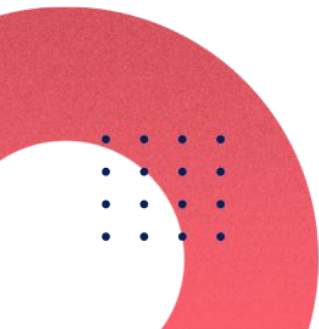


Part 2/2



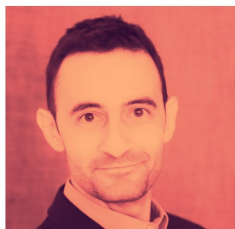
No Mercy for Manual Entry

29/Sep/2021
Workshop @ AMLD2021



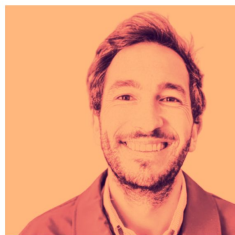
Samurai

Authors



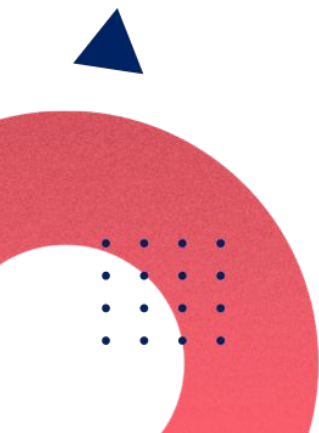
Valerio Rossetti, PhD

Co-founder of SamurAI
Senior Data Scientist



Giulio Grossi, PhD

Senior Quantitative Portfolio Manager
at ONE swiss bank



Extract text from documents

We will revise **different techniques** to extract information from documents, that cover a good variety of business cases.

Different level of Difficulty based on the type of document and on the information to extract:

- 1: Documents with Items in fixed positions ●
- 2: Scanned Documents ●
- 3: Documents Taken From Camera ●
- 4: Documents with Items in variable position ●

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

To
Name: James Buti
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbuti@gmail.com

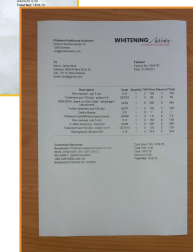
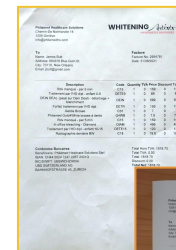
WHITENING *Artists*
HYGIENISTES DENTAIRES

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dipl.	DETR	1	0	159	0	159
Geldis Brosse	C31	1	0	7	0	7
Philamed Gold4White brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par l'HD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonnées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: USSWCH2Z80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70





N.B.

Each business problem has its own formulation and specificity.
Therefore, the techniques applied will vary and adapt to each case.

Not all of them require Deep/Machine Learning!
aka

Keep the process as easy as possible

The libraries we'll be using



Alongside many others ...



1: Document with items in fixed position

● Difficulty : Padawan

Business case:

Giulio Cornelio Grossi, a young asset manager just passed a trade to his broker, a big Swiss brokerage firm. The brokerage firm sends to ONE swiss bank back-office a confirmation ticket with the information regarding the transaction. The information on the ticket is always in the same position.

ONE swiss bank wants to automate the process of database feed.

The solution:

Define Regions of Interest (Rols).

Use Tesseract to extract the text in the Rols.

BOURSE

VENTE "TO CLOSE"

Vente -5 FUT FT100 10 IFF 21/09/17 à GBP 0.00

SORTIE

FUT FT100 10 IFF 21/09/17

No ISIN: GB00JRR84H81 No Telekurs: 81574497

EFFET CASH

Montant brut	GBP	0.00
Courtage	GBP	-25.00
Montant net	GBP	-25.00

Veuillez noter:

agit pour cette transaction en son propre nom en tant qu'agent, mais en votre nom et à vos propres risques.

INFORMATIONS COMPLÉMENTAIRES

Général	Date de transaction	10.09.2021
	Date valeur	13.09.2021
	Date comptable	10.09.2021
Référence	Date de l'ordre	10.09.2021 à 16:33:34
	Type d'ordre	Au marché
	Date d'expiration	17.09.2021 à 23:59:59
Exécution	Type d'exécution	Vente
	Direction	To close
	Place de bourse	ICE Futures Europe - LIFFE, Financials Products, Londres (IFLL)
Dérivé	Genre valeur	Future - Indice
	Type d'option	Special
	Expiration	Short
	Taille du contrat	10
	Sous-jacent	INDICE FTSE 100
	Style	Européen
	Type d'exercice	Uniquement à la date d'expiration
	Date d'expiration	17.09.2021
	Type de conversion	Cash settlement
	Date de conversion	17.09.2021
	Quantité engagée	50
	Prix	GBP 7'024.50
	Montant engagé	GBP 351'225.00
Autre	Dépositaire	

S.E. & O.

