



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	3
1.1	State of the Art	3
1.2	Dataset	4
2	General Information Useful for Training	7
2.1	Data Augmentation	7
2.2	Regularization	7
2.3	Dropout	8
2.4	Activation Functions	8
2.5	Optimizers	8
2.6	Genetic Algorithms	8
2.6.1	Genotype	9
2.6.2	Population	9
2.6.3	Fitness Function	9
2.6.4	Selection Algorithm	9
2.6.5	Crossover Algorithm	9
2.6.6	Mutation Algorithm	10
2.6.7	Elitism	10
3	CNN from Scratch	11
3.1	Standard CNN	11
3.1.1	Simple model	11
3.1.2	Deeper network and dropout	12
4	Pre-Trained Models	13
4.1	VGG16	13
4.1.1	Test 1: Classical VGG16 (Feature Extraction)	13
4.1.2	Test 2: Adding Dropout to Test 1	14
4.1.3	Test 3: Finetuning One Convolutional Layer	15
4.1.4	Test 4: test 3 with dropout and different optimizer	16
4.1.5	Test 5: Finetuning Two Convolutional Layers	16
4.1.6	Test 6: Finetuning One Convolutional Layer and Weights Regularization	17
4.1.7	Test 7: Finetuning Two Convolutional Layers and Weights Regularization	18
4.1.8	Test 8: Genetic Algorithm for Hyper-parameters and Architecture Optimization	18
4.2	ResNet50V2	18
4.2.1	Test 1: Classical ResNet50V2 (Feature Extraction)	18
4.2.2	Test 2: Finetuning 1 block	19
4.2.3	Test 3: Finetuning 2 blocks	20
4.2.4	Test 4: Finetuning with One Block and Adding Two Dense layers	20
4.2.5	Test 5: Finetuning with Two Blocks and Adding Two Dense layers	21
4.3	ResNet101V2	21
4.3.1	Test 1: Classical ResNet101V2	22
4.3.2	Test 2: Finetuning One Sub-Block	22
4.3.3	Test 3: Finetuning the Entire Block 5	23
4.3.4	Test 4: Finetuning Half Block 4	23
4.3.5	Test 5: Test 4 and Dropout	24
4.3.6	Test 6: Adding Dense Layers to Test 4	25

4.3.7	Test 7: Going Deeper and Deeper	25
4.4	InceptionV3	26
4.4.1	Test 1: Classical InceptionV3	27
4.4.2	Test 2: Finetuning 1 Block	27
4.4.3	Test 3: Finetuning 2 Blocks	28
4.4.4	Test 4: Finetuning 3 Blocks	29
4.4.5	Test 5: Finetuning 3 Blocks with Dropout	29
4.4.6	Test 6: Finetuning 3 Blocks and ExponentialDecay Learning Rate	30
5	Siamese Architectures	32
5.1	Siamese Network	32
5.2	Convolutional Siamese Network	32
6	Ensemble Network	33

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 since they correspond to two different periods.

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 network; networks based on comparison with baseline input and using an ensemble network made of by the best classifiers found.

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*, *Banksy* 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 style* in Fine-Art Painting².

The first time the problem of artist identification was really tackled was with J. Jou and S. Agrawal³ 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

¹<https://www.artsy.net/categories>

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

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

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

⁵The 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.

in the Rijksmuseum in Amsterdam (the Netherlands). For the artist classification challenge, the paper said they read a test accuracy of about 60%. The year later, Saleh and Elgammal’s paper⁶ 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 Kaggle⁸. 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 Viswanathan⁹ 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 Challenge* 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 dataset¹⁰ 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,

⁶B. 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

⁷In 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>

⁹Nitin 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:

```

1 import os
2 import numpy as np
3 import shutil
4
5 rootdir= '/content/images/images' #path of the original folder
6 classes = os.listdir(rootdir)
7
8 for i, c in enumerate(classes, start=1):
9     if c not in artists_top_name.tolist():
10         shutil.rmtree(rootdir + '/' + c)
11         continue
12     if not os.path.exists(rootdir + '/train/' + c):
13         os.makedirs(rootdir + '/train/' + c)
14     if not os.path.exists(rootdir + '/test/' + c):
15         os.makedirs(rootdir + '/test/' + c)
16
17     source = os.path.join(rootdir, c)
18     allFileNames = os.listdir(source)
19
20     np.random.shuffle(allFileNames)
21
22     test_ratio = 0.10
23     train_FileNames, test_FileNames = np.split(np.array(allFileNames),
24                                                [int(len(allFileNames)*
25                                                    (1 - test_ratio))])
26
27     train_FileNames = [source+'/'+ name for name in train_FileNames.tolist()]
28     test_FileNames = [source+'/'+ name for name in test_FileNames.tolist()]
29
30     for name in train_FileNames:
31         shutil.copy(name, rootdir + '/train/' + c)
32
33     for name in test_FileNames:
34         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:

```

1 import tensorflow as tf
2
3 training_images = tf.keras.preprocessing.image_dataset_from_directory(
4     TRAIN_DIR, labels='inferred', label_mode='categorical',
5     class_names=None, color_mode='rgb', batch_size=BATCH_SIZE,
6     image_size=(IMAGE_HEIGHT, IMAGE_WIDTH), shuffle=True, seed=RANDOM_SEED,
7     validation_split=VALIDATION_SPLIT, subset='training',
8     interpolation='bilinear', follow_links=False
9 )
10
11 val_images = tf.keras.preprocessing.image_dataset_from_directory(
12     TRAIN_DIR, labels='inferred', label_mode='categorical',
13     class_names=None, color_mode='rgb', batch_size=BATCH_SIZE,
14     image_size=(IMAGE_HEIGHT, IMAGE_WIDTH), shuffle=True, seed=RANDOM_SEED,
15     validation_split=VALIDATION_SPLIT, subset='validation',
16     interpolation='bilinear', follow_links=False
17 )

```

```
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 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:

```
1 data_augmentation = ks.Sequential(  
2     [  
3         layers.RandomFlip('horizontal'),  
4         layers.RandomRotation(0.1),  
5         layers.RandomZoom(0.2),  
6         layers.RandomHeight(0.1),  
7         layers.RandomWidth(0.1)  
8     ]  
9 )
```

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
- Population
- Fitness Function

- Selection Algorithm
- Crossover Algorithm
- Mutation Algorithm
- Elitism

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

2.6.1 Genotype

The *genotype* is a collection of genes that are grouped into chromosomes. 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.

2.6.2 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. Since for each individual we train a model and evaluate its performance, we decided to use only 20 individuals per generation to limit the computation.

2.6.3 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. 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.

