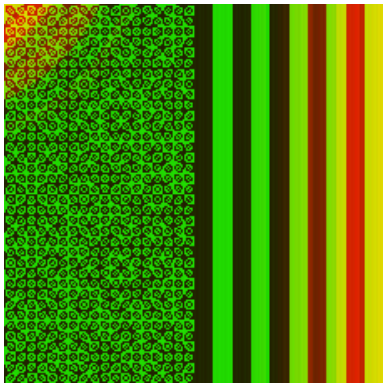

Seeing the beauty in Neural Networks

Hendra Hadhil Choiri (4468457)
David van den Berg (4487575)
Philipp Schwarz (4370058)
Thijs Boumans (4214854)

July 2016

Contents

1 Introduction

One of the major criticisms of artificial neural networks are that their being black boxes since little satisfactory explanation of their behavior and interpretation exists (Benitez et al., 1997). In recent years several tools for analysing, interpretation and visualizing neural networks have been developed some will be adopted in this study. Although, artificial neural networks are only loosely based on how biological neurons work, it is interesting that also individuals often have difficulty to reason why they perceive an image or feature of an image as beautiful or attractive. This paper aims to shed light on what a supervised machine has learned to assess image beauty by trying to find plausible interpretations of the neural network and its multiple layers.

Human assessment of the beauty of images and aesthetics in general is subjective in nature. That is because there is no unanimous measurement for beauty and different people have different perceptions and appreciation for aesthetics. Literature on photography and psychology suggest several properties which are commonly considered to make images more attractive. Nevertheless, the underlying basis of what is visual aesthetics remains too vague to create a general model. In this document we present an experiment that tries to give a broader understanding of what beauty is by examination of neural networks trained for beauty recognition in images. Moreover, the problem we face is that when a human being is confronted with an image, the brain only returns a final aesthetic judgement and not the entire process that gives rise to this judgement; we try to solve this by mimicking the process that gives rise to aesthetic judgement in a neural network, and inspect the network to find out what defines beauty.

This paper discusses two experiments that build upon the following design. A multi-layer convolutional neural network is trained on a big data set of aesthetically rated pictures; networks of this kind have been shown to be capable of aesthetic judgement (Veerina, 2012). To better understand the network and ultimately beauty we extract data not only from the final layer but also from the layer precursing the final layer. The second to last layer delivers 4096 individual neuron output values. The activation values of the second to last layer for a certain picture can be thought of as a vector in a 4096 dimensional space $\mathcal{S}$. We hypothesize that certain regions of this space are related to beauty.

In a first experiment we feed the trained network with a set of 3000 images and we record the activation values of the second to last layer. The activation data is examined using principal component analysis (PCA). Next, we take a closer look at the first few components calculated with PCA which span a subspace of $\mathcal{S}$, images that are located in the direct vicinity of one the Principal component vectors are manually examined. In this way we explore the semantics of the base vectors which correspond to the most decisive regions in $\mathcal{S}$ for beauty judgement. Ideally we find a clear meaning or image feature for the individual components. A

second experiment examines the same first few components generated with PCA, but rather then picking images in the vicinity of the principle components we apply an evolutionary art program to generate images that lie within $\mathcal{S}$ close to a given component.

2 Related Works

Considerable research has already been performed with respect to image beauty assessment from different perspectives. One of the primary objectives has been to develop the ability to reorganize high volume databases by aesthetic criteria.

Bhattacharya et al. (2010) use best practice derived from the field of photography and assess with a supervised learning algorithm preset aesthetic feature, based on the premise that following few simple guidelines enhances the quality of photos significantly. First, the 'Rules of Thirds, that is the normalized Euclidean distance between the center of mass of the foreground to each of four symmetric stress points in the image. Second, optimal visual weight balance, which is the ratio between sky region and ground region. The datasets are then used as training data for support vector regressor. This approach achieves 86% accuracy in predicting the attractiveness of testing images. However, this research is only utilizing geometric composition and for our purpose the approach is not directly transferable since training does not involve a neural network.
Likewise focused on photos, Murray et al. (2012) have developed a database with 250.000 images and corresponding meta-data specially targeted at beauty research. The images carry more meta-data compared to images commonly used in the field. The database is however tailored to beauty as conceived by professional photographers. An image in the AVA-database is ranked on aesthetic quality using a vote system, where votes are given by trained and amateur photographers. Further, the images are given one or more semantic labels and are subdivided on their photographic style. The database has been subjected by the authors to a machine learning algorithm to perform content-based aesthetic categorization. The described database was also utilized by, Lu et al. (2014) who designed a deep convolutional neural network. Their approach demonstrated to be able to outperform other state of the art beauty recognition algorithms.
Veerina (2012) shows that various multi-layer convolutionary neural networks can distinguish whether an image is beautiful after extended supervised learning. Using state of the are networks running in the caffe framework Veerina (2012) reports accuracies ranging from 67 to 77 per cent. It is noted that these networks where designed for categorizing images.
Enquist et al. (1994) have done investigations into the relation of symmetry to beauty using a neural network, the paper suggests a strong correlation and links this correlation to natural evolution. It is hypothesised that animals prefer symmetric objects over asymmetric objects as it eases the process of recognizing them from different angles.
In the context of evolutionary art Li and Hu (2010); Li et al. (2013) have made attempts to learn how human judge aesthetics. Evolutionary art is a branch of generative art and refers to art created with the use of computer algorithm that undergoes a evolutionary process (survival of the fittest over generations with inherited properties from both parents). The paper demonstrates that certain features in color ingredients image complexity and image order underlie human judgment of aesthetic. In Li et al. (2013) the authors compare multi-layer perceptrons, a feedforward artificial neural network model with a C4.5 decision tree classifiers and conclude that their approach is a promising pathway for this problem class.

3 Methods

The basis of the research is connected to the use of the BLVC Refference CaffeNet Krizhevsky et al. (2012), further techniques are principle component analysis and an evolutionary program considered with art generation.

3.1 Dataset

As reference of the beauty images, AVA (Aesthetic Visual Analysis) Murray et al. (2012) dataset is used. The dataset consists of 255.530 images. Each image has 10 rating slots with number of votes given in each

rating. Higher rating means the image is considered as more beautiful. The overall label for an image is extracted by using this rule: if the number of votes for rating 6-10 is higher than votes for rating 1-5, then the image is labeled as beautiful. Otherwise, it is not beautiful.

From this dataset, 30,000 images will be used to train the neural network system and 12,307 images are used as testing dataset. 45% images in the training set are labeled as beautiful.

3.2 Neural network

The reference CaffeNet is a multilayer convolutional neural network. The network was introduced by Krizhevsky et al. (2012) for a categorizing problem, working with a dataset of 1.5 million pictures belonging to 22000 categories. In the task of labeling the images the network performance was state-of-the-art for the given task with an error rate of 17 %, 8 % better than the second best at the time of writing.

For the task of beauty recognition the system has a comparative performance (Veerina, 2012).

Architecture The architecture proposed by Krizhevsky et al. (2012) is an eight layer network consisting of five convolutional layers and three fully connected neuron layers. The first two neuron layers contain 4096 neurons each, the last fully connected layer contains 1000 neurons to get a probability distribution across 1000 labels.

