Automated Emotion Recognition in Songs

CSCI 357 AI & Cognitive Science

Christina Yu, Swarup Dhar, Hannah Shin, Lawrence Li
Professor Christopher Dancy

Project Description

This project aims to improve the emotion recognition of music through the training of neural networks with user feedback. The neural network will first train a preliminary set of music from the Image-Music Affective Correspondence (IMAC) dataset, which takes advantage of the Million Song dataset that provides emotion classification for positive, neutral, and negative. Then it will predict the emotion of the user input music and the user will give feedback on the accuracy of the classification. If the prediction does not meet the user's expectations of the song, they can choose to add the music to the training dataset so that the neural network can learn and have a better prediction on how users rate the emotion of the music. In the end, the neural network prediction will be personalized based on the user's subjectivity.

Initiatives

We chose the emotion recognition of music due to our preferences in listening to music to adjust our mood. We use music apps like Spotify, YouTube music, and SoundCloud to find the music we love. During the discussion, it was found interesting that we value the emotion of certain music differently, which drove us to consider a music emotion recognition intelligence as the goal of the project. Through initial research, we found existing research in the relation between human emotion and music, as well as existing machine learning based on the study, and we would like to modify the machine learning process such that each person can have its own neural network prediction model. Based on our initiatives of the problem, instead of training the neural network to recognize music using a general database, we want the users to actually improve the learning by inserting their own dataset into the desired emotional category.

Research Background

Researches indicate that music causes stimulation through specific brain circuits to produce emotions (Verma et.al, 2019). There were several researches focused on the emotion recognition of music and the relationships between human emotion and the music, with machine learning applied to their study. Currently existing machine learnings include music emotion recognition with the extraction of audio features (e.g. Music Information Retrieval Toolbox in Matlab) (Huq et.al, 2010). A lot of research was done about music recommendations on behalf of the user's

mood with the integration of neural network training and learning, and provided a publicity available datasets with download links under attribution non-commercial copyright permissions (Verma et.al, 2019). We will use the IMAC dataset for our neural network training. IMAC dataset is a dataset of music for emotion training by a set of emotion scales: positive, neutral, and negative. Each song corresponds to a score of three numbers with the given emotion classification. Here is one example of the IMAC emotion dataset:

TRAAAJN128F428E437	Welcome To The Pleasuredome	Frankie Goes To Hollywood	1_0_0
TRAAAJO12903CAAC69	Stranger in Paradise	Tony Bennett	0_1_0
TRAAAZF12903CCCF6B	Break My Stride	Matthew Wilder	2_0_0
TRAAAGR128F425B14B	Into The Nightlife	Cyndi Lauper	1_0_0
TRAAAED128E0783FAB	It's About Time	Jamie Cullum	0_2_0
TRAAAYX128F4263BC0	Popular Modern Themes	Chokebore	0_0_1
TRAABIG128F9356C56	Walk the Walk	Poe	1_0_0
TRAABJS128F9325C99	Auburn and Ivory	Beach House	0_1_0
TRAABHO12903D08576	I Know You	Mike Stern	0_1_0
TRAACNS128F14A2DF5	Spanish Grease	Willie Bobo	1_0_0

