Genre Classification of Album Covers

Minda Chen | Qiumin Zhang
{chen.mind, zhang.qium}@husky.neu.edu
INFO7390, Spring 2018, Northeastern University

ABSTRACT

The topic of this project is to teach a machine to predict the genre of an album, merely taking its cover as input. Inspired by how human beings distinguish albums genres visually, we believe that there is some information hidden among these pixels that somehow represents the genre of the album. To verify that, we decided to apply some commonly used image classification methods (CNN basically), as well as other uncommonly used but reasonable approaches (extensions) to tackle the problem and get a conclusion on the relationship between albums genres and covers.

In image classification part, we conducted a transfer learning by using a pre-trained CNN model, Inception_v3, to implement image classification. The results turned out that the model can reach an accuracy of approximately 30%. While a human being can reach an 40% accuracy on our dataset. However, in intuition, we believe that CNN only recognizes elements in the cover, the real problem is how the combination of these elements define the genre. This is also the reason why we did the extension part.

In extension part, we're going to tag out all elements in every cover and use these tags to train a fully-connected neural network and a SVM model and then apply ANOVA (analysis of variance) on the tags to measure how much influence each element has on certain genre. Detailed results will be listed in the RESULTS part.

1. INTRODUCTION

According to the behavior of playing music, sometimes human pick the music subconsciously by having a quick look at the corresponding album cover. Because human can use intuition to determine if the tracks belong to the genre they are into. What about training machine to have the ability of identifying genre of given album cover? Image classification or recognition method has been widely used and proved very effective in many areas. However, many of those projects tried to distinguish if the image indicates a real object. For example, if the image is a cat or a dog? These objects are very realistic, concrete and vivid. Yet, in art, things become more imaginative, abstract and vague, of which albums covers are great representatives. The album cover is usually composed of both real-life objects and artistic symbols which are always abstract. Human, after being

trained continuously from birth, can easily understand these abstract things using their intuition and make decisions. Can we enable machine to have the intuition? This project will try to make contributions to the answer of this question.

2. DATASET

2.1 SOURCE

The dataset is collected from several main music streaming applications: discogs.com, Spotify and allmusic.com.

2.2 PREPROCESSING

Due to multiple abstract factors of designing an album cover matter, we tried to preprocess and scale the dataset to genre bias. That is, raise the portion of genre taken among factors like album genre, album theme, time of the album and design style. For doing so, tagged album covers that had already classified by genre using Clarifai API[1], ranked the top 10 labels for each genre. As the result, we moved out the images with quite less high-frequency labels in each genre to make sure the dataset is genre bias.

3. CONCLUSION

Approach	Accuracy
Regular CNN	25.0%~34.4%
SVM using tags	39.2%
ANN using tags	39.6%

Figure 3-1 Accuracy of different approaches

Figure 3-1 generally shows the final accuracy of different approaches we applied in this project. Apparently, approaches using tags seem to have a stronger ability to classify album genre by inputting the cover.

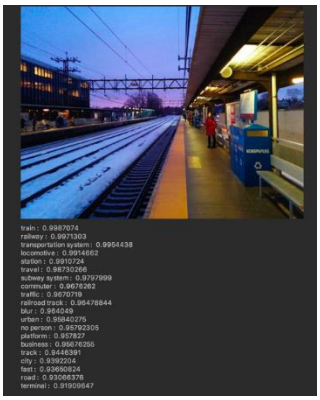
CNN:

30% accuracy of CNN models proved our assumption which is that the cover of an album somehow represents its genre, since it's much better than 14.3% -- the result of a random guess (we have 7 genres in total, so it's 7/100=14.3%). However, there're still some deficiency in the training. For instances,

1. The data set is far less that enough, we don't think the model can critically capture enough features to become robust.
2. The accuracy cannot maintain stable. It might be caused by lack of training. (hardware restricts)

Extension:

We used Google Cloud Vision to tag out elements in each cover. For example,



(the number following the tag is the possibility)

Then, we built a fully-connected neural network which takes the tags of a certain cover (one-hot encoded) as input and predict its genre. After 5 epochs of training, the accuracy of the model levelled out at around 39.6% which seems to be a considerable improvement on the accuracy of the CNN model mentioned above. This support our hypothesis that the genre of an album is, to some extent, defined by the combination of elements in its cover. We also trained an SVM model which reached a similar accuracy of 39.19%.

