Convolutional Neural Networks for Scene Image Recognition

Yulian Li[1], Chao Luo[1], Hao Yang[1,2], and Tao Wu[1(✉)]

[1] School of Computer Science, Chengdu University of Information Technology, Chengdu 610025, China
wut@cuit.edu.cn
[2] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China

Abstract. Words are the most indispensable information in human life. It is important to analyze and understand the meaning of words. Compared with the general visual elements, the text conveys rich and high-level meaning information, which enables the computer to better understand the semantic content of the text. With the rapid development of computer technology, the research on text information recognition has made great achievements. However, in the face of dealing with text characters in natural scenes, there are certain limitations in the recognition of natural scene images. Because scene images have more interference and complexity than text, these factors make the identification of natural scene image texts facing many challenges. This paper focused on the recognition of natural scene image texts, and mainly studied a text recognition method based on deep learning in natural scene images. Firstly, text recognition is based on Kares using the Dense Convolutional Network (DenseNet) network model by using the existing standard test data set. Secondly, each character is classified using Softmax outputs to achieve the use of automatic learning. The characteristics of the context replace the manually defined features, which improve the recognition efficiency and accuracy. Lastly, the text recognition of the natural scene image is realized. And the method is suitable for problems encountered in medical images.

Keywords: Recognition · Deep learning · DenseNet · Softmax

1 Introduction

Words, as the crystallization of human wisdom, are one of the most important symbols of human civilization. Since ancient times, words have played an indispensable role in our lives. Text contains rich and accurate semantic information that is widely used in visual understanding-based tasks, so natural scene text detection and recognition becomes more and more important and becomes.

A research hotspot in computer vision and document analysis. In recent years, a large amount of research results and great research progress have been made in this field, but there are still many challenges, such as noise, blur and distortion, for text extraction and recognition in natural scene images. With the rapid development of computer technology and handheld mobile camera equipment and the popularity of

X. Sun et al. (Eds.): ICAIS 2019, LNCS 11632, pp. 463–474, 2019.
https://doi.org/10.1007/978-3-030-24274-9_42

web 2.0 technology, we have ushered in the era of media, everyone can shoot and upload pictures anytime, anywhere, resulting in a surge in the number of images containing various natural scenes. These massive natural scene images greatly exceed the intensity of manual processing, and automation of image management and retrieval is urgently required. In natural scene images or videos, it often contains a lot of text information, such as notices, road signs, and captions. The text contains rich and accurate semantic information that is widely used in visual understanding, such as: autopilot, blind navigation, license plate recognition, and text-based image retrieval. Extracting text information from natural scene images can help us to further understand the image and retrieve the required information from massive data, saving time and improving efficiency. Therefore, text detection and recognition in natural scenes are becoming more and more important. It has gradually become a very active research topic in computer vision and document analysis. Internationally, the emphasis on the detection and recognition of natural scene texts has gradually increased, and many well-known journals and conferences have included them in the agenda. To promote the development of this field, since 1991, the International Conference on Document Analysis and Recognition (ICDAR) has been held every two years. In recent years, researchers have done a lot of research and made great progress in this field. Although the traditional document text detection and recognition has matured, there are still many insurmountable text detection and recognition for natural scene images. Challenges such as noise, blur, distortion, and non-uniform illumination. Natural scene text detection and recognition belongs to the intersection of pattern recognition, computer vision, machine learning, etc., and the research results can promote the theoretical development of these fields. Therefore, text detection and recognition in natural scenes have important theoretical research and practical application value. Based on the in-depth analysis of the current progress and challenges in this field, this paper combines the current popular deep learning techniques to design and implement a text recognition method based on deep learning.

Text recognition is to obtain semantic information in a picture, and images containing semantic information are generally pre-cut. Since the natural scene image is very different from the document images, the conventional character recognition method cannot be directly applied to the text recognition in the scene image. In recent years, researchers have done a lot of research on text recognition in natural scenes. Text recognition is a process of transforming image information into a sequence of symbols that can be represented and processed by a computer. The text recognition task can be considered as a special translation process in which the image signal is translated into a natural language. This is similar to speech recognition and machine translation. From a mathematical point of view, they convert a set of input sequences containing a lot of noise into an output sequence of a given set of labels.

