People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University M'Hamed BOUGARA – Boumerdes

Institute of Electrical and Electronic Engineering

Department of Electronics

Final Year Project Report Presented in Partial Fulfilment of the Requirements for the Degree of

MASTER

In **Electronics**

Option: **Computer Engineering**

Title:

Algeria license plate recognition system using Faster-RCNN and YOLO models

Presented by:

- **BOUDISSA Mehieddine**
- **KISSOUM Malik**

Supervisor:

Pr. Abdelhakim KHOUAS

Registration Number:……../2020

Abstract

In some institutions, office buildings, or government facilities the flow of incoming and outgoing traffic of people and cars needs to be monitored and recorded for security purposes as well as practicality and automation of entry pass for vehicles. Over the last years, many techniques have been proposed in an attempt to solve the Automatic License Plate Recognition System (ALPRS) problem. These techniques rely mainly on hand-crafted approaches and basic computer vision algorithms such as edge detection with Sobel filter. These approaches are not accurate enough for real-world applications, nor are they robust enough to changes in size, shape, and rotation of the license plates. Recently, deep learning techniques have been shown to be a strong tool for solving computer vision and object detection problems, such as ALPRS.

In this project, we propose a solution based on convolutional neural networks (CNN). A data set containing 1000 car images has been collected, labeled, and then split into a training set and testing set. The size of this data set would allow for a transfer learning approach and fine-tuning of models. In the next step, various models belonging to the "You Only Look Once" (YOLO) CNN and "Faster Region-based CNN" (Faster RCNN) families are trained to perform plate detection task only. Once the models are trained and optimized, they are used to crop images of plates from the original car images. These cropped images are used to train models to perform the digit recognition task, similar to those trained for plate detection. The training process was repeated for different structures and parameters of the models to obtain the best performance possible.

Evaluating these models relies on the use of the mean average precision (mAP) used in the original papers of YOLO and Faster-RCNN. The evaluation of the final model (plate detection and digit recognition) relies on the accuracy of performing the identification of the license plate numbers. The end result is an application that achieved an accuracy of 81.36% with real-time video processing capabilities and robust to changes in size, shape, color, and rotation of the license plates.

This project provides users of the application with a reliable and practical security tool. It would also supply Algerian academics and software developers with a benchmark data set for further research on the topic and evaluation of future models.

Dedication

“ This work is wholeheartedly dedicated to my beloved parents Mokhtar and Fouzia, and my brothers Ahmed and Seddik and my sister Meriem for the support on all levels, who have been my source of inspiration; whose affection, love, encouragement and prays of day and night make me able to get such success and honor. I also extend my gratitude and congratulations to my partner Malik for the good work that he did despite facing hard personal challenges as well as my friends, relatives, mentor, teachers, classmates who have been all along. ”

Mehieddine

“ I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Mohammed and Ouiza whose words of encouragement and push for tenacity ring in my ears. My sisters Dihia, and Lydia have never left my side and are very special. I also dedicate this dissertation to my many friends and classmates who have supported me throughout the process. I will always appreciate all they have done, especially Djamel Eddine Zidane and AbdelMalek Bouaoune who reminded me, all along, with the importance of our religion and kept motivating and pushing me to read Quran and never neglect it whatever how desperate the situation may seem. I dedicate this work and give special thanks to my partner Mehieddine for being there for me throughout my struggle and kept faith with me. ”

Malik

Acknowledgments

In the name of Allah, the Most Beneficent and the Most Merciful, all gratitude goes to Him before and after. For the strength and the preservation from all sickness. Secondly, we would like to express our special thanks of gratitude to our teacher and supervisor Pr. Khouas who gave us the opportunity to fulfill this project and for his guidance and professionalism all along. We thank Dr. Andrew NG from coursera as well, who provided insights and expertise that greatly assisted the development of the project through his excellent courses and occasional tweets.

Contents

List of Figures

List of Tables

List of Abbreviations

Adam Adaptive moment estimate

ALPRS Automatic License Plate Recognition System

BCE Binary Cross-Entropy

BFLOP Billion Floating Operations

CNN Convolutional Neural Networks

COCO Common Objects in Context

CPU Central Processing Unit

CUDA Compute Unified Device Architecture

DSC Depth-wise Separable Convolution

FC Fully Connected

FPS Frames Per Second

GD Gradient Descent

GPU Graphical Processing Unit

ILSVRC ImageNet Large Scale Visual Recognition Challenge

IoU Intersection over Union

mAP mean Average Precision

MLE Maximum Likelihood Estimate

MLP Multi Layer Perceptron

PASCAL Pattern Analysis, Statistical modeling and Computational Learning

RCNN Region Convolutional Neural Networks

RGB Red Blue Green

ROI Region Of Interest

RPN Region Proposal Network

SGD Stochastic Gradient Descent

VOC Visual Object Classes

YOLO You Only Look Once

Introduction

The world nowadays is moving towards automation at an unprecedented rate. Almost every month, a breakthrough is made in the path towards creating smart computer systems with the ability to learn and make decisions of their own. This progress does not only manifest itself in theoretical work and research papers, but also in real-world applications such as the latest voice recognition software found in Amazon Alexa, or the computer vision systems used by Tesla autonomous cars. Most experts and speculators predict that soon enough, all aspects of ordinary life will be dependent upon the use of these artificially intelligent machines. Some experts are optimistic and consider this as a practical solution for a wide range of problems in various fields such as medicine, transportation, telecommunication, and even politics and economics. Other experts express fears that this technology may produce more problems than it would solve.

This work is an attempt to bring a contribution to the field, and specifically to the topic of CNN, by developing a real-world ALPRS to identify license plate numbers using image inputs of the cars. ALPRS help efficiently identify vehicle license plates without the need for major human resources. Recently, ALPRS have become more and more important. It can be used by government agencies to find cars that are involved in crime, look up if annual fees are paid, or identify persons who violate the traffic rules. Many countries (e.g. U.S., Japan, Germany, Italy, U.K, France, . . .) have successfully applied ALPRS in their traffic management. Several private operators are also benefiting from ALPRS.

Recognition tasks are mere for humans, whereas, for computers, it is an extremely tedious work since all that a computer can understand are numbers. Recent advancements in computer vision have given computers the ability to extract semantically meaningful information from images.

In this project, we take advantage of this ability to develop an ALPRS. The system is mainly split into three major stages: data collection and labeling, license plate detection, and lastly character (digit) detection. The two last stages use various deep learning techniques and are further elaborated in chapter three and four alongside with the first stage. Traditional computer vision techniques employ features chosen by humans to represent the underlying features of the image. These techniques require sophisticated human-designed models to translate raw input pixels into useful recognition responses. There are few main issues facing the development of a practical ALPRS. The first one, is

that most systems out there do not yield an accuracy that is satisfying enough for a real world useful application. The second one, is that these same systems are susceptible to changes in the shape, size, color, or rotation of the license plates. In the work done by M. Sarfraz, M.J. Ahmed and S.A. Ghazi entitled "Saudi Arabian license plate recognition system" [35], the authors relied on the template matching technique, which basically performs a sweeping operation all over the image, and comparing each region with a template of a certain character or number. This technique only works if the plates are facing forward to the camera, and the frames would have nothing in them that is similar in shape to a character. But when it comes to the state-of-the-art concerning ALPRS, the work done by INGA Astawa, I Gusti Ngurah Bagus Caturbawa, Sajayasa and Ari Dwi Suta Atmaja entitled "Detection of License Plate using Sliding Window, Histogram of Oriented Gradient, and Support Vector Machines Method" [26], shows the best results that we know of. The work relied on the sliding window segmentation technique and has shown an accuracy of 96%. This result has been achieved on a very restricted test images where all plate images were facing forward with good lighting conditions. Using a deep learning approach, these underlying human-engineered features are automatically selected by the algorithm. Both works showed practical state-of-the-art results in the tasks of object detection. This project also includes the creation of a labeled data set of license plates, which is the first one ever of this kind. This data set will allow for other students or researchers to conduct similar works and will be used as a benchmark to track advancements in ALPRS. This report is divided into four chapters. Chapter 1 describes basic concepts of deep learning and neural networks. Chapter 2 explains basic CNN concepts and how they are derived from plain neural networks. Chapter 3 presents Faster RCNN and YOLO model theory. Chapter 4 discusses the implementation and design of ALPRS using Faster RCNN and YOLO models.

Chapter 1

Feed-Forward neural networks

To understand the techniques used in this report, it is necessary to understand basic neural networks functioning. Let us begin with a classification example. Classification is the process of organizing data into groups or categories. Practical examples are spam filters that assign a given email to the "spam" of "non-spam" class, or assigning a diagnosis to a a given patient based on observed characteristics (sex, blood pressure, presence or absence of certain symptoms, ...). The observed characteristics are called **features**, and the classes or categories are referred to as **labels**. The collection of the features and corresponding labels is called **training set**. A sample from the training set is called **an example**. Given a scenario with a training set of labeled data $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ denotes the training example composed of multiple features, say $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, and $\mathbf{y}$ the corresponding label, a 'perceptron' is a basic computational function that can be used to perform classification. Perceptrons are the building blocks of neural networks, and the best way to get familiar with them is with an example. Assume that at at some university's admission office the students are evaluated with two pieces of information, the results of an entrance test, and their grades during their previous school years. Observe a 2D plot of some sample students' information, see fig. 1.1, where the abscissa represents the grades at the entrance test, and the ordinate represents their previous average grade. The blue dots represent the students who got accepted and the red dots represent those who got rejected. Our job is to develop a model that can separate the two groups and predict which students will be accepted and which ones will be rejected.

The two groups of points on the figure can be easily separated with a line, where most students above the line get accepted and most students under the line get rejected, see fig. 1.2. Therefore, this line can be our model for accepting and rejecting students.
The model makes a couple of mistakes since there are a few blue points that are under the line and few over the line, but they are considered as noise (they might have been accepted or rejected by mistake) and add no new information to our model. The problem is how to find the best line that does the separation.

We start by labeling the axis $\mathbf{x} = \{x_1, x_2\}$, where x_1 is TEST axis and x_2 is GRADES

Figure 1.1: *Test vs grades [1].*

Figure 1.2: *Boundary line separating accepted students represented with blue points and rejected students represented with red points [1].*

axis. The boundary line separating the students has a linear equation specifically: $2x_1 + x_2 - 18 = 0$. Passing the grades in the equation gives rise to a score, if the score is positive (the student gets plotted above the line), the student gets accepted. On the other hand, if the score is negative (the student is plotted under the line) the student gets rejected. This is called a prediction.
In a more general case, our boundary will be an equation of the following form:

$$w_1 x_1 + w_2 x_2 + b = 0. \tag{1.1}$$

Abbreviating this equation into vector notation:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{1.2}$$

Where $\mathbf{w} = \{w_1, w_2\}$. We refer to $\mathbf{x}$ as the input, $\mathbf{w}$ as the weights and b as the bias. Here $\mathbf{y} = \{0, 1\}$ is the label, where 0 indicates the student being rejected whereas 1 indicates the student being accepted. Finally, our prediction is going to be called $\hat{y}$ and it will be what the algorithm predicts; what the label will be, namely:

$$\hat{\mathbf{y}} = \begin{cases} 1 & \mathbf{w} \cdot \mathbf{x} + b \geq 0 \\ 0 & \mathbf{w} \cdot \mathbf{x} + b < 0 \end{cases} \tag{1.3}$$

The goal of the algorithm is to have $\hat{y}$ resembling to y as closely as possible. Reorganizing the equations in a graph and generalizing for n features, gives rise to fig. 1.3. The bias is considered a dummy input with value 1 to the Perceptron with weight w.

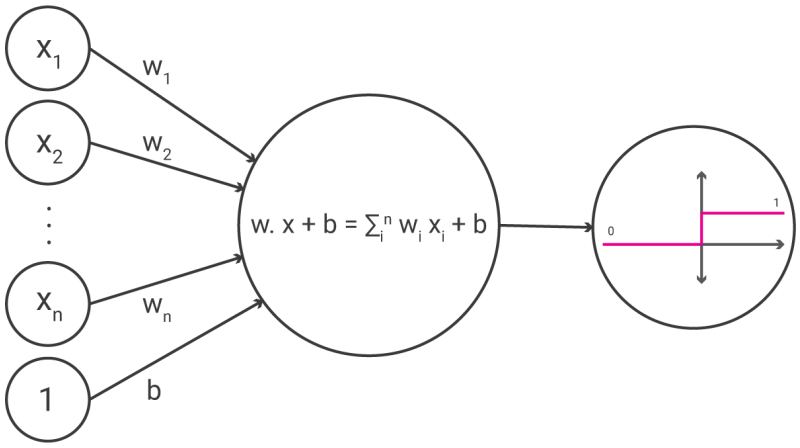


Figure 1.3: *Perceptron graph representation.*

1.1 Cost function

In order to estimate the accuracy of the algorithm, or otherwise stated, determine how well a certain prediction, given by the algorithm is, we may establish a cost function, which

measures the error that the algorithm makes on some prediction (the cost function is often referred to as the error function or the loss function). There are more than one choice for such a function. Equation (1.4) can be used, it is called "The Mean Squared Error".

$$L(\mathbf{w},\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2. \tag{1.4}$$

This function, becomes large when our algorithm approximates y badly, and small when the approximation is accurate. Additionally, notice that if we set $L_{x_i} = \frac{1}{2}(y_i - \hat{y}_i)^2$ we have that:

$$L(\mathbf{w},\mathbf{b}) = \sum_{i=1}^{n} L_{x_i}. \tag{1.5}$$

This property will be important in the algorithm described in section 1.4.

Another way of defining a cost function is using the the maximum likelihood estimation (MLE) technique, since the sigmoid function deals with probabilities. We take the joint probability of an entire training set, assuming the training examples being independent events:

$$L(\mathbf{w},\mathbf{b}) = p(y^{(1)},y^{(2)},\ldots,y^{(n)}|x^{(1)},\ldots,x^{(n)}) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)}) \tag{1.6}$$

where $x^{(i)}$, $y^{(i)}$ represent the i^{th} training example and label respectively. Thus by maximizing the joint probability, or respectively minimizing the $(-\log)$ of the likelihood, we can get an estimate of the parameters $\mathbf{w}$ *and* $\mathbf{b}$. Now, in our worked example, the neural network can be treated as a random variable having a Bernoulli distribution, therefore eq. (1.6) can be rewritten as follows [25]:

$$L(\mathbf{w},\mathbf{b}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i). \tag{1.7}$$

Equation (1.7) is usually the cost function used for Bernoulli distributed labeled data. It is often referred to as **binary cross-entropy** (BCE). For multi-class classification (predicting multiple classes, say k classes), a similar idea can be used considering a multinoulli distribution on the data set where $p(y|\mathbf{x}) = \prod_{i=1}^{k} p_i^{[y=i]}$, where $[y=i]$ evaluates to 1 if $x = i$, 0 otherwise. This leads to following cost function using MLE

$$L(\mathbf{w},\mathbf{b}) = -\sum_{j=1}^{k}\sum_{i=1}^{n} y_{i,j}\log(\hat{y}_{i,j}). \tag{1.8}$$

1.2 Gradient descent

In order to minimize the cost function we rely on optimization algorithms from numerical methods as it is unpractical to minimize manually. The technique used in neural networks

is the gradient descent. The gradient of a differentiable function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ at a point $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a vector in $\mathbb{R}^n$ of the form [24]:

$$\nabla f(x) = (\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x})) \tag{1.9}$$

It is a well known result that, given a point $x \in \mathbb{R}^n$, the gradient at that point indicates the direction of steepest ascent. Given that f is differentiable at x, the vector $-\nabla f(x)$ indicates the direction of steepest descent of the function f at the point x. In order to obtain the minimum value of the function, the gradient descent strategy tells us to start at a given $x_0 \in \mathbb{R}^n$, calculate the value of $\nabla f(x_0)$, and then proceed to calculate a new point $x_1 = x_0 - \alpha \nabla f(x_0)$, where $\alpha > 0$ is called the **learning rate**. We then repeat this process, creating a sequence $\{x_i\}$ defined by our initial choice of x_0, the learning rate α, and the rule: $x_{i+1} = x_i - \alpha \nabla f(x_0)$. This sequence continues until we approach a region close to our desired minimum.

The method of gradient descent when taken continuously over infinitesimally small increments (that is, taking the limit $\alpha \to 0$) usually converges to a local minimum. However, depending on the location of the initial x_0, the local minimum achieved may not be the global minimum of the function. Furthermore, since when carrying out calculations on an unknown function we must take discrete steps (which vary in length depending on the learning rate), we are not even guaranteed a local minimum but rather may oscillate close to one, or even 'jump' past it altogether if the learning rate is too big. Still, even with these possible complications, gradient descent is a surprisingly successful method for many real life applications and is the most standard method of training for feed-forward neural networks. Given that our cost function indicates how poorly our neural network approximates a given function, by calculating the gradient of the cost function with respect to the weights and biases of the network and adjusting these parameters in the direction opposite to the gradient, we will decrease our error and therefore get closer to an adequate network (in most cases) [24].

1.2.1 Gradient calculation

Before applying the gradient descent technique, we can clearly see that the an output of 0 or 1 is problematic since the derivatives would be 0. Therefore, the gradient descent technique will not work. To remedy this, following the MLE, a Bernoulli distribution has been defined on y, therefore the neural net needs to predict $\hat{y} = p(y = 1|x) = \sigma(x)$. For this number to be a valid probability, it must lie in the interval $[0, 1]$.

A good approach would ensure the existence of a strong gradient whenever the model has the wrong answer. For consistency with the perceptron's decision rule (eq. (1.3)), a very positive linear combination of the input x has to have a probability close to 1 and vise versa (see fig. 1.4), otherwise [25]:

$$\lim_{x\to+\infty}\sigma(x)=1, \qquad \lim_{x\to-\infty}\sigma(x)=0. \tag{1.10}$$

This approach is based on the sigmoid function:

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{1.11}$$

This function is suitable for the problem at hand, namely binary classification. However, depending on the output $\hat{y}$ other functions might be used. For example if a neural network is used to predict continuous non-bounded function that takes values in the interval $]-\infty,+\infty[$, then a more clever choice is the linear activation function, which is defined as: $f(x)=x$. Another example is the multi-class classification, where a multinoulli distribution is defined over the training data. The function used is the softmax defined as

$$softmax(x_i)=\frac{e^{x_i}}{\sum_j e^{x_j}} \tag{1.12}$$

Where x_i represents a training example from class i [25]. Here the output consists of j outputs rather than a single one. See fig. 1.5 to illustrate the output layer.

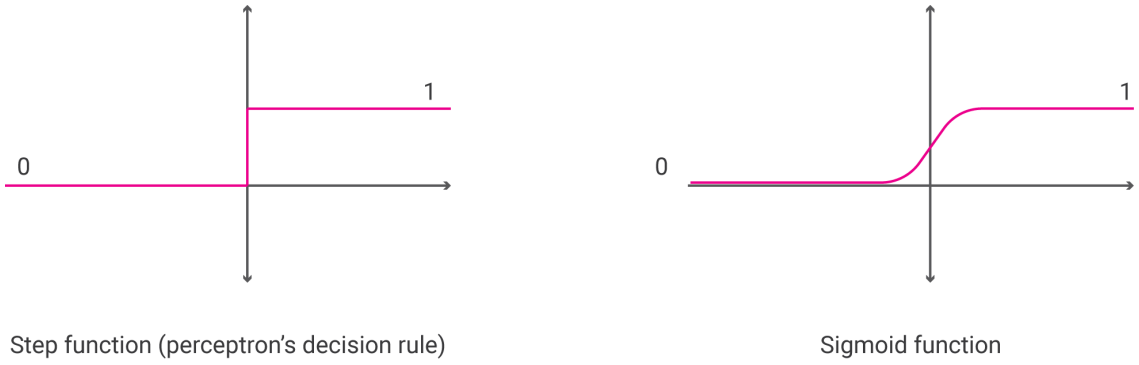


Figure 1.4: *Sigmoid vs step function.*

Now, let us apply the gradient descent technique to our network. Our goal is to calculate the gradient of L at a point $x=(x_1,\ldots,x_n)$ given by the partial derivatives, see eq. (1.9). In addition, the property mentioned in eq. (1.5) now become important: we are only going to calculate the value of ∇L_x for a given labeled data point and then add the values of the gradient together, see below.

$$\nabla L(w,b)=\nabla(\sum_{i=1}^{n}L_x)=\sum_{i=1}^{n}\nabla L_{x_i}. \tag{1.13}$$

The error produced by each point is simply: $L_x=-y\log(\hat{y})-(1-y)\log(1-\hat{y})$. In order to calculate the derivative of this error with respect to the weights, we will first calculate $\frac{\partial}{\partial w_j}\hat{y}$, where $\hat{y}=\sigma(w\cdot x+b)$ [1].

$$\frac{\partial}{\partial w_j}\hat{y} = \frac{\partial}{\partial w_j}\sigma(w\cdot x+b) \tag{1.14}$$

$$= \sigma(w\cdot x+b)(1-\sigma(w\cdot x+b))\cdot\frac{\partial}{\partial w_j}(w\cdot x+b)^1 \tag{1.15}$$

$$= \hat{y}(1-\hat{y})\cdot\frac{\partial}{\partial w_j}(w\cdot x+b) \tag{1.16}$$

$$= \hat{y}(1-\hat{y})\cdot\frac{\partial}{\partial w_j}(w_1x_1+\ldots+w_jx_j+\ldots w_nx_n+b) \tag{1.17}$$

$$= \hat{y}(1-\hat{y})\cdot x_j. \tag{1.18}$$

Now we can go ahead and calculate the derivative of the error L at a point x, with respect to the weight w_j.

$$\frac{\partial}{\partial w_j}L_x = \frac{\partial}{\partial w_j}[-y\log(\hat{y})-(1-y)\log(1-\hat{y})] \tag{1.19}$$

$$= -y\frac{\partial}{\partial w_j}\log(\hat{y})-(1-y)\frac{\partial}{\partial w_j}(1-\hat{y}) \tag{1.20}$$

$$= -y\frac{1}{\hat{y}}\cdot\frac{\partial}{\partial w_j}\hat{y}-(1-y)\frac{1}{1-\hat{y}}\cdot\frac{\partial}{\partial w_j}(1-\hat{y}) \tag{1.21}$$

$$= -y(1-\hat{y})\cdot x_j+(1-y)\hat{y}\cdot x_j \tag{1.22}$$

$$= -(y-\hat{y})x_j \tag{1.23}$$

A similar calculation will show that :

$$\frac{\partial}{\partial b}L_x = -(y-\hat{y}) \tag{1.24}$$

Therefore, since the gradient descent step simply consists in subtracting a multiple of the gradient of the error function at every point, then this updates the weights in the following way [1]:

$$w_i' = w_i+\alpha(y-\hat{y})x_i \tag{1.25}$$

$$b' = b+\alpha(y-\hat{y}) \tag{1.26}$$

[1]The sigmoid has a nice derivative: $\sigma'(x) = \sigma(x)(1-\sigma(x))$

1.3 Neural network architecture

In our work example, the target function was a simple linear function. However, in real world situations the input data is much more complex and often cannot be separated with a line. That is where neural networks shine. Neural networks also referred to as **feedforward** neural networks or **multilayer perceptron** (MLPs) are stacks of perceptrons, where each **unit** receives the input x, calculates the inner product with a set of weights and apply a non-linearity to the result, then these results are fed to a next layer of units that does the same calculations and so on. The overall length of the chain gives the **depth** of the model. The final layer of such a network is called the **output layer**, whereas the intermediate layer are referred to as **hidden layers**. The goal of the feed-forward network is to approximate some function f^*. The training example specify directly what the output layer must do at each point x; it must produce a value that is close to y. Therefore, the function computed after the linear combination is important. This function and the functions used in the hidden layers are referred to as **activation functions**. For example for a classifier, the function maps an input x to a category y, a natural choice of activation is the sigmoid; whereas in a regression problem, where the output is continuous non-bounded that takes values in the interval $]-\infty,+\infty[$, a more clever choice is the linear activation function, which is defined as: $f(x) = x$ [25]. The figure below depicts the architecture described above.

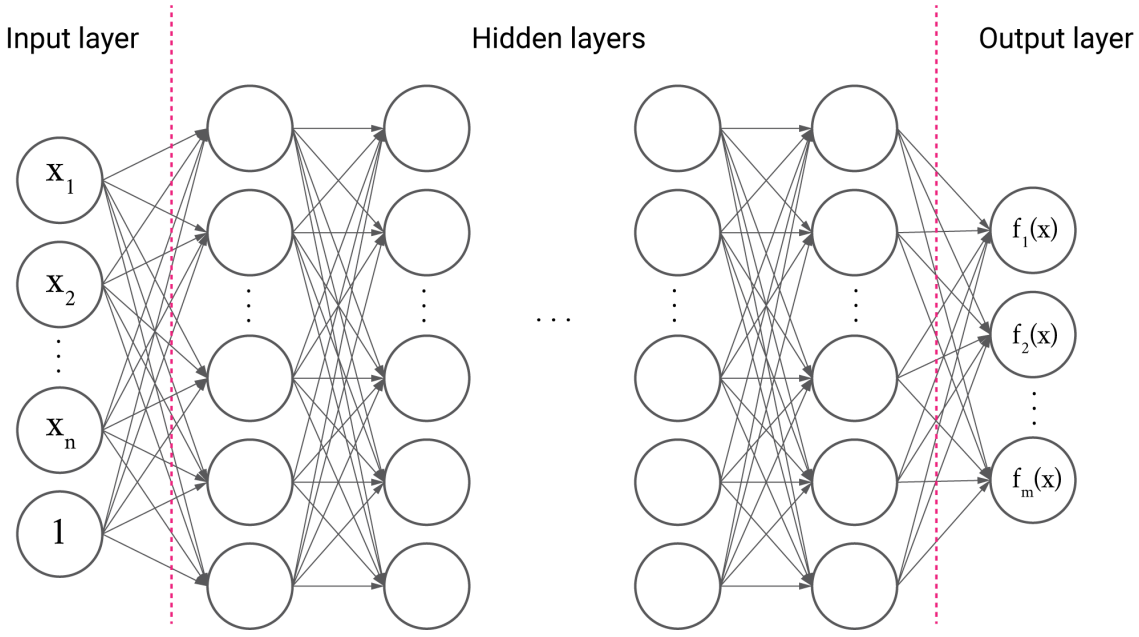


Figure 1.5: *A visual representation of a feed-forward network which approximates some function f.[24].*

For notation, set $w^n_{a,b} \in \mathbb{R}$ as the weight between the a^{th} unit in the $(n-1)^{th}$ layer with k units to the b^{th} unit in the n^{th} layer with j units.

$$w_n = \begin{pmatrix} w^n_{1,1} & \cdots & w^n_{1,k} \\ \vdots & \ddots & \vdots \\ w^n_{j,1} & \cdots & w^n_{j,k} \end{pmatrix} \tag{1.27}$$

The bias can be added as a dummy unit with input $x_{n+1} = 1$, which is a constant $b \in \mathbb{R}^j$. In order to calculate the output a_n of the n^{th} layer, we use the formula:

$$a_n = \sigma(w_n \cdot a_{n-1} + b_n). \tag{1.28}$$

In the above equation, the activation function σ is applied element-wise to each element of the resulting vector. As the computations are carried out along the network's layers, the final function f calculated by a network of depth N is [24]

$$f(x) = \sigma(w_N \sigma(\ldots \sigma(w_2 \sigma(w_1 \cdot x + b_1) + b_2)\ldots) + b_N) \tag{1.29}$$

1.4 Back-propagation

In order to train a neural network, the same techniques are used as in section 1.2.1. First we define a cost function (which is the same as in the perceptron algorithm eq. (1.7),but with a much more complex $\hat{y}$), we calculate the feed-forward pass (we calculate the output $\hat{y}$), and then calculate the the gradient of the cost function L with respect to every single weight and bias in the network, we get the following gradient vector $\nabla L = (\ldots, \frac{\partial}{\partial w^l_{i,j}} L, \ldots)$. Then applying the gradient step using eq. (1.25):

$$w'^l_{i,j} = w^l_{i,j} - \alpha \frac{\partial}{\partial w^l_{i,j}} L \tag{1.30}$$

$$b'^l_j = b_j - \alpha \frac{\partial}{\partial b^l_j} L \tag{1.31}$$

The big challenge of applying gradient descent to neural networks is calculating these partial derivatives. This is where back-propagation comes in. This algorithm first tells us how to calculate these values for the last layer of connections, and with these results then inductively goes "backwards" through the network, calculating the partial derivatives of each layer until it reaches the first layer of the network. Hence the name "back-propagation" [24].

