XLI CILAMCE
XLI Ibero-Latin American Congress on Computational Methods in Engineering
16-19 November, 2020 FOZ DO IGUAÇU PARANÁ - BRAZIL

Artificial Intelligence usage for identifying automotive products

Leandro M. Gonzaga[1], Gustavo M. de Almeida[1]

[1]*Dept. of Control and Automation Engineering, Federal Institute of Espírito Santo*
Rodovia ES-010 – KM 6,5 – Manguinhos, 29173-087, Espírito Santo, Brazil
leandromoreiragonzaga2015@gmail.com, gmaia@ifes.edu.br

Abstract. The Computational Vision process has been presenting a huge development in the recent years. This is occurring thanks to the development in the field of Artificial Neural Networks, specially the Convolutional Neural Networks. These networks are capable of training to identify patterns in a large set of images, for latter identifying these same patterns in other images. A very common architecture used nowadays, due to its high accuracy, is the Mask R-CNN. This architecture not only classifies and identifies objects, but also realizes its segmentation pixel by pixel. In this present work, Mask R-CNN was used for segmentation of automotive products (windshields, headlights, tail lights, bumpers and rearview mirrors) in an aftermarket organization. In its evaluation, the algorithm presented a significantly high mAP and accuracy – checked through a confusion matrix, even reaching a val_loss of 2.413, demonstrating a satisfactory result for its proposed applications: a system filter for preventing human error and a premise for future works of identifying defects in the mentioned products.

Keywords: automotive products, computational vision, Mask R-CNN, artificial intelligence.

1 Introduction

Computational Vision is a research area that has been presenting a huge development. According to Gonzales and Woods [1], this occurs due to two main reasons: the possibility of information improvement for human interpretation, and image data processing for storage, transmission and representation for implementation in autonomous machines. Studies involving this area, specifically object recognition, are basically focused on trying to resemble the capacity that the human brain has in recognizing three-dimensional objects based only in a bi-dimensional image, as stated by Hogendoom [2].

Images contains huge amounts of information that are perceptible by human eyes and, due to this, Carvalho [3] quoted that images have become data sources for researches in areas involving Computational Vision.

In a short period, Computational Vision reached a rapid development in the field of detection and recognition. Image detection and classification for objects or groups of objects compose some useful, interesting and hard challenges for machine vision. In his work, Bisneto [4] states that a lot of progress has been reached during the last decade: the modeling projects that capture the image and its natural objects geometrical characteristics, the development of algorithms that rapidly correlate these models to the images, and the improvement in learning techniques that can estimate these models from a training and limited supervision.

Referring to learning techniques, He *et al.* [5] recently presented one of the most used nowadays: Mask R-CNN. This is a Convolutional Neural Network, region-based, capable of object detection, identification and segmentation, being the State of the Art in relation to the Computational Vision applied to machine learning, and is the architecture used in this article.

According to Bisneto [4], the classical problem of Computational Vision and Image Processing is determining if an image contains an object, characteristic or a pre-determined activity. It is an easily solved task for humans, but not yet satisfactory done by autonomous equipment, in which objects, situations, illumination and positioning are completely arbitrary. Machado [6] completes it stating that Computational Vision process main problem is related to quantifying visual information presented in images, in other words, for a given object

recognition it is necessary to find some image characteristics that distinguish it from other objects in its same universe.

The actual evolution of the economy and the global society made the cars selling grow rapidly. At the same time, the vehicle insurance market developed substantially, growing its customer chart. Aftermarket companies that supplies the insurance market, due to higher demands for quality, must provide even better services and products, meeting customer wishes. One of these services is the mobile inspection, in which the owner of the vehicle can photograph the parts of the car, including the damaged ones, for register and recording into the company files. However, many clients are not used to the name of the car parts and end up sending the wrong part pictures (the not-damaged ones, for fraud evaluation, for example). This generates a lot of rework in the process, because the policy holder has to resend the correct pictures or even go to the workshop for having an inspection done by a technician.

