Learning the Deep Emotional Intelligence of Artists

Masoud Charkhabi
Stanford University
masoudc@stanford.edu

Arturo Garrido
Stanford University
agarrido@stanford.edu

Abstract

Creators of content attempt to provoke certain emotions in viewers. In this project we attempt to learn this mapping of art (through pixels) to emotion (through labels), with deep learning. Our project has two goals; first construct a classifier to predict emotion classes of art images, and second; build GANs to generate novel art, which can be tested by our classifier.

1 Introduction

Classifying art into emotion categories has been challenging for researchers even with state-of-the-art deep learning. The challenges stem from subjective labeling, user perspective, and that emotions can be invoked by global and local features of an image [11] [2]. The main challenges with generating images in this domain relate to significant data and computation requirements [1]. Our motivation to study emotion with deep learning was that it has been less explored with pure computation due to the aforementioned challenges, and that progress could have many applications. From discussions with our project sponsor, industry experts and literature review [5] [6], we see applications in education, human-robot interaction and media and entertainment.

2 Related work

Our literature review were in three areas; emotion classification, GANs on emotion data, and general Computer Vision (CV) and psychology research on emotion. [15] is the main paper that motivated our classification work. The authors highlight the class imbalance challenge with facial expression data. Multiple ML techniques were used including SVMs for classification, and t-SNE to understand emotion class similarities. The authors achieved within-class accuracy of more than 90% on three benchmark datasets. What was novel about this paper is that they attempted to augment their data with GANs which motivated us to do so. [1] is a successful GAN implementation on artworks rather than facial expression data. The authors train class-conditioned GANs which are used to map classes to emotions. Our work on GANs was not satisfactory, from a naturalness perspective, to the point of generating images for either augmentation such as [15], or for class specific art generation such as [1]. We believe this is related to our data, compute and time constraints. We present our work on GANs as interesting work for the future. Our interaction with the dataset through our classifier and GANs motivated us to research the deep history on emotion and images from CV and psychology, and gain a deeper understanding of art data through [7]. [3], [4] and [13] achieved explainable results for emotion classification with carefully designed features motivated by CV. The authors used their models to alter emotional reactions. We used the Hard-Soft metric [4] based on Perlin noise [9] and structural similarity [14] to define a texture score that along with a brightness score motivated by [8], and color maps from [3] were used to create Fully Connected (FC) neural network.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3 Dataset and Features

Our data are 40k artistic images from a sponsor, 3k of which have emotion labels and other tags. Most of the artwork is medieval with highly subjective labels. There are no ground truth labels for our data; a painting may transmit different emotions to different people. With this intuition, from [15], and some analysis of label distributions we grouped 16 emotion classes into 4. Nearly 39% of our data were labeled "neutral" or "NA". Data from each of our four classes are shown and described in Figure 1.

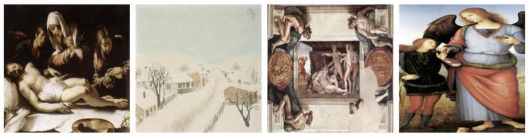

Figure 1: Four grouped classes from left to right: 0:[Sadness, Fear, Disgust, Anger], 1:[Neutral, NA], 2:[Lust, Envy, Surprise], 3:[Optimism, Joy, Love].

After some unsuccessful preliminary models we filtered our data to only portraits. This was motivated from error analysis using Class Activation Maps. We observed the last layer of the model has not learning features that we could interpret as meaningful or relating to emotion. This is shown in Figure 2.

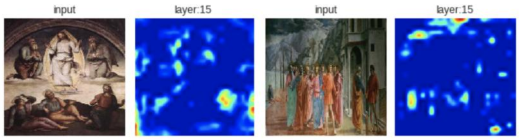

Figure 2: Error analysis on 64 artworks with CAMs and CMVs lead us to believe the VGG16+TL model is over-fitting to noise, since the high gradient regions do not provoke any emotional reaction. Left: class 3, Right: class 2.

Our model in this case was VGG16 as the starting point for Transfer Learning (TL). We describe the refined VGG16+TL model in 4 Methods. We also noticed that our labels seem to be either related to local features such as objects, or global features such as color, texture or brightness. This is shown in Figure 3.

