DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2019

A comparative analysis of CNN and LSTM for music genre classification

Gabriel Gessle and Simon Åkesson

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Degree Project in Computer Science, DD142X
En jämförande analys av CNN och LSTM för klassificering av musikgenrer

Authors

Gabriel Gessle and Simon Åkesson
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
June 2019

Examiner

Örjan Ekeberg
KTH Royal Institute of Technology

Supervisor

Erik Fransén

KTH Royal Institute of Technology

Abstract

The music industry has seen a great influx of new channels to browse and distribute music. This does not come without drawbacks. As the data rapidly increases, manual curation becomes a much more difficult task. Audio files have a plethora of features that could be used to make parts of this process a lot easier. It is possible to extract these features, but the best way to handle these for different tasks is not always known. This thesis compares the two deep learning models, convolutional neural network (CNN) and long short-term memory (LSTM), for music genre classification when trained using mel-frequency cepstral coefficients (MFCCs) in hopes of making audio data as useful as possible for future usage. These models were tested on two different datasets, GTZAN and FMA, and the results show that the CNN had a 56.0% and 50.5% prediction accuracy, respectively. This outperformed the LSTM model that instead achieved a 42.0% and 33.5% prediction accuracy.

Keywords

Bachelor thesis, music genre classification, GTZAN, FMA, CNN, LSTM

Sammanfattning

Musikindustrin har sett en stor ökning i antalet sätt att hitta och distribuera musik. Det kommer däremot med sina nackdelar, då mängden data ökar fort så blir det svårare att hantera den på ett bra sätt. Ljudfiler har mängder av information man kan extrahera och därmed göra den här processen enklare. Det är möjligt att använda sig av de olika typer av information som finns i filen, men bästa sättet att hantera dessa är inte alltid känt. Den här rapporten jämför två olika djupinlärningsmetoder, convolutional neural network (CNN) och long short-term memory (LSTM), tränade med mel-frequency cepstral coefficients (MFCCs) för klassificering av musikgenre i hopp om att göra ljuddata lättare att hantera inför framtida användning. Modellerna testades på två olika dataset, GTZAN och FMA, där resultaten visade att CNN:et fick en träffsäkerhet på 56.0% och 50.5% tränat på respektive dataset. Denna utpresterade LSTM modellen som istället uppnådde en träffsäkerhet på 42.0% och 33.5%.

Nyckelord

Kandidatexamensarbete, klassificering av musikgenrer, GTZAN, FMA, CNN, LSTM

Acknowledgements

We would like to thank our supervisor and everyone else that have supported and helped us throughout this project.

Abbreviations

CNN Convolutional neural network

LSTM Long short-term memory

MFCCs Mel-frequency cepstral coefficients

Contents

1 Introduction

The amount of data that is available to us is rapidly increasing every day, to the point where manual curation is becoming infeasible and classification using automated systems a necessity. The music industry is no exception. Automating the process of music tagging would result in better organization of the data and thereby making further development using this data easier, such as creating themed playlists or recommending songs to users. Machine learning can be used to find the subtle patterns in the data, which would otherwise be very difficult to explicitly code algorithms for. One such case is determining what genre a song belongs to, which is the use case this report will cover. However, finding patterns in audio is not only useful for musical analysis. It is possible that the results of this study could find use within other fields that incorporate audio.

The purpose of this study is to provide insight into the strengths and weaknesses of two methods of music genre classification using machine learning, namely supervised learning using convolutional neural networks (section 2.5.1) and long short-term memory networks (section 2.5.2).

1.1 Problem

Around 28% of Internet users have tagged photos, news stories or blog posts online where the tags for these multimedia objects are referred to as ”social tags”. There is an issue when it is the users themselves that contribute with tags, which is that the trendy posts are the ones getting most of the attention which leaves a lot of posts unnoticed [5]. As the availability of data increases, the need for categorization of said data becomes necessary in order to make good use of it. There are multiple methods of classifying data, but the reasoning behind why one method should be chosen over another is still unclear. This report aims to help this problem by analyzing and comparing the aforementioned methods. The research thesis in this report can thereby be formulated as, *which of CNN and LSTM is best suited for music genre classification?*