Yao et al. obtained a series of middle-level features by learning the sub-blocks of characters: strokelets, and then obtained the characters by Hough voting. Then, the random forest classifier is trained by the strokelet and HOG features for character recognition. Wang et al. trained a CNN character recognizer to obtain a character position confidence map by sliding the input image, and based on the confidence map, the characters were segmented and then identified by Bissacco et al. The two segmentation methods perform character segmentation on words, and then combine word

search (Beam Search) and HOG feature-based character classifier to realize word recognition. This method is faster and more robust, but requires a lot of training data and can only recognize horizontal text lines.

Goel et al. used the entire text image to identify words. They used gradient-based feature maps to compare pre-made word images, and used dynamic k-nearest neighbors to determine the words contained in the current image [1], which relies on a fixed dictionary and pre-generated word images. In 2013, Rodriguez-Serrano et al. used an integrated fisher vector and a structured support vector machine framework to establish the relationship between pictures and the entire word encoding [2].

Multi-digit Number Recognition from Street View released by Google in 2013 [3]. This paper described a system for extracting numbers from streetscapes. The system used an end-to-end neural network, and the authors continued to explain how the same network could beat Google's own CAPTCHA SYSTEM with human-level accuracy, and Google's architecture have proven to work well for CAPTCHAS. In the original text, Goodfellow et al. first proposed using Maxout [4] as a nonlinear activation unit to construct a deep CNN to encode the entire image, and using multiple position-sensitive character-level classifiers [5] for text recognition. They have achieved great success in the identification of the street view number. They also applied the model to the 8-bit verification code recognition task and trained the model using synthetic training data. This method achieved a recognition rate of more than 96% in the Google Street View number recognition task. At the same time, it has obtained more than 99% recognition rate for the Google verification code recognition task, and then obtained the state-of-the-art result in the text classification [6]. The disadvantage of Deep CNN [7] is that the maximum length of the predictable sequence is selected in advance, which is more suitable for the house number or license plate number (a small number of characters, and each character can be regarded as independent).

The two results published by Jaderberg et al. in 2014 made minor changes to the above model which the classifier for predicting the length of the character was eliminated, and the terminator was used to indicate the end of the text [8]. They then demonstrated that models trained using synthetic training data can be successfully applied to real-world identification problems. Encoding words [9] into vectors is a viable dictionary word recognition method, but in the case of no constraints, characters can be arbitrarily combined. When the number of characters is sufficient, the performance of the method based on fixed length vector coding is significantly degraded. However, there are still some shortcomings: some studies use deep learning techniques in the single-character identification step, but the overall framework still follows the traditional processing flow design, so the problems described in the introduction will still be encountered in other steps. Goodfellow et al.'s [10] research uses a pure neural network to directly complete the entire identification process and achieve industry-leading results. However, since they need to use a fixed-size image as an input and encode the input image as a fixed-length feature vector, the recognition accuracy of the model is significantly reduced in the case of many characters in the image. On the other hand, since their models do not explicitly position [11] and segment [12] the image, it is impossible to know where each character is located in the original image.

2 Related Work

There is a lot of work to study how to define a good set of text features [13], but most of the features of practical applications are not universal. In extreme cases, many features are almost ineffective or even impossible to extract, such as stroke features, shape features, and so on. On the other hand, defining and extracting artificial features [14] is also an extremely time consuming and labor intensive task. The pictures that usually need to be identified are divided into two categories. The simple texts with a clear background and the complex texts with blurred backgrounds and complex texts are very stressful and challenging [15].

This paper attempts to combine deep learning neural network technology with text recognition technology, and proposes an effective text recognition method to find a feasible new way to solve the difficulties in text recognition in natural scene images. The focus and difficulty of image recognition in natural scene images is the feature extraction of characters and the design of classifiers. In the research of character extraction method of characters, the effective feature information of characters is extracted by studying the characters, and these features are combined with traditional character statistical features and structural features to identify the characters in the corpus system.

