jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2014, VOL. 60, NO. 4, PP. 321–326
Manuscript received September 1, 2014; revised December 2014. DOI: 10.2478/eletel-2014-0042

Deep Image Features in Music Information Retrieval

Grzegorz Gwardys and Daniel Grzywczak

***Abstract*—Applications of Convolutional Neural Networks (CNNs) to various problems have been the subject of a number of recent studies ranging from image classification and object detection to scene parsing, segmentation 3D volumetric images and action recognition in videos. CNNs are able to learn input data representation, instead of using fixed engineered features. In this study, the image model trained on CNN were applied to a Music Information Retrieval (MIR), in particular to musical genre recognition. The model was trained on ILSVRC-2012 (more than 1 million natural images) to perform image classification and was reused to perform genre classification using spectrograms images. Harmonic/percussive separation was applied, because it is characteristic for musical genre. At final stage, the evaluation of various strategies of merging Support Vector Machines (SVMs) was performed on well known in MIR community - GTZAN dataset. Even though, the model was trained on natural images, the results achieved in this study were close to the state-of-the-art.**

***Keywords*—music information retrieval, deep learning, genre classification, convolutional neural networks, transfer learning**

I. Introduction

THE enormous growth of unstructured data, including music data, encourages to searching for methods of effective indexing, classification or clustering. In music data case, these tasks are in interest of Music Information Retrieval. The first studies were performed in the 60s of 20th century [1], however intensive development can be notified at the beginning of the 21st century - in year 2000 International Society of Music Information Retrieval (ISMIR) was established and in the year 2005 the competition in MIR tasks called Music Information Retrieval Evaluation eXchange (MIREX) was launched. At now (MIREX 2014) algorithms are evaluated in 17 different categories such as melody extraction, cover song identification, or query by singing [2]–[5]. It is worth to mention that MIR is a highly interdisciplinary field that involves not only many parts of computer science and signal processing, but also non-technical disciplines such as music theory, musicology or psychology. As mentioned before, giant amounts of unstructured data also apply to music databases. Spotify[1] posses about 24 million users, every fourth pays $10 monthly for unlimited, free of advertisements, access to music database containing about 20 million songs. Spotify adds approximately 20 000 songs per day[2]. This business's need of dealing with a huge amount of audio files and academic interest in Signal Processing, Machine Learning and lastly Feature Learning resulted with many interesting systems.

Authors are with the Institute of Radioelectronics, Faculty of Electronics and Information Technology, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland (e-mail: g.gwardys@ire.pw.edu.pl, d.grzywczak@ire.pw.edu.pl).

[1]http://spotify.com

[2]http://press.spotify.com/us/information/

II. Music Representation

A. Engineered Features

Engineered features or low-level features are simple representation of complex, unstructured data (which include images, videos or audio). They describe features of signal in time or frequency, rather than its semantic content. These kind of features are quite universal, because the same feature extraction algorithm can be used to many types of data. There are plenty of different kinds of engineered features for music representation, the most popular are: spectrogram, zero crossing rate, spectral centroid, fundamental frequency [6], [7]. There are also more sophisticated features for music representation such as Chromagrams or Mel Frequency Cepstrum Coeffcients (MFCC).

Because MFCC is perceptually motivated (Mel scale is a perceptual scale of pitches), they are better adopted to represent audio signal and they are commonly used in speech recognition [8]. But still this is a ready 'recipe' for feature extraction, not a tailor-made one for a given data distribution or task.

Chromagram (or Harmonic Pitch Class Profile) describes intensiveness of each of 12 semitones in octave, basing on frequency spectrum, so it is also perceptually motivated method for music description. This feature can be used for chord recognition [9], but also in music similarity [10] or cover identification [11].

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are specific types of Artificial Neural Networks. They were introduced by Kunihiko Fukushima in 1980 in [12], [13] and later improved, especially by Y. LeCun [12], [14]. They are very successful, because of:

- taking into account correlations of neighbouring data
- weight sharing that allows to perform effective training
- large amounts of data and GPU implementations
- bunch of tricks such as dropout, ReLU, local contrast normalization, max pooling

One of most famous CNN architecture was LeNet5 [14] which was used for handwritten character recognition on MNIST database with great result 0.7% error rate.