2.6.4 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. The selection algorithm used by us 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.

2.6.5 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. There exists different 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)¹².

¹¹<https://deap.readthedocs.io/en/master/>

¹²<https://content.wolfram.com/uploads/sites/13/2018/02/09-2-2.pdf>

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.

2.6.6 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. 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.

2.6.7 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, in our case 5) 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.

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 using a Genetic Algorithm(see Section 2.6)

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.

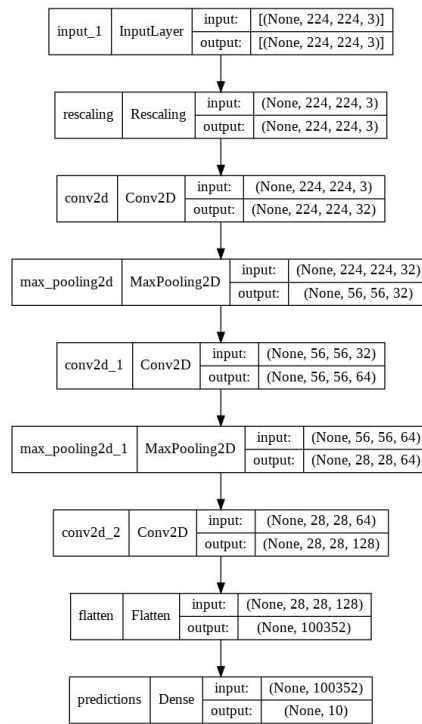


Figure 2: Customized Standard CNN Architecture

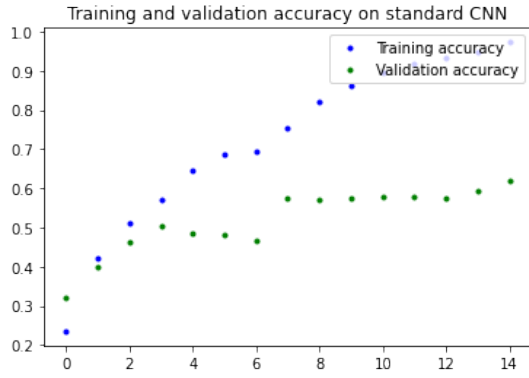
This model is trained using these default hyperparameters:

- optimizer: *ADAM*
- dropout rate :0.0
- learning rate: optimizer's default
- batch size: 128
- learning rate decay: none

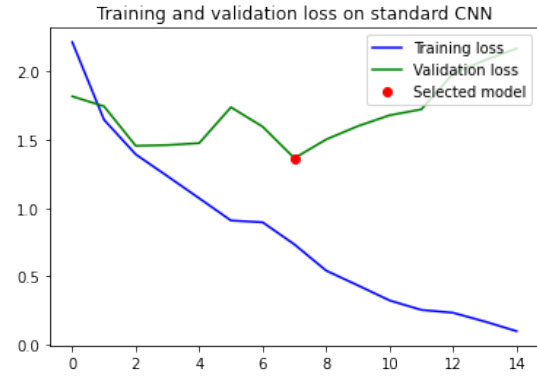
In particular, we set a large value for batch size both because our main goal is to maximize the accuracy and also (as presented in the Introduction) to face off with the great variability of paintings with very dissimilar style but belonging to the same author. In this way, we increase the probability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

The results obtained are the following:

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



(a) Standard CNN Accuracy



(b) Standard CNN Loss

The network reaches an accuracy value equal to 57.3% in just 8 epochs, then it 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

4 — Pre-Trained Models

This section describes the results obtained using different pre-trained architecture and strategies¹³. The pre-trained networks here tested are:

- VGG16
- ResNet50V2
- ResNet101V2
- InceptionV3

4.1 VGG16

VGG16⁴ 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.

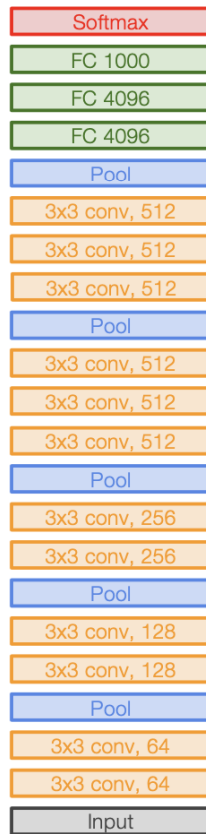


Figure 4: VGG16 Architecture

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

¹³The 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 neurons⁵.

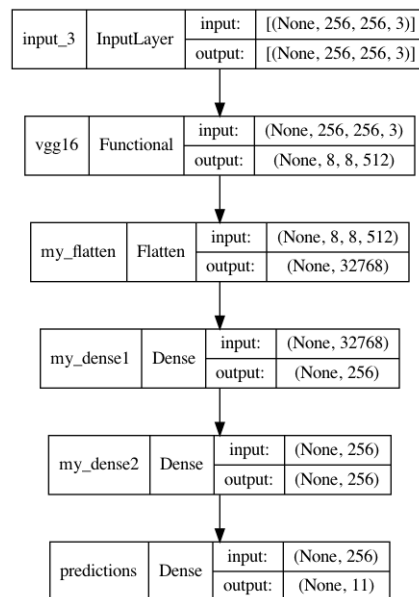
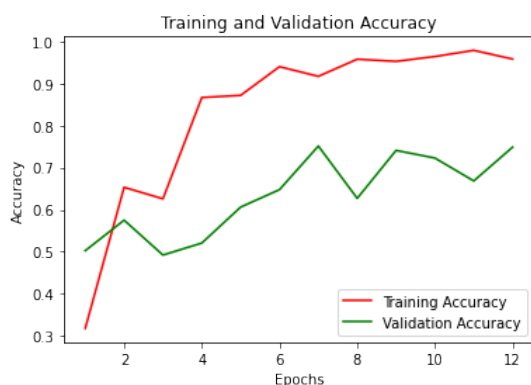


Figure 5: Our Feature Extraction Network

The result obtained, using RMSprop as optimizer, are:

Feature Extraction				
Epoch stopped	Validation Accuracy	Testing Accuracy	Validation Loss	Testing Loss
12	0.7409	0.7057	5.1959	5.7

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 Figure7. 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.

