Exploring emotional representation in deep neural network image emotion classifiers

Konrad Cybulski

March 2019

1 Introduction

For centuries artists have been extremely talented at creating pieces of artwork that convey a range of emotions to those who view them. Extensive research has been conducted into how visual features affect humans emotionally and how these can be used to predict and detect the emotional content of images and text (Machajdik and Hanbury, 2010; Zhao et al., 2014). Due to the subjective and qualitative nature of human emotion, assigning a quantitative measure of emotion to an image is no easy task. Furthermore the ability to computationally recognise the emotional content of an image has wide-ranging applications from classifying posts on social media, to the creation of images, text, and even physical spaces in an emotionally quantifiable way.

Methods for representing emotion in a quantifiable way has been explored thoroughly in the domain of psychology and medicine, with continuous multi-dimensional representations being used in lieu of a single emotional label (Russell, 1980). The circumplex model of emotion introduced a two-dimensional space characterised by valence, and arousal; respectively representing positivity or negativity, and the level of excitement associated with it. Such a continuous model is not without flaw, failing to accurately capture more complex emotions that often represent concurrent conflicting sides of a given axis (Larsen and Diener, 1992). The complexity of such a continuous representation of emotion has been extended by Bradley and Lang (1994) through the addition of a third dimension: dominance; which is particularly of interest within social dynamics.

The domain of emotion classification has had a particular focus on facial expressions and text (Cambria, 2016; Warriner et al., 2013). Image emotion classifiers have been explored (Kim et al., 2018; Machajdik and Hanbury, 2010; Chen et al., 2015, 2014) yet their use has been limited. While humans ability to recognise, label, and discuss the emotive content of an image is not lacking, the capability to computationally classify image emotion, and the underlying patterns learnt by such classifiers is.

2 Aims

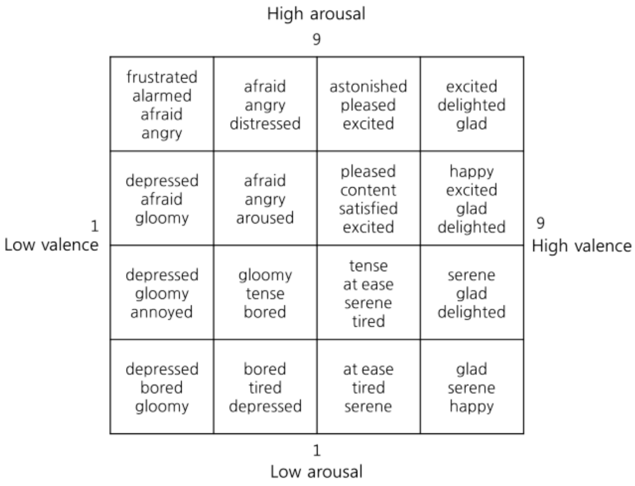
The aim of this research is to better understand the visual patterns associated with various emotions conveyed in images. This will primarily leverage image emotion recognition architectures explored by Kim et al. (2018), in combination with the dataset produced by Zhao et al. (2016) containing over 1.4 million images with assigned valence, arousal, and dominance levels derived from their descriptions. Producing an architecture with which valence-arousal (VA) values can be assigned to images forms the basis for this research. Such a platform allows further exploration into the way in which it assigns VA values to more complex, multi-faceted and multi-layered images, and the efficacy with which this is done. Furthermore this process, and the patterns learned by it can be better understood through the use of generative processes which maximise given target features in a quality-diverse way (Nguyen et al., 2015b,a), or through a generative adversarial approach (Tan et al., 2017). In the context of this research, such features include valence, arousal, dominance, happiness, sadness, etc. This will allow a great understanding of the visual patterns that such an architecture learns, and any psychological or artistic parallels that can be drawn. The combination of such a generative system with a classifier of emotion can be further extended to domains such as generative art and text-to-image synthesis, with the focus of emotion-driven image generation.

3 Background

3.1 Image emotion recognition

- □ Sentiment classification: text & images.
- □ Emotion classification in images: facial expression, general imagery.
- □ Methods of classifying emotion in images: single target emotion, discrete categorical likelihood, decomposition into continuous vector (valence-arousal).