For the purpose of this section it is useful to consider the values of each layer before the activation function step. Consider:

$$z_j^l = \sum_k w_{j,k}^l a_k^{l-1} + b_j^l \qquad \text{so that} \qquad a_j^l = \sigma(z_j^l). \tag{1.32}$$

Additionally, we define the following:

$$\delta_j^l = \frac{\partial}{\partial z_j^l} L \tag{1.33}$$

This value will be useful for propagating the algorithm backwards through the network and directly related to $\frac{\partial}{\partial w_{i,j}^l} L$ and $\frac{\partial}{\partial b_j^l} L$ by the chain rule. since we have:

$$\frac{\partial L}{\partial w_{i,j}^l} = \frac{\partial L}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{i,j}^l} = \delta_j^l a_i^{l-1} \tag{1.34}$$

$$\frac{\partial L}{\partial b_j^l} = \frac{\partial L}{\partial z_j^l} \frac{\partial z_j^l}{\partial b_j^l} = \delta_j^l. \tag{1.35}$$

The value a_j^{l-1} has already been calculated through the forward pass. The only remaining term to calculate is δ_j^l and we obtain our gradient. Our first step is calculating this value for the last layer of the network, that is, δ_j^N for a network with N layers. Since $a_j^N = \sigma(z_j^N)$, again using the chain rule:

$$\delta_j^N = \frac{\partial L}{\partial a_j^N} \frac{\partial a_j^N}{\partial z_j^N} = \frac{\partial L}{\partial a_j^N} \sigma'(z_j^N) \tag{1.36}$$

which can be easily calculated by a computer if we know how to calculate σ' (which should be true for any practical activation function). Now we will only need to "propagate" this backwards in the network in order to obtain δ_j^{N-1}. In order to do so, we apply the chain rule once again [24]:

$$\delta_j^{N-1} = \frac{\partial L}{\partial z_j^{N-1}} \tag{1.37}$$

$$= \sum_i^k \frac{\partial L}{\partial z_i^N} \frac{\partial z_i^N}{\partial z_j^{N-1}} \tag{1.38}$$

$$= \sum_i^k \delta_i^N \frac{\partial z_i^N}{\partial z_j^{N-1}} \tag{1.39}$$

If we focus on the term $\frac{\partial z_i^N}{\partial z_j^{N-1}}$, we find that:

$$\frac{\partial z_i^N}{\partial z_j^{N-1}} = \frac{\partial(\sum_k w_{i,k}^N a_k^{N-1} + b_i^N)}{\partial z_j^{N-1}} \tag{1.40}$$

$$= \frac{\partial(w_{i,j}^N \sigma(z_j^{N-1}))}{\partial z_j^{N-1}} \tag{1.41}$$

$$= w_{i,j}^N \sigma'(z_j^{N-1}) \tag{1.42}$$

which, again, can be easily calculated by a computer given the network. Therefore:

$$\delta_j^{N-1} = \sum_i^k \delta_i^L w_{i,j}^N \sigma'(z_j^{N-1}). \tag{1.43}$$

This formula tells us how to calculate any δ_j^l in the network, assuming we know δ^{l+1}. We finally have a way to calculate all the δ_j^l 's, given that we know what the values of δ_j^{l+1} are. Thus, by propagating this method backwards through the layers of the network, we are able to find all our desired partial derivatives, and can therefore calculate the value of ∇L as a function of the weights and biases of the network and execute the method of gradient descent [24].

1.5 Batch and stochastic gradient descent

Batch gradient descent is just another name for the gradient descent (GD) discussed so far. It involves calculations over the full training set to take a single step as a result of which it is very slow on very large training data due to the size of the weight matrices that take up large memory portions. Thus it become very computationally expensive to do batch GD. One can take advantage of the property mentioned in section 1.1, eq. (1.5). Therefore, instead of going through the entire data-set, at each iteration, we select a few elements from the training set, commonly selected by randomly sampling from all the available labeled data, calculate the gradient, update the network's weights and repeat the process until the network arrives at satisfactory results. The gradients computations are faster as there is much fewer data to manipulate in a single time. This technique is referred to as **stochastic** gradient descent (SGD). One downside though of SGD is, once it reaches close to the minimum value it does not settle down, instead bounces around which gives us a good value for model parameters but not optimal. This can be solved by reducing the learning rate at each step which can reduce the bouncing and SGD might settle down at global minimum after some time [13].

1.6 Adam optimizer

Adam is a an optimization algorithm and is an extension to the stochastic gradient descent. It is the most preferred optimizer within the deep learning community because it almost always work faster than SGD. The method computes individual adaptive learning rates for different parameters from estimates of first (mean) and second (variance) moments of the gradients; the name Adam is derived from adaptive moment estimation [18]. The algorithm updates exponential moving averages of the gradient (m_t) and the squared gradient (v_t) where two parameters $\beta_1, \beta_2 \in [0,1)$ control the exponential decay rates of these moving averages. The moving averages themselves are estimates of the first moment (the mean m_t) and the second raw moment (the variance v_t) of the gradient [18]. The update rule of the adam optimizer is as follow, first calculate the running average of the weights as follows (eq. (1.44)):

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot \frac{\partial L}{\partial w^l_{i,j}} \tag{1.44}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot (\frac{\partial L}{\partial w^l_{i,j}})^2 \tag{1.45}$$

Where m_0 and v_0 are initialized to zero vectors. This leads to moment estimates that are biased towards zero, especially during the initial timesteps. To counteract this bias, both m_t and v_t are divided by $(1-\beta^t)$ (eq. (1.46)) [18]. The square operation in the second equation is actually an element-wise square not the ordinary square operation.

$$\hat{m}_t = m_t/(1-\beta_1^t), \qquad \hat{v}_t = v_t/(1-\beta_2^t) \tag{1.46}$$

Finally, the weights are updated according to eq. (1.47)

$$w_t = w_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{1.47}$$

The same equations apply to the bias term b. The reason why adam is effective is because it is invariant to the scale of the gradients; rescaling the gradients with a factor of c will scale $\hat{m}_t$ with a factor c and $\hat{v}_t$ with a factor of c^2, which cancel out [18]. The other reason is that if the gradients using SGD with a high learning rate oscillates a lot in some dimensions, the adam optimizer has the effect of dumping those oscillations since it is taking the mean of the gradient in that direction, summing up positive and negative numbers making the gradients move faster towards the minimum. Also in eq. (1.47), we can notice that if $\sqrt{\hat{v}_t} + \varepsilon$ is large due to large variation of the gradients, then the term $\hat{m}_t$ will be divide by a large number making it small and this has the effect of dumping as well the oscillations. See figure fig. 1.6. The ε term is added for numerical stability in case

$\hat{v}_t = 0$ [3].

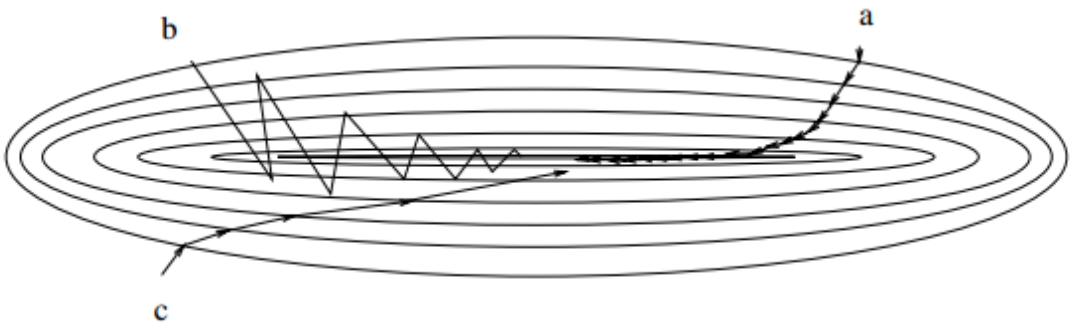


Figure 1.6: *The descent in weight space [13].*

The concentric ellipsis represents the counters of the cost function. a represents SGD steps with a small learning rate, b represents SGD with a large learning rate, and c represents Adam with a large learning rate.

1.7 Problems related to neural nets

Neural networks are extremely powerful function approximators, but during the design of the architecture care should be taken since there are many parameter one can tune (depth, number of units in each layer, . . .). Therefore, a complex design namely high number of units in each layer and a deep network can lead to **overfitting**. Over-fitting is the case where the overall cost is really small (The network is doing very well on the training set) but the generalization of the model to unseen data is poor and unreliable. There are many solutions proposed to break this effect such as *dropout* which consists of randomly zeroing the output of some units in each layer to force the algorithm to take different routes through the network. This has the effect of training smaller portions of the network, and thus smaller functions with reduced complexity are learned. It has the effect of reducing the high variance of the overall neural net. This is referred to as **regularization**.
Another famous problem neural nets suffer from is **local minimum** problem, The error surface of a complex network is full of hills and valleys. Because of the gradient descent, the network can get trapped in a local minimum when there is a much deeper minimum nearby. A suggested solution is to increase the number of hidden units. This technique works because of the higher dimensionality of the error space, making the chance to get trapped smaller [13].
Another issue in deep neural nets is the **vanishing gradients** problem. As we learned from back-propagation, each of the neural network's weights receive an update proportional to the partial derivative of the error function with respect to the current weight in each training iteration. The problem is that in some cases, the gradient will be vanishingly small, eventually preventing the weight from changing its value. In the worst case, this

may completely stop the neural network from further training. As the network trains, the weights can be adjusted to very large values. The total input of a hidden unit or output unit can therefore reach very high (either positive or negative) values, and because of the sigmoid activation function the unit will have an activation very close to zero or very close to one [13]. And since back-propagation computes gradients using the chain rule, this has the effect of multiplying N of these small numbers to compute gradients of the "front" layers in an N-layer network, meaning that the gradient decreases exponentially with N while the front layers train very slowly.

To remedy this problem other activation functions might be used in the hidden layers. The behavior of the hidden layers is not directly specified by the training data. The learning algorithm must decide how to use those layers to produce the desired output, but the training data do not say what each individual layer should do. Instead, the learning algorithm must decide how to use these layers to best implement an approximation of f. Therefore the choice of the activation function in those layers is irrelevant, which makes the use of other activation possible. Many functions have been proposed to escape the trap of vanishing gradients, namely the ReLU function is of popularity in deep learning. The ReLU stands for rectified linear unit defined as $ReLU(x) = \max(0, x)$.

Chapter 2

Convolutional Neural Networks

Convolutional neural networks (CNN) are a specialized kind of neural network for processing data that has a known grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels. CNNs have been tremendously successful in practical applications. The name "convolutional neural network" indicates that the network employs a mathematical operation called convolution. Convolution is a specialized kind of linear operation. CNNs are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers [25].

2.1 Convolution operation

The convolution operation is well known in the engineering terminology, which, in its most general form, is an operation on two functions of a real-valued argument. defined as:

$$s[n] = y[n] * x[n] = \sum_{k=-\infty}^{k=\infty} y[k]x[n-k] \tag{2.1}$$

We are interested in the discrete convolution operation, since data on a computer is presented as discrete values rather than continuous signals. The eq. (2.1) presented above is for discrete time signals.

In convolutional neural network terminology, the first argument to the convolution is often referred to as **the input**, and the second argument as **the kernel**. The output is sometimes referred as the **feature map**. The input is usually a multidimensional array of data (Red Green Blue channel (RGB) images), and the kernel is usually a multidimensional array of parameters that are adapted by the learning algorithm. These multidimensional arrays are referred to as tensors. Finally, we often use convolution over more than one axis at a time. For example if we use a two-dimensional image I as our input, we probably also

want to use a two-dimensional kernel K:

$$S[m,n] = I[m,n] * K[m,n] = \sum_i \sum_j I[i,j]K[m-i,n-j]. \tag{2.2}$$

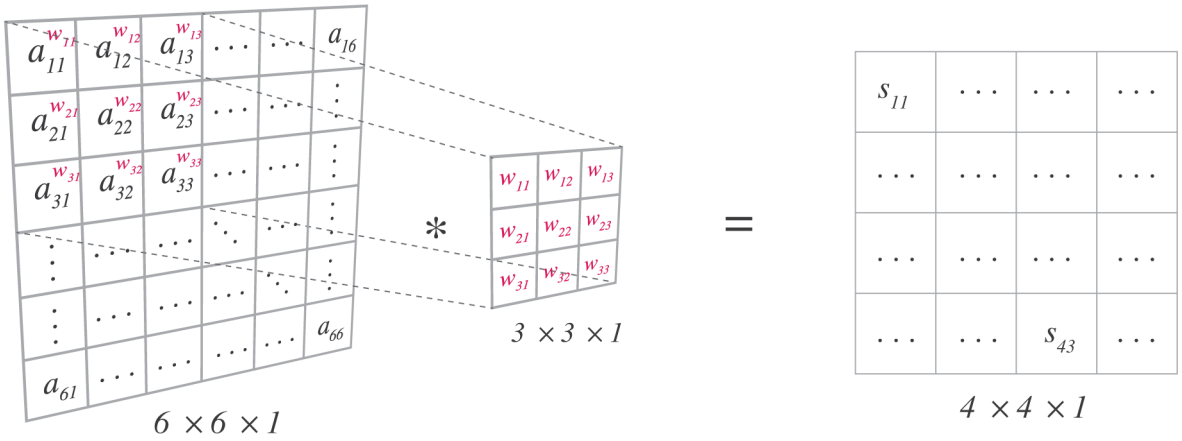
Convolution is commutative, meaning we can equivalently write:

$$S[m,n] = K[m,n] * I[m,n] = \sum_i \sum_j I[m-i,n-j]K[i,j]. \tag{2.3}$$

While the commutative property is useful for writing proofs, it is not usually an important property of a neural network implementation. Instead, many neural network libraries implement a related function called the cross-correlation, which is the same as convolution but without flipping the kernel:

$$S[m,n] = I[m,n] * K[m,n] = \sum_i \sum_j I[m+i,n+j]K[i,j]. \tag{2.4}$$

Many machine learning libraries implement cross-correlation but call it convolution. See fig. 2.1 for an example of convolution applied to a 2D tensor (gray-scale image).



$$s_{11} = a_{11}\,w_{11} + a_{12}\,w_{12} + a_{13}\,w_{13} + a_{21}\,w_{21} + a_{22}\,w_{22} + a_{23}\,w_{23} + a_{31}\,w_{31} + a_{32}\,w_{32} + a_{33}\,w_{33}$$

$$s_{43} = a_{43}\,w_{11} + a_{44}\,w_{12} + a_{45}\,w_{13} + a_{53}\,w_{21} + a_{54}\,w_{22} + a_{54}\,w_{23} + a_{63}\,w_{31} + a_{64}\,w_{32} + a_{65}\,w_{33}$$

Figure 2.1: *An example of 2-D convolution [25].*

2.2 CNN architectures

Convolution leverages three important ideas that can help improve a machine learning system: sparse connectivity, parameter sharing and equivariant representations.

Traditional neural network layers use matrix multiplication by a matrix of parameters with a separate parameter describing the interaction between each input unit and each output unit. This means that every output unit interacts with every input unit, see fig. 2.2. CNNs, however, typically have sparse interactions (also referred to as **sparse connectivity** or sparse weights). This is accomplished by making the kernel smaller than the input. For example, when processing an image, the input image might have thousands or millions of pixels, but we can detect small, meaningful features such as edges with kernels that occupy only tens or hundreds of pixels. This means that we need to store fewer parameters, which both reduces the memory requirements of the model and improves its statistical efficiency. It also means that computing the output requires fewer operations. These improvements in efficiency are usually quite large. If there are m inputs and n outputs, then matrix multiplication requires $m \times n$ parameters, and the algorithms used in practice have $O(m \times n)$ runtime (per example). If we limit the number of connections each output may have to k, then the sparsely connected approach requires only $k \times n$ parameters and $O(k \times n)$ runtime. For many practical applications, it is possible to obtain good performance on the machine learning task while keeping k several orders of magnitude smaller than m [25]. For graphical demonstrations of sparse connectivity, see fig. 2.3 and fig. 2.4. Rearranging each vector as a matrix, the relationship between the nodes in each layer are more obvious, see fig. 2.4.

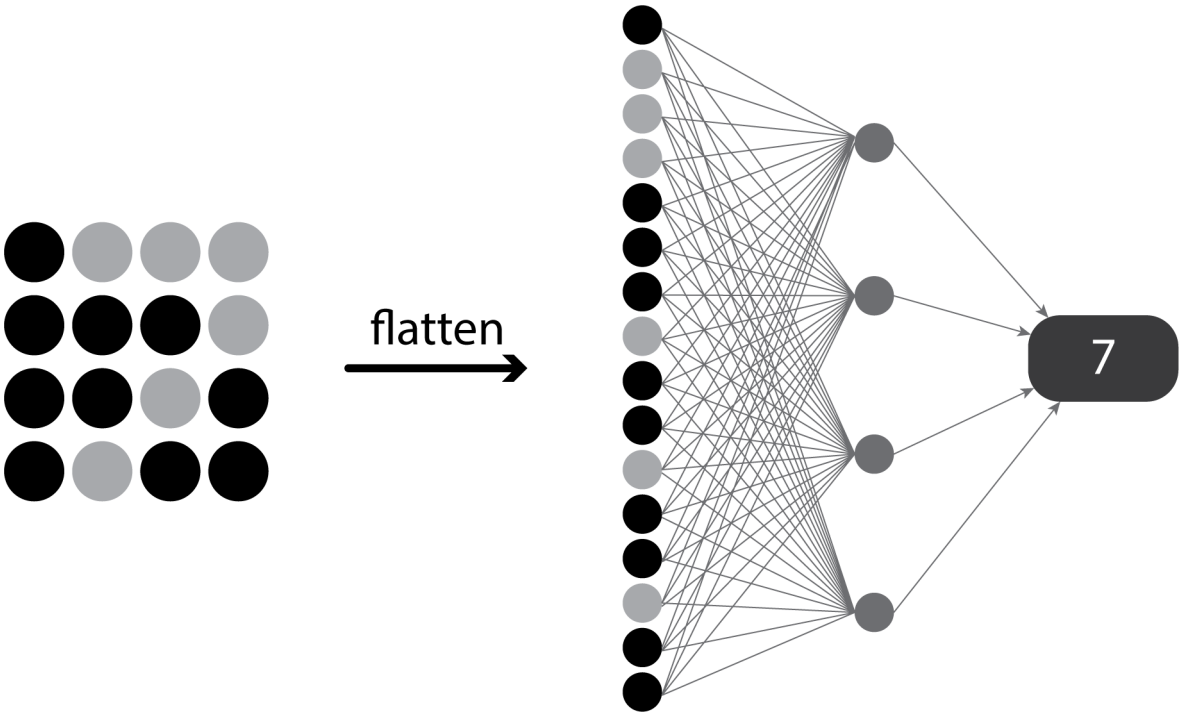


Figure 2.2: *Traditional neural network connections. The last layer has been replaced by a black box for simplicity*

Parameter sharing refers to using the same parameter for more than one function in a model. In a traditional neural net, each element of the weight matrix is used exactly once

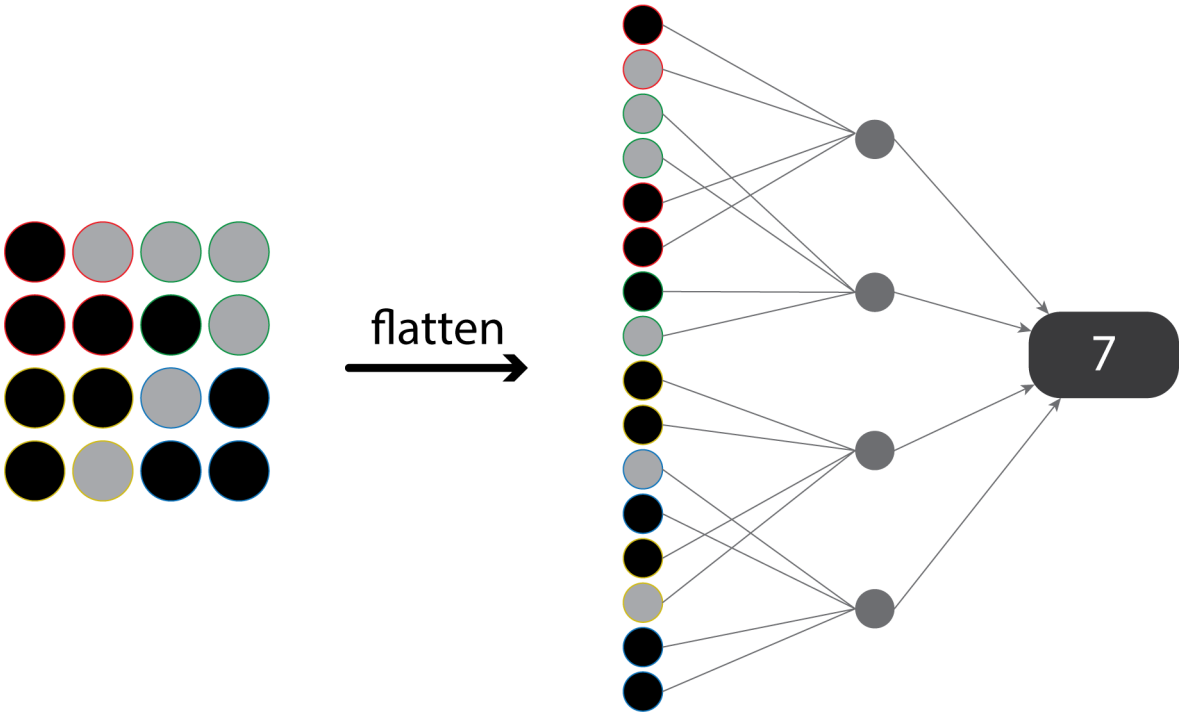


Figure 2.3: *Sparse connectivity example.*

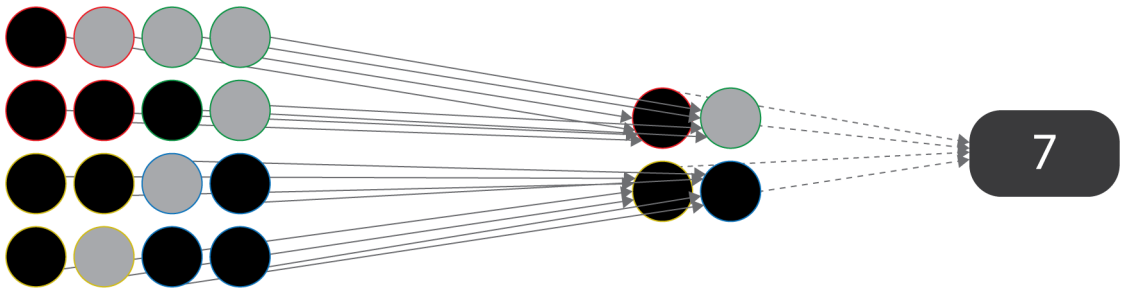


Figure 2.4: *Sparse connectivity after rearrangement.*

when computing the output of a layer. It is multiplied by one element of the input and then never revisited. As a synonym for parameter sharing, one can say that a network has tied weights, because the value of the weight applied to one input is tied to the value of a weight applied elsewhere fig. 2.2. That is the reason, traditional nets are referred as to fully connected (FC) networks or dense networks.

In a CNN, each member of the kernel is used at every position of the input (except perhaps some of the boundary pixels, depending on the design decisions regarding the boundary). The parameter sharing used by the convolution operation means that rather than learning a separate set of parameters for every location, we learn only one set. In fig. 2.4, each of the color coded image quarters are connected to a single color coded node in the next layer. All of these connections have exactly the same shared weights, see fig. 2.1, the weights w_{11} through w_{33} do not change as the filter slides through the image. This does not affect the runtime of forward propagation —it is still $O(k \times n)$— but it does further reduce the storage requirements of the model to k parameters. The particular form of parameter sharing causes the layer to have a property called **equivariance to translation**.

To say a function is equivariant means that if the input changes, the output changes in the same way. Specifically, a function $f(x)$ is equivariant to a function $g(x)$ if $f(g(x)) = g(f(x))$. In the case of convolution, if we let g be any function that translates the input, that is, shifts it, then the convolution function is equivariant to g. For example, let I be a function giving image brightness at integer coordinates. Let g be a function mapping one image function to another image function, such that $I' = g(I)$ is the image function with $I'(x,y) = I(x-1,y)$. This shifts every pixel of I one unit to the right. If we apply this transformation to I, then apply convolution, the result will be the same as if we applied convolution to I, then applied the transformation g to the output. With images, convolution creates a 2-D map of where certain features appear in the input. If we move the object in the input, its representation will move the same amount in the output. This is useful for when we know that some function of a small number of neighboring pixels is useful when applied to multiple input locations. For example, when processing images, it is useful to detect edges in the first layer of a convolutional network. The same edges appear more or less everywhere in the image, so it is practical to share parameters across the entire image.

Convolution is not naturally equivariant to some other transformations, such as changes in the scale or rotation of an image. Other mechanisms are necessary for handling these kinds of transformations (section 2.7). To illustrate these principles in action, we shall use a hand picked filter that is used to detect edges in a image, see below.

$$K = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix} \tag{2.5}$$

These filters are called high pass filters. They enhance high frequency components in an image. Frequency in images just like in signals is the rate of change of the intensity, which areas in neighboring pixels that rapidly changes for example from very dark to very light (in grayscale images). See fig. 2.5 to see the effect of applying the above filter to a grayscale image.