Due to the characteristics of neural network and its potential in the field of image text recognition, this paper conducts in-depth research on the natural scene image text detection and recognition technology based on neural network, and proposes a natural scene image text detection based on neural network. And the recognition algorithm [16, 17], using DenseNet network [18] model based on Kares [19] to construct text recognition [20], using Softmax output to classify each character [21], combined with the corpus to find the word that matches the character feature [22]. The classification of the character is obtained, thereby improving the recognition efficiency and precision, and realizing the text recognition of the natural scene image. Experiments show that this method can obtain better natural scene image text detection and text recognition effects, and achieve an efficient deep learning framework. The framework can support a variety of neural network structures and provide a series of effective training. Strategy, using this framework to preliminarily verify the effectiveness of natural scene text recognition algorithm based on deep learning.

3 System Introduction

In this study, the data set was preprocessed by CTPN [23], then the text recognition model in the image was established by DensNet, and finally the text was classified by Softmax classification.

3.1 Data Reprocessing

This research is mainly aimed at the problem of text recognition in natural scene images. Firstly, the standard data set of VOC is made by using data set. The data

needed by this research is obtained by CTPN method, as shown in Fig. 1. Then one is passed. A Chinese corpus containing 5990 characters is used to mark the obtained data set for subsequent model training.

Fig. 1. Training data

3.2 Basic Model

In this study, using a DensNet network model based on the Convolutional Neural Network (CNN) [24, 25], CNN is the basic framework of most network models. CNN is a standard feedforward multi-layer neural network inspired by biological processes. It uses a sparse connection and shared weighting strategy. It consists of a series of hidden convolutional layers and pool sampling layers, optionally following the fully connected layer, and is good at extracting useful. Local and global features are trained to classify. The standard CNN contains the layout of the convolutional layer and the pool sampling sublayer. Through this series of hidden layers, CNN uses the BP algorithm to train the weights. The main components of the network are shown in Fig. 2.

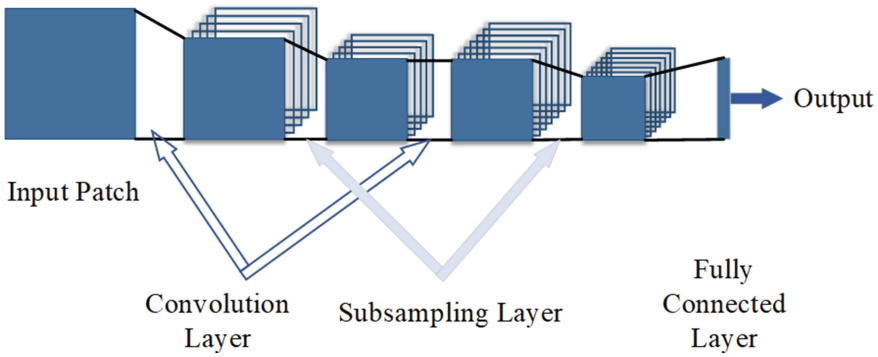

Fig. 2. CNN network main components

The main components of the hidden layer of the CNN network are described in detail below,

(1) Convolutional layer: Layer parameters consist of a set of learnable convolution kernels (filters). In the forward process, each convolution kernel moves incrementally over the width and height of the image matrix, and at each position, a dot product is computed between the number of convolution kernels and the local image matrix until convolution transformation (the mapping is completed), the completion of the original image. The BP algorithm is then applied to the gradient calculation.
(2) Activation function: This function increases the nonlinear nature of the decision function. Three common activation functions are commonly used, namely Sigmoid, hyperbolic tangent (tanh) and corrected linear unit (ReLU) functions.
(3) Pool sampling layer: This layer gradually reduces the amount of space represented by the image to reduce the number of parameters and computational load in the network. In practice, several nonlinear functions are often used to implement pools, such as maximum pools, mean pools, and random pools.
(4) Fully connected layer: Each node of the fully connected layer is connected to all nodes of the previous layer to combine the features extracted from the front. Due to its fully connected nature, the parameters of the fully connected layer are also the most common.

