UNIVERSITÀ DI PISA

Artificial Intelligence and Data Engineering

Computational Intelligence and Deep Learning

Artist Identification with Convolutional Neural Networks

Project Documentation

TEAM MEMBERS:
Edoardo Fazzari
Mirco Ramo

Academic Year: 2020/2021

Contents

1 — Introduction

Artist identification is traditionally performed by *art historians* and *curators* who have expertise and familiarity with different artists and styles of art. This is a complex and interesting problem for computers because identifying an artist does not just require object or face detection; artists can paint a wide variety of objects and scenes. Additionally, many artists from the same time period will have similar styles, and some such as **Pablo Picasso** (see figure 1) have painted in multiple styles and changed their style over time.

Figure 1: Both of these paintings were created by Pablo Picasso, but they have vastly different styles and content.

The aim of this project is to use Convolutional Neural Networks for the identification of an artist given a painting. In particular, the CNN networks will be modeled using multiple techniques: from scratch; via pretrained networks and using an ensemble network made of the best trained classifiers.

1.1 State of the Art

As mentioned, artist identification has primarily been tackled by humans. An example of that is the Artsy's Art Genome Project [1], which is led by experts who manually classify art. This strategy is not very scalable even if it is highly precise in the classification (the site is a marketplace of fine-arts for collects, you can find Pisarro, Bansky and other famous artists).

Most prior attempts to apply machine learning to this problem have been feature-based, aiming to identify what qualities most effectively distinguish artists and styles. Many generic image features have been used, including scale-invariant feature transforms (SIFT), histograms of oriented gradients (HOG), and more, but with the focus on *discriminating different styles* in Fine-Art Painting[2].

The first time the problem of artist identification was really tackled was with J. Jou and S. Agrawal[3] in 2011, they applied several multi-class classification techniques like Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, K-Means and SVMs and achieve a maximum classification accuracy of 65% for an unknown painting across 5 artists. Later on, the problem of identifying artists was retackled by the *Rijksmuseum Challenge*[4]. The objective of the challenge was to predict the artist, type, material and creation year (each of them was a different challenge) of the 112,039 photographic [5] (containing different viewpoints of an artwork, and different types of them: sculptures, paintings, saucers, etc.) reproductions of the artworks exhibited in

[1] https://www.artsy.net/categories

[2] T. E. Lombardi. The classification of style in fine-art paint- ing. ETD Collection for Pace University, 2005

[3] J. Jou and S. Agrawal. Artist identification for renaissance paintings.

[4] T. Mensink and J. van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. 2014

[5] The dataset contains 6,629 artists in total, with high variation in the number of pieces per artist. For example, Rembrandt has 1,384 pieces, and Vermeer has only 4. There are 350 artists with more than 50 pieces, 180 artists have around 100, and 90 artists have 200 pieces.

the Rijksmuseum in Amsterdam (the Netherlands). For the artist classification challenge, the paper said they reached a test accuracy of about 60%. The year later, Saleh and Elgammal's paper[6] was the first attempt to identify artists with a large and varied dataset, but still using generic features. The collection they used has images of 81,449 fine-art paintings from 1,119 artists ranging from fifteen centuries to contemporary artists, reaching an accuracy of 59%[7].

More recent attempts are related to the *Painter by Numbers*, a **Playground Prediction Competition** by *Kaggle*[8]. This competition used a pairwise comparison scheme: participants had to create an algorithm which needs to examine two images and predict whether the two images are by the same artist or not. Thus, it is not our same objective, however it can be consider the first application of Deep Learning to the problem. The real deal was taken by Nitin Viswanathan[9] in 2017. Viswanathan, using the same dataset of the mentioned *Kaggle Challenge*, proposed the use of ResNet with transfer learning (he first held the weights of the base ResNet constant and updated only the fully-connected layer for a few epochs). This trained network reached a train accuracy of 0.973 and a test accuracy of 0.898.

1.2 Dataset

Unfortunately, the dataset provided by the *Kaggle Challange* is huge thus unfeasible to be used in Colab: in fact it is about 60GB: unbearable on the free version of Colab, which provides only about 30GB of disk. Stated that, we decided to use a different dataset[10] with only 2GB of data and about 8k unique images.
The data downloaded from Kaggle has the following directories and csv file:

```
/
├── images
│   └── images
│       ├── Albrecht_Durer
│       ├── Alfred_Sisley
│       ├── Amedeo_Modigliani
│       ├── Andrei_Rublev
│       ├── Andy_Warhol
│       ├── Camille_Pissarro
│       ├── Caravaggio
│       ├── Claude_Monet
│       ├── Diego_Rivera
│       ├── Diego_Velazquez
│       ├── Edgar_Degas
│       ├── Edouard_Manet
│       ├── Edvard_Munch
│       └── an many others (total of 50 different artists)
├── resized
└── artists.csv
```

The *resized* directory is not useful for our studies, hence we deleted it to save space on the disk. On the other hand, we first use the *csv* file to select only the artists with at least 200 pieces, this operation was done to reduce the number of classes to a number per which the ratio between the number of artists and images was reasonable for learning. Even done that, the dataset was still unbalanced, e.g. Van Gogh's paintings are 877 against the 239 of Chagall's,

[6] B. Saleh and A. M. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. CoRR, abs/1505.00855, 2015

[7] In the paper they tried to use also CNN, but reaching only an accuracy of 33.62%

[8] https://www.kaggle.com/c/painter-by-numbers/data

[9] Nitin Viswanathan, Artist Identification with Convolutional Neural Networks

[10] https://www.kaggle.com/ikarus777/best-artworks-of-all-time

thus we consider to compute **class weights** in order to use them in the *fit function*:

$$\text{class\_weights} = \frac{\text{Total number of paintings considered}}{\text{Number of artists considered} \cdot \text{Number of paintings per author}}$$

Then, we modified the structure of the *images/images* directory in order to create two directories, **train** and **test**, containing 90% and 10% of the images from each different artist's directory respectively (considering only the artists with at least 200 paintings). The newly created directories have the same structured of *images/images*. This was done in *python* in this way:

```
import os
import numpy as np
import shutil

rootdir= '/content/images/images' #path of the original folder
classes = os.listdir(rootdir)

for i, c in enumerate(classes, start=1):
  if c not in artists_top_name.tolist():
    shutil.rmtree(rootdir + '/' + c)
    continue
  if not os.path.exists(rootdir + '/train/' + c):
    os.makedirs(rootdir + '/train/' + c)
  if not os.path.exists(rootdir + '/test/' + c):
    os.makedirs(rootdir + '/test/' + c)

  source = os.path.join(rootdir, c)
  allFileNames = os.listdir(source)

  np.random.shuffle(allFileNames)

  test_ratio = 0.10
  train_FileNames, test_FileNames = np.split(np.array(allFileNames),
                                                        [int(len(allFileNames)*
                                            (1 - test_ratio))])

  train_FileNames = [source+'/'+ name for name in train_FileNames.tolist()]
  test_FileNames = [source+'/' + name for name in test_FileNames.tolist()]

  for name in train_FileNames:
    shutil.copy(name, rootdir +'/train/' + c)

  for name in test_FileNames:
    shutil.copy(name, rootdir +'/test/' + c)
```

After that we created the train/validation/test-sets using the *image_dataset_from_directory* function provided by **Keras** in the following way:

```
import tensorflow as tf

training_images = tf.keras.preprocessing.image_dataset_from_directory(
    TRAIN_DIR, labels='inferred', label_mode='categorical',
    class_names=None, color_mode='rgb', batch_size=BATCH_SIZE,
    image_size=(IMAGE_HEIGHT,  IMAGE_WIDTH), shuffle=True, seed=RANDOM_SEED,
    validation_split=VALIDATION_SPLIT, subset='training',
    interpolation='bilinear', follow_links=False
)

val_images = tf.keras.preprocessing.image_dataset_from_directory(
    TRAIN_DIR, labels='inferred', label_mode='categorical',
    class_names=None, color_mode='rgb', batch_size=BATCH_SIZE,
    image_size=(IMAGE_HEIGHT, IMAGE_WIDTH), shuffle=True, seed=RANDOM_SEED,
    validation_split=VALIDATION_SPLIT, subset='validation',
    interpolation='bilinear', follow_links=False
)
```

```
18
19 test_images = tf.keras.preprocessing.image_dataset_from_directory(
20     TEST_DIR, labels='inferred', label_mode='categorical',
21     class_names=None, color_mode='rgb', batch_size=BATCH_SIZE,
22     image_size=(IMAGE_HEIGHT, IMAGE_WIDTH), shuffle=True, seed=RANDOM_SEED,
23     interpolation='bilinear', follow_links=False
24 )
```

Where *VALIDATION_SPLIT* is equal to 0.1.
Obtaining in this way:

- 3478 files for training (belonging to 11 classes).
- 386 files for validation (belonging to 11 classes).
- 438 files for testing (belonging to 11 classes).

Hence, we have a total of 4299 different pictures.

2 — General Information Useful for Training

In the following chapters we will make use of different strategies:

- Class Weights (already talked about)
- Data augmentation
- Regularization
- Dropout
- Multiple activation functions
- Multiple optimizers
- Genetic Algorithms

In order to allow a better and faster reading of the tests done, in the following paragraph the mentioned strategies are discussed.

2.1 Data Augmentation

Data augmentation takes the approach of generating more training data from existing training samples by augmenting the samples via a number of random transformations that yield believable-looking images. The goal is that, at training time, your model will never see the exact same picture twice. This helps to expose the model to more aspects of the data so it can generalize better. In Keras, this can be done by adding a number of data augmentation layers at the start of your model. In our model, we included the following transformation:

```
data_augmentation = ks.Sequential(
    [
        layers.RandomFlip('horizontal'),
        layers.RandomRotation(0.1),
        layers.RandomZoom(0.2),
        layers.RandomHeight(0.1),
        layers.RandomWidth(0.1)
    ]
)
```

2.2 Regularization

Regularization techniques are a set of best practices that actively impede the model's ability to fit perfectly to the training data, with the goal of making the model perform better during validation. This is called "regularizing" the model, because it tends to make the model simpler, more "regular", its curve smoother, more "generic"; thus it is less specific to the training set and better able to generalize by more closely approximating the latent manifold of the data. A common way to mitigate overfitting is to put constraints on the complexity of a model by forcing its weights to take only small values, which makes the distribution of weight values more regular. This is called *weight regularization*, and it's done by adding to the loss function of the model a cost associated with having large weights. This cost comes in two flavors:

1. *L1 regularization*—The cost added is proportional to the absolute value of the weight coefficients (the L1 norm of the weights).

2. *L2 regularization*—The cost added is proportional to the square of the value of the weight coefficients (the L2 norm of the weights).

3. *L1_L2 regularization*—Combine L1 and L2.

2.3 Dropout

Dropout is one of the most effective and most commonly used regularization techniques for neural networks; it was developed by Geoff Hinton and his students at the University of Toronto. Dropout, applied to a layer, consists of randomly dropping out (setting to zero) a number of output features of the layer during training.

In **keras** can be set using the *layers.Dropout* function passing as parameter the *dropout rate*. We tried different values for the dropout rate during our studies, anyway for the *Pre-Trained Models* chapters it is always set to 0.5 if not otherwise specified.

2.4 Activation Functions

In the studies done in the following chapters we used three different activation functions:

- $ReLU$: $\max(0, x)$
- ELU: $\max(0.2x, x)$

They will be useful in the genetic algorithm analysis done fore the *scratch architecture* and the *VGG16*

2.5 Optimizers

An optimizer is the mechanism through which the model will update itself based on the training data it sees, so as to improve its performance. In our project we make use of:

- *RMSprop*: the gist of RMSprop is to:
  - Maintain a moving (discounted) average of the square of gradients
  - Divide the gradient by the root of this average
  - It uses plain momentum, not Nesterov momentum.
- *Adam*: stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments.

2.6 Genetic Algorithms

Genetic algorithms are a family of search algorithms inspired by the principles of evolution in nature. By imitating the process of natural selection and reproduction, genetic algorithms can produce high-quality solutions for various problems involving search, optimization, and learning. At the same time, their analogy to natural evolution allows genetic algorithms to overcome some of the hurdles that are encountered by traditional search and optimization algorithms, especially for problems with a large number of parameters and complex mathematical representations. Thus, they come in handy for optimizing our networks. In order to make use of genetic algorithms we must decide some components, which are:

- *Genotype*: the *genotype* is a collection of genes that are grouped into chromosomes.
- *Population*: at any point in time, genetic algorithms maintain a population of individuals (i.e., chromosomes)– a collection of candidate solutions for the problem at hand.

- *Fitness Function*: at each iteration of the algorithm, the individuals are evaluated using a fitness function (also called the target function). This is the function we seek to optimize or the problem we attempt to solve.
- *Selection Algorithm*: after calculating the fitness of every individual in the population, a selection process is used to determine which of the individuals in the population will get to reproduce and create the offspring that will form the next generation.
- *Crossover Algorithm*: to create a pair of new individuals, two parents are chosen from the current generation, and parts of their chromosomes are interchanged (crossed over) to create two new chromosomes representing the offspring.
- *Mutation Algorithm*: the purpose of the mutation operator is to periodically and randomly refresh the population, introduce new patterns into the chromosomes, and encourage search in uncharted areas of the solution space.
- *Elitism*: described in the following paragraph.

All of these components are implemented using the python library **deap**[11].

2.6.1 Elitism

While the average fitness of the genetic algorithm population generally increases as generations go by, it is possible at any point that the best individual(s) of the current generation will be lost. This is due to the selection, crossover, and mutation operators altering the individuals in the process of creating the next generation. In many cases, the loss is temporary as these individuals (or better individuals) will be re-introduced into the population in a future generation.

However, if we want to guarantee that the best individual(s) always make it to the next generation, we can apply the optional elitism strategy. This means that the top n individuals (n being a small, predefined parameter) are duplicated into the next generation before we fill the rest of the available spots with offspring that are created using selection, crossover, and mutation. The elite individuals that were duplicated are still eligible for the selection process so they can still be used as the parents of new individuals.

Elitism is made possible in our code thanks to the function *eaSimpleWithElitism*, which is a modification of the function *eaSimple* present in the **Deap** framework.

[11] https://deap.readthedocs.io/en/master/

3 — CNN from Scratch

This chapter shows the results of the training of several custom architecture, that have been defined in order to solve the classification task. Starting from a very simple model, we start to analyze how to improve it and what modifications to apply in order to improve performance, taking into account mainly the accuracy on the validation test, but also considering other metrics like training time or number of parameters. The overall strategy is the following:

- test of different custom architectures defined from scratch
- analysis of the level of fitting, try of different techniques to fight possible underfitting/overfitting
- experiment with the addition of Batch Normalization
- hyperparameters optimization on the best model so far

The objective of the presented procedure is not the total exploration and exploitation of the search space, but it aims at finding good results in a reasonable time exploiting an ad-hoc heuristic search. The tested models are the following:

3.1 Standard CNN

3.1.1 Simple model

The first experiment has been conducted using a customized standard CNN that exploits Convolutional Layers and Max Pooling to process input images. To start, we defined a very simple model, whose structure is reported in the image 2. This network is mainly a starting point of our trial-and-error approach and will give us a first approximation of what is going to be our prediction power on the considered task.

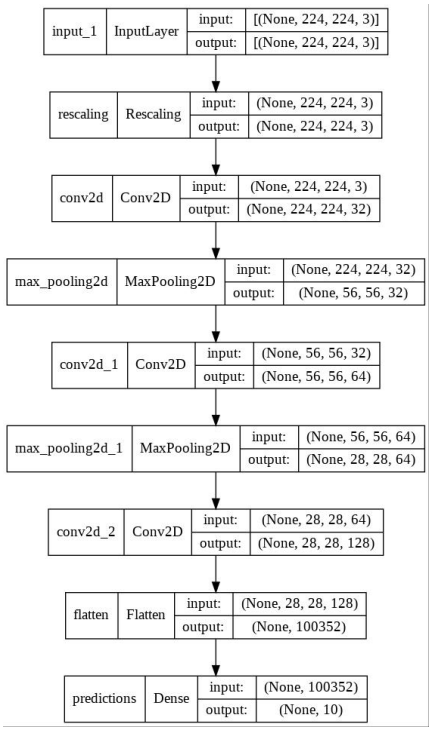


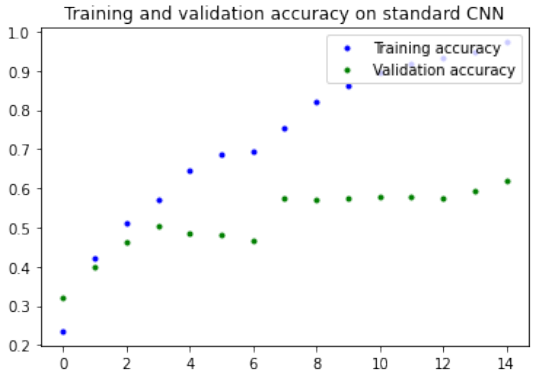
Figure 2: Customized Standard CNN Architecture

This model is trained using these default hyperparameters:

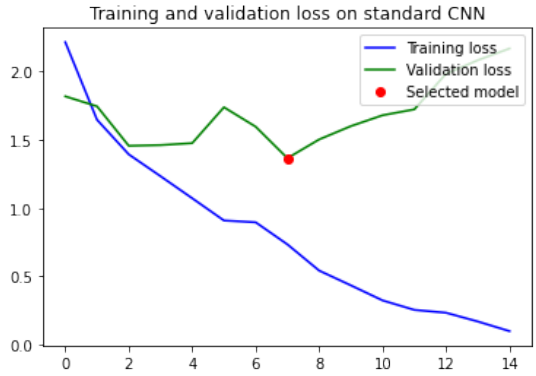
- optimizer: *ADAM*
- dropout rate: 0.0
- learning rate: optimizer's default
- batch size: 128
- learning rate decay: optimizer's default

In particular, we set a large value for batch size both because our main goal is to maximize the accuracy and also (as presented in the *Introduction*) to face off with the great variability of paintings with very dissimilar style but belonging to the same author. In this way, we increase the probability of a batch to be "complete", thus being representative of this variability.
The results obtained are the following:

StandardCNN				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
15	0.5730	0.786	1,3656	0.6542



(a) Standard CNN Accuracy



(b) Standard CNN Loss

The result on test set is stunning, however we believe that it has been a lucky coincidence more than an actual prediction power, given the fact that the relative training and validation accuracy were respectively 0.68 and 0.57. Both the discrepancy between validation and test accuracy and the fact that 79% is actually higher than training accuracy prove that this value is just a happy coincidence, probably given a high similarity between the training samples correctly learned by the network and the test samples. The maximum validation accuracy value is equal to 57.3% and it is reached in just 8 epochs, then the network overfits very quickly. It can be caused by mainly 3 factors:

- Data available is insufficient, so the network loses the ability to generalize
- Lack of regularization techniques, such as Dropout or L1/L2 regularizations
- **The spatial extent of the feature map of the last layer is still quite large, so the fully connected layers have too many parameters operating in a quite shallow representation.**

3.1.2 Deeper network and dropout

The highlighted problem can be tackled exploiting a deeper architecture, in which we add more convolutional layers in order to reduce the spatial extent of the extracted features, together with the reduction of parameters and the further processing applied to the original image: in this way we expect the network to have higher capacity and to need more epochs to overfit, thus incrementing accuracy. To better process the feature vector, we also added a 32-neurons *Fully-Connected* layer, right before the *Softmax* classifier. In order to avoid overfitting that could be caused by those additional parameters we introduce *Dropout* regularization to the latter layer. The final resulting network is the one reported in the figure below4.

