Neural Networks 121 (2020) 512–519

Contents lists available at ScienceDirect

Neural Networks

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

Deep CovDenseSNN: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment

Qi Xu [a], Jianxin Peng [b], Jiangrong Shen [a,c], Huajin Tang [a], Gang Pan [a,d,*]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China
[b] College of Computer Science, Sichuan University, Chengdu, 610065, China
[c] Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, 310027, China
[d] State Key Lab of CAD & CG, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 12 May 2019
Received in revised form 20 August 2019
Accepted 25 August 2019
Available online 28 September 2019

Keywords:
Spiking neurons
Encoding
Feature extraction
Noisy environment

ABSTRACT

Neurons in the brain use an event signal, termed spike, encode temporal information for neural computation. Spiking neural networks (SNNs) take this advantage to serve as biological relevant models. However, the effective encoding of sensory information and also its integration with downstream neurons of SNNs are limited by the current shallow structures and learning algorithms. To tackle this limitation, this paper proposes a novel hybrid framework combining the feature learning ability of continuous-valued convolutional neural networks (CNNs) and SNNs, named deep CovDenseSNN, such that SNNs can make use of feature extraction ability of CNNs during the encoding stage, but still process features with unsupervised learning rule of spiking neurons. We evaluate them on MNIST and its variations to show that our model can extract and transmit more important information than existing models, especially for anti-noise ability in the noisy environment. The proposed architecture provides efficient ways to perform feature representation and recognition in a consistent temporal learning framework, which is easily adapted to neuromorphic hardware implementations and bring more biological realism into modern image classification models, with the hope that the proposed framework can inform us how sensory information is transmitted and represented in the brain.

There are various conventional methods to implement pattern recognition, such as maximum entropy classifier, naive Bayes classifier, decision trees, support vector machines and fuzzy control systems (Qiu, Sun, Wang, & Gao, 2019; Sun, Mou, Qiu, Wang, & Gao, 2019). These computational models, despite inspired by neuroscience in a way, are lacking several aspects of biological property of neuron system, one being the presence of spike, which is the fundamental information unit for neural computation (Bengio, Lee, Bornschein, Mesnard, & Lin, 2015).

How information is represented in the brain still remains far from understanding. On one hand, deep learning provides comprehensive computational models to encode and extract hierarchically organized features from arbitrary images. (Wen, Shi, Zhang, Lu, & Liu, 2016) have adopted convolutional neural networks (CNNs) for studying some open questions of neuroscience, although, the prediction of neuronal responses has a good performance, the final output of that CNN model is bringing dense computations conducted in many layers, which may not be relevant to the biological underpinnings of information processing in the brain. On the other hand, there is strong evidence (Hung, Kreiman, Poggio, & DiCarlo, 2005) to believe that spikes are an optimal way for transmission and information representation. Unlike neurons in CNNs, which communicate via real values, neurons in a spiking neural network (SNN) communicate via spikes. Moreover, SNNs are event-driven as computation in synapses and neurons are triggered by incoming spikes. SNNs are advantageous to deal with spatio-temporal information patterns through spike-based learning and memory mechanisms (Buonomano & Maass, 2009; Liu & Buonomano, 2009).

However, typical SNNs are surely at a great disadvantage about feature extraction because they consist of just only a fully-connected layer with biologically based neurons. In contrast, deep CNNs have a great ability of feature extraction at the pixel level (LeCun, Bengio, & Hinton, 2015), in addition, recent studies show that the convolutional filters are similar to the receptive fields of the retinal neurons (Maheswaranathan, et al., 2018; Yan et al., 2018). Serre, Oliva, and Poggio (2007) explored the visual system using the hierarchical simple cell and complex cell feed-forward model, and showed that there is a high resemblance of the feature extraction process between the model and biological

* Corresponding author at: College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China.
E-mail addresses: xuqi123@zju.edu.cn (Q. Xu), pjx1234567@foxmail.com (J. Peng), jrshen@zju.edu.cn (J. Shen), htang@scu.edu.cn (H. Tang), gpan@zju.edu.cn (G. Pan).

https://doi.org/10.1016/j.neunet.2019.08.034
0893-6080/

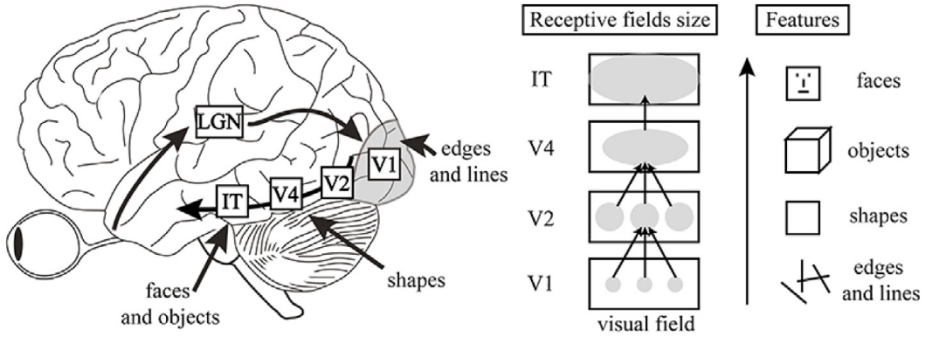

Fig. 1. Left: A typical hierarchical, feedforward model about how vision forms in human brain. External stimuli are processed at the retina, proceeds to the LGN, then to V1, V2, V4 and IT. Decisions about external stimuli are made in the frontal cortex. **Center:** Lower visual areas have smaller receptive fields, while neurons in higher areas gradually increasing receptive field sizes, integrating information over larger and larger regions of the visual field. **Right:** Lower visual areas, such as V1, are sensitive to basic features such as edges and lines. Higher-level neurons pool information over multiple low-level neurons with smaller receptive fields and code for more complex features.

brain. Nevertheless, the previous model (Serre et al., 2007) does not fully account for the recognition in a biological realistic way.

