ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

DEMOKRITOS

AUTOMATED INFORMATION EXTRACTION FROM WEB PAGES

by

Stergios Giannios

Submitted
in partial fulfilment of the requirements for the degree of
Master of Artificial Intelligence
at the
UNIVERSITY OF PIRAEUS

Athens, February 2022

Author . . . . Stergios Giannios. . . . .

II-MSc "Artificial Intelligence"
February 20, 2022

Γ. Πετάσης

Certified by . . . .

Georgios Petasis,
Researcher,
Thesis Supervisor

Certified by . . . .

Georgios Giannakopoulos,
Researcher,
Member of Examination
Committee

Certified by . . . .

Theodoros Giannakopoulos,
Researcher,
Member of Examination
Committee

AUTOMATED INFORMATION EXTRACTION FROM WEB PAGES

by

Stergios Giannios

Submitted to the II-MSc "Artificial Intelligence" on February 20, 2022,
in partial fulfillment of the
requirements for the MSc degree

Abstract

The continuous growth of the world wide web (WWW) has resulted in enormous amounts of information. Specific data, contained in webpages, can be extracted and leveraged in numerous applications. A semi-automatic/automatic approach to retrieving data from webpages is needed since manual extraction is very time-consuming and does not scale well. However, because of the heterogeneity and semi-structured nature of webpages, the automatic extraction of data is a non-trivial task. The task of web information extraction (WIE) is most commonly addressed with wrapper induction (WI). In WI, the goal is to learn a set of extraction rules by using manually labelled examples. The primary issue with WI is that the learned rules are frequently incapable of dealing with even slight variations in a webpage's template, and cannot generalize to other websites. In this thesis, the WIE problem is reframed as an object detection task. For this purpose, a dataset was built, with news articles that were collected and annotated. A state-of-the-art detector, YOLOv5, was used to extract specific attributes such as the news articles' title, metadata, author, date, main image, text, and keywords. The model yielded 90% mAP (over all classes) in stratified (based on website domain) 5-fold cross-validation. One-shot learning capabilities of the model were also explored by using transfer learning to fine-tune the model to unseen news websites in English but also in another language (Greek) achieving 79% mAP and 90% mAP respectively. A dataset with books' product details from Amazon.in, with extracting targets the books' title, author, and price was used to compare our approach with a state-of-the-art approach where a previous version of YOLO (version 2) was utilized. The mAP of our approach yielded 95% mAP compared to the state-of-art approach which yielded 74% mAP.

Acknowledgments

I would like to thank my supervisor, Dr. Georgios Petasis, for his guidance and feedback throughout the process of this thesis. I would also like to thank my family for their continuous support throughout the years.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Description

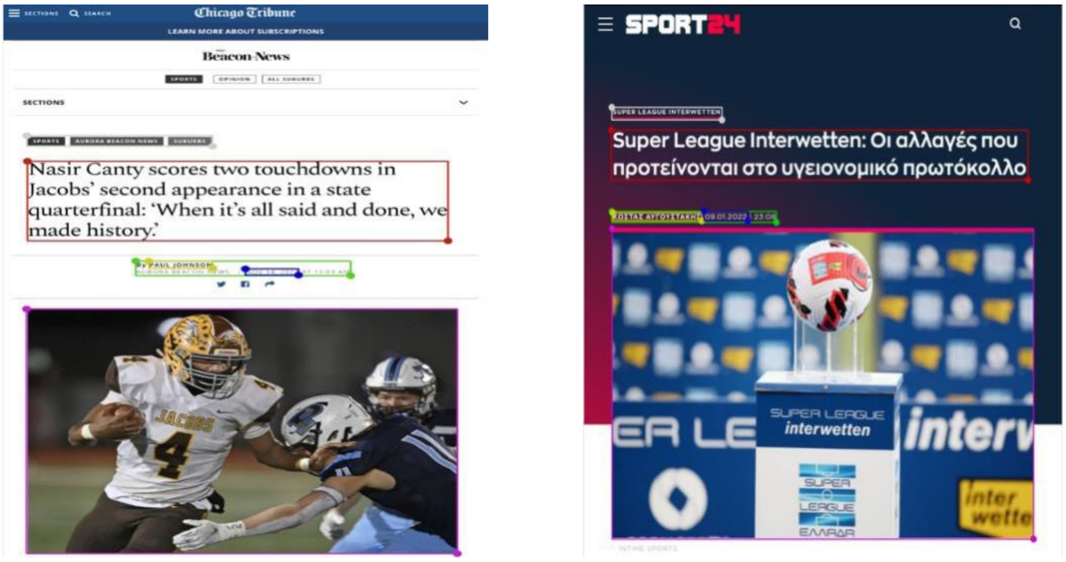
The continuous growth of the world wide web (WWW) has resulted in enormous amounts of information. Specific data, contained in webpages, can be extracted and leveraged in numerous applications such as products price comparison, vacancies monitoring, and news aggregation just to name a few. A semi-automatic/automatic approach to retrieving data from webpages is needed since manual extraction is very time-consuming and does not scale well. However, because of the heterogeneity and semi-structured nature of webpages, the automatic extraction of data is a non-trivial task. Moreover, the main components of the WWW are HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript which are more user-centric (designed for presentation purposes) and not easily processable by machines.