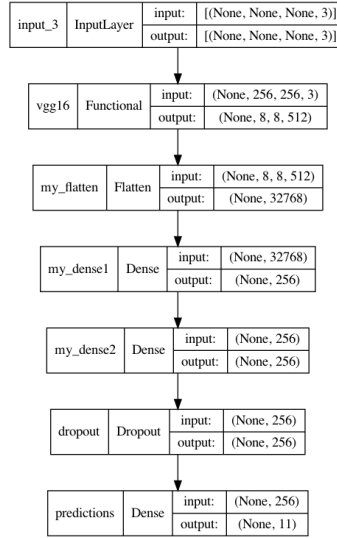
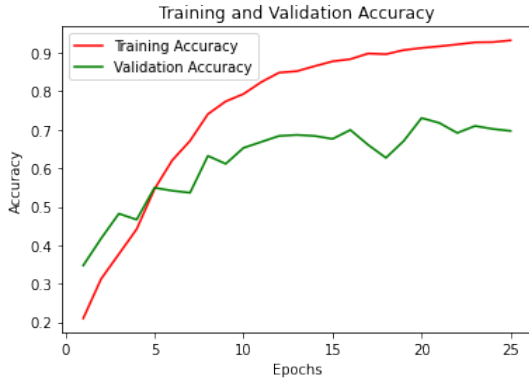


Figure 7: Our Feature Extraction Network + Dropout

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



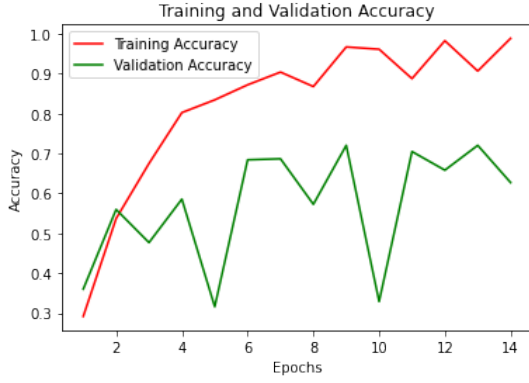
(b) Test 2 Loss

As expected, dropout mitigated the magnitude of overfitting, however our network perform slightly worst (now the validation accuracy is 0.73 and before was 0.74) than without the dropout layer.

4.1.3 Test 3: Finetuning One Convolutional Layer

Using the model defined in test 1, the 3rd Conv2D layer in the 5th block is un-frozen 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



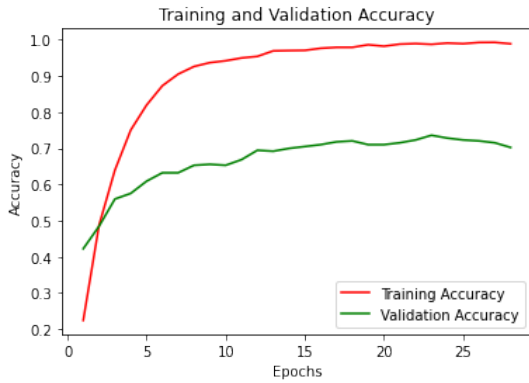
(b) Test 3 Loss

Again the performance are worst than in test one, but 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.

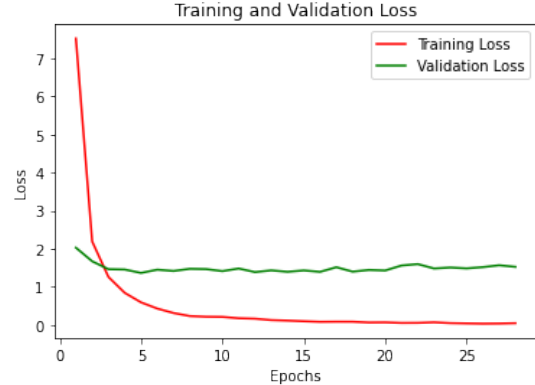
4.1.4 Test 4: test 3 with dropout and different optimizer

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

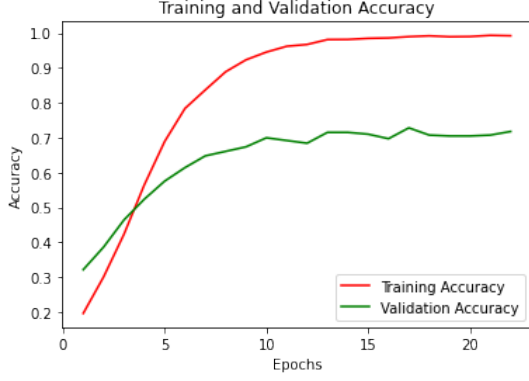
Adding dropout and a different optimizer with a little learning rate, we finally obtained what we were aiming. Anyway, the result is now comparable to the test 1, however now we are using a more complex network which is not good if the results are the same.

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.

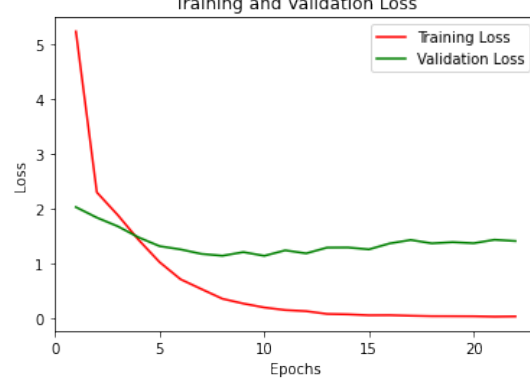
4.1.5 Test 5: Finetuning Two Convolutional Layers

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



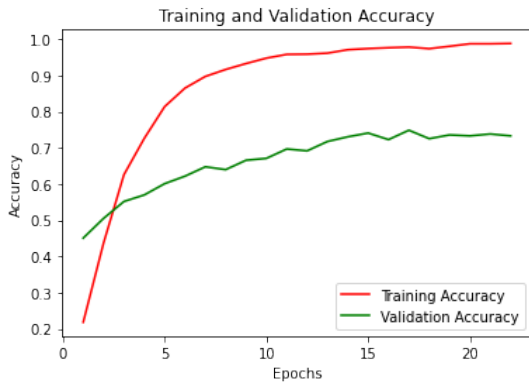
(b) Test 5 Loss

Following the approach used in the previous test we obtained also here more smoothed graphs, but the performance is little lower than before. 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.

