National College of Ireland

A drug identification model developed implementing instance segmentation

MSc Research Project
MSc in Science Cloud Computing

Alvaro Ricardo Corral Paramo
Student ID: 19190565

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Alvaro Ricardo Corral Paramo …….…… **Name:**

Student ID: 19190565……

Programme: MSc in Science Cloud Computing…… **Year:** 2021……

Module: MSc Research Project ……..……
Supervisor: Vikas Sahni……..……
Submission
Due Date: 16th August of 2021 ……..……

Project: A drug identification model developed using instance segmentation

Word Count: ……5327…… **Page Count**……20……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……  ……

Date: 16th August of 2021 ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□

You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A drug identification model developed using instance segmentation

Alvaro Ricardo Corral Paramo
19190565

Medicine boxes recognition is an important process in several sectors and industries, as pharmacies and the pharmacy industry or hospitals. A single failure in this procedure can have consequences that affect both on a human level and on a large-scale legal level. Due to these factors, it is necessary to address the problem of how to identify medicine packages based on deep learning techniques, in concrete, using the Convolutional Neural Network. However, several of the algorithms previously applied have failed to obtain accurate results when objects or logos were too small or the scene was not completely clear. The paper has been focused on the improvement of efficiency in object detection and recognition algorithms based on Mask detection. Also, it has been covered the need to find a dataset based on small objects and pharmacist products. The model is implemented with COCO dataset and a custom dataset of medicine packages. The evaluation carried out on a cloud platform, compares the algorithm Yolov4 and Mask-RCNN combined with the backbone ResNet50, as a result, Mask R-CNN needs less time for training and has higher performance. The effectiveness raised for the mask segmentation architecture was higher than 95%, in consequence, it is an acceptable score for the real world.

Keywords: Deep Learning, Convolutional Neural Networks, Object Detection

1 Content Table

2 Introduction

Deep Learning is one of the most promising technology still to discover the maximum power on industry, companies and society in general. Indeed, attempting to adjust the way that a human being obtains his erudition could be an impeccable bridge with the machine world. This technology became popular as a recognition methodology to spot patterns and features.

A high percentage of precision has been achieved by deep learning and one of the most powerful features is that it is scalable, verbalizing about computing and databases. Undoubtedly, it can be applied in different fields as language learning, speech analysis, or image recognition like in this case. Especially, in computer vision, it has superseded other approaches and techniques in detection and classification as it is underlined in Magalhães et al.[2].

Over the recent years, computing capacity, like enhancements in GPUs and the widespread use of cloud computing and big data, let deep learning perpetuate growing illimitable. Systems based on this technology trend to imitate the nervous system of the human brain for training.

With regards to Convolutional Neural Networks(CNN), the concept was invented in 1989 by Yann Lecun and others but it is nowadays when it has been possible to start applying it in practical environments [8]. Most importantly to explain this in terms of programming, there is an input, called tensor, which for example is an image of a medicine box and is going to go through several layers. CNNs work by dividing an image on several number matrixes, which, later, will be analysed and its values will be simplified too.

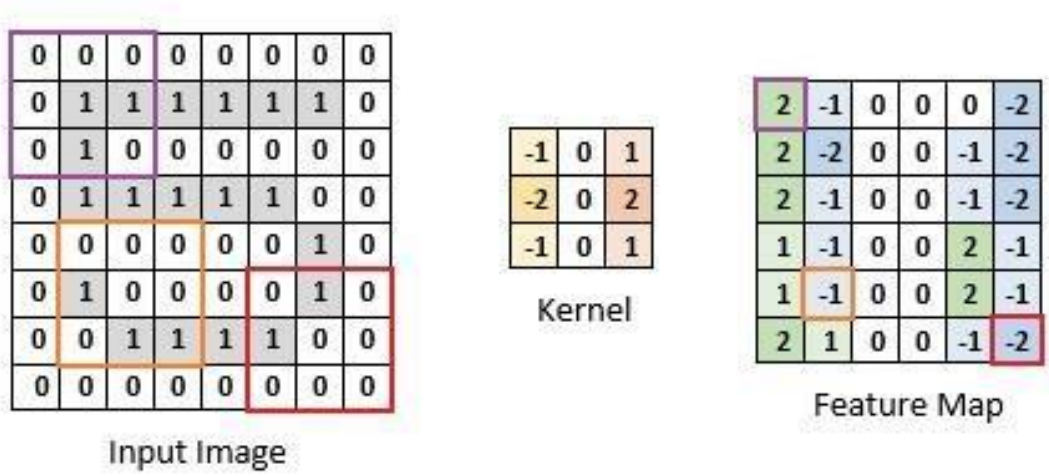


Figure 1: Image Matrix – Filter – Feature Map[4].

A *layer* is a conjunction of operations(*neurons*) or filters applied to the input to find a possible identification and extract features. They can be classified as convolutional, pooling and fully connected layers. The process of *activation*, or called also *normalization*, set all the

negatives values on the matrix to zero. Another important concept is *overfitting*. It occurs when the path followed to get and output is very similar and it using the same layers, so the accuracy level will be lower. To avoid the *overfitting* consequences is utilized the *dropout*, switching off some layers randomly to get different approaches.