The task of Web Information Extraction (WIE) that is explored in this thesis differs from traditional Information Extraction (IE) tasks, in that in traditional IE, data is mainly extracted from unstructured free text whereas in WIE data is extracted from semi-structured webpages that are usually generated in an automated way by a server-side application. Recent research in IE leverages neural models for extracting entities and relations from plain texts. However, these methods do not work well on webpages which contain both free text and markup information [1].

In the bibliography, the WIE task is tackled with various approaches. Based on the type of information that is used to extract the data, these methodologies can be Document Object Model (DOM)-based, text-based, vision-based, or a combination of the aforementioned methods. Programs that perform the task of WIE are called extractors or wrappers [2]. A pattern matching method is generally implemented by a wrapper using a set of extraction rules [3]. Most commonly, WIE is addressed with Wrapper Induction (WI). In WI the goal is to learn a set of extraction rules by using manually labeled examples. The primary issue with WI is that the learned rules are frequently incapable of dealing with even slight variations in a webpage's template, and can not generalize to other websites [4].

In recent years, there has been a vast development of neural network architectures for computer vision (CV) tasks with immense success. In this thesis, the problem of WIE is formulated as an object detection task where the desired attributes or blocks of information can be identified and located within the webpage. Intuitively, we try to mimic the way humans detect objects based on appearance and context (e.g., the news articles' title is usually written in black bold font and positioned above the news articles' main image). Moreover, these visual features are somewhat independent of the websites' language as illustrated in Figure 1.

For this purpose, 728 English news webpages from 31 different domains as well as 45 Greek news webpages from 2 different domains were collected and annotated. The desired classes/attributes that were targeted for extraction were the news articles' title, metadata, author, date, main image, text, and keywords. The one-stage object detector YOLOv5 was utilized which is the latest iteration of the YOLO family of detectors and yielded 90% mAP (over all classes) in stratified (based on domain) 5-fold cross-validation. One-shot learning capabilities of the model were also explored by using transfer learning to fine-tune the model to unseen news websites in English but also in another language (Greek) achieving 79% mAP and 90% mAP respectively. A dataset with books' product details from Amazon.in, provided by the authors of [5], with classes title, author, date was used to compare our approach with a state-of-the-art approach where a previous version of YOLO (version 2) was utilized. The mAP of our approach yielded 95% mAP compared to the state-of-art approach which yielded 74% mAP.



Nasir Canty scores two touchdowns in Jacobs' second appearance in a state quarterfinal: 'When it's all said and done, we made history.'

Super League Interwetten: Οι αλλαγές που προτείνονται στο υγειονομικό πρωτόκολλο

Figure 1. On the left, an example of the top segment of the news webpage (https://www.chicagotribune.com/) in the English language. On the right an example of the top segment of the news webpage (https://www.sport24.gr/) in the Greek language. The bounding boxes colors correspond to the following blocks: red: title, green: metadata, yellow: author, blue: date, purple: main image, gray: keywords.

1.2 Thesis Structure

The rest of the thesis is organized as follows. In the second chapter, the theoretical background regarding object detection is given. Specifically, a high-level overview of how one-stage-detectors such as YOLO and two-stage-detectors such as R-CNN work and have evolved during the recent past years as well as the most common evaluation metrics for object detection are presented. Related work is also discussed in this chapter. In the third chapter, we discuss about the dataset collection process, the annotation strategy, the architecture of the YOLOv5 model that was utilized as well as the training process. In the fourth chapter, the experimental results are presented and evaluated. Lastly, we conclude our research with a brief summary of our contributions and give some directions and thoughts for future work.

2 Background

2.1 Evolution of Object Detection Techniques

Object detection is a challenging and fundamental problem in the field of computer vision and has been a topic of research for quite some time. With object detection, the aim is to examine if there are any instances of objects from specific predefined classes (such as vehicles, animals, etc.) in an image and, if there are, to return the specific location of each object (e.g., with a bounding box). Object detection has now been successfully utilized in many real-world applications, such as robot vision, video surveillance, and autonomous driving just to name a few.

Hand-crafted features (e.g., HOG, SIFT) [6] were mostly used for the early object detection algorithms. However, the performance of these methods required significant effort from domain experts and reached a plateau after 2010. R. Girshick et al. [6] say: "... progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods". Most recently, deep learning methodologies have been used for learning feature representations automatically from the data. Krizhevsky et al. [7] proposed a Deep Convolutional Neural Network (DCNN) which they named AlexNet. The model achieved significantly higher image classification accuracy in the Large-Scale Visual Recognition Challenge (ILSVRC). Deep learning methods have, since that time, predominately been the research focus for most computer vision-related tasks. R-CNN [6] combines region proposals with CNNs and was one of the first successful integrations of CNNs in an object detection pipeline. The successes of those deep learning models can be mostly attributed to the readily available, huge amounts of training data and also the increasingly powerful computing capabilities (GPUs especially) that enable the training of large networks with millions or even billions of parameters.

In the deep learning era, object detection algorithms can be divided into two categories: "two-stage detection" and "one-stage detection" [8]. In general, two-stage detectors achieve better classification accuracy and localization whereas one-stage detectors are very efficient and therefore faster. We will give a high-level overview of how two representative detectors (RCNN for two-stage detection and YOLO for one-stage detection) work.

2.2 R-CNN Family of Detectors

2.2.1 R-CNN

R-CNN [6] was one of the earliest examples of convolutional neural networks being successfully applied to the problem of object detection, and it served as a foundation for the creation of additional advanced object detection systems.