2 Background

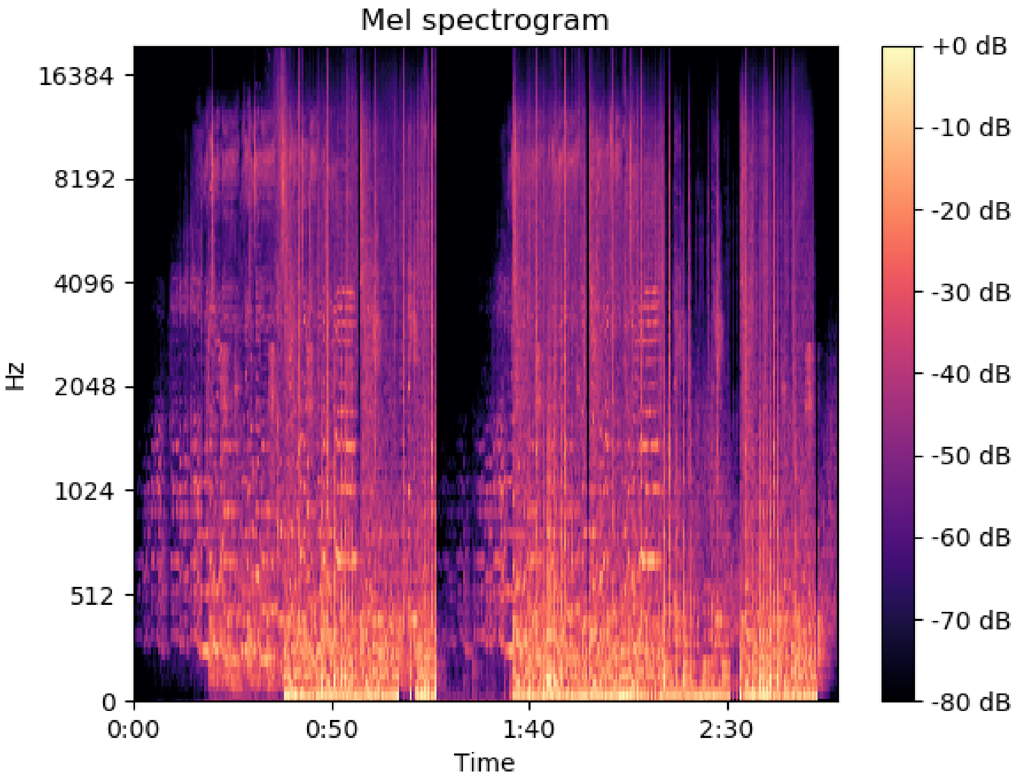
In order to understand how a computer can classify music with the methods used in this report, it is recommended to get introduced to the following concepts.

2.1 Genre classification

Genre classification can be seen as a way to classify music categorically by seeing it as belonging to a shared tradition or a set of different conventions [8]. What is important to note about this is that there is no explicit definition of what a genre should sound like, seeing as it is a very traditional way to look at music. There is no right or wrong in saying that a song should be classified a certain way, rather it is a subjective opinion of what that song makes the person feel and how they relate to it. However, there has to be some sort of consistency when it comes to what a genre sounds like, since classifying songs is not a random process for humans.

2.2 Mel-spectrogram

A spectrogram is a visual way to represent signal strength of different frequencies over time. One way to create this from an audio file is to use a Short-time Fourier transform, which transforms the audio signal into the frequency domain using a Fourier transform on shorter segments of the signal [1]. A mel-spectrogram is a spectrogram that has been mapped to the mel scale. This leads to lower frequencies gaining more importance, which is how humans interpret sound. This visual representation can then be viewed and analyzed by pattern recognition tools [6].



(a) Mel spectrogram created using librosa

The figure shows the frequencies present at a specific time of the song, where the scale on the right symbolises the different values. It can give you information such as when certain instruments start to play, like the bass as an example. In this figure, the yellow at the very bottom tells the spectator that something with a very low frequency has started playing.

2.3 MFCC

Mel-frequency cepstral coefficients (MFCCs) is a compressible representation of a mel-spectrogram. It is a set of features that describes the shape of the frequency spectrum. This is often used in audio analysis as an enhancement of the Fourier transform [11].

2.4 Supervised learning

Supervised learning is the task of finding a function that maps inputs to desired outputs, based on example input-output pairs. The idea is to learn patterns using the examples to then be able to predict outputs for new (previously unseen) inputs [7].

2.5 Neural network

A neural network is a network of nodes organized in layers. It defines a function between the nodes in the input layer and the nodes in the output layer. By training on samples with expected labels, it can learn to approximate the function that maps new samples to the correct label [3].

2.5.1 CNN

A CNN is a neural network especially suited for image recognition tasks [3].

2.5.2 LSTM

LSTM is a kind of recurrent neural network, which is a family of neural networks that specialize in sequential data, such as text and speech [3].

2.6 GTZAN dataset

The GTZAN dataset was created by sampling CDs, radio and microphone recordings to reproduce different environments a recording could be done within. All the data within the set was gathered between the years 2000 and 2001 and consists of 1000 audio tracks that are 30 seconds long, categorized into 10 genres of 100 tracks each. These genres are blues, classical, country, disco, hip hop, jazz, metal, pop, reggae and rock [12].

