UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Leonardo de Araujo Monte

Semantic Segmentation for People Detection on Beach Images

Recife

2021

Leonardo de Araujo Monte

Semantic Segmentation for People Detection on Beach Images

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Valmir Macário Filho

Recife

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

M772s Monte, Leonardo de Araujo

Semantic segmentation for people detectionon beach images / Leonardo de Araujo Monte. - 2021.
25 f. : il.

Orientador: Valmir Macario Filho.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2021.

1. Segmentação semântica. 2. Detecção de pessoas. 3. Aprendizado profundo. I. Filho, Valmir Macario, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Leonardo de Araújo Monte às 14 horas do dia 01 de março de 2021, no link https://meet.google.com/oqm-wciq-djy, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado Semantic Segmentation for People Detection on Beach Images, orientado por Valmir Macário Filho e aprovado pela seguinte banca examinadora:

____

Valmir Macário Filho
DC/UFRPE

____

Filipe Rolim Cordeiro
DC/UFRPE

Resumo

As câmeras de monitoramento estão sendo cada vez mais aperfeiçoadas com o uso de sistemas de visão computacional capazes de identificar situações de risco. Este trabalho faz parte de um sistema de rastreamento automático de monitoramento de praias na região metropolitana do Recife, com o objetivo de evitar que banhistas ultrapassem os limites seguros na região de banho de praia. A segmentação semântica tem ganhado força em diferentes tarefas de visão computacional. Geralmente a meta-arquitetura de uma rede de segmentação semântica consiste em dois módulos: codificador (backbone) e decodificador. Este trabalho realiza um estudo combinando um conjunto de redes de segmentação semântica, U-net, Xnet, LinkNet e Unet++ com os backbones pré-treinados VGG16 e VGG19, com o objetivo de detectar banhistas em imagens de praia. Nós utilizamos a nossa própria base de dados, constituída de diferentes imagens da praia de Boa Viagem, Recife-Brasil. Os algoritmos foram avaliados com a métrica MIoU utilizando toda a cena da imagem, e apenas a faixa de mar. O melhor resultado de MIoU com relação à imagem completa foi 80.87%, e foi obtido pela XNet com o backbone da VGG19. O melhor MIoU na detecção de banhistas na faixa de mar obteve 85.56% e foi alcançado com a LinkNet com os backbones da VGG16 e VGG19.

Palavras-chave: Segmentação Semântica, Detecção de Pessoas, Aprendizado Profundo.

Abstract

Cameras monitoring are increasingly aided by computer vision systems that identify risk situations. This work is part of an automatic track system to monitor beaches in the metropolitan area of Recife in order to prevent bathers to trespass the boundaries of the safe region for swimming. Semantic segmentation has gained strength in several computer vision tasks. Usually, the meta-architecture of a semantic segmentation network consists of two modules: encoder (backbone) and decoder. This work does a study combining a set of semantic segmentation networks, U-net, Xnet, LinkNet and Unet++ with the pre-trained backbones VGG16 and VGG19, to detect swimmners in beach images. We have used our own dataset, made by several images taken at the Boa Viagem beach, Recife-Brazil. The algorithms are evaluated with MIoU metric regarding the entire image scene and just in the water area. The best MIoU regarding all image was 80.87best MIoU in detecting swimmers at the beach was 85.56obtained by the LinkNet algorithm with both VGG16 and VGG19 backbones.

Keywords: Semantic Segmentation, Person Detection, Deep Learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CEMIT	Shark Incident Monitoring Committee of the state of Pernambuco
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CSP	Center and Scale Prediction
Faster R-CNN	Faster Region Based Convolutional Neural Networks
FCN	Fully Convolutional Network
FCN8	Fully Convolutional Network8
MIoU	Mean Intersection Over Union
NAS	Neural Architecture Search
NDNet	Narrow Deep Network
IoU	Intersection Over Union
PCA	Principal Component Analysis
R-FCN	Region-based Fully Convolutional Networks

Sumário

1 Introduction

According to the statistic report published by the Shark Incident Monitoring Committee of the state of Pernambuco (CEMIT) (CEMIT, 2021) in Brazil, 62 incidents with sharks happened in Pernambuco, 24 of them having occurred in the Boa Viagem beach, in Recife. Along all the coast of Boa Viagem beach, there are signs advertising bathers to stay within the borders of natural occurring reefs or up to waist-high waters, due to the risk of shark attacks. However, some swimmers trespass the safe limit of the beach, increasing the probability of an accident to happen to them. Due the large extension of the beaches, requiring human effort to supervise the entire coast is impractical, making a machine learning-guided solution, on the condition that it produces fast and accurate detection, ideal in this situation.

Cameras monitoring are increasingly aided by computer vision systems that identify risk situations(CHEN; SURETTE; SHAH, 2020). As some of these areas need to be continually monitored for dangerous situations, an automated system would be an effective risk control measure. The most significant challenges for this problem are variable scene illumination, partial occlusion and distant camera position (Chevtchenko et al., 2018). Another limitation is the acquisition of positive samples, in our case images of people in the water: as instructed by the warning signs, most of the bathers avoid bathing. This work is part of an automatic track system, in order to prevent bathers to trespass the boundaries of the safe region for swimming. In case of a bather trespass a safe line, an alert is emitted for the responsible authorities, in order to take necessary precautions and avoid unwanted accidents.

Some classic automatic track systems has three steps: segmentation, classification and tracking (ROUGIER et al., 2013). The first step is the preprocessing and segmentation of the beach images, the second step is responsible to decide whether or not the segmented object is a person and the third step is where the people classified is tracked. A semantic segmentation network (YU et al., 2018) is able to do both, the segmentation and the classification steps simultaneously, classifying image pixels into semantic classes such as sky, road, person, vehicle, and other objects on the scene.

