Republic of Tunisia
Ministry of Higher Education and Scientific Research
University of Carthage

Higher School of Communication Tunis

SUP'COM
Higher School of Communication of Tunis

Engineering Internship Project report

Prepared by : Ayed Houssem (INDP2 - SYSTEL)

Analysis of vocal folds videos by deep neural network

SUPERVISED BY : PROF BENAZZA AMEL
COSIM LAB SUP'COM

CoSIM
Communications, Signals and Images

Acadamic year : 2019/2020

Acknowledgement

I would like to express my deepest appreciation to all those who provided me the possibility to complete this project. A special gratitude I give to our signal and image processing teacher in SUP’COM : prof Benazza Amel, whose contribution in stimulating, suggestions and encouragement, she helped me to accomplish my project especially in making a plan for this report.

Furthermore I would also like to acknowledge with much appreciation the crucial role of the team of images processing of COSIM Lab, who gave the help to use all required documentation and references to complete the task “Image segmentation and making binary masks for images”.

Finally I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Abstract

The larynx is located in the throat, in the middle of the neck, at the crossroads of the respiratory and digestive tracts. It directs air to the lungs and food to the stomach. The larynx contains the vocal cords, small elastics stretched at the top of the airways, which allow the emission of sounds, the phonation.

Laryngeal dysfunction can cause voice impairment, eating and / or breathing difficulties which their factors are various : age, malmenage, overwork, misuse of voice, smoking, exposure to dust and pollutants …

To deal with vocal folds disorders there are several techniques like testing voice quality (speech analysis) or capturing the vocal folds vibration (videos). While processing images of a video the main tasks are : segmentation, calculating features and attributes to make a classification (healthy/unhealthy vocal folds conditions) at the last step. Delivering critical informations about the shapes and volumes of these organs is crucial in analysing some diseases or picking up where is located a probable infection based on laryngeal images.

Medical image segmentation, identifying the pixels of organs or lesions from medical images is one of the most challenging tasks in medical image analysis. Many segmentation methods have been designed.

In the 2000s, owing to hardware improvement, deep learning approaches came into the picture and started to demonstrate their considerable capabilities in image processing tasks. The promising ability of deep learning approaches has put them as an interesting alternative for image segmentation, and in particular for medical image segmentation. Especially in the previous few years, image segmentation based on deep learning techniques has received a great attention.

The goal of the internship is to use such interesting neural networks to segment and analyze images and videos from stroboscopie operation then compare the performance with the conventional methods.

Table of Contents

Liste of figures

General Introduction

There are two vocal cords located in the duct formed by the larynx. Within this duct, the vocal cords are located about 8mm from the lower edge of the thyroid cartilage. They extend from front to back, and form a V-shaped structure pointing forward.

The vocal cords are composed of several elements : The mucosa of the vocal cords is composed of an epithelium and a chorion. The latter has bundles forming the vocal ligament or lower thyro-arytenoid ligament. The vocal process is a cartilaginous structure that fixes the vocal ligament at the level of the arytenoid cartilage. The muscles of the vocal cords are the vocal muscle, located in the thickness of the vocal cords, as well as the cricothyroid muscle. Consisting of two beams, it intervenes in the rocking movement of the arytenoidal cartilages, thus allowing the tension of the vocal cords.

In case of persistent vocal folds disorder, an examination of the larynx and vocal cords is required. Indirect laryngoscopy ; It allows to observe the larynx with a small mirror placed at the bottom of the throat. Direct laryngoscopy ; The larynx is studied using a rigid and flexible tube introduced by the nose. This procedure may also allow for a sample (biopsy) if the examination requires it. Laryngopharyngography. This radiological examination of the larynx can be performed to complete the diagnosis. A small camera is slipped into a nostril, or in the mouth, and allows to see directly the larynx, its anatomy, its functioning and mobility of vocal folds.

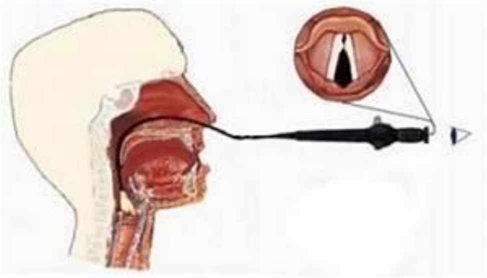


Figure 1: Laryngoscopy (via https://www.lalanguefrancaise.com/dictionnaire/definition-laryngoscopie/).

This is laryngoscopy, performed easily in consultation. It is a painless but very unpleasant examination.

To ameliorate the quality of segmentation and thus the analysis of vocal folds problems the idea that cames in mind : why not exploiting artificial intelligence and deep learning exponential thrive to obtain a segmented images based on neural networks ? as it is invading our lives and giving a lot of solutions for many problems even complicated ones.

In COSIM Lab there is 2 teams the first one, which is supervised by prof Benjbara Sofia, is focusing on the speech anaysis field and the other one, which is supervised by prof Benazza Amel, is focusing on video analysis and image processing.

I’m part of the second team and the objective of my internship is to investigate the potential of deep neural networks in order to segment vocal folds images.

Chapter 1 : Background

1. Neural network :

A neural network is a method of computing based on the interaction of multiple connected processing elements. It's highly recommended to solve many real world problems and it has the ability to learn from historic experience in order to get better performance.

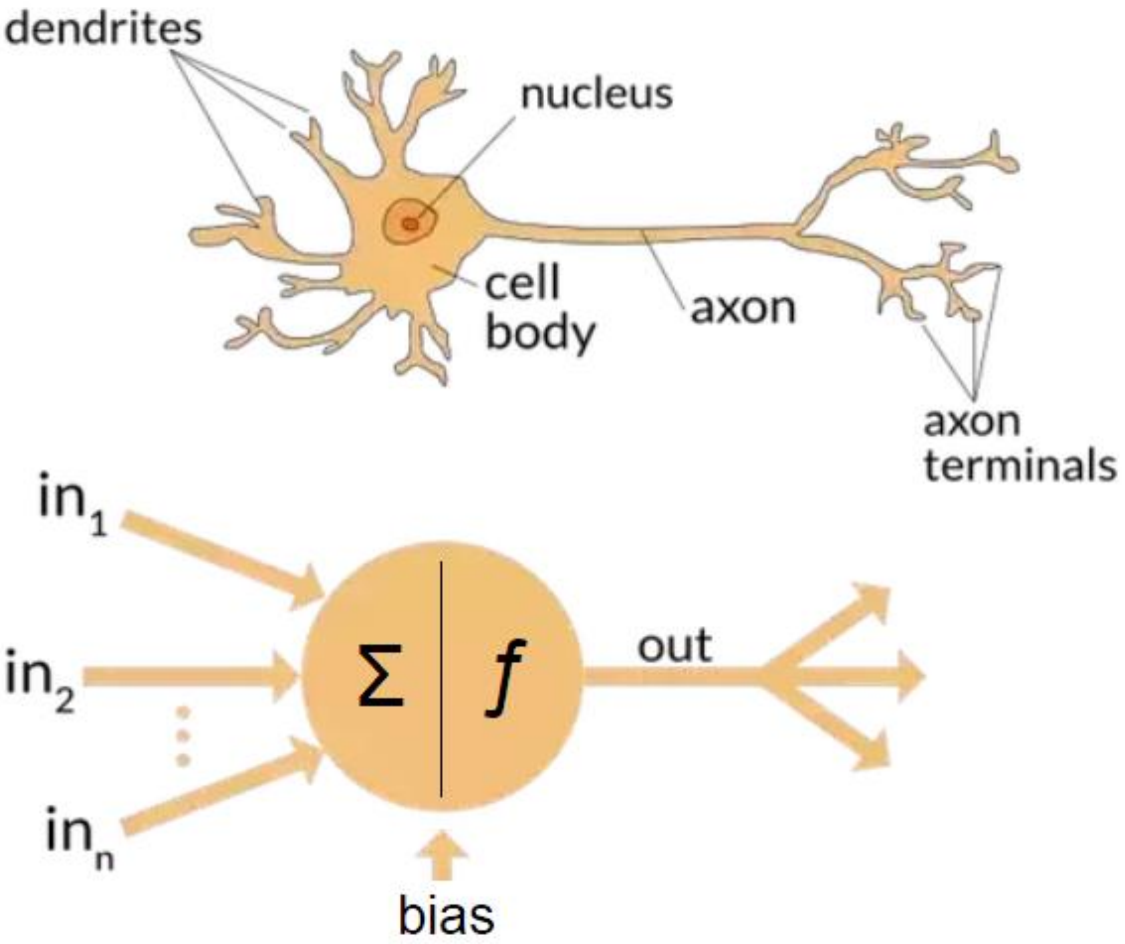


Figure 2 :A biological and an artificial neuron (via https://www.quora.com/What-is-the-differences-between-artificial-neural-network-computer-science-and-biological-neural-network).

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways.

