LAB 2 REPORT

Music Genre Classification - CNN

Manuel Castillo Obregón

Universitat Pompeu Fabra

Machine Learning for Sound and Music

Google Colab Notebook

GitHub Repository

1. INTRODUCTION

This lab explores convolutional neural networks (CNN in the rest of the document) for music genre classification, using Mel Spectrogram images as input. The following are implemented, as per the instructions for Task 1 and Task 2 as well as for analysis:

1. Task 1: A CNN trained from scratch and optimized through experimentation with convolutional depth, filters, dropout and augmentation.
2. Task 2: fine tuning is implemented for a pretrained model (VGG16) for the same dataset.
3. Energy and CO2 impact analysis using CodeCarbon.
4. Comparison with Lab 1 where models used manually extracted features instead of images.

The dataset consists of 10 genres, divided in two folders with audio and images. There's 1 Mel Spectrogram per audio file, totaling 1000 images.

The data was split in 80-20, with 20% of the training set used as validation.

2. DATASETS

We used 1000 Mel Spectrogram images across 10 balanced genres. The dataset was split into 80% training and 20% testing, with 20% of the training data used for validation. Additional feature CSV files from Lab 1 were used only for comparison.

- Train set: 640 images
- Validation set: 160 images
- Test set: 200 images
- 10 balanced classes

3. METHODS

3.1 Data preprocessing

All audio files from the dataset were converted into Mel Spectrogram images. Each spectrogram was resized to 128x128x3 for the CNN and 224x224x3 for VGG16.

Pixel values were normalized using:

- Rescale 1/255 for the CNN.
- VGG16 preprocess input for transfer learning.

A unified split was used across all models:

- 80% training and 20% validation
- 20% test
- Stratified by labels
- Same split for Task 1 and Task 2

Small data augmentation was applied in Task 1 to reduce overfitting.

3.2 Algorithms

3.2.1 CNN from scratch: Task 1

A compact convolutional architecture was designed.

- Three convolutional blocks.
- Filters doubled per block.
- Dense 256 and Dropout 0.3
- Adam optimizer
- EarlyStopping
- Categorical cross entropy

3.2.2 VGG16 Transfer learning: Task 2

A two phase fine tuning pipeline was used.

- Feature extraction phase
- Fine tuning phase

3.3.3 Environmental impact

To evaluate the computational footprint of each model, the following was measured:

1. Wall-clock runtime of the cross-validation procedure.
2. Peak RAM usage during model evaluation.
3. Estimated CO2 emissions (kg CO2eq) using CodeCarbon.

4. TASK 1

A compact CNN with three convolutional blocks was trained using: mild data augmentation, ReduceLROnPlateau, and EarlyStopping. The model achieved:

- Test accuracy: 0.55
- Test loss: 1.411
- CO2 emissions: 0.000153 kg
- Training time: 287.6 s

The CNN showed moderate generalization and better performance in some classes although significant confusion could be noticed in others, specially for genres with overlapping spectral patterns.

5. TASK 2

A VGG16 model was pretrained on ImageNet was fine tuned in two stages. The model achieved:

- Test accuracy: 0.67
- Test loss: 1.069
- CO2 emissions (head and fine tune): 0.000389 kg
- Training time (head and fine tune): 732.6s

VGG16 significantly outperformed the CNN from scratch. This shows the effectiveness of transfer learning even with small datasets.

6. GENERAL CONCLUSSIONS

In this lab we explored convolutional and pretrained neural networks for classifying music and genre using 2D Mel Spectrograms. Two deep learning strategies were implemented: a CNN trained from scratch and a VGG16 for transfer learning.

6.1. Key findings:

1. CNN from zero: Moderate success.
    a. The CNN reached a 55% accuracy. This shows that learning is possible but data limited.
    b. CNNs trained from zero typically require a large dataset.
    c. Despite the low accuracy, the CNN captured patterns for specific genres, specifically in genre 1 and 6, which shows the potential of the algorithm.
2. VGG16 fine tuning: It was a significant boost.
    a. Transfer learning achieved 67% accuracy, outperforming the algorithm from Task 1 by 12 percentage points.
    b. This confirms that pretrained visual features are advantageous, even with small datasets.
3. Comparison with Lab 1: Traditional ML seems more efficient.
    a. Feature engineered models from Lab 1 outperform deep models from Lab 2.
    b. This could be expected as MFCC based features encode timbre, spectral envelope and energy. These are highly meaningful descriptors for genre classification.
    c. When datasets are small, classical ML and strong features will outperform deep networks.
4. Energy impact:
    a. Although VGG required longer training time than the scratch CNN, both produced low CO2 emissions.
    b. Deep learning audio pipelines can be environmentally sustainable when properly optimized.

6.2. Overall Conclusion

With small datasets, pretrained CNN will outperform algorithms executed from zero but classical featured based Machine Learning still achieves a better accuracy.

Deep learning seems to be competitive only when we have a lot of data or when we need strong augmentation.

Transfer learning looks like the most promising approach when transitioning from handcrafted features to spectrogram based learning.