RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2019 Volume 27, Number 45

CONVOLUTIONAL NETWORKS USED TO CLASSIFY VIDEO AND AUDIO DATA

Marcel NIKMON[1], Roman BUDJAČ[1], Daniel KUCHÁR[1],
Peter SCHREIBER[1], Dagmar JANÁČOVÁ[2]

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA,
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA,
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS,
ULICA JÁNA BOTTU 25, 917 24 TRNAVA, SLOVAK REPUBLIC
e-mail: marcel.nikmon@stuba.sk, roman.budjac@stuba.sk, daniel.kuchar@stuba.sk,
peter.schreiber@stuba.sk

[2]TOMAS BATA UNIVERSITY OF ZLIN, FACULTY OF APPLIED INFORMATICS,
DEPARTMENT OF AUTOMATION AND CONTROL ENGINEERING,
NAD STRÁNĚMI 4511, 760 05 ZLÍN, CZECH REPUBLIC
e-mail: janacova@utb.cz

Received 26 August 2019, accepted 7 October 2019, published 29 November 2019

Abstract

Deep learning is a kind of machine learning, and machine learning is a kind of artificial intelligence. Machine learning depicts groups of various technologies, and deep learning is one of them. The use of deep learning is an integral part of the current data classification practice in today's world. This paper introduces the possibilities of classification using convolutional networks. Experiments focused on audio and video data show different approaches to data classification. Most experiments use the well-known pre-trained AlexNet network with various pre-processing types of input data. However, there are also comparisons of other neural network architectures, and we also show the results of training on small and larger datasets. The paper comprises description of eight different kinds of experiments. Several training sessions were conducted in each experiment with different aspects that were monitored. The focus was put on the effect of batch size on the accuracy of deep learning, including many other parameters that affect deep learning [1].

Key words

Convolutional neural network, classification, AlexNet, audio-visual data, deep learning

INTRODUCTION

At present, the artificial intelligence is used for solving not only simple logical problems, but also complicated tasks, for example weather forecast, image and the text or voice commands

classification. The main aspects of the artificial intelligence comprise ability to learn, improve and derive the optimal solutions. In the previous years, the undergraduates in this Faculty worked with the neural networks in various classification tasks. This paper presents findings and results from eight of them. Different types of convolutional neural networks had been used for a classification of audio-visual data sets according to various specific requirements. Some of the theses used transfer learning for AlexNet, the others had been developed by their authors. The results demonstrate the application of the neural networks with small datasets.

METHODOLOGY OF EXPERIMENTS AND ATTAINED RESULTS

Deep learning in musical instruments sound recognition

The aim of this thesis was design of a Matlab application for classification of 10 musical instruments by means of the AlexNet deep neural network (transfer learning). The input for the application is an audio .mp3 file, although a .jpg picture – spectrogram serves as an input for the neural network. A Matlab script was used for conversion purposes from a musical track into a spectrogram with Fourier transformation. All the spectrograms showed dimensions of 227x227 pixels irrespective of the input track length.

Dataset

Ten musical instruments: banjo, cello, contrabassoon, double-bass, flute, guitar, mandolin, saxophone, trumpet and violin. Each of them was represented by 60 tracks of different lengths from 2 to 10 seconds. The tracks were split in the following manner: 40+10 for training and validation in Matlab; the last 10 for a special testing purpose.

Network training

All of the neural networks were prepared by transfer learning. At the beginning, a small number of musical instruments of a different acoustic resemblance were used. Each of the musical instruments had always the same number of tracks – for training and testing [2].