For the task we subject the net we alter the architecture to include only two fully connected layers. Elimination of one layer was done to limit training time, and overcome over-fitting. The final output layer has two instead of 1000 neurons, respectively, the neurons represent beautiful and not-beautiful. Instead of a two neuron system a single neuron with binary output can be used. A two neuron system was chosen since it was proven to work by Veerina (2012).

3.3 Principle Component Analysis

Principle component analysis is applied to activation data from large neuron layers like the second to last layer in our network. These activation data live in a space $\mathcal{S}$, spanned by the activation values of the second to last neuron layer (for a more complete description see section 4). PCA is used for finding vectors for which the variance of the data is maximal, or; finding vectors that span a basis with minimal distance to the data. These linearly independent vectors lie in directions that are most important in discriminating beautiful from non-beautiful images.

PCA has the advantage of reducing the dimensionality while keeping important information. The decision of whether an image has aesthetic value likely comes from a cascade of neurons firing or having high activation values in a certain pattern, these patterns can be caught in a single vector in "neuron space" $\mathcal{S}$.

3.4 Evolutionary algorithm

Evolutionary algorithms are algorithms that attempt to optimize a given value by generating a number of initial values. It then tests these values, find the best once and uses them to generate the next generation.

4 Architecture

The main component of the system is an altered version of the reference CaffeNet convolutional neural net depicted in figure 1. Several python scripts are designed for interaction with the network using the PyCaffe framework. After conducting some experiments, we found that below configuration works good for the training.

- Initial learning rate: 0.01
- Stepsize: 5000
- Gamma (LR discount): 0.5
- Dropout ratio: 0.6

By default, the net takes in RGB square images of 256 pixels in size. Since the network is trained for this image size, all further script feeding images to the net are altered to this size.

The net is connected to a python script that given a picture can return any requested layer output. Our interest lies mainly in the last and second to last layers as these layers should encompass high level information. The size of the last layer output with two neurons is very clear and easy to interpret, the second to last layer with 4096 neurons is much harder to interpret. After training, the 4096 neurons of the second to last layer account for much of the high level decision making. The beauty concepts that are trained into this layer are not evident nor confined to a single neuron. It is useful to interpret the neuron data vectors in a continues 4096 dimensional space $\mathcal{S}$. Using principle component analysis we can find partitions of $\mathcal{S}$ that are most relevant for beauty perception (see section 3.3).

The first few components gathered from PCA span a subspace $\mathcal{C} \in \mathcal{S}$. Since the components are the principle components of $\mathcal{S}$, a beautiful image generally has a high "score" or projection along these axis. Further, these components are linearly independent and correspond to isolated groups of neurons.

We hypothesize that the principle components correspond to specific characteristics that make up the concept of aesthetics. Therefore, we are set out to get a better understanding of the principal components, for this we have developed two experiments; a PCA image analysis experiment and an evolutionary art experiment.

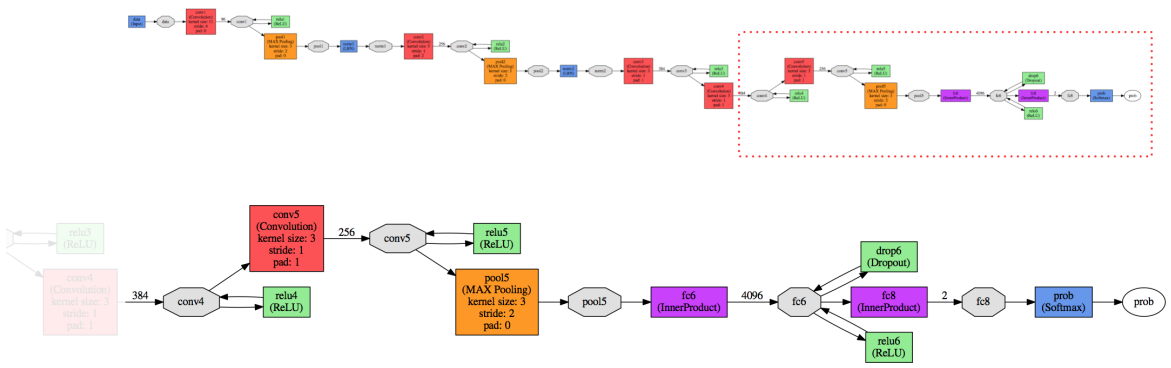

Figure 1: Schematic of the architecture of the used neural network. The fully connected neural layers are expanded. The original reference CaffeNet features an other 4096 neuron fully connected layer "fc7", situated between "fc6" and "fc8".

4.1 PCA image analysis experiment

This experiment is conducted to look for images that are producing similar neurons in the last two layers to one of the principal components (we consider first and second principal components). This is done by projecting the neural output of the training images to the selected principal component and check the correspondence. Images that score high values in the direction of a given principal component are stored and compared.

4.2 Evolutionary art experiment

The principal components are vectors that are hard to interpret for humans. To get a better idea of their meaning we decided to wholly create images with a good fit with the Principal components, complimentary to the images found in the first experiment. Generating an image with high scores is hard to do by hand, therefore, we decided on using an evolutionary algorithm to find images. For this we started with the evolutionarily art system called Jene [1]. This system can evolve images based on the fitness that is calculated by the neural network.

[1] https://github.com/wolfmanstout/jene

Some modifications are done in the Jene. Every time a new generation of images is created, the images are inputted into the neural network to compute the fitness. We increased the population to 25 images per generation so as to offer more variation to the network in each iteration. We decreased the mutation rate so as to better exploit the search space, we suspect it was set to eight in order to deal with the limits of human ratings which are not present in our system. We then made it so that in each generation it selects the top 4 images for reproduction. Lastly, we added a rule that creates five random images each generation to get the system moving in case it gets stuck in a local optimum.

5 Results

5.1 Built network

After training our version of CaffeNet by using specified dataset with 30,000 iterations, we achieved 59.46% accuracy on the test set. This performance is considered sufficient recalling that the original CaffeNet uses 200,000 training dataset and have more layers. As a note, the accuracy was around 55% with 10,000 iteration and around 58% after 20,000 iterations. By increasing dataset size and adding more iterations, the accuracy of our system can be improved. However, due to the constraint of time and computer's capability, we didn't manage to go further.

The main objective in this report is to analyze the neurons in the built network and find the information regarding to the beauty of image. Our current network is used for this analysis.

The convolution filter of the first layer is shown in Figure 2. As the comparison, the visualization from the original CaffeNet is also shown. If these the first 36 of these filters are applied to sample cat image in Figure 3, the result can be seen in Figure 4.

As for the second last layer, the activation values of the neurons can be seen in Figure 5.

5.2 Principle component analysis

A PCA of the second-to-last layer data generated from running 2500 images through the network shows that most variation occurs along three dimensions, as can be seen in the scree plot depicted in figure 6.

What is interesting is that while the original layer of conv1 looks very smooth the same layer in our network is far more noisy. When we also look at figure 4 we see that while the old network appears to focus on edges and corners our network seems to focus on textures instead. Suggesting that perhaps texture matter more then edges for beauty detection.