Recently, CNN are used in many tasks, such as segmentation 3D volumetric images [15], scene parsing [16], action

recognition from videos [17] or object detection [18]. The most spectacular results CNN achieved in image classification competition.

Large Scale Visual Recognition Challenge (ILSVRC) is an annual competition in image recognition [19]. In year 2012 one of tasks was to choose 5 most probably classes of image form 1000. Whole image dataset which was available in this competition contained more than 1 million images. Winners in this category were A. Krizhevsky, I. Sutskever and G. E. Hinton as SuperVision Team. They used large convolutional neural network for image classification, achieving spectacular success with 15,3% error rate compared to 26,2% error rate being the second best result [20]. This result began a revolution in image classification and again drew attention to the neural network methods. In year 2012 SuperVision was the only team using CNN for image classification, in year 2013 many other implementations of neural networks appeared and in ILSVRC 2014 almost all teams used CNN for image classification.

Winning CNN, which is presented in figure 1, has eight layers. First five layers are convolutional ones and the remaining ones are fully connected. Whole network has 650 000 neurons and 60 million parameters. To train such a huge network, several tricks have been proposed [21]. First improvement is non-saturation activation function $f(x) = max(0, x)$ which is called Rectified Linear Unit (ReLU). This significantly decreases number of iteration during learning process. The second trick is dropout technique consisting of randomly removing out output of neurons. This method makes training longer, but prevents overfitting, and makes network to learn robust features. The next idea is local response normalization. In this approach, activity of neuron (after applying ReLU function) is normalized by adjacent activation neuron activities, which reduces error rate. Another trick is artificially extending dataset by extracting random patches and making reflection. Last trick is overlapping maxpooling layers between convolutional layers, which slightly reduces error rate and overfitting. Input signal for this network is raw RGB image with size 256 x 256 pixels with subtracted mean of the dataset. Whole network was implemented on two GPUs (GTX 580 3GB) and training of whole network took about a week [21].

SuperVision not only created great classification algorithm, but also some hierarchical representation of images. Each layer of network represent some features of images, from low-level signal features for low layers to high-level semantic features for high layers. These set of features create some universal representation of images, which can be used in many tasks.

III. Transfer Learning

In classical machine learning techniques training data and test data are from the same dataset and have similar distribution. As an example, we can check in MNIST database [22], [14], one can not recognize if given image belongs to test or training set. But in most real-world applications it is difficult or even impossible to collect enough training data, because of high cost of data collecting, small number of potential training samples, short period of validity of data or long time of model training. To solve this problem transfer learning is used.

Basing on S. Jialin Pan and Q. Y. Survey [23], definition of transfer learning helps to improve the learning of the target predictive function f_T for target task T_T in target domain D_T using the knowledge in source domain D_S and source task T_S, where $D_S \neq D_T$ or $T_S \neq T_T$. According to [23], transfer learning can be divided into three categories:

- inductive transfer learning is used when target task T_T is different from source task T_S, labeled data in target domain is available regardless of whether the labeled source data is available or not;
- transductive transfer learning is used when target task T_T is equal to source task T_S, labeled data in target domain is not available and source domain D_S and target domain D_T are different;
- unsupervised transfer learning is used when task T_T is different from source task T_S, there is no labeled data in target and source domain.

There are several different approaches in transfer learning:

- instance-transfer - reusing source data during learning target predictive function f_T, it can be used in inductive and transductive transfer learning;
- feature-representation-transfer - creating good feature representation in target domain, it can be used in every type of transfer learning;
- parameter-transfer - reusing parameters or distribution of model for source task, it can be used in inductive transfer learning;
- relational-knowledge-transfer - using relations between source and target data, it can be used in inductive transfer learning.

Transfer learning can be used in many real applications such as text classification [24], spam filtering [25], image classification [26] or even face verification [27].

