Neural Networks 103 (2018) 118–127

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Spiking neural networks for handwritten digit recognition—Supervised learning and network optimization

Shruti R. Kulkarni, Bipin Rajendran *

Department of Electrical and Computer Engineering, New Jersey Institute of Technology, NJ, 07102, USA

ARTICLE INFO

Article history:
Received 2 December 2017
Received in revised form 13 February 2018
Accepted 27 March 2018
Available online 6 April 2018

Keywords:
Neural networks
Spiking neurons
Supervised learning
Pattern recognition
Approximate computing
Neuromorphic computing

ABSTRACT

We demonstrate supervised learning in Spiking Neural Networks (SNNs) for the problem of handwritten digit recognition using the spike triggered Normalized Approximate Descent (NormAD) algorithm. Our network that employs neurons operating at sparse biological spike rates below 300 Hz achieves a classification accuracy of 98.17% on the MNIST test database with four times fewer parameters compared to the state-of-the-art. We present several insights from extensive numerical experiments regarding optimization of learning parameters and network configuration to improve its accuracy. We also describe a number of strategies to optimize the SNN for implementation in memory and energy constrained hardware, including approximations in computing the neuronal dynamics and reduced precision in storing the synaptic weights. Experiments reveal that even with 3-bit synaptic weights, the classification accuracy of the designed SNN does not degrade beyond 1% as compared to the floating-point baseline. Further, the proposed SNN, which is trained based on the precise spike timing information outperforms an equivalent non-spiking artificial neural network (ANN) trained using back propagation, especially at low bit precision. Thus, our study shows the potential for realizing efficient neuromorphic systems that use spike based information encoding and learning for real-world applications.

1. Introduction

The superior computational efficiency of biological systems has inspired the quest to reverse engineer the brain in order to develop intelligent computing platforms that can learn to execute a wide variety of data analytics and inference tasks (NAE, 2009). Artificial neural networks (ANNs), inspired by the network architecture of the brain, have emerged as the state-of-the-art for various machine learning applications. In particular, inspired by the Nobel prize winning work of Hubel and Weisel on elucidating the mechanisms of information representation in the visual cortex (Hubel & Wiesel, 1968), multi-layer convolutional neural networks have shown impressive performance for a wide variety of applications such as image recognition, natural language processing, speech recognition and video analytics (Ciregan, Meier, & Schmidhuber, 2012; Goldberg, 2016; Goodfellow, Bengio, & Courville, 2016; Goodfellow, Warde-Farley, Mirza, Courville, & Bengio, 2013; Hinton et al., 2012; Hinton, Osindero, & Teh, 2006; Karpathy et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012; Lecun, Bottou, Bengio, & Haffner, 1998).

Nevertheless, the neurons in ANNs implement a memoryless nonlinear transformation of the input synaptic signals to create real-valued output signals. This is vastly different from the behavior of neurons in the brain, which encode information in the timing of binary signals, called action potentials or spikes based on the timing of incoming spike signals from upstream nodes. The third generation of artificial neural networks, also called spiking neural networks (SNNs), have been introduced to mimic this key aspect of information processing in the brain (Maass, 1997). There is growing evidence that SNNs have significant computational advantages as a result of their higher information representational capacity due to the incorporation of the temporal dimension (Brader, Senn, & Fusi, 2007; Crotty & Levy, 2005; Gutig & Sompolinsky, 2006; Hopfield & Brody, 2004). Furthermore, SNNs issue spikes sparsely – the observed spike rate in biological networks is in the range of 0.1 to 300 Hz – and they operate in an event-driven manner (Gabbiani & Metzner, 1999; Roxin, Brunel, Hansel, Mongillo, & van Vreeswijk, 2011; Shoham, O'Connor, & Segev, 2006; Wang et al., 2016). Therefore, highly energy efficient neuromorphic systems can be realized in hardware based on SNNs, as is evidenced by recent demonstrations (Benjamin et al., 2014; Furber, Galluppi, Temple, & Plana, 2014; Gehlhaar, 2014; Merolla et al., 2014; Qiao et al., 2015).