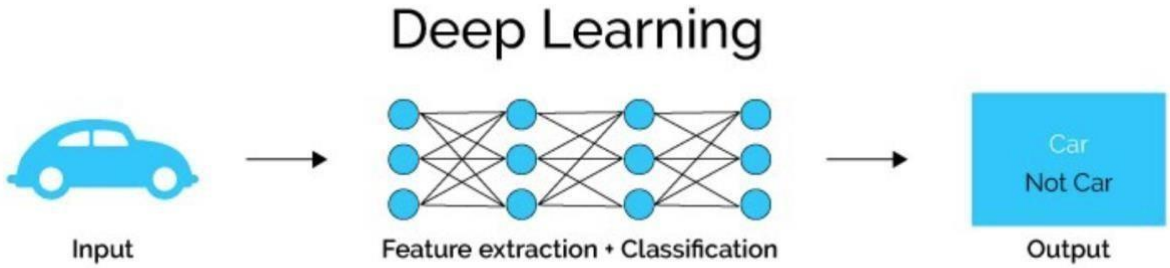


Figure 2: Deep Learning[5].

Applying the repetition of all these processes, it is possible learning new features automatically and without incipient training for the network, so it could be updated by itself. Training is another factor to consider because it can be delayed a ready network as soon as possible for testing, which is one of the objectives of this work and it can be expensive. Domain adaptation methods allow to classified another dataset taking advantage of a previous dataset from the training phase as described S. S. Sarwar et al. [21].

Nevertheless, the obstacle of this approach involves background noise, lighting variations, imaging blurs, and hardware requirements. CNN starts to check the features from the package of the drug and in which layers should include a process to compare if there are big differences with the pattern from the training. In that case, the neural network can investigate whether it is a product makeover or a counterfeit. Habitually, counterfeit pharma products may contain dangerous components or lack the appropriate active main ingredient[11]. Then, it could avail to eschew the distribution and early detection by professionals and end-users.

Inspired by all the benefits of deep learning could contribute, this work is focused on defining an efficient identification and classification approach for medicine boxes, from images as an input, applying algorithms more efficiently with specific training for updating the convolutional network.

The rest of the content is structured as follows: after a review of related work (section 3) with three sub-sections(3.1 CNN,3.2 Object Detection, and Recognition and 3.3 Dataset) where the background is exposed, the Methodology and Specification (section 4) describes how is working the object recognition algorithm and extended in three sub-section more Sections 4.1–4.3 talking about Equipment, Dataset, Experiments(section 5) goes through different metrics comparing Yolov4 and Mask R-CNN methods and extended in three sub-section more and finally 6)Conclusions.

3 Literature Review

3.1 Convolutional Neural Networks

The world of Deep Learning has never stopped to discover new methods and even more when Convolutional Neural Networks are considered. With sufficient data[3], they attain the ability to learn expeditiously rules, although the problem is if the complexity is excessive like in [12] experiments with the dataset ImageNet[33] and CIFAR-10[34]. Deep CNNs have reached a high target in computer vision and they have engaged the interest due one of the characteristics is learning new features, cognate to updating the medicine box image. Frequently, they have been composed by an input layers, multiple convolution layer ,down-sampling layer and output layer[5]. .In Three-CNN[20], where the model grows like a three reducing the training effort and in Sarwar[21] which exposes a DCNN with the ability to learn new tasks, through a "clone-branch", both introduce neural networks that acquire new knowledge from a new dataset, retaining the data acquired in previous trainings. In this case, it is not necessary that the network learn new tasks, just the second one. Tommasi et al. [15] displayed a model with different sources, and each source had different weights, where was possible to appreciate the first dataset had been the biggest score or weight and a source not contrasted had been the smallest one. ImageNet classification has been inspired by the definition of the classification function could not be determinate just only by the information from the training, the CNN should involve an scalable knowledge [13].

3.2 Object Detection and Recognition

First of all, there had been different methods for object detection for CNN, like RCNN[16] including CNN features that required to do repetitive processing which is enhanced on SPPNet[27]. Later, Fast R-CNN decreased the time for training and testing, when there has been remaining a dependency of Candidate Region Selection(CRS) algorithm[27]. That previous issue was eliminated on Faster RCNN[6]. With SSD(Single Shot MultiBox Detector) had been appeared the multi-scale feature map conceiving an enhancement on the speed nevertheless it had been not effective efficient with small objects as it explained in Krüger, J. *et al.* [16]. In this case, the network requested to find diminutive features as logos, or images on the box to discern between an original package and a counterfeit. Additionally, there are several versions for YOLO(You Only Look Once), especially for real-time detection, notwithstanding , like in [14][16] was demonstrated, Faster- R-CNN had more accuracy and better performance with small objects and, in [8] was revealed was not useful if objects were held in the hand. Conclusively, there has been another characteristic that required to be analysed, the text detection, in [1] has been investigating for the best framework to apperceive text and shapes together. Mask R-CNN[28] was accentuated due to its precision and the blind spot are the high-performance detection of arbitrary-shape text, therefore with the typography of the packages it should not affect the result and the detection speed[36]. Also, it had solved the problem of instance-segmentation with an extra-branch to predict object masks in parallel[25].

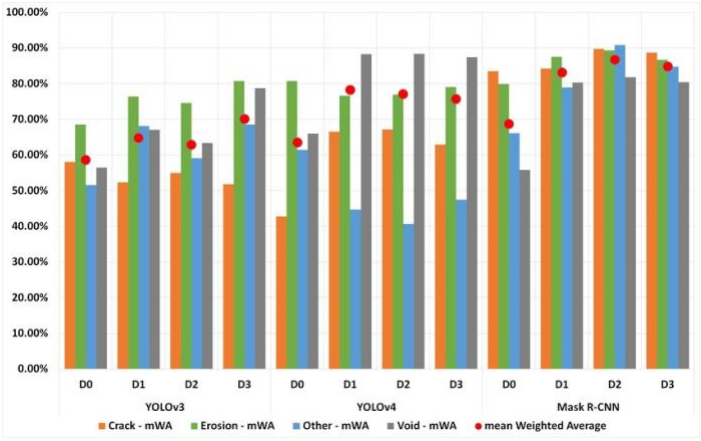


Figure 3 - Zhang, Jet Al. [36] Mean Weighted Average Performance Evaluation

Besides, Magalhães et al.[2] included a Barcode and Optical Channel Recognition(OCR) without using Neural Networks, the percentage of accuracy was almost 100% of success, however, this solution was only suitable with photos in a uniform background [10], under those circumstances, this option has been rejected.