Next, we applied ANOVA on the tags to see the difference between groups of covers that contain certain label or otherwise, which, in other words, is to measure the influence of certain tag on certain genre. Then we set a threshold of 0.05 for the p value to filter out most related tags for each genre. Here are parts of the results (tags are sorted by p value in ascending order):

Pop: ‘gold', 'danish pastry', 'pc game', 'dolphin', 'iceberg'
Rock: 'tradition', 'bachelorette party', 'temporary tattoo', 'floppy disk', 'floral design'
Hiphop-rap: 'bedtime', 'writing', 'plan', 'pigeons and doves', 'square'
Ambient: 'insect', 'sea anemone', 'theatre', 'musk deer', 'horse harness'
Folk: 'knitting', 'stalactite', 'crêpe', 'rotorcraft', 'glass bottle'
Dubstep: 'crêpe', 'crow', 'stalactite', 'carrack', 'statue'
Jazz: 'planet', 'werewolf', 'swimsuit top', 'wire', 'demon'

Notice that these are just very few elements that exist in my particular dataset. They are NOT ‘general symbols’ of their genres.

4. RESULTS

CNN results are visualized by TensorBoard.

Inception_V3

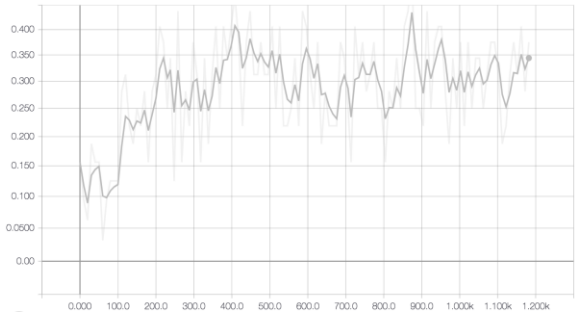


Figure 4-1 Inception_V3 Accuracy
（x: Global Step y: Accuracy）

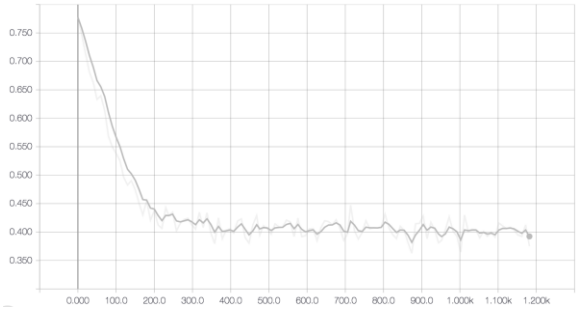


Figure 4-2 Inception_V3 Loss
（x: Global Step y: Loss ）

In **Figure 4-1,** we can see that the accuracy fluctuates a lot even after a long time of training, which is abnormal. This implies that the model is either being overfitting and sensitive to noises or is not complicated enough to capture general features of the input.

In **Figure 4-2**, the loss steadily goes down and then gets stuck at certain value, this is the most ordinary scene of loss in CNN.

