WikiArt Analysis Using Facial Emotion Analysis

Ariel Rakovitsky
Princeton University
arielr@princeton.edu

Julianne Knott
Princeton University
jrknott@princeton.edu

Abstract

This project focuses on the comparison of the emotions expressed by the subjects of portraits to the emotions evoked by the artworks that they are in. We used the FER-2013 [1] and WikiArt Emotions [3] datasets to train a model to predict the emotions expressed by the subjects of a painting and compare these results with the emotions that painting evokes in its audience. While the emotions observed in the subjects by the audience may not always be the same as those the audience experiences while viewing a given painting, the patterns between the two are especially intriguing in order to better understand the relationship between sympathy, empathy, and the artistic merit of an artwork.

1. Introduction

The WikiArt dataset contains thousands of paintings spanning the last several hundred years annotated with the emotions that those paintings invoked in the annotators. Our aim was to implement a novel solution to art-emotion classification. We hoped to draw a conclusion about art using computer science and computer vision analysis tools.

Going off of cultural idioms (the eyes are the windows to the soul), we produced the thesis that the key source of emotion in artwork lies in the faces of the subjects of a painting. A man in a painting with a sad look on his face is likely to inspire sadness in those observing the art just as a smiling woman is likely to inspire happiness.

We present here our investigation of this thesis. For a dataset, we cropped the WikiArt dataset, taking only images with faces. We built an emotion classification system using the FER-2013 emotion dataset and a convolutional neural network. We then found faces in paintings using an off-the-shelf face detector and predicted painting emtotions using the emotions of the faces in the painting. We further performed several ablation studies to better understand our model.

2. Previous Work

2.1. Emotion Classification

We base our emotion classification approach on the work of Ngo and Yoon in "Facial Expression Recognition on Static Images." In their paper, Ngo and Yoon propose a no frills approach to emotion classification using a single unmodified convolutional neural Inetwork (ResNet-50). Ngo and Yoon use a ResNet-50 model pretrained on ImageNet, arguing that the transfer learning aspect of using a pretrained model boosts performance.

Ngo and Yoon operate on static images, meaning there is no context given before or after the single picture. This is one class of emotion classification problems, the other being dynamic emotion classification. Here, a short video clip of someone experiencing an emotion is given. We chose to model our system off of a static approach as paintings are static images.

2.2. Datasets

To train our model, we used the FER-2013 [1] dataset. The FER-2013 training set consists of 28,709 images of faces, each of which is gray-scale and 48 x 48 pixels in size. The test set consists of 3,589 examples. Each example is labeled with one of the following emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

To evaluate our model when applied to artworks, we used the WikiArt Emotions [3] dataset. The WikiArt dataset contains of 1718 images of artworks which are both labeled as containing faces and are labeled with emotions based on the image only (and not the title). From these images, we were able to detect 1574 faces which we used for our evaluation. Each of these images was labeled with one or more of the following emotions: agreeableness, anger, anticipation, arrogance, disagreeableness, disgust, fear, gratitude, happiness, humility, love, optimism, pessimism, regret, sadness, shame, shyness, surprise, trust. Each of these emotions was grouped into one of 3 categories: Positive, Negative, and Other / Mixed. The images span a wide range of time periods (1415 - 2012) and include 4 categories of art: Renaissance, Post-Renaissance, Modern, and Contemporary.

Understandably, more images from the Renaissance period both had faces and had faces that were realistic enough to be recognized as faces by our face detector.

3. Design and Implementation

3.1. Emotion Classifier

The first progress step in our project was to build an effective emotion recognition system. We chose to build this without the use of external emotion-detection libraries. We follow the approach of Ngo and Yoon in using a convolutional neural network as the centerpiece of our system.

Our system uses ResNet-50 [2] as its convolutional neural network. Unlike Ngo and Yoon, we chose not to use a pretrained neural network, as we were operating with grayscale images. We added a dropout layer (with parameter .05) and a fully connected layer (with 7 output nodes) to ResNet. We modified the first convolutional layer of ResNet to accommodate grayscale images with one color channel instead of three. Softmax was used as a final non-linearity. Thus, we used a single convolutional network that output confidence scores between 7 possible emotions.

To implement this network, we used PyTorch for its training and data loading functionality. To train, we used the Stochastic Gradient Descent optimizer, with an initial learning rate of 0.01 and a momentum of 0.9. We also used a slight weight decay (L2 penalty) of 0.0005. Cross Entropy Loss served as our loss function. We reduced the learning rate by a factor of 10 every 10 epochs. The network was trained for 30 epochs using a Tesla K40C GPU and a mini-batch size of 64.