3.3 Dataset

In this section, the most optimum dataset for a deep learning method to recognize medicine packages has been sought in past papers. The information and the set of images should be optimized to improve the algorithm and obtaining maximum precision[4].

Firstly, there are datasets just for training[1] our network with different kinds of objects, animals, humans, etc and in this way it can identify faster where is the box, because one of the goals is to obtain positive results in an irregular background, in the worst case. ImageNet belongs to this first category that has been exposed and is one of the most extensive, with around 15 million labelled HD images, in 22.000 categories. MJSynth, for scene recognition, 8.9 million has been synthesized images, using Google fonts, and the images from ICDAR03[35]. All the labels were just in English, assuming precision diminishes with drugs in different languages. SynthText was designed thinking on scene text detection, nonetheless, a combination with MJSyth has been good results[13]. CIFAR-10, a collection of natural images classified into 10 classes, although the images were not in HD[18][16]. Chen et al. [24] exposed a combination of SUN and Microsoft COCO dataset for small objects, which could be interesting for our research due to identify logos or small images on drugs packages. MNIST has presented an accuracy close of perfection, 98.08% with Deep Neural Networks[11]. 70,000 images have been gathered, including examples to endeavour to cheat the network[17]. Secondly, datasets with a concrete feature purpose, in this case, medicine boxes. PharmaPack[11], filled the gap of a dataset with images from 1.000 drugs from different angles, and it is possible to access the files for free. All the samples are sharp without any noise or background which affects a quick match, hence, they are impeccable as a basic pattern for training.

Secondly, datasets with a concrete feature purpose, in this case, medicine boxes. PharmaPack[11], filled the gap of a dataset with images from 1.000 drugs from different angles, being this compilation, public and free of charges. Part of this dataset were sharp samples without any noise or background which could affect a quick match, and the other part included a hand holding them equivalent to the real world.

After reviewing the literature review, we have found the best algorithm which has not been applied for medicine boxes recognition was Mask R-CNN following the work from Fu, C.-Y et al. [32] where compares the accuracy of Yolov3 and several backbones with Mask RCNN. It could obtain the precision level required for this research. Furthermore, having a combination of several datasets for training with PharmaPack could achieve the goal of identifying medicine boxes in a shorter time in no ideal situations.

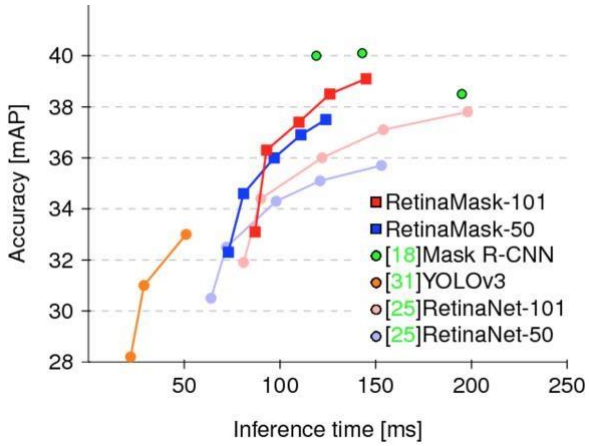


Figure 5 - Fu, C.-Y et al. [32] Comparations between several algorithms

3.4 Framework

Mask R-CNN uses the PyTorch packages for simplicity [29], as PyTorch was chosen in numerous new studies to get the best performance with this algorithm[30].

4 Research Methodology and Specification

In this paper, PyTorch has been adopted as a deep learning framework for development. In terms of CPU usage, it is not very popular, but talking about GPU, it achieves good results on performance[40]. GPUs demand more resources than CPUs when deep learning is the goal of an operation. They work faster in a simple parallel process and the price is lower[6].

4.1 Algorithms

4.1.1 Mask R-CNN

The object recognition algorithm chosen as the favourite, as seen in the previous section of the literature review, is Mask R-CNN. A description of the architecture and its components can be found below

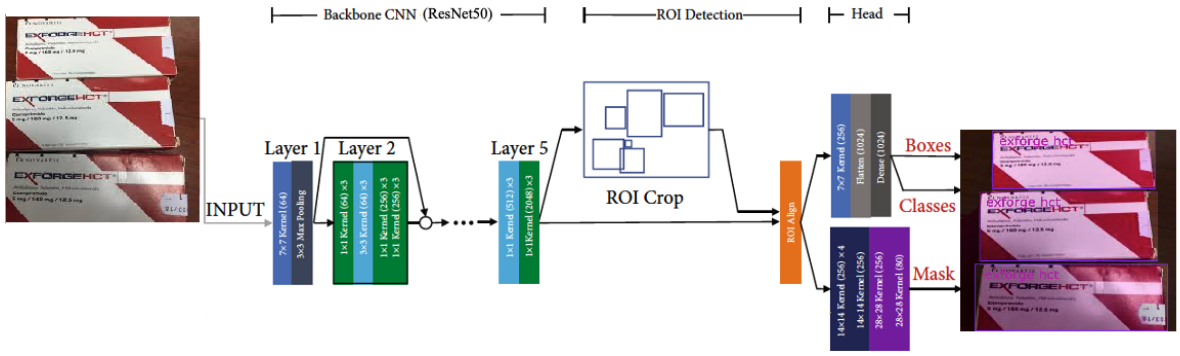


Figure 6 - Mask R-CNN Architecture. Adapted [36]

- *Backbone*: a CNN that obtains features from the inputs, converting them into a feature matrix. Afterwards, the data will be sent to the next layer. The backbone selected for our experiments is ResNet50 pre-trained using COCO dataset.