IV. Our Approach

In this paper, inductive transfer learning using feature representation transfer was performed. Because of lack of infrastructure (GPU with at least 6GB memory) we decided to reuse already trained ILSVRC CNN model computed using CAFFE framework [28]. We generated image of frequency spectrogram for each music track. To achieve better results we decided to use harmonic/percussive separation, which is described in subsection IV-A, and generate frequency spectrograms for each component. To fit ILSVRC model to new data we decided to use two different methods: fine-tuning and Support Vector Machines (SVM) features classification which are described respectively in subsection IV-B and IV-C.

A. Harmonics and Percussion Separation

Harmonics and Percussion separation was performed in proposed system, due to the significant improvement in classification accuracy. Ono et al [29] made an intuitive observation that percussive components form vertical ridges with a broadband frequency response, while stationary components or stable harmonic make horizontal ridges on the spectrogram.

The harmonic events can be treated as outliers in the frequency spectrum at a given time frame. Similarly, the

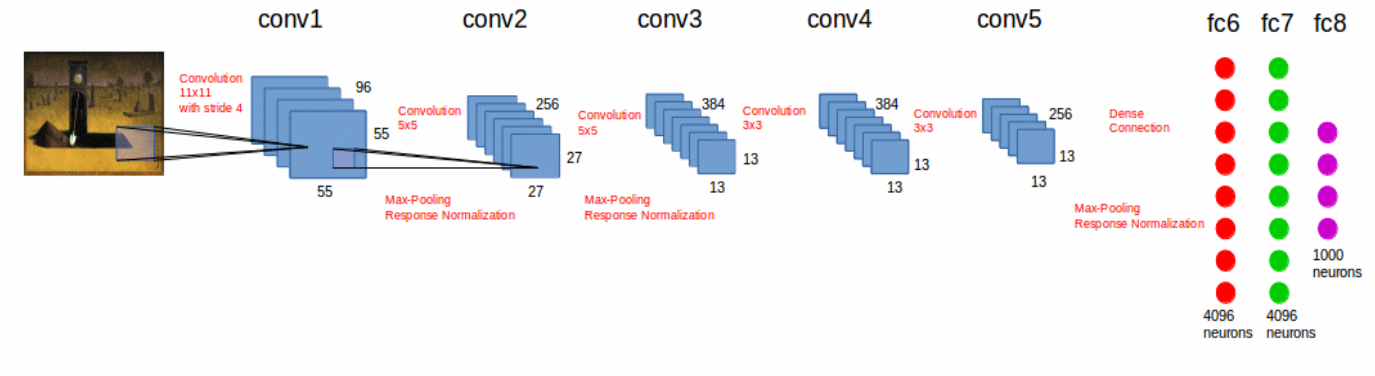

Fig. 1. ILSVRC winning CNN topology

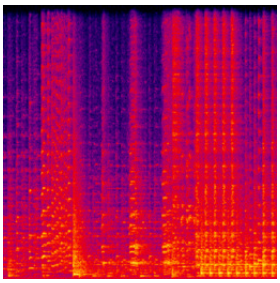

Fig. 2. Original Spectrogram

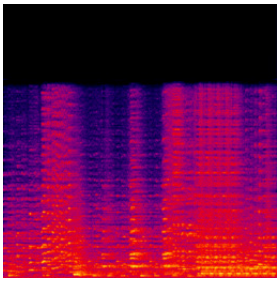

Fig. 3. Harmonic Spectrogram with visible horizontal ridges

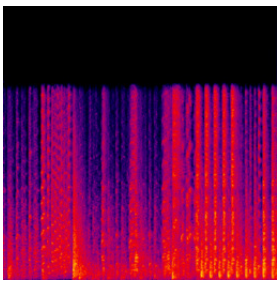

Fig. 4. Percussive Spectrogram with visible vertical ridges

percussive events can be treated as outliers across time in a given frequency bin. This leads to a concept of using median filters separately in the horizontal and vertical directions for harmonic/percussive separation.Examples of spectrogram for are presented in figures 2, 4, 3

