25th International Conference in Pattern Recognition
Milan, 10-15 January 2021

Mood detection analyzing lyrics and audio signal based on deep learning architectures

K. Pyrovolakis, P. Tzouveli, G. Stamou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

konpyro@gmail.com, tpar@ntua.gr, gstam@cs.ntua.gr

Contents

1. Introduction
2. From Audio and Lyrics to Mood
3. Data Preparation
4. System Architecture
5. Results

Introduction

The terms music and mood are two concepts strongly connected.

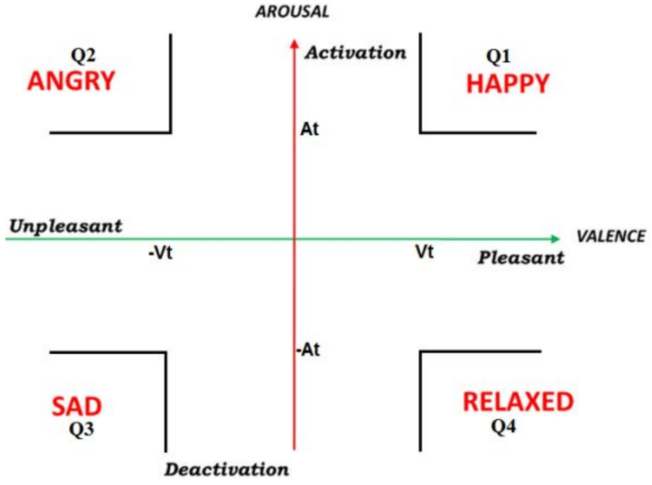
In our paper we investigate how to detect the mood of a music track applying Deep Learning techniques.

What was our approach?

- A Lyric analysis model
- An Audio analysis model
- A Multichannel model
- Compare the results

The Emotional Model

- Rusel's Circumplex is the emotional model we used in our work
- According to Circumplex all human emotions are distributed in a two-dimensional space with axes of valence and arousal
- Each quadrum represents a mood class



Valence (V) and arousal (A) values	Mood
$A > A_t$ and $V > V_t$	Happy
$A > A_t$ and $V < -V_t$	Angry
$A < -A_t$ and $V < -V_t$	Sad
$A < -A_t$ and $V > V_t$	Relaxed

From Audio to Mood

Association between structural features of music and emotion

Structural Feature	Definition	Associated Emotion
Tempo	The speed or pace of a musical piece	Fast tempo:happiness, excitement, anger. Slow tempo: sadness, serenity.
Mode	The type of scale	Major tonality: happiness, joy. Minor tonality: sadness.
Loudness	The physical strength and amplitude of a sound	Intensity, power, or anger
Melody	The linear succession of musical tones that the listener perceives as a single entity	Complementing harmonies: happiness, relaxation, serenity. Clashing harmonies: excitement, anger, unpleasantness.
Rhythm	The regularly recurring pattern or beat of a song	Smooth/consistent rhythm: happiness, peace. Rough/irregular rhythm: amusement, uneasiness. Varied rhythm: joy.

Features extracted from audio that we experimented with:

- Spectogram
- Mel Spectogram
- Log-Mel Spectogram
- MFCCs
- Chroma features
- Centroid tonal features
- Spectral contrast

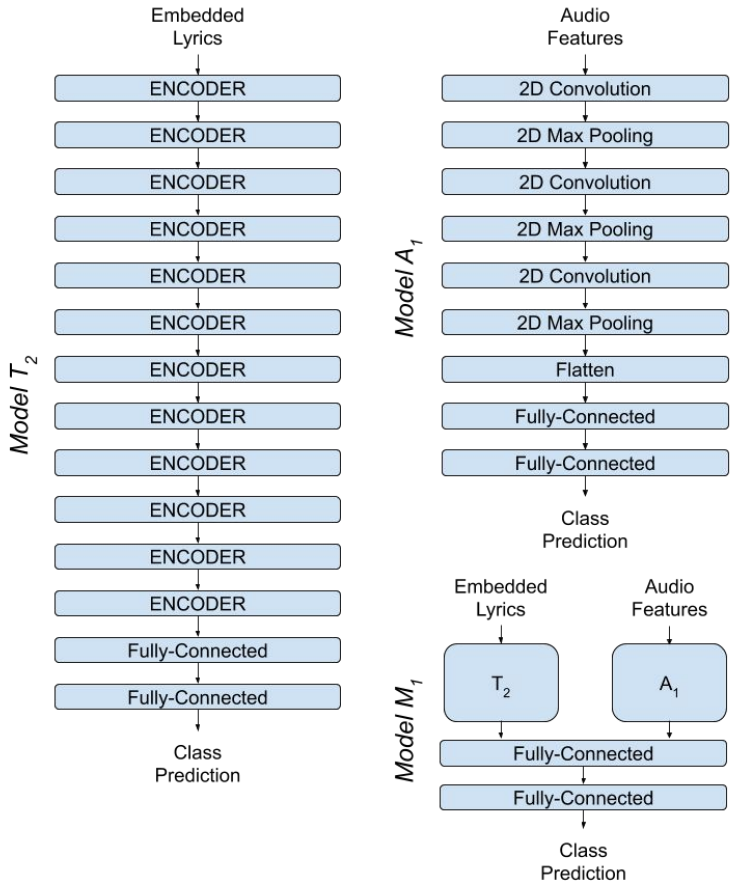
From Lyrics to Mood

- Each world in lyrics is attributed to pair of valence and arousal values
- The set of values is computed with the help of dictionaries which contain emotional information
- A general pair of valence and arousal values is computed for each song