- *Region Proposal Network(RPN):* consists of a fully connected convolutional network to predict the main objects on images. For example, a person and a medicine package of Fig.4. It is trained on the frames of the general areas to detect the scope and the boundary of an object. Also, it is compounded by two extra layers named Reg and CLs. The first one determines the coordinates of the proposal. and the second one is where the decisions about background foreground are taken. It is integrated inside the backbone and the last step of its output[6].

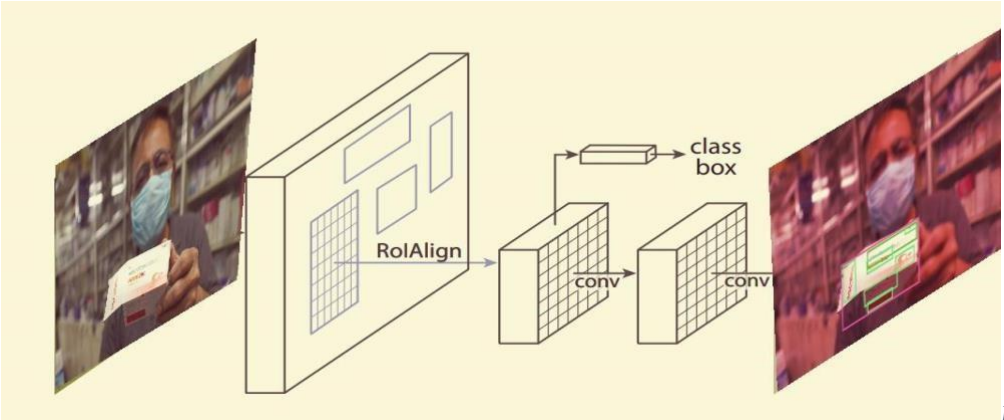
- *ROI Pooling*: involves the main functions to align an object candidate in a picture. to the patch related in the feature map. and transfer these patches to the next level, to the fully connected layer into a data of uniform size[28].

- *Object Classifying component*: a class for each object detected on the RPN is defined in this element [27].

- *Bounding box component*: it draws a coloured box around the identified objects on the picture[28].

Figure 7: Mask-R-CNN [26].

The candidate areas, called anchor boxes, are evaluated. They are distributed in the whole picture. These anchor boxes are of different types of sizes and ratios, and it depends on the targeted object. The anchor boxes will have scores designated by the Region Proposal Network. It indicates whether the candidate is in the front or not. A high classification score may indicate that an anchor box includes a portion of an item. Furthermore, another function of RPN is to redefine the anchor boxes. Using the process named bounding box refinement, RPN bounds better the object [29].



Figure

Figure 8: RoIAlign preserves exact locations. adapted[26].

Finally, after the network processes the selected foreground areas, with the Object Classifying component, it generates masks for each of them.

4.1.2 Yolov4

In the experiments, the comparison with the algorithm selected was using Yolov4. As in the literature review was shown, there are several versions of this architecture. The version 4 included improvements about the accuracy and real time detection[31].

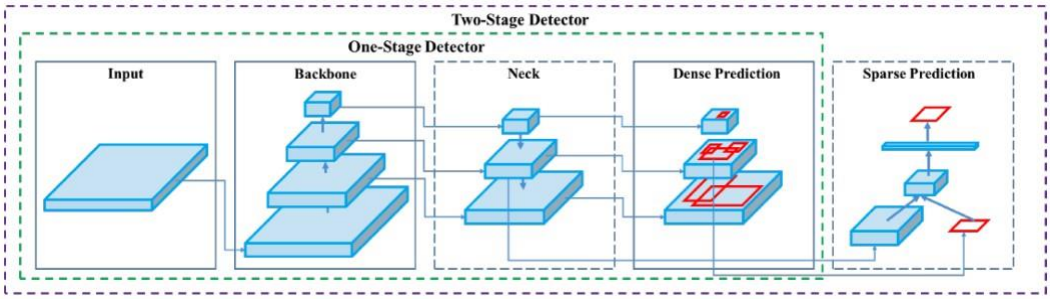


Figure 9 - Yolov4 Architecture [31]

There are two types of object detection models: one stage models and two stage models. A one-stage model may be used to identify objects without the need for a pre-processing phase. A two-stage detector, on the other hand, uses a preliminary phase in which regions of importance are defined and then classified to determine whether or not an object has been detected in these areas. A one-stage detector has the advantage of being able to provide predictions quickly, allowing for real-time use.

YOLOv4 divides the object identification task into two parts: regression and classification. Regression determines the item's location using bounding boxes, while classification determines the object's class. This approach is similar to the one used for YOLOv3.The backbone network for an object detector is pre-trained on COCO categorization in terms of architecture. The weights of the network have already been modified to identify significant characteristics in a picture, however, they will be altered for the additional purpose of object recognition. Second, the neck is compounded by layer flow up and down, with only a few layers at the convolutional network's end connecting. Finally, for detection, YOLOv4 uses the same YOLO head as YOLOv3, with three levels of detection granularity and anchor-based detection stages

4.2 Dataset

The data source for training is compounded by several sources. First, COCO Microsoft has been developed to gather images from complex familiar scenes containing common objects in their natural context[37]. Using COCO, the algorithm avoids interference when the picture´s background contains other objects that are not medicines. On the other hand, the designated dataset, which contains the information regarding medicine packages, is a custom dataset generated from the PharmaPack[11] dataset. This dataset has been converted to a newer format because Matlab is not compatible with our algorithm. In addition, it has been extended, using image and bounding box augmentation methods as is shown in fig.2[39]. CNN has learned how to identify the features in a way more exhaustive with this training set.