The area of image emotion and sentiment classification has been explored in a number of ways, primarily through image feature analysis derived from art and psychological factors (Machajdik and Hanbury, 2010); and more recently using techniques such as deep neural networks (Chen et al., 2015; Kim et al., 2018). Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal (den Heijer and Eiben, 2010a,b, 2011) and as an emotional feature vector for sentiment classification (Machajdik and Hanbury, 2010). Due to the artistic and psychological underpinnings used by Machajdik and Hanbury (2010), the low-level features extracted from images can be understood at a high level. The relationship between an image's emotion and its core artistic components such as balance, harmony, and variety was further explored by Zhao et al. (2014), which uses a comparably small feature vector to Machajdik and Hanbury (2010), resulting however in a 5% classification increase to state-of-the-art approaches at the time.



High arousal

9

1

Low valence

frustrated alarmed afraid angry	afraid angry distressed	astonished pleased excited	excited delighted glad
depressed afraid gloomy	afraid angry aroused	pleased content satisfied excited	happy excited glad delighted
depressed gloomy annoyed	gloomy tense bored	tense at ease serene tired	serene glad delighted
depressed bored gloomy	bored tired depressed	at ease tired serene	glad serene happy

9

High valence

1

Low arousal

Figure 1: Distribution of emotions associated with levels of valence and arousal determined by DNN classifier produced by Kim et al. (2018)

Deep neural networks in this domain provide less transparency to the process with which emotions and sentiment are classified compared to feature analysis. The emotional content of an image can be decomposed in various ways. Image databases with singular emotion labels, and adjective-noun pairs (ANP) have been used for the training of deep neural network classifiers (Chen et al., 2014; Yang et al., 2018) with up to 200% performance gains over support vector machine classifiers.

3.2 Computational image synthesis

☐ Image generation methods explored: evolutionary algorithms, neural networks, line drawing, etc.

☐ Measuring *goodness* of a generated image: realism, abstractness, aesthetic appeal, etc.

☐ What is the generative adversarial network, and differences in common architectures

☐ Why are GANs advantageous over the use of target feature analysis (aesthetics, etc.)

☐ Text-to-image

☐ Style transfer & image-to-image

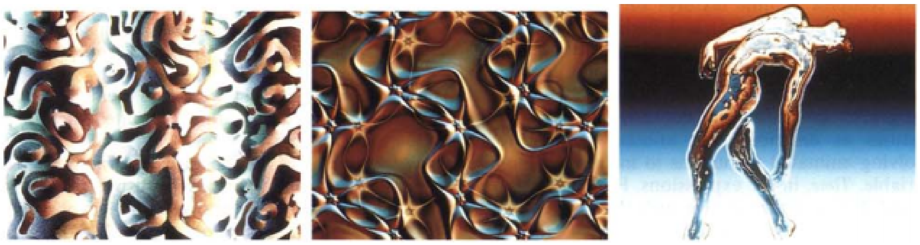


Figure 2: Images generated through the process of interactive evolution introduced by Sims (1993)

Generative systems have been explored in various domains with numerous techniques. Examples range from the generation of images and art using evolutionary methods (Sims, 1993; Machado and Cardoso, 2000), to the synthesis of text to describe a given image through the use of deep neural networks (Mathews et al., 2016). Some of the first *human-in-the-loop* image synthesis systems such as *NEvAr* (Machado and Cardoso, 2000) produced greatly impressive images through evolutionary techniques. Evolutionary art leveraged methods introduced and exemplified by Sims (1993) such as those shown in Figure 2. Sims (1993) proposed using *Lisp* expressions for genotype definitions, which accepted a coordinate (x, y) which could be evaluated into a grayscale or RGB value producing images. This genotype expression has been used in numerous further research into the process of both supervised and unsupervised image synthesis through evolutionary techniques (Machado and Cardoso, 2000; Sims, 1993; den Heijer and Eiben, 2011, 2013; Ross et al., 2006).

Despite the slow nature of the interactive process, Sims (1993) and Machado and Cardoso (2000) were able to produce images with visually striking characteristics. Ross et al. (2006) investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph (2006), that the distribution of colour gradients in fine art tend towards normal. While the images produced through this method did not meet the level of intricacy and detail as the results of Sims (1993) or Machado and Cardoso (2000), it represented a self-contained system able to generate appealing imagery without human interaction.

