Music Genre Classification on the FMA Dataset

ENGG*6500 Fall 2022

Ben Chapman-Kish

Outline

- Context of music genre recognition
- Justification for dataset selection
- Literature review of historical approaches to MGR
- Advanced music feature selection
- Methodology of proposed novel approaches
- Implementation and results
- Conclusion

Background: what is MGR?

- ”Musical genres are categorical labels created by humans to characterize pieces of music”
- Music Genre Recognition/Classification (MGR): automatically recognizing the genre of a particular song
- First algorithms were designed in 2002 by Tzanetakis and Cook

Choice of dataset

- Historically, single point of truth was GTZAN (2002)
  - Analogous to MNIST for digits or CIFAR for object recognition
- GTZAN: 1000 songs spanning 10 genres
  - Of course, all the songs are over 2 decades old
  - Numerous issues pointed out by Sturm (2013): incorrect labels, heavy artist repetition, and copyrighted songs
- Along comes Free Music Archive (FMA) in 2016 to fix these issues
  - Choice of 8000 songs and 8 genres or 106574 songs and 161 genres

Previous classifier approaches

- Originally, MGR was performed with basic ANNs
- Accuracy of 65% was achieved

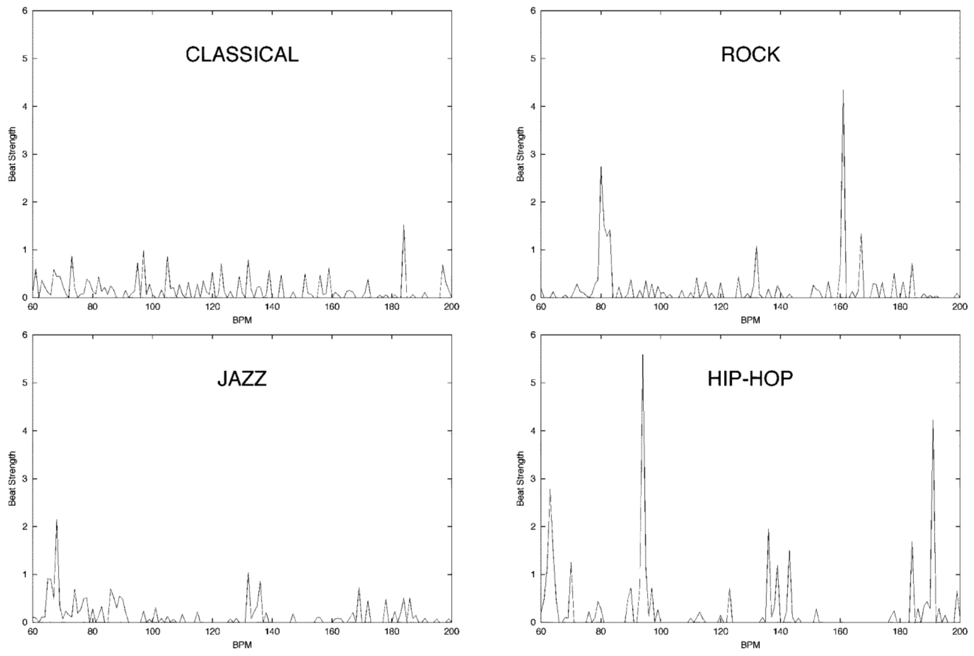
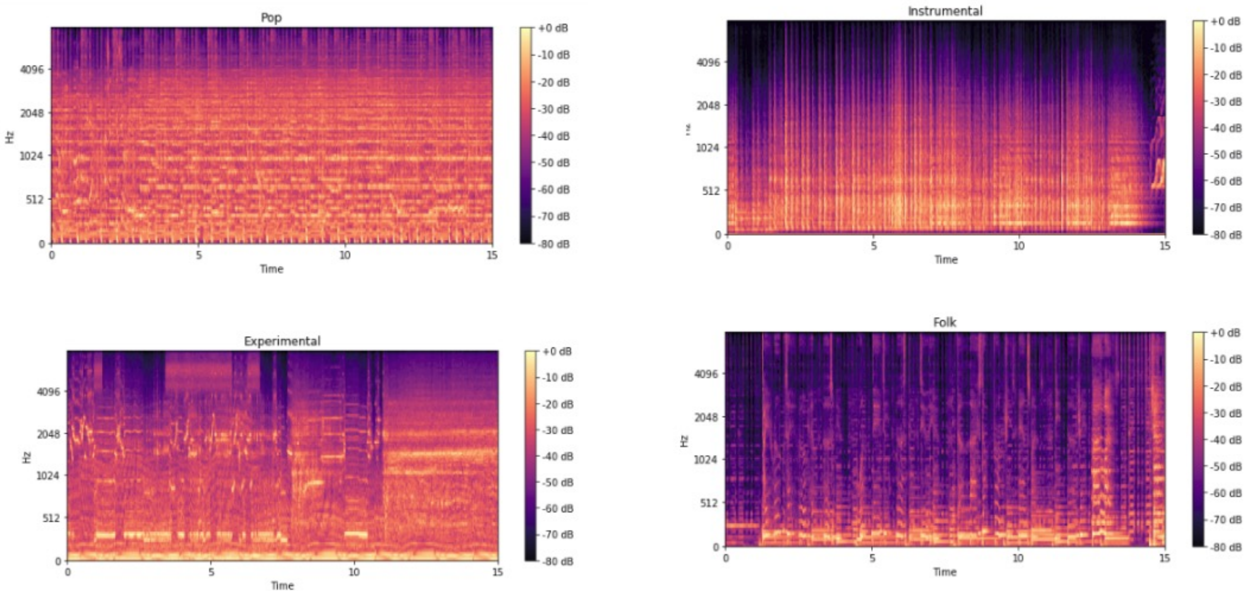