Semantic segmentation has gained strength in several computer vision tasks, like autonomous driving (MICLEA; NEDEVSCHI, 2019), human-machine interaction (WONG et al., 2017), handwritten text segmentation (JO et al., 2020), and other applications (GARCIA-GARCIA et al., 2017). Usually, the meta-architecture of a semantic segmentation network consists of two modules: encoder and decoder. The encoder module execute a scale-decreased network, which is commonly referred to a backbone model (e.g., VGG, ResNets). Then a decoder network is applied to the backbone to recover the feature resolutions (e.g., Unet, LinkNet networks). A typical decoder network consists of a combination of low-level and high level features from a backbone to generate strong multi-scale feature maps.

This work is a study of combining a set of semantic segmentation networks, U-net(RONNEBERGER; FISCHER; BROX, 2015), Xnet(BULLOCK; CUESTA-LÁZARO; QUERA-BOFARULL, 2018), LinkNet(CHAURASIA; CULURCIELLO, 2017) and Unet++(ZHOU et al., 2018), with the backbones VGG16 and VGG19, to detect bathers in beach images. Each semantic segmentation network was choose using two premises, either the network was made for perform well on small datasets or the network has a faster training and inferring time, in comparison with other semantic segmentation networks. We have used our own dataset, made by several images taken at the Boa Viagem beach, Recife-Brazil. The algorithms are evaluated with MIoU metric in two different approaches. In the first, are only evaluated the network using the entire image scene, including the beach and coast and in the second, the evaluations performance is computed just in the water area.

2 Related Works

2.1 Person detection on beach images

The task of people detection in beach scenarios has been tackled in the literature as can be seen in Green at al. (Green et al., 2005), where the authors developed a people detection system using a database formed by images of persons and non-persons (objects that can be misunderstood as persons). The objects were segmented using the breadth first search and Canny edge detector. The results using a Multi-layer Perceptron were of 91% of true positives and 13% of false positives. Following the same idea, Silva et al. (Luna da Silva et al., 2017) proposed a system for people detection in beach scenarios using a different set of feature descriptors and classifiers. The authors used a database formed by images of persons and non-persons taken in the Boa Viagem, Recife-Brazil beach. The best recognition rate found was 90.31% using the PCA technique, an combination of the descriptors HOG and LBP and the radial-based Support Vector Machine classifier. The work of Chevtchenko et al. (Chevtchenko et al., 2018) use Deep Learning meta-architectures Faster R-CNN, R-FCN and SSD combined with Classification algorithms pretrained with COCO dataset. The authors founded that SSD detectors were an order of magnitude faster than R-FCN and Faster R-CNN, but had struggled to detect distant objects. In other hand, they found that the Faster R-CNN with Resnet 101 provided a significant better detection, but at 5.6 frames per second with a GTX 1080 Ti graphic card.

2.2 Semantic segmentation

Despite the bathers detection move forward to using deep learning methods, the methods used CNN meta-architectures to feature detection combined with classification algorithms. The results of these algorithms are placed in bounding-boxes, losing location precision of the person while needs more time of execution and training to obtain good results. The semantic segmentation algorithms have been used in recent literature to get good results in pixel-wise labeled outputs in a more natural way, significantly reducing the difficulty in training, being faithful to the person location in real time (SIAM et al., 2018; LIU et al., 2019b; YANG et al., 2020). The work of Siam et al. (SIAM et al., 2018) made a study with the trade-off between accuracy and computational efficiency with combination of meta-architectures of different methods of Backbones and Encoder methods to segment objects on cityscape dataset. The Liu et al. (LIU et al., 2019b) looks for central points where there are pedestrians proposing a new method called

Center and Scale Prediction (CSP). The proposed method has two components, i.e. the feature extraction module and the detection head. The CSP algorithm achieved the new state-of-the-art performance on two challenging pedestrian detection benchmarks, Citypersons and Caltech. The Narrow Deep Network (NDNet) was proposed by Yang et al. (YANG et al., 2020), that use a separable convolution based bottleneck structure modifying the fully convolutional network8 (FCN8) (LONG; SHELHAMER; DARRELL, 2015) structure with a learned score fusion and a small object augmentation to identify more small objects in Cityscapes dataset.

Our work is motivated by the recent success of semantic segmentation algorithms for person detection, including approaches that detects small objects in real time. This work is a study of semantic segmentation networks with the premises of performing well in a small dataset or with a fast training and inference time to detect bathers in beach images.

3 Background

In this work are evaluated four recently proposed Convolutional Neural Networks for semantic segmentation to detect people in beach images. In the next subsections, is introduced the meaning of semantic segmentation and those CNNs.

3.1 Semantic Segmentation

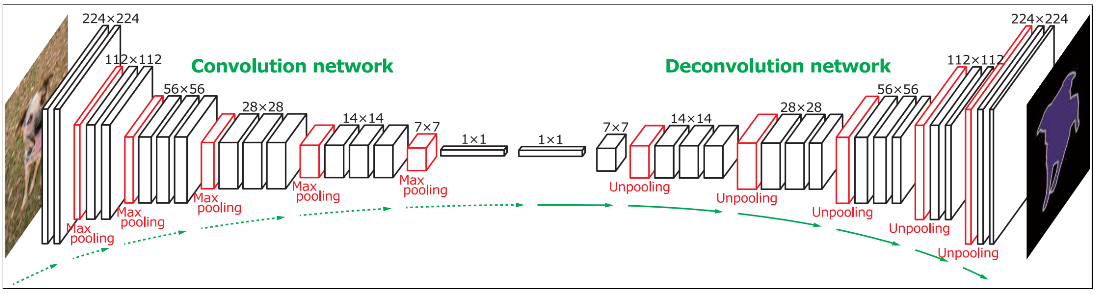


Figura 1 – Semantic segmentation architecture (NOH; HONG; HAN, 2015).