Figure 6 shows the scree plot of the PCA. A scree plot is a plot that shows how much of the variance of the input data is explained by each of the principal components. In this case we can see that the first 3 principal components explain over 60% of all variance. Suggesting that we should focus on these when trying to explain the meaning of each principal component.

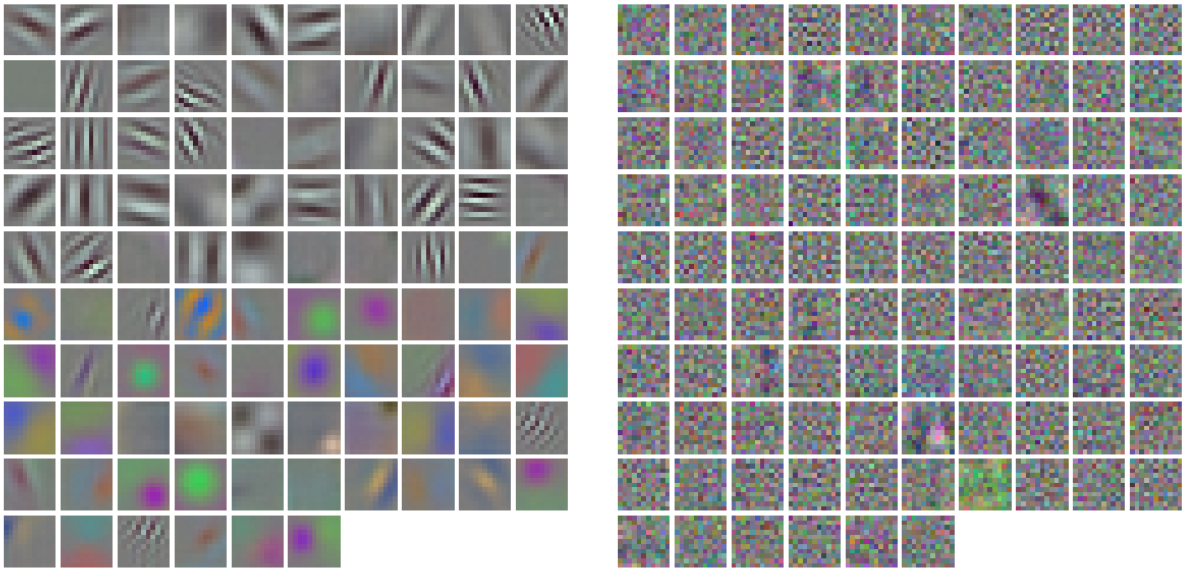

Figure 2: The visualized parameters of conv1 in original CaffeNet network (left) and our network (right).

Figure 3: Cat image as example to test the convolution filter

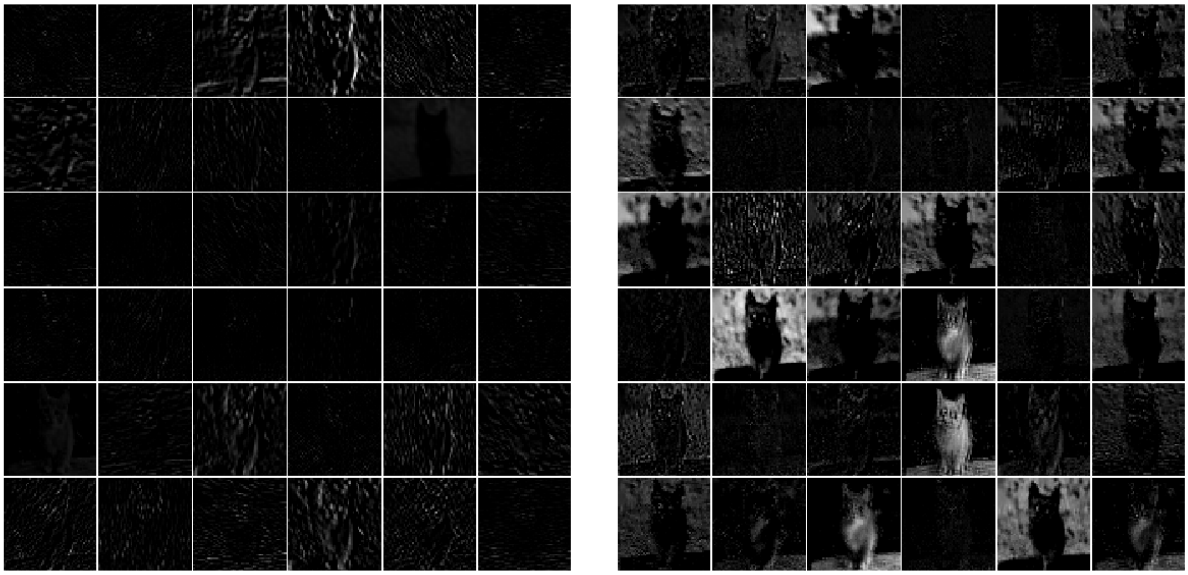

Figure 4: Response of the filters (conv1) from original CaffeNet (left) and our network (right) applied to the sample cat image