Table 1 Accuracy of instrument classification

Net	Number of instruments	Number of samples	Type of test	Failures	Accuracy [%]
Net_1	4	40	Matlab	3	92.5
			Manual	3	92.5
Net_2	5	50	Matlab	5	90
			Manual	5	90
Net_3	6	60	Matlab	9	85
			Manual	4	93.3
Net_4	10	100	Matlab	17	83
			Manual	8	92

Deep learning in music genre recognition

The aim of the thesis was classification of musical genres by means of the AlexNet deep neural network (transfer learning). The secondary aim was design of a training set in the shape of standardized (the same file format) and normalized (the same qualitative parameters) musical tracks. Once the sound processing was done, a conversion from a sound file into a picture–spectrogram followed. The spectrograms depicted the training set for AlexNet.

More quality of the processed files means higher ability of the neural nets for registering enough features of the musical genres. The .mp3 file format was chosen with respect to adequate quality and the minimal memory requirement: bit rate of 320kbit/s. The processing was done in order to achieve a united length of 60s. The tracks were transformed into the spectrograms of .jpeg format with 227x227 pixels dimensions.

Dataset

The emphasis was put on the strong difference between musical genres: classical music, rock, disco, Latin and rap. For each of the genres, 190 tracks-spectrograms were prepared, inclusive the augmentation.

Network training

At the beginning, three genres (classical music, Latin and rock) at the length of 30s and later 60s were used. Each of the genres included successively 50, 100 and 150 spectrograms with 80% portion for training and 20% for validation. In the case of longer tracks, the net was able to register more features of appropriate genres [3].

Table 2 Accuracy of genre classification for 3 classes

Number of musical genres	Samples within a genre	Accuracy [%] – 30s	Accuracy [%] – 60s
3	50	93.33	90.00
	100	93.33	96.67
	150	96.67	100.00

The fourth genre was rap with the length of 60s:

Table 3 Accuracy of genre classification for 4 classes

Number of musical genres	Samples within a genre	Accuracy [%] - 60s
4	100	90.00
	150	90.00
	190	95.00

The last, fifth, genre was disco:

Table 4 Accuracy of genre classification for 5 classes

Number of musical genres	Samples within a genre	Accuracy [%] - 60s
5	190	94.00

Deep learning in music bands recognition

The aim of the thesis was investigation of the possibilities for the use of AlexNet deep neural network fitted with feature extract and layer replacement methods for the band classification by means of the musical tracks. The same training set was used for both methods.

Dataset

Six bands of different kinds of music were chosen: Behemoth, David Guetta, Katy Perry, Linkin Park, Muse and Tina Guo. Fifty tracks of .mp3 format with a bit rate of 320 kb/s were chosen for each interpreter.

Network training

The data set was split in the following manner: 70% of the tracks for training and 30% for validation. In the first case, feature of extraction method was used. The network learned from fc6 layer. In the second case, the layer replacement method with transfer learning was adopted. For each of the bands, 5 songs (so far not used) were randomly chosen - altogether 30 songs of different length (15s, 30s, 45s, 60s, 90s, 120s and entire song) [4].

Table 5 Accuracy of music bands classification		
Sample	**Feature Extraction [%]**	**Layer Replacement [%]**
15s	43	33
30s	50	50
45s	53	47
60s	67	47
90s	53	60
120s	80	70
Entire song	93	93

The use of deep learning in recognizing the voices of a group of people

In the thesis, the author solved the problem of voice classification using deep neural networks. The resulting application was used to determine the author of the voice on the recording. In the neural network architecture design, convolutional network layers with the input vector of 343x434 were used. This input represented spectrogram images that were created from the sound recording after applying the Fourier transform.

Dataset

The training set was prepared by recording the audio recordings of eight people where the male voice and the female voice had the same representation in the number of samples. The diversity of the voice font of individual persons was ensured by the distribution of recordings from the age of 23 to 57. Twenty unique recordings with a recording sequence length of 6 seconds were recorded by each person. After recording the recordings in expected audio format, a spectrogram was made from each, and subsequently it served as an input to the neural network. One hundred thirty-six spectrograms were used in the training set, and 17 spectrograms were used in the test set.