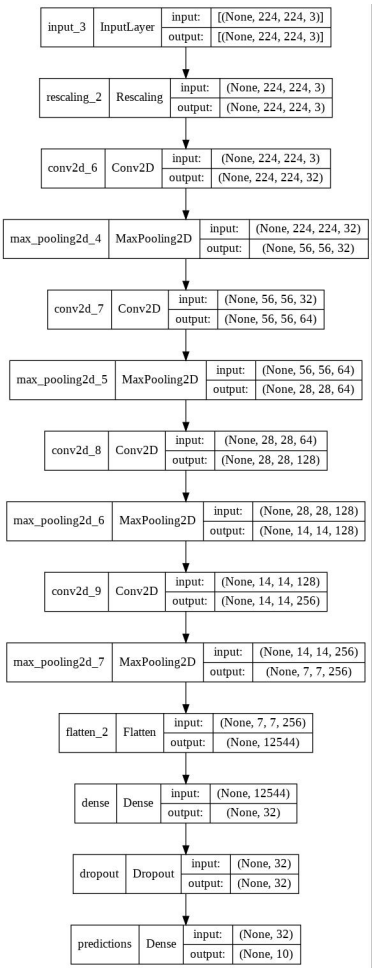
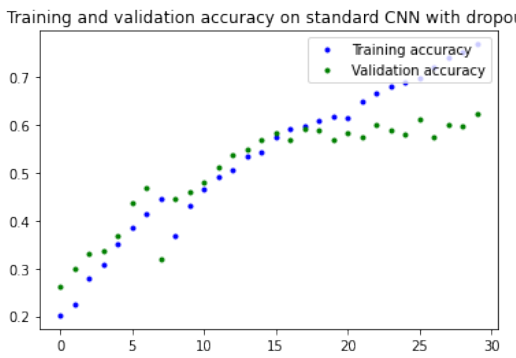


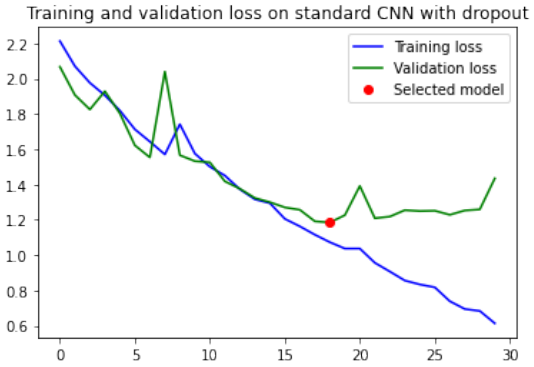
Figure 4: Deeper CNN with Dropout Architecture

Again, we train the network for at least 15 epochs. To optimize the training process, we actually tune dynamically the number of epochs and we save the history of the training itself. In this way we can interrupt and restart the learning phase, visualize the intermediate results and save in an external file the model that, during training, achieved best validation accuracy. Whenever possible, we train the network till the convergence of training loss, thus visualizing when and how fast the model started to overfit and trying to understand what can be the causes.
The results obtained are the following:

Deeper CNN with Dropout				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
45	0.590	0.6244	1.1996	1.1516



(a) Deeper CNN with Dropout accuracy



(b) Deeper CNN with Dropout loss

Analyzing these graphs, we realize that the validation accuracy is actually improved, and also test accuracy and loss are coherent with validation ones. The increased network capacity improved its ability to generalize, but we now have a problem, that is the model couldn't converge (not shown in the graphs above), starting to oscillate once reached 1.2 of training loss. We can argue that is due a too high learning rate value, but we actually use ADAM as optimizer which is able to dynamically tune it to proceed with learning. Thus the cause could be the underfitting: since we deepened the network, we are providing the few dense neurons with a feature vectors whose variability is now much bigger than before as result of the further processing. The phenomenon is exacerbated by the Dropout, that reduces Fully-Connected capacity by 20% at each epoch. If we could train further, we might have found a model that also improves validation accuracy.

3.1.3 Bigger Network and Early Stopping

If the problem is the overfitting, we can try to go further and increase network dimensions at FC layers. With this experiment we apply that, expecting to see our network to converge. If our hypothesis is correct, we expect the accuracy value to be more or less the same or at least to increase, while if it is not, we expect the validation loss to diverge quickly. The network we are going to test is the following6:

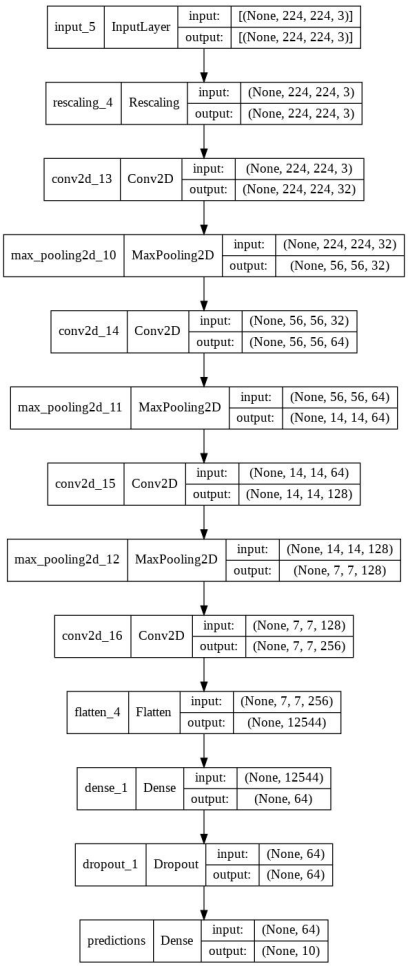
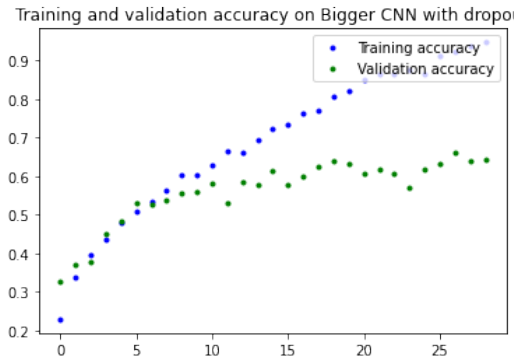


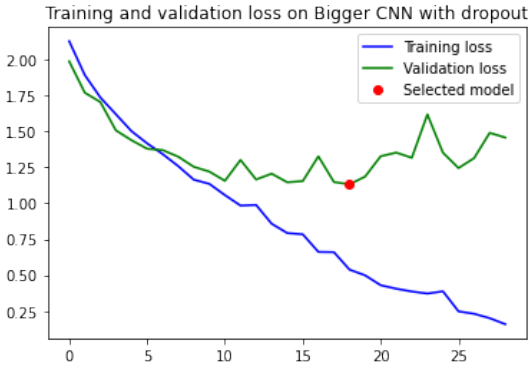
Figure 6: Customized architecture with bigger FC layers

This time we introduced the *Early Stopping* to the training process: Early Stopping is a mechanism offered by Keras among the possible callbacks, that consist on halting the training process if some condition is fired. In our case we monitor the validation loss (not the accuracy, the primer is more robust since it takes into account also class weights and how much a network is confident on its classification, while the latter only considers the number of correctly classified samples), whenever the model does not improve it for a certain number of epochs (patience value), training process is stopped. Thanks to the previously explained mechanism, we are still able to resume it manually if we believe that the validation values may actually improve.
The outcome of the learning is the following:

Bigger CNN				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
30	0.638	0.6443	1.1321	1.1742



(a) Bigger network accuracy



(b) Bigger network loss

Accuracy and loss are actually improved, and the convergence problem is now solved. Increasing capacity but equipping the network with the Dropout regularization allowed it to increase its learning power, but avoiding neurons to overspecialize themselves, and this had a beneficial effect on the architecture.

3.1.4 Batch Normalization

The previous model showed quite good results, however we wonder if we can improve it. For this purpose we try to use the *Batch Normalization*. Batch Normalization is a technique that is frequently used to make the training process more stable, usually gaining also better accuracies especially on very big and deep networks. Batch normalization is not formally proved, but the idea is quite simple: given a batch of input training samples, we subtract each by their mean and divide them by their standard deviation, thus obtaining a Gaussian Normal distribution of the inputs. To give an oversimplified view of the intuition, given a 2d linearly-separable dataset and a linear activation, to find the best dividing straight is much easier if all the points are near the origin rather than if they are far away: the found line will be much more robust. Batch Normalization tries to apply such transformation to all the samples of each batch, and it's usually exploited right before the non-linear activation (ReLU in our case). At test time, BN parameters are not computed on the test batch, but the average training ones are used. Taking as base the previous network (Bigger CNN), which is the one that performed best so far, and adding the Batch Normalization layers we obtain the following architecture:8

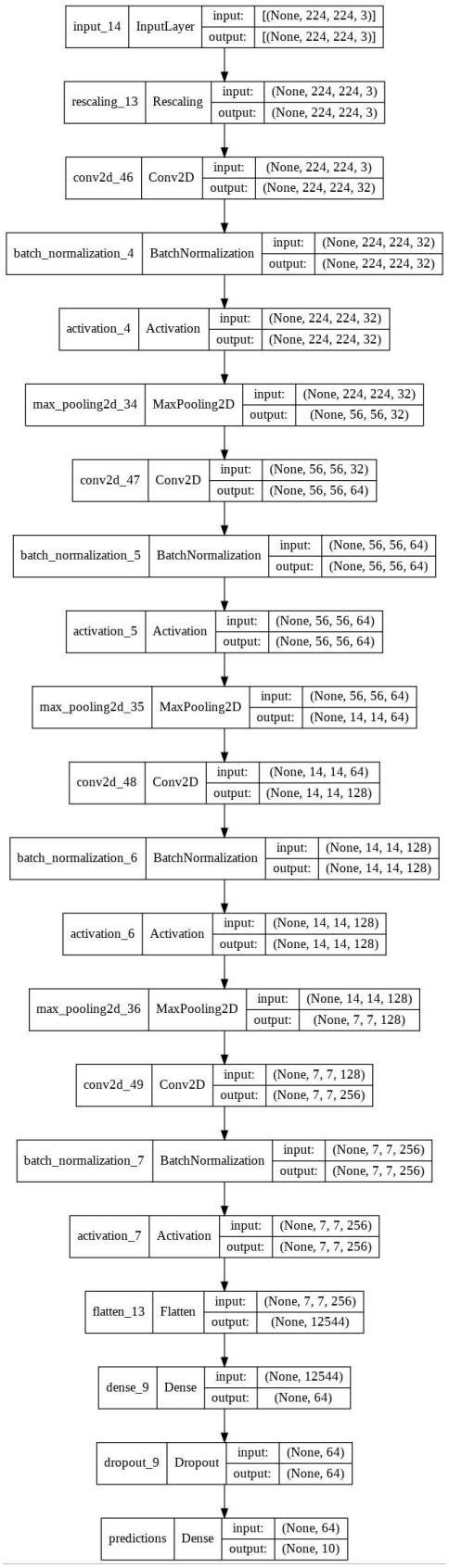
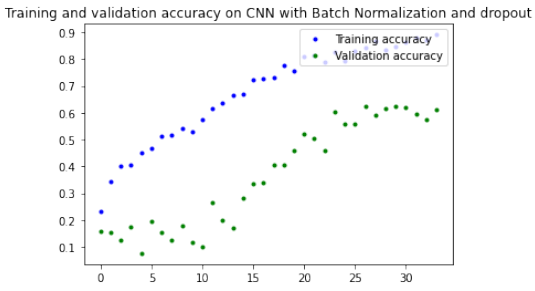


Figure 8: CNN architecture with batch normalization

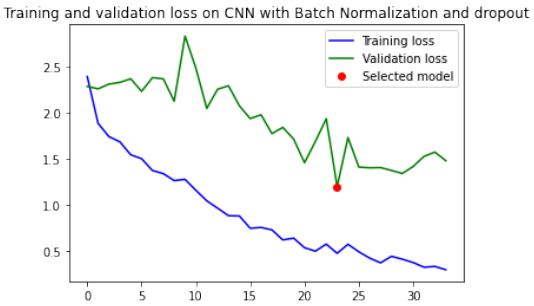
This figure could be a little bit misleading because, once separated activations from Convolutional Layers, the network seems to be deeper. Actually it is not.

The results of the training process are:

CNN with Batch Normalization				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
35	0.604	0.6144	1.1978	1.2942



(a) Bigger CNN with Batch Normalization accuracy



(b) Bigger CNN with Batch Normalization loss

Batch Normalization is very useful for big networks, however it showed no improvement in our experiment, adding noise to the learning process. This is also due to the limited number of training examples: mean and standard deviation are not very representative of the whole application domain, so they could differ very much from the ones of validation or test set

3.1.5 Data Augmentation

Last experiment conducted on standard CNNs exploits Data Augmentation. Data Augmentation is a techinique that consists on artificially expanding labeled training dataset by applying affine transformation or elastic deformations[12]. The idea is to try to increase accuracy by reducing more the overfitting, and it could be particularly suitable in our case where we have a very small dataset with several classes. In the following, we apply data augmentation to the model that performed best so far (paragraph 3.1.3)10.

[12]PEREZ, Luis; WANG, Jason. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

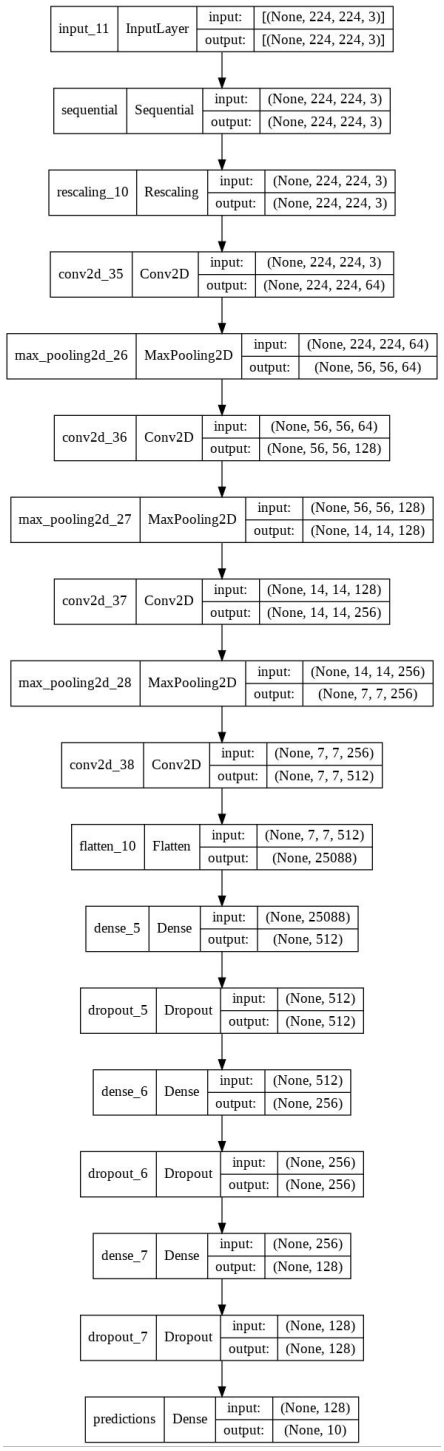
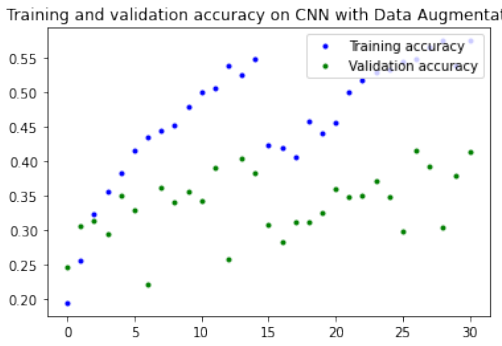


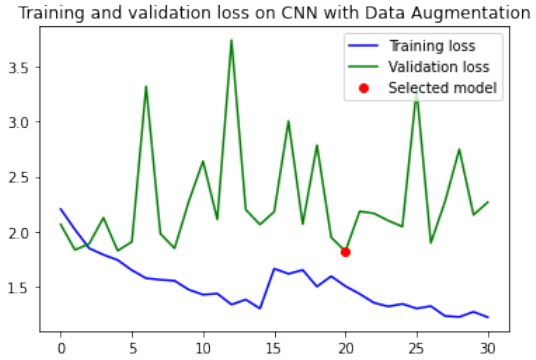
Figure 10: CNN architecture with data augmentation

The results of the training process are:

CNN with Data Augmentation				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
31	0.360	0.4055	1.8218	1.7557



(a) Bigger CNN with Data Augmentation accuracy



(b) Bigger CNN with Data Augmentatio loss

The training process is very very noisy and the model severely underfits. The noise can be explained by the high randomness amount that has been inserted in the learning process, both due to droupout and the random transformations of data augmentation itself. The underfitting was actually predictable: data augmentation is used mainly to contrast overfitting, but the 3.1.3 model actually overfitted very late. The only solution would be the training of a larger, deeper network, thus increasing parameters that can be trained and reducing overfitting: we tried such approach and it was actually promising, however Colab limitations halted the training and we concluded that it was impossible to train such a big model from scratch.

3.2 Aggressive Downsampling

Aggressive Downsampling consists on designing a Convolutional neural network such that, first of all, it shrinks very rapidly the spatial extent of the input image: in this way the number of initial parameters are drastically reduced, allowing for deeper networks and more FC layers. The idea is that in the input image (especialy for high-resolution ones), contiguous pixels have similar values, so even such an aggressive initial downsampling does not destroy much information. GoogleNet and AlexNet are examples of networks exploiting Aggressive Downsampling. In the following section, we define and train some network with aggressive downsampling, comparing performance with the previous ones.

