Part 2/2

No Mercy for Manual Entry

29/Sep/2021
Workshop @ AMLD2021

ONE swiss bank

SamurAI

AMLD EPFL

Authors

Valerio Rossetti, PhD
Co-founder of SamurAI
Senior Data Scientist

Giulio Grossi, PhD
Senior Quantitative Portfolio Manager
at ONE swiss bank

ONE swiss bank

SamurAI

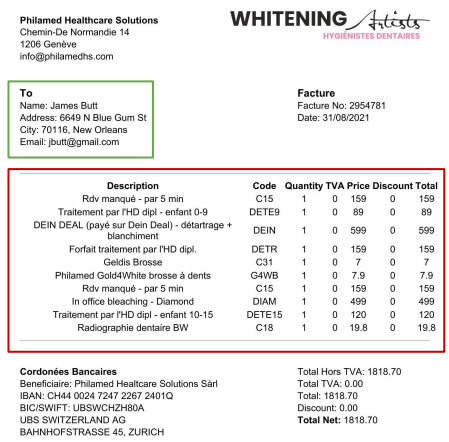
AMLDEPFL

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:

- Documents with Items in fixed positions
- Scanned Documents
- Documents Taken From Camera
- Documents with Items in variable position



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

WHITENING Artists
HYGIÉNISTES DENTAIRES

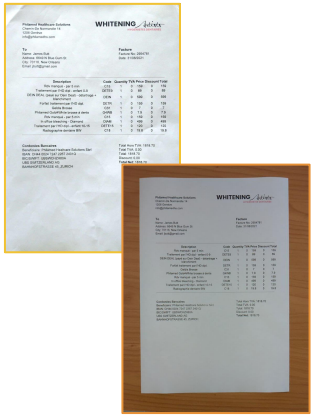
To
Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

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 Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
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



ONE swiss bank

SamurAI

AMLDEPFL

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

ONE swiss bank
SamurAI
AMLD EPFL

The libraries we'll be using

Alongside many others ...

Document with items in fixed position

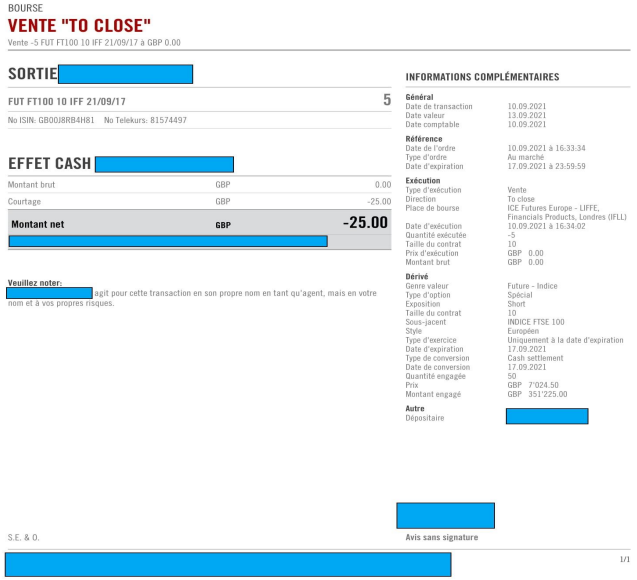
Business case:

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 so engages SamurAI to solve the problem.

The solution proposed by SamurAI:

Define Regions of Interest (RoIs).
Use Tesseract to extract the text in the RoIs.



BOURSE

VENTE "TO CLOSE"

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

SORTIE [redacted]

FUT FT100 10 IFF 21/09/17 5

No ISIN: GB00J8RB4H81 No Telekurs: 81574497

EFFET CASH [redacted]

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

[redacted]

Veuillez noter: [redacted] 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)
Date d'exécution	10.09.2021 à 16:34:02
Quantité exécutée	-5
Taille du contrat	10
Prix d'exécution	GBP 0.00
Montant brut	GBP 0.00

Dérivé

Genre valeur	Future - Indice
Type d'option	Spécial
Exposition	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 [redacted]

[redacted]

S.E. & O. **Avis sans signature**

[redacted]

1/1

ONE swiss bank

SamurAI

AMLDEPFL

Document with items in fixed position

Define RoI:

