Multi-scale Embedded CNN for Music Tagging

Nima Hamidi, Mohsen Vahidzadeh, Stephen Baek

The University of Iowa

Introduction

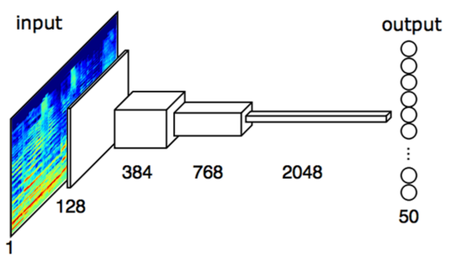
Abstract

Convolutional neural networks (CNN) recently gained notable attraction in a variety of machine learning tasks: including music classification and style tagging. In this work, we propose implementing intermediate connections to the CNN architecture to facilitate the transfer of multi-scale/level knowledge between different layers. Our novel model for music tagging shows significant improvement in comparison to the proposed approaches in the literature, due to its ability to carry low-level timbral features to the last layer.

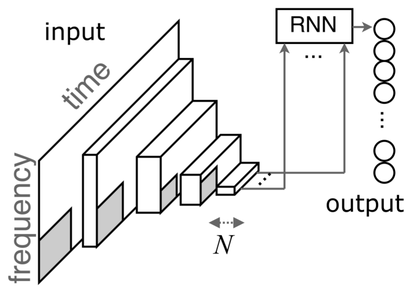
Objectives

Music Tagging

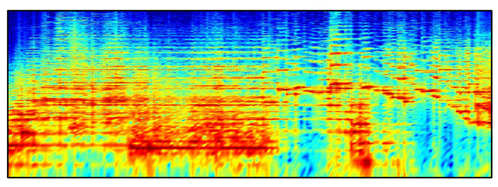
Music classification and style tagging is a traditional problem in music information retrieval (MIR) which entails predicting specific tags of a song, including genre, emotion, and instrumentation. Two main types of neural networks used in music classification algorithms include convolutional neural networks (CNN) and recurrent neural networks (RNN), both of which were initially designed to facilitate different problems in image and language processing.