3.2.1 Standard CNN with aggressive downsampling

In this paragraph we analyze a model very similar to the 3.1.1 one, but it exploits aggressive downsampling. Thanks to that, the network is deeper but also has less parameters than the standard model, for sure reducing overtraining. The architecture is the following12:

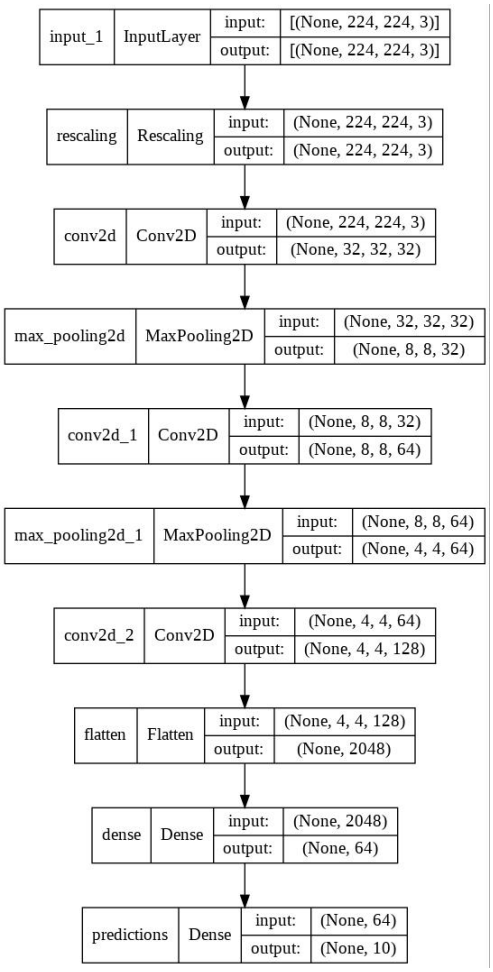
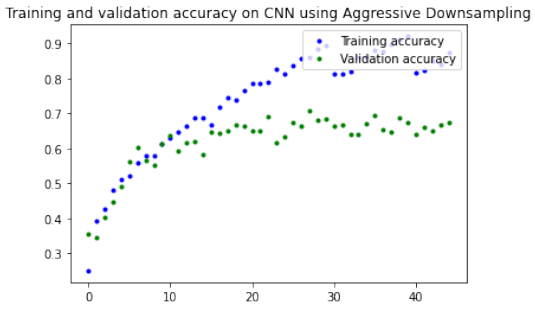


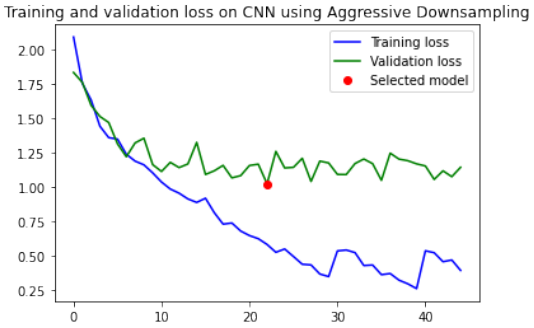
Figure 12: standard CNN architecture with aggressive downsampling

As we can observe, the number of parameters is reduced from more than 1 million to just 225 thousands, but despite this it has a fully connected layer more. The training process took much less computational time and the results are:

CNN with Aggressive Downsampling				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
45	0.691	0.6194	1.0244	1.0796



(a) CNN with aggressive downsampling accuracy



(b) CNN with aggressive downsampling loss

The choice of aggressive downsampling has sensational effects on the classification power of the network: with very few GPU memory consumed, this model outperforms the 3.1.1 one,

especially on validation accuracy and loss. Since the test results of the standard CNN model were not reliable, we cannot compare them, but with respect to the other models presented in the section 3.1, this network has similar accuracy and better loss. The only problem came up around iteration 30: the network cannot converge but starts oscillating around the loss value of 0.5, also training accuracy oscillates while validation loss is stuck around 1.2. The first hypothesis could be that the learning rate is too large in that training phase, but we used as optimizer Adam which computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients, and its convergence is proved[13], so the only possibility is that the network has too few parameters to learn all the train samples. In the following experiment we try to increase the network size to face such problem.

3.2.2 Bigger CNN with aggressive downsampling

As anticipated in the previous paragraph, in this experiment we test a bigger Convolutional NN that exploits aggressive downsampling, trying to solve the problem of missing convergence. Note that, even if it will be solved, it might not involve that validation and test accuracy will increase. The architecture is the following14:

[13]KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

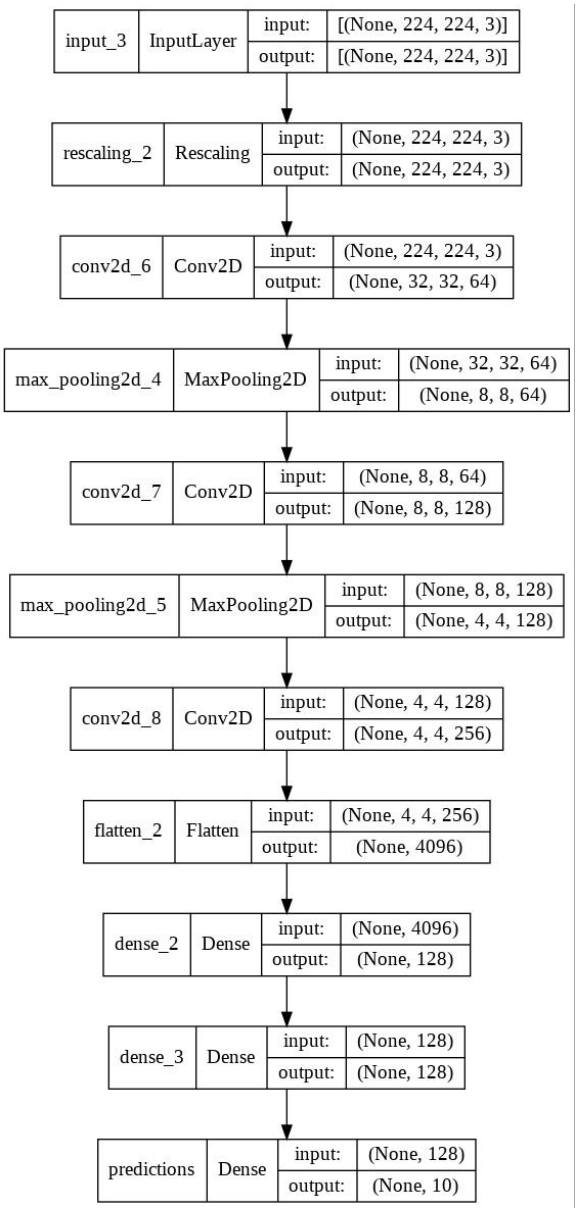
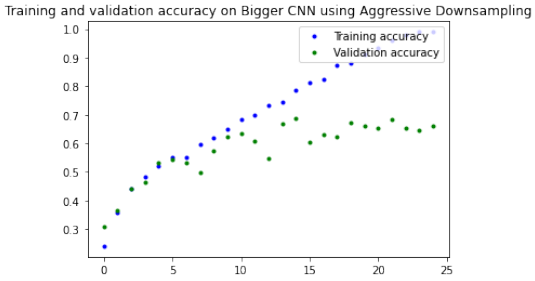


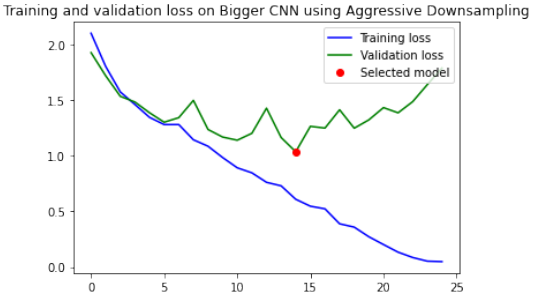
Figure 14: Bigger CNN architecture with aggressive downsampling

Number of parameters is much higher, almost 1 million. The training results are:

Bigger CNN with Aggressive Downsampling				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
25	0.688	0.7910	1.0349	0.6340



(a) Bigger CNN with aggressive downsampling accuracy



(b) Bigger CNN with aggressive downsampling loss

Increasing parameters actually helped the network to reach convergence, and it showed comparable results on the validation test and great results on test set. However, as discussed in 3.1.1, it is very likely that with this random seed training and test set are quite more similar than training and validation ones, so the test metrics could be inflated by this fact.

3.2.3 Bigger CNN with aggressive downsampling and data augmentation

Despite the previous results, the graphs are still characteristics of overfitting: the best validation values where reached quite soon, and then they deteriorate quickly while the training loss continues to decrease. In this experiment we will try to fight overfitting introducing again data augmentation16:

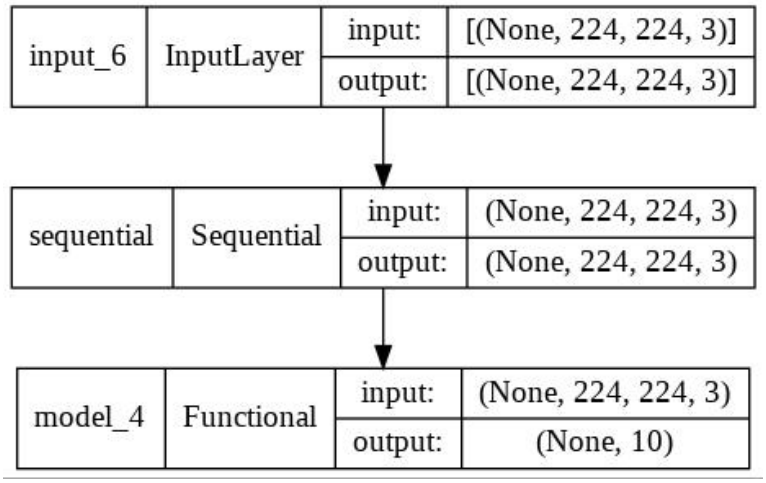
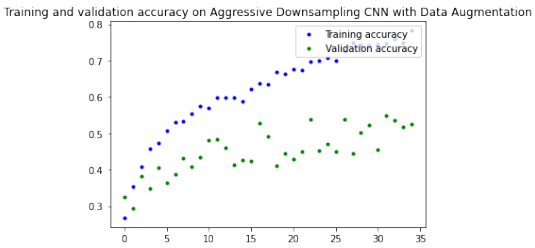


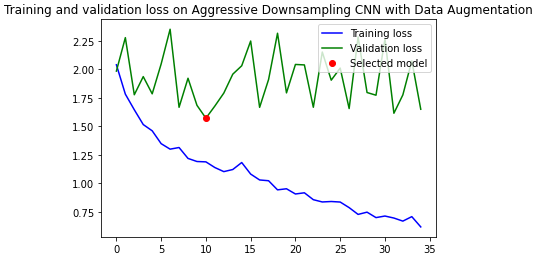
Figure 16: Bigger CNN architecture with aggressive downsampling and data augmentation

The central hidden model is the 3.2.2 one. Results of training are:

Bigger CNN architecture with aggressive downsampling and data augmentation				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
34	0.4803	0.5199	1.5720	1.4629



(a) Bigger CNN with aggressive downsampling and data augmentation accuracy



(b) Bigger CNN with aggressive downsampling and data augmentation loss

Also this time data augmentation did not help, the training loss convergence is smoother and slower but validation curves are very noisy and fang-shaped, showing very bad results and in some iterations almost random predictions. We should increase the network and use some regularization technique like L1 or L2 regularization, but again we cannot train from scratch a network bigger than this one in a reasonable time due to Colab limitations.

3.3 Inception Blocks

Inception modules are special convolutional blocks in which we do not define just one layer, but we concatenate more layers with different kernel sizes, together with a max pooling. The idea is that choosing which specific filter to use at each convolutional layer is an hyper-parameter, so it can be source of losses of performance. On the contrary, horizontally concatenating such layers we let the network "learn and decide" which one to use in order to minimize its training loss. Google started exploiting inception modules from GoogleNet(2014), and a finer version of them is present in the modern Inceptionv3 and Inceptionv4. In this section, we try to design and train a Convolutional Neural network equipped with Inception Layers, and we are going to measure its accuracy.

3.3.1 Standard CNN with Inception Layers

In this paragraph we build a model like the 3.1.1 one, but that also exploits custom-defined inception modules. At every module, the network can decide if to use a 3x3 or 5x5 convolution, or a 3x3 max-pooling with stride 1. The architecture is the following:

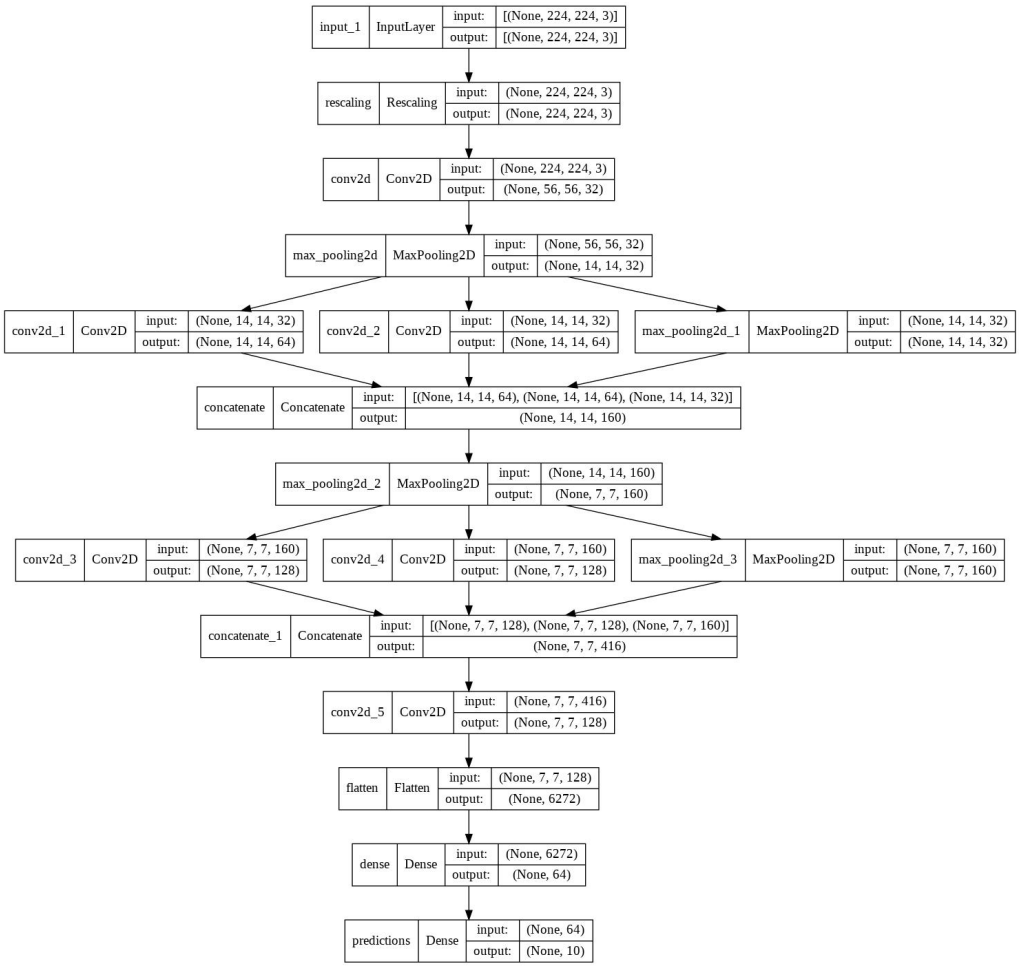
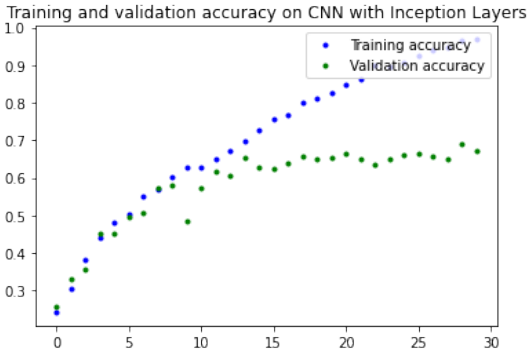
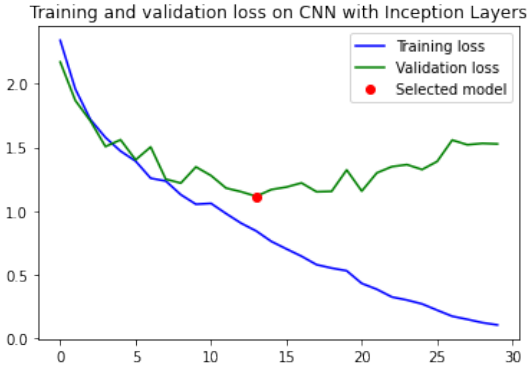


Figure 18: CNN architecture with Inception Layers

CNN with Inception Layers				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
30	0.652	0.6915	1.1157	0.8206



(a) CNN with inception layers accuracy



(b) CNN with inception layers loss

The network performed really well, showing quite good accuracy on validation set and great results on test set, both comparable with the ones reached by the 3.2.1 but overcoming the missed convergency problem.

3.3.2 Larger CNN with Inception Layers

In this experiment we try to increase the number of hidden neurons in the fully connected layers and we measure the resulting accuracy. This approach is usually exploited when there is underfitting: in our case there is not. However to increment the size of the dense layer turned out to be very effective in the previous experiments, so we will try to do that, but of course mitigating with dropout to avoid overfitting:

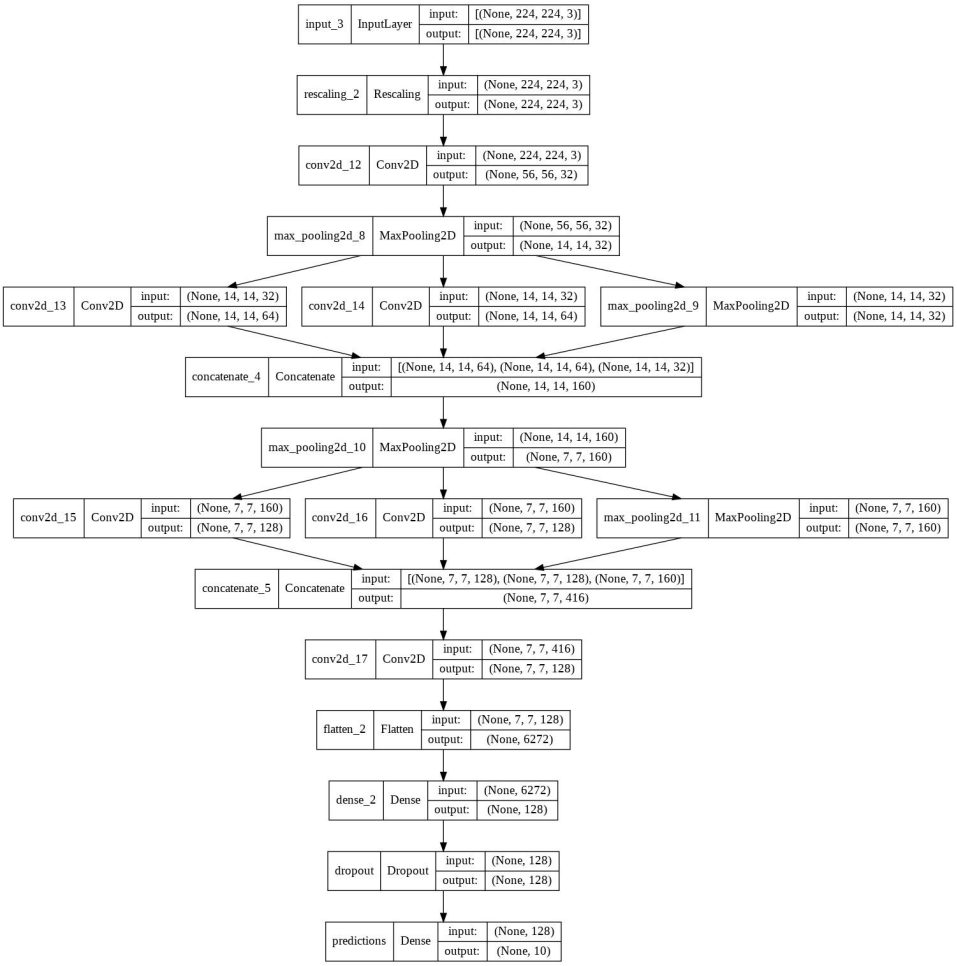
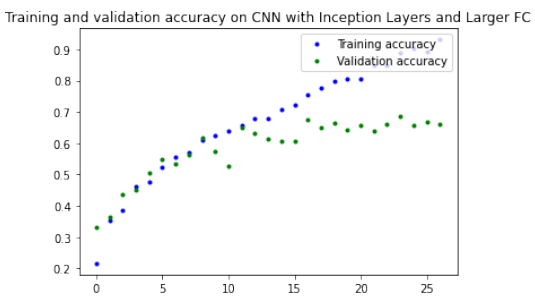


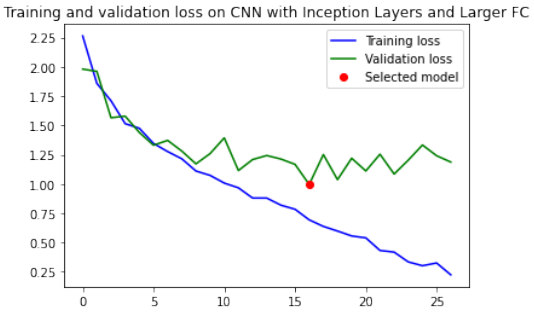
Figure 20: Bigger CNN architecture with Inception Layers

Number of parameters is pretty high, more than 2 millions. The training outcome is the following:

Bigger CNN with Inception Layers				
Epoch stopped	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
25	0.677	0.6219	0.9961	1.1133



(a) Bigger CNN with Inception Layers accuracy



(b) Bigger CNN with inception layers loss

Performances are more or less the same w.r.t. the previous model. Accuracy on validation set is slightly worse but loss is lower, which individuates more "confident", thus robust, predictions on this set. On test set results are not so good as the previous experiment.