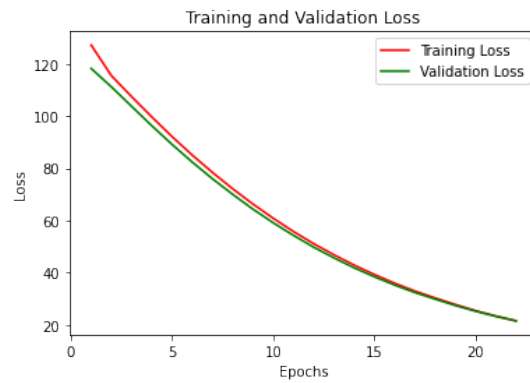
4.1.6 Test 6: Finetuning One Convolutional Layer and Weights Regularization

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

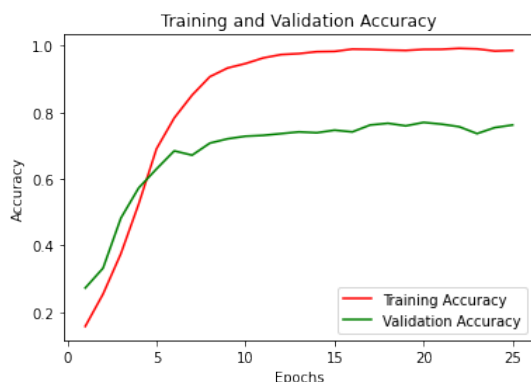
The shapes of the graphs are more or less the same of the two had in test 4, the only things that changes are the values of the losses which are here higher and more curvilinear due to the weight regularization approach used. Anyway, this heavy regularization helped us to surpass the result in test 1, not by much but it is an improvement that, maybe, can be exploited

finetuning more. Following this lead, in the next paragraph we use weight regularization with two convolutional layers finetuned.

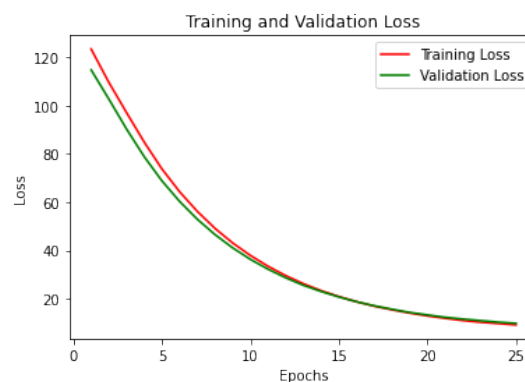
4.1.7 Test 7: Finetuning Two Convolutional Layers and Weights Regularization

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

4.2 ResNet50V2

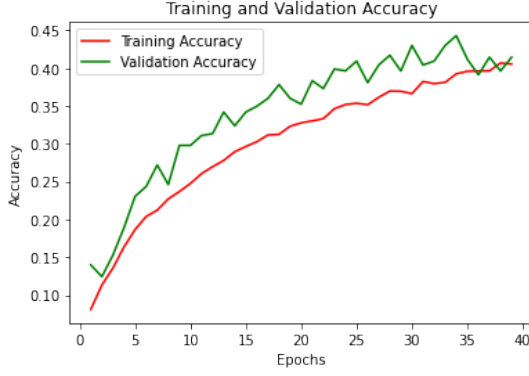
ResNet stands for Residual Network. It is an innovative neural network that was first introduced by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in their 2015 computer vision research paper titled "Deep Residual Learning for Image Recognition". This model was immensely successful, as can be 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)

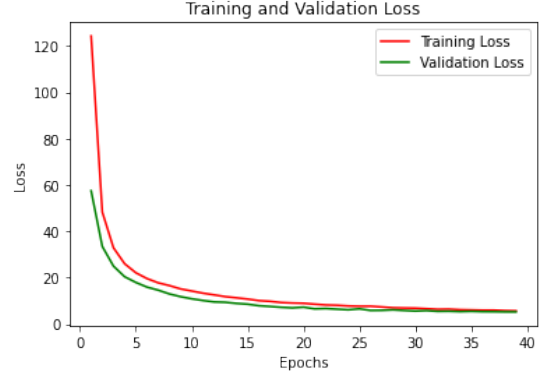
The original ResNet50 comes with a GlobalAveragePooling2D and a prediction layer soon after. In this test we used the same approach, resizing the prediction layer to the number of classes we have.

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

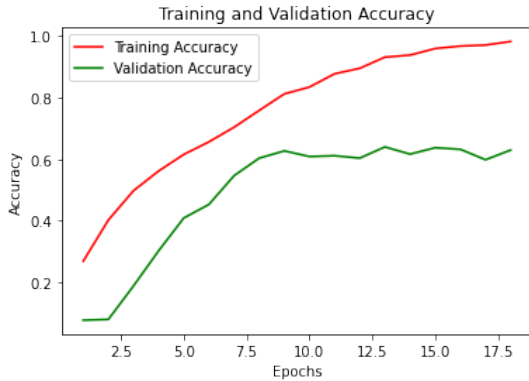
The network poorly performed compared with the results obtained using VGG16. This can be explained since the *imagenet dataset* is not specialized in distinguish paintings' artists, thus the frozen weights of our networks are very far to be the ones we need. To overcome this problem we can fine tune or add more layers after the *GlobalAveragePooling2D* layer.

Another problem can be noticed 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 worst accuracy.

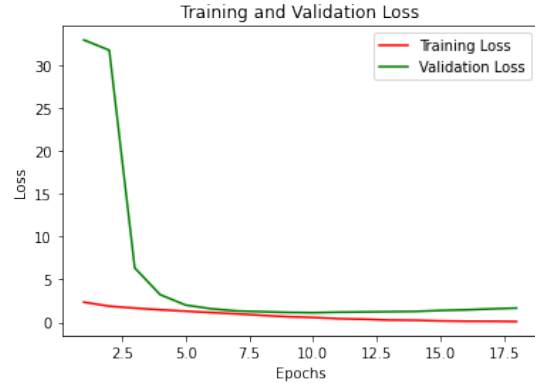
4.2.2 Test 2: Finetuning 1 block

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 input¹⁴. 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

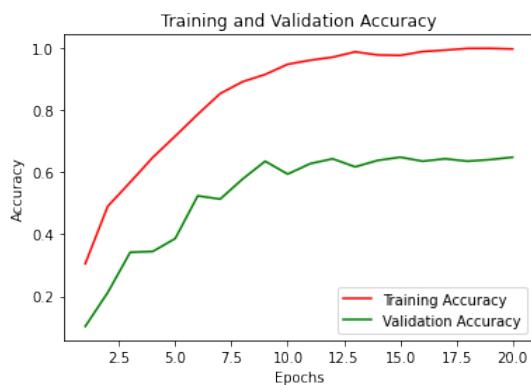
¹⁴Sometimes the residual is downsampled using a max pooling layer, this is done in order to match the actual size of the feature map obtained at that level of the network.

