GENERATION OF ART WORKS
USING DEEP NEURAL NETWORKS

Armand Vilalta Arias

Directors: René Alquezar Mancho, Enrique Romero Merino
Department of Computer Science

MASTER IN ARTIFICIAL INTELLIGENCE

FACULTAT D’INFORMÀTICA DE BARCELONA (FIB)
FACULTAT DE MATEMÀTIQUES (UB)
ESCOLA TÈCNICA SUPERIOR D’ENGINYERIA (URV)
UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) – BarcelonaTech
UNIVERSITAT DE BARCELONA (UB)
UNIVERSITAT ROVIRA I VIRGILI (URV)

5 February 2016
A la Mare i a la Nerea
pel recolzament per començar,
to MAI friends
for their support to do it.
Abstract

Art creation has always been one of the most popular and difficult challenges of Artificial Intelligence. This work uses state of the art Artificial Intelligence techniques to create artistic images. First the concept of what art is is analysed along its history. It is related to its capacity to show a conceptual and subjective representation of reality that is able to stimulate our minds. The representative power of Deep Neural Networks is used to create, from photographs of reality, artworks.

Along the Art history, from Paleolithic to modern art, Art has been used to communicate to other humans (or gods) ideas in human minds. From ancient times, Art studies the relations that already exist in nature, like the golden ratio, and applies them to its works to provide a pleasant feeling to observers. Art finds the ways to express ideas as they are stored in minds, not like perfect reproductions of reality. These ideas are affected by subjective point of view, preconceptions, emotions or the artist’s whole understanding of the world. The artwork is able to communicate author’s ideas to other humans.

In modern times Art moved away from politics, religion and society and started looking for its own reasons. The concept of aesthetic is born as the philosophy of the art. Kant states that beauty is not a property of an artwork or natural phenomenon, but is instead a consciousness of the pleasure that attends the ‘free play’ of the imagination and the understanding. From different attitudes, the artists expected to interpret the reality creating feelings of vitality, pleasure, imagination and reflection, giving birth to different styles or artistic movements. Some philosophers and artists call nowadays the Death of Art period, since everything can be considered Art.

The process of creating an artwork starts with the perception of reality. In our brain and mind a representation of reality is coded in the form of neural impulses and ideas. These are combined with previous state of mind, ideas and emotions. The imagination employs the same circuits used by perception to create a mental representation of an artwork that produces in the mind the stimulus that the artist wants to communicate. During the creation of the artwork the artist keeps modifying it until it matches the imaginary image, until the artwork produces in the artist the desired effect. The algorithm presented emulates, up to a certain point, this process.

Deep Neural Networks have proved its capacity to learn to build a representation of an image useful for complex tasks like image classification. Their training on real world examples provides them capacity of "understanding" real images somehow similar to ours.

The process of creating an artwork starts with the encoding of the images presented in the form of features that correspond to the activations of the pre-trained Very Deep Neural Network at different layers. A starting image is provided as a basis for the creation of the artwork. It is also fed to the same Neural Network to build its feature representation. These feature representations are used as they are, combined in more elaborate features, or modified. These combinations and modifications are defined in the form of losses on the features. The gradients of the errors are back-propagated through the Neural Network until the image layer. Optimisation is performed on the image to obtain a new image that matches better the desired representation in the activations of the neurons. This process is repeated until the images matches perfectly the desired representation, it can not evolve to a better solution or the user stops it.

The different errors used correspond to different models that pretend to focus on different features of the images and different effects. Here are presented: Content model, represents the same image that is fed and tries to make the generated image to be a reconstruction of
it. Texture model: represents the texture of the image presented; is based on combinations of features that use to be present in same place. Maximum Activation: maximises the sum of the activations of the neurons. Minimum activation: minimises the sum of activations. The last two have a version where the features to maximise or minimise need to be present too in an auxiliary image. All of them can be combined working on the same framework having as input an arbitrary number of images represented at different layers. Relative importance of different models is controlled using sets of weights for each model, each image and each layer.

The number of possibilities that this technique offers is unlimited. Several experiments are done showing the effects of each model and the main parameters of them. Also combinations of models are tested. The results are interpreted from the perspective of what do they represent and their analogy to human processes and the practical use of the algorithm for art creation as well.

Up to this point the algorithm is able to produce artistic works. Nevertheless the images produced usually have some noise or small artifacts. To provide a painting finish they are processed with a special Convolutional Neural Network. This convolutional neural network was originally designed for super-resolution. It is trained using low-res inputs and high-res outputs. For this work I am using a set of weights trained with Japanese anime images. They encode a plain colours effect while keeping sharp edges, quite similar to the style of painting artworks.

This work and some of the artworks produced were presented with Cadi Art Natura and Espai d’Art La Duana to a group of artists. They recognised the artistic quality of the artworks and admitted that never would think that they were made by a computer. They explained interpretations and similarities with other styles. They considered this an interesting tool for their work and suggested uses and possibilities that I did not consider before. There is a world of possibilities to explore in the use of Deep Neural Networks for Art creation.
# Contents

Abstract 6

1 Introduction, Motivation and Objectives 9
  1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
  1.2 Motivation and Objectives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2 Art 11
  2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
  2.2 Prehistoric art . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
  2.3 Classical Antiquity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
  2.4 Middle Age . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
  2.5 Renaissance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.6 Modern Age . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.7 Contemporary art . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
  2.8 Creation process . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
    2.8.1 Creation process in humans . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
    2.8.2 Creation process proposed . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

3 Methods 25
  3.1 Convolutional Neural Network . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
    3.1.1 VGG19 Architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
    3.1.2 Modifications on the network . . . . . . . . . . . . . . . . . . . . . . . . . 25
    3.1.3 VGG19 Trained features . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
    3.1.4 Other Neural Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
  3.2 Image generation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
  3.3 Models and Losses . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
    3.3.1 Content model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
    3.3.2 Texture model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
    3.3.3 Activation maximisation model . . . . . . . . . . . . . . . . . . . . . . . . . 30
    3.3.4 Activation maximisation model on image . . . . . . . . . . . . . . . . . . . 30
    3.3.5 Activation minimisation model . . . . . . . . . . . . . . . . . . . . . . . . . 31
    3.3.6 Activation minimisation model on image . . . . . . . . . . . . . . . . . . . 32
    3.3.7 Other models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
  3.4 Artistic Finishing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

4 Experiments 35
  4.1 Hardware used . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
  4.2 Images used . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
  4.3 Content model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
  4.4 Texture model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
    4.4.1 Weights . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
    4.4.2 Layers . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
  4.5 Texture and content models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
    4.5.1 Weights . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
    4.5.2 Use content image as initial image . . . . . . . . . . . . . . . . . . . . . . . 43

7
Chapter 1

Introduction, Motivation and Objectives

1.1 Introduction

Art creation has always been one of the most popular and difficult challenges of Artificial Intelligence. Creativity has long been regarded as an essential part of intelligence and so it is also a challenge for artificial intelligence. Last year, Prof Riedl proposed a new type of test to see whether artificial intelligence was on a par with that of humans. Rather than requiring a machine to have a human-like conversation, as proposed by the Turing Test, his LoveLace 2.0 test would ask a machine to create a convincing poem, story or painting [28].

The idea that computers cannot create art makes intuitive sense. Art is abstract, expressive, cultural; it depends on mood, nuance and emotion—the very opposite of the rational, logical qualities we ascribe to computers.

In recent years Deep Neural Networks (DNN), specially Deep Convolutional Neural Networks (CNN), have been used successfully for difficult perceptual tasks like object recognition. They are the first artificial systems that rival biology. At the same time their hierarchical architecture and computational properties have a fundamental similarity to real neural systems [10]. There are evidences of the similarity between DCNN representations and those in the ventral visual pathway [16]. These properties make them a good candidate for studying visual information processing in the brain. In fact, it was recently suggested that textures generated from the representations of performance-optimised convolutional networks “may therefore prove useful as stimuli in perceptual or physiological investigations” [22].

Several approaches have been used to obtain representations of the information encoded in the features [29], [26] [21], [7]. These techniques developed provide the possibility of generating images from the representation of reality encoded in the features of a DNN. Obtaining these images can be an interesting tool to understand the features and how the information is represented in the neural network.

Recent works have showed the possibility of creating artistic images using the representational power of deep neural networks."A Neural Algorithm of Artistic Style"[9] introduces an artificial system based on a Deep Neural Network that creates artistic images of high perceptual quality. The system uses neural representations to separate and recombine content and style of arbitrary images, providing a neural algorithm for the creation of artistic images. In "Inceptionism: Going Deeper into Neural Networks" [27], instead of exactly prescribing which feature they want the network to amplify, they let the network make that decision. They simply feed the network an arbitrary image or photo and let the network analyse the picture. They then pick a layer and ask the network to enhance whatever it detected. Each layer of the network deals with features at a different level of abstraction, so the complexity of features we generate depends on which layer they choose to enhance. If they choose higher-level layers, which identify more sophisticated features in images, complex features or even whole objects tend to emerge.

In this work I will explore the possibilities that the Deep Neural Networks open for the
creation of artwork. First the concept of what art is is analysed along its history. It is related to its capacity to show a conceptual and subjective representation of reality that is able to stimulate our minds. Then, representative power of Deep Neural Networks is used to create, from photographs of reality, artworks.

1.2 Motivation and Objectives

From my personal point of view the best of art is that it can get rid of anything it has too. There are no prior constraints in the art creation. The creativity can grow boundless. This makes possible to art to explore unknown frontiers of human knowledge. In this thesis I adopted this point of view when exploring the landscape and afterwards reasoning processes tried to explain some of the things found.

Neural networks used are trained for object recognition. The final output is a label assigning the processed image to one of the possible categories in which the database is divided. With this output precise performance parameters are computed and the model is evaluated. First thing art gives to this work is the possibility to get rid of scientific performance measures. It is not the aim of this work to get a 1% improvement on a contest, but to understand how things work.

It is not focused on the exploitation of the actual knowledge on neural networks but in the exploration, getting out of regular works’ piece of space.

Since this work is concerned about the image representation inside the convolution features learnt and not the final classification results, the final fully-connected layers are not required. Actually they are substituted by a better interpreter of the solution obtained, our mind, whose sensible and understanding capabilities are still beyond best computer performance, not to mention that is not constrained to a bounded, small set of possible outputs.

The first objective of this work is to learn. To learn about the state of the art of deep neural networks, one of the today’s most powerful tools used in artificial intelligence. It involves the use of high performance algorithms like Caffe [15] and the deepest and best performing architectures for image recognition like VGG19 [30]. It also uses state of the art methods for image reconstruction from the neural networks features [29], [26] [21], [7].

The second objective is to have and idea of what makes a work Art. The art can be a tool to understand better the working processes in our mind and its correlation with the process in neural networks. This will be achieved via experimentation and the results will be, like it uses to happen with art, subjective.

And the final goal is, of course, to generate images that can be considered Artworks.
Chapter 2

Art

2.1 Introduction

In this section I am going to do an overview on some aspects of the art over its history that I consider important for the further work exposed in this thesis. This does not pretend to be an exhaustive study on art, including all its styles, periods or works nor influences, uses or goals; obviously there is no space for that and it’s neither the aim of this work. I just want to remark certain aspects of it that are present in very different times and forms along its history that are a common and basic ground of art as I understand it.

2.2 Prehistoric art

In the history of art, prehistoric art is all art produced in preliterate, prehistorical cultures beginning somewhere in very late geological history, and generally continuing until that culture either develops writing or other methods of record-keeping, or makes significant contact with another culture that has, and that makes some record of major historical events.

One of the best examples of the prehistoric paintings are the Altamira caves, near Santander in the north of Spain. The images of animals spread over the ceiling with bright colours. One interesting feature of this painting is that they adapt to the shape of the ceiling, using the natural irregularities to give the painting a sense of volume. In figure 2.1 we can see that the paintings are adapted to the surface. This fact let us know that the painters were able to see those volumes and adapt the paints to them, that is imagine the final figures on the surface before starting the work and adapt the idea in their imagination to fit the volumes on the caves. This shows a mental process where it is possible to have a mind representation of the idea to paint interacting with the surface, being modified to adapt to it in the process of creation, and maybe even the image to be painted is formed in mind after the volumes are observed in a similar way we do when we find figures in the clouds.

Apart from this adaptation to the paintings support there is another constraint in the paintings realisation: the tools available. The tools used to paint and the colours available are restricted and they constrain the possible representations that can be done. Again the mind capabilities can overcome this difficulty keeping the essence of the reality represented, using the available tools, so the image painted does not look realistic but keeps the essence of the ideas that have to be expressed.

