Department of Computer Science

Submitted in part fulfilment for the degree of MEng

Computer learning of what is memorable from images

Predicting Visual Memory Schema using Conditional General Adversarial Networks

Anurag Mathur

18 June 2020

Supervisor: Dr Adrian Bors

ACKNOWLEDGEMENTS

Thank you to Dr Adrian Bors for supervision on this project.

STATEMENT OF ETHICS

There were no legal, social, ethical, professional or commercial issues to mention as the project was undertaken solely by me. Supporting work/research is properly accredited in the references and text.

TABLE OF CONTENTS

Executive summary .....
1 Introduction **1**
1.1 Aims and Objectives **1**
1.2 Structure **2**
2 Literature Review **3**
2.1 Section Overview **3**
2.2 Memorability Research **3**
2.2.1 Early days 3
2.2.2 Adoption of computing techniques 4
2.2.3 Most relevant research 5
2.3 Motivation **6**
3 Model Architecture Background **8**
3.1 Building Block 1: Artificial Neural Networks **8**
3.1.1 Perceptron Structure 8
3.1.2 Neural Nets (ANNs) 9
3.1.3 Backpropagation Learning 9
3.2 Building Block 2: Convolutional Neural Networks **10**
3.2.1 Structure and Layers 10
3.3 Building Block 3: General Adversarial Networks **12**
3.3.1 Framework and Explanation 12
4 Method **14**
4.1 Model and its use **14**
4.2 Actions taken **15**
4.3 Justifications **16**
4.4 Reproducibility and Ethics **17**
5 Results, Analysis and Critique **19**
5.1 Near-perfect examples **19**
5.2 Average examples **20**
5.3 Surprising results **21**

5.4 Poor results 23
5.5 Major Trends and Implications 23
5.6 Model Loss Graphs 24
5.7 Critique and Improvements 25
6 Conclusion and Further Work 26
7 Bibliography 27

TABLE OF FIGURES

Figure 2-1: Visual Memory Schema example from [10] 5
Figure 2-2: GANalyze memorability scroller example from [13] 6
Figure 3-1: Perceptron structure from [17] 8
Figure 3-2: Neural Net skeleton structure from [18] 9
Figure 3-3: High level CNN view from [20] 10
Figure 3-4: Conv. kernal matrix dot product visual from [21] 11
Figure 3-5: Visual of how CNNs learn at different layers [22] 11
Figure 3-6: GAN framework from [25] 12
Figure 4-1: High level view of how the model works (made for this project) 14
Figure 4-2: Pix2Pix Generator encoder-decoder architecture from [27] 15
Figure 5-1: Theme park example (near perfect) 19
Figure 5-2: Building example (near perfect) 20
Figure 5-3: Kitchen (average) 20
Figure 5-4: Living room (average) 21
Figure 5-5: View of a landscape 21
Figure 5-6: Golf court with identical bright regions 22
Figure 5-7: Park (poor result) 23
Figure 5-8: G and D loss graphs 24

Executive summary

The study of human memory has proven to be a real challenge throughout human history. However, it was not until the 18th century until a scientific approach to study human memory was devised by a young German philosopher, Herman Ebbinghaus. Researchers began treating the brain as a black box, where inputs were fed, and outputs analysed. This was done for many decades with some consensus on what makes humans remember. Then came the advent of Artificial Specific Intelligence (ASI), where it became possible for computers to use networks based on biological neurons to mimic human capabilities to a certain extent. Memorability research progressed with this new wave, and studies made use of the big datasets available in conjunction with convolutional neural networks in order to ascertain more concrete and reproducible results. However, the neural networks of today still cannot accurately reproduce the memory of a human. In order to create the next AI wave, the one of Artificial General Intelligence (AGI), it may be crucial to crack the mystery of human memory.

The aim of this work is to further previous research done in the field of machine learning and memorability. The objective for this work was to have a model reproduce an aspect of human memory. Instead of just studying data after the fact, the goal was to create new data without human input that resembled the intrinsic qualities of a person's memory. The item being created by the network was a Visual Memory Schema (VMS) of an image. This is a 2D histogram that represents the internal memory model a person may have of an image. Visually, the VMS would have rectangles drawn over the most memorable regions of an image. For example, an image of a child's birthday party would have the corresponding VMS showing a bright white box over the scary clown (brighter regions are more memorable). The real data containing image-VMS pairs is obtained from the VISCHEMA dataset. The network used would be a variant of a General Adversarial Network (GAN). These are based on neural networks, a core building block in AI, and allow for the generation of new data that mimics the statistics of the training data. The use of GANs is unique in the sense that no explicit memorability programming is done and so prior bias' are omitted.

The overall success criteria is simple in theory but difficult in practice: have a machine learning model successfully output a VMS of an image from VISCHEMA that has some resemblance to the

corresponding real VMS. In simple terms, this would mean that the network can recreate a similar internal memory model present only in humans (to date).

The methodology is straightforward. This is an image-image translation task, where the input is a real image and the output is a corresponding (fake) VMS. The Pix2Pix variant of GANs was adapted to make this possible. The original Pix2Pix model was fine-tuned to make this possible. This meant adjusting its structure to avoid common issues that GANs face (discussed) and hyper-tuning parameters for better results. The dataset was split into a training set (~700 images) and a test set (~100 images). No transfer learning was used despite the small training dataset due to the lack of applicable technology (discussed). Once the results were obtained, the output to the models (fake VMS) were visually compared to the real VMS on the test set.

The results indicated that the network was mostly successful. For many images, the regions identified by the fake VMS were nearly identical to those identified by the real VMS. This is to say that the model could identify which regions of an image are likely to be memorable, which is remarkable given the complex nature of that task. In the most extreme cases, the brightest regions were also identical, meaning that the most memorable region was identified separately to the rest. There were very few images that were misclassified entirely. The most common trend was that the network was not precise enough in pinpointing smaller regions. However, most of the regions identified in the test set did correspond to regions identified by the humans as seen in the real VMS, meaning that the network was not predicting randomly.

The work done here proves that it is possible for a machine to accurately identify memorable regions of images to an extent. Further work can be done on improving the model to improve its precision. A simple way to do this is to collect more training data using the methodology outlined by the VISCHEMA researchers. Another way would be to add a “memorability” constraint on the latent space of a GAN to get more accurate results. However, determining what this constraint should be and getting results without mode collapse may be challenging and require the use of newer GAN variations (discussed). Overall, the project was a success since the predicted images had a lot of resemblance to the real images despite the tough domain.

1 Introduction

The advent of Artificial Intelligence (AI) and Big Data has changed the way research is being conducted in many fields. Climate science is now closer to being able to predict natural disasters before they happen than ever before [1], machines can predict personality traits from Facebook user data [2] and people can unlock their phone using their face [3]. All this is possible due to the Artificial Neural Network (ANN), the core technology behind most modern AI systems. However, as ANNs become more embedded in our daily lives, the discussion of the next wave of Computing deepens. This is the wave of Artificial General Intelligence (AGI), systems that can optimise in tasks that are complex and cross-domain. To create such a system, inspiration is being sought from species that can do such tasks, i.e. humans.

However, the human brain is still largely a mystery. The work detailed in this report aims to further research done on human memorability using the most recent machine learning breakthroughs, namely that of General Adversarial Networks (GANs). These are generative models that have not been commonly used in memory research, despite their enormous potential to offer powerful visual insights into how we as a species remember.

The memorability aspect of this report will be concerned with Visual Memory Schema (VMS), which are 2D histograms that represent the internal memory model humans have when they look at images. A more thorough look into this topic and how GANs can be relevant will be provided.

1.1 Aims and Objectives

The primary aim of this project is to assess the feasibility of using General Adversarial Networks to predict VMS. This includes understanding current state-of-the-art GAN variations and their potential, while understanding the importance of VMS and how they may be different from prior research.

The objectives for this project are as follows:

- Understand memorability research conducted in the past few decades and the potential applications of computing methods can have to further such work
- From the research, devise a method that will illustrate the feasibility of using GANs for predicting VMS. This will involve explaining and justifying the technology used.