Inspired from neurobiology the neural network is a network of many very simple processors, each possibly having a local memory, then those units are connected by unidirectional communication channels, which carry numeric data between "neurons" in order to perform the desired function or task and thus the models are build to imitate the reaction of human brain neural system not like the serial computer with traditional algorithms

There are many categories of neural network models based on topolgy :

- Single layer

- Multilayer

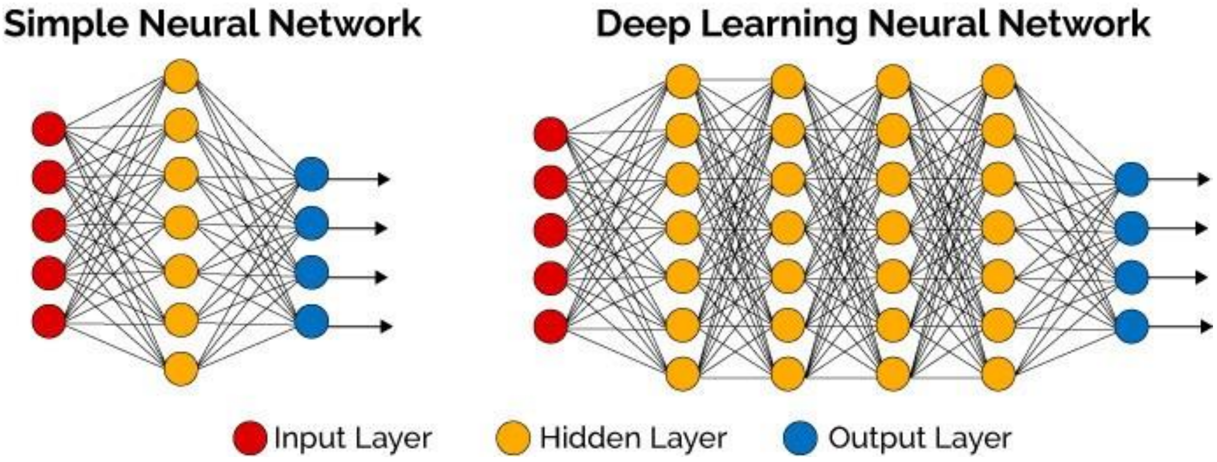


Figure 3: Topolgies of neural network (via https://thedatascientist.com/what-deep-learning-is-and-isnt/).

The neural network is the most used methods in many applications such as : patterns recognition, investment analysis, control system and monitoring, mobile computing, marketing and financial applications, forecasting (sales/market/research/meteorlogy) ...

Deep learning is the name we use for "stacked neural networks"; that is, networks composed of several layers. The layers are made of nodes. A node is just a place where computation happens, loosely patterned on a neuron in the human brain, which fires when it encounters sufficient stimuli. A node combines input from the data with a set of coefficients, or weights, that either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn; e.g. which input is most helpful is classifying data without error? These input-weight products are summed and then the sum is passed through a node's so-called activation function, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome, say, an act of classification. If the signals passes through, the neuron has been "activated."

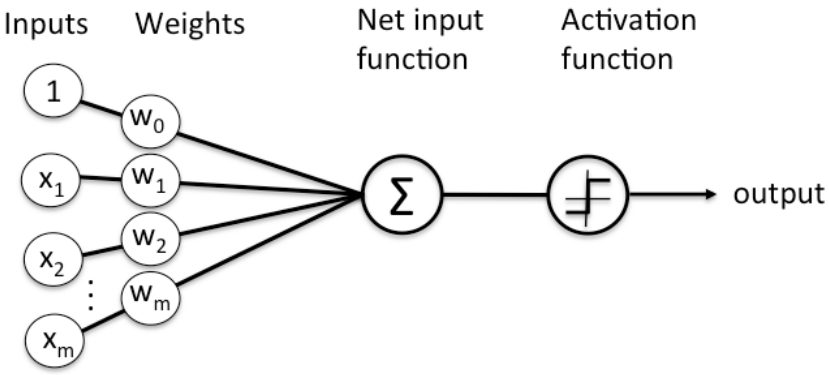


Figure 4: Node or neuron (via https://skymind.ai/wiki/neural-network).

One of those applications is the image segmentation field which is one of the fields in which deep learning is the key to resolve many issues. The need of newest application to delimitate the object shape in the image, it becomes a great challenge to develop more complex models in order to allow machines to keypoint detection, action recognition, video captioning, visual question answering and so on.

There's many architectures that can be used in image segmentation task.

a. Fully Convolutional Network (FCN) :

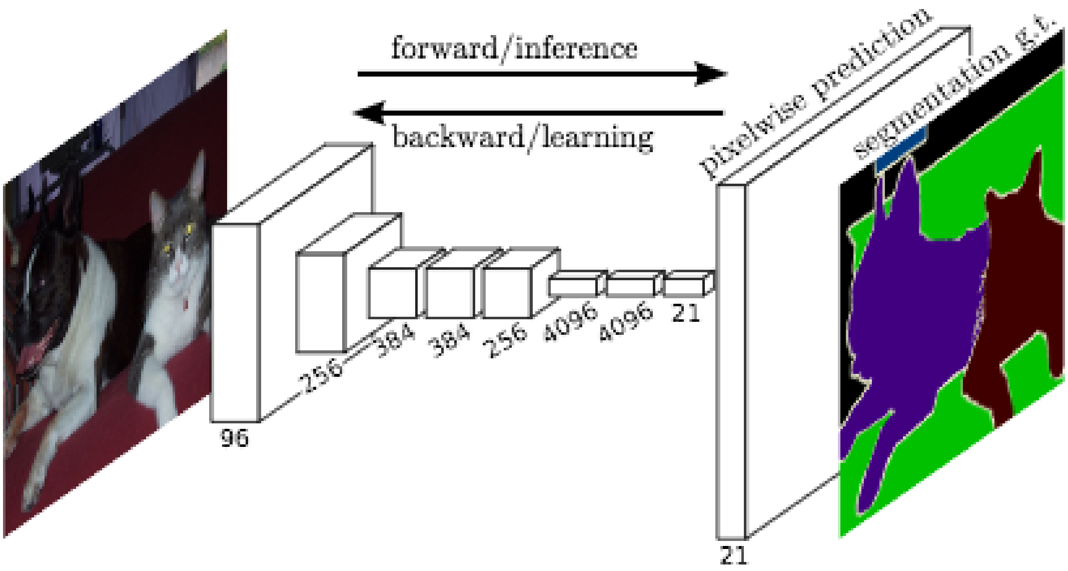


Figure 5: Architecture of the FCN. Note that the skip connections are not drawn here. Source: J.Long et al. (2015).

Advantages :

- Fully Convolutional Networks (FCNs) are being used for semantic segmentation of natural images, for multi-modal medical image analysis and multispectral satellite image segmentation, it offers learns features from all the combinations of the features of the previous layer.
- FCNs have great reduction in the number of parameters.

Disadvantages :

- The network is a bit too slow and complicated if you just want a good pre-trained model.
- The fact of being "fully connected" model cause the loss of spatial information

b. ParseNet :

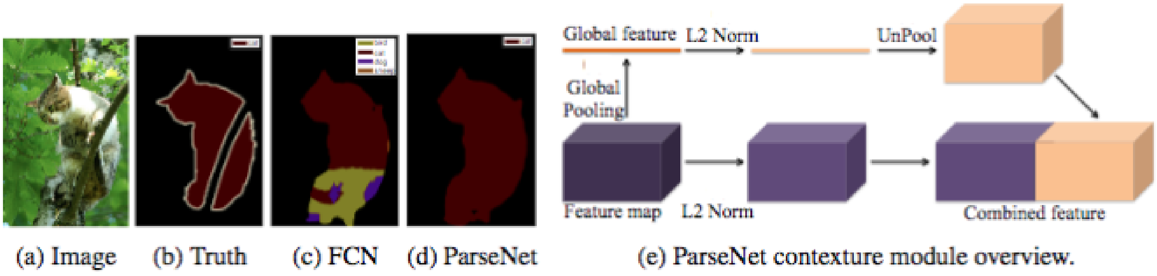


Figure 6: Comparaison between the segmentation of the FCN and the ParseNet and architecture of the ParseNet module. Source: W. Liu al. (2015).

Advantages :

- The model is based on adding intermediate layer features in order to improve the results in term of accuracy
- The network is more robust to local changes by the context defined by the aggregated features.

Disadvantages :

- The network doesn't learn in a fast manner, it's not adapted for real time applications.

c. Convolutional and Deconvolutional Networks :

The first part is processing and transformation by a **convolutional network** to generate a vector of features. The second part is a **deconvolutional network** taking the vector of features as input and generating a map of pixel-wise probabilities belonging to each class.