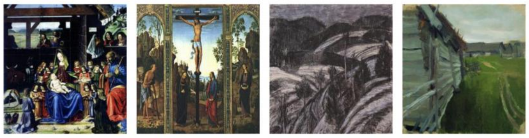

Figure 3: Labels for the first two images from the left driven by local features (classes 3 and 1 respectively) vs. two images on the right driven by global features (classes 1 and 3 respectively).

We validated this hypothesis by running our data through the YOLO pre-trained object detector, and found many images to have a "person" object Figure 4. Also from our dataset tags, we found many had a tag "portrait" that suggested a person in the art work. We limited our data to only these art works. Two other analyses led us to filtered data left us with nearly 500 labeled artworks for our classifier which we augmented to 2k. Flips and rotations were used but not blur since it would impact texture which would interfer with one our theories that will be described next. For GANs, we used 7k unlabeled portraits. With motivation from [3] and [4] we built a classifier based engineered features. The model was fairly simple. Here we describe the features; 13 global hand engineered features were designed to capture the global features color, texture and brightness. The texture features are based on Perlin noise [9] which forms the Hard-Soft (HS) scale presented by [8]. We use a standard method to measure image brightness by calculating the square root of a linear combination of the RMS values for each color channel from [8]. For color features we used the method in [3]; color features we first extract the top k color centroids from the example image, then compare them to the standard color maps we created for each emotion class. The color features are an error vector, for the closest color map, the value represents the euclidean distance between the example image and that closest color map, and for the other color maps, the value is set to an arbitrarily large maximum distance. These feature are shown and described in Figure 5.

Figure 4: Object detection revealed 20 percent of artworks included a “person” or “portrait”.

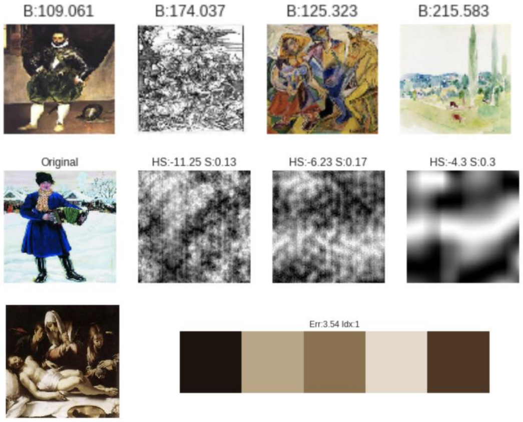

Figure 5: Hand engineered features, from top to bottom: Perceived Brightness where B is the brightness value, Texture Similarity Structure where HS is the Hard-Soft value and S is the structural similarity of the input image to our texture maps, Colour Map Distance where the 5 primary colors of the input image are extracted and the distances to our landmark colour maps are calculated.

4 Methods

4.1 Classification

We began with a CNN with the same architecture as we later describe in 4.2.2. We refer to this model as ShallNet. We used the adam optimizer with categorical cross-entropy loss. Performance was not satisfactory on the train and test sets and is shown in Table 1. After consultation, we attempted to train a FC layer appended to VGG16 for feature extraction. We froze the parameters of the VGG16, replaced the last maxpooling layer with a global average pooling layer and used one FC layer rather than two with a softmax on top. We refined our classifiers based on their performance on the test set, as well as analyzing Class Activation Maps (CAM) and Class Model Visualization (CMV). This improved results but was still not satisfactory. We refer to this model as VGG16+TLv1. By observing the CAMs from our VGG16+TL model, we concluded that many labels are driven by global features such as colour and texture rather than local ones such as shapes and objects that VGG16 was meant to learn. VGG16 was also trained on images rather than art. This is when we decided to filter our data, as described in 3 Dataset and Features, and tune another VGG16+TL which we call VGG16+TLv2. Performance of these models are shown in Table 1. Given the global and local nature of features described earlier we then created CV+FC which was as FC neural network on 13 hand engineered features we described in detail. This model was competitive with VGG16+TLv2 as seen in Table 1. [4] used similar features to achieve 70-80% accuracy on their data.

