MUSIC GENRE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS

Mateusz Matocha, Sławomir K. Zieliński

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: The aim of this study was to develop a music genre classifier using convolutional neural networks and to compare its performance with a traditional algorithm based on support vector machines. A distinct feature of the proposed approach was to utilize two-channel stereo signals at the input of the convolutional network. The proposed method yielded similar results compared to those obtained with the traditional approach, demonstrating the potential of the proposed method and indicating the need for its further optimization. Using two-channel stereo signals at the input of the algorithm showed no improvements over the baseline method exploiting single-channel recordings, suggesting that monaural signals fed to the convolutional network might be sufficient to undertake the task of music genre recognition. According to the results, the network 'prioritized' the temporal changes over the frequency variations of the signals. This observation tentatively implies that the classifiers specifically designed to account for temporal changes might potentially better serve the task of music genre recognition than the convolutional neural networks.

Keywords: automatic music genre recognition, convolutional neural networks, music information retrieval

1. Introduction

The growing popularity of music-on-demand services in the Internet gave rise to the situation where manual labelling of audio recordings according to their styles is no longer practical. There is also a need to automatically organize music content and to make intelligent recommendations to the listeners based on their preferences [10]. Therefore, automatic music genre recognition has recently become one of the most prominent research topics within a broader field of music information retrieval [27], [31], [9].

The remarkable success of the machine learning algorithms based on the convolutional neural networks (CNNs) in the field of image classification [17] suggests that

Advances in Computer Science Research, vol. 14, pp. 125-142, 2018. DOI: 10.24427/acsr-2018-vol14-0008

such algorithms might also exhibit competitive performance when applied to the task of music genre recognition. On the other hand, inherent properties of the convolutional neural networks, outlined in more detail in the next section, may prevent them from achieving as good results as those obtained using 'sound-oriented' algorithms, that is the algorithms which better account for temporal variations of sound, e.g. the recurrent neural networks [5], [12]. Recently, Muraurer and Specht [22] demonstrated that even a traditional algorithm employing hand-engineered features and XGBoost classifier performed better than the CNN. Hence, more research is needed in order to evaluate the applicability of the CNNs in the area of automatic music genre recognition.

The way in which stereophonic recordings are produced is genre specific. For example, in the case of classical music recordings the stereophonic panorama (distribution of audio objects in space) is determined by the position of the musicians and the microphone technique used by a sound engineer (e.g. XY, ORTF or Decca Tree). By contrast, for pop music recordings, stereophonic panorama is typically 'created' artificially, using amplitude panning algorithms in mixing consoles (see [26] for a comprehensive review of audio recording techniques). Hence, it was hypothesized that exploiting two-channel stereophonic signals, as opposed to monophonic ones, could enhance the performance of a music genre classification method.

The purpose of this study is twofold. First, we want to validate the suitability of convolutional neural networks to undertake the task of automatic music genre recognition by comparing its performance with a traditional algorithm based on the hand-crafted features and support vector machines (SVM). Second, we want to check whether there is any merit in using two-channel stereo sound at the input of the convolutional networks, compared to the standard approach exploiting single-channel monaural signals.

2. Related Work

The studies in the area of automatic music genre recognition were pioneered in 1995 by Matityaho and Furst [18] and then followed, among other researchers, by Tzanetakis and Cook [32], McKay and Fujinaga [20], Silla et al. [28], and Bhalke et al. [3]. An interested reader is referred to [27], [31] for a comprehensive literature review in this field. Most of the methods developed so far involved a two-stage approach. Music signals were subject to a procedure of hand-crafted feature extraction first, and then the extracted data were fed to the input of a classifier, such as a k-NN algorithm, random forest, logistic regression or support vector machines. A meaningful comparison of the results across the studies was impeded by the fact that the researchers

used a different number of genre categories, they applied diverse classification metrics, and employed various music corpora. Therefore, to increase the comparability of research, several datasets with manually labelled music recordings were developed and made publically available, most notably GTZAN [32], LMD [29], MSD [2], and FMA [8]. The number of songs in these databases ranged from 1 000 to 1000 000.

As mentioned above, the researchers in the area of music genre recognition use diverse classification metrics, which hinders a consistent comparison of the results across the studies. The average classification accuracy A and the $F1$ metric appear to constitute the most commonly exploited measures. They can be defined using the following two equations, respectively [30]:

$$A = \sum_{i=1}^{L} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i} / L, \tag{1}$$

$$F1 = 2\frac{pr}{p+r}. \tag{2}$$

The parameter L in Eq. (1) denotes the number of music genres (classes) used in a given study. The variables p and r in Eq. (2) represent classification precision and recall [30], respectively.

Convolutional neural networks are artificial neural networks intended for processing data that have a grid-like topology [13]. As already mentioned, they proved to be particularly effective in image classification [17]. Therefore, in order to apply them to the task of classification of audio, most of the researchers convert sound signals into images first and then feed them to the input of the CNN. This conversion is typically accomplished by calculating two-dimensional spectrograms [6], although other forms of visual representation of sound, such as tonnetz-plots and tempograms, could also be used [12].