3.4 Hyper-parameters optimization

Once we found the results of several experiments, we would like to run a hyper-parameter optimization algorithm on the best models found so far, in order to try to tune them to reach even better accuracies in a more stable manner. As we know, there are several algorithms with different complexities and efficiencies, among them we can list Grid Search, Random Search, Bayesian Model and Hyperband. The latter has been chosen to perform our hyper-parameters optimization: hyperband is an algorithm that relies on a principled early-stopping strategy to allocate resources, allowing it to evaluate orders-of-magnitude more configurations than black-box procedures like Bayesian optimization methods.[14] Hyperband is based on the Successive Halving algorithm, which starts from a budget B (e.g. total training epochs or training time ore GPU memory) and a value n of different configurations and at each iteration it discards the worst half. In successive halving, a difficult task is to choose B but more important to choose n: the trade-off is between n (that will act as exploration value) and B/n (that will act as exploitation value). Hyperbands overcomes the problem allocating several combinations of n and B/n, keeping B constant at each call of Successive Halving. The input parameters to the algorithm were:

- Models: 3.2.2 and 3.3.1 that were 2 of the most performant ones we trained from scratch
- R=20. Because those 2 models needed less than 20 epochs to be trained, this can be a good upper bound for training
- η=3 as default
- $s_{max} = \lfloor \log_\eta R \rfloor$=2
- $B = (s_{max} + 1)R$=80

With this configuration we run Hyperband on the 2 models, aiming at optimizing the following hyper-parameters:

- *n_neurons*: Number of neurons in the FC layer, right after the Flatten and before the classification, chosen as integer value between 0 and 256 with a step of 16
- *dropout_rate*: Dropout rate in this layer, as float value between 0 and 0.6

[14] LI, Lisha, et al. Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 2017, 18.1: 6765-6816.

- Activation function in the convolutional layers. Till now we always considered only ReLU (Rectified Linear Unit), now we introduce the possibility to choose also from ELU (Explonential Linear Unit) or GELU (Gaussian Error Linear Unit)
- lr: Initial Learning rate for the optimizer. Till now we always used default one for ADAM (1e-3), now the tuner will choose a float value between 5e-5 and 5e-3

The actual Hyperband optimization we used is the Keras Tuner, which provides automatic APIs to run Hyperband and other algorithms in avery fast and easy way, however consuming much GPU memory and computational time to find the best model. After running the tuner, we saved best models into 2 files and we visualized the best combinations of hyper-parameters, which are:

- First model: $n\_neurons$=208, $dropout\_rate$=0.03 activation=ReLU, lr=0.0012
- Second model: $n\_neurons$=176, $dropout\_rate$=0.09 activation=ReLU, lr=0.0004

As we notice, the optimal choice is to adopt a large number of hidden neurons (between 150 and 225), low or no dropout rate, ReLU as activation function and learning rate that is near the default one (1e-3). The models' performance is the following:

Performance comparison				
Model	**Validation Accuracy**	**Test Accuracy**	**Validation Loss**	**Test Loss**
Custom Aggressive Downsampling(3.2.2)	0.688	0.7910	1.0349	0.6340
Optimized Aggressive Downsampling	0.6130	0.6020	1.0670	1.2014
Custom CNN with Inception Layers(3.3.1)	0.652	0.6915	1.1157	0.8206
Optimized CNN with Inception layers	0.664	0.6990	0.9695	0.9797

Comparing the performance of the models found by the tuner w.r.t their original counterparts, we actually observe few differences: the first optimized model is slightly less performant than the custom one, while the optimized CNN with Inception Layers achieved slightly better results, but on test set it didn't overcome the loss value of the other model.

3.5 Visualization techniques

In this chapter we will adopt some technique to visually analyze what the network focuses on, in order to understand if our CNN is robust and is looking to the correct features to recognize the author of each painting. Taking again the 3.2.2 and 3.3.1 models, we will evaluate:

- The heatmap of class activations, that is starting from the activation level of a certain layer of our network with respect to a specific class, we apply a weighted gradient to highlight in the input image what portions of it were considered the most in order to classify the image with that label[15]
- A partial exclusion mask, that is we try to occlude different portions and positions of the input image, to understand if the network is still able to classify it correctly

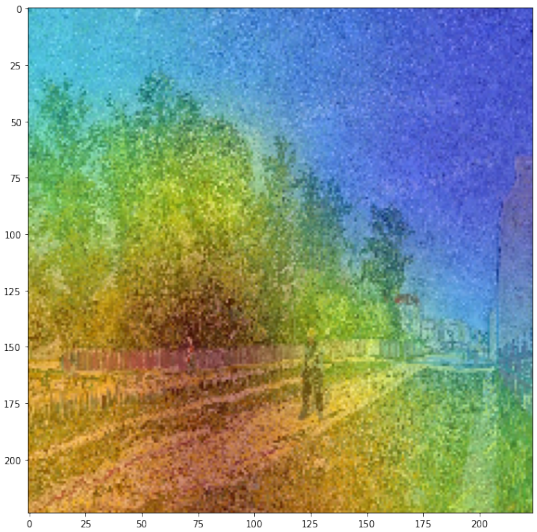
[15]SELVARAJU, Ramprasaath R., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618-626.

3.5.1 Class activations heatmap

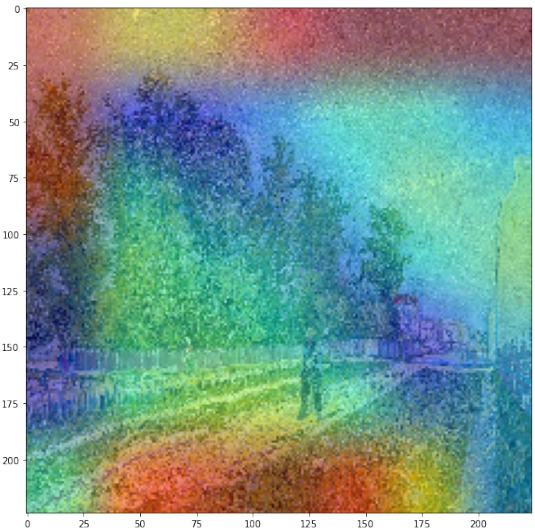
The idea is, given an input of the correct class, to map the gradients of the top-predicted class w.r.t. the activations of the convolutional layer we are analyzing, and then superimpose this heatmap to the input in order to understand which portions were mostly important for the final decision. We took as input image, a test sample of Vincent Van Gogh, which had been correctly classified by the two considered network at testing phase. The image is the following: 22

Figure 22: Test image

Once selected our models, we replace the last layers with new ones in order to observe more clearly what the back-prop gradients are, then after submitting the test image we record the locations of the activation map of the last layer that were "excited" the most. In this way we obtained our heatmap, which can be superimposed to the test image to visualize which locations were considered as more interesting for the monitored convolutional layer. Applying it for the last convolution of both the models we obtained:



(a) Class activation map of the last layer of 3.2.2



(b) Class activation map of the last layer of 3.3.1

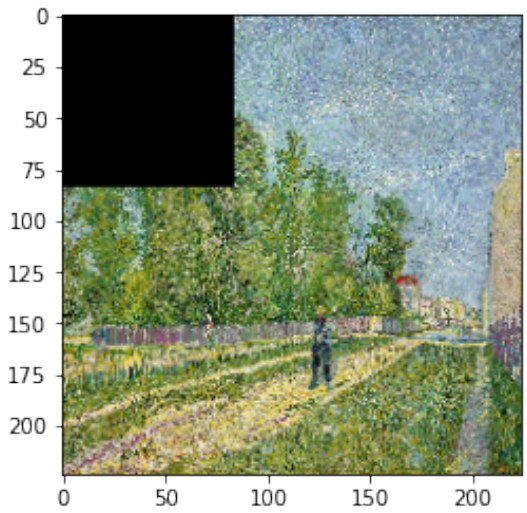
As we can see, the first model tends to look more "at the center" of the image, maybe seeking for author's typical characters/objects, or maybe recognizing the style or the typical color pattern of the objects. Remember also that this network aggressively downsamples the input image, thus the remaining information at the center of the image will be very various and precious for this network. On the other hand, the second model seems to "look at the edges", activating the most this heatmap in correspondence of the sky and the road: this network maybe is able to recognize the style of the author, and classifies images, distinguishing how he/she represents common elements and maybe which colors and patterns he/she uses to characterize the background.

3.5.2 Occlusion masks

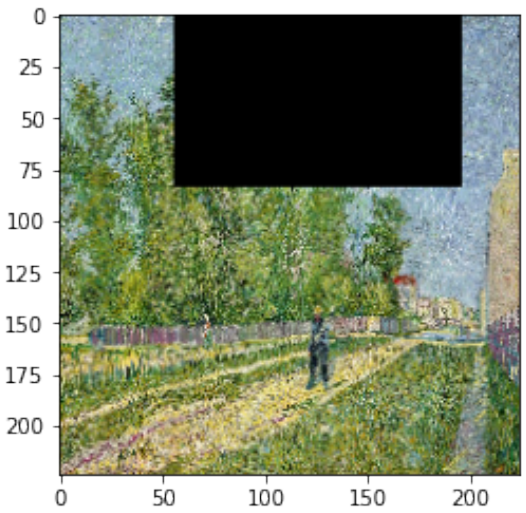
The idea of occlusion masks is very simple and intuitive: in order to see what portions of the test image are considered to take the decision of the correct class, we "hide" with a black box a part of it and we submit the resulting image to the classifier, annotating the response. Iterating for different positions and different sizes, we can estimate the robustness of the network and the portions of the input image that it considers as "fundamental" to take the correct decision. We iterated the algorithm with 3 occlusion mask sizes: 10, 28 and 35 pixels; we applied it every size*3 pixels in width and height, and the results are the following:

- With size=10, 9 masks are produced, both the networks are still able to correctly predict *Vincent Van Gogh* independently from the covered zone
- With size=35, 4 masks are produced, both the networks miss all prediction, outputting *Rembrandt*
- With size=28, 4 masks are produced, while the first network is still always able to classify the image as *Vincent Van Gogh*, the second network mispredicts twice.

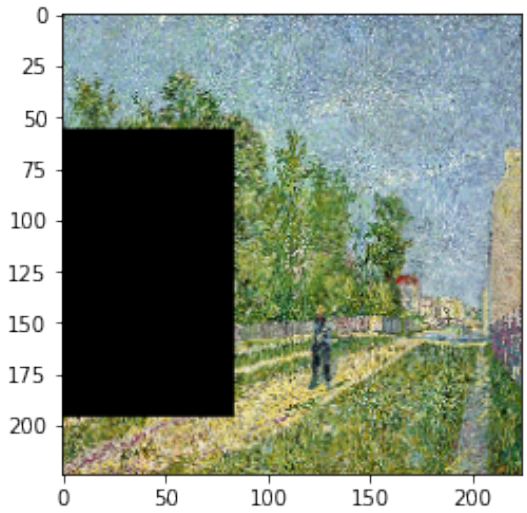
Applied occlusion masks and relative predictions with size=28 are:



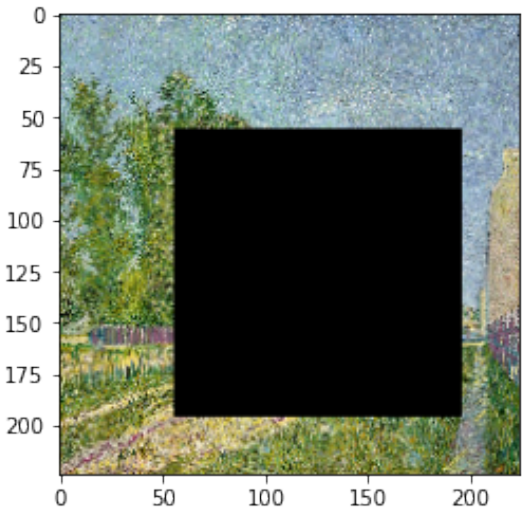
(a) Vincent Van Gogh – Vincent Van Gogh



(b) Vincent Van Gogh – Alfred Sysley



(c) Vincent Van Gogh – Vincent Van Gogh



(d) Vincent Van Gogh – Rembrandt

From these results we can infer very interesting knowledge. First of all, the 3.2.2 network proved to be very robust and to be able to correctly classifying the image even in the most adverse situation, i.e. when the center is completely covered. This is absolutely not trivial if we consider that, as shown in the 3.5.1 heatmap, it tends to look exactly at the center to classify the painting. Even more interesting what concerns the second model: as we observed in the 3.5.1 heatmap, it tends to focus on background characteristics, especially sky or ground style, and in fact when the sky or the center of the image (including the road) are occluded, the classification is wrong. In spite of it, we can say that it is "not so wrong": the first misprediction (b) outputted *Alfred Sisley*, which was an impressionist (*Van Gogh* was post-impressionist) whose style differs a bit from the Dutch painter, but the biggest differences are actually in the sky which is usually calm and regular for *Sisley*, dramatic for *Van Gogh*; without this information the error is comprehensible. A similar consideration holds for the second misprediction (d): we can argue that from the center of the image the network considers very much the color, and in fact when we substituted the center with a black box, it classified the test image as *Rembrandt*, who is known for very dark paintings, with a lot of black in the center. Anyway, from those results, the model 3.2.2 showed to be more robust and precise with respect to the 3.3.1 one.

4 — Pre-Trained Models

This section describes the results obtained using different pre-trained architecture and strategies[16]. The pre-trained networks here tested are:

- VGG16
- ResNet50V2
- ResNet101V2
- InceptionV3

4.1 VGG16

VGG1625 is a convolutional neural network model proposed by Simonyan et al., with several 3x3 convolutional layers in cascade occasionally interleaved with 2x2 max-pooling layers forming the so called *blocks*. Developed for the ILSVRC2014 challenge, it was able to achieve a top-5 accuracy of 92.7 on ImageNet.

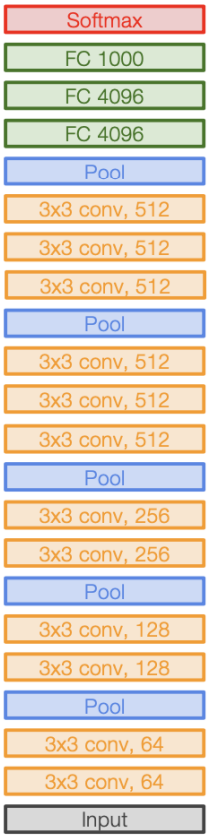


Figure 25: VGG16 Architecture

4.1.1 Test 1: Classical VGG16 (Feature Extraction)

The original VGG16 comes with a couple of 4096 FC layers followed by 1000 softmax neurons, which is alright for ImageNet but definitely oversized for our purpose. Hence, the convolutional

[16]The data augmentation strategy is always used since we have very little data

base is left as it is, and the fully-connected block is replaced by the a shrunk version with only 256 neurons per layer, followed by our prediction layer made of 11 neurons26.

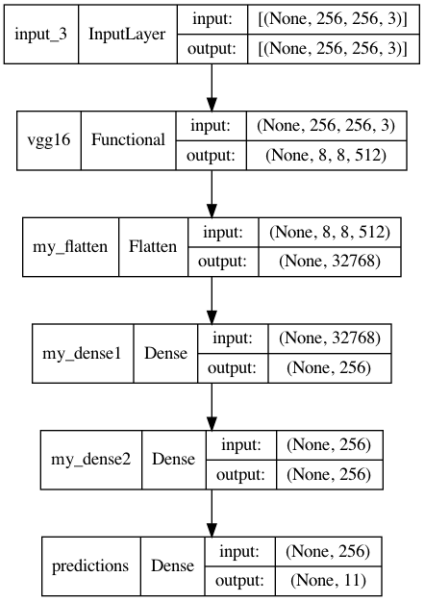
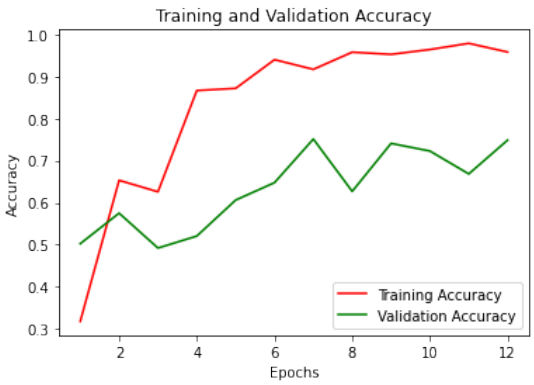


Figure 26: Our Feature Extraction Network

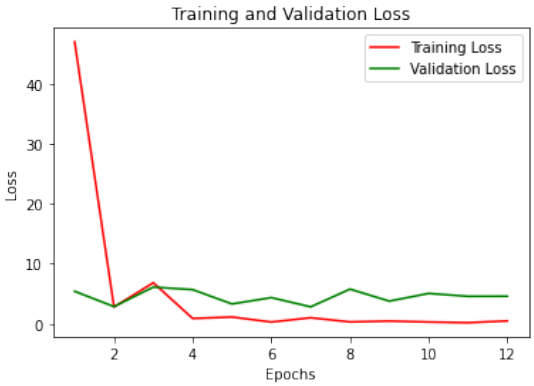
The result obtained, using RMSprop as optimizer, are:

Feature Extraction				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
12	0.7409	0.6391	5.1959	5.7002

The network begin to overfit very fast, hence some regularization methods are needed.



(a) Simple VGG16 Feature Extraction Accuracy



(b) Simple VGG16 Feature Extraction Loss

4.1.2 Test 2: Adding Dropout to Test 1

We have two possible positions to use the dropout layer in our network and they are after each 256-dense layer, but we decided to use just one layer at the end of the second 256-Dense layer (*my_dense1*) as shown in Figure28. We didn't use a dropout layer between the two 256-dense layers, since this type of architecture led to worst performance, this mainly because we would have less units to fully train our topic-specific network.