Earlier efforts to build learning algorithms for SNNs were inspired by recent discoveries from neuroscience that shed light on the synaptic (neuronal interconnections) mechanisms of adaptation based on the difference in the times of issue of pre- and

* Corresponding author.
E-mail address: bipin@njit.edu (B. Rajendran).

https://doi.org/10.1016/j.neunet.2018.03.019
0893-6080/

post-synaptic spikes. The most prominent among them is the Remote Supervised Method (ReSuMe) (Ponulak & Kasinski, 2010), that adjusts the synaptic weights based on the precise timing differences of the input and output neurons, inspired by the spike timing dependent plasticity (STDP) rule. Other spike based learning algorithms that have been proposed include the SpikeProp algorithm (though it was restricted to single spike learning) (Bohte, Kok, & La Poutre, 2002), SPAN and PSD, which converted spikes to smoothened analog signals and defined a continuous time cost function for training (Mohemmed, Schliebs, Matsuda, & Kasabov, 2012; Yu, Tang, Tan, & Li, 2013). Another important spike based supervised learning rule was the Chronotron rule which used piece-wise gradient descent and was demonstrated to be efficient in identifying different classes of random spike trains (Florian, 2012). Recently, the reward modulated STDP or R-STDP learning has shown superior performance on several benchmark problems compared to STDP SNNs and even traditional CNNs, even though training was limited to a single layer in the network (Mozafari, Kheradpisheh, Masquelier, Nowzari-Dalini, & Ganjtabesh, 2017). A variant of ReSuMe algorithm, called the Delay Learning (DL)-ReSuMe, in addition to the synaptic weights, made use of the transmission delays of synapses interconnecting the neurons as parameters to train the network (Taherkhani, Belatreche, Li, & Maguire, 2015). This algorithm has been shown to be superior in terms of accuracy and speed of convergence compared to the basic ReSuMe algorithm. The accurate synaptic efficiency adjustment method is another spike-error triggered supervised learning rule based on STDP, which optimizes a cost function defined in terms of membrane potential differences (Xie, Qu, Yi, & Kurths, 2017). This method has been used to demonstrate excellent performance in several UCI datasets with few training parameters. The Synaptic Kernel Inverse Method (SKIM) (Tapson et al., 2013) evaluates the weights analytically rather than learning them iteratively and has been applied to the problem of speech based digit recognition in a small network with 50 neurons. Based on the SKIM method, the convex optimized synaptic efficiencies (CONE) algorithm was developed (Lee, Kukreja, & Thakor, 2017) and was used for the problem of gait detection. The generalization capability of this algorithm and the noise tolerance of a variation of the algorithm called CONE-R has also been demonstrated.

Our work focuses on applying a precise spike based supervised learning algorithm to the MNIST (Modified National Institute of Standards and Technology database) handwritten digit classification problem and optimizing the network in terms of the number of learning parameters for implementation in energy and memory constrained hardware.

In addition to the above mentioned learning methods, unsupervised learning algorithms for SNNs have also been explored, based on the biological spike timing dependent plasticity (STDP) rule (Allred & Roy, 2016; Diehl & Cook, 2015; Kheradpisheh, Ganjtabesh, Thorpe, & Masquelier, 2017; Masquelier & Thorpe, 2007; Panda & Roy, 2016; Roy & Basu, 2017; Tavanaei & Maida, 2017). While these networks use multi-layered convolution architectures with more than one million parameters and have achieved over 98% accuracy on the MNIST dataset (Kheradpisheh et al., 2017; Tavanaei & Maida, 2017), we demonstrate similar accuracy with 13× fewer parameters.