Avis sans signature

1/1



1: Document with items in fixed position

● Difficulty : Padawan

Define Rol:

A Rol is a simple set of pixel coordinates that define a region of an image. Usually is a rectangle:

`(x_top_left,y_top_left), (x_bottom_right,y_bottom_right)`

Select Rol:

Remember: an image is a numpy array of size `(h,w,3)` so selecting a Rol is as easy as slicing a numpy array!

Extract the text using Tesseract:

We can use the `image_to_string()` method. There are others, sometimes more effective methods, we will be covering in the practise session.

`(x_start,y_start)`

I am a simple text to be
extracted with
Tesseract

`(x_end,y_end)`



```
roi = image[y_start:y_end,x_start:x_end]
text = pytesseract.image_to_string(roi)
```



1: Document with items in fixed position

● Difficulty : Padawan

1

Read First Name and Last Name from Registration Form

WHITENING *Artists*
DENTISTES DENTAIRES

Personal Info

First Name
James

Last Name
Butt

Street and N.
6649 N Blue Gum St

ZIP
70116

City
New Orleans

Email
jbutt@gmail.com

Phone
504-621-8927

James

Butt

2

Create folder



JamesButt



Read First Name and Last Name from Invoice and copy to the corresponding folder

3

Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philameds.ch

WHITENING *Artists*
DENTISTES DENTAIRES

To
Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dipl.	DETR	1	0	159	0	159
Gaëlle Brosse	C31	1	0	7	0	7
Philamed Gold&White brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par l'HD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC: SWIFT: UBSWCH33HAN
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZÜRICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70





2: Scanned Documents

● Difficulty : Novice

What if the document is scanned?

We can still apply the same methodology of extracting Rols, as we did for 1. We will encounter some difficulties though, that will make the method more unstable and less efficient.

This part is left for exercise in the Notebook, to be taken home or if we still have time after the last exercise.

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING *Artist*
HYGIENES DENTAIRES

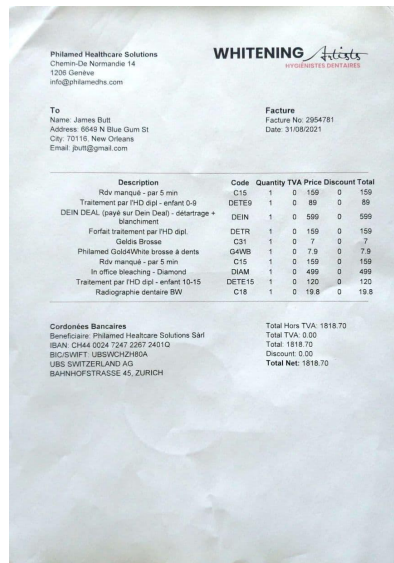
To
Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN IDEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dipl.	DETR	1	0	159	0	159
Gelis Brosses	C31	1	0	7	0	7
Philamed GoldWhite brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par l'HD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonnées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2287 2401 Q
BIC: SWIFT: UBSWCH2H8A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70





● **Difficulty : Master**

Things gets way more complicated. The document always changes position and orientation. We need to find a way to detect the document and align it.

The solution:

- Detect edges in the image.
- Find the shape corresponding to the document.
- Apply a perspective transformation to the shape.





3: Document from Camera

Difficulty : Master

Edge Detection:

We can use the OpenCV Canny algorithm. The algorithm calculates the pixel intensity variation along the x and y axes, and keep only the pixels which intensity is between a lower and an upper threshold (hyperparameters). Returns a 'mask' (an image of only 0s and 1s)

0	0	0	0	0
255	255	255	255	255
0	0	0	0	0

Find the shape corresponding to the document:

We can use the OpenCV `findContours()` method, to grab all the shapes in the edge mask and retain only the shape with the maximum area with 4 edges.



```
# convert the image to grayscale, blur it, and find edges
# in the image
gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
gray = cv2.GaussianBlur(gray, (5, 5), 0)
edged = cv2.Canny(gray, thresh_low, thresh_high)

# find the contours with max area
cnts = cv2.findContours(edged.copy(), cv2.RETR_LIST, cv2.CHAIN_APPROX_SIMPLE)
cnts = sorted(cnts[0], key = cv2.contourArea, reverse = True)[:5]

# init output contour
screen = None

# loop over the contours
for c in cnts:
    # approximate the contour with the closest polygon
    peri = cv2.arcLength(c, True)
    approx = cv2.approxPolyDP(c, 0.02 * peri, True)
    # if our approximated contour has four points, then we
    # can assume that we have found our screen
    if len(approx) == 4:
        screen = approx
```



3: Document from Camera

● Difficulty : Master

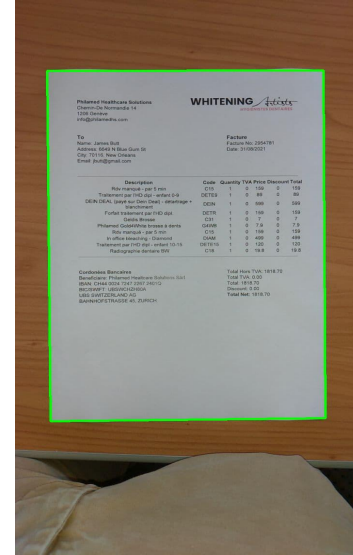
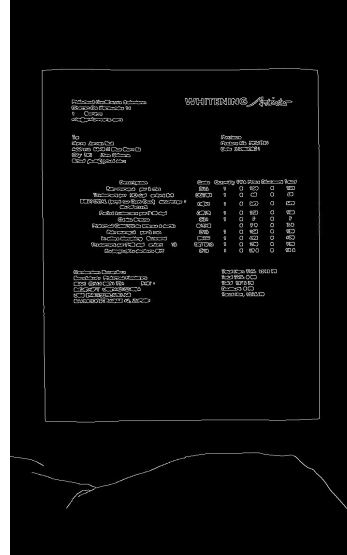
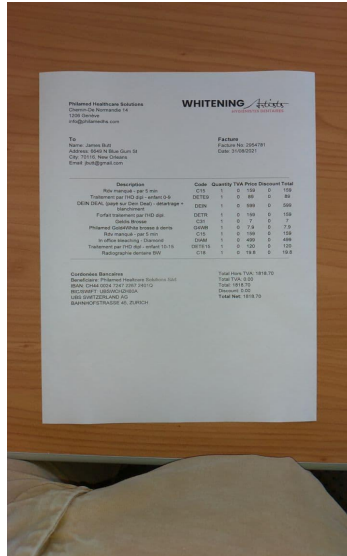
Input

>>>

Edge mask

>>>

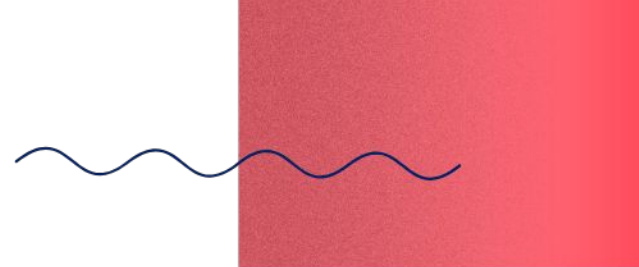
Find contour





3: Document from Camera

● Difficulty : Master



Perspective transformation:

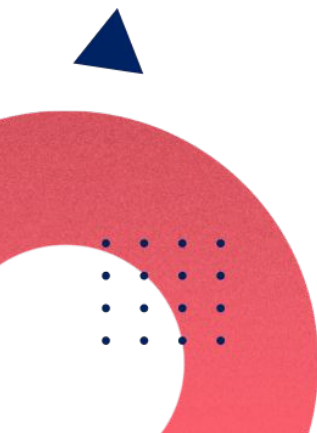
The name sounds more complicated than reality:
mapping an x, y point to a point x', y' using a
transformation matrix.