- Implement the method using a deep learning framework (with justification), following best practices where possible
- Prove that the model used completed training in an effective manner
- Have ways to discuss, given the results, if the aim of the project was met. For this project, this would involve having a qualitative visual comparison in tandem with some statistical analysis to show that the GAN works with memorability data

The success criteria will be to meet the aims and objectives stated. An additional success criterion would be to have a functioning VMS predictor generate images that resemble the internal memory model (VMS) of a human.

1.2 Structure

.

Firstly, there will be a literature review on the field of memorability and its eventual crossover with AI, followed by the motivation for pursuing this project. This order matters, as a lot of the motivation came from the literature.

Secondly, the building blocks of the model will be detailed. This is a crucial aspect of the report as it highlights the machine learning research done in the past decades to make this project possible.

Thirdly, the methodology will be stated. A description of how the model works will be present with reference to the building blocks explained in the prior section. A list of actions taken to meet the project objectives will be present. The justification the actions will be provided. A short subsection will show that the project is reproduceable and a few comments on the ethical and professional standard will be made.

Fourthly, the results will be given and analysed in relation to the success criteria. An overall evaluation of the project will be present, and improvements concisely stated.

Finally, there will be a conclusion which offers a synopsis of the project and further applications of this research will be mentioned.

2 Literature Review

2.1 Section Overview

It was important to read a lot of the literature published on memorability in order to find the different trajectories explored and get a feel on what should be done next. Although every paper I had read had some impact on choosing this project, only the most impactful papers are mentioned for the sake of clarity. This section will detail some memorability research done over the past few decades in chronological order, up to the most recent work. From this, some motivating factors for this work will be discussed with reference to the research. It will also become clear that the first objective has been met.

2.2 Memorability Research

2.2.1 Early days

Although the quest to understand human memory dates further than the era of Aristotle, the focus here will be on the work of the 20th century that is relevant to this project.

The first relevant paper is titled, “Short-term memory for complex meaningful visual configurations: a demonstration of capacity” [4]. This was published in 1965 by Raymond Nickerson who sought to investigate the ability of people to recollect images. This study is relevant as it was first to introduce the episodic memory test for images, a method that was commonly used in future research. As mentioned in the paper, the 56 participants were shown 200 photos with no duplicates followed by 400 photos containing some duplicates from the prior segment. During the second segment, they had to indicate which of the 400 photos were “old” and which were “new”. There was a delay of 5 seconds between each photo. This was defined as an episodic memory test which tested the ability of participants to recall images given extra noise (additional new images added into the mix).

A few years later in 1973, this method was adopted by Lionel Standing in his paper, “Learning 10,000 pictures” [5]. Standing used 4 variations of this method to conduct 4 mini-experiments on the capacity of a brain to remember images. The results were remarkable, as they suggested that our long-term memories can recall images from thousands seen previously. Albeit, the relevance here is in the fact that pre-computational memory research was very limited in scope. The data collected from these experiments were to do with recall times and recognition rates (% of images recalled correctly). There was no concrete analysis on why participants remembered certain images over others, rather simply a narrow statistical analysis using the data gathered. This is not a critique of Standing, as he was a product of his

time and the intention of the paper was clearly stated. However, the paper does resemble the kind of research done in this period, i.e. pre-computational techniques.

It is important to note that a years prior to Standing's paper, there was an attempt at a qualitative analysis that concluded that familiarity for a "class of pictorial stimuli improves recognizability of new members of this class" [6], but there were only 90 pictures and 46 participants used. The idea of big data and machine learning techniques to provide more concrete results was not possible until much later.

2.2.2 Adoption of computing techniques

The first paper of relevance here is "What makes an image memorable?", published by Isola et al. in 2011 [7]. The aim of this paper was to identify which features of an image make it memorable and train memorability predictors on those features. The same episodic memory test explained previously was re-adopted here in the methodology. A memorability score was assigned to each image, determined by the percentage of times the picture was correctly identified in the second sequence of the episodic memory test given to participants. Statistical and computational techniques were used to see correlations between image features and memorability scores. For example, it was found that the mean hue was a poor predictor of memory. This example alone explains the relevance of citing this paper. In the paper, through computer vision techniques paired with rank coefficient correlation graphs, it was found that as the hue transitions from red to green to blue to purple, the memorability decreases. A suggestion was that this may be due to "blue and green landscapes being remembered less frequently than more warmly coloured faces and indoor scenes". This kind of inference was only possible with the tools made available from the advancement of computing, and progressed human memory research significantly.

This second paper highlights the relevance of AI models in memorability research. "Understanding and Predicting Image Memorability at a Large Scale", published by Khosla et al. in 2015, was the first culmination of big data and convolutional neural networks (CNN) used in memorability research [8]. A detailed explanation on what CNNs are is available below as they are used in this project, but for now it is important to know that they are essentially machine learning models that are capable of understanding patterns in data. The dataset and model for this paper are referred to as LaMem and MemNet respectively. LaMem consists of 60,000 images and their corresponding memorability scores. MemNet was a CNN trained on

this dataset, and so understood the relationship between images and their memorability scores. Although the methodology of this approach is not entirely indicative of how humans remember images (discussed later), the paper gained traction online as people could upload their own images and get a predicted memorability score for it using MemNet [9]. The use of AI allowed for the first accessible real-world application of memorability research to be possible.

2.2.3 Most relevant research

The most relevant paper was titled, “Defining Image Memorability using the Visual Memory Schema”, by Bors et al. published in 2019 [10]. This paper provided the groundwork for my project as it offered an alternative, more realistic representation of human memory called the Visual Memory Schema.

A VMS at its core is a depiction of the internal memory model of a person. It is synthesised when after a standard episodic memory test, participants were asked to select parts of the image that made them remember it. The regions that were accurately remembered contributed to the true VMS, and the falsely remembered regions to the false VMS. Note that the VMS construction involved overlapping each participants selection with respect to a single image, which also meant that the most overlapped pixels were the brightest and hence most memorable regions. A combined VMS was constructed by overlapping the true and false schema, as can be seen in Figure 2-1.

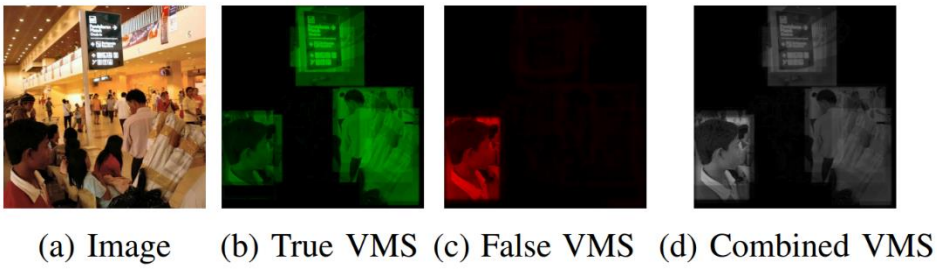

Figure 2-1: Visual Memory Schema example from [10]

Following this research, the VISCHEMA dataset was made available publicly that contained 800 image-VMS pairs [11].

Another relevant paper that acted as inspiration for this project came from MIT in 2019, titled, “GANalyze: Toward Visual Definitions of Cognitive Image Properties” [12]. This research uses a General Adversarial Network (GAN) variant called ‘StyleGAN’, to make an image more/less memorable. Although the results look realistic (as shown in Figure 2-2), memorability constraints embedded in the latent code of the GAN is based on MemNet [8] as an assessor. In light of

the VMS research done [10], it is now clear that the global image properties offer less of an insight into human memory than human-annotated local regions. This was a key reason for the focus of this project to be on the VISCHEMA dataset as opposed to other data and models.