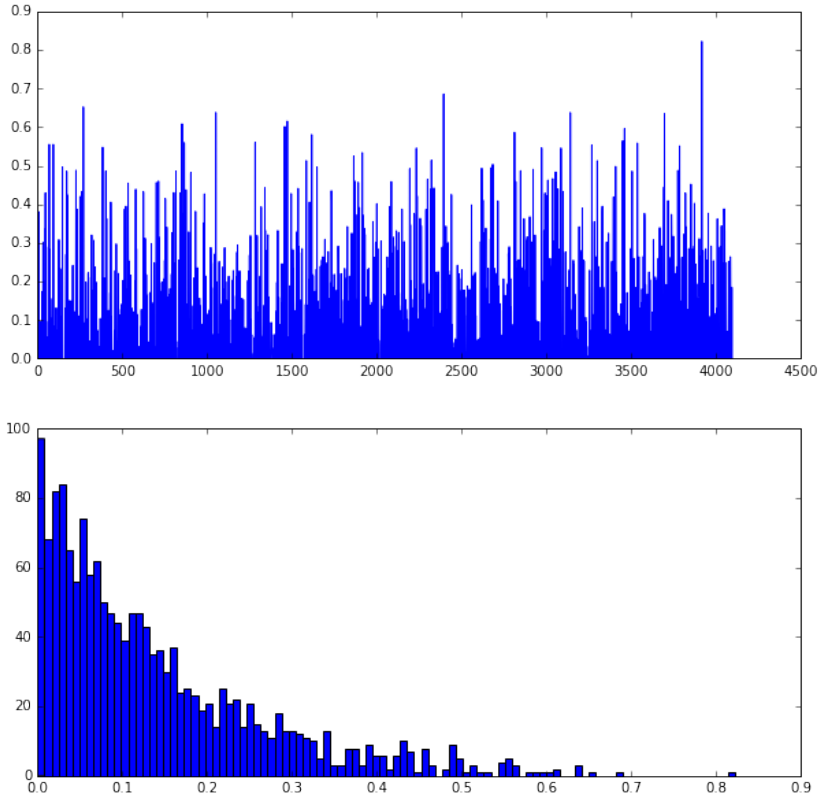

Figure 5: The neuron activations in layer fc6 and the histogram

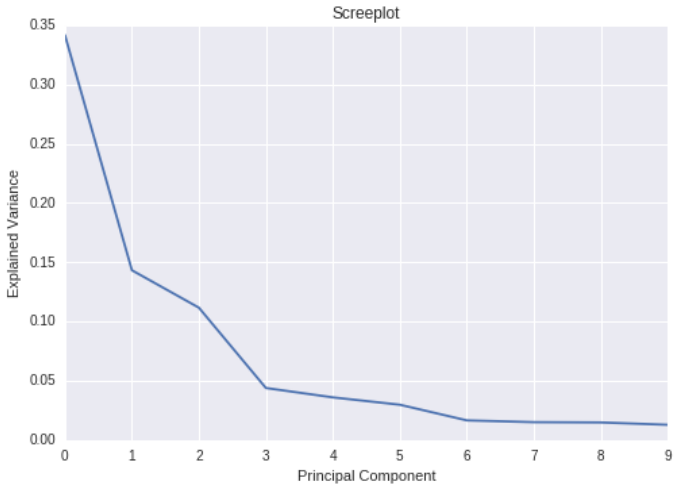

Figure 6: Scree plot of the activation distribution in the second to last layer. The plot provides us that three components within $\mathcal{S}$ are dominant.

5.3 PCA image analysis experiment

When examining the pictures in the dataset that maximise certain principal components as done in figure 7 to 11. We can clearly see a visual similarity within the images in each principal component. Especially the second principal component in figure 8 shows a strong visual cohesion. Another interesting finding we did is all 3 of the maximum principal components in layer 6 resulted in beautiful images showing that clearly the principal components matter in terms of the final result. We also see several images that were in the top 9 for the 3 principal components also show up in the top 9 overall further strengthening this relation.

5.4 Evolutionary art experiment

When looking at the various principal components various things become clear, when looking at the evolutionary results of maximising or minimising the last layer we can see the pictures of what the network thinks is the most beautiful or ugly. If we look at the result of maximising the principal component direction of one of the earlier layers then we see what feature the network pays a lot of attention too.

First experiment is analyzing the second last layer. Figure 12 shows the average and top fitness for each iteration until 100. Although the average fluctuates per iteration, but it has tendency to be increasing and often finds an image with higher fitness.

In figure 14 we can see a number of images that attempt to maximize the first principal component of the image of the last layer. These are the most important features that the network pays attention to on the highest level.

What is interesting is that the top two of these images show a combination of organic shapes and band of discoloration in the middle. This suggest that the network is looking for both organic shapes and objects of interest that exist in the center of mass. While the second two instead show a high degree of symmetry and a focus again a strong focus on the middle of the image. When we compare this with the images in 8 then that suggests that the neural network is looking for certain texture especially in the middle of the image

Figure 15 and 16 illustrate the PCA of the second and third component. The images generated based on maximizing the second component have large bluish areas and diverse other features. The third component produces apparent non-aesthetic pictures with only fuzzy structures and blurry lines.

In figure 17 we can see a number of images generated by our evolutionarily algorithm trying to maximize the principal component of the last layer, generating as beautiful as possible images. The plots of average and top fitness are shown in figure 13 (it seems that after 40^{th} generation, the algorithm couldn't find image with higher fitness anymore).

This is interesting as it shows not only images that are beautiful to people but also images that are not actually that beautiful. This can be useful as labeling these images can create a feedback loop where the

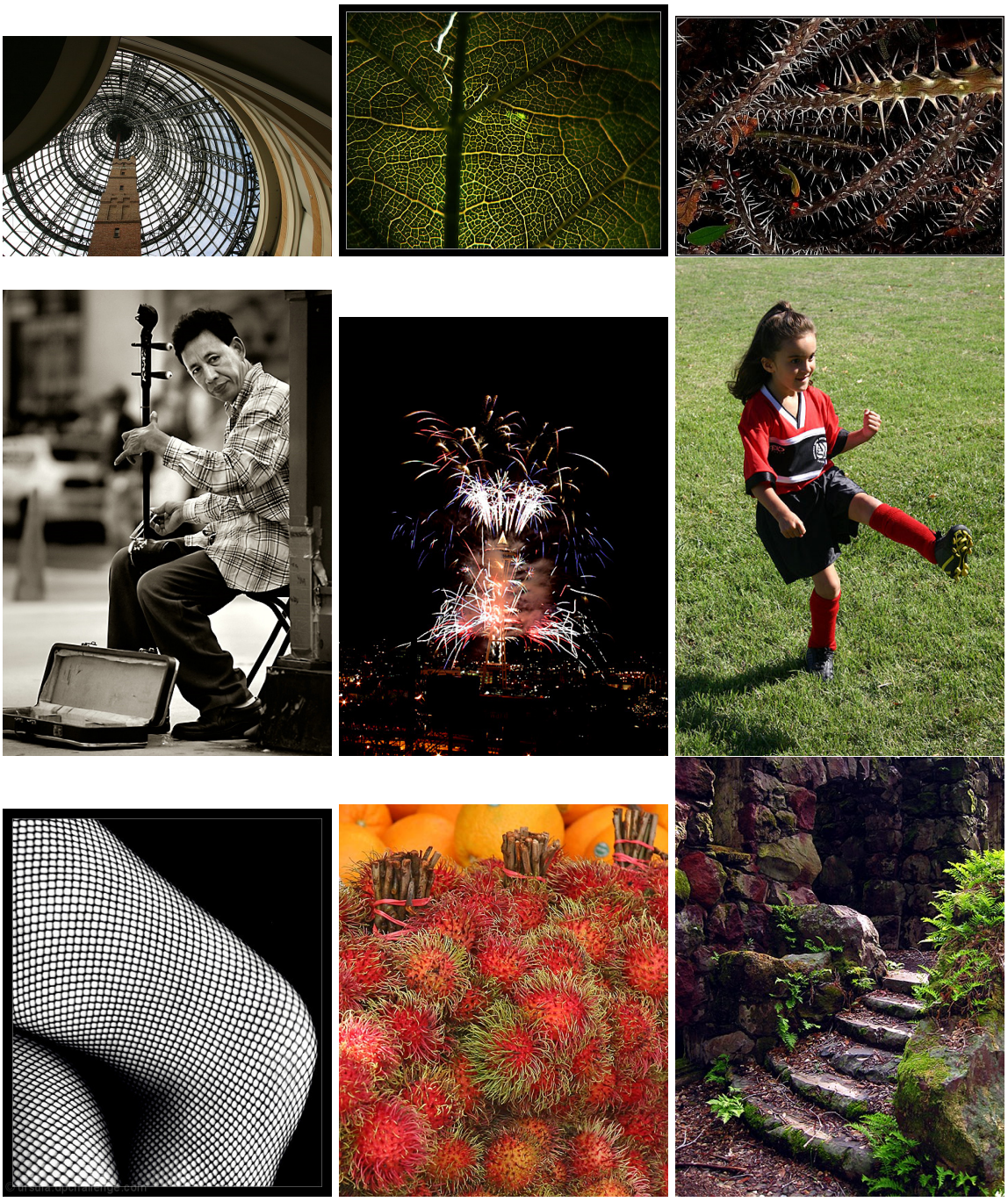

Figure 7: highest PCA score for first principal component of layer 6 amongst all images

