Movies and Pop Songs Recommendation System by Emotion Detection through Facial Recognition

Movies and Pop Songs Recommendation System by Emotion Detection through Facial Recognition

Jingye Zhang

Shanghai Starriver Bilingual School, Shanghai, China

hzhc_wjdi@163.com

Abstract. In this paper, a convenient, accurate, and widely applicable recommendation system of films and pop songs, using deep learning as an efficacious tool, is propounded. We devised a refined Deep Residual Network (ResNet-38) for the emotion detection of the users, which achieves an accuracy of 64.02% for the testing set of Kaggle-fer2013. Other traditional methods including the use of Classifier SVM or four-layer CNN produce the average accuracies of 40.7% and 60.7% respectively. Thus, it is cogent to conclude that our model outperform other traditional models. The usual emotion-based movies or songs Recommendation systems include web scrawling of the user's personal information, like his or her recent comments, based on which researchers construct "interest models" to achieve an understanding of the user's emotion. However, this type of method, though comprehensive, infringes the users' privacy, presents biased results, and lacks instantaneity and interactivity. Thus, this research paper serves to introduce a novel mode of Recommendation system so as to counter theses drawbacks.

1. Introduction

The investigation in Recommendation System has been around for years, and this area has evolved considerably over the past decade. Nevertheless, these is still limited utilization of user's emotion recognition in Recommendation System; even if there is, there exist different kinds of shortcomings. To specify, flaws of the previous emotion-based films and songs are enumerated above along with the solutions provided by the system provided by this paper.

2. Materials

2.1. Kaggle-fer2013 dataset

The facial expression database for training this ResNet-38 is the fer-2013 public dataset from the Kaggle website. Used for the Kaggle facial expression recognition challenge, this dataset consists of 28,709 examples for the training set and 3,589 examples for the test set. The final test set used for determining the winner of the competition contains another 3,589 examples. All the example images were derived from Google's face recognition API, later undergone Boundary-Processing and de-duplication, and eventually trimmed to 48*48 gray-scale images.

Table 1. Categories and Distributions of Fer-2013 Dataset

Number	Emotion	Amount in Dataset
0	Anger	3,995
1	Disgust	436
2	Fear	4,097
3	Happy	7,215
4	Sad	4,830
5	Surprise	3,171
6	Neutral	4,965

2.2. Deep Residual Network

The ResNet Structures is invented to solve the degradation problem and vanishing/exploding gradient. According to Kaiming He and his colleagues' paper, facing the problem that a deeper network sometimes is unable to perform as well as a shallower network is on the training set because of the phenomenon that when the network depth increase, accuracy of the outputs inevitably saturates or even decreases, they created Residual Network based on the foundation of VGG19, which is modified by adding residual units through shortcut mechanism, structure of which is shown in the Fig.1.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 1. Structure of RestNet

For a stack structure, when the input is x, the feature it learns is H (x). Now ResNet makes it learn the residual F (x) = H (x) - x, so the original learning feature will become F (x) +x. The reason for this is that residual learning of original features is easier for machine training than direct learning of them. When the residual is zero, the stack only does identity mapping at that specific moment time, so at least the network performance will not decline. In fact, it is impossible for the residual to be 0, which will also enable the stack to learn new features based on the input characteristics, thus having better performance. The graphical representation of the structure of residual learning is shown in the Fig.2.

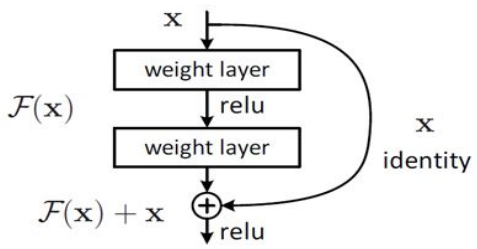

Figure 2. Shortcut mechanism

2.3. IMDB Movies Dataset

IMDb (Internet Movie Database) is an online data base of information related to films, television programs, home videos, and video games. The content of IMDb includes a great amount of information of the film, actor, length, plot introduction, classification, and comments. IMDb has collected date on 4,734,693 works and 8,702,001 characters so far.