Semantic segmentation is a computer vision task that paves the way towards complete scene understanding. It can be seen as a process of classification on pixel level, where an entry image is labeled pixel by pixel to the correspondent label image. The pixel by pixel problem can be reduced to the following formulation: Find a way to assign one label from the label set $L = \{l_1, l_2, \dots, l_k\}$ for each one of the pixels of a 2D image with dimensions W x $H = N$ pixels. Each label l represents a different class or object, that can be as an example, a plane, a car, a traffic light or in the case of this paper, a bather. The space of the labels set has k possible classes, where is commonly extended to $k + 1$ been l_0 the background or the empty class(GARCIA-GARCIA et al., 2017).

Many state-of-the-art deep learning semantic segmentation techniques follow the basic structure of the Long et al. (LONG; SHELHAMER; DARRELL, 2015) proposed Fully Convolutional Network (FCN). The basic idea is to take advantage of existing deep learning classification networks, by replacing the fully connected layers with a probability map in the same size to input image, indicating probability of each pixel that belongs to one of the predefined classes. Those maps are upsampled using deconvolutions operations, for example, bilinear interpolation, to produce dense per-pixel labeled outputs (GARCIA-GARCIA et al., 2017; NOH; HONG; HAN, 2015). Next subsections are described some state-of-the-art deep learning semantic segmentation techniques: Unet, Xnet, LinkNet and Unet++.

3.1.1 U-net

(RONNEBERGER; FISCHER; BROX, 2015) is a Convolutional Neural Network for semantic segmentation built for medical images segmentation. Due to the small size of the medical images dataset, this network was made with the purpose of work efficiently with a small quantity of data. The network receive that name because of the U shape, which is due to the fact that its configuration of layers, where the first half is called contracting path and the second half is called expansive path. Contracting path is the first step of the network which performs a series of convolutions followed by maxpoolings in the data, in order to extract information on different levels from the image. The expansive path is the second part of the network which is performed convolutions followed by upsamplings, to recover the original size of the image and extract useful information in the process.

3.1.2 Xnet

(BULLOCK; CUESTA-LÁZARO; QUERA-BOFARULL, 2018) is a Convolutional Neural Network just like the U-net, created for medical images segmentation, more precisely for x-ray images. In the first quarter of the transformations, the network performs convolutions followed by maxpoolings. After that process, is performed the upsampling which is extracted a precise localization of the image information. This process is repeated one more time and at the end is applied a activation layer for the pixel by pixel classification of the segmented image.

3.1.3 LinkNet

(CHAURASIA; CULURCIELLO, 2017) is a Convolutional Neural Network for semantic segmentation created with the premise of achieve similar results to the state of the art networks, with a low computational cost and without necessarily increase the network parameters quantity. The encoder consists in convolutions followed by dimension reductions of the image by a factor of two, while the decoder consists in convolutions followed by dimension increase of the image also by a factor of two, so that is possible to extract information in several scales and regain the original dimension for the pixel by pixel classification.

3.1.4 Unet++

(ZHOU et al., 2018) is a Convolutional Neural Network for semantic segmentation created from the Unet architecture, which consists in the application of a contracting path and an expansive path, just like in the Unet. In the contracting path are performed convolutions followed by maxpoolings so it can be possible to extract different types of information from the image, while reducing your original size. In the expansive path are performed convolutions followed by upsamplings so it can be possible to extract information as same as in the contracting path, while increasing the image to the original size. Between each stage of the contracting path and expansive path it is performed a series of dense convolutions, so that the network can propagate the activation from one part to the other with a smaller semantic difference. Each layer has a skip connection function that consists in propagate the layers activation's from the contracting path to the expansive path.

4 Experimental Evaluation

This section presents the methodology and results of evaluating the performance of bathers detection in beach images. The semantic segmentation algorithms, Unet, Unet++, Xnet and Linknet are evaluated with different backbones: VGG16 and VGG19. The backbones are pretrained on a large ImageNet dataset (DENG et al., 2009) containing 1000 classes of objects, including people with different levels of scale and occlusion.

4.1 Dataset

The image dataset consists of photos taken from life-guard lookouts at Boa Viagem beach, Recife-Brazil. Our dataset consists of 300 images, with two possible classes, person and background, where everything that is not a person in the image is considered from the background class. The dataset contain a total of 14023 persons labels manually obtained by using the LabelMe (WADA, 2016) online tool. The persons are presented in the images in different levels of occlusion, where objects like beach umbrellas are occluding some persons body parts, other important fact is that some persons in the image are partly immerse in the water, sometimes only been visible one body part, for an example the head, what makes the people detection problem in beach images challenging. You can see a label image containing persons labels, where some people are in the water and some are in the sand part of the beach as shown in the 2, the original image in the 3 and the image 4 containing the sea, sand and sky labels used for the evaluation of the networks taking in regard different parts of the image.