4.2 GANs

The goal of building GANs was to train them with portraits from a specific emotion class, and then test newly generated portraits on our classifier. However, because of the reduced amount of labeled data, we first start by training our models on all the unlabeled portraits (7k) and analyze their performance. The loss functions from the generator and discriminator are respectively:

Table 1: Model Performance

Model	Description	Train Acc.	Test Acc.
ShallNet	4conv + 1FC	29.0%	27.6%
VGG16+TLv1	noisy data, 50 epochs	45.1%	39.3%
VGG16+TLv2	refined data, 50 epochs	56.1%	57.2%
CV+FC	13 CV motivated features + 1FC	55.2%	56.3%

$$L_G^{NSGAN} = -E_{\hat{x}\sim p_g}[log(D(\hat{x}))]$$

$$L_D^{NSGAN} = -E_{x\sim p_d}[log(D(x))] - E_{\hat{x}\sim p_g}[log(1 - D(\hat{x}))]$$

We tried different architectures for the generator and discriminator, and combined them in different ways. We present here the basic generator-discriminator architecture that we used as a baseline, a customized more complex architecture and an architecture extracted from the literature. For our baseline model, the Generator architecture was a shallow network with FC layers. The network receives a 100-d random code, and generates an array of dimension w * h * c through 3 FC layers. Our Discriminator architecture is a shallow network with FC layers. The input image is flattened and fed to 3 FC layers. As presented in Section 5, this model does not yield great results because of the low complexity of the architectures. In our Complex Customized Model, the Generator architecture had deconvolution layers, he random code input, after going through a FC layer, is fed to 3 deconvolutions. The Discriminator architecture was a shallow CNN network with 2 convolutional layers (with max pooling) and a FC layer. As shown in Figure 6. In DCGAN, Here our Generator architecture has 4 deconvolution layers and Discriminator architecture 4 convolutional layers. To test this model, we use open code from a github repository [12] as a baseline, and change it to adapt it to our data.

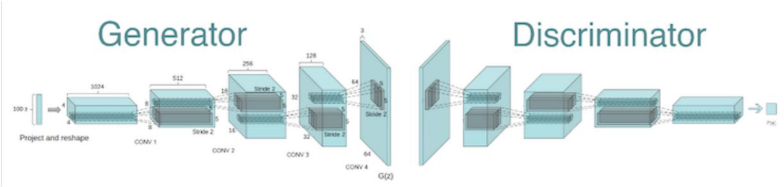

Figure 6: DCGAN architecture [10]

5 Results and Discussion

Results for our classification models are summarized in Table 1, and described in detail in 4.1 Classification. We believe that with more data VGG16+TLv2 will outperform CV+FC. An ensemble of these two models or a model with a concatenated final FC layer may outperform each individual model by capturing both local features (VGG16+TLv2) and global features (CV+FC).

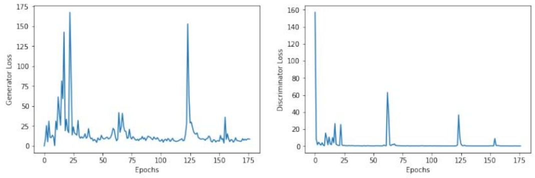

Figure 7: Costs for the (a, left) Genarator, (b, right) Discriminator, Baseline Model

For GANs, as observed in Figure 7, the discriminator converges fast while the generator cost has more noise. This can be explained by the fact that the generator fails to trick the discriminator most of the times. We tried tuning the learning rates of both generator and discriminator, as well as training the generator more often, but the results were similar. Therefore, for the next iteration, a more complex architecture of the generator was the primary requirement. We plot some of the images generated by the trained generator, and observe that although some patterns (such as more brightness in the center of the image, were the person would be in the portrait) are learned, they are mostly noise.

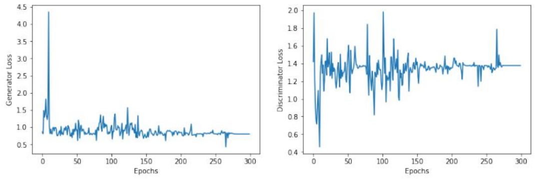

Figure 8: Costs for the (a, left) Genarator, (b, right) Discriminator, Complex Customized Model

For our complex customized GAN, as observed in Figure 8, the cost functions are different from the previous case in that the discriminator takes much longer to converge. This suggests that the generator is making it more difficult for the discriminator to detect fake images. However, when both converge, most of the images generated by the generator are labeled as fake by the discriminator, which is superior again. Either more training or more complex architectures for the generator are required. Again, we plot images generated by the generator, and now the pattern of a face is recognized. However, the images are still very noisy and lack global cohesion. For our DCGAN, this architecture first presented in [10], has the characteristic that it is stable in most settings. However, its training requires a lot of computational power. We could only run some epochs, because of the huge amount of time that the training took, so the results are not realistic paintings.

