Predicting emotion ratings from color statistics of images

Claire Mallela and Clayton Greenberg

Abstract

Which features in art affect emotional reactions? In order to answer this question, we took a variety of abstract images and included people's emotional ratings for each image, for eight different emotions. We utilized this data to see how visual elements affect our everyday lives by engineering features of images and running that through a variety of neural networks. Out of all the networks we ran our data through, the linear regression model performed the best with a mean average error of around 1.4. All the other neural networks, including a convolutional neural network, had a mean average error of at least 1.7. The full-size linear regression model with all engineered features worked best in predicting the emotional values associated with each image. That is, all specific features were significant and we did not find meaningful differences among the models for different emotion ratings.

1. Introduction

Our research question is: what features about art affect emotions? In order to answer this question, we used a dataset that was composed of two parts: a set of abstract images – in color and with different subjects – and a set of eight ratings for different emotions for each image. This type of problem was working with labeled data, taking one set and attempting to predict something based on the other, therefore it was a supervised regression model. The image data was numerical when broken down to its essentials – though it also had visual elements – and the emotion data was numerical as well. We also isolated features from our data, such as overall color values and brightness. Our data output was, essentially, a list of predicted ratings for "amusement", "anger", "awe", "content", "disgust", "excitement", "fear", and "sad" on a scale from 0 to 13.

2. Background

Combinations of subject matter and style affect how people view art. More specifically, it affects their emotions. Kemp and Cupchik (2007) found that positive subjects in art evoked pleasant feelings, while negative subjects required thought and challenge to process. The painting's style, defined as a set of traits shared by one group, artist, or era, also described the structure and expression of an art piece. Emotions people feel while looking at art are subjective and depend on the stimulus they receive (and also how they view it).

The relationships people have with art may vary depending on technical or aesthetic standing and expertise, but emotional aspects of art appreciation are consistent throughout people despite skill or expertise level (van Paasschen, Bacci, and Melcher, 2015). Observers with no training in art awarded high valence and arousal ratings to abstract art. Valence and arousal can be related to visual features of art – line, shape, color – consistent with Chatterjee's model of aesthetics.

Overall, previous research indicates that the emotional reaction to art pieces can be influenced by a combination of line, shape, color, and object, and a person's history and thought process.

3. Dataset

The data we were working with were divided into two main sets: one dataset of 280 colored images with varying sizes, and another corresponding dataset with 8 emotions – “amusement”, “anger”, “awe”, “content”, “disgust”, “excitement”, “fear”, and “sad” – and their specific values ascribed per image. These values per emotion ranged from 0 to 13. Each image in isolation was rated by some number of participants.

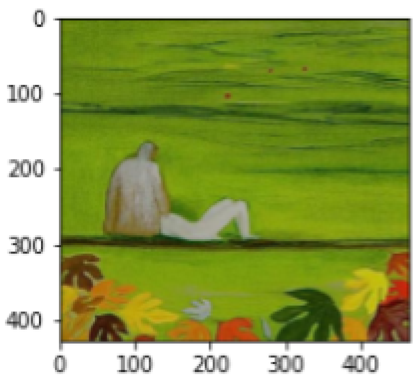

Fig. 1: Example of an image from the dataset: IMG 71, with ratings “content” [10], “excitement” [2], “sad” [1], and “amusement”, “awe”, “disgust”, “fear”, and “anger” [0].