The images seem sometimes a little deformed, as it is usual in this new-born art, but the deformations tend to make the animals look more massive, more strong than they are in reality. Somehow the preconceptions on elaborated ideas, like strength or power of the animal are modifying the idea of the animal that one may obtain only with the observation of the animals. We will see that this influence of the high level ideas on the representation of the subject of the artwork, making it less realistic, will be a common fact in the further art.

Here are already present the main mental abilities that make possible for humans to make a representation of the reality. This skill is the main proof of the ability of humans to elaborate
mind representations of reality, modify them in mind according to higher concepts and adapt them to the reality itself when creating a representation.

The exact purpose of the Paleolithic cave paintings is not known. Evidence suggests that they were not merely decorations of living areas since the caves in which they have been found do not have signs of ongoing habitation. They are also often located in areas of caves that are not easily accessible. Some theories hold that cave paintings may have been a way of communicating with others, while other theories ascribe a religious or ceremonial purpose to them. The paintings are remarkably similar around the world, with animals being common subjects that give the most impressive images. Humans mainly appear as images of hands, mostly hand stencils made by blowing pigment on a hand held to the wall[34].

My theory about their purpose is: they did it because they could. They could because they had the ability to have mind representations of the reality that they could use to create, using
the tools available, a representation of the reality in their mind and, they were able to identify in that produced image the representation of the reality that they had in their mind. The main reason to express ideas is the fact of having ideas and that others can understand the ideas from their expression. This can be done by several methods, paintings, music, language, dance, etc. All of them have in common that they can change the state of mind of the receiver, or create in the mind new concepts, and that they are done with this purpose. So, at the end it is always a purpose of communication, being the receiver the other members of the tribe or figures of Gods that are assumed to have human like communication capabilities.

I find also interesting the particular purposes, i.e. why paint a bull instead of a lizard. The point is not what is the purpose, the point is that it usually exists. Does not matter if it is to persuade Gods to favour hunting, show off the greatness of the tribe or just paint your hand to show the world that you exist, it is always there as uses to happen with any of our actions. We can plan the effects of our actions and we have a will. We find some incidents to be more desirable than others because we have desires and we plan our actions so the expected consequence of them is according to our wishes. This is another key factor in any action, communication or Art manifestation, even when we are not aware of it because it is hidden in our unconscious mind.

2.3 Classical Antiquity

The concept of aesthetically appealing has been evolving in the western culture from ancient Greece to nowadays. The Greek thinking understood the beauty as a concept objective, logical and rational, related to the laws of the nature. For a man work to be considered beauty should match the laws of the nature. This canon expressed the essential, the proportionate harmonic and aspire to perfection, a perfection based on a universal unity. The man constructions were guided by a mathematical geometrical point of view [3].

The mathematical relations pursued the target of mimicking the nature from an unbiased point of view. The art was to extract the structure present in the nature, make it explicit in mathematical and geometrical language and apply it to design human constructions. The relations expressed mathematically were not from a simple model, that is they were not the reproduction of a particular instance of a nature element but a general common relation, a middle point over the particular relations of observed instances. This is a simplification of the reality that keeps the common average features over instances of same kind.

But they took it further and tried to find the common relations in different models, for instance, the relations between parts of human body were also valid for relations between parts of other animals or plants. This way they achieved a mathematical representation of general relations in nature. The best known of these relations is the golden ratio. First written definition of it is found in Euclid’s Elements but it is usually represented with a Greek letter phi (φ) in honour to Phidias, the sculptor of Athena Parthenos in the Parthenon (fig. 2.4), who already used it in 447 BC [13]. It has been observed that the Egyptian Pyramids proportions are built according to the Golden ratio and is thought it had a sacred meaning for them as well as an aesthetic value. From golden ration it is constructed the logarithmic golden spiral that is found in many diverse places in nature as is shown in figure 2.3.

Figure 2.3: Different correlations with golden ratio in nature.
The artists used these mathematical relations then in their works, but not to reproduce the same models where they were found, they used them in very different works in many different ways, where usually an evident similarity with the model where the relation is originally present can not be found. Even in those cases the use of the mathematical relations found in nature helps to create appealing artworks.

![Parthenon and golden spiral, Grece.](image)

The eagerness to get closer to reality produced in the latest Hellenistic sculpture exaggeration of the shapes. Here the equilibrium and harmony in seek of beauty were lost, leading to another concept of beauty conceived as imitation: mimesis of the reality [23]. In the sculptures of this period we see that the shapes were somehow deformed, being even impossible in reality, to empathise the ideas that were transmitting like strength, violence, drama. Here we see that the rules of the nature had been slightly modified to be more appealing to an impression, that is a subjective point of view, that makes a stronger impact on the observer.

The Romans were not so prone to the elaboration of reflexions over arts. They were more focused on productivity so the functionality was imposed over the beauty. Even though, they keep the Greek aesthetic concepts but the function gained importance over the harmony. It is interesting the work of Vitruvius, a Roman author, architect, civil engineer and military engineer during the 1st century BC, known for his multi-volume work entitled De architectura. This work is the only surviving major book on architecture from classical antiquity. According to Vitruvius, architecture is an imitation of nature.[33] In the first book, out of 10 that compose De architectura, Vitruvius defines that the architecture is composed of:

- **Order**: Proportion of the elements of a work, isolated and as a group.
- **Arrangement**: The correct layout of the elements.
- **Eurythmic**: Arrangements of the elements to give perception of beauty in the observer. Notice that the proportion is not objective but subjective to the observer. Sometimes, to achieve it, some optical corrections are required to make them not beauty or harmonious themselves but from the point of view of the observer.
- **Symmetry**: harmony of the parts for aesthetic purposes, not functional.
- **Ornament**: correct according to the work.
- **Distribution**: Economy, design.

The third book explains the correct arrangement of temples for Gods, that will be based on symmetry and this on proportion, like the parts of a well formed human body [33]. This sense of proportion will be the basis for the Renaissance’s man of Vitruvius. Again we see in these works the interest for the mathematical relations and proportions observed in nature as a basis for the art creation, but also the importance of the subjective observer that can introduce modifications on them.
2.4 Middle Age

The Christianity during the Middle Age used the art as a pedagogical tool to support the power of religion on feudal society [3]. During this time, even the maps, works whose goal is to be truthfully representations of reality, were deformed to match the ideas of the people supporting religion.

(a) Beato de Gerona. Year 950 aprox.  
(b) Beato del Burgo de Osma. Year 1078 aprox.

Figure 2.6: Mapamundis from Middle Age. [36]

In the examples of mapamundis from Middle Age in figure 2.6 we can see how the world is represented as three parts of earth surrounded by water that correspond to the 3 sons of Noe. The orientation is always the same, Jerusalem is in the centre of the world and heaven is on top (next to India). The geographical locations are mixed with saints and other religious figures
The artwork mixed a representation of the physical world with the religious understanding of it, the structure that in that moment the world had in the minds of the people, resulting in a deformation of the reality to match the preconception of it. This representation was used to transmit and spread the religious conception of the reality.

2.5 Renaissance

With the irruption of bourgeoisie at the end of Middle Age the liberal arts came back from their stagnation. From the philosophical currents of Renaissance and the confirmation of the rational method afterwards, the arts split from science. Mathematical concepts become important inside any aesthetic work and canons based on equilibrium, measure, sobriety, serenity and harmony between the parts and the whole are established [3]. Again mathematical relations and reality knowledge structures become a scaffolding for artworks. It is world famous the man of Vitruvius from Leonardo Da Vinci (fig.2.7) as a symbol of proportion based on man shape.

![Figure 2.7: Leonardo Da Vinci’s Man of Vitruvius. Galery of Academy of Venice [32].](image)

2.6 Modern Age

At the end of the XVIII century the establishment of the aesthetic or the philosophy of the art as a new discipline gave philosophical backing to the autonomy of the artistic fact. This was the starting point of what we call modern art. The art moved away from politics, religion and society and started looking for its own reasons [3].

The German idealism, in opposition to the theoretical realism, moved the way of thinking to a knowledge complemented with the metaphysical theory. The beauty unchained from the objective point of view and gets subjective, understanding the subjectivity as something real and not only abstract [12].

The aesthetic is informally a synonym of beauty but its origin is philosophical. It was born as a concept in Ancient Greece but it was not until XVIII century that it became a philosophical science. Its creator was the German philosopher Alexander Gottlieb Baumgarten. The root of
the word is the Greek adjective aisthetike that means sensation, perception. In 1750 Aesthetica was the title of his unfinished work, where he institutionalised the word as the knowledge of the sensible, the sentiments and the perception of beauty [12].

The systematisation that Immanuel Kant did on Aesthetics made it a philosophical discipline. In his main work, the Critique of Pure Reason, Kant argued for a Transcendental Aesthetic, an approach to the problems of perception in which space and time are argued not to be objects but ways in which the observing subject's mind organises and structures the sensory world. This was further developed in the third Critique, Critique of Judgement, whose first part Critique of Aesthetic Judgement, discusses the four possible "reflective judgements": the agreeable, the beautiful, the sublime, and the good. In reflective judgement we seek to find unknown universals for given particulars; whereas in determinative judgement, we just subsume given particulars under universals that are already known [11]. It is easy to observe the perfect matching of the Kant's judgements' principles and the basis of machine learning and neural networks algorithms used today.

The agreeable is a purely sensory judgement, based on inclination alone, e.g. "The food at FIB bar is good".

The good is essentially a judgement that something is ethical. That conforms with moral law, a coherence with a fixed and absolute notion of reason. It is in many ways the absolute opposite of the agreeable, in that it is a purely objective judgement: things are either moral or they are not, according to Kant.

The remaining two judgements, the beautiful and the sublime, differ from both the agreeable and the good. They are what Kant refers to as "subjective universal" judgements. This means that, in practice, the judgements are subjective, and are not tied to any absolute and determinate concept. However, the judgement that something is beautiful or sublime is made with the belief that other people ought to agree with this judgement, even though it is known that many will not.

Kant states that beauty is not a property of an artwork or natural phenomenon, but is instead a consciousness of the pleasure that attends the 'free play' of the imagination and the understanding. Even though it appears that we are using reason to decide what is beautiful, the judgement is not a cognitive judgement, "and is consequently not logical, but aesthetical". A pure judgement of taste is in fact subjective insofar as it refers to the emotional response of the subject and is based upon nothing but esteem for an object itself: it is a disinterested pleasure, and we feel that pure judgements of taste, i.e. judgements of beauty, lay claim to universal validity. It is important to note that this universal validity is not derived from a determinate concept of beauty but from common sense [11].

In a quite free and inaccurate reinterpretation, I consider Kant's "free play" between imagination and understanding an inspiration for this work. During the training of the neural network on natural images, we are trying to build understanding of the images in the neural network when we pursue the correct classification of the images in different classes. When we propagate an image through previously learnt filters, the neural network is understanding that image accordingly to its previously formed structures of knowledge. Since the neural network has been trained according to common knowledge (the ground truth of the classification task), the representation it uses can be understood as a form of basic common sense with universal validity, although we are not talking about beauty itself as Kant states.

On the other hand we have the imagination. Imagination is the ability to form new images and sensations in the mind that are not perceived through senses. When the neural network produces a new image using the representation stored inside we can say that is imagining the image. Properly, the neural network imagination process would be the activation of the different neurons in the net, not the back-propagation of it to the image generated itself. The imagination is the representation of the image in the activation of the neurons that we store after the forward propagation from the data layer to the different layers of the net. This process, in the model used, is not a neural memory itself, so the matching of it to the imagination process is not perfect.

In this section of the critique Kant also establishes a faculty of mind that is in many ways the inverse of judgement: the faculty of genius. Whereas judgement allows one to determine whether something is beautiful or sublime, genius allows one to produce what is beautiful or sublime [11].
According to that, we could say that our algorithm is a genius. This is the final part of the algorithm, working in the opposite direction, where from the understanding and the imagination it generates the beautiful object. The process is carried by little changes in the random noise starting point to make it match the different imagined features of the work in its understanding, which may be up to certain point not compatible, finding a compromise between them.

Kant further distinguishes between free and adherent beauty. Whereas judgements of free beauty are made without having one determinate concept for the object being judged (e.g. an ornament or well-formed line), a judgement of beauty is adherent if we do have such a determined concept in mind (e.g. a well-built horse that is recognised as such). The main difference between these two judgements is that purpose or use of the object plays no role in the case of free beauty. In contrast, adherent judgements of beauty are only possible if the object is not ill-suited for its purpose [11].

In this work both kinds of beauty are present. Some of the pictures produced try to keep information on the original object represented, being it a portrait or a landscape, while other just try to find beauty in abstract shapes and lines. At some point the limits between a representation of a known object and an appealing combination of shapes appear diffuse and I consider that the idea of a fuzzy limit is more suitable than a clear distinction. I consider pictures in this fuzzy limit can boost the free play between imagination and knowledge.