Figure 10 - Roboflow augmentation methods[39]

Finally, a set of images from counterfeits were added to the custom dataset. In case, some of the features are detected by the CNN, the output will be identified as a fake package.

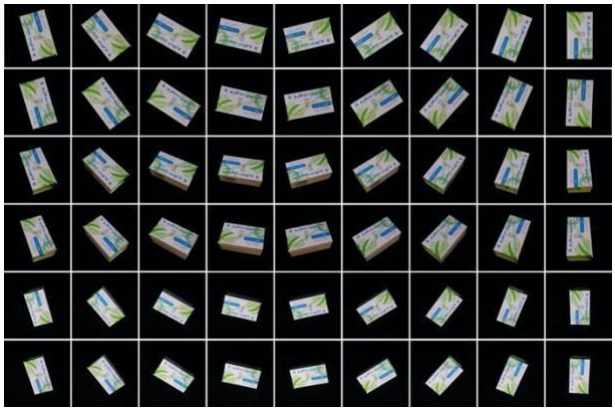


Figure 11: PharmaPack Dataset.

4.3 Training

Training is one of the most important areas to enhance in Neural Networks[23]. In Pytorch, the drug information and the image are going to become a tensor(input). Thereupon, it loads the configuration for the training process. Second, CNN sends the tensors to be analysed by the detection algorithm.

To illustrate, training specifications have been exposed below:

1. COCO [11][17] dataset is indicated to train the CNN for general image recognition and small object detection. It can be concluded that, in this step, the network has achieved the goal of notice if there is a hand in the photo. Mask R-CNN Resnet50 integrates the functionality to pretrained the model by COCO dataset. For instance, it is not required to obtain a huge amount of data with a set of images [28].

2. Second, a custom dataset can be pre-trained accessing the Roboflow services. As a result, transferring the weights reduces the training time by several hours. In fact, the choice to obtain the dataset is available on their platform through a simple link[39]. Also, combining

this dataset with COCO, domain adaptation methods are applied, retaining the information without losing the first training[21].

3. Typically, the directory structure will be divided into three folders: *train*, *validation*, and *test*. *Train* is the main folder, containing clear samples for each class of medicine. *Validation folder* compiles images that are already included in the *train* directory. In the end, pictures from the real world are included inside the *test* folder.

5 Experiments

The experiments have been consisted of comparing the training time and the precision of the two methods. The first method implementing Yolov4 and the second algorithm applying the Mask R-CNN combined with the ResNet50 backbone.

5.1 Equipment

In the implementation for this paper, the free of charge services from the platform Google Colab[38] were consumed, which is a convenient tool based on the cloud, for accessing and executing the algorithms instantly.

Item	Specifications
CPU	2 Processors Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Nvidia Tesla T4
RAM	13GB RAM
HDD	30 GB

Table 1 - Google Colab Free tier specs[38]

5.2 Metrics

Accessing the equipment described above, the training phase with the two algorithms has been applied using the custom dataset. A total of 9,000 images with 256 classified products. The learning time was approximately 26 hours when Yolov4 was executed and almost 25 hours for the Mask R-CNN algorithm. The Mask RCNN runs at 67.47 ms per image and Yolov4 obtains the mark of 72.69 ms.

The images, to carry out the trial test, have been divided into four types:

- *Type 1*: The drug is the only object in the image.
- *Type 2*: The medicine is held by one hand.
- *Type 3*: The drug appears with more objects in the scene, such as an advertisement.
- *Type 4*: In the scene appears several medicines.

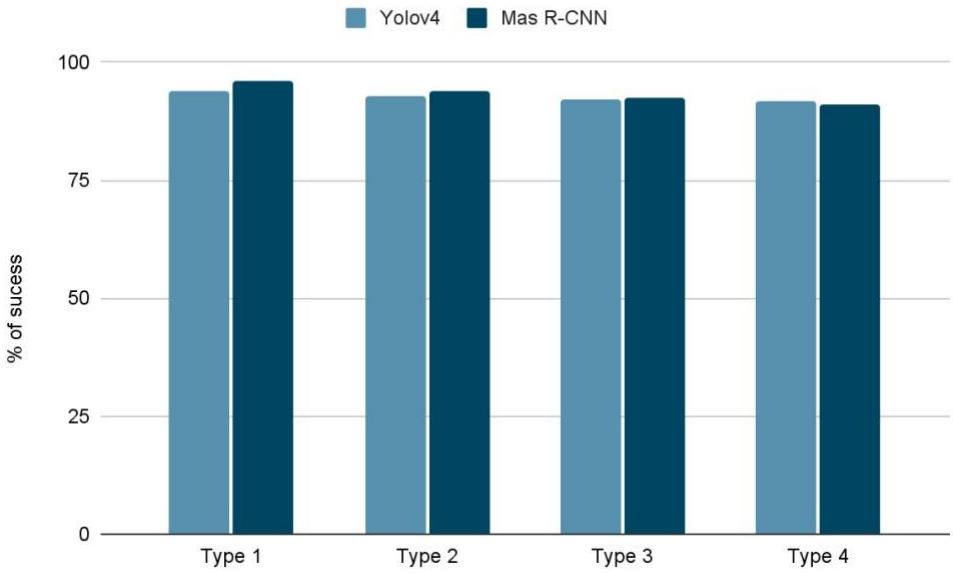


Figure 12: Percentage of success by type of pictures.