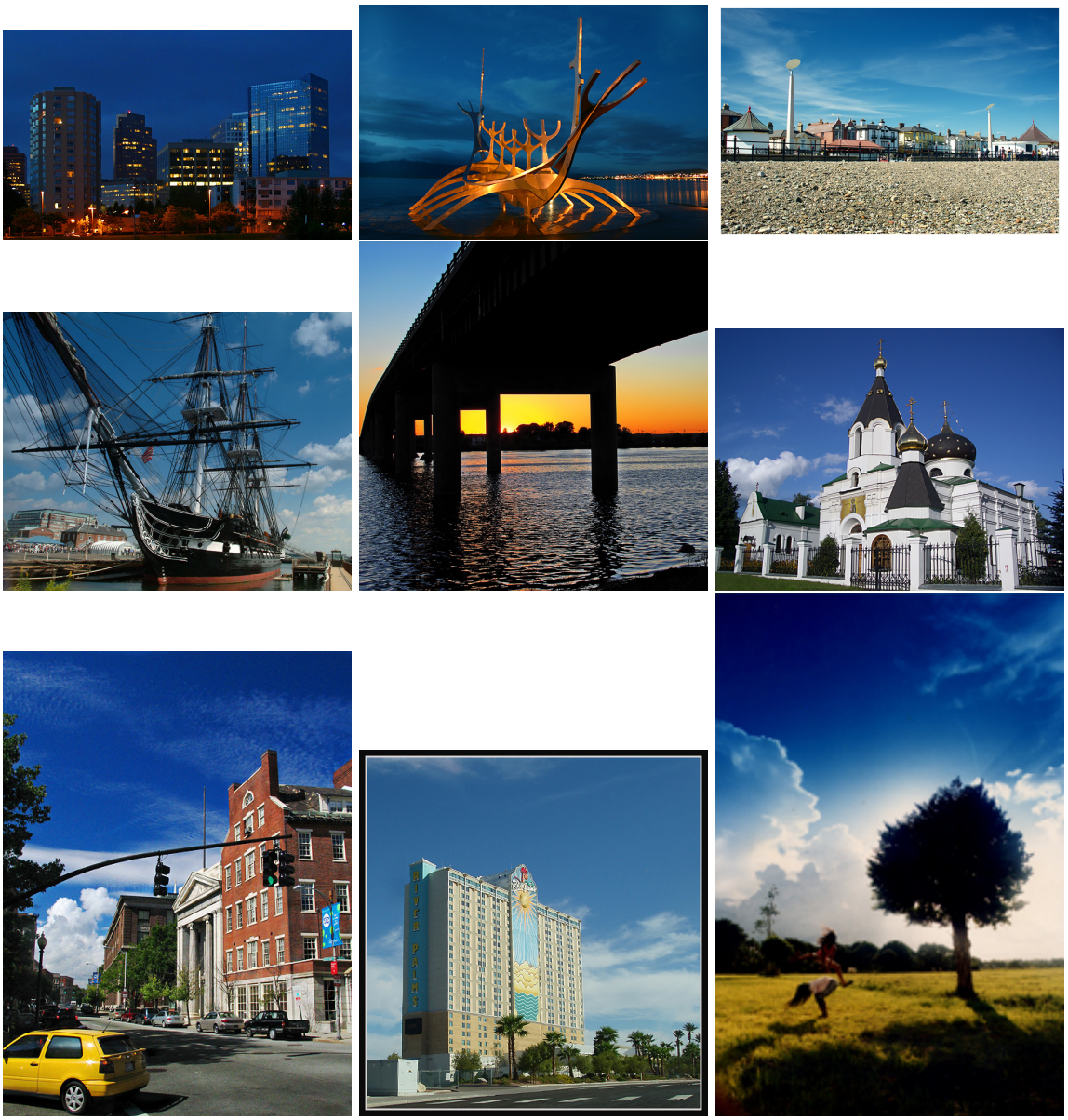

Figure 8: highest PCA score for second principal component of layer 6 amongst all images

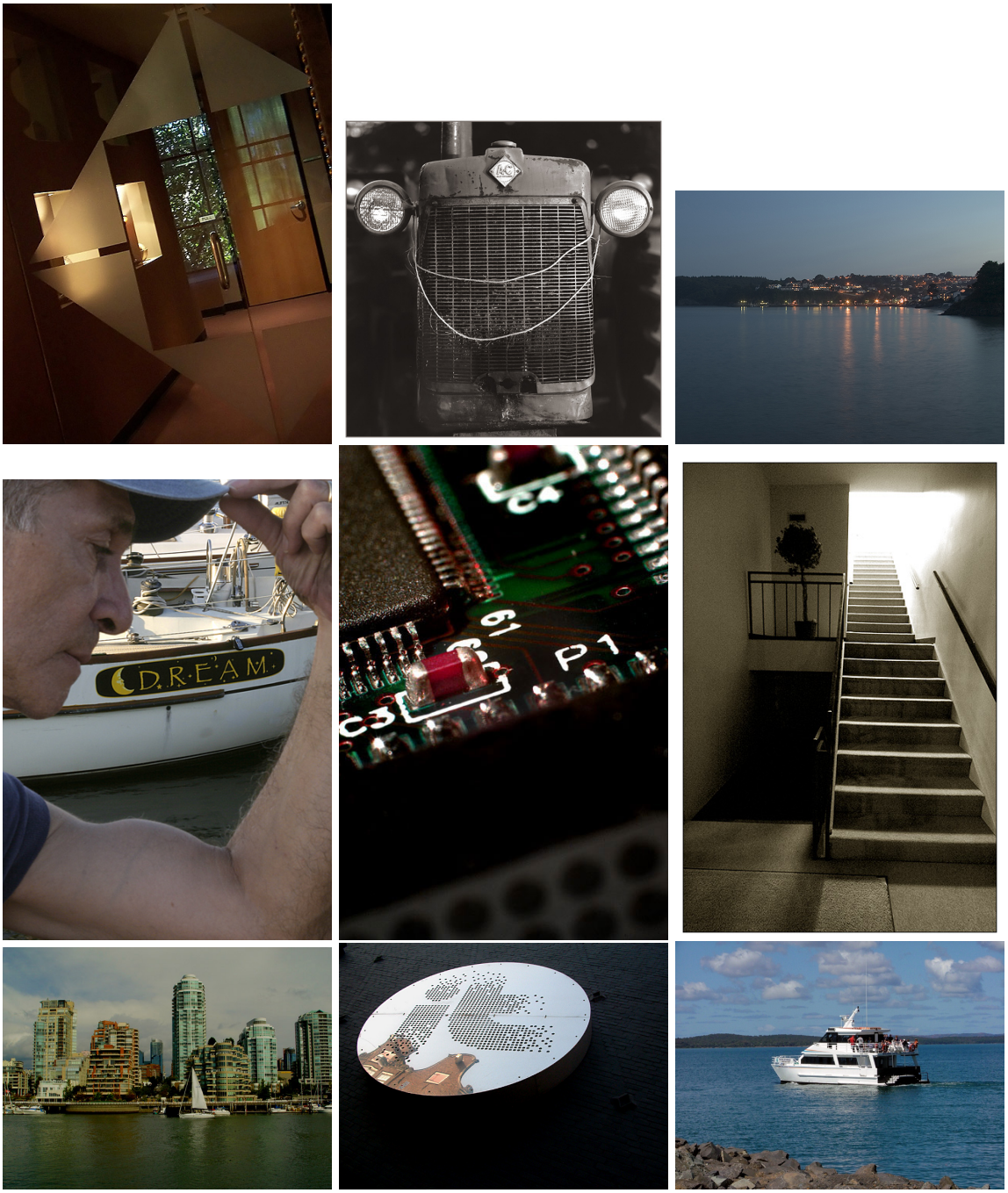

Figure 9: highest PCA score for third principal component of layer 6 amongst all images

