COMPUTATIONAL AESTHETIC EVALUATION USING CONVOLUTIONAL NEURAL NETWORKS

Teddy Knox

Adviser: Professor Christopher Andrews

A Thesis

Presented to the Faculty of the Computer Science Department

of Middlebury College

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

May 2012

ABSTRACT

The accuracies of state-of-the-art machine learning techniques have exceeded the accuracy of humans at certain limited tasks, suggesting the computers will soon be able to complete many sophisticated tasks competently. This study tested the accuracy of state-of-the-art convolutional neural networks for modeling the statistically-defined visual aesthetic tastes of randomly generated triangle art. Our trained model demonstrated no predictive power at this task. Further research will be needed to interpret this result.

ACKNOWLEDGEMENTS

I dedicate this paper to science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Machine learning is a broad discipline of computer science that has seen rapid advances in recent years, and has the potential to impact all human enterprises. The latest algorithms are capable of competitive or even better-than-human performance [citation needed] on certain tasks, and their applications are increasing in proportion to the availability of training data. One such application will be in the more field of generative art and design. Generative art is art that is created procedurally, often with the help of a computer. Until recently, the use of computers in art and design has been limited to rote tasks such as preparing renderings and generating randomness, but advances in machine learning may soon lend computers to more high-level aspects the creative process, such as creative exploration and Computational Aesthetic Evaluation. Computation Aesthetic Evaluation (CAE) is the problem in generative art research of programming a computer to subjectively differentiate between aesthetically pleasing and displeasing content. This problem is important to the field of generative art for its potential to automate searches for aesthetic solutions, and would also be useful for building personalized recommendation systems. This research investigates the applicability of state-of-the-art general-purpose pattern recognition algorithms for Computational Aesthetic Evaluation. This experimentation is motivated by an interest in discovering the limits of these impressive models. By evaluating their accuracy on unconventional datasets of unknown complexity, we hope to gain a better understanding of both the model's capabilities and the hidden structure of the dataset. This paper first provides background and formulates the problem, then explains the experimentation in detail, and concludes with a discussion of results and future uses of machine learning models for the analysis of artistic content.

Convolutional Neural Networks (CNNs) are a type of model inspired by the biology of the human visual processing center, and have been shown to be the most effective

model for image recognition to date. CNNs are a type of feed-forward artificial neural network architecture formed by layers of are convolutional filters, trained to derive high-level features from low-level pixel data. These high-level features are powerful and versatitle differentiators of the training data, and are typically fed into a non-linear classifier or regressor to produce final predictions. The network's parameters are learned through an iterative training process, where batches of examples are fed into the network and network parameters are adjusted to reduce prediction errors. Iterations of adjustments are made until the network's prediction accuracy on its training data converges or surpasses a threshold. The CNN training process is computationally intentensive, with a requiring training time that grows quickly with the width and depth of layers. The main model we used for classification is a replica of the Google Research entry into the 2014 Imagenet Large-Scale Visual Recognition Challenge (ILSVRC 2014), dubbed GoogLeNet. Imagenet is an image database organized according to the wordnet hierarchy, and the ILSVRC 2014 challenge tested the ability of models to accurately tag photos with up to 5 tags out of a body of 1000. The GoogLeNet architecture won the ILSVRC challenge in 2014, achieving an impressive top-5 error rate of 6.67%. GoogLeNet is an extremely deep network comprised of 24 layers, [how many convolutional layers?], and operates with [how many parameters?]. A full discussion of the qualities and mechanics of convolutional neural networks will come in the theory section.

Accurate Computational Aesthetic Evaluation algorithms would be useful in the fields of generative art and recommendation systems. With the invention of modern computers, the limiting factor on the increasing prevalance and complexity of generative art has become the artist, specifically their expertise in creative decision making and aesthetic evaluation. Despite their enormous computational power, the primary use of modern computers in generative art falls into the category of rendering procedures into their realized form, rather than participating in the search for aesthetic beauty. Advances

in machine learning for pattern recognition have proved useful in many different types of well-defined tasks. In this study we apply modern image classification techniques using convolutional neural networks to the problem of computational aesthetic evaluation, observing whether they are capable of learning the tastes evident in a dataset of randomly generated triangle art, labeled with aesthetic quality ratings.