Therefore, CSNN (Xu et al., 2018) and S1C1-SNN (Yu, Tang, Tan, & Li, 2013) try to construct hierarchical cognitive models to address the pattern recognition tasks in a biologically plausible way. They both adopt one layer based feature extractor, which means the feature extraction ability is still limited compared to deeper and more complex structures. There still exist big challenges about robust pattern recognition that demands the invariant representation of visual features and training methods.

In such kind of feature encoding and representation rules, only spatial information is obtained and transmitted, and it is not able to represent spatial–temporal information. Spatial–temporal feature encoding mechanisms and effective training methods remain as open problems, which are not only key ways for rapid visual recognition tasks (Hung et al., 2005) in neural systems, but also important solutions to achieve highly efficient pattern recognition in dynamic visual scenes (Orchard et al., 2015). For examples, they adopt supervised training rule (Tempotron) to adjust the parameters of the models, where after presenting an input example, each neuron receives its specific error signal to update the weights. Compared to unsupervised learning rules, updating parameters by supervised learning rules means involving tremendous labeled training samples. It is normally a difficult task with high workload to get large size training datasets. Moreover, it seems unlikely that such a explicit teaching error signal would be implemented in the brain O'Reilly and Munakata (2000), instead, more evidence is pointing toward unsupervised learning rules such as spike-timing-dependent plasticity (STDP) based learning rules (Bi & Poo, 2012) in neuronal system.

In this paper, we take advantage of the hierarchical model as the fundamental framework. In addition, we adopt spiking neurons to construct the classifier for making the final decision. The parameters of this deep CovDenseSNN model are updated by unsupervised learning rules. As a single neuron can be regarded as a dynamic arithmetic operation unit (Silver, 2010), we design algebraic transformation on feature to spikes relation, these rules are embedded into our framework to transfer the real-value based features to specific incoming spatial–temporal spike trains. This structure and functional units are expected to improve the representation invariance and enhance the information representation in neuromorphic visual systems which suggests applicability in heterogeneous biological neural networks.

Furthermore, this paper adopts a novel experimental setting to test the generalization ability of networks named mixed training–testing method, which is implemented by two parts: (1), training on clean-datasets, testing on clean-datasets and noisy-datasets, (2), training on noisy-datasets, testing on clean-datasets and noisy-datasets. We evaluate the deep CovDenseSNN framework on the MNIST and its variations, including learning capabilities, robustness to noisy stimuli and its classification performance. Our model contributes to a better understanding of how the brain builds up a feedforward temporal encoding and learning model based on more biologically feasible principles.

1. Overview of deep CovDenseSNN model

The retina is a functional part in the brain and its structures are remarkably well known. It is widely believed that the retina receives the external stimuli, and extracts the features through neuronal spikes. It is therefore understood that information transmitted from retina to brain encodes the visual stimuli at each specific receptive field.

Inspired by the mechanisms of vision information in biological brain as shown in Fig. 1 (Fig. 1 was adopted from Manassi, Sayim, & Herzog, 2013), we propose a brain-inspired deep CovDenseSNN model which is a mixture of a partial CNN and an SNN as shown in Fig. 2. This model is a neuromorphic visual system which consists of the feature extractor part and decision-making part. The feature extractor part is a partial CNN which acts as the V1–V4 part of visual cortex, which is an interesting sensory area to study neural information processing, since its functional organization and structure are relatively well known. The decision-making part suggests a role which IT (Inferior Temporal) part plays in visual information. The deep CovDenseSNN model is a unified systematic model with feature extraction, consistent encoding, learning and readout parts.

1.1. CNN based feature extractor

Focusing on imitating the information processing in visual sensory system, we use a continuous-valued neuron, partial CNN for extracting features. This CNN acts as the feature extractor.

For the deep CovDenseSNN model in which the image information is captured and filtered by the feature extractor within convolutional and pooling parts. Convolutional layers in this model act as similar to the lateral geniculate nucleus (LGN) part of the brain, since the filters in convolutional layers are believed to mimic how neural processing in the retina of the eyes extract the important information from external stimuli (Krizhevsky, Sutskever, & Hinton, 2012). For the visual system, the LGN is used as the first layer of the cortex to collect information from stimuli, after that, the Lateral Cingulate Cortex (LCC) maintains the dimensions of the features from the local regions in a whole image produced by the retina.

