

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**INFORMATION EXTRACTION FROM
SALES FLYERS USING SEMI-SUPERVISED
LEARNING**

HARLINTON PALACIOS MOSQUERA
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR
ASSOC. PROF. DR. YAKUP GENÇ

GEBZE
2019

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YARI DENETİMLİ ÖĞRENME
KULLANILARAK SATIŞ
BROŞÜRLERİNDEN BİLGİ ÇIKARIMI

HARLINTON PALACIOS MOSQUERA
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
DR. ÖĞR. ÜYESİ YAKUP GENÇ

GEBZE
2019

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 03/07/2019 tarih ve 2019/30 sayılı kararıyla oluşturulan jüri tarafından 24 / 09/2019 tarihinde tez savunma sınavı yapılan Harlinton Palacios Mosquera'nın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Dr. Öğr. Üyesi Yakup GENÇ

ÜYE

: Prof. Dr. Yusuf Sinan AKGÜL

ÜYE

: Dr. Öğr. Üyesi Ayşe Betül OKTAY

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve/..... sayılı kararı.

SUMMARY

The detection of objects and texts as well as the extraction of them from images has been an enormous challenge that we are still trying to perfect for the past three decades. The current progress in deep learning and machine learning techniques in image classification and object localization particularly for images, which has a lot of information as sales flyers, has been a topic of discussion in the last half-decade and we have seen some brilliant advancements in recent times which try to solve at least one of these sub-tasks text and object localization and recognition from images and digital documents as sale flyers.

In this thesis we presented an automatic learning algorithm for information extraction from sales flyers, which consists of 3 steps: the first is, the detection of objects (Products) in the sales flyers inspired by the faster region convolutional neural network (Faster RCNN) framework and Residual Neural Network as feature map. The second step is, the detection and extraction of the text in the sale flyers by optical character recognition and finally the merge algorithm to merge each product with its information such as price and description. We presented three different merge algorithms and compared them: The function algorithm (1) (merge by distances between object and text) achieved a precision 74%, Recall 64%, F1 Score 69% and Accuracy 67%. The function algorithm (2) (merge by distances between Object, text and text on the object) achieved a precision 76%, Recall 70%, F1 Score 73% and Accuracy 71%. The main contribution of this thesis been the combination of Convolutional neural network (CNN) and Multilayer Perceptron (MLP) achieved a precision 85%, Recall 80%, F1 Score 82% and Accuracy 81%.

Keywords: Deep Learning, Faster RCNN, Machine Learning, Sales Flyers, Object Detection, Text Detection.

ÖZET

Nesnelerin ve metinlerin algılanması ve bunların imgelerden çıkarılması, son 30 yılda hala mükemmelleştirmeye çalıştığımız büyük bir zorluk oldu. Görüntü öğreniminde Derin Öğrenme ve Makine Öğrenimi tekniklerindeki mevcut ilerleme ve özellikle satış broşürleri gibi birçok bilgiye sahip olan görüntü için amaç lokalizasyonu, amaç son yarım on yılda bir tartışma konusu olmuştur ve son zamanlarda çok parlak gelişmeler görmüştük. bu alt-görevlerden en az birini çözmeye çalışmakta ve yerleştirmeyi ve resimlerden ve dijital belgelerden satış broşürleri olarak tanınmasını hedeflemektedir.

Bu tezde, satış broşürlerinden bilgi çıkarımı için 3 adımdan oluşan otomatik bir öğrenme algoritması sunduk: birincisi, daha hızlı bölge evrimsel sinir ağı (Faster RCNN) çerçevesinden ilham alan satış broşürlerinde nesnelerin (Ürünler) tespiti ve Özellik haritası olarak Artık Sinir Ağı. İkinci adım, satış broşürlerindeki metnin Optik karakter tanıma yöntemiyle algılanması ve çıkarılması ve son olarak her bir ürünün fiyat ve açıklama gibi bilgilerle birleştirilmesi amacıyla birleştirme algoritmasıdır. Üç farklı birleştirme algoritması sunduk ve bunları aralarında karşılaştırdık: İşlev algoritması (1) (nesne ve metin arasındaki mesafelerle birleştirme)% 74,% 64 Geri Çağırma, F1 Puanı% 69 ve Doğruluk% 67 kesinliğe ulaştı. İşlev algoritması (2) (Nesne, metin ve nesne üzerindeki metin arasındaki mesafelerle birleştirme)% 76,% 70'i Geri Çağırma, F1 Skoru% 73 ve Doğruluk% 71 kesinliğe ulaştı ve ana katkımız Convolutional sinir ağı CNN'in birleşimi oldu. ve Çok Katmanlı Perceptron (MLP)% 85, Geri Çağırma% 80, F1 Skoru% 82 ve Doğruluk% 81'i buldu.

Anahtar Kelimeler: Derin Öğrenme, Daha Hızlı RCNN, Makine Öğrenimi, Satış El ilanları, Nesne Tespiti, Metin Tespiti.

ACKNOWLEDGEMENTS

First of all, I want to thank my omnipotent God for giving me the wisdom and intelligence necessary to carry out my thesis.

Second, I would like to thank Dr. Yakup Genç for his guidance, constant support and motivation through the entire process of my dissertation development. He was the one who introduced me to the area of Deep Learning and Machine Learning and generously agreed to be my faculty advisor for this thesis. Since then he has been instrumental in developing my concepts in this field of study. He helped me to build my research skills in a structural manner while giving me the freedom to innovate and illustrate my own ideas.

Last but not least, I would also like to thank all members of my family and friends for always being there for me when I needed them and believing in me despite the circumstances and distances.

TABLE of CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDMENTS	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS and ACRONYMS	x
LIST of FIGURES	xi
LIST of TABLES	xii
1. INTRODUCTION	1
2. BACKGROUND	3
2.1. Approaches Used in the Thesis	4
2.1.1. Constrained Clustering	4
2.1.2. Convolutional Neural Network	4
2.1.3. Convolution Layer	5
2.1.4. Pooling Layer	5
2.1.5. Fully-Connected Layer	6
2.1.6. Resnet	6
2.1.7. Multilayer Perceptron	6
2.1.8. Dropout Regularization	6
2.1.9. Cross-Validation	7
2.1.10. Optical Character Recognition (OCR)	7
2.1.11. Common Evaluation Metrics	7
2.1.12. Word2Vec	9
2.2. Alternative Approach for Object Recognition	9
2.2.1. R-CNN	9
2.2.2. Fast RCNN	10
2.2.3. YOLO (You Only Look Once)	10
3. MULTIMODEL FUSION	11
3.1. Product recognition and location: Faster R-CNN	12

3.1.1.	Region Proposal Networks on Faster RCNN	13
3.1.2.	Classifier Fast-RCNN	13
3.1.3.	Data Details for Product Detection	13
3.1.4.	Performance Evaluation Metrics for Object Detection	14
3.1.5.	Evaluation Results of Object Detection	17
3.2.	Text Location and Extraction	21
3.2.1.	Neural Network for Text Localization	21
3.2.2.	OCR with API Microsoft	21
3.2.3.	Result of Text Location and Extraction	22
3.3.	Merge Algorithms	25
3.3.1.	By distances Between Object and Text	25
3.3.2.	By distances Between Object, Text and Text on the Object	27
3.3.3.	CNN and Multi-Layer Perceptron	29
3.3.4.	Evaluation of CNN and Multi-Layer Perceptron Results	30
4.4.	Discussion	35
5.	CONCLUSIONS	39
	REFERENCES	41
	BIOGRAPHY	44

LIST of ABBREVIATIONS and ACRONYMS

<u>Abbreviations</u>	<u>Explanations</u>
<u>and Acronyms</u>	
<i>ANN</i>	: Artificial Neural Network
<i>CNN</i>	: Convolutional neural network
<i>CTPN</i>	: Connectionist Text Proposal Network
<i>CV</i>	: Computer Vision
<i>GPU</i>	: Graphics processing unit
<i>ILSVRC</i>	: ImageNet Large Scale Visual Recognition Competition
<i>IOU</i>	: Intersect over Union
<i>KNN</i>	: K Nearest Neighbors
<i>LSTM</i>	: Long Short-Term Memory
<i>mAP</i>	: Mean Average Precision
<i>mAR</i>	: Mean Average Recall
<i>MLP</i>	: Multilayer Perceptron
<i>MSE</i>	: Maximally Stable Extremal Regions
<i>OCR</i>	: Optical Character Recognition
<i>PC</i>	: Computer
<i>RAM</i>	: Random-access memory
<i>R-CNN</i>	: Region Convolutional Neural Network
<i>RNN</i>	: Recurrent Neural Networks
<i>RPN</i>	: Region Proposal Network
<i>SSD</i>	: Single Shot Multi Box Detector
<i>SVM</i>	: Support Vector Machines
<i>VGG16</i>	: Model of Convolutional Neural Network
<i>YOLO</i>	: You Only Look Once

LIST of FIGURES

<u>Figure No:</u>	<u>Page</u>
3.1: Architecture of an Algorithm dedicated to automatic extraction of information from sales flyers.	11
3.2: Faster-RCNN Architecture.	12
3.3: Total Loss.	14
3.4: The Formula with Which the Intersection of The Union (Iou).	15
3.5: Sample IoU Scores.	15
3.6: Precision-Recall Curve for A Product Classifier.	18
3.7: Precision-Recall Curves Calculated at Various IoU Thresholds.	18
3.8: Product Detection Results Faster R-CNN With Our Data.	20
3.9: Text Detection Result with CTP Network.	23
3.10: Text Detection Result with API Microsoft.	24
3.11: Result of Merge Algorithm by Distances Between Object and Text.	26
3.12: Result of Merge Algorithm by Distances Between Object, Text and Text on The Object.	28
3.13: Our Combined Method Has Multi-Input + Mixed Data Model.	29
3.14: Training Measure Accuracy Over Epochs With 4 Cross Validation.	30
3.15: Learning Curves of Model Accuracy on Train and Test Dataset. Over Each Training Epoch With 3 Cross Validation.	31
3.16: Model Loss on Training and Validation Data With 3 Cross Validation.	32
3.17: The Roc Curve of The Our Model Evaluation.	33
3.18: Result of CNN and Multi-Layer Perceptron (MLP).	34

LIST of TABLES

<u>Table No:</u>	<u>Page</u>
3.1: Accumulating evaluation results.	19
3.2 Cross Validation Performance Result of the 3 End-to-End Methods.	37
3.3 Average Performance Result of the 3 End-to-End Methods using Cross Validation.	37

1.INTRODUCTION

In present years text and object localization from images and digital documents as sale flyers is an interesting task that finds applications in several commercial areas where autonomous systems could replace human tasks by performing tedious repetitive data entry tasks[1].