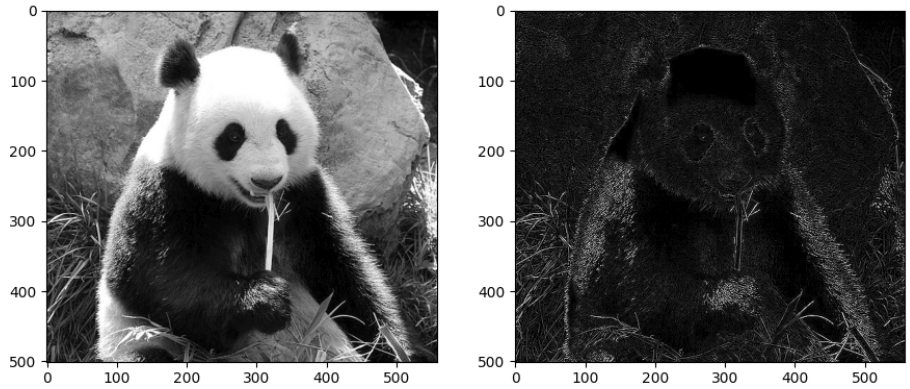


Figure 2.5: *2D convolution on a practical example.*

As we can see in fig. 2.5, where there is no change or little change of intensity in the original picture, the high pass filter block those areas out and turn the pixels black. But in the areas where a pixel is way brighter than its immediate neighbors, the high pass filter enhance the change and create a line. This has the effect of emphasizing edges. Edges are just areas in an image where the intensity changes very quickly. This images has been obtain by convolving the filter K with the image in the left, as we can see the three principles discussed above apply to this filter. The values of K didn't change while convolving (shared parameters), sparse connectivity where the filter looks only to a small portion of the image at a time, and the equivariant translation, where we clearly see that no matter the position of the edge in the image the filter successful highlight it.

2.3 Convolutional layer

The convolutional layer is produced by applying a series of many different image filters, also known as convolutional kernels, to an input image.

In the example shown in fig. 2.6, 4 different filters produce 4 differently filtered output images. When we stack these images, we form a complete convolutional layer with a depth of 4, see fig. 2.7.

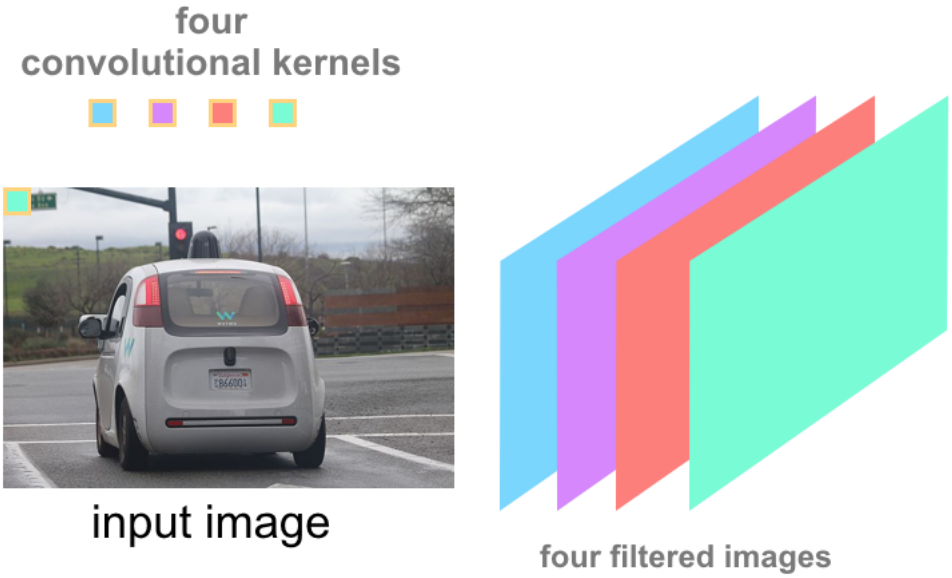


Figure 2.6: *Multiple filters for multiple pattern detection [1].*

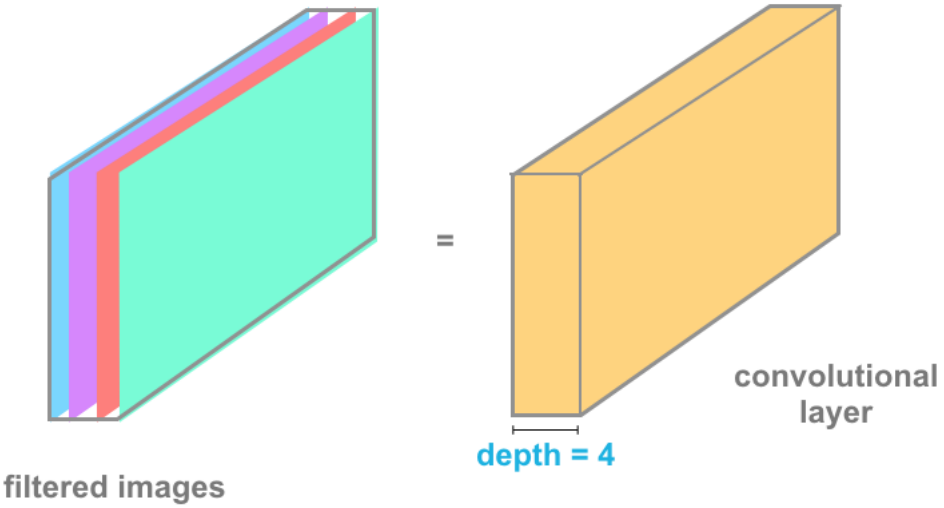


Figure 2.7: *A complete convolutional layer with 4 filters [1].*

In the case of colored images, the computer interprets them as 3-D tensors $(Height \times width \times channels)$. Here, channels are the RGB channels. When performing convolution, the kernel K is itself chosen to be a three dimensional tensor. A typical kernel K would be $3 \times 3 \times 3$. The resulting output feature map would be $(Height \times Width)$. In order to depict multiple patterns in the image, instead of having a single kernel, multiple kernels are defined. Each resulting output feature map can be considered as an image channel and when stacked together a 3 dimensional array is obtained. This 3D array can be used as input to another convolutional layer to discover patterns within the patterns that we discovered in the first convolutional layer. This operation can be repeated multiple times to discover various patterns within the input image.

In CNNs, inference works the same way as in the old plain neural networks. Both convolutional and dense layers have weights and biases that are initial randomly generated. Therefore, in the case of CNNs, where the weights take the form of convolutional kernels or filters, those kernels are randomly generated as well as the patterns that they are initially designed to detect. As with FC networks, when we construct a CNN, we will always specify a loss function. In the case of multiclass classification, this will be categorical cross-entropy loss (eq. (1.8)). Then, as we train the model through back propagation, the filters are updated at each iteration to take on values that minimizes the loss function. In other words, the CNN determines what kind of patterns it needs to detect base on the loss function.

2.4 Stride and padding

The behavior of a CNN can be controlled by specifying the number of filters and the size of each filter, these are referred to as **hyper-parameters**. For instance, to increase the number of nodes in a convolutional layer, you could increase the number of filters. To increase the size of the detected patterns, you could increase the size of the filters. But there are more hyper-parameters than we can tune. One of these hyper-parameters is referred to as the stride of the convolution. The stride is just the amount by which the filter slides over the image. In the previous example of fig. 2.1, the stride was one. We move the convolution window horizontally and vertically across the image one pixel at a time [1]. With an input image of $n \times n$, and an $f \times f$ filter, the width W and height H of the output of the convolution is given by eq. (2.6):

$$(W,H) = (n-f+1) \times (n-f+1) \tag{2.6}$$

If we introduce the stride parameter s, eq. (2.6) can be rewritten as follow:

$$(W,H) = \left(\lfloor \frac{n-f}{s} \rfloor + 1 \right) \times \left(\lfloor \frac{n-f}{s} \rfloor + 1 \right) \tag{2.7}$$

Where $\lfloor . \rfloor$ is the the floor function. It takes as input a real number, and gives as output the greatest integer less than or equal to that input. One downside of the convolution operation is the shrinking input dimensions. Indeed, according to eq. (2.6), the input dimension shrinks each time by few pixels which can be an undesirable effect in very deep networks, where the image can shrink to very small dimensions. Another downside of the convolution is, the top left pixel (or corners of an image in general) is only involved in one pass of the filter, whereas if we take a pixel in the middle, then many 2×2 regions will overlap that pixel. It is as if the pixels at the corners are used much less in the output, so information is thrown away near the edge of the image. Therefore to solve both of this problems, before applying the convolution we can pad the image with additional borders p. In fig. 2.8, $p = 1$ pixel has been used. Therefore the width and height of the output feature map is calculated as:

$$(W,H) = \left(\lfloor \frac{n-f+2p}{s} \rfloor + 1 \right) \times \left(\lfloor \frac{n-f+2p}{s} \rfloor + 1 \right) \tag{2.8}$$

Now with this additional border of zeros, the output feature maps' dimensions can be made equal to the input's dimension by setting the appropriate padding value. And the corner pixels contribute more in the output feature map.

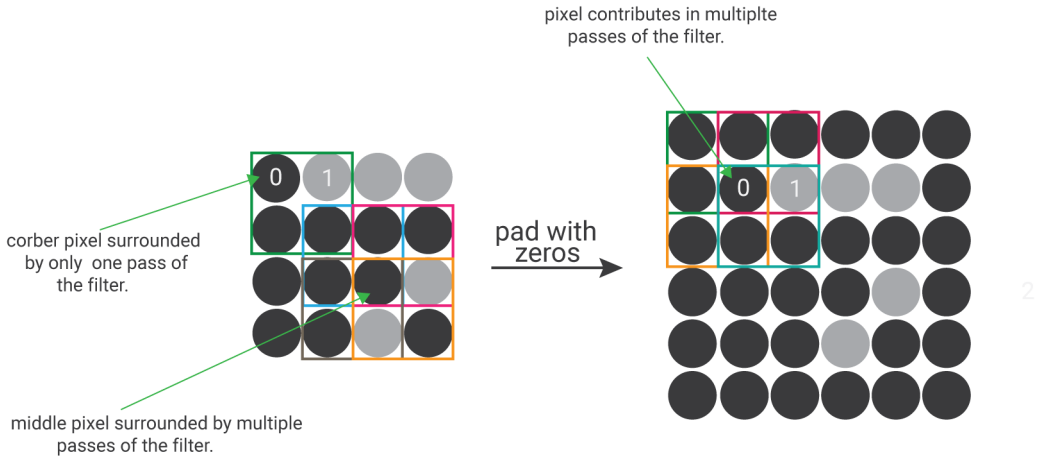


Figure 2.8: *Padding example.*

2.5 Pooling

Pooling function is the next type of layer in CNNs. It replaces the output of the net at a certain location with a summary statistic of the nearby outputs. For example, the max pooling operation reports the maximum output within a rectangular neighborhood. Other popular pooling functions include average pooling of a rectangular neighborhood [25]. See fig. 2.9 on how to perform max pooling.

We clearly see from fig. 2.9, As in the convolution operation, we slide a window

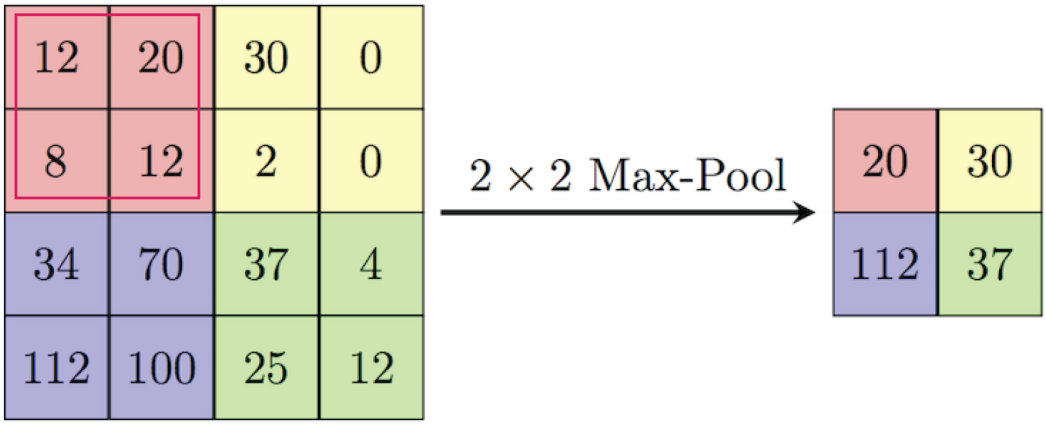


Figure 2.9: *Maxpooling example.*

across the image typically a 2×2 window. The value of the corresponding node in the max pooling layer is calculated by just taking the maximum of the pixels contained in the window. The pooling function is applied independently on every feature map in the input stack. The output is a stack with same number of feature maps with width and height reduced by a factor of two.

In all cases, pooling helps to make the representation approximately invariant to small translations of the input. Invariance to translation means that if we translate the input by a small amount, the values of most of the pooled outputs do not change. Invariance to local translation can be a useful property if we care more about whether some feature is present than exactly where it is. For example, when determining whether an image contains a face, we need not know the location of the eyes with pixel-perfect accuracy, we just need to know that there is an eye on the left side of the face and an eye on the right side of the face. Another improvement that pooling brings is the computational efficiency of the network. The reason being is that pooling reports summary statistics for regions spaced with stride s (typically 2 is used), therefore the next layer has roughly s times fewer inputs to process and reduces the memory requirements for storing parameters [25].

Therefore, most CNNs are composed of only those two layers: Pooling and convolution. We begin with convolution layers which detects regional patterns in an image using a series of filters. Typically, just like fully connected networks, an activation function is applied to the output feature maps. ReLU activation function is used as it has proven to be extremely efficient in object classification tasks. Then pooling layers follow the convolutional layers to reduce the dimensionality of their input tensors. CNNs are designed with the goal of taking an input image and gradually making it much deeper than it is tall or wide. As the network gets deeper, it is actually extracting more and more complex patters and features that help identify the content and objects in an image. CNNs are usually referred to as **feature extractors**. Another issue that rises when training CNNs, is the input image

dimensions. Since training requires large data-sets of thousands of images, it is no surprise that these images are of different sizes and shapes. Therefore CNNs requires a fixed sized input due to batch training. Indeed, instead of passing one image at a time through the network, we usually pass batches of images which are just stacks of images. But in order to do that, all the images have to have the same width and height. So, we have to pick an image size and resize all of our images to that same size before doing anything else.

2.6 Transfer learning

Usually, training very deep networks from scratch is a very tedious task; huge data-sets are required for the task to better generalize to real life situations. Modern CNNs usually take 2-3 weeks to train across multiple GPUs. However, it has been revealed that deep networks trained on natural images exhibits a curious phenomenon in common: on the first layer they learn general features similar to color blobs and edges. Such first layer features appear not to be *specific* to a particular data-set or task, but *general* in that they are applicable to many data-sets and tasks [27]. This means it may be useful to transfer this knowledge to other similar tasks. This technique is referred to as **transfer learning**. Deep CNNs are good candidates for this task because they are usually trained on general tasks (like image classification of daily life objects) and have many adjustable layers. As reference [27] states, the transferability of features decreases as the distance between the base task and target task increases, but that transferring features even from distant tasks can be better than using random features. A final surprising result is that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after "fine-tuning" to the target data-set. One of the strategies used when using transfer learning is referred to as **fine-tuning**. This simply means retraining the whole or parts of the pre-trained CNN. This is done by retraining with the new data-set without changing the architecture or reinitializing the weights (but some new layers might be added or changed depending on the task at hand). The existing weights are said to be *fine-tuned* to the new task at hand [28].

2.7 Data augmentation

Data augmentation is a strategy that enables practitioners to increase the diversity of data available for training models, without actually collecting new data. This is achieved by applying random (but realistic) transformations such as image rotation, cropping, padding, and horizontal flipping .... This is a good practice, since augmenting the size of the training set means more data to learn from, which makes the network even robust.

Chapter 3

CNN application: Object detection

The concept of convolution and convolutional neural networks has been applied to many real life problems: including object classification, object detection, speech recognition, disease depiction in medical images, self driving cars, and many more. Object detection is the task of detecting, meaning classifying and localizing instances of semantic objects of a certain class (in our case, car license plates along with their digits). An object detection algorithm should not only be able to classify an object but as well as localizing it in an image by drawing a bounding box around it, see fig. 3.1.

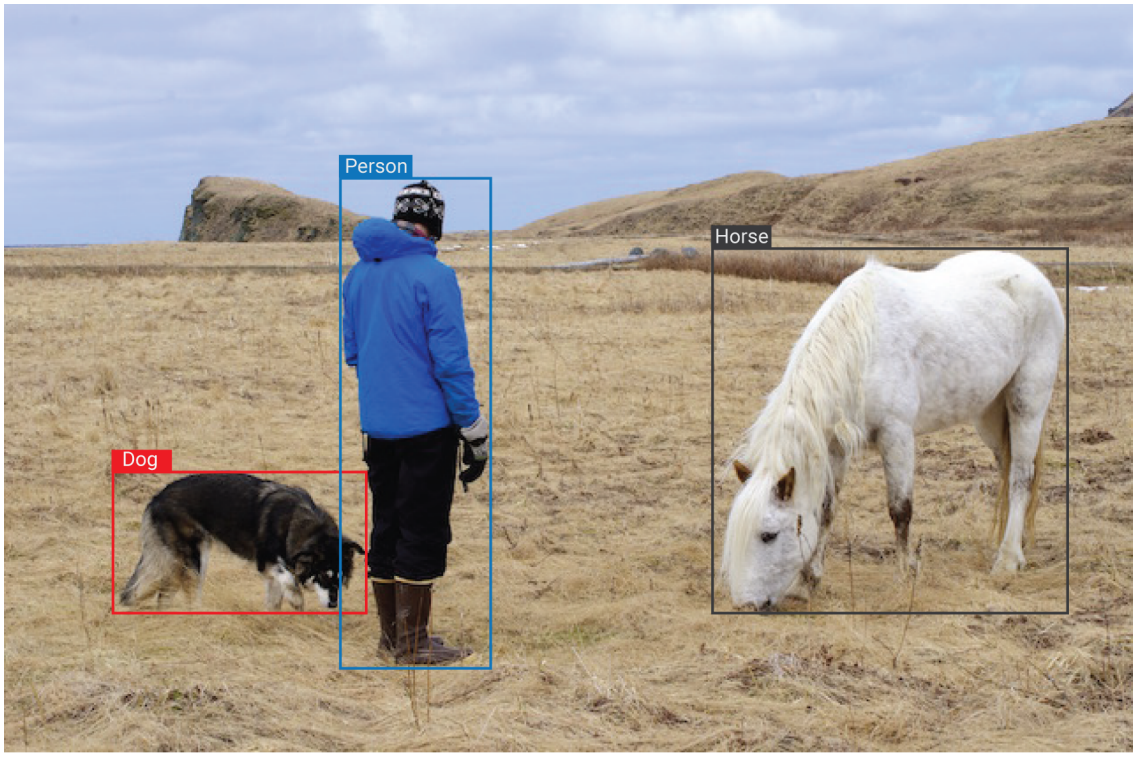


Figure 3.1: *Example of what an object detection system should accomplish.*

In this is chapter, we will focus on the theory behind the state-of-the-art detection systems: you only look once (YOLO) and Region-CNN (RCNN).

3.1 YOLO: you only look once

Over the past few years, the YOLO algorithm have evolved quite a lot going; from YOLOv1 all through version four. The different improvements that this algorithm went through are just the fruits of many research developments in the deep learning field incorporated into YOLO algorithm to make it more robust and less prone to errors. In this section, we shall present the version three of YOLO. Version four has only been developed in April 2020 during the middle of the pandemic. Many techniques have been included in this last paper [20] which makes a bit difficult since we have to go through all the new details. Therefore we shall only present version three.

3.1.1 Bounding boxes

The YOLO algorithm divides the input image into an $S \times S$ grid. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object [31]. Each grid cell predicts B bounding boxes, using anchor boxes. Anchor boxes are predefined boxes of certain width and height. They are defined to capture the scale and aspect ratio of specific object classes you want to detect. Anchor boxes are typically chosen based on object sizes in the training data [4], see fig. 3.2. Anchor boxes have been introduced to solve two issues (second issue will be discussed in section 3.1.2). Objects in the YOLO algorithm are associated with grid cells that their centers fall into. If two objects' centers fall into the same grid cell, we will not be able to predict both objects. Therefore, we can associate each grid cell with multiple anchor boxes, each responsible to detect only one object in that cell. A typical number of boxes used is three, see fig. 3.2. That is the first issue anchor boxes solves.

Each bounding box is associated with a confidence score p_c, which reflects how confident the network is that the bounding box contains an object (also called objectness) [31]. This should be ideally 1 if there is an object otherwise 0 [30]. The Bounding box is defined by 5 parameters: the box center coordinates b_x and b_y, the height b_h, the width b_w, and the class confidence score p_c [5]. For instance, if we are building a self driving car object detection system, we may want to detects cars, pedestrians and motorcycles. Therefore, each grid cell will be associated with an $((5 + number\ of\ classes\ to\ detect) \times number\ of\ anchor\ boxes)$ dimensional vector. As we can see on fig. 3.2, the anchor boxes capture the scale and aspect ratio of cars and pedestrians. Indeed, most cars and humans will have approximately the same scale and aspect ratio. The vector y is composed of the objectness score as well as the bounding boxes and the class probabilities repeated for each anchor box. Here two anchor boxes have been used. YOLOv3 uses 3 anchor boxes. The image has been divided into a 3×3 grid just for illustration. The vector y represents the manual labeling for the central cell. Anchor box 1 is associated with the pedestrian

while the second one is associated with the car.

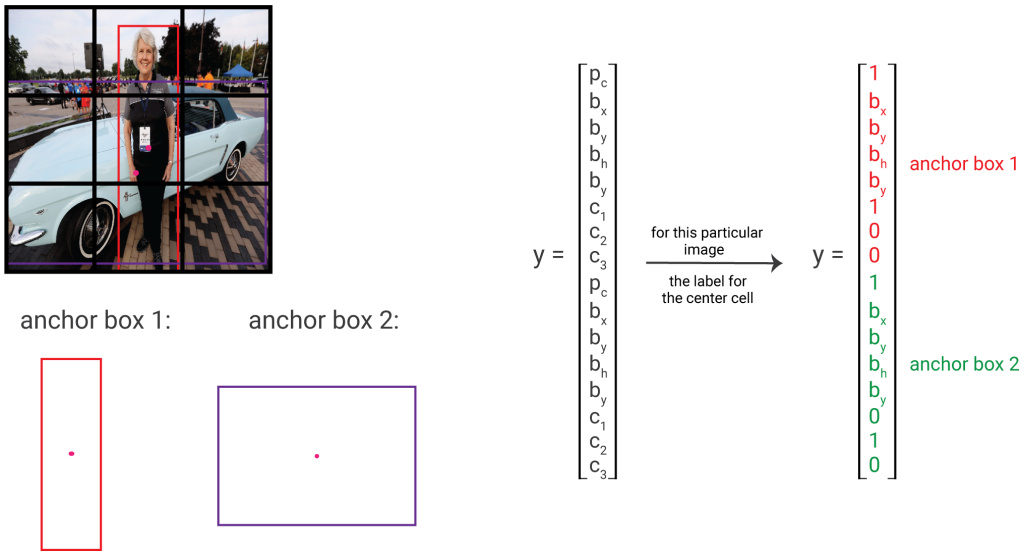


Figure 3.2: *Example of anchor boxes.*

3.1.2 Network design

The network is a series of convolutional and pooling layers chosen so that the network eventually maps the input image $W \times H \times 3$ to an output volume $S \times S \times ((5 + number\ of\ classes\ to\ detect) \times number\ of\ anchor\ boxes)$. YOLO's convolutional layers down-sample the image by a factor of 32, 16, 8. The YOLOv3 network has therefore 3 outputs instead of one, but we will be focusing on only one as the same calculation happen at each scale. The exact architecture is discussed in chapter 4. Now, to train the convolutional neural network, we pick an image size of 416×416. This number has been chosen because we want an odd number of locations in our feature map so there is a single center cell. Objects, especially large objects, tend to occupy the center of the image so it's good to have a single location right at the center to predict these objects instead of four locations that are all nearby [29]. By using an input image of 416×416 we get an output feature map of 13×13. The second issue anchor boxes address is the training instability [29]. In fact, during the early epochs of training if b_x and b_y are randomly initialized, the network struggles to converge to the right ground truth box's center. To overcome this problem, YOLO predicts location coordinates b_x and b_y relative to the grid cell. This bounds the ground truth to fall between 0 and 1. We use sigmoid activation to constrain the network's prediction to fall in this range. The network predicts B bounding boxes at each cell in the output feature map. The network predicts 5 coordinates for each bounding box $t_x,\ t_y,\ t_h,\ t_w\ and\ t_0$, see fig. 3.3. If the cell is offset from the top left corner of the image by (c_x, c_y) and the anchor box has width and height p_w and p_h, then the predictions correspond to [29]:

$$b_x = \sigma(t_x) + c_x \tag{3.1}$$

$$b_y = \sigma(t_y) + c_y \tag{3.2}$$

$$b_w = p_w e^{t_w} \tag{3.3}$$

$$b_h = p_h e^{t_h} \tag{3.4}$$

Since we constrain the location prediction, the parametrization is easier to learn, making the network more stable [29], see fig. 3.4. The question that naturally rises is: how, at the beginning, do we get p_w and p_h? Otherwise, how to assign an anchor box to a ground truth object ? The answer to this question is given is section 3.1.3 as we need to define an important function to proceed, see section 3.1.3.