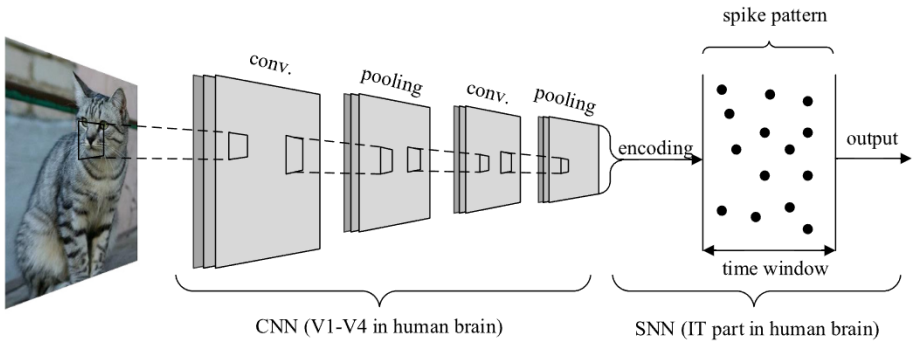

Fig. 2. Visual processing in deep CovDenseSNN model. Deep CNN mimics feature extraction from low-level analysis (edges and lines) to complex figural processing (shapes and objects) and SNN acts as final decision-making classifier.

There is a similarity between the pooling layers in deep CovDenseSNN model and the roles played by LGN layer. A pooling layer applies a nonlinear max pooling operation to its input to achieve invariance. Max pooling over different directions, different scales and different local positions offers contrast scale invariance, reverse invariance and position invariance, respectively. Yu, Giese, and Poggio (2002) have proposed biophysically plausible implementations of the MAX operation. Biological evidence (Lampl, Ferster, Poggio, & Riesenhuber, 2004) of neurons performing MAX-like operation have been found in visual system.

Following the MAX operation, the activation function would trigger the value of the feature maps produced by the pooling layer. The pixel would more easily be activated if its value is larger, whereas numerical small ones would be activated weakly. In this paper, we choose ReLu (Jarrett, Kavukcuoglu, Ranzato, & Lecun, 2010) as CNN's activation function.

1.2. Encoding mechanisms of deep CovDenseSNN model

In deep CovDenseSNN model, we consider the actual values of neurons' activations to generate spikes. In particular, we choose rate based encoding rules as our feature to spike transfer mechanism. Each image is presented to the encoding layer and converted into spatiotemporal pattern. The encoding rule is essential as a mapping between numerical values and spikes.

The rate based encoding (Peter, Daniel, Liu, Tobi, & Michael, 2013) is used to encode images into dense spikes, represented as the average number of spikes counting within a temporal encoding window. A higher firing rate gives higher sensitivity. The rate based encoding always uses dense spikes (the Poisson spike trains) to represent the neurons' firing rate. Compared with the temporal encoding which encodes a pixel to a firing time, rate encoding tends to generate a spike train.

Eq. (1) illustrates the temporal encoding rule which is adopted by CSNN and S1C1-SNN, T_{spike} is the firing time which is calculated from time window T and features A of row pixel. And Eq. (2) gives a example about how to encode a pixel to a spike train in a time window T, generally, the value of the pixel is treated as firing rate r in a Poisson Distribution and $s(t)$ is a Heaviside function to denote the spike firing or not. Temporal encoding is a one pixel to one spike (one to one) rule which uses the accurate firing time to represent a pixel, compared with that, rate based encoding can be regarded as a one pixel to many spikes (one to many) rule which uses a spike train to denote a pixel.

Although larger real value would impose higher computational load on downstream spiking neurons, it has high fault tolerance. For example, if an image has the Gaussian noise, other encoding rules (temporal encoding, sparse encoding and so on) may transfer disparate results compared with rate based encoding, because this rule maps a real value to a spike train and slight noises would not influence spike patterns drastically.

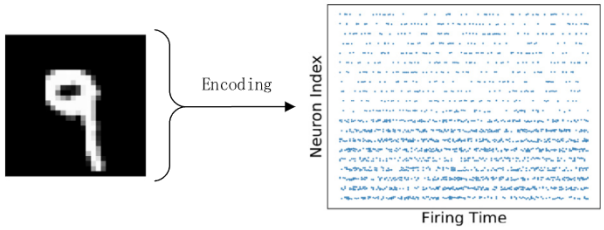

Fig. 3. An image (Digit 9) is encoded to spikes via rate based encoding rule.

$$T_{spike} = T - T * A, \tag{1}$$

$$r = \frac{n_{spikes}}{T} = \frac{1}{T}\int_0^T s(t)dt, \tag{2}$$

$$s(t) = \begin{cases} 0, & t \le 0, \\ 1, & t \ge 0, \end{cases} \tag{3}$$

The activation values are nonlinearly mapped to firing rate of spikes. The raw images are presented to the network for constant time span in the form of Poisson-distributed spike trains, with firing rates proportional to the intensity of the pixels of the pictures.

The deep CovDenseSNN model is capable to extract the basic information from an input image and encode it to a spatio-temporal spiking pattern. A sparse representation is obtained through the encoding process as shown in Fig. 3. The x-axis denotes the time window and the y-axis represents the indices of the neurons firing spikes. This spatio-temporal can translate the image into a sparse spiking pattern, to some extent, is compatible with the biological observations in the visual system of the brain.