B. *Fine-tuning*

Fine-tuning in neural networks is a process where already trained model parameters are adjusted to data. Fine-tuning can be performed on the whole network or on some subset of parameters (e.g. one networks layer) without changing. This approach can be used to adjust network parameters to new, unseen data with different distribution than training set or when network architecture have to be partially changed (e.g. number of output labels have to be increased). Using this idea, new model, adopted to new data or task can be trained with low computational effort, because model weights are well initiated and only some parts of model have to be updated. In this paper fine-tuning of the last layer of the ILSVRC network model was performed.

C. *SVM Features Classification*

This approach features extracted from spectrogram images from penultimate layer of ILSVRC network model were used for SVM training. The proposed system, presented in figure 5, can be summarized to next steps:

1) Audio file normalization - simple data normalization
2) Harmonics and Percussive Separation - to improve classification rate, we increase number of data
3) Performing Short Time Fourier Transform (STFT) for all three versions (original, harmonic and percussive) - to achieved tree spectrogram images
4) Spectrograms are forwarded through CNN and features extraction from penultimate layer of network - 4096 dimensional vector for each image
5) Training separate SVMs for all three versions
6) Merging results, final classification

V. Experiments and Results

The evaluation was performed on 1000 music tracks from GTZAN dataset, that includes 10 musical genres (100 music tracks for each genre):

- Metal
- Jazz
- Pop
- Reggae
- Blues

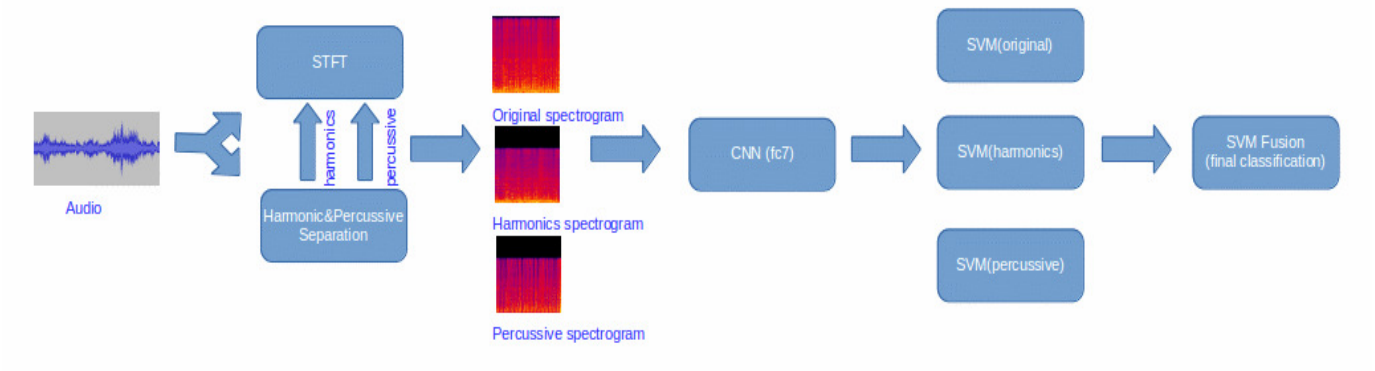

Fig. 5. Authors architecture of system for genre classification

- Classical
- HipHop
- Rock
- Country
- Disco

GTZAN dataset was divided into training dataset (900 music tracks) and testing dataset (100 music tracks). Both training and testing datasets are balanced - that means there are equal numbers of music tracks from each musical genre. Fine-tuning was performed for each type of spectrogram (original, harmonic, percussive). In each configuration, fine-tuning was stopped after 3000 iterations and model with the best mode with best result on training data set was chosen. Results and confusion matrix from fine-tuning test set evaluation are presented respectively on table I and II. Results from each classifier were combined according to the formula $result = 0.5l_p^{ft} + 0.5l_{all}^{ft}$ where l_p^{ft} are likelihoods from percussive fine-tuned network and l_{all}^{ft} are likelihoods from whole song fine-tuned network. Best result for fine-tuning was 72%, which was achieved for combining percussive and original likelihood where harmonic likelihood weight is equal to 0.