Introduction of the generative adversarial network architecture (GAN) by Goodfellow et al. (2014) allowed the process of image generation to be completely unsupervised. Common GAN application has involved the generation of realistic images, such as has been done by Bao et al. (2017), where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. (2017) and Reed et al. (2016) have recently explored text to image synthesis, in which detailed descriptions of birds and flowers have been converted into photo-realistic images using the GAN model. Such an architecture has been applied to the area of art synthesis by Tan et al. (2017) in which images were generated according to a target genre and artist. Learning from a dataset of countless artworks under various categories, styles, and artists, it was hugely successful in generating images that were stylistically similar to existing art of the target artist/genre. Due to the competitive relationship of the generator and discriminator networks, the patterns learned by the discriminator propagate through the generator network. The discriminator network of the GAN architecture aims to learn patterns and styles from the dataset on which it is trained in order to discriminate between the generated and existing images. As a result, the images generated tend to resemble closely those in the training dataset, which represents a benefit in successfully producing images closely matching the target domain, and a detriment with regard to the networks potential creativity.

Recent work by Nguyen et al. (2015b) and Nguyen et al. (2015a) investigated the use of quality-diverse algorithms for image generation particularly to better understand the patterns learned by deep neural network image classifiers. Quality-diverse (QD) evolutionary algorithms such as Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) (Mouret and Clune, 2015) and Novelty Search (Lehman and Stanley, 2008, 2011) have been developed to address the need for a high quality, yet diverse solution space in related optimisation domains. The use of such QD algorithms has shown great promise in its efficiency and accuracy on a number of hard optimisation problems (Pugh et al., 2016) such as maze navigation (Lehman and Stanley, 2011). Nguyen et al. (2015a) and Nguyen et al. (2015b) use MAP-Elites in conjunction with a pre-trained deep neural network (DNN) image classifier; assigning individual image fitness according to the accuracy with which it is classified. Using the MAP-Elites framework in this context, each dimension of the feature-space represents a classification label, and as such the generated images allow the exploration of label representative patterns and shapes learnt by the classifier. Nguyen et al. (2015a) leverages such an architecture to show the shallowness with which an image classifier recognises images. Assigning the label of *school bus* to alternating yellow and black lines is a prime example of the way such a network has learnt to differentiate one class from the others. Thus enabling exploration into the inner workings of the DNN classifier by uncovering features that maximise the separation of one label to another. In contrast, Nguyen et al. (2015b) uses the same architecture to explore the novelty-driven evolutionary path taken by generated images and the potential for such a system in the field of content synthesis. While the conclusions derived from Nguyen et al. (2015b) and Nguyen et al. (2015a) contrast greatly, the quality-diverse generative method used to understand the visual components learnt by the classifier show such an architecture's exploratory abilities.

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis is limited. *Emotion transfer* was explored by Ali and Ali (2017) which involved the transformation of an image's colour and style with the aim of altering its conveyed emotion.

4 Methodology

- □ Emotion profile representation
  - □ Discrete feature vector
  - □ Continuous valence-arousal space
- □ System architecture (training/evaluation method)
  - □ Generator: input type (related to emotion profile representation)
  - □ Generator: base architecture e.g. use ArtGAN (Tan et al., 2017)
  - □ Discriminator: use output image's emotion classification and error from target emotional profile as error function (autoencoder method)
  - □ Discriminator: use image label as target emotional profile and use standard logistic discrimination.
- □ Any interaction made between human and generator e.g. text-to-image synthesis using text emotion classification fed through to the generative model.

4.1 Datasets

There exist a few datasets in which non-facial images have labels of their affective emotion on the viewer. One particularly of interest to this research is the *WikiArt Emotions* dataset (Mohammad and Kiritchenko, 2018) in which images of artwork were labelled with an emotional profile. Each of the more than 4000 images in the dataset has an associated vector representing the proportion of viewers assigning a given emotion to the artwork. Emotions such as gratitude, happiness, anger, and arrogance are represented among the twenty emotions assigned to the images. Along with their respective emotional profile, each image is classified according to its artistic category (Impressionism, Baroque, etc.) and other desirable measures such as viewer rating, artwork title, artist name, and year of creation. Due to the level of detail relating to each artwork's affective emotional profile, and its classified style and category, this dataset will be investigated initially for creating the generative system detailed below.

4.2 Emotion profile representation

The method with which an emotional classification can be represented has been investigated as mentioned in the background section. With options ranging from a single target label (happy, sad, etc.), to the continuous two-dimensional circumplex model (valence-arousal) representation first introduced by Russell (1980). A model for representing emotion commonly used in classification tasks is that of a single label target, due its simplicity. However due to the subjectivity involved with the emotional classification of an image, a floating-point vector models the relative proportions with which an image's emotion is classified.