With the fast development of deep learning, the object detection method based in deep convolutional neural networks has been widely used. In the presented works until now, the problem of vehicle recognition stayed focused on the vehicle detection, vehicle type (passenger car, SUV's, trucks) and the plate identification, while the vehicle components recognition is still left aside. Qianqian *et al.* [7] presented the most recent work related to vehicle components recognition. In their work, three networks have their accuracy and precision checked for external vehicle components segmentation, using three different types of dataset (panoramic view, close view and an integration of both types) for turning them the most accurate possible, making it possible to be used in future applications.

Therefore, this present work aims to present a model proposal for vehicle external components segmentation, specifically windshields, rearview mirrors, bumpers, headlights and tail lights, for using as a systemic filter in wrong pictures sending made by clients from insurance companies. It can also be used as a filter for stock correction made by the employee in charge of the distribution center, that sometimes relate the picture to a wrong product or defect (in the cases of the Quality Control inspections). Also, the vehicle external components recognition is particularly important as a premise for the defect detection in those component to locate the damage in the Quality Control process in the entry of the product.

This article is divided into five chapters. The first one contains a brief summary with the contextualization and the proposed problem. In the second chapter the theoretical reference on which the research is developed is presented. Chapter three describes the methodology used in the research, describing its activities and parameters so it can be reproduced. In the fourth chapter the detailed work results are presented. The last chapter presents the conclusions about the work, with a result evaluation and proposal for future works.

2 Theoretical references

2.1 Convolutional Neural Networks

In a Convolutional Neural Network, each image pixel is converted in a characterized representation, by a series of mathematical operations. Images can be represented by an order 3 tensor, with height, width and color channels. According to Ferguson [8], the input sequentially passes by some processing steps, commonly referred as layers, realizing a random transformation, providing a feature map as output, as can be seen in Fig. 1. In most modern Convolutional Neural Networks the first convolutional layers extract features like edges and textures. The deeper convolutional layers can extract features that span a greater spatial area of the image, such as object shapes.

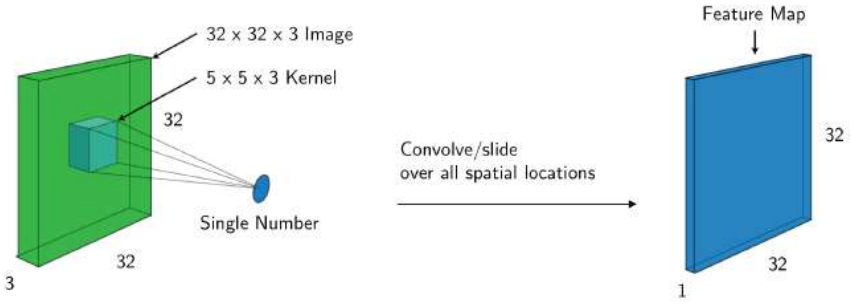

Figure 1. Image convolution with a kernel to produce a feature map.

Wu [9] mentions that by combining multiple layers it is possible to develop a complex nonlinear function, mapping data such as images, providing outputs as classification, identification and segmentation of objects

Deeper neural networks are, by design, parametrized nonlinear functions, as presented by Wu [9] in his work. An activation function is applied in the output of a layer to introduce this nonlinearity. He *et al.* [10] also presents the pooling layers. The primary function of theses layers is to reduce progressively the spatial size of the representation to reduce the number of parameters in the network, avoiding overfitting.

The training of a neural network is made by reducing a loss function, which Wu [9] defines as a measure between the neural network output and the ground truth. As long as each layer of the neural network is differentiable, it is possible to calculate the loss function gradient, in respect to dome parameters. Werbos [11] also defines that the backpropagation algorithm allows the numeric gradients to be calculated efficiently.

2.2 Mask R-CNN

Mask R-CNN architecture, as defined by He *et al.* [5], like other architectures, has two outputs for each potential object: a classification and the bounding box. Beyond that, it was added a third branch, bringing the object segmentation as an output. The additional output distinguish from the others due to the need of extracting a much finer spatial layout of an object.

