Neurocomputing 277 (2018) 237–248

Contents lists available at ScienceDirect

Neurocomputing

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

On the application of reservoir computing networks for noisy image recognition

Azarakhsh Jalalvand[a,*], Kris Demuynck[a], Wesley De Neve[a,b], Jean-Pierre Martens[a]

[a] IDLab, Ghent University–imec, B–9052, Ghent, Belgium
[b] Image and Video Systems Lab, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

ARTICLE INFO

Article history:
Received 16 June 2016
Revised 19 September 2016
Accepted 12 November 2016
Available online 24 August 2017

Keywords:
Reservoir computing networks
Recurrent neural networks
Text recognition
Image classification
Image denoising

ABSTRACT

Reservoir Computing Networks (RCNs) are a special type of single layer recurrent neural networks, in which the input and the recurrent connections are randomly generated and only the output weights are trained. Besides the ability to process temporal information, the key points of RCN are easy training and robustness against noise. Recently, we introduced a simple strategy to tune the parameters of RCNs. Evaluation in the domain of noise robust speech recognition proved that this method was effective. The aim of this work is to extend that study to the field of image processing, by showing that the proposed parameter tuning procedure is equally valid in the field of image processing and conforming that RCNs are apt at temporal modeling and are robust with respect to noise. In particular, we investigate the potential of RCNs in achieving competitive performance on the well-known MNIST dataset by following the aforementioned parameter optimizing strategy. Moreover, we achieve good noise robust recognition by utilizing such a network to denoise images and supplying them to a recognizer that is solely trained on clean images. The experiments demonstrate that the proposed RCN-based handwritten digit recognizer achieves an error rate of 0.81 percent on the clean test data of the MNIST benchmark and that the proposed RCN-based denoiser can effectively reduce the error rate on the various types of noise.

1. Introduction

Thanks to the advances in the structure of the neural networks since the early 80's, such as introducing the concept of Deep Neural Networks (DNNs) [1,2] and Convolutional Neural Network (CNN) [3] along with the more powerful computational hardware, image processing have become more elegant than ever. For instance, in recent devices the traditional keyboards are being replaced with modern interfaces such as touchscreens that input text via handwriting (e.g., TV touchpad remotes). Due to the different handwriting styles, there is a lot of variability in the images of the same character, making automatic handwriting recognition (HWR) a challenging task.

The presence of background noise is another source of variability the HWR system may have to deal with. In fact, in applications such as address recognition on parcels or full text recognition from digital scans of old manuscripts or typed documents, noise corrupted images such as the one depicted in Fig. 1 are the norm. In this paper, we show that reservoir computing networks (RCNs) have great potential for achieving good performance in HWR from noise corrupted images. We demonstrate this on the MNIST [3] dataset, a handwritten digit recognition task (HDR) used by many research groups to benchmark their technologies.

More than two decades ago, Multilayer Perceptrons (MLPs) [4] were among the first classifiers that were tested on MNIST. In [3], an MLP with two computational layers of neurons was reported to reach a digit error rate (DER) of 2.95%, and a later study [5] reported a DER of 1.60%. Employing MLPs with more layers was long time believed to yield no significant improvement. However, new training methods, permitting a better exploitation of multiple hidden layers, were recently discovered and gave rise to the emergence of Deep Neural Networks [6]. Differences in the details lead to DNNs of various types. Two of them, Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) have also been tested on MNIST (see Table 1). Roughly speaking, they achieve a DER of about 1%.

It is, however, generally acknowledged that conventional and modern neural networks such as DNNs perform well, but that they are still hard to train: it takes a lot of time and the hyperparameters of the training process must be set properly. More recent approaches such as *dropout* [7] and *maxout* [8] are examples of efforts to both facilitate improved training and improved effectiveness of the models.

* Corresponding author.
E-mail address: azarakhsh.jalalvand@ugent.be (A. Jalalvand).

http://dx.doi.org/10.1016/j.neucom.2016.11.100
0925-2312/

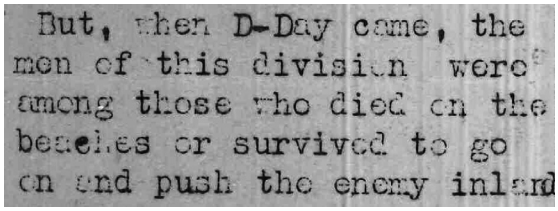

Fig. 1. Part of the military newspaper "The Stars and Stripes" published in 1944.

Table 1
Reference results on MNIST using the original training set and using an expanded version of the training set (for example, by applying deformation). The presented DERs are accompanied by a reference to the paper introducing the technique that was used.