4.2 Evaluation method

The metric evaluation used in the experiments of this paper is the Mean Intersection Over Union (MIoU). The MIoU metric calculate the division between the true positives (intersection) per the sum of the true positives, false negatives and false positives (union). Due to the imbalance of the classes in the problem, where the number of persons pixels is considerably smaller than the number of background pixels, the IoU is calculated for each class, person and background (non-person), and then, the mean of the two classes is computed. The MIoU is calculated using the following equation, where there is $k+1$ classes from l_0 to l_k, being l_0 the background class, p_{ii} represents the true positives, p_{ij} are the false positives and p_{ji} are the false negatives (GARCIA-

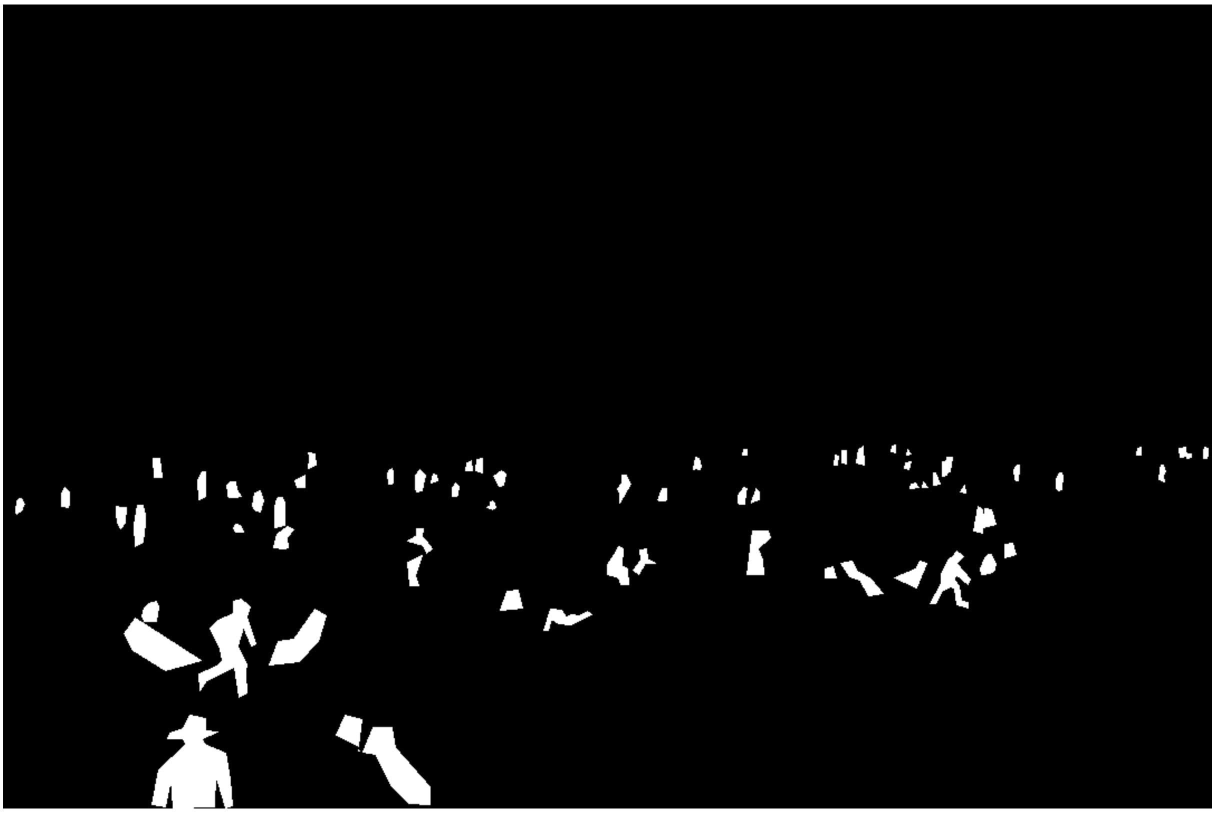


Figura 2 – Example of the image labeled by persons in the dataset.

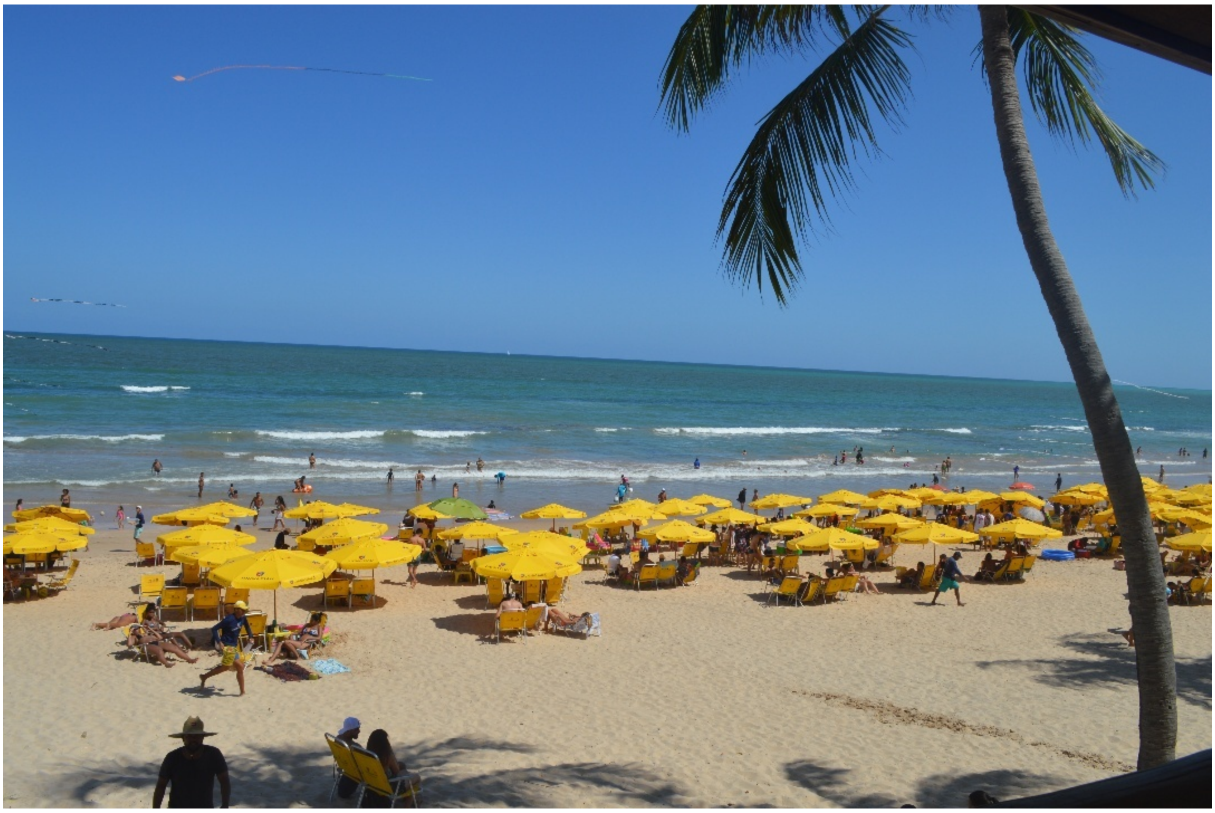


Figura 3 – Example of the image in the dataset.