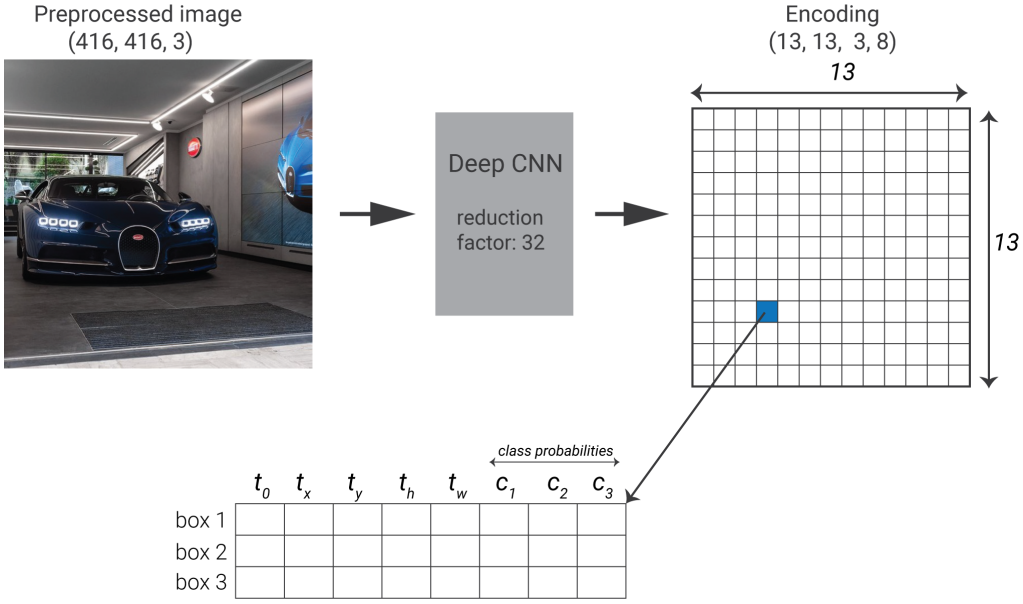


Figure 3.3: *The true output of YOLOv3 after introducing the training instability issue.*

In fig. 3.3, the network outputs a $13 \times 13 \times (8 \times 3)$ in this case, or simply put $13 \times 13 \times 24$ output volume. Each grid cell outputs three bounding boxes.

During training, we optimize the multi-part loss function L (eq. (3.5)).

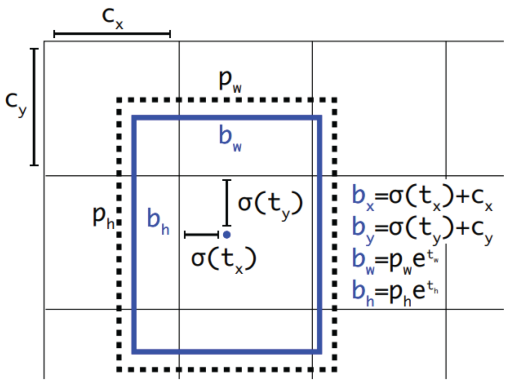


Figure 3.4: *Bounding box calculation [29].*

$$
\begin{aligned}
L(w,b) = \sum_{scales} \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} \left[(t_x - \hat{t}_x)^2 + (t_y - \hat{t}_y)^2 + (t_w - \hat{t}_w)^2 + (t_h - \hat{t}_h)^2 \right] \\
+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} \left[-\log(\sigma(t_o)) + \sum_{k=1}^{C} BCE(\hat{y}_k, y_k \right] \\
+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{noobj} \left[-log(1 - \sigma(t_o)) \right] \quad (3.5)
\end{aligned}
$$

As we can see in eq. (3.5), the first sum is over scales, meaning different regions of the network. Indeed, the network used in YOLOv3 does have only one output but three. The architecture of the network is discussed further in chapter 4.

where $1_{i,j}^{obi}$ denotes if object appears in cell i and that the j^{th} anchor box in cell i is "responsible" for that prediction. If an anchor box is not assigned to a ground truth object, it incurs no loss for coordinate or class predictions, only objectness. In cells that contain an object, the bounding box coordinates are calculated using the sum-squared loss function. Each box predicts the classes the bounding box may contain using multi-label classification using BCE.

3.1.3 Processing the algorithm's output

After training, the network will infer multiple detections. In fact, for each cell in the $S \times S$ grid, using B anchor boxes, the algorithm will infer B bounding boxes for each cell, which makes a total of $B \times S^2$. Therefore an object can be detected multiple times. **Non-max suppression** is an algorithm that cleans up those detections and makes sure each object gets detected only once. Before discussing it though, let us introduce an important function called **Intersection over Union** (IoU) that calculates how much a box or a rectangle overlaps another. So, IoU calculates the area defined by the intersection of the two boxes and divide it by the area defined by their union, see fig. 3.5.

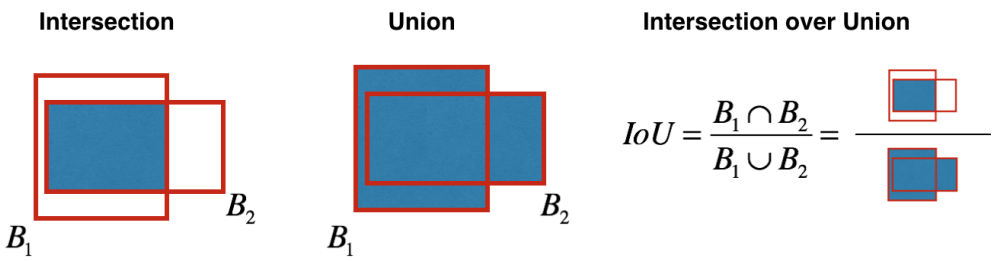


Figure 3.5: *Intersection over union metric.*

IoU is an evaluation metric used to measure the accuracy of an object detection system on a particular data set. Indeed, most object detection algorithms will judge a detection to be correct if the IoU between the ground truth box and the detected box is more than 0.5, see fig. 3.6. We often see this evaluation metric used in object detection challenges such as the popular PASCAL VOC challenge [6].

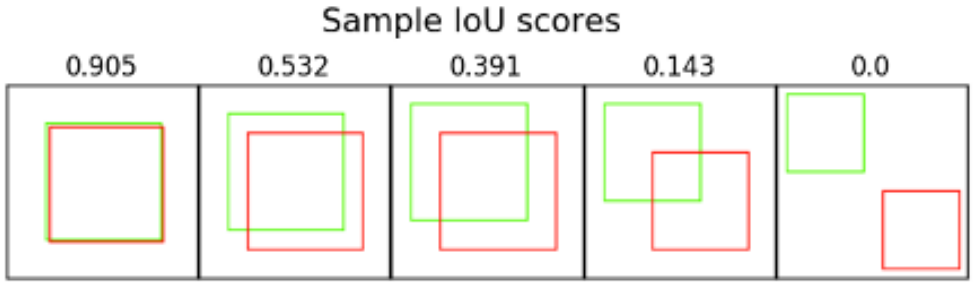


Figure 3.6: *Sample IoU scores.*

Back to our original question, how non-max suppression works. First, all the boxes having an $objectness \times the\ class\ probability$ less than or equal to some threshold are discarded (typical value of threshold is .6). While there any remaining boxes, we pick the box with the largest $objectness \times the\ class\ probability$ and output it as a prediction. Then we discard any remaining box with $IoU \geq 0.5$ with the box outputted in the previous step. This algorithm ensures that each object is detected only once.

In section 3.1.2, we discussed how the bounding boxes are being computed, and we finished it with a question: How does the anchor boxes being assigned to ground truth objects at the beginning? YOLOv3 assigns the anchor with the highest IoU overlap with a ground truth box.

3.2 Faster R-CNN

Several object detection techniques and models have been developed over the years; each with its benefits and drawbacks. In this section, we shall explore the faster region-CNNs technique to tackle this task. The Faster R-CNN model is composed of two networks: region proposal network (RPN) for generating region proposals and a network using these proposals to detect objects [40]. The main difference here with its' predecessor Fast R-CNN is that the later uses an algorithm called "selective search" to generate region proposals [23]. The time cost of generating region proposals is much smaller in RPN than selective search, since the RPN network does a significant part of computation which overlaps with the computation needed for the object detection network. In short, RPN ranks region boxes (called anchors) from most likely to less likely to contain an object and proposes the ones most likely containing objects [40]. The architecture is shown in fig. 3.7.

3.2.1 Anchors

In the default configuration of Faster R-CNN, it considers 9 anchors at each position of an image. fig. 3.8 shows 9 anchors at the position $(320, 320)$ of an image with size $(600, 800)$. The colors represent three scales or sizes: $128 \times 128,\ 256 \times 256,\ 512 \times 512$. For each color we have three boxes that have height width ratios $1:1,\ 1:2$ and $2:1$ respectively. These two parameters called "scales" and "aspect ratios", have a significant effect on the performance of our model. The RPN selects a position in a given image at every stride of 16 where it generates those 9 anchors. In an image of the same size as fig. 3.8 there will be 1989 (39×51) positions, leading to 17901 (1989×9) boxes to consider. This number of anchors is hardly smaller than the technique of of sliding window and pyramid. The advantage here is that we can use region proposal network to significantly reduce the number of boxes that will be considered by the classifier network [40].

These anchors work well for Pascal VOC [21] data set as well as the COCO data set [42]. We have the freedom to design different kinds of anchors/boxes. If for example, you are designing a network to detect passengers/pedestrians, you may not need to consider the very short, very big, or square boxes. A uniform set of anchors may increase the speed as well as the accuracy.

3.2.2 Region Proposal Network

The input to the RPN module is the feature map of an image. The RPN generates centers on the original image for each "pixel" in a feature map obtained from a forward pass through a pre-trained CNN. It then generates 9 anchors around each center according to the specified scales and aspect ratios [40]. The output of a RPN is a set of probabilities for

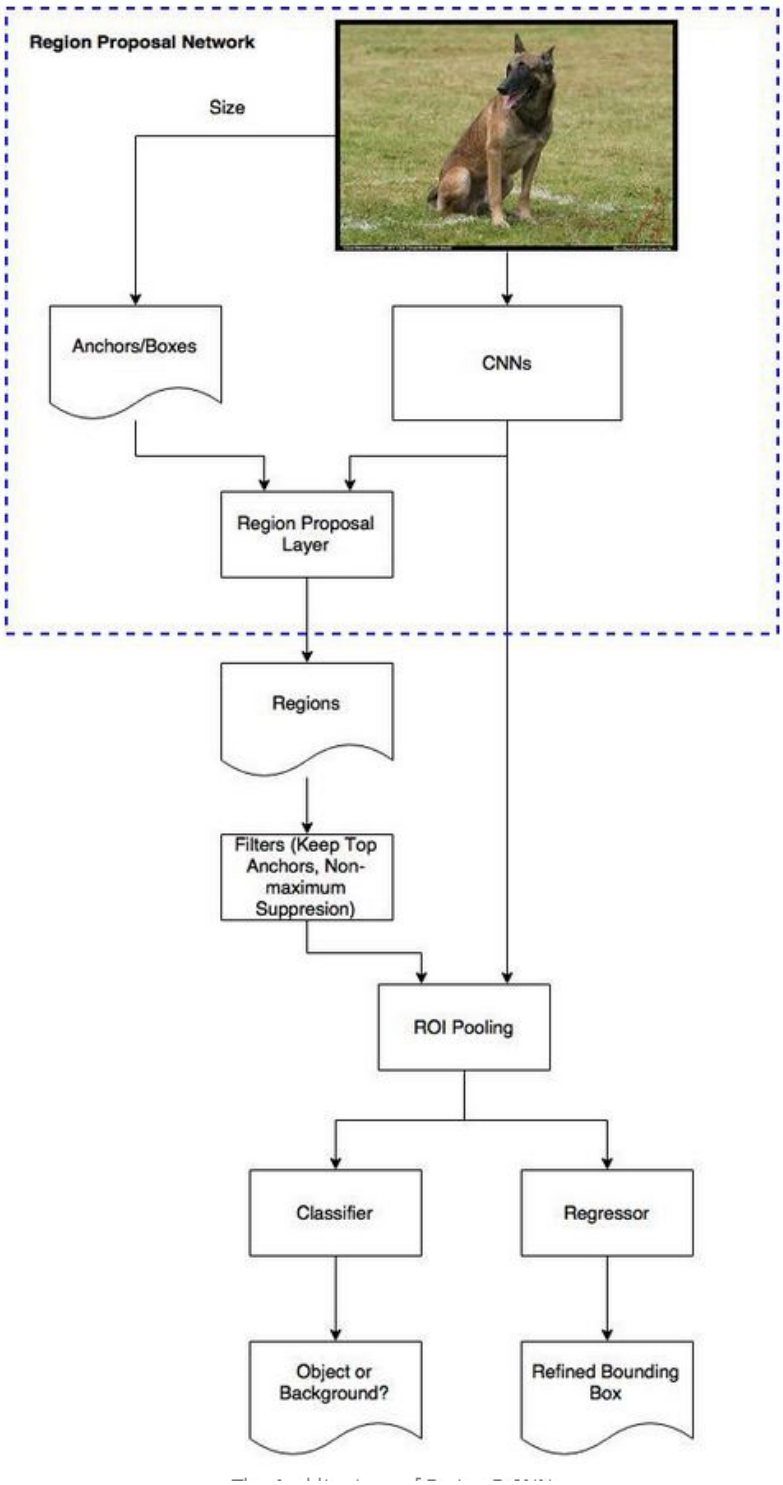


Figure 3.7: *Faster R-CNN model structure.*

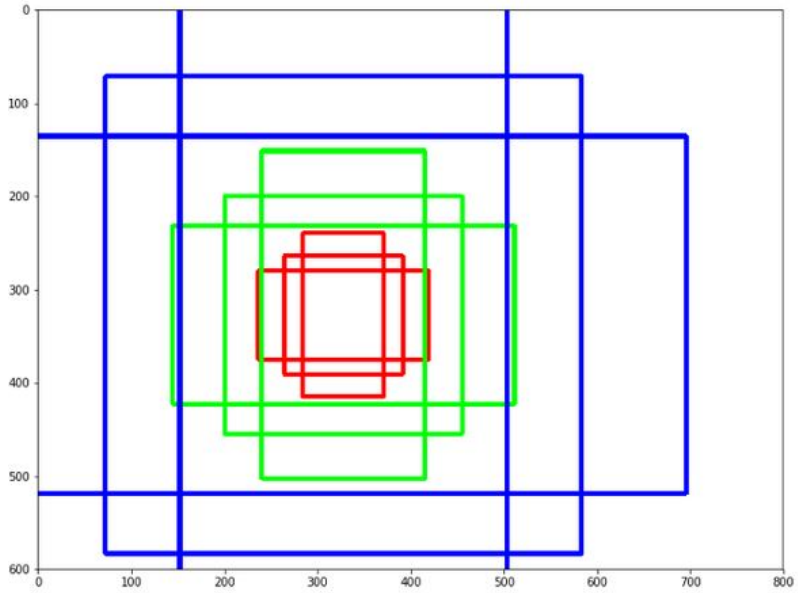


Figure 3.8: *Example of anchors at single location.*

each anchor that determine the probability of a certain anchor being an object or not. It also outputs a set of error estimations for the anchors which overlap with a ground truth box. These outputs will be examined by a classifier and regressor to eventually check the occurrence of objects. To be more precise, RPN predicts the possibility of an anchor being background or foreground, and refines the dimensions of an anchor [40]. Since the RPN performs a classification task, it will go through a training process for which we must have a clear definition of the data-set and the labels. In this case, our data-set is the anchors defined for each image. As for the labels; the basic idea is that we want to label the anchors having the higher overlaps with ground-truth boxes (the bounding box surrounding the object we wish to detect as foreground), and the ones with lower overlaps as background. For this we use the IOU function. If the value of the IOU is higher than a certain threshold then it would be labeled as foreground otherwise it is labeled as background [40].

The RPN also performs a regression task on the same anchors in order to correct the dimensions and location of these same anchors. For each anchor, it computes an estimation of error on the width t_w, the height t_h, and the location of the anchor center t_x and t_y such that

$$t_w = log\left(\frac{w}{w_a}\right) \tag{3.6}$$

$$t_h = log\left(\frac{h}{h_a}\right) \tag{3.7}$$

$$t_x = \frac{x - x_a}{w_a} \tag{3.8}$$

$$t_y = \frac{y - y_a}{h_a} \tag{3.9}$$

x, y, w, h are the ground truth box center coordinates, width and height, and $x_a, y_a, h_a and w_a$ are the anchor boxe center coordinates, width and height, respectively[40].

The final and most important component of the training process is the loss function

$$L(p_i, t_i) = (\frac{1}{N_{cls}}) \sum_i L_{cls}(p_i, p_i^*) + \lambda (\frac{1}{N_{reg}}) \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3.10}$$

where p_i is the predicted probability of objectness and p_i^* is the actual score. t_i and t_i^* are the predicted anchor dimensions and coordinates and actual ones, respectively. The ground-truth label p_i^* is 1 if the the anchor is positive and 0 if the anchor is negative [40].

3.2.3 ROI Pooling

The purpose of the region of interest pooling (ROI) is to perform max pooling on inputs of non-uniform sizes to obtain fixed-size feature maps (e.g. 7×7). This layer takes two inputs:

- A fixed-size feature map obtained from a deep convolutional network with several convolutions and max-pooling layers.
- An $N \times 5$ matrix of representing a list of regions of interest, where N is the number of ROIs. The first column represents the image index and the remaining four are the co-ordinates of the top left and bottom right corners of the region.

For every ROI from the input list, it takes a section of the input feature map that corresponds to it and scales it to some predefined size (e.g., 7×7). The scaling is done by:

- Dividing the region proposal into equal-sized sections (the number of which is the same as the dimension of the output).
- Finding the largest value in each section.
- Copying these max values to the output buffer.

The result is that from a list of rectangles with different sizes we can quickly get a list of corresponding feature maps with a fixed size. Note that the dimension of the RoI pooling output does not actually depend on the size of the input feature map nor on the size of the region proposals, but is determined solely by the number of sections we divide the proposal into.

One of the benefits of ROI pooling is a reduction of processing speed. If there are multiple object proposals on the frame (and usually there'll be a lot of them), we can still use the same input feature map for all of them. Since computing the convolutions at early stages of processing is very expensive, this approach can save us a lot of time [40]. Figure 3.9 shows the working of ROI pooling.

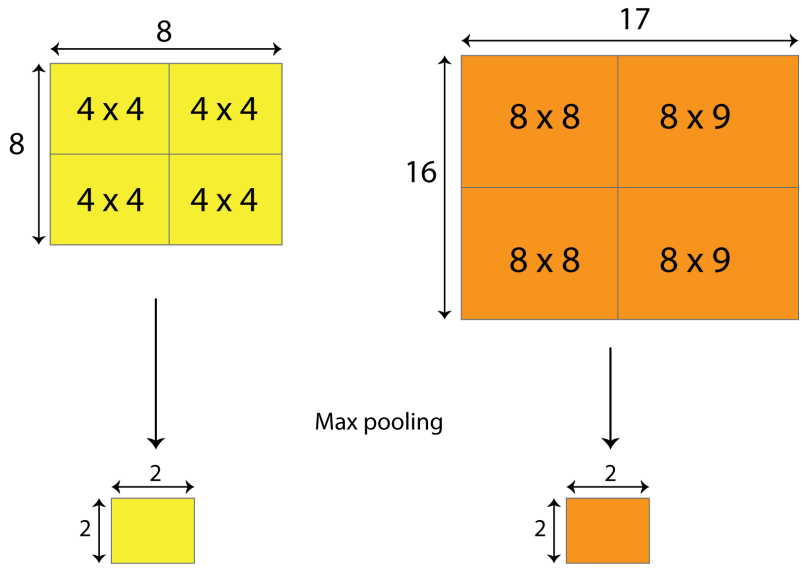


Figure 3.9: *ROI pooling operation.*

The output of ROI pooling will be the input to a classifier network, which is a copy of the pre-trained backbone network we used to obtain the feature map, and which will be referred to as "Fast RCNN classifier network". This network will further branch out to a classification head and regression head. The loss function for the Fast-RCNN network is defined in the same way as the RPN loss, except for the significance of the variables. p_i is the predicted class scores for every class of objects we want to detect and p_i^* is the actual score. t_i and t_i^* are the predicted cooridinates and actual coordinates, respectively. The ground-truth label p_i^* is 1 for a certain class of objects if the region outputted by the ROI layer contains that object and 0 if it does not.

3.2.4 Faster RCNN training

For training the entire Faster RCNN model, there must be a well defined loss function which encapsulates all of the losses mentioned before. It can be considered as an estimation

for error in the model, regardless of where the error occurs. The total loss is defined as the sum of the RPN loss and the Fast RCNN classifier network loss.

$$\text{Total loss} = \text{RPN loss} + \text{Fast RCNN classifier loss} \tag{3.11}$$

Chapter 4

Design and implementation of ALPRS

4.1 Workflow of the ALPRS implementation

For building the ALPRS application, we used a part-by-part approach rather than an end-to-end approach. The reason being is the lack of labeled data sets designed for this specific application, especially when it comes to license plates for which there are no published data sets. Needless to mention that the task at hand is relatively less tedious since Algerian plates contain no characters other than digits. Our system consists of a "detect plate", then "detect digits" pipeline. Each step has a dedicated module that runs sequentially and independently. The first module, called the “plate network”, takes the full raw image as input, detects plates in the image, crops the detected plates based upon their bounding boxes then passes them to the second module as inputs. The detected plates are cropped with an extra margin around them to avoid the exclusion of important details (maybe digits) for digit detection. The second module, called “the digits network”, receives the detected plates from the first module, detects and recognizes the 10 digit classes from 0 to 9, which are specific to the license plates. The position of the bounding boxes outputted by the previous module determines the order of each digit and a final prediction is given. To achieve this purpose, we divided our work into several small steps highlighted in the diagram shown in fig. 4.1.

The block of data collection is done using a phone camera. The tasks of labeling are done using software tools. The blocks of training, testing, and ALPR system implementation are done using a Python package called "Pytorch" [7] which provides functions and tools used to build deep learning models. The block of image extraction is done using the Python library OpenCv [8] which is a library specialized in the manipulation and handling of images. For each block, there is a Python script. Training blocks for Faster RCNN models use the same script with some changes made to the training process parameters. The testing blocks for the Faster RCNN models also use the exact same Python code, it passes examples from the testing and training set through the trained model, and computes

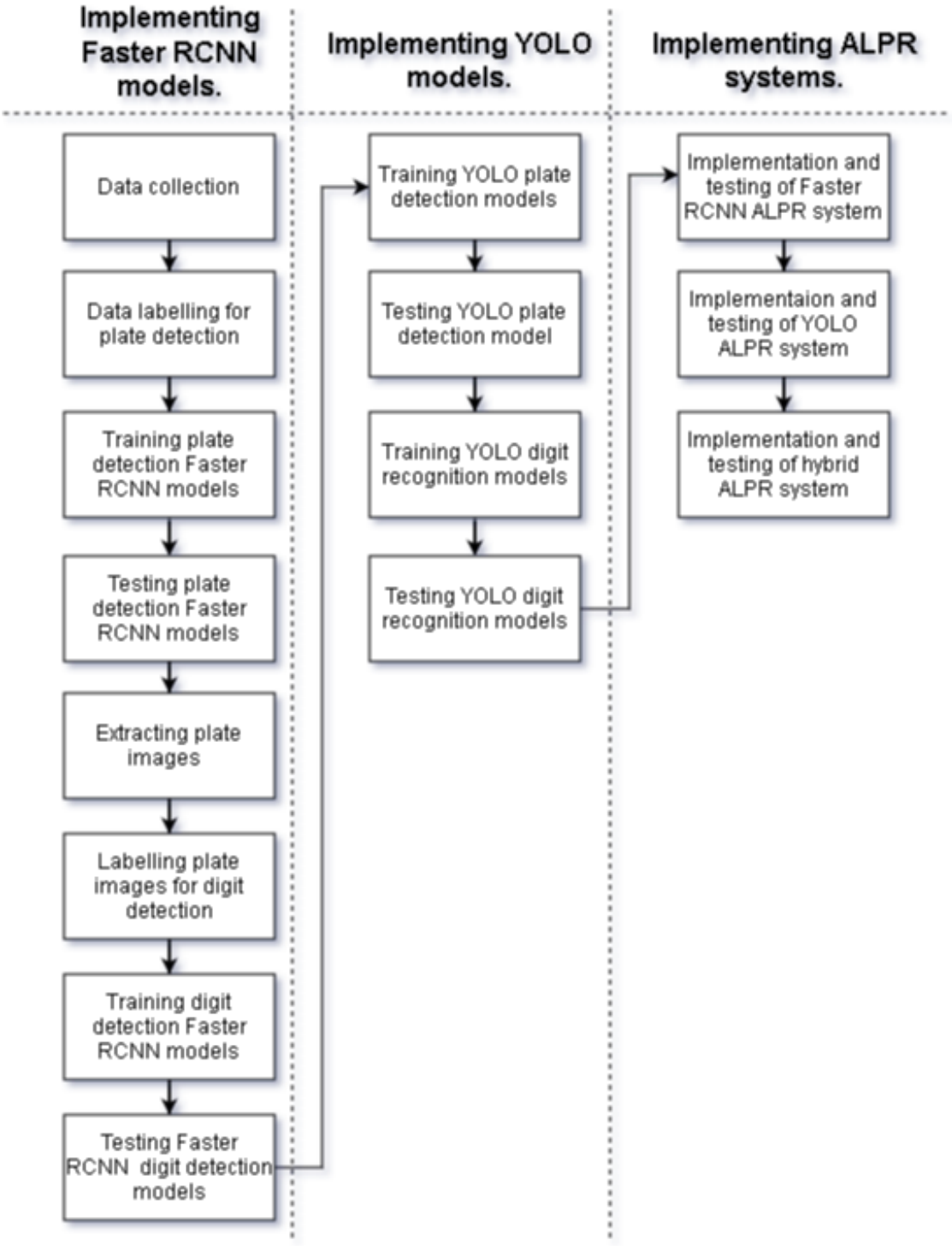


Figure 4.1: *Block diagram illustrating the workflow.*

the mean average precision (mAP). All of the tasks inside the training and testing processes have been coded from scratch using built-in Pytorch tools that perform convolution and similar basic operations, except for the back-propagation and cost computation tasks for which Pytorch provides built-in functions. The training and testing blocks of the YOLO models are done using a framework called Darknet [9].

4.2 Data collection and labeling

Data collection includes the manual collection of images of license plates in different positions, angles, lighting, distance, size, color .... We used the Huawei P10 phone camera to collect digital images. The process consisted of taking pictures of random cars in the streets, parking lots, and in some instances, we made visits to license plate shops and took pictures of isolated license plates. After a few hundred images, we noticed that the great majority of license plates numbers ended with 35 or 19 because the cities where these pictures were taken were Boumerdes and Setif. This raised a concern that there might be an over-representation of the digits composing these numbers, which would make the CNN disproportionately better at them than the remaining digits. If this happens, it might give us good results during evaluation but they would only mean that the CNN is very good at detecting the digits 3, 5, 9, and 1 and not the entire set of digits. The solution we came up with is, instead of taking a trip to every wilaya to take the pictures, we visited the international airport in Algiers. There, we found that the cars parked there had plate numbers that are diverse and somewhat balanced. We managed to take close to 450 pictures. Another advantage of visiting the airport parking-lot is that we had access to a large number of license plates all in the same place. The only caveat to this solution is the fact that all the cars were under the same lighting conditions as all the pictures were taken during the day. The overall data set contains close to 1000 images that are set to be used for labeling. Figure 4.2 shows some examples.

The second CNN which performs digit recognition will be trained on cropped images of plates only, which will be obtained by passing the original set of images through the plate detector and using the obtained plate bounding boxes to crop the plates.