2.7 FMA dataset

The FMA dataset was made for musical analysis with up to 161 genres consisting of 106,574 untrimmed tracks from 16,341 artists and 14,854 albums in total. All of the data is MP3-encoded and offers the dataset in different sizes, where there is also a possibility to get 8,000 tracks with 8 balanced genres. These are electronic, experimental, folk, hip hop, instrumental, international, pop and rock. It was created for the Music Information Retrieval field, due to the lack of large datasets available for feature and end-to-end learning [2].

2.8 Related Work

An analysis has been done between using a CNN based model or four more traditional machine learning classifiers that utilizes hand-crafted features, where the conclusion was that the CNN outperformed them all. They also combined the CNN model with XGBoost to get an even better model. It is not something that will be tested in this report, but it is worth noting that combining several models can yield a greater result. Another thing worth noting is that the dataset used was *Audio Set*, which can come to affect the performance differences between these studies [1]. One thing this study did not do is compare CNN to another deep learning method, which is the focus of this report. The other algorithm studied is LSTM. It has been used for this task before, using seven LSTMs arranged in a novel tree structure [10]. We would like to repeat this experiment using a single conventional LSTM network to make the conditions as similar as possible to that of the CNN.

3 Method

3.1 Preprocessing

To analyze the audio files from the two datasets, a python library called Librosa was used. This makes it possible to load the audio file as a floating point time series and resampled to a specific sampling rate. In order to use the data now available in a CNN and LSTM, this report creates 20 MFCCs that the model can be trained on. There are several different types of ways to visualize audio files, where spectrograms and chromagrams are two examples that display what frequencies are being played over a period of time. This report only handles the usage of MFCCs.

The GTZAN and FMA datasets were split into one training and one validation set each, the first of which was used for training and the other (unseen by the model) for performance evaluation [9]. For both datasets random audio files were chosen from the different genres in such a way that there was the same amount of songs in each genres in both the training and validation set. The validation set for the GTZAN dataset consists of 200 songs, where there are 20 songs of each genre. While the validation set for the FMA dataset consists of 800 songs, where there are 100 songs of each genre. This left 800 and 7200 songs left respectively for each training set.

3.2 Models

3.2.1 CNN

MFCCs for each audio file in both datasets were created using a sampling rate of 22050Hz. In order to get good accuracy on the frequency domain together with decent performance a hop length of 512 was used together with a FFT window of length 2048. These were then standardized in order to accelerate the learning rate of the model.

The result was retrieved using a network built with Keras. It has five convolutional layers of 32 nodes each (window size 3x2) followed by one fully connected layer and the output layer, 128 and 10 nodes respectively. It was trained with a learning rate of 0.001. To prevent overfitting, a dropout of 25% on the convolutional layers and 50% on the fully connected layer was used. Categorical crossentropy served as the loss function and the activation function of the final layer was softmax in order for the song to only be classified under one genre.

3.2.2 LSTM

From each audio file 20 MFCCs were retrieved with a sampling rate of 22050Hz with the same parameters as the MFCCs created for the CNN. Reasoning behind only using 20 rather than a larger number is due to the model then becoming too complex.

This network was also built with Keras. It has 5 LSTM-layers of 128, 32, 32, 32, and 32 nodes respectively. The first layer has 50% dropout and the rest 30%. Again, categorical crossentropy and softmax were used.

3.3 Benchmarks

To get comparable results without trying to incorporate the quality of the file too much a sampling rate of 22050Hz was used for both datasets. The GTZAN dataset is the older of the two and is in a 22050Hz Mono 16-bit format, which is the cause of the chosen rate.

In order to benchmark the two different models they were tested on both datasets, where one has fewer samples than the other. This allows for comparison of performance when it comes to the amount of samples available. It is also interesting to note which of the models has the greater performance increase, as this can spark discussion as for which one might perform better for even larger datasets.