Data Preparation

- The dataset we used is the MoodyLyrics Dataset
- 2.000 song titles with their corresponding mood label
- Mood labels = {happy, angry, sad, relaxed}
- Audio data
  - Collect audio files from web
  - Augment samples (37.989 audio samples)
  - Extract audio features
- Lyrics data
  - Collect lyrics from web
  - Augment samples (18.115 lyrics samples)
  - Compute BERT Embeddings

System Architecture



How the multichannel system (M_1) is developed?

- Train BERT-base uncased model (T_2) on lyrics
- Train CNN model (A_1) on audio signal
- System M_1 is implemented as the fusion of A_1 and T_2 with a common classifier of two fully connected layers

Results

- <u>Lyric Analysis Subsystem</u>
  We trained BERT model (T_2) and compared its results with several text analysis techniques

Model	Embedding Method	Loss	Accuracy %
T_1'	BoW	1.287	65.49
T_1'	TF-IDF	1.381	67.98
T_1	Word2Vec	1.262	41.66
T_1	GloVe	1.064	53.33
T_2	Bert	1.353	69.11

Results

- Audio Analysis Subsystem
  We trained CNN model (A_1) and experimented with different possible feature combinations

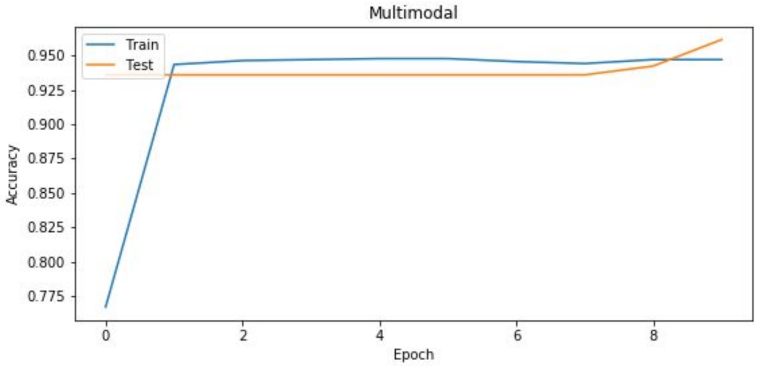
Feature Combination	Accuracy %
Mel	64.97
Mel, Log-Mel	68.38
Mel, Chroma, Tonnetz, Spectral Contrast	60.86
Log-Mel, Chroma, Tonnetz, Spectral Contrast	58.96
MFCC, Chroma, Tonnetz, Spectral Contrast	65.36
Mel, Log-Mel, MFCC, Chroma, Tonnetz	69.77
Mel, Log-Mel, MFCC, Chroma, Tonnetz, Spectral Contrast	70.34

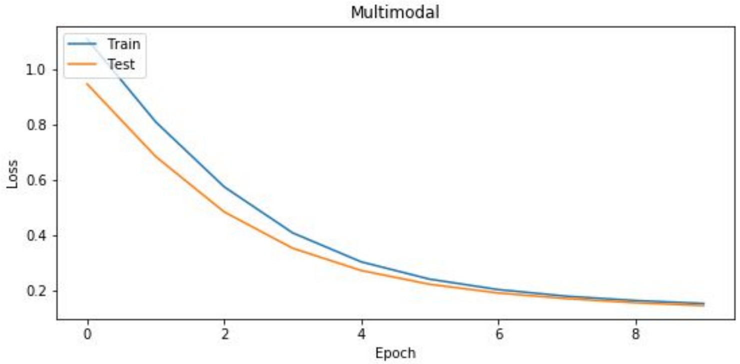
Results

- <u>Fuse Analysis System</u>
  We used the already trained subsytems to train our multichannel model (M_1)
  And compared its results with the previous models

Model	Loss	Accuracy %	Computational Time
$T_1^{'}$	1.381	67.98	0m 25.391s
T_2	1.353	69.11	18m 12.444s
A_1	0.743	70.51	80m 13.064s
M_1	0.156	94.58	3m 38.551s

Results


Multimodal
Train
Test
Accuracy
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0
2
4
6
8
Epoch


Multimodal
Train
Test
Loss
1.0
0.8
0.6
0.4
0.2
0
2
4
6
8
Epoch

Conclusion

- BERT outperforms simple text analysis techniques
- The combination of all six audio features has the best performance on the task
- Fusing the two subsystem into one complex system achieves huge improvement in performance and outperforms single channel systems