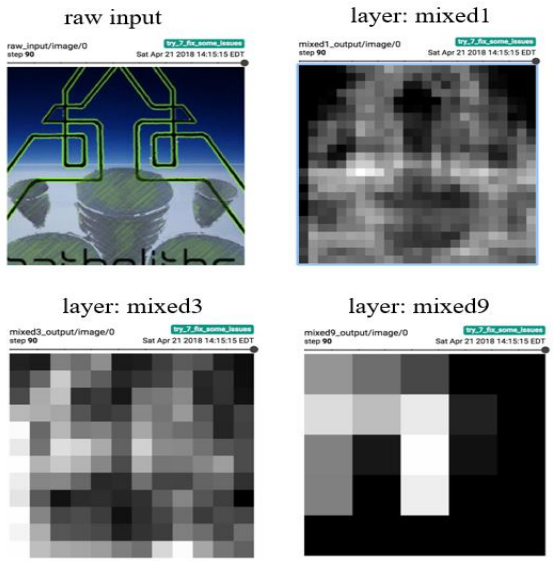
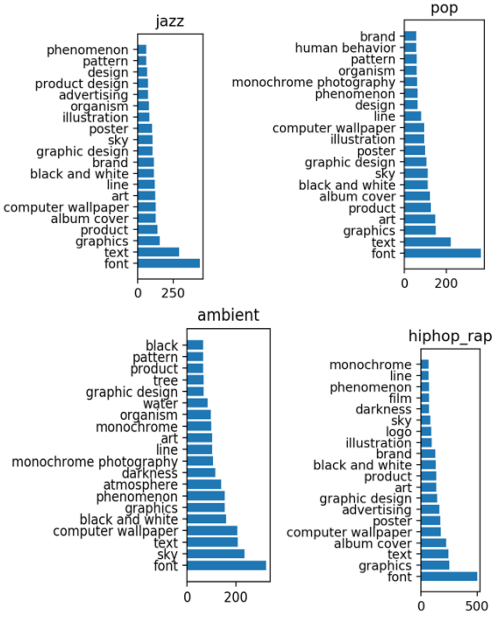


Figure 4-3 Filter Visualization

Figure 4-3 are some intuitively visualization of outputs of certain layers.

Extension:

Google Cloud Vision:



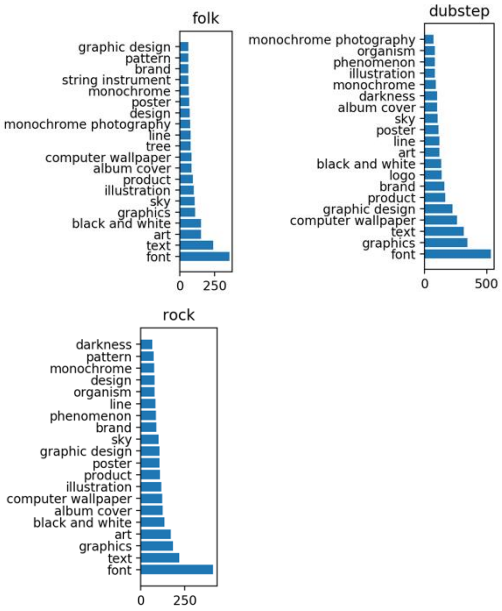


Figure 4-4 Tags Histogram by genre

Though many tags seem to be meaningless, like 'font' or 'text'. However we can still find something interesting in the figure above. Like 'string instrument' in folk, 'water' in ambient, 'human behavior' in pop, these all make sense.

ANOVA:

The entire ANOVA results is shown in appendix.

5. REFERENCES

[1].https://clarifai.com/developer/guide/

[2]. Karen Simonyan, Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition" airXiv preprint arXiv: 1409.1556(2014)

[3]. Christian Szegedy, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[4]. Kaiming He, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

6. CODE LINK

https://github.com/qiuminzhang/INFO7390_Final_Final

7. APPENDIX

	pop	rock	hiphop_rap	ambient	ambient	dubstep	jazz
0	outdoor_shoe	kitten	clock	sofa_bed	viol	viol	male
1	geological_phenomenon	larch	sliced_bread	cupboard	mangaka	car_subwoofer	squash
2	goggles	sky	electronic_device	violone	car_subwoofer	boudin	outdoor_play_equipment
3	road	automotive_exterior	number	supersonic_transport	kitten	mangaka	natural_foods
4	battleship	tower	kitten	kitten	metropolis	cliff	musk_deer
5	mangaka	bicycle_part	postage_stamp	room	bulldog	blazer	tower
6	nose	red_meat	mellophone	estate	cliff	winter_storm	larch
7	love	prophet	living_room	fedora	male	laughter	automotive_exterior
8	musical_theatre	prunus	tartan	dreadlocks	musk_deer	bass_guitar	walkway
9	astronaut	liquid	boxing_glove	mellophone	flock	muroidea	viol
10	larch	mellophone	wood_flooring	cliff	factory	helicopter_rotor	supermodel
11	disciple	sofa_bed	blazer	polar_ice_cap	mellophone	english_billiards	trumpet
12	houseplant	steeple	larch	clock	sky	bayonne_ham	sky
13	small_to_medium_sized_cats	rodent	cliff	bog	squash	sahara	stork
14	fast_food_restaurant	local_food	bass_guitar	orange_lily	material	woody_plant	dawn
15	automotive_exterior	estate	tower	home_fencing	drummer	stained_glass	laughter
16	bread	amusement_park	marine_biology	whiskers	slot_machine	kitten	sailing_ship
17	factory	factory	garden_roses	monument	clock	infant_bed	desert
18	viol	trumpet	middle_ages	liquid	goats	trumpet	sliced_bread
19	facial_expression	pancake	musk_deer	car_subwoofer	church	male	mangaka
20	cliff	fedora	supersonic_transport	jewelry_making	sliced_bread	tarte_flambée	ant
21	black_cat	dreadlocks	tundra	american_food	granite	dance	bicycle_part