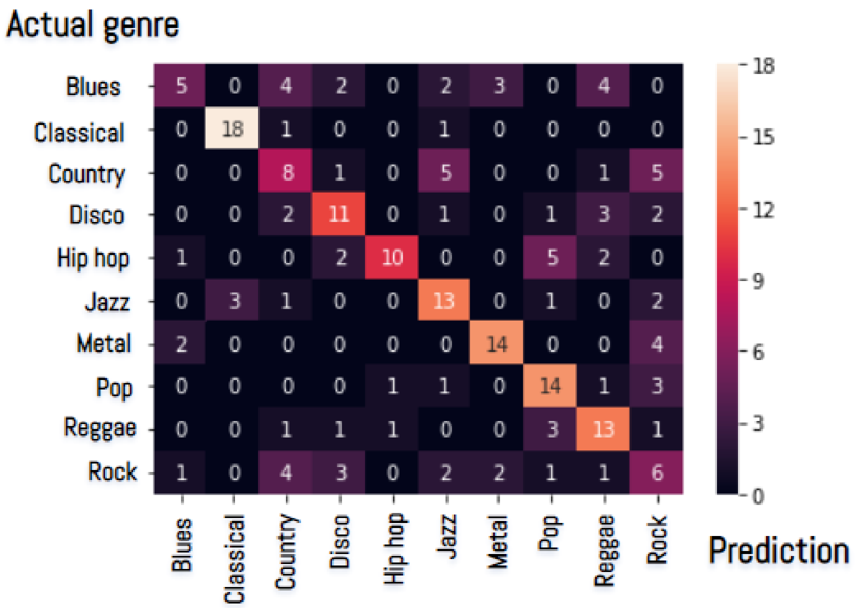
4 Results

To show the results a confusion matrix is given for each dataset the model was trained on, where the x-axis is the prediction of the model while the y-axis displays the true genre. This is accompanied by a diagram showing the model loss over time.

4.1 CNN

4.1.1 GTZAN

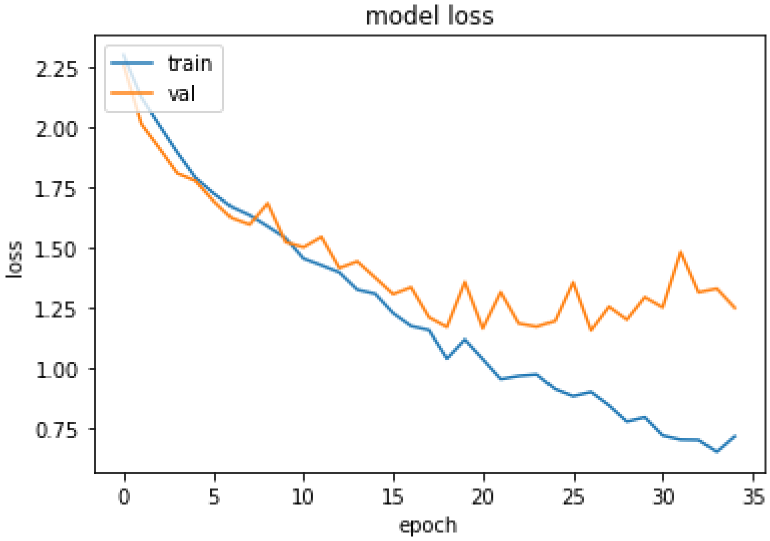
The following displays the model accuracy when trained on the GTZAN dataset resulted in a 56.0% accuracy.



	Blues	Classical	Country	Disco	Hip hop	Jazz	Metal	Pop	Reggae	Rock
Blues	5	0	4	2	0	2	3	0	4	0
Classical	0	18	1	0	0	1	0	0	0	0
Country	0	0	8	1	0	5	0	0	1	5
Disco	0	0	2	11	0	1	0	1	3	2
Hip hop	1	0	0	2	10	0	0	5	2	0
Jazz	0	3	1	0	0	13	0	1	0	2
Metal	2	0	0	0	0	0	14	0	0	4
Pop	0	0	0	0	1	1	0	14	1	3
Reggae	0	0	1	1	1	0	0	3	13	1
Rock	1	0	4	3	0	2	2	1	1	6

(a) Confusion matrix of the CNN model trained on the GTZAN dataset using MFCCs

The model managed to predict at least 50% correct in six out of the ten genres. The genres that it was struggling the most with were blues, country and rock. It had a really easy time predicting classical songs.