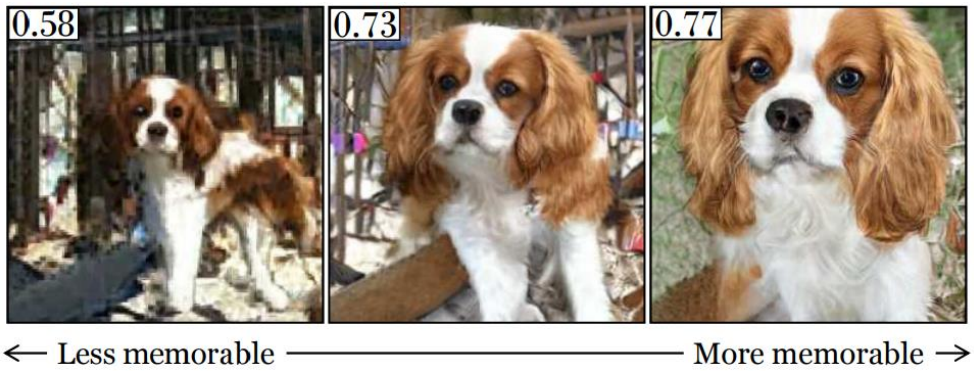

Figure 2-2: GANalyze memorability scroller example from [13]

2.3 Motivation

There were several motivating factors for pursuing a project in this field. The research published on memorability was the most influential. I had assumed that they were close to understanding the human brain and that an aspect like human memory was already close to being solved. From reading the literature, it became apparent that this was not the case. Simple black box experiments such as those of Standing's and Nickerson's [4] [5], although significant and necessary at the time, offered a very small piece to the overall puzzle. Fast-forward to after the first AI revolution and we can see that ANNs and CNNs added a plethora of fresh air into the mix. For the first time, memorability research was being applied; one could go online and get a % score for how memorable their picture is [9]. The idea that this is even somewhat accurate was a huge motivation.

I was also aware of what GANs were and how they worked due to media coverage of Deepfakes [13]. It became apparent that this generative modelling technique could be applied to memorability as memorability datasets such as LaMem were available [14]. The 'GANalyze' paper reinforced this belief as it showed one application of this technology in the field [12]. The VISCHEMA research then highlighted that the underlying basis of previous technologies, such as MemNet and GANalyze, may be flawed in the sense that they are not tailored enough to human memory [10]. The explanation of how local image properties noted by humans was a more accurate depiction of an internal memory model than computer generated memorability

maps convinced me that GANs could be applied to produce better and more memorable scenes in the long run [10]. Nonetheless, a crucial step along the way is to be able for a model to recreate the memory models present in humans, and so my project is based on this.

3 Model Architecture Background

This section will focus on explaining how the machine learning model used functions. The Pix2Pix Conditional GAN that is used is at the cutting-edge of current machine learning techniques, and so the section will focus on explaining some of the precursors to the technology (i.e. building blocks) and conclude with the framework that has been adopted to make Pix2Pix. Note that the building blocks are mostly explained at a high level in order to provide a clarity of explanation to the final model architecture shown in the Method. Understanding the concepts mentioned in this section is paramount to have a full grasp on the eventual methodology used and how the results were obtained.

3.1 Building Block 1: Artificial Neural Networks

Artificial Neural Networks (ANNs) aim to mimic parts of the human nervous system on a computer in order to learn the way humans do. They simplify the functioning of biological neurons and can achieve effective and remarkable results. The first computational model was created by McCulloch and Pitts (1943), who formulated the “all-or-none law” in their paper [15]. They abstracted away from the functioning of a biological neuron and created the same idea in propositional logic. This new artificial neuron, often called a perceptron, is the foundational block for most AI systems as it allows for the creation of ANNs.

3.1.1 Perceptron Structure

A perceptron can be visualised as a black box, where the input is finite, and the output is a single value as shown in the figure below.

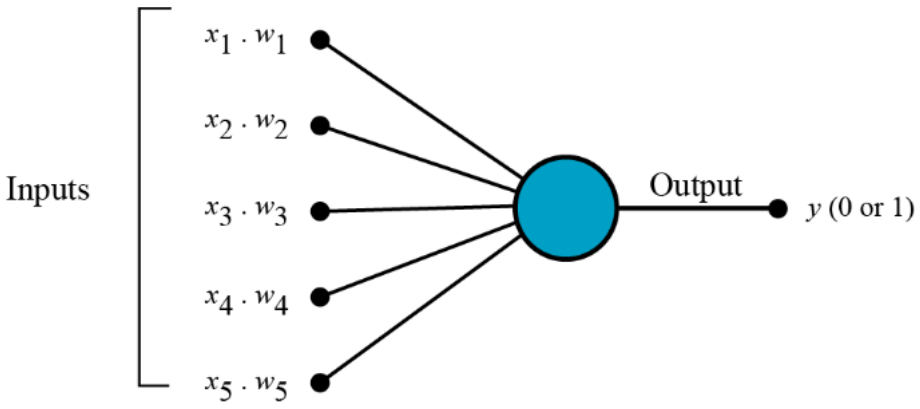

Figure 3-1: Perceptron structure from [17]

The inputs for the perceptron shown above are $X_1, \ldots, X_5$ and the output is binary. The weights $W_1, \ldots, W_5$ are learned by the perceptron through many data points and are changed using a learning algorithm such as backpropagation. They influence which of the inputs has a greater impact on what the output is.

3.1.2 Neural Nets (ANNs)

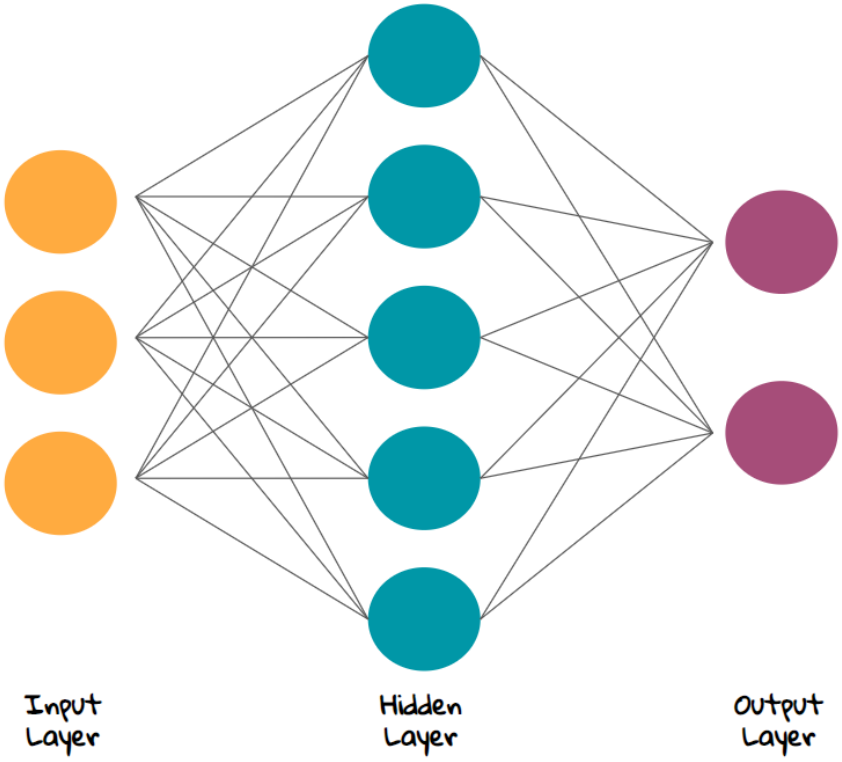

Figure 3-2: Neural Net skeleton structure from [18]

Fully connected ANNs are structures of perceptron's combined in a way where they form multiple layers. The first layer takes inputs and the last layers gives outputs. Each perceptron in every layer (except the output layer) is connected to perceptron's in the next layer, as shown in the figure above.

The idea of forming many layers of perceptron's is that the knowledge from one layer is passed onto the next, where a higher-level construct of the original input can be understood.

3.1.3 Backpropagation Learning

Backpropagation is the learning algorithm used in order to train these networks. It requires a solid understanding of calculus (mostly derivative math) and a full explanation is out of the scope of this work

but the original algorithm is available here [16]. The algorithm is discussed below in general terms for the purposes of understanding how the “learning” aspect of ANNs works.

ANNs have a cost function, which mathematises the difference between the model output given certain inputs and the real outputs of those inputs (part of the data). This function can be thought of as a curve. A method such as Stochastic Gradient Descent seeks to minimise this cost function by adjusting the weights of the model (initially random) to produce new outputs that move towards the point of the curve where the cost is minimal. Backpropagation is the algorithm that can iteratively calculate how much the weights need to be adjusted at each iteration to minimise the cost. Parameters such as the learning rate are set to determine how drastic of a change the algorithm can make to the weights at each iteration. If the learning rates are too high, the weight adjustments may be too large for the minimal cost point to be found; the optimal weights would be “skipped over”. This algorithm is key and is used throughout machine learning, including in this project.