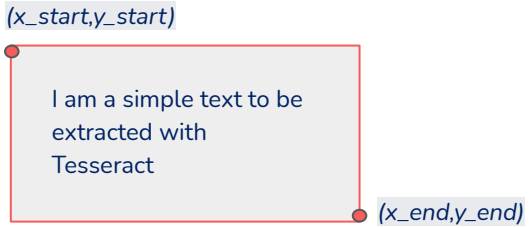
A RoI is a simple set of pixel coordinates that define a region of an image. Usually is a rectangle: *(x_top_left,y_top_left), (x_bottom_right,y_bottom_right)*

Select RoI:

Remember: an image is a numpy array of size *(h,w,3)* so selecting a RoI 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.




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

ONE swiss bank

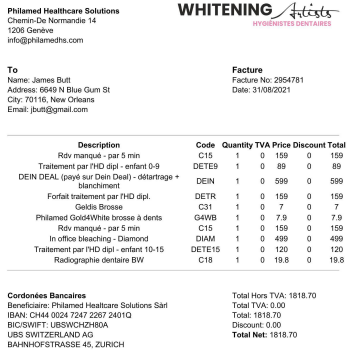
SamurAI

AMLDEPFL

Scanned documents

What if the document is scanned?

The methodology gets more difficult. We will explore and analyse what's different during the practice session.



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

WHITENING Artists
HYGIÉNISTES DENTAIRES

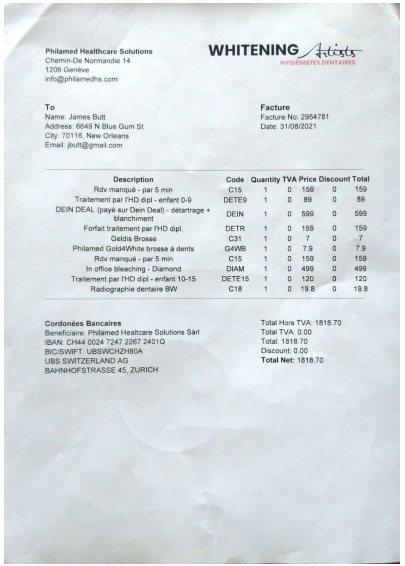
To
Name: James Butt
Address: 6649 N Blue Gum St
City: 70116, New Orleans
Email: jbutt@gmail.com

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 Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
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
HYGIÉNISTES 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 Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
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

ONE swiss bank

SamurAI

AMLDEPFL

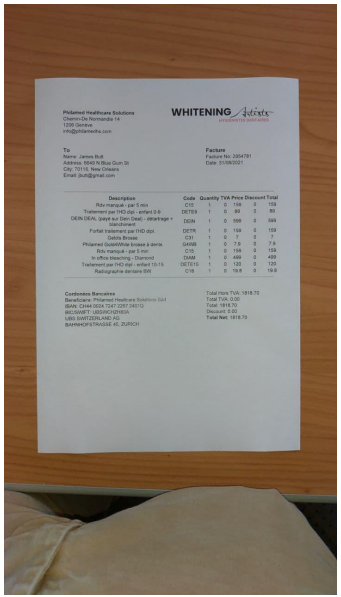
Document from camera

What if the document is taken from a camera?

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 proposed by SamurAI:

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



ONE swiss bank

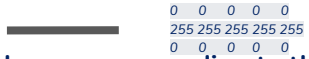
SamurAI

AMLD EPFL

Document from camera

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)



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.

```python
# 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
```