The R-CNN architecture consists of four components as illustrated in Figure 2:

- *Regions of interest (RoI) component.* The first component extracts class-independent region proposals. Regions that have a high likelihood of containing an object are extracted using an algorithm named selective search [9]. Since the proposed RoIs may vary in size, they are resized to acquire predefined fixed dimensions; since CNNs can only process images with a fixed input size.
- *Feature extraction component.* A pretrained CNN (e.g., AlexNet on ImageNet dataset) is commonly used to extract features from each RoI.
- *Classification component.* A set of class-specific linear SVM classifiers are trained based on the extracted features from the previous component.
- *Bounding-box (BB) regressor component.* This component is responsible for predicting and refining the bounding box's location as well as size for each object. Four values are predicted by the regressor; (x,y): coordinates of the center of the box and (h,w): height and width of the box.

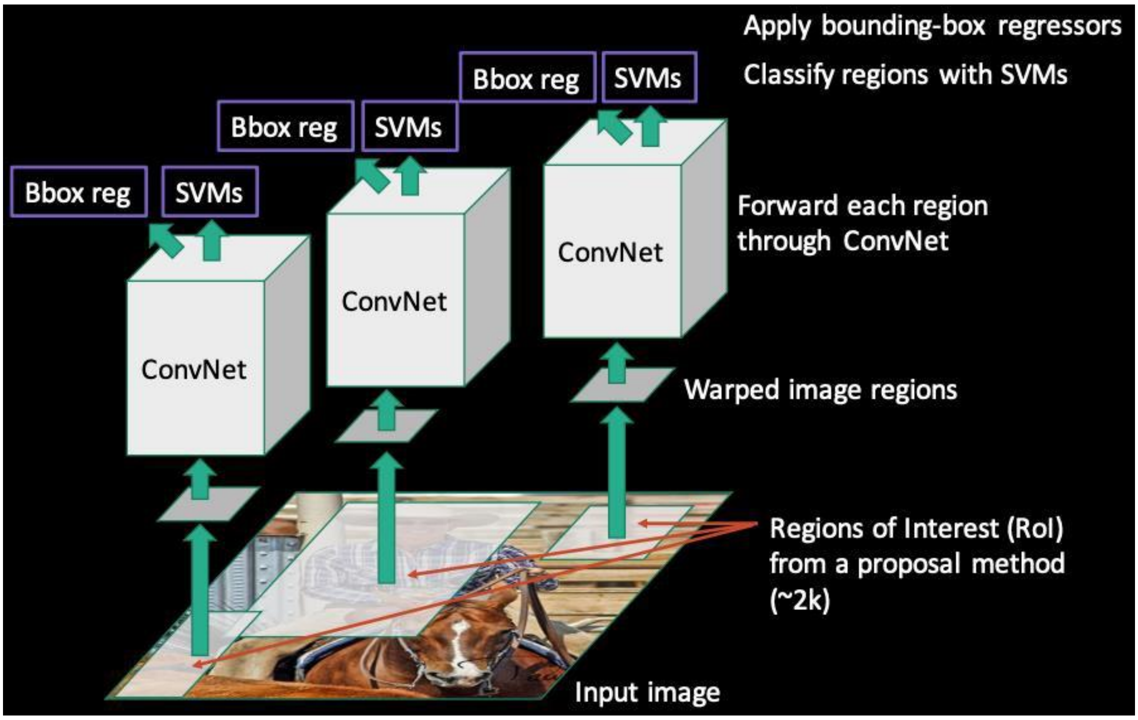

Apply bounding-box regressors
Classify regions with SVMs
Bbox reg
SVMs
Bbox reg
SVMs
Bbox reg
SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

Figure 2. R-CNN architecture [10].

Disadvantages of R-CNN:

- Training is a multi-phase pipeline. Three separate training components (CNN, SVMs, BB regressor) make the training process hard to be optimized in an end-to-end manner.
- Training is computationally expensive both in time as well as in space.
- Very slow object detection since for each image the selective search algorithm proposes about 2,000 RoIs to be furtherly processed.

2.2.2 Fast R-CNN

Fast R-CNN [11] introduced two novel ideas, increasing both detection speed and accuracy:

- Application of the CNN feature extractor to the entire image first and then proposal of regions. By doing so, the CNN is utilized only once instead of running over each of the 2000 candidate regions.
- Replacement of the SVMs with a softmax layer. This way, the CNN performs both feature extraction as well as object classification.

The Fast R-CNN architecture consists of the following components as shown in Figure 3:

- *Feature extraction component.* To extract features from the whole image, a pretrained CNN is used.
- *RoI extraction component.* 2000 region candidates are proposed per image via the selective search algorithm.
- *RoI pooling layer component.* Extracts fixed-size segments (since fully connected layers require fixed-size input) from the proposed regions by applying max pooling.
- *Two-head output layer.*
    - Softmax classifier for classification. Predicts class probabilities.
    - BB regressor for localization. Predicts offsets from the original RoI.