The judgement that something is sublime is a judgement that it is beyond the limits of comprehension. The feeling of the sublime describes two subjective moments that concern the relationship of the faculty of the imagination to reason. The mathematical sublime is situated in the failure of the imagination to comprehend natural objects that appear boundless and formless, or appear "absolutely great". In the dynamical sublime there is the sense of annihilation of the sensible self as the imagination tries to comprehend a vast might [11].

I relate this sense "absolutely" to the activation of neurons. They can be overexcited, they can have an absolute response to a very impressive stimulus. In neural networks this phenomenon has a parallelism with the activation of the artificial neurons. Their output can be correlated to the firing rate in human neurons. A maximisation of their output may be a correspondent to an absolute. An image that produces this response may have an impact in us too. The opposite can also be stated: the annihilation can correspond to the negation of impulses, the vanishing of activations.

I want to make clear that the correspondence between Kant theories and the models used in this work is not accurate in the way I stated them, so it is not possible to say, according to them, that the machine is creating art, but the ideas behind the theory are close enough to make us think about it.

A more accurate interpretation is that, if the work produced with these algorithms is able to produce in the subjective human mind that observes its creations the free play relation between understanding and imagination that can be generalised, then it is Art. Given the loose correspondence between the model for the evaluation just mentioned and the model for creation, I consider plausible that this process effectively results in the production of art works.

The autonomy that this philosophical thinking gave to Art, now in capital letters, provides the basis for the artworks as a chance for aesthetic judgement. From different attitudes, the artists expected to interpret the reality creating feelings of vitality, pleasure, imagination and reflection, giving birth to different styles or artistic movements. In the XIX century it starts the period of artistic movements: neoclassicism, romanticism, realism, impressionism, symbolism, art nouveau, modernism,...[3]

In the XX century the emancipation of the beauty as perfectly controlled idea gained strength. In the vanguards, the limits of the beauty grow towards the limits of the language [3]. Among the movements which flowered in the first decade of the 20th century were Fauvism, Cubism, Expressionism, and Futurism. Many other movements flourish during the rest of the century including constructivism, suprematism, dadaism, surrealism, rationalism, pop art, happening, minimalism, hyper-realism, conceptual art, land art,...

Some of these currents try to express ideas or feelings that are not in the conscious part of the mind. Sometimes their greatest achievement is to find an artwork that stimulates connections in our brain that we are not aware of. The kind of these connections can be completely different, from surrealism or impressionism to abstract art, they target to hidden connections in our brain producing this way a stimulus to our mind.
One of the results of this work uses existing art pictures and tries to mix their style with the content from a real photography. As we will see later, this approach uses the textures representation in a neural network to represent the style of the painting. Although many information is lost, and probably no expert would recognise the work produced this way as a part of the original style, the impression in the general public is that the work produced really matches the style of the original. I let the interpretation of that phenomenon on art experts.

2.7 Contemporary art

Today the contemporary art uses the beauty or the ugly equally with the objective of producing aesthetical emotion, going further of the sensible, getting in the field of the aesthetic experience, the emotivity, the play and the surprise. With the raise of the conceptual and popular, the art today evaporates in any human action, the ugly can be aesthetic and the artistic can be anything. The aura that always tried to have the art exists no more. Everything has become invaluable, illogical and indeterminate. During the 80’s the US philosopher Arthur C. Danto called this moment the death of the art [1]. The "end of art" refers to the beginning of our modern era of art in which art no longer adheres to the constraints of imitation theory but serves a new purpose. Art began with an "era of imitation, followed by an era of ideology, followed by our post-historical era in which, with qualification, anything goes... In our narrative, at first only mimesis (imitation) was art, then several things were art but each tried to extinguish its competitors, and then, finally, it became apparent that there were no stylistic or philosophical constraints. There is no special way works of art have to be. And that is the present and, I should say, the final moment in the master narrative. It is the end of the story."[1].

According to this point of view, anything produced in this work would be art, but this is not my personal point of view. Getting deeper in Danto’s point of view, his definition of art has been glossed as follows according to the Stanford Encyclopedia of Philosophy [1]:

“Something is a work of art if and only if:

1. it has a subject
2. about which it projects some attitude or point of view (has a style)
3. by means of rhetorical ellipsis (usually metaphorical) which ellipsis engages audience participation in filling in what is missing, and
4. where the work in question and the interpretations thereof require an art historical context.”

I like to think about the first condition as what is it about?, what are we talking about? Similarly, the second would be what do we say about it?, how do we feel about it? What are the important features of it that we do want to highlight? I consider that these interpretations are included in the definition itself. But, expressing them this way we may realise that we are talking, that we are communicating something, so, the art needs a subject to talk about and a style to say something about it. As we will see later, these two parts of an art work are clearly differentiated in the approach described in the work.

The third condition is the most fascinating. It reminds me about the free play between understanding and imagination of Kant. I think that both are talking about the same idea, difficult to express with words but present in every modern artwork, not to say just any artwork. I don’t know better ways to express it, but the way I understand it is related to the way information is represented in our mind. Our mind has a natural ability to fill the gaps, to provide the missing information according to the context, that relies on our neural structure and the representation of the information in it. Filling this ellipsis, these gaps, can be engaging and I consider this a key to love the art.

Oppositely, I strongly disagree with the fourth condition. It is a condition about cultural background, but is written in the form of restricted art historical context. It is closing the existence of art, so the appreciation of it, to the subset of well formed minds aware of art historical context. I also consider that the cultural background is a key factor of the art. but I would extend it to the more general possible cultural background. Even more, art appreciation
can extend over different cultures and times, and our appreciation of ancient and foreign art is a proof on that. Then, the art background must be something, at least, available for the entire human race. Maybe, it could even be open to other sensible beings.

After opening the range of the fourth condition, I would add another loose condition for the art: originality. I consider that one of the main characteristics of the art, that distinguishes it from craft-work and industrial production, is that is a way to explore new horizons. Art is usually on the edge, exploring new ways of expression, new subjects. Art is not about repeating once and again the same things, it is about doing them in a different way or doing new things, it is a kind of research for expression.

2.8 Creation process

2.8.1 Creation process in humans

It is difficult to define the creative process followed when creating an artwork. According to what I explained before about art I developed my own. The main elements involved in the process of creating an artwork can be roughly modelled according to the diagram in figure 2.8.

![Figure 2.8: Creative process in humans](image)

In blue we can see the mental processes. These processes involve the nervous system for transmitting and manipulating the information. The processes involving manipulation of information are separated in 3 boxes according to their different roles:

- Perception: It bridges the gap between the real world and the representation of it in our mind. It involves not only the senses but also the pre-process of the information collected to build a good representation of it. It receives the inputs from the real world in general, but more specifically for art creation from the tools used and the support that will become the artwork itself when finished.
• Mind: Is the center of the system. It works with high level representations of reality. It generates high level concepts and abstract concepts that do not need correlate directly to perceptions of the real world. It has the ability to imagine a reality that is not perceived. To create this reality it makes use of part of the perception system, to give a sense of real to the concepts generated at high level. Mind is also the responsible of controlling the body, and through it actuate in the real world. The mind, to perfection this control, has the feedback via proprioception.

• Emotions, will: Are the driving force of the mind. Freud called them the Superego and the It. They are affected by the ideas in the mind and the perceptions directly. They are who decide what to do. They many times play a key role in the creative process, being the art itself a manifestation of these forces. The style in the art, the point of view is a way to express them. They are many times the reason that stimulates the creation of art.

The working of the whole system start in the real world. To create an artwork the first thing is to gather the inputs from the real world. These inputs need to be mentally represented, so are encoded in the form of higher ideas.

These ideas can be then processed by the mind and combined with the subjectives point of view coming from the will and the emotions. The ideas are altered and are no more a realistic representation of the world to become a subjective vision.

The mind manipulate the real world to create a representation of the subjective ideas via the hands, or feet or any other way it has to act on the world. These actuators are part of the body but their capabilities can be extended using tools that are part of the world.

The actuation is focused on a part of the world itself, the support for the artwork. The support is going to change during the creation of the artwork to become the artwork itself. The perception that arrives to the mind of the support can modify the artwork to be created. In the same way the perception of the artwork itself during the process of creation can change the idea of it and then modify the final result. When the perception received from the artwork fulfils the requirements of the mind and will, or when the mind recognises that can not find a way to improve it, the artwork is finished.

2.8.2 Creation process proposed

In this work a modular structure based on deep neural networks (DNN) to create artworks is proposed. In figure 2.9 it is shown roughly the general schema used in the generation of artworks as it is coded.

In the lowest part of the schema there are the input images. The left block with 3 images, representing a generic array of N images, are the auxiliary images used for the composition. On top of each image a pre-trained CNN takes this images as inputs to build a representation of them in the activations of their neurons. The architecture, given the adjustments for different sizes of the images, and the weights are the same for all networks in the schema, so they are represented as a unique neural network. The activations on top of the networks are gathered at several levels of depth in the neural networks, not only in the final layer. In the right part we have the image that is being generated at two different stages. The same CNN is used to extract the set of features at the same layers used by any of the other networks for auxiliary images.

On top of the activations we have the error measures. These errors are computed according to the different models available. The models used can be all different or several instances of the same error. The errors are fed by the activations of each auxiliary image and the activations of the image in process of creation. Some proposed models’ error measures do not require an auxiliary image, so they are fed only by the generative CNN. The errors output a loss measure and a gradient according to its definition.

Each of the errors is weighted using specific weights for each layer, each image and each kind of model. The weighted gradients are combined and fed to the generative neural network. Since the errors depend on the depth of the layer they are really combined inside the generative neural network. The gradients of same layer are vectorially added, then they are propagated to shallower layers and there their propagated value combined with other errors of that layer.
This point is not showed in this general schema for the sake of simplicity. The gradient is finally propagated until the image layer is reached.

The arriving gradient is used as input for L-BFGS numerical optimization of the image in process of generation.

Figure 2.10: General schema of the algorithm proposed with finish CNN.

To complete the schema I have to add the finishing process as is shown in figure 2.10. The final image resulting from the previous schema is fed in a special neural network that encodes knowledge about paintings appearance. This Finish CNN removes the noise generated during the image generation process and gives a painting appearance to the final artwork. This process shares a similar structure with the human process of artwork creation. We can fit the main components in the previous schema with little changes, like deleting some of the information flows, now in white, as we can see in figure 2.11.

The point that is out of the system here defined are the emotions, the will. The subjective part of the artwork creation is not included in the schema working in the computer. This part is provided by the human operator. The user defines which images are to be used, which error measures and their relative importance. The user is also the final judge of the artwork created. From this point of view it is clear that an important part of the creation process is done by the
algorithm.

Another, more simple, point of view is to restrict the algorithm in the Tools box. Then all the creation process is done by the human user and the algorithm is just a tool. I consider more appropriate the first possibility since I consider that the role of the AI is capable of getting its own representation of the world. It is obtained through the images presented thanks to its previous training which gave it this ability. It can go beyond the capabilities of the user’s imagination producing completely unexpected results.

I understand this can be a very controversial point subject to different interpretations.
Chapter 3

Methods

3.1 Convolutional Neural Network

The neural network used to represent the features of the original images is the same network that is used for the image generation. In this work it is called composition network. The architecture of convolutions and pooling layers and the weights must be the same. The size is the same in the case of the content model, since all the features are used to compute the error. In the case of the texture model the size is not required to be the same since the information in the Gram matrix is stationary, not dependent of the positions in the image.

3.1.1 VGG19 Architecture

I use the VGG-19 network, a convolutional neural network trained on object recognition that was introduced and extensively described previously [30]. I use the feature space provided by the 16 convolutional and 5 pooling layers of the VGG-19 network and do not use any of the fully connected layers. The network’s architecture is based on two fundamental computations:

1. Linearily rectified convolution with filters of size $3 \times 3 \times k$ where $k$ is the number of input feature maps. Stride and padding of the convolution is equal to one such that the output feature map has the same spatial dimensions as the input feature maps.

2. Maximum pooling in non-overlapping $2 \times 2$ regions, which down-samples the feature maps by a factor of two.

These two computations are applied in an alternating manner. A number of convolutional layers is followed by a max-pooling layer. After each of the first three pooling layers the number of feature maps is doubled. Together with the spatial down-sampling, this transformation results in a reduction of the total number of feature responses by a factor of two. Next table provides a schematic overview over the network architecture and the number of feature maps in each layer. Since only the convolutional layers are used, the input images can be arbitrarily large. The first convolutional layer has the same size as the image and for the following layers the ratio between the feature map sizes (rows and columns) remains fixed. Generally each layer in the network defines a non-linear filter bank, whose complexity increases with the position of the layer in the network. In figure 3.1 we can see the architecture for a typical image used in the tests.