1.3. SNN based classifier

In the SNN part of the deep CovDenseSNN model, we adopt leaky integrate-and-fire (LIF) neuron model (Hu, Tang, Tan, Li, & Shi, 2013) as this framework's fundamental unit. LIF neuron model is adopted in modeling an SNN because of its strong biology support and effective computation. There are several variants of the LIF. The simplest and the most widely used variant is the current-based LIF model, which is voltage based. The membrane potential V of a LIF neuron is governed by the following

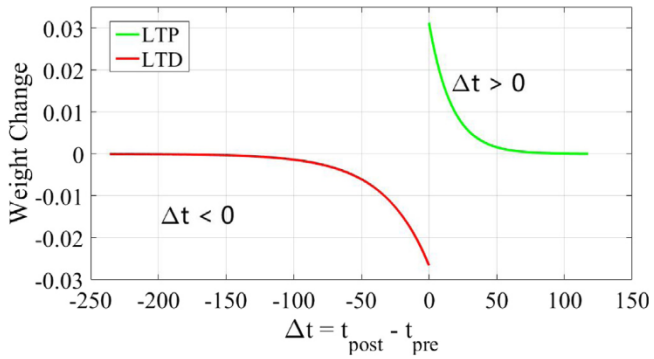

Fig. 4. The learning window of the STDP rules.

equations.

$$C_m \frac{dV}{dt} = g_l(E_l - V) + I, \tag{4}$$

$$V = V_{rest}, \quad if\ V \geq V_{th}, \tag{5}$$

where, C_m denotes membrane capacitance, g_l denotes conductance (inverse of resistance) of the leakage channels, E_l denotes the equilibrium potential of the leakage channels, and I denotes total input current. The membrane potential $V(t)$ of a LIF neuron is weighted sum of postsynaptic potentials (PSP) from all afferent stimuli.

Unsupervised learning based on Hebb's rule, is stated informally as: Neurons that fire together, wire together. More formally, an increase in synaptic efficacy arises from the presynaptic neuron's repeated and persistent stimulation of the post-synaptic neuron. Spiking Timing Dependent Plasticity (STDP) is a Hebbian learning rule for SNNs: for two connected neurons, if the pre-synaptic neuron A always fires within a small time window before the post-synaptic neuron B fires, which means firing of B is correlated with firing of A, then the strength of the synapse between A and B is increased (long-term potentiation, LTP); vice versa, so called long-term depression (LTD), as shown in Fig. 4.

$$\Delta W_{ij} = \begin{cases} M_+ exp(\frac{t_j - t_i}{\tau^+}), & if\ t_j < t_i\ (LTP), \\ M_- exp(\frac{t_i - t_j}{\tau^-}), & if\ t_j > t_i\ (LTD), \end{cases} \tag{6}$$

The STDP rule can be described as Eq. (6), where τ^+ and τ^- control the ranges of pre-post synaptic intervals which affect synaptic strengthening (LTP) or weakening (LTD). M_+ and M_- are the learning rates to determine the maximum amounts of synaptic changes for LTP and LTD, respectively.

The SNN part is similar to the previous framework (Diehl & Cook, 2015). As shown in Fig. 5, this figure comes from Diehl and Cook (2015), the SNN composes a layer with excitatory neurons and a layer with inhibitory neurons. The excitatory neurons are connected to inhibitory neurons through one-to-one connections and each inhibitory neuron is connected to all excitatory neurons, but apart from the one which receives a connection from. This connectivity provides lateral inhibition and leads to competition among excitatory neurons.

2. Experimental results

In this section, we evaluate deep CovDenseSNN model on three benchmark datasets: basic MNIST (Lecun, Bottou, Bengio, & Haffner, 1998), background MNIST (Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007) and background-random MNIST (Larochelle et al., 2007) as shown in Fig. 6. For convenience, their names are abbreviated as basic, bg and bg-rand, respectively. Each dataset consists of 28 × 28 grayscale images of handwritten digits from 0 to 9 and the dataset is divided into two parts: training set (50,000 training samples) and test set (10,000 test samples).

In order to show the networks' generalization ability in the noisy environment caused by bg and bg-rand MNIST, we divide the sizes of the training set and test set to verify that deep CovDenseSNN can achieve better performance on small-size training sets than other cognitive models. For examples, when the training samples are 500 and the test samples are 100, which means we choose the 500 training samples from the whole 50,000 training samples randomly and they are evenly distributed in ten classes. All of the experimental results are obtained from ten independent replicates.

The training method of this deep model follows two steps: Firstly, training a full CNN with the Stochastic Gradient Descent (SGD) algorithm such as the Backpropagation (BP). After that, only the convolutional and pooling layers of the trained CNN are kept, while the fully-connected layers are discarded. Then, training the SNN part of the model according to the unsupervised rule (STDP). During the training phase of the SNN, parameters of the CNN are fixed.

2.1. Experimental settings

The experiments are conducted on a server equipped with two-processor Intel(R) Xeon(R) Core CPU and two NVidia GeForce GTX 1080Ti GPUs. The operating system is Ubuntu 16.04. We use Tensorflow Abadi, Agarwal, Barham, Brevdo, Chen, Citro, Corrado, Davis, Dean, Devin, et al. (2016) and Brain Dan and Brette (2008) for training and testing the proposed deep CovDenseSNN model.

For basic, bg and bg-rand MNIST datasets, we trained a CNN as the feature extractor and adopted its convolutional and pooling layers as feature extractor. Its architecture is 6C6@28x28-12C5-24C5-P and the SNN architecture which has been described above. The partial CNN in deep CovDenseSNN model consists of several convolutional and pooling layers as the aforementioned description. Furthermore, in order to alleviate the overfitting (Xu, Zhang, Gu, & Pan, 2019) in SNN, especially when the training samples are limited and noisy. We adopt two different SNN sizes, to be specific, when the training samples are less than 10,000, the excitatory neurons and inhibitory neurons are 100 respectively. And when the training samples are equal to or greater than 10,000, the excitatory/inhibitory neurons in SNN are 400. Besides the CNN size, the SNN part in the proposed system is also changeable according to the size of training samples.