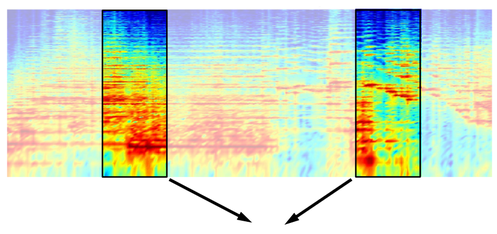
Automatic Music Tagging Using CNN



Convolutional Recurrent Network For Music Tagging



Receptive field in CNN

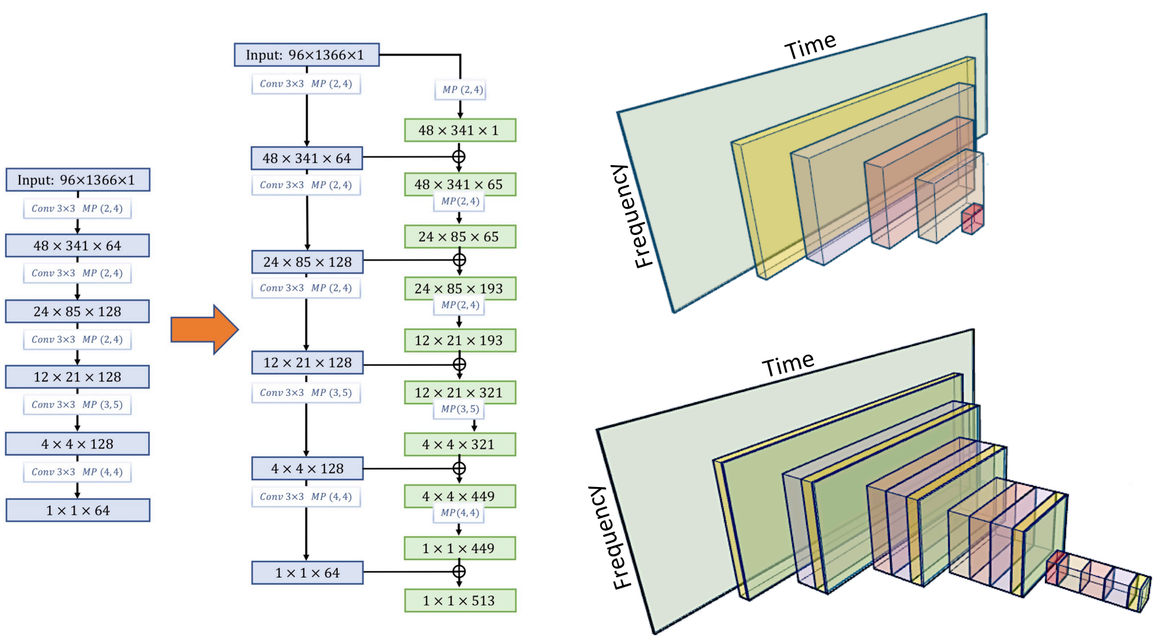


Receptive field in RNN

Methods

MsE-CNN

We propose that, by utilizing intermediate connections in our CNN architecture, we can carry important multi-scale features to the last layer for improved classification. We claim that such an approach will allow the model to learn low-level features such as musical texture and timbre as well as high-level temporal characteristics to improve the classification.



Results

Experiment

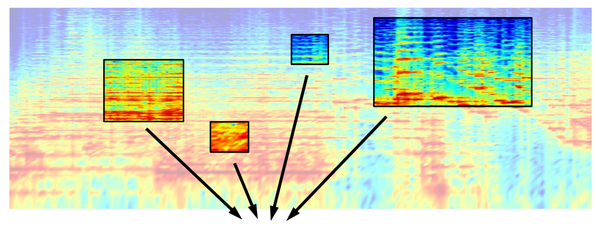
In this study we used Magna Tag Tune (MTT) dataset to tag songs and evaluate the performance of the model using both Area under the receiver operating characteristics (ROC-AUC) and Area under the precision-recall curve (PR - AUC).

Model	ROC-AUC	PR-AUC
MsE-CNN (ours)	0.914	0.423
FCN-5 (reproduced)	0.897	0.404
FCN-4 [2,3]	0.894	0.376
End to End Learning [2]	0.904	0.381
Timbre CNN [2,3]	0.893	0.349
Bag of features and RBM [2]	0.888	-
1D convolutions [2]	0.882	-
Transferred learning [2]	0.88	-
Multi-scale approach [2]	0.898	-
Pooling MFCC [2]	0.861	-

Conclusions

The change of timbral quality over time affects our perception of music, which is critical to the understanding of the genre, mood, instrumentation. We propose that in spectral representation of music, timbre is equivalent to texture and color in an image while long-term temporal structures are equivalent to shapes such as eye, hand, and nose. We believe a traditional CNN model learns long-term structures; however, it forgets textual features in the final classification step. Similar to the intermediate connection in U-Net and FPN, one can improve music classification by transferring low-level characteristics via such links.

Similar to a vision task, learning musical textures and timbre as well as long-term characteristics is crucial in audio classification.



Multi-scale receptive field
Multi-level features from different layers

Our model, MsE-CNN, is an experiment to support the idea that timbral fluctuation over time is disregarded while going deeper in a CNN model. Due to the larger receptive field in later layers in a CNN, the model starts to forget low-level features that carry the textual details which are indeed crucial for audio classification.

References

Choi, K., Fazekas, G., Sandler, M. Automatic Tagging Using Deep Convolutional Neural Networks. *In 17th International Society of Music Information Retrieval Conference*.

Choi, K., Fazekas, G., Sandler, M., Cho K. Convolutional Recurrent Neural Networks for Music Classification. *In 2017 IEEE International Conference on Acoustics, Speech, and Signal*.

Dieleman, S. and Schrauwen, B. Multiscale approaches to music audio feature learning. *In 14th International Society for Music Information Retrieval Conference*.

Dieleman, S. and Schrauwen, B. End-to-end learning for music audio. *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. *In CVPR 2017*.

Ronneberger, O., Fischer, Ph., Brox Th. U-Net: Convolutional Networks for Biomedical Image Segmentation. *In MICCAI 2015*

Contact Information

Nima Hamidi

www.nimahamidi.com

nima.hamidi.g@gmail.com