3.2. Art Emotion Classification

To detect faces in the artwork, we used the face_recognition 1.3.0 python library [1]. While this is a very effective tool for detecting adult faces in photographs, we found that it worked significantly worse for artworks. Of the 1718 artworks labeled as having faces, face_recognition was only able to detect 855 images. Because some of these images contained multiple faces, we were able to detect 1574 faces in total. While we would have liked the face detector to work better on paintings and other artworks, it is difficult to say if its poor performance is because the face detector is inherently insufficient for detecting faces in artworks or if the labeling of the artworks didn't accurately reflect whether a human recognizable face was in fact present. Upon examining many of the images where the face detector couldn't find a face, it became clear that WikiArt's definition of "face" was very loose, including both images where only the back of someone's head was shown 1, or the face belonged to an animal 2, or where the supposed "face" was more an abstract conception of the idea of a face and not something that would be recognizable, even to the average human, as a face 3. There were, however many images where there were faces that were obvious to a human viewer that the face detector could not detect 4. Notably, these included many post-renaissance, impressionistic artworks where the faces

[1] https://pypi.org/project/face-recognition/

Figure 1: An example of an Image with an obscured face.

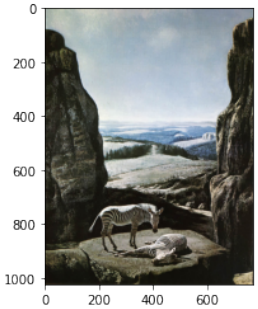


Figure 2: An example of an Image with an animal face.

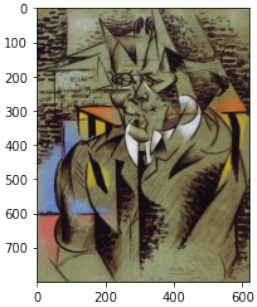


Figure 3: An example of an Image with an abstract face.

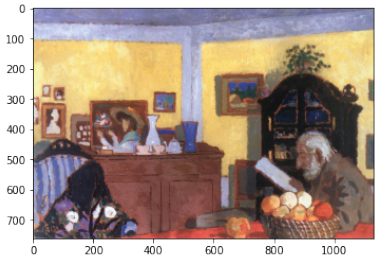


Figure 4: An example of an Image with a face that should have been detected, but was not.

were done in thick, gestures of color and texture instead of the photo-realistic precision of renaissance works.

Once the faces had been detected, they were assigned the FER-2013 label corresponding to their WikiArt label. Since the FER-2013 labels consisted of a subset of the labels used in the WikiArt data set, when we applied our model to the WikiArt data set, we needed to create a mapping from the WikiArt labels to the FER-2013 labels. To do this, we mapped each WikiArt label to the FER-2013 label that we deemed to be closest. The FER-2013 labels contained only one label that was positive ("happy") and one that was other / mixed ("neutral"), so each label in one of these WikiArt emotion categories got mapped to the corresponding FER-2013 label accordingly. Mapping negative labels was somewhat more nuanced, as FER-2013 contained 5 labels that WikiArt categorized as negative. For this, we chose to map each WikiArt label without a direct correspondent to the most common negative label ("sadness"). When an image had multiple WikiArt emotion labels, the FER-2013 equivalent was found for each of the labels and the FER-2013 label that appeared most frequently in that image was chosen with ties broken arbitrarily. This was effective because in most of the images with multiple labels, the labels all fell within the same category of positive, negative, or other and would be mapped to the same FER-2013 labels. Finally, the face images were converted to grayscale, resized to be 48 by 48 pixels and normalized before we applied our emotion recognition model to them.

4. Results

4.1. Emotion Classifier

4.1.1 Quantitative

The following are the accuracies of the various models emotion-classification models we produced on the FER-2013 test set.

Implementation	Acc.
ResNet-50; Dropout .02, Weight Decay .0005	**52.9%**
ResNet-18; Dropout .02, Weight Decay .0005	51.5%
ResNet-50; Dropout .2, Weight Decay .05	52.7%

This accuracy of 52.9% lands us in about the top-20 on the FER-2013 leaderboard. It does not, however, match the 71.1% top accuracy on the leaderboard. However, because our end goal was ultimately to draw a conclusion about the art, and not the emotion classifier, we determined that our performance was strong enough for use in later sections.

(a) Happy (correct)

(b) Angry (correct)

(c) Neutral (incorrect, correct was surprise)

(d) Sad (correct)

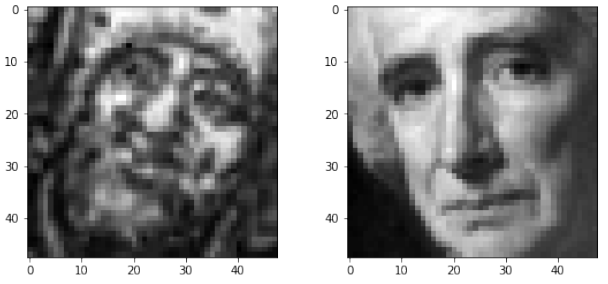
Figure 5: Qualitative Examples of Emotion Classification System

4.2. Art Emotion Classification

4.2.1 Quantitative

When we applied our model to all of the faces we detected in the WikiArt artworks, we found that our model detected the same emotion in the subjects' faces as was evoked in the audience by the image with 16.58% Accuracy.

