CS230

CS230 Final Report
Generating Realistic Emotional Pictures with Matched Visual and Emotional Attributes

Jinxiao Zhang, Insub Kim
Department of Psychology
Stanford University
jzhang18@stanford.edu, insubkim@stanford.edu

1 Introduction

Extracting emotional values in the natural scene is a fundamental aspect human visual perception. Emotional stimuli are prioritized in human visual processing, and it has been argued that humans selectively attend more to emotional pictures than non-emotional pictures. However, it is still unclear how humans recognize emotional information from the scene in a fast and efficient manner. It has been suggested that emotion perception is mediated by low-level natural statistics (color, contrast, luminance, spatial, and phase information) of the emotional eliciting content [4, 6]. For example, fear inducing images generally contains more high-spatial frequency component than happy or neutral emotion inducing images [2]. Therefore, simple low-level features of the image such as 2D Fourier spectra can be indicative of the emotional sentiment of the scene [1].

The dissociation between low-level image features and emotional content is difficult because it is inevitably intertwined in the natural image. In the field of psychology, researchers often need to make a specific claim about emotional responses that correspond only to emotional representations while controlling for the low-level visual features across different emotional categories. Yet, emotional image datasets that are currently being used in the psychological researches fail to control for the low-level visual features across different emotional categories. In this project, we aim to create a new database of emotional images that is controlled for low-level visual features using generative adversarial networks (GANs).

As a first step, we created the GAN model based on the pix2pix implementation [5] to generate emotional pictures with normalized low-level visual features (D_{low}). Secondly, we plan to add another Discriminator (D_{emo}) that matches emotional values of the pictures. To ensure the generated images contain the same emotional attributes. We will further test the performance of the sentiment analysis neural network given the newly created dataset. The newly generated emotional pictures will be a significant addition to the stimuli sets used in psychological research. Also other researchers will be able to use our model to generate image dataset of their own if needed.

2 Dataset

We used five publicly available databases of emotional pictures, namely the International Affective Picture System (IAPS, n= 956), the Nencki Affective Picture System (NAPS, n= 1356), the Open Affective Standardised Image Set (OASIS, n= 900), the Socio-Moral Image Database (SMID, n= 2941) and The Image Emotion Dataset (IED, n= 10766). Example pictures are shown in Figure 1. These pictures are in .jpg, .png, or .bmp formats that have RGB values. Each picture has a human

CS230: Deep Learning, Fall 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

rated valence (negative-positive) and arousal (calm-excited) score which is useful for matching the emotional attributes. In total, the final dataset consists of approximately 17k pictures.

Using the collected dataset, we generated low-level feature controlled target images using a MATLAB shine color toolbox [3]. The spatial frequency values were normalized by using the average of Fourier transformed amplitude spectra. The color and luminance were matched using a color space transformation. RGB values of the image were first converted to HSV color space (hue, saturation, value) and their color channel values were rescaled to matcth the mean responses of the training dataset. These target images served as a ground-truth data for the neural network to learn about the feature mapping between the original and low-level controlled images. Finally, images were re-scaled to have identical matrix size across different datasets. Some examples are in Figure 6.

Figure 1: Examples of original emotional pictures with human-rated emotional valence and arousal.

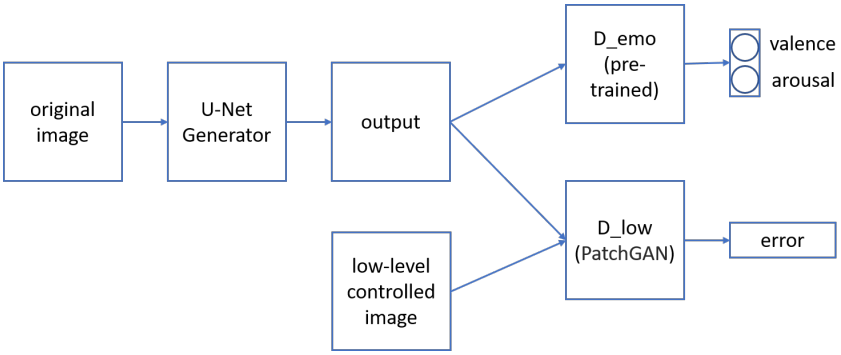

Figure 2: Emotional pix2pix model architecture.