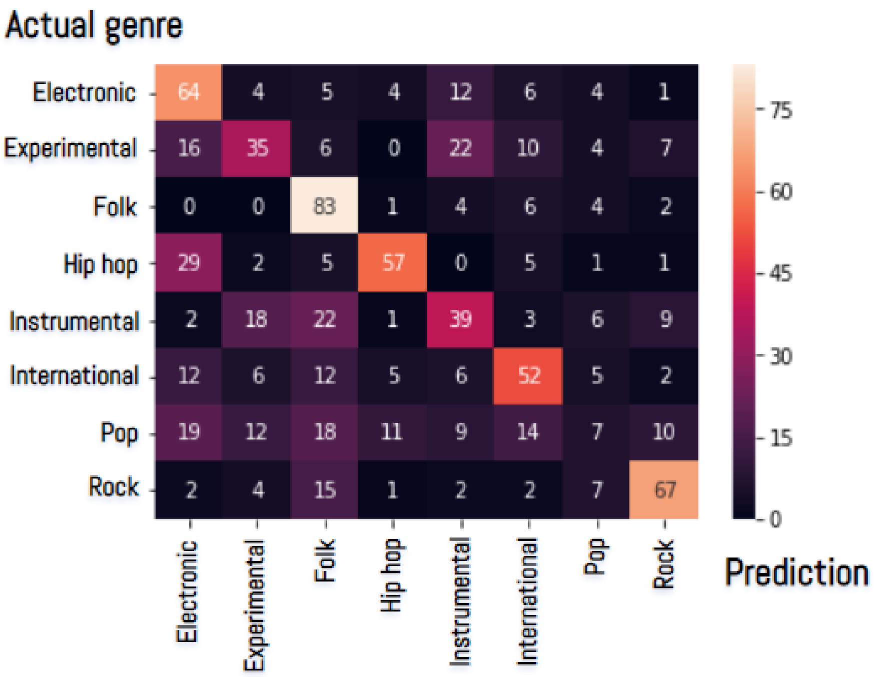


(a) Model loss of the CNN model trained on the GTZAN dataset using MFCCs

The validation set loss stopped getting better a little bit before 20 epochs.

4.1.2 FMA

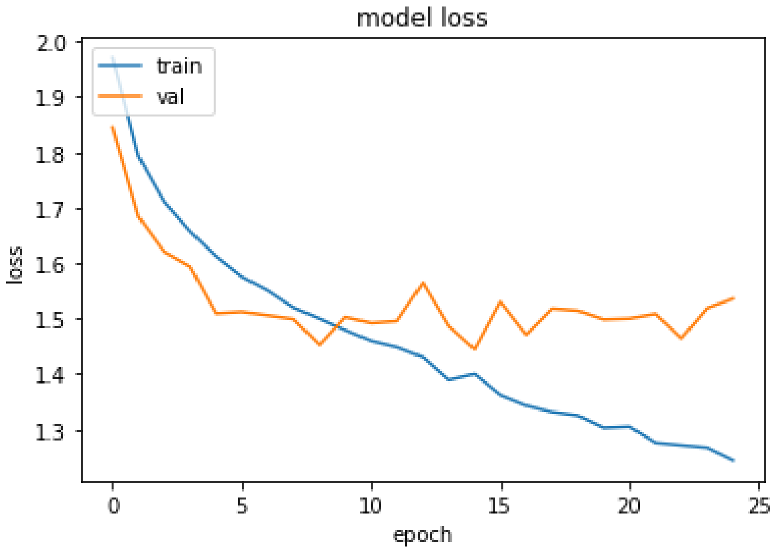
Training the model on the FMA dataset provided a different result with a 50.5% prediction accuracy, as shown in the following confusion matrix.



Actual genre	Electronic	Experimental	Folk	Hip hop	Instrumental	International	Pop	Rock
Electronic	64	4	5	4	12	6	4	1
Experimental	16	35	6	0	22	10	4	7
Folk	0	0	83	1	4	6	4	2
Hip hop	29	2	5	57	0	5	1	1
Instrumental	2	18	22	1	39	3	6	9
International	12	6	12	5	6	52	5	2
Pop	19	12	18	11	9	14	7	10
Rock	2	4	15	1	2	2	7	67

(a) Confusion matrix of the CNN model trained on the FMA dataset using MFCCs

Out of the eight genres, the model managed to get a prediction accuracy of at least 50% for five of them. It had an extremely difficult time to classify pop songs, where it only managed to predict 7 out of 100 songs correctly. Folk music was a different story, however, where it managed to get 83 out of 100 correctly.



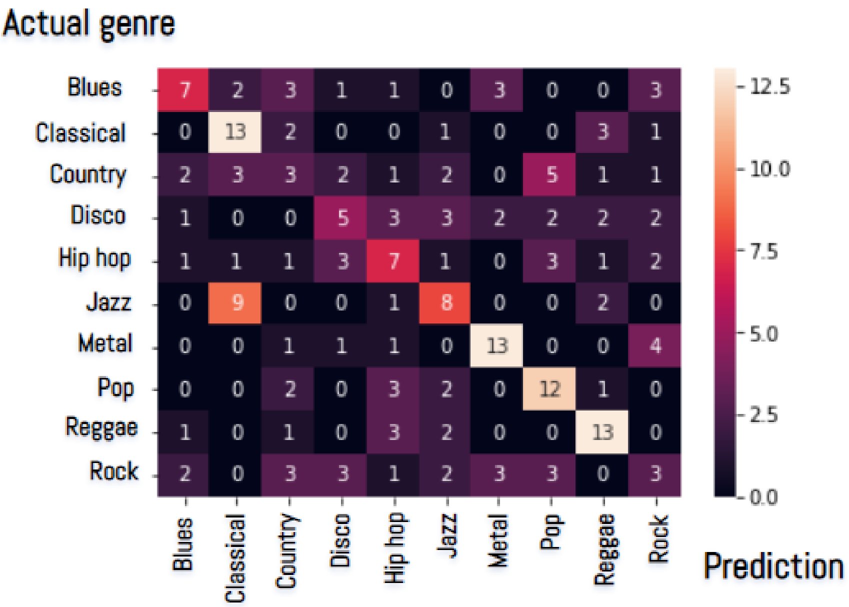
(a) Model loss of the CNN model trained on the FMA dataset using MFCCs