The next step is to label these images by assigning a bounding box to each license plate in the images.“Labeled” data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of it with informative tags. For example, a label might indicate whether a photo contains a horse or a cow, which words were uttered in an audio recording, what type of action is being performed in a video, what the topic of a news article is, what the overall sentiment of a tweet is, or whether a dot in an X-ray is a cancer.

Labels can be obtained by asking humans to make judgments about a given piece of unlabeled data (e.g., "Does this photo contain a horse or a cow?"), and are significantly

Figure 4.2: *Image collection example*

more expensive to obtain than the raw unlabeled data. After obtaining a labeled data set, machine learning models can be applied to the data so that new unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data.

For the plate detection CNN, the data will be labeled manually by defining a rectangular bounding box around every license plate in each image. This process consists of manually drawing bounding boxes around the license plates in all of the 1000 images in our data set using a software tool called "LabelImg" [10], shown in fig. 4.3. It provides an interface to pass through the images saved in a directory, as well as some tools for drawing boxes and assigning labels to them. LabelImg saves the label for each image in an XML file holding the name of the image file in the form of coordinates of the top-left and bottom right-corner of every box along with the label assigned to it. The total number of boxes drawn is about 1000 since most images contained a single car, which was a tedious task that took few weeks.

The labeling process for the CNNs of digit detection is done after the CNNs of plate detection are trained and tested. The plate detection CNNs were used to crop out images of plates. To perform this task we coded a Python script using the OpenCv library which provides functions for the manipulation and editing of image files. The script runs the entire data set through a selected plate detection CNN. For each image, the script takes the output of the CNN, which is the coordinates of the predicted bounding box around plates, and uses them to crop the part of the image which contains the license plate and saves them in a separate directory, see fig. 4.4.

For the labeling process a similar tool is used to define a rectangular box around each digit in each plate image and assign a corresponding label to each bounding box. Note that the number of digits in our data set is over 10000 digits. This means that the process

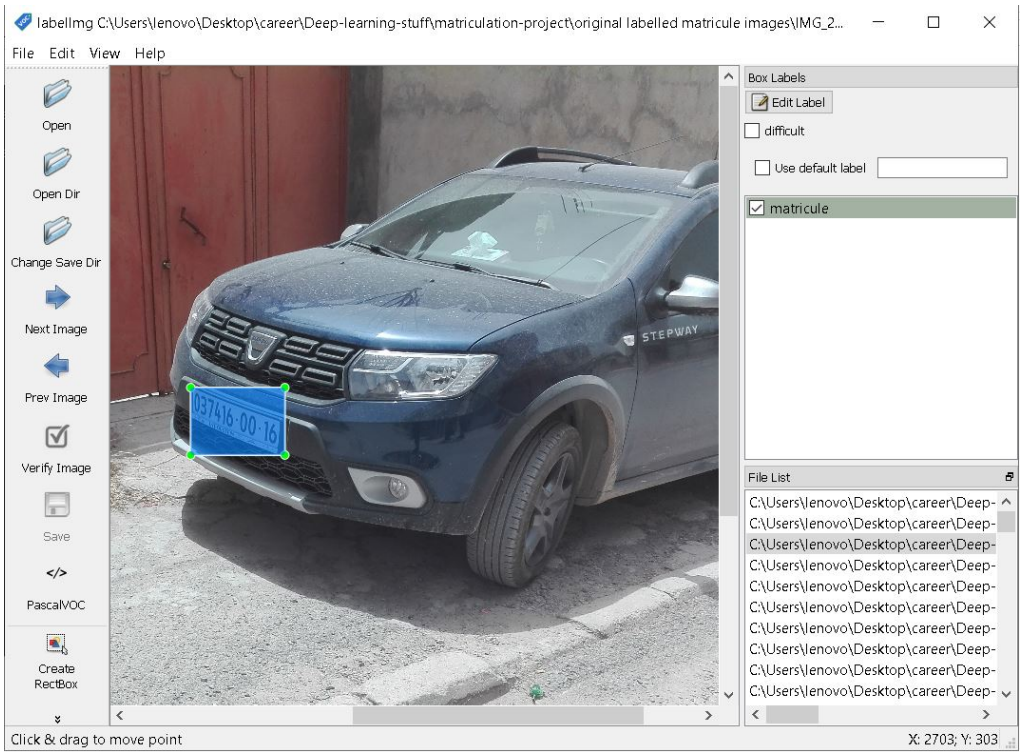


Figure 4.3: *Example of plate labeling.*

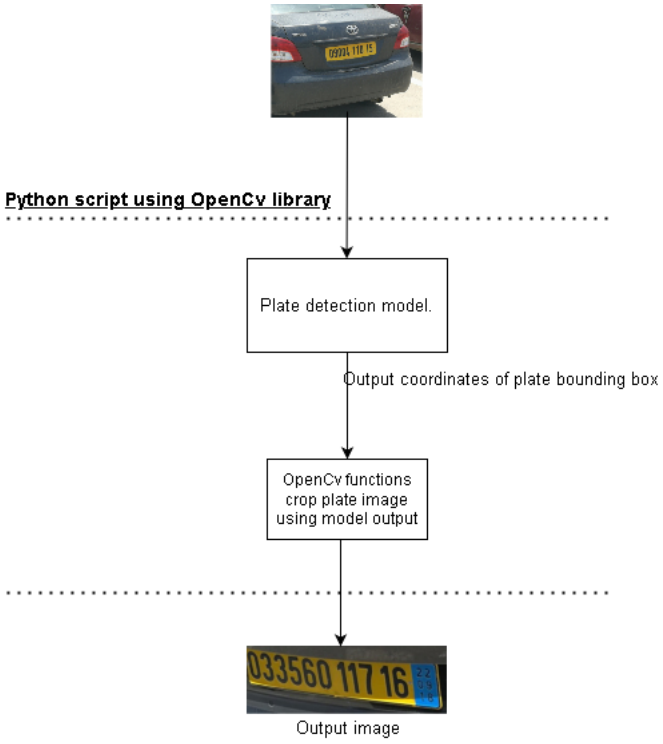


Figure 4.4: *Process of extracting plate images from the original images.*

consists of manually drawing and labeling 10000 rectangular boxes which took a few months. This process must be done carefully and the progress must be kept in secure storage. Labeling data for a machine learning project presents the advantage of eliminating the need for data cleaning and complicated pre-processing since the data set can be built in which ever form is suitable for training. Figure 4.3 illustrates the labeling tools used for both CNNs and fig. 4.5 illustrates an example of labeling.

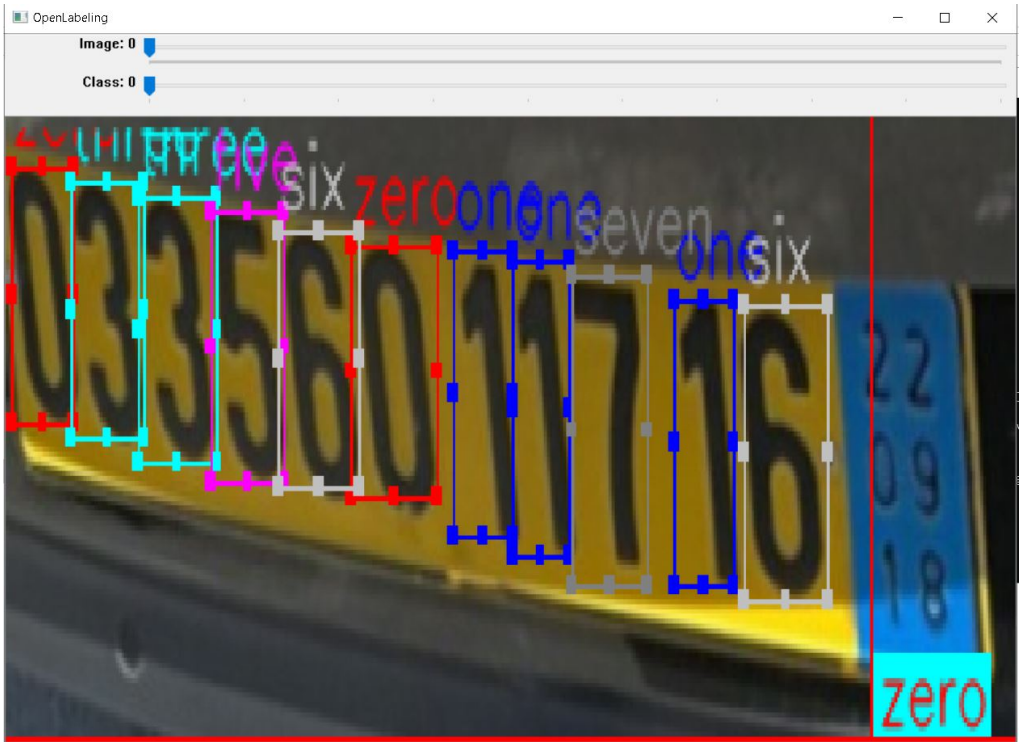


Figure 4.5: *Example of digit labeling.*

4.3 Training Faster RCNN models

Plate and digit detection requires training a set of models to detect and localize a license plate numbers in any given image. The training process of a Faster-RCNN model is summarized in fig. 4.6. The training process is the same for both plate detection and digit recognition. The only difference is in the data set used and the number of object classes to detect.

First, an image from the training set is passed through a CNN model which is referred to as the backbone network. The output of the backbone network is a 3D tensor which is the concatenation of the results of the CNN last convolutional layer. As discussed previously, for every "pixel" in the feature map, 9 anchors with different scales and aspect ratios, are generated in the corresponding location on the original image. These anchors are labeled into two classes based on their area of overlap with any ground truth box: if they have an IoU larger than 0.7 they are labeled as positive anchors, otherwise they are

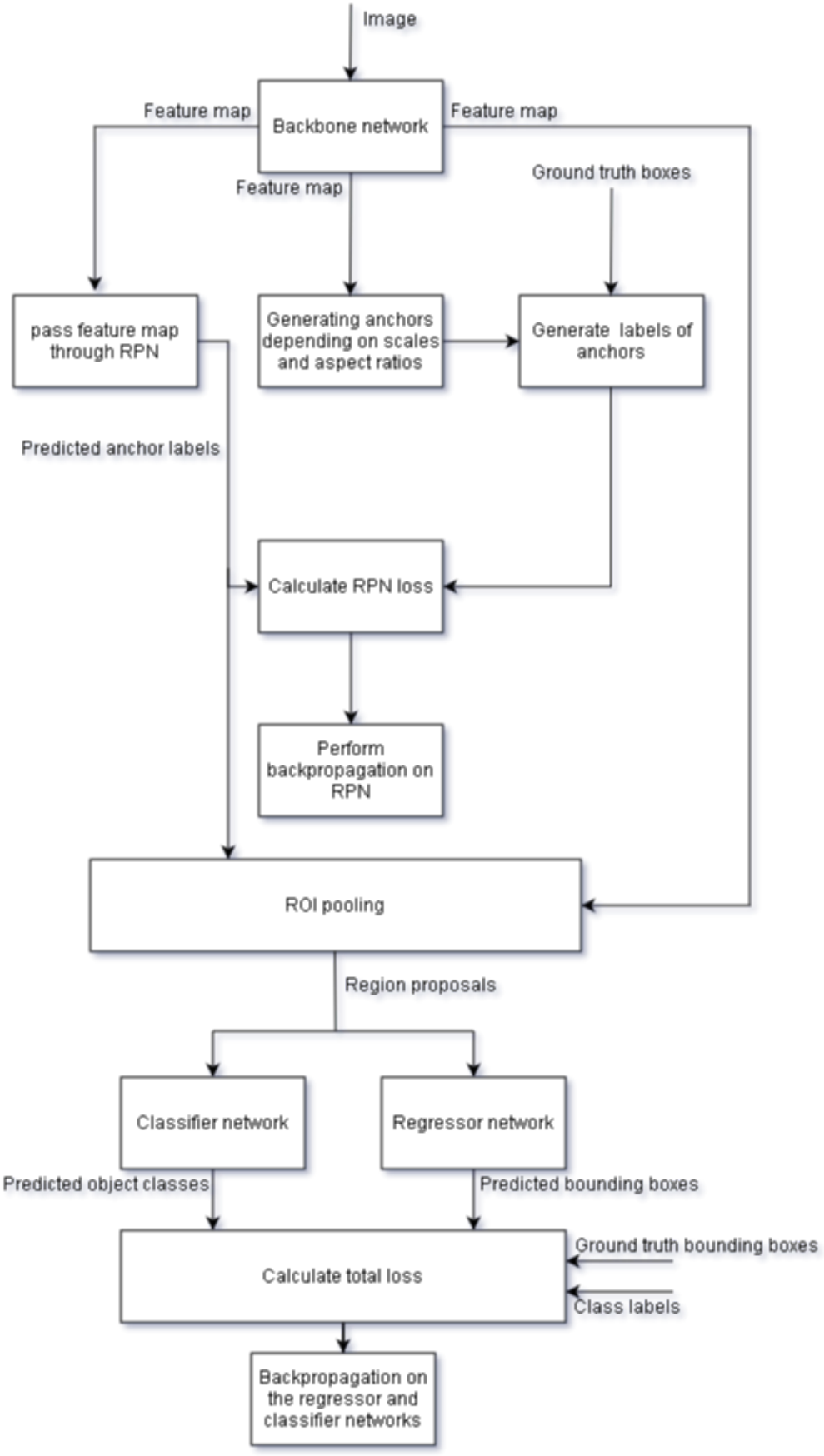


Figure 4.6: *Block diagram summarizing the Faster RCNN training process.*

labeled as negative.

The feature map is also passed through the RPN which is supposed to predict the label of each anchor. The RPN also performs a regression task on the same anchors in order to correct the dimensions and location of these same anchors. For each anchor, it computes an estimation of the error on the width t_w, the height t_h, as well as on the anchor center location t_x and t_y according to the following equations

$$t_w = log\left(\frac{w}{w_a}\right) \tag{4.1}$$

$$t_h = log\left(\frac{h}{h_a}\right) \tag{4.2}$$

$$t_x = \frac{x - x_a}{w_a} \tag{4.3}$$

$$t_y = \frac{y - y_a}{h_a} \tag{4.4}$$

x, y, w, h are the ground truth box center coordinates, width and height, whereas, x_a, y_a, h_a and w_a are the anchor boxes center coordinates, width and height [40].

These values are used to compute a cost function for the RPN as follows:

$$L(p_i, t_i) = \left(\frac{1}{N_{cls}}\right)\sum_i L_{cls}(p_i, p_i^*) + \lambda\left(\frac{1}{N_{reg}}\right)\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{4.5}$$

where p_i is the predicted probability of objectness and p_i^* is the actual score, t_i and t_i^* are the predicted coordinates and actual coordinates respectively. The ground-truth label p_i^* is 1 if the the anchor is positive and 0 if the anchor is negative [40]. The next step is to perform back propagation through the RPN network and optimize its weights and biases.

As discussed previously, RoI pooling selects relevant areas of the feature map covered by the positive anchors. It passes these areas to a classifier and a regressor networks. The classification and the regression networks use the same input, which is the coordinates of the proposed regions of the image. These tasks are independent from each other, and do not need to be performed in series. Both tasks of classification and regression happen in parallel. The classification network is concerned with outputting probabilities of the existence of specific objects in the proposed region. The regression network is concerned with outputting adjustments to the proposed regions based on the features shown in that region. The output of the classification network is compared to the true object classes of the training example, whereas, he output of the regression network is compared to the coordinates of the ground truth bounding box in the training example. The output of these networks is used to compute the total loss of the model. Finally, back-propagation is performed on the regressor and classifier networks.

There are three main parameters which can be adjusted in order to obtain optimal

results. The following parameters were chosen for both the digit and plate networks:

- Anchor scales;
- Backbone network;
- Number of training epochs.

This choice is justified by the fact that the other parameters as specified by the original paper [40], have been proven to be optimal in a number of previous works regardless of the application or the data set. The optimization algorithm used is the same for all instances of training. The used anchor scales and corresponding aspect ratios parameters are:

- scales : (32, 64, 128), aspect ratios : $(0.5, 1.0, 2.0)$
- scales : (64, 128, 256), aspect ratios : $(1.0, 2.0, 4.0)$

The used backbone networks are:

- VGG16
- Mobilenet
- Inception
- ResNet

The used numbers for training epochs are:

- 10 epochs
- 30 epochs
- 50 epochs

The list of parameters to be modified implies that the number of training processes launched is 24 for both plate detection and digit recognition, since we have 2 different aspect ratio, 4 backbone networks, and 3 numbers of epochs ($2 \times 3 \times 4$).

The main coding tool is Pytorch library which is a Python package developed by Facebook [10]. The platform used is Google COLAB [11] which is a free cloud service offered by Google. It provides Python programming environments equipped with all the tools and packages needed for building deep learning models, including Pytorch and GPU accelerators. Note that Pytorch is distinguished by its object oriented approach to machine learning models. Whereas every model, design, or process is represented by a class.

The process of training a Faster RCNN model using these tools includes the following steps:

- Implementing a data set class called "LicencePlateDataset";
- Implementing a backbone network class;
- Implementing a Faster RCNN model class;
- Implementing a trainer function.

LicencePlateDataset class implementation: The class “LicencePlateDataset” inherits from the Pytorch class called "Dataset". The main function of this class is to instantiate objects which represent the training examples. The *Dataset* class is characterized by the function “__getitem(index)__” which takes an index as an argument and returns a training

example as an image along with its corresponding label. The function *__getitem(index)__* is overridden to match the specific format of data that the model requires. The *Dataset* class is also characterized by the function "process()" which access the data set and modifies it in a way that makes it accessible by the *__getitem(index)__* function.

The attribute "transforms" is an object from the "torchvision.transforms" class which contains specification for dimentionality and data type modifications applied on each training example. The size of the training images needs to be $500 \times 500 \times 3$ as in the original paper [23]. The data type of the image tensor (images are stacks of three 2D matrices representing the RGB channels) needs to be a "torch.Tensor", which is the data type required by Pytorch models.

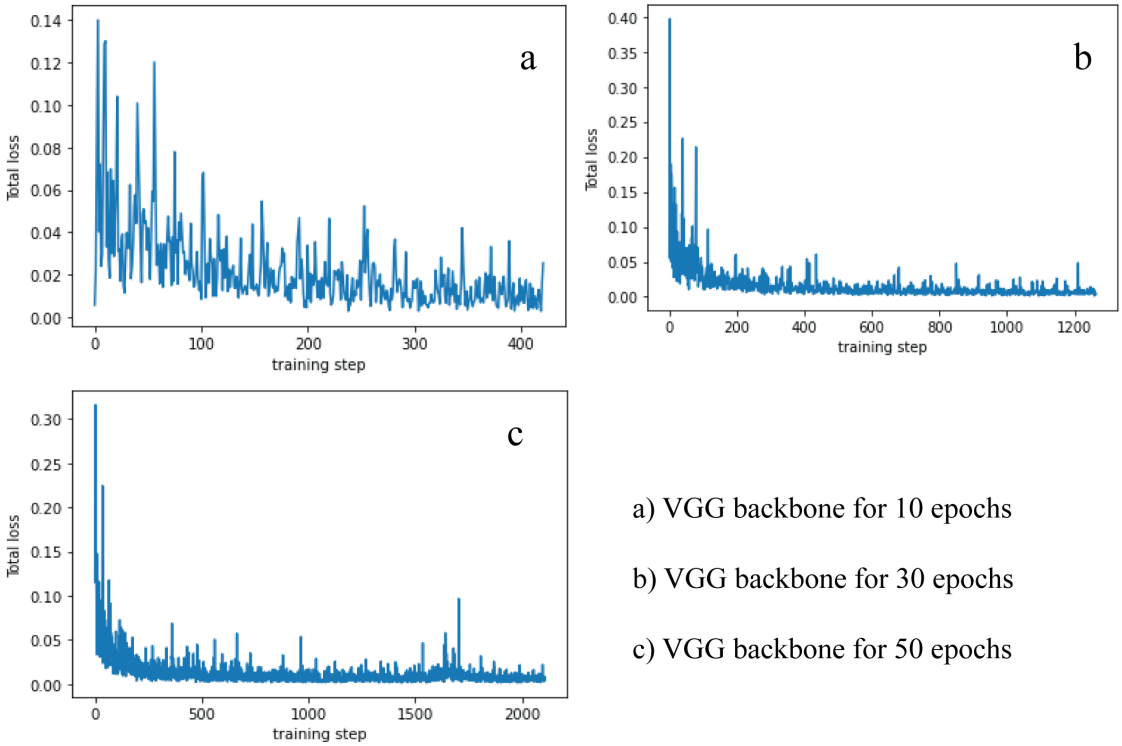
Backbone class implementation: The Backbone class instantiates backbone networks as objects. Objects from this class are characterized by the method "feed_forward(X)" which passes a given image X through the backbone network model and returns the output from the final convolutional layer. Objects from the same class are also characterized by the method "nn_base()" which defines the layers of the convolutional neural network. For each backbone network, a separate class is implemented because they each have different architectures and different helper methods and special operations. The full code is available in a Google drive [10]. These backbone networks are designed initially with random weights and biases, but the Faster RCNN model loads up pre-trained weights and biases in order to use them for transfer learning.

Faster RCNN model class implementation: The "FasterRCNN" class defines the different parts of the Faster RCNN model which are shown in fig. 3.7. The FasterRCNN class inherits from the class "GeneralizedRCNN" which is a built in class of Pytorch. GeneralizedRCNN provides some useful methods to models from the RCNN family. In addition to the FasterRCNN class, there are two additional classes implemented which are "TwoMLPHead" and "FastRCNNPredictor". TwoMLPHead is used to create the classifier and regressor networks that come after the RoI pooling layer.

Trainer function implementation: The training function contains the training loop in which the training process will take place. The first network to train is the RPN. Afterwards, the regressor and the classifier networks are trained on the output of the RoI layer output. The training function starts by feeding a training example through the network. A Pytorch built-in function calculate the gradients and optimizes the weights. Another Pytorch built-in function calculates the losses and the training process saves them in a list for plotting.

4.3.1 Results of training

The value of the total loss for each training process for plate detection was plotted with respect to the training steps. The total loss graphs for digit recognition are not presented because they have the same characteristics relevant for analysis as those for plate detection.



a) VGG backbone for 10 epochs

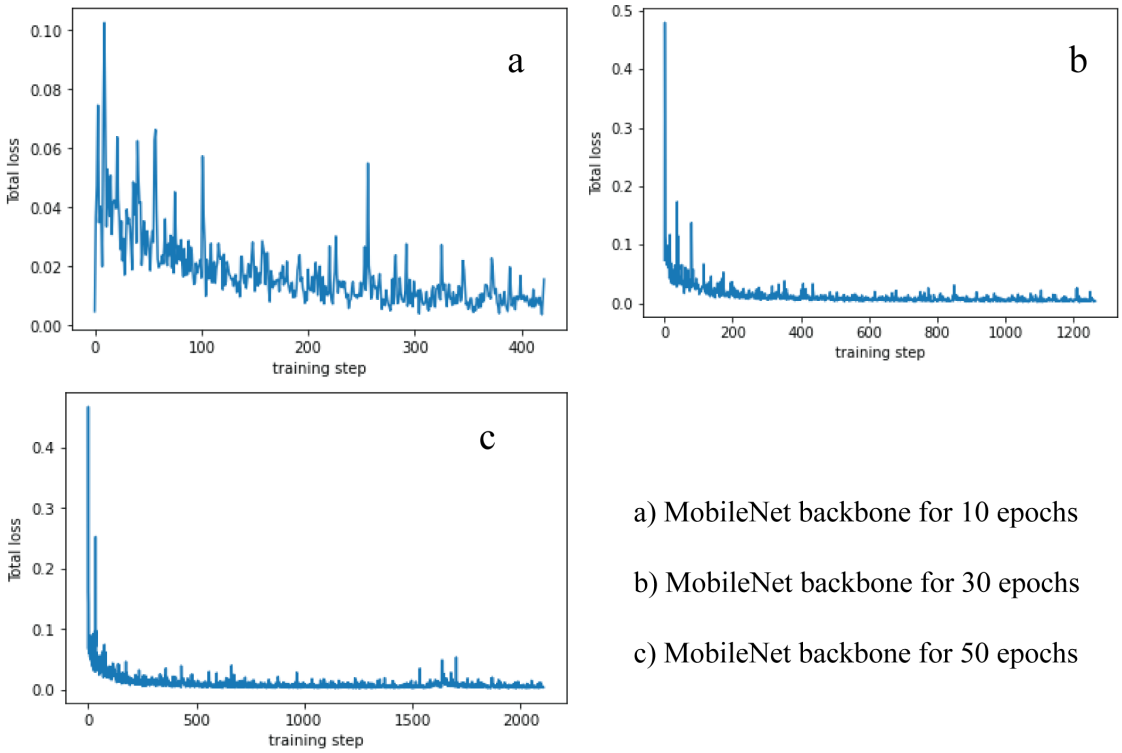
b) VGG backbone for 30 epochs

c) VGG backbone for 50 epochs

Figure 4.7: *VGG-16 for the first scales and aspect ratios.*

A first glance at the graphs from fig. 4.7 to fig. 4.14, indicates that the first set of scales and aspect ratios for the RPN anchors produce a cost function that converges towards a certain minimum; whereas the second set of scales and aspect ratios produce a cost function that seemingly diverges, and keeps oscillating around the initial cost value. This is due to the fact that the first set of scales and aspect ratios produce anchors of similar shape and scale to those in both training data sets. Using the second set was an attempt to find a better suited one for the tasks at hand, but it seems like there needs to be a grid-search [39] operation in order to possibly find one, which requires more advanced hardware resources and more time. The first set of scales and aspect ratios was used by the authors of the original paper and obtained the best results on very large and diverse data sets of common objects such as COCO [42], PASCAL [21], and Image-NET [16].