3.3 Text Recognition Model

This study uses the DenseNet network model to build our text recognition model based on Kares [26]. The biggest advantage of using DenseNet is to enhance the feature propagation and encourage feature reuse. The core idea is to create a cross-layer connection to connect the network. The middle and back layers are ideal for scene text recognition.

After DenseNet was proposed by Resnet [27], the variant network of ResNet emerged in an endless stream, each with its own characteristics, and the network performance also improved. DenseNet is mainly compared with ResNet and Inception networks. It has a new idea, but it is a new structure. The network structure is not complicated, but it is very effective. It completely surpasses ResNet in Cifar [28] indicators. It can be said that DenseNet absorbs the best of ResNet. Partly, and doing more innovative work here, the network performance is further improved. The connection is dense, the gradient disappears, the feature propagation is enhanced, feature reuse is encouraged, and the parameter quantity is greatly reduced.

DenseNet is a convolutional neural network with dense connections. In this network, there is a direct connection between any two layers. That is to say, the input of each layer of the network is the union of the output of all the previous layers, and the feature map learned by the layer is also directly transmitted to all layers behind it are used as input. Figure 3 is a map of DenseNet's dense block. The structure inside a block is as follows, which is basically the same as BottleNeck in ResNet: BN-ReLU-Conv (1 × 1)-BN-ReLU-Conv (3 × 3), and one DenseNet consists of multiple such blocks. The layer between each DenseBlock is called transition layers and consists of BN −> Conv(1 × 1) −> average Pooling(2 × 2).

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{1}$$

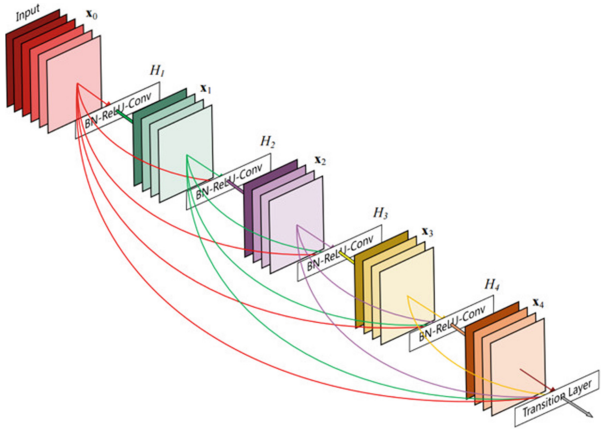

Fig. 3. DenseNet network model

The above formula is an expression of ResNet, where x_l represents the output of the l layer and H_l represents a nonlinear change, for ResNet, the output of layer $l-1$ is the output of layer $l-1$ plus a nonlinear transformation of the output of layer.

$$\mathrm{x}_l = H_l([\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_{l-1}]) \tag{2}$$

The above formula is the expression of DenseNet, where $[\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_{l-1}]$ means that the output feature map of the 0 to $l-1$ layer is concatenation, concatenation is the channel merge, and is the same as Inception [29], while the former ResNet is the sum of the values, the number of channels is unchanged, H_l includes convolution of BN, ReLU and 3 * 3.

3.4 Classifier

This experiment uses the Softmax classifier as the final output layer [7]. The Softmax function is based on Softmax regression, which is a supervised learning algorithm based on loss function of Softmax formula,

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} I\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] \tag{3}$$

Where $I\{\bullet\}$ represents an exponential function indicating whether $x^{(i)}$ is a class j. The function calculates the crossover loss based on the probability of the output of the Softmax function.

4 Experiment

In this experiment, we used the standard data used in CTPN, randomly divided all the data into 5 parts, and produced training sets, verification machines, and test sets in a ratio of 3:1:1. The training process iterates 50epoches, and the 1epoch test is performed after each 1epoch training iteration. Finally, the test set is used to test the accuracy of the training model.

