Stanford | ENGINEERING
Computer Science

Learning the Deep Emotional Intelligence of Artists

Masoud Charkhabi, Arturo Garrido, Deepak Bansal {masoudc, agarrido, deepakb7}@stanford.edu

Motivation

Creators of content attempt to provoke certain emotions in viewers. In this project we attempt to learn the mapping of pixel to emotion with deep learning. Our project has two goals:

1. Construct a classifier to predict emotion classes of art images.
2. Build GANs to generate art, which can be tested by our classifier.

Data and Problem Definition

Our data are ~40k artistic images from a sponsor, ~3k of which have emotion labels and other tags. Most of the art 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, and some analysis of label distributions we grouped 16 emotion classes into 4. Nearly 39% of our data were labeled "neutral" or "NA". After some preliminary models we filtered our data to only portraits (further explanation in Approach). This filtered data left us with nearly 500 labeled artworks for our classifier which we augmented to 2k. For GANs, we used 7k unlabeled portraits.

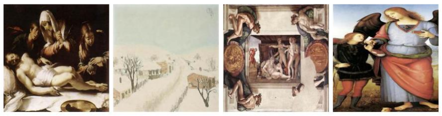

Figure 1. Four grouped classes from left to right: 0:[Sadness, Fear, Disgust, Anger], 1:[Neutral, NA], 2:[Lust, Envy, Surprise], 3:[Optimism, Joy, Love].

GAN Models

The loss functions of the generator and discriminator are respectively:

$$\mathcal{L}_{\mathrm{G}}^{\mathrm{NSGAN}} = -\mathbb{E}_{\hat{x}\sim p_g}[\log(D(\hat{x}))]$$

$$\mathcal{L}_{\mathrm{D}}^{\mathrm{NSGAN}} = -\mathbb{E}_{x\sim p_d}[\log(D(x))] - \mathbb{E}_{\hat{x}\sim p_g}[\log(1 - D(\hat{x}))]$$

Generator architectures:
1. Shallow network with Fully Connected (FC) layers: first we consider a network that receives a 100-d random code, and generates an array of dimension w * h * c through 3 FC layers.
2. Deconvolution layers: the random code input, after going through a FC layer, is reshaped and goes through 3 deconvolutions.

Discriminator architectures:
1. Shallow network with FC layers: input image is flattened and fed to 3 FC layers.
2. CNN: network with 2 convolutional layers (with max pooling) and a FC layer.

Different combinations of generator and discriminator architectures are used, while varying parameters (filter sizes, channels, etc.) and hyperparameters.

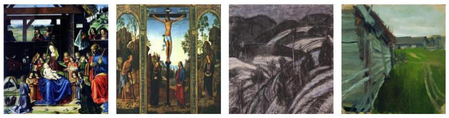

Figure 2. Labels for the first two images from the left driven by local features (classes 3 and 1 respectively) vs. two images on the right driven by global features (classes 1 and 3 respectively).

Classifier Models with Attribution

We refined our classifiers based on their performance on the test set, as well as analysing Class Activation Maps (CAM) and Class Model Visualization (CMV). We first attempted a shallow network with the same architecture used in our GANs. We used the adam optimizer with categorical cross-entropy loss. Performance was not satisfactory on the train and test sets, hence we attempted to train a FC layer appended to VGG-16 for feature extraction. This improved results but was still not satisfactory. By observing the CAMs from our Transfer Learning (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 VGG was meant to learn.

To build on the above finding we examined the additional tags and ran our data through the YOLO pre-trained object detector, and found many images to have a "person" object, and also tags such as "portrait" that suggested a person in the art work. We limited our data to only these art works.

Figure 3. Object detection revealed ~20% of artworks included a "person" or "portrait".

We then built two models; a VGG feature extractor with a trained FC layer to classify based on the local features, and a shallow FC model to classify based on 13 global hand engineered features motivated by Computer Vision (CV) theory [3]. Since our data was now filtered we used data augmentation to increase it. Flips and rotations were used but not blur since it would impact texture.