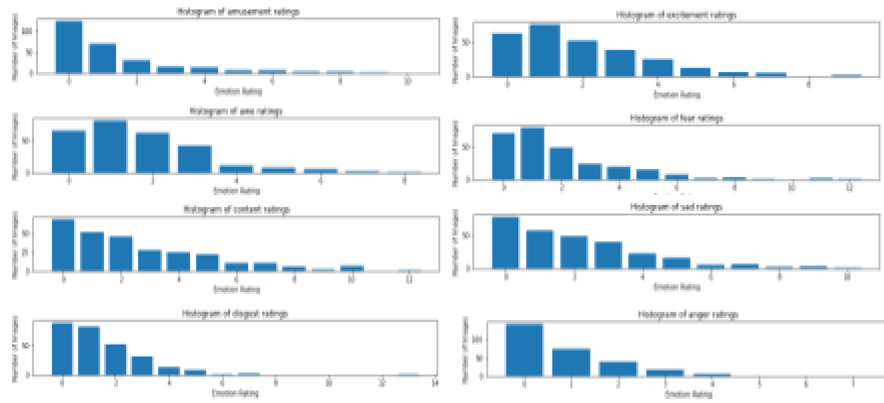

Fig 2. Histogram of ratings for the eight emotions in the image dataset: amusement, awe, content, disgust, excitement, fear, sad, and anger. Ranges varied across emotional categories, but the highest concentration of ratings were in the lower numbers 0-2.

With the image data, we could go deeper into the dataset to identify each individual rgb value for every pixel in the image. Our image data was numerical, with lists for rgb within the lists for the image. However, our label data was in both strings and integers, with the emotions labeled per each image and the corresponding ratings given in integer format.

We analyzed our image data by creating histograms of the ratings for each emotion to see what the distribution was. We also calculated the average rating for each emotion for all the datasets, which fell between 0 and 2. According to our dataset, “anger” had the lowest average ratings, while “sadness” had the highest – but they were, again, relatively evenly distributed. We also calculated the standard deviation of the emotion ratings across all the images, and the values fell between 1 and 2.

4. Methodology/Models

In order to answer our research question, we first had to engineer features each image had that could affect the emotion values ascribed to it. We innovated nine features to include in our models: average rgb values for red, orange, yellow, green, cyan, blue and purple throughout the entire image, as well as average brightness and standard deviation of brightness. To do this, we combined the rgb values using the proportion definitions of the feature colors, i.e. red is (1,0,0) and orange is (⅔,⅓,0). To emphasize, these features were calculated globally across the image. That is, there was one yellow value for the entire image. We used the grayscale values of the image to determine the average brightness, and then we calculated the standard deviation of these brightness values. We repeated this process for each of the 280 images in the dataset.

We used these features for prediction (independent variables). The dependent variables were the emotion ratings. We split the data into training and testing data at an 80%/20% ratio, and then we trained a linear regressor, a random forest, a ridge regressor, a decision tree regressor, and a CNN.

In the linear regression model, we simply attempted to predict emotion values for each image using only the average color and brightness values we determined earlier; we imported the linear regression function from a sklearn library and fit a model to the aforementioned data. In order to calculate the prediction error, we also imported the mean absolute error function from sklearn. By inputting both the predicted values from our linear regression model and our actual values we were able to determine the mean absolute error of our model. We followed the same process with the random forest, ridge regressor, and the decision tree models. We also imported each model from the sklearn library, and applied the same method to determine mean absolute error in order to see how accurate our model was. As an extra step, we also ran the image dataset through multiple loops of the linear regression model, ablating out one feature at a time, to identify any features that may not be contributing significantly to the models.

We ran three sets of data through our CNN: the unedited images themselves and two different resized versions. We used two methods of modifying the images; first, we cropped the images to a desired length and height – these values were the shortest length and the shortest height of all the images in the dataset. The other method was resizing the images to the largest length and height of all the images in the dataset. We filled the extra padding on the images with the average rgb value of all the pixels in the images so as to not skew the color values of each image. Once we prepared our training data, we ran it through the CNN and predicted the accuracy using the same methods used for the regression models.

5. Results and Discussion

The linear regression model consistently gave a mean absolute error value of 1.4. The other baseline models we ran our training data through – a random forest model, a ridge regressor model, and a decision tree regressor model – consistently gave mean absolute error values of averages of 1.5, 1.45, and 1.9 respectively. None of these regression models worked as well as the original linear regression model, with all the original features we determined.

With the linear regression model performing the best, we turned our attention back to the features that we were using as training data. Running the model by taking out a singular feature and seeing how the model performed without it returned varying results, with some runs indicating secondary colors like orange and cyan hindered the accuracy of the model, and other runs indicating that the model performed the best with all the features. Ultimately, the data we received from these individual runs was not consistent enough to determine whether one specific feature was affecting the emotional ratings of the image.