The Mask R-CNN architecture adopts a two stage execution. The first one, called Region Proposal Network (RPN), proposes bounding boxes for the object. In the second stage, parallel to the class and the box prediction, Mask R-CNN also provides as an output a binary mask for each Region of Interest (RoI). This differs Mask R-CNN from other modern systems, which depends of mask predictions, like the ones designed by Pinheiro *et al.* [12], Dai *et al.* [13] and Li *et al.* [14].

Qianqian [7] explains that the Mask R-CNN architecture uses a RoI alignment layer to correct pixels from the images, then uses a network to classify the targets and execute regressions in the boxes of possible candidates. Parallel to the implementation problem solving of classification and regression, a prediction branch is added in Mask R-CNN, and each pixel in the RoI can be identified as belonging do the class of a given object.

Formally stated by He *et al.* [5], during training, a multitask loss function is defined in each Region of Interest, defined by:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

The classification loss, L_{cls}, and the box loss, L_{box}, are identical to those defined by Li *et al.* [14]. The mask branch has an output for each RoI, for each class. To this is applied a sigmoid function pixel by pixel, defining L_{mask} as the average binary cross-entropy loss.

According to He *et al.* [5], as can be seen in Fig. 2, the image detection system is composed by four modules. The first module is a feature extraction module that generates a representation of the input image with its high-level features. The second module is a Convolutional Neural Network that proposes Regions of Interest in the image, based on the features map. The third module is a Convolutional Neural Network that tries to classify the objects in each RoI. The fourth module realizes image segmentation, with the goal to generate a binary mask for

each region.

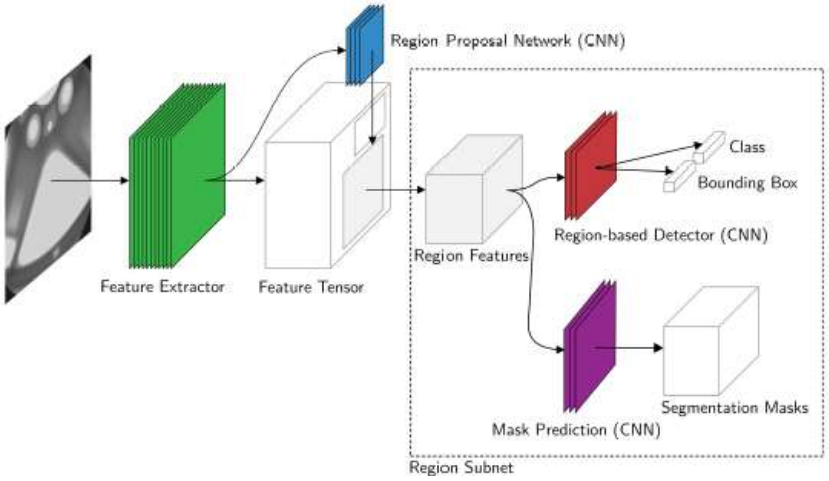

Figure 2. Neural Network architecture for object segmentation

2.3 Metrics

For measuring the results of the Mask R-CNN applied to the classification and segmentation of vehicle external components it is usual to apply the Mean Average Precision (mAP) as defined by Tan [15]. Beyond that, to solve the problem of mistaken pictures and use the model to act as a filter, Accuracy, Precision, Recall and F-Measure is used to evaluate the classifications made by the network.

3 Methodology

In this section, a classification, detection and segmentation system for automotive parts (headlights, tail lights, bumpers, rearview mirrors and windshields) is proposed, based in the Mask R-CNN architecture, using the Transfer Learning technique available in Github by Abdulla [16]. The implementation was realized using the 1.14 version of Tensorflow and the 2.2.4 version of Keras library.