TRAAAJN128F428E437	AbCEF4sYSiY	SVDC6kPCkWA	IrnvPCSP6Q8	WfHKgcTaU_4	puFcUSfyTJM	jOmHcBVA-BU	yIlK4xjIsIE	pEBXuhqVSws	U7cyVSnlpVM	hBVfb0x2q6U
TRAAAJO12903CAAC69	jzA8gwfJr9I	WFrUsa5SUv0	otJXFz5EN0o	u3qSNHela0c	_TYk-pbciFg	N-0TsjHGIG4	XkYNqqa0JP4	kAG1rXrauaQ	_uyexSA3fjA	yFkirV2NE5c
TRAAAZF12903CCCF6B	cy46iOwWQiE	B4c_SkROzzo	Y4jxylREJCg	x4GZJIZRaAk	gx0lWmerL_s	P5LMuvGFMUA	_k4Bp6hI1aw	#NAME?	wuZlnzETGKE	hHZWTAC_oLA
TRAAAGR128F425B14B	O5UCUusl9jo	Sj0kntF94Oo	vhmvxK9M0RE	5IQL8hcGe2Q	ktwQA6lMuwQ	9fKF_ixdfMU	DKWbLC5qUNA	lFPqhtrlXRE	IafrcJRJWik	sMLT00LlYa0
TRAAAED128E0783FAB	p_tkNQSGV30	eErTL5_3oK4	XQK5CdVar3Y	C55lnH_PK3w	dmqCsXOlobo	flj8ssFn2iw	Hpikb7vqLYs	xOq2QTaZAq0	3f-cqyIKZT4	2Dcv36QRDRw
TRAAAYX128F4263BC0	KBPwK07tH3E	6U0bfn4lTkc	WDHkbkDtiac	bcDloRRPsTc	sw8SC4wOTDM	viYQSkly3P8	KqigSUAAIlo	ZcHVz_YSpXU	GNDp9e1bvao	
TRAABIG128F9356C56	lEEunvd7ZX8	juAU77-Hqkl	fs11WSu4nEA	s0eDLvEFhzI	9hB_R9jWzns	VAajzK728nw	VF_gYfuTie0	KymZ3dlRUA0	7OYO2P6pvSw	Y-flIRs6qCl
TRAABJS128F9325C99	cVpu7OtRUGg	eSy2jbj74Hw	gqjpUDmHt7E	UF2dj6d-wgQ	X2TccA18Psw	iv3lfTfR0Xk	Ays6Ditxg5M	oleDaqs9jLQ	GKlP1i6bNNA	
TRAABHO12903D08576	B5OdOeegiio	vq-VbMw-lkk	6U9ir6LdPZk	SAsXsRKrOY4	45RcfJwds8M	p6wu14Xji4g	HOijJFgBAH4	hAa3ncUrM0M	hlyL6l6cEjw	GQNf-En7yrY
TRAACNS128F14A2DF5	7DgsvfiyUYo	o34TbQOzhkU	x21w1wK4iFQ	e_VhfwpUxLI	ZUIrXl8bTZQ	n9BPvlvmVfo	My3W9PIFSvc	Mt__-9QFgUw		

Figure 1: The music emotion score for each song with corresponding download link

Figure 1 shows the title of the song. The emotion score is given in the right column. A score of 2_0_1 indicates that the song has a positive score of 2, neutral score of 0, and negative score of 1. Every song has a corresponding ID given in the left column, and provides all possible download links for that song file under attribution non-commercial copyright permissions. The music emotion dataset provides a good starting point for neural network training, which can then be used for the implementation of music emotion recognition.

However, individuals have different tastes in music, which suggests that even an identical music piece can have various emotional interpretations by different people. This makes a generic prediction model impractical and indicates a need for personalized, correspondence machine learning designed for individuals (Brownlee, 2020). Hence, we create a neural network that can predict the music emotion based on the IMAC datasets as a starting point and make an option for users to take in their own songs, evaluate the correctness of the prediction, and improve the learning of the neural network. We based the neural network implementation on the tutorial of

neural network training on music genre recognition, which can be modified to recognize music emotion (Vaidya, 2020, 2021).

Implementation

We chose the convolutional neural network model, a deep learning algorithm which can take in input images, assign importance with learnable weights and biases with each image, and be able to differentiate each other. To apply the image prediction with audio we first split each audio into ten 3-seconds audio and then convert each splitted audio to mel spectrograms image, a two-dimensional image representation of any music. Each music has different mel spectrograms so it is achievable and practical to use mel spectrograms to train the network. We build a 5-layer convolutional neural network with corresponding MaxPooling2D layers, one input layer, one dropout layer, and a flatten layer based on a tutorial about training music datasets. The classification of music emotion is based on the IMAC dataset, which are positive, neutral, and negative. About 1000 images of mel spectrograms are trained in 80 epochs with 3 classes as initial training. After the training, the user will input their own music for emotion prediction and will later be evaluated by the user for future learning improvements through a simple guide program. For user input song, it will also be splitted into ten 3-seconds excerpts and convert each to a mel spectrogram. The program will count the total number of occurrences of a prediction of a positive, neutral, or negative and calculate the probability of each emotion class. For instance, positivity: 60%, neutrality: 20%, negativity: 20% with a total of 100%.