CHAPTER 2

RELATED WORK

The fields of Computational Aesthetics and Machine Learning are both growing rapidly within the field of computer science. There is a massive of body of literature within both fields, and gaining a firm footing of the basic principles in both of these fields would be an ambitious task. Given the worldwide academic computer science interest, I have learned that with in the natural world of time constraints, a trade-off must always be made between one's awareness of the peripherals of their field and one's productivity in research. I would speculate that remaining an expert in a sufficiently narrow field is easier than becoming an expert in that field, and so it would seem that the standard mode of a scientist would be to be both extremely informed in a small field while still making significant contributions to their field. This observation on the nature of scientific study suggests that as a the popularity of a field of study increases, each researcher will find it advantageous to become more specialized in order to maintain a high level of expertise of their field's immediate surroundings.

The research I did for my thesis was unusual in that I developed an idea for an experiment with Professor Andrews before developing a strong understanding of the field of computational aesthetics at large. For this reason, positioning my work in the field post-facto became was daunting task, because I wanted to simultaneously gain a good high-level understanding of the field and as well as focus down on a specific part. I spent a lot of time filtering through tangentially related papers on Google Scholar before deciding it would be a better use of my time to begin experimenting. To give you a sense of the scale of the search, there are about 17,200 results for the search query "computational aesthetic evaluation using neural networks". After an hour of filtering, I had curated a list of 60 possibly relevant papers of variable length to continue to process. Within my time constraints I was unable to find a study that was deeply related to my

research. This may have been because of the novelty or the naiveté my hypothesis of study, or both. Despite this, I was able to gain a distinct appreciation for the non-trivial degree of time-management and strategy involved with developing expertise in a new field of study.

2.1 Related Work in Computational Aesthetics

The field of computational aesthetics is remarkably young. In doing high-level survey research in this field I felt a distinct tension between the ease in which we are able to program computers to produce impressive generative art and the difficulty we face in offloading more of the creative burden to the computers. The field of computational aesthetics has no shortage of high-level problem formulations, but currently lacks any theoretical advances to justify its ambitions. I was unable to unearth a landmark "exciting" study to form the foundation of this field, that paper has yet to come. Within the field of Computational Aesthetics, the most-cited papers have titles like "Computational aesthetic evaluation: Past and future", "Defining computational aesthetics", and "Experiments in computational aesthetics". These papers are either concerned with developing a modern connectionist model of aesthetics or with conducting aesthetic classification experiments to validate that accordance with art school guidelines really does correlate with subjective aesthetic ratings [9, 11, 14].

The top-down work in computational aesthetics aims to develop an enlightened philosophical framework for understanding the human aesthetic system, and is indispensible for thinking clearly about bottom-up experimental work. Galanter and Hoenig grapple head-on with the field's conceptual entanglement, with an awareness of the "baby steps" that any progress will look like. I suspect, and I think these researchers would agree, that generalizable advances in this field will be deeply related to machine learning and neuroesthetics (the biological study of aesthetics), rather than more practi-

cal areas of aesthetics offering guidelines like the Gestalt laws. Given that philosophers have been devising high-level heuristics for judging aesthetics since antiquity, it seems unlikely that the increased availability of computational power will directly impact our conception of the problem.

For the sake of providing context, I will now briefly summarize the history and foundational ideas of computational aesthetics.

One of the most significant advances in the greater philosophy of aesthetics is owed to Mathematician G.D. Birkoff, who spent a year traveling the world to study beauty across cultures, before proposing a speculative psychological model for quantifying it. In his oft-referenced 1933 paper *Aesthetic Measure*, Birkoff proposes that the perception of beauty is the pleasurable resolution of apparent complexity of a stimulus into a compact mental representation. He codified this relationship into a formula that relates degree of beauty B to the balanced ratio of order O to complexity C.

$$B = \frac{O}{C}$$

The plausibility of Birkoff's measure as a guiding philosophy of aesthetics earned it significant attention in later experimental studies. The recent work of Ragau et. al. and Koshel et. al. draws from information theory to approximate the parameters of Birkoff's formula. They used zeroth-order measures such as Kolmogorov complexity and Shannon entropy [16, 13], to demonstrate encouraging correlations between Birkoff's measure and human-rated beauty.

These results are significant enough to validate the direction Birkoff's core idea — that the human aesthetic response is tuned to respond to stimuli that are sophisticated yet digestible. Novel ideas in the fields of machine learning, neuroesthetics (the study of aesthetics from a biological perspective), and the emerging field of evolutionary philosophy (The top-down study of evolutionary adaptations), may be a way forward. Scientist