$$\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$

OpenCV calculates this mapping matrix from 4
points of a starting rectangle and 4 points of a
target rectangle using the method
`getPerspectiveTransform()`. You can then apply the
transformation to the entire image using the
`warpPerspective()` method.

```
# use the target image shape as destination point of the transformation
h, w, c = target.shape
dst = np.array([[0, 0], [w - 1, 0], [w - 1, h - 1], [0, h - 1]], dtype = "float32")

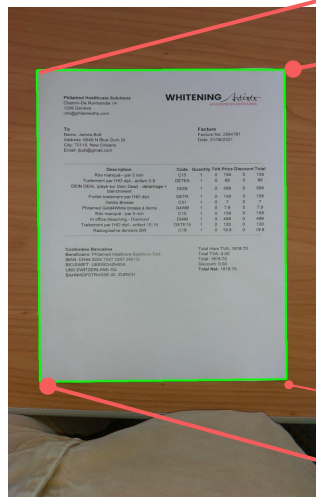
# compute the perspective transform matrix and then apply it
M = cv2.getPerspectiveTransform(rect, dst)
warped = cv2.warpPerspective(image, M, (w, h))
```





3: Document from Camera

Difficulty : Master



$$\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

Philamed Healthcare Solutions
Chemin De Normandie 14
1204 Genève
info@philamedhs.com

WHITENING *Active*
BIOBIOSTYLA 2007/2015

To: Name: James Bull
Address: 5545 N Blue Gum St
City: 70116, New Orleans
Email: jbull@gmail.com

Facture
Facture No: 2064781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dgi - enfant 0-9	DETE9	1	0	89	0	89
DEIN (DEAL) (jeu) sur Don Deal - démarrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dgi	DETR	1	0	159	0	159
Déjà Brosse	C31	1	0	7	0	7
Philamed GoldWhite brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	D4AM	1	0	499	0	499
Traitement par l'HD dgi - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Conditions Bancaires
Bénéficiaire: Philamed Healthcare Solutions Ltd
IBAN: CH44 024 7247 2287 7401 Q
BIC: SWIFT: UBSWCH33HAN
UBS SWITZERLAND AG
BARNHOFSTRASSE 45, ZÜRICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70



4: Document with items in variable position

● Difficulty : Sith Lord

Problem:

What if the items in the document are not in the same position every time? In this particular example the product list in the invoice table always changes length, making the elements in red move in each invoice.

We need to find a method that is able to spot a RoI independently of its position and orientation.

The solution:

Bounding Box Regression using Keras and Deep Learning. Instruct a Deep Learning Algorithm to understand where is the position of the RoI in each document.

Philamed Healthcare Solutions
Chemin-De Normandie 14
1205 Genève
info@philamedhs.com

WHITENING *Artists*
HYGIENISTES DENTAIRES

To
Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par IHD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par IHD dipl.	DETR	1	0	159	0	159
Geldis Brosse	C31	1	0	7	0	7
Philamed Gold4White brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par IHD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCH2HBA
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70



4: Regression vs. Classification

● Difficulty : Sith Lord

Classification:

The task of predicting labels or values in a discrete range. ['dog','cat'] or [0,1]. Usually a 'cross entropy' loss is minimized during training.

Regression:

The task of predicting values in a continuous range. Usually an error metric is minimized during training. I.e. mean squared error:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Bounding Box Regression:

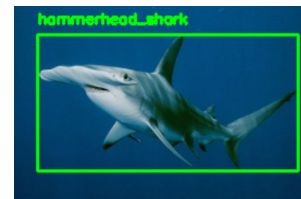
The task of predicting the position of the pixels of the rectangle surrounding a particular object.

Ingredients for Bounding Box Regression:

Dataset: A Representative dataset. Needed for training our model.

Annotations: A file containing the coordinates of the pixels of the bounding box for each image in our dataset. Needed to tell our model where is the position of the object we're looking for

Model: a proper neural network that will be able to accomplish the task.





4: Dataset Annotations

● Difficulty : Sith Lord

Dataset:

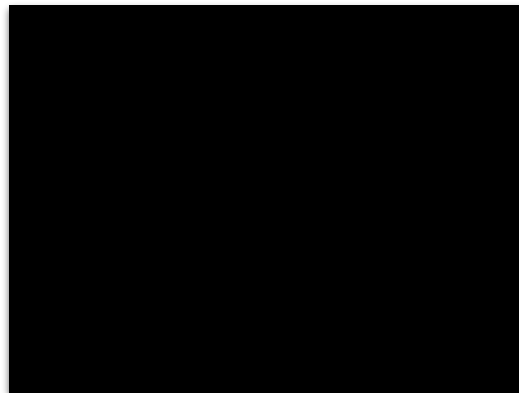
The 430 fake invoices used in Part 1.