One of the first attempts to apply convolutional neural networks to the task of automatic classification of music genres was undertaken by Gwardys and Grzywczak [14]. In contrast to a traditional procedure of hand-engineered extraction of signal features, they used the convolutional neural network to automatically generate the features of music recordings. These features were subsequently used as input data of the support vector machines. When applied to the GTZAN [32] dataset, their method yielded a classification accuracy approaching 78%. A year later Rajanna et al. [24] reported another attempt to employ a deep learning technique to the task of music genre classification. They compared a range of traditional procedures of feature extraction followed by two hidden layered feed-forward neural network, yielding rather poor classification results with an accuracy below 39%. Since that time, a markedly

increased interest of the research community in convolutional networks and deep learning could be observed. For example, Kim et al. [16] reported that CNNs might be employed to classify music genres with an $F1$ metric of 0.6571, which was demonstrated using the FMA [8] database. Similarly, promising results were also reported by Ghosal and Kolekar [12], Costa et al. [7] as well as by Bahuleyan [1].

Despite a growing body of research univocally supporting a high potential of the convolutional neural networks in the area of music genre recognition, there are some properties of the convolutional neural networks theoretically inhibiting their performance when applied to the classification of audio signals. CNNs are known to be '*approximately invariant to small translations*' [13]. This property is advantageous when the networks are applied to two-dimensional images which have the same interpretation of both dimensions. However, the dimensions of the spectrograms of audio signals have a different physical interpretation (time and frequency). Therefore the above property may be detrimental in terms of music genre classification, as highlighted by Medhat et al. [21]. Moreover, other classification algorithms, such as the recurrent convolutional networks, may be better at 'capturing' the temporal changes of the music signals compared to CNNs [12], [5].

To the best of the authors' knowledge, all of the algorithms used in the area of music genre recognition exploit single-channel signals at their input. While most of the publically available datasets of music excerpts contain the two-channel stereo recordings, they are typically down-mixed to mono (by averaging the stereo signals), before being fed to the input of the classification algorithms. Consequently, potentially important information is discarded. In their work regarding the acoustic scene classification (a different field of research compared to music genre recognition), Ham et al. [15] have recently demonstrated that exploiting two-channel stereo signals could enhance the performance of the audio classification algorithms. Hence, it can be hypothesized that including spatial information conveyed by the two-channel stereo signals could improve the performance of the music genre classification algorithms. This hypothesis was verified in this study.

3. Experiments

A 'small' version of the recently developed FMA [8] corpus of music recordings was used as a basis for the research described in this paper. It contained a set of 8 000 music excerpts of 30 seconds in duration, representing the following eight genres: electronic music, experimental, folk, hip-hop, instrumental, international, pop, and rock.

In the first and the second experiment 1999 music recordings were selected from the FMA database. Then, 70% of the recordings were used for **training** purposes whereas the remaining 30% of the excerpts were exploited for **testing**. The following metrics were used to evaluate the classification performance: accuracy, $F1$ metric, and the area under the curve (AUC). In the third experiment, 7994 recordings were selected from the FMA database. They were split in the proportion of 75% to 25% for the training and testing purposes, respectively. For the test employing SVM, the division between the training and testing datasets was slightly different. Namely, 70% of the recordings were used for training and 30% for testing. This difference was taken into consideration in the statistical test comparing the results between the methods. Due to long computational time, a hold-out validation technique was employed. In order to detect a potential problem with over-fitting, the classification learning curves were visually inspected. They were plotted as a function of training epochs both for the training and the testing datasets, respectively.

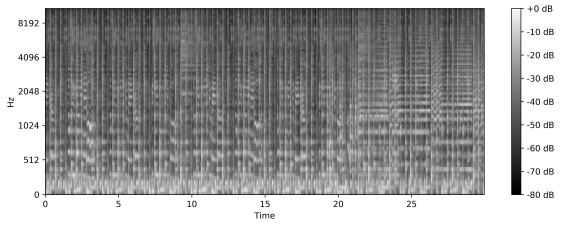
Due to the inconsistencies in sampling rate between the recordings, all the excerpts were down-sampled to 22.05 kHz. The pilot tests and the literature reports [32], [1], [12] confirmed that such sampling rate was adequate for the purpose of automatic genre recognition. The recordings were processed using 1024-samples long time frames with a 50% overlap. A Hanning window was applied to each frame.

Prior to exploiting the CNN, the musical signals were converted into the standard spectrograms (images) using a bank of 128 Mel-frequency filters. Example spectrograms obtained for the selected hip-hop and folk recordings were depicted in Fig. 1.