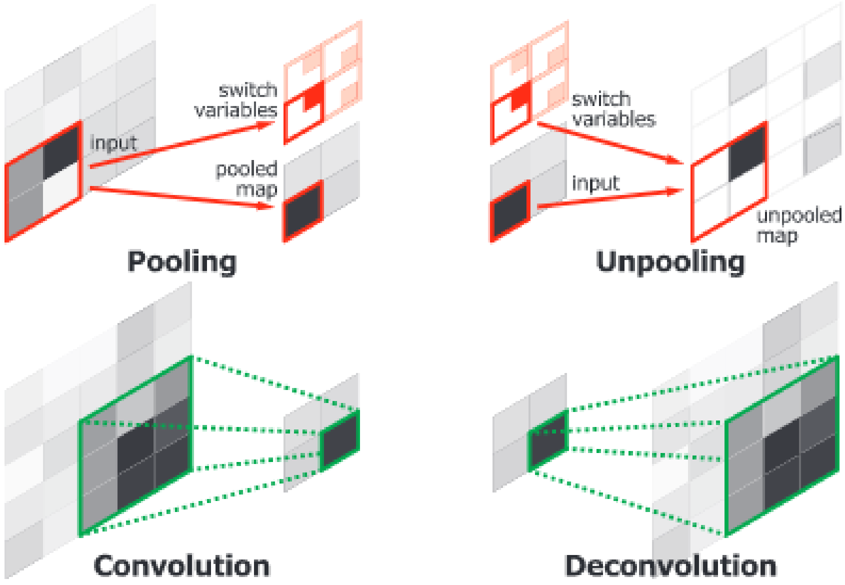


Figure 7: Comparaison of the convolutional network layers (pooling and convolution) with the deconvolution network layers (unpooling and deconvolution). Source: H. Noh et al. (2015).

Advantages :

- The model can take a image of an object, and generate a new image where the object isolated is observed from its background which is required for image segmentation.

Disadvantages :

- Computations are little bit more expensive because the deconvolution is considered as a new convolution (in the other way) which will slow down the training time.

d. U-Net :

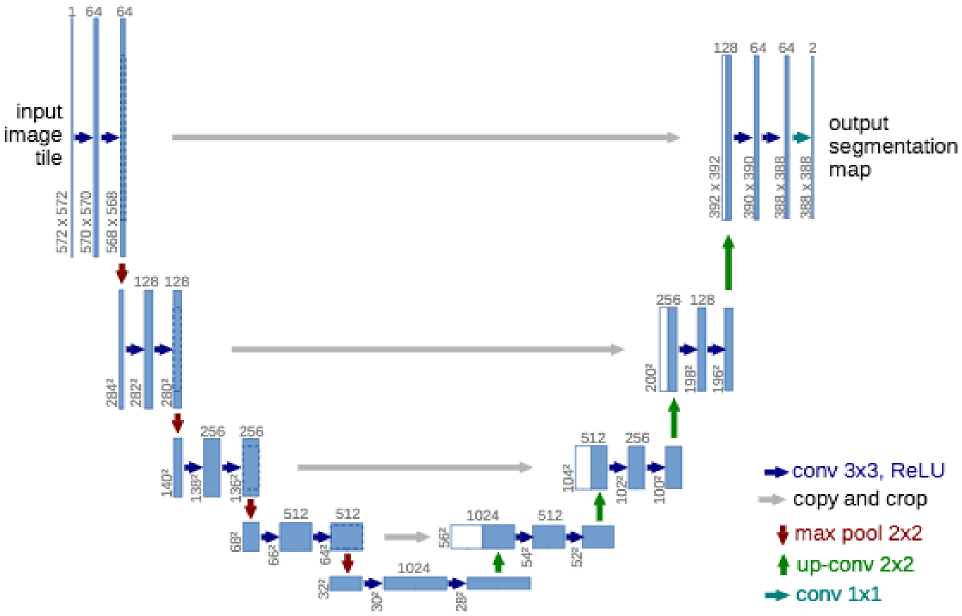


Figure 8: Architecture of the U-net for a given input image. The blue boxes correspond to feature maps blocks with their denoted shapes. The white boxes correspond to the copied and cropped feature maps. Source: O. Ronneberger et al. (2015).

Advantages :

- The use of massive data augmentation is important in domains like biomedical segmentation, since the number of annotated samples is usually limited.
- The U-Net combines the location information from the downsampling path with the contextual information in the upsampling path to finally obtain a general information combining localisation and context, which is necessary to predict a good segmentation map.
- No dense layer, so images of different sizes can be used as input (since the only parameters to learn on convolution layers are the kernel, and the size of the kernel is independent from input image' size).

Disadvantages :

- The model is based on convolutions and deconvolutions so it consumes a huge computational ressources, soi t slow down the model rapidity and increase the cost of deployment (number of GPUs needed, training time).
- The model is not standard to have many pre-trained models available (it's too task specific).

e. Mask R-CNN :

Mask R-CNN is an extension over Faster R-CNN. Faster R-CNN predicts bounding boxes and Mask R-CNN essentially adds one more branch for predicting an object mask in parallel.

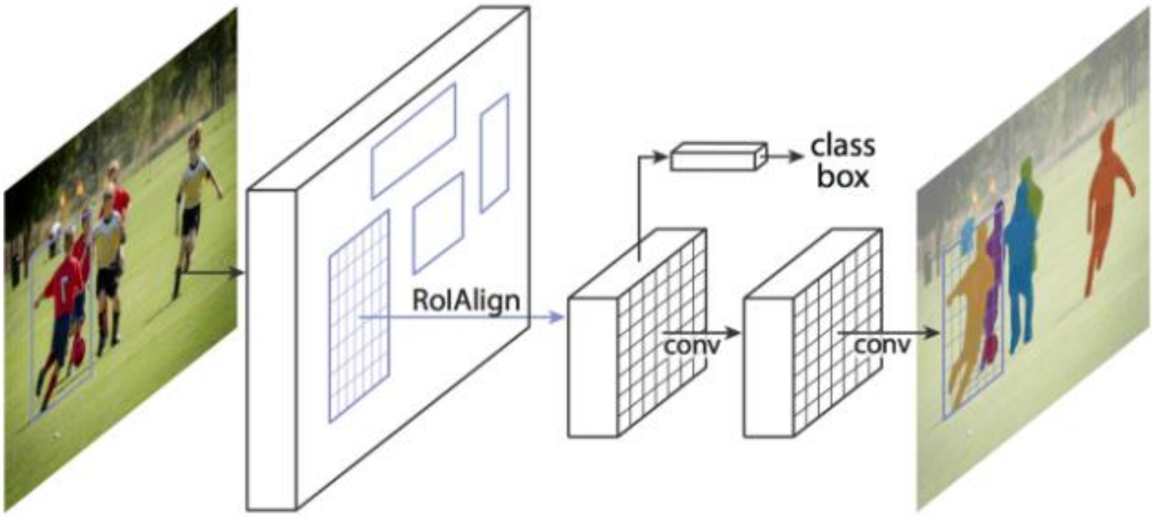


Figure 9: Mask R-CNN framework for instance segmentation. Source: http://arxv.org/abs/1703.06870.

Advantages :

- It outputs a binary mask that says whether or not a given pixel is part of an object with high precision.

Disadvantages :

- The model requires too much computational resources due to convolution steps.

Intersection over Union : (Evaluation Method)

The Intersection over Union (IoU) metric, also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the target mask and our prediction output. This metric is closely related to the Dice coefficient which is often used as a loss function during training.

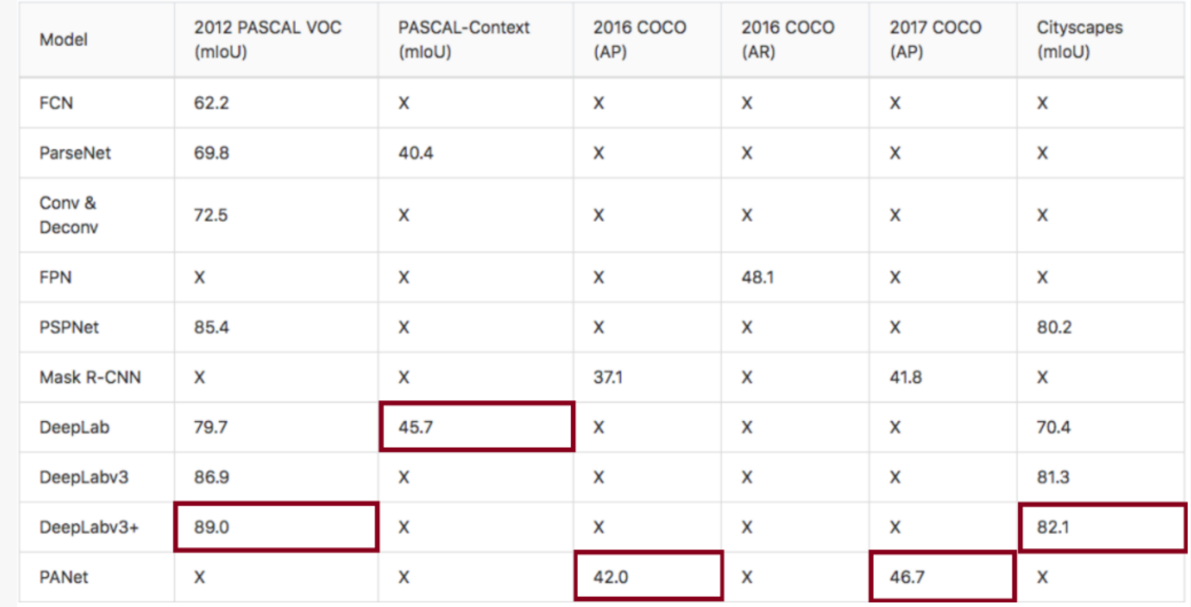
Quite simply, the IoU metric measures the number of pixels common between the target and prediction masks divided by the total number of pixels present across bothmasks.

$$\text{IoU} = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}}$$

The higher the IoUs, the more accurate the bounding the box.