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

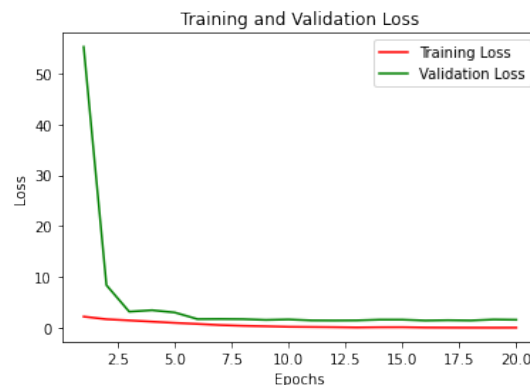
4.2.3 Test 3: Finetuning 2 blocks

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



(a) ResNet50V2 Test 3 Accuracy



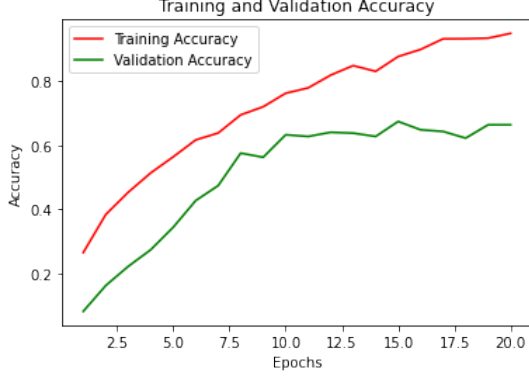
(b) ResNet50V2 Test 3 Loss

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 improved (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.

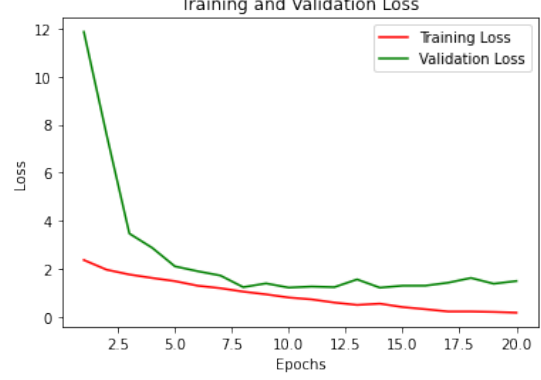
4.2.4 Test 4: Finetuning with One Block and Adding Two Dense layers

Here we took the same approach of test 3, but adding two dense layers followed by a dropout layer (as used in figure7).

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



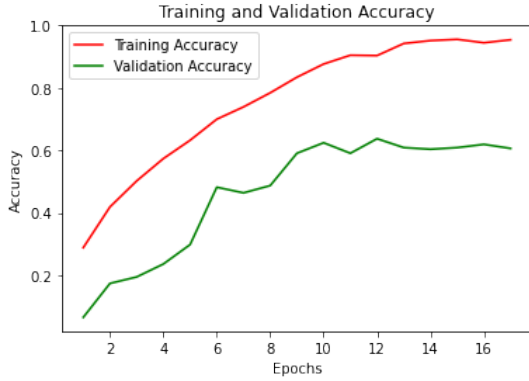
(b) ResNet50V2 Test 4 Loss

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 *GlobalAveragePooling2D*, keeping the problem for more time to a higher dimensionally than the one used for prediction.

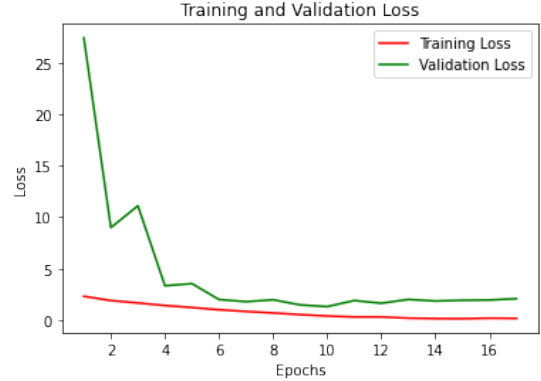
4.2.5 Test 5: Finetuning with Two Blocks and Adding Two Dense layers

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 this 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 blocks¹⁵. ResNet101

¹⁵Referring to the notation used before, we intend as a sub-block a block between two adds.

proved to be slightly more accurate than ResNet50.¹⁶

4.3.1 Test 1: Classical ResNet101V2

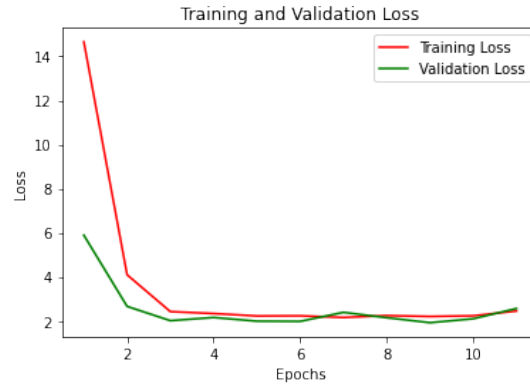
The original ResNet50 comes with a GlobalAveragePooling2D and a prediction layer soon after. In this test we used the same approach, resizing the prediction layer to the number of classes we have.

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



(a) ResNet101V2 Feature Extraction Accuracy



(b) ResNet101V2 T Feature Extraction Loss

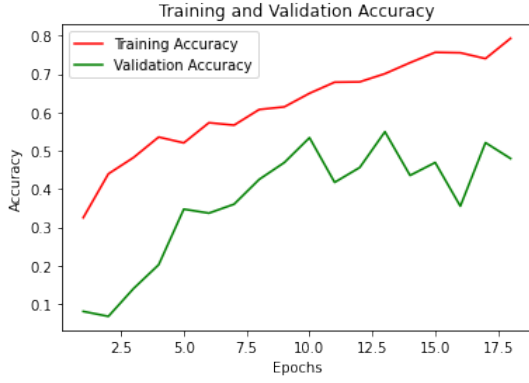
The network under-fitted our training, performing drastically poorly. The same considerations done in chapter 4.2.1 apply here.

4.3.2 Test 2: Finetuning One Sub-Block

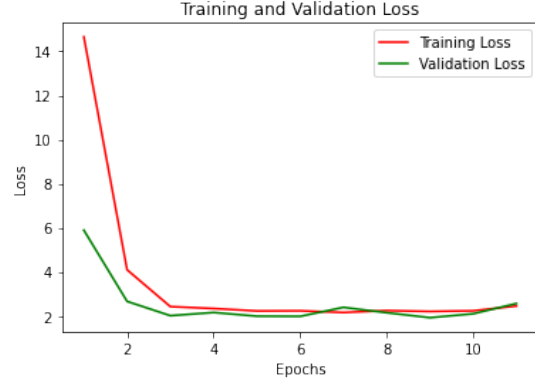
As done for ResNet50, we tried to improve our ResNet101 performance using finetuning. In this paragraph we just finetuned the last sub-block of the network.