There isn't, nowadays, any public repository with an image dataset for automotive products recognition. Therefore, it was necessary to build my own dataset. Thus, we gathered 945 files, consisting in 1320 images of car parts for training (244 headlights, 214 tail lights, 170 bumpers, 526 windshields and 166 rearview mirrors) and 325 images for validation (58 headlights, 51 tail lights, 55 bumpers, 111 windshields and 51 rearview mirrors). The dataset consists of images of the products fitted in the vehicles and out of the packing in the moment of the entry (Fig. 3). Images resolutions go from 640 x 480 to 5184 x 3456 pixels. The annotations for object recognition and segmentation were made with VGG Image Annotator 1.0.6, available by Dutta [17].

Figure 3. Examples of the dataset, mounted in the vehicle (left) and outside the packing (right).

There were too many attempts to find the optimum hyper parameters for this implementation until finding

the one with the best results (the last presented in Tab.1).

Table 1. Hyper parameters configuration atempts

Backbone	Epochs	Augmentation	Layers Trained	Image Max Dim	Image Min Dim	Loss Weights	Train ROIs Per Image	Learning Rate
ResNet101	50	Flip, Crop, Contrast, Normalization, Multiply, Scale, Rotate	Heads	512	256	(1, 1, 1, 1, 1)	100	0,0001
ResNet101	100	N/A	Heads	512	512	(1, 1, 1, 1, 1)	200	0,0001
ResNet50	50	N/A	3x	1024	800	(1, 1, 2, 3, 2)	200	0,0001
ResNet50	51	N/A	Heads	1024	800	(1, 1, 2, 3, 2)	200	0,001
ResNet50	60	N/A	Heads	1024	800	(1, 1, 2, 3, 2)	200	0,001

All other hyper parameters were used in the default mode of Mask R-CNN.

To compute the accuracy, precision, recall and F1 score, a Confusion Matrix was made using another dataset of images and these metrics were extracted.

4 Results

After the algorithm execution it was possible to evaluate the performance of Mask R-CNN for this dataset. Comparing the evolution of loss and val_loss during 60 epochs, as shown in Fig. 4, we can conclude that the network presented an overfitting close to epoch 25 (being the L_{box} the loss that contributed the most for this result).

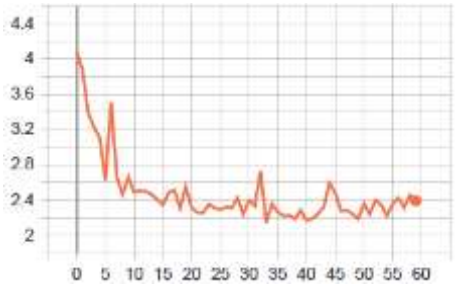

Figure 4. Val_loss over 60 epochs of training

Besides that, we present the Confusion Matrix of the classification in Fig. 5, from where the metrics were calculated and presented in Tab. 2.

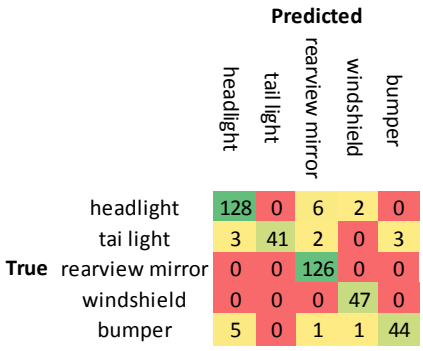

Figure 5. Confusion Matrix from classification using Mask R-CNN

Table 2. Results of the metrics calculated for the dataset

Product	Accuracy	Precision	Recall	F-Measure
headlight	0,9609	0,9412	0,9412	0,9412
tail light	0,9804	1,0000	0,8367	0,9111
rearview mirror	0,9780	0,9333	1,0000	0,9655
windshield	0,9927	0,9400	1,0000	0,9691
bumper	0,9756	0,9362	0,8627	0,8979

The result for mAP, considering an Intersection over Union of 0.5, is 0.4919. The visualization of the segmentation can be seen in Fig. 6 as follows.