2.2. Evaluation of hierarchical structures

The CNN part of deep CovDenseSNN model consists of convolutional and pooling layers, channels of convolutional kernels will produce multifarious feature maps. The scale of the partial CNN can be adjusted according to the size of the training datasets. We compare the proposed deep CovDenseSNN model with one fully-connected layer based SNN named Un-stdp (Diehl & Cook, 2015) to show that the hierarchical structures can extract more important information which is helpful for a robust training of the SNN. From the structure of the convolutional feature maps, the CNN can extract a higher degree of abstract features than the Un-stdp.

As shown in Fig. 7, for both networks, we train them on basic MNIST and test the performance on its corresponding test sets. We observe that the deep CovDenseSNN model can achieve better classification accuracies than Un-stdp in all cases except the case that training samples are 5000 and test samples are 1000, but the

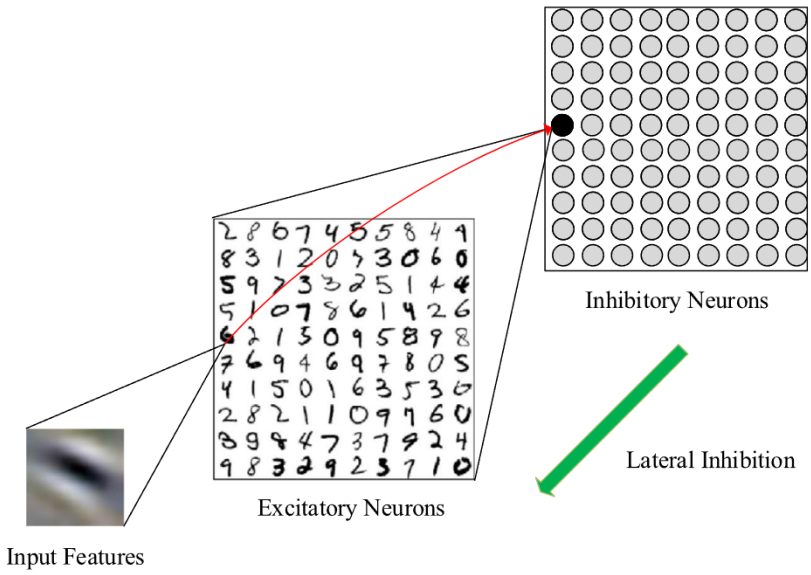

Fig. 5. Network architecture of the SNN part in deep CovDenseSNN.

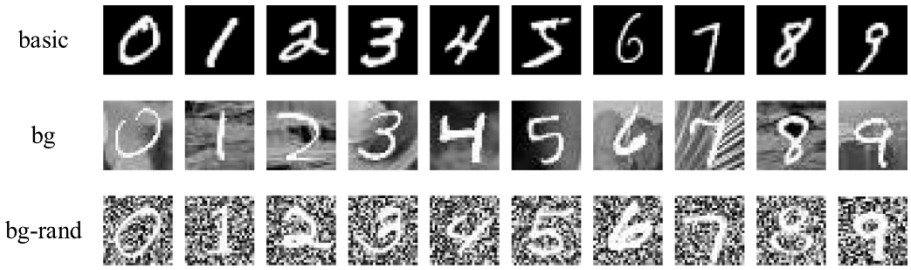

Fig. 6. MNIST and its variations.

Table 1
Classification accuracy (%) of Un-stdp on clean/noisy MNIST test sets.

Train Test	Basic Basic	Basic bg	Basic bg-rand
500/100	**76.0**	12.3	11.2
2500/500	**82.0**	13.4	12.6
5000/1000	**88.0**	14.8	13.2
10000/2000	**90.4**	15.7	15.2

disparity is not significant between deep CovDenseSNN (86.2%) and Un-stdp (88.0%). Furthermore, when training samples are limited (500, 1000), the gap of accuracies between deep CovDenseSNN and Un-stdp is big. The deep CovDenseSNN can achieve 82.0% and 85.5% when the training samples are 500 and 1000, meanwhile, the Un-stdp only has 76.0% and 79.0%, respectively. But the gap is closing with the samples of training set increasing. It is reasonable that our framework is a hierarchical structure compared to Un-stdp (one fully-connected layer), the hierarchical system can extract more important information which is helpful for training the following classifier. With the training samples increasing, the accuracies of deep CovDenseSNN have some slight fluctuations and peaks at 92.2% when trained on a training set with 40,000 samples, and the result of Un-stdp is 90.6%. Although the gap between Un-stdp and deep CovDenseSNN is closing, the presented deep CovDenseSNN model behaves better than Un-stdp as a whole.

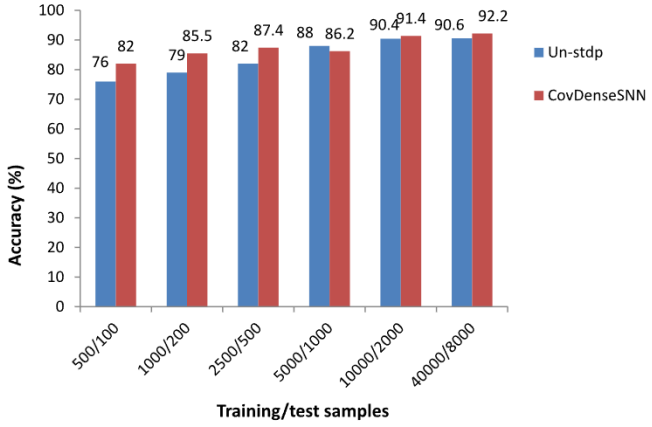

Fig. 7. Test accuracy (%) of basic MNIST from deep CovDenseSNN and Un-stdp.