3 Model architecture

We built our model (Emotional pix2pix model; Figure 2) based on the original pix2pix model [5]. In particular, we feed the original images to a U-Net Generator (a encoder-decoder structure) with skipped connections. The Generator generates output images of the same size as the input images. To reach the goals of matching low-level visual as well emotional attributes in the output image, we use two discriminator in the model. First, we used one Discriminator (D_{low}) to match the output image with the MATLAB generated low-level controlled image [3]. We adopted the PatchGAN architecture as in the pix2pix model for D_{low}.

Second, we used another Discriminator (D_{emo}) to retain the emotional attributes of the output image to the original image. This procedure ensured the emotional values of the images to be preserved

while low-level features are being altered. Specifically, we pre-trained a Convolutional Neural Network (CNN) to predict the human rated emotional valence and arousal from the original images. We use this pre-trained CNN with fixed parameters as D_{emo}. The loss for D_{emo} is the Mean Squared Error (MSE) between the valence/arousal from the output image and the original image. The total loss for the Generator is a linear combination of the D_{emo} loss, the D_{low} loss (see formula below), and an L1 regularization term. We experimented how the hyperparameter λ_{emo} influences the output image.

$$L(G, D_{low}, D_{emo}) = L_{low}(G, D_{low}) + \lambda_{emo} L_{emo}(G, D_{emo}) + \lambda_{L1} L1(G)$$

4 Results

Low-level controlled image generation without D_{emo}

We have implemented the model with a Generator and a Discriminator to match low-level visual attributes (D_{low}). The model parameters were initialized as the pre-trained $day2night$ model from the pix2pix paper [5]. We trained the model on a sub-section of the data over 200 epochs. The output images are shown in supplemental Figure 6. They look very similar to the target low-level controlled images. However, these output images may look unrealistic to humans, and they are likely to have altered emotional attributes to humans. In Figure 6 (second row), the original picture has positive values with a sunshiny background while the generated and the target picture has overall a gloomy background textures. Other examples can be found in Figure 4. On the first row of the left panel, the baby in the original picture is healthy while the low-level feature controlled one looks sick. On the last row of the right panel, the paper in the original picture is clean while the low-level feature controlled one looks dirty.

Pre-trained Emotion Discriminator D_{emo}

We initialized the emotion Discriminator D_{emo} with a pre-trained 18-layer resnet with adapted last layers to output a 2-element vector (valence and arousal). The D_{emo} was trained to fit the human rated valance and arousal for each image with an mean squared error (MSE) loss. The performance of D_{emo} is shown in Figure3. The correlations between the predicted and the ground-truth emotional valence and arousal are about .99, suggesting a precise estimation of emotional valence and arousal of an image by D_{emo}. We used it with fixed parameters in the Emotional pix2pix model.

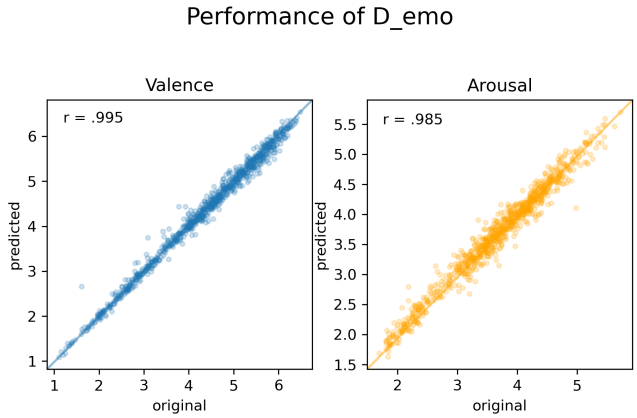

Figure 3: Performance of the Emotion Discriminator D_{emo}.

Emotion loss hyperparameter search

To generate more naturalistic and emotionally matching images additional emotional discriminator was implemented. With this additional implementation, we expected the generated images to preserve emotional attributes of the original image while controlling for the low-level image statistics.

The effect of the emotional discriminator (D_{emo}) on the generated images was evaluated using different ranges of loss weights (λ_{emo}) were tested. Specifically, three different loss weights (λ_{emo}=

0, 10, 20) were initialized to generate images, where weighting of 0 indicates the image generation without emotional matching component (only low-level controlled), and weighting of 10 and 20 indicate more contributions of emotional discriminator on the image generation procedure.

