Synthesising emotion-driven images using image emotion classifiers

Konrad Cybulski

April 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 (19; 39). 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.

This research aims to explore the use of neural network image emotion classifiers in synthesising emotionally-driven images. Primarily exploring the patterns and visual characteristics of such generative systems in their ability to create emotionally-targeted images. In doing so, compare categorical (happy, sad, etc.) and continuous (valence-arousal-dominance) representations for emotion in the context of both image classification, and synthesis. Through the generative process, gain a better understanding of visual patterns learnt by such a classifier by maximising classification confidence through the use of conditional generative adversarial networks (12), and evolutionary techniques explored by (24).

Methods for quantifiably representing emotion have been explored thoroughly in the domain of psychology, with continuous multi-dimensional models being used in lieu of a single emotional label. 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 (30). Such a continuous model is not without flaw, failing to accurately capture more complex emotional states, often those that represent concurrent conflicting sides of a given axis (15). The complexity of such a continuous representation of emotion has been investigated in depth and extended in various ways such as by Bradley and Lang (3), 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 (4; 35). Image emotion classifiers have been explored (14;

19; 5; 6) 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.

As with the use of deep neural network image classifiers such as ResNet, AlexNet, and Inception, the shapes and patterns learnt by such deep learning systems are difficult to extract. Recent work has looked at the use of quality-diverse generative algorithms with such deep classifiers (24; 25). The underlying patterns learned by the classifier can be surfaced by synthesising images that maximise various desirable features. This method allowed the exploration of images to which the classifier assigns a given label such as *school bus* or *lighthouse*, enabling a deeper understanding of the visual characteristics inherent to each class.

2 Aims and motivation

The aim of this research is develop an image emotion classifier in order to produce a generative system to synthesise emotionally-driven images to better understand the visual patterns associated with emotions conveyed in images as learnt by such a classifier. This will primarily leverage image emotion recognition architectures explored by Kim et al. (14), in combination with the dataset produced by Zhao et al. (41) containing over 1.4 million images with assigned valence, arousal, and dominance levels derived from their descriptions. Producing and comparing architectures with which valence-arousal (VA) values, or emotion labels can be assigned to images forms the basis for this research.

This base architecture enables the exploration into how VA values are assigned to more complex, multi-faceted and multi-layered images, and methods for incorporating this classifier into an image synthesis system. A generative process for synthesising emotionally-driven images will be explored that utilises the image emotion classifier produced. This generative system, and the patterns learned by it, will be better explored by maximising given target features in a quality-diverse way (25; 24), and through a conditional generative adversarial approach (32; 11). In the context of this research, such features include valence, arousal, dominance, happiness, sadness, etc. This will allow a greater 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 a focus on emotion-driven image generation.

3 Research questions

How can an image emotion classifier be used to synthesise emotion-driven images?

- How can the emotional content of images be represented?
- What methods can be used in combination with an image emotion classifier to synthesise images?
- How can patterns and visual characteristics learnt by an image emotion classifier be explored?

4 Background

4.1 Image emotion recognition

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 (19); and more recently using techniques such as deep neural networks (5; 14). Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal (7; 10; 8) and as an emotional feature vector for sentiment classification (19). Due to the artistic and psychological underpinnings used by Machajdik and Hanbury (19), 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. (39), which uses a comparably small feature vector to Machajdik and Hanbury (19), resulting however in a 5% classification increase to state-of-the-art approaches at the time.

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 (6; 36) with up to 200% performance gains over support vector machine classifiers.

Given the large amount of work conducted into image object classification, producing neural networks such as AlexNet, ResNet, and Inception, transfer learning has been heavily relied upon in various domains of classification and synthesis from images. Techniques used by image emotion recognition have been used extensively in domains such as image sentiment analysis (37), and image-to-text synthesis (33). The subjectivity and human-dependent nature of emotion has naturally resulted in only small datasets of emotion-labelled images. Through transfer learning, pre-trained image object classifiers can and have been used in the domain of image emotion (14; Wang and Lewis).