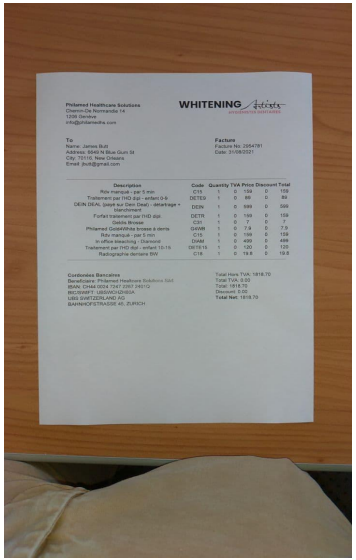
ONE swiss bank

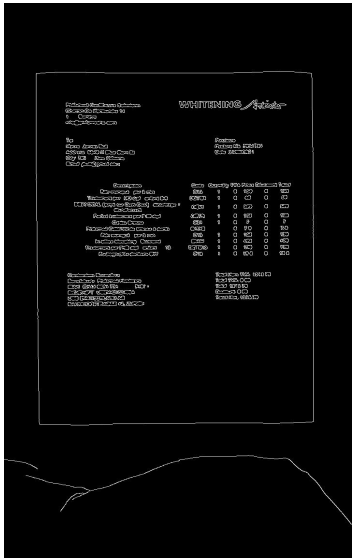
SamurAI

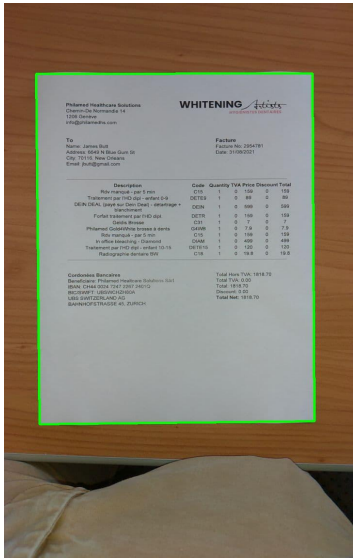
AMLD EPFL

Document from camera

Input >>> Edge mask >>> Find contour







ONE swiss bank

SamurAI

AMLDEPFL

Document from camera

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.

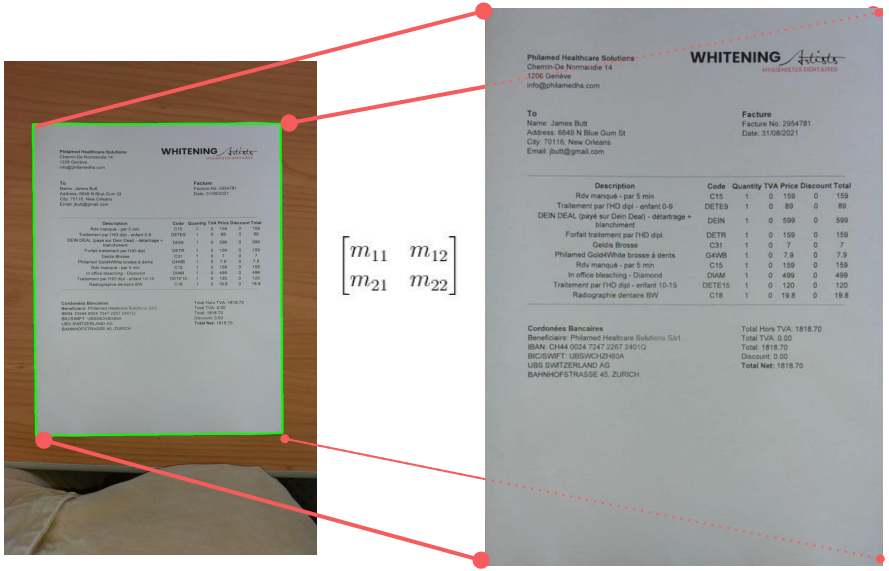
```
# use the target image shape as destination point of the transformation
h, w, c = target.shape
dst = np.array([[0, 0],[w - 1, 0],[w - 1, h - 1],[0, h - 1]], dtype = "float32")

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

ONE swiss bank

SamurAI

AMLDEPFL

Document from camera



ONE swiss bank

SamurAI

AMLDEPFL

Document with items in variable position

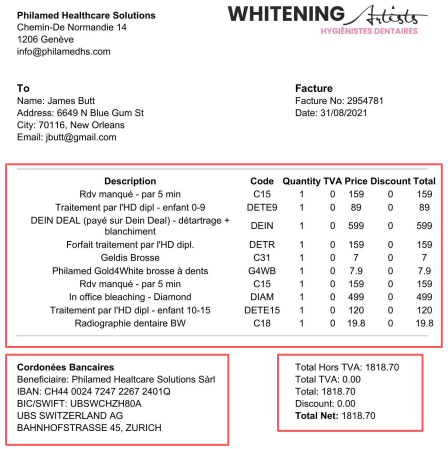
Problem:

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 proposed by SamurAI:

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
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES 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 Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
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

ONE swiss bank

SamurAI

AMLDEPFL

Regression vs. Classification

Classification:
The task of predicting labels or values in a discrete range. *['dog','cat']* or *[0,1]*

Regression:
The task of predicting values in a continuous range.

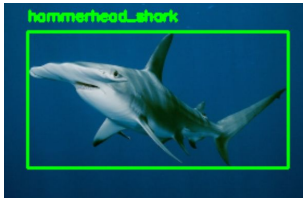
Bounding Box Regression:
The task of predicting the position of the pixels of the rectangle surrounding a particular object.

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.



ONE swiss bank

SamurAI

AMLDEPFL

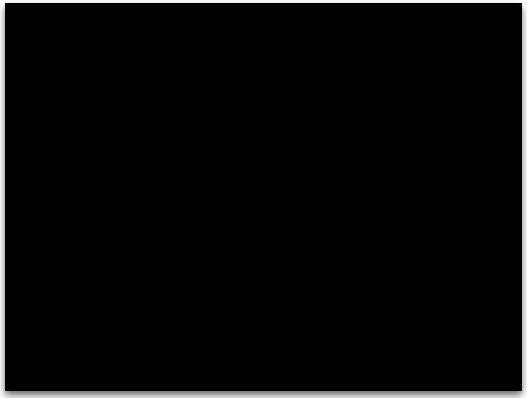
Dataset Annotation

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.

You may watch **this(link)** 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: {}
  },
```