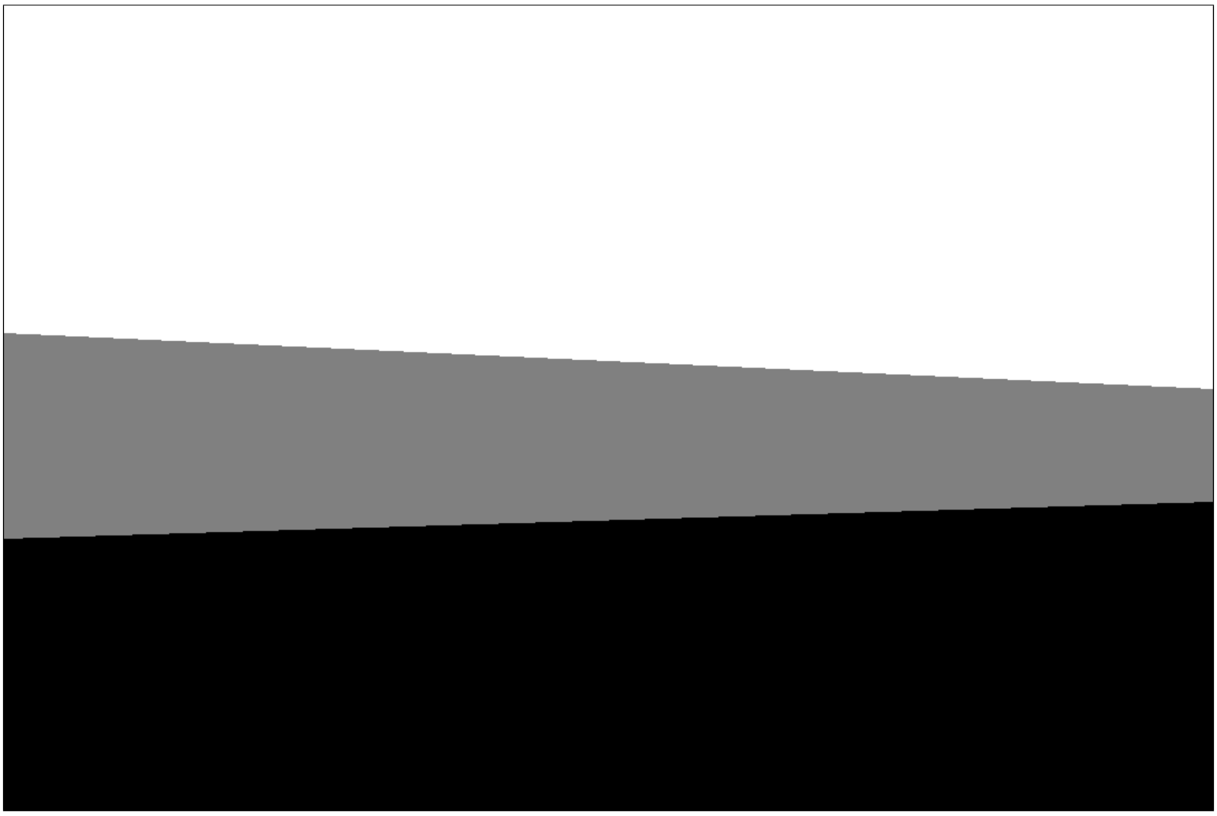


Figura 4 – Example of the image labeled by sky, sea and coast in the dataset.

GARCIA et al., 2017).

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{4.1}$$

4.3 Parameters selection

Deep neural networks has a set of hyper-parameters that can be tuned in order to find the best possibles results (KOUTSOUKAS et al., 2017). For each network a set of hyper-parameters was chosen based initially in the original article implementation and then refined throughout the best results in the experimentation tests. The following items describe each hyper-parameter used for the semantic segmentation networks.

- **Learning rate:** For the U-net the learning rate used was 0.001, for the Xnet was used 0.0001, for the Linknet it was 0.001 and for the Unet++ it was 0.0003. All the semantic segmentations networks used learning rate decay with the decay value 0.0001.
- **Backbone:** Each of the semantic segmentation network was evaluated using a backbone CNN architecture. The backbones used in the experiments were the VGG16 and the VGG19 (SIMONYAN; ZISSERMAN, 2015).