Annotations:

A json file containing the coordinates of the Bounding Box corresponding to the 'total invoice' region.

Annotations are made with a specific tool called VIA ([VGG Image Annotation](#)). You will need it in the practice session to make your own annotations. Please, [download it here](#).

You may watch [this 5 minute tutorial](#) to get started.



```
{
  "invoice_0.jpg252520": {
    "filename": "invoice_0.jpg",
    "size": 252520,
    "regions": [
      {
        "shape_attributes": {
          "name": "rect",
          "x": 1049,
          "y": 1309,
          "width": 416,
          "height": 264
        },
        "region_attributes": {
          "region_label": "invoice_total"
        }
      }
    ]
  },
  "file_attributes": {}
}
```



4: Model and Training

● Difficulty : Sith Lord

Transfer Learning:

We will exploit a formidable technique to train our Bounding Box regression model.

Transfer Learning is based on two principles:

Network surgery: take an already trained neural network, modify a small part of it (usually the output layer) leaving the rest untouched.

Finetune: train only the layers we modified and (maybe) some inner layers to accomplish the specific task we're aiming to.

Why is this technique so powerful?

Usually a very big and complex network is used. This network was already trained on huge datasets to be able to accomplish tasks on a vast variety of different images.

The inner layers of the network are thus already capable of extracting very meaningful features from any kind of image. Those features are the building block of any image, so that it is necessary just to train few layers on a custom dataset to apply it to a specific problem.

Useless to say that it would be impossible with our own computational means to train such a network from scratch!





4: Model and Training

● Difficulty : Sith Lord

VGG16 Network:

Trained on [ImageNet](#) Dataset. (14M images and 1000 classes) .

Procedure:

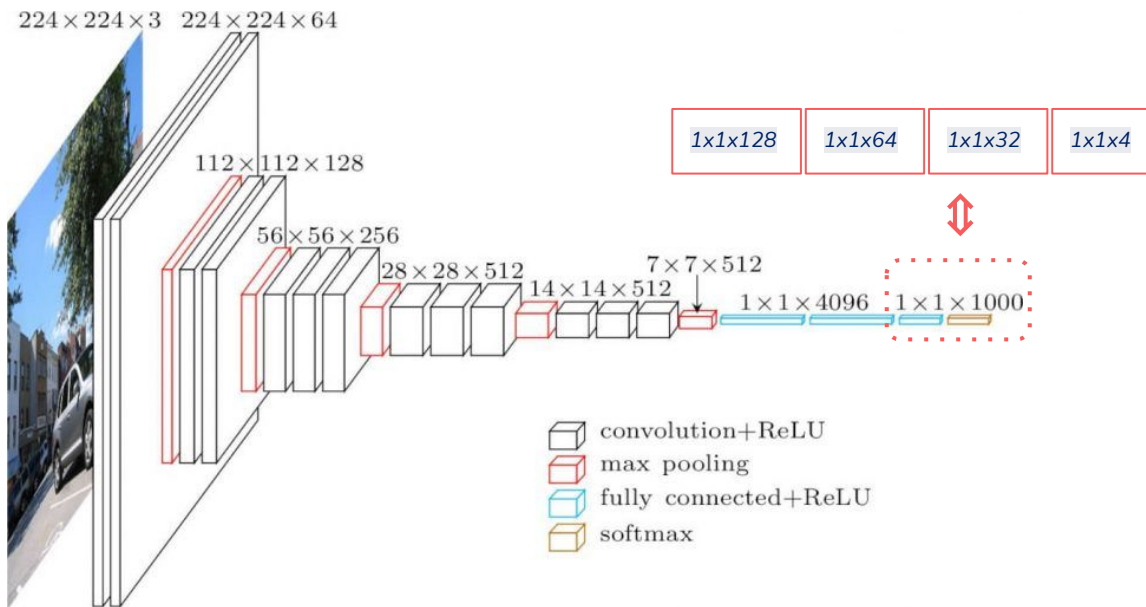
Download the VGG16 with ImageNet weights.

Chop the Fully Connected output.

Replace with Dense Layers with 4-neuron output.

'Freeze' (make not-trainable) the inner Convolutional layers

Train only the Dense Layers we added.