SVM method was applied in similar way. Separate SVM classificator was trained for each type of spectrogram. Best results were achieved for Radial Basis Function kernel. Two methods of merging results from three separate SVMs were evaluated. The adding class likelihoods turned out to be slightly better (by 2%) than multiplying. Results from each classifier were combined according to the formula $result = 0.1l_h^{svm} + 0.4l_p^{svm} + 0.5l_{all}^{svm}$ where l_h^{svm} are likelihoods from harmonics SVM classifier, l_p^{svm} are likelihoods from percussive SVM classifier and l_{all}^{svm} are likelihoods from whole song SVM classifier.Finally, the classification rate for merging all three SVMs reached 78% - 10% improvement comparing to version trained only on original spectrograms.

VI. Conclusion

This study presents a successful application of CNNs to a MIR task such as Genre Recognition. The presented system can be classified as inductive transfer learning, because model trained on more than 1 million natural images (ILSVRC-2012) was used.

In the first stage of research fine-tuning approach was used to fit last layer of CNN to new data with 68 % accuracy. To improve results harmonic/percussive separation was performed. Results from each fine-tuned model were merged which gave 4 % improvement in accuracy.

In the second stage of research SVM features classification was used, instead of fine-tuning. Separate classificator for each type of spectrogram was trained and results from each classificator were merged. This approach achieved 78 % accuracy which is already close to the state-of-the-art results, despite the fact there is used model trained on natural images was used, not on specific spectrograms.

In further works, authors would like to explore CNNs in other Music Information Retrieval tasks, both by finetuning ILSVRC-2012 model and making new CNN model

References

[1] M. Kassler, "Toward Musical Information," *Perspectives of New Music*, vol. 4, no. 2, pp. 59–66, 1966. [Online]. Available: http://www.jstor.org/discover/10.2307/832213?uid=3738840\&uid=2134\&uid=2\&uid=70\&uid=4\&sid=21103750075213

[2] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," *9th International Symposium on Computer Music Modeling and Retrieval*, 2012. [Online]. Available: http://www.mendeley.com/research/survey-music-recommendation-systems-future-perspectives-1/

[3] J. Futrelle and J. S. Downie, "Interdisciplinary Communities and Research Issues in Music Information Retrieval," *Library and Information Science*, pp. 215–221, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.9456\&rep=rep1\&type=pdf

[4] J. S. Downie, K. West, A. F. Ehmann, and E. Vincent, "The 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview," *International Conference on Music Information Retrieval*, no. Mirex, pp. 320–323, 2005.

[5] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The music information retrieval evaluation exchange: Some observations and insights." in *Advances in Music Information Retrieval*, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wieczorkowska, Eds. Springer, 2010, vol. 274, pp. 93–115. [Online]. Available: http://dblp.uni-trier.de/db/series/sci/sci274.html\#DownieEBJ10

[6] P. Rao, "Audio signal processing," in *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, ser. Studies in Computational Intelligence, B. Prasad and S. Prasanna, Eds. Springer Berlin Heidelberg, 2008, vol. 83, pp. 169–189.

[7] D. Grzywczak and G. Gwardys, "Audio features in music information retrieval," in *Active Media Technology*, ser. Lecture Notes in Computer Science, D. lzak, G. Schaefer, S. Vuong, and Y.-S. Kim, Eds. Springer International Publishing, 2014, vol. 8610, pp. 187–199.