There are also several efforts directed towards developing architectures with adaptive and evolving network structures (Kasabov, 2014; Kasabov et al., 2016; Takuya, Haruhiko, Hiroharu, & Shinji, 2016; Wang, Belatreche, Maguire, & McGinnity, 2015, 2017). SpikeTemp and SpikeComp are algorithms where neurons are progressively added in the classifier layer as the training algorithm approaches the optimal point (Wang et al., 2015, 2017). The recently developed evolving architecture called NeuCube, directly inspired by the brain Kasabov (2014), incorporates weight adjustments based on supervised and unsupervised rules and additionally, adds new network neurons as per training requirements.

Besides the above-mentioned approaches for designing learning algorithms for SNNs that operate directly in the spike domain, several authors have proposed to convert ANNs trained with the well-established backpropagation algorithm to SNNs so that the latter can be used as inference engines (Cao, Chen, & Khosla, 2015; Diehl et al., 2015; Hunsberger & Eliasmith, 2016; Hunsberger, Eric, 2018; Rueckauer, Lungu, Hu, & Pfeiffer, 2016; Rueckauer, Lungu, Hu, Pfeiffer, & Liu, 2017). ANN-to-SNN conversion imposes that the firing rate of a spiking neuron in the SNN be proportional to the activation output of a non-spiking neuron in the ANN. Various techniques such as approximating the response of a spiking neuron with a smooth differentiable ReLU-like function, weight normalization, noise addition, lateral inhibition or spiking rate based pooling masks, which is similar to max pooling operation, have been employed to this end. Using these approaches, state-of-the-art inference accuracies have been demonstrated in spike domain equivalent of deep learning networks such as VGG-16 and Inception-V13 for ImageNet classification problem, and close to 2× reduction in the number of operations needed compared to CNNs for smaller problems such as MNIST and CIFAR-10 (Rueckauer et al., 2017). Recently, a more biologically plausible algorithm called the Feedback Alignment (FA) has been proposed, which unlike the standard backpropagation uses two different sets of weights in the feed-forward and feedback paths (Lillicrap, Cownden, Tweed, & Akerman, 2016). This method has also been demonstrated in SNNs, using approximate differentiable functions of leaky integrate and fire (LIF) spiking neurons to train them in an online manner. However, the FA rule has lower performance compared to the standard backpropagation rule (Hunsberger, Eric, 2018).

Towards the goal of demonstrating a learning SNN capable of high accuracy and efficiency, we use the recently proposed Normalized Approximate Descent (NormAD) algorithm to train the output layer weights of a three-layered network with fixed convolutional kernel weights in the hidden layer. This spike-triggered weight update rule frames the learning task as a supervised optimization problem aimed at tuning the membrane potential to create spikes at desired time instants. Compared to other deterministic learning algorithms in the spike domain such as ReSuMe, at least 10× faster convergence characteristics have been demonstrated using this algorithm for generating arbitrarily desired spike streams (Anwani & Rajendran, 2015).

Prior SNN based demonstration of handwritten digit recognition using spiking versions of backpropagation of errors has achieved 98.7% based on a fully connected 4−layer network and 99.31% with convolutional spiking networks, but also with more than 4× higher number of trainable synapses compared to our network (Lee, Delbruck, & Pfeiffer, 2016). The training algorithm employed in that work has a cost function that is continuous in time defined in terms of the low pass filtered spike trains (both input and output). Compared to the state-of-the-art networks which have shown over 99% accuracy, our SNN trained with NormAD shows an accuracy of 98.17% on the test set of the MNIST database, with 4× fewer synaptic learning parameters (Ciregan et al., 2012; Goodfellow et al., 2013; Lecun et al., 1998; Lee et al., 2016). Furthermore, if the network architecture and number of synaptic parameters are kept the same, we show that the accuracy and performance of the NormAD trained SNN is slightly better than that of an equivalent ANN trained using backpropagation.

This paper is organized as follows. We introduce the basic units of SNNs in Section 2. Section 3 describes the architecture of our network, the spike encoding at the input and output of the network, and the training algorithm used for weight updates. Section 4 describes several hyper-parameter tuning experiments and the