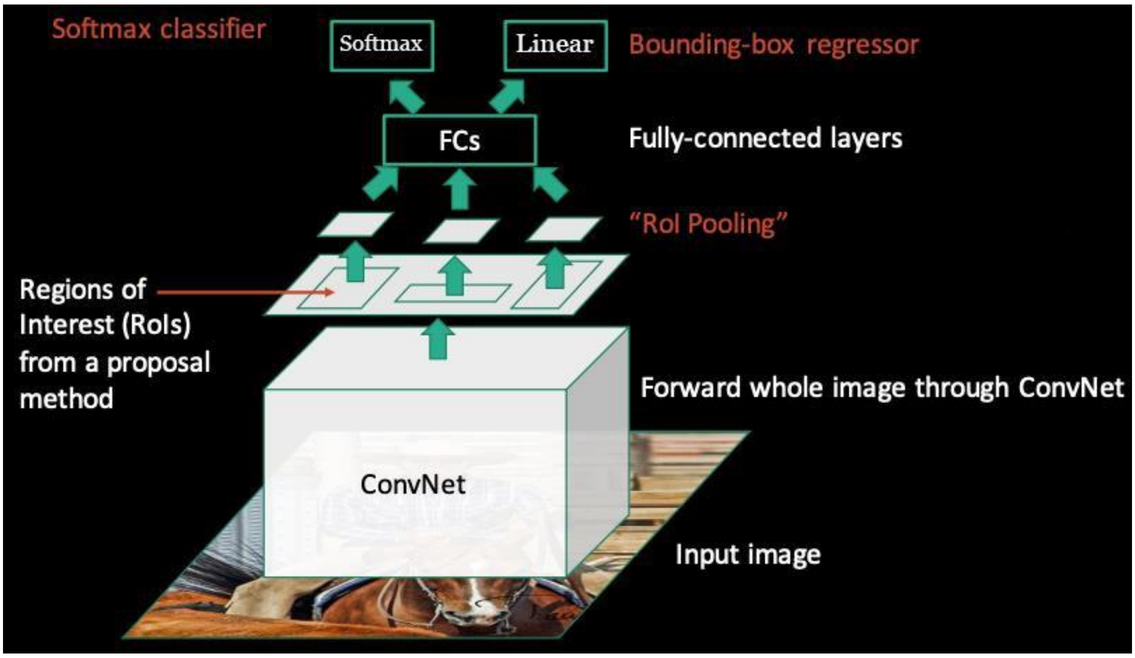


Figure 3. Fast R-CNN architecture [10].

Fast-RCNN is much faster in regards to inference time since the CNN is only run once per image. Training is faster as well because of the unification of many of the components to a single CNN. The only bottleneck that remains is the proposal of candidate regions which are produced by a separate model.

2.2.3 Faster R-CNN

In Faster R-CNN [12] the selective search algorithm is substituted with a region proposal network (RPN) that is also part of the same network, resulting in a fully trainable end-to-end deep learning object detection system.

The Faster R-CNN architecture consists of the following networks as shown in Figure 4:

- *RPN.* Proposes RoIs by using the last feature map provided by the feature extractor. Two outputs are produced by the RPN: the objectness score (indicates if an object is present) and the box location.
- *Fast R-CNN.* It consists of the same aforementioned components of the Fast-RCNN architecture.

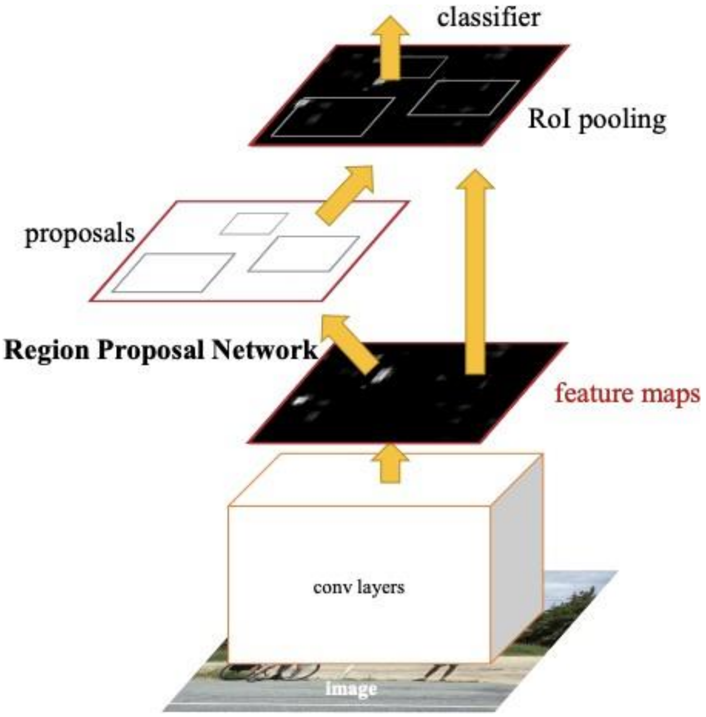


Figure 4. Faster R-CNN architecture [12].

A novel method for multi-scale object detection was also introduced with no need for multiple scales of input images or feature maps. This is accomplished, by using predefined boxes as references called anchor boxes. The BB regressor will then predict offsets from these boxes adjusting the anchor boxes to better fit the objects as illustrated in Figure 5. By using a sliding window, the RPN generates k (hyperparameter) anchor boxes for each location in the feature map. The anchors are at the center of their respective windows and vary in scale and aspect ratio to support a wide range of objects. The anchor boxes that are used are dataset specific and hence should be treated as hyperparameters. The authors of Faster-RCNN used nine anchor boxes with three different aspect ratios and three different scales.