Nowadays, the increase in market competition both physical and online has generated for years the massive use of sales flyers[2] which are an excellent means to increment the perceivability of physical supermarkets, being an advertising tool used very often due to its low production cost and easy distribution as well as generating an easy interaction between the customer and the supplier of products, this way they can frequently announce a new product or new offers[3]. Everything has to do with transmitting information to the customer otherwise, the unstructured offers showing up on those documents can't be effectively contrasted neither be easily comparable with online offers, making it difficult for a client to understand if it is more appropriate to buy a product online or in the physical store, also another drawbacks is that, usually in recommendation systems customers are only interested in a specific product but the sales flyers have more than one product therefore, the way is to give the user the specific recommendation of a product, the exact product they are looking for or what they previously were interested in, therefore, it increases the need for an automatic learning algorithm for information extraction from sales flyers to be used.

Reading [4] and text extraction in commercial images has been a task that has generated broad interest in deep learning and machine learning [5]. such as an image on OCR, translation of language, etc. Thus, there are many different models of neural networks which have been used previously for object detection such as CTPN (Connectionist Text Proposal Network)[4], Faster R-CNN[5], YOLO [6], SSD [7], Maximally Stable Extremal Regions (MSER) have been modified to detect scene texts and those methods had been proposed to deal with this issue. However, because texts in sales flyers have various colors, font, and size, it is difficult to detect all kinds of text[8].

Correspondingly, the detection of objects has had a massive growth during the past years, starting with the classification of a simple image, until locating or detecting multiple objects in the same image is a more challenging labor that demand more

complex process to work out because detection requires the precise location of objects that exist in processed number of candidate objects and these candidate objects put up just approximate location that has to be refined to obtain precise localization[5].

We introduce an automatic learning algorithm, shown in figure 3.1 that detects products and its information such as description, price and extracts them, finally classifies them by location. On our implementation, we will be focused on information extraction from sales flyers, which consists of three parts: the first is the detection of objects (Products) in the sales flyers inspired by the faster region convolutional neural network (Faster RCNN) framework and Residual Neural Network as feature map. The second step is the detection and extraction of the text in the sale flyers by optical character recognition. The text detection and extraction are more challenging than the recognition task carried on word images where the main challenge is the different text patterns backgrounds of the extracted text. For carrying out the text detection task on sales flyers we applied Connectionist Text Proposal Network (CTPN) implemented by [4] with VGG16 model that accurately localizes and recognizes text sales flyers. The Text Network detects the text in a string of fine-layover text proposals in convolution feature maps and even for some task we applied an extract text with OCR from a sales flyer by using Computer Visions API Microsoft. Towards Object Detection on sales flyers, we applied a Deep Residual Learning[9]. Towards Image Recognition implementation for Faster-RCNN [11] architecture. This method allows training of deep networks [10] accuracy and reduces computational cost and finally the merge algorithm to merge each product with its information such as price and description.

Finally, after obtaining the Object (the Product) and its information we presented the merge algorithms intending to join each product with its respective description and price. We presented three different merge algorithms and compared them: (1) merge by distances between Object and text, (2) merge by distances between Object, text and the text on the object and our main contribution is the combination with (3) Convolutional neural network CNN and Multilayer Perceptron (MLP).

2.BACKGROUND

In this section, we reviewed works that relate to our representation: objects recognition, localization, and text detection and extraction methods and discussed in detail solution performances which will be an essential guide for the solution of extracting important information from sales flyers.

Diverse papers have proposed ways to use deep networks and have employed machine learning techniques for locating specific bounding boxes for texts and objects. The initial usability of Deep Convolutional Neural Networks was handwritten in the 1990s but it was not possible to remain fashionable when SVM appeared. The interest in the convolutional neural network was revived by [10] showing a considerable improvement in image classification accuracy [11] [12].

Some text recognition approaches [13] [4] [14] [12] [15] treat text as a set of characters, thus they first localize the characters and then group characters into a text line. There are several methods for dealing with texts indifferent orientations, like [16] [17] [18] grouping of multi-oriented lines is done through rule-based methods or a more complex graphical model. Generally, all systems for performing recognition of image-text use two processes[19]. The first is from the input to detect regions of text and extract these regions of interest. Second, are recognition of the text and return of the extracted text [20] [21] have proposed using an artificial neural network as Faster R-CNN for locating text bounding boxes. In the method [22], proposed a text instance by two forms: text center block and word stroke region. These objects are detected by two FCN. Finally, the detections of texts are produced by the text region as a surrounding box.

Fortunately, [3] [2] presented a method that is a bit similar to our approach, they introduced a Content Extraction from Online flyers by the next steps: Step(i) single tokens are extracted from the processed marketing flyer; step(ii) each token gathered is classified using an adequately trained CNN classifier from AlexNet architecture[21]; step(iii) neighboring words having the same outcome are merged into semantically interrelated paragraphs; step(iv) paragraphs representing associated product titles, descriptions and prices are further merged all together to identify the deals contained within the processed sales flyer.

2.1. Approaches Used in the Thesis

In this section, an overview is given about the important findings from literary research on machine learning and deep learning approaches used in this thesis. The worldwide increment of machine learning and deep learning use and applications brings up the issue: Which method should be used for our specific application? To answer that question, the section deals with types of machine learning and deep learning algorithms, followed by techniques as well as a detailed elaboration in general.

2.1.1. Constrained Clustering

In a semi-supervised grouping setting constrained clustering is a class of semi-supervised learning algorithms. Typically, constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or both, with a Data clustering algorithm. Both a must-link and a cannot-link constraint define a relationship between two data instances.

2.1.2. Convolutional Neural Network

Convolutional Neural Networks (CNN) are very similar to ordinary Neural Networks which were inspired by biological processes. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. CNN architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture which makes the forward function more efficient to implement and vastly reduce the number of parameters in the network. Once the desired characteristics have been extracted after applying a suitable number of convolutions and pooling, the last pooling layer output is connected to a CNN to perform a classification of the image.

2.1.3.Convolution Layer

Convolution is an essential linear filtering operation in computational vision, which uses different filters or kernes which are applied for tasks such as blurring, sharpening, accentuating relief or edge detection in an image[27].

When speaking of convolution, there are three main concepts to consider: the input, usually an image which will be represented as I , the kernel, which is represented by the letter K and the output, the feature map. The convolution of a kernel K of size $m \times n$ placed in the pixel (i, j) of an I image is represented by the following expression:

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n) \quad (2.1)$$

The convolutional layers of the networks usually apply several convolutions in parallel at different points (i, j) , this can be better understood by visualizing a kernel as a sliding window that runs through the whole image and calculates the resulting feature map in each jump.

2.1.4.Pooling Layer

It is common to periodically insert a Pooling layer in-between successive Conv layers in a ConvNet architecture. Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. The Pooling Layer operates independently on every depth slice of the input and resizes it spatially, using the MAX operation. The most common form is a pooling layer with filters of size 2×2 applied with a stride of 2 down samples every depth slice in the input by 2 along both width and height, discarding 75% of the activations. Every MAX operation would in this case be taking a max over 4 numbers (little 2×2 region in some depth slice).[28].

2.1.5.Fully-Connected Layer

Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset. See the Neural Network section of the notes for more information.

2.1.6.Resnet

Residual Network developed by [18] was the winner of ILSVRC 2015. It features special skip connections and a heavy use of batch normalization. The architecture is also missing fully connected layers at the end of the network.

2.1.7.Multilayer Perceptron

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network.

2.1.8.Dropout Regularization

Dropout refers to ignoring neurons during the training phase of certain set of neurons which is chosen at random. By “ignoring”, I mean these units are not considered during a particular forward or backward pass. More technically, at each training stage, individual nodes are either dropped out of the net with probability $1-p$ or kept with probability p , so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed. Dropout regularization is useful to prevent of overfitting in deep networks due a fully connected layer occupies most of the parameters, and hence, neurons develop co-dependency amongst each other during

training which curbs the individual power of each neuron leading to over-fitting of training data [24].

2.1.9. Cross-Validation

In this technique, known as cross-validation, the original corpus is subdivided into k subsets called folds. For each of these subsets, we train a model using all the data except the data belonging to that subset, and then the model is tested with the data of the subset in question. The choice of the value k will depend on the size and characteristics of the data set.

2.1.10. Optical Character Recognition (OCR)

OCR, also referred to as text recognition, is technology that transforms characters such as numbers and letters from printed or written documents into an electronic form more easily recognized and read by computers and other software programs. OCR allows to extract information from images, which splits the letters on the image, and then puts them in words and the words in a sentence, which allows us to access and extract and edit text.

2.1.11. Common Evaluation Metrics

The metrics that you choose to evaluate your machine learning algorithms are very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. They influence how you weight the importance of different characteristics in the results and your ultimate choice of which algorithm to choose. In this section, we discuss basic evaluation metrics commonly used for classification problems.

- Accuracy measures the percentage of entries in the evaluation set that the classifier correctly classifies.
- The confusion matrix is a handy presentation of the accuracy of a model with two or more classes. This table presents predictions on the x-axis and accuracy

outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm.

To define Precision, recall and F-score measures, we will consider the following groups of elements in which an instance can be classified:

- True Positives (TP) - This shows that a model correctly predicted positive cases as positive.
- True Negatives (TN) - This shows that a model correctly predicted Negative cases as positive.
- False Positives (FP) This shows that a model incorrectly predicted negative cases as positive.
- False Negatives (FN) This shows that an incorrectly model predicted positive cases as negative.
- Precision is the measure that indicates how many of the elements we have identified are positive. Precision can be empirically calculated using [25]:

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$

- Recall is the measure that indicates how many of the positive elements we have identified. Recall can be empirically calculated using [25]:

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

- F1 score is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. F1 score can be empirically calculated using [25]:

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \quad (2.4)$$

2.1.12. Word2Vec

Word2vec is a two-layer neural net that processes text. Its input is a text corpus and its output are a set of vectors: feature vectors for words in that corpus, the usefulness of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. Word2Vec is seemingly the most famous of the word embedding models. Word embeddings are a sort of word representation that allows words with comparative importance to have a similar representation and words are represented in a predefined vector space. Every word is mapped to just one vector and the values of the vector are in a manner that is a resemblance to a neural system. In any case, Word2vec's applications reach out past parsing sentences in nature. It tends to be applied similarly also to qualities, code, likes, playlists, web-based life charts and other verbal or emblematic arrangement in which examples might be observed. So, in single word `gene2vec`, `like2vec` and `devotee 2vec` are all possible. Word2Vec can be implemented utilizing multilayer perceptron (MLP), RNN's, long short-term memory (LSTM's).

2.2. Alternative Approaches for Object Recognition

This section explains those alternative techniques that could be used to carry out the extraction of information from sales flyers and we refer to the region-based strategies for object detection.

2.2.1.R-CNN

R-CNN is one of the state-of-the-art CNN-based deep learning object detection approaches, was first introduced by Ross [13] in 2014. This type of convolutional network is forced to focus on a single region at a time, that way interference is minimized because it is expected that only a single object of interest will dominate in a given region. The regions in the R-CNN are detected by a selective search algorithm followed by resizing so that the regions are of equal size before they are fed to a CNN

for classification and bounding box regression. The general procedures are: First, R-CNN uses selective search to generate about 2K region proposals, i.e. bounding boxes for image classification, Then, for each bounding box, image classification is done through CNN. Finally, each bounding box can be refined using regression.