In the graph, the percentage of success identifying objects in the Test training set is shown. Mask R-CNN has been exceeded in almost all types of images. The exception was the Type 4 where several medicines are displayed at the same time. It seems if there are more than three classes involved in the same image, it affects the operation of the classification of instance segmentation.

Furthermore, the measures that were utilised to compare the performance of the two methods will be described:

- Precision: In classification tasks, it assesses the model's quality. In classification tasks, it assesses the model's quality.
- Recall: describes the amount that the model is capable of identifying. It is calculated by dividing the number of True Positives by the number of True Positives and False Negatives.
- F1 Score: used to aggregate precision and recall metrics into a single number to make comparing two approaches simpler

$$\text{F1 score} = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Algorithm type	Yolov4	Mask R-CNN
Epochs	50	50
Precision	95.93%	97.11%
Recall	96.20%	97.24%
F1 Score	95.41%	96.63%

Table 2 - Precision, Recall and F1 scores

The precision and recall percentage of the Mask R-CNN(97.11% and 97.24%) were higher than those of the Yolov4 algorithm (95.93 and 96.20%). Furthermore, the F1 Score is 1.22% higher with the second algorithm, meaning Mask R-CNN Resnet50 is preferred.

- Loss value: The loss value is composed of the classification loss, bounding-box loss from the RPN structure, classification loss, bounding-box loss,
  and mask loss from the backend of the model[43].

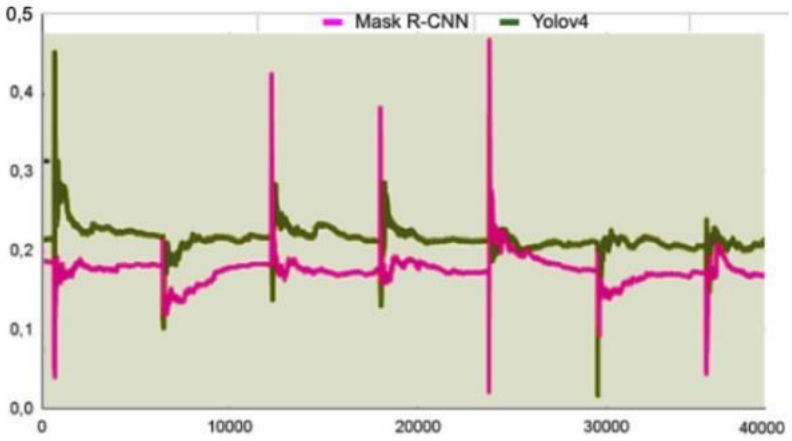


Figure 13: Loss curve comparing Yolov4 and Mask R-CNN algorithms where Y-axis is the total loss and X-axis the number of iterations.

It can be appreciated from the graph, Yolov4 starts to have a higher loss rate at the beginning of training with fewer iterations. However, after 5,000 iterations it acquires a severe decrease, where a turning point can be observed. Mask R-CNN remains below the indexes of the other algorithm until from 12,000 iterations it tends to a large increase. In the middle part of the graph, it remains above the Yolov4 results. In the last phase, there is a tendency to descend and stay below the other sample to stabilize. Therefore, it can be deduced that Mask R-CNN makes fewer losses than Yolov4.

Besides, the peak memory usage of the training process has been measured in relation to the number of iterations. This shows that memory is another point in favour of the Mask R-CNN algorithm.

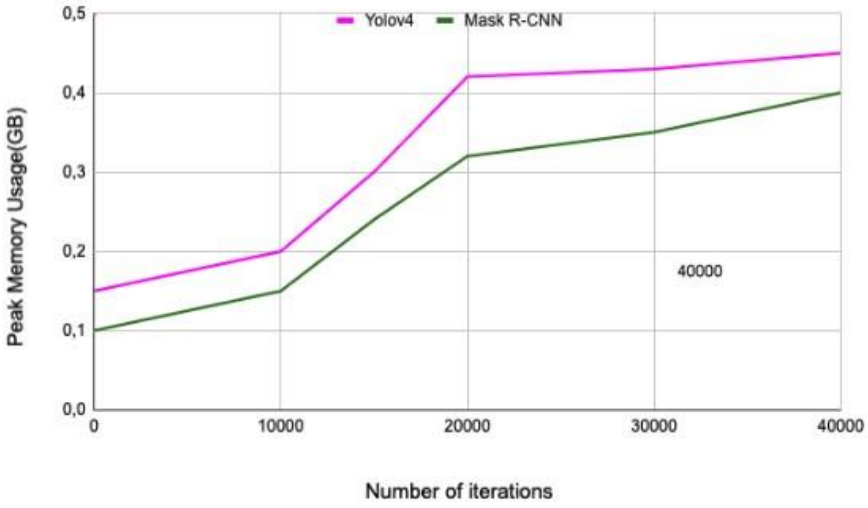


Figure 14: Memory usage comparing Yolov4 and Mask R-CNN algorithms.

These metrics are widely used and authoritative indicators to judge a deep network model's performance in object detection and instance segmentation[43].

Below you can see an extract of the percentages of success of various products and the total. Mask R-CNN continues to get the highest percentages. Where the percentages appear lower, it is because the characteristics of the product were more difficult to recognize.

Drug	Yolov4	Mask R-CNN
Vi-de3	91.4%	94.3%
pariet	92.3%	95.3%
tadenan	94.3%	96.1%
tramundin	92.9%	93.2%
viburcol	95.6%	96.1%
voltadvance	92.2%	93.2%
exforge hct	94.8%	95.4%
Total	**93.7%**	**95.5%**

Table 3 - Success rate regarding sample from Test training set.