ONE swiss bank

SamurAI

AMLDEPFL

Model and Training

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!

ONE swiss bank

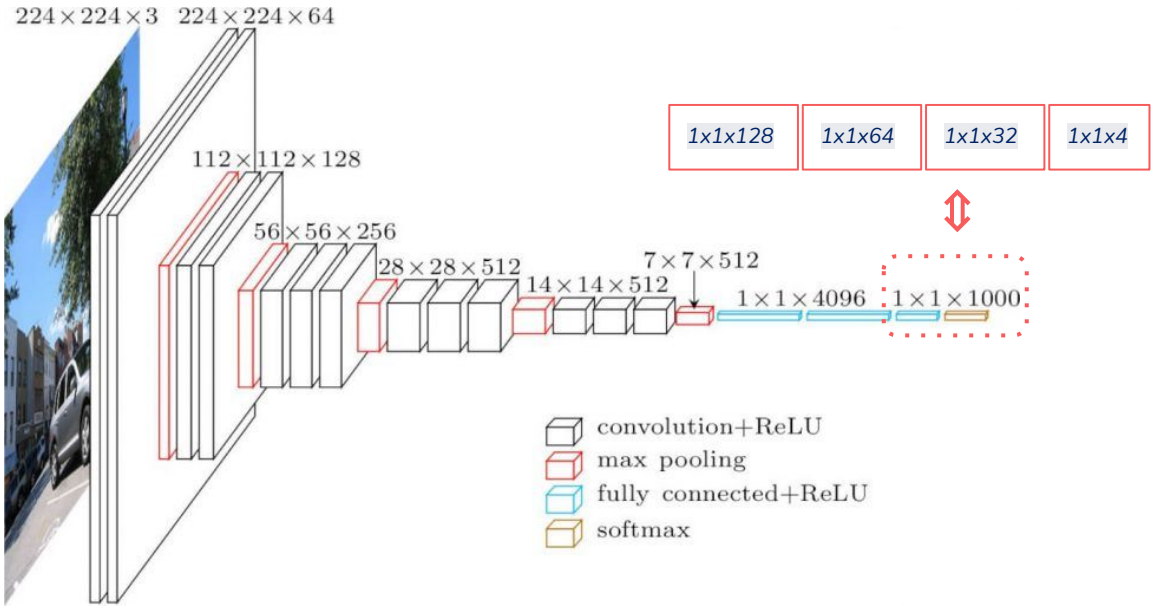
SamurAI

AMLD EPFL

Model and Training

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.



ONE swiss bank
SamurAI
AMLDEPFL

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 OCR.ipynb notebook and let’s have fun with Python, OpenCV and Keras!

Before Starting:
There are 3 Exercises in the Notebook. For timing reasons, I would suggest you to 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 !')
```

ONE swiss bank

SamurAI

AMLDEPFL