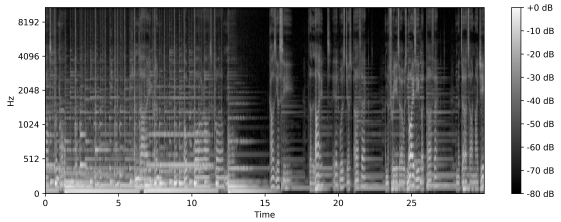
The proposed architecture of the CNN was implemented in Python programming language using the *keras* and *tensorflow-gpu* libraries. Moreover, the *sklrean* package was used to standardize the data, to split them into the training and testing datasets as well as to run the SVM classifier. The *librosa* package was used in order to generate the required spectrograms. The simulations were accelerated using the NVIDIA graphical processing unit GTX 1080Ti with 11 GB of memory.

3.1 Experiment 1 – Initial selection of the network topology

The purpose of the first experiment was to establish the number and the shape of the convolutional filters. The proposed CNN model consisted of three convolutional layers and two fully-connected layers. The three convolutional layers were interleaved by three average pooling layers, as shown in Fig. 2a. Due to the restrictions of available memory space in the graphical processing unit, the number of filters in the first convolutional layer was limited to 32. To reduce the risk of overfitting, a dropout technique (0.5 rate) was applied to the fully-connected layers. A stochastic gradient



(a)



(b)

Fig. 1: Example spectrograms of the selected recordings: (a) hip-hop and (b) folk.

descent (SGD) optimization algorithm was used. In order to reduce the data size at the output of the last average pooling layer, a technique of data reduction in time-domain was performed by calculated average values along the time-axis. A similar approach was proposed by Dieleman [10]. Table 1 shows a set of parameter values used during the optimization in Experiment 1. The obtained results were summarized in Table 2. Since accuracy A and $F1$ metric are commonly quoted classification performance measures in the field of music genre recognition, they were used together to evaluate the models developed in this study.

The best performance was seen with the network topology using 128 filters of a shape of 14×4 ($length \times height$). In this case, the $F1$ metric was equal to 0.345 which constitutes a mediocre outcome compared to the state-of-the-art algorithms [9]. Nevertheless, in line with the present results, the filters of a size of 14×4 were employed in the next experiment (see the next section).

The common feature of the best models was an irregular shape of the filters, with a shorter height y, and a longer length x. This outcome indicates that information conveyed by temporal changes of the signals (horizontal axis of the spectrograms) was more important than information represented by frequency changes (vertical axis of the spectrograms).

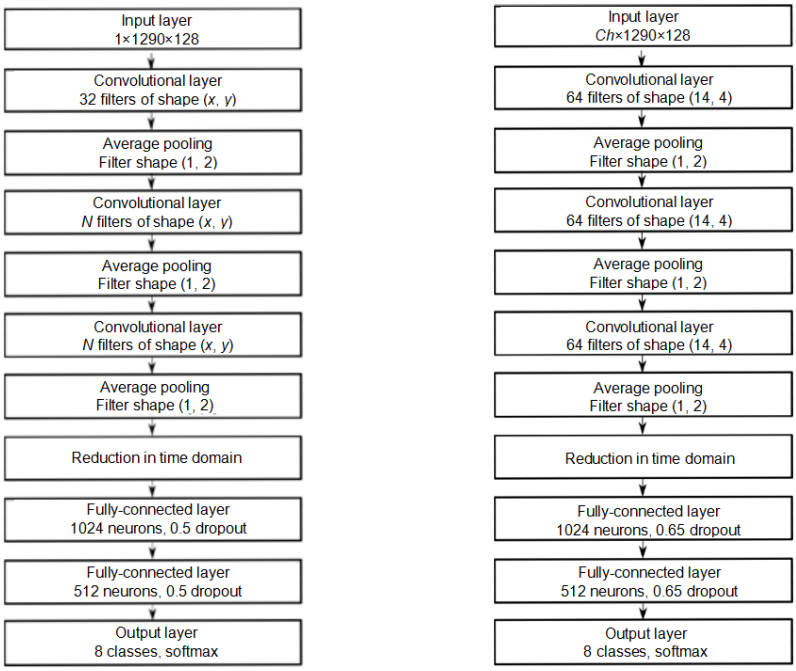


Fig. 2: A model of a neural network used in: (a) Experiment 1, (b) Experiment 2. A symbol *Ch* denotes the number of channels (*Ch* = 1 for monaural signals and *Ch* = 2 for stereophonic signals).

3.2 Experiment 2 – Utilizing stereophonic sound

The aim of the second experiment was to assess the merit of using two-channel stereophonic signals at the input of the CNN. To this end two networks were designed, one for the monaural signals and another one for the two-channel stereophonic signals. Their topology was similar to each other, the only difference being the number of the input layers *Ch* (see Fig. 2b). Monaural signals used in this experiment were obtained by averaging the two-channel stereophonic signals.

The number of the convolutional filters was set to 64, preserving the best shape identified in the first experiment (14×4). The reason for using 64 convolutional filters instead of 32 ones, as in the previous experiment, was due to a redesigned architecture of the network and graphical processing unit memory capacity limitations. In the first experiment, the usage of more than 32 filters required more memory than

Table 1: The parameters under optimization in Experiment 1.