4.1 Learning Rate Strategy

In this experiment, we set the initial learning rate to 0.01, and the learning rate decreased by 10 times for each epoch training. On the one hand, we know that if the learning rate is set too high, it will cause oscillation, that is, during the training process, the accuracy will be high and low, which will lead to the network not learning the characteristic information smoothly. On the other hand, if the learning rate is set too low, the network will not converge steadily. Therefore, we have adopted a strategy of decreasing learning rates.

4.2 Experimental Configuration

To complete this experiment, we used precision as the unit of measure. All experiments were implemented in Python 2.7 with the Tensorflow [30] and Keras frameworks. We trained the network M40 GPU on NVIDIA Tesla using the SGD [31] solver and model to perform the best test data set preservation for further analysis.

4.3 Experimental Procedure

4.3.1 Production Data Set

The VOC dataset format is a standard data format that is now used by many deep learning frameworks. We made all the images into a dataset in VOC format, which uses the folders Annotation, Image Sets, and JPEG Images. The folder Annotation mainly stores xml files, each xml corresponds to an image, and each xml stores the location and category information of each target of the mark, and the naming is usually the same as the corresponding original image; while Image Sets we only need Use the Main folder, which is stored in some text files, usually train.txt, test.txt, etc. The content of the text file is the name of the image that needs to be trained or tested (no suffix without path); In the JPEG Images folder, put the original image we have named according to the uniform rules. Input the prepared data set into the CTPN model for training, iterate 50epoches, input 30 pictures each time (batch size = 30), verify each time it is completed, and finally save the best model as the prediction model, and pass the prediction. The model gets the data set needed for text recognition. All the images obtained by CTPN are only intercepted into the demand area, and then combined with the Chinese corpus to make a training data set.

4.3.2 Training and Experimental Results

The prepared data set is input into the network for training, iterating 50epoches, inputting 30 pictures each time (batch size = 30), verifying each time it is completed, and finally saving the best model as a predictive model. The results of different model experiments are shown in Table 1. The experimental comparison results show that the LeNet result is 87.2, the NinNet test result is 82.4, the VGGNet test result is 93.3, and the DenseNet test result is 94.0. From the experimental results, the experimental results of VGGNet are good, but compared with the DenseNet experiment. The result is still a bit worse, and the experimental results will change due to environmental influences. The experimental results are shown in Fig. 4.

Table 1. The accuracy of predictive models of scene images

Model	Training sets	Testing sets	Accuracy/(%)
LeNet	254800	10000	87.2
NinNet	254800	10000	82.4
VGGNet	254800	10000	93.3
DenseNet	254800	10000	94.0


```
Detection took 5.327s for 3 object proposals
Mission complete, it took 6.225s

Recognition Result:

Humpsfor
60tards
```


```
Detection took 1.914s for 3 object proposals
Mission complete, it took 2.121s

Recognition Result:

M00SE
(
CROSSINC
```


```
Detection took 1.334s for 2 object proposals
Mission complete, it took 1.488s

Recognition Result:

EXAMROOM10
Thera0yH00m
```


```
Detection took 1.044s for 4 object proposals
Mission complete, it took 1.514s

Recognition Result:

!269
Dr.ChisGreen
ztftifery
CASCOBAY//CHIROPRACTIC
```


```
Detection took 0.192s for 2 object proposals
Mission complete, it took 0.268s

Recognition Result:

MARKEISIREET
```


```
Detection took 0.257s for 3 object proposals
Mission complete, it took 0.444s

Recognition Result:

AVALON
HOUSE
STYLEEXPRESSION
```

Fig. 4. Experimental results

5 Conclusions

In this study, the relatively simple background of the text area can be well recognized, the background is relatively complicated, especially in the background, there are many pictures that interfere with the text, and it is easy to mark the text area with heavy background, which leads to misidentification and do optimization work. For text recognition, according to the data set produced above, the accuracy of the experiment can reach about 94.0%. The experiment proves that the text recognition model based on DensNet can well realize the text recognition in natural scene images.