Although other neural networks can be used with the same code I consider that for the purpose of this thesis is better to use only one neural network not to increase the number of parameters to consider in the tests.

3.1.2 Modifications on the network

According to Gatys, [10] replacing the max-pooling operation by average pooling improves the gradient flow and obtains slightly cleaner results. Also, for practical reasons, in [10] the weights in the network are rescaled such that the mean activation of each filter over images and positions
Figure 3.1: List of layers that are used in my experiments and their features. The data dimensions are produced on a base of a 254 × 190 pixels RGB image used in the experiments.

<table>
<thead>
<tr>
<th>layer</th>
<th>weights dimension</th>
<th>activations dimension</th>
<th>kernel size</th>
<th>stride</th>
<th>size reduction factor</th>
<th>influence size</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>(1, 3, 254, 190)</td>
<td>(1, 3, 254, 190)</td>
<td></td>
<td></td>
<td>(1,1)</td>
<td>1</td>
</tr>
<tr>
<td>conv1_1</td>
<td>(64, 3, 3, 3)</td>
<td>(1, 64, 254, 190)</td>
<td>3</td>
<td>1</td>
<td>(1,1)</td>
<td>3</td>
</tr>
<tr>
<td>conv1_2</td>
<td>(64, 64, 3, 3)</td>
<td>(1, 64, 254, 190)</td>
<td>3</td>
<td>1</td>
<td>(1,1)</td>
<td>5</td>
</tr>
<tr>
<td>pool1</td>
<td></td>
<td>(1, 64, 127, 95)</td>
<td>2</td>
<td>2</td>
<td>(2,2)</td>
<td>6</td>
</tr>
<tr>
<td>conv2_1</td>
<td>(128, 64, 3, 3)</td>
<td>(1, 128, 127, 95)</td>
<td>3</td>
<td>1</td>
<td>(2,2)</td>
<td>10</td>
</tr>
<tr>
<td>conv2_2</td>
<td>(128, 128, 3, 3)</td>
<td>(1, 128, 127, 95)</td>
<td>3</td>
<td>1</td>
<td>(2,2)</td>
<td>12</td>
</tr>
<tr>
<td>pool2</td>
<td></td>
<td>(1, 128, 64, 48)</td>
<td>2</td>
<td>2</td>
<td>(4,4)</td>
<td>13</td>
</tr>
<tr>
<td>conv3_1</td>
<td>(256, 128, 3, 3)</td>
<td>(1, 256, 64, 48)</td>
<td>3</td>
<td>1</td>
<td>(4,4)</td>
<td>17</td>
</tr>
<tr>
<td>conv3_2</td>
<td>(256, 256, 3, 3)</td>
<td>(1, 256, 64, 48)</td>
<td>3</td>
<td>1</td>
<td>(4,4)</td>
<td>19</td>
</tr>
<tr>
<td>conv3_3</td>
<td>(256, 256, 3, 3)</td>
<td>(1, 256, 64, 48)</td>
<td>3</td>
<td>1</td>
<td>(4,4)</td>
<td>21</td>
</tr>
<tr>
<td>conv3_4</td>
<td>(256, 256, 3, 3)</td>
<td>(1, 256, 64, 48)</td>
<td>3</td>
<td>1</td>
<td>(4,4)</td>
<td>23</td>
</tr>
<tr>
<td>pool3</td>
<td></td>
<td>(1, 256, 32, 24)</td>
<td>2</td>
<td>2</td>
<td>(8,8)</td>
<td>24</td>
</tr>
<tr>
<td>conv4_1</td>
<td>(512, 256, 3, 3)</td>
<td>(1, 512, 32, 24)</td>
<td>3</td>
<td>1</td>
<td>(8,8)</td>
<td>28</td>
</tr>
<tr>
<td>conv4_2</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 32, 24)</td>
<td>3</td>
<td>1</td>
<td>(8,8)</td>
<td>30</td>
</tr>
<tr>
<td>conv4_3</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 32, 24)</td>
<td>3</td>
<td>1</td>
<td>(8,8)</td>
<td>32</td>
</tr>
<tr>
<td>conv4_4</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 32, 24)</td>
<td>3</td>
<td>1</td>
<td>(8,8)</td>
<td>34</td>
</tr>
<tr>
<td>pool4</td>
<td></td>
<td>(1, 512, 16, 12)</td>
<td>2</td>
<td>2</td>
<td>(16,16)</td>
<td>35</td>
</tr>
<tr>
<td>conv5_1</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 16, 12)</td>
<td>3</td>
<td>1</td>
<td>(16,16)</td>
<td>39</td>
</tr>
<tr>
<td>conv5_2</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 16, 12)</td>
<td>3</td>
<td>1</td>
<td>(16,16)</td>
<td>41</td>
</tr>
<tr>
<td>conv5_3</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 16, 12)</td>
<td>3</td>
<td>1</td>
<td>(16,16)</td>
<td>43</td>
</tr>
<tr>
<td>conv5_4</td>
<td>(512, 512, 3, 3)</td>
<td>(1, 512, 16, 12)</td>
<td>3</td>
<td>1</td>
<td>(16,16)</td>
<td>45</td>
</tr>
<tr>
<td>pool5</td>
<td></td>
<td>(1, 512, 8, 6)</td>
<td>2</td>
<td>2</td>
<td>(32,32)</td>
<td>46</td>
</tr>
</tbody>
</table>

is equal to one. Such re-scaling can always be done without changing the output of a neural network if the non-linearities in the network are rectifying linear 1 . In this work I use these two tricks modifying the VGG19 pretrained neural network. In some of the experiments the network with MAX pooling has also been used providing interesting results.

### 3.1.3 VGG19 Trained features

The trained convolutional network is publicly available [30] and its usability for new applications is supported by the caffe-framework [15] used in this thesis. This CNN has been trained extensively for the ImageNet classification task. The images used for training are photographs.
of objects including a wide range of different classes [4].

We can visualise the filters at each layer of the network in figure 3.2. If we visualise the first filter we can have a visual interpretation of their meaning since they try to match visual features present in the image, like edges or gradients. But when we try to visualise filter deeper than the first layer their interpretation concerning image representation is not straightforward. They rely on previous layer filters so, even though they are clear in relation to previous layer, their final impact on image level is usually impossible to visualise.

### 3.1.4 Other Neural Networks

The model presented in this thesis is not restricted to one specific CNN. It should work well with other architectures and other trainings. Other architectures may give more or less representative power to the network. Other trainings may learn specific features according to the images shown and the works obtained may change.

The only requirement is that the network is trained and is capable of building features representative of the images. Experiments done with random features for the VGG19 have shown that only noise like results can be obtained for texture generation [10]. I consider that these results can be generalisable to other filters and models presented.

The interpretation of this phenomenon is that the neural network used needs to be able to build a representation of the image in the features encoded in the activations of the layers. A network with no training is not encoding any knowledge when an image is passed forward.

It would be interesting to use Deep Convolutional Autoencoders as the neural network used. I consider that this may give an image representation more unbiased since there are no specific classes to focus on during training.

From a philosophical point of view, it makes sense that we need to have a prior understanding about how the world is before trying to mix the ideas that should be build on that understanding.

### 3.2 Image generation

To generate a new image an initial image $\vec{x}_0$ is required. The size of the generated image will be the same of $\vec{x}_0$. Usually the initial image is a random noise image but any image can be used.

The initial image $\vec{x}_0$ is passed through the generation convolution neural network and the activations at each layer $l$ in the network are obtained. Since each layer in the network can be understood as a non-linear filter bank, its activations in response to an image form a set of filtered images $F_l$. A layer with $N_l$ distinct filters has $N_l$ feature maps each of size $W_l \times H_l$ being $W_l$ the number of pixels in the horizontal dimension and $H_l$ the number of pixels in the vertical dimension. The images are represented at layer $l$ by a tensor $F_l$ with dimensions $N_l \times W_l \times H_l$.

The representations at different layers $F_l$ are fed to specific error measures, or distances, for different effects. The different models and errors used are explained in the following sections. These error measures output a gradient for each layer $l$ and each error $D_{el}$. These errors are already weighted by particular weights depending on the layer and error and general weights depending on the kind of error.

For the deepest layer $l_{end}$ where errors are defined, all the gradients from different errors are vectorially added:

$$\vec{D}_l = \sum_{e} \vec{D}_{el}$$

The gradient of a certain layer with respect to the activations of another layer $F_l$ or with respect to the pixels $\vec{x}$ can be readily computed using standard error back-propagation [18].

At each layer $l$ below the top layer all the gradients from different errors and propagated from deeper layers are vectorially added:

$$\vec{D}_l = \sum_{e} \vec{D}_{el} + prop(D_{el+1})$$
The gradient that is propagated to the image (to the pixels) can be used as input for some numerical optimisation strategy. In this work I use L-BFGS [37] which seems a reasonable choice for a high dimensional optimization problem.
This process is repeated until optimization is obtained or, more usually, a certain maximum number of iterations is reached.

3.3 Models and Losses

3.3.1 Content model

To model the content of an image I use the activations of the neural network at a certain layer when the image is presented.

To characterise the content of the image \( \hat{x} \) in this model, I follow the same procedure that was used to get a representation of the generated image passing the image \( \hat{x} \) through the convolutional neural network and computing the activations for each layer \( \hat{F}^l \) with dimensions \( N_l \times W_l \times H_l \).

The error used to compute the content similarity is the mean squared distance between the tensor representing the content and the corresponding tensor computed from the image to generate.

\[
E^l = \frac{1}{2} \sum_{i,j,k} (F^l_{ijk} - \hat{F}^l_{ijk})^2
\]

and the total loss is

\[
\text{Loss}(\bar{x}, \hat{x}) = \sum_{l=0}^L \omega_l E^l
\]

where \( \omega_l \) are weighting factors of the contribution of each layer to the total loss. The derivative of \( E^l \) with respect to the activations in layer \( l \) can be computed analytically:

\[
D_l = \frac{\partial E^l}{\partial F^l_{ij}} = (F^l - \hat{F}^l)_{ij}
\]

This loss and gradient are fed to the generation network. Note that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

3.3.2 Texture model

For the texture modelisation I use the strategy proposed by Leon A. Gatys, Alexander S. Ecker and Matthias Bethge [10]. This strategy relies on the idea of using Gram matrices on top of Convolutional Neural Network (CNN) features as a descriptor for the texture. The same idea has also been proposed by T. Lin, A. RoyChowdhury and S. Maji in [20], where they call it Bilinear CNN (B-CNN) or later [19] Bilinear features. In this work I am going to use the Gram name since it was in their work [9] that it was first proposed to be used with a content descriptor to generate images, although [19] proposes a similar approach.

The basic idea to generate a texture from a given source image is: first extract features of different sizes homogeneously from this image. Next compute a spatial summary statistic on the feature responses to obtain a stationary description of the source image.

In this work I use the feature space provided by the activations of the neurons of a high performance convolutional neural network and the spatial summary statistic are the correlations between neuron activations in each layer of the network.

To characterise a texture \( \hat{\bar{x}} \) in this model, I do same process that in previous section but now the vectorized version of the features \( \hat{F}^l \) will have dimensions \( N_l \times M_l \) being \( M_l = W_l \times H_l \) the feature maps vectorised. \( \hat{F}^l_{jk} \) is the activation of the \( j \) filter at position \( k \) in layer \( l \). A summary statistic that discards the spatial information in the feature maps is given by the correlations between the responses of different features. These feature correlations are, up to a constant of proportionality, given by the Gram matrix \( \hat{G}^l \) of dimensions \( N_l \times N_l \), where \( \hat{G}^l_{ij} \) is the inner product between feature map \( i \) and \( j \) in layer \( l \):

\[
\hat{G}^l_{ij} = \sum_k \hat{F}^l_{ik} \hat{F}^l_{jk}
\]
A set of Gram matrices $G^1, G^2, ..., G^L$, from some layers $1, ..., L$ in the network in response to a given texture provides a stationary description of the texture, which fully specifies a texture in this model.

Let $\hat{x}$ and $x$ be the original model image and the image that is generated, and $\hat{G}^l$ and $G^l$ their respective Gram-matrix representations in layer $l$. The contribution of layer $l$ to the total loss for this model is then:

$$E^l = \frac{1}{4N_i^2M_i^2} \sum_{i,j} (G^l_{ij} - \hat{G}^l_{ij})^2$$

and the total loss is

$$\text{Loss}(\hat{x}, x) = \sum_{l=0}^{L} \omega_l E^l$$

where $\omega_l$ are weighting factors of the contribution of each layer to the total loss. The derivative of $E^l$ with respect to the activations in layer $l$ can be computed analytically:

$$D_l = \frac{\partial E^l}{\partial F^l_{ij}} = \frac{1}{N_i^2M_i^2} (F^l_{ij}^T (G^l - \hat{G}^l))_{ji}$$

This loss and gradient are fed to the generation network. Note again that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

### 3.3.3 Activation maximisation model

This model is inspired in the Google Inception [27]. Unlike the previous models it is not applied on an image to characterise but on the same input image, starting point for image generation. Like the other models used it is applied successive times, every time on the output of the previous iteration.