Optimized parameters	A set of tested values
Number of filters (N)	32, 64, 96, 129
Filter height (y)	1, 2, 3, 4, 5
Filter length (x)	2, 4, 6, 8, 10, 12, 14, 16, 18, 20

Table 2: Overview of the best results obtained in Experiment 1.

Number of filters	Shape of filters (x, y)	$F1$ metric (Accuracy)
32	(20, 2)	0.255 (0.337)
32	(14, 3)	0.305 (0.353)
64	(20, 2)	0.313 (0.363)
128	(18, 3)	0.329 (0.378)
64	(14, 3)	0.332 (0.375)
128	**(14, 4)**	**0.345 (0.389)**

it was available during the tests of the largest filters (e.g. 20×4). In the second experiment, the network was redesigned according to the previously obtained results. The change of the filters' shape allowed to increase the number of filters in the first layer from 32 to 64. Due to the encountered problems with overfitting, the dropout rate was increased to 0.65, compared to the previous experiments. The following optimization algorithms were trialled in the pilot test (not reported in the paper): SGD, ADAM, and ADADELTA [25]. Since the ADAM algorithm produced the best results, it was employed in this experiment.

In this experiment the corpus was split into four datasets (quarters), each containing the equal number of tracks per genre. The training process was repeated four times, once per each quarter. Then, the average results were computed. Every model was trained for 50 epochs.

Contrary to the expectation, the performance level of the network exploiting the two-channel stereophonic recordings was slightly worse compared to that obtained using the monaural signals. The average $F1$ metrics achieved for the monaural and stereophonic algorithms were equal to 0.431 and 0.408, respectively.

3.3 Experiment 3 – Network optimization and comparison with the SVM classifier

The aim of the last experiment was to undertake the final optimization of the network and to compare its performance with the traditional method based on the support vector machine (SVM). Since the classical network layout, exploiting monaural signals at its input, outperformed the algorithm using the stereophonic signals, the standard topology employing a single-channel input was adopted in this experiment (see Fig. 3). The values of the parameters taken into account during the optimization procedure were presented in Table 3. Due to a large number of all possible combinations of the values (over 4.7×10^{10}) and limited computational resources, it was not practical to run an exhaustive search. Instead, a heuristic method was used in which a grid search algorithm was executed iteratively. In each iteration, the parameter values which gave rise to a significant deterioration in model accuracy were removed from the parameter grid.

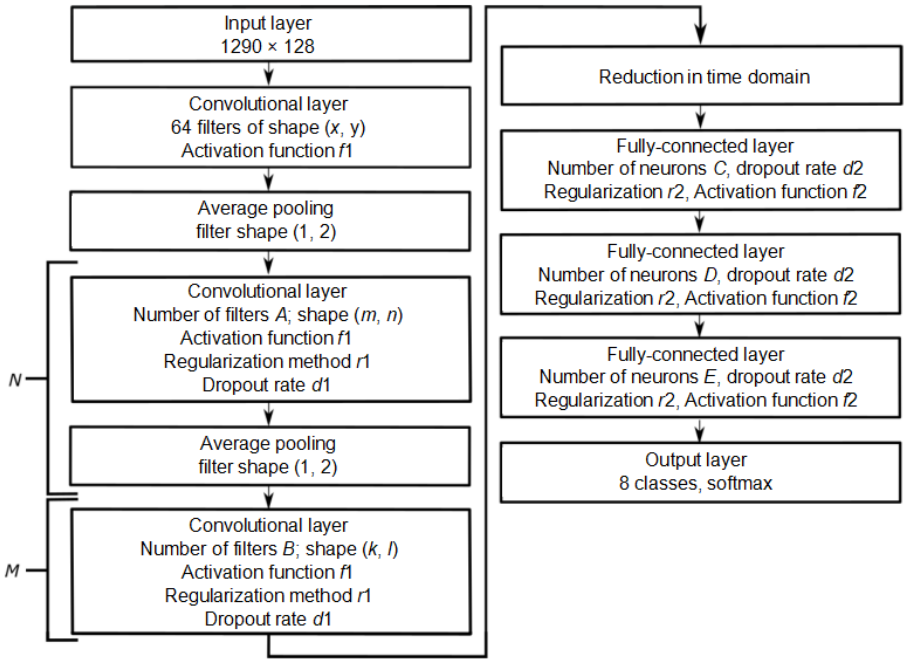


Fig. 3: A model of the neural network used in Experiment 3.

Table 3: The parameters taken into account during the final optimization. See Fig. 3 for explanation of the symbols.