3.2 Building Block 2: Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are more complex ANNs that are used for computer vision tasks. They are capable of complex pattern recognition tasks on images and can learn the relationship between images and their given output (with a large enough dataset). They are components of the GAN used for this project, so it is important to understand how they work in the general sense.

3.2.1 Structure and Layers

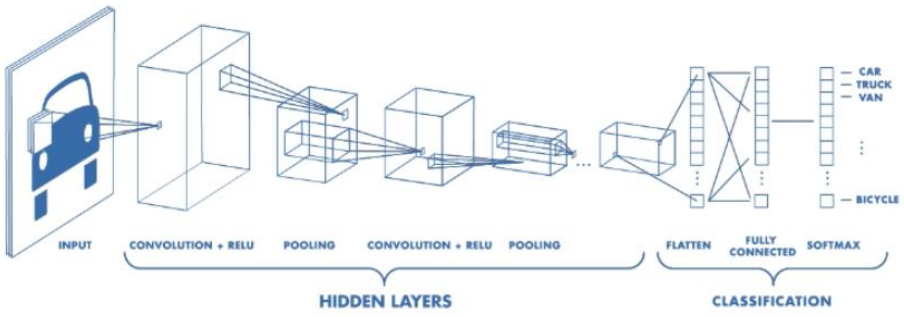

Figure 3-3: High level CNN view from [20]

The figure above shows that an input image matrix is fed into the network and is assigned a classification. Similar to ANNs, the CNN will initially assign a random classification if the weights for each perceptron were randomised initially. A cost function will be based on a function of the difference between the target output (real label classification, in this case a car) and the model output (fake output). This will then be optimised on by backpropagation in order to determine the optimal weights for the model to minimise the cost, similar to what happens in an ANN. The difference in the structure of a CNN lies at the layer-level.

The first convolutional layer passes a convolutional kernel (also known as a feature map) over the matrix to extract high-level features such as lines and edges. This is by done by using matrix dot products as shown in Figure 3-4. The consequent convolutional layers repeat the process of extracting different high-level features (using different feature maps) until the low-level commonalities between images (Figure 3-5)

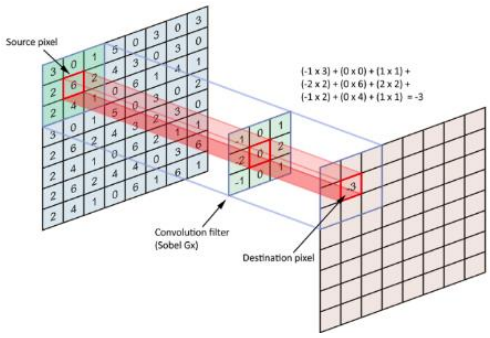

Figure 3-4: Conv. kernal matrix dot product visual from [21]

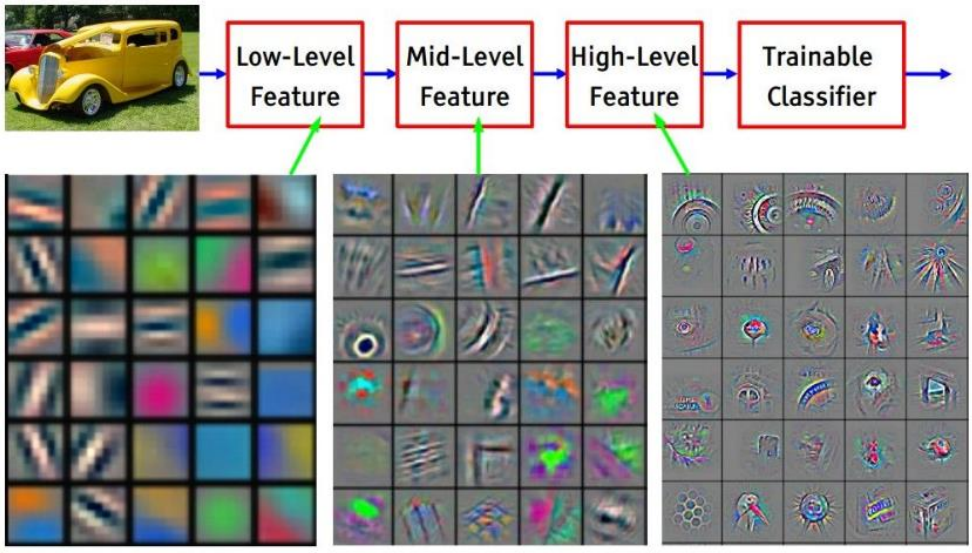

Figure 3-5: Visual of how CNNs learn at different layers [22]

Between each Convolutional layer, a Pooling and ReLU operation is typically done. The Pooling layer decreases the computational power needed to process data through dimensionality reduction [17]. Moreover, it extracts dominant features that are not affected by their position/rotation. ReLU stands for Rectified Linear Unit. It is an activation function used to make the model converge faster and make the entire operation cheap to compute [17].

3.3 Building Block 3: General Adversarial Networks

A GAN is a generative model created by Goodfellow et al. in 2014 that uses game theory and CNNs to generate new data based on training data remarkably well [18]. Although there are many variations of a GAN nowadays, the core concept behind them all is important to understand for this project.

3.3.1 Framework and Explanation

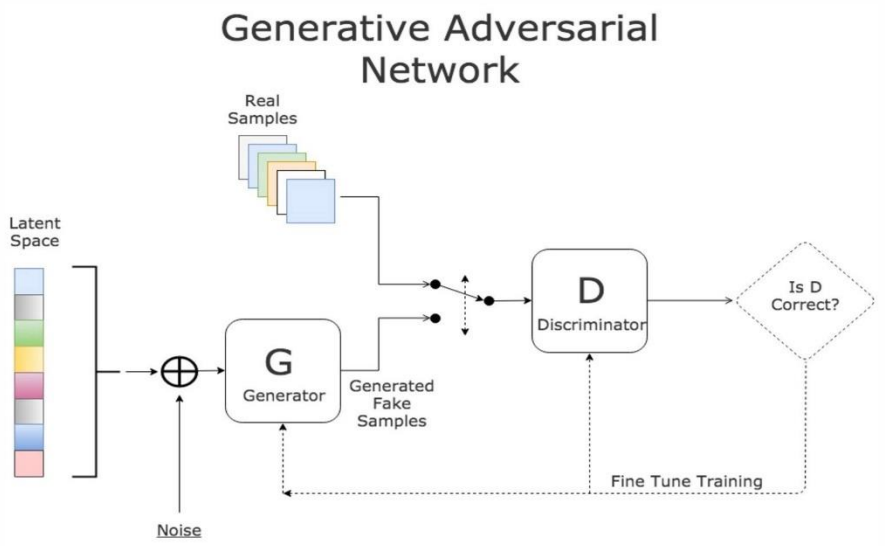

Figure 3-6: GAN framework from [25]

Key terms:

- **Latent Space (Z):** a fixed-length vector initialised from a Gaussian distribution that acts as a seed for the generative process.
- **Generator (G):** CNN with Z as the input and a fake image as output.
- **Discriminator (D):** CNN that takes an image as input and classifies it as being either real/fake.
- **Real Samples:** Images from the dataset
- **Fake Samples:** Output images from the G

G and D are trained separately. Initially, D's weights are adjusted by having D classify images as either being real/fake. For example, the real samples will be labelled '1' and the fake samples from G labelled '0'. This stage is tantamount to training a single CNN through backpropagation. Note that only D's weights are updated for this round.

Next, the weights of D are frozen, and the process is re-run. The difference at this stage is that the weights of G are updated in order to fool D into misclassifying the fake images as real. Therefore, the network is adversarial, since D is being fooled by adversarial examples. The loss from D is calculated and backpropagation is done through both G and D in order to obtain the gradients, but only the weights of G are updated [18].