2.2.2.Fast RCNN

Fast R-CNN moves one step forward. Instead of applying 2,000 times CNN to proposed areas, it only passes the original image to a pre-trained CNN model once. Search selective algorithm is computed based on the output feature map of the previous step. Then, ROI pooling layer is used to ensure the standard and pre-defined output size. These valid outputs are passed to a fully connected layer as inputs. Finally, two output vectors are used to predict the observed object with a SoftMax classifier and adapt bounding box localizations with a linear regressor[26].

Faster R-CNN makes further progress than Fast R-CNN. Search selective process is replaced by Region Proposal Network (RPN). As the name revealed, RPN is a network to propose regions. For instance, after getting the output feature map from a pre-trained model (VGG-16), if the input image has 600x800x3 dimensions, the output feature map would be 37x50x256 dimensions [19].

2.2.3.YOLO (You Only Look Once)

Yolo is yet another neural network designed for object detection. The object detection task consists of determining the location on the image where certain objects are present, as well as classifying those objects. Previous methods for this, like R-CNN and its variations, used a pipeline to perform this task in multiple steps. This can be slow to run and also hard to optimize because each component must be trained separately. YOLO does it all with a single neural network.

3.MULTIMODEL FUSION

This section presents details of an algorithm dedicated to the automatic extraction of information from sales flyers. The algorithm detects product information such as their descriptions and prices. This information can be in text as well as in pictures. The algorithm has three parts: the first two involve text and the product's picture detection whiles the third part merges these text and pictures to identify each item in the given sales flyer. The whole system is summarized in Figure 3.1. In the rest of this section, each of these components are described in detail.

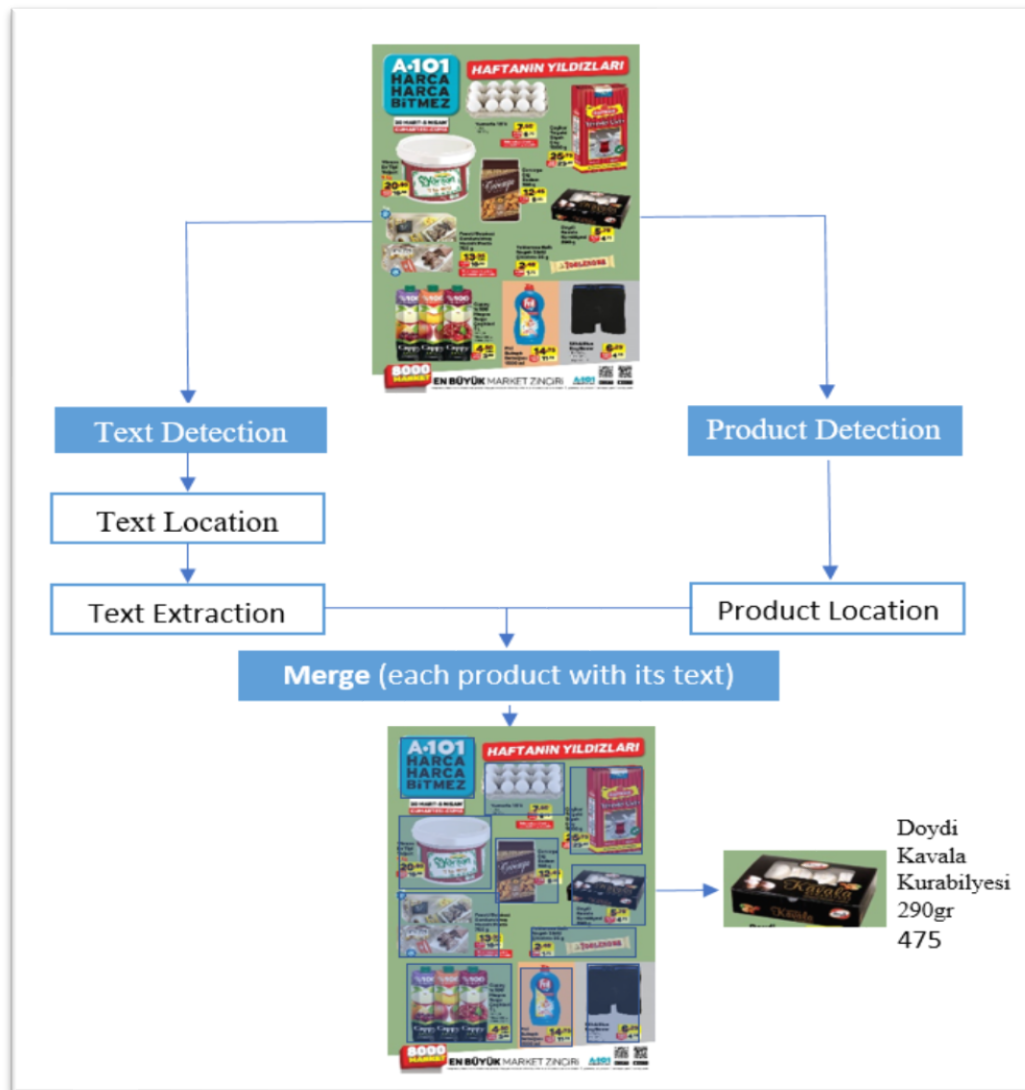


Figure 3.1: Architecture of aalgorithm dedicated to automatic extraction of information from sales flyers, the right side represents the process for product. detection and the left side represents the process for extracting text, both are joined by an algorithm.

3.1. Product recognition and location: Faster R-CNN

Our product detection neural network is inspired by the original Faster R-CNN, the algorithm takes the full image as the input. Following this, the input is propagated through all the convolution layers of the network (Deep Residual Learning [18]). The main improvement of the technique is that the region proposed network (RPN [27]) is placed on top of the convolution network. Faster RCNN could just ‘look’ at the last convolutional feature map and produce the region proposals. It then classifies the input as ‘object’ or ‘not object’, using binary classification loss. Finally, if recognized as an ‘object’, the classification decision and the bounding box for the object is produced [7].

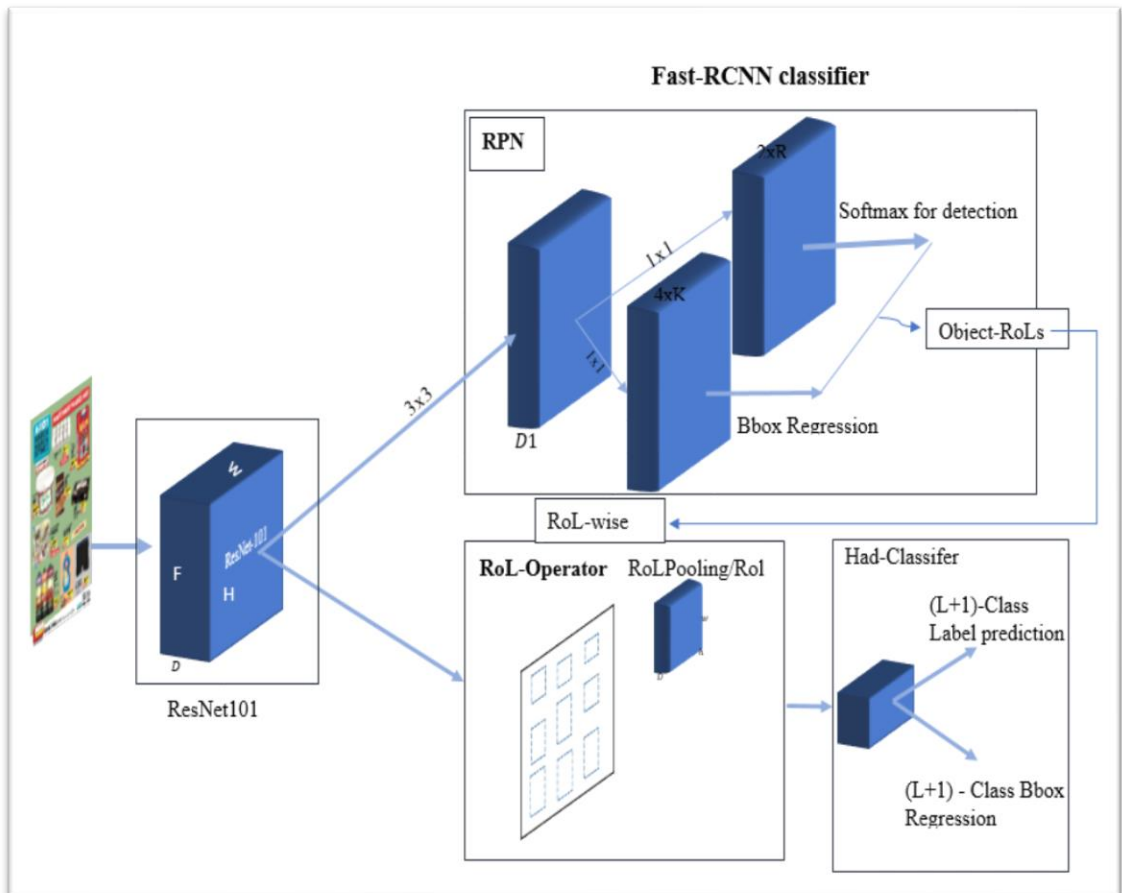


Figure 3.2: Faster-RCNN Architecture, It is composed of three parts, (1) ResNet Convolutional Neural Network, (2) Region Proposal Network and (3) Two Fully Connected Layers.

3.1.1.Region Proposal Networks on Faster RCNN

The region proposal network (RPN) in the faster region-based convolutional neural network (Faster R-CNN) is used to decide “where” to look in order to reduce the computational requirements of the overall inference process. The RPN quickly and efficiently scans every location in order to assess whether further processing needs to be carried out in a given region. It does that by outputting k bounding box proposals each with 2 scores representing probability of object or not at each location [10]. The anchors boxes are just referencing, they are selected to have different aspect ratios and scales in order to accommodate different types of objects, elongated objects like buses, for example, cannot be properly represented by a square bounding box. In Faster R-CNN they used $k = 9$ representing 3 scales and 3 aspect ratios. Each regressor in the RPN only computes 4 offset values (w , h , x , y) to the corresponding reference anchor box.

3.1.2.Classifier Fast-RCNN

The RPN uses a 3×3 window that slides over a high-level conv feature map, the effective size of that small window is actually 177×177 when reprojected back to the input layer, so the RPN is actually using a lot of context when making the proposals. This 3×3 window is resampled to a 256-dimensional vector before feeding into two fully connected layers, a box regression layer (reg), that computes the box offsets, and the box classification layer (cls) that computes the confidence scores that are related to probability of objectness. The regression layer has $4k$ outputs while the cls layer has $2k$ outputs making the total RPN output per position to $4k + 2k$.

3.1.3.Training Result and Data Details for Product Detection

Since our aim was to use the state-of-the-art Faster RCNN framework for product recognition, we needed a dataset that was suitable to use in it. The Faster RCNN framework requires a dataset with images with bounding box information for localization of object in the image, and a label for the object inside the bounding box.

Hence, images with just the label(s) of its content and no bounding box information was not useful for our purposes.