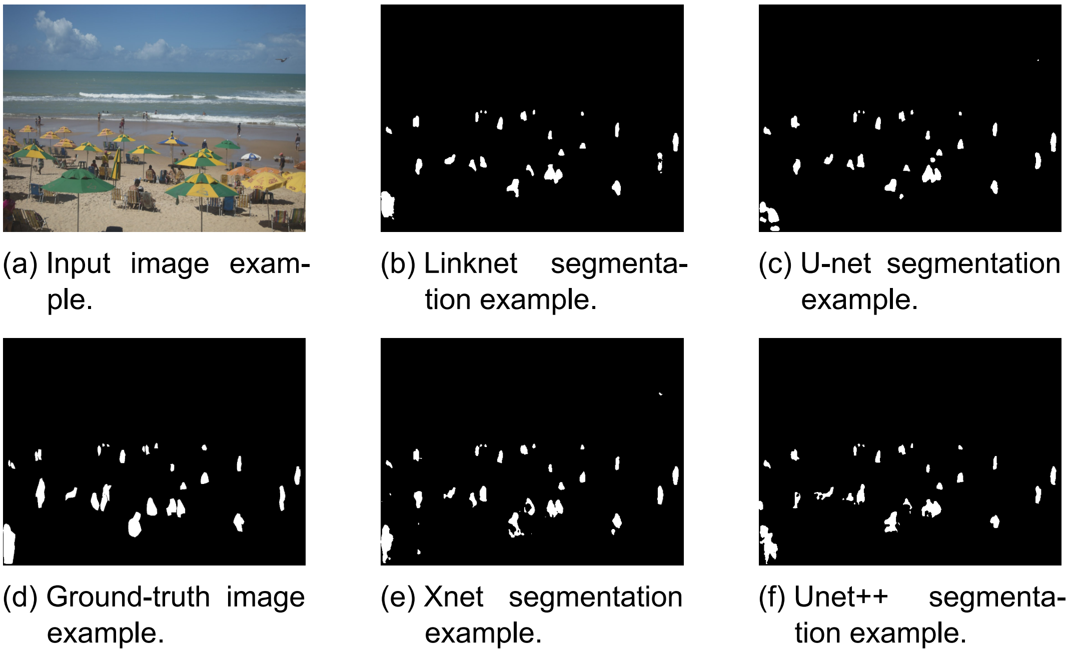
- **Activation Function of the output layer:** The activation function used for the result experiments was the sigmoid function.
- **Loss function:** The loss function used in the semantic segmentation networks was the Jaccard coefficient, also known as mean intersection over union (BERMAN; TRIKI; BLASCHKO, 2018).
- **Optimizer:** For the U-net, Xnet and Unet++ the optimizer used was the Adam and for the LinkNet it was used the RMSprop, based in the experimental results.
- **Number of epochs:** For the results experiments it was chosen the number 200 for the number of epochs, it was also used the early stopping technique with the number of epochs choose to stop in case of not any improvement in the performance of 25 epochs to avoid overfitting.
- **Weight initialization:** Each of the semantic segmentation networks were evaluated using two types of weight initialization. The first one was using the weights of the network pretrained in the imagenet(KRIZHEVSKY; SUTSKEVER; HINTON, 2012) dataset and the second one had the weights initialized using Glorot Uniform initialization (HANIN; ROLNICK, 2018) that is the default parameter for weight initialization of the Keras(CHOLLET et al., 2015) library, therefore non pre-trained weights.

4.4 Evaluation method

For the experimental evaluation it was used the 5-fold cross validation technique executed 6 times with different dataset configurations. So, the mean and standard deviation of the MIoU, and IoUs of the person and background classes. These results were computed using two types of evaluations: the first evaluation was made taking in regard the persons of the entire scene, sand and sea and the second evaluation was made computing the same metrics of just the sea area. The second one was computed because the focus of this work is to detect the swimmers in the sea part of the beach.

5 Results

A qualitative comparison between an input image example, the ground-truth and the networks segmentation outputs can be seen in the Figure 5 and a comparative summary is presented in the bellow tables, presenting the mean and standard deviation for the MIoU and the IoU of the person and background classes. The networks have two possible types, pre-trained (pre) and non pre-trained (nopre). The pre-trained is the one with the weight initialization from the pre-trained network in the Imagenet and the non pre-trained is the network with the default weight initialization of the Keras library. For each one of the results there are the values from the mean and the standard deviation respectively. The Table 1 and the Table 2 show the results of the networks evaluation in the entire image scene, where are evaluated the network performance both in the sea and in the shore part. The Table 3 and the Table 4 show the results from the evaluation of the networks taking in regard only the sea area of the image.


(a) Input image example.

(b) Linknet segmentation example.

(c) U-net segmentation example.

(d) Ground-truth image example.

(e) Xnet segmentation example.

(f) Unet++ segmentation example.

Figura 5 – Qualitative comparison between the networks.

The results of the non pre-trained networks evaluated in the entire image scene, showed that the Unet with VGG16 backbone achieved the best results in the MioU, the person class and background IoUs, however the Linknet had the worst results, where the in the experiment with the VGG16 backbone achieved only 12.41% for IoU from the person class.

Differently from the non pre-trained networks evaluation, in the entire scene evaluation, the pre-trained networks had better results overall, where all them achieved

Tabela 1 – Mean and standard deviation from the experiments results from the non pre-trained networks on the evaluation of the entire image scene.

Modelo	MIoU	person	Background
Unet-nopre-VGG16	**79.63**(**0.006**)	**60.94**(**0.012**)	**98.33**(**0.0009**)
Xnet-nopre-VGG16	77.19(0.100)	56.27(0.019)	98.12(0.0011)
Linknet-nopre-VGG16	54.63(0.130)	12.41(0.252)	96.85(0.0078)
Unet++-nopre-VGG16	71.22(0.117)	44.63(0.228)	97.80(0.0068)
Unet-nopre-VGG19	79.57(**0.006**)	60.82(**0.012**)	98.32(**0.0009**)
Xnet-nopred-VGG19	77.10(0.009)	56.09(0.017)	98.12(0.0011)
Linknet-nopre-VGG19	57.79(0.146)	18.55(0.283)	97.03(0.0088)
Unet++-nopre-VGG19	77.77(0.012)	57.35(0.023)	98.19(0.0012)

Tabela 2 – Mean and standard deviation from the experiments results from the pre-trained networks on the evaluation of the entire image scene.