Parameters	A set of tested values
Shape of filters (x, y)	(20, 2); (14, 3); (14, 4); (5, 1); (8, 1); (10, 1); (13, 1); (15, 1)
Activation function ($f1$)	Tanh; ReLU; Exponential Linear Unit (ELU)
Number of filters (A)	256; 128; 96; 64; 32
Shape of filters (m, n)	(20, 2); (14, 3); (14, 4); (3, 1); (5, 1); (8, 1); (10, 1)
Regularization type ($r1$)	L1; L2
Regularization value ($r1$)	1e-7; 1e-6; 1e-5; 1e-4
Dropout rate ($d1$)	0; 0.01; 0.05; 0.1; 0.2; 0.5
Number of filters (B)	256; 128; 96; 64; 32
Shape of a filter (k, l)	(14, 3); (14, 4); (3, 1); (5, 1); (8, 1); (10, 1)
Number of neurons (C)	2048; 1024; 512
Dropout rate ($d2$)	0.2; 0.3; 0.5; 0.6; 0.7
Regularization type ($r2$)	L1; L2
Regularization value ($r2$)	0.025; 0.03; 0.035; 0.04
Activation function ($f2$)	Tanh; ReLU; ELU
Number of neurons (D)	2048; 1024; 512
Number of neurons (E)	2048; 1024; 512
Number of layers (N)	4; 3; 2; 1
Number of layers (M)	4; 3; 2

The optimized version of the network yielded the best results compared to the previous experiments. The $F1$ metric and the accuracy of the best model reached the values of 0.6 and 0.605, respectively.

The hyper-parameters yielding the best performance of the network were gathered in Table 4. It was interesting to observe that the shape of the filter in the first convolutional layer of the best model was 'reduced' to a single dimension (10, 1). This outcome indicates that the initial layers of the network tended to ignore frequency information and to prioritize temporal information. Note, that a similar effect, although less pronounced, was also observed in the first experiment. In that case a shape of the best filter was 14×4 (see Table 2). Hence, it might be tentatively concluded that in order to obtain the best results in sound classification, irregular shapes of the convolutional layers, prioritizing time-axis of the spectrograms, might be beneficial. This supposition is supported by the recent study of Ghosal and Kolekar [12] who deliber-

Table 4: Overview of the best parameters in Experiment 3. See Fig. 3 for explanation of the symbols.

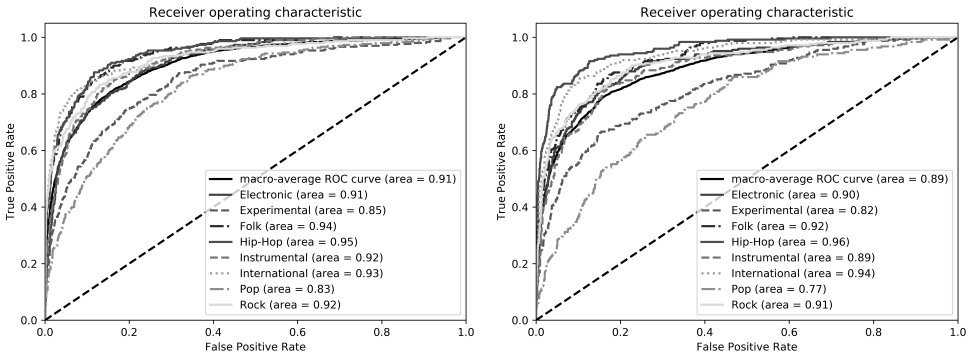
Optimized parameter	Value	Optimized parameter	Value
Shape of filters (x, y)	(10,1)	Number of neurons (C)	2048
Activation function ($f1$)	ELU	Dropout rate ($d2$)	0.5
Number of filters (A)	64	Regularization type ($r2$)	L2
Shape of filters (m, n)	(8, 1)	Regularization value ($r2$)	0.04
Regularization type ($r1$)	L1	Activation function ($f2$)	ReLU
Regularization value ($r1$)	1e-6	Number of neurons (D)	2048
Dropout rate ($d1$)	0	Number of neurons (E)	512
Number of filters (B)	64	Number of layers (N)	2
Shape of a filter (k, l)	(5, 1)	Number of layers (M)	3

ately restricted the filters to '1D convolution' along the time-axis in their music genre classification algorithm.

In order to compare the performance of the optimized CNN with a traditional method, it was decided to use a set of 518 hand-engineered features and to feed them to the input of the support vector machine (SVM). The rationale for choosing the SVM classifier was related to its well-known generalizability property. The above-mentioned features were calculated by the authors of the FMA music dataset [8]. They contained the standard metrics commonly used in music information retrieval applications, such as zero crossing, spectral centroid or spectral bandwidth. Several kernels were trailed in the classifier (not reported in the paper) indicating that the one employing the radial basis function (RBF) produced the best results. The hyper-parameters C and gamma of the support vector machine were equal to 10 and 0.3, respectively.

According to the obtained results, the classifier based on the hand-engineered features yielded an increase in accuracy by 0.1%. However, according to the binomial test of proportions, the above increment was statistically insignificant. Hence, it cannot be concluded that the traditional method outperformed the CNN.

