DISCLAIMER

My prediction data in my google drive somehow disappeared so I had to run all the predictions again tonight (28.feb).

So I ran my analysis on not all the data and I will be updating the numbers and discussion of the results if the results change a lot.

Numbers in this paper are generated with;

Musicnn: 434/574 chunks predicted

VGGish: 265/574 chunks predicted

I will submit it not fully finished because I want to submit on time and like I said I will be updating it, when prediction data is finished

Audio And Music Processing Lab

Module 1 - Large audio datasets assignment

For this task I will annotate music data by hand. I was given 574 audio chunks to annotate. I have eleven characteristics for annotating. The characteristics are;

- If the song is electric
- If the song is acoustic
- If the song is aggressive
- If the song is relaxed
- If the song is happy
- If the song is sad
- If its a party song
- If the song is tonal or atonal
- Danceability
- If the song is instrumental
- Finally if the song is not instrumental then what gender is the singing voice

One of the main purposes for annotating data by hand is to transform that data into a form, that is suitable for computer-aided analysis. I will take my annotated data and compare it to two different pre-existing models, musicnn and vggish, that predict these characteristics automatically and analyse the accuracy of those models.

My process was pretty straight forward. I used the programming environment Colab because it is the easiest way to use Essentia (I have not figured out how to install Essentia on my macBook computer). I installed all the prediction models using !wget. I iterated over all the sound chunks and for each sound chunk I used each model to predict all the characteristics and saved the results in a json file so it was formatted the same way as the data which was annotated by hand. The predictions from the models gave me an array of probabilities for each characteristic, so I find what index gives me the best probability and that is what the model classified. I then calculate the accuracy for each model given that my annotation are ground truth and calculate confusion matrices for the results. I also noticed that some data was flipped and calculated the accuracies with that in mind.

Accuracy table for musicnn model:

	mood_acoustic	mood_electronic	mood_aggressive	mood_relaxed	mood_happy	mood_sad	mood_party	tonal_atonal	danceability	voice_instrumental	gender
Accuracy	0.7575057736720554	0.7304147465437788	0.7603686635944701	0.5184331797235023	0.684331797235023	0.42857142857142855	0.836405529953917	0.5783410138248848	0.6889400921658986	0.7972350230414746	0.9838709677419355

Accuracy table for vggish model

	mood_acoustic	mood_electronic	mood_aggressive	mood_relaxed	mood_happy	mood_sad	mood_party	tonal_atonal	danceability	voice_instrumental	gender
Accuracy	0.8295454545454546	0.8409090909090909	0.7765151515151515	0.6212121212121212	0.6439393939393939	0.6363636363636364	0.803030303030303	0.3371212121212121	0.7272727272727273	0.8409090909090909	0.9545454545454546

Confusion matrices for vggish model

		Actual Values	
	mood_acoustic	Positive	Negative
Predicted Values	Positive	66	38
	Negative	7	153

		Actual Values	
	mood_electronic	Positive	Negative
Predicted Values	Positive	133	28
	Negative	14	89

		Actual Values	
	mood_aggressive	Positive	Negative
Predicted Values	Positive	29	46
	Negative	13	176

		Actual Values	
	mood_relaxed	Positive	Negative
Predicted Values	Positive	97	99
	Negative	1	67

		Actual Values	
	mood_happy	Positive	Negative
Predicted Values	Positive	31	57
	Negative	37	139

		Actual Values	
	mood_sad	Positive	Negative
Predicted Values	Positive	136	92
	Negative	4	32

		Actual Values	
	mood_party	Positive	Negative
Predicted Values	Positive	172	40
	Negative	12	40

		Actual Values	
	tonal_atonal	Positive	Negative
Predicted Values	Positive	86	90
	Negative	85	3

		Actual Values	
	danceability	Positive	Negative
Predicted Values	Positive	75	10
	Negative	62	117

		Actual Values	
	voice_instrumental	Positive	Negative
Predicted Values	Positive	165	33
	Negative	9	57

		Actual Values	
	gender	Positive	Negative
Predicted Values	Positive	144	66
	Negative	7	47

The vggish model gave me better results. It has better accuracy in 7 out of 11 characteristics. One interesting result is tonality in both models, it looks like I misrepresented what tonality is. You can see when notating objective characteristics like acoustic, electronic and voice/instrumentals both models give a decent score there and for me that is expected. For the more subjective characteristics the accuracy score fluctuates a little bit. I think its a bit tough to get an accurate model for predicting such subjective characteristics, the feeling of sadness and happiness and just mood in general can be pretty different individually I think.