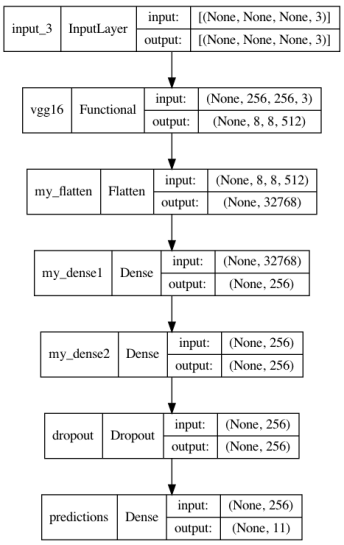
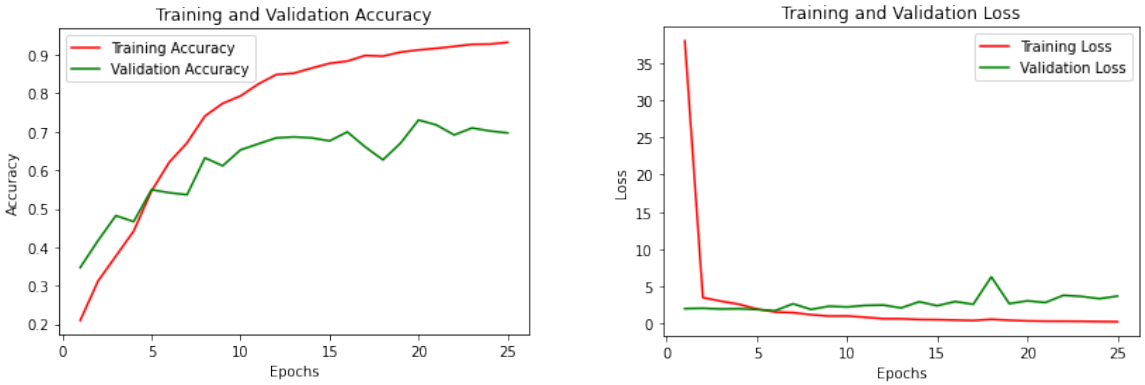


Figure 28: Our Feature Extraction Network + Dropout

The result obtained, using RMSprop as optimizer, are:

Feature Extraction w/ dropout				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
25	0.7306	0.7195	3.0616	3.0035



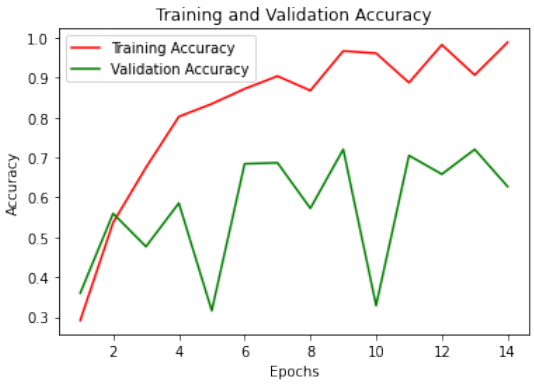
(a) Test 2 Accuracy **(b)** Test 2 Loss

As expected, dropout mitigated the magnitude of overfitting, however our network perform slightly worst on accuracy (now the validation one is 0.73, before was 0.74) than without the dropout layer, but shown quite good improvement in the loss value and the test accuracy.

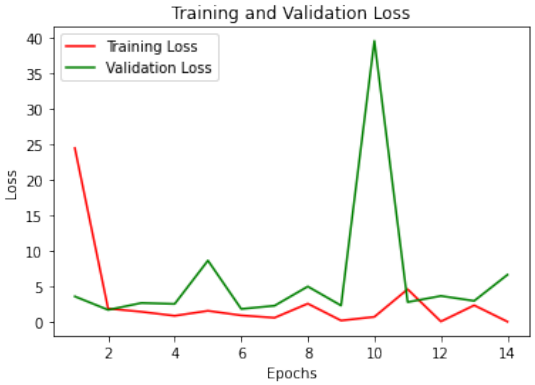
4.1.3 Test 3: Finetuning One Convolutional Layer

Using the model defined in test 4.1.1, the 3rd Conv2D layer in the 5th block is un-fronzen and the network is trained. The result obtained using RMSprop as optimizer are the following:

Finetuning one convolutional layer				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
28	0.7202	0.6253	2.3687	8.1807



(a) Test 3 Accuracy



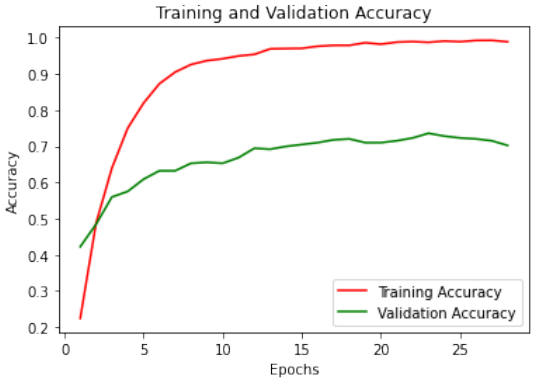
(b) Test 3 Loss

Again the performance are deterioted especially on test set, moreover looking at the graphs it can be seen that not only our network overfitted very fast, but it forms also few fang-shaped changes in direction.

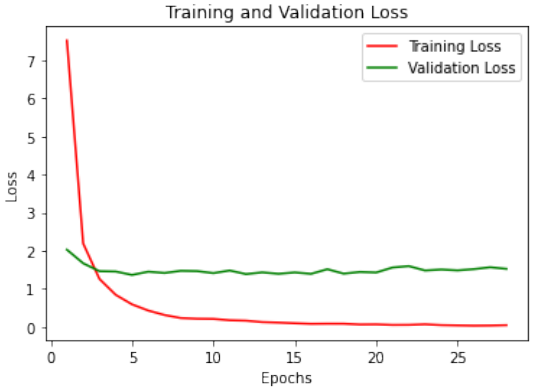
4.1.4 Test 4: test 3 with dropout and different optimizer

To overcome the previous problems, the overfitting and the strange shape behavior, in this test we opt to use **Adam** as an optimizer, changing its default learning rate (i.e., 0.001) to 0.0001 in order to slowly learn and hoping to have a smoother accuracy and loss functions.

Finetuning one conv layer w/ dropout and Adam				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
28	0.7358	0.7218	0.9871	1.1792



(a) Test 4 Accuracy



(b) Test 4 Loss

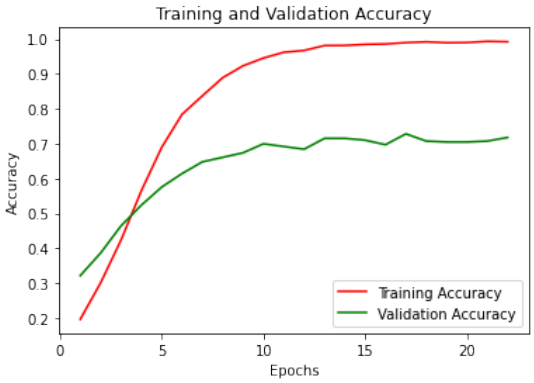
Adding dropout and a different optimizer with a little learning rate, we finally obtained what we were aiming, with results slightly better than ones on test 4.1.1, but now we our using a more complex network.

However we can still do something, the training accuracy increases very rapidly even if dropout is applied. To decrease this effect in test 6 we use weight regularization techniques.

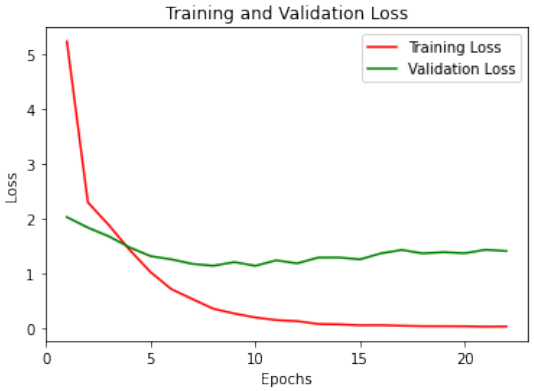
4.1.5 Test 5: Finetuning Two Convolutional Layers

This test is basically test 4, but finetuning the last two convolutional layers of VGG16.

Finetuning two convolutional layers				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
22	0.7280	0.7379	1.4253	1.2589



(a) Test 5 Accuracy



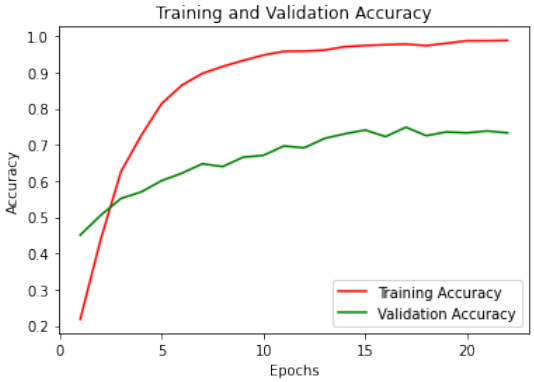
(b) Test 5 Loss

Following the approach used in the previous test we obtained also here more smoothed graphs, but the performance is little lower than before (especially on loss values). The problem here is that we are propagating to the second layer in block 5 gradients that are not really improvements of the ones set by the *imagenet* default configuration.

4.1.6 Test 6: Finetuning One Convolutional Layer and Weights Regularization

In this paragraph we exploited the conclusion mentioned in test 4, introducing here *L1_L2 weight regularization* on the one convolutional layer finetuned network.

Finetuning one conv layer and weights regularization				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
22	0.7487	0.7471	21.5590	9.8097



(a) Test 6 Accuracy

(b) Test 6 Loss

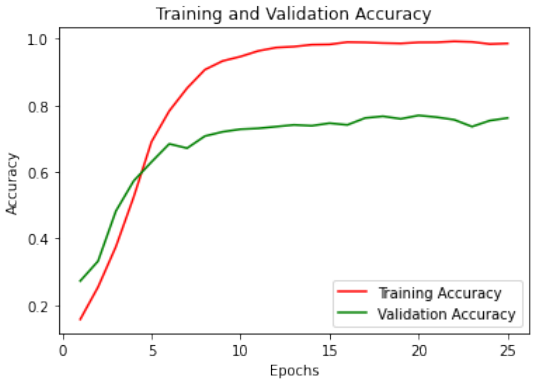
The shapes of the graphs are more or less the same of the two had in test 4, the only things that changes are the values of the losses which are here higher and more curvilinear due to the weight regularization approach used. Anyway, this heavy regularization helped us to surpass the result in previous tests , not by much but it is an improvement that, maybe, can be exploited finetuning more. The model could potentially been trained more, restoring

then best weights in case of no improvements, but due to Colab limitations it was impossible to go further. Following this lead, in the next paragraph we use weight regularization with two convolutional layers finetuned.

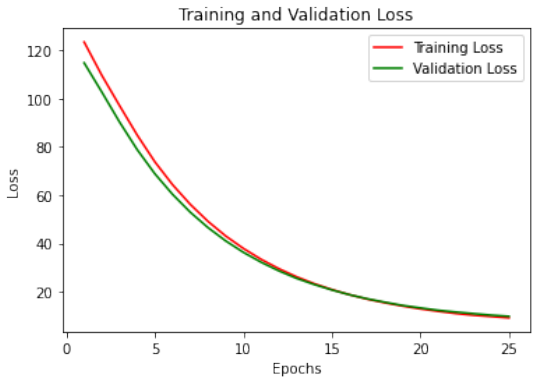
4.1.7 Test 7: Finetuning Two Convolutional Layers and Weights Regularization

Following the good result obtained in test 6, in this paragraph we add, to the network used in test 5, *L1_L2 weight regularization*.

Finetuning two conv layers and weights regularization				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
25	0.7694	0.7494	15.6742	9.8097



(a) Test 7 Accuracy



(b) Test 7 Loss

Even if the increase the flexibility of our network finetuning more layers, as hoped, we achieved a better accuracy which is also the best obtained so far.

4.1.8 Test 8: Genetic Algorithm for Hyper-parameters and Architecture Optimization

Following the elements presented in chapter 2.6 regarding the components of genetic algorithms, we here introduce the specific cases used here.

4.1.8.1 Genotype

In our specific case, our genes are bounded real-valued encoded and they represent three different parameters:

- *activation_function*: bounded between 0 and 1.999. The integer part stands for ReLU (0) and ELU (1);
- *optimizer*: bounded between 0 and 1.999. The integer part stands for rmsprop (0) and adam (1);
- *learning rate*: bounded between 0.001 and 0.1.

4.1.8.2 Population

Since for each individual we train a model and evaluate its performance, we decided to use only 5 individuals per generation to limit the computation.

4.1.8.3 Fitness Function

In our problem, the fitness function is the the function which calculate the maximum validation accuracy reached by the model. Our objective is maximizing the validation accuracy.

4.1.8.4 Selection Algorithm

The selection algorithm used in this case is *tournament selection*. In each round of the tournament selection method, two individuals are randomly picked from the population, and the one with the highest fitness score wins and gets selected. We decided to select only two individuals since our population is small and selecting more could cause an abuse in exploitation.

4.1.8.5 Crossover Algorithm

There are multiple algorithms applicable to real-value encoded individuals, we decided to use the *Simulated Binary Bounded Crossover*, which is a bounded version of the Simulated Binary Crossover (SBX)[17].

The idea behind the simulated binary crossover is to imitate the properties of the single-point crossover that is commonly used with binary-coded chromosomes. One of these properties is that the average of the parents' values is equal to that of the offsprings' values. When applying SBX, the two offspring are created from the two parents using the following formula:

$offspring_1 = \frac{1}{2}[(1+\beta)parent_1 + (1-\beta)parent_2]$

$offspring_2 = \frac{1}{2}[(1-\beta)parent_1 + (1+\beta)parent_2]$

Here, β is a random number referred to as the *spread factor*.
This formula has the following notable properties:

- The average of the two offspring is equal to that of the parents, regardless of the value of β.
- When the β value is 1, the offspring are duplicates of the parents.
- When the β value is smaller than 1, the offspring are closer to each other than the parents were.
- When the β value is larger than 1, the offspring are farther apart from each other than the parents were.

The probability to mate is set equal to 0.9[18].

4.1.8.6 Mutation Algorithm

As mutation algorithm we decided to use the *Polynomial Bounded* method, which is a bounded mutation operator that uses a polynomial function for the probability distribution. The probability to mutate is set equal to 0.5.

4.1.8.7 Elitism

We set the number of individuals for the elitism mechanism to 1 since our population is only made by 5 individuals. Anyway, to speed up even more the computation of the algorithm we introduced a system that saves the results of individuals already seen, in this way we do not need to recompute the accuracy again for those individuals.

[17] https://content.wolfram.com/uploads/sites/13/2018/02/09-2-2.pdf

[18] Value suggested by the book "Hands on Genetic Algorithms with Python" by Eyal Wirsansky, also the other values of probabilities are taken by the suggestions of the book

4.1.8.8 Results

Even if the code was uploaded on the directory, we was unable to run it till completion due to the limitations of Colab. In fact, even the strong limitation we made on the population, for each epoch the algorithm takes about 40 minutes (since all the first epochs are done at the same time, due to the parallelism introduced by *deap*), thus producing the final result (i.e., the result at the end of the last generation) after few tens of hours[19].

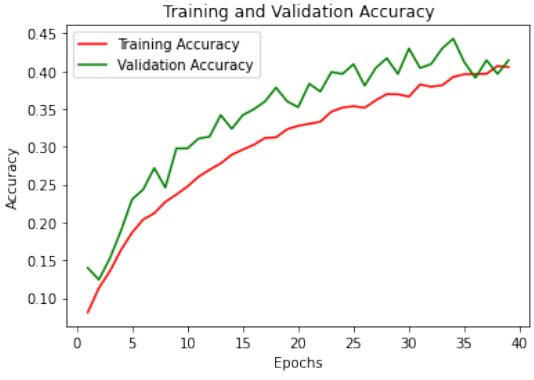
4.2 ResNet50V2

ResNet stands for Residual Network. It is an innovative neural network that was first introduced by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in their 2015 computer vision research paper titled "Deep Residual Learning for Image Recognition". This model was immensely successful, as can be ascertained from the fact that its ensemble won the top position at the ILSVRC 2015 classification competition with an error of only 3.57%. Additionally, it also came first in the ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in the ILSVRC & COCO competitions of 2015.

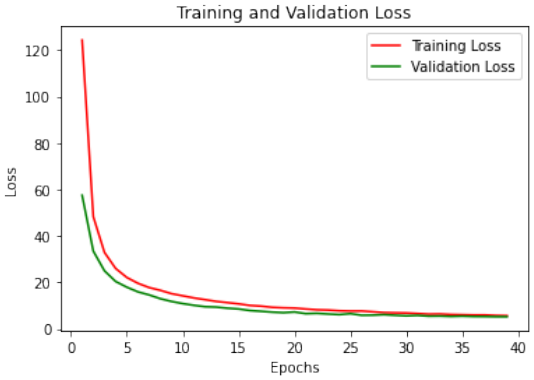
4.2.1 Test 1: Classical ResNet50V2 (Feature Extraction)

The original ResNet50 comes with a GlobalAveragePooling2D and a prediction layer soon after. In this test we used the same approach, resizing the prediction layer to the number of classes we have.
The results obtained after training are:

Feature Extraction				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
39	0.4430	0.3241	5.3733	6.8513



(a) Simple ResNet50V2 Feature Extraction Accuracy



(b) Simple ResNet50V2 Feature Extraction Loss

The network poorly performed compered with the results obtained using VGG16. This can be explained since the *imagenet dataset* is not specialized in distinguish paintings' artists, thus the frozen weights of our networks are very far to be the ones we need: the only prediction layer cannot learn properly those distinctions alone and as result the network underfits. To overcome this problem we can fine tune or add more layers after the *GlobalAveragePooling2D* layer.

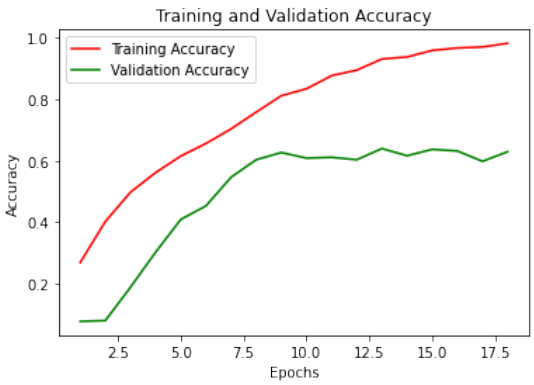
[19]Consider 40 minutes per epoch and an average of 18 epochs, we get that we need 720 minutes for training a generation, since we have 5 generations, the computation end after approximately 3600 minutes, which are 60 hours.

Another problem can be notice looking at the accuracy plot, the training accuracy is less than the validation accuracy. To understand this issue, it is important to consider the difference of the number of images in the training and validation set. The validation set is made of about few hundreds of pictures against the few thousands of the training set, the latter may involve, with greater probability, paintings being part of a different period of the artists (e.g., the example shown in Image1), which can lead to a worse accuracy.