Sentiment is typically represented as a binary classification of either positive or negative (36; 6). Facial and image emotion classification however face a higher complexity problem, where research has typically focused on a subset of potential emotions ranging in size from 7 (1), 8 (19), 11 (Wang and Lewis), and even 19 (21). Other methods for representing emotion include use of the circumplex model of affective emotion (30; 3). This model represents emotion

as a continuous multi-dimensional space. The original definition by Russell (30) introduces a two-dimensional space defined by valence, and arousal. Valence represents the positive and negative aspect of an emotion; arousal, the excitatory component. In such a model, each emotion label used in common categorical classification resides within the space as seen in 1. The circumplex model has been applied and extended in various ways. The addition of a third dimension, dominance, introduced by Bradley and Lang (3) allowed the representation of emotion relating to feelings of situational control. Despite the increased freedom of representation in a continuous space, problems exist with the circumplex model (15). Some of the key issues with such a model is its dichotomous nature, resulting in a comparative failure to capture more complex emotions, particularly emotional states representing conflicting sides of a dimension.

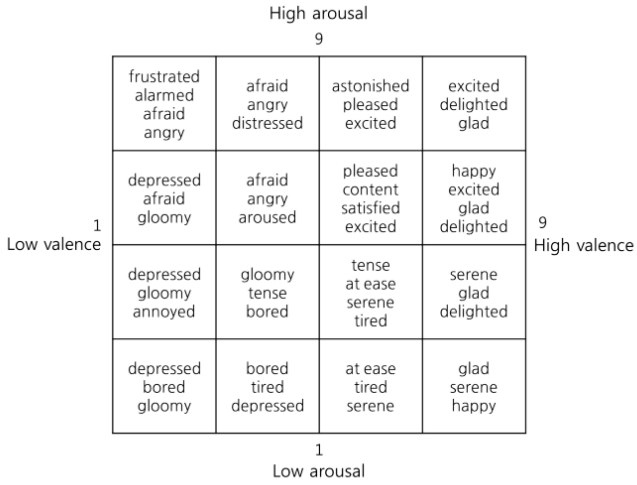

Figure 1: Distribution of emotions associated with levels of valence and arousal determined by the DNN classifier produced by Kim et al. (14)

The use of continuous emotion representations, particularly relating to the circumplex model have been explored in image emotion recognition tasks (14; 41; 40). Regression models produced to predict the valence-arousal (VA) values of given images have shown high accuracy on various datasets. In leveraging pre-trained image classification networks through transfer learning, even smaller datasets (10,000 images) can have high accuracy classification results (14). Recent datasets produced for image emotion recognition have used the valence-arousal-dominance model due to its continuity (41). While categorical classification is more easily verified by humans, training predictive models with data that has an element of noise and uncertainty benefits from both continuity and volume. This is a key advantage of the circumplex model as applied by Kim

et al. (14) and Zhao et al. (41) in image emotion recognition over categorical classification.

4.2 Computational image synthesis

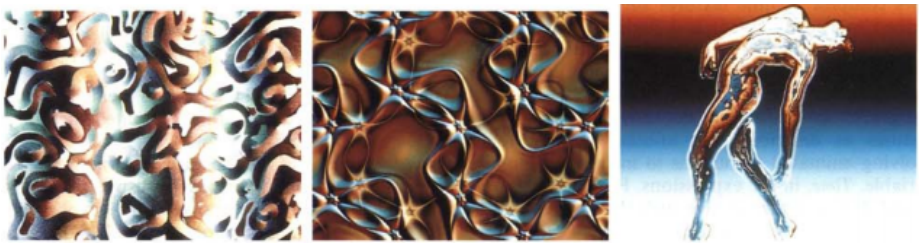

Figure 2: Images generated through the process of interactive evolution introduced by Sims (31)

Generative systems have been explored in various domains with numerous techniques. Examples range from the generation of images and art using evolutionary methods (31; 18), to the synthesis of text to describe a given image through the use of deep neural networks (20). Some of the first *human-in-the-loop* image synthesis systems such as *NEvAr* (18) produced greatly impressive images through evolutionary techniques. Evolutionary art leveraged methods introduced and exemplified by Sims (31) such as those shown in Figure 2. Sims (31) 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 (18; 31; 8; 9; 29).