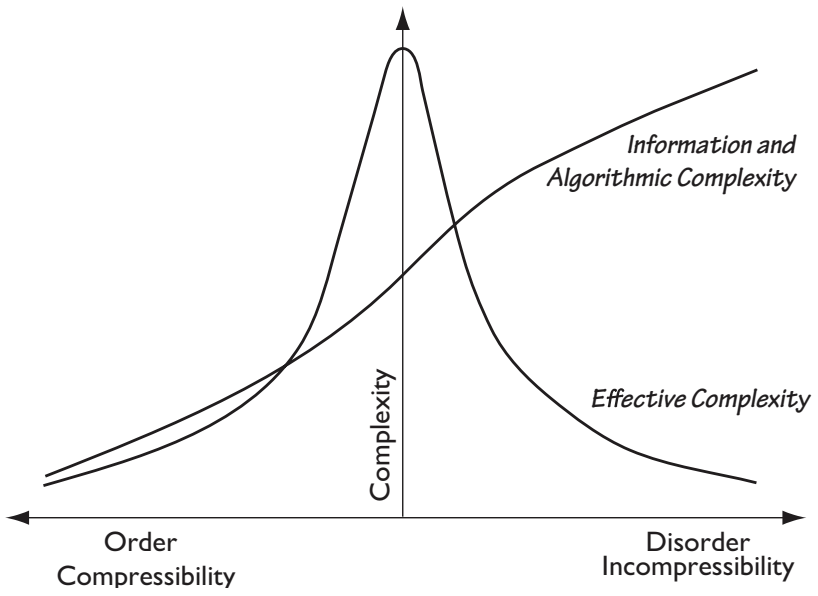

Figure 2.1: Relationship between effective complexity and order

Philip Galanter stands at the intersection of these fields, and subscribes to the view that humans find stimuli with the greatest "effective complexity" to be the most engaging, and thus the most rewarding. Within this framework, the problem of computational aesthetics can be expressed as the study of how humans judge "effective complexity" [8, 6, 5, 7]. Effective complexity is a theoretical measure conceived by Gell-Man et. al. [10] to roughly correspond to the colloquial meaning of the word "complex". Effective complexity measures resistance to entropic forces rather than resistance to signal compression. For example, a symphony by Mozart possesses greater effective complexity than white noise, because it presents a manageable amount of harmonious sophistication for us to process.

In seeking to develop a computational aesthetic evaluator this research is touching onto topics of information complexity and effective complexity, as developed by these researchers. These improving measures will likely align closely with the neurology of human sensory perception. It may already be possible to draw parallels between the principles underlying machine learning systems and Birkoff's measure, since the suc-

cess of these systems already depend on learning of compact, high-level represenations of low-level inputs. This is one of the reasons we became interested in applying the latest advances in image recognition to the field of computational aesthetics. Convolutional neural networks are models inspired by the structure of the human visual cortex, and as such may have a higher chance of making accurate predictions.

Several types of machine learning techniques have been used in computational aesthetic evaluation algorithms with increasing success. Penousal Machado et al [14] implemented an iterative artificial artist system by combining a neural network with an evolutionary programming system. The neural network served as an aesthetic evaluator, and was trained on manually extracted features for every step of the iteration. The genetic programming system would generate populations of artwork, using the aesthetic evaluator as its fitness function. In observing the creative process as alternation between creative and evaluative stages, their system repeatedly produces populations of "drafts" and then evaluates how well they match a constant set of target images. The target images they chose were pieces of artwork. The drafts that most closely matched the target images would then seed the next generation of drafts, and the aesthetic evaluator would be retrained to recognize all of the drafts as non-target images. This feedback loop was designed to ensure that every generation of drafts would be different from the last, and hopefully resulting in greater image complexity. The entire algorithm produces interesting results until the 12th generation, when classification error increases steeply. The system's aesthetic evaluator uses 41 features based metrics like fractal complexity and zipf distribution, each applied over a variety of image channels such as hue or saturation. Although these features provided some predictive value, none were discovered to be orders of magnitude more predictive than the others. Even as more evolved drafts are added to the non-target training set, the classifier consistently trains to more than 99% accuracy. This is to say that when asked to differentiate between images of paintings and

computer generated images of previous generations, the classifier is accurate. The accuracy of the classifier on the next generation of drafts is much less stable. As the system's iterations progress, the neural network's accuracy for classifying the next generation of drafts based on previous generations steadily decreases, and then drops around iteration 12. This experiment demonstrates how neural networks are being applied in novel ways to explore computer capabilities for generating and recognizing aesthetically pleasing art.

Many have also put genetic algorithms to the task of coming up with novel solutions to design problems, and have incorporated aesthetic features into their fitness function by manually extracting heuristics of good design as features for some sort of classifier. These approaches demonstrate acceptable results, and are generally limited by their set of handmade low-level features for prediction.