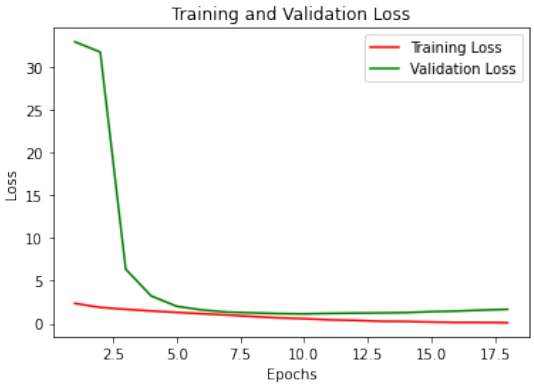
4.2.2 Test 2: Finetuning 1 block

ResNet50V2 is made of multiple big blocks (i.e., 5) so called *conv* in the model. These blocks are made of sub-blocks connected by each other by an add layer which connects the processed input (e.g., processed by Conv2D, Padding, Pooling, BatchNormalization) and the residual input[20]. We finetuned considering this under-blocks, hence in this paragraph we finetuned the *conv5_block3*.

Finetuning 1 block				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
18	0.6399	0.5793	1.2453	1.7916



(a) ResNet50V2 Test 2 Accuracy



(b) ResNet50V2 Test 2 Loss

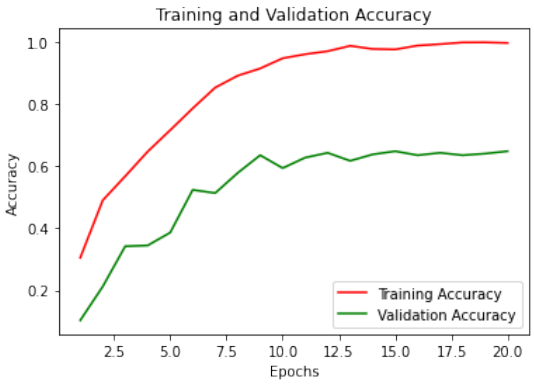
As expected, the network has definitely reached a more satisfactory result. Also the behavior of the curves is now is more reasonable.

4.2.3 Test 3: Finetuning 2 blocks

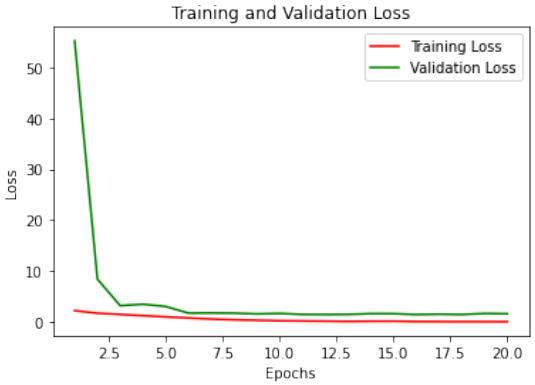
Since finetuning a block led to better results, the next step we did was to tuning also the previous sub-block, thus starting to tuning from *conv5_block2*.

Finetuning 2 blocks				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
20	0.6477	0.6092	1.5918	1.7734

[20]Sometimes the residual is downsampled using a max pooling layer, this is done in order to match the actual size of the feature map obtained at that level of the network.



(a) ResNet50V2 Test 3 Accuracy



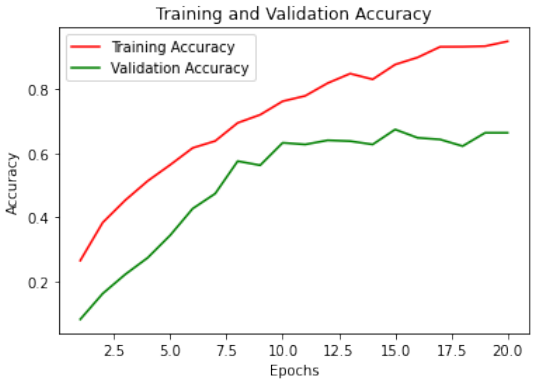
(b) ResNet50V2 Test 3 Loss

This results proves that our idea of finetuning was good since the network improved a little. Anyway, finetuning more can be risky since the computational power and time needed can be very high and becoming unbearable to the limits imposed by Colab. Thus, we tried to improve (without satisfaction, as the nexts to paragraphs describe) our network adding dense layers between our *GlobalAveragePooling2D* and our prediction layer, following the structure used in the VGG16 study.

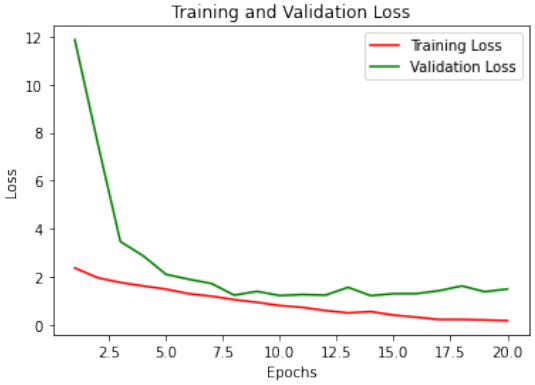
4.2.4 Test 4: Finetuning with One Block and Adding Two Dense layers

Here we took the same approach of test 3, but adding two dense layers followed by a dropout layer (as used in figure28).

Finetuning one block and dense layers				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
20	0.6736	0.6368	1.6655	1.7734



(a) ResNet50V2 Test 4 Accuracy



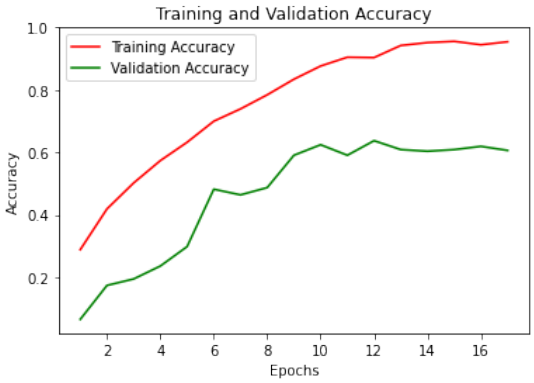
(b) ResNet50V2 Test 4 Loss

Enlarging in this way the network resulted in better performance. This can be because the dense layers before the softmax can better exploit the information of the *GlobalAverage-Pooling2D*, keeping the problem for more time to a higher dimensionally than the one used for prediction.

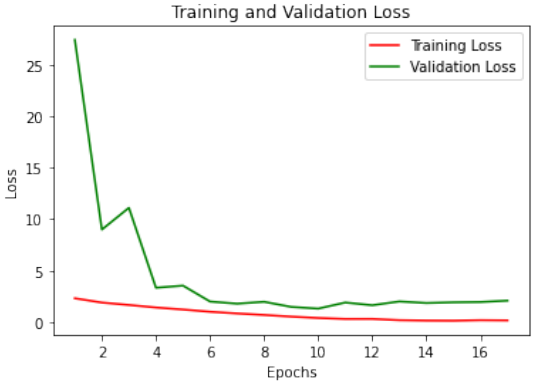
4.2.5 Test 5: Finetuning with Two Blocks and Adding Two Dense layers

In this paragraph we did the same thing done in test 4, but finetuning two blocks as in test 3.

Finetuning two blocks and dense layers				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
17	0.6373	0.5954	1.6404	2.1089



(a) ResNet50V2 Test 5 Accuracy



(b) ResNet50V2 Test 5 Loss

Unfortunately, finetuning too much with these additional layers led to worst performance. Perhaps, this is due to backpropagation, which changed the weights (which are 9M, against the 14M that are not trainable) to a not-optimal solution. Anyway, this result is also too poor compared to test 3, which without having the newly added dense layers still perform better.

4.3 ResNet101V2

ResNet101 was implemented as one of the model used in the ensemble method cited in the ResNet50V2 introduction. ResNet101 is an extension of ResNet50 that goes deeper reaching 101 layers in depth, simply adding more convolutional 'sub'-blocks in the blocks[21]. ResNet101 proved to be slightly more accurate than ResNet50.[22]

4.3.1 Test 1: Classical ResNet101V2

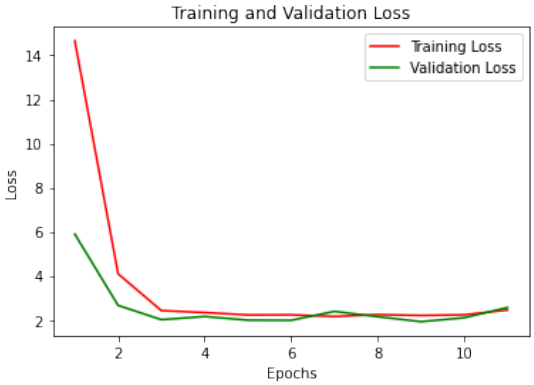
The original ResNet50 comes with a GlobalAveragePooling2D and a prediction layer soon after. In this test we used the same approach, resizing the prediction layer to the number of classes we have.
The results obtained after training are:

Feature Extraction				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
11	0.3756	0.3195	1.9798	2.6044

[21] Referring to the notation used before, we intend as a sub-block a block between two adds.

[22] It is not only additional network used in the ensemble method, an extension of it was also used, ResNet152. This model is the more accurate of the three, but because of it size we decide to limit our studies to the 101-layers version.

(a) ResNet101V2 Feature Extraction Accuracy



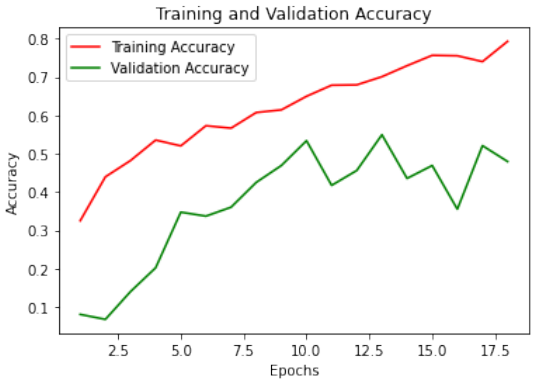
(b) ResNet101V2 T Feature Extraction Loss

The network under-fitted our training, performing drastically poorly. The same considerations done in chapter 4.2.1 apply here.

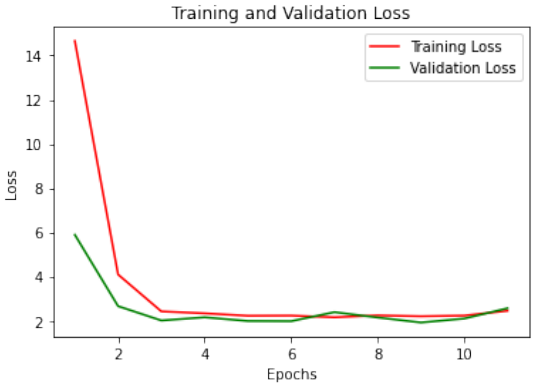
4.3.2 Test 2: Finetuning One Sub-Block

As done for ResNet50, we tried to improve our ResNet101 performance using finetuning. In this paragraph we just finetuned the last sub-block of the network.

Finetuning One Sub-Block				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
18	0.5492	0.5218	1.7628	2.3036



(a) ResNet101V2 Test 2 Accuracy



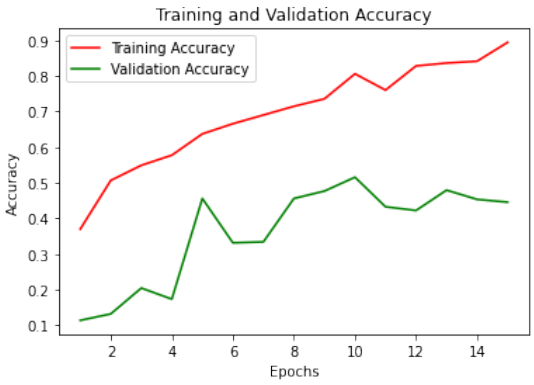
(b) ResNet101V2 Test 2 Loss

As expected, the performance are better compered to test 1. Anyway, the network still cannot converge: finetuning just one sub-block is not enough for a neural network made of 101 layers.

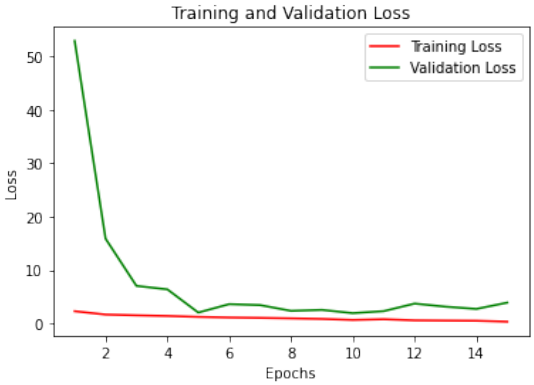
4.3.3 Test 3: Finetuning the Entire Block 5

Aiming to overfitting, we finetuned the entire block 5, which is made of 3 sub-blocks.

Finetuning Entire Block 5				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
15	0.5155	0.4805	1.9216	3.4421



(a) ResNet101V2 Test 3 Accuracy



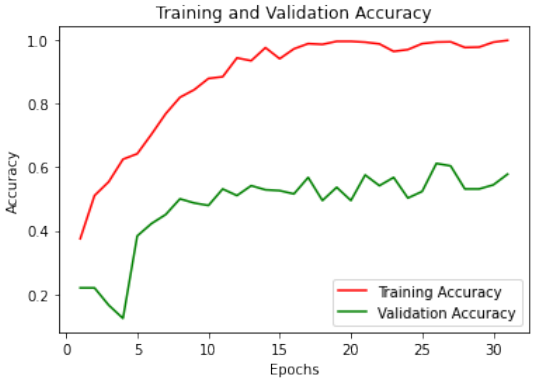
(b) ResNet101V2 Test 3 Loss

Unfortunately, we have little less performance and we still don't reach convergence: the training accuracy is less than 0.9, meaning that we are not finetuning enough but we are following the right lead since the training accuracy gain 0.1 point.

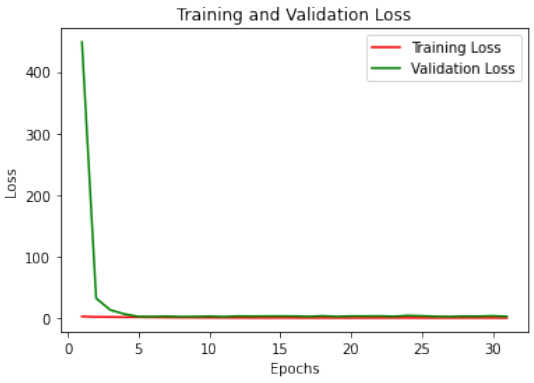
4.3.4 Test 4: Finetuning Half Block 4

In this paragraph, heavy finetuning is done setting as trainable parameters all of them after *conv4_block13_out* for a total of 13 sub-blocks finetuned (i.e., 10 in block 4 and 3 in block 5).

Finetuning Half Block 4				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
31	0.6114	0.5977	2.3865	2.3642



(a) ResNet101V2 Test 4 Accuracy



(b) ResNet101V2 Test 4 Loss

Finetuning these many layers helped the network to be more topic-specific. In fact, the neural network finally overfit properly our training data as the accuracy plot shows, also even the loss is higher than before. However, even if we improve from the previous test it is not enough if we consider the best result obtained using ResNet50 or VGG16.

4.3.5 Test 5: Test 4 and Dropout

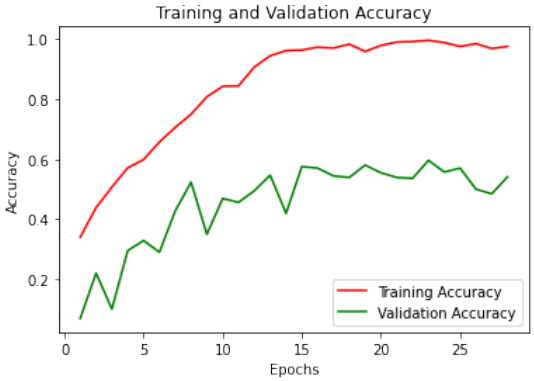
In test 4 the network overfitted, even if not much, we try to use dropout to mitigate this problem. This approach can result in two possible outcomes:

1. We perform more or less the same as in test 4: this is the goal of dropout, neutralize neurons but performing the same as before, reducing in this way the number of computations and performing better on the test set.

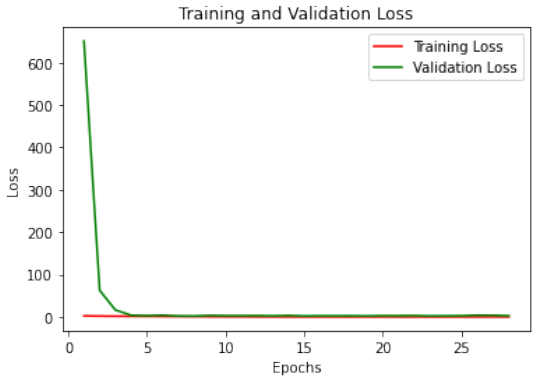
2. We perform worst than test 4: dropout neutralizes too many units, performing worst both on the validation set and the test set. This is a problem, the dropout rate should be reduced or other regularization techniques apply.

The results obtained are the following:

Finetuning Half Block 4 and Dropout				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
28	0.5959	0.5356	2.2455	2.8590



(a) ResNet101V2 Test 5 Accuracy



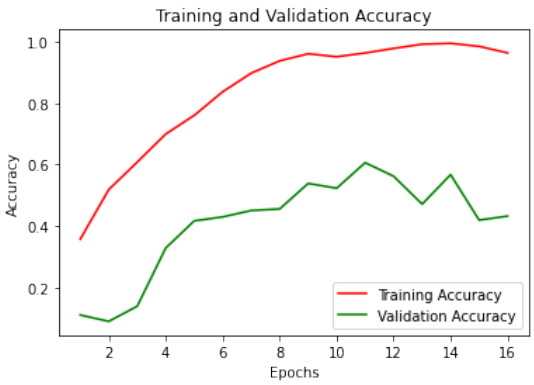
(b) ResNet101V2 Test 5 Loss

The plots and the table show that the situation this test inclines is the second outcome presented. Even if the validation accuracy losses just 0.2 points, the test accuracy has a gap 0.6 points with the one obtained in test 4, which is a lot considering that the overall performance is not good (it barely arrives to 0.6, which is very low to be considered a good classifier). Hence, for the following test we decided to drop the use of dropout until better performances are met.