Now we have seen how D and G are trained. To train a GAN, D must train for one/more epochs (iteration) followed by the training of G for one/more epochs. The weights must be frozen (as discussed above) with respect to which of G/D is being trained [18].

After the training is complete, the idea is that G and D will have a similar loss and the fake outputs of G generated from a latent vector Z will mimic the statistics of the original real sample set [18].

4 Method

4.1 Model and its use

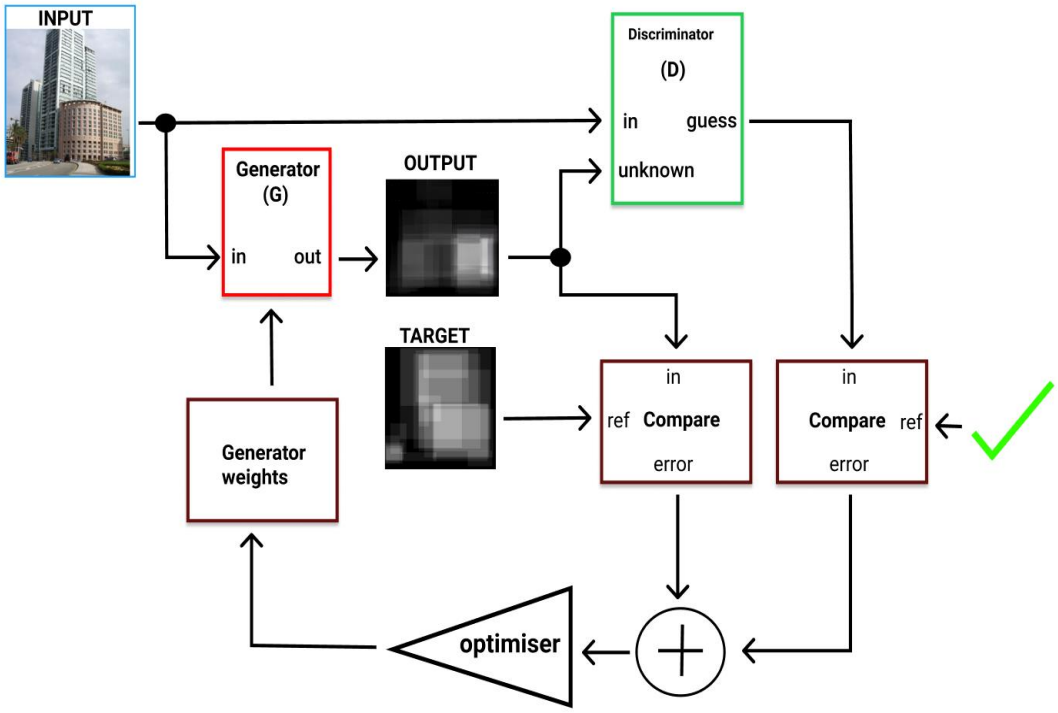

Figure 4-1: High level view of how the model works (made for this project)

As discussed above, GANs are very effective tools at generating images to mimic a dataset. However, image-image translation tasks require additional complexity since there are pairing relationship must be understood. The Pix2Pix framework [19] builds on top of previous work into Conditional GANs but is uniquely non-generic, i.e. can be applied to any dataset. This is the reason it was used for this project.

As shown in Figure 4-1, the model is essentially a Conditional GAN.

The aim of G is to transform the input image into the target image. It does so with an encoder-decoder architecture as shown in Figure 4-2

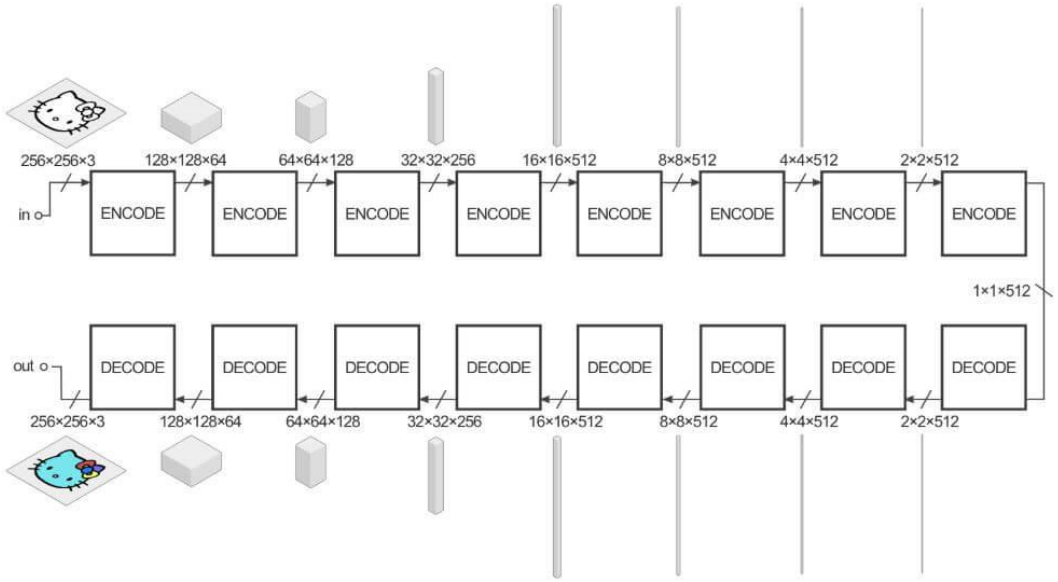

Figure 4-2: Pix2Pix Generator encoder-decoder architecture from [27]

As stated in the paper, the encoder-decoder architecture are as follows:

encoder: C64-C128-C256-C512-C512-C512-C512-C512

decoder: CD512-CD512-CD512-C512-C256-C128-C64

where C [NUM] stands for a Convolution-BatchNorm-ReLU layer with [NUM] filters and CD[NUM] denotes a Convolution-BatchNorm-Dropout-ReLU layer with a 50% dropout rate [19].

In this case, G took as input an image from VISCHEMA and generated a fake VMS.

The aim of D was to take as input either a real/fake VMS and determine whether it was real/produced by G. The training for this model was done in the standard way GANs are trained (as mentioned earlier), essentially by training the Discriminator and Generator step-by-step ensuring that the opposite CNNs weights are frozen while the other is being trained.

This is evidence that the research from both memorability and machine learning was combined in order to devise an appropriate methodology.

4.2 Actions taken

Here is a full list of actions I had to take in order to understand and use the model for the VISCHEMA dataset:

- Understand a deep learning framework (Pytorch) completely so that I could make use of cutting-edge technology and research

- Immersing myself in the memorability literature to understand the potential that such work could have

- Experiment with different GAN architectures to select the appropriate model.

- Data cleansing. The VISCHEMA data had to be put into specific training/testing folders in order to work.

- Once the Pix2Pix CGAN was selected, going through the source code and making the appropriate changes was necessary for this project to work. Some of the changes included:
    1. Cropping was originally random and was not the same for target and input image. This was made consistent, so the same random crop is generated for both images.
    2. Similarly, flipping was not consistent for target and input image. This was made consistent.
    3. The data loader had a bug where it was not loading the target image corresponding to the input image. This was made consistent.
    4. Edited the script to output images in the desired format.
    5. There were some instances of model collapse, so hyper-parameter tuning such as adjusting the learning rate were made.
    6. Graphs of G vs D loss were plotted to ensure the model is stable during training

- Once the training had finished, using G to predict on the test images from VISCHEMA

This methodology was important to state in order to demonstrate that the implementation objective has been met.

4.3 Justifications

The reason for choosing to code in Pytorch was because of personal preference. It seemed more intuitive than other libraries (TensorFlow) and offered more functionality.

The decision to not use transfer learning was also deliberate. Transfer learning is good for small datasets, where pre-determined models trained on larger datasets can be applied to smaller datasets to achieve results that could not otherwise be achieved. However, the weights of the model are largely a factor of what big data they were trained on. In this case, the VISCHEMA data [11] differs drastically from other memorability datasets such as LaMem in the sense that it has human annotations. This meant that no previously trained CGAN was directly applicable to this data. The only point of using transfer learning would be if the dataset was too small to achieve sensible results. However, in this case, good results were achieved regardless, and the use of transfer learning may have had no/negative impact.