Network training

During network training, six experiments were carried out with the convolutional neural network architecture. An iterative approach was chosen for experimentation. In the first experiment, a simple six-layer network was used. The success rate of the network classification was 41.67% after adjustment of the hyper parameters. Experiment 2 was used with two convolution layers. The number of network parameters was 2094848 and 180 epochs. This architecture achieved an accuracy of 62.5%. In all experiments, the softmax activation function at the neural network output was used. With this architectural modification, a net accuracy of 79.17% was achieved. Four convolution layers were used in the experiment 4, and 64 filters and stride = 2 were used since the third convolution. The achieved network accuracy was 91.67%. In the experiment 5, five layers were used. Also, the filter from the second layer on stride = 2 was modified. In this way, an accuracy of 100% on the training set was achieved. In the final experiment, 10 convolution layers with a filter size of 10x10 and in layers 5 and 6 at a size of 4x4 were used. The maximum number of filters was 32. This architecture performed best on the new voice recordings that were not part of the testing. Therefore, the author decided to implement the software. Significant were the last three experiments, when the network achieved the best results owing to the designed architecture [5].

Table 6 Classification accuracy for identity recognition from voice		
Experiment	**Number of convolution layers**	**Accuracy**
1.	1	41.67%
2.	2	62.50%
3.	3	75.00%
4.	4	87.5%
5.	5	100%
6	6	100%

Deep learning in voice commands recognition

The thesis solved the problem of controlling a virtual vehicle by ten voice commands, using deep neural networks. The length of each audio recording was 1.5 seconds. The chosen format was.wav, .mp3, .ogg / .opus. The input to the neural network was a set of spectrograms created from the recorded audio recordings. A convolutional neural network was used to classify individual commands.

Dataset

The training set contained 20 audio commands: forward, backward, right, left, brakes, lights, remote, left and right. The remaining 10 were numbers recorded from 0-10. The total number of voice recordings in the data set was 850. The data set was divided into the training, testing and validation sets. The test set represented 10% and the validation set represented 10% of the total number of audio recordings. All audio recordings were transformed into spectrograms. Subsequently, the images were adjusted by shifting in both axes [6].

Table 7 Classification error of multiple voice commands									
	Number of commands								
Order	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1.Train	0.95%	0.92%	1.11%	1.82%	1.60%	4.29%	2.17%	1.54%	0.66%
2.Train	2.91%	0.68%	5.45%	1.83%	3.089%	1.83%	1.25%	1.54%	0.96%
1.Train	3.12%	9.09%	10.14%	9.85%	16.04%	16.09%	8.88%	11.95%	10.10%
2.Train	4.61%	7.04%	5.63%	10.81%	13.33%	1.98%	16.48%	13.71%	12.24%

Deep learning in voice recognition for the operating system

The main aim of the thesis was to design a convolutional neural network enabling recognition og voice commands on Windows PC. Five commands were selected. AutoHotkey was used to run individual processes that evaluate the neural network. The author used the AlexNet pre-trained neural network.

Dataset

The data set was designed in a lossless .wav format with a 16000Hz sampling rate of 16 bits per sample and with a single channel recording. The sound signal was modified to remove silence from recordings and noise by applying the Savitz-Golay filter. Subsequently,

spectrograms were formed from individual recordings of 227x227 pixels. The training/validation set ratio was 70:30.

Network training

The Transfer Learning process was accomplished by locking layers in the original AlexNet neural network, except the last three layers. As we can see in the Table below, the experiment 1 for two commands achieved the accuracy rate of 90.74%. After removing this feature, the accuracy reached 96.3% [7].