The receiver operating curves (ROC) obtained for the traditional method and the one based on the CNN were illustrated in Fig. 4. It can be seen that the overall performance of the CNN and the traditional algorithm was similar. However, while the traditional algorithm was better at classification of pop, electronic and experimental music, the CNN-based algorithm was a 'winner' in terms of hip-hop and international music classification. Therefore, these two algorithms seem to complement each other

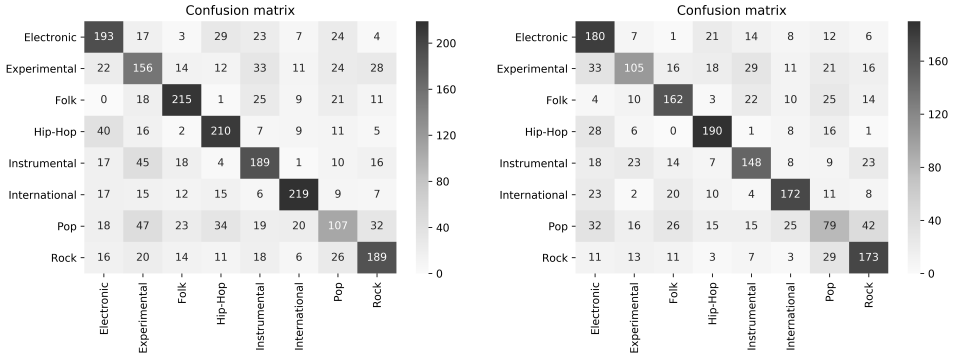


(a) (b)

Fig. 4: ROC curves: (a) SVM with RBF kernel; (b) CNN

and theoretically might be used in parallel in an ensemble of classifiers. Verification of this conclusions was beyond the scope of the study and was left for future work.

A confusion matrix obtained for the SVM-based method was presented in Fig. 5a. It can be seen that prediction accuracy varied between genres. The worst results were obtained for pop music which was often misclassified as rock or electronic music. A method incorporating CNN also exhibited difficulty with the classification of pop music recordings (see Fig. 5b).



(a) (b)

Fig. 5: Confusion matrices: (a) SVM with RBF kernel; (b) CNN. Predicted classes are listed at the bottom of the chart.

Table 5 illustrates how the obtained results compare to those obtained by other researchers. The data presented in the table were limited to the studies involving CNNs and has to be treated with some caution, as the direct comparison between the studies is hindered by the inconsistencies in terms of the music datasets and the evaluation metrics used by various authors. While the method proposed in this paper did not match the state-of-the-art algorithms described in the most recent literature, the obtained results indicate a high potential of the CNNs in the area of music genre recognition. If the results gathered in the table were further limited to those employing the FMA dataset, the CNN-based algorithm proposed in the present paper was only marginally worse compared to the one developed by Kim et al. [16].

Table 5: Comparison of the results in the literature (limited to the studies using CNNs for automated music genre recognition).

Authors	Source	Music dataset	Accuracy	$F1$	AUC
Ghosal and Kolekar (2018)	Tab. 2 in [12]	GTZAN [32]	0.942	—	—
Costa et al. (2017)	CNN in Tab. 3 in [7]	LMD [29]	0.83	0.836	—
Costa et al. (2017)	CNN in Tab. 8 in [7]	ISMIR [4]	0.859	0.863	—
Gwardys and Grzywczak (2014)	[14]	GTZAN [32]	0.78	—	—
Kim et al. (2018)	Tab. 4 in [16]	FMA [8]	—	0.6571	—
Bahuleyan (2018)	VGG-16 CNN F. Tun. in Tab. 2 in [1]	Audio Set [11]	0.64	0.61	0.889
Matocha and Zieliński	**Present study**	**FMA [8]**	**0.605**	**0.6**	**0.89**
Murauer and Specht (2018)	CNN in Tab. 5 in [22]	FMA [8]	—	0.48	—
Oramas et al. (2018)	Audio (A) in Tab. 6 in [23]	MuMu [19]	—	—	0.888
Oramas et al. (2018)	CNN_Audio in Tab. 2 in [23]	MSD [2]	—	0.336	—
Choi et al. (2017)	Fig. 3 in [5] blue dashed line	MSD [2]	—	—	0.83

4. Conclusions

The aim of this paper was to develop a music genre classifier using a convolutional neural network (CNN) and to compare its performance with a traditional algorithm based on hand-engineered audio metrics and support vector machines. A novel feature of the proposed approach was to utilize two-channel stereo signals at the input of the convolutional network. To the best of the authors' knowledge, no-one has attempted to employ stereophonic signals at the input of CNN to classify music genres yet. According to the results, using two-channel stereo signals showed no improvements over the baseline method exploiting single-channel recordings. On one hand, this outcome tentatively suggests that monaural signals fed to the convolutional network might be sufficient to undertake the task of music genre recognition. On the other hand, one may not exclude a possibility that the best topology of a network, handling two-channel stereo sounds, has not been identified yet. The latter conclusion is supported by the recent work of Ham et al. [15] in the area of acoustic scene classification. Similarly to this study, they observed that the algorithm exploiting the two-channel stereo sounds at the input of the CNN, used in isolation, produced worse results than the standard method. However, when the 'stereophonic' algorithm was employed as a part of an ensemble of classifiers, the overall performance of the method markedly improved. Therefore, further work is required before one could dismiss the usefulness of spatial information conveyed by two-channel stereo sounds in the area of deep learning and automatic classification of music genres.