To evaluate the proposed approach, we prepared our own data set, because many of products that the system needs to detect are not available on the internet. A total number of 10.000 images with approximately a total number of 25.000 products have been obtained from commercial marketing flyers. The collected sales flyers come from different domains of industries e.g. electronic, foods and present different design styles. Each product has been manually labelled by us using software called LabelImg which is a graphical image annotation tool, that allows us to draw a bounding box for each item on the sales flyers and it provides XTML document with bounding boxes coordinates.

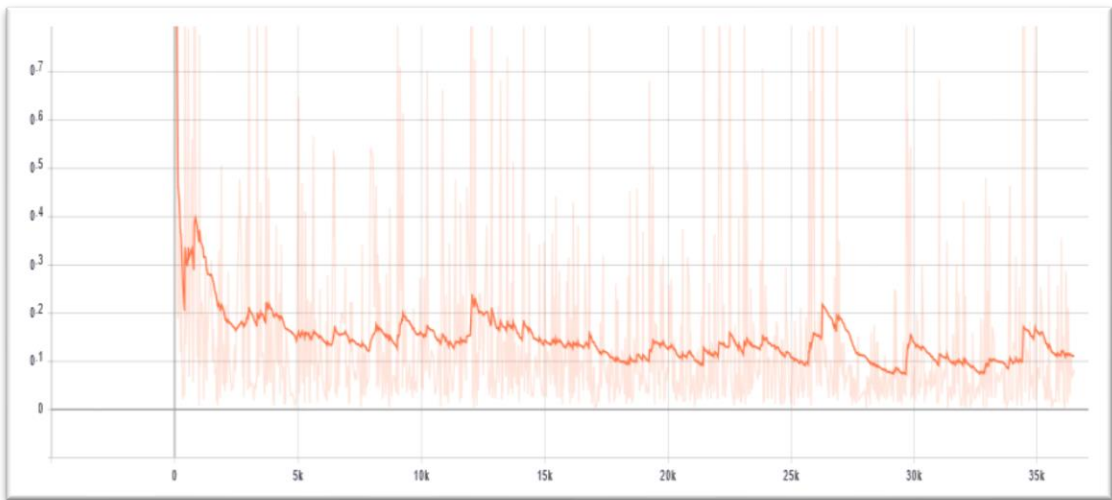


Figure 3.3: Total loss: The orange line shows the values applied to the training examples and represents the cost on the training data, where it can be seen that the loss function has a great decrease in the 5k interaction then it becomes a bit stable.

3.1.4. Performance Evaluation Metrics for Object Detection

In object detection, evaluation is not trivial, because there are two distinct tasks to measure: Determining whether an object exists in the image (classification) and determining the location of the object (localization, a regression task).

Furthermore, in a typical data set there will be many classes and their distributions are non-uniform. So, a simple accuracy-based metric will introduce biases. It is also important to assess the risk of misclassifications. Thus, there is the need to associate a “confidence score” or model score with each bounding box detected and to assess the model at various level of confidence.

In order to address these needs, the Average Precision (AP) was introduced. To understand the AP, it is necessary to understand the precision and recall of a classifier.

- Intersect Over Union (IoU)

Intersect over Union (IoU) is a metric that allows us to evaluate how similar our predicted bounding box is to the ground truth bounding box. The idea is that we want to compare the ratio of the area where the two boxes overlap to the total combined area of the two boxes. The formula for calculating IoU is as follows:

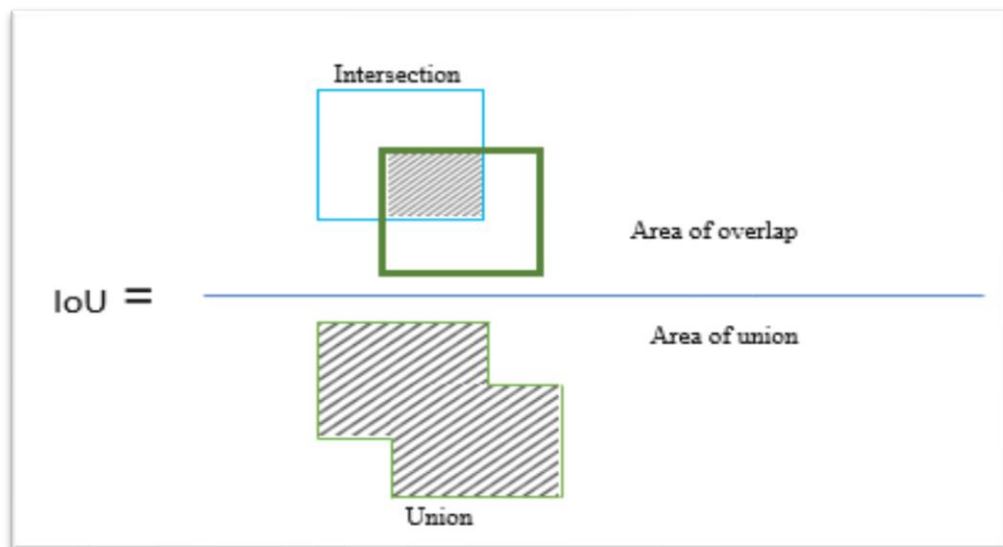


Figure 3.4: The formula with which the Intersection of the Union (IOU) which represents the areas of intersection and union in the pair detection-ground-truth.

- Interpreting IoU scores

IoU is quite intuitive to interpret. A score of 1 means that the predicted bounding box precisely matches the ground truth bounding box. A score of 0 means that the predicted and the true bounding box do not overlap at all.

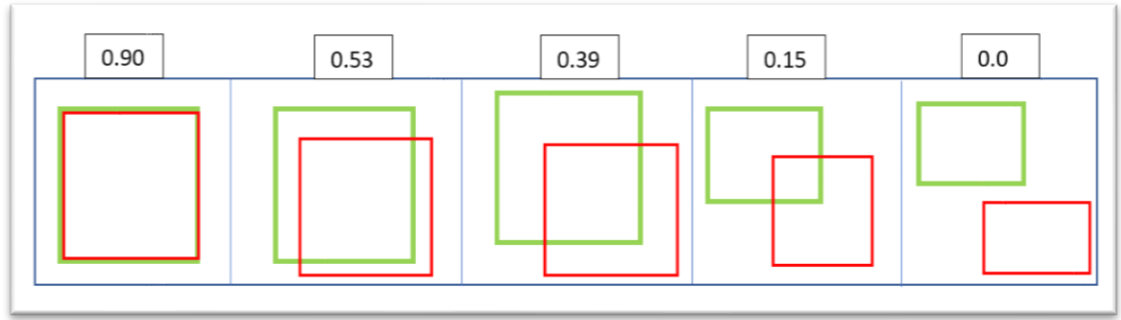


Figure 3.5: Sample IoU scores: the bounding box with the green border represents ground truth and the red one represents the predicted bounding box.

- Precision x Recall curve

The Precision $\times$ Recall curve is a good way to evaluate the performance of an object detector as the confidence is changed by plotting a curve for each object class. An object detector of a particular class is considered good if its precision stays high as recall increases, which means that if you vary the confidence threshold, the precision and recall will still be high. Another way to identify a good object detector is to look for a detector that can identify only relevant objects (0 False Positives = high precision), finding all ground- truth objects (0 False Negatives = high recall).

- Average Precision

Although the precision-recall curve can be used to evaluate the performance of a detector, it is not easy to compare among different detectors when the curves intersect with each other. It would be better if we have a numerical metric that can be used directly for the comparison. This is where average precision (AP), which is based on the precision-recall curve, comes into play. In essence, AP is the precision averaged across all unique recall levels.

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) P_{interp}(r_{i+1}) \quad (3.1)$$

- Mean Average Precision

The calculation of AP only involves one class. However, in object detection, there are usually $K > 1$ classes. Mean average precision (mAP) is defined as the mean of AP across all K classes:

$$mAP = \frac{\sum_{i=1}^k AP_i}{K} \quad (3.2)$$

- Average Recall

Like AP, average recall (AR) is also a numerical metric that can be used to compare detector performance. In essence, AR is the recall averaged over all $IoU \in [0.5, 1.0]$ and can be computed as two times the area under the recall-IoU curve:

$$AR = 2 \int_{0.5}^1 recall(o) do \quad (3.3)$$

where o is IoU and $recall(o)$ is the corresponding recall.

- Mean Average Recall

Mean average recall is defined as the mean of AR across all K classes:

$$mAR = \frac{\sum_{i=1}^k AR_i}{K} \quad (3.4)$$

- Coco Detection Metrics

This metric report additional statistics such as mAP at IOU thresholds of .5:.95, and precision/recall statistics for small, medium, and large objects.

- Weighted PASCAL VOC Detection Metric

The weighted PASCAL metric computes the mean average precision as the average precision when treating all classes as a single class.

3.1.5.Evaluation Results of Object Detection.

In this section we presented model evaluation, for this purpose we used exclusively two Precision-Recall curves obtained from the Weighted PASCAL VOC detection metric and we used the COCO detection metrics that generate a table that contains the average metrics of precision and recall for different sizes of objects.

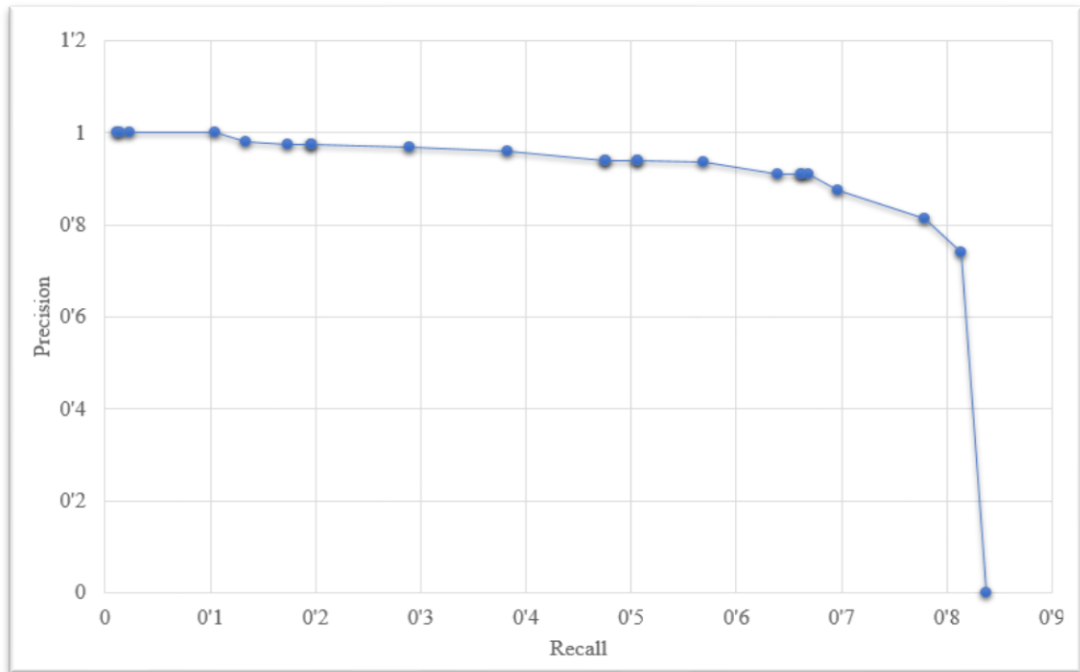


Figure 3.6: Precision-Recall curve for a product classifier with an IOU used for our system. This precision-recall curve was determined by considering all objects above a model score threshold as a positive prediction.