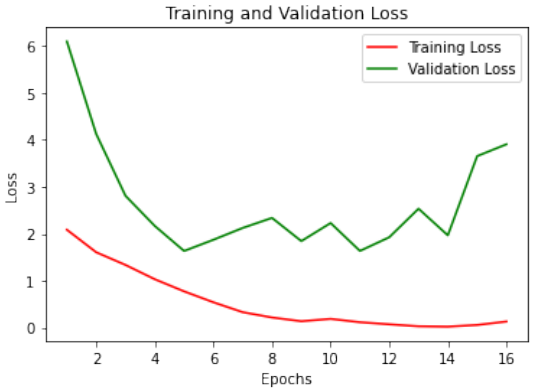
4.3.6 Test 6: Adding Dense Layers to Test 4

Following the tests performed with VGG16 and ResNet50V2, we tried also here to add dense layers at the top of the *GlobalAveragePooling2D* layer. Doing this experiment, the learning rate of *Adam* is change from the default value (i.e, 0.001) to 0.0001.
The result obtained are the following:

Finetuning Half Block 4 and Adding 2 Dense Layers				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
16	0.6062	0.4552	1.6339	3.5964



(a) ResNet101V2 Test 6 Accuracy



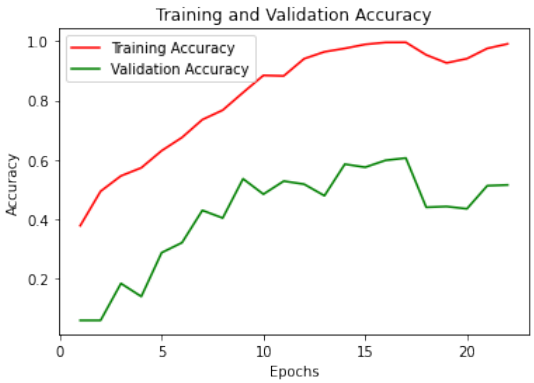
(b) ResNet101V2 Test 6 Loss

The introduction of these layers worsen the performance of test 4, but they helped achieving better loss values. Anyway, they had a bad effect on the test accuracy which is now one of the lower values obtained for ResNet101. This can be explained, perhaps, since the network now shrinks in two layers before predicting: layers specialized on the training set. Hence, they ended up discovering features that are not helpful for predicting correctly the test set.

4.3.7 Test 7: Going Deeper and Deeper

Since increasing the finetuning level led to better result in the previous tests, we finetuned ResNet101 till *conv3_block4_out* including the parameters of all conv4 (i.e., block4). The network in this way is really unbalanced in terms of trainable parameters, most of them were trainable (i.e., about 80%). Even that, we tested it and what we obtain was a failure, the network didn't learn well an ended up with a tremendous low accuracy. Thus, we decided to limit our going deeper to the *conv4_block10_out*: adding three more sub-blocks to the ones tested in test 3 (which is still our current optimum). The result obtained are the following:

Finetuning till conv4_block10_ou				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
22	0.6062	0.5310	2.1705	2.5997



(a) ResNet101V2 Test 7 Accuracy

(b) ResNet101V2 Test 7 Loss

These results gained the second best position for our ResNet101 tests. Anyway, the performance are still too poor and the computational power required for training is very high (i.e., we have about 30M parameters). It is important to notice that the same result for validation was obtained in test 6, but with a lower score for the test accuracy. This indicates that the point is a pitfall, where our validation gets caught and wherever it moves it finds only worst values.

Anyway, we have no way to escape it: the accuracy for the training set is curvilinear without frequent zig-zags, thus changing the learning rate of Adam will let to worst or the same result:

1. if we decrease the learning rate we increase the possibility to stop earlier or to fall in the same pitfall;

2. if we increase the learning rate we will overfit very fast jeopardizing the training.

4.4 InceptionV3

InceptionV3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. The model is the culmination of many ideas developed by multiple researchers over the years. It is based on the original paper: "Rethinking the Inception Architecture for Computer Vision" by Szegedy, et. al.

The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Batchnorm is used extensively throughout the model and applied to activation inputs. Loss is computed via softmax.

A high-level diagram of the first model proposed is shown below:

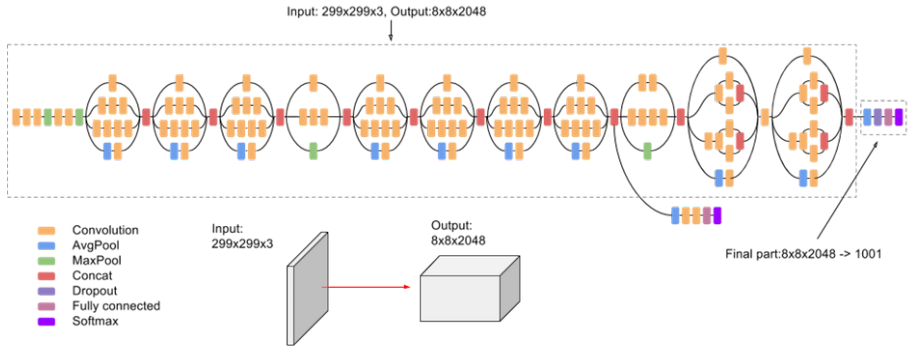


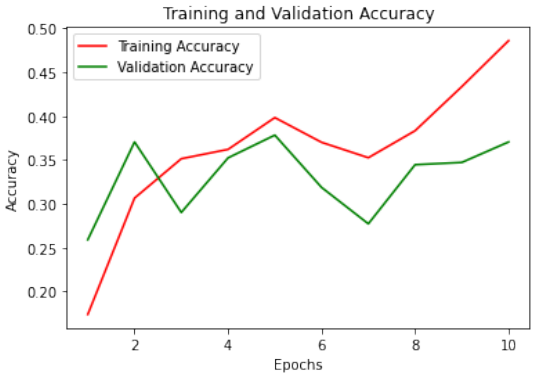
Figure 47: Inception v3 First Proposal

Anyway, the model presented in figure is different to the one present in Keras, which at the end of the network has simply a *GlobalAveragePooling* layer and a prediction layer. The model depicted has a more complex output and it make use of an *auxiliary classifier*: training using loss at the end of the network didn't work well since the network is too deep and the gradients don't propagate cleanly. As a hack, the idea of the time was to attach "auxiliary classifiers" at several intermediate points in the network that also try to classify the image and receive loss. All of that just because the network was proposed before *batch normalization*, with batch normalization there is no longer need to use this trick, In fact, the current version of InceptionV3 implemented in Keras uses BatchNorm.

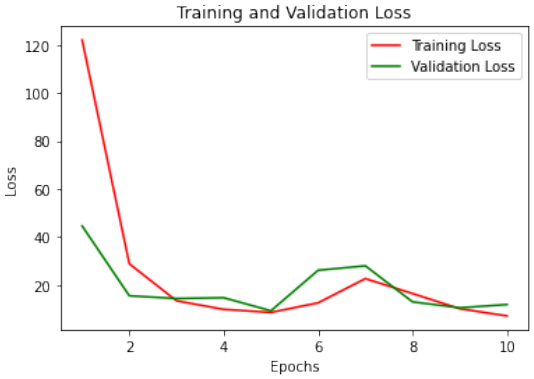
4.4.1 Test 1: Classical InceptionV3

The base InceptionV3 provided by Keras comes with a GlobalAveragePooling2D and a prediction layer after that. In this test we used the same approach, resizing the prediction layer to the number of classes we have.
The results obtained after training are (learning rate=0.01, Adam):

Feature Extraction				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
10	0.3782	0.4092	9.3051	10.9338



(a) InceptionV3 Test 1 Accuracy



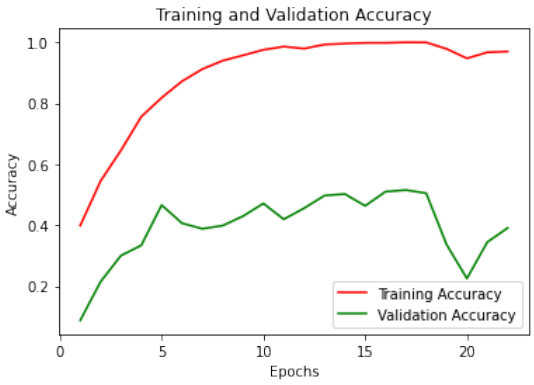
(b) InceptionV3 Test 1 Loss

The network under-fitted our training, performing drastically poorly. The same considerations done in chapter 4.2.1 apply here.

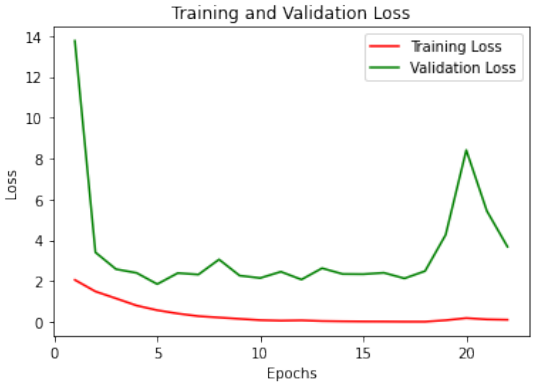
4.4.2 Test 2: Finetuning 1 Block

As done for the previous networks, we finetuned one block at a time, where a block is an inception block.
The results for finetuning only the last block are (learning rate=0.001, Adam):

Finetuning 1 Block				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
22	0.5155	0.4437	2.1245	3.3541



(a) Inception Test 2 Accuracy



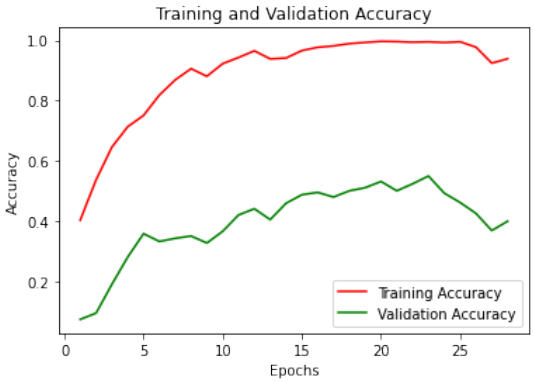
(b) Inception Test 2 Loss

As expected, the network finally overfitted and the accuracy of the validation set increased as well. Anyway, the curves in the last epochs have a strange behavior, the validation accuracy and loss suddenly go worse. Perhaps, this is due to the fact that we are overfitting by few epochs at that point and the iteration moves the curves to points that are in the close to an high accuracy, but which features are far different from the one of the validation set.

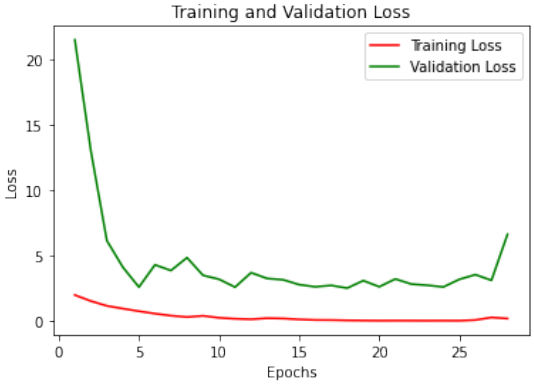
4.4.3 Test 3: Finetuning 2 Blocks

Since we want to find first a good number of blocks to set trainable and than optimize it, we continued our test finetuning the last 2 blocks.
The results obtained are (learning rate=0.001, Adam):

Finetuning 2 Blocks				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
28	0.5492	0.4184	2.7196	6.3845



(a) Inception Test 3 Accuracy



(b) Inception Test 3 Loss

The plots are more or less similar to the ones in the previous test, but the validation accuracy is greater by 0.03 points and the test accuracy is less than 0.03 points and with doubled loss. As said before, this can be addressed to the fact that the test and validation set can be different to each other in term of style, thus same feature discovered by a feature map can apply correctly to one case and not in the other. Anyway, since we are supposed to know nothing about the test set and our conclusion should only be driven by the result of the validation, the model can be consider an improvement of the one obtained in the previous test.

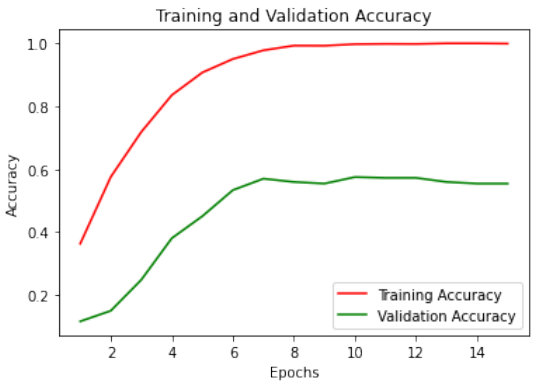
Also in this case, the accuracy (loss) decreases (increases) suddenly, but the behavior seems here to be alleviated. Thus, suggesting that finetuning more blocks can help preventing this strange behavior.

4.4.4 Test 4: Finetuning 3 Blocks

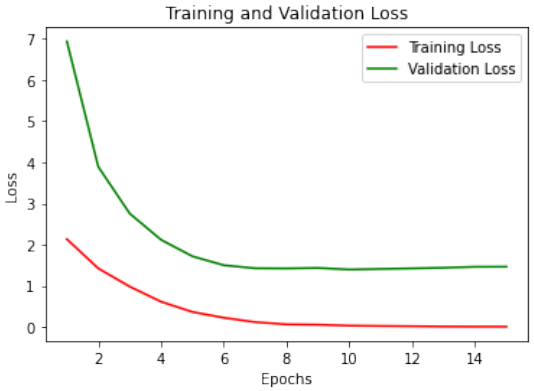
Finetuning till the third last block was our maximum in finetuning this type of network, since finetuning more led to worse performance[23].
The results obtained are (learning rate=0.0001, Adam):

Finetuning 3 Blocks				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
15	0.5725	0.5908	1.4111	1.4673

[23]Here not reported, but the code on GitHub provide the possibility to finetune other 2 more layers.



(a) Inception Test 4 Accuracy



(b) Inception Test 4 Loss

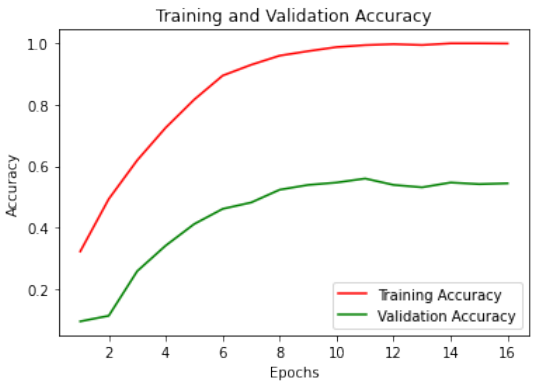
First thing to notice is the absence of the strange behavior underlined in the previous paragraphs, having more parameters to train help our network to discover a better solution. The validation accuracy gained other 0.03 points and the testing accuracy increases a lot reaching 0.59: the model become more precise in understanding the differences between authors, differences that in the testing loss are more concise reading this result.

Anyway, we overfit very fast so we decided to try preventing it using dropout.

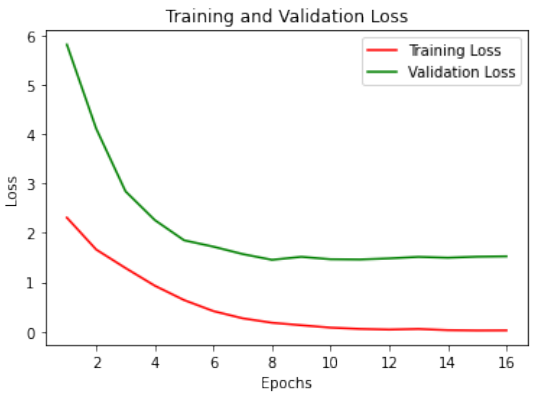
4.4.5 Test 5: Finetuning 3 Blocks with Dropout

In this test, we added a dropout layer after the *GlobalAveragePooling2D*
The results obtained are (learning rate=0.0001, Adam):

Finetuning 3 Blocks with Dropout				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
16	0.5596	0.6069	1.4603	1.3284



(a) Inception Test 5 Accuracy



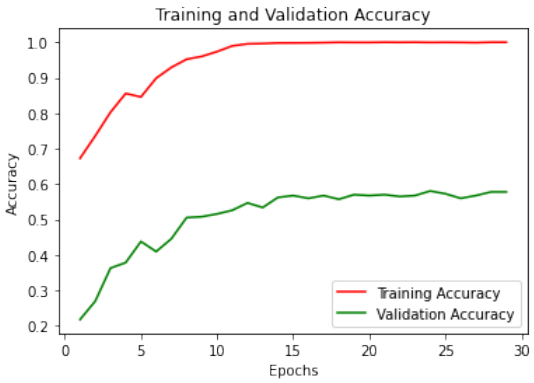
(b) Inception Test 5 Loss

The plots are quite the same of the test without dropout, but we need an additional epoch to stop the training. The dropout effect is very low, this is quite reasonable since the most of units are inside the InceptionV3 architecture. Anyway, by pure luck we achieve our best test accuracy till now, but since, at this point of the study, we care only about the validation accuracy, this result is meaningless.

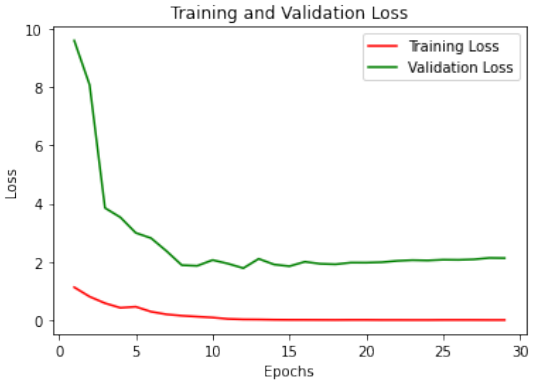
4.4.6 Test 6: Finetuning 3 Blocks and ExponentialDecay Learning Rate

In this test, we perform the same strategy of test 4 (finetuning 3 blocks), but we changed the optimizer learning rate. In order to try different approaches, here we used Adam with a *exponential learning rate decay*, which can be used through the **ExponentialDecay** class provided by Keras. The constructor takes 3 parameters[24]: the initial learning rate, the decay step and the decay rate, which we set to 0.01, 25 and 0.9 respectively.

Finetuning 3 Blocks and ExponentialDecay Learning				
Epoch stopped	**Validation Accuracy**	**Testing Accuracy**	**Validation Loss**	**Testing Loss**
29	0.5803	0.6253	2.0511	2.3217



(a) Inception Test 6 Accuracy



(b) Inception Test 6 Loss

It is important to notice that this method led us to reach high accuracy values on the training very quickly, but at the same time increasing the validation accuracy rapidly, and then stabilizing and growing slightly (almost imperceptibly looking at the plot), leading to improve more and more the accuracy obtaining the reported value. Moreover, we have a very high value for the test set as well, making this model the best obtained for InceptionV3.

An important thing to take into consideration is the behavior of the loss, which appears here much higher than in the fourth test, this means that even if we achieved a more precise model, we pay in the robustness of it.

[24]For the others we used the default values

5 — Ensemble Network

Ensembling consists of pooling together the predictions of a set of different models to produce better predictions. Nowadays, in every machine learning competitions, in particular on Kaggle, the winners use very large ensembles of models that inevitably beat any single model, no matter how good, hence we decided to use it also in our project.

Ensembling relies on the assumption that different well-performing models trained independently are likely to be good for different reasons: each model looks at slightly different aspects of the data to make its predictions, getting part of the "truth" but not all of it[25]. By pooling the perspective of different models together, a far more accurate description of the data can be obtained.

5.1 Our approach

In our project we decided to use a simple ensembling approach, which is *average voting*: average voting computes the average of the scores (i.e., softmax output) of the base classifiers and select the class with the highest score.

Although we decided to use this simple approach, even if it is very common, we decided to optimize the choice of models to be used in order to maximize the accuracy obtained on the validation set. In fact, all the models described in the chapter "pre-trained models" have been saved as *".keras"* models so that they can be reloaded and used to directly evaluate the validation set in the ensemble, so that they do not need to be trained again (thus increasing the speed of test execution).

The main problem was to choose which of these models combined together obtained the best accuracy. Since the number of models was more than 20 a grid search seemed too slow as a method of solving this problem, so we decided to use a genetic algorithm to solve the problem automatically.