Fig. 3. Beat histogram examples.

- Key differentiating qualities of genres: timbral texture, rhythmic content, pitch content
- By 2010, the most successful models according to Fu et al:
  - KNNs, SVMs, and GMMs
- Deep learning networks started to take over after 2012

But what are we training on exactly?



- Basic premise is familiar to us electrical & computer engineers:
  - Looking at the frequency spectrum of a signal, yay!
- Industry standard for MGR is called Mel-frequency cepstral coefficients (MFCCs):
  1. Calculate the short-term Fourier transform of the audio wave
  2. Map it onto the Mel scale (of melody frequencies)
  3. Take the logs of the powers at each Mel frequency
  4. Find the discrete cosine transform
  5. The amplitudes of this are our features
- MFCCs provide significant insight into differences between music genres

Modern classifier approaches

- Sigtia and Dixon (2014) used Random Forest, max voting, and dropout to get 83% accuracy on GTZAN
- Dieleman and Schrauwen (2014) were first to use CNNs on MGR
- This was taken further by Choi et al. in 2017, cross-breeding CNNs with RNNs to get a CRNN that out-performs all old models

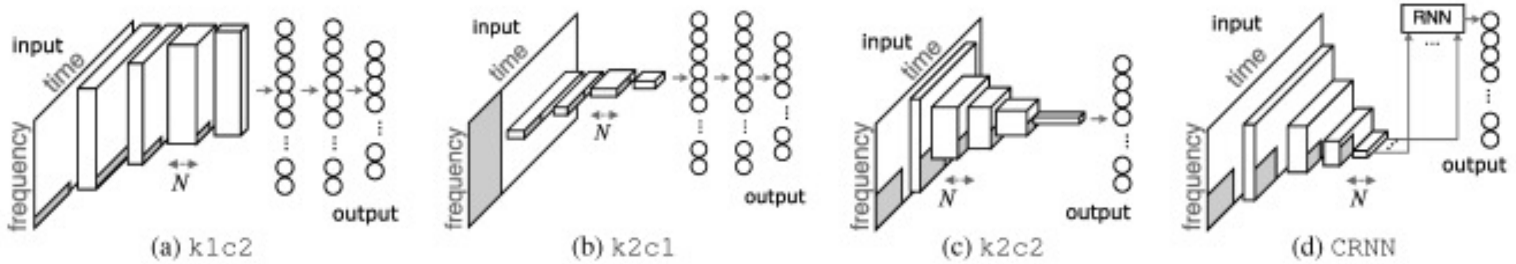


Fig. 1: Block diagrams of k1c2, k2c1, k2c2, and CRNN. The grey areas illustrate the convolution kernels. N refers to the number of feature maps of convolutional layers.

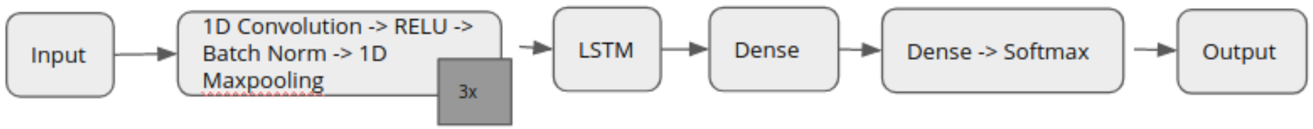
Putting it all together

```
import librosa

def compute_mfcc(filepath) -> np.ndarray:
    y, sr = librosa.load(filepath)

    mel = librosa.feature.melspectrogram(
        y=y, sr=sr, n_fft=2048, hop_length=1024)