In the same way of content model in section 3.3.1, from an input image $\hat{x}$ the activations are computed for each layer $F^l$ with dimensions $N_i \times W_i \times H_i$.

The error used to maximise the activation of the features in the framework of a minimisation problem, compatible with other models, is minus the squared value of the activations of the features:

$$E^l = -\frac{1}{2} \sum_{i,j,k} (F^l_{ijk})^2$$

and the total loss is

$$\text{Loss}(\hat{x}) = \sum_{l=0}^{L} \omega_l E^l$$

where $\omega_l$ are weighting factors of the contribution of each layer to the total loss. The derivative of $E^l$ with respect to the activations in layer $l$ can be computed analytically:

$$D_l = \frac{\partial E^l}{\partial F^l_{ij}} = -(F^l)_{ij}$$

This loss and gradient are fed to the generation network. Note again that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

### 3.3.4 Activation maximisation model on image

This model is again inspired in the Google Inception [27] but in the case of the features to be enhanced are to be found in an auxiliary image instead that in the network itself. It is also applied on the same input image, starting point for image generation. Unlike the previous model it requires an auxiliary image where the features appear. The auxiliary image is like an index of features that can be used for the maximisation of the features appearing in the main
image. Like the other models used it is applied successive times, every time on the output of the previous iteration.

In the same way of content model in section 3.3.1, from an input image $\vec{x}$ and auxiliary model image $\hat{\vec{x}}$ the activations are computed for each layer $F^l$ and $\hat{F}^l$ with dimensions $N_l \times W_l \times H_l$ and $\hat{N}_l \times \hat{W}_l \times \hat{H}_l$.

To compute the error first we compute the similarity of the features of both images using the inner product. Then we select the features that give a maximum similarity independently of their positions in the images.

$$FA = \arg \max((F^l)^T \cdot \hat{F}^l)_i$$

The error used to maximize the activation of the features in the framework of a minimization problem, compatible with other models, is minus the squared value of the activations of the features that maximized the inner product:

$$E^l = -\frac{1}{2} \sum_{j,k} (FA^l_{jk})^2$$

Note that this error is independent of the position of the features in the auxiliary image but not in the generated image.

The total loss, as usual is weighted in every layer:

$$\text{Loss}(\vec{x}, \hat{\vec{x}}) = \sum_{l=0}^{L} \omega_l E^l$$

where $\omega_l$ are weighting factors of the contribution of each layer to the total loss. The derivative of $E^l$ with respect to the activations in layer $l$ can be computed analytically:

$$D_l = \frac{\partial E^l}{\partial F^l_{ij}} = -(FA^l)_{ij}$$

This loss and gradient are fed to the generation network. Note again that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

### 3.3.5 Activation minimisation model

This model make a behaviour opposite to previous one in section 3.3.3. The model wants to reduce to a minimum the activation of the layers in a certain layer. This could be understood as a regularisation used in training but the effects here are different. It is applied on the same input image, starting point for image generation. Like the other models used it is applied successive times, every time on the output of the previous iteration.

In the same way of content model in section 3.3.1, from an input image $\vec{x}$ the activations are computed for each layer $F^l$ with dimensions $N_l \times W_l \times H_l$.

The error used is the activation of the features itself squared:

$$E^l = \frac{1}{2} \sum_{i,j,k} (F^l_{ijk})^2$$

and the total loss is

$$\text{Loss}(\vec{x}) = \sum_{l=0}^{L} \omega_l E^l$$

where $\omega_l$ are weighting factors of the contribution of each layer to the total loss. The derivative of $E^l$ with respect to the activations in layer $l$ can be computed analytically:

$$D_l = \frac{\partial E^l}{\partial F^l_{ij}} = (F^l)_{ij}$$
This loss and gradient are fed to the generation network. Note again that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

### 3.3.6 Activation minimisation model on image

Combining the previous models we can apply the minimisation model on auxiliary image. This time we have two versions of the model that make some sense. The initial approach is the same, minimising the activations that produce a high response in both, generated and auxiliary image. The second one is to minimise the activations that already produce a low correlation. While the first should tend to make the image more different, minimising what they have in common, the second should make them more similar reducing what is not similar.

In the same way of content model in section 3.3.1, from an input image $\vec{x}$ and auxiliary model image $\hat{\vec{x}}$ the activations are computed for each layer $F_l$ and $\hat{F}_l$ with dimensions $N_l \times W_l \times H_l$ and $\hat{N}_l \times \hat{W}_l \times \hat{H}_l$.

To compute the error first we compute the similarity or dissimilarity of the features of both images using the inner product. Then we select the features that give a maximum or minimum similarity independently of their positions in the images.

For maximum similarity we have:

$$FA_{max} = \text{argmax}((F_l^i)^T \cdot \hat{F}_l^i),$$

and for minimum

$$FA_{min} = \text{argmin}((F_l^i)^T \cdot \hat{F}_l^i),$$

The error used to maximise the activation of the features in the framework of a minimisation problem, compatible with other models, is minus the squared value of the activations of the features that maximised or minimised the inner product:

$$E_l = -\frac{1}{2} \sum_{j,k}^{} (FA^l_{jk})^2$$

Note that this error is independent of the position of the features in the auxiliary image but not in the generated image.

The total loss, as usual is weighted in every layer:

$$\text{Loss}(\vec{x}, \hat{\vec{x}}) = \sum_{l=0}^{L} \omega_l E_l$$

where $\omega_l$ are weighting factors of the contribution of each layer to the total loss. The derivative of $E_l$ with respect to the activations in layer $l$ can be computed analytically:

$$D_l = \frac{\partial E_l}{\partial F_{ij}^l} = -(FA^l)_{ij}$$

This loss and gradient are fed to the generation network. Note again that in the case the activation of a specific neuron is 0 the gradient will not be propagated for that neuron.

### 3.3.7 Other models

Given the modular structure of the approach presented it is clear that other models can be developed. The models can rely or not on auxiliary images. New models can find other kinds of relations between features that have not been explored yet. Also it is possible to think on combinations of presented models.

### 3.4 Artistic Finishing

One of the problems of the system used for generation of images from a DNN representation is that during the generation of the image some noisy artifacts may be included.
Some authors [19] suggest using an error soft constraint on the image pixels during the generation of the image. This is usually called image prior $\Gamma(x)$. The most common value for it is the TV norm:

$$\Gamma(x) = \sum_{ij} (x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2$$

which encourages nearby pixels to have similar values and give smoother reconstructions.

Although this approach would be easy to introduce in the general schema used in this work, I decided not to use it. I considered that sometimes this normalisation may make the picture generated appear blurred, lose strength, making the images obtained more dull. I did not make any tests on that, just checked the pictures presented by other authors. It is also a personal point of view. From my point of view it is worth to try something different.

CNN had been used for image super-resolution achieving good performance on the details [6]. They are also a good denoiser, another work used this property to build a compression denoiser [5]. I considered to use a similar approach using a neural network trained not on real images but on paintings to denoise the final images, giving a painting style finishing at the same time.

![Figure 3.3: Schema of the CNN architecture used for super-resolution [6]](image)

As we can see in figure 3.3, given a low resolution image $\tilde{y}$, the first convolutional layer of the Super-Resolution CNN (SRCNN) extracts a set of feature maps. The second layer maps these feature maps nonlinearly to high-resolution patch representations. The last layer combines the predictions within a spatial neighbourhood to produce the final high-resolution image $\tilde{F}(\tilde{y})$.

The hardware restrictions made that idea not practical since the training time will be too long for this thesis. Fortunately I found a shortcut. A Japanese software engineer had been training a CNN on Japanese anime and made it publicly available [24]. It can be found as a ready to use app in the web under the name Waifu2X [25].

Finally I used the pretrained neural network on anime with different degrees of denoising according to the finishing I wanted to give to the final pictures. In some images I used it several times to obtain smooth images, in others I used the real images filter to obtain less noise filtering. This process worked very well, keeping sharp edges and giving continuity to the colour at the same time.
Chapter 4

Experiments

4.1 Hardware used

For all the computations done in this thesis, a laptop with a CPU Intel Core i5-4200 and a GPU Nvidia Geforce 820M has been used. The use of a generic laptop demonstrates that the use of Deep Neural Networks is not constrained to research labs equipped with powerful processors but, on the other hand, sets restrictions on the computations carried on.

From the first tests it was clear that the use of GPU for the processing of forward and backward propagations in the neural network was the fastest option, but even in this case the time required to compute one model with 2000 iterations is around one hour.

The memory available in the GPU, 2Gb, is partially used for error correcting and visualization, leaving only around 1,2Gb of free memory for nets. This restricted the size of the images to use to 2202 pixels. This value was obtained after several tentatives with bigger images and different models. Obviously the number of constraints and models used made the maximum size variable, but to ensure that experiments could be finished this size was chosen.

The low resolution is not a big problem for the experiments since the networks used are usually trained with 2562 pixels, so the features trained are mainly contained in the image size used. The flexibility of caffe to adapt to different input sizes is also key for the success of this work. From the artistic point of view it is obvious that bigger images would be preferable but, since the cost of obtaining them in time would result in less experiments, I consider it is better to keep them small and do more tests.

To overcome these constraints a schedule of the experiments was required. I made a version of the code able to run several experiments sequentially. After first trials I used a slightly faster version performing around 15 experiments in each run, spending a total time of about 10-12 hours.

4.2 Images used

The choice of what images are more suitable for preliminary experiments and art generation is not trivial. In order to make easier the evaluation of the results obtained and achievement of appealing art works the images used are carefully selected.

One of the most studied problems in computer vision is face recognition. It has been studied for long time and includes all kind of classification problems ranging from gender recognition to age estimation and properly face recognition for identification purposes. It has also been demonstrated that the ability that humans have for face recognition is learnt in early stages of development and that is higher than any other visual task. Evolution favoured individuals with higher ability recognising the small differences from one face to another and the differences that show particular emotions. This natural training makes us experts understanding faces and, at the same time, finding strange elements in a face or detecting faces in images where there is no face actually. This is the reason why I use face images in many of the experiments. I used Geoffrey Hinton’s photo as a homage for his work in developing neural networks, which are the basis of this work.
I also use some landscapes to see the impression produced for images generated in the case of a image of reality that is natural too but we are not so much used to it.

For textures I use different kind of images that have a texture appearance. I specially focus on images with straight lines. Similarly to the case of the faces explained before we also have a special ability to recognise straight lines and regular geometric patterns. I think it is interesting to use them to have a better understanding of the behaviour of the texture algorithms used.

I also use images of gradients of colours, like the colour palettes, to give the neural networks a representation that includes all kind of colours for the images reconstruction.

### 4.3 Content model

One of the first experiments to do is to test the content model. In the following figures 4.1 and 4.2 we can see the images obtained from the representation encoded in the features of only one layer. In all these tests the initial image used is random noise. From this initial image the algorithm tries to find an image that minimises the error in feature representation at specified layer.

![Image Reconstruction](image)

Figure 4.1: Image reconstruction from content features. For all images: `contentweight = 10^4`, maximum number of iterations = 2000

In figure 4.1 we can see that the reconstruction of the original image is decreasing its quality as deeper layers are used. The images are very similar to the original image for lower layers, up to 'pool3'. In 'pool3' the smallest features appear blurred but the shapes very similar, with some noise. In the reconstruction built only with features from 'pool4' the quality is not that good. The details in the borders of the coins and in the figures of the coins are not restored properly. Nevertheless, the general appearance of the image is quite good and the objects can be easily identified as coins. In 'pool5' reconstruction lots of artifacts appear, the shapes are very distorted and it is hard to recognise the coins if that image is presented independently. However, in the context of the figure we can find it similar to the original and the general texture is kept.

In figure 4.2 we do the same experiment using a face as the image to be reconstructed. The behaviour is similar to the previous experiment. Up to 'pool3' layer the reconstructed images are very similar to the original one. In 'pool4' some artifacts and noise start to appear. It is not specially noticeable but in the eyes. In 'pool5' the image obtained is completely weird. It does not keep any trait of the original image. Unlike the previous test, now it is completely unrecognisable the original figure.
Figure 4.2: Image reconstruction from content features. For all images: \( \text{contentweight} = 10^3 \), maximum number of iterations = 2000

In the paper [26] the authors explain this phenomenon of multiple possible inputs that minimise the error in the feature representation of a CNN but have no apparent similitude with the original.