Table 8 Classification accuracy of two voice commands

	Epochs	**Iteration**	**Data augmentation**	**Training time**	**Accuracy**
1.	6	72	Yes	8 min. 47s	90.74%
2.	10	120	Yes	13min. 7s	90.74%
3.	6	72	No	7min. 52s	96.30%

Table 9 Classification accuracy of multiple voice commands

Num. of commands	**Epochs**	**Iteration**	**Data augmentation**	**Training time**	**Accuracy**
2	6	84	No	11 min. 28s	98.33%
3	6	126	No	20 min. 56s	100%
4	6	168	No	35 min. 21s	100%
5	6	210	No	45 min. 13s	98%

After implementing the application, neural network accuracy was tested on 150 test samples. The commands 1, 3, 4 and 5 achieved classification reliability between 80-100%. Command 2 achieved a successful classification of only 46-73.3% when testing the application.

Deep learning in paint styles recognition

The aim of the thesis was to use deep learning in the recognition of individual art styles. The total number of recognized styles of art will be five. The recognized art styles were cubism, pop art, op-art, surrealism and impressionism. The thesis was focused on experimenting with classification of art styles by the AlexNet transfer learning with network and comparison with the basic neural network.

The experiment 1 found difference in the classification success between the conventional neural network and the pre-trained AlexNet convolutional network by changing the batch size. Our trained neural network distinguishes only two basic styles, pop-art and op-art. The research results suggest that as the batch size increased, the success of the training process decreased. Furthermore, the convolutional network architecture provides the classification accuracy of 85% for a batch size of 10. In another experiment, the effect of training set size on the success of the classification was examined. The effects can be seen in Figure 1 [8].

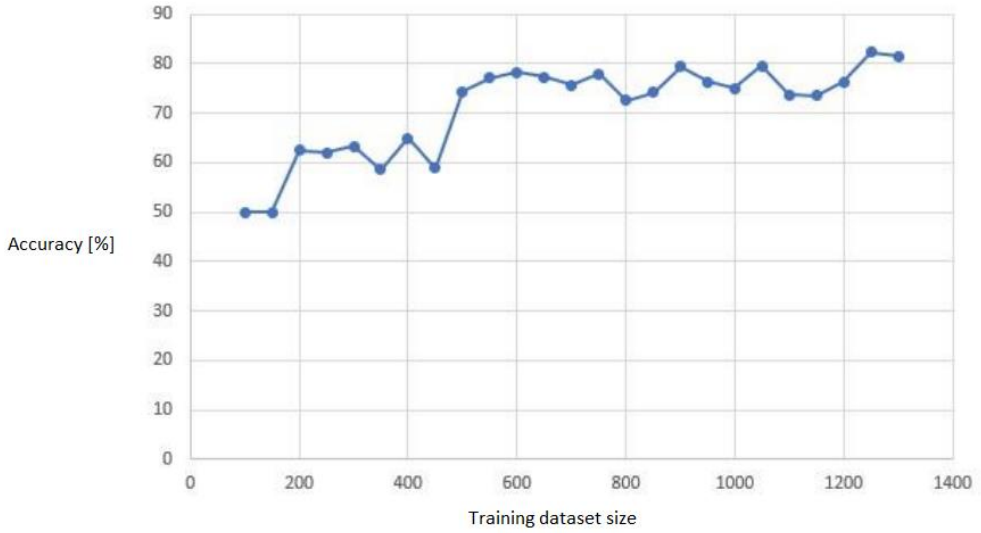

Figure 1 *Influence of the training dataset size on classification accuracy*

Subsequently, a third art style, Cubism, was added. Again, the experiment was performed with different batch sizes; the most successful was a batch size 10. Of course, this increased the data set at each step. The results show that the detection rate dropped by 3.3% at the addition of the classification class. More experiments added more art styles. For four styles, the success rate dropped by 9%. The training of all five styles and the size of the training set was 2000, accuracy dropped to 64.24%. It would be advisable to have a larger training set for multiple grades so that the success of the classification does not decrease so much.