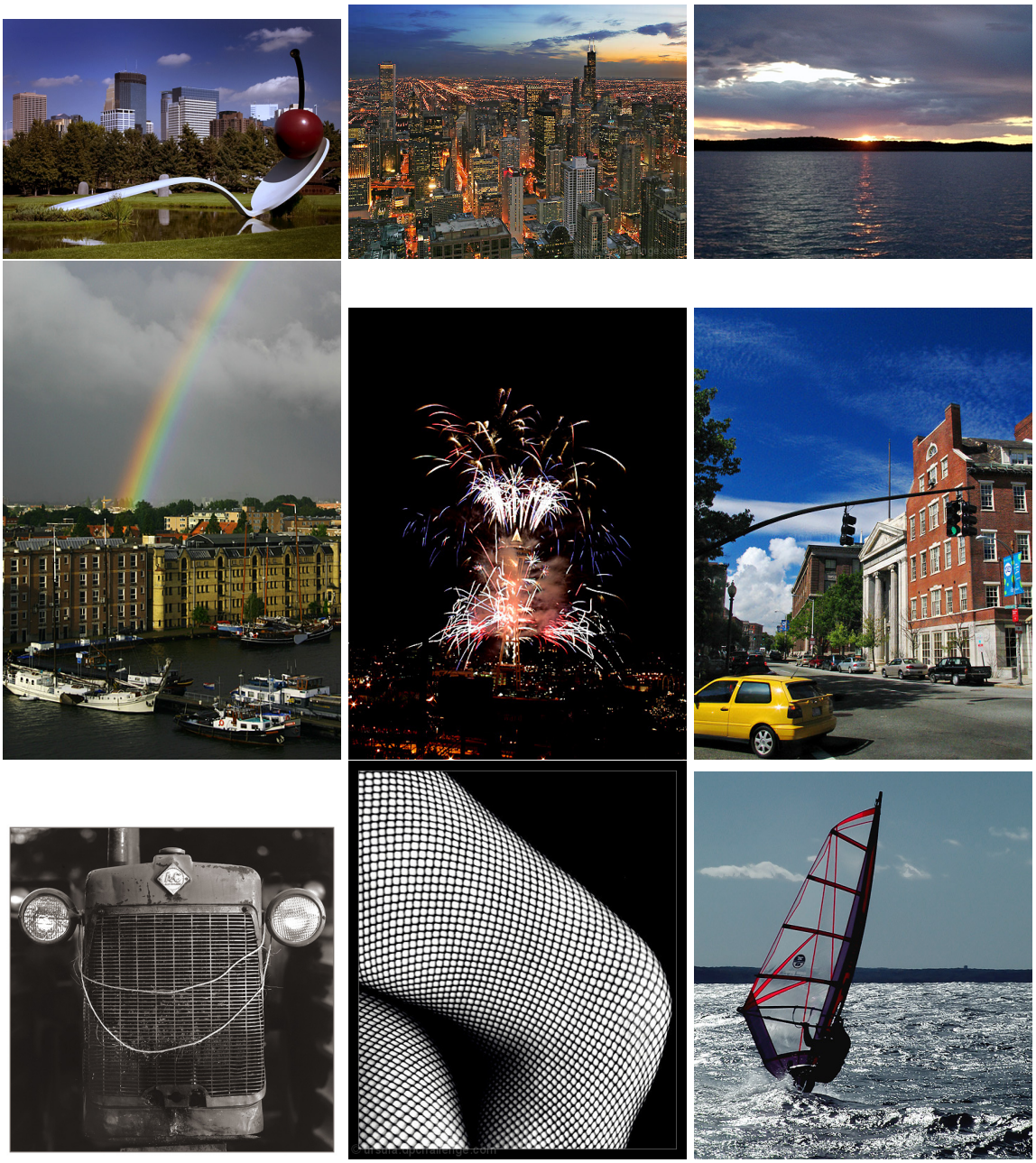

Figure 10: highest PCA score for first principal component of the last layer (2 neurons), most beautiful pictures

Figure 11: Lowest PCA score for first principal component of the last layer (2 neurons), Least beautiful pictures

network can improve itself based on these example images.

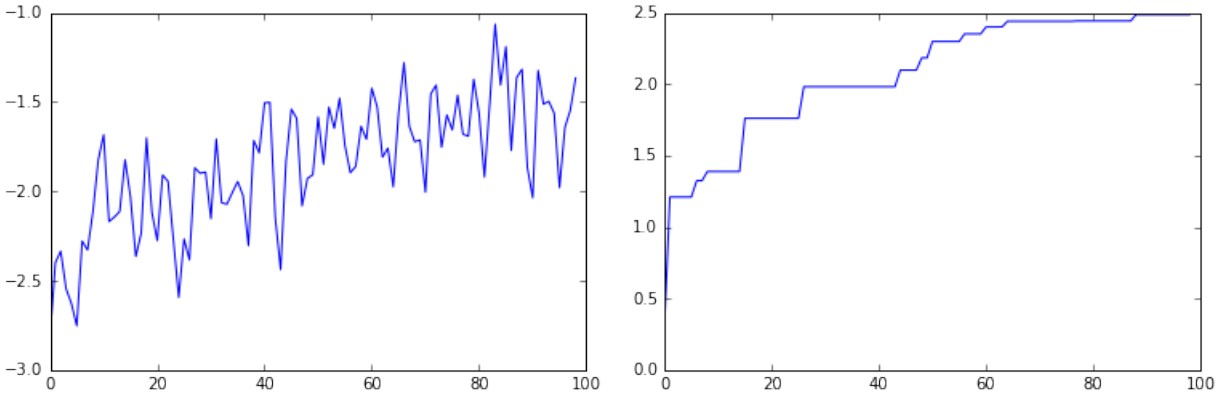

Figure 12: Average and top fitness per iteration of evolutionary algorithm in the second last layer