The second characteristic to notice is the fact that less complex models start at a higher loss value but converge towards a minimum faster than more complex ones. This confirms that models with less layers and less parameters per layer train faster than ones with more layers and more parameters. This is explained by the fact that the data sets at hand are too small to influence very deep models like Res-Net. This phenomenon is

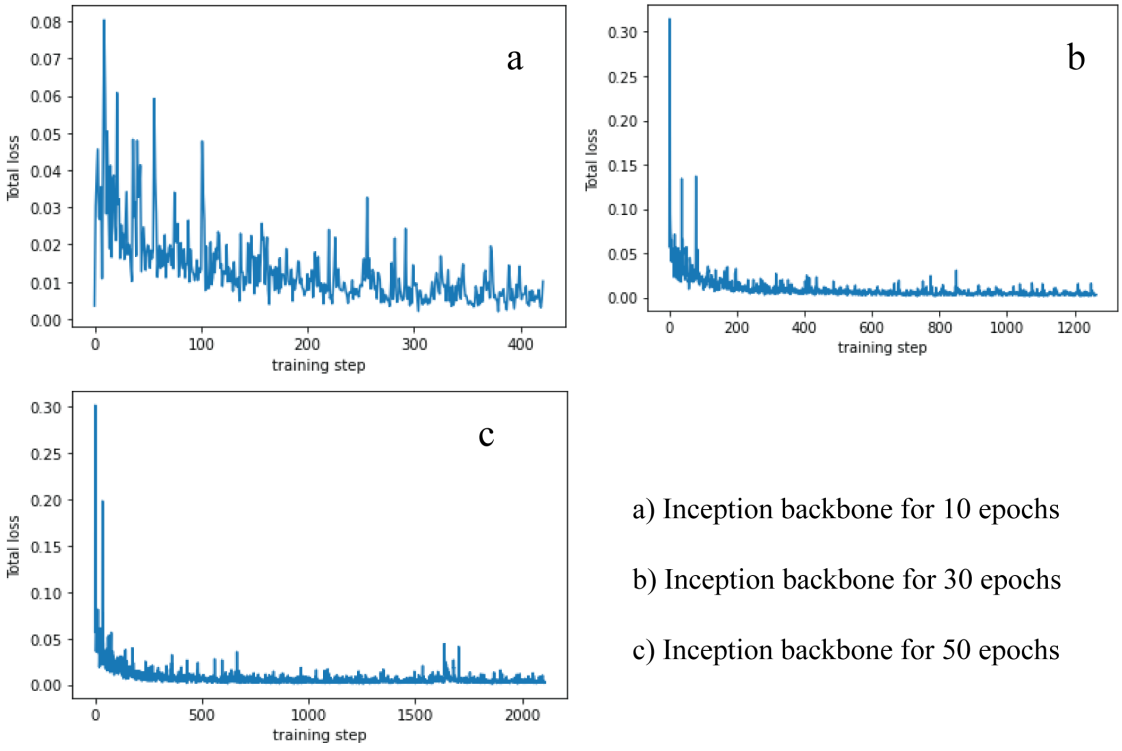


a) MobileNet backbone for 10 epochs

b) MobileNet backbone for 30 epochs

c) MobileNet backbone for 50 epochs

Figure 4.8: *Mobilenet for the first scales and aspect ratios.*

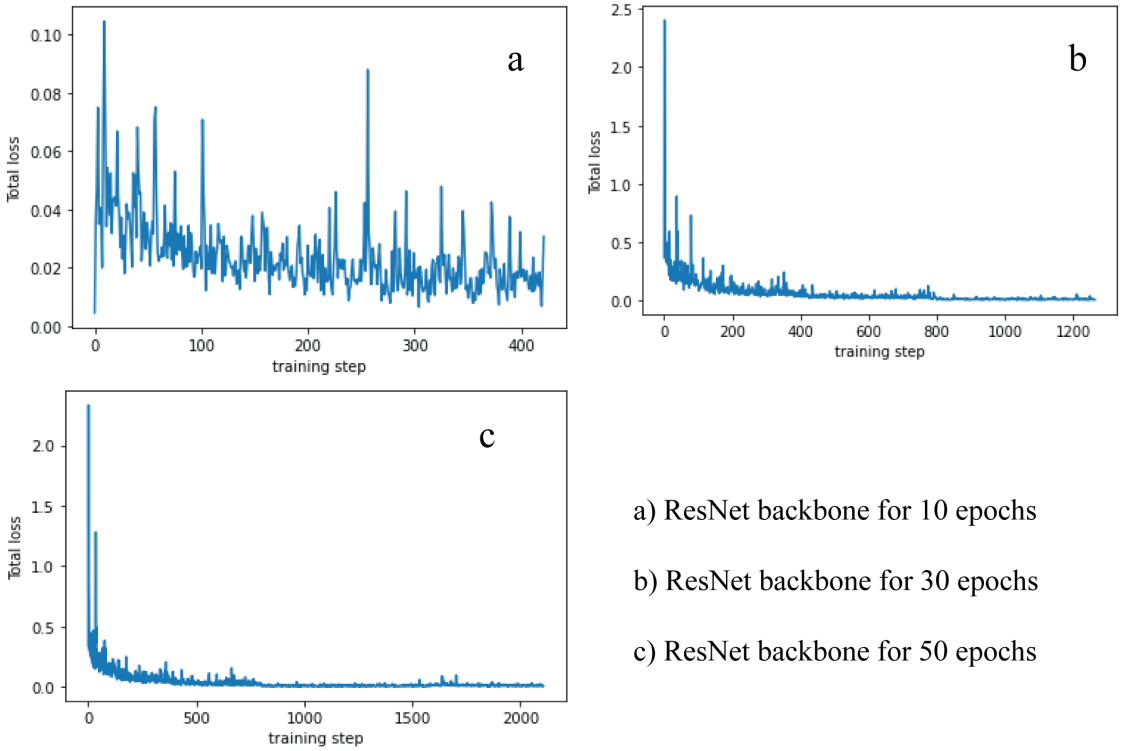


a) Inception backbone for 10 epochs

b) Inception backbone for 30 epochs

c) Inception backbone for 50 epochs

Figure 4.9: *Inception for the first scales and aspect ratios.*

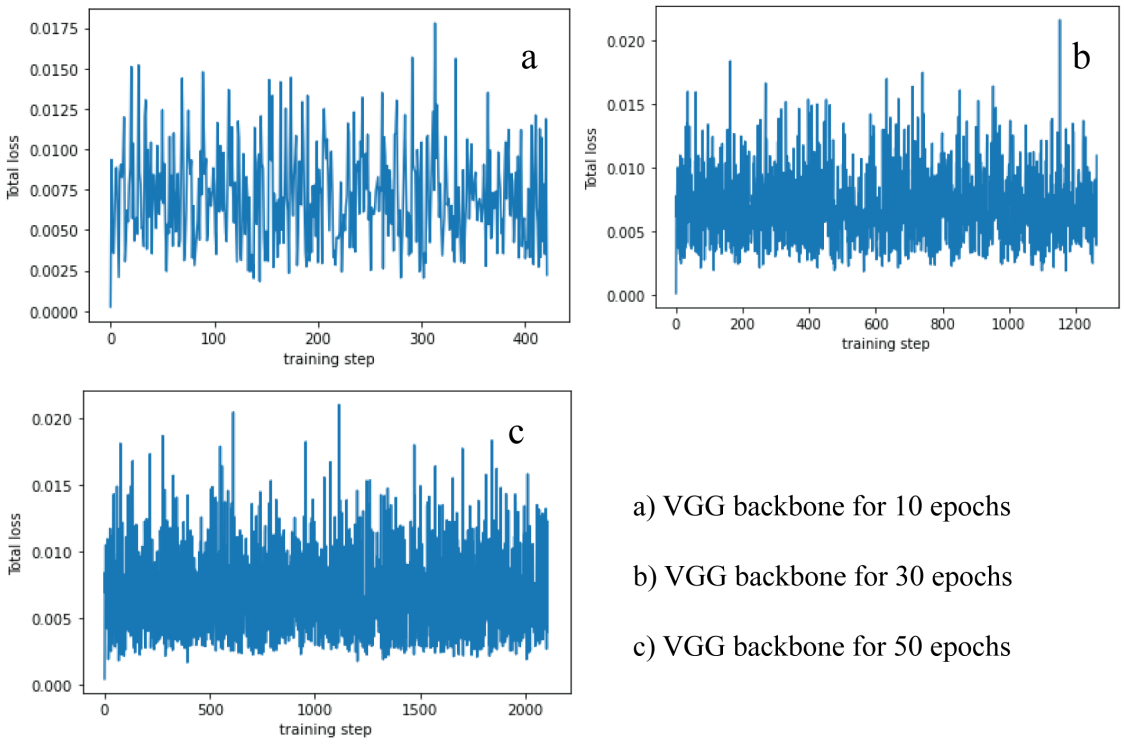


a) ResNet backbone for 10 epochs

b) ResNet backbone for 30 epochs

c) ResNet backbone for 50 epochs

Figure 4.10: *Resnet for the first scales and aspect ratios.*

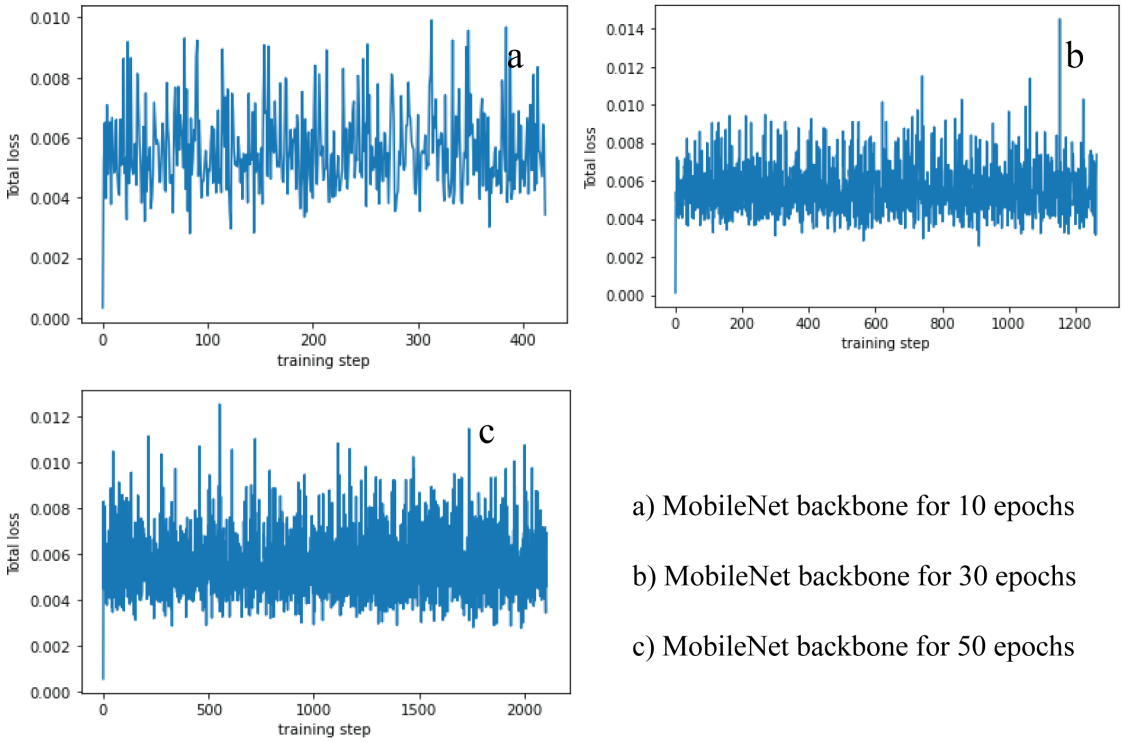


a) VGG backbone for 10 epochs

b) VGG backbone for 30 epochs

c) VGG backbone for 50 epochs

Figure 4.11: *VGG-16 for the second scales and aspect ratios.*

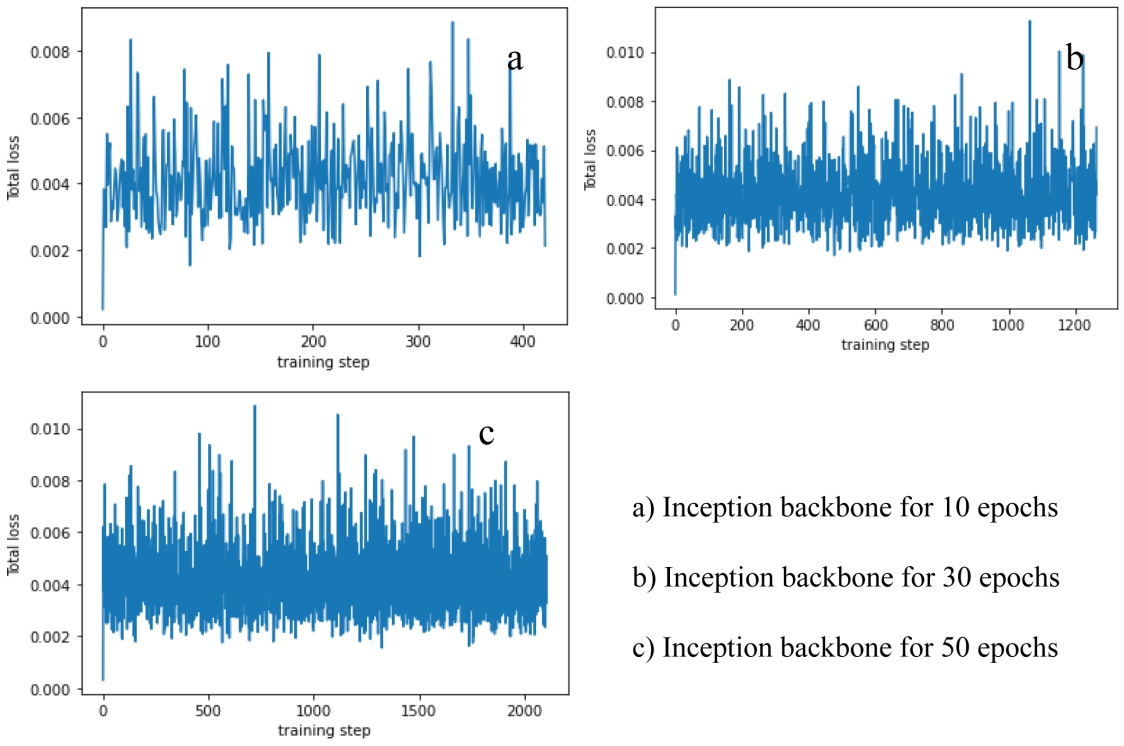


a) MobileNet backbone for 10 epochs

b) MobileNet backbone for 30 epochs

c) MobileNet backbone for 50 epochs

Figure 4.12: *Mobilenet for the second scales and aspect ratios.*



a) Inception backbone for 10 epochs

b) Inception backbone for 30 epochs

c) Inception backbone for 50 epochs

Figure 4.13: *Inception for the second scales and aspect ratios.*

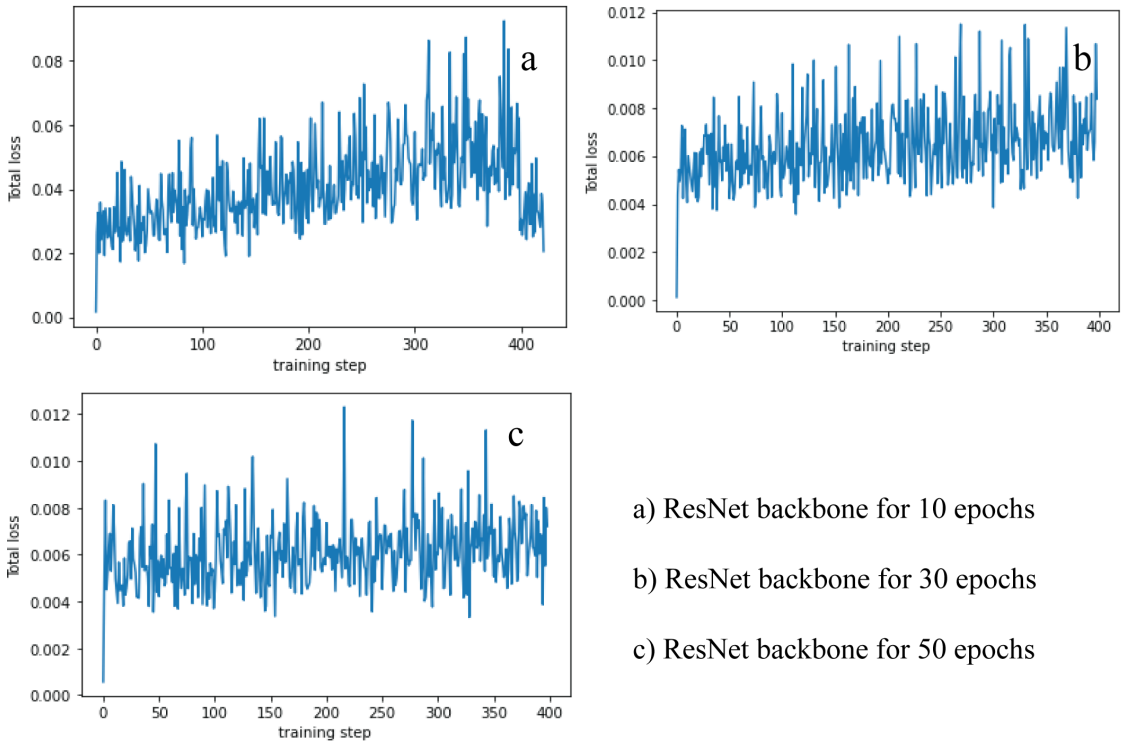


Figure 4.14: *Resnet for the second scales and aspect ratios.*

called "Over-parameterization" [33], where a machine learning model contains too many parameters to train on the data set at hand. This also explains why the Res-Net model converges towards a minimum cost which is higher than that of the less complex models.

4.4 Testing Faster RCNN models

After these models are trained, they need to be tested on both the training and testing sets for analysis. The criterion chosen is the mean average precision (mAP, see appendix B). It is a performance evaluation formula which takes into account the accuracy of the classification as well as the precision of the localization of objects. The testing process is shown in fig. 4.15. The same code for the training process is used for the testing process except for the new added block calculating the mAP. The mAP is calculated using Pytorch built-in functions.

The test results for plate detection are recorded in the following manner: for each set of scales and aspect ratios, the mAP on the training and test sets is tabulated in table 4.1, table 4.2, table 4.3, and table 4.4. Similarly, the same set of results are recorded for the digit recognition models.

The mAP tables shows that the models trained using the first set of anchors and aspect ratios perform far better than the ones using the second set of scales and aspect ratios, which is an obvious result based on the total loss curves. The best result was obtained by

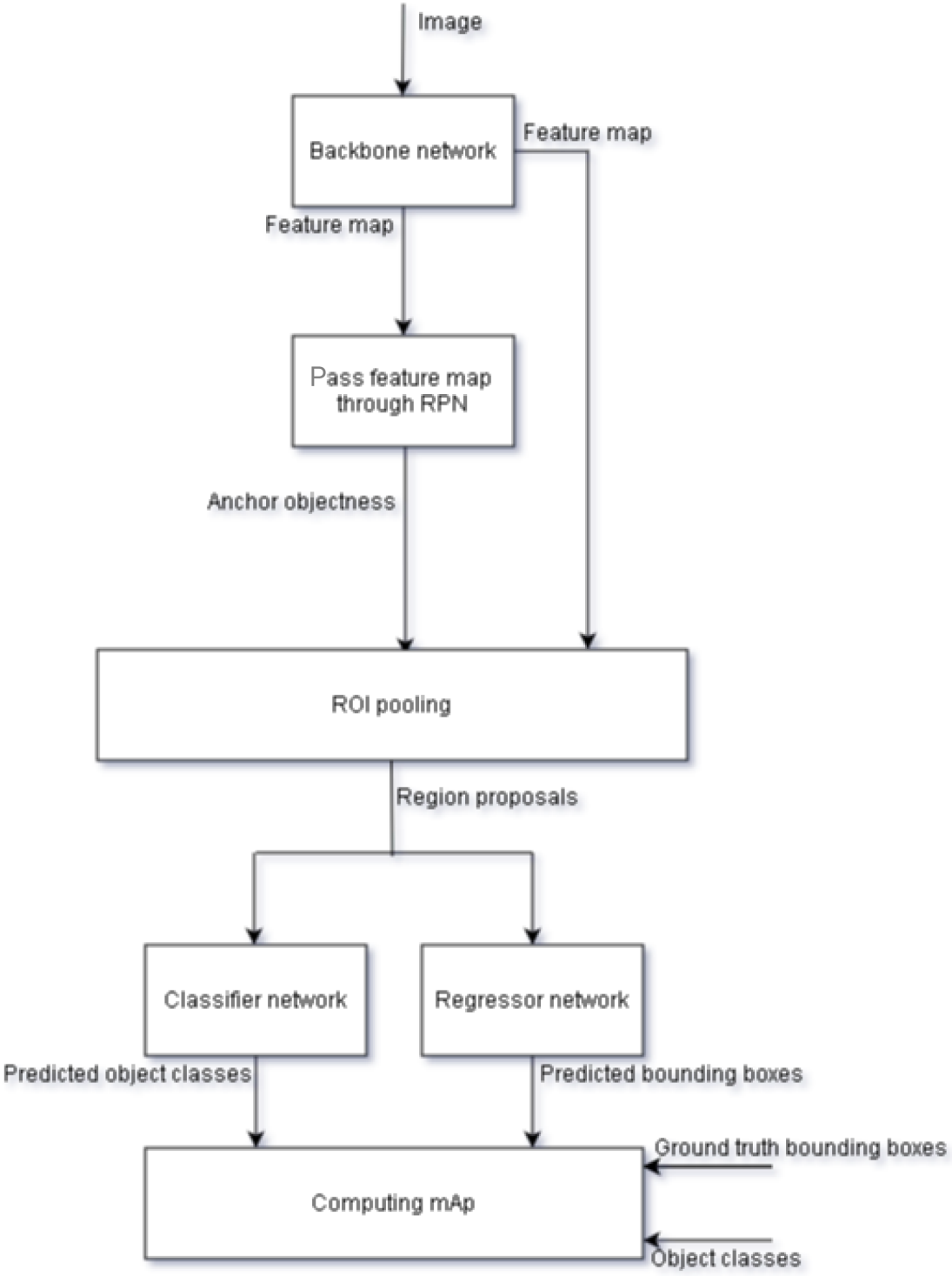


Figure 4.15: *Block diagram summarizing the testing process for the Faster RCNN models.*

Table 4.1: *mAP for plate detection on training set for the first set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	58%	65.1%	77.3%	78.4%
30 epochs	58.2%	65.1%	77%	78%
50 epochs	58%	65%	77%	78%

Table 4.2: *mAP for plate detection on training set for the second set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	27.5%	25.1%	27%	28.4%
30 epochs	27.5%	25.1%	27%	28%
50 epochs	27%	25%	27%	28%

Table 4.3: *mAP for plate detection on test set for the first set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	57.5%	64.1%	75%	70.3%
30 epochs	57.2%	65%	75.1%	70%
50 epochs	57%	65%	75%	71%

Table 4.4: *mAP for plate detection on test set for the second set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	28%	25.1%	27%	28.4%
30 epochs	28.2%	25.1%	27%	28%
50 epochs	28%	25%	27%	28%

Table 4.5: *mAP for digit recognition on training set for the first set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	57.8%	64.9%	77.1%	77.4%
30 epochs	57.5%	64.9%	77.1%	77%
50 epochs	57.6%	64.8%	77%	77%

Table 4.6: *mAP for digit recognition on training set for the second set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	25.5%	23.1%	25%	26.4%
30 epochs	25.5%	23.1%	25%	26%
50 epochs	25%	23%	25%	26%

Table 4.7: *mAP for digit recognition on test set for the first set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	57%	64%	74.4%	70%
30 epochs	57%	65%	75%	69.9%
50 epochs	57%	65%	75%	69.9%

Table 4.8: *mAP results for digit recognition on test set for the second set of scales and anchor ratios.*

Number of Epochs	VGG-16	Mobilenet	Inception	Res-Net
10 epochs	26%	23.1%	25%	26.4%
30 epochs	26.2%	23.1%	25%	26%
50 epochs	26%	23%	25%	26%

the model using Inception network as a backbone which is 75% and 74.4% for test sets of plate detection and digit recognition, respectively. For Res-Net, the performance seems to have dropped, which is explained by the fact that Res-Net needs more data to learn the specific features of the objects. The difference between the performance on the training set and testing set is expected since the test set is data which the model has never seen before. What is noticeable is that this difference is bigger for Res-Net backbone models than it is for the other models. This indicates that the Res-Net backbone model has "Over-fitted" [39] to the training set. Since the other models did not over-fit, it is only reasonable to deduce that the over-fitting was due to the higher complexity of the Res-Net based models, which tends to happen with highly complex machine learning models.

4.5 Speed performance evaluation for RCNN models

A Python script has been implemented to infer the processing time of each model. This has been achieved using built-in Python functions. The time complexity of these models is recorded in number of seconds per frame. Keep in mind that the speed of the model does not depend on any hyper-parameter except for the size of the model itself and the input image size. The tests were run on a Windows 10 computer equipped with an i5-6200u

processor and an NVIDIA GPU 930MX. Table 4.9 shows the results for the plate detection models and table 4.10 shows the results for the digit recognition models.

Table 4.9: *One frame processing time of each plate detection model.*

VGG-16	Mobilenet	Inception	Res-Net
0.3 s	1.1 s	1.6 s	3 s

Table 4.10: *One frame processing time of each digit recognition model.*

VGG-16	Mobilenet	Inception	Res-Net
0.3 s	1.1 s	1.6 s	3 s

4.6 Training YOLOv3 models

Both the plate and digit networks are based on the YOLOv3 architecture (see fig. 4.16) with some modification and parameter choices to fit in with the new data sets. The first change to the base architecture is done to fit the new number of output classes. The original implementation was built to predict the 80 classes from the COCO data set [30]. The plate and digit networks predict one and ten possible classes respectively. As discussed in section 3.1.1, the network divides the input image into an $S \times S$ grid. We must choose the value of S carefully. Indeed, a small value would be good, but then in a data set with overlapping objects, we must choose a higher number of anchors B which are difficult to calculate. A large value of S is preferable since we want each object to fall into its grid cell to be detected alone. This way, we can use a small number of anchor boxes B because many objects are unlikely to fall into the same cell. But doing so will result in a large output volume that makes the training more difficult and very slow. To remedy these problems, the YOLOv3 writers were pretty genius. Instead of having one output, they put three. Each one divides the input image into a different $S \times S$ grid. The outputted S value depends on the input image shape. For an input image shape of 416×416, the three outputs would result in a 13×13 grid at the first output, a 26×26 at the second, and a 52×52 at the third. Objects that have not been detected in one of the outputs are likely to be in the others.

The CNN Darknet-53 (see appendix A) pre-trained on the ImageNet data-set is used as a backbone for both networks. The backbone represents the 52 first layers in the architecture (see table A.2). Since the network is huge, it is impractical to retrain; that would take ages to train due to low computational power available. The last FC layer of Darknet-53 is removed and replaced with seven additional convolutional layers, where

the last one represents the first output. Next, the feature map from the 2 previous layers is taken and upsampled (resized) by 2. A feature map is also taken from earlier in the network and merged by concatenation with the upsampled features. This method enables us to get more meaningful semantic information from the upsampled features and finer-grained information from the earlier feature map. Then, few more convolutional layers are added to process this combined feature map, and eventually predict a similar tensor, although now twice the size. The same design is performed one more time to predict boxes for the final scale. Thus our predictions for the 3rd scale benefit from all the prior computation as well as fine-grained features from early on in the network [30], see fig. 4.16.
Both the plate and digit networks have been trained with a 0.001 learning rate using Adam optimizer with a momentum of 0.9. The plate network has been trained for 4500 iterations whereas the digits network for 8200 iterations. The number of iterations depends on the number of classes the network is trained on. The more classes there are the more the number of iterations are required. Data augmentation has been used a lot through the training process. Specifically, for every 10 iterations, both networks randomly choose a new image dimension size, changes the saturation, the exposure, and the hue of the images. This regime forces the networks to learn to predict well across a variety of input dimensions and gain in robustness. Both networks have been trained in the cloud using the free google service called COLAB. It offers 12 hours access to a virtual machine equipped with an NVIDIA Tesla K80 GPU [11].