System	DER% (Original training set)	DER% (Enriched training set)
LSTM [9,10]	1.80	0.32
2-layer MLP [5]	1.60	0.70
MLP + dropout [11]	1.05	–
DBN [12]	1.03	–
DBM [13]	0.95	–
CNN [3]	0.95	0.80
MLP + maxout + dropout [8]	0.94	–
ELM [14]	0.86	–
DCN [15,16]	0.83	0.35
DBM + dropout [7]	0.79	–
Large CNN [17]	0.60	0.39
CNN + maxout + dropout [8]	0.45	–
Multi-CNN [18]	–	0.23

Nonetheless, long before deep neural networks became successful, significant improvement over a standard 2-layer MLP was achieved by means of a Convolutional Neural Network (CNN) [3] that acts like a feature extractor. In fact, one of the main points of criticism raised against an MLP was that its hidden neurons see the whole image and are therefore bound to overlook the local topological relations between adjacent pixels undoubtedly present in sub-regions of the image [3]. Hence, the idea was to scan the image, to filter the pixels appearing in the emerging sub-regions by means of trainable filters and to down-sample the filtered outputs so as to create a rich and compact feature representation that constitutes a more suitable input to the MLP-based classifier. The first results obtained with the CNN approach were already mentioned in [3]. With a DER of 0.95%, CNN-based systems can still be considered as state-of-the-art for HDR and the CNN-based features are being used in many complex visual understanding models [19].

Obviously, there is no reason why the concepts of a CNN and a DNN could not be combined. Deep Convolutional Neural Networks are the exponent of that idea, leading to a DER of 0.83%.

Another idea that induced a significant boost in HDR, was to enrich the original training dataset with new images, obtained by deforming the raw training images. In [5], for instance, elastic deformations were applied to the raw images achieving a convincing drop in DER from 1.6% to 0.7%. Since then, basically all novel methods have shown to benefit from such an enrichment of the training set (see right column of Table 1). By also introducing separate DNNs for different digit widths (6 classes), it was even possible to achieve human-competitive performance (0.23%) [18].

In spite of the spectacular performances achieved in clean conditions, all aforementioned approaches fail dramatically when recognizing digits from noisy samples. In [12], for instance, it was shown that a DBN trained on clean samples, fails completely when recognizing noisy samples. The DER raises to 33.8% when the digits are partially masked by square blocks and to 66.1% when the digits are surrounded by a black border (see Fig. 2). Consequently, new research has been directed towards improving the robustness of HDR against the presence of noise.

In general, one can roughly distinguish three approaches: (1) add noise to the training examples and perform a so-called *multi-conditional training* of the neural network, (2) make the classifier intrinsically more robust against the effects induced by noise, for example, by using a sparsely connected DBN rather than a conventional fully connected one [12], and (3) remove a large part of the noise from the input image before presenting it to the classifier. In [12], it was argued that due to the noise, a lot of neurons are driven into saturation and are therefore not contributing to the recognition anymore. By training it on noisy images, the standard DBN could be made much more effective. The DER could be reduced from 33.8% to 8.7% for the case of block noise and from 66.1% to 1.9% for the case of border noise.

In some other work [20,21], a stacked sparse DBN-based denoising auto-encoder (SSDA) is trained to denoise the images. In such a system, one SSDA per noise type was trained and the denoised image is obtained as a linear combination of the individual SSDA outputs. Feeding these images to a DBN trained on clean samples induced a dramatic improvement. The average error rate was reduced from 34.3% (an average over five noise types) to 2.4%. Examples of the noise types are depicted in Fig. 2 and Table 5 lists the improvements per noise type. As the combination weights are determined by a weight prediction module, the latter system was called an adaptive multi-column SSDA (AMC-SSDA) system.

Besides the noise-robustness, the incapability of processing temporal information is another challenge in expanding the application of these techniques to process sequential data such as continuous text and videos. Long short-term memory systems (LSTM) and some more complex CNNs have been proposed and used to address this weakness with the purpose of motion picture classification [9,10,22].

In this paper, the focus is on reservoir computing networks (RCNs) [23], which are a special type of recurrent neural networks. As was shown in [24,25], RCNs in combination with the proposed simple but effective training procedure, can provide adequate solutions in the field of speech recognition and noise robust speech processing. The aim of this paper is to investigate if the same training procedure is applicable in the domain of image recognition and if the strong points of RCNs such as good temporal modeling and noise robustness transfer to the new domain as well.

Although developed in parallel, on the conceptual level, an RCN can be considered as an extension of the Extreme Learning Machine (ELM) proposed in [26]. An ELM is a two-layer MLP with a randomly initialized and afterwards fixed (i.e. non-trained) hidden layer of non-linear neurons followed by an output layer of linear neurons whose weights are determined so as to minimize the mean squared difference between the computed and the desired outputs. Under these constraints, there exists a closed-form solution for the weights which can be obtained by inverting a squared matrix and performing some additional matrix multiplications.

ELMs have been proven to be effective, efficient and robust algorithms for pattern classification. In the past years, several versions of ELMs have been introduced to tackle the different challenges in the field of machine learning. For instance, due to the difficulty in obtaining the labeled data, Huang et, al. [27] proposed a semi-supervised ELM for classification and an unsupervised ELM for clustering. Other solutions to the dilemma of insufficient labeled data are domain adaptation and transfer learning [28]. In this respect, Zhang and Zhang [29] extended ELMs to handle domain adaptation problems for improving the transferring capability of ELM between multiple domains with very few labeled guide instances in target domain, and overcome the generalization disadvantages of ELM in multi-domains application.