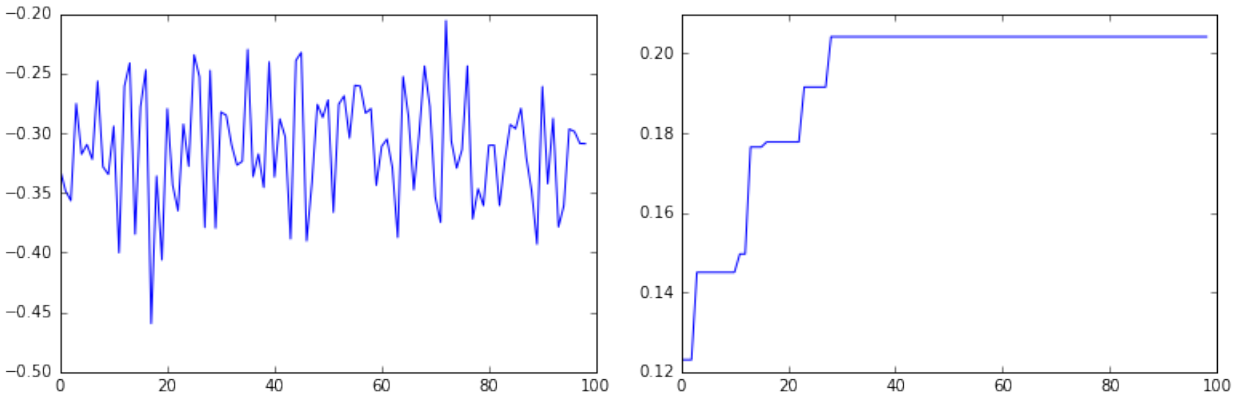

Figure 13: Average and top fitness per iteration of evolutionary algorithm in the last layer

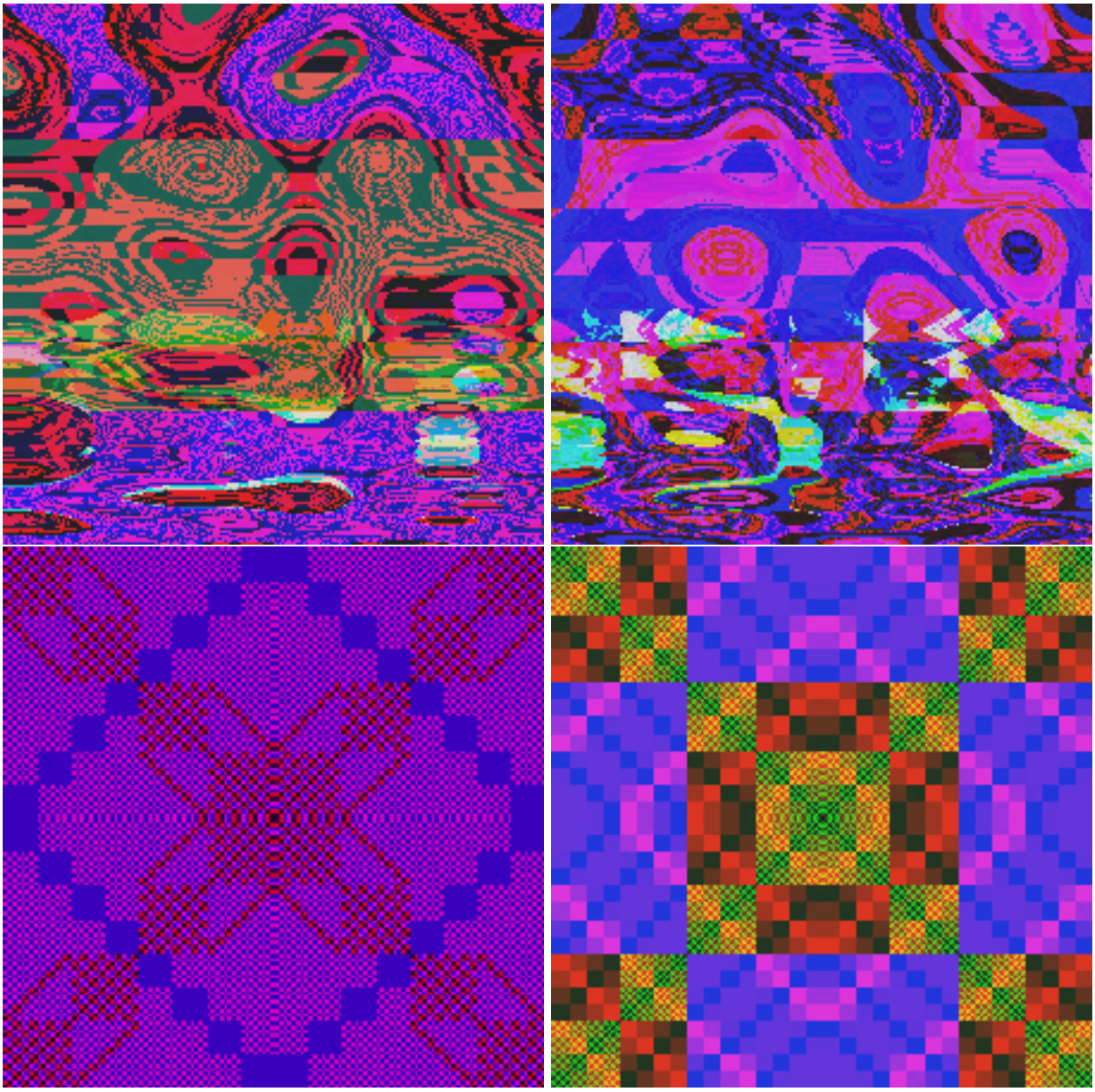

Figure 14: Maximised values for the first principal component of the second last layer

Figure 15: Maximised values for the second principal component of the second last layer

