold *et al.*,"Spiking neural network equalization for IM/DD optical communication," in *Proc. Advanced Photonic Congress: Signal Processing in Photonic Communications (SPPCom)*, Maastricht, NL, July 2022.
5. E.-M. Bansbach, A. von Bank and L. Schmalen, "Spiking neural network decision feedback equalization", in *Proc. ITG WSA-SCC*, Braunschweig, Germany, Feb. 2023
6. A. von Bank, E.-M. Edelmann, L. Schmalen, "Spiking neural network decision feedback equalization for IM/DD systems," in *Proc. Advanced Photonic Congress: Signal Processing in Photonic Communications (SPPCom)*, Busan, South Korea, Jul. 2023.
7. C. Pehle and J. Pedersen, "Norse – A deep learning library for spiking neural networks," Jan. 2021, doi:10.5281/zenodo.4422025, Documentation: https://norse.ai/docs/.
8. A. Cherni *et al.*, "SPOQ l_p-Over-l_q regularization for sparse signal recovery applied to mass spectrometry," in *IEEE Trans. Signal Process.*, vol. 68, pp. 6070-6084, 2020
9. E. O. Neftci, H. Mostafa and F. Zenke , "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 51-63, Nov. 2019
10. Intel, "Taking neuromorphic computing to the next level with Loihi 2," 2021. [Online]. Available: https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf, (accessed on: 29.09.23).