From Fig. 7, we can conclude that the proposed deep CovDenseSNN model has strong robustness on different sizes of training sets. And the Un-stdp cannot perform well if the training samples are limited. In addition, we trained Un-stdp on clean datasets (basic) and test it on basic, bg and bg-rand MNIST, the results are reported in Table 1, all of the classification accuracies from the noisy test sets (bg and bg-rand) are poor (below 16.0%). And the performance from the clean test set (basic) is significantly

Table 2
Comparison of training methods and feature extraction ability between S1C1-SNN (unsupervised training) and CSNN (supervised training) and deep CovDenseSNN (unsupervised training). Each single table cell has three classification accuracies from S1C1-SNN (**Left**), CSNN (**Center**) and deep CovDenseSNN (**Right**).

Training	Basic		
Test	Basic	bg	bg-rand
500/100	75.0/81.0/82.0	30.0/27.0/11.0	19.0/38.0/9.3
1000/200	78.5/87.0/85.5	29.0/26.0/12.5	14.5/23.5/12.2
5000/1000	77.3/84.7/86.2	32.4/23.5/14.6	15.6/25.9/12.3
10000/2000	77.4/86.1/91.4	31.3/22.3/16.2	12.3/24.0/13.4
40000/8000	76.0/83.8/92.2	29.2/34.7/33.3	14.3/24.4/25.6
Training	bg		
Test	Basic	bg	bg-rand
500/100	45.0/43.0/52.0	28.0/60.0/26.0	15.0/64.0/30.0
1000/200	58.0/47.0/52.6	34.0/33.5/31.0	15.5/38.5/37.5
5000/1000	64.5/52.2/52.7	25.6/39.8/33.3	12.2/40.6/44.8
10000/2000	60.2/52.9/52.4	19.8/41.8/48.1	10.8/39.9/44.9
40000/8000	59.2/52.6/53.5	20.2/39.6/52.2	11.3/41.6/59.6
Training	bg-rand		
Test	Basic	bg	bg-rand
500/100	21.0/9.0/12.0	20.0/44.0/20.0	20.0/62.0/40.4
1000/200	19.0/24.0/11.5	21.0/34.5/33.2	15.5/42.5/43.6
5000/1000	10.9/27.3/17.5	14.8/29.4/39.2	12.2/39.0/47.3
10000/2000	11.4/23.3/45.6	15.4/27.2/49.0	12.4/39.2/52.1
40000/8000	9.8/21.5/61.3	13.1/26.3/60.5	11.6/37.0/66.4

better (from 76.0% to 90.4%) within the different sizes of training samples than those from the noisy test tests. It leads to another conclusion, although neurons from SNNs are more biological than artificial neurons from ANNs, the noise immunity of the whole network also depends on the other factors such as hierarchical structures.

2.3. Comparison of training methods between supervised and unsupervised learning rules

To show the advanced unsupervised training method and feature extraction ability adopted by deep CovDenseSNN, we implement and compare our deep CovDenseSNN model with the other two hierarchical networks S1C1-SNN (Yu et al., 2013) and CSNN (Xu et al., 2018) which were trained by supervised training rule.

S1C1-SNN is an SNN based visual system which chooses Gabor filter and max operation as feature extractor, CSNN is also a neuromorphic image classification framework. Since they are both novel hierarchical SNN based framework, besides comparing the classification performance on basic MNIST, we adopt an experimental setting to test the generalization ability of networks named mixed training–testing method which consists of two parts: (1), training the networks on clean-datasets, testing the performance on clean-datasets and noisy-datasets, (2), training them on noisy datasets, testing the performance on clean and noisy datasets. These conditions are expected to help test the different degree of generalization ability. Hence, there are 9 cases: training set is basic MNIST, test sets are basic, bg and bg-rand MNIST and so on.

Test accuracies of all 9 conditional cases are shown in Table 2. The left result in the table cell is the classification accuracy from S1C1-SNN, the middle one is from the CSNN and the right figure is from our proposed model the deep CovDenseSNN. From this table, we can observe that when the training set is clean (basic), the test accuracies of the deep CovDenseSNN are significant better than those from the other two models. For example, when they are all trained on basic MNIST and tested on its corresponding test set, the deep CovDenseSNN reaches nearly 87.5% no matter how the size of the dataset changes, compared to that, S1C1-SNN and CSNN only achieve about 77% and 86% which depend on the amount of the training samples obviously. The other two cases training on basic MNIST and testing on bg and bg-rand MNIST report the similar results, although in a few conditionals, S1C1-SNN and CSNN perform better than deep CovDenseSNN, by and large the deep CovDenseSNN behaves better than the other two networks. Because the SNNs in S1C1-SNN and CSNN are fully connected, and the SNN in our proposed framework has more complex structure and different connection types (excitatory and inhibitory mechanisms), although they are trained via supervised learning rules the Tempotron.