    log_spect = librosa.power_to_db(mel, ref=np.max)
    mfcc = librosa.feature.mfcc(S=log_spect, n_mfcc=20)

    return mfcc

tracks = pd.read_csv("tracks.csv")
features = [compute_mfcc(f) for f in tracks]
```

- Let’s start by getting the dataset and processing the audio files to get the MFCCs
- All the usual data processing can happen too
    - For every model I applied standard scaling and represented the genres with one-hot encoding
- MFCCs are higher-dimensional, but for a basic test we can flatten the data and pass it through a SVM
    - Not a great idea – test accuracy is 47%

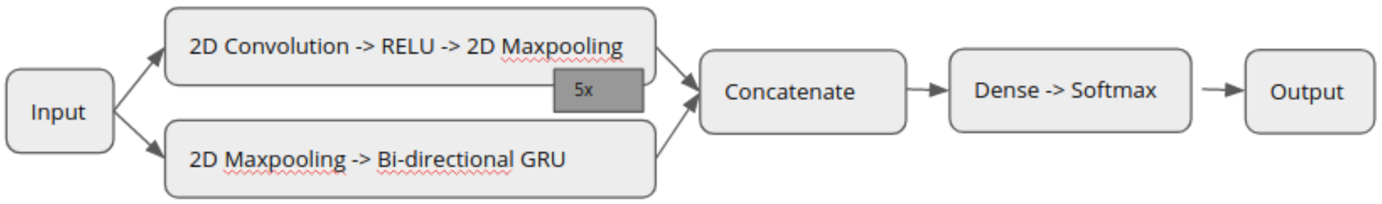
Stepping it up a notch

- Better idea: using a contemporary deep learning network designed for MGR
  - None of those networks have been used on FMA yet
- Let’s try both the CRNN and a parallel network (PRCNN)
- Keep in mind that the best classifiers on FMA struggle to reach above 80% accuracy



The CRNN architecture

- The CRNN performs relatively well on FMA, achieving around 66% accuracy



The PRCNN architecture

- The PRCNN does even better though, at 71% accuracy

Conclusion?

- Music is a very complex subject, even humans struggle to identify genres for many songs
- The best ways to differentiate genres appear to be a combination of pitch, timbre, and rhythm – all of which can be represented on the Mel frequency spectrum
- Recurrent networks are well-suited for the temporal aspects and convolutional networks are excellent at summarizing higher dimensional audio features
- Parallel techniques show the most promising performances so far


That’s all for today, rock on!