Anyway, due to the fact that we have different inputs for each type of network (ResNet, VGG and Inception) we decided to divide the ensembling networks into three types, one for each type of network.

After each genetic solution, we tested the network obtained on the test set in order to see the improvement we have on an unseen set, as we did in the previous Sections.

5.1.1 The Ensemble Classes

To perform the ensemble in an orderly fashion, what was done was to create three classes called respectively *EnsembleVGG*, *EnsembleResNet* and *EnsembleInception*. The classes allows to construct an ensemble method based on a subset or the entire set of network provided to it, and evaluate the obtained network using the *Keras* function *evaluate*.

The class for VGG16 models is as follows:

```
class EnsembleVGG:
    def __init__(self, set_vgg):
        self.set_vgg = set_vgg

        # create list of models
        self.model_list = []

        base_dir = '/content/drive/MyDrive/Fazzari_Ramo/Models/'
        # VGG16
        for name in vgg16:
            tmp = ks.models.load_model(base_dir + 'vgg16/' + name)
```

[25] This can be see in detail in the "data visualization" chapter

```
            tmp._name = name
            for i, layer in enumerate(tmp.layers):
                layer.trainable = False
                layer._name = 'ensemble_' + str(i + 1) + '_' + layer.name
            self.model_list.append(ks.models.load_model(base_dir + 'vgg16/' +
                                        name))


    def ensembleModel(self, active_models):
        models = []
        for i, model in enumerate(self.model_list):
            if active_models[i] == 1:
                model._name = str(i)
                models.append(model)
        model_input = ks.Input(shape=(IMAGE_WIDTH_VGG, IMAGE_HEIGHT_VGG, 3))
        model_outputs = [model(model_input) for model in models]
        ensemble_output = ks.layers.Average()(model_outputs)
        ensemble_model = ks.Model(inputs=model_input, outputs=ensemble_output)
        ensemble_model.compile(metrics=['accuracy'])
        return ensemble_model

    def evaluate(self, ensemble_model):
        loss, acc = ensemble_model.evaluate(self.set_vgg)
        return acc

    def ensemble_and_evaluate(self, active_models):
        if np.sum(active_models) == 0:
            return 0.0
        elif np.sum(active_models) == 1:
            print(active_models.index(1))
            return self.evaluate(self.model_list[active_models.index(1)])
        else:
            return self.evaluate(self.ensembleModel(active_models))

    def models_len(self):
        return len(self.model_list)
```

The class related to the ResNet family of networks is as follows:

```
class EnsembleResNet:
    def __init__(self, set_resnet):
        self.set_resnet = set_resnet

        # create list of models
        self.model_list = []

        base_dir = '/content/drive/MyDrive/Fazzari_Ramo/Models/'
        # ResNet
        resnet = resnet50 + resnet101
        for i, name in enumerate(resnet):
            if i < len(resnet50):
                tmp = ks.models.load_model(base_dir + 'resnet50/' + name)
            else:
                tmp = ks.models.load_model(base_dir + 'resnet101/' + name)
            tmp._name = name
            for j, layer in enumerate(tmp.layers):
                layer.trainable = False
                layer._name = 'ensemble_' + str(j + 1) + '_' + layer.name

            self.model_list.append(tmp)



    def ensembleModel(self, active_models):
```

```
        models = []
        for i, model in enumerate(self.model_list):
            if active_models[i] == 1:
                model._name = str(i)
                models.append(model)
        model_input = ks.Input(shape=(IMAGE_WIDTH_RESNET, IMAGE_HEIGHT_RESNET,
                                      3))
        model_outputs = [model(model_input) for model in models]
        ensemble_output = ks.layers.Average()(model_outputs)
        ensemble_model = ks.Model(inputs=model_input, outputs=ensemble_output)
        ensemble_model.compile(metrics=['accuracy'])
        return ensemble_model

    def evaluate(self, ensemble_model):
        loss, acc = ensemble_model.evaluate(self.set_resnet)
        return acc

    def ensemble_and_evaluate(self, active_models):
        if np.sum(active_models) == 0:
            return 0.0
        elif np.sum(active_models) == 1:
            print(active_models.index(1))
            return self.evaluate(self.model_list[active_models.index(1)])
        else:
            return self.evaluate(self.ensembleModel(active_models))

    def models_len(self):
        return len(self.model_list)
```

Finally, the ensemble class for Inception is:

```
class EnsembleInception:
    def __init__(self, set_inception):
        self.set_inception = set_inception

        # create list of models
        self.model_list = []

        base_dir = '/content/drive/MyDrive/Fazzari_Ramo/Models/'
        # VGG16
        for name in inception:
            tmp = ks.models.load_model(base_dir + 'inception/' + name)
            tmp._name = name
            for i, layer in enumerate(tmp.layers):
                layer.trainable = False
                layer._name = 'ensemble_' + str(i + 1) + '_' + layer.name
            self.model_list.append(ks.models.load_model(base_dir + 'inception/'
                                    + name))


    def ensembleModel(self, active_models):
        models = []
        for i, model in enumerate(self.model_list):
            if active_models[i] == 1:
                model._name = str(i)
                models.append(model)
        model_input = ks.Input(shape=(IMAGE_WIDTH_INCEPTION,
                                      IMAGE_HEIGHT_INCEPTION, 3))
        model_outputs = [model(model_input) for model in models]
        ensemble_output = ks.layers.Average()(model_outputs)
        ensemble_model = ks.Model(inputs=model_input, outputs=ensemble_output)
        ensemble_model.compile(metrics=['accuracy'])
        return ensemble_model

```

```
    def evaluate(self, ensemble_model):
        loss, acc = ensemble_model.evaluate(self.set_inception)
        return acc

    def ensemble_and_evaluate(self, active_models):
        if np.sum(active_models) == 0:
            return 0.0
        elif np.sum(active_models) == 1:
            print(active_models.index(1))
            return self.evaluate(self.model_list[active_models.index(1)])
        else:
            return self.evaluate(self.ensembleModel(active_models))

    def models_len(self):
        return len(self.model_list)
```

The constructor takes as input the testset for which the ensemble class will do the evaluation, and what it does is load the different models. Next, in each three we have four functions:

1. *ensembleModel*: given an array of equal size to the number of different models, the function creates an ensemble model using the models that are in the position where in the input array there is a one.
2. *evaluate*: calls the keras evaluate function on the ensemble model
3. *ensemble_and_evaluate*: calls both the ensembleModel and the evaluate functions. In cases the input is all made of zeros, it returns 0 without further computations, on the other hand if the input has only one 1 in it the function evaluates it directly.
4. *models_len*: retrieves the number of models.

Notice that in all functions we modified the name of the layers and models in order to prevent every form of name duplication in the ensemble network.

5.1.2 Genetic Algorithm Workflow

Following the guidelines in Section 2.6, we need to decide how to develop the different components and decide on the hyperparameters in order to define the flow of the genetic algorithm.

5.1.2.1 Genotype

In our specific case, our chromosomes are binary encoded and each gene represents a different model. Hence, we have individuals made of a number of genes equal to the number of different models that we have.

```
# create an operator that randomly returns 0 or 1
toolbox.register('zeroOrOne', random.randint, 0, 1)

# define a single objective, maximizing fitness strategy:
creator.create('FitnessMax', base.Fitness, weights=(1.0,))

# create the Individual class based on list:
creator.create('Individual', list, fitness=creator.FitnessMax)

# create the individual operator to fill up an Individual instance:
toolbox.register('individualCreator', tools.initRepeat, creator.Individual,
                                      toolbox.zeroOrOne, INDIVIDUAL_LENGTH)
```

5.1.2.2 Population

The population is composed of 8 individuals, we decided to keep the number of individuals low so as to reduce the number of evaluations done and, therefore, the total duration of the algorithm, in order to obtain a satisfactory result by waiting less than an hour.

```
toolbox.register('populationCreator', tools.initRepeat, list, toolbox.
                                        individualCreator)
```

5.1.2.3 Fitness Function

The fitness function is the accuracy obtained through the *predict_and_evaluate* function of the Ensemble class. It was registered in the *toolbox* as follows:

```
def ensembleAccuracy(individual):
    return ensemble.predict_and_evaluate(individual),

toolbox.register('evaluate', ensembleAccuracy)
```

Notice that the function returns a tuple, this is done because the framework accept only tuple values as result of the evaluation function.

5.1.2.4 Selection Algorithm

The selection algorithm used in this case is *tournament selection*. In each round of the tournament selection method, two individuals are randomly picked from the population, and the one with the highest fitness score wins and gets selected. We decided to select only two individuals since our population is small and selecting more could cause an abuse in exploitation.

```
# Tournament selection with tournament size of 2:
toolbox.register("select", tools.selTournament, tournsize=2)
```

5.1.2.5 Crossover Algorithm

As crossover algorithm we decided to use the *two-point crossover*. In the two-point crossover method, two crossover points on the chromosomes of both parents are selected randomly. The genes residing between these points are swapped between the two parent chromosomes.

The following diagram demonstrates a two-point crossover carried out on a pair of binary chromosomes, with the first crossover point located between the third and fourth genes, and the other between the seventh and eighth genes:

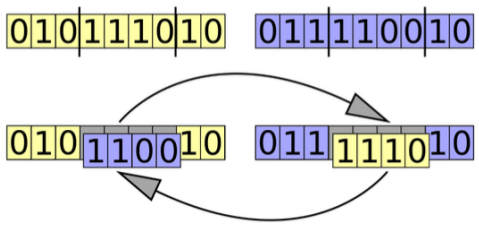


Figure 54: Two point crossover example.

The code for that is:

```
toolbox.register("mate", tools.cxTwoPoint)
```

5.1.2.6 Mutation Algorithm

As mutation algorithm we decided to use the *Multiple Flip bit mutation*. When applying the flip bit mutation to a binary chromosome, one gene is randomly selected and its value is flipped (complemented), as shown in the following diagram:

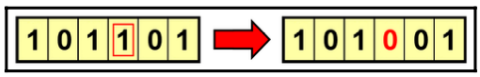


Figure 55: Flip bit mutation example.

This can be extended to several random genes being flipped instead of just one, obtaining the multiple flip bit mutation that we use. The code for that is:

```
toolbox.register("mutate", tools.mutFlipBit, indpb=1.0/INDIVIDUAL_LENGTH)
```

5.1.2.7 Elitism

We set the number of individuals for the elitism mechanism to 1. This value was decided to have a minimum of elitism, but without abusing it too much. In fact, the number of individuals is so low that having a higher value could lead to a problem, therefore, in order not to risk bad results (i.e., not global optimum results), we chose to use this conservative approach and limit the number to only one.

5.1.2.8 GA Flow

The first thing to do is to define the initial population, this is easily done by the following line of code:

```
population  = toolbox.populationCreator(n=POPULATION_SIZE)
```

After that, we created a statistical object. This object has served to have a report of the flow of the genetic algorithm, allowing us to save for each generation the maximum and average fitness values obtained, so that we can show them once the generations are over.

```
stats = tools.Statistics(lambda ind: ind.fitness.values)
stats.register("max", np.max)
stats.register("avg", np.mean)
```

Then, the last object we need to create for the *eaSimple* is the **HallOfFame**, which can be done through the following line of code:

```
hof = tools.HallOfFame(HALL_OF_FAME_SIZE)
```

The main flow is done using the *eaSimple* in the way as follow:

```
population, logbook = eaSimpleWithElitism(population,
                    toolbox,
                    cxpb=P_CROSSOVER,
                    mutpb=P_MUTATION,
                    ngen=MAX_GENERATIONS,
                    stats=stats,
                    halloffame=hof,
                    verbose=True)
```

Where the P_CROSSOVER is equal to 0.9, P_MUTATION 0.1.[26]

[26]Value suggested by the book "Hands on Genetic Algorithms with Python" by Eyal Wirsansky, also the other values of probabilities are taken by the suggestions of the book

5.1.3 GA Results for VGG16 Ensemble

The following graph shows the results obtained for each generation from the population (i.e., the maximum and mean accuracy):

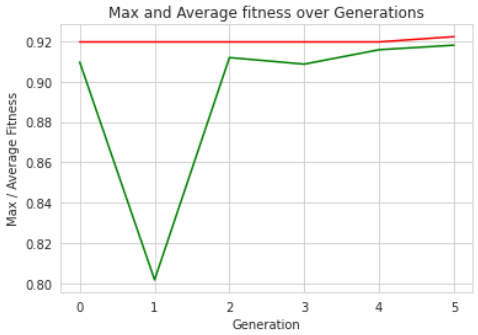


Figure 56: Maximum and average accuracy in the population for each generation.

As it can be noticed from the graph the elitism allows the maximum value to never go down. Moreover, it is possible to notice that the maximum result obtained is much higher than the one obtained with VGG16, which reached almost 80%, instead here we reach 92.23%. This was obtained from the genetic algorithm by combining the following models together:

- VGG16 feature extraction
- VGG16 feature extraction with dropout
- VGG16 with 2 layers finetuned and with dropout
- VGG16 with 1 layer finetuned

Chosen by the chromosome [0, 0, 1, 1, 1, 1, 0], our winner in the final population.

Looking closely at the graph we can see a very slight growth at the end of the fifth generation, this with later testing turned out to be the global optimum.

Testing the network obtained by choosing the above models with the test set it was possible to obtain an accuracy of 0.9586. Very high and even higher than that obtained on the validation set, this can be related to two reasons:

1. The features of the frameworks in the test set are much more similar to those in the training, possible but not that much;
2. The quantity of pictures present in the test set is greater than that of the validation as reported in chapter 1.2, therefore the weight of mistaking a picture in the case of validation has a higher weight in percentage than in the test set.

5.1.4 GA Results for ResNet Ensemble

The following graph shows the results obtained for each generation from the population (i.e., the maximum and mean accuracy):

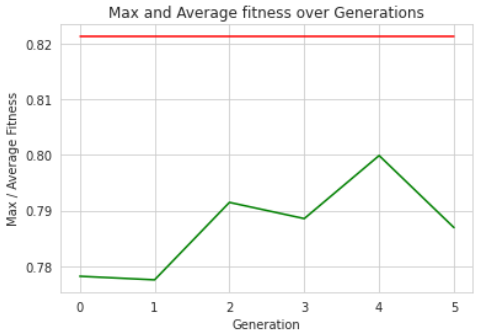


Figure 57: Maximum and average accuracy in the population for each generation.

The maximum accuracy value for ResNet was obtained in Section 4.2.4 using fine tuning of one block and two additional dense layers at the end of the network, reaching an accuracy score of 0.6736. Also here we exceed the value reached by the single model obtaining a much higher value, reaching 0.8212, satisfactory value even if lower than the one obtained with the ensemble of VGGs. This was obtained from the genetic algorithm by combining the following models together:

- ResNet50 with 1 block finetuned
- ResNet50 with 2 blocks finetuned
- ResNet50 with 1 block finetuned plus two dense layers at the end of it
- ResNet50 with 2 blocks finetuned plus two dense layers at the end of it
- ResNet101 finetuned till block 4

Chosen by the chromosome [0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0], our winner in the final population.

Anyway, it is important to notice here that the value for the average forms a mountain scape shape, this is a problem related to the GA hyperparameters and operators, which led to unravel the result obtained by the previous population.

Testing the network obtained by choosing the above models with the test set it was possible to obtain an accuracy of 0.9517. Very high and much higher than that obtained on the validation set, hence we can take into consideration the hypothesis done, related to this behavior, in the previous paragraph.

5.1.5 GA Results for ResNet Inception

The following graph shows the results obtained for each generation from the population (i.e., the maximum and mean accuracy):

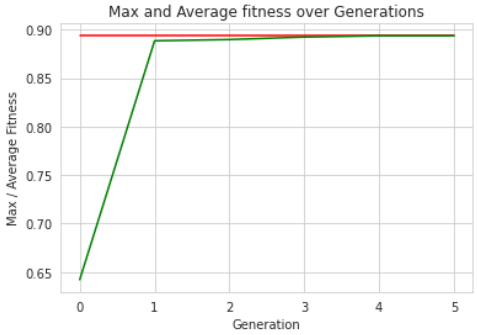


Figure 58: Maximum and average accuracy in the population for each generation.

The maximum accuracy value for Inception was obtained in Section 4.4.5 fine tuning 3 blocks and using exponential decay for the learning rate, reaching an accuracy score of 0.5803. Also here we exceed the value reached by the single model obtaining a much higher value, reaching 0.8938, satisfactory value even if lower than the one obtained with the ensemble of VGGs. This was obtained from the genetic algorithm by combining the following models together:

- Inception with 1 block finetuned
- Inception with 3 blocks finetuned

Chosen by the chromosome [0, 1, 0, 1, 0], our winner in the final population.

Testing the network obtained by choosing the above models with the test set it was possible to obtain an accuracy of 0.9264 and a loss equal to 0.3 (in the other ensembles it was really close to 0 that the evaluation function of keras printed directly 0). Very high and much higher than that obtained on the validation set, hence we can take into consideration the hypothesis done, related to this behavior, in the previous paragraph.

6 — Conclusion

The project reports a series of possible solutions for the artist recognition from a given paintings dataset. The solutions are therefore based on the use of multiple CNN networks, considering the current state of the art and the limitations of Colab. First of all, we impemented several custom models, evaluating for each of them the validation and test accuracy and loss, and taking choices with the purpose of maximising performance and reacting to possible training issues. Starting from a very classical CNN, we also tested the effects of Droupout, Batch Normalization and Data Augmentation, we then defined and tested new architectures to tackle the artist classification task (namely Aggressive Downsampling techniques and custom Inception Modules). Most performing networks have been input to a hyper-paramters optimization procedure, and they have been analyzed to explainably visualize what they focus on and what is their robustness to occlusions. Secondly we considered pre-trained models where we were able to obtain more satisfactory results than scratch and the best results were obtained thanks to VGG16, which with its simplicity manages better to classify than the larger networks that are trained with greater accuracy on *imagenet*, which has a dataset very different from ours and perhaps this is the reason why we obtained slightly worse results for ResNets and Inception. Much more satisfactory results have been obtained thanks to the ensemble, reaching thanks to the aggregation of VGG16s models about 96% of accuracy for the test set and about 92% for the validation set, the highest values reached in all our tests, bringing the model obtained to be our final model for the identification of the painters of the paintings. One thing to keep in mind talking about the ensemble is the fact that the genetic algorithm finds the conclusion immediately or almost immediately, this thing is related to a lucky starting point, but also, above all, to the very small number of genes that are part of the different chromosomes. As for the state-of-art, we said in the introduction that the highest value obtained on the test accuracy was found by Nitin Viswanathan, and it was about 90%, just few points below us. This makes our final result higher than the state-of-the-art.

References

[1] T. E. Lombardi. The classification of style in fine-art paint- ing. ETD Collection for Pace University, 2005.

[2] J. Jou and S. Agrawal. Artist identification for renaissance paintings.

[3] T. Mensink and J. van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. 2014

[4] B. Saleh and A. M. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. CoRR, abs/1505.00855, 2015.

[5] Nitin Viswanathan, Artist Identification with Convolutional Neural Networks

[6] Eyal Wirsansky, "Hands on Genetic Algorithms with Python" book

[7] Perez, Luis; Wang, Jason. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

[8] KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[9] SELVARAJU, Ramprasaath R., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618-626.