The loss graph shows us that the model loss started to converge around epoch 7.

4.2 LSTM

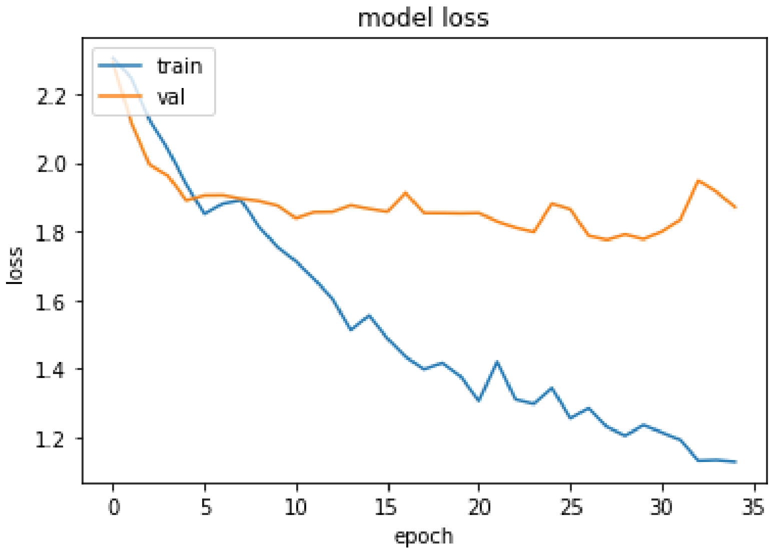
4.2.1 GTZAN

The following displays the model accuracy when trained on the GTZAN dataset which resulted in a 42.0% accuracy.



(a) Confusion matrix of the LSTM model trained on the GTZAN dataset using MFCCs

This model resulted in four genres having a over 50% prediction accuracy which were classical, metal, pop and reggae. None of those had any significant signs of being very easy to predict as the best one predicted 13 out of 20 correct.

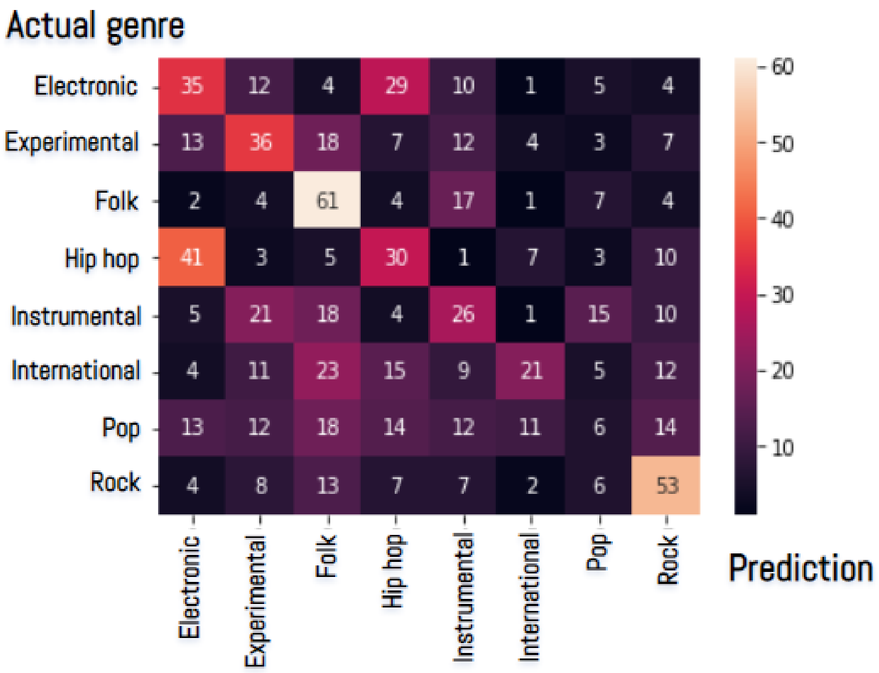


(a) Model loss of the LSTM model trained on the GTZAN dataset using MFCCs

In the loss graph it can be seen that the validation set loss started converging very early but kept getting smaller for quite some time, until it started ascending a bit around epoch 30.