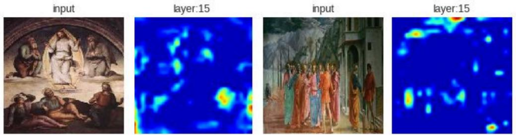

Figure 4. Error analysis on 64 artworks with CAMs and CMVs lead us to believe the VGG16-TL model is overfitting to noise, since the high gradient regions do not provoke any emotional reaction. Left: class 3, Right: class 2.

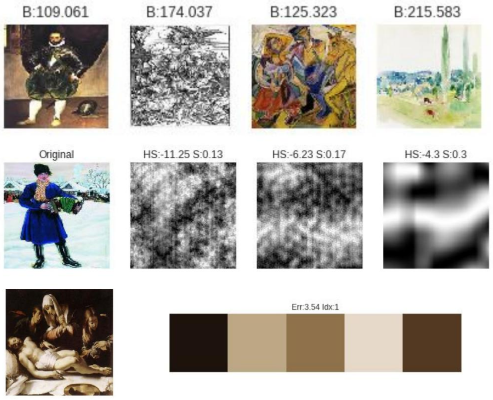

Figure 5. Hand engineered features, from top to bottom: Perceived Brightness where B is the brightness value, Texture Similarity Structure where HS is the Hard-Soft value and S is the structural similarity of the input image to our texture maps, Colour Map Distance where the 5 primary colors of the input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

Results and Analysis

We examine results on a 25% test set of our artworks. Below are the results of local and global feature models under different architectures and hyperparameters:

Model	Train Acc.	Test Acc.
ShallNet-Arch1	29.0%	27.6%
ShallNet-Arch2	30.0%	25.4%
VGG16-TL-Arch1	45.1%	40.0%
VGG16-TL-Arch2	47.0%	48.1%
VGG16-TL-Arch2	56.1%	57.4%
CV-FC	55.2%	56.3%

For deep generators, the artworks are too noisy. We think more training epochs are needed, as there are many parameters. For our 3-deconvolutions model, although the generated images are not coherent, some patterns (face shape) are reproduced.

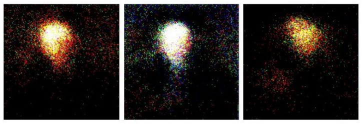

Figure 6. GAN generated portraits, face area is brighter.

Discussion and Future Work

This task was challenging due to various reasons. The VGG models were pre-trained on real images (ImageNet), whereas here we are working with works of art. In addition, labeling tended to be ambiguous, with very similar images labeled as different emotions. Despite this we believe our classifier has managed to learn some utility.

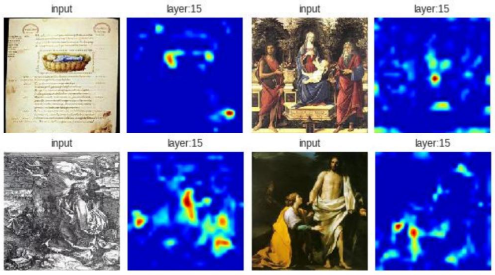

Figure 7. CAM/CMV analysis shows our classifier may have learned some useful local features, from top to bottom: "Baby Detector", "Submission (kneel and reach) Detector".

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.

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. Lucassen MP, Gevers T, and Gijsenij A. Texture Affects Color Emotion. Color Research & Application 36(6):426-36. December 2011.
2. Liu S and Pei M. Texture-aware Emotional Color Transfer Between Images. IEEE Access PP(99):1-1. June 2018.
3. MACHAJDIK, Jana; HANBURY, Allan. Affective image classification using features inspired by psychology and art theory. En Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010. p. 83-92.
4. A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In ICML, pages 2642–2651, 2017.
5. Odena, Augustus, Dumoulin, Vincent, and Olah, Chris. Deconvolution and checkerboard artifacts. http://distill.pub/2016/deconv-checkerboard/, 2016.
6. YOU, Quanzeng, et al. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. En AAAI. 2015. p. 381-388.