Table 10 Classification accuracy of paint styles

Classification classes	Training dataset	Accuracy
2	100	91.91%
3	1050	88.61%
4	1600	79,51%
5	2000	64.24%

Video post-processing by deep learning

The thesis solved the issue of adding audio to video using the Deep Learning methods. The goal of the thesis was to develop an application that can recognize the video scene and add the sounds of the environment accordingly. GoogleNet and RESNET50 are compared to classify the video environment. The networks were trained using ADAM, rmsprop and sgdm algorithms in Matlab environment. More than 300 pictures of each category were used for training [9].

Table 11 Classification accuracy of video scene

Architecture	Training algorithm	Training duration (minutes)	Accuracy
GoogleNet	ADAM	55	93,09%
GoogleNet	RMSPROP	66	94.24%
GoogleNet	SGDM	67	95.39%
RESNET50	ADAM	248	94.74%
RESNET50	RMSPROP	227	90.46%
RESNET50	SGDM	215	98.85%

DISCUSSION

In this paper, various experiments with neural networks are presented. The difficulty of implementation and use varied from low to medium. All the neural networks used above had an image input, and, in some cases, image as a result of the Fourier transformation of the audio recording was used. The results suggest that lower number of classification classes increases the success of classification in convolutional networks. The results of the classification accuracy are difficult to compare because the theses used different types of input data distribution, mostly 80:20 or 70:30, but also the validation methods were different. Theses showed the classification of voice commands, music bands, genres and instruments, the identity of the person from the voice, but also the type of the scene from the video. Some functional networks could be used to get metadata from videos, while others can help control devices using voice. For real-world application, a larger training kit would be needed.

CONCLUSION

This paper describes the application of pre-trained AlexNet or other known nets. All of the tasks were basically simple. A human being is able to solve them, but, at the same time, they are complicated enough to be solved only by common classification techniques. In case of forthcoming research, some nets could be applied in the future only with larger training datasets. The use of neural networks does not end when classifying video or audio data. Research showed that traditional methods such as SVN, KNN, random forest, boosting tree and others give worse results in data prediction than neural networks. These are the reasons why neural networks will be an everyday part of our lives in the future [10].

Acknowledgement

This publication is the result of the VEGA Project 1/0272/18: "Holistic approach to knowledge discovery from production data in compliance with Industry 4.0 concept" supported by the VEGA, and KEGA project 0009STU-4/2018: "The innovation of the subject Intelligent Control Methods at the Faculty of Materials Science and Technology of Slovak University of Technology" supported by the KEGA.

This publication has been written thanks to support of the Operational Program Research and Innovation for the project Research, modeling and simulation of industrial production processes using progressive technologies, ITMS code: NFP313010T589 co-financed by the European Regional Development Fund.

References

[1] KIM, P. 2017. *Matlab deep learning: With machine learning, Neural networks and Artificial Intelligence*. Apress media, 2 p. ISBN-13: 978-1-4842-2844-9

[2] HORSKÝ, D. 2018. *Deep learning in musical instruments sound recognition*. Diploma thesis.

[3] KONCZ, I. 2019. *Deep learning in music genre recognition*. Diploma thesis.

[4] MAJDAFA, V. 2018. *Deep learning in music groups recognition*. Diploma thesis.

[5] DUGA, S. 2019. *The use of deep learning in recognizing the voices of a group of people*. Diploma thesis.

[6] BACIGÁL, L. 2019. *Deep learning in voice commands recognition*. Diploma thesis.

[7] TKÁČ, T. 2019. *Deep learning in voice recognition for the operating system*. Diploma thesis.

[8] BENEDIKTOVIČ, M. 2018. *Deep learning in paint styles recognition*. Diploma thesis.

[9] BENKA, D. 2019. *Video post-processing by deep learning*, 2019. Diploma thesis.

[10] VAŽAN, P., JANÍKOVÁ, D., TANUŠKA,P., KEBÍSEK, M., ČERVEŇANSKÁ, Z. 2017. Using data mining methods for manufacturing process control. In*: IFAC-PapersOnLine*, **50**(1), pp. 6178-6183.