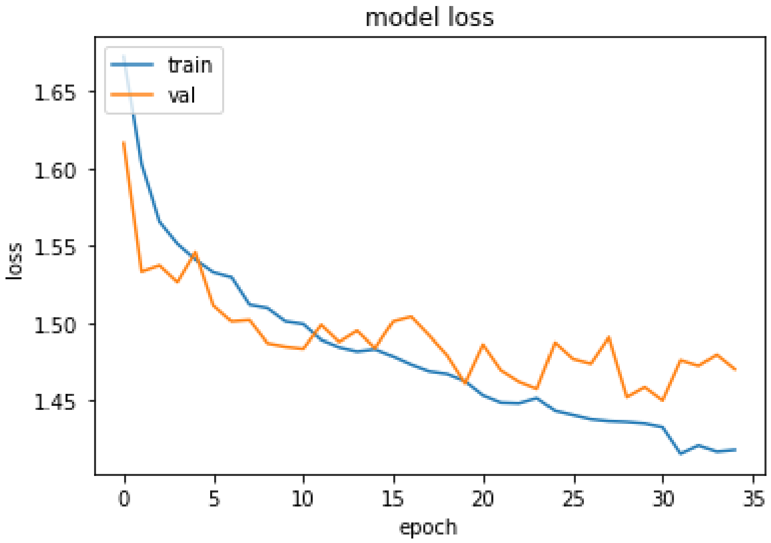
4.2.2 FMA

Training the model on the FMA dataset resultet in the following confusion matrix where the prediction accuracy was 33.5%.



(a) Confusion matrix of the LSTM model trained on the FMA dataset using MFCCs

Similarly to when the CNN model was trained on the same dataset, pop was difficult to classify correctly as this was only the case for 6 out of 100 songs. It only managed to predict two genres with an accuracy better than 50%, however, which were rock and folk.



(a) Model loss of the LSTM model trained on the FMA dataset using MFCCs

The loss graph shows that the model didn't get substantially better at predicting the genres around epoch 20, with a dip around epoch 30.

5 Discussion

There is an interesting topic worth discussing before going into the results and that is how musical genres evolve over time.

Music is constantly changing and what is now classified as pop music is not the same as pop music fifty years ago. Changes in genres have always been around, but now they're happening at a faster rate due to the availability that different services provide us. Considering genres are very subjective, an issue to look at when it comes to the classification of these is what would happen if they started sounding more and more like each other. The chances of this happening does most likely increase because of the influx of music availability, but does this mean we need new methods to classify songs? Probably not. It can, however, result in genres becoming more and more similar to each other and thus becoming a single genre. It is important to keep this in mind while discussing the results, considering that the datasets were created during different years and circumstances.

5.1 Results

When the models were trained on the GTZAN dataset we can see that the CNN outperformed the LSTM model. It was also a lot more consistent in its classifications, since we can see that there are a lot more genres never being predicted in the confusion matrix. This might be due to the CNN finding patterns in the data more easily since it had less parameters to train.

They were pretty similar in regards to what genres they had an easier time classifying, for example classical music. The reason why we believe that classical got such a good accuracy is due to it being subjectively distinct, since the arrangement of instruments is different compared to other genres. Another factor may be the lack of drums in classical pieces, which otherwise tend to fill up quite a bit of space in the frequency spectrum.

Rock, however, seems to be very difficult for the LSTM model to learn. It only managed to predict 3 out of the 20 songs correctly, whereas the CNN model managed to predict 6 out of 20. This is not a fantastic number either, but it is better. A reason for rock to have such a low score overall can be due to it having a lot of different sub-genres, which can result in it sounding like other genres in the dataset.

The FMA dataset didn't actually increase the performance of our models even though the sample size is significantly larger, they actually got worse. There can be several reasons behind this, such as human error when classifying the songs, the environment of the recordings or the genres being used.

Our CNN model did get a higher prediction accuracy than the LSTM model once again and the patterns we can see are very similar to when trained on GTZAN. We can see that the consistency in classifications is a lot better when it comes to the CNN model, also the fact that it had five genres it could predict with a better than 50% accuracy makes it the superior model according to this study.

An interesting observation that can be made is that folk was the easiest genre to predict for both models, which also has quite a distinct sound just like classical music that was discussed earlier. The unique instruments being used and the lack of electronic production might have a lot to do with this.