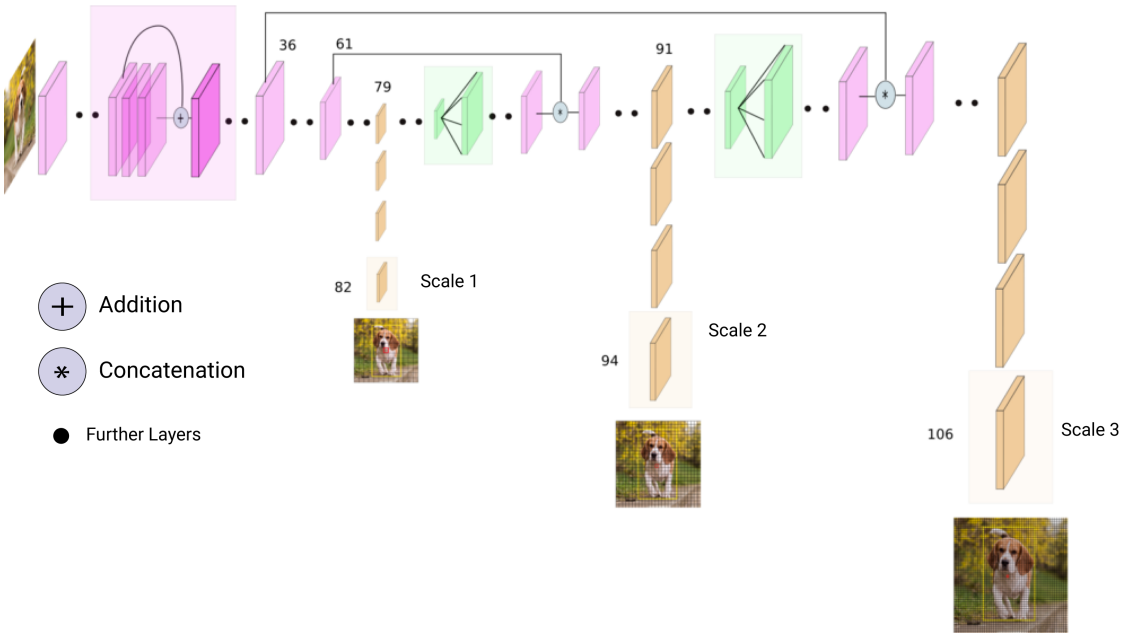


Figure 4.16: *YOLOv3 network architecture [2].*

The training process of both the plate and digit network has been realized using a deep learning framework called Darknet [9]. Darknet has been developed by the YOLO designers in order to train their own YOLO model. The framework is an open-source

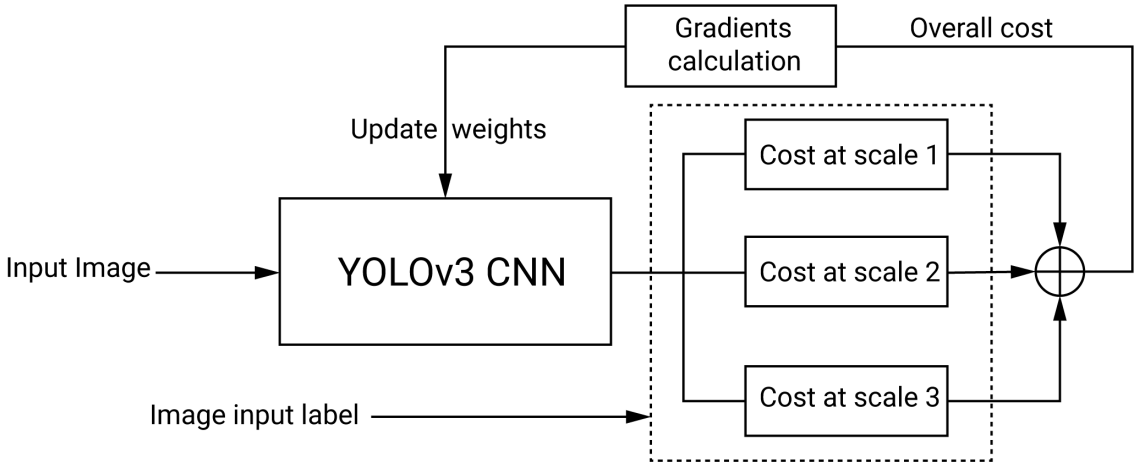


Figure 4.17: *Block diagram illustrating the training process for YOLO models.*

implementation hosted at `github.com`. Along with the framework, they provide detailed instructions for developers to train their own customized YOLO models.

Before showing how to use the framework, we shall first discuss the process of training. As illustrated by the diagram in fig. 4.17, an input image is fed to YOLOv3 network. It passes through the different convolutional layers and eventually outputs a prediction. Along with the ground truth labels, these predictions are used to compute a cost function at each scale using eq. (3.5). The costs are summed up and used to compute the gradients and update the weights and biases of the network as described in chapter one.

In order to use the framework, some specific files must be changed to fit the new developers model configuration. By configuration, we mean the number of classes that the model at hand is supposed to predict. Since the original implementation was built to predict 80 classes, the output volume predict by the network is $S \times S \times (3 * (5 + 80))$, whereas for the digit network for example, the number of classes to predict is 10, therefore the network must predict an $S \times S \times (3 * (5 + 10))$ output volume. The Darknet developers provide a ".cfg" file that contain the model configuration; all the layers of the network are listed in this file. Namely, to indicate a convolutional layer with its input parameters: size, filters, stride, pad, and activation; where "size" represent the size of the filters applied (e.g. 3×3, 1×1, ...), "filters" represents the number of output filters constituting the output feature map, "stride" is the stride to use, "pad" is a boolean that indicates to either use padding or not. Finally, activation represents the activation function to use. In this case, only the leaky ReLU (defined by $f(x) = max(0.1x, x)$) and the linear activation functions are used. The majority of the layers are not to be changed since they represent the implementation of YOLOv3. The only layers to change are the layers designated by "yolo" and the convolutional layer just before it to fit our own custom network. The convolutional layer just before the *yolo* layer represents one of the three outputs of the network. Therefore, the filters parameter has to be changed in order to fit the new number of output classes. In our case, for the digit network for example, the output volume as

discussed previously is $S \times S \times (3 * (5 + 10))$, meaning $S \times S \times 45$; the number of output filters outputted by this convolutional layer is therefore 45. This implies that the filter parameter should be set to 45 and this in all three output layers of the network.
The *yolo* layer in the other hand contains the following parameters:

- **anchors**, represent the initial hand-picked anchor boxes to be adjusted during training. Each pair represents the width and the height of the anchor box respectively.
- **mask**, is a boolean array representing the set of anchor boxes to use at each output.
- **classes**, represent the number of output classes. In our example it is ten.

After downloading the Darknet framework into the COLAB environment, compiling it, and configuring all the necessary files (the one presented above is the most important one, other files are just system configuration), the training can start.

4.7 Training and testing results

4.7.1 Plate detection network

The plate network has been trained on 630 manually labeled images. It has been trained for 1400 iterations. The loss went down to 0.123 with a mAP of 97.0% (see fig. 4.18) on the test set, where 270 images have been used. But after inspection, we find out that the model struggled with the case of multiple plates in the same image and some fuzzy images. We restarted training for another 2500 iterations from the same point in the hope to get the loss under 0.03, but unfortunately, the training took too long without any noticeable improvements, therefore we stopped it at 0.052 loss with mAP of 97.4% (see fig. 4.19). In fact, these are pretty good results, and further improvements can be made with more training data and hopefully our initial problems where solved. Table 4.11 summarizes the different results mentioned above.

Table 4.11: *Plate network results using YOLOv3 model.*

Number of Epochs	Loss	mAP
1400 epochs	0.123	97.0%
4000 epochs	0.052	97.4%

Figure 4.20 shows some examples of plate detection using our network: the ground truth boxes are colored with violet, whereas the network predictions are yellow. As we can see in fig. 4.20, the plate in the example label *b* has not been detected at all. The problem encountered here is probably due to the shape and the color of the plates which are rather unusual; they don't appears often in the training data set.

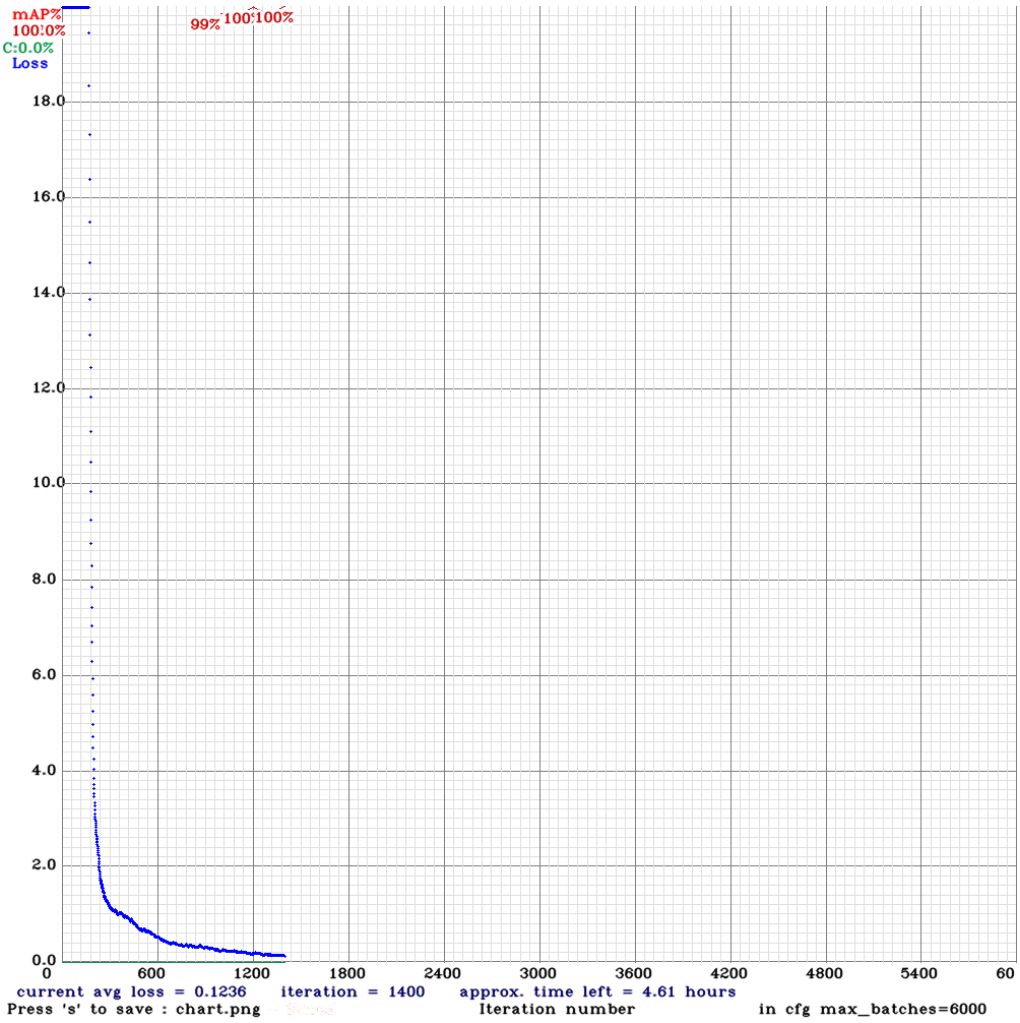


Figure 4.18: *Loss plot for the plate network after 1400 iterations.*

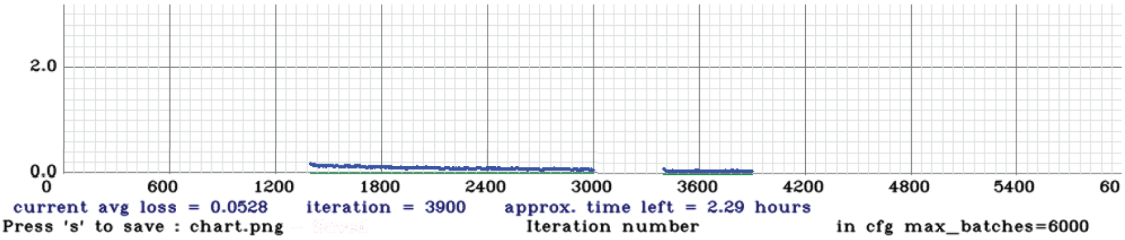


Figure 4.19: *Loss plot for the plate network after 4000 iterations. The erased blue portion was due to unstable internet connection.*

Figure 4.20: *Plate detection examples.*

4.7.2 Digit recognition network

The digit network has been trained on 845 images containing 10,402 hand annotated digits. As the plate network, it has been trained for 4000 iterations at the beginning which brought the loss to 0.8205 with an mAP of 64.2% on a test set containing 145 images with 1453 annotations. The network took so long to get there, but unfortunately doing horribly on the test set. We restarted training for another 4000 iterations and finally got the network to 0.567 loss with a 65.1% mAP on the test set (the plots of the plate and digit network are very similar, therefore there is no need to show them). Table 4.12 summarizes the results mentioned above. Figure 4.21 shows some examples of digit detection using our network. Our model struggles a lot with diagonal images due to the lack of examples.

Table 4.12: *Digit network results using YOLOv3 model.*

Number of Epochs	Loss	mAP
4000 epochs	0.8205	64.2%
8000 epochs	0.567	65.1%

Figure 4.21: *Digit detection examples.*

4.8 Speed performance evaluation for YOLO models

The networks have been tested in the cloud on an NVIDIA Tesla K80 GPU. The speed of networks mainly depends on the input image size; the bigger the image the slower the processing time. Our networks both resize the input images to 416 by 416 pixels keeping their aspect ratios. All speed tests performed in this project disregard the time it takes to save images to the disk and load the networks into the memory. Only the processing time used by the networks is measured, see table 4.13.

Table 4.13: *Speed test summary.*

Network	Average Time
Plate network	26.7 ms
Digit network	26.9 ms
Entire system	52.6 ms

The plate network is very good at detecting plates, even better than the RCNN model and way faster. It achieves around 27 ms per image processed, equivalent to about 37 FPS (frames per second) when running on video streams. This is within the requirements to run on real-time video streams without dropping frames. To achieve such high processing speeds, a relatively modern GPU is required. This is not problematic for real-world applications running on desktop; however for applications running on embedded devices, another lighter architecture should be consider which would reduce mAP significantly.

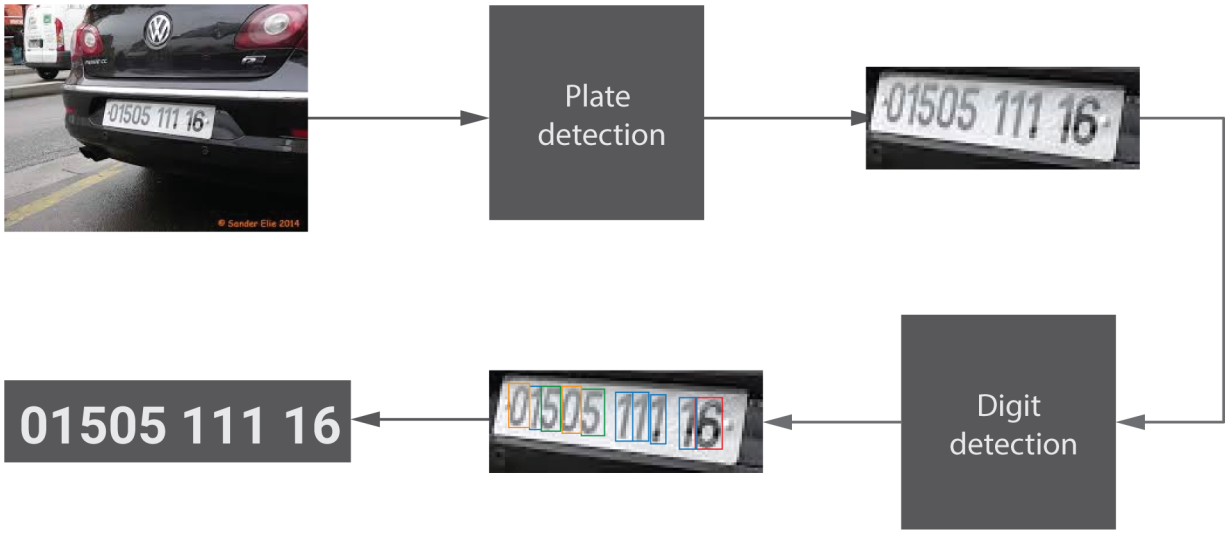


Figure 4.22: *Diagram illustrating the top level structure of the ALPRS.*

The digit network on the other hand, achieved real-time video processing speeds as well, but with a relatively low mAP which is way far from human performance. This is due to the unbalance in the data set. Although the number of training examples is relatively high, some digits are under represented; which is problematic. Therefore, a more balanced dataset would fix the problem. This can be achieved providing more data.

4.9 ALPRS implementation

Implementing the ALPRS consists of concatenating a plate detection model and a digit recognition model to perform the task of license plate recognition from a raw image. Figure 4.22 illustrates this process.

4.9.1 Faster RCNN based ALPRS

The ALPRS application is implemented using a Python script that takes an image as input and passes it through the plate detection model, then uses the output to crop the regions containing the license plates and passes them through the digit recognition model. Both models use the Inception network as backbone as it has the best mAP. The Inception based model is slightly more robust to errors that can be made by the plate detection network, because it checks if the digit recognition model output contains a number of digits that is not consistent with reality. Note that license plates contain from 9 to 11 digits.

For the evaluation of this model, the only relevant characteristics is whether or not it reads the license plate number correctly. After testing on 100 test images containing 114 license plates, 96 license plates were read correctly, therefore, the accuracy is **84.21%**.

4.9.2 YOLOv3 based ALPRS

A similar ALPRS application using the YOLO model is built. It uses a Python script that can be found at `https://github.com/netvor-73/Lpd`. An image is loaded into the YOLO plate detection model, crops the license plates and passes them to the YOLO digit recognition model. It detects the individual digits of each license plate and sorts them according to their location on the images. The final system using YOLOv3 has been tested on 100 images containing 114 license plates, 81 were correctly identified, yielding an accuracy of **71.05%**. Figure 4.23 are some examples of plate and digit detection using the final system.

Figure 4.23: *Examples of ALPRS detection.*

4.9.3 Hybrid model based ALPRS

Both proposed methods in this work achieved decent accuracies as well as acceptable processing speeds. But as it has been shown, the RCNN method despite its better accuracy (84.21%) suffers from relatively low processing speeds. On the other side, disregarding its relatively low accuracy (71.05%), YOLO achieves real times video processing capabilities. Therefore, with the aim to compensate for each other's weaknesses, a hybrid model between both methods seems the way to go. Indeed, the YOLO plate detector has been proven to be really efficient in terms of accuracy and speed, whereas the RCNN digit

network has proven its robustness to extremely challenging conditions. Given these two pieces of information, these last networks have been combined together in a final Python pipeline implementing the flow diagram illustrated in fig. 4.22. The hybrid method runs on real-time video streams, with an accuracy of **81.36%** which almost meets human performance.

The proposed method is fairly general. It could be easily adapted to many real-life scenarios without changing the data set. It has been proven to work in very harsh conditions. Indeed, with a relatively controlled environment; good camera placement, and decent lighting conditions, the method would achieve super human-performance quite easily. In addition, the method could, also, be adapted to different data sets without major changes. The system is generic given that it does not use any hand-crafted features but rather learn them, which is quite powerful.

The accuracy of 81.36% obtained in the application testing is considered a good result for a first attempt. Although there are many ways to improve this result by working on:

- Building a bigger data set;
- Using more modern techniques and deep learning models;
- Encoding the models as C++ data structures to improve inference speed;
- Using CUDA programming language to optimize computations on GPU [17];
- Using unsupervised learning techniques to ensure that the model keeps learning from its mistakes even after the application deployment;

There are other ways to correct the short-comings of the application without any further elaboration on the model. For instance, the application can provide the model with many frames of the same car, thereby making sure that if the model makes a mistake in one frame, it can correct it in another. The accuracy can also be boosted, in a human-assisted environment with the addition of a warning system. It can be easily achieved by thresholding the number of detected digits; whenever this number is less than a certain threshold, the system warns the human-assistance of a wrongly detected plate.

Conclusion

This work was an attempt to create a practical application to be used for license plate detection and recognition. First, the data set was built from the ground up, including both data collection and data labeling processes. Second, the Faster-RCNN and YOLO models were selected as main tools for this task. Afterwards, a number of different backbone models with different structures and parameters were trained and analyzed in order to explore the effects of these parameters on the outcomes of the model, and to figure out the best methods in the building of the application. These models were tested on new real world data and have shown accuracies of 84.21% and 70.3% for Faster-RCNN and YOLO respectively. Finally, in the strive to achieve real world applications requirements, both methods have been combined.

In the future, this project can be improved by working on many aspects of the development such as the ones mentioned in the last chapter. It can also be refined to be a useful tool in the hands of different institutions, businesses, and even law enforcement or security organizations.

In fact, this work is being optimized and integrated into a real-world application used for monitoring garbage trucks under the supervision of a tech start-up in Algiers called Brainiac. A direct quote from a Co-Founder of this company states the following: "Introducing similar technologies into the Algerian market is not only a savvy business idea but also a great opportunity for researchers and young software developers to turn their innovations into real-world applications". And finally, it would also be insightful to mention the following quote from Ray Kurzweil - American inventor and futurist -: “Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, and we will have multiplied the intelligence a billion-fold.”

Bibliography

[1] https://classroom.udacity.com/courses/ud188/lessons.

[2] https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b.

[3] https://www.coursera.org/learn/deep-neural-network?skipBrowseRedirect=true.

[4] https://www.mathworks.com/help/vision/ug/anchor-boxes-for-object-detection.html.

[5] https://www.coursera.org/learn/convolutional-neural-networks?

[6] http://host.robots.ox.ac.uk/pascal/VOC/.

[7] https://pytorch.org/.

[8] https://opencv.org/.

[9] https://github.com/AlexeyAB/darknet#how-to-train-to-detect-your-custom-objec

[10] https://drive.google.com/drive/folders/16K68eRaGJHb3WB7Tfx-0s0h1gVQmmMSW?usp=sharing.

[11] https://colab.research.google.com.

[12] Josh Patterson Adam Gibson. *Deep Learning: A Practitioner's Approach*. O'Reilly, 2017.

[13] Patrick van der Smagt Ben Krose. *An introduction to neural networks*. The University of Amsterdam, November 1996.

[14] Sergey Ioffe Jonathon Shlens Zbigniew Wojna Christian Szegedy, Vincent Vanhoucke. Rethinking the inception architecture for computer vision. *arXiv:1512.00567*, 11 Dec 2015.

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

[16] Shane Cook. *CUDA Programming: A Developer's Guide to Parallel Computing with GPUs*. Newnes, 2012.

[17] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[18] Jimmy Lei Ba Diederik P. Kingma. Adam: A method for stochastic optimization. *arXiv:1412.6980v9*, January 2017.

[19] Jarek Duda. Sgd momentum optimizer with step estimation by online parabola model. *arXiv:1907.07063*, Dec 2019.

[20] Alexey Bochkovskiy et al. Yolov4: Optimal speed and accuracy of object detection. *arXiv:2004.10934v1*, April 2020.

[21] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[22] *et al.* G. Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 17 Apr 2017.

[23] Ross Girshick. Fast r-cnn. *arXiv:1504.08083*, Sep 2015.

[24] Leonardo Ferreira Guilhoto. An overview of artificial neural networks for mathematicians. *http://math.uchicago.edu/ may/REU2018/REUPapers/Guilhoto.pdf*.

[25] Aaron Courville Ian Goodfellow, Yoshua Bengio. *Deep learning*. MIT Press, 2017.

[26] Sajayasa Ari Dwi Suta Atmaja INGA Astawa, I Gusti Ngurah Bagus Caturbawa. Detection of license plate using sliding window, histogram of oriented gradient, and support vector machines method. *J. Phys.: Conf. Ser. 953 012062*, 2018.

[27] Yoshua Bengio Jason Yosinski, Jeff Clune and Hod Lipson. How transferable are features in deep neural networks? *arXiv:1411.1792v1*, November 2014.

[28] Hogne Jorgensen. *Automatic License Plate Recognition using Deep Learning Techniques*. PhD thesis, Norwegian University of Science and Technology, Department of Computer Science, July 2017.

[29] Ali Farhadi Joseph Redmon. Yolo9000: Better, faster, stronger. *arXiv arXiv:1612.08242v1*, December 2016.

[30] Ali Farhadi Joseph Redmon. Yolov3: An incremental improvement. *arXiv arXiv:1804.02767v1*, April 2018.

[31] Ross Girshick Ali Farhadi Joseph Redmon, Santosh Divvala. You only look once: Unified, real-time object detection. *arXiv arXiv:1506.02640v5*, May 2016.

[32] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. *arXiv:1512.03385*, Dec 2015.

[33] Zheng Xu W. Ronny Huang Tom Goldstein Karthik A. Sankararaman, Soham De. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. *arXiv:1904.06963*, 15 Apr 2019.

[34] Jude Shavlik Lisa Torrey. Transfer learning. *University of Wisconsin, Madison WI, USA*, 2009.

[35] S.A. Ghazi M. Sarfraz, M.J. Ahmed. Saudi arabian license plate recognition system. *IEEE*, 2012.

[36] Michael Nielsen. *Neural Networks and Deep Learning*. Online book: http://neuralnetworksanddeeplearning.com/, December 2019.

[37] D. Renuka devi and D. Kanagapushpavalli. Automatic license plate recognition. In *3rd International Conference on Trendz in Information Sciences Computing (TISC2011)*, pages 75–78, 2011.

[38] Trevor Darrell Jitendra Malik Ross Girshick, Jeff Donahue. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524v5*, October 2014.

[39] Xiuwen Liu Shaeke Salman. Overfitting mechanism and avoidance in deep neural networks. *arXiv:1901.06566*, 19 Jan 2019.

[40] Ross Girshick Shaoqing Ren, Kaiming He and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv:1506.01497v3*, June 2016.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, Sep 2014.

[42] Serge Belongie Lubomir Bourdev Ross Girshick James Hays Pietro Perona Deva Ramanan C. Lawrence Zitnick Piotr Dollár Tsung-Yi Lin, Michael Maire. Microsoft coco: Common objects in context. *arXiv:1405.0312*, 1 May 2014.