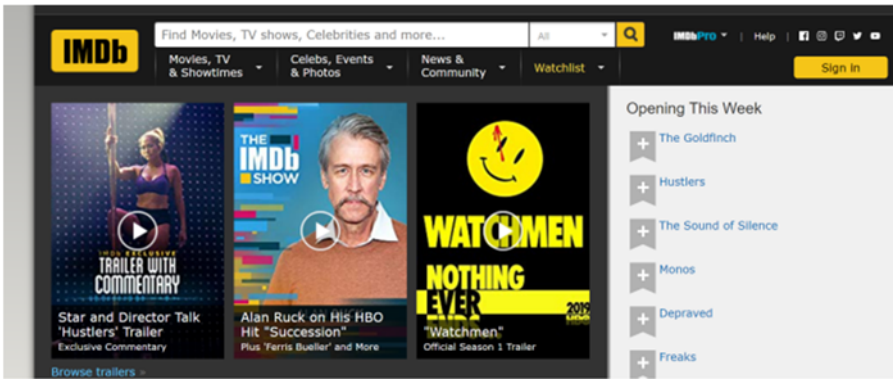

Figure 3. Home page of IMDb

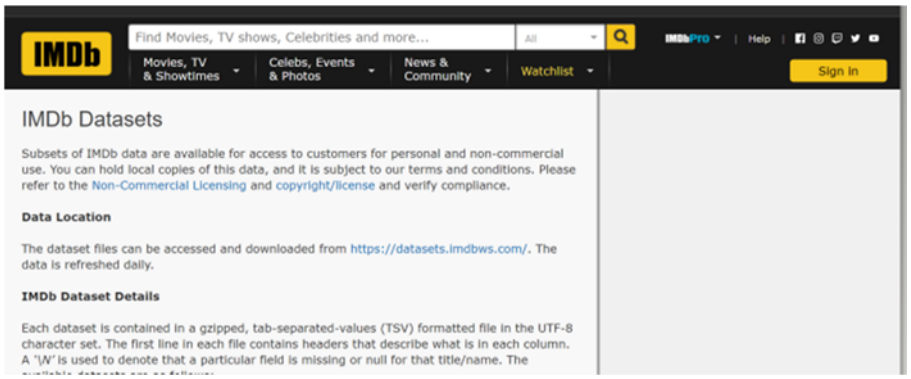

Figure 4. Interfaces of IMDb Dataset

3. Experiment Design

3.1. ResNet-38 for facial expression recognition

The model employed here is a self-constructed Deep Residual Network, the detailed structure of which is displayed in Fig.6. First, the data go through the Input Layer and then two Convolution Layers for the feature extraction of user's countenance. Each Convolution Layer is followed by a Batch-Normalization Layer intending to guarantee the data identically distributed. The activation function employed here is Relu. Next, there are four Residual Structures. As the depth of the model increases, these serve to protect the data error from amplifying. Each Residual Structure contain two lines: for the first line, it consists of the following layers in sequence: Seperable_Conv2D, Batch-Normalizaton, Activation (Relu), Seperable_Conv2D, Batch-Normalizaton, and finally Max_Pooling2D; for the second layer, it contains a Convolutional Layer and Batch-Normalization Layer. These two lines converge to a Add Layer, which ensures the error doesn't influence the functioning of the model, as previously mentioned in Materials $f(x) = F(x) + X$. The Last part contains a Convolution Layer, a GlobalAveragePooling layer, and a final Activation Layer(Softmax), which outputs a the probability distribution of a person's facial expressions and take the expression with the highest probability score as the final emotional category of the image. The parameters of our network and the output tensor of each layer are shown in fig.7. Our input layer receives a 48*48 gray scale image, which is processed by ResNet-38, the output is a one-dimensional

vector of size 7. The total parameters of the model are 58,423(Trainable parameters 56,951 and Non-trainable parameters 1,472).

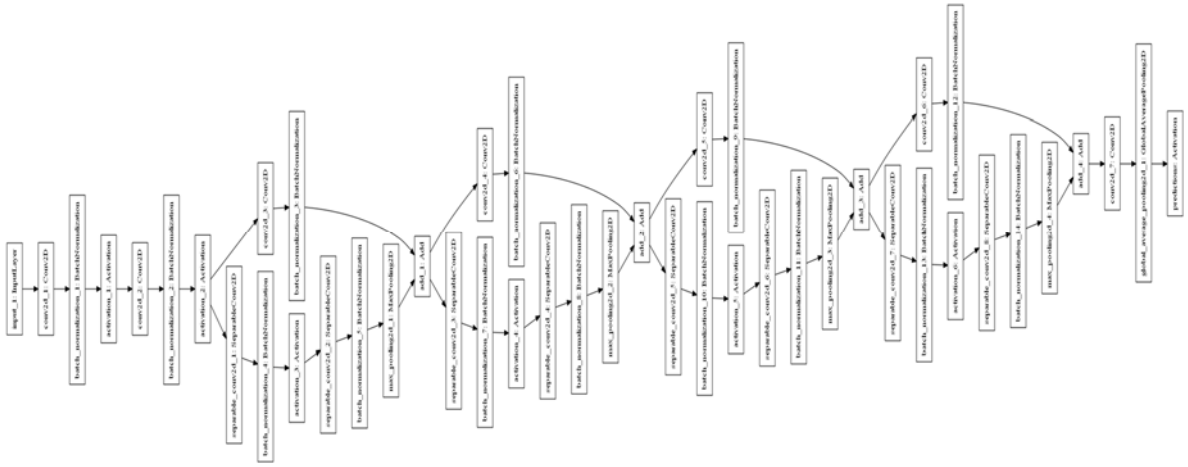

Figure 5. Structure of ResNet-38

3.2. Movie Recommendation System

After ResNet-38 yields the result, the real-time emotion of the user, the emotion tag will serve as an index for recommendation use, as shown in the form 2. This strategy builds a one-to-one relationships between user's emotion and movies' genre. To get the perdoctions of movies that the user wants to see, we first go through a local set of txt files which include categorized film names. The program will randomly select three from the targeted list. Secondly, we use the crawler to find another list of recommended genre of movie from www.imbd.com. These two parts construct a complete list of recommended movies.

4. Results

4.1. Train Process

Based on the accuracy graph shown in Fig.6, there is a detectable trend that the overall accuracy of the training accuracy and testing accuracy increases as the epoch increases. To specify, from epoch 0 to epoch to 20, the accuracy of the training set and testing test increase together in a relatively rapid speed. Although the latter fluctuates more severely, they all stabilize around the accuracy 62.5%, and experience a minor increase afterwards.

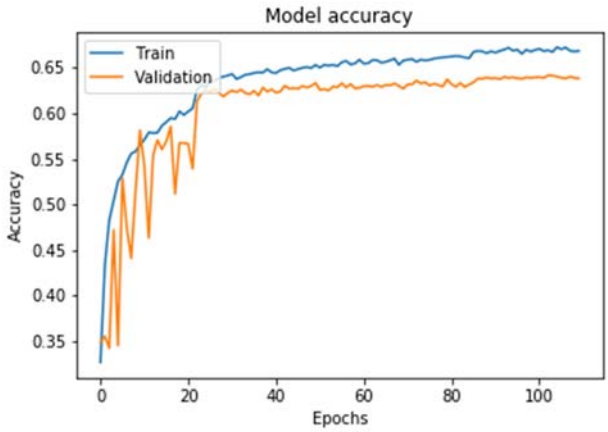

Figure 6. The accuracy graph for training set and testing set

For the loss graph, as shown in Fig.7, the loss of training set and testing set tends to decrease as the epoch increase. To elaborate, the loss of both sets decreases in a comparatively fast rate from epoch 0 to 20. While the blue line, which stands for the loss of the training set, displays a curved feature, the yellow line, which stands for the loss for the loss of the testing set, undulates strongly with a decreasing trend. They all stabilize around epoch 20 and experience a minor decrease in the remaining epochs.

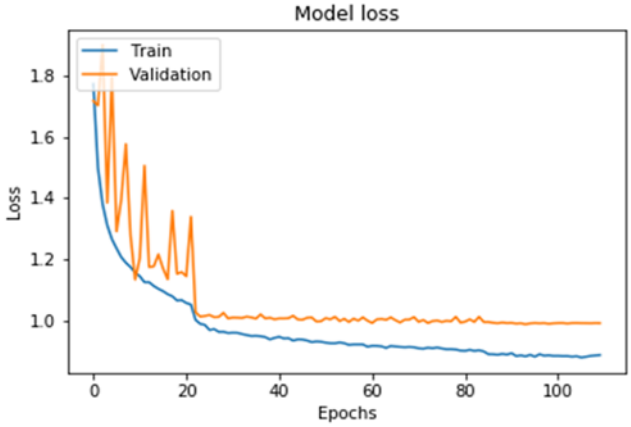

Figure 7. The loss graph for training set and testing set

4.2. Model Performances Comparison

Compared to other deep learning models for facial expression recognition on the Kaggle-fer2013 dataset, our ResNet-38 stands out with better performance. The detailed comparison is shown in the Table 2.

Table 2. Different Methods Compared to our ResNet-38

Method	*Anger*	*Disgust*	*Fear*	*Happy*	*Sad*	*Surprise*	*Neutral*	*Average Accuracy*
SVM	*57.7%*	*12.6%*	*23.0%*	*67.0%*	*44.0%*	*56.0%*	*43.0%*	*40.7%*
CNN	*60.0%*	*21.0%*	*33.0%*	*81.0%*	*52.0%*	*74.0%*	*54.0%*	*60.7%*
ResNet-38	*64.5%*	*23.0%*	*36.0%*	*80.0%*	*58.0%*	*77.0%*	*52.7%*	*64.15%*

As shown above, the results of SVM and CNN are from Xinhui Song's Master Dissertation, "Facial expression recognition based on deep learning". To contextualize the comparison, the first method is a traditional one that first extracts the Gabor Wavelet feature of the face and later uses SVM to classify the images. The second method utilizes Convolutional Neural Networks. It structure is Conv(3,3,32)-MaxPool(2,2)-Conv(3,3,64)-MaxPool(2,2)-Conv(3,3,96)-MaxPool(2,2)-Conv(3,3,128)-MaxPool(2,2)-FC(1,200)-FC(1,7). Based on the content of the form Ⅴ, ResNet-38 generally has a higher accuracy for different kinds of emotion and higher average accuracy, so it is reasonable to conclude ResNet-38's superiority over other models.

5. Conclusion

This research incorporates facial expression recognition to construct a Movie or Music recommendation. Through the utilization of deep residual network, the accuracy of this model achieves drastic increase. The implementation of OpenCV creates a convenient access for real-time interaction between user and the computer. User's emotion here is deemed as an index, later paired with a specific genre of movie/music, for web crawling. In this way, not only the possible intrusion into the user's privacy is averted, but also our recommendation system can present names of movie or music that potentially cater to the user's taste in real life situations to which conventional recommendation system cannot apply.

In future research, still many extensions of our research can possibly be realized. Firstly, the accuracy of this model (ResNet-38) have space to increase, basically in two aspects. Because of the limited size

of Kaggle-fer2013 dataset, Generative Adversarial Network can be employed to enlarge both the training set and testing set for better adjustments of the model. Secondly, various possible modules can be added to preprocess the images. In reality, many factors like lightning, angle, and occlusions may affect the input of our model. Therefore, researchers can further make use of preprocessing modules to catch the feature of the user's face more accurately. 3D-modeling may be applied to reconstruct a complete and clear human face based on a image where part of its information is lost in certain situations. Thirdly, since it is a real-time emotion recognition, besides facial expressions, other movements like body language can be utilized to yield a combined result, which achieves better comprehensiveness.

References

[1] Xinhui Song, "Facial Expression Recognition based on Deep Learning", January 2017
[2] Wu Shihao, "Research of Deep Learning For Facial Expression Recognition," April 2018
[3] Li Jiang, "The research and implementation of facial expression recognition based on deep learning," May 2017
[4] Guo Fangliang, "Research on Video Facial micro-expression Recognition Based on Deep Learning", January 2017
[5] Luo Xiangyun, Zhou Xiaohui, Fu Kefu, "Facial Expression Recognition Based on Deep Learning," December 2016
[6] Lei Ming, Zhu Ming, "Application of sentiment analysis in movie recommendation system," February 2017
[7] Alicia Esquivias Román, "Recommendations by Emotions Detection through Facial Recognition", August 2017
[8] Xia Mingxing, "Comments polarity classification based on emotion analysis and movie recommender system design and implementation," March 2016