This experiment shows that in low to high layers ('pool4' is layer 16th) it is possible to obtain a high precise Reconstruction of the original image from random noise. Using only very high layers ('pool5' is layer 21st) it is not possible unless additional constraints are provided. The data compression in these levels is so high that many images can produce the same feature representation. Most of these images are not natural at all, so are unrecognisable for us. This result can give some clues about the way the information is encoded in CNN trained for image recognition and their difference with the way we do in our brains.

From the art creation perspective, it is interesting the different level of abstraction that each level provides on the content represented. When we want to combine different "ideas" from different images, it is important that the representation of the content allows different solutions that still are representative of the content. In this case the results obtained for layers 'pool5' and 'pool4' are specially interesting. The non perfect, or "wild", reconstruction obtained show that there is space for more that one solution that still matches the main features represented in those layers. This fact may allow a better combination of the content with other features, and so give more space for the "fair play" mentioned in section 2.6.

Another interesting result of this experiment was found in the early stages of evolution by an artist consulted. When he saw the images said that the first images were showing the most important structure of the content image. Besides the interest he found from an artistic point of view, latest studies reveal that the same kind of images can be used to make Saliency Maps for classification tasks that can be used for automatic segmentation [29].
4.4 Texture model

4.4.1 Weights

The implementation includes a general factor for the weights of gradients for all layers and sources. In the algorithm the gradient is not normalised but each loss is multiplied by a different weight before the back-propagation.

It has been observed that for small weights for the features the minimisation problem can be solved with only one iteration. This produces images that are almost random noise. I consider that this is a problem of numerical stability, since with small values for the gradients they can be so close to zero that after propagation no optimisation can be practically done.

In the images of figure 4.3 it can be seen that a very large weight can make the gradient drive optimisation to extreme values in pixels. However after a few iterations the results are practically the same.

As a rule of thumb, minimum weights of the order of $10^3$ are required. Higher weights do not improve final image and do not speed up optimisation process.

4.4.2 Layers

The selection of layer used for the texture reconstruction has an important effect on the results. Their representation is based on the features encoded by different layers and it was shown in previous section that this representation is different according to the layer. Figures 4.4 and 4.5 show the images obtained for different combinations of layers for a particular texture. The images used in the experiments in figures 4.4 and 4.5 where chosen because they have geometric shapes that can be easily identified visually.

The first interesting thing we can see in these images is the size of the features that they are able to encode. As we increase the depth of the layers used, the size of the features increase according to the area covered by those features. As we have seen in section 3.1.1 the size of the area influencing one feature increases with the depth of the layer. A feature of layer 'conv1_1' gathers information from a 3x3 pixels area, while a feature from 'pool5' is based on a 46x46 pixels square.

Nevertheless, we can see that the area of the visual features of the experiments is bigger than this size. According to its formulation in section 3.3.2, the Gram matrix encodes the spatial correlation between the different features. That is which features tend to have a similar activation in all places across the image. If in all the places feature $i$ has a strong activation and feature $j$ has also a strong activation and vice-verse, then the Gram matrix value $G_{ij}$ will be close to 1. If they are not related they will be close to 0. But the features share some overlapping area, in our case up to $2/3$ of their total area of influence. So their activations are not completely independent. This interdependence in layers features increases the size of the features that Gram matrices apparently encode.

In the images of coins in figures 4.4 and 4.5 we can see that the complete ellipsoid/rounded shape of the coins is only achieved in 'pool4' and 'pool5' experiments. For lower layers the features are too small.

In the case of the mosaic at 'pool3' the sizes and square form are kept. In pool2 we can see that the vertical and horizontal lines are coded, forming clear rectangles, but not the proportion. In previous layers the lines are more blurred, although they keep the direction.

In the gravel images we see again that the hexagonal line orientation is present from the most simple representations. In 'pool2' it starts to appear the idea that different colours should group together and a cell wall should be placed between. In successive images this grouping feature is more evident. The cell size also tends to be similar to the original. Nevertheless any of the generated images has been able to evolve to a point where the two colours are separated in only two groups or where the hexagonal structure is coherent in all the image. This can have two explanations: the random starting point made the network evolve to a local minimum and is not able to find a better solution, or the size of the features is not big enough to penalise these errors as happens in the lower lever models.

In the last of the examples, the colours images, we can see again the size of the scope of the features in the first two images. It is interesting to notice that in general the bright colours are
Figure 4.3: Effect of order of magnitude of weights in image generation. A certain minimum weight is required. Very high weights do not affect results after a few iterations. Only texture model is used. Layers used: ‘pool4’, ‘pool3’, ‘pool2’, ‘pool1’, ‘conv1_1’
Figure 4.4: Effect of the layers used in image generation starting from lowest layers. Each row includes the layer stated and also the previous ones. i.e for last row are used layers: ’pool5’, ’pool4’, ’pool3’, ’pool2’, ’pool1’ and ’conv1_1’. We can see that the complexity and size of the features increase when using deeper layers.

not touching each other but placed in a grey background. If we look close to the original image we realise that there are shadows between colours. The small scope of the network may be using this feature as a general border between colour blobs of different random sizes, more related to initialisation than to their representation area. In the third image the idea of vertical lines gains
Table 4.5: Effect of the layers used in image generation starting from lowest layers. Each row includes the layer stated and also the previous ones. i.e for last row the used layers are: 'pool5', 'pool4', 'pool3', 'pool2', 'pool1' and 'conv1_1'. We can see that the complexity and size of the features increase when using deeper layers.

importance but is not until the 'pool2' that is associated with colour changes. The horizontal relations between colours are also kept at this level disappearing the grey background. From this point the change in successive deeper models is the increasing vertical size of the colours, that I consider related to the size of the scope of the features in deeper layers.

Figure 4.5: Effect of the layers used in image generation starting from lowest layers. Each row includes the layer stated and also the previous ones. i.e for last row the used layers are: 'pool5', 'pool4', 'pool3', 'pool2', 'pool1' and 'conv1_1'. We can see that the complexity and size of the features increase when using deeper layers.
We can also see the effect of choosing a deep layer alone, without any other of lower complexity. The general appearance of the results for higher layer alone is very similar to the one obtained in the previous section using only one layer for content generation. It is obvious that texture features based only on one layer can not code information that is out of what this layer is capable to provide.

![Figure 4.6: Effect of the layers used in image generation starting from higher layers. We can see that when only using deeper layers details are lost, degenerating the images and appearing artifacts.](image)

On the other hand, we have seen in figures 4.4 and 4.5 not all the layers are required to obtain a visually representative reconstruction of the original features in a high level. Choosing only some spaced layers we can code all the information required for a high quality reconstruction of the texture of the image. For the highest level reconstruction, from deepest layer ‘pool5’, only 5 additional layers out of 20 are used to obtain the images presented in the figure 4.5.

### 4.5 Texture and content models

In this section we are going to combine the two models, for content and texture using different source images.

#### 4.5.1 Weights

The ratio of the weights used for content and texture models is critical for the results obtained. For that reason, besides the general weight applied to all layers and the weight for each particular
layer, the algorithm implements a general texture/content ratio. It is usually expressed in this text like $\frac{\text{tex}}{\text{cont}} = \frac{10^7}{10^4}$ where the $10^7$ is the total texture weight and $10^4$ the total content weight. For these experiments the layer particular weight is usually set to 1.

In all experiments done the texture weight needs to be several orders of magnitude higher than the content weight. However the result is image and layer dependant, so a general rule cannot be established for all combinations of content images and textures and layers. In figure 4.7 some examples are shown on the layer dependence of the weights. In figure 4.8 there are examples of the different effects that same weights have depending on the images used as a seed for content and texture.

It is important to notice that the similarity of the images used by the composition is important for the results obtained. Very different images possibly will produce strange compositions like the two first columns of figure 4.8. In the other hand the use of textures that can match the features of the image will tend to produce images like the fourth column in figure 4.8.

Other experiments showed that the use of black and white content image with coloured textures usually result in poor matching. Best results are usually obtained being both colur images.

<table>
<thead>
<tr>
<th>Tex/Cont</th>
<th>Content</th>
<th>Texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^7/10^4$</td>
<td>'pool5'</td>
<td>'conv1_1'</td>
</tr>
<tr>
<td>$10^{11}/10^2$</td>
<td>'pool5'</td>
<td>'conv1_1'</td>
</tr>
<tr>
<td>$10^{10}/10^3$</td>
<td>'pool1'</td>
<td>'poolI'</td>
</tr>
<tr>
<td>$10^{11}/10^2$</td>
<td>'pool4'</td>
<td>'conv1_1'</td>
</tr>
</tbody>
</table>

Figure 4.7: Image reconstruction from content and texture features. The columns correspond to different combinations of content and texture layers. Different texture/content ratios effect is shown in each row. Initial image is random noise. Maximum number of iterations = 1500.

### 4.5.2 Use content image as initial image

One of the phenomenons observed in the experiments mixing content and texture from different images is that the random initialisation can make the generated image fall in a local minimum driven by a texture that matches well the original random. The generating process is not able to jump out of this minimum and the mixture of content and image is not the one that reflects best the content. To overcome this problem one option is to initialise the network with the content image. The constraint on content and texture are the same and the image generated correspond equally to a minimum combined error. I repeated some of the images in figure 4.7 but now initialising with the content image instead of random noise.
Figure 4.8: Image reconstruction from content and texture features. Effect of different images and textures on weights effect. Different texture/content ratios effect is shown in each row. Maximum number of iterations = 1500.

| content: 'pool4' | content: 'pool4' | content: 'pool4' |
| texture: 'conv1_1' | texture: 'conv1_1' | texture: 'conv1_1' |

Figure 4.9: Image reconstruction from content and texture features. Initial image is the content image. The columns correspond to different combinations of content and texture layers. Different texture/content ratios effect is shown in each row. Maximum number of iterations = 1500

We can see the results in figure 4.9. It is clear that the faces look more natural and many of the distortions are now removed. In the image using 'pool2' 'pool1' and 'conv1_1' for texture and a tex/cont $= 10^{11}/10^2$ we can see now a face, that in the previous image was not possible to see.
4.5.3 Evolve texture on content image

Another possibility that this algorithm gives is to use only the texture model on an initial image. The content is not used in the optimisation of representation, only to initialise the texture reconstruction. The results of this technique in figure 4.10 show that can achieve a perfect texture reconstruction. The price is that the original subject may be lost or heavily deformed in the way.

The method has the advantage of being usually faster that when using also the content model for the loss computation. This is because the algorithm is faster and also because the minimum is reached sooner.

The other strong point is that during the optimisation it creates textures that can be even more interesting than the final result.

![Initial image](image1.png) ![Texture image](image2.png)

Figure 4.10: Generation of images from an initial content using only texture model. The sequence is formed by images at different iterations of the evolution.

4.5.4 Texture and content from same image

The reconstruction of the original image using only the content representation at a certain layer has been shown in section 4.3. In this section the information encoded in the content representation is complemented with the information encoded in the texture representation of the same image.

I consider it is interesting to know if with a combination of content features and texture features from the same image is possible to obtain a good reconstruction of the image. It can be useful to get a better knowledge of the kind of information encoded in each level. And it can also provide interesting visual effects. In figure 4.11 some of the experiments done are shown.

If the results of the figure 4.11 and the previous figure 4.2 are compared we see that the images in the first figure for content models up to layer ‘pool3’ are very similar. Adding the texture does not improve the reconstruction and, in content ‘pool3’ they are even slightly worse if we look at the eyes. The images for content from ‘pool4’ are of similar quality but different strange errors appear in each of the images corresponding to different texture models. In these cases it is clear that the additional information provided by the texture is not improving the reconstruction of the original image.

Completely different is the case of the content supplied by layer ‘pool5’ features. We saw in 4.2 that the images generated using only content can be very wild versions of the original image. Now part of this big space of images that minimise the loss in the content model is restricted by the texture model. The images obtained in all cases have a range of colours similar to the original image. In the direction the texture representation uses deeper layers, the features of the image tend to be bigger and represent understandable features of the original image. Using ‘pool3’ and ‘pool4’ texture layers features recognisable as parts of the face appear. Their positions are usually also close to what should be. Similarly it happens in the background and the shoulders. The image gains similitude with the original.
In the image combining content from 'pool5' and textures from 'pool4' and below recognisable part of the face are more clear although their position and composition in approximate. From an artistic perspective, it seems a nightmare representation of a face in the limit of our compression capabilities. An artist pointed that it is interesting how similar it is to Francis Bacon’s work [2].