TABLE I
CLASSIFICATION RATES FOR FINE-TUNING

Type	Classification Rate (%)
Original Spectrograms	68
Harmonic Spectrograms	55
Percussive Spectrograms	65
Harmonic + Percussive Spectrograms	65
Original + Harmonic + Percussive Spectrograms	72

TABLE II
CONFUSION MATRIX FOR FINE-TUNING, WHERE THE INPUT CONSISTS OF ORIGINAL SPECTROGRAMS/HARMONIC SPECTROGRAMS/PERCUSSIVE SPECTROGRAMS/HARMONIC AND PERCUSSIVE SPECTROGRAMS/ORIGINAL AND HARMONIC AND PERCUSSIVE SPECTROGRAMS

Genre	Metal	Jazz	Pop	Reggae	Blues	Classical	HipHop	Rock	Country	Disco
Metal	**9/9/9/9/9**	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	1/1/0/0/0	0/0/0/0/0	0/0/1/1/0	0/0/0/0/1
Jazz	0/0/0/0/0	7/5/**8/8/8**	0/0/0/0/0	1/1/0/0/0	1/1/0/0/1	1/2/1/1/1	0/0/0/0/0	0/0/0/0/0	0/1/1/1/0	0/0/0/0/0
Pop	0/1/1/1/0	0/0/0/0/0	5/3/7/7/**8**	0/1/0/0/0	0/0/0/0/0	0/0/0/0/0	4/3/2/2/1	0/1/0/0/0	1/0/0/0/1	0/1/0/0/0
Reggae	0/0/0/0/0	0/0/0/0/0	0/0/1/1/0	**7**/5/6/6/**7**	0/1/1/1/0	0/0/0/0/0	2/2/0/0/1	0/0/0/0/0	0/0/0/0/0	1/2/2/2/2
Blues	1/1/2/2/2	0/1/1/1/1	0/0/0/0/0	1/0/1/1/0	**5**/4/4/4/**5**	1/0/0/0/1	0/1/0/0/0	0/2/2/2/0	0/1/0/0/0	2/0/0/0/1
Classical	0/0/0/0/0	0/0/1/1/1	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	**10**/8/8/8/9	0/0/0/0/0	0/1/0/0/0	0/1/1/1/0	0/0/0/0/0
HipHop	1/1/1/1/1	0/0/0/0/0	0/1/3/3/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	**9**/8/6/6/**9**	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0
Rock	0/0/1/1/1	0/0/0/0/0	0/0/0/0/0	1/1/1/1/1	0/0/0/0/0	0/0/0/0/0	1/3/1/1/1	**6**/5/5/5/**6**	0/0/0/0/0	2/1/2/2/1
Country	0/0/0/0/0	0/1/1/1/0	0/0/0/0/0	3/4/2/2/2	1/0/0/0/1	0/0/0/0/0	2/1/1/1/2	0/0/0/0/0	4/4/**5/5/5**	0/0/1/1/0
Disco	1/1/0/0/0	0/0/0/0/0	0/0/0/0/0	2/2/1/1/1	0/0/0/0/0	0/0/0/0/0	0/1/1/1/2	1/1/1/1/1	0/1/0/0/0	6/4/**7/7**/6

TABLE III
CLASSIFICATION RATES FOR SVM FEATURES CLASSIFICATION

Type	Classification Rate (%)
Original Spectrograms	68
Harmonic Spectrograms	59
Percussive Spectrograms	64
Harmonic + Percussive Spectrograms	69
Original + Harmonic + Percussive Spectrograms	78

TABLE IV
CONFUSION MATRIX FOR SVM FEATURES CLASSIFICATION, WHERE THE INPUT CONSISTS OF ORIGINAL SPECTROGRAMS/HARMONIC SPECTROGRAMS/PERCUSSIVE SPECTROGRAMS/HARMONIC AND PERCUSSIVE SPECTROGRAMS/ORIGINAL AND HARMONIC AND PERCUSSIVE SPECTROGRAMS