Acknowledgments. This study was supported by the Scientific Research Foundation (KYTZ201718) of CUIT. (KYTZ201718).

References

1. Goel, V., Mishra, A., Alahari, K., et al.: Whole is greater than sum of parts: recognizing scene text words. In: International Conference on Document Analysis and Recognition, pp. 398–402. IEEE Computer Society (2013)
2. Rodriguez-Serrano, J.A., Perronnin, F.: Label embedding for text recognition. In: BMVC (2013)
3. Goodfellow, I.J., Bulatov, Y., Ibarz, J., et al.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. Comput. Sci. (2013)
4. Goodfellow, I.J., Warde-Farley, D., Mirza, M., et al.: Maxout networks. arXiv preprint arXiv:1302.4389 (2013)
5. Yaeger, L.S., Lyon, R.F., Webb, B.J.: Effective training of a neural network character classifier for word recognition. In: Advances in Neural Information Processing Systems, pp. 807–816 (1997)
6. Jaderberg, M., Simonyan, K., Vedaldi, A., et al.: Synthetic data and artificial neural networks for natural scene text recognition. Eprint Arxiv (2014)
7. Zhang, K., Zuo, W., Chen, Y., et al.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
8. Jaderberg, M., Simonyan, K., Vedaldi, A., et al.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vis. **116**(1), 1–20 (2016)
9. Burden, V., Campbell, R.: The development of word-coding skills in the born deaf: an experimental study of deaf school-leavers. Br. J. Dev. Psychol. **12**(3), 331–349 (1994)
10. Goodfellow, I.J., Bulatov, Y., Ibarz, J., et al.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013)
11. Harrington, S.J., Klassen, R.V.: Subpixel character positioning with antialiasing with grey masking techniques: U.S. Patent 5,701,365 23 December 1997
12. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
13. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Text feature selection using ant colony optimization. Expert Syst. Appl. **36**(3), 6843–6853 (2009)
14. Tian, J., Chen, D.M.: Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition. Int. J. Remote Sens. **28**(20), 4625–4644 (2007)

15. Fang, W., Zhang, F., Sheng, V.S., Ding, Y.: A method for improving CNN-based image recognition using DCGAN. CMC Comput. Mater. Continua **57**(1), 167–178 (2018)
16. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015)
17. Coates, A., Carpenter, B., Case, C., et al.: Text detection and character recognition in scene images with unsupervised feature learning. In: International Conference on Document Analysis and Recognition, pp. 440–445. IEEE (2011)
18. Huang, G., Liu, Z., Van Der Maaten, L., et al.: Densely connected convolutional networks. In: CVPR, vol. 1, no. 2, p. 3 (2017)
19. Bien, Z., Chung, M.J., Chang, P.H., et al.: Integration of a rehabilitation robotic system (KARES II) with human-friendly man-machine interaction units. Auton. Rob. **16**(2), 165–191 (2004)
20. Zhu, Y., Newsam, S.: DenseNet for Dense Flow, pp. 790–794 (2017)
21. Salakhutdinov, R., Hinton, G.E.: Replicated softmax: an undirected topic model. In: International Conference on Neural Information Processing Systems, pp. 1607–1614. Curran Associates Inc. (2009)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
23. Tian, Z., Huang, W., He, T., He, P., Qiao, Yu.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
24. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
25. Zhang, Y., Wang, Y., Li, Y., Wu, X.: Sentiment classification based on piecewise pooling convolutional neural network. CMC Comput. Mater. Continua **56**(2), 285–297 (2018)
26. Krizhevsky, A., Hinton, G.: Convolutional deep belief networks on cifar-10. Unpublished Manuscript **40**(7) (2010)
27. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE (2015)
28. Huang, F., Ash, J., Langford, J., et al.: Learning Deep ResNet Blocks Sequentially using Boosting Theory (2018)
29. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
30. Abadi, M., Barham, P., Chen, J., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI, vol. 16, pp. 265–283 (2016)
31. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT 2010, pp. 177–186. Physica, Heidelberg (2010). https://doi.org/10.1007/978-3-7908-2604-3_16