Another contribution of this study was to quantify the accuracy yielded by the CNN when applied to the task of music genre recognition. Only a few papers devoted to this research domain have been published so far. To the best of the authors' knowledge, there are only two papers in which the researchers applied the CNN to the new FMA (2018) corpus [8]. While the results obtained in this study were 5% worse than those achieved by Kim et al. [16], they proved to be 12% better than the results published by Murauer and Specht [22]. The classification outcomes obtained in the area of machine learning do not always depend on the type of a classification method employed but also on "the match" between the database characteristics and the properties of a chosen classifier. Hence, the results presented in this paper could help other researchers to make an informed choice regarding a classification method for the task of the music genre recognition, particularly in relation to the new FMA dataset [8].

The proposed method yielded similar results compared to those obtained with the traditional approach, indicating that these methods could be used interchangeably or, which constitutes the subject of a future verification, in an ensemble of two algo-

rithms working in parallel. While the method proposed in this paper did not match the state-of-the-art algorithms described in the most recent literature, the obtained results demonstrated a high potential of the CNNs in the area of automatic music genre recognition.

The implemented CNN put higher importance to information carried by the temporal changes rather than to the frequency variations of the processed signals. This observation implies that the classifiers specifically designed to account for temporal changes, such as the recurrent neural networks, might better serve the task of music genre recognition than the convolutional neural networks. This conclusion is in accordance with the preliminary results obtained recently by Choi et al. [5] as well as by Ghosal and Kolekar [12].

Acknowledgments

This work was supported by a grant S/WI/3/2018 from Białystok University of Technology and funded from the resources for research by Ministry of Science and Higher Education.

References

[1] H. Bahuleyan: Music Genre Classification using Machine Learning Techniques, arXiv preprint arXiv:1804.01149, [https://arxiv.org/abs/1804.01149v1], Access time: November 5, 2018.

[2] T. Bertin-Mahieux, D. Ellis, B. Whitman and P. Lamere: The Milion Song Dataset, In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011.

[3] D.G. Bhalke, B. Rajesh and D.S. Bormane: Automatic Genre Classification Using Fractional Fourier Transform Based Mel Frequency Cepstral Coefficient and Timbral Features, Archives of Acoustics, vol. 42(2), pp. 213–222, 2017.

[4] P. Cano, E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S.Streich and N. Wack: ISMIR 2004 audio description contest, Technical report, Music Technology Group – Universitat Pompeu Fabra, 2006.

[5] K. Choi, G. Fazekas and K. Cho: Convolutional recurrent neural networks for music classification, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[6] Y.M.G. Costa, L.S. Oliveira, A.L. Koericb and F. Gouyon: Music genre recognition using spectrograms, In Proceedings of the 18th International Conference on Systems, Signals and Image Processing, 2011.

[7] Y.M.G. Costa, L.S. Oliveira and C.N. Silla Jr.: An evaluation of convolutional neural networks for music classification using spectrograms, Applied Soft Computing, vol. 52, pp. 28–38, 2017.

[8] M. Defferrard, K. Benzi, P. Vandergheynst and X. Bresson: FMA: A Dataset for Music Analysis, In Proceedings of the 18th International Society for Music Information Retrievel Converence (ISMIR), 2017.

[9] M. Defferrard, S.P. Mohanty, S.F. Carroll and M. Salathé: Learning to Recognize Musical Genre from Audio: Challenge Overview, In Companion of the Web Conference 2018. Lyon, France, April 23–27, 2018.

[10] S. Dieleman: Recommending music on Spotify with deep learning, Site [http://benanne.github.io/2014/08/05/spotify-cnns.html], Access time: November 6, 2018

[11] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.Ch. Moore, M. Plakal and M. Ritter: Audio set: An ontology and human-labeled dataset for audio events, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780, 2017.

[12] D. Ghosal and M.H. Kolekar: Music Genre Recognition using Deep Neural Networks and Transfer Learning, In Proceedings of Interspeech, September, 2018.

[13] I. Goodfellow, Y. Bengio, and A. Courville: Deep Learning, MIT Press, 2016.

[14] G. Gwardys and D. Grzywczak: Deep Image Features in Music Information Retrieval, Intl Journal of Electronics and Telecommunications, vol. 60(4), pp. 321–326, 2014.

[15] Y. Ham, J. Park and K. Lee: Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification, Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, 2017.

[16] J. Kim, M. Won, X. Serra and C.C.S. Liem: Transfer Learning of Artist Group Factors to Musical Genre Classification, In Companion of the the Web Conference 2018, Lyon, France, April 23–27, 2018.