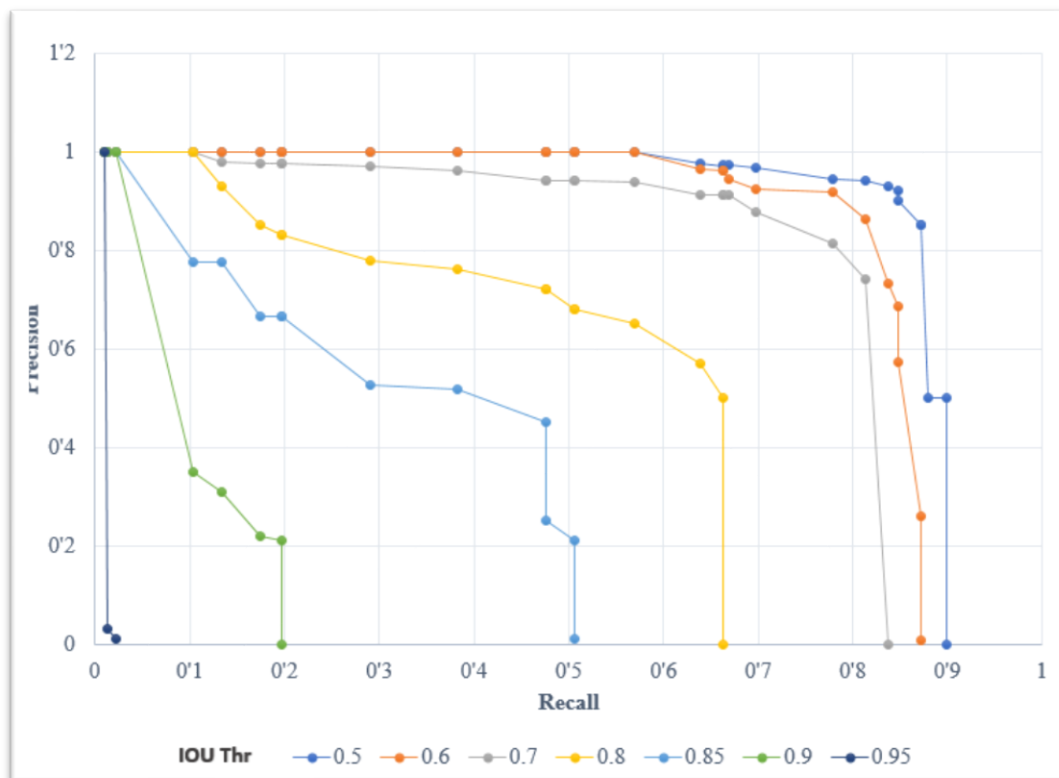


Figure 3.7: Precision-Recall curves calculated at various IoU thresholds. Dashed lines correspond to equally spaced recall values where the AP is calculated.

The most common IoU threshold for detection systems is $\text{IoU} = 0.5$. but as the main tasks are to locate the product and draw a very precise bounding box, it is therefore necessary in our case to use detections with an overlap of $\text{IoU} > 0.7$ as it acquired a good accuracy as shown in Figure 3.6.

In figure 3.7 Seven precision and recall curves have been drawn, one for each IoU threshold, first taking the standard IoU threshold, that is, 0.5 to 0.95. As it can be seen in the figure, the Faster RCNN model with the Resnet 101 extractor shows a good precise localization.

Table 3.1 Accumulating evaluation results

Average Metrics	IoU	Area	Maximum Detections	mAP
Average Precision (AP)	0.50 - 0.95	all	300	0.675
Average Precision (AP)	0.50	all	300	0.856
Average Precision (AP)	0.75	all	300	0.704
Average Precision (AP)	0.50 - 0.95	small	300	-1.000
Average Precision (AP)	0.50 - 0.95	medium	300	0.029
Average Precision (AP)	0.50 - 0.95	large	300	0.766
Average Recall (AR)	0.50 - 0.95	all	300	0.431
Average Recall (AR)	0.50 - 0.95	all	300	0.634
Average Recall (AR)	0.50 - 0.95	all	300	0.651
Average Recall (AR)	0.50 - 0.95	small	300	-1.000
Average Recall (AR)	0.50 - 0.95	medium	300	0.075
Average Recall (AR)	0.50 - 0.95	large	300	0.745

As you can see Table 3.1, reports mAP at IOU thresholds of .5:.95, and precision/recall statistics for small, medium, and large objects, where small and medium objects had a very bad mAP score because, in our system the images of training and testing do not contain small or medium objects, everything is considered large. On the contrary, in large objects the results vary depending on the value of the IOU, we can see that when the IOU value is 0.5 the mAP value is much better than the rest for all objects of different sizes.

3.2. Text Location and Extraction

This section puts up two distinct algorithms proposed for text extraction. One is a Novel Connectionist Text Proposal Network (CTPN)[4] that directly localizes text sequences in convolutional layers. We leverage the advantages of strong deep convolutional features and sharing computation mechanism and on the other hand, we presented the Microsoft API application for text extraction and compared the results.

3.2.1. Neural Network for Text Localization

Similar to Region Proposal Network RPN [5], the CTPN[4] is essentially a fully convolutional network that allows an input image of arbitrary size. It detects a text line by densely sliding a small window in the convolutional feature maps, and outputs a sequence of fine-scale text proposals.

We take the very deep 16-layer VGG16 [4] as an example to describe our approach, which is readily applicable to other deep models. We use a small spatial window, 3×3 , to slide the feature maps of last convolutional layer. The size of conv5 feature maps is determined by the size of input image, while the total stride and receptive field are fixed as 16 and 228 pixels, respectively. Both the total stride and receptive field are fixed by the network architecture. Using a sliding window in the convolutional layer allows it to share convolutional computation, which is the key to reduce computation of the costly sliding-window[4]. In the end we applied OCR [28] for each text proposal.

3.2.2. OCR with API Microsoft

The Computer Vision API provides state-of-the-art algorithms to process images and return information. Optical Character Recognition (OCR) detects text in an image and extracts the recognized characters into a machine-usable character stream. The OCR results in the hierarchy of region/line/word. The results include text, bounding box for regions, lines and words. Bounding box of a recognized region, line, or word, depending on the parent object. The four integers represent the x-coordinate of the left

edge, the y-coordinate of the top edge, width, and height of the bounding box, in the coordinate system of the input image.

3.2.3.Result of Text Location and Extraction

The Neural Network for Text Localization for training on GTX1070 for 20k iterations on ICDAR 2013/2015 data. It takes about 4s per iteration. So, it will take about 3.5 hours to finished 30k iterations obtaining precision 0.74, recall 0.52 and F-measure 0.61. This method is efficient when text is in a line sequence. But for our case it could not work perfectly because this method detects text sequentially and can unite some products and price descriptions in places that do not correlate, see result in Figure 3.9. Therefore Microsoft API for text recognition and extraction has a better result, shown in Figure 3.10, which recognizes word by word and for each of them generates its limited box evading the previous problems.

A.101
HARCA
HARCA
BITMEZ

2 MAYIS
PERŞEMBE

ÜLKER

ALDIN ALDIN!

Ülker Kremalı Bisküvi 8x76 g 6,95

Ülker Çikolatalı Gofret 5x36 g 3,95

Ülker Saklıköy Fındık Kremalı Bisküvi 3x88 g 3,45

Ülker Çokokrem Pankek 40 g 1,00

Ülker Hanımeller Kurabiyem 180 g 4,45

Ülker Canpare Kremalı Bisküvi 81 g 1,00 Adet

Ülker Fındıklı Kremalı Rulokat 170 g 4,95

Ülker Albeni Çikolata Kaplı Kek 43 g 1,00

Çerezos Fox Baharatlı Mısır Cipsi 27 g 0,90

PD Foods Kuru Meyve Bari Çeşitleri 20 g 1,00 Adet

Nari Osmanlı Şerbeti İçecek 1 L 2,95

8000 MARKET

EN BÜYÜK MARKET ZİNCİRİ

A.101

QR codes and social media links at the bottom.

Figure 3.10: Text Detection Result with API Microsoft has the better result, this method recognizes word by word and for each of them generates its bounding box.

3.3. Function Algorithms

In this section, we presented in detail the three different algorithms dedicated to automatic extraction of information from sales flyers. Our algorithms started with a general procedure. This procedure consists of generating a new bounding box for the entire description of the products since the detection and extraction of text generates a text separated by words as shown in Figure 3.10 and it is necessary to unite each product description in a single text set. To generate this new bounding box for each product description, we used clustering task that collects text based on similarity distances.

3.3.1.By distances Between Object and Text

To merge each product with its respective text information which are prices and the description, this algorithm uses the outputs of both products and texts location coordinates generated for both Faster RCNN and API Microsoft as shown in Figure 3.10. The algorithm starts by selecting the bounding box of a random product and merge it with its closest description bounding boxes to form a set of offers. The Same process is repeated for all the remaining bounding boxes in the sales flyers to form a finite set of products with its descriptions, this whole process is repeated multiple times for each item number that has the sales flyer and finally, we paint a new bounding box which contains the product, price and description, as shown in Figure 3.10.

This algorithm has a good performance when the products are separated by a considerable distance, but when there are many products in the sale flyers and they are very close together the algorithm tends to get confused and merge some product with the incorrect description.

A.101

HARCA HARCA BİTMEZ

2 MAYIS*
PERŞEMBE

ALDIN ALDIN!

Ülker Kremalı Bisküvi 8x76 g

6,95

Ülker Çikolatalı Gofret 5x36 g

3,95

Ülker Saklıköy Fındık Kremalı Bisküvi 3x88 g

3,45

Ülker Çokokrem Pankek 40 g

1,00

Ülker Hanımeller Kurabiyem 180 g

4,45

Ülker Canpare Kremalı Bisküvi 81 g

1,00

* Fındıklı
* Çikolatalı

Ülker Fındıklı Kremalı Rulokat 170 g

4,95

Ülker Albeni Çikolata Kaplamalı Kek 43 g

1,00

Çerezos Fox Baharatlı Mısır Cipsi 27 g

0,90

PD Foods Kuru Meyve Barı Çeşitleri 20 g

1,00

* Kayısı & Badem
* İncir & Ceviz

Nari Osmanlı Şerbeti İçecek 1 L

2,95

EN BÜYÜK MARKET ZİNCİRİ

* Kampanya 2 Mayıs tarihinden itibaren geçerlidir. Bu alışverişte her bir ürün 8.500 adetle sınırlıdır. | Çağrı Merkezi: 0800 099 2101 | İletişim: 0212 250 00 00 | www.a101.com.tr

Figure 3.11: Result of Merge algorithm by distances between Object and text; On the right side is the text extracted by each item on the sales flyer, but this algorithm has a bad classification when the products are very close.