6 Conclusion and Future Work

Emotion classification and generation was challenging due to reasons described in 1 Introduction and 2 Related work. The VGG16 model was pre-trained on real images (ImageNet), whereas here we are working with works of art. Despite lower classification accuracy, and less natural generated images, we believe our classifier has managed to learn some utility. We show some of the interesting features that our VGG16+TLv2 model has discovered in Figure 9. GANs struggle to generate globally coherent images from datasets with high variability, such as our's (different styles, colours, textures, etc.). The main issue we ran into is that the generator failed to "trick" the discriminator most of the times. However, it did learn some patterns, as seen in Figure 6. As future work, we would have liked to try deeper generator architectures, and train the model for more epochs. Some interesting generated images from our GAN are shown in Figure 10.

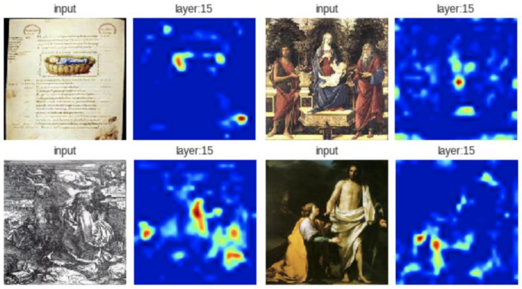

Figure 9: CAM/CMV analysis shows our classifier may have learned some useful local features, from top to bottom: "Baby Detector" finds the bright round head of a baby, "Submission (kneel and reach) Detector" finds the kneel and reach posture common in art from this period.

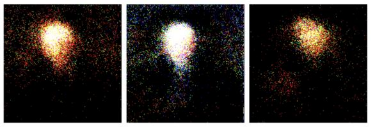

Figure 10: GAN generated portraits resemble the data, face area is brighter and collar is noticeable.

7 Contributions

Masoud Charkhabi: Literature review, Classifier, Transfer Learning, Object Detection, NST, GANs, data pipeline, write-up, meetings.

Arturo Garrido: GANs, shallow CNN classifier, data pipeline, write-up, active in meetings with advisor and sponsor.

Deepak Bansal: set up shared drive, but otherwise has not been involved in the project.

Chez Mana, our sponsor, is a digital content creator. They aim to augment their content library using AI techniques. The high level goals of the project were developed jointly by our working team, CEO Mana Lewis, and science advisor Bill Jarred.

References

[1] David Alvarez-Melis. The emotional gan : Priming adversarial generation of art with emotion. 2017.

[2] A. Hurlbert and Yazhu Ling. Biological components of sex differences in color preference. *Current Biology*, 17:R623–R625, 2007.

[3] S. Liu and M. Pei. Texture-aware emotional color transfer between images. *IEEE Access*, 6:31375–31386, 2018.

[4] Marcel P. Lucassen, Theo Gevers, and Arjan Gijsenij. Texture affects color emotion. 2010.

[5] S. Malik, Sungchul Kim, and Eunyee Koh. Perceptual similarity ranking of temporal heatmaps using convolutional neural networks. In *EE-USAD'18*, 2018.

[6] Daniel J. McDuff, Abdelrahman N. Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana El Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *CHI Extended Abstracts*, 2016.

[7] Saif Mohammad and Svetlana Kiritchenko. Wikiart emotions: An annotated dataset of emotions evoked by art. In *LREC*, 2018.

[8] R. M. H. Nguyen and M. S. Brown. Why you should forget luminance conversion and do something better. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5920–5928, July 2017.

[9] Ken Perlin. Improving noise.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[11] Paul J. Silvia. Emotional responses to art : From collation and arousal to cognition and emotion. 2005.

[12] carpedm20 Taehoon Kim. Tensorflow implementation of "deep convolutional generative adversarial networks. `https://github.com/carpedm20/DCGAN-tensorflow`, 2018.

[13] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 47–56, New York, NY, USA, 2014. ACM.

[14] Hamid Rahim Sheikh Eero P. Simoncelli Zhou Wang, Alan Conrad Bovik. Image quality assessment.

[15] Xinyue Zhu, Yifan Liu, Zengchang Qin, and Jiahong Li. Data augmentation in emotion classification using generative adversarial networks. *CoRR*, abs/1711.00648, 2017.