Genre	Metal	Jazz	Pop	Reggae	Blues	Classical	HipHop	Rock	Country	Disco
Metal	9/**10**/9/9/9	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	0/0/1/1/0	1/0/0/0/1
Jazz	0/0/0/0/0	**6/6/6/6/6**	0/0/0/0/0	0/0/0/0/0	2/1/0/1/2	1/1/1/1/1	0/0/0/0/0	1/1/1/1/1	0/1/2/1/0	0/0/0/0/0
Pop	0/0/0/0/0	0/0/0/0/0	**8**/7/7/7/**8**	0/1/0/1/0	0/0/0/0/0	0/0/0/0/0	1/1/1/1/1	0/1/1/1/0	1/0/0/0/1	0/0/1/0/0
Reggae	0/0/0/0/0	0/0/0/0/0	2/0/0/0/1	5/5/6/6/**7**	0/1/0/0/0	0/0/0/0/0	1/2/2/2/1	0/0/1/1/0	1/2/1/1/0	1/0/0/0/1
Blues	1/2/0/0/0	1/2/1/1/1	0/0/0/0/0	0/0/0/0/0	**8**/5/**8/8/8**	0/0/0/0/0	0/0/0/0/0	0/1/1/1/1	0/0/0/0/0	0/0/0/0/0
Classical	0/0/0/0/0	0/0/2/1/1	0/0/0/0/0	0/0/0/0/0	0/0/0/0/0	**9**/8/7/8/**9**	0/0/0/0/0	1/1/0/0/0	0/1/1/1/0	0/0/0/0/0
HipHop	1/1/1/1/1	0/0/0/0/0	2/2/2/1/2	0/0/1/1/0	1/0/0/0/0	0/0/0/0/0	6/**7**/5/**7/7**	0/0/0/0/0	0/0/0/0/0	0/0/1/0/0
Rock	0/0/1/1/1	0/1/0/0/0	1/0/0/0/0	0/1/1/1/0	1/2/0/1/0	0/0/0/0/0	1/1/1/1/1	6/4/5/5/**8**	0/1/1/1/0	1/0/1/0/0
Country	0/0/0/0/0	1/1/1/1/1	0/0/0/0/0	1/1/0/0/1	2/3/0/0/0	0/0/0/0/0	0/1/2/1/0	0/1/2/2/0	6/3/5/6/**8**	0/0/0/0/0
Disco	0/0/0/0/0	0/0/0/0/0	2/2/0/0/0	2/0/1/1/1	0/1/0/0/0	0/0/0/0/0	0/0/1/1/0	1/2/1/1/1	0/1/1/0/0	5/4/6/7/**8**

[8] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the mfcc in speech and speaker recognition." in *INTERSPEECH*. ISCA, 2000, pp. 487–490.

[9] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile," in *ICMC Proceedings*, 2006.

[10] X. Yu, J. Zhang, J. Liu, W. Wan, and W. Yang, "An audio retrieval method based on chromagram and distance metrics," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*. IEEE, 2010, pp. 425–428.

[11] J. Serr, E. Gmez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech, and Language Processing*, 2008.

[12] J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 3642–3649. [Online]. Available: http://dl.acm.org/citation.cfm?id=2354409.2354694

[13] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[15] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[16] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene parsing," *arXiv preprint arXiv:1306.2795*, 2013.

[17] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 140–153.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[19] "Ilsvrc 2014," http://image-net.org/challenges/LSVRC/2014/index, accessed: 2014-08-31.
[20] "Ilsvrc 2012 results," http://image-net.org/challenges/LSVRC/2012/results.html, accessed: 2014-08-31.
[21] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information*, pp. 1–9, 2012. [Online]. Available: http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf
[22] "Mnist dataset," http://yann.lecun.com/exdb/mnist/, accessed: 2014-08-31.
[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.
[24] W. Dai, G. rong Xue, Q. Yang, and Y. Yu, "Transferring naive bayes classifiers for text classification," in *In Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 2007, pp. 540–545.
[25] J. na Meng, H. fei Lin, and Y. hai Yu, "Transfer learning based on svd for spam filtering," in *Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on*, June 2010, pp. 491–494.
[26] H. Wang, F. Nie, H. Huang, and C. Ding, "Dyadic transfer learning for cross-domain image classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 551–556.
[27] *A Practical Transfer Learning Algorithm for Face Verification*. International Conference on Computer Vision (ICCV), 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=202192
[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[29] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. EUSIPCO*, 2008.