Such a model is used explicitly by Ali and Ali (2017) as a target emotion profile, and is the representation used by (Mohammad and Kiritchenko, 2018) due to statistical methods used in gathering data. The dataset created by Mohammad and Kiritchenko (2018) uses this representation of emotion to label the artwork available on WikiArt, and given it's relevance to both the domain of art, and emotion, this will be the representation first investigated.

The circumplex model of emotion will be investigated further with both the WikiArt Emotion dataset, which will use the valence-arousal (VA) decomposition presented by Kim et al. (2018) to evaluate the VA equivalence for each artwork. The process to convert the existing dataset's floating-point vector labels for emotion, to their respective VA values will involve applying the VA decomposition network to each of the dataset's images. Having both original vector and VA labels allows a direct comparison of the system's performance with both representations. A dataset has been created by Zhao et al. (2016) in which over 1,400,000 images are labelled according to the valence, arousal, and dominance values using image description text analysis. This dataset can be used in training an predictor architecture according to Kim et al. (2018) that can determine the VA values of a given image. This can then be applied to both the WikiArt dataset,and extended to a larger number of images available on WikiArt.

4.3 Generative architecture

The architecture through which images will be generated represents a core component of the entire creative system. Tan et al. (2017) specifies a GAN architecture which allows the propagation of a target style vector, and a process through which the assigned discriminator style label can be backpropagated to the generator network. Using such an architecture further allows the incorporation of emotion into the generator input.

5 Expected Outcomes & Contributions

The outcomes of this project will include a generative system with which art can be synthesized according to a target emotional profile. This system will be the combination of methods for the representation of emotion for use in a generative model, and an architecture with which such a model can be trained. Due to the exploratory nature of the project with respect to both system architecture and emotional profile representation, this research will have tested and analyzed various options and any comparative differences.

The proposed generative system will be used to create a collection of art, categorised by the target emotional profile with which they were seeded. The verification proposed involves the public exhibition of produced images, providing feedback to the generative process and pairing the generated images with a human assigned emotion label for use in any further system training.

References

Ali, A. R. and Ali, M. (2017). Emotional filters: Automatic image transformation for inducing affect. *arXiv preprint arXiv:1707.08148*.

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2017). Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Chen, M., Zhang, L., and Allebach, J. P. (2015). Learning deep features for image emotion classification. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4491–4495. IEEE.

Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

den Heijer, E. and Eiben, A. (2010a). Using aesthetic measures to evolve art. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE.

den Heijer, E. and Eiben, A. (2011). Evolving art using multiple aesthetic measures. In *European Conference on the Applications of Evolutionary Computation*, pages 234–243. Springer.

den Heijer, E. and Eiben, A. (2013). Maintaining population diversity in evolutionary art using structured populations. In *2013 IEEE Congress on Evolutionary Computation*, pages 529–536. IEEE.

den Heijer, E. and Eiben, A. E. (2010b). Comparing aesthetic measures for evolutionary art. In *European Conference on the Applications of Evolutionary Computation*, pages 311–320. Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Kim, H.-R., Kim, Y.-S., Kim, S. J., and Lee, I.-K. (2018). Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992.

Larsen, R. J. and Diener, E. (1992). Promises and problems with the circumplex model of emotion.

Lehman, J. and Stanley, K. O. (2008). Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336.

Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223.

Machado, P. and Cardoso, A. (2000). Nevar–the assessment of an evolutionary art tool. In *Proceedings of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, Birmingham, UK*, volume 456.

Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM.

Mathews, A. P., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mohammad, S. and Kiritchenko, S. (2018). Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Mouret, J.-B. and Clune, J. (2015). Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*.

Nguyen, A., Yosinski, J., and Clune, J. (2015a). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Nguyen, A. M., Yosinski, J., and Clune, J. (2015b). Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 959–966. ACM.

Pugh, J. K., Soros, L. B., and Stanley, K. O. (2016). Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40.

Ralph, W. (2006). Painting the bell curve: The occurrence of the normal distribution in fine art.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Ross, B. J., Ralph, W., and Zong, H. (2006). Evolutionary image synthesis using a model of aesthetics. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1087–1094. IEEE.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sims, K. (1993). Interactive evolution of equations for procedural models. *The Visual Computer*, 9(8):466–476.

Tan, W. R., Chan, C. S., Aguirre, H. E., and Tanaka, K. (2017). Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Yang, J., She, D., Sun, M., Cheng, M.-M., Rosin, P. L., and Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., and Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM.

Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., and Chua, T.-S. (2016). Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1385–1394. ACM.