The hyper-parameters were kept mostly the same and thus are not mentioned much. The model was made to work on a variety of datasets and so no significant changes were needed. A full list of the parameters used is available via running the code attached with this project.

For the scope of this work, only the Global VMS were predicted. As stated earlier, the aim of this work is to assess the feasibility of applying GANs to predicting VMS. The model will not change based on which type of VMS are used, and hence only 1 type is used.

The Pix2Pix GAN was chosen due to its appropriateness for the data. That is, it can perform image-image translation on paired training data. Were the data to be unpaired, a CycleGAN would have been used. All the other GAN variations were considered but not used due to their lack of suitability [20]. An added benefit to using Pix2Pix was the thorough documentation available with the official implementation [21], which could not be said of the other GAN implementations.

4.4 Reproducibility and Ethics

It was important to me that this work be reproducible given the number of parameters and ambiguity involved in training GANs. For this reason, a Python notebook is attached with this report that can be run on Google Collab. Only an internet connection and Google account are required to run the code. The code will do the following:

1. Clone my GitHub repository containing the Pix2Pix CGAN used with the amendments I have made.
2. Download the VISCHEMA dataset (may take a while)

3. Extract and prepare the dataset in the required format
4. Train the model
5. Test the model
6. Generate graphs of the training losses

All the hyper-parameters are available in this Python notebook and I highly recommend running the code to properly understand the methodology.

It is important to note that I have not created the model from scratch nor am claiming to. The framework and Pytorch implementation for Pix2Pix already existed and has been adopted for the purpose of this project. The credit for the concept has been given [19] and the credit for the code is in the paper [21]. Nonetheless, understanding how it works and adapting the code to make it work on the VISCHEMA dataset to produce meaningful results in the field of memorability required significant time and effort on my part. The changes made to the original code can be found on my GitHub by looking through the commit history [22].

5 Results, Analysis and Critique

Attached to this project are 50 real life images from VISCHEMA and 50 corresponding fake VMS generated by my adapted model. More fake data can be generated by running the cells in the attached Python notebook on Google Collab. For the purposes of this section, I will showcase some of the results and provide analysis.

5.1 Near-perfect examples

Some of the fake VMS generated were almost identical to the VMS made from the human data.

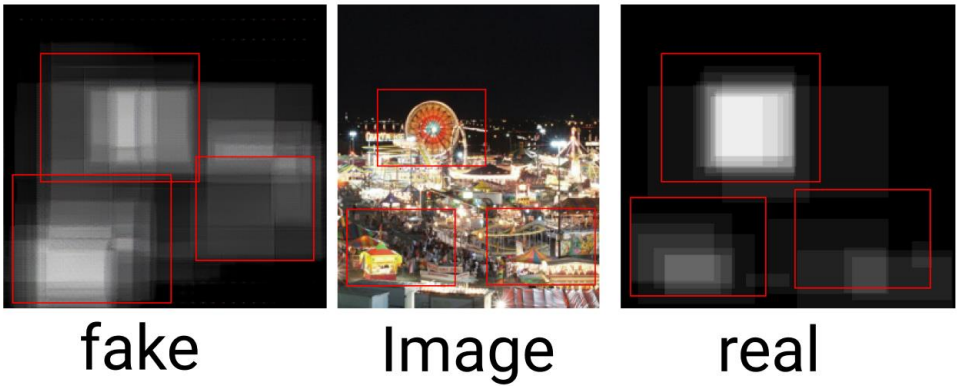

Figure 5-1: Theme park example (near perfect)

From Figure 5-1, we can see the real potential of this technology working at its peak. The regions identified on the fake VMS are identical to those on the real VMS. From the image, we can that the Ferris wheel (most memorable region) is clearly identified in both the schema as well as the stall in the bottom left corner. Intuitively, this makes sense as they clearly stand out as being 2 bright regions with a notable attraction. The region on the right-hand side is in a slightly different position although it is not as bright as the other regions and is hence less memorable anyway. This could suggest that the model has learned a trade-off between positioning its schema bounding boxes relative to the brightest (most memorable) region.

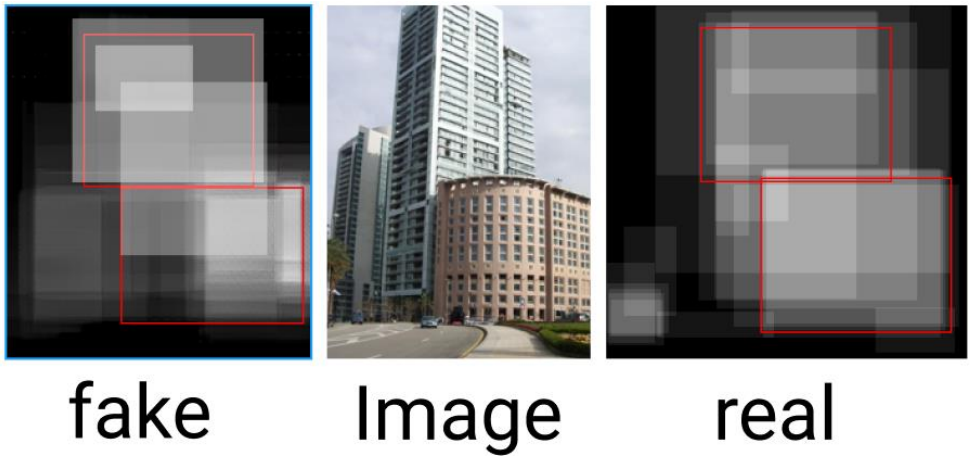

Figure 5-2: Building example (near perfect)

Figure 5-2 shows that the fake VMS has identified the 2 buildings present in the real image. Traces of the far-left building are picked up as being memorable by the model whereas only the bottom left corner was picked up by humans. The level of brightness is somewhat similar in both VMS which suggests that the degree of memorability has been learned by the model.

5.2 Average examples

Most of the fake schema generated follow a similar trend. They identify some (not all) memorable regions of an image but cannot identify the brightest spots. Moreover, the identified regions are generally larger and less precise than their real counterparts.

Figure 5-3: Kitchen (average)

For this first example (Fig. 5-3), the right-hand side of the image is memorable compared to the left. The fake VMS shows us evidence that the model has understood this. The real VMS indicates that humans found the top open cupboard and the bottom open drawer the most interesting in the image. However, the fake output cannot

differentiate well enough to separately identify the top and bottom region. Instead, it just identifies a broader segment of the image as being memorable.

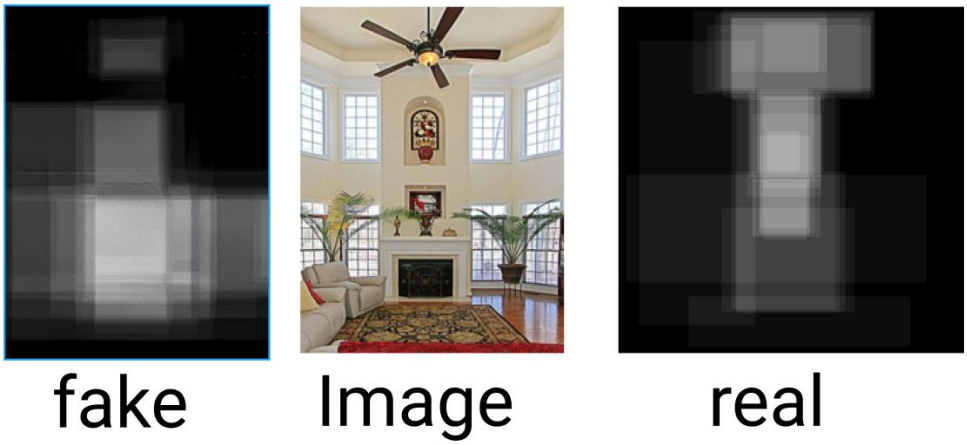

Figure 5-4: Living room (average)

In the above living room example, we can see yet another average example. Once again, the trend continues. The model identifies the middle regions as memorable but cannot precisely pinpoint the objects in the room that cause the memorability in the same way that humans can. Humans most noticed the fan and the artwork above the fireplace as well as the fireplace itself. In comparison, the model identified much broader horizontal regions around the fireplace and only some vertical regions.