We anticipate there being significant error as a result of our emotion classifier being trained on photographs whereas the faces in the WikiArt data set are primarily paintings. However, since each of the faces used was "life-like" enough to be identified by a face detection algorithm which was also trained exclusively on photographs, this points to other sources of error. For one, we believe that our results suggest that most art is not empathetic in nature, meaning that the emotions expressed by the subject(s) are not those meant to be experienced by the audience. Additionally, we believe our results point to facial expressions only comprising a small subset of the elements of an artwork, such as color, setting, and historical / cultural context, which collectively contribute to the emotions it evokes. This is supported by examining some of our qualitative examples 6 where a human likely would not have chosen correctly if they were given only the gray-scale face to examine.



(a) Sad (correct)

(b) Sad (incorrect, correct was Happy)

(c) Neutral (incorrect, correct was Happy)

(d) Sad (incorrect, correct was Fear)

Figure 6: Qualitative Examples of Emotion Classification of Artworks

5. Analysis

5.1. Ablation Studies

5.1.1 Emotion Classifier

In the implementation section, we noted that the emotion classifier system was trained for 30 epochs. We observed, however, the the validation accuracy did not increase significantly between the twentieth and the thirtieth epochs. To better understand whether this was a limitation of our training or our model, we performed several ablation studies.

First, we trained for 250 epochs, with a learning rate decrease factor of 10 every 10 epochs. Again, after about 20-25 epochs, no significant change in validation accuracy occured, even though training accuracy reached near-90% after 250 epochs. This is a clear example of overfitting. We speculate that this overfitting is due to the tremendous number of free parameters in ResNet50 and the relatively small (48x48) image size we used for training.

To attempt to mitigate this, we performed an experiment using RestNet-18 instead of ResNet-50. We hoped that because ResNet-18 had fewer free parameters, it would not overfit to the same degree. The training accuracy of ResNet-18 increased more slowly (measured in number of images seen) than that of ResNet-50, but nevertheless reached about 90% after 250 epochs. The validation accuracy, however, peaked once again at about 20 epochs. This time, however, our validation accuracy was about a percent lower than with ResNet-50. Thus, using the smaller ResNet-18 did not help combat overfitting, but rather forced the overfitting to occur more slowly.

Finally, we tried using a greater dropout layer value of 0.2, a greater weight decay value of .05, and ResNet-50. As wtih ResNet-18, we saw training accuracy increase more slowly, but cross the 90% threshold after 250 epochs. Again, validation accuracy peaked early, at about 25 epochs.

From these ablation studies, we conclude that our training scheme is not the problem in overfitting, but rather the model used, the small-to-moderate amount of data (28,000 images every epoch), and the small image size (48x48). We believe that the path to mitigating overfitting lies not in training-parameter adjustment but rather data manipulation. This means applying transformations to existing images to produce more data. Further, color images would be of great help so that our convolutional network could learn color features.

5.1.2 Art Emotion Classification

Anticipating that some of our error derived from inaccuracies in how we converted WikiArt labels to FER-2013 labels, we tried running our model on only the WikiArt artworks with labels that directly corresponded to the FER-2013 labels. We found that our accuracy for this selection of data was slightly less, at 15.73%. This indicated that our mapping of emotions had a very slight positive impact on the accuracy.

5.2. Strengths and Weaknesses

The the strength of our classification system is the emotion classifier. Our classifier performs very well on photographs from the FER-2013 test set. Additionally, based on observations of our qualitative results from running our classifier on the WikiArt faces, we believe it performs better on the artistic renditions of faces than the measured accuracy would lead us to believe. This is because, outside of the context of the overall image, which the ground truth label is based on, many of the faces's emotions seem to more closely reflect the predicted emotion than the ground truth emotion 6. In particular, many artworks rely heavily on historical and cultural illusions, details from the background, and the overall choice of colors and textures to create a very specific emotion. Additionally, while renaissance art encompassed the majority of our face examples because the faces in these paintings tended to be the most photorealistic, the style of renaissance art tends to portray people with very neutral, serious expressions. It is rare to see a renaissance

painting, even of a very happy scene, in which someone is smiling. This means that the ground truth labels of these paintings were influenced more by the context of the painting than the subjects' expressions, suggesting that these labels may not be the best metric by which to evaluate our model.

The major weakness of our art emotion classification system is its tendency to ignore vital information. Our thesis requires we look only at faces in an art piece and our face emotion classifier, which was trained on black and white data, considers only one color channel. We thus ignore the spatial features of the art piece as a whole and the color features of both the face and the art piece as a whole. We found that the information found only in the face was not strong enough to ignore the other information.

Further, our system must find a face to predict emotions inspired by an art piece. This constrains us to photorealistic art pieces. However, non-photorealistic art also invokes emotion that our system cannot even attempt to predict. This is a structural weakness of our approach.

6. Conclusion

We successfully implemented a face emotion classification system and applied this system to artwork with visible human faces. We found that faces alone are not a strong predictor of emotions invoked by art. Nevertheless, this paper presents a novel use of computer science tools to study artwork.

References

[1] Challenges in representation learning: Facial expression recognition challenge — kaggle, 2013.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[3] Saif M. Mohammad and Svetlana Kiritchenko. An annotated dataset of emotions evoked by art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.