And so, this is one way to map localization, to accuracy where you just count up the number of times an algorithm correctly detects and localizes an object where you could use a definition like this, of whether or not the object is correctly localized.

Comparaison between methods based on IOU metrics explained previously:



Model	2012 PASCAL VOC (mIoU)	PASCAL-Context (mIoU)	2016 COCO (AP)	2016 COCO (AR)	2017 COCO (AP)	Cityscapes (mIoU)
FCN	62.2	X	X	X	X	X
ParseNet	69.8	40.4	X	X	X	X
Conv & Deconv	72.5	X	X	X	X	X
FPN	X	X	X	48.1	X	X
PSPNet	85.4	X	X	X	X	80.2
Mask R-CNN	X	X	37.1	X	41.8	X
DeepLab	79.7	45.7	X	X	X	70.4
DeepLabv3	86.9	X	X	X	X	81.3
DeepLabv3+	89.0	X	X	X	X	82.1
PANet	X	X	42.0	X	46.7	X

COCO dataset :

The COCO dataset is an initiative to collect natural images, the images that reflect everyday scene and provides contextual information. ... A machine learning practitioner can take advantage of the labeled and segmented images to create a better performing object detection model.

2. Choice of the deep learning model :

The best performance are observed when using a DeepLabv3+ model but those performances are obtained while using a different versions of COCO dataset.

As we're using our own dataset (small size between under 200 images) and we have to detect object from medical images so according to my point of view the **UNet model** is a good choice to make a good detection of our object (Vocal folds). The input of our model may be a video or a list of images.

Chapter 2 : UNet Model :

1. Presentation :

U-Net is a convolutional neural network developed for the segmentation of biomedical images in the computer science department of the University of Freiburg in Germany. The network is based on the fully convolutional network and its architecture has been modified and extended to work with fewer training images and to allow for more accurate segmentation.

2. Architecture :

- The network consists of a contracting party and an expansive path, which gives it a U-shaped architecture. The contracting party is a typical convolutional network that consists of a repeated application of convolutions, each followed by a rectified linear unit (ReLU) and a maximum pooling operation.
- During the contraction, the spatial information is reduced while the information on the features is increased. The expansive path combines geographical and spatial feature information through a sequence of ascending convolutions and concatenations with high resolution features from the contracting path.

3. Mechanism of Unet :

Goal : implementation of deep learning model for image segmentation and testing it using a specified dataset.

The network architecture is symmetric, having an Encoder that extracts spatial features from the image, and a Decoder that constructs the segmentation map from the encoded features.
The Encoder follows the typical formation of a convolutional network. It involves a sequence of two 3×3 convolution operations, which is followed by a max pooling operation with a pooling size of 2×2 and stride of 2.
This sequence is repeated six times, and after each downsampling **the number of filters in the convolutional layers are doubled**.
An intermediatery steps is a progression of two 3×3 convolution operations connects the Encoder to the Decoder.
On the contrary, the Decoder first up-samples the feature map using a 2×2 transposed convolution operation, reducing the feature channels by half. Then again a sequence of two 3×3 convolution operations is performed. Similar to the Encoder, this succession of up-sampling and two convolution operations is repeated five times, halving the number of filters in each stage. Finally, a 1×1 convolution operation is performed to generate the final segmentation map.
All convolutional layers in this architecture except for the final one use the ReLU (Rectified Linear Unit) activation function ; the final convolutional layer uses a Sigmoid activation function.

Dataset :

Based on the Laryngeal Endoscopic Images for Semantic Segmentation dataset that is found in github and the several photos from videos of Larynx from www.entusa.com. The dataset that will feed the UNet model, is build with approximately 160 images with their annotations.

The images are shown in the figure 10 below.

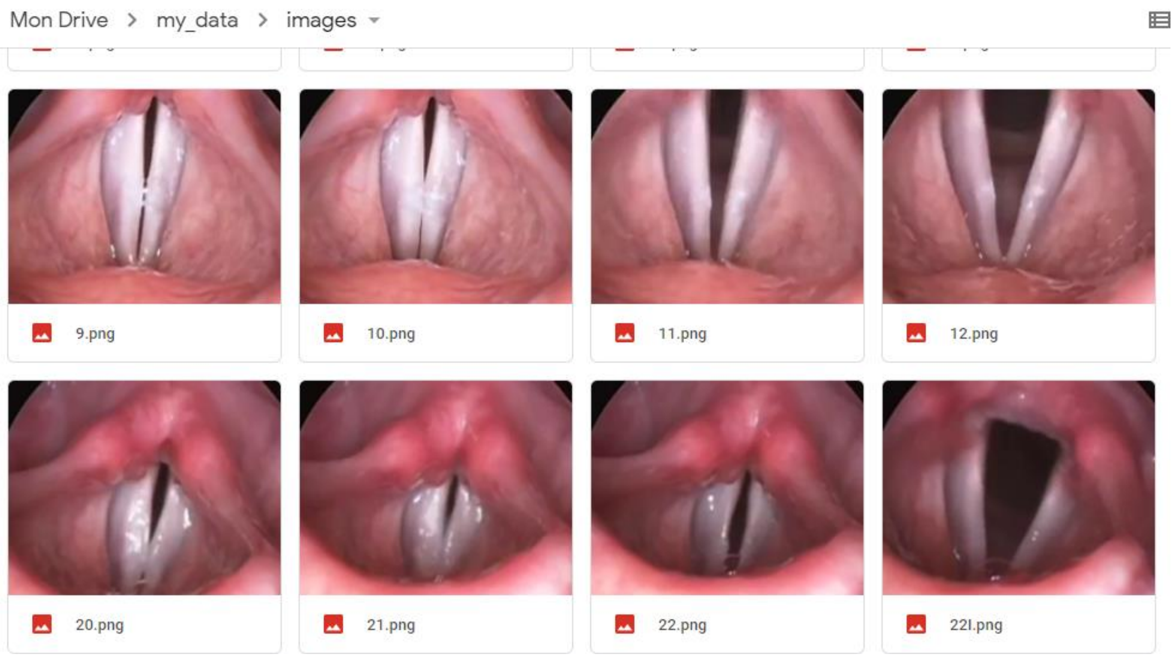


Figure 10: Images of larynx region.

And the annotations are shown in the figure 11 below.

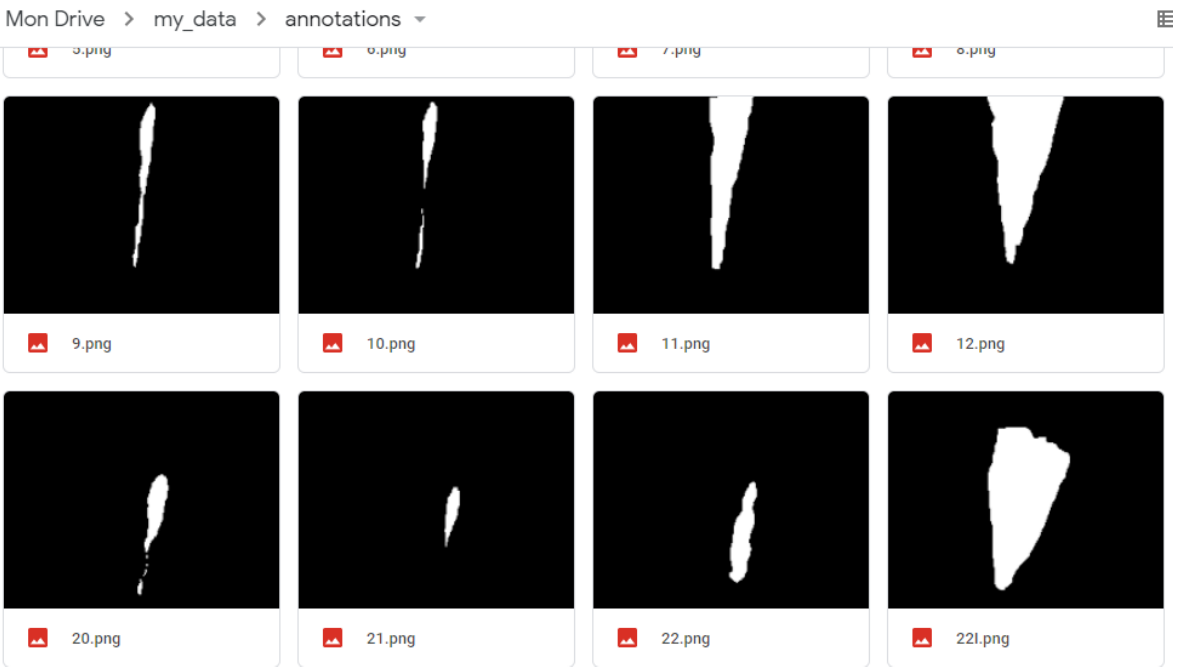


Figure 11: Annotations of larynx images.