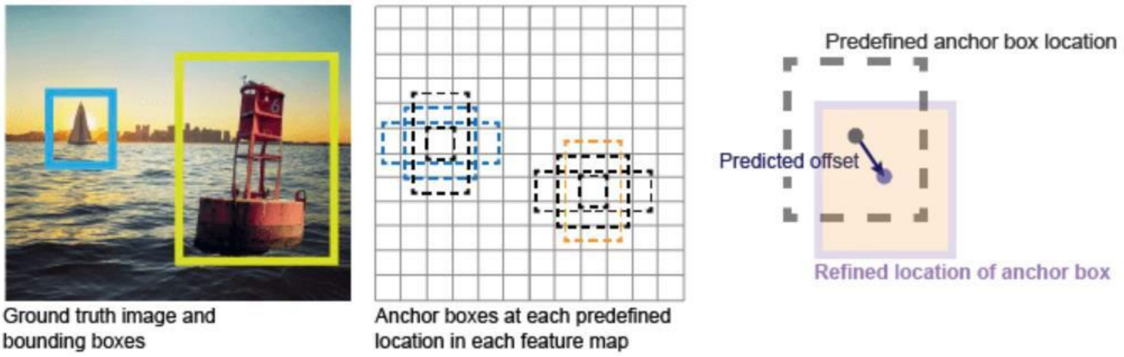

Ground truth image and bounding boxes
Anchor boxes at each predefined location in each feature map
Predefined anchor box location
Predicted offset
Refined location of anchor box

Figure 5. Offset coordinates predictions from initial anchor box [13].

2.2.4 Limitations of R-CNNs

Great improvement in performance both in object detection speed and accuracy has been shown with each new iteration of the R-CNN family of detectors. However, these approaches have drawbacks such as that training consists of multiple phases (region proposal and detection) which makes the network hard to optimize, and training and inference time is quite slow. Fortunately, one-stage detectors such as the YOLO family of detectors have addressed some of these challenges with varying degrees of success with each iteration.

2.3 YOLO Family of Detectors

YOLO detectors reframe the object detection task as a regression problem by localizing and classifying objects with a single end-to-end neural network which can be more easily handled, optimized, and is significantly faster. Furthermore, when making predictions, YOLO detectors consider the image as a whole.

Contrary to region proposal and sliding window-based approaches, YOLO detectors examine the entire image during training and inference, which allows them to encode contextual information about each object implicitly. The accuracy of YOLO models is close to that of R-CNN and the detection speed is significantly faster [8].

2.3.1 YOLOv1

YOLOv1 [14] splits the input image into a grid of S x S cells, each of which is responsible for identifying an object if the center of the bounding box of that object falls into that grid cell. Each grid cell predicts:

- *Coordinates of B bounding boxes.* (x,y): coordinates of the center of the box and (h,w): height and width of the box.
- *Objectness score.* The probability of an object being present in the cell. The objectness score is normalized as a probability with a value ranging from 0 to 1 using a softmax layer.
- *Class (C) prediction.* The model predicts the probability for K number of classes if the bounding box contains an object, where K is the total number of classes for a given problem.

It is important to note that even though YOLOv1 can predict for each cell multiple boxes and objectness scores, it can only predict one class. The final output of YOLOv1 is a tensor of shape $S \times S \times (5 * B + C)$. Finally, YOLOv1 applies non-maximum suppression (NMS) to reduce the number of candidate boxes to only one bounding box per object. NMS examines all of the overlapping boxes that surround an object to determine which box has the highest prediction probability and suppresses the others. Boxes with an Intersection over Union (IoU) and a confidence loss below an adjustable threshold are discarded as shown in Figure 6.

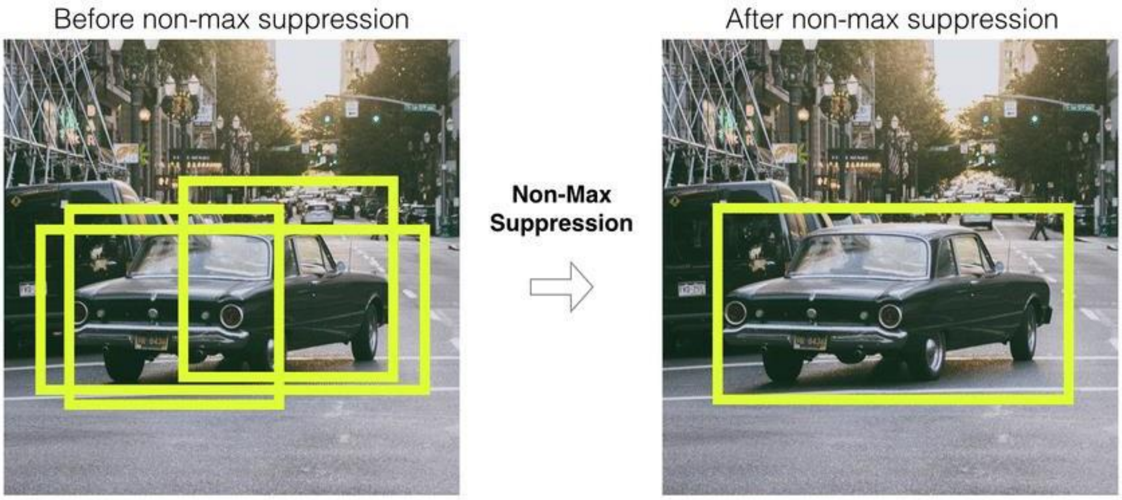


Figure 6. Non-maximum suppression [15].