2.2 Related Work in Image Classification

Machine learning presents some of the most promising opportunities for advances in computational aesthetics. The accuracy of state-of-the-art pattern recognizers have been increasing steadily from year to year since the mid-2000s. Researchers think that these encouraging advances in data science have been the result of a combination of the increasing availability of huge datasets, the decreasing cost of computation, and the influx of new ideas and interest.

Of particular interest to me was a powerful image recognition architecture that Google developed to win the 2014 Imagenet Large Scale Visual Recognition Challenge (ILSVRC) [18]. ILSVRC 2014 challenged participants to correctly classify a set of test images derived from 1000 different Imagenet categories. Some of the categories look extremely similar, and demanding a near human level of recognition ability to differentiate. For example, the Siberian huskey and the Eskimo dog look very similar, and both were in-

cluded as categories for the Imagenet challenge. With an appreciation of the ability of GoogLeNet to recognize patterns, I wondered if this same architecture could demonstrate a similar level of ability at aesthetic evaluation tasks.

CHAPTER 3

THEORY

Theory introduction

3.1 Compuational Aesthetic Evaluation - Convolutional Neural Network

We will use various CNN models that come with Caffe öut of the boxẗo test their efficacy, rather than attempt to customize a model to this task. The first model we will test is called GoogLeNet, which won the ImageNet Large Scale Visual Recognition Challenge in 2014.

CHAPTER 4

IMPLEMENTATION

There were three motivations the informed the design of this experiment.

The first was an interest in whether convolutional neural networks might prove more effective than hand-programmed feature extraction techniques for computational aesthetic evaluation. In my literature review I noticed that a majority of experiments in aesthetic evaluation rely on tens of hand-programmed feature extractors and standard machine learning classifiers to produce modest results. These approaches are based on rigid assumptions on the appearance of aesthetic value. One of these experiment extracted a feature corresponding to the dynamic contrast of colors in the image, and another corresponding to the diversity of colors in a patch of an image. Using a bunch of hand-programmed feature extractors in these experiments may be an efficient way to try to discover an ultra-effective feature extractor, but if this search is not fruitful it may make sense to turn towards more versatile feature extraction techniques. In the task of image recognition, convolutional neural networks are effective for recognizing extremely different patterns; could a CNN do for aesthetic evaluation what it did for image recognition?

The second motivation informing the design of this study was an interest in the separability of subjective aesthetic judgements of arbitrary complexity. I wanted to throw a dataset of unknown difficulty at our model and see what it would give me back. There is much hype surrounding the capabilities of this type of model, some of which I had admittedly bought into. In an effort to challenge my expectations of these models, I asked whether such a model could remain effective on a type dataset it was not designed for.

The third motivation was an interest in the relationship between the accuracy of our classifier for a given image and the type of image provided to the classifier. If we used

a simple type of art for our experiment, it might be easier to identify pattern or types of images that the classifier finds it difficult to classify.

With these factors in mind, I designed the experiment. The body of artwork for rating would be randomly generated triangle art. I would build a training interface for rating these images, where I would assign each presented image a completely subjective binary rating of attractiveness. I would click the green button if an image looked good to me, and the red button if it did not. In order to ensure the internal consistency of my ratings, I would make sure to give each image more than one rating during the training proces. I represented a "thumbs up" rating with a 1, and a "thumbs down" rating with a 0. If an image received conflicting ratings, its recorded label would become 0, to avoid introducing controversial images to our classifier. With a dataset of consistently rated images, we would train GoogLeNet on a big portion of the data, and test GoogLeNet on the rest.

To perform this experiment I would need to assemble a relatively small dataset of images with ratings using a training interface, and I would need an instance of the GoogLeNet classifier.

4.1 Training Interface

Assembling a dataset would be relatively straightforward. Using python, the Flask web framework, sqlite3, and the Pillow graphics package, I built a website, database, and image generation algorithm, respectively.

The website would algorithmically decide for each rating whether to have me rate a new image or re-rate an existing image. It would do so in a random manner, with a bias towards rating existing images with the fewest ratings. This interface was compatible with mobile-browsers, and I added swipe gesture support to allow for me to easily judge images on-the-go.

The algorithm for generating each image worked as follows. First a blank 256x256 canvas would be filled with a random color. Then a random number of triangles (2-6) would be drawn on top of each other. Triangles' corners would be randomly placed, and their fill color would be random. All of the randomness used was uniform, and all of the colors were specified in HSL format, so that the lightness of each randomly selected color could be restricted to an attractive range of values.