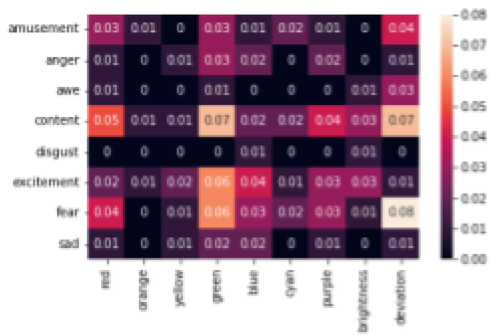

Fig. 3: A heatmap of the coefficient values obtained from the linear regression model. Values highlighted with a color on the lighter scale indicate a higher effect a specific feature had on a specific emotion

However, according to another feature of the linear regression model, we were able to pinpoint features that would have affected the emotion predictions. The regression coefficient values for each feature and emotion provided some insight into which features were contributing more to the emotion prediction: the more variance in the value of the coefficient, the more that the feature affected the prediction. Consistently, the two coefficient sets that varied the most belonged to the features “average green” and “brightness deviation”. These two seemed to be the most linked to emotional responses, but the specific emotions they affected the most are still inconclusive.

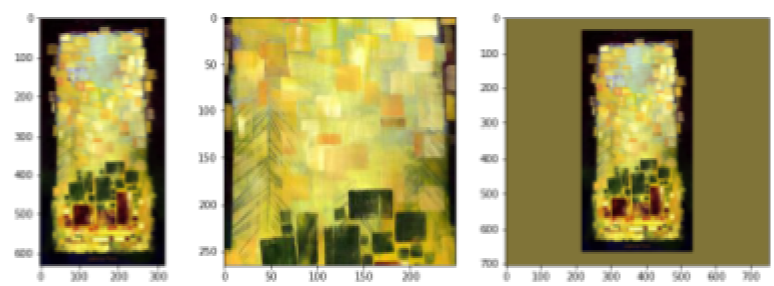

Fig. 4: An example of an unedited image from the dataset (left), what it would look like after being cropped (center), and what it would look like with the padding (right).

After this analysis, we ran the images through a CNN in two ways. First, we cropped the images to make the image size uniform and equal to the smallest possible size of all the images. For the second way, we added a border in the average rgb value to make the image size uniform and the largest possible size of all the images. Each time, the CNN performed with a mean absolute error value that was higher than the one we consistently got with the linear regression model.

Given that the CNN focuses on local features such as lines and shapes, the fact that it was not working as well on global features such as overall color values throughout the image provides insight into what types of features affect how we perceive art. Our results suggest that emotional responses to images are not affected by individual shapes as much as they are influenced by the “bigger picture”.

6. Conclusion

Our goal was to identify specific features that contributed to emotional reactions people had to viewing a piece of art. We engineered nine novel features that we hypothesized would be related to emotional responses and analyzed them using a variety of machine learning models.

The model that consistently performed best was the linear regression model with all the features combined. By calculating the correlation between multiple independent variables (color composition, lighting, etc) and one dependent variable (emotional reaction to the painting) we were able to infer that the combination of all features in a painting affects it the most, not just one specific aspect. The CNN, which focused on local features in the image itself, was not as effective as the linear regression model. Some features were recorded to affect the painting more than others, such as overall green and tone variance, but overall it seemed like the total combination of all these features across the entire image led to the most accurate model. That is, all features were significant.

We would like to try out different image datasets that focus on specific objects as subjects to really test the limits of the CNN, and figure out if the overall "big picture" of each painting affects emotions or if it's the specific subjects. We also would like to refine the linear regression model to see if it produces a more consistent result, and/or determine fully whether any specific colors or features affect emotional reaction. The trends we discovered with these models have a lot of room to explore further.

References

Stephen W.P. Kemp, Gerald C. Cupchik. 2007. "The Emotionally Evocative Effect of Paintings" in *Visual Arts Research* Vol. 33, University of Illinois Press. https://www.jstor.org/stable/20715435#metadata_info_tab_contents

Jorien van Paasschen, Francesca Bacci, David P. Melcher. 2015. "The Influence of Art Expertise and Training on Emotion and Preference Ratings for Representational and Abstract Artworks" *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134241