4: Image Preprocessing

● Difficulty : Sith Lord

Image Preprocessing:

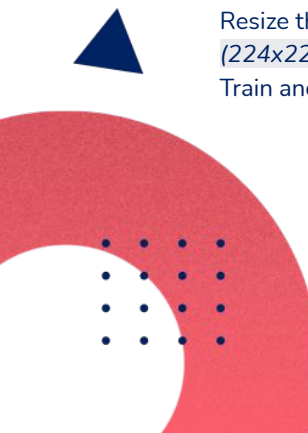
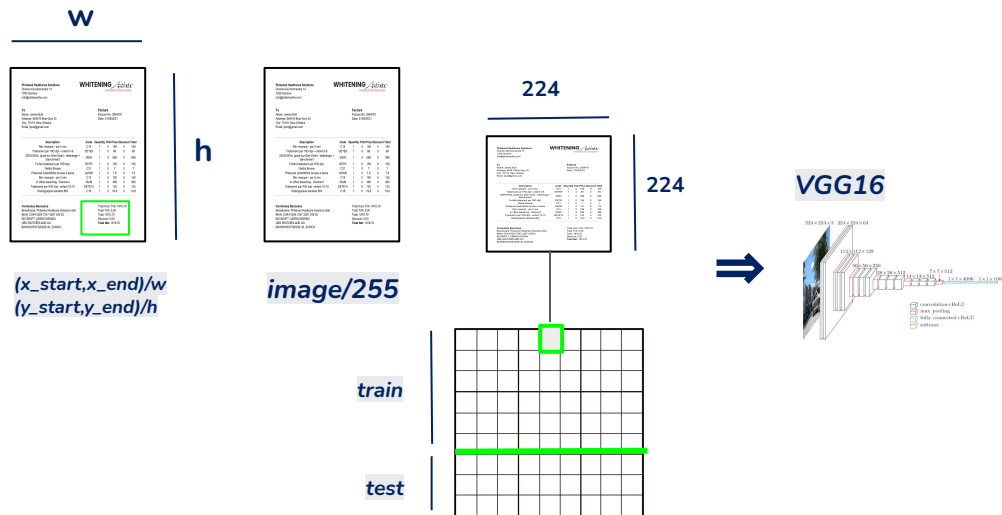
The images that we present to the network should pass through a series of manipulations (preprocessing) both when **training** and when **making predictions**:

Scale the bounding box coordinates in the range $[0,1]$

Scale the pixel intensities in the range $[0,1]$

Resize the image to the VGG16 input size (224x224x3)

Train and test split





4: Practice

● Difficulty : Sith Lord



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYPOHISTES DENTAIRES

To: Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dipl.	DETR	1	0	159	0	159
Geldis Brosse	C31	1	0	7	0	7
Philamed Gold4White brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par l'HD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC:SWIFT: UBSWCH2H80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYPOHISTES DENTAIRES

To: Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

Facture
Facture No: 2954781
Date: 31/08/2021

Description	Code	Quantity	TVA	Price	Discount	Total
Rdv manqué - par 5 min	C15	1	0	159	0	159
Traitement par l'HD dipl - enfant 0-9	DETE9	1	0	89	0	89
DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment	DEIN	1	0	599	0	599
Forfait traitement par l'HD dipl.	DETR	1	0	159	0	159
Geldis Brosse	C31	1	0	7	0	7
Philamed Gold4White brosse à dents	G4WB	1	0	7.9	0	7.9
Rdv manqué - par 5 min	C15	1	0	159	0	159
In office bleaching - Diamond	DIAM	1	0	499	0	499
Traitement par l'HD dipl - enfant 10-15	DETE15	1	0	120	0	120
Radiographie dentaire BW	C18	1	0	19.8	0	19.8

Cordonées Bancaires
Beneficiaire: Philamed Healthcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC:SWIFT: UBSWCH2H80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 1818.70
Total TVA: 0.00
Total: 1818.70
Discount: 0.00
Total Net: 1818.70



my_annotations.json



⇒ Train and Predict



⇒ your_annotations.json

In the Notebook:
We will detect the 'total invoice' Rol

In the Exercise:
You will annotate the box corresponding to the 'table of products' Rol



Samurai

It's time for practice!

We are going to see how these concepts are realized in practice with Python.

Please make a copy of the **AMLD2021-NoMercy-p2.ipynb** notebook and let's have fun with Python, OpenCV and Keras!

Before Starting:

There are 3 Exercises in the Notebook. For timing reasons, We will skip the first 2 and concentrate on the last one. You can come back on the others if there's time left or even at home!

```
print('Thank you for your attention !')
```