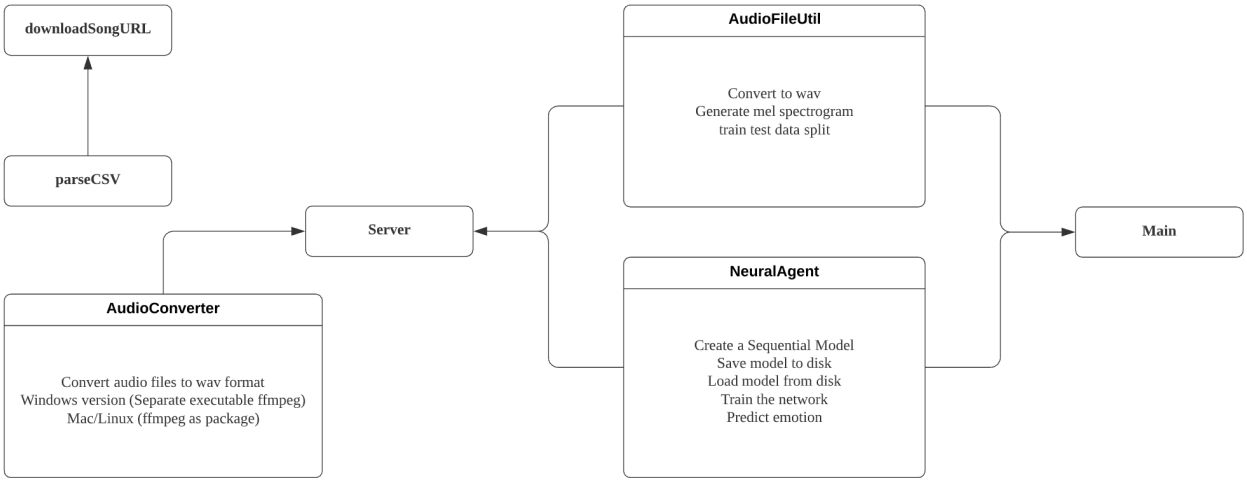

Figure 2: UML class diagram of implementation

Figure 2 shows a simple UML diagram showing how we utilize object-oriented principles to structure the implementation by organizing all methods we might use in the project.

Results

The results of training are satisfactory but nowhere near perfection. Through one training for 80 epochs, the training accuracy is 98.69% but the validation accuracy is only 80.53%. The training loss using cross-entropy function is 0.0382 whereas the validation loss is 0.8013. The results of the training and validation accuracy do not suggest the presence of overfitting issue due to the dropout layer added in the neural network for regularizing overfitting. This will be the starting point of our neural network for personalized learning improvement based on users. When users choose to add their own music to the training list, the program will give an option for retraining the neural network for a better, personalized emotional recognition for them.

The high train accuracy resulted from the use of audio splitting. The project only trained the first 30 seconds of the song and split that 30 seconds excerpt to ten equal 3-seconds parts. The identical training shape makes a high training accuracy. The project also splits train and test data based on a 7:3 scale respectively. The loss diagram of the training is shown in figure 3 below.

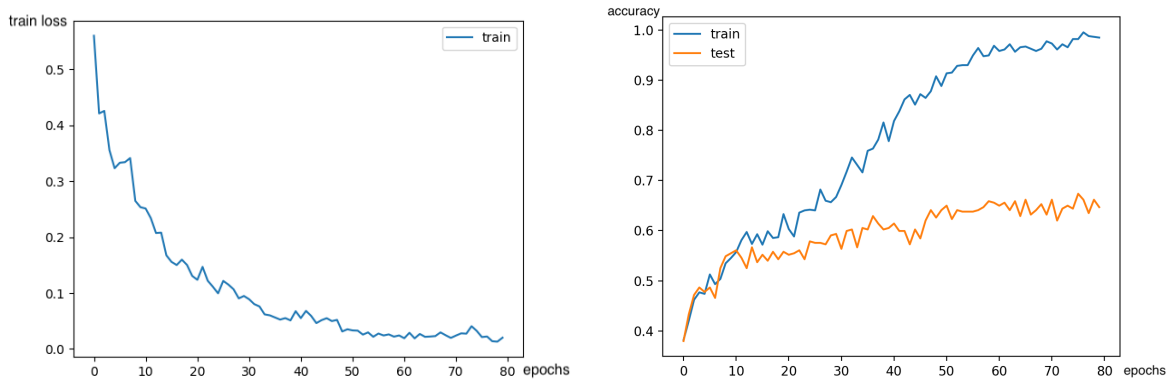

Figure 3: The train accuracy and loss through the first 80 epochs

The train loss graph has an inversely proportional relation with the epochs, which is satisfied in terms of the neural network training accuracy of this project. The train and test (validation)

accuracy deviates from 10 epochs, suggesting a big gap between both accuracies. The following table shows a general prediction of songs from the initial neural network.