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

¹⁶It 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 its size we decided to limit our studies to the 101-layers version.



(a) ResNet101V2 Test 2 Accuracy



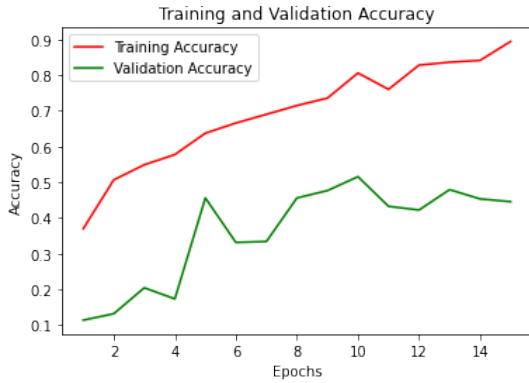
(b) ResNet101V2 Test 2 Loss

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

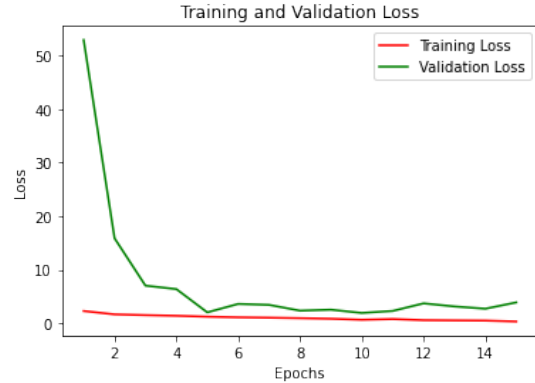
4.3.3 Test 3: Finetuning the Entire Block 5

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



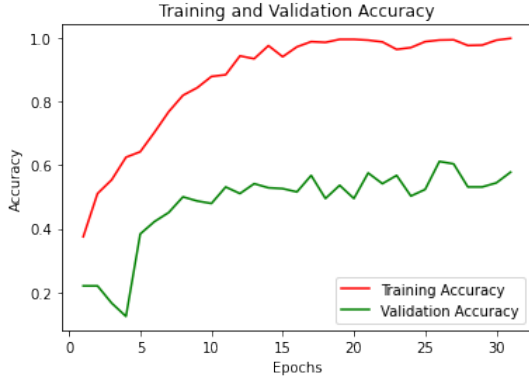
(b) ResNet101V2 Test 3 Loss

Unfortunately, we have little less performance and we still not completely overfitting: 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.

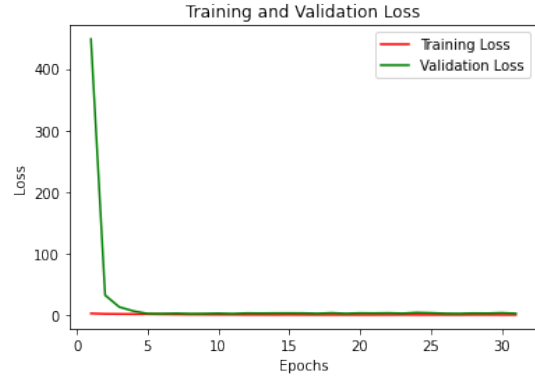
4.3.4 Test 4: Finetuning Half Block 4

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 our topic-specific. In fact, the neural network finally overfit properly our training data as the accuracy plot shows, also even the loss is lesser 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.

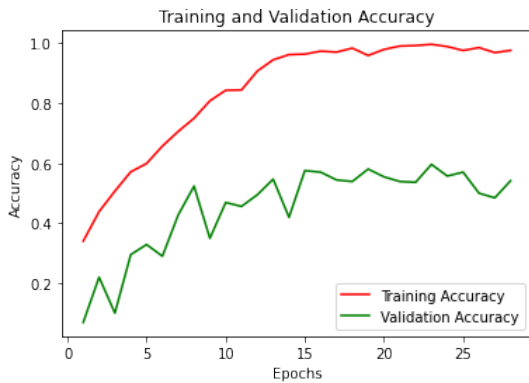
4.3.5 Test 5: Test 4 and Dropout

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:

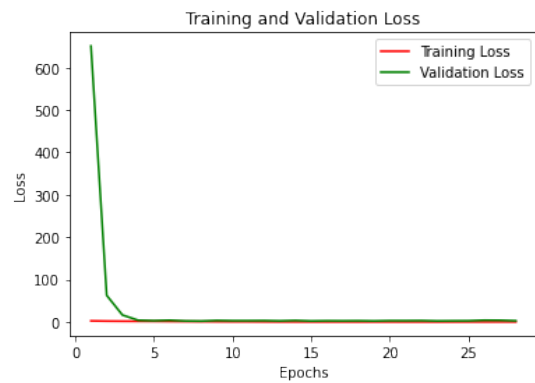
1. We perform more or less the same as in test 4: this is the goal of dropout, neutralize neurons but performing the same as before, reducing in this way the number of computations and performing better on the test set.
2. We perform worst than test 4: dropout neutralizes too many units, performing worst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

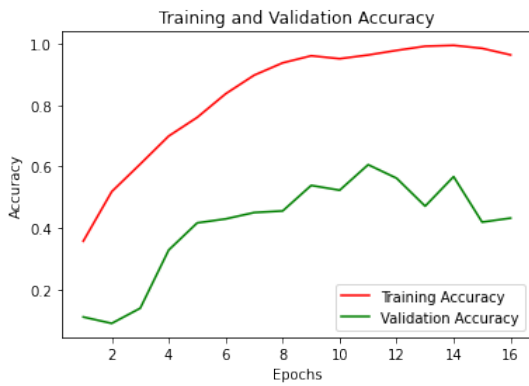
points with the one obtained in test 4, which is a lot considering that the overall performance is not good (it barely arrives to 0.6, which is very low to be considered a good classifier). Hence, for the following test we decided to drop the use of dropout until better performances are met.

4.3.6 Test 6: Adding Dense Layers to Test 4

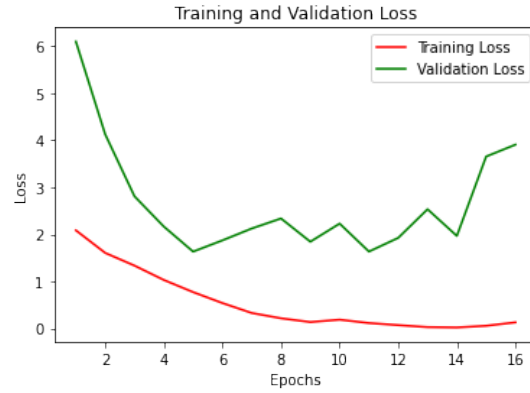
Following the tests performed with VGG16 and ResNet50V2, we tried also here to add dense layers at the top of the *GlobalAveragePooling2D* layer. Doing this experiment, the learning rate of *Adam* is change from the default value (i.e, 0.001) to 0.0001.