[17] A. Krizhevsky, I. Sutskever and G.E. Hinton: ImageNet classification with deep convolutional neural networks, In Advances in neural information processing systems, vol. 25(2), pp. 1097–110, 2012.

[18] B. Matityaho and M. Furst: Neural network based model for classification of music type, In Proceedings of the Convention of Electrical and Electronics Engineers in Israel, pp. 1–5, March, 1995.

[19] J. McAuley, C. Targett, Q. Shi and A. Van Den Hengel: Image-based recommendations on styles and substitutes, In Proceedings of the 38th Interna-

tional ACM SIGIR Conference on Research and Development in Information Retrieval, 2015.

[20] C. McKay and I. Fujinaga: Music genre classification: is it worth pursuing and how can it be improved?, In Proceedings of the ISMIR, Victoria, Canada, October, 2006.

[21] F. Medhat, D. Chesmore and J. Robinson: Automatic Classification of Music Genre using Masked Conditional Neural Networks, In Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 979–984, 2017.

[22] B. Murauer and G. Specht: Detecting Music Genre Using Extreme Gradient Boosting, In Companion of the Web Conference 2018. Lyon, France, April 23–27, 2018.

[23] S. Oramas, F. Barbieri, O. Nieto and X. Serra: Multimodal Deep Learning for Music Genre Classification, Transactions of the International Society for Music Information Retrieval, 1(1), pp. 4–21, 2018.

[24] A.R. Rajanna, K. Aryafar, A. Shokoufandeh and R. Ptucha: Deep Neural Networks: A Case Study for Music Genre Classification, In Proceedings of the 14th International Conference on Machine Learning and Applications, 2015.

[25] S. Ruder: An overview of gradient descent optimization algorithms, Site: [https://arxiv.org/pdf/1609.04747.pdf], 2017. Access time: November 6, 2018.

[26] F. Rumsey and T. McCormick: Sound and Recording, Focal Press, 2014.

[27] N. Scaringella, G. Zoia and D. Mlynek: Automatic Genre Classification of Music Content. A survey, IEEE Signal Process. Mag., vol. 23(2), pp. 133141, 2006.

[28] C. Silla, C. Kaestner and A. Koerich: Automatic Music Genre Classification Using Ensemble of Classifiers, IEEE International Conference on Systems, Man, and Cybernetics, pp. 1687–1692, 2007.

[29] C. Silla, A. Koerich and C. Kaestner: The Latin Music Database, In Proc. of the 9th International Conference on Music Information Retrieval (ISMIR), 2008.

[30] M. Sokolova and G. Lapalme: A systematic analysis of performance measures for classification tasks, Information Processing and Management, vol. 45, pp. 427–437, 2009.

[31] B.L. Sturm: A Survey of Evaluation in Music Genre Recognition, In A. Nürnberger et.al. (eds) Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation. AMR 2012. Lecture Notes in Computer Science, vol. 8382. Springer, Cham. 2014

[32] G. Tzanetakis and P. Cook: Musical genre classification of audio signals, IEEE Trans. Speech Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

ROZPOZNAWANIE GATUNKÓW MUZYCZNYCH Z UŻYCIEM SPLOTOWYCH SIECI NEURONOWYCH

Streszczenie Celem niniejszej pracy było opracowanie klasyfikatora gatunków muzycznych z użyciem splotowych sieci neuronowych i porównanie go z tradycyjnym algorytmem opartym na maszynie wektorów wspierających. Wyróżniającą cechą zaproponowanego podejścia było wykorzystanie dwu-kanałowego dźwięku stereofonicznego na wejściu sieci splotowej. Zaproponowana metoda dała podobne wyniki do rezultatów otrzymanych z użyciem podejścia tradycyjnego, demonstrując potencjał zaproponowanej metody oraz wskazując na potrzebę jej dalszej optymalizacji. Wykorzystanie dwu-kanałowego dźwięku stereofonicznego na wejściu algorytmu nie poprawiło wyników w porównaniu z metodą bazową wykorzystującą nagrania jednokanałowe, sugerując, iż zastosowanie dźwięków monofonicznych na wejściu splotowej sieci neuronowej jest adekwatne do celów rozpoznawania gatunków muzycznych. Zgodnie z uzyskanymi wynikami, sieć 'potraktowała priorytetowo' zmiany czasowe w porównaniu ze zmianami częstotliwościowymi sygnałów. Obserwacja ta pozwala wstępnie przypuszczać, że klasyfikatory specjalnie zaprojektowane, by uwzględnić zmiany czasowe, potencjalnie mogłyby lepiej służyć celom rozpoznawania gatunków muzycznych niż neuronowe sieci splotowe.

Słowa kluczowe: automatyczne rozpoznawanie gatunków muzycznych, splotowe sieci neuronowe, pozyskiwanie informacji w muzyce

Badania zostały zrealizowane w ramach pracy S/WI/3/2018 sfinansowanej ze środków na naukę MNiSW.