Table 1: Example showcase of emotion recognition of various songs

Music name	Positivity (%)	Neutrality (%)	Negativity (%)
Billie Jean	50	30	20
Boku no Sensou	40	40	20
Never Gonna Give You Up	80	0	20

As shown above, the neural agent has the ability to predict probabilities of each emotion class given a song file based on the IMAC dataset. (All songs tested here are not included in IMAC) There are three major downsides for this project. The first is lack of music available for machine training with valid attribution license. The second is the limitation of disk spaces, where generating lots of mel spectrograms for each music piece requires a significant amount of memory. The first two downsides limited the number of music pieces we can use to train from the IMAC dataset. The third one is the time costs for neural training. Since the differences of mel spectrograms between songs are far less identifiable than any other image classifications, it would require a higher resolution image recognition for neural networks, which significantly increase the time cost for training even given a high performance computer.

As for the downsides of user improvement, it would be time costly since the user has to input songs with given emotion one by one and would require retraining the neural network for better accuracy according to user preference.

Conclusion

The general approach to the improvement of machine learning is appropriate as we integrate the user evaluation and feedback where we give the option to put the music into the training set if the prediction does not meet the user's interpretation. However, the approach that we initialize the neural network with a given training dataset from IMAC contradicts our notion that different

people have different emotional interpretations of music, where a general machine learning approach is considered impractical since we cannot decide a ubiquitous emotional interpretation for all people. But at the same time we want the neural network to have a certain prediction ability at the start. We believed that it would be better to initialize the neural network in this way than nothing. The approach for user input music for model training fits most into the central problem of the project for training and predicting personalization for each individual user.

However, although we made it possible for neural networks to have basic emotion recognition, improving such networks by user evaluation could become time consuming and would require lots of user's efforts to input the songs and emotion to the dataset and require retraining, which is a major downside of the approaches adopted in this project.

Another downside of our approaches is the way we approach the song files by YouTube URL, where we downloaded the song from URL which harm the copyright content from the original creator. This approach may be abused by people who just want to download copyrighted songs without permission. The only way we could approach this is to hide all the downloaded files from the public and the user can only play the song from the given program. In other words, users are not allowed to get the wav song file.

References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset

Brownlee, J. (2020, August 27). How to save and load your keras deep learning model. Retrieved March 22, 2021, from https://machinelearningmastery.com/save-load-keras-deep-learning-models/

Huq, A., Bello, J. P., & Rowe, R. (2010). Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, *39*(3), 227-244.

Juthi, J. H., Gomes, A., Bhuiyan, T., & Mahmud, I. (2020). Music emotion recognition with the extraction of audio features using machine learning approaches. In Proceedings of ICETIT 2019 (pp. 318-329). Springer, Cham.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... & Turnbull, D. (2010, August). Music emotion recognition: A state of the art review. In *Proc. ismir* (Vol. 86, pp. 937-952).

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., & Yang, Y. H. (2013, October). 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (pp. 1-6).

Vaidya, K. (2020, December 09). Music genre recognition using convolutional neural networks (CNN) - Part 1. Retrieved March 22, 2021, from https://towardsdatascience.com/music-genre-recognition-using-convolutional-neural-networks-cnn-part-1-212c6b93da76

Vaidya, K. (2021, February 11). Music genre recognition using convolutional neural networks-part 2. Retrieved March 22, 2021, from https://towardsdatascience.com/music-genre-recognition-using-convolutional-neural-networks-part-2-f1cd2d64e983

Verma, G., Dhekane, E. G., & Guha, T. (2019, May). Learning affective correspondence between music and image. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3975-3979). IEEE

Yang, Y. H., & Chen, H. H. (2011). *Music emotion recognition*. CRC Press.