Over the course of a few days I made roughly 4000 ratings on 2000 images. In rating triangle art for several hours I was surprised by how quickly I would grow tired of judging the images. In computational aesthetic literature user fatigue is a recognized difficulty with interactive training systems, as it dramatically limits the scale of data collection. In order to reach 4000 high-quality ratings, I found it useful to take breaks every 10 minutes, and try not to adjust my taste in the triangle images. I decided to make 4000 ratings rather unscientifically. At this stage in the experiment I was anxious to begin classifying my data, and unsure of whether a dataset of, say, 10,000 example would be significantly better than a dataset of 4,000 examples. My intuition, based on the published results of companies using large datasets, was that unless I could produce at least an order of magnitude more training examples, my result's accuracy would not vary by much. I labeled 75% of the images with low aesthetic value and the remaining 25% with high aesthetic value.

4.2 Classifier

Preparing an instance of the GoogLeNet classifier would turn out to be less straightforward. I decided to use a popular neural network framework called "Caffe", developed at the Berkeley Language and Vision Center only two years ago. Caffe comes packaged with some example models, and one of those models was conveinently the GoogLeNet model. This meant that once Caffe was intalled I had no problem booting up an exact

replica of very sophisticated model Google built.

Training convolutional neural networks is computationally intensive, and doing so in a reasonable amout of time requires machines multiple times more powerful than my laptop. In order to take advantage of the enticing GPU speedups available with Caffe I ended up renting an AWS g2.2xlarge instance, complete with a graphics card sporting 1,536 CUDA cores. This setup was nice, because after installing all of the necessary software and running my tests I was able to stop and start the instance at-will to reduce billing costs.

With Caffe installed on our instance, I retrofitted the GoogLeNet architecture to produce predictions on 2 labels, instead of 1000. Next, I divided my database of labeled images, placing 75

CHAPTER 5

RESULTS

The classifier converged at a test accuracy of about 75%, which is immediately troubling for its alignment with the dataset's priors. 75% of the images in the test set are "ugly", which means that a fake model could achieve a 75% accuracy by predicting "ugly" every time. The F1 score for the classifier on the test dataset is 0.16, which is very low.

CHAPTER 6

FUTURE WORK

CHAPTER 7

CONCLUSION

Given the difficulty of defining art itself, it is no surprise that the criticism and reception of art itself is as nebulous. Yet despite the unclear underpinnings of aesthetic tastes, it is clear that aesthetic judgements are an integral part of everyday life. Short of machine intelligence, computerized judgements of aesthetics will likely never supercede our own, but in combination with our own, there is the potential for computers to make artists out of us all, or at least reduce the cost of aesthetically pleasing design.

The outcome of this experiment may not be clear and but the field of computational aesthetic evaluation will only gain in relevance as we strive to make our world more beautiful.

Qualitative results here

BIBLIOGRAPHY

[1] G.D. Birkhoff. *Aesthetic Measure*. Harvard University Press, 1933.

[2] Hartmut Bohnacker. *Generative design : visualize, program, and create with processing*. Princeton Architectural Press, New York, 2012.

[3] Roberto Cipolla. *Machine learning for computer vision*. Springer, Berlin New York, 2013.

[4] Gary William Flake. *The Computational Beauty of Nature*. The MIT Press, 1998.

[5] Galanter. Computational aesthetic evaluation: Steps towards machine creativity.

[6] Philip Galanter. What is generative art? complexity theory as a context for art theory. In *In GA2003–6th Generative Art Conference*. Citeseer, 2003.

[7] Philip Galanter. Complexity, neuroaesthetics, and computational aesthetic evaluation. *13th International Conference on Generative Art (GA2010)*, 2010.

[8] Philip Galanter. The problem with evolutionary art is ... 2010.

[9] Philip Galanter. Computational aesthetic evaluation: Past and future. In *Computers and Creativity*, pages 255–293. Springer, 2012.

[10] Murray Gell-Mann and Seth Lloyd. Effective complexity. *Nonextensive entropy*, pages 387–398, 2004.

[11] Florian Hoenig. Defining computational aesthetics. In *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 13–18. Eurographics Association, 2005.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] Misha Kosheleva, Vladik Kreinovich, and Yeung Yam. Towards the use of aesthetics in decision making: Kolmogorov complexity formalizes birkhoff's idea. 1998.

[14] Penousal Machado, Juan Romero, and Bill Manaris. Experiments in computational aesthetics. In *The Art of Artificial Evolution*, pages 381–415. Springer, 2008.

[15] Jon McCormack. *Computers and creativity*. Springer, Berlin New York, 2012.

[16] Jaume Rigau, Miquel Feixas, and Mateu Sbert. Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. In *Computational Aesthetics*, pages 105–112, 2007.

[17] Daniel Shiffman. *The nature of code*. D. Shiffman, United States, 2012.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[19] Hideyuki Takagi. Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation.