Moods in Classical Paintings: An AI based Classification Approach

Shuqiao Huo*
Glenelg Country School, Maryland, USA

*Corresponding author: hzhc_wjdi@163.com

Abstract. According to WHO, globally there are at least 2.2 billion people who have a vision impairment or blindness. While they are known as the disabled in our society, rarely have they benefited from the development of technology and introduction of new products. That is how I developed this idea of constructing a medium that would transfer visual information to audio forms, which would help this disadvantaged group gain access to “visual experience”. In order to achieve this goal, several essential problems should be solved: how can one establish a technological formula to convert visual input to audio output? Extracting visual features from paintings, I could set up my own data collection based on some core features, which can capture the mood of a painting. I then use Convolutional Neural Network to train the AI to link visual and audio information. Loss functions are introduced to minimize the error of the product. After I obtain the mood of my pictures, I will use it to formulate a definitive sound, in the process of which the “mood” gets preserved, such as joy, sorrow, or any artistic theme that is shared both in visual and audio experience. Towards the end of this paper, several experiments will be introduced to test the efficacy of my project. The first part is on the functional efficacy of my AI system. The second part is about my product hypothesis: I will give to two distinct participants one pair of visual input and audial output. Then I will check whether they can recognize the exact theme inherent in the corresponding media. Their successful identifying the same “mood” testifies the success of the project goal of transferring visual experience to audio.

1. Introduction
There are many disabled people who have vision problems, and this physical deficiency severely interferes their lives. In order to help these group of people with their special needs, and to better help them to sense and receive visual information, an AI inspired device can help these people to appreciate various emotions that the paintings are expressing.

In recent years, researchers learn the nervous system within the human body, and have tried to imitate and apply the basic logic to solve physical problems. This logic helps with the development of deep learning which was efficiently used to enhance the classification of pictures. The appearance of Vgg, GoogleNet, ResNet [1], and other AI algorithms lays a solid foundation for the future development of AI algorithms that can help solving various visual problems. Vgg network is implemented by Visual Computing Group in Oxford, which aims to achieve better prediction performance on ImageNet, the biggest image database at that time [2]. However, it limits the
architecture of the neural network to the linear version, which hinders the training process. The GoogleNet is developed under the inspiration of networks in networks [3]. It proposed a much more sophisticated structure, which broaden the width of the neural network. It also suggests that small kernel could replace larger kernel without losing accuracy. The ResNet is developed by research organization of Microsoft, which greatly increase the depth of the neural network and provides design paradigm for future [4]. The fundamental principles behind the neural network is using a convolutional neural network (CNN) to detect features or patterns and learns to classify a massive number of pictures. Although there are many institutions, schools, or volunteers that are trying to help blind people, a computer-based system can be more approachable and applicable to the minority groups in their daily life.

2. Materials
The training process of the proposed system needs plenty of pictures with their classified category. The pictures that are used in the training process are chosen from online collections of paintings, illustrations, portraits. After the gathering of pictures, they are divided into different categories with subjective cognition of their emotions. For example, some pictures about glorious sceneries with low-saturated colors can be classified into “peaceful” pictures, pictures with dark colors and strange patterns can be classified into “horrifying” pictures, etc. All pictures are then cropped with a dimension of 300 x 300 and can be used for the training process.

3. Overview

3.1. Training Process based on ResNet
The core of the training process is divided into several stages. Firstly, pictures are classified and organized in different categories, and all pictures are inputted in the process that assembles the dataset. All pictures are cropped with a dimension of 300 x 300 and changed into a standard file type of jpg. Next, in order to build a model, torchvision is introduced and executed by code. Then, by calculating the loss, a loss function is generated to show the disparity between the real data with the ideal data. Lastly, an optimizer is used to optimize the result and reduce the loss.

The basic logic of code is to continuously set up dataset, generate output, compare the result, and update the model.

In the initial phase of programming, pictures are collected to form a dataset, and all data are inputted into different stages. Pictures go into multiple convolutional layer to experience convolution operation. Each picture has specific RGB values, height, width and depth, and each convolutional layer (convolutional kernel) would put more weights on it to output a more abstract image with specific features.

The mathematical principle behind the convolutional operation is the operation of matrix. By calculating the image with a matrix with random number, the output would finally generate an image with smaller width and length but with more height.

Each convolutional kernel has a specific dimension of width and length, and it would go through the mathematical principle with the data on pictures with same dimension. Similar to the operation of neurons in human body, the input pictures go through multiple layers and finally generate an output.

During the convolutional process, pooling is needed to reduce the dimension of inputted data. This process collects the information in a specific range, and average or take the maximum of the data.

4. Method
In order to construct a relationship between the original or the inputted data with its classified category, the core ideas involves in a complicated training process. The whole training process is divided into several stages: data collection and modification, construction of model, application of loss function, and application of optimizer. The experiment is a repetitive circulation of the convolution, calculation
for loss, and improvement for parameters. The picture below shows the core steps of the experiment, and the arrow represents the direction that the data is processed.

![Backpropagation Diagram]

**Figure 1** The whole process of backpropagation

### 4.1. Data

Firstly, data need to be collected and modified into a congruent format. All data are collected from online open sources as jpg files, and they are all painting, portraits, and pictures. Pictures are classified into different categories with their corresponding characteristics of emotions. The classification process is judged from subject point of view, but the standard is common and understandable. Next, when using pictures, all data needed to be decoded from jpg format and stored in internal storage. By creating a txt file that displays the file path in computer, file name, and category, data are more clearly organized and could be used in the next step. All pictures are cropped into a dimension of 300 x 300 pixels.

During the classification stage, there are mainly four categories that differentiate the emotion of pictures. In general, pictures are weather happy, sad, peaceful or horrifying. Under each category, there are more than 50 photos.

Pictures should also be organized and inputted in batches. The batch size that the training process involves is 32 pictures. It could more efficiently use the computer to accelerate the computing speed. In a batch, data will not interdependent. They will not affect each other. To display the abilities of computer hardware to the full.

At the initial stage, Dataloader is used to extract data from the dataset, and batch size defines how many data would be extracted every time. The batch size for this experiment is 32, which is an appropriate size for the process. However, if data is extracted in an arranged order, the result of the experiment would be affected to an extent. In order to prevent the randomness of the result and improve its reliability, data would be shuffled and extracted randomly.

### 4.2. Model

The construction and training of the model involves in convolution, pooling, and other steps that are commonly used when constructing a neural network.
In the model, when calculating, all parameters are randomly generated. In the beginning all parameters have no meaning and value. Those numbers only define what calculation is needed and will be modified. The basic principle of the process is to train the parameters and generate a new set of parameters that can connect the input data with the desired output. After the training process, those parameters could make sense.

In this model, it uses a ResNet structure. ResNet is one of the most often used neural structure, and its basic elements involves in convolution, pooling, etc.

The core principle of ResNet is displayed in the picture below:

![Figure 2 The basic architecture of ResNet](image)

Moreover, relu is an activation function, and its basic features are shown in picture below:

![Figure 3. The figure of Relu function](image)
In the picture above, when x (input value) is nonzero, the output would be its original value; else, it would be zero.

The classical model of ResNet is embedded in the torchvision library. The downloading of the torchvision would benefit the whole experiment. There are many kinds of ResNet structures. In this experiment, ResNet 18 is used. As one of many parameters inside ResNet 18, the “pretrained=true” is set at the initial stage to speed up the training process. The disadvantage of the pretrained parameter is that the model structure has to be fixed, but in this case, a fixed structure would not affect the overall training process.

4.3. Loss Function
Loss function is introduced in the process to compare the input and output and generate the difference. It works like a simple control system which detects the difference and would help minimize it. For example, an incubator would detect the current temperature within, calculate the difference between current temperature with the desired one, and regulate the temperature to the desired value. By knowing the difference, the use of loss function could better help to improve the system and generate a better relationship.

The loss function that is used in this experiment is cross entropy loss which is often used in tasks that involve with classification.

4.4. Optimizer
An optimizer is also introduced in the process. It would cooperate with loss function to improve the model. An optimizer defines clearly how to use the loss function to advance the model. It will use the calculated difference to update the parameters used in the model.

The optimizer that is used in this experiment is SGD (Stochastic gradient descent). SGD is a basic optimizer that could compare the output result with the label and update related parameters. For the parameters of the optimizer, the learning rate is set to 0.001, and momentum is 0.9.

4.5. Overall
The complete pathway of the data would go through calculation using the model, comparison of the difference using the loss function, and upgrade the parameters of the model by analyzing the difference detected by the loss function. The whole system involves in the continuous circulation of all steps.

The parameters of the model would frequently be modified and would gradually be more meaningful. When the difference calculated by the loss function is really small, and most data could be outputted with a desired result, the training process of the model is completed. Since the loss function compares the difference between the experiment and the ideal data, when the value calculated by the loss function is small, the output generated for each input data is close to its real label. Under this circumstance, the model is fully constructed and useful.

5. Experiments

5.1. Resources
The training and testing are implemented under MacOS with multi-core CPUs.

In the overall experiment, PyTorch is used as an open source machine learning library to further the deep learning process. In details, incorporated by PyTorch, torchvision is introduced to the system to process pictures. Since torchvision covers many classical models that can be used for deep learning studies, it is important to be imported. Other modules, like numpy, sklearn and PIL are also used to help the image preprocessing.
5.2. Results

In the graph above, the relationship of epoch and loss is represented. The proposed trendline shows the result that the value of loss is gradually decreasing as the number for epoch increases. The loss curve would tend to converge to the x axis after 300 epochs. The convergence of the loss indicates that the model gradually fits to the interested mapping relationship.

**Figure 4** The trends of loss during training

| optimizer | Accuracy | Training epoch |
|-----------|----------|----------------|
| SGD       | 87.16%   | 450            |
| Ada       | 91.08%   | 300            |
| Adam      | 93.7%    | 300            |

The SGD algorithm is less efficient than other two, which takes longer training and converging time. Ada and Adam uses carefully designed learning rate adjusting algorithm, so the whole training process time decreased and the final accuracy improved.

6. Conclusion

In this project, a novel algorithm to identify the moods expressed in classical paintings, based on neural network is proposed and achieved promising results of 93.7% accuracy. The model architecture utilized in this project is ResNet, which greatly increase the depth of the neural network, thus the increasing of the classifying performance. All the data are collected and labeled by experts in related fields, to make the labeling process more and more subjective. The next step of this work is to migrate the whole method into a vaster area, including the abstract paintings of paintings of modernism.
References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *computer vision and pattern recognition.

[2] Simonyan, K., & Zisserman, A. (2014). VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. *computer vision and pattern recognition.

[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. *computer vision and pattern recognition.

[4] Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Mao, M. Z., ... & Ng, A. Y. (2012). Large Scale Distributed Deep Networks. *neural information processing systems.*