3.3.2.By distances Between Object, Text and Text on the Object

This algorithm is similar to the previously explained one, unlike the previous algorithm, before merging the product with the text, a new procedure is added that consists of detecting and extracting the text on the object (generally the text that is on the product is the product's name) and then comparing by distances that text with product descriptions, in case its similar it immediately considers that the product description corresponds to the product, on the other hand, if the algorithm cannot extract or calculate the similarity matrix enters the product descriptions and the text that is on the product, we proceed to return to the previous algorithm calculating the similarity matrix between the bounding boxes of both the description and the product. The same procedure is continued for each text and object. This whole process is repeated multiple times for each item number that has the sales flyer and finally, we paint a new bounding box which contains the product, price and description.

A.101

HARCA HARCA BİTMEZ

HAFTANIN YILDIZLARI

17-23 AĞUSTOS*
CUMARTESİ-CUMA



Danet Dana Fermente Sucuk 500 g

~~28.95~~
24.95 (%14 indirim)



Biryah Homojenize Yoğurt 1500 g

~~6.90~~
5.75 (%17 indirim)



Mest Bitki Çay Cesitleri 20'li

~~3.50~~
2.50 (%28 indirim)



Dr. Oetker Guseppe Dondurulmuş Karışık Pizza 405 g

~~14.75~~
11.95 (%19 indirim)



Yoremce Koçbaşı Nohut 1 kg

~~7.25~~
6.25 (%14 indirim)



Dr. Oetker Dark Kakao 100 g

~~5.45~~
4.25 (%22 indirim)



Twix Xtra Karamelli Bar Çikolata 75 g

~~2.95~~
1.95 (%34 indirim)



Orkide Ayçiçek Yağı 2 L

~~21.50~~
18.50 (%14 indirim)



Çerezya Tuzlu Karışık Çerez 200 g

~~6.95~~
5.95 (%14 indirim)

8000

MARKET

EN BÜYÜK MARKET ZİNCİRİ

A101
MOBİL UYUMLAMA




* Kampanya 17-23 Ağustos tarihleri arasında geçerlidir. Ayrıca geçerli olan diğer indirimli fiyatlarla birleştirilerek 2 adet satış yapılabilir. | Çiğdem Halkalı 0452 836 2 101 | iletisim@al101.com.tr | www.al101.com.tr

Figure 3.12: Result of Merge algorithm by distances between Object, text and text on the object; Each green box represents the classification of the products with their respective information.

3.3.3.CNN and Multi-Layer Perceptron

Our model is composed of convolutional neural network and multi-layer perceptron with multiple inputs, as the structure shown in Figure 3.13. To build our combined method which has multi-input, we needed two branches: The first branch is composed of an Turkish Pre-trained Word2Vec Model followed by simple Multi-layer Perceptron (MLP) designed to handle the input of the product description, the second branch is a CNN to operate over the product image data. These branches are concatenated together followed by a fully connected layer to form the final output which is the probability of 0 and 1.

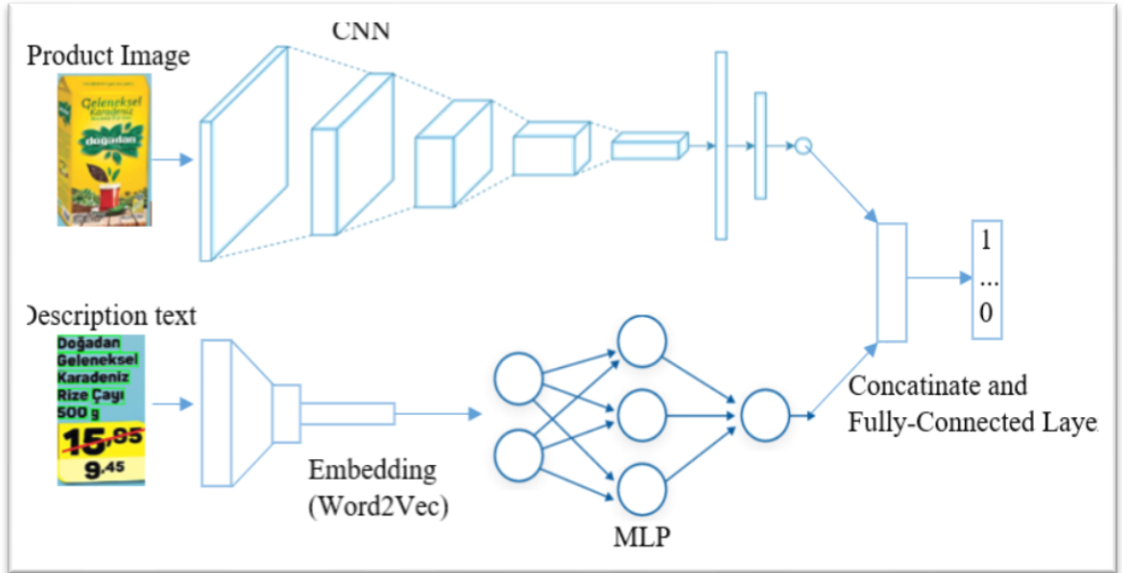


Figure 3.13: Our combined method has multi-input + mixed data model has one branch that accepts the descriptions of the product's data (below) and another branch that accepts the product's image (above).

Our model received two inputs of data, the description of the product and the image of the product, which have already been obtained previously by Faster RCN; each bounding box of the products is cut off from the sales flyer, OCR with API Microsoft; each bounding text box is clustered with the closest boxes to obtain the complete description of the product as explained in section 3.3, its output is generated by the Sigmoid function that gives us the probability between 1 and 0, which corresponds if the output is greater than 0.5, otherwise it does not correspond.

3.3.4. Evaluation of CNN and Multi-Layer Perceptron Results

We trained our combined method with 8000 products and their descriptions. This data is composed of different types of products such as electronics, appliances, clothing and accessories, toys, supplies for schools, makeup, perfumes, personal care, drinks, detergents, canned goods, cosmetics, etc.

To get a better performance, we carried out a heavy sequence argumentation of data which consists of crop + horizontal + flips + Gaussian blur, leading to significant changes in our data, increasing it to 18000 number of images. To evaluate our method in the best way, we resort to the use of the Cross-Validation method in which our original data was subdivided into k subsets called folds. For each of these subsets, we trained a model using all the data except the data belonging to that subset, and then the model tested with the data of the subset in question.

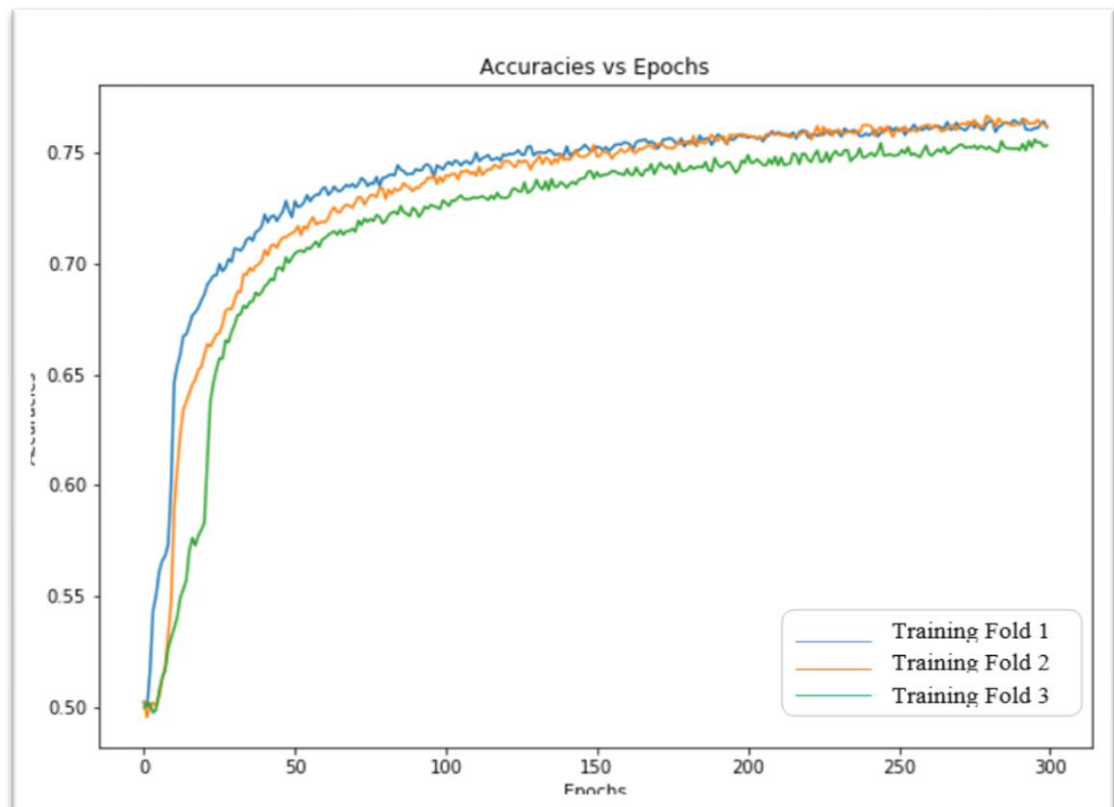


Figure 3.14: Training measure accuracy over epochs with 3 Cross Validation.

Figure 3.14 shows that 3 lines plot are created showing the learning curves for the model accuracies on the train sets over each training epoch, we can see that the model achieved an average of approximately 75% overall training folds results on the training dataset on 300 epochs.

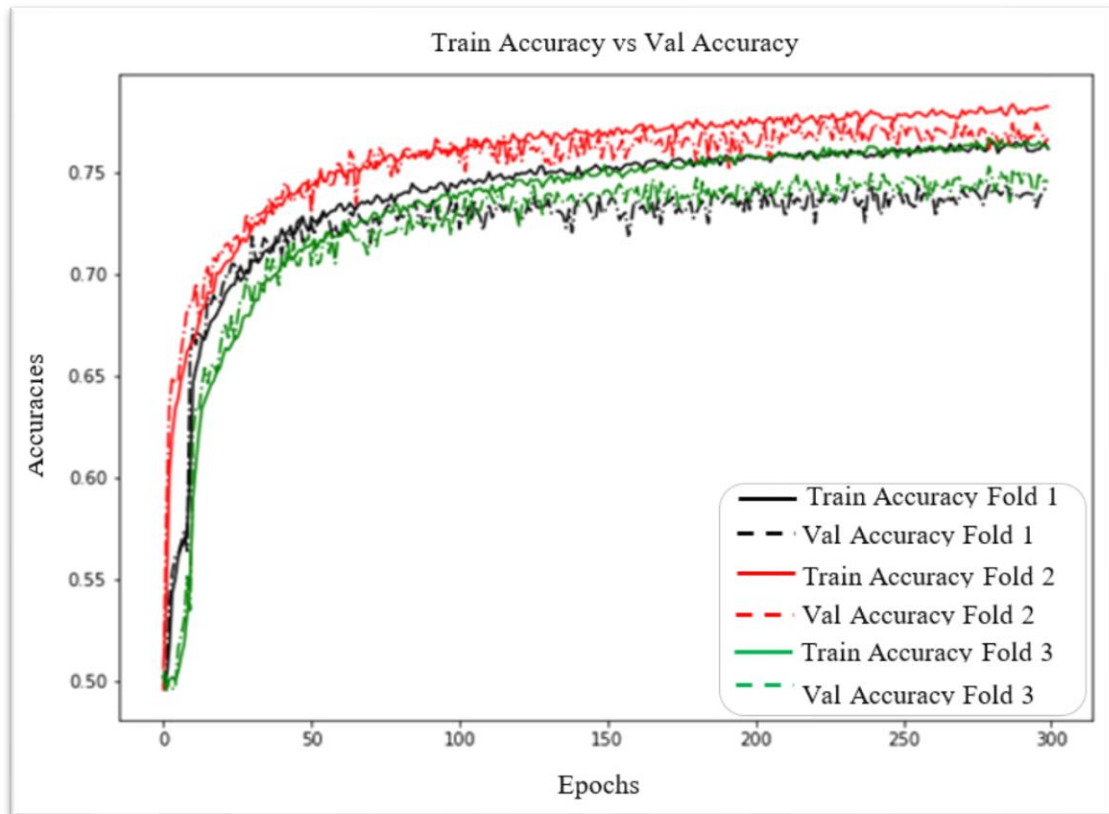


Figure 3.15: Learning Curves of Model Accuracy on Train and Test Dataset over Each Training Epoch with 3 Cross Validation.