For the annotations we're using SegmentTool, that launches a UI-driven tool of MATLAB for trying different approaches to make the annotations (binary masks) of our images.

This tool gives the opportunity to try different edge detection algorithms, modifying all possible parameters ; implement global or local thresholds ; detect regional or extended minima or maxima ; find lines or circles. See the results of different approaches and inputs immediately. The image segmentor allow us to make annotation of the images in our dataset manually by creating a mask that delimitate the region of vocal folds opening from the rest of the image.

Evaluation criteria : (Mean_IOU and/or Dice_index metrics)

The Intersection over Union (IoU) metric, also referred to as the Jaccard index, is essentially a method to quantify the percent **overlap between the target mask and our prediction output**. This metric is closely related to the Dice coefficient which is often used as a loss function during training.
Quite simply, the IoU metric measures the number of pixels common between the target and prediction masks divided by the total number of pixels present across bothmasks.

$$\mathrm{IoU} = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}}$$

➔The higher the IoUs, the more accurate the bounding the box.
And so, this is one way to map localization, to accuracy where you just count up the number of times an algorithm correctly detects and localizes an object where you could use a definition like this, of whether or not the object is correctly localized.

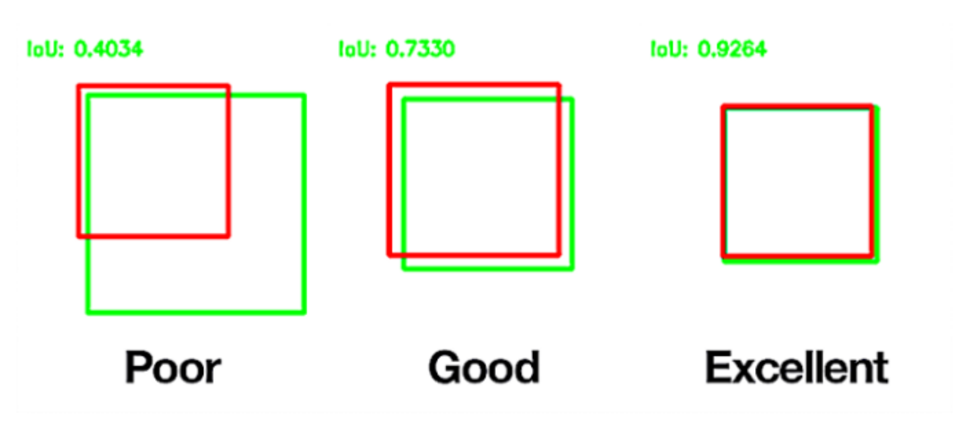


Figure 12: IoU metrics evaluation.

If we want to use a Dice index we can deduct it from the iou metrics with the formula :

$$\mathrm{IoU} = \frac{\mathrm{Dice}}{2 - \mathrm{Dice}}$$

Training the UNet model :

The task of semantic segmentation is to predict whether the individual pixels represents a point of interest, or is merely a part of the background. Therefore, this problem ultimately reduces to a pixel-wise binary classification problem. Hence, as the loss function of the network we simply took the **binary_crossentropy** function and minimized it.

Hyperparameteres :

Batch size	The batch size is an hyperparameter that defines the number of samples to work through before updating the internal model parameters. The batch may be observed as a for-loop iterating over one or more samples and making predictions. At the end of the batch, the predictions are compared to the expected output variables and an error is calculated. From this error, the update algorithm is used to improve the model.
Split size	The indicates the percentage of the data that should be held over for testing. It’s usually around 80/20 or 70/30.
Epochs	The number of epochs is a hyperparameter that defines the number times that the learning algorithm will work through the entire training dataset. One epoch means that each sample in the training dataset has had an opportunity to update the internal model parameters. An epoch is comprised of one or more batches.
Learning rate	The learning rate or step size in machine learning (deep learning) is a hyperparameter which determines to what extent newly acquired information overrides old information. The learning rate is often denoted by the character η or α. A too high learning rate will make the learning jump over minima but a too low learning rate will either take too long to converge or get stuck in an undesirable local minimum.

Parameters of the model

Number of layers	6 in the downsampling and 5 in the upsampling then a mapping layers to make the classification.
Optimizer	Rmsprop
Loss function	binary_crossentropy
Metrics	Mean_iou metrics or Dice index

Evaluation of performance on training and validation set :

The figure below is showing how loss value and mean_iou metric (a method used to calculate accuracy in image segmentation) for training and in validation.

Begining with loss funtion in the training and in validation :

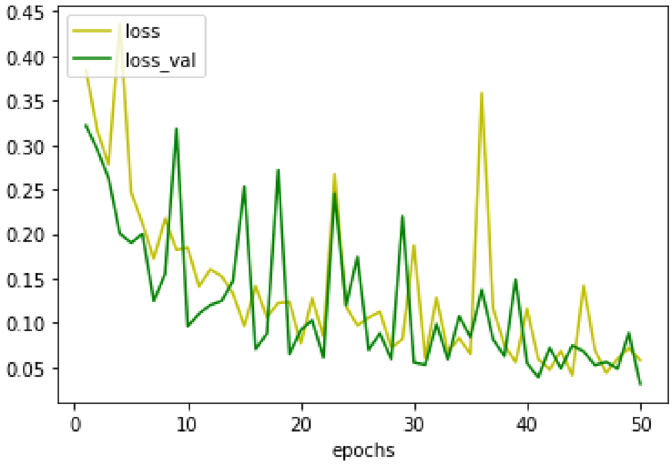


Figure 13: Loss function for training and validation sets.

The figure 13 shows the evolution of the loss function defined in our model during the training and the validation steps. The curve in yellow « loss » is the curve of the loss function in the training set and the curve in green « loss_val » is the curve of the loss function in the test set. The model is improving during the training and the loss function is decreasing in a significant way that means that the rate of misclassified pixels is reducing while the model is fitting. Another important point is that the two curves are decreasing together, this is a sinn of a performant model as there isn't an overfitting which is the problem in many neural network models. The fluctuation which is observed in the two curves is caused by the limited number of images (160) but it doesn't affect the model as the two curves are converging and the fluctuation is minimized by the end of the training.

Moving to the accuracy throw the mesure of mean_iou metrics :

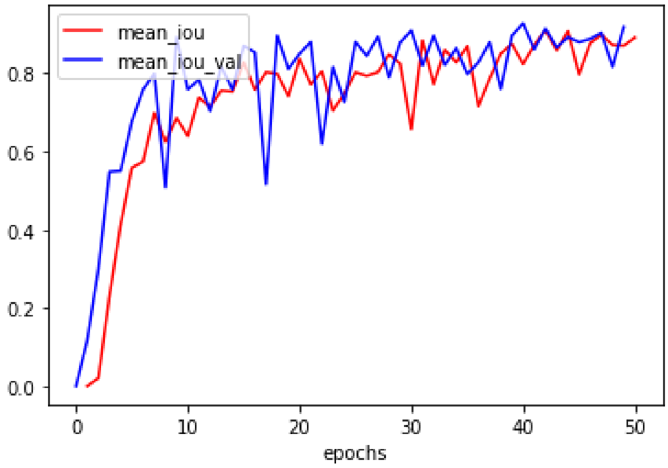


Figure 14: Mean_iou metrics for training and validations sets.

The figure 14 shows the evolution of the mean_iou metrics, which is the evaluation criteria choosed to mesure the model performance, In the figure 17 two curves are plotted : the curve « mean_iou » colored in red is the evolution of the mean_iou metric in the training set and the other one « mean_iou_val » colored in blue is the evolution of the mean_iou metric in the validation set. The figure shows that the mean_iou metrics is increasing till it reachs the edge of 0.91 by the 30th epoch for the training part (in red) for the validation part (in bleu) it has the same behaviour of the previous one but with a lower mean_iou value around 0.88 which is logic as it's validated with images which are never seen before in the training part.

Once the model performance reachs a constant value for the mean_iou metrics there won't be another amelioration and we risk an overfitting if the model continue training (it means that if the model have a problem of overfitting, the model cannot segment perfectly images from training set but can't generalize for images never seen before which is not good) So the training is for 50 epochs to avoid this issue **as we observe that the mean_iou metrics is converging and the fluctuation is decreasing**.