GoogLeNet [16] inspired the architecture of YOLOv1, as shown in Figure 7. The network uses 24 convolutional layers followed by 2 fully connected layers and instead of inception modules, 1 × 1 followed by 3 × 3 convolutional layers are used. A pretrained CNN (on ImageNet dataset) is fine-tuned at half the resolution and then the resolution is doubled for more accurate detection. The authors called this architecture Darknet.

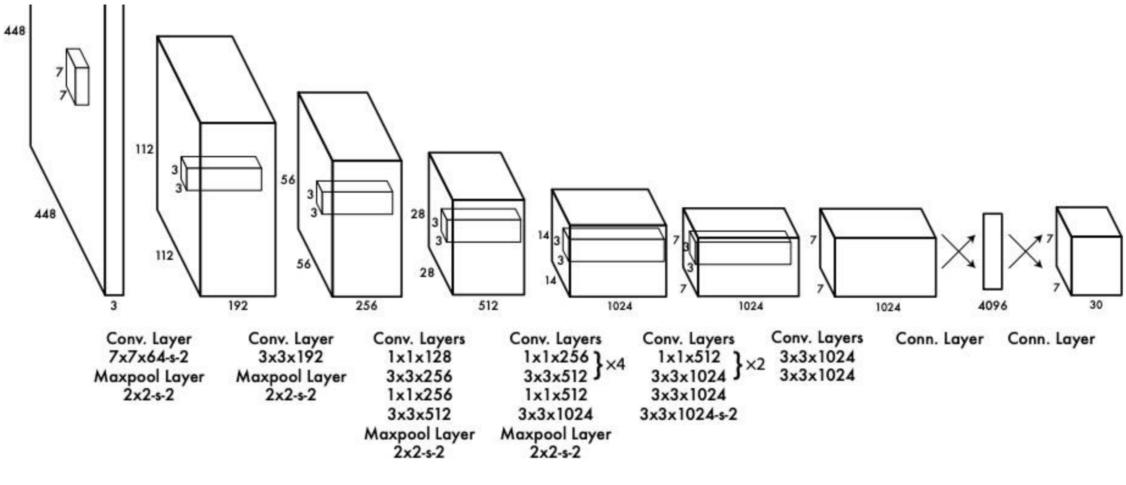


Figure 7. YOLOv1 architecture [14]

2.3.2 YOLOv2

YOLOv2 [17] introduced some novel features to improve performance and speed:

- *Batch normalization [18].* Batch normalization is used to regularize the model and improve its convergence. The addition of batch normalization resulted in a 2% increase in mean Average Precision (mAP).
- *High-resolution classifier.* The original YOLOv1 trains the CNN at 224 × 224 pixels and then the resolution is increased to 448 x 448 pixels for detection. YOLOv2 fine-tunes the CNN for 10 epochs first, using the full 448 × 448 pixels resolution. This gives the network time to adapt to higher resolution instead of suddenly switching to it for the detection phase. The addition of the high-resolution classifier improved the mAP by 4%.
- *Predictions with anchor boxes.* Rather than predicting the coordinates of the bounding boxes directly, anchor boxes are used in a similar manner as in Faster R-CNN. Predicting offsets rather than coordinates makes the task easier and allows the network to learn quicker. Instead of manually selecting priors, k-means clustering is utilized on the bounding boxes of the training set to automatically identify suitable priors. Multiple classes can now be predicted by the same cell.

2.3.3 YOLOv3

YOLOv2 often failed to predict accurately small objects. This was attributed to the successive downsampling of the input which led to the loss of fine-grained features. YOLOv2's architecture was still missing some of the most significant features that were part of most state-of-the-art algorithms: no skip connections, no residual blocks, and no upsampling. YOLOv3 [19] integrates all of these modifications allowing predictions at three different scales per cell by upsampling and concatenating feature maps at different stages within the network as shown in Figure 8. The model used Darknet-53 architecture with a 53-layer network for training the feature extractor. For the detection phase, an additional 53 layers are added, giving a total of a 106-layer fully convolutional model.