As for training on noisy datasets (bg and bg-rand MNIST), the performance gap between the deep CovDenseSNN and the other two models is large. From Table 2, we can see that when the training size is limited, the deep CovDenseSNN achieves significant better accuracies than S1C1-SNN and CSNN. For example, the accuracy from deep CovDenseSNN is 52.0% when adopting 500 bg MNIST training samples and testing on basic MNIST, meanwhile the figures from the S1C1-SNN and CSNN are just 45.0% and 43.0%, respectively. Furthermore, the gap is closing with the increase of training samples. When training set is bg MNIST, these three models nearly behave the same. When the training set is bg-rand MNIST, the deep CovDenseSNN still performs robustly and better than them. This is due to that the encoding mechanisms of the S1C1-SNN and CSNN are just temporal encoding rules, compared to rate based encoding rules of deep CovDenseSNN model, they are poor in noisy environments. Another important reason why the proposed framework behaves better is that its structure is more reasonable and deeper than S1C1-SNN and CSNN. These results also mean that the proper and deep structure is better than shallow one as mentioned in the aforementioned section. And all test accuracies from the cases trained on noisy datasets (bg and bg-rand MNIST) are still low except the results from using their corresponding test sets.

From the above experimental results, we observe that although other networks such as S1C1-SNN and CSNN are hierarchical systems and they adopt supervised learning rules to train the classifier, they are still shallow frameworks and their classification performance lies on various factors. The proposed deep CovDenseSNN network can behave better in spite of that training a network via unsupervised learning rules, the deep structure could enhance the feature extraction ability of the networks. Additionally, the encoding rules of the deep CovDenseSNN (one-to-many rated encoding) are more suitable than one-to-one temporal encoding in deep structures, especially when the training set is noisy. With the advantages of more powerful structures and encoding mechanisms, the deep CovDenseSNN is a more powerful in feature extractor than the other two hierarchical models. Furthermore, the proposed framework is trained by the unsupervised learning rule STDP, which is more biologically plausible.

2.4. Classification performance comparison with other models

The presented deep CovDenseSNN network achieves good classification on the MNIST and its variations with the partial CNN made of hierarchical structures and the SNN trained by unsupervised learning rules. A comparison of SNN based cognitive models for benchmark basic MNIST is shown in Table 3. We compare our deep CovDenseSNN model to some state-of-the-art models: S1C1-SNN, CSNN, Spiking RBM (Merolla et al., 2011), Dendritic Neurons (Hussain et al., 2014), Un-stdp (Diehl & Cook, 2015) and Multi-layer hierarchical network (Beyeler et al., 2013).

Because the scale of SNN is not fixed, the size of each model is various. The training and test samples are adjusted according

Table 3
Comparison of classification accuracy of spiking neural networks on basic MNIST test set.

Network type	Encoding methods	Training rules	Training/test samples	Acc (%)
S1C1-SNN (Yu et al., 2013)	Temporal	Tempotron (sup.)	500/100	78.0
CSNN (Xu et al., 2018)	Temporal	Tempotron (sup.)	10000/2000	87.0
Spiking RBM (Merolla et al., 2011)	Rate-based	Contrastive divergence (sup.)	60000/10000	89.0
Dendritic Neurons (Hussain, Liu, & Basu, 2014)	Rate-based	Morphology learning (sup.)	10000/5000	90.3
Un-stdp (Diehl & Cook, 2015)	Rate-based	STDP (unsup.)	40000/8000	90.6
Multi-Net (Beyeler, Dutt, & Krichmar, 2013)	Temporal	STDP with calcium (sup.)	2000/1000	91.6
Deep CovDenseSNN	**Rate-based**	**STDP (unsup.)**	**500/100**	**82.0**
Deep CovDenseSNN	**Rate-based**	**STDP (unsup.)**	**2000/1000**	**88.0**
Deep CovDenseSNN	**Rate-based**	**STDP (unsup.)**	**10000/2000**	**91.4**

to the network capacity. Table 3 illustrates the details of test accuracies of the basic MNIST from the different cognitive models. From this table, we observe that when the training samples are limited (500 training samples), the deep CovDenseSNN achieves 82.0%, whereas S1C1-SNN only achieves 78% under the same experimental conditionals.

As for CSNN and Multi-Net, they all adopt temporal encoding methods and supervised learning rules. Due to the limit of the network capacity (300 neurons), CSNN only achieves 87.0% on 10000 training-sample datasets. Multi-Net performs best in all items and gets 91.6% classification with only 2000 training samples.

Other models such as Spiking RBM, Dendritic Neurons and Un-stdp can achieve 89.0%, 90.3% and 90.6% under the different sizes of training samples, respectively. The reason why they perform well lies in the fact that they either adopt supervised learning rules or employ large number of training samples.

Although the deep CovDenseSNN model cannot get the best performance (91.4%) compared to result from Multi-Net (91.6%), it can behave well on small-size training sets. Compared to supervised learning rules adopted by Multi-Net, deep CovDenseSNN is trained by unsupervised which is widely believed to close to the nature of neural computation in the brain. And the Multi-Net holds a large quantity of parameters (71,026 neurons) than deep CovDenseSNN holds (the maximum 800 neurons, the minimum is 200 neurons), more parameters mean higher cost at computation power, storage, network bandwidth, power consumption and so on. Besides, the proposed framework can behave better with the increase of training samples which benefit by the scalability of its structure.

3. Conclusion

In this paper, our ConvDenseSNN: a hybrid spike-based learning framework uses a continuous valued CNN for feature extraction and an SNN for classifying is proposed. This recognition model combines feature extraction ability of CNNs and biological plausibility of SNNs. With the help of the feature extraction, robust encoding mechanisms and unsupervised learning rules (STDP), this visual system can encode the external stimuli (images) to spatiotemporal patterns. These advantages also make the whole system become more robust especially in the noisy environment. We show the performance of the presented network applied to MNIST and its variations is comparable to the other novel cognitive models: S1C1-SNN, CSNN, spiking RBM, Dendritic Neurons, Un-stdp and Multi-Net, but computed in a more efficient way. We argue that the structure and learning methods adopted by deep CovDenseSNN can help to extract more important features and lead to train a more robust cognitive model and efficient recognition.