The same figure is obtained when we use the Dice index :

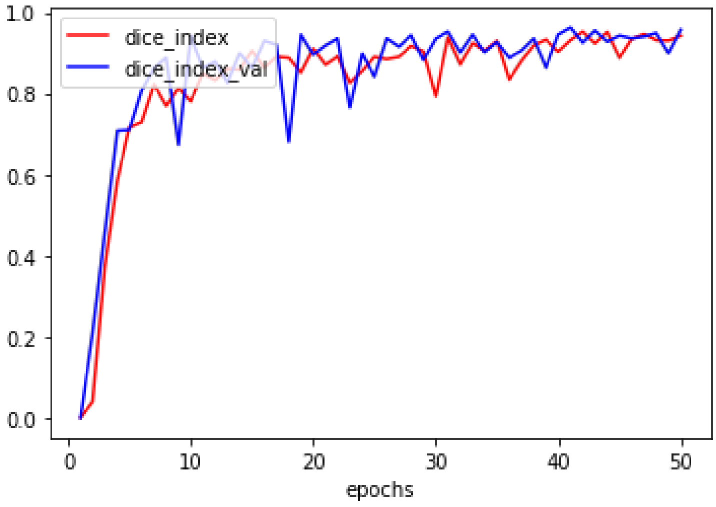


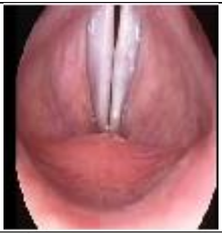
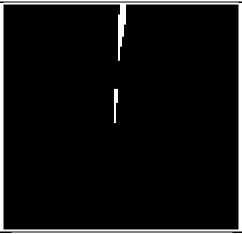
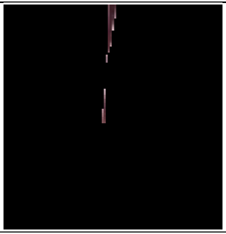
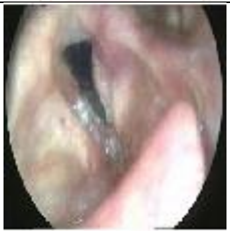
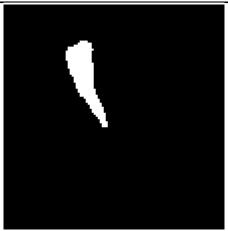
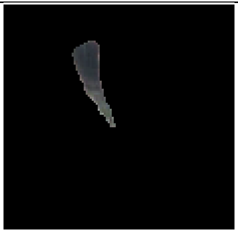
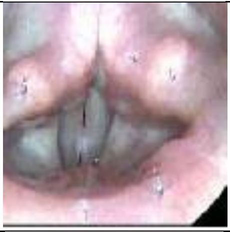
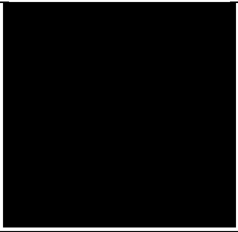
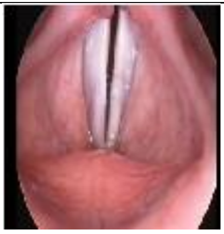
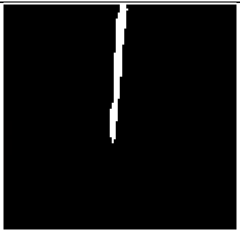
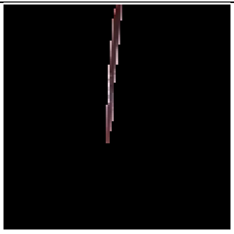
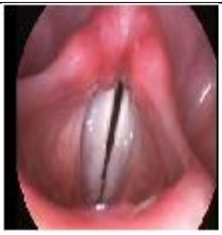
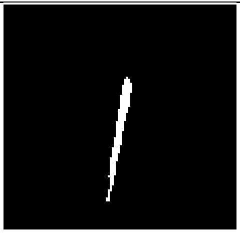
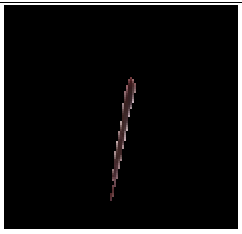
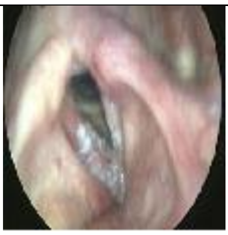
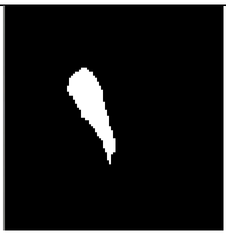
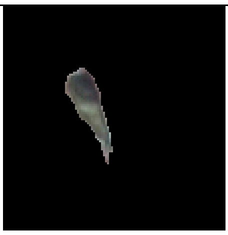
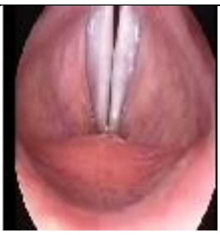
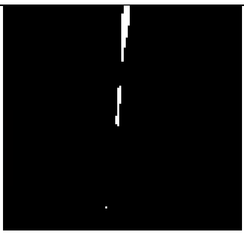
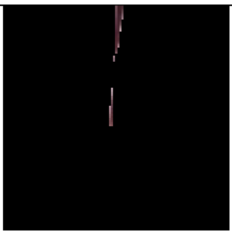
Figure 15: Dice index for training and validation sets.

The figure 15 shows the evolution of the dice index (another evaluation metric).

The curve « dice_index » is the evolution of the dice index metric in the training set and the « dice_index_val » is the evolution of the dice index metrics in the validation set. This figure is similar to the figure 14. Furthermore, it's clear that the model is converging by the end of the fitting for the training set and the validation set.

The table 1 below shows how the model is predicting the mask of an image :

Table1 : Model predicted images

image	predicted mask	combination (image+mask)
		
		
		
		
		
		
		

4. Comparaison between UNet Model and Meanshift :

In this part of the chapter, there is an evaluation on some images of the vocal folds not existing in the dataset of the Unet model. Applying the model and another algorithme « Meanshift » to compare how the model is performing when dealing with images which are never seen.

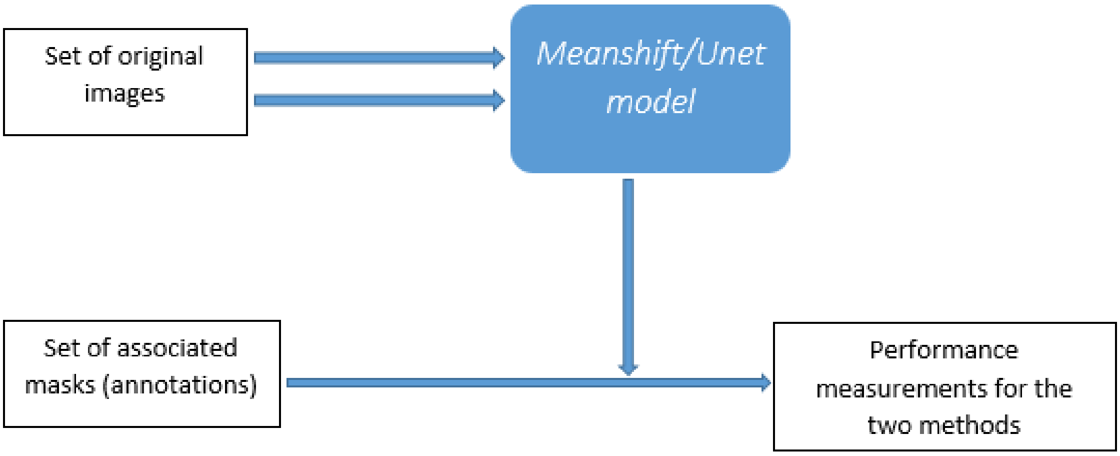

Set of original images
Meanshift/Unet model
Set of associated masks (annotations)
Performance measurements for the two methods

Figure 16: Performance mesurements scheme.

To mesure that score there's a set of images with their true annotation (masks) and they will be passed through a Unet model and a MeanShift algorithm after that there's a mesure of accuracy of object delmitation based on mean_iou metrics (or Dice index).

In the table below the results are calculated with mean_iou formula :

Table 2 : Mean_iou score for unet model and meanshift algorithm

Img n°	UNet iou Score	Meanshift iou Score
21	0.90	0.03
22	0.84	0.10
24	0.87	0.0
26	0.89	0.05
27	0.90	0.06
28	0.81	0.11
29	0.86	0.0
Average_score	0.86	0.05

And in the table below, the score is calculated with the Dice index formula :

Table 3 : Dice index score for unet model and meanshift algorithm

Img n°	UNet Dice Score	Meanshift Dice Score
21	0.95	0.06
22	0.91	0.18
24	0.93	0.01
26	0.94	0.09
27	0.95	0.12
28	0.89	0.19
29	0.92	0.0
Average_score	0.93	0.10

The score is showing a big gap between results of Unet model and Meanshift algorithm. Unet is much more suitable for medical image segmentation as it's fed with a large set of images and there's a features extraction of images to classify whether the part of the image belong to the background or to the object (vocal folds). The variety of images make the model capable of extracting features in many positions and configurations, thus the delimitation of the object is done according to each image. On the other way MeanShift algorithm is configured to delimitate the object (based on some mathematical operations and logical combinations) in the same way for all images and that make it too worse in image segmentation task in comparaison with Unet model, it needs a human intervention to modify the configurations of the algorithm for each image which is exhaustive task and not preferred in the automatic tasks.

Chapter 3 : Pathologies Detection :

1. Introduction :

Voice pathology detection is a field of important research area in voice and speech processing as it may affect the quality of life of the population, especially in people who use voice extensively in their professional activity, as speakers, singers, actors, lawyers, broadcasters, priests, teachers, call center workers, etc ...

The success in treating voice pathologies depend on their early detection, and as such simple yet powerful inspection procedures are desirable. Among those procedures patient's voice inspection is a simple, low cost and fast method to obtain an estimation of the presence of pathology, which can be used as a screening routine to decide if other specialized inspection methods –as videoendoscopy- are to be used, as these being more precise in pathology classification, are at the same time less comfortable, more expensive and complicate, and their use should be obliviated if a simple inspection could help in screening patients before being subject to full inspection procedures.