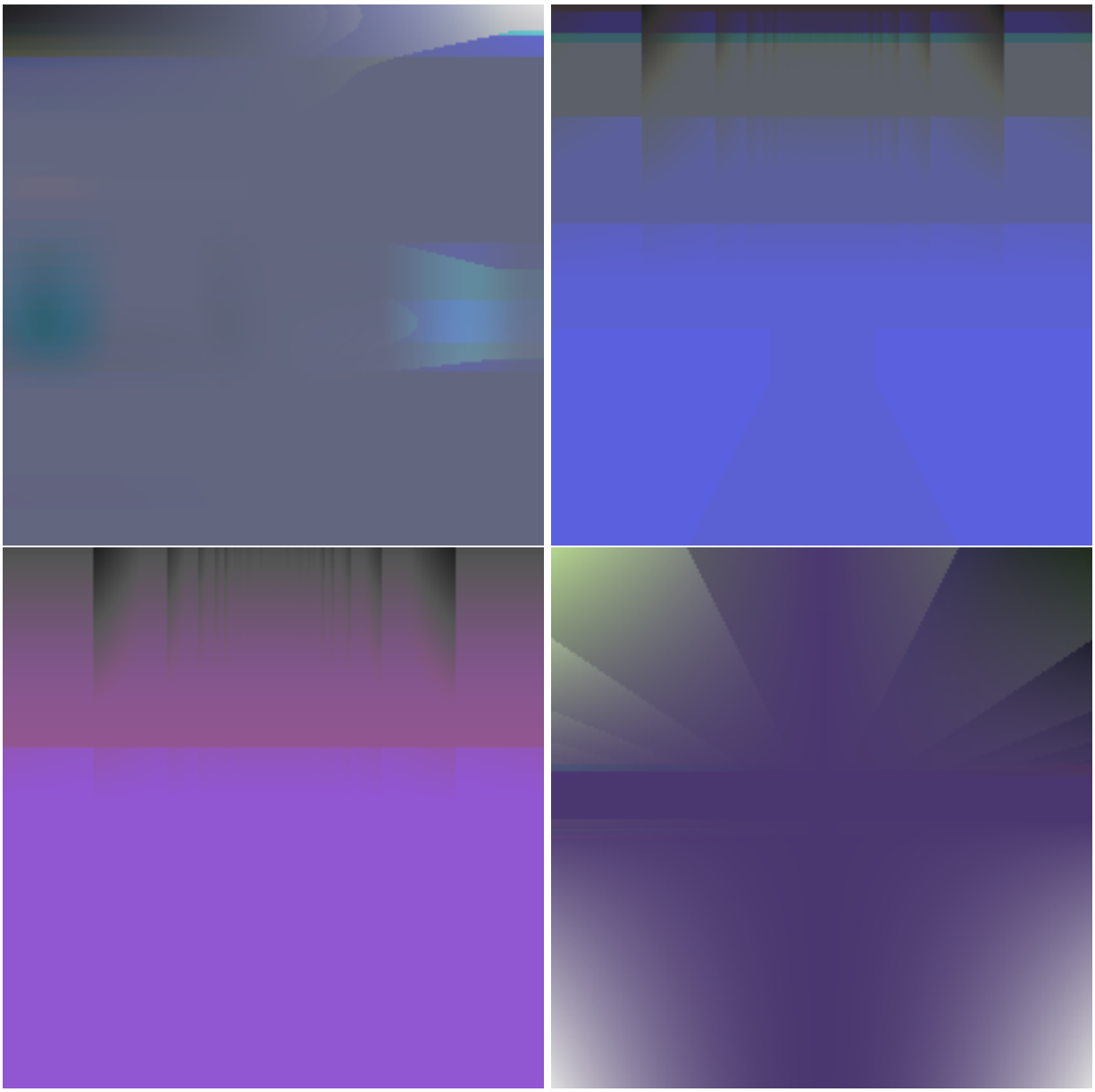

Figure 16: Maximised values for the third principal component of the second last layer

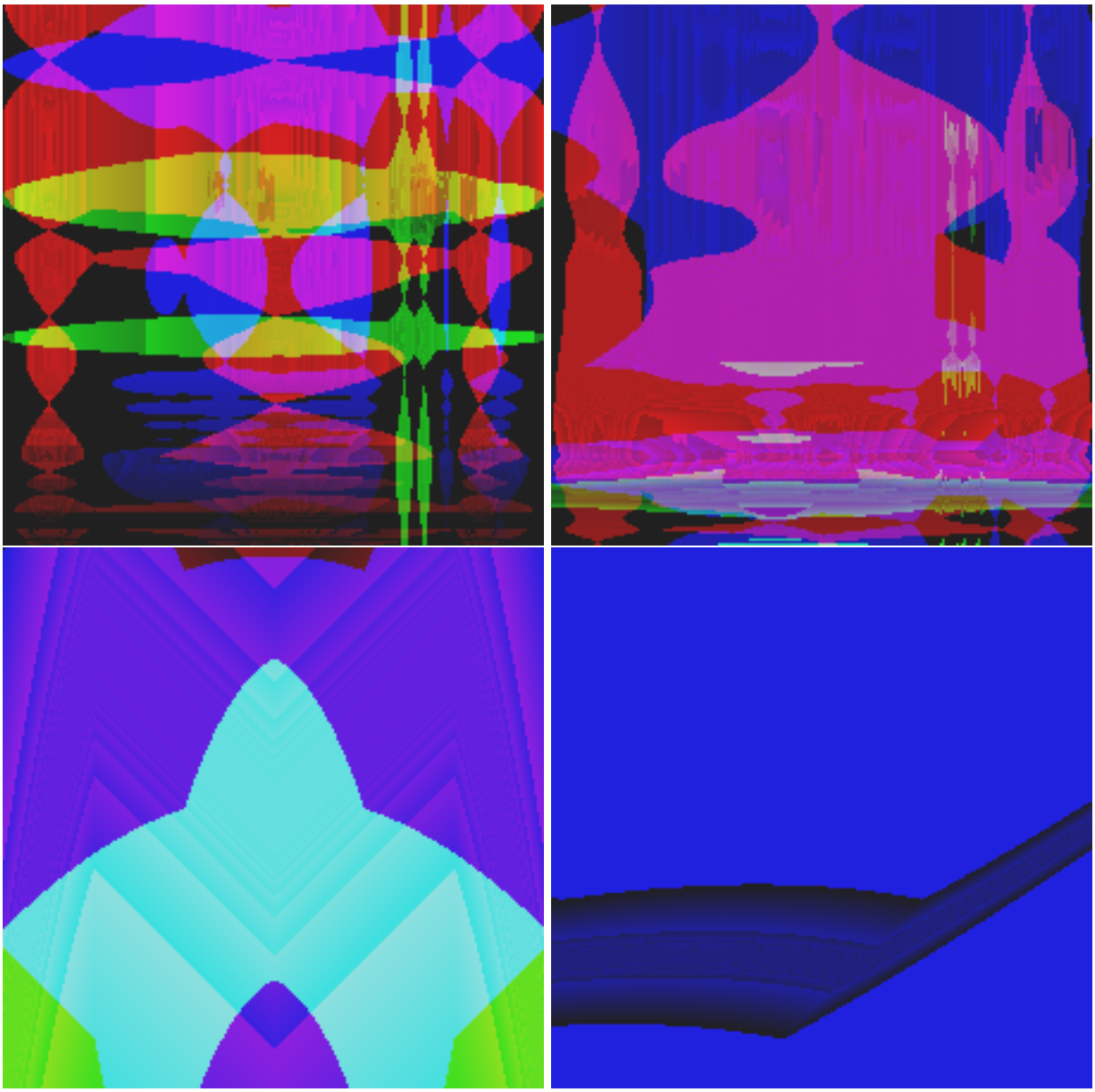

Figure 17: Maximised values for the principal component of last layer

References

Benitez, J. M., Castro, J. L., and Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, 8(5):1156–1164.

Bhattacharya, S., Sukthankar, R., and Shah, M. (2010). A framework for photo-quality assessment and enhancement based on visual aesthetic. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 271–280. ACM.

Enquist, M., Arak, A., et al. (1994). Symmetry, beauty and evolution. *Nature*, 372(6502):169–172.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, Y., Hu, C., Minku, L. L., and Zuo, H. (2013). Learning aesthetic judgements in evolutionary art systems. *Genetic Programming and Evolvable Machines*, 14:315–337.

Li, Y. and Hu, C.-J. (2010). Aesthetic learning in an interactive evolutionary art system. In Chio, C. D., Brabazon, A., Caro, G. A. D., Ebner, M., Farooq, M., Fink, A., Grahl, J., Greenfield, G., Machado, P., O'Neill, M., Tarantino, E., and Urquhart, N., editors, *Applications of Evolutionary Computation*, pages 301–310. Springer Berlin Heidelberg.

Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466. ACM.

Murray, N., Marchesotti, L., and Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE.

Veerina, P. (2012). Learning good taste: Classifying aesthetic images.