Modelo	MIoU	person	Background
Unet-pre-VGG16	80.13(0.012)	61.87(0.024)	98.38(0.0010)
Xnet-pre-VGG16	80.79(**0.004**)	63.18(**0.008**)	98.40(**0.0007**)
Linknet-pre-VGG16	80.54(0.006)	62.67(0.011)	**98.42**(0.0008)
Unet++-pre-VGG16	79.96(0.005)	61.58(0.010)	98.35(0.0009)
Unet-pre-VGG19	80.06(0.006)	61.74(0.011)	98.38(0.0009)
Xnet-pre-VGG19	**80.87**(**0.004**)	**63.31**(0.009)	**98.42**(**0.0007**)
Linknet-pre-VGG19	80.63(**0.004**)	62.84(**0.008**)	98.41(0.0008)
Unet++-pre-VGG19	80.05(0.006)	61.75(0.012)	98.35(0.0009)

values above 60% for the IoU from the person class, as shown in the Table 2, having the best results been achieved by the Xnet with VGG19 backbone.

Tabela 3 – Mean and standard deviation from the experiments results from the non pre-trained networks on the evaluation of the sea part of the image.

Modelo	MIoU	person	Background
Unet-nopre-VGG16	84.24(**0.005**)	69.42(**0.010**)	**99.06**(**0.0005**)
Xnet-nopre-VGG16	82.38(0.009)	65.81(0.017)	98.94(0.0006)
Linknet-nopre-VGG16	55.82(0.145)	13.96(0.284)	97.67(0.0072)
Unet++-nopre-VGG16	75.00(0.134)	51.40(0.263)	98.61(0.0064)
Unet-nopre-VGG19	**84.28**(0.006)	**69.49**(0.013)	**99.06**(0.0006)
Xnet-nopre-VGG19	82.21(0.008)	65.48(0.017)	98.93(0.0007)
Linknet-nopre-VGG19	59.44(0.164)	21.03(0.321)	97.85(0.0080)
Unet++-nopre-VGG19	82.57(0.011)	66.17(0.023)	98.96(0.0007)

About the evaluation of non pre-trained networks taking into account only the sea area of the image, the Unet with VGG19 backbone achieved the best results and similarly to the experiments presented in the Table 1, the Linknet had the worst results, with 13.96% for the person class IoU.

The pre-trained networks evaluated only in the sea area of the image had similar results to the non pre-trained networks, where the Linknet with both, VGG16 and

Tabela 4 – Mean and standard deviation from the experiments results from the pre-trained networks on the evaluation of the sea part of the image.

Modelo	**MIoU**	**person**	**Background**
Unet-pre-VGG16	84.49(0.005)	69.90(0.011)	**99.08**(**0.0005**)
Xnet-pre-VGG16	84.03(0.005)	69.02(0.010)	99.04(**0.0005**)
Linknet-pre-VGG16	**84.56**(**0.004**)	**70.05**(**0.009**)	**99.08**(**0.0005**)
Unet++-pre-VGG16	83.51(0.005)	68.02(0.010)	99.00(0.0006)
Unet-pre-VGG19	84.41(0.006)	69.74(0.011)	99.07(**0.0005**)
Xnet-pre-VGG19	84.10(**0.004**)	69.16(**0.009**)	99.04(**0.0005**)
Linknet-pre-VGG19	**84.56**(0.005)	70.03(0.011)	**99.08**(**0.0005**)
Unet++-pre-VGG19	83.57(0.005)	68.13(0.010)	99.01(0.0006)

VGG19 backbones achieved the best results for all metrics, as can been seen in the Table 4.

If the objective is detect persons in the entire image of the beach, the XNet with VGG19 backbone obtained the best results for all metrics. Meantime, if the objective is to detect the swimmers in the sea, as this project, the best results were obtained by the Linknet with both, VGG16 and VGG19 backbones for all metrics. The evaluation using only the sea part of the image had better results in comparison with the ones evaluated using the entire image scene, one possible explanation is the fact that the network has a bigger quantity of pixels from diverse objects in the shore part of the image, like beach umbrellas, beach chairs, balls and other possible elements that can interfere in the network segmentation, while in the sea part of the image the majority of the pixels are composed only by the water and the persons in the water.

The pre-trained networks achieved better results overall in comparison with the non pre-trained, the difference is bigger in the person class IoU evaluation. The Linknet had the biggest gap between the values of the IoU from the person class in the pre-trained network and the non pre-trained network, where in the Table 1 the Linknet with the VGG16 backbone acheived 12.41% and in the Table 2 with the same backbone and the same evaluation method, however pre-trained with the imagenet dataset achieved 62.67%. In comparison of backbones the Xnet, Linknet and Unet++ using the VGG19 backbone achieved better results than using the VGG16 backbone and in the case of Unet that improvement did not happened in the evaluation of the entire image scene, where the result with the VGG16 backbone achieved a better result than the one with the VGG19 backbone, however in the evaluation of the non pre-trained networks with the imagenet dataset in only the sea part of the image the Unet with the VGG19 backbone improved in comparison with the VGG16 backbone.

The qualitative results in the Figure 5 has shown that the networks were capable to segment all the persons in the image, but some pixels were lost in the process that made the IoU metric results decay. On the other hand the semantic segmentation networks efficiently segmented all the background pixels, as an example the Linknet with the VGG19 backbone achieved 99.08% of IoU for the background class, what can be interpreted as an low false positive rate, due to the fact that the network usually classify correctly what is not an pixel of the person class.

6 Conclusions and Future Work

This study analysed four semantic segmentation networks in the people detection problem in the context of beach environment, for each one it was used two different backbones, VGG16 and VGG19, where all the networks with the pre-trained weights on the Imagenet dataset had achieved better results in comparison with the ones with the weights initialized using the default Keras function for weight initialization.