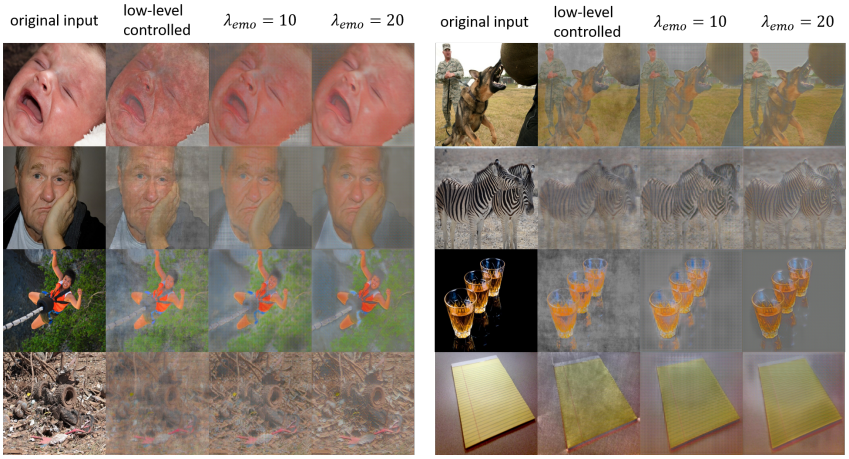

Figure 4: Examples of different emotional loss weightings

The implementation of emotional discriminator produced more realistic images. Figure 4. As more weights were given to the emotional loss the pictures became more perceptually realistic. If the given weights were too high (e.g. λ_{emo} = 100) the generator failed to produce any meaningful images. Thus, the optimum value of the weight was chosen to be λ_{emo} = 20, which successfully controlled for the low-level visual characteristics while preserving the original emotional attributes.

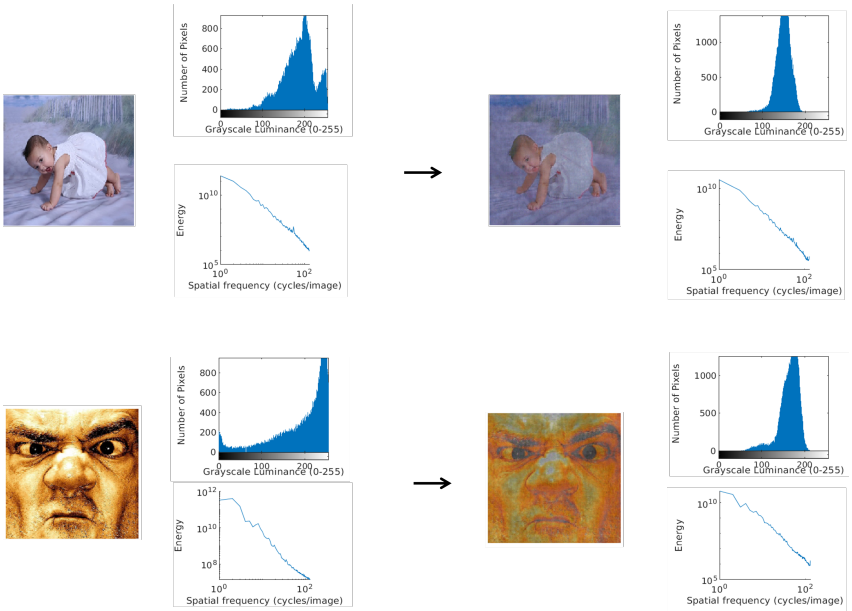

Figure 5: Low-level image statistic matching performance

Model performance evaluation

The low-level statics of the generated images across different emotional categories were evaluated. Figure 5. First, the pixel-wise luminance distribution was computed for each of the generated image. While the original input images had very distinct distributions of its pixel-wise luminance values, the generated images showed more uniformly distributed luminance values with similar mean and variance. Second, spatial frequency for each images were computed using the Fourier transformation.

Again we found that the generated images produced more matching spatial frequency values between emotional categories. In Figure 5, the lower spatial frequency energy contribution for the angry face was reduced to match the overall spatial frequency distribution of the entire emotional image dataset. As a result, the generated spatial frequency of the angry face now closely matches spatial frequency of the baby image.