### 4.5.5 Drawings

I wanted to test the capability of the system to deal with more idealised forms of expression. I consider that drawings are more ideal since many of the features that appear in a natural image have been removed and only a simplification of them is present. I consider that a very simplified kind of drawing is the manga or anime or many other cartoons. In the case of charcoal drawings
the representation includes lots of shadows and tend to have an appearance close to real images but the manga just uses simple lines while keeping a good expressivity.

For this reason I chose a manga drawing to do some tests with. In the figures 4.12 and 4.13 we can see the results. In general the assimilation of textures from a manga content is poor. The general appearance of the images gets more realistic, closer to what would be a charcoal drawing style. Although in the more shadowed areas of the hair in 4.12 A and B we can see some assimilation, the rest of the generated image only gets a general background colour from the texture real image. In figure 4.12 C we can see that if we use another cartoon image for the texture, the transfer of plain colour without figure knowledge is more clear.

The main reason for this behaviour is that the network used was originally trained on real world images. That is to say it never saw a drawing before. It had no opportunity to learn features that represent the features existent in drawings and no connection of these features to real features. It has no implicit knowledge over the world linking shapes and natural colours. The features do not disentangle the shapes and the colours and show a poor performance on symbolic representations like manga drawings. Two possible conclusions are that either this knowledge is in more abstract levels of image understanding or it requires a specific training.

The case of 4.12 D is slightly different. Here the content is coded in the last layer of the network, trying to get the deepest representation possible of the manga. In this case we see an interesting behaviour in the left eye of the figure. It seems that the CNN coded the information of this area as an "eye" feature, so the output tries to reconstruct the idea of eye that the network learnt instead of the real manga shape. But this similitude does not appear in any other place, so we can not get conclusions from it.

The opposite experiment, to give manga textures to a real image, is shown in 4.13. In this case we can see that the final appearance of the generated image is far from having a manga style or even a similitude with a drawing.

Actually experiments have shown the difficulties of the algorithm even to match a grey-scale content with a colourful texture. The idea behind this experiments was to find a way to give colour to old grey-scale images. I would expect that the shape features could lead to a proper colour filling but the results were disappointing.

4.5.6 Using MAX pooling instead of AVE pooling

As we stated in section 3.1.1 we use the VGG19 CNN with the modification of the pooling layer from MAX pooling to AVE (average) pooling. According to [10] this change helps to obtain smoother less noisy images. I did some tests using the original VGG19 with MAX pooling to see the importance of this effect. In figure 4.14 I repeat some of the previous experiments of figure 4.8 using MAX pooling while keeping all the other parameters, except the last one that now is initialised with the content image.

We can see in this images in figure 4.14 that a very little noise is added respect to corresponding ones in previous figure 4.8. There is a clear increase in the sharpening of the images, clearer edges appear. The unexpected change is in the parts of the original image that appear almost unchanged. In the second image appears a complete face appearing where there was nothing before. But the effect is even more dramatic in the last image with an eye in the middle of blue squares.

I tried to use the MAX pooling version in the experiment using the content and texture models on the same image. We can see the result in figure 4.15.

It is clear that the quality of the images is not too much affected and the results are surprising. This effects are not possible to achieve with the AVE version as far as I tested it. The use of MAX pooling opens another universe of possibilities for the algorithm to produce different kind of creations.

4.6 Activation maximisation model

Inspired in Google Inception [27] this algorithm tries to maximise the sum of the activations of the neurons at a certain layer. The algorithm has some important changes compared to the original. Some modifications are required to be able to work with the other models included in
Figure 4.12: Generation of images from a manga content and real textures. **Image A,B,C:** tex/cont $= 10^4$; content layers = 'conv5_1'; texture layers = 'pool3', 'pool2', 'pool1', 'conv1_1'. **Image D:** tex/cont $= 10^4$; content layers = 'pool5'; texture layers = 'pool4', 'pool3', 'pool2', 'pool1', 'conv1_1'.
Figure 4.13: Generation of an image from a real content and manga textures. \( \text{tex/cont} = 10^4 \); content layers = 'conv5_1'; texture layers = 'pool3', 'pool2', 'pool1', 'conv1_1'.

Figure 4.14: Image reconstruction from content and texture features using MAX pooling. For all images \( \text{tex/cont} = 10^{11}/10^2 \). Maximum number of iterations = 1000.

The algorithm. Others are done to obtain a different result. This algorithm does not intend to create dreams from the images that end up in a representation of the neural network knowledge, not related to the image used. The intention is to emphasise the features present in the image that are significant for the neural network.

One of the changes that showed to have a significant impact in the results is the neural network used. Like in all the rest of the experiments the neural network used is the VGG19 modified with average pooling. The information this network encodes in each layer is different from the GoogLenet used by Google. This has a major impact in the features generated.

Besides the impact of the neural network architecture and training, the use of average pooling have also a big impact in the results. In figure 4.18 we can see the results when we use MAX pooling instead of AVE pooling. The result is more different when we use the model in pooling layers.

From an artistic point of view it is clear that this images achieve the goal of maximising the activations of our neurons as well as artificial ones. The images are impacting but the interesting thing is that the shapes they emphasise have been always there. Even in the last images of the sequences we can see, in some cases, similarities with the original one. They are still a representation of the same subject.

The images at first iterations show a subtle increase of certain features already present in the image. This effect remembers me the Greek sculptors that made the muscles of the bodies more sharp that they were in reality, actually, more sharp that they can never be. This way they emphasised that feature. This experiments show that this model can be used in the same
Figure 4.15: Image reconstruction from content and texture features from same image using MAX pooling. \( \frac{tex}{cont} = 10^7/10^4 \). Maximum number of iterations = 1000.

### 4.7 Activation maximisation on image model

This group of experiments perform, like the previous one, a maximisation of the sum of activations at a certain layer of the network. In this case an auxiliary image is provided and the features to maximise must be strong in that image as well. The results of the experiments are divided in two figures, 4.19 and 4.20, for convenience of presentation.

In this set of experiments we test different weights for the loss and the gradient. We can see that the range of weights that work well in these experiments depend on the layer used. It actually depend on the strength of the features to maximise. And this strength depend on the correlation of features between principal and auxiliary images. In any case we see that for large weights we obtain stronger results.

In the case that there is almost no relation between the two images in that layer, as happens in layer 'pool5', no image is computed or if obtained is almost the original one. This algorithm use to stop after a few iterations since at some point the features stop being similar because the comparison is always done on the generated image at each iteration.

For understanding of neural networks it is interesting this algorithm because it produces a clear strong output when the features of both images at a certain level are related and almost no output otherwise. It visually answer the question "what do have in common this pair of images at this lever of representation". It is also possible to locate the features in the original image and visualise their target. In the final image we see also the shape of their preferred output, so it is very clear what do they represent.

From the artistic point of view this maximisation of activation can emphasise features present in both images. This deformation of reality can be visually attractive, and in some cases even shocking. Personally, I consider some of the images created with this technique and the previous one, very artistic and visually attractive.

This tool can also be useful for previous drafts of the artwork to be created, avoiding the colourful explosion it usually generates. The fact of restricting the features used to be present
in another image give an incredible analysis power. It can also be used to emphasise features that unconsciously we find in other images in the same way the CNN does.

4.8 Activation minimisation model

As we have seen previously maximisation of the activation can produce interesting images that are at the same time informative of the process inside the neural network. In this set of experiments the method tested is the opposite. The goal is to minimise the sum of the activations of the neurons at a certain layer. The results for different layers and iterations are showed in figures 4.21 and 4.22. Some of the images are very similar at same iterations, so different iterations have been chosen for representation.

It is interesting to see in figures 4.21 and 4.22 that lower layers tend to make a completely grey image, while highest layers can be in a minimum with a recognisable image. In all layers above 'conv1_1' the final image is not completely grey, it has some patterns although they are almost unrecognisable. In layers 'conv4_4' and 'pool4' we can almost see the face in the image.

Another feature of this images is that each layer draws a special pattern in the image. This should be the visually opposite of the pattern that this layer recognises. The visually opposite is not the negative of the pattern but an independent one, so it produces no activation.

The images in the sequence of iterations look phantasmagorical modifications of the original. Contrary to the opposite algorithm that draws bright colours, this tend vanish the images. The deformations that occur to the face while vanishing also have a phantasmagorical distortion. The effect is the best seen on video of the sequence of images in both directions. There we can also appreciate better that the dissolution of the original shape generates waves around it. I consider that this waves are generated when gradient in the image vanish generating smaller gradients around which in the next iteration also vanish repeating the cycle. This process makes the original face look smaller every time, being the process faster in the areas where there are less features.

It is very interesting to realise that the images that we associate with live are the ones that maximise the activations of the CNN while the ones we associate with death are the ones that minimise the activations. Also the process of vanishing produces shapes with small mouths and big empty eyes that are the one we usually use to represent phantoms.

4.9 Activation minimisation on image model

The same process used in section 4.8 can be used in the case of a minimisation filter. As we will see, now makes sense to try two different approaches.

The initial approach is the same, minimising the activations that produce a high response in both, generated and auxiliary image. The second one is to minimise the activations that already produce a low correlation. While the first should tend to make the image more different, minimising what they have in common, the second should make them more similar reducing what is not similar.

A problem with this solutions is the initial similarity of both images. As we confirmed in the previous experiment in section 4.7 the process tend to make little changes in the images when they are not so similar. Here we are trying to reduce the activations that are used in that comparison so a little change in the images may happen.

When I did the experiments I confirmed that for high specialised layer there is no effect or very small. In lower layer, more generic there are more similarities so the algorithm can produce something. In figure 4.23 we can see the comparison of the outputs of both algorithms for minimisation of activations of 'pool3' using both approaches.

We can see in the images of figure 4.23 that the features minimised in both cases are very different. Any of them is similar to the features minimised in previous experiment with same layer and image in figure 4.22.
4.10 Iteration process

4.10.1 Fastening boundaries

If we look at the films of the evolution of the texture pictures we can realise that the boundaries of the picture start to get the appropriate shape before the middle parts of the picture. It is a phenomenon that had been usually observed in the experiments conducted. The interpretation of that phenomenon is based on the constraints during the optimisation process.

We know that the features used for the optimisation have a spatial area of influence. The higher level the feature is, the bigger area of influence it has. The texture features are actually the spatial correlation of the activation of layer features. A pixel placed close to the edge of the image is connected to less feature influence areas, because there are less areas overlapping on that pixel, than in the case of a pixel in the centre. When optimisation is conducted the general optimisation gradient sum the influence of all the features that overlap their influence areas on that pixel. These gradients are based on all the area of each feature, so each of them, according to its influence area, may be pushing the pixel value in a different direction. In the sides of the images there are less features pushing, so it is easier to find a coordinated direction for all of them, a common agreement, and the values of all affected pixels move to that common minimum faster.

This phenomenon has a strong effect in the final image. When boundaries are fastened to their local minima, it is more difficult to find a common minimum for the centre of the image. The fact of having a poor matching in the middle has an impact on the general loss computed but this is, in many cases, smaller than the impact that would have to change the value of all neighbouring pixels, increasing for each of them their contribution to the total loss. The usual result of this is that the image get stuck in a local minimum with some “unmatched” areas instead of evolving to a global minimum where all the parts find a global organisation that matches the patterns represented in the full extent of the image.

The experiments with highly structured images in textures using high level layers make clear to our vision this phenomenon, while in less structured images it is not so easy to see. This ability to realise the disorganisation in some kind of images while not in others could be discussed from a neuroscience point of view.

To overcome this limitation the optimisation process should be carried gradually, starting from the centre of the image and gradually increasing to the boundaries. This approach could be implemented:

1. Starting image smaller that target image.
2. Generate image
3. Extend image with randomly generated boundaries
4. Use extended image as a seed for next generation step.
5. Repeat from 2 to 4 until the desired size is achieved.

In the code used in this thesis this could be done defining progressively decreasing non-optimisation masks in the boundaries.

Another possibility is to assign bigger weights to the optimisation losses according to their position in the image. This would have two effects:

1. Increase the magnitude of the gradients in the central area, speeding up the optimisation in this areas.
2. Make possible for central areas with bad matching to surpass low losses of well matching surrounding areas.

To introduce this solution in the code used in this thesis is not as much straightforward as the previous one. If we want to apply it to the final loss in the final layer we have the problem that the Gram matrix computed is spatial independent. An indirect approach would be to act directly on the gradients applied to the image. This, actually, would be quite similar to the first solution applying a weighted mask instead of a none/all mask.
4.10.2 Change waves and the end of a generation

We can see in some parts of the images a wave texture. We can see it clearly in the sequences of images of figure 4.24. The images in the first row show the generation of an image of only texture. While the images in the second row combine the same texture with a content. In this case the ratio of weight is \( \frac{\text{tex}_w}{\text{cont}_w} = \frac{10^{10}}{10^2} \).