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



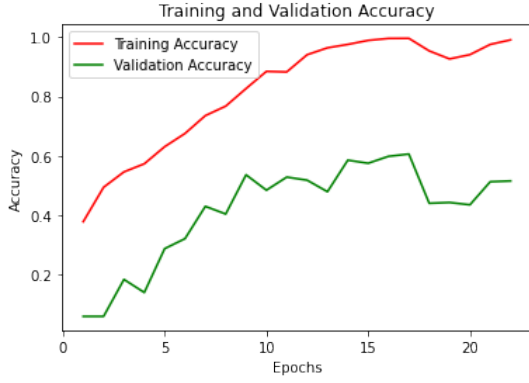
(b) ResNet101V2 Test 6 Loss

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.

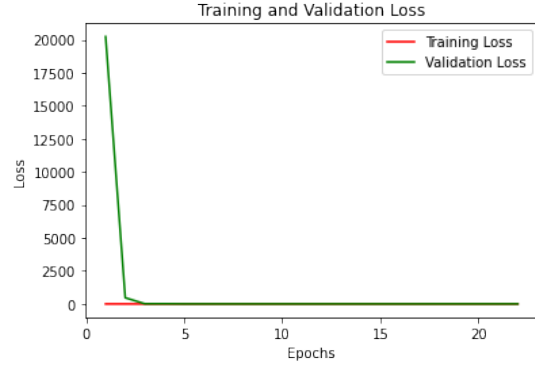
4.3.7 Test 7: Going Deeper and Deeper

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 has 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:

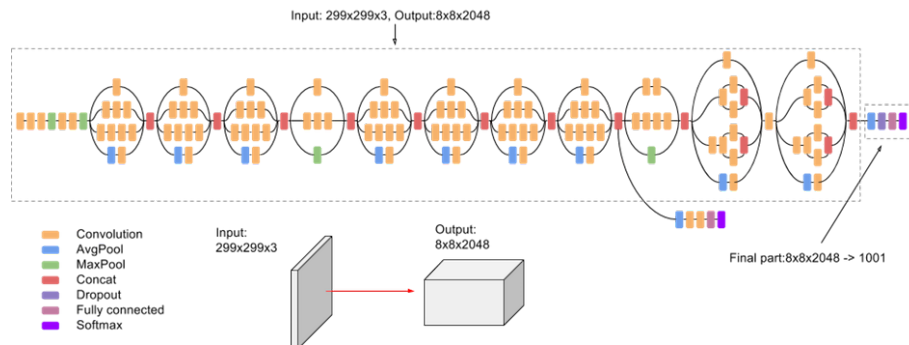


Figure 26: Inception v3 First Proposal

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

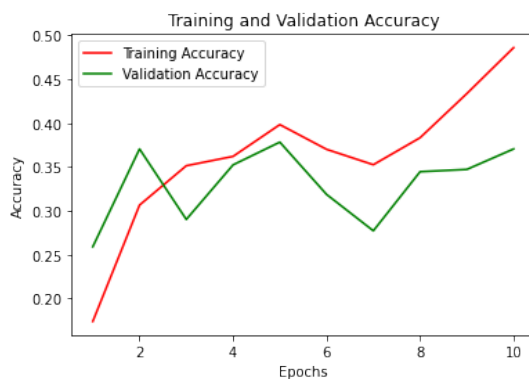
model depicted has a more complex output and it make use of an *auxiliary classifier*: training using loss at the end of the network didn't work well since the network is too deep and the gradients don't propagate cleanly. As a hack, the idea of the time was to attach "auxiliary classifiers" at several 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

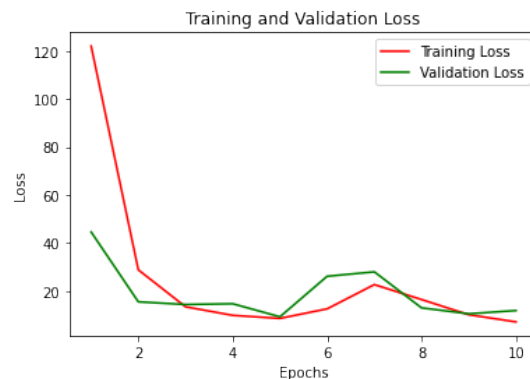
The base InceptionV3 provided by Keras comes with a GlobalAveragePooling2D and a prediction layer after that. In this test we used the same approach, resizing the prediction layer to the number of classes we have.

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

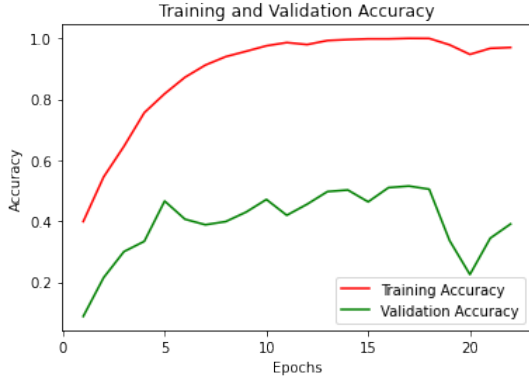
The network under-fitted our training, performing drastically poorly. The same considerations done in chapter 4.2.1 apply here.

4.4.2 Test 2: Finetuning 1 Block

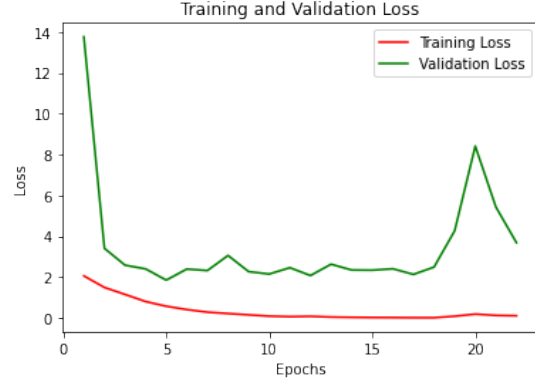
As done for the previous networks, we finetuned one block at a time, where a block is an inception block.