Mask R-CNN ResNet50 showed a better object detection performance over Yolov4. The Mask R-CNN had room to fill up the shortcomings of Yolov4 implementing instance segmentation improving the object detection performance. Mask R-CNN had a stronger feature extraction capability for drug packages.

In this study, the algorithm of Mask R-CNN is available to address the identification of the drug name in 95% of the cases. The cases for which it fails are tightly related to uneven colour backgrounds on which the text is written.

6 Conclusions

The goal of this paper was to solve the challenging problem of identifying a successful method for the recognition of medicine packages. With an accuracy greater than 95%, it surpasses previous algorithms applied in this field. The findings of this study have the potential to be implemented in the real world, assisting hospitals in categorising medications and preventing human errors. This method is able to achieve success whether the product appears in an image completely or partially and in a regular or irregular background. Future work aims to improve the algorithm when, in the testing scenario, there are several medicines in the scene and try to obtain a higher percentage of accuracy. In addition, it will be possible to modify the code to apply TPU(Tensor Processing Units) technology. TPUs improve the speed of training and focus the resources on specific loads of work[44].

References

[1] Raisi, Z., Naiel, M., Fieguth, P., Wardell, S. and Zelek, J., 2020. *Text Detection And Recognition In The Wild: A Review*. [online] arXiv.org. Available at: <https://arxiv.org/abs/2006.04305> .

[2] L. Magalhães, B. Ribeiro, N. Alves and M. Guevara, "A three-staged approach to medicine box recognition," 2017 24º Encontro Português de Computação Gráfica e Interação (EPCGI), Guimaraes, 2017, pp. 1-7, doi: 10.1109/EPCGI.2017.8124317.

[3] Dong T., Qi X., Zhang Q., Li, W. and Xiong, L. (2019) Overview on Vision-Based 3D Object Recognition Methods. In: Zhao Y., Barnes N., Chen B., Westermann R., Kong X., Lin C. (eds) Image and Graphics. ICIG 2019. Lecture Notes in Computer Science, vol 11902. Springer, Cham

[4] Benjamim, X. C. *et al.* (2012) 'Visual identification of medicine boxes using features matching', *2012 IEEE International Conference on Virtual Environments HumanComputer Interfaces and Measurement Systems (VECIMS) Proceedings, Virtual Environments HumanComputer Interfaces and Measurement Systems (VECIMS), 2012 IEEE International Conference on*, pp. 43–47. doi: 10.1109/VECIMS.2012.6273190.

[5] Wang, L., Wen, M. and Ying Wu (2019) 'Object Detection Based on Deep Convolutional Neural Network', *Acta Microscopica*, 28(2), pp. 331–338.

[6] W. Chang, L. Chen, C. Hsu, C. Lin and T. Yang, "A Deep Learning-Based Intelligent Medicine Recognition System for Chronic Patients," in *IEEE Access*, vol. 7, pp.

4444144458, 2019, doi: 10.1109/ACCESS.2019.2908843.

[7] Krüger, J. *et al.* (2019) 'Deep learning for part identification based on inherent features', *CIRP Annals - Manufacturing Technology*, 68(1), pp. 9–12. doi: 10.1016/j.cirp.2019.04.095.

[8] Ting, H. *et al.* (2020) 'A drug identification model developed using deep learning technologies: experience of a medical center in Taiwan', *BMC Health Services Research*, 20(1), pp. 1–9. doi: 10.1186/s12913-020-05166-w.

[9] Wang, J. S. , Ambikapathi, A., Han, Y., Chung, S., Ting, H. and Chen, C. "Highlighted Deep Learning based Identification of Pharmaceutical Blister Packages," *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, Turin, 2018, pp. 638-645, doi: 10.1109/ETFA.2018.8502488.

[10] Liu, X. *et al.* (2020) 'DLI-IT: a deep learning approach to drug label identification through image and text embedding', *BMC Medical Informatics & Decision Making*, 20(1), pp. 1–9. doi: 10.1186/s12911-020-1078-3.

[11] Taran, O. et al. (2017) 'PharmaPack: Mobile fine-grained recognition of pharma packages', 2017 25th European Signal Processing Conference (EUSIPCO), Signal Processing Conference (EUSIPCO), 2017 25th European, pp. 1917–1921. doi: 10.23919/EUSIPCO.2017.8081543.

[12] He, K. *et al.* (2016) 'Deep Residual Learning for Image Recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[13] Krizhevsky , Ilya Sutskever and Geoffrey E. Hinton (2017) 'ImageNet classification with deep convolutional neural networks', *Communications of the ACM*, 60(6), pp. 84–90. doi: 10.1145/3065386.

[14] Xia Zhao, Haihang Jia and Yingting Ni (2018) 'A novel three-dimensional object detection with the modified You Only Look Once method', *International Journal of Advanced Robotic Systems*, 15. doi: 10.1177/1729881418765507.

[15] Patricia, N. and Caputo, B. (2014) 'Learning to Learn, from Transfer Learning to Domain Adaptation: A Unifying Perspective', *2014 IEEE Conference on Computer Vision and Pattern Recognition, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 1442–1449.
doi: 10.1109/CVPR.2014.187.

[16] Nguyen, N.-D. *et al.* (2020) 'An Evaluation of Deep Learning Methods for Small Object Detection', *Journal of Electrical & Computer Engineering*, pp. 1–18. doi: 10.1155/2020/3189691.

[17] Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W., 2020. *Textboxes: A Fast Text Detector With A Single Deep Neural Network*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1611.06779>

[18] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316.