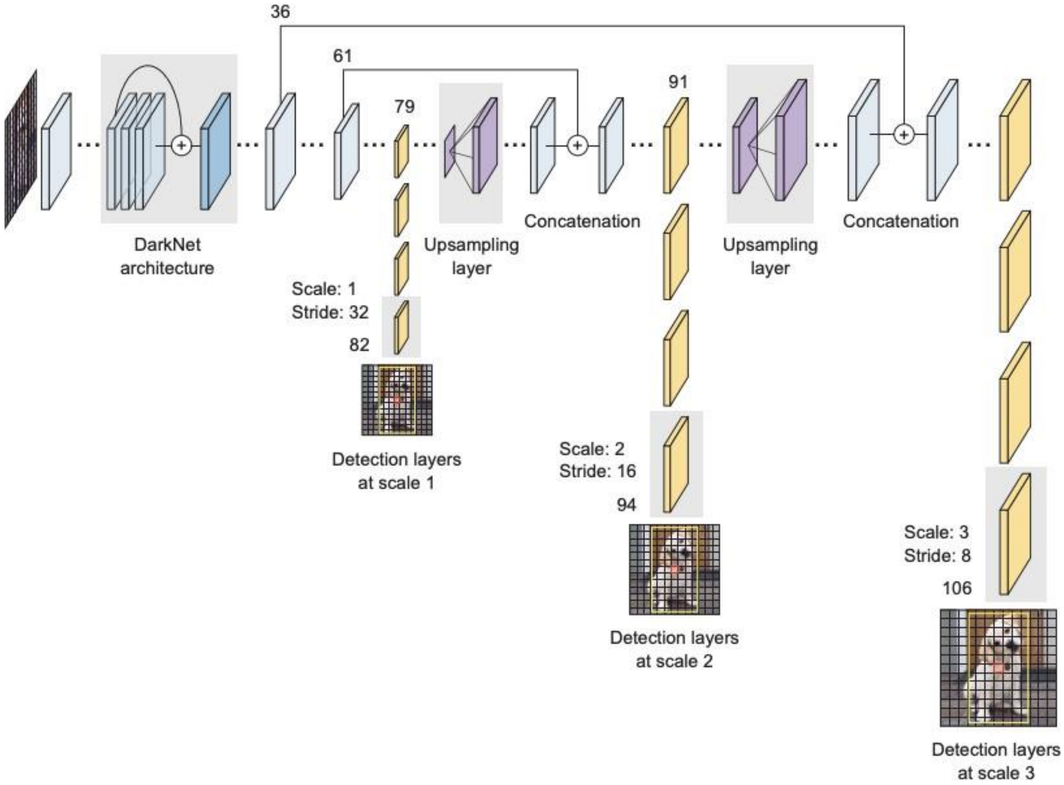


Figure 8. YOLOv3 architecture [20].

2.3.4 YOLOv4 and YOLOv5

The backbone of YOLOv4 [21] is CSPDarknet53. CSP stands for Cross Stage Partial and is derived from DenseNet [22]. DenseNet was designed to connect layers in a very deep neural network in order to solve the problem of vanishing gradients (as ResNet [23]). CSP preserves fine-grained features through forward propagation, enables the network to reuse features, and reduces the number of network parameters. A spatial pyramid pooling [24] (SPP) block is added since it considerably improves the receptive field and separates the most important contextual features. PANet [25] is used for the aggregation of features from different backbone levels for multi-scale detection.

The architecture of YOLOv5 [26] closely resembles that of YOLOv4 but is considerably faster. Moreover, YOLOv5 is written in Python using the PyTorch framework which makes it more accessible. Specific details about the architecture of YOLOv5 will be discussed in the third chapter.

2.4 Evaluation Metrics for Object Detection

Object detector performance is most commonly measured using two metrics: mean average precision (mAP) for detection accuracy and frames per second (FPS) for detection speed. The mAP is a percentage raging from 0 to 100, with a higher percentage indicating better detection performance. The FPS metric is used to measure how many frames can be inferenced in a second with higher values indicating faster detection. IoU measures the overlap between two bounding boxes: the ground truth bounding box and the predicted bounding box. Based on a threshold, the IoU score is then used to determine whether a detection is valid (True Positive (TP)) or not (False Positive (FP). This threshold is tunable but a value of 0.5 (a value greater than 0.5 detonates TP) is commonly used. In Figure 9 examples of various IoU values are presented.

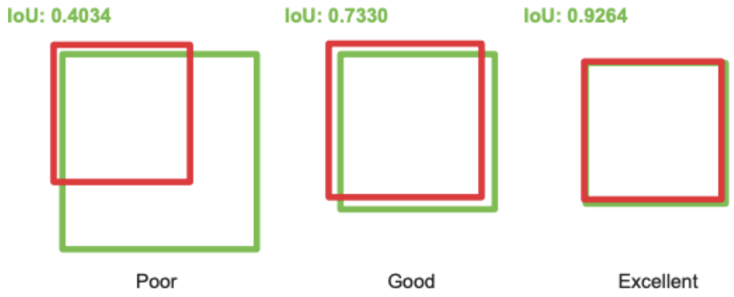


Figure 9. IoU examples. The higher the overlap, the bigger the IoU value [27].

Using the IoU to define TP and FP, precision and recall can be calculated as follows:

$$Precision = \frac{TP}{(TP + FP)} \text{ [8]}$$

$$Recall = \frac{TP}{(TP + FN)} \text{ [8]}$$

The PR curve can be drawn when the precision and recall for all classes has been calculated. Computing the area under the curve (AUC) yields the average precision (AP). The mAP is the average of all the APs obtained across all classes.

2.5 Related Work

The first WIE systems utilized manual wrappers (e.g., TSIMMIS [28], Web-OQL [29]) to extract structured information. A pattern matching is usually performed by a wrapper using a set of extraction rules. However, this approach is hard to scale and maintain since programming expertise would be required to write custom scripts for each website or even each webpage. For these reasons, WIE has been mainly addressed with WI. WI methods (e.g., WHISK [30], STALKER [31]) take as input labeled webpages with examples of the data to be extracted and produce a wrapper. Thus, the user interaction has gone from writing custom extraction rules to labeling specific fields that are to be extracted. However, the main criticism of WI systems is that they tend to break if the webpages change, and that they cannot generalize to different webpages of the same domain (e.g., news websites). Efforts have been made in wrapper induction methods to induce extraction rules that are more robust to small changes, but with not much success [4].