5.3 Surprising results

The biggest surprise came from the fact that landscape images, typically found to be the hardest to classify [10] [7] [8], did surprisingly well.

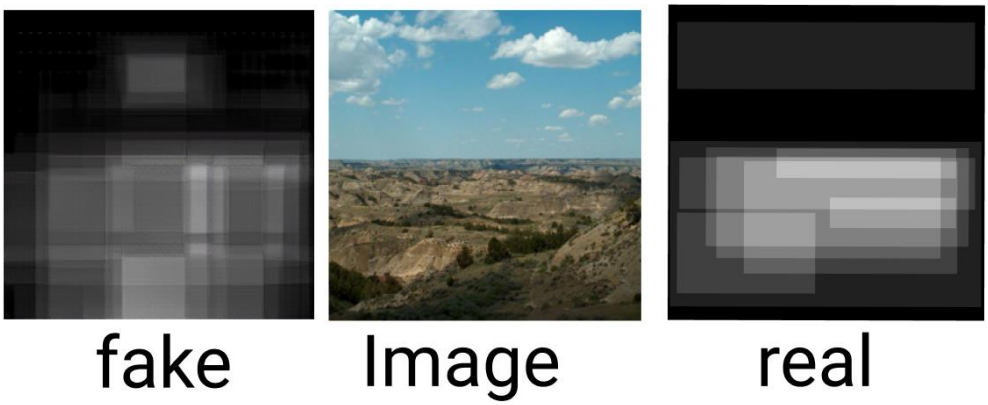

Figure 5-5: View of a landscape

Although not precise, Figure 5-5 shows some similarity between real and fake. The top region, identified in both, is also segmented in both. From the image we can see this to be the clouds being separated from the lower identified regions. There is some ambiguity, however. The real VMS can draw a clear distinction between the land further away being more memorable than that close by, whereas the fake VMS cannot. Nonetheless, the accurate segmentation on a class of images that is hard for humans to classify was somewhat surprising.

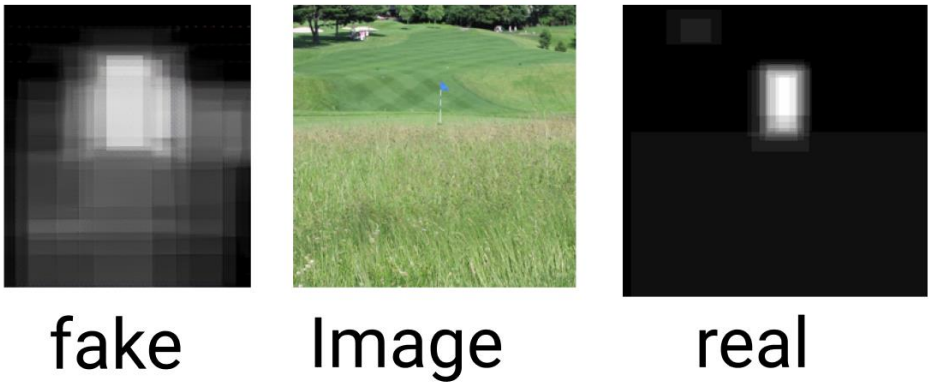

Figure 5-6: Golf court with identical bright regions

The golf court images above show the same trend. There is very little to notice let alone remember from the image. The real VMS shows that almost all the participants noticed only the faded blue flag in the middle of the picture. Although the fake output looks messy in the sense that many regions are identified, the brightest region is distinct from the rest and directly correlates with what was identified by the humans. This is interesting as it shows that the model can make the correct prediction but identify non-relevant regions at the same time.

5.4 Poor results

Although rare, it is important to note the fact that some results were completely random and had no resemblance to the real image / real VMS. Only a single example is provided since there is not much to analyse given the random nature of detection. A few more can be seen from the attached results of 50 images.

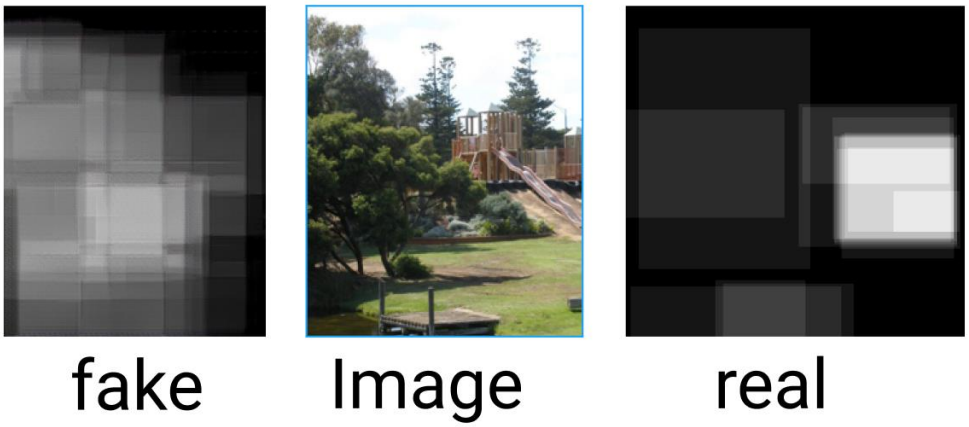

Figure 5-7: Park (poor result)

5.5 Major Trends and Implications

1. Memorable regions for landscape images were often over-detected although if a bright region existed it was almost always detected as most memorable by the model.
2. Cluttered indoor scenes were detected accurately but imprecisely whereas non-cluttered indoor scenes were detected both precisely and accurately
3. Distinctive outdoor buildings and objects that were separated were identified both precisely and accurately. If they were nearby then the boxes would overlap and detect their midpoints.
4. Poor results, although rare, occurred across all categories and did not share common traits.

   The results of this project illustrate that human annotated memorability maps of images can be reproduced by GAN technology with a high degree of accuracy in a lot of cases, albeit varying degrees of precision. Contrary to previous experiments [8], the memorability classification of landscape

images seems to be relatively accurate. Although there are many regions across many images which are not precisely classified, the results of this work may encourage others to apply GAN-related technology in the field of memorability. Perhaps some previously held common notions on the ability of machines to mimic human memorability are dispelled.

5.6 Model Loss Graphs

Proving that the model works effectively was an objective of this project.

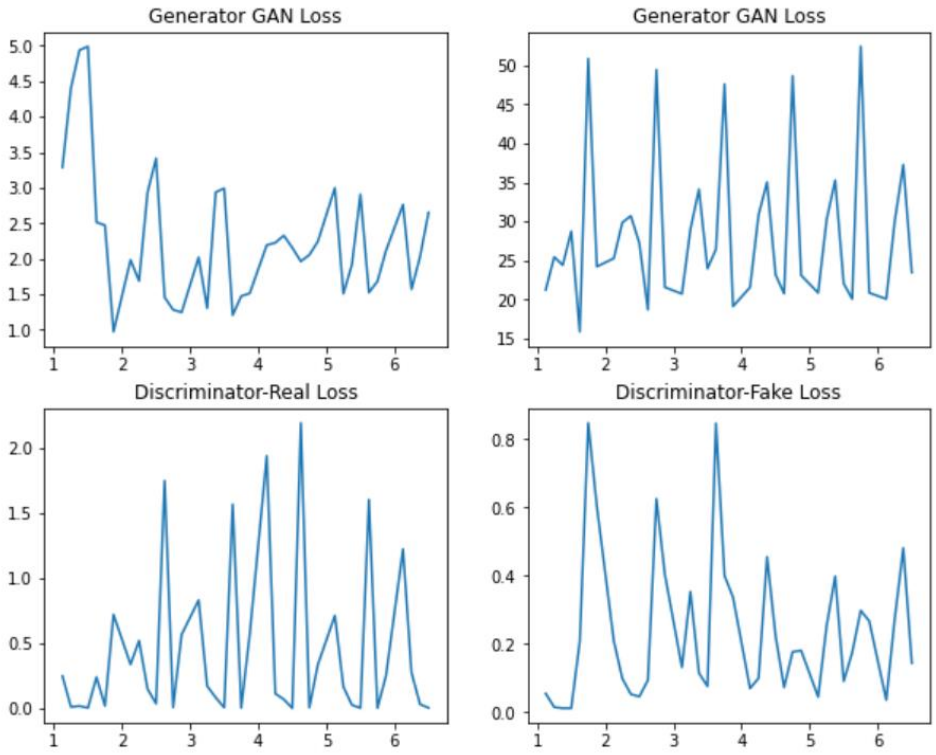

Figure 5-8: G and D loss graphs

As mentioned earlier when explaining how GANs work, the job of G is to fool D into classifying the fake outputs as real images. From the graphs above, we can see the model was converging through the epochs (shown on the x-axis) and so there was no model collapse and that G's weighted outputs are based on the training data and not random.