In the first row, in the first image, we can see a strong wave effect over all image, while in the second it almost disappeared over all the image. In the third we can not see any of these features. It is possible to observe that between the first and second images there are important differences in the images. Mainly some small features disappeared and the image is more consistent according to high level features. The second and third images are very similar. It is important to notice the larger number of iterations between the second and third images than between the first and second.

In the second sequence we can see that, in the first image, forms are still being generated and there are no visible waves. In the second one there are a lot in the central area that is changing at this moment. In the third one they are reduced but we can still see some in the boundaries of the shapes.

After many experiments I realised that these waves appear in areas that are changing in that moment and their absence over all the image is a sign of stabilisation of the optimisation process. Nevertheless, sometimes they don’t disappear even when the image is clearly stable and almost no noticeable changes occur for 1000 iterations. This phenomenon happens more often and is more pronounced in images that combine different features.

In figure 4.25 we can see that the size of the waves is similar to the size of the features coded in the first layer of the network. They are not just random changes pixel by pixel, noise like, but of the size of the features of the lowest level layer, ‘conv1_1’. This fact points in the direction that there is relation with this layer activations.

Actually, we don’t see the size of the gradients in this layer but the size of the plain areas. In other experiments we see that waves of this size are stable when the texture used has many abrupt changes in the colour. My explanation is that the neural networks knows that it should be a plain area but there is contradictory information on its colour. I consider that these waves are caused by a low level texture layer whose minimum loss correspond to them and that gets stuck in a local minimum. When the higher layer gradient propagation makes the image displace from this minimum point and change the features activated in lower layers for others corresponding to a consistent solution they disappear. Usually they still remain close to some edges probably because the higher level gradients in these areas are not strong enough to jump over the local minimum for lower layers.

The main importance of this fact is that it becomes, with the lack of important changes in last iterations, a visual indicator that the result is stable and more iterations would make no significant difference in the image produced.

4.11 Finish

In this section we show the effect of applying the finish network to upscale the images to double size while at the same time reducing noise and giving a painting appearance. In the examples in figures 4.26 4.27 we can see that the effect sometimes is too strong and some interesting features of the image are lost. In other cases is preferable a more smooth result. We can see in the Maximisation of activation images that the process use to keep the sharp edges even in high frequency images.
Figure 4.16: Generation of images maximising the activation of neurons at certain layer. The different columns show different number of iterations. For first 4 columns the number of iteration is different since interesting features appear at different times depending on the layer used. The lists of iterations are [layer: A,B,C,D]: ['conv1_1': 50,100,200,300], ['pool1': 100,250,333,450], ['pool2': 100,300,500,800], ['pool3': 10,20,30,50], ['pool4': 15,25,45,80], ['pool5': 19,33,75,170]. The unnormalized and AVE pooling version is used. The general weight applied in all cases is 1
Figure 4.17: Effect of gradient normalisation in generation of images maximising the activation of neurons at certain layer. We can see that different layers represent different kind of features. The difference between the normalised and unnormalised gradient versions of the algorithm is presented. The AVE pooling network is used. The general weight applied in all cases is 1.
Figure 4.18: Effect of MAX pooling in CNN architecture in generation of images maximising the activation of neurons at certain layer. We can see that different layers represent different kind of features. For first 4 columns the number of iteration is different since interesting features appear at different times depending on the layer used. The lists of iterations are [layer: A,B,C,D]: [pool4: 5, 10, 15, 50], [conv4_1: 3, 5, 8, 12], [conv4_2: 3, 6, 10, 100], [conv4_3: 3, 6, 11, 22], [conv4_4: 3, 15, 44, 100], [pool5: 3, 10, 66, 120]. Use normalise. General weight = 1
Figure 4.19: Maximisation of activations that are present in an auxiliary image (part 1). In all the experiments the network stopped before reaching the maximum number of iterations. The number of iterations achieved are [layer: $w = 10^{-1}, w = 1, w = 10^1, w = 10^2$]: ['pool3': 6, 6, 22, 20], ['pool4': 2, 12, 15, 15], ['pool5': 0, 0, 0, 0]. Notice that for layer ‘pool5’ the CNN did not produce a single image for $w \leq 10^1$. 
Figure 4.20: Maximisation of activations that are present in an auxiliary image (part 2). In all the experiments the network stopped before reaching the maximum number of iterations. The number of iterations achieved are [layer: $w = 10^3$, $w = 10^4$, $w = 10^5$, $w = 10^6$]: ['pool3': 20, 19, 20, 20], ['pool4': 15, 16, 13, 16], ['pool5': 0, 1, 1, 1]. Notice that for layer 'pool5' the CNN did not produce a single image for $w \leq 10^3$. 
Figure 4.21: Minimisation of the sum of neurons activation at a certain layer. Different iterations are chosen at different layers [layer: A,B,C]; ['conv1_1': 5, 12, 178], ['conv1_2': 12, 16, 1032], ['pool1': 6, 12, 2000], ['conv2_2': 10, 20, 2000], ['pool2': 2, 6, 2000]. C layer is always the last. The AVE pooling network is used. The general weight applied in all cases is 1.
Figure 4.22: Minimisation of the sum of neurons activation at a certain layer. Different iterations are chosen at different layers [layer: A,B,C]: ['conv3_4': 7, 20, 2000], ['pool3': 5, 33, 2000], ['conv4_4': 22, 111, 2000], ['pool4': 10, 200, 2000], ['pool5': 100, 500, 2000]. C layer is always the last. The AVE pooling network is used. The general weight applied in all cases is 1
Figure 4.23: Minimisation of the sum of neurons activation at layer 'pool3' with maximum and minimum correspondence between original and auxiliary images. The AVE pooling network is used. The result correspond to the last iteration before algorithm finish. The general weight applied in all cases is $10^{-3}$. 
Figure 4.24: The waves that appear in the image tend to disappear when the image stabilises. The first row of images use only texture while the second combines the same texture with content.

Figure 4.25: Relation of the size of the waves with the size of the features coded in layer 'conv1_1'
Figure 4.26: Effect of the finish neural network filter. All images are upscaled 2X in the process. The level of noise reduction is different in each column.
Figure 4.27: Effect of the finish neural network filter. All images are upscaled 2X. The level of noise reduction is different in each column.
Chapter 5

Discussion and conclusions

5.1 CadiArtNatura’s presentation

In order to have an independent evaluation of this work from an artistic point of view I contacted a group of artists called CadiArtNatura [17]. This group share a studio in Berga, organise different artistic events and keep a permanent exposition in Espai d’Art La Duana. Most of the members of the group are painters. Some of them work professionally in art creation and teaching and the majority sell regularly their works.

The responsible of organizing the expositions gave me the chance of making a presentation of this work in Espai d’Art La Duana on 15th of January. The presentation was open to everybody although it was not announced outside the artists group. I presented this work, with a brief introduction on Artificial Intelligence and Neural Networks. Some of the works produced were printed and distributed among the audience.

After the presentation we had an interesting debate. The general opinion was that the works presented are art. Most people recognised that if I did not tell before the would never think that they were done by a computer. A professional painter stated that in the actual art period, that he called the Death of Art, everything is art; furthermore, the works presented can be considered even classical.

Most of the assistants found very interesting the capacity of mix the shapes and textures from two images. The ones that had been working on it were amazed how easy was for the computer to do it, and the results obtained. An artist said he once tried to make a picture similar to one of the presented and remembers how difficult was to get the draft combining the ideas. Another said it could help him with one project involving a woman and a tree for which he could not find a fully satisfactory solution yet.

Other artists were amazed by the process of evolving of the images. They were more focused on simplification in their work. They considered that the images formed in the early iterations were very interesting because they were just the essence of the forms. I have to recognise that I did not consider this point of view before.

Some pointed to the idea of combining other painters textures to their compositions. I explained the work of A. Gatys et. al. [9]. and they found it amazing and suggested other variations.

They did not find especially interesting the images formed by the excitation or suppression of neural activations. Which I consider extremely interesting because of the relation of the impressions they produce and their origin.

They considered that this can be a very useful tool for making drafts for posterior paintings. This tool can allow them to try different things before starting to paint, saving a lot of time. A painting teacher said that with this tool you can do 90% of the preparation work automatically.

Talking about the reactions and evaluations on specific pictures:

- Most of the assistants recognised an impressionist style in “Girasols”. It liked in general to everybody. The version where it is not finished did not like so much but was also thought to be impressionist.
• One of the presents suggested that “Hinton by VGG” remembers Francis Bacon’s portraits. In general nobody liked it.

• Some cubist influence was found in “Blue Hinton”. Artist liked it more that general public.

• Approximately half of the presents recognised the shape of a nude woman in the abstract picture “Nu” (did not know the title).

• Most of the public liked the game of colors in the serie “Colors” although nobody thought that the image used was the same for the three pictures.

• Among the presented printed works the Public Award was undoubtedly for “Violí” (fig.5.1.

![Image of a violin](image.png)

Figure 5.1: Violí. The Public Award in the presentation in Espai d’Art La Duana.

With this feedback I have to consider that this work raise interest in the art world beyond the academic or philosophical research.

5.2 Conclusion

In the review of history and philosophy of art I remarked the importance of the representative function of art. It can be the most powerful tool in man history to encode and transmit the knowledge about the world and ourselves. It has been proven useful to represent objective and subjectives visions of reality. It can express the mathematical relations found in nature and our exaggerated ideals driven by unconscious impressions. It combines reality and well structured
ideas, the input of the senses and preconception of the mind. And, successfully, transmits all that to other people, being able to introduce simple concepts or to evoke feelings.

Kant states that beauty is not a property of an artwork or natural phenomenon, but is instead a consciousness of the pleasure that attends the 'free play' of the imagination and the understanding. This idea is connected to the representation the deep neural networks provide. The process of the process of creation of an artwork have been studied and a parallelism with the process of image generation described here established.

The framework proposed for art creation is modular. Its elements are well defined and can be combined in multiple forms to produce different results. New elements and combinations can be created in the future to expand its possibilities. The different models can be combined in parallel or sequentially.

The content model is able to produce a good reconstruction of the original image but, when deepest layers are used, it is also possible to produce free representations that stimulate the 'free play' in our mind.

The texture model provides a good representation of the textures in the original image, keeping their most important elements. It also have flexibility enough to establish a compromise with the content model mixing the information of both representations intricately.

The maximisation of activations model generates striking images. The images that produce high activations of artificial neurons seem to be capable to stun human minds as well. In the other hand, they prove to be an effective visualisation of the features encoded in the neural networks and the inputs they are connected to. Even more, they can be used to compare images according to each feature level.

The minimisation of activations model was already expected to show a behaviour opposite to the maximisation, but the phantasmagorical figures that appeared are amazing. All the characteristics of these images: lack of colour, edges, blurred shapes are associated with death, but the change of the shapes of the face ended in a mask of phantom. This results denote that phantoms really exist in computers' minds, their characteristics are the ones that their neurons perceive when their activation in front of a real image is reduced. Is that true for humans as well? Do we see phantoms when our neurons slow down? And another philosophical thought: the final minimum activation is not for any colour, nor for black, nor for white, is for grey.

All this results have been produced without any special training of the neural network on art. The neural network used have only been trained on natural images from ImageNet, so it does not have any specific knowledge beyond the reality. This proved to be a handicap when drawings were presented. This may be a sign that the ability to draw is not directly related to the ability to understand the images in the same way that object recognition is.

To overcome this handicap a specially trained neural network have been used. This last module completes the schema that is able to generate art in a form that is similar to paintings. Here the neural networks show their capacity to encode implicit knowledge acquired during their training.

After the feedback from the artists of Cadi Art Natura I feel that the algorithm and models presented are a first step in the use of the representative power of Deep Neural Networks in the creation of Artworks, and envision that it provides an exciting new tool for artists.

The evaluation of the quality of the artworks presented is subjective, and all opinions are welcome. I can only say that I am satisfied.


References


Appendix A

Selection of artworks produced
Figure A.1: Hinton by VGG.
Figure A.2: Blue Hinton.
Figure A.3: Nu.
Figure A.4: Colors1.
Figure A.6: Colors3.
Figure A.7: Nothing?
Figure A.8: eyes.
Figure A.9: Daisy Hinton.
Figure A.10: MAX3.
Figure A.11: Fantasma.
Figure A.12: Hinton color.
Figure A.13: Violí. The Public Award in the presentation in Espai d’Art La Duana.
Figure A.14: Feelings1.
Figure A.15: Feelings2.
Figure A.16: Feelings3.
Figure A.17: Girasols.