The networks were evaluated using two different types of evaluation and the MIoU metric, the first evaluation type was applying the MIoU in the entire image scene, the network that had the best performance was the Xnet using the VGG19 backbone and pre-trained with the imageset dataset. The second evaluation type was applying the MIoU only in the sea part of the image, to see how well the networks had performed only in the sea and the network that had the best performance was the Linknet using either the VGG16 and the VGG19 backbone both pre-trained in the imagenet dataset. The focus of this project is to track beachgoers in the sea, thus being capable to avoid accidents, due that the network that had the best possible result to be used in the people tracking system was the Linknet.

As future work, new techniques like Conditional random fields (CRF) (Zhou et al., 2016) to refine the output of the segmentation networks, NAS (Neural Architecture Search) (LIU et al., 2019a) to search other architectures variations that improve the performance of the segmentation and other metrics capable to count the number of persons segmented in comparison with the number of persons present in the image and prediction time can be taken in consideration to chose the best algorithm.

Referências

BERMAN, M.; TRIKI, A. R.; BLASCHKO, M. B. *The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks*. 2018. Citado na página 18.

BULLOCK, J.; CUESTA-LÁZARO, C.; QUERA-BOFARULL, A. Xnet: A convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets. *CoRR*, abs/1812.00548, 2018. Disponível em: <http://arxiv.org/abs/1812.00548>. Citado 2 vezes nas páginas 9 e 13.

CEMIT. *Statistics of Shark Incidents in the state of Pernambuco-Brazil*. 2021. Disponível em: <https://www.sds.pe.gov.br/cemit/52-cemit/196-estatisticas>. Citado na página 8.

CHAURASIA, A.; CULURCIELLO, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. *CoRR*, abs/1707.03718, 2017. Disponível em: <http://arxiv.org/abs/1707.03718>. Citado 2 vezes nas páginas 9 e 13.

CHEN, C.; SURETTE, R.; SHAH, M. Automated monitoring for security camera networks: promise from computer vision labs. *Security Journal*, Feb 2020. ISSN 1743-4645. Disponível em: <https://doi.org/10.1057/s41284-020-00230-w>. Citado na página 8.

Chevtchenko, S. et al. Deep learning for people detection on beach images. In: *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2018. p. 218–223. Citado 2 vezes nas páginas 8 e 10.

CHOLLET, F. et al. *Keras*. [S.l.]: GitHub, 2015. <https://github.com/fchollet/keras>. Citado na página 18.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255. Citado na página 15.

GARCIA-GARCIA, A. et al. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017. Citado 3 vezes nas páginas 8, 12 e 17.

Green, S. et al. The detection and quantification of persons in cluttered beach scenes using neural network-based classification. In: *Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05)*. [S.l.: s.n.], 2005. p. 303–308. Citado na página 10.

HANIN, B.; ROLNICK, D. *How to Start Training: The Effect of Initialization and Architecture*. 2018. Citado na página 18.

JO, J. et al. Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications*, Springer, v. 79, n. 43, p. 32137–32150, 2020. Citado na página 8.

KOUTSOUKAS, A. et al. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, v. 9, n. 1, p. 42, Jun 2017. ISSN 1758-2946. Disponível em: <https://doi.org/10.1186/s13321-017-0226-y>. Citado na página 17.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012. p. 1097–1105. Disponível em: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>. Citado na página 18.

LIU, C. et al. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. Citado na página 23.

LIU, W. et al. High-level semantic feature detection: A new perspective for pedestrian detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 5187–5196. Citado na página 10.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3431–3440. Citado 2 vezes nas páginas 11 e 12.

Luna da Silva, R. et al. Detecting people from beach images. In: *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. [S.l.: s.n.], 2017. p. 636–643. Citado na página 10.

MICLEA, V.-C.; NEDEVSCHI, S. Real-time semantic segmentation-based stereo reconstruction. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 21, n. 4, p. 1514–1524, 2019. Citado na página 8.

NOH, H.; HONG, S.; HAN, B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 1520–1528. Citado 2 vezes nas páginas 4 e 12.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. Disponível em: <http://arxiv.org/abs/1505.04597>. Citado 2 vezes nas páginas 9 e 13.

ROUGIER, C. et al. 3d head tracking for fall detection using a single calibrated camera. *Image and Vision Computing*, Elsevier, v. 31, n. 3, p. 246–254, 2013. Citado na página 8.

SIAM, M. et al. Rtseg: Real-time semantic segmentation comparative study. In: IEEE. *2018 25th IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2018. p. 1603–1607. Citado na página 10.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA,* *May 7-9, 2015, Conference Track Proceedings*. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1409.1556>. Citado na página 17.

WADA, K. *labelme: Image Polygonal Annotation with Python*. 2016. <https://github.com/wkentaro/labelme>. Citado na página 15.

WONG, J. M. et al. Segicp: Integrated deep semantic segmentation and pose estimation. In: IEEE. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. [S.l.], 2017. p. 5784–5789. Citado na página 8.

YANG, Z. et al. Small object augmentation of urban scenes for real-time semantic segmentation. *IEEE Transactions on Image Processing*, IEEE, v. 29, p. 5175–5190, 2020. Citado 2 vezes nas páginas 10 e 11.

YU, H. et al. Methods and datasets on semantic segmentation: A review. *Neurocomputing*, Elsevier, v. 304, p. 82–103, 2018. Citado na página 8.

Zhou, H. et al. Image semantic segmentation based on fcn-crf model. In: *2016 International Conference on Image, Vision and Computing (ICIVC)*. [S.l.: s.n.], 2016. p. 9–14. Citado na página 23.

ZHOU, Z. et al. Unet++: A nested u-net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. [S.l.]: Springer, 2018. p. 3–11. Citado 2 vezes nas páginas 9 e 13.