This is evidence that the implementation aspect of my objectives has been met. It is also clear that a thorough visual analysis has been conducted with the major trends identified.

5.7 Critique and Improvements

Given that the project had an enormous scope, i.e. the application of cutting-edge machine learning technology to the most recent human memorability research, it was handled well. A lot of skills had to be learned and a lot of papers had to be read. Nonetheless, given what I know now, here are some possible improvements and critique:

- Add a memorability constraint to the GAN. This is perhaps the most influential improvement one could make. Currently, the model is not optimised to learn memorability. The model as it stands just looks for patterns in the images, in the way traditional CNNs do. The fact that the data being used is more human-orientated and this is the first GAN-related project to do with the data is an obvious upside for my work. However, having a constraint on the latent space on the GAN would severely improve the results. Although it can be expected that the results may be more unstable. One way to get around this would be to research Wasserstein GANs and the loss metrics they use. These improvements would most likely improve the results significantly.
- Have a larger dataset. This is an obvious improvement since the more data allows the model to find more patterns and will most likely lead to better results. This may solve the problem of precision within regions.
- There was more application of machine learning that could have been done. A loss function could have been made between the fake and real VMS to provide further insight into accuracy or perhaps even act as a memorability parameter for the first idea mentioned.
- The comparison between the real and fake VMS was done manually, which is an obvious flaw since some details will be missed. Although this is a critique, the improvement remains unclear.
- For this project, only the global VMS from VISCHEMA were used [10]. If the false/positive schema were also analysed they may have provided a deeper insight into memorability. However, visually finding patterns across the fake/real versions of these images will take a lot more time.

6 Conclusion and Further Work

The first success criteria for this project was to meet all the objectives specified. As has been made clear by the work above, these have been met. A thorough literature review spanning the past few decades was conducted with respect to memorability. The advancement of early technology and the eventual implementation of machine learning techniques in the field was also noted. The technology I planned to use was clearly explained from its foundations in neural networks to an eventual high-level overview. A clear explanation on how the model would interact with the memorability data I was using was provided. The original model borrowed from literature was adapted for the dataset I used, and each step of the procedure was noted. Ethical and professional standards were discussed in suitable detail. The procedure was justified at every opportunity. The results were presented in a clear and succinct fashion and an appropriate and thorough analysis was conducted. It was clearly illustrated that GANs make good predictors of VMS and show surprising results in some cases. These results also confirmed my second success criterion, as they showed realistic predictions in many cases. However, many critiques and improvements were mentioned, such as imposing a memorability constraint on the GAN's latent space to get better results. In conclusion, given that the project used cutting-edge machine learning technology and successfully applied it on the most current memorability research, I note this project as a success.

There is a lot of further work that can be done before. The memorability predictor used in this project can be used for a lot of other projects. It can also be further optimised by using another GAN and creating an auxiliary loss function between the fake and target output schema. If more data containing image-VMS pairs is released, the model could be significantly improved. If a similar GAN is trained on another dataset containing human annotated data, transfer learning can be used to improve the results of this model. All in all, the field of memorability and machine learning is changing so rapidly that the use cases/improvements for this project is only going to increase.

7 Bibliography

[1] J. Snow, “National Geographic,” 19 July 2019. [Online]. Available: https://www.nationalgeographic.co.uk/science-and-technology/2019/07/how-artificial-intelligence-can-tackle-climate-change. [Accessed May 2020].

[2] T. Tandera, “Personality prediction systems from Facebook Users,” *ICCSCI,* 2017.

[3] Apple. [Online]. Available: https://support.apple.com/en-gb/HT208108. [Accessed March 2020].

[4] R. Nickerson, “Short-term memory for complex meaningful visual configurations: A demonstration of capacity,” in *Canadian Journal of Pyschology* , 1965.

[5] L. Standing, “Learning 10,000 pictures,” in *Quarterly Journal of Experimental Psychology*, 1973.

[6] A. Goldstein, “Visual memory recognition for complex configurations,” in *Perception and Physchophysics*, 1970.

[7] P. Isola, A. Oliva, A. Torralba and X. Jianxiong , “What makes an image memorable?,” in *CVPR*, 2011.

[8] A. Khosla, A. Raju, A. Oliva and A. Torralba, “Understanding and Predicting Image Memorability at a Large Scale,” in *ICCV*, 2015.

[9] A. Khosla, “LaMem Demo,” [Online]. Available: http://memorability.csail.mit.edu/demo.html. [Accessed March 2020].

[10] A. Bors, K. Evans and E. Akagunduz, “Defining Image Memorability using the Visual Memory Schema,” *IEEE transactions on pattern analysis and machine intelligence,* 2019.

[11] “Defining Image Memorability using,” University of York, [Online]. Available: https://www.cs.york.ac.uk/vischema/. [Accessed May 2020].

[12] A. Oliva, P. Isola, A. Andonian and L. Goetschalckx, “Toward Visual Definitions of Cognitive Image Properties,” MIT CSAIL, 2019.

[13] I. Sample, “Guardian,” [Online]. Available: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them. [Accessed February 2020].

[14] MIT, “LaMem Dataset,” [Online]. Available: http://memorability.csail.mit.edu/explore.html. [Accessed March 2020].

[15] W. McChulloch and W. Pitts, “A Logical Calculus of Ideas Immanent in Nervous Activity,” *Bulletin of Mathematical Biophysics,* 1943.

[16] G. Hinton, D. Rumelhart and R. Williams, “Learning representations by back-propagating errors,” *Nature,* 1986.

[17] “An Intuitive Explanation of Convolutional Neural Networks,” [Online]. Available: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/. [Accessed May 2020].

[18] I. Goodfellow, “Generative Adversarial Networks,” *NIPS,* June 2014.

[19] P. Isola, . T. Zhou, A. Efros and J.-Y. Zhu, “Image-to-Image Translation with Conditional Adversarial Networks,” *CVPR,* 2016.

[20] E. Lindernoren, “GitHub Pytorch GANs,” [Online]. Available: https://github.com/eriklindernoren/PyTorch-GAN#conditional-gan. [Accessed March 2020].

[21] P. Isola, “GitHub,” [Online]. Available: https://github.com/phillipi/pix2pix. [Accessed May 2020].

[22] A. Mathur, “GitHub PRBX Project,” [Online]. Available: https://github.com/anura-g/PRBX.

[23] [Online]. Available: http://ganalyze.csail.mit.edu/.

[24] S. Sharma, “What the hell is a Percptron?,” [Online]. Available: https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53. [Accessed April 2020].

[25] “Neural Network,” [Online]. Available: http://mindforcesys.com/neural/. [Accessed April 2020].

[26] S. Saha, “A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way,” [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-

convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed May 2020].

[27] D. Cornelisse, “An intuitive guide to Convolutional Neural Networks,” [Online]. Available: https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/. [Accessed May 2020].

[28] A. Sagar, “Convolutional Neural Network for Breast Cancer Classification,” [Online]. Available: https://towardsdatascience.com/convolutional-neural-network-for-breast-cancer-classification-52f1213dcc9. [Accessed March 2020].

[29] D. G. Mosquera, “GANs from Scratch 1: A deep introduction.,” [Online]. Available: https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f. [Accessed May 2020].

[30] “Pix2Pix – Image-to-Image Translation Neural Network,” [Online]. Available: https://neurohive.io/en/popular-networks/pix2pix-image-to-image-translation/. [Accessed May 2020].