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

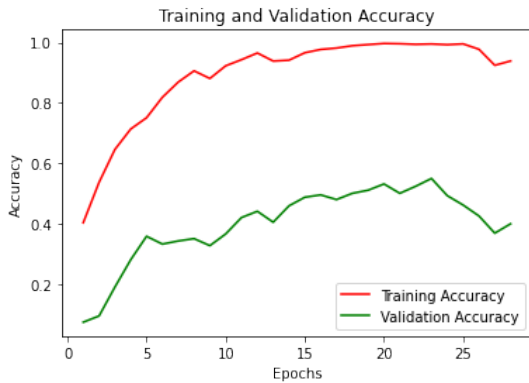
As expected, the network finally overfitted and the accuracy of the validation set increased as well. Anyway, the curves in the last epochs have a strange behavior, the validation accuracy and loss suddenly go worst. Perhaps, this is due to the fact that we are overfitting by few epochs at that point 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

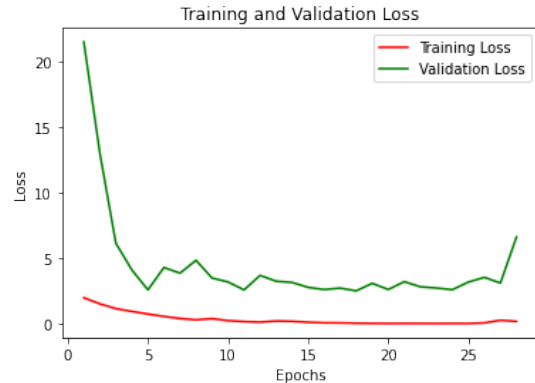
Since we want to find first a good number of blocks to set trainable and than optimize it, we continued our test finetuning the last 2 blocks.

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 0.3 points and the test accuracy is less than 0.3 points. 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

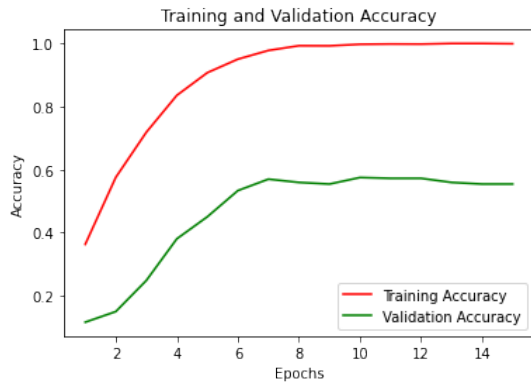
strange behavior.

4.4.4 Test 4: Finetuning 3 Blocks

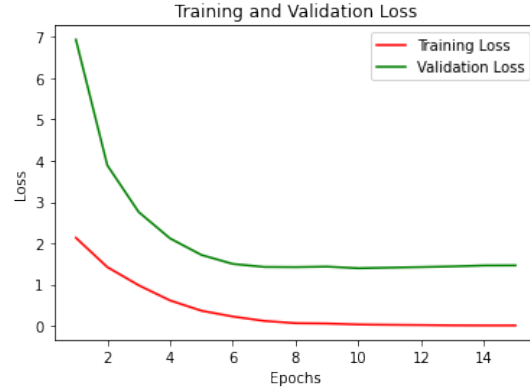
Finetuning till the third last block was our maximum in finetuning this type of network, since finetuning more led to worst performance¹⁷.

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



(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 0.3 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.

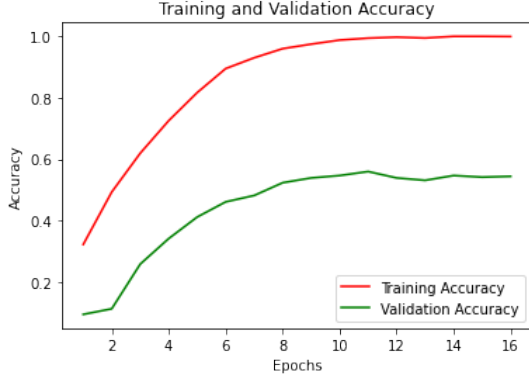
4.4.5 Test 5: Finetuning 3 Blocks with Dropout

In this test, we added a dropout layer after the *GlobalAveragePooling2D*

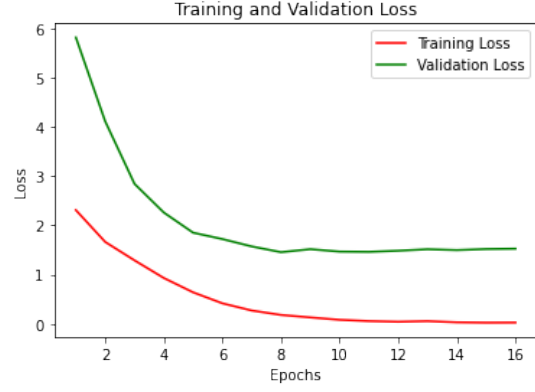
The results obtained are (learning rate=0.0001, Adam):

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

¹⁷Here not reported, but the code on GitHub provide the possibility to finetune other 2 more layers.



(a) Inception Test 5 Accuracy



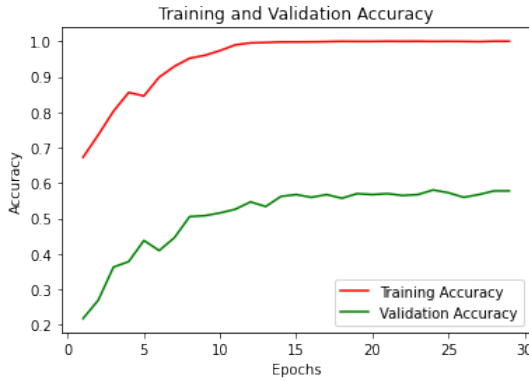
(b) Inception Test 5 Loss

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.

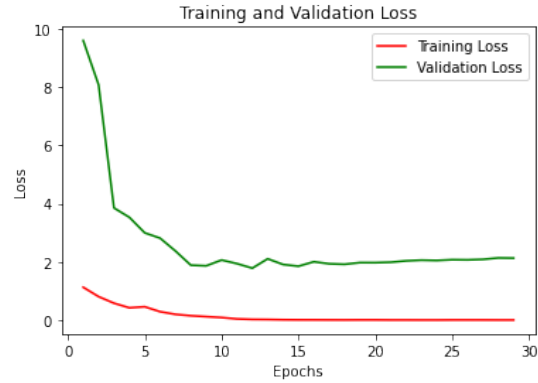
4.4.6 Test 6: Finetuning 3 Blocks and ExponentialDecay Learning Rate

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 parameters¹⁸: 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.

¹⁸For the others we used the default values

An important thing to take into consideration is the behavior of the loss, which appears here much higher than in the fourth test, this can be explained perhaps by a different direction that our system takes in the search for convergence.

5 — Siamese Architectures

5.1 Siamese Network

5.2 Convolutional Siamese Network

6 — Ensemble Network