Despite the slow nature of the interactive process, Sims (31) and Machado and Cardoso (18) were able to produce images with visually striking characteristics. Ross et al. (29) investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph (27), 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 (31) or Machado and Cardoso (18), 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. (13) allowed the process of image generation to depend only on collecting a sufficiently large dataset. Common GAN application has involved the generation of realistic images, such as has been done by Bao et al. (2), where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. (38) and Reed et al. (28) 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. (32) 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. (25) and Nguyen et al. (24) 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) (22) and Novelty Search (16; 17) 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 (26) such as maze navigation (17). Nguyen et al. (24) and Nguyen et al. (25) 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. (24) 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. (25) 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. (25) and Nguyen et al. (24) contrast greatly, the quality-diverse generative method used to understand the visual components learnt by the classifier show such an architecture's exploratory abilities. This technique for understanding the patterns learnt by such a classifier has not been explored in the context of regression.

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis, is limited. Sentiment-driven examples of generative systems include image captioning according to target sentiment (20). The task of describing an image was extended from a

traditional GAN approach through the addition of an sentiment target input. The method used to train such a generative system involved the conditional GAN architecture as described by Gauthier (12). A similar technique was used in style-driven image captioning (factual, romantic, humorous) in combination with a long short-term memory (LSTM) neural network model Gan et al. (11). In the context of image-to-image synthesis, *emotion transfer* was explored by Ali and Ali (1) which involved the transformation of an image's colour and style with the aim of altering its conveyed emotion.

5 Methodology

5.1 Emotion representation and prediction

Method for representing emotion in this research have been limited to three options based on the dataset: a single categorical label (happy, sad, etc.); a two-dimensional circumplex model (valence-arousal); and a three-dimensional circumplex model (valence-arousal-dominance). A model commonly used for representing emotion in classification tasks is that of a single categorical label, due its simplicity. While this model represents a reductionist view of emotion, omitting a great deal of complexity, it will be the emotional representation used in this investigation. Investigating the applicability of this simplistic model enables comparison between categorical and continuous representations of emotion explored in this work. The more complex models investigated in this research include both the two and three dimensional circumplex models: valence-arousal, and valence-arousal-dominance respectively. Each of these three models will be used as the classification/regression target of a deep neural network (DNN) model.

The DNN model architectures tested will be based on commonly used techniques such as those explored by Kim et al. (14); Chen et al. (5). This technique, known as transfer learning, involves the repurposing of pre-trained DNN image classifiers by replacing or feeding through the final layers to another network. The altered topology has the desired output shape and is able to proceed with further training in the given problem domain. Initial investigation into the classification of image emotion will involve the use of such pre-trained DNN classifiers as ResNet, AlexNet, Inception, and VGGNet. Further feature extraction and ensemble methods may be explored, particularly those investigated by Kim et al. (14), Chen et al. (5), and Chen et al. (5). The accuracy of architectures will be compared within the context of each emotional representation, and between them.

5.2 Image synthesis

Having produced an image emotion classifier, this research will then focus itself on developing an emotion-driven image generation system to both generate images, and explore the underlying patterns and shapes learnt by the predictor

network. Following from work by Nguyen et al. (25) and Tan et al. (32), this research will explore both the evolutionary quality-diverse algorithm MAP-Elites, in addition to neural networks for image synthesis.