Features of web documents that can be used for the WIE task are:

- *Text-based features* e.g., RAPIER [32].
    - Extracting data from the HTML source code by, for example, string or regex-based delimiters.
    - Punctuation marks, frequency of links e.g., when compared to typical texts, advertisements have a higher link distribution but a lower punctuation mark rate [33].
- *DOM-based features* e.g., Thresher [34]. The DOM is a tree structure that allows the selection of elements, based on their relative positions to other elements of the DOM (e.g., parent-of or sibling-of).
- *Visual-based features* (e.g., VIPS [35]). Exploiting visual features, such as borders, colors, font size, etc. to identify elements.

It's important to note that most WI approaches use text and/or DOM information, making the assumption that the semantics of the webpage are accurately depicted by the HTML/DOM structure which is not always the case.

Most recently, machine learning approaches have been used to tackle the WIE task. The authors of [36] use multiple features such as fonts, links, and positions by utilizing the DOM tree node properties to train a machine learning model in order to extract the main content and remove other elements such as branding banners, navigational elements, advertisements, copyright, etc. The authors of [1] create a two-stage neural approach named FreeDOM. The first stage combines text and markup information to learn a representation for every DOM node in the webpage. Using a relational neural network, the second stage enables the capturing of longer-range semantic relationships. The authors of [5] use YOLOv2 to extract product information of books from Amazon.in such as the books' author, title, and price. They then use the tesseract optical character recognition (OCR) engine to convert the extracted segments into a machine-readable format.

3 Methodology

3.1 Datasets

To the best of our knowledge, there is no large-scale dataset for WIE that is annotated for object detection. For that reason, we created a new dataset based on screenshots from news articles. NewsCatcher API [37] is used to collect links from news articles written in English. Keywords are provided to the API to search articles that contain these terms in either the title or the summary. We used simple/general keywords such as ['the', 'a', 'to', 'and'] which are more likely to be in the title or the summary of the articles. The links were aggregated from 336 different news sources which were published up to 1 week before the API call. We collected 50000 links with this method. We removed duplicate links and then selected 92 sources; each of which had at least 5 links associated with it.

To get the screenshots, Selenium [38] was used. Selenium provides a simple API via a python library to control well-known web browsers e.g., Chrome, Firefox, and allows for web browser automation. A headless chrome driver is used for better performance; meaning that the browser runs in the "background" without a user interface. In order to render most of the websites correctly, to get a screenshot, we have to bypass pop-ups (e.g., cookies, advertisements). For that purpose, we wrote custom scripts for each website that it was required, to simulate the action of a user closing the pop-ups. Finally, once the content was loaded a screenshot was taken and saved as a PNG file. 10000 screenshots were saved from 92 different sources. From the 10000 screenshots, 728 screenshots from 31 different sources were selected and annotated. Domains with different visual features were prioritized in the selection process in order to capture as much variation as possible.

Furthermore, to test the capabilities of our model for transfer/one-shot-learning for news articles written in a different language we manually collected 45 links from 2 sources of news articles written in Greek. We saved the screenshots using the same method as mentioned above.

To compare our proposed approach with a state-of-the-art approach, the authors of [5] provided us with a subset of the dataset that they used for their approach. The dataset includes 138 screenshots of books product details from amazon.in. The images had no labels so we had to annotate them as well.

3.2 Annotation Strategy

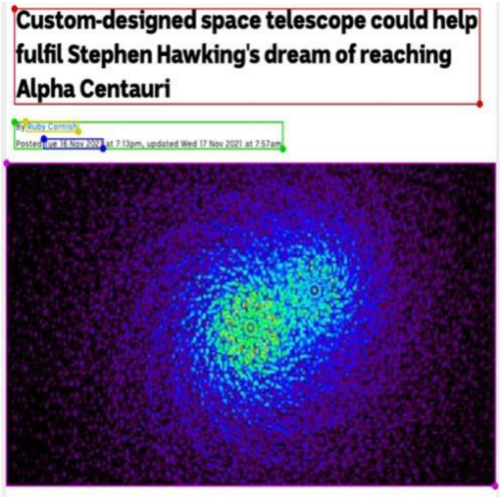
The datasets were annotated by drawing bounding boxes around each object within each image using the open-source annotation tool Labelme [39].

The labels for the news article dataset are title, metadata, author, date, main image, text, and keywords. Annotated examples from the news article dataset are presented in Figure 10. Here are some metadata and rationale for labeling each attribute:

- *Title.* The title is almost always positioned on the top segment of the page and above the main image. There may be metadata or keywords above or below the title. One very visually discriminative feature of the title is that the text is usually in bold, most specifically in black color.
- *Metadata.* The metadata block includes information about the article. It usually also includes the date and author blocks. It is usually positioned on the top segment of the page, most commonly below the title and above the main image.
- *Author.* One or more authors of the article can be selected. Positioned usually inside the metadata block.
- *Date.* Only the original publication day of the article is selected. For example, if there is a date related to the latest update, it is not selected. Positioned usually inside the metadata block.
- *Text.* Refers to the main text blocks of the article. Positioned usually below the top segment of the page. Text blocks that include ads or redirect to other articles are not selected if they are visually distinguishable from the

main content and are visually consistent within the same domain. Text blocks are continuous; meaning that if there is an image, a visual separator, an advertisement, or just a large blank block between texts then only the continuous text blocks are selected.