[19] Seif, G., 2020. *Deep Learning For Image Recognition: Why It'S Challenging, Where We've Been, And What'S Next*. [online] Medium. Available at: https://towardsdatascience.com/deep-learning-for-image-classification-why-itschallengingwhere-we-ve-been-and-what-s-next-93b56948fcef>

[20] Roy, D., Panda, P., & Roy, K. (2020). Tree-CNN: A hierarchical Deep Convolutional Neural Network for incremental learning. Neural Networks, 121, 148–160. doi:10.1016/j.neunet.2019.09.010

[21] S. S. Sarwar, A. Ankit and K. Roy, "Incremental Learning in Deep Convolutional Neural Networks Using Partial Network Sharing," in *IEEE Access*, vol. 8, pp. 46154628, 2020, doi: 10.1109/ACCESS.2019.2963056.

[22] Shi, Y., Zhang, Y. and Harik, R. (2020) 'Manufacturing feature recognition with a 2D convolutional neural network', *CIRP Journal of Manufacturing Science and Technology*. doi: 10.1016/j.cirpj.2020.04.001.

[23] Zhao, Y., Barnes, N., Chen, B. and Westermann, R., 2020. *IMAGE AND GRAPHICS*. [Place of publication not identified]: SPRINGER NATURE, p.245.

[24] Chenyi Chen et al. (2017) 'R-CNN for Small Object Detection', Computer Vision – ACCV 2016 : 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V. (Lecture Notes in Computer Science. 10115), p. 214. doi: 10.1007/978-3-319-54193-8_14.

[25] Hwejin Jung, Bilal Lodhi and Jaewoo Kang (2019) 'An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images', *BMC Biomedical Engineering*, 1(1), pp. 1–12. doi: 10.1186/s42490-019-0026-8.

[26] He, K. *et al.* (2020) 'Mask R-CNN', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(2), p. 386. Available at: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edb &AN =141230570&site=eds-live&scope=site.

[27] Girshick, R. (2015) 'Fast R-CNN', *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448. Available at:

https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edb&A N=114288187&site=eds-live&scope=site

[28] Shivajirao, S. *et al.* (2019) 'Mask R-CNN End-to-End Text Detection and Recognition', *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Machine Learning And Applications (ICMLA), 2019 18th IEEE International Conference On*, pp. 1787–1793. doi: 10.1109/ICMLA.2019.00289.

[29] Suh, S, Park, Y, Ko, K, Yang, S, Ahn, J, Shin, J-K & Kim, S 2021, 'Weighted Mask RCNN for Improving Adjacent Boundary Segmentation', Journal of Sensors, pp. 1–8, <https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=ed b&AN=148281046&site=eds-live&scope=site>.

[30] Wu, M. et al. (2020) 'Object detection based on RGC mask R-CNN', IET Image Processing (Wiley-Blackwell), 14(8), pp. 1502–1508. Available at: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edb &AN=148084336&site=eds-live&scope=sit.

[31] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs, eess].

[32] Fu, C.-Y., Shvets, M., Berg, A.C., 2019. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free.

[33] ImageNet [WWW Document], n.d. URL https://www.image-net.org/ .

[34] CIFAR-10 and CIFAR-100 datasets [WWW Document], n.d. URL https://www.cs.toronto.edu/~kriz/cifar.html.

[35] ICDAR 2003 Robust Reading Competitions - TC11 (no date). Available at: http://www.iaprtc11.org/mediawiki/index.php/ICDAR_2003_Robust_Reading_Competitions

[36] Zhang, J., Cosma, G. and Watkins, J. (2021) 'Image Enhanced Mask R-CNN: A Deep Learning Pipeline with New Evaluation Measures for Wind Turbine Blade Defect Detection and Classification', *Journal of Imaging*, 7, p. 46.

[37] Lin, T.-Y. *et al.* (2014) 'Microsoft COCO: Common Objects in Context', in Fleet, D. et al. (eds) *Computer Vision – ECCV 2014*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 740–755

[38] Google Colaboratory Available at: https://colab.sandbox.google.com › notebooks

[39] Roboflow Available at: https://roboflow.com/

[40] Pointer, I. (2020) ‘5 reasons to choose PyTorchfor deep learning’, *InfoWorld.com*, 24 February. Available at: https://www.infoworld.com/article/3528780/5-reasons-to-choosepytorch-for-deep-learning.html

[41] Brownlee, J. (2014) ‘Classification Accuracy is Not Enough: More Performance Measures You Can Use’, *Machine Learning Mastery*, 20 March. Available at: https://machinelearningmastery.com/classification-accuracy-is-not-enough-moreperformance-measures-you-can-use/

[42] Yin, K. (2019) *Overcome Overfitting During Instance Segmentation with Mask-RCNN, Medium*. Available at: https://towardsdatascience.com/overcome-overfitting-duringinstance-segmentation-with-mask-rcnn-32db91f400bc.

[43] Wu, Q.; Feng, D.; Cao, C.; Zeng, X.; Feng, Z.; Wu, J.; Huang, Z. (2021) Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images. Available at: https://doi.org/10.3390/s21082618

[44] You, Y., Zhang, Z., Hsieh, C.-J., Demmel, J., & Keutzer, K. (2019). Fast Deep Neural Network Training on Distributed Systems and Cloud TPUs. IEEE Transactions on Parallel and Distributed Systems, 1–1. doi:10.1109/tpds.2019.2913833

[45] Dorrer, M. G. and Tolmacheva, A. E. (2020) ‘Comparison of the YOLOv3 and Mask R-CNN architectures’ efficiency in the smart refrigerator’s computer vision’, *Journal of Physics: Conference Series*, 1679, p. 042022. doi: 10.1088/1742-6596/1679/4/042022.

[46] RUGERY, P. (2020) Explaining YoloV4 a one stage detector, Medium. Available at: https://becominghuman.ai/explaining-yolov4-a-one-stage-detector-cdac0826cbd7