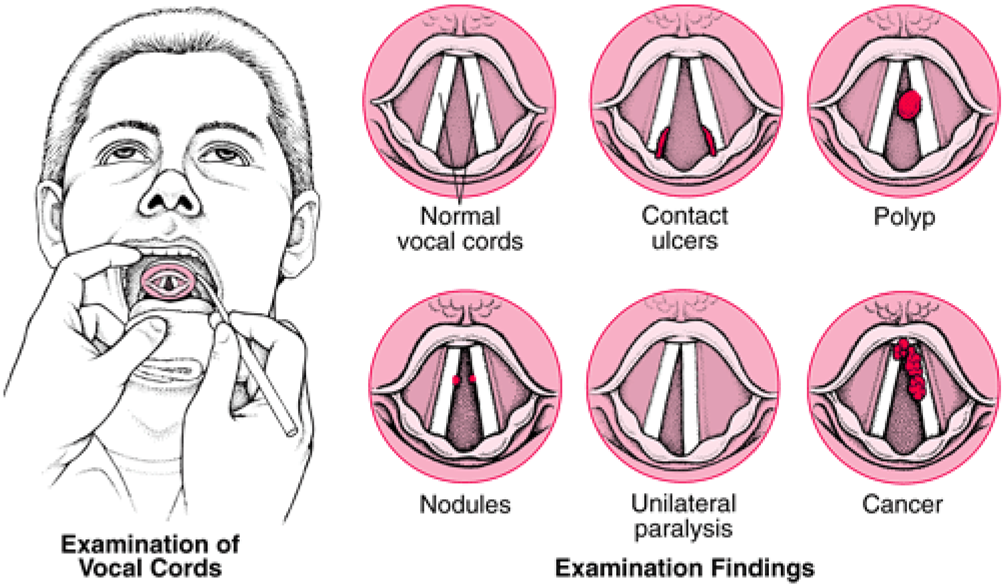


Figure 17: Pathologies of vocal folds. (via https://www.msdmanuals.com/en-nz/professional/ear,-nose,-and-throat-disorders/laryngeal-disorders/overview-of-laryngeal-disorders).

In our case, the task to perform is to decide whether a patient has a vocal folds problems or not through the use one more time of artificial intelligence and machine learning. The classification in our application is binary so the output expected is the existence of a probable disease while capturing some vocal folds sequences.

2. Pathologies detection operation :

After the segmentation of the images with Unet model and extracting the region of vocal folds for a sequence of images, that construct the video recorded in acquisition step, the glottal are waveform (GAW) is calculated for each image and plotted for the whole sequence.

Going back to the GAW obtained with a normal vocal folds condition (Documentation), to see what are the caracteristics of its plot, there’s a cyclic pattern which is repeated through time while patient is saying « I » :

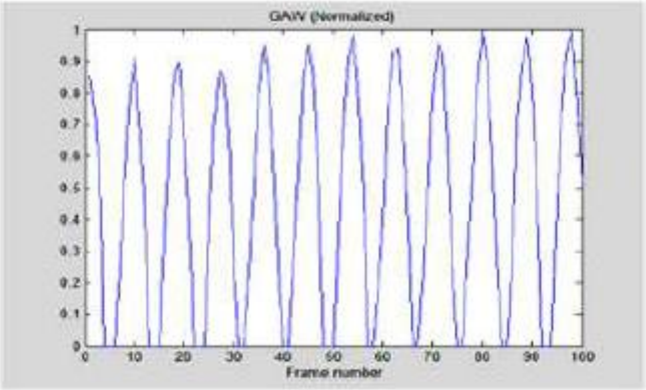


Figure 18: GAW for normal vocal folds condition. Source : chen_yan_IEEETransbiomed_2006.

In the other hand when dealing with a patient having a phenomenon, referred to as voice diplophonia : the image sequence represents two glottal cycles. Clearly a bimodal vibratory pattern is revealed from the GAW, which corresponds to the production of two simultaneous tones. This dynamic behavior of the vocal folds results from an asynchronization in the vibrations at different spatial locations of the vocal folds.

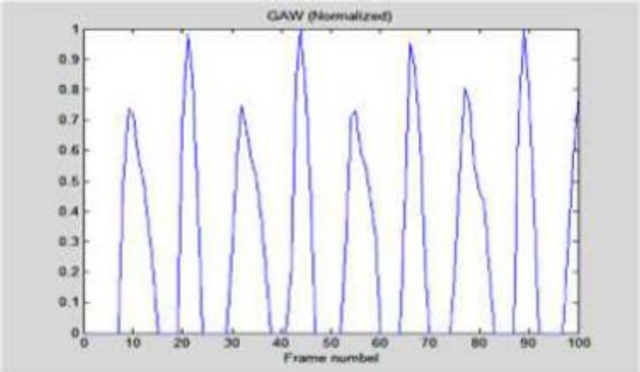


Figure 19: GAW for a patient having diplophonia. Source : chen_yan_IEEETransbiomed_2006.

Having many sequences for healthy patients and others with vocal folds problems. They will be passed through the model and the GAW (glottal ared waveform) is calculated and shown in a figure. For each sequence we’ll get a list of numbers representing the area for that cycle and we have the class of it (normal vocal folds / pathological vocal folds).

This will be the dataset of our model that will classify patients. Consequently, to decide whether the GAW plot is for a normal vocal folds condition patient or presenting some disorders we build an SVM binary classifier to distinguish that based on the dataset we obtain before.

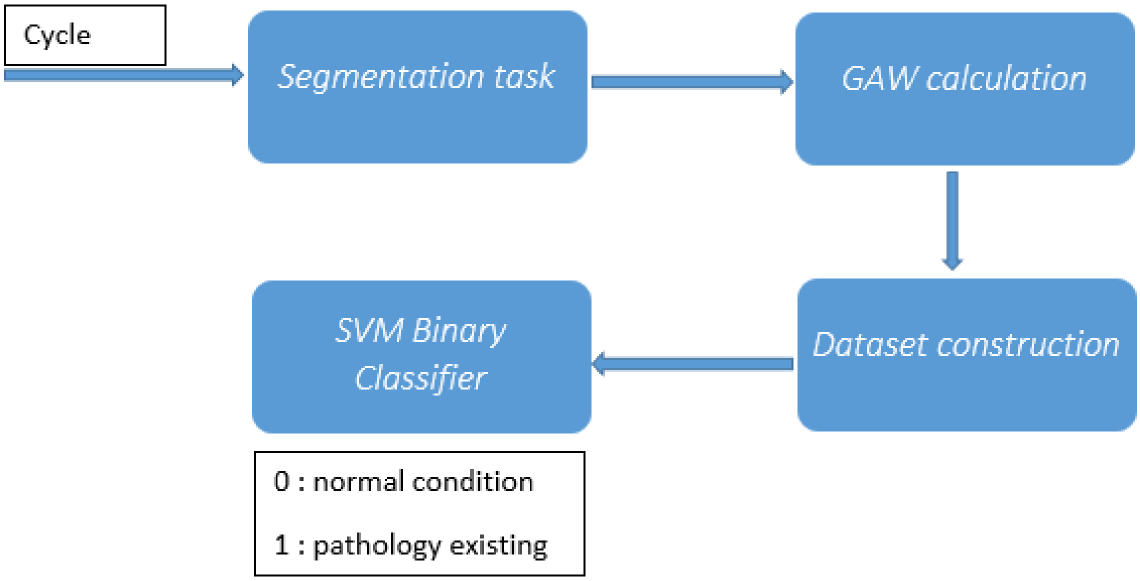


Figure 20: Pathology detection methodolgy with intelligent algorithms.

<u>What's SVM :</u>

Support Vector Machines are supervised learning models for classification and regression problems. They can solve linear and non-linear problems and work well for many practical problems. The idea of Support Vector Machines is simple: The algorithm creates a line which separates the classes in case e.g. in a classification problem. The goal of the line is to maximizing the margin between the points on either side of the so called decision line. The benefit of this process is, that after the separation, the model can easily guess the target classes (labels) for new cases.

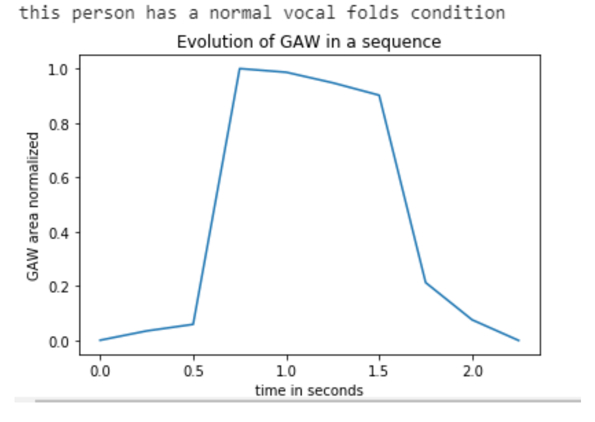
Having the result of classification of SVM model let's make another comparaison between the two models based on those results.