In exploring the use of MAP-Elites for image generation, the system architecture used bases itself heavily on that introduced in Nguyen et al. (25). The feature-space is explored through the evolution of direct image representations, allowing crossover and mutation of individual and collections of pixel values, evaluating fitness according to the accuracy with which the image is classified. Given the quality-diverse focus of the MAP-Elites algorithm, the output will be a multi-dimensional space containing generated images distributed according to their label and classification accuracy. The multi-dimensional regression model spanning the valence-arousal-dominance (VAD) coordinate space however cannot be explored in the same way since there is no classification accuracy with which to assign fitness. The MAP-Elites algorithm will be applied in a similar way to investigate the use of such an evolutionary image synthesis method for later comparison. In this context, fitness evaluation will be computed using the error between target VAD values and those determined by the classifier.

The other method explored for image synthesis involves training a conditional generator neural network following from work by Tan et al. (32) and Gauthier (12). This process involves the training of a network that accepts as input: noise, and a condition. The condition with which this network is trained represents the target emotion of the generated image. This proposed architecture is trained to minimise error between the target emotion and that predicted by the classifier. There exists the option to use networks pre-trained on other datasets for the generator network, as is explored by Nguyen et al. (23). In doing so, enabling the generation of emotionally-driven images specific to a given domain.

In order to verify the system for emotion-driven image synthesis, the images generated by it will be involved in a human-validation process. This will involve a sample of generated images having VAD values assigned to them using the self-assessment manikin (3) and the analysis of how human-attributed values compare to those intended by the system.

5.3 Exploring emotion-specific patterns learnt by the image classifier

A process for image synthesis conditioned on a target emotion enables the exploration of generated images with regard to their emotion-differentiating patterns and visual characteristics. Techniques used by Nguyen et al. (24) and Nguyen et al. (25) to explore DNN image classifiers involved the MAP-Elites algorithm; generating images that maximise classification accuracy as discussed in *Background*. While these techniques used the MAP-Elites algorithm, the process of exploring the feature-space through image synthesis can also be completed by the neural network approach mentioned. The generative system produced as part of this research aims to generate images within a feature-space, that minimise the classification/regression error in order to better understand some

key visual patterns and characteristics learned by the predictive network. Since generative systems trained on given sets of data tend toward generating images similar to its training data, this generative exploration aims to incorporate as little prior learning into the process.

6 Timeline

Semester One	
Week 6	Research proposal.
Week 9	Literature review draft.
Week 12	Literature review.
Week 13	Compilation of data, ready for experimentation.
Week 14	Interim presentation.
Semester Two	
Week 3	Emotion representation predictive models trained.
Week 4	Comparison of image emotion predictive models completed.
Week 6	Image synthesis systems created and content produced for human-validation.
Week 8	Human validation of synthesised images complete.
Week 10	Further exploration of patterns learnt by classifier through generative exploration.
Week 12	Thesis completion.
Week 14	Final presentation.

7 Expected Outcomes and Contributions

The initial outcome of this project is an image emotion classification system. Multiple types of emotional representation are explored with reference to the classifier, both categorical emotion labels, and continuous multi-dimensional representations of valence, arousal, and dominance. Comparisons between the accuracy of representations with respect to the classification domain will be explored, as will the comparative performance of various network architectures used in the transfer learning approach. This section of the research will produce a usable, generalised, and extensible image emotion classifier.

The primary outcome of this research is the creation of an emotion-driven image synthesis system. It will have been developed exploring various generative methods as discussed in *Methodology*, and comparisons between them will be produced. The generative process will be computationally validated in its ability to generate images according to a target emotion, in addition to being validated by humans using methods for measuring affective emotion. The extensibility of techniques used for developing the image generator, particularly related to transfer learning from pre-trained image generators, will also be explored. The content produced by the generative system in order to explore and better understand the visual components learnt by the classifier will be qualitatively analysed with respect to technical, artistic and psychological image features as introduced by Machajdik and Hanbury (19). The synthesised images from the system developed will be exhibited to validate the efficacy with which it is able to produce targeted affective images.