It would also be beneficial for implementations of neuromorphic chips with the aid of its structure. Furthermore, this work proposes a more biological realistic framework which could be applied into pattern recognition tasks such as image classification, with the hope that this model can help us understand how the mammalian neocortex is performing computations especially in high-level vision task.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2017YFB1002503), Ten Thousand Talent Program of Zhejiang Province (No. 2018R52039) and Zhejiang University Academic Award for Outstanding Doctoral Candidates (No. 2018085).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. CoRR, abs/1502.04156 . URL: http://arxiv.org/abs/1502.04156. arXiv:1502.04156.

Beyeler, M., Dutt, N. D., & Krichmar, J. L. (2013). Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule. *Neural Networks*, *48*(10), 109–124.

Bi, G., & Poo, M. (2012). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, *18*(24), 10464–10472.

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*(2), 113.

Dan, G., & Brette, R. (2008). Brian: A simulator for spiking neural networks in python.. *BMC Neuroscience*, *9*(1), 1–2.

Diehl, P. U., & Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, *9*, 99.

Hu, J., Tang, H., Tan, K. C., Li, H., & Shi, L. (2013). A spike-timing-based integrated model for pattern recognition.. *Neural Computation*, *25*(2), 450–472.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Hussain, S., Liu, S.-C., & Basu, A. (2014). Improved margin multi-class classification using dendritic neurons with morphological learning. In: *IEEE international symposium on circuits and systems* (pp. 2640–2643).

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & Lecun, Y. (2010). What is the best multi-stage architecture for object recognition? In: *IEEE international conference on computer vision* (pp. 2146–2153).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In: *International conference on neural information processing systems* (pp. 1097–1105).

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, *92*(5), 2704.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In: *International conference on machine learning* (pp. 473–480).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Liu, J. K., & Buonomano, D. V. (2009). Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner. *Journal of Neuroscience the Official Journal of the Society for Neuroscience*, *29*(42), 13172–13181.

Maheswaranathan, N., McIntosh, L. T., Kastner, D. B., Melander, J., Brezovec, L., Nayebi, A., Wang, J., Ganguli, S., & Baccus, S. A. (2018). Deep learning models reveal internal structure and diverse computa- tions in the retina under natural scenes. bioRxiv, . URL: https://www.biorxiv.org/content/early/2018/06/14/340943. http://dx.doi.org/10.1101/340943. arXiv:https://www.biorxiv.org/content/early/2018/06/14/340943.full.pdf.

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding.. *Journal of Vision*, *13*(13), 10.

Merolla, P., Arthur, J., Akopyan, F., Imam, N., Manohar, R., & Modha, D. S. (2011). A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm. In: *Custom integrated circuits conference* (pp. 1–4).

Orchard, G., Meyer, C., Etienne-Cummings, R., Posch, C., Thakor, N., & Benosman, R. (2015). Hfirst: A temporal approach to object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(10), 2028–2040.

O'Reilly, R. C., & Munakata, Y. (2000). Computational explorations in cog- nitive neuroscience: Understanding the mind by simulating the brain., . 5, 225–225.

Peter, O., Daniel, N., Liu, S. C., Tobi, D., & Michael, P. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in Neuroscience*, *7*, 178.

Qiu, J., Sun, K., Wang, T., & Gao, H. (2019). Observer-based fuzzy adaptive event-triggered control for pure-feedback nonlinear systems with prescribed performance. *IEEE Transactions on Fuzzy Systems*, *PP*(99), 1.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization.. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–6429.

Silver, R. A. (2010). Neuronal arithmetic. *Nature Reviews Neuroscience*, *11*(7), 474–489.

Sun, K., Mou, S., Qiu, J., Wang, T., & Gao, H. (2019). Adaptive fuzzy control for non-triangular structural stochastic switched nonlinear systems with full state constraints. *IEEE Transactions on Fuzzy Systems*, *PP*(99), 1.

Wen, H., Shi, J., Zhang, Y., Lu, K. H., & Liu, Z. (2016). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12).

Xu, Q., Qi, Y., Yu, H., Shen, J., Tang, h., & Pan, G. (2018). CSNN: An augmented spiking based framework with perceptron-inception. In: *Twenty- seventh international joint conference on artificial intelligence* (pp. 3560–3566).

Xu, Q., Zhang, M., Gu, Z., & Pan, G. (2019). Overfitting remedy by sparsifying regularization on fully-connected layers of cnns. *Neurocomputing*, *328*, 69–74.

Yan, Q., Zheng, Y., Jia, S., Zhang, Y., Yu, Z., Chen, F., Tian, Y., Huang, T., & Liu, J. K. (2018). Revealing fine structures of the retinal receptive field by deep learning networks. CoRR, abs/1811.02290 . URL: http://arxiv.org/abs/1811.02290. arXiv:1811.02290.

Yu, A. J., Giese, M. A., & Poggio, T. A. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Computation*, *14*(12), 2857.

Yu, Q., Tang, H., Tan, K. C., & Li, H. (2013). Rapid feedforward computation by temporal encoding and learning with spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(10), 1539–1552.