Table 4 : SVM error probabilities when using unet model

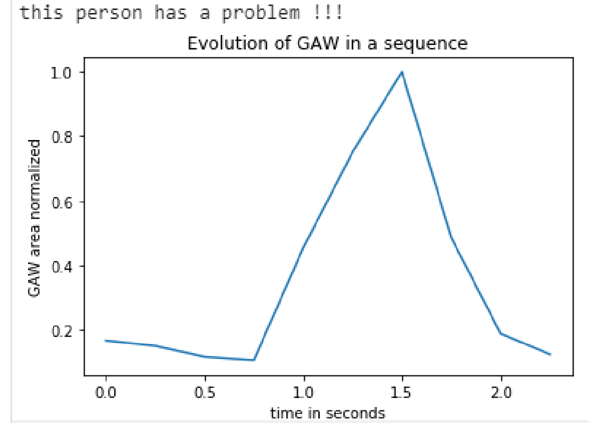
True label predicted	Sick : 0	Healthy : 1
Sick : 0	0.34	0.07
Healthy : 1	0.12	0.45

In the table below, the GAW plot is represented as well as the result of SVM classifier for some sequences choosed from the dataset.

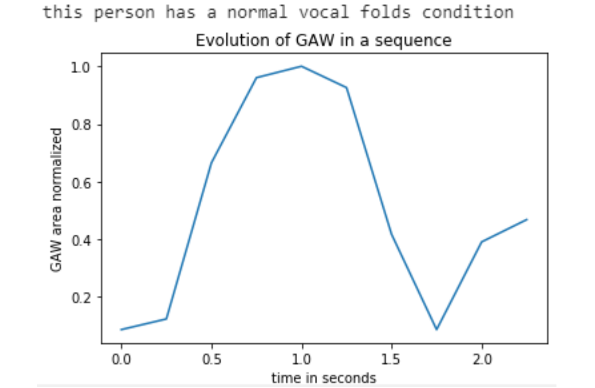
Table 5 : Results of classification (SVM with unet model for segmentation)



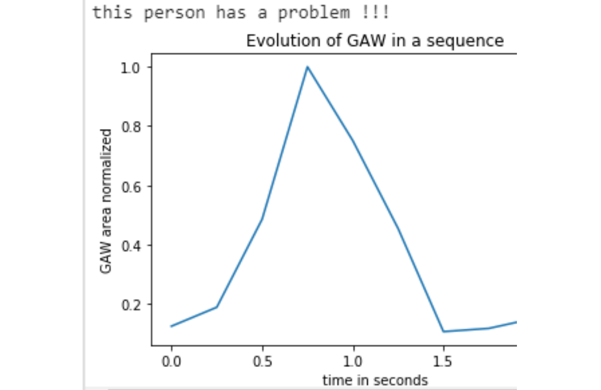
this person has a normal vocal folds condition



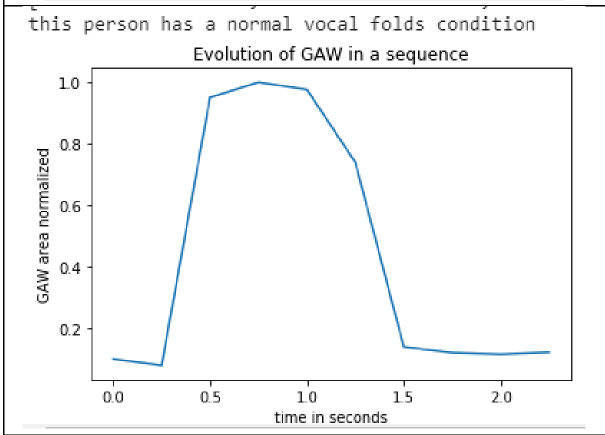
this person has a problem !!!



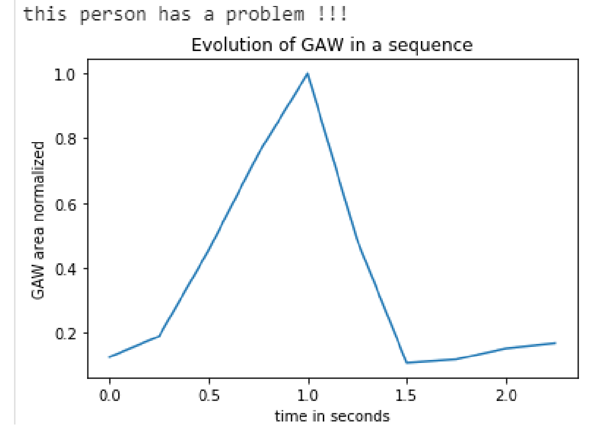
this person has a normal vocal folds condition



this person has a problem !!!



this person has a normal vocal folds condition



this person has a problem !!!

The result of meanShift algorithm are shown in the table below:

Table 6 : SVM error probabilities with meanshift algorithm

True label predicted	Sick : 0	Healthy : 1
Sick : 0	0.04	0.331
Healthy : 1	0.56	0.069

The results that we have are based on GAW calculation as it's done in the previous researchs and it gives valuable informations that can be helpful in the next days to automatically detect pathologies without the need of human intervention.

3. Ameliorations :

For the previous task the model predict a binary mask to classify the region of vocal folds from the background. In the end of the layers of Unet model there is a final layer to make the classification and generate the mask.

The idea that come in mind and probably may improve the model is that in place of generating a mask can the model return the layer just before the classification step : It means that the model will generate a vector of values that will be the set of features generated for each image. So the idea isn't to calculate a GAW and plot the result, but to extract the vector of features, which in our case a tensorflow object, for each image and plot its absolute value or its average …

In the end, for each sequence we'll generate from the values calculated from the vector of features a dataset that will feed the SVM binary classifier.

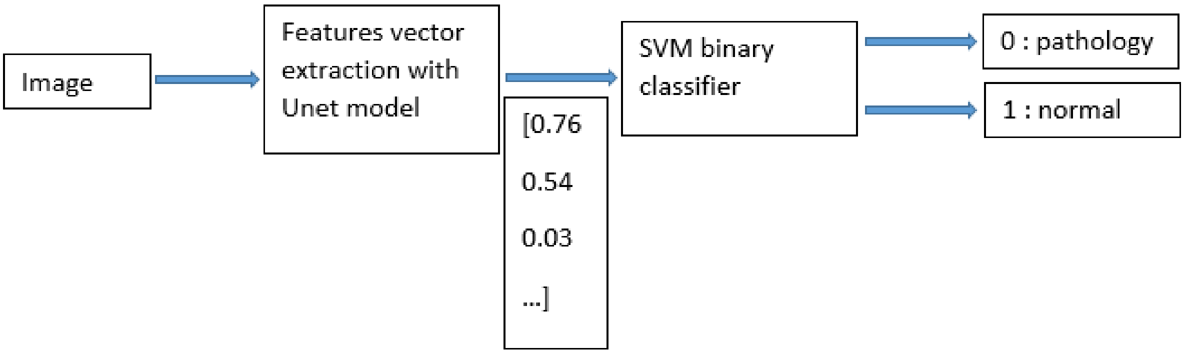


Figure 21: Other perspective.

The new way of dealing with this problem is something new in this field as the features of pathologies are not human made. This can be an introduction of machines in replacing the human interpretation and intervention. But it still suffer from the accuracy which is not too close to 100% and it cannot be applied alone.

Conclusion

In the end of the internship project, there is a deep thrives in the way of thinking, interpreting the result and making decision. With the teacher's help I acquire a capabality of convincing and talking about the project and highlighting the work that I do.

Deep learning is the key to solve a large set of problems as the medical imagery field that become essential in detecting patterns, organs and infections … The artificial intelligence make the machines capable of imitating humans work and performing in a more precise way : The extraction of features of images at a pixel level is done through the intervention of deep neural network which allow detection of patterns, thus detection of existence of the object and locating it in the image.

In this project I learn how to choose the appropriate model to satisfy our needs and how to get the best performance of it. And I have a detailed knowledge of image processing tasks like image segmentation and generating masks. Last but not least The internship in COSIM Lab allows me to get a general idea abou research field and how it develop innovative solution for the improvement of future.

References

https://androidkt.com/tensorflow-keras-unet-for-image-image-segmentation/

https://medium.com/coinmonks/learn-how-to-train-u-net-on-your-dataset-8e3f89fbd623

https://missinglink.ai/guides/tensorflow/tensorflow-image-segmentation-two-quick-tutorials/

https://nanonets.com/blog/how-to-do-semantic-segmentation-using-deep-learning/

https://www.hopkinsmedicine.org/health/conditions-and-diseases/vocal-cord-disorders

https://www.ncbi.nlm.nih.gov/pubmed/16676550

http://www.entusa.com

https://scikit-learn.org/stable/modules/svm.html

https://medium.com/@arthur_ouaknine/review-of-deep-learning-algorithms-for-image-semantic-segmentation-509a600f7b57

https://towardsdatascience.com/yolo-you-only-look-once-real-time-object-detection-explained-492dc9230006

http://deeplearning.net/tutorial/unet.html

https://github.com/ldenoue/keras-unet