Figure 6. Image segmentation for the automotive parts proposed in this work

5 Conclusions

The developed model, even having a small number of images in the dataset, showed appropriate results for recognizing the automotive products shown in this work (headlight, tail light, rearview mirror, windshield and bumper), while still there is a need for evolving with the bounding boxes and segmentation generation, that still does not fit perfectly to the objects (Fig. 6).

The val_loss of approximately 2.4, as can be seen in Fig. 4, showed that the model presented an overfitting behavior. This can be solved by increasing the number of images of each product in the dataset. Even reaching a

relatively high loss, the results has proven the model to be satisfactory, principally in the classification of the objects, reaching high values of the metrics exposed in Tab. 2. For the applications of acting like a filter for the pictures sent by the insurance companies' clients or uploaded by the stock corrections, the main metric to observe is recall, because it is really important to have a high number of true positives amongst the real quantity of that specific product. Analyzing the recall for each automotive product, the values go from 0,8367 to 1,0000, which are good results for the products classification. For raising the recall for tail lights (0,8367) and bumpers (0,8627) we must increase images of these products in the dataset.

Evaluating the segmentation, the object recognition worked in an appropriate way, attending the goal to work as a premise for learning about Mask R-CNN. The next step would be to apply this knowledge to develop a model to recognize defects in these products, like scratch, spot, breaking, among others.

For future works, we would suggest to enlarge the products portfolio, including other products sold by the company, like radiator, suspension, rear and lateral window, hood and bumper grill.

Acknowledgements. We would like to thanks Autoglass for the assignment of the products' images for building the training, validation and test datasets.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] R. C. Gonzales, R. E. Woods. *Digital Image Processing*. Tom Robins, 2002.
[2] H. Hogendoom. *The State of the Art in Visuall Object Recognition*. PhD thesis, GLA University, 2006.
[3] A. A. C. Carvalho. *Fundamentação teórica para Processamento Digital de Imagens*. Graduation conclusion work, Federal University of Lavras, 2003.
[4] C. R. B. Bisneto. *Reconhecimento de objetos utilizando Redes Neurais Artificiais e Geometria Fractal.* PhD thesis, SENAI CIMATEC, 2011.
[5] K. He, G. Gkioxari, P. Dollár, R. Girshick., "Mask R-CNN". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2018.
[6] D. S. Machado. *Reconhecimento de objetos de formas variadas*. PhD thesis, Federal University of Minas Gerais, 2008.
[7] Z. Qianqian, L. Sem, G. Weiming. "Research on vehicle appearance component recognition based on Mask R-CNN". *Journal of Physics: Conference Series*, vol. 1, n. 1335, 2019.
[8] M. Ferguson, R. Ak, Y. Lee, K. Law. "Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning". *Smart and Sustainable Manufacturing Systems*, vol. 2, n. 1, pp. 137–164, 2018.
[9] J. Wu. *Convolutional neural networks*. Class notes, Nanjing University, 2017.
[10] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (VCPR)*, pp. 770–778, 2016.
[11] P. K. Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of IEEE*, vol. 78, pp. 1550-1560.
[12] P. O. Pinheiro, R. Collobert, P. Dollár. "Learning to segment objects in candidates". In: *28th International Conference on Neural Information Processing Systems*, vol. 2, pp. 1990-1998, 2015.
[13] J. Dai, K. He, J. Sun. "Instance-aware semantic segmentation via multi-task network cascades". In: *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 980-988, 2016.
[14] Y. Li, H. Qi, J. Dai, Y. Wei. "Fully convolutional instance-aware semantic segmentation". In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 711-720, 2017.
[15] R. J. Tan. *Breaking down Mean Average Precision (mAP)*. Graduation conclusion work, NUS MTech, 2019.
[16] W. Abdulla. *Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow*. Github, 2017.
[17] A. Dutta, A. Gupta, A. Zisserman. *VGG Image Annotator (VIA)*. University of Oxford, 2018.