22	blazer	polar_ice_cap	presentation	sky	praline	squash	concert_dance
23	granite	supermodel	male	cool	phragmites	crêpe	stone_carving
24	aquatic_plant	garden_roses	viol	mangaka	horse	praline	factory
25	arctic	sawmill	automotive_exterior	amusement_park	geological_phenomenon	crocodilia	bulldog
26	african_chameleon	laughter	mouth	visual_effects	jellyfish	romance	wing
27	tower	guitarist	local_food	boudin	larch	facial_expression	violone
28	traffic	orange_lily	bass_drum	infant_bed	prunus	musk_deer	music
29	boudin	monument	structure	bicycle_part	outdoor_play_equipment	supermodel	business_executive
30	canis_lupus_tundrarum	goggles	orange_lily	prunus	subshrub	tundra	percussionist
31	prairie	bog	sofa_bed	embroidery	crocodilia	timbales	bus
32	spruce	music	room	old_world_flycatcher	blazer	vehicle	number
33	presentation	sahara	crocodilia	computer_disk	boxing_glove	stork	rodent
34	pollution	system	cupboard	system	boudin	metropolis	red_meat
35	mouth	mouth	singer_songwriter	gadget	timbales	electronic_device	
36	local_food	old_world_flycatcher	bulldog	factory	mouth	bicycle_part	
37	polar_ice_cap	computer_terminal	american_food	shoe	indian_elephant	walkway	
38	afro	cupboard	cowboy	music	singer_songwriter	jewelry_making	
39	mythology	american_fod	boudin	squash	afro	windshield	
40		embroidery	plant_community	sliced_bread	sahara	bulldog	
41		ant	crêpe	wing	stone_carving	tiple	
42		arctic	metropolis	bear	swan	graphics	
43		stained_glass	protective_equipment_in_gridiron_football	mane	temple	swan	
44		percussionist	prunus	bicycle_saddle	residential_area	boxing_glove	
45		tarte_flambée	hard_dough_bread	media_player	institution	supersonic_transport	
46		helicopter_rotor	phragmites	viol	muroidea	geological_phenomenon	
47		boxing_glove	cheetah	blazer	stained_glass	horse	

48		viol	bus	tundra	accordion	tomato	
49		dance	indian_elephant	street_light	infant	shack	
50		pizza_cheese	tank	herd	bicycle_wheel	indian_elephant	
51		supersonic_transport	flautist	graphics	radiography		
52		cliff	squash	lingerie_top	jewelry_making		
53		tiple	battleship	windshield	snapshot		
54		middle_ages	stony_coral	presentation	violone		
55		room	subshrub	tartan	fluvial_landforms_of_streams		
56				electronic_device	liquid_bubble		
57				crêpe	flightless_bird		
58				african_chameleon	major_appliance		
59				pizza_cheese	red_meat		
60				musk_deer			
61				marine_biology			
62				dance			
63				male			
64				helicopter_rotor			
65				uniform			
66				garden_roses			
67				muroidea			
68				liquid_bubble			
69				church			
70				romance			
71				desert			
72				green_algae			
73				trumpet			

74				screen			
75				seahorse			
76				pancake			
77				number			