We believe that managing to predict a song's genre 56.0% of the time is still a decently good number, considering the diffuse nature of genres. When using both the smaller and larger sample size the CNN seems to outperform the LSTM. A lot of things could improve it such as using more features and a larger sample size. Another thing one could do is group genres, to decrease the amount of possible categories.

5.2 Method

The two datasets used in the research do not have the same genres, which could explain the discrepancy of the results. This could be due to the genres in a certain dataset being more similar to each other than in the other. The overall quality of the samples might also have had an affect, even though we sampled them with the same sampling rate, because of other factors such as how the recording was performed. This means that comparing the performance of the two models on the

GTZAN and FMA datasets might not only depend on the sample size, but also how they were collected.

If an even larger dataset was available, it would probably have been better to test performance on a smaller slice of it and then later doing a test on the entire collection. Not only would the genres be the same, but the human factor would also be minimized, since every song would be recorded and classified similarly.

As mentioned previously in the report, there are several types of features you can retrieve from an audio file. In this report we only touched upon MFCCs, but other features such as chromagrams could also be used to train the models. Comparing the results of training using different kinds of features would grant a better overview of the performance benefits for one model over the other.

6 Conclusions

According to this study, the CNN model outperformed the LSTM model regardless of which of the two datasets was used and that leads to the conclusion that CNN is better suited for music genre classification.

As mentioned in the background, there has been studies done on combining models with each other which might be the better thing to actually analyze. The CNN model showed to give a better result in this report however, which also leads to the conclusion that this might be a good model to build more complex ones with. Considering that it also was a lot more consistent in its classifications shows that it found patterns in the songs easier.

It certainly is possible to find patterns in songs related to their genre, but simply using a CNN or LSTM model might not be sufficient in getting a good enough model. This holds true to both datasets tested. For an automatic system to be put in place to tag all data that passes through, it would need a better prediction accuracy. In general, it is very difficult to say what a good enough prediction accuracy for an automatic system would be since even humans are not perfect at classifying songs. However, it could be useful for a system where human confirmation is used, as it would decrease the workload.

6.1 Future Work

This study only compared how well a LSTM model would compare to a CNN model when trying to classify the genre of a song, but it would also be interesting to see how well they would perform at classifying the emotions of a song. Being able to say if the song is energetic or happy as an example would give another way to help classify songs when genres start sounding like each other. A follow up study comparing these two models with different features would also help the understanding of which model works best depending on the information given.

There was also a report done on music classification that called genres out of date and that an alternative way that could be more accurate is to describe music through feelings. This is due to several factors, such as that music nowadays is so easy to publish and browse, which results in that the diversity in music increases and our concept of genres evolves [4]. This could mean that classifying music becomes a harder task, since a lot more types of music will be created. How will this affect classification of musical genres as a whole?

References

[1] Bahuleyan, Hareesh. "Music genre classification using machine learning techniques". In: *arXiv preprint arXiv:1804.01149* (2018).

[2] Defferrard, Michaël et al. "FMA: A Dataset for Music Analysis". In: *18th International Society for Music Information Retrieval Conference*. 2017. URL: `https://arxiv.org/abs/1612.01840`.

[3] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[4] Greenberg, David M. *Musical genres are out of date – but this new system explains why you might like both jazz and hip hop*. Aug. 2016. URL: `http://www.econotimes.com/Musical-genres-are-out-of-date---but-this-new-system-explains-why-you-might-like-both-jazz-and-hip-hop-244941` (visited on 04/29/2019).

[5] Law, Edith and Von Ahn, Luis. "Input-agreement: a new mechanism for collecting data using human computation games". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 1197–1206.

[6] Prahallad, Kishore. *Speech Technology: A Practical Introduction*. URL: `http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf` (visited on 04/28/2019).

[7] Russell, Stuart and Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 3rd. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN: 0136042597, 9780136042594.

[8] Samson, Jim. *Genre*. Jan. 2001. URL: `http://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040599`.

[9] Schneider, Jeff. *Cross Validation*. Feb. 1997. URL: `https://www.cs.cmu.edu/~schneide/tut5/node42.html` (visited on 06/07/2019).

[10] Tang, Chun Pui et al. “Music genre classification using a hierarchical long short term memory (LSTM) model”. In: *Third International Workshop on Pattern Recognition*. Vol. 10828. International Society for Optics and Photonics. 2018, 108281B.

[11] Tjoa, Steve. *Mel Frequency Cepstral Coefficients (MFCCs)*. July 2015. URL: `https://musicinformationretrieval.com/mfcc.html` (visited on 04/27/2019).

[12] Tzanetakis, G. *GTZAN Genre Collection*. 2015. URL: `http://marsyas.info/downloads/datasets.html` (visited on 04/20/2019).

TRITA-EECS-EX-2019:372