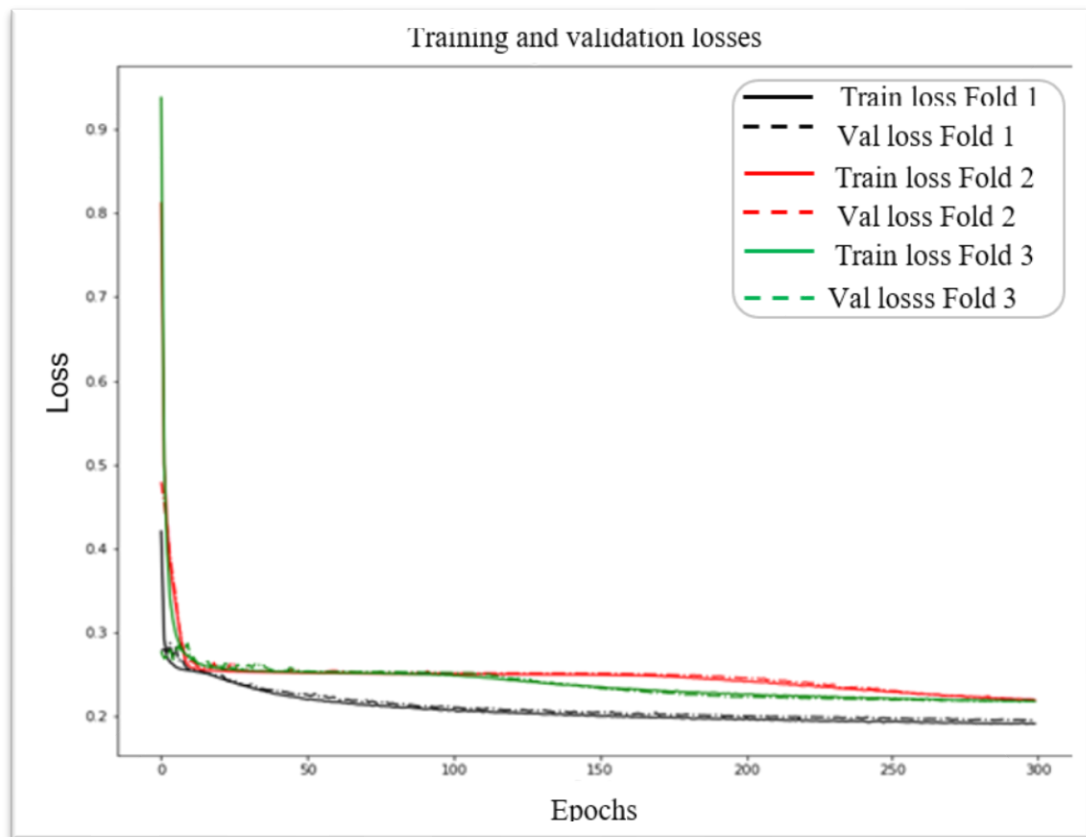


Figure 3.16: Model Loss on Training and Validation Data with 3 Cross Validation.

As you can see in Figure 3.15 The 3 lines plot represent the learning curves for the model accuracy on the training and test sets over each training epoch. In this case, we can see that the model achieved an average of approximately 75% accuracy on the training dataset, which we know is optimistic, and about 71% on the test dataset, which we would expect to be more realistic.

In Figure 3.16 of loss, we can see that the model has comparable performance on both training and validation datasets where there are 3 different Folds, which has its respective training loss validation, fold 1 is identified with the black line and the same color with a dash-dot line style of the same color represents the loss of validation, in the same way fold 2 is red and fold 3 is green. We can see that our model has good fit learning curves in each folds. Both the training and validation loss decrease in an exponential fashion as the number of epochs is increased, suggesting that the model gains a high degree of accuracy as our epochs are increased. so we can end by saying that our algorithm has good loss.

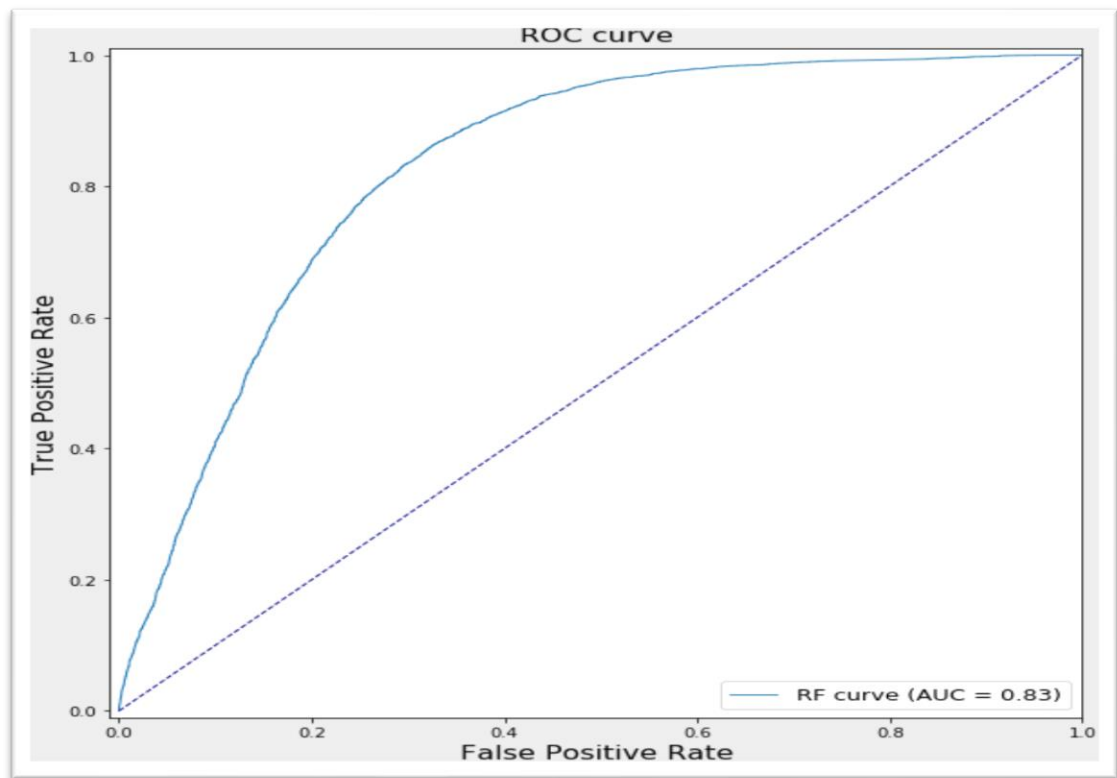


Figure 3.17: The ROC curve of the our model evaluation.

Figure 3.17 is a hypothetical ROC curve demonstrating the tradeoff between sensitivity and specificity. Particularly, sensitivity and specificity are inversely related, as the sensitivity increases, the specificity decreases, and vice versa. This figure shows that our model has excellent discrimination based on the general rule of thumb for interpreting AUC because our model is in the $0.85 \geq \text{AUC} > 0.7$ classification range, with a score of AUC 0.83.

Çerezzya Dakota Çekirdeği 250 g 0.7889216393232346
 Çerezzya Kavulmuş Fındık İçi %8 indirim 875 0.9987443116260692
 Çerezzya Kavulmuş Ana Cıtlak Antep Fıstığı %14 3295 indirim 0.5466873347759247
 Çerezzya Çiğ Badem .150 0.6514876186847687
 Coca Cola Gazlı İçecek m1 0.9999999613595385

A.101

HARCA HARCA BİTMEZ

17-26 TEMMUZ

TARİHLERİ ARASINDA



Çerezzya Kavulmuş Ana Cıtlak Antep Fıstığı 400 g

~~38,45~~
32,95
%14 indirim



Çerezzya Çiğ Badem 150 g

~~12,85~~
10,95
%15 indirim



Çerezzya Kavulmuş Fındık İçi 150 g

~~9,50~~
8,75
%8 indirim



Çerezzya Dakota Çekirdeği 250 g

~~5,95~~
4,75
%20 indirim



Coca Cola Gazlı İçecek 4x200 ml

~~6,95~~
5,45
%22 indirim

8000 MARKET

EN BÜYÜK MARKET ZİNCİRİ





Figure 3.18: Result of CNN and Multi-Layer Perceptron (MLP); Each product on the sales flyer has a green box, which lists the product detected, the price and the description. At the top the text strings extracted by each product and their respective probability show that the product corresponds to the product description.

3.4. Discussion

Our multi-model function involves different steps to carry out automatic extraction of information from sales flyers, one of them is product detection with Faster RCNN, Table 1 illustrates mAP at IOU thresholds between 5.0 to 9.5, and precision/recall statistics for different object sizes, where small and medium objects have a very bad mAP score because, our Faster RCNN model images of training and testing do not contain small or medium objects, in our model every object is considered large. On the contrary, in large objects the results vary depending on the value of the IOU as figure 3.7 shows, we can see that when the IOU value is 0.5 for all objects with different sizes, the mAP value is much better than the rest.

On the other hand, text location and extraction part consisted of a neural network for text localization getting precision 74%, recall 52% and F-measure 61%. but for our case, it could not work perfectly because this method detects text sequentially and can unite some product and price descriptions in places that do not correlate as a result, see Figure 3.9. Therefore, Microsoft API for text recognition and extraction has the better result, as shown in Figure 3.10, which recognizes word by word and for each of them generates a bounding box.

Otherwise, Table 3.2 shows a comparison of the 3 methods developed in this thesis from end to end, to obtain this comparison, the 4 most popular evaluation metrics were used. Function Algorithm 1 has a good performance when the products are separated by a considerable distance, but when there are many products in the sale flyers and they are very close together the algorithm tends to get confused and merge some product with the incorrect description.