Appendices

Appendix A

YOLOv3 network architecture

A.1 Feature extractor

YOLO developers didn't use any of the already pre-trained backbone networks on image classification. Instead they trained their own classifier on the ImageNet data-set. The network uses successive 3×3 and 1×1 convolutional layers with some shortcut (skip) connections. It has 53 convolutional layers, therefore it has been named Darknet-53. Darknet is the deep learning framework used to train it. See table A.2

Darknet-53 network is a pretty good one compared to other backbones that are even deeper which makes it way faster at inference time. See table A.1. Each network is trained with identical settings and tested at 256×256, single crop accuracy. Run times are measured on a Titan X at 256×256. Thus Darknet-53 performs on par with state-of-the-art classifiers but with fewer floating point operations and more speed. Darknet-53 is better than ResNet-101 and $1.5\times$ faster. Darknet-53 has similar performance to ResNet-152 and is $2\times$ faster. Darknet-53 also achieves the highest measured floating point operations per second. This means the network structure better utilizes the GPU, making it more efficient to evaluate and thus faster. That's mostly because ResNets have just way too many layers and aren't very efficient [30].

Table A.1: *Comparison of backbones. Accuracy, billions of operations, billion floating point operations per second, and FPS for various networks [30].*

Backbone	**Top-1**	**Top-5**	**Bn Ops**	**BFLOP/s**	FPS
ResNet-101	77.1	93.7	19.7	1039	53
ResNet-152	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	**78**

Table A.2: *Darknet-53 structure[30].*

	Type	Filters	Size	Output
	Convolutional	32	3×3	256×256
	Convolutional	64	3×3 /2	128×128
	Convolutional	32	1×1	
$1 \times$	Convolutional	64	3×3	
	Residual			128×128
	Convolutional	128	3×3 /2	64×64
	Convolutional	64	1×1	
$2 \times$	Convolutional	128	3×3	
	Residual			64×64
	Convolutional	256	3×3 /2	32×32
	Convolutional	128	1×1	
$8 \times$	Convolutional	256	3×3	
	Residual			32×32
	Convolutional	512	3×3 /2	16×16
	Convolutional	256	1×1	
$8 \times$	Convolutional	512	3×3	
	Residual			16×16
	Convolutional	1024	3×3 /2	8×8
	Convolutional	512	1×1	
$4 \times$	Convolutional	1024	3×3	
	Residual			8×8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Appendix B

mAP (mean Average Precision) for Object Detection

Average precision (AP) is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, SSD, etc. To understand it, its better to recap some related concepts.

B.1 Precision and recall

Precision measures how accurate are your predictions. i.e. the percentage of your correct predictions. Recall measures how good you find all the positives. For example, we can find 80% of the possible positive cases in our top K predictions. Here are their mathematical definitions:

$$Precision = \frac{TP}{TP+FP} \tag{B.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{B.2}$$

where TP stands for true positive, FP for false positive, TN for true negative, and FN for false negative.

B.2 Average Precision

Let's create an over-simplified example that demonstrates the calculation of the average precision. In this example, the whole dataset contains 5 apples only. We collect all the predictions made for apples in all the images and rank it in descending order according to the predicted confidence level, see fig. B.1. The second column indicates whether the prediction is correct or not. In this example, the prediction is correct if the $IoU \geq 0.5$.

Rank	Correct?	Precision	Recall
1	True	1.0	0.2
2	True	1.0	0.4
3	False	0.67	0.4
4	False	0.5	0.4
5	False	0.4	0.4
6	True	0.5	0.6
7	True	0.57	0.8
8	False	0.5	0.8
9	False	0.44	0.8
10	True	0.5	1.0

Figure B.1: *Example of precision and recall values.*

Let's take the row with rank 3 and demonstrate how precision and recall are calculated first. Precision is the proportion of TP = 2/3 = 0.67. Recall = 2/5 = 0.4. Recall values increase as we go down the prediction ranking. However, precision has a zigzag pattern — it goes down with false positives and goes up again with true positives, see fig. B.2.

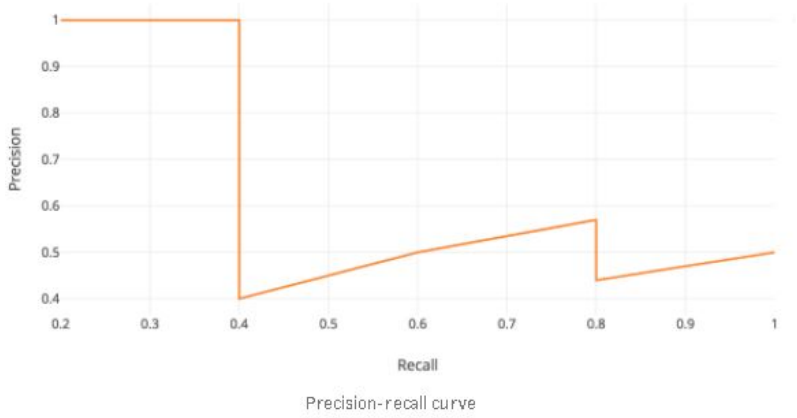


Figure B.2: *Plot of precision vs recall.*

The general definition for the Average Precision (AP) is finding the area under the precision-recall curve.

$$AP = \int_0^1 p(r)dr$$

Precision and recall are always between 0 and 1. Therefore, AP falls within 0 and 1 also. Before calculating AP for the object detection, we often smooth out the zigzag pattern first. Graphically, at each recall level, we replace each precision value with the maximum precision value to the right of that recall level, see fig. B.3.

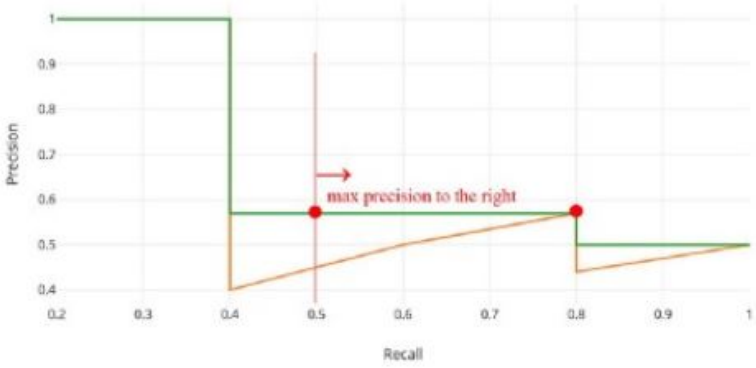


Figure B.3: *Elimination of zigzag pattern.*

Appendix C

Backbone models

In the few last chapters, we introduced the basic building blocks of CNNs such as convolutional layers, pooling layers and fully connected layers. It turns out a lot of the past few years of computer vision research has been on how to put together these basic building blocks to form effective convolutional neural networks, focusing on the *object classification task*. One of the best ways to get intuition on how to build CNNs is to read or to see other examples of effective CNNs. It turns out that an architecture that works well on one computer vision task often works well on other tasks also. Indeed the same networks called **backbones**, discussed in this section are used as feature extractors for object detection networks since an object detection requires the classification of objects and their locations.

C.1 VGG-16

VGG16 is a CNN that achieves 92.7% top-5 test accuracy in ImageNet [41], which is a data-set of over 14 million images belonging to 1000 classes [17]. It was one of the famous model submitted to one of the most prestigious computer vision competition. It makes the improvement over state of the art at that time by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks using graphical processing technology. see fig. C.1.

The ConvNet configurations are outlined in fig. C.2. The nets are referred to their names (A-E). All configurations follow the generic design present in architecture and differ only in the depth: from 11 weight layers in the network A (8 convolutional layers and 3 fully connected layers) to 19 weight layers in the network E (16 convolutional layers and 3 fully connected layers). The width of convolutional layers (the number of channels) is rather small, starting from 64 in the first layer and then increasing by a factor of 2 after each max-pooling layer, until it reaches 512. See fig. C.2.

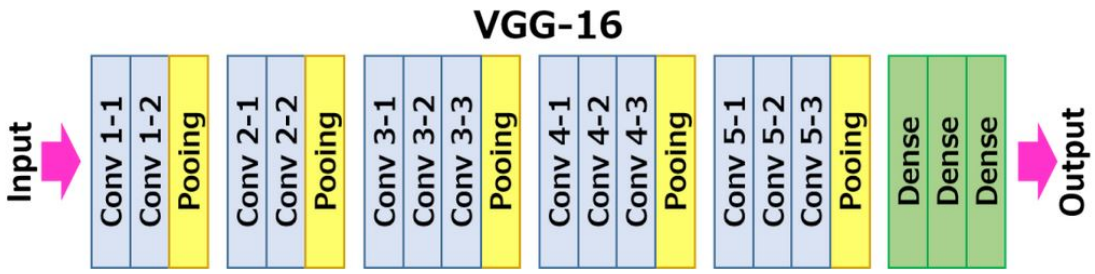


Figure C.1: *VGG-16 structure by layers.*

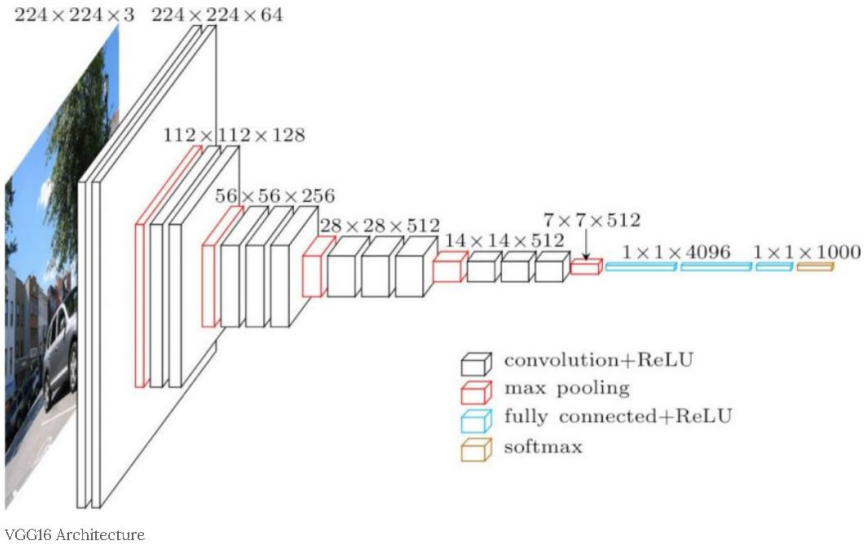


Figure C.2: *VGG-16 network configuration.*

C.2 Mobilenet

MobileNet is a CNN architecture model used for object detection and image classification, generally used in small applications.There exists a variety of models designed for the same purpose but the reason why MobileNet stand out is that it requires much less computation power to run or apply transfer learning to.This characteristic is what makes it optimal to run on embedded systems in general, computer systems without GPU or low computational efficiency, as well as Mobile devices which don't have the necessary hardware to run more costly models. Needless to say, the use of Mobilenet comes with a significant compromise in the accuracy of the results. It is also best suited for web browsers as browsers have limitation over computation, graphic processing , and storage. Mobilenet architecture is distinguished by an essential features know as "Depth-wise Separable Convolution" [22]. Before we get into the definition of "Depth-wise Separable Convolution" we need to go over some aspects of the convolution operation. Let's consider an input matrix of shape $D_f \times D_f \times M$ as shown in fig. C.3. If our input was an RGB image then M would be equal to 3. If we apply a convolution using a filter of shape $D_k \times D_k \times M$ we would obtain an output of size $D_G \times D_G \times 1$ if we apply the same convolution using N filters of the same shape and concatenate the results we would obtain an output shape of $D_G \times D_G \times N$, see fig. C.4.

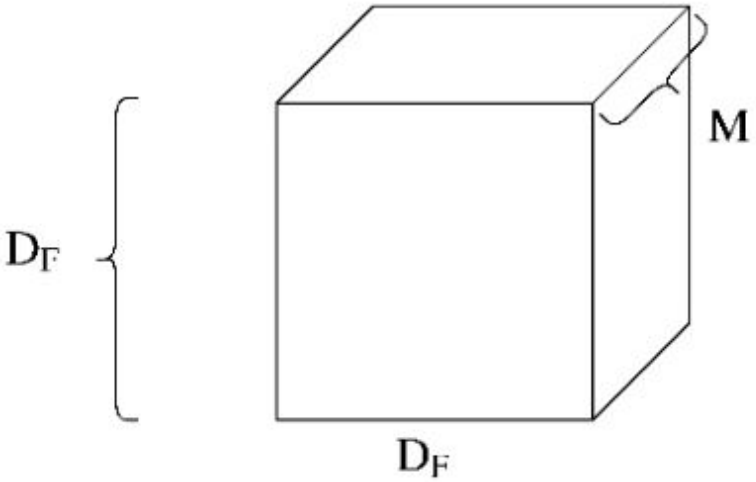


Figure C.3: *Convolution input.*

Since the multiplication operation is more expensive relative to the addition, let's consider the cost of the convolution operation with regards to the number of multiplications. For one convolution step for one kernel the number of multiplications is $D_k \times D_k \times M$, for and entire convolution step for one filter the number of multiplications is $D_G \times D_G \times D_k \times D_k \times M$ therefore when we account for N filters the number of multiplication for a

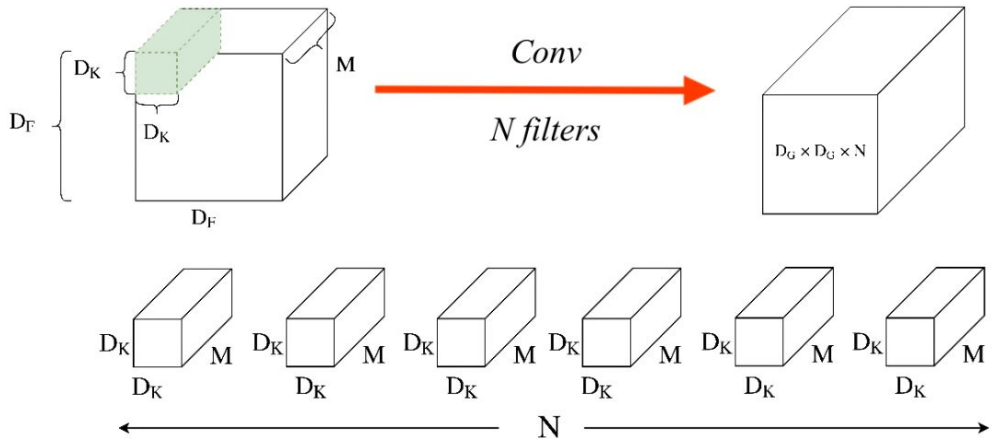


Figure C.4: *Convolution operation.*

convolutional layer is $D_G^2 \times D_k^2 \times M$.

With this in mind, we can introduce the concepts of "Depth-wise Convolution" and "Point-wise convolution", which when put together yield a " Depth-wise Separable Convolution ". Unlike simple convolution, Depth-wise Convolution applies convolution to single input channel at a time, using M filters of shape $D_k \times D_k \times 1$ see fig. C.5. Point-wise convolution applies N filters of shape $1 \times 1 \times M$ to the output of the Depth-wise convolution and by concatenating the results we obtain the same output shape as simple convolution, see fig. C.6 [22].

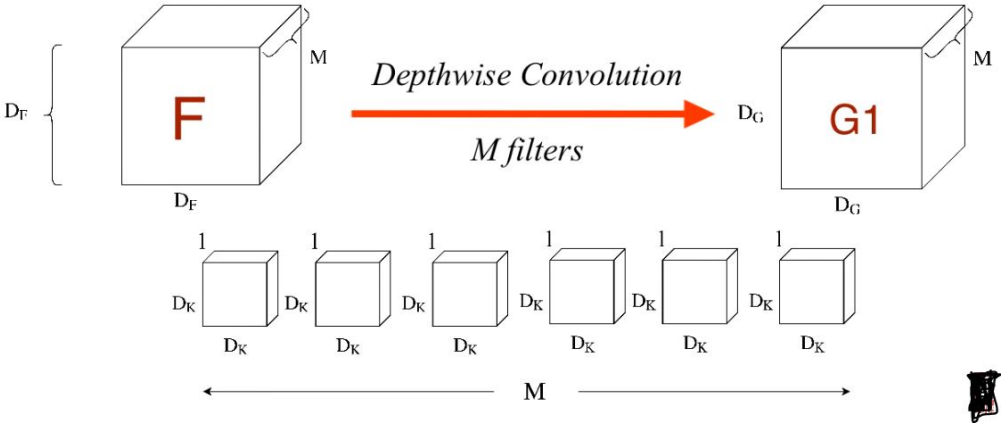


Figure C.5: *Depth-wise Convolution.*

Computing the number of multiplications for the entire process gives $M \times D_G^2 \times (D_k^2 + N)$; which is less than the cost of simple convolution. But to get a an understanding of how much computational power is reduced, we should compute the ratio

$$\frac{number\ of\ multiplications\ for\ DSC}{number\ of\ multiplications\ for\ simple\ conv}$$

which is found to be

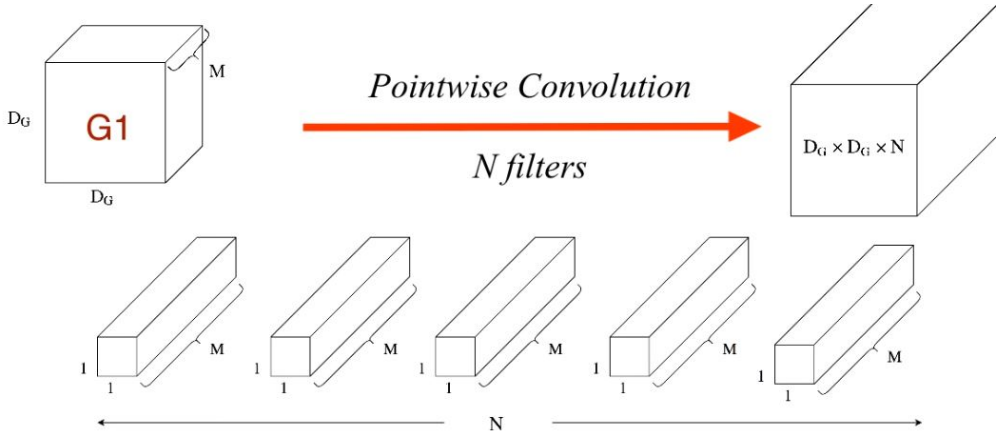


Figure C.6: *Point-wise Convolution.*

$$\frac{1}{N} + \frac{1}{D_k^2} \tag{C.1}$$

By taking an example of $D_k = 3$ and $N = 1024$, we get a ratio of approximately $\frac{1}{9}$ which signifies a substantial decrease in computational requirements.

C.2.1 Mobilenet model structure

The Mobilenet model is composed of convolutional and Max Pool layers where the full structure is demonstrated in table C.1 [22].

C.3 Inception

The Inception network was an important milestone in the development of CNN classifiers. Prior to its inception (pun intended), most popular CNNs just stacked convolution layers deeper and deeper, hoping to get better performance.The Inception network, on the other hand, was complex (heavily engineered). It used a lot of tricks to push performance; both in terms of speed and accuracy. Its constant evolution lead to the creation of several versions of the network. The popular versions are : Inception v1, Inception v2, Inception v3, and Inception Res-Net. Each version is an iterative improvement over the previous one. Understanding the upgrades can help us to build custom classifiers that are optimized both in speed and accuracy.

C.3.1 Inception V1

The problem addressed by the developers of this model is the extreme large variation in the size of the salient parts in the image. For instance, an image of a dog can have any of the forms shown in fig. C.7. The area occupied by the dog is different in each image. This

Table C.1: *mobilenet structure table.*

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

significant variation in the location of the relevant features of the object we wish to detect and classify requires choosing the right kernel size for the convolution operation, which becomes a complicated task. A larger kernel is preferred for information that is distributed more globally, and a smaller kernel is preferred for information that is distributed more locally. Considering the fact that very deep networks are prone to over-fitting in addition to the difficulty they pose in performing back-propagation across the layers it goes without saying that naively stacking large convolution operations is computationally expensive and will not improve network performance on new data.

The authors of the original paper suggested the use of multiple filters of different sizes in one layer rendering the network "wider" rather than “deeper” [15].

Figure C.8 explains the core idea of this model. It performs convolution on an input, with 3 different sizes of filters $(1 \times 1,\ 3 \times 3,\ 5 \times 5)$. Additionally, max pooling is also performed. The outputs are concatenated and sent to the next inception layer.

As stated before, deep neural networks are computationally expensive. To make it cheaper, the authors limit the number of input channels by adding an extra 1×1 convolution before the 3×3 and 5×5 convolutions. Though adding an extra operation may seem counter intuitive, 1×1 convolutions are far less expensive than 5×5 convolutions, and the reduced number of input channels also help (similar to the method used in Mobilenet). Note that, however, the 1x1 convolution is introduced after the max pooling layer, rather than before [14], see fig. C.9

Figure C.7: *Examples of dog images.*

Using the dimension reduced inception module, a neural network architecture was built. This was popularly known as GoogLeNet (Inception v1). The architecture is shown in fig. C.10.

GoogLeNet has 9 such inception modules stacked linearly. It is 22 layers deep (27, including the pooling layers). It uses global average pooling at the end of the last inception module.

Needless to say, it is a pretty deep classifier. As with any very deep network, it is subject to the vanishing gradient problem.

To prevent the middle part of the network from “dying out”, the authors introduced

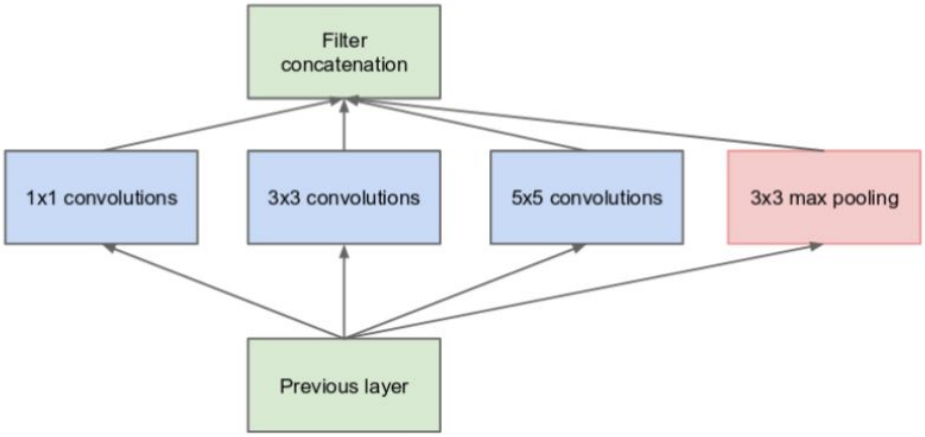


Figure C.8: *Inception network layer general structure.*

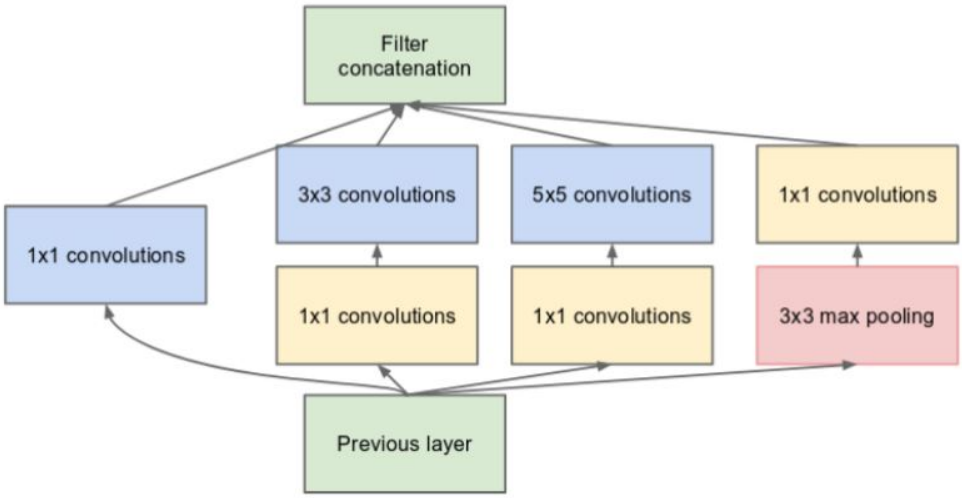


Figure C.9: *Inception layer with added* 1×1 *convolution operation.*

two auxiliary classifiers (The purple boxes in fig. C.10). They essentially applied softmax to the outputs of two of the inception modules, and computed an auxiliary loss over the same labels. The total loss function is a weighted sum of the auxiliary loss and the real loss. The weight value used in the paper was 0.3 for each auxiliary loss [14]. Needless to say, auxiliary loss is purely used for training purposes, and is ignored during testing.

$$totalloss = real\ loss + 0.3\ aux\ loss\ 1 + 0.3\ aux\ loss\ 2 \tag{C.2}$$

C.4 Res-Net

ResNet, short for Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015.

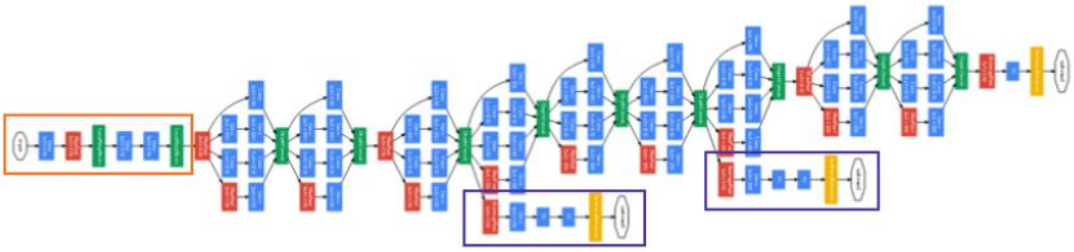


Figure C.10: *Inception network structure.*

The fundamental breakthrough with ResNet was that it allowed us to train extremely deep neural networks with 150+layers successfully [32]. Prior to ResNet, training very deep neural networks was difficult due to the problem of vanishing gradients.

AlexNet, the winner of ImageNet 2012 and the model that apparently kick started the focus on deep learning had only 8 convolutional layers, the VGG network had 19 and Inception or GoogleNet had 22 layers and ResNet 152 had 152 layers.

However, increasing the network depth does not work by simply stacking layers together. Deep networks are hard to train because of the notorious vanishing gradient problem as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient extremely small. As a result, as the network goes deeper, its performance gets saturated or even starts degrading rapidly. For this reason the creators of Res-Net introduced the idea of "Skip Connections"

C.4.1 Skip Connection

ResNet first introduced the concept of skip connection fig. C.11. illustrates skip connection. The figure on the left is stacking convolution layers together one after the other. We still stack convolution layers as before but we now also add the original input to the output of the convolution block. This is called skip connection. We must note that the addition operation occurs before the output goes through the ReLU (Rectification linear unit) function [32]. The main two reasons why skip connections work are:

- Skip connections mitigate the problem of vanishing gradient by allowing this alternate shortcut path for gradient to flow through.
- They allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer, and not worse.

Figure C.11: *Skip connection.*