5 Discussions and Future Directions

The psychology research community has long been used emotional images that are unmatched for low-level visual features as stimuli in experiments. This may have fundamentally biased the neurophysiological signals from these experiments. The overarching goal of this project is to produce a new dataset of low-level matched emotional images. We trained a pix2pix GAN with an additional emotion Discriminator in order to retain the emotional attributes of images while they are matched for low-level visual features. Our dataset offers a potentially advantageous set of stimuli for psychological research related to emotion.

With some hyperparameter search, we fine-tuned the model to achieve both of our goals. However, the emotion estimation was done by our pre-trained D_{emo}, which we presume to evaluate the emotional valence and arousal of an image like humans. To fully confirm the model's performance, we plan to further evaluate the model's performance by involving human participants to rate the emotional attributes for the model generated images. Participants will evaluate the emotional valence and arousal of the both original, low-level controlled, and generated images. Our prediction is that the original and generated images will be rated similarly while the low-level controlled images will be perceived to have different emotion levels by human raters. Finally, we plan to create and release a publicly available dataset based on this model which will be readily usable for the researchers.

References

[1] Antonio Torralba and Aude Oliva. "Statistics of natural image categories". In: *Network: computation in neural systems* 14.3 (2003), p. 391.

[2] Patrik Vuilleumier et al. "Distinct spatial frequency sensitivities for processing faces and emotional expressions". In: *Nature neuroscience* 6.6 (2003), pp. 624–631.

[3] Verena Willenbockel et al. "Controlling low-level image properties: the SHINE toolbox". In: *Behavior research methods* 42.3 (2010), pp. 671–684.

[4] Devpriya Kumar and Narayanan Srinivasan. "Emotion perception is mediated by spatial frequency content." In: *Emotion* 11.5 (2011), p. 1144.

[5] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. 2017.

[6] L Jack Rhodes et al. "The role of low-level image features in the affective categorization of rapidly presented scenes". In: *PloS one* 14.5 (2019), e0215975.

6 Supplement

Figure 6: Examples of input and output images.

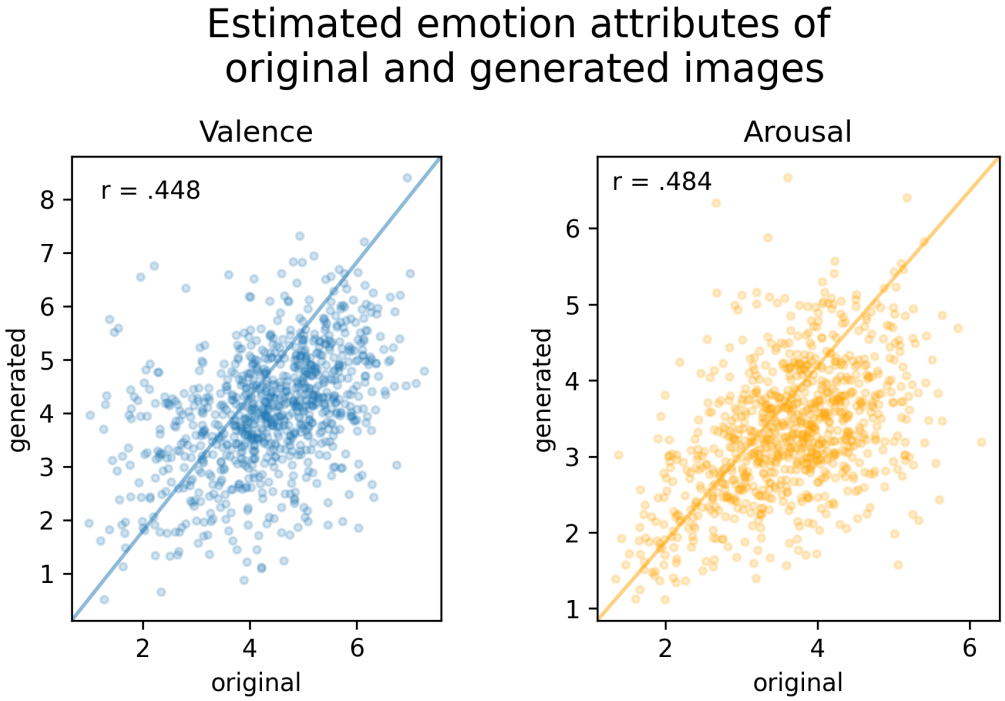

Figure 7: Estimated emotional valence and arousal of the original and output images.