Compared to the function Algorithms, Function Algorithm 2 has a better performance than the function Algorithm 1, due to the new added procedure where we get the text that is on the object and generally this text is the products name, which allows at the time of merge, for the product and its description to be merged correspondingly and not depending only on the distance between the objects and the text, consequently decreasing the issue of the previous algorithm, but in the case of not being able to extract the text on the product or the case of not having text on it, unfortunately, we return to the procedure of the previous algorithm using only the distance between text and product.

As we can see the two previous methods have a problem with sales flyers when there are many products and they are not well separated. To avoid the problem of distances presented in the previous algorithms, we combined CNN and MLP with multiple inputs. This method provides us with better performance where we carry out a heavy sequence argumentation of data which consists of crop + horizontal + flips + Gaussian blur, leading to significant changes in our data. we used Cross-Validation method in which our original data was subdivided into 3 folds.

We observed that the model has a comparable performance on both training and validation datasets where there are 3 different folds, which has its respective training loss validation. We can also see that our model has Good Fit Learning Curves in each of the folds. Both the training and validation loss decrease in an exponential fashion as the number of epochs is increased, suggesting that the model gains a high degree of accuracy as our epochs are increased so, we can conclude by saying that our algorithm has good accuracy.

In this way, we can realize the differences between the three methods, where the similarity of performance is evident when using only the distances to classify the sales flyers, on the other hand, it can be seen that the combination method between Convolutional Neural Network and Multi-Layer Perceptron has a better performance.

Table 3.2 Cross Validation Performance Result of End-to-End Methods.

N° Fold	Method	Precision	Recall	F1 Score	Accuracy
1	Function Algorithm 1	67%	77%	71%	70%
	Function Algorithm 2	70%	79%	74%	73%
	CNN and MLP	78%	85%	81%	80%
2	Function Algorithm 1	65%	73%	69%	67%
	Function Algorithm 2	70%	80%	75%	73%
	CNN and MLP	80%	86%	83%	82%
3	Function Algorithm 1	62%	71%	66%	64%
	Function Algorithm 2	75%	66%	70%	68%
	CNN and MLP	76%	84%	80%	79%
4	Function Algorithm 1	66%	75%	70%	68%
	Function Algorithm 2	68%	77%	72%	71%
	CNN and MLP	80%	86%	83%	82%
5	Function Algorithm 1	67%	77%	71%	70%
	Function Algorithm 2	70%	80%	74%	73%
	CNN and MLP	87%	82%	84%	84%

Table 3.3 Average Performance Result of the End-to-End Methods using Cross Validation.

Methods	Recall	Precision	F1 Score	Accuracy
Function Algorithm 1	65%	74%	69%	67%
Function Algorithm 2	70%	76%	73%	72%
CNN and MLP	80%	85%	82%	81%

As shown in Table 3.2, cross validation performance results of end-to-end methods is reported. To carry out the calculation of end to end results presented, 100 flyers were used, which contained a total of 540 items, our original data testing was subdivided into 5 subsets called folds, for each of these subsets, we evaluate the performance of our model, then proceeded to calculate the variables TP, FP and FN where TP is the items correctly classified, that is, each product has been classified with their respective description, FP is the items that only the product without the total description and FN is the Items that have not been correctly classified, that is, products

that have been classified with the wrong description. We calculated each of the evaluation matrices shown.

In Table 3.3 in table 1, we calculated the average of each of the metrics: precision, Recall, F1 Score and Accuracy for each of the algorithms with respect to the 5 subdivision of the evaluation data used, we can see which CNN method and MLP has a better performance in each of the measurements.

4.CONCLUSIONS

In this thesis we presented an automatic learning algorithm for the information extraction from sale flyers, which consists of 3 steps: The first is the detection of objects (Products) in the sales flyers inspired by the faster region convolutional neural network (Faster R-CNN) is incorporated of two principal parts. One is a deep fully convolutional network that put forward, regions using Deep Residual Learning 101, and the other is Fast Regions-CNN detector that uses the proposal regions. The Resnet model produces the whole image with several convolutional (conv) and max-pooling layers to produce a convolution feature map, then the next process is for each object to generate a product proposal region of interest. Each feature vector is fed on a sequence of fully connected layers that finally are divided into two output layers: one which generates SoftMax probability approximations over k product classes and another layer which outputs four real-valued numbers for each of the k product classes. Our product location model provides us a high precise product location on sales flyers with an Average Precision of 85.6%.

The second step is the detection and extraction of the text in the sale flyers by optical character recognition, which generates two data outputs, one is the location coordinates of the text on the sales flyer and the other is the text string. OCR API provides us a precise detection and high quality in exaction of the detected text. Finally, we presented 3 function algorithms with the aim of joining the products with its respective description. Both of the steps of detecting objects and extracting text from sales flyer provide the inputs for the function algorithms, which are the location of the bounding boxes of texts and objects: The performance of the function algorithm (1) (merge by distances between object and text) achieved 74% Precision, 64% Recall, 69% F1 Score and 67% Accuracy. The function algorithm (2) (merge by distances between Object, text and text on the object) achieved a 76% precision, 70% Recall, 73% F1 Score and 71% Accuracy whiles our main contribution been the combination of Convolutional neural network CNN and Multilayer Perceptron (MLP) achieved a precision of 85%, 80% Recall, 82% F1 Score and 81% Accuracy. Although our third method has an excellent precision performance for information extraction from sale flyers, this method would have limitations when two or more products in the sales flyer have the same name and similar descriptions where only the size of the product will

be different, in this case our classifier might classify the products incorrectly with their respective descriptions, for this reason our future work will consist of recognition of patterns of the sales flyers so that our method can recognize and predict the location of the products and the side where their respective price and description are located, thus avoiding the limitation of lousy classification of the aforementioned method. On the other hand, in the third method of merge algorithms we planning to replace embedding and multilayer perceptron neural network with Long short-term memory (LSTM).

REFERENCES

- [1] Zamberletti A., Gallo I., Noce L., (2016), "Augmented text character proposals and convolutional neural networks for text spotting from scene images", *Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*, (August), 196–200.
- [2] Gallo I., Zamberletti A., Noce L., (2015), "Content extraction from marketing flyers". *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9256, 325–336.
- [3] Calefati A., Gallo I., Zamberletti A., Noce L., (2016), "Using convolutional neural networks for content extraction from online flyers", *DocEng 2016 - Proceedings of the 2016 ACM Symposium on Document Engineering*, 127–130.
- [4] Zhi T., Weilin H., Tong He. P. H., (2018), "Detecting Text in Natural Image with Connectionist Text Proposal Network", *Zhi In Analytica Chimica Acta* (1035).
- [5] Ren S., He K., Girshick R., Sun J., (2017), "DL-物体検出04-2_2017_Faster R-CNN(IEEE)", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.
- [6] Redmon J., Divvala S., Girshick R., Farhadi A., (2015), "You Only Look Once: Unified, Real-Time Object Detection",
- [7] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C. Y., Berg A. C., (2016), "SSD: Single shot multibox detector". *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS, 21–37.
- [8] Zhong Z., Jin L., Zhang S., Feng Z., (2016), "DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images", 1-12.
- [9] He K., Zhang X., Ren S., Sun J., (2015), "Deep Residual Learning for Image Recognition" 770-778.
- [10] Alex K., Ilya S., Geoffrey E. H., (2012), "ImageNet Classification with Deep Convolutional Neural Networks Alex", 60-1-60–16.
- [11] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Fei-Fei L., (2015), "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, 115(3), 211–252.
- [12] Jia D., Wei D., Socher R., Li-Jia L., Kai L., Li F., (2009), "ImageNet: A large-scale hierarchical image database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- [13] Girshick R., Donahue J., Darrell T., Malik J., (2014), "Rich feature hierarchies

for accurate object detection and semantic segmentation", In Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587.

- [14] Ren S., He K., Girshick R., Sun J., (2015), "Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks", 1–14.
- [15] Gorabal J., (2017) "Text Detection and Extraction from Images –A Comprehensive Survey", *Int. J. Mod. Trends Eng. Res.*, 4, 4, 95–98.
- [16] Zhang Z., Zhang C., Shen W., Yao C., Liu W., Bai X., (2016), "Multi-Oriented Text Detection with Fully Convolutional Networks", 4159-4167.
- [17] Huang W., Lin Z., Yang J., Wang J., (2013), "Text Localization in Natural Images using Stroke Feature Transform and Text Covariance Descriptors", 2013 IEEE International Conference on Computer Vision, 1241–1248.
- [18] Chen H., Tsai S. S., Schroth G., Chen D. M., Grzeszczuk R., Girod B, (2011), "Edge-Enhanced Maximally Stable Extremal Regions", 2657–2660.
- [19] He W., Fei X. Z., & Liu Y. C., (2017), "Deep Direct Regression for Multi-Oriented Scene Text Detection", 745-753 .
- [20] Yao C., Bai X., Liu W., Ma Y., Tu Z., (2012), "Detecting Texts of Arbitrary Orientations in Natural Images", 8, 1083–1090.
- [21] Kang L., Li Y., Doermann D., (2014), "Orientation Robust Text Line Detection in Natural Images", 4034-4041 .
- [22] Bartz C., (2017), "STN-OCR: A single Neural Network for Text Detection and Text Recognition", 1-9.
- [23] Nagaoka Y., Miyazaki T., Sugaya Y., Omachi S., (2017), "Text Detection by Faster R-CNN with Multiple Region Proposal Networks". 15–20.
- [24] Abitha A., Assistant L. K., (2018), "A Faster RCNN Based Image Text Detection and Text to Speech Conversion". 5(5), 11–14.
- [25] Liu, Z., Zhou, W., & Li, H. (2019). Scene text detection with fully convolutional neural networks. *Multimedia Tools and Applications*, 1-23
- [26] Qin H., Zhang H., Wang H., Yan Y., Zhang M., Zhao W., (2019), "An Algorithm for Scene Text Detection Using Multibox and Semantic Segmentation", *Applied Sciences*, 9(6), 1054.
- [27] Szeliski R., (2010), "Computer Vision: Algorithms and Applications", 1–25.
- [28] De Vylder J., Donne S., Luong H., Philips W., (2016), "Image Restoration Using Deep Learning", *Benelux Bioinformatics Conference*, 2011–2013.
- [29] Srivastava N., Hinton G., Sutskever Krizhevsky A., Ilya Salakhutdinov R.,

- (2017). "Dropout: A Simple Way to Prevent Neural Networks from Overfittin". 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering, ICAEES 2016, 15, 520–525.
- [30] Flach P. A., (2003), "The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics", Proceedings, Twentieth International Conference on Machine Learning, 1, 194–201.
- [31] Girshick R., (2015), "Fast R-CNN", Proceedings of the IEEE International Conference on Computer Vision, 2015 International Conference on Computer Vision, ICCV 2015, 1440–1448.
- [32] Bergeron B. P., (2009), "Optical character recognition", Postgraduate Medicine, 103(3), 37–39.

BIOGRAPHY

Harlinton Palacios Mosquera was born in 1992 in Canton del San Pablo, Colombia. from 2010 to 2015, he studied Computer Engineering at the University of Chocó “Diego Luis Córdoba”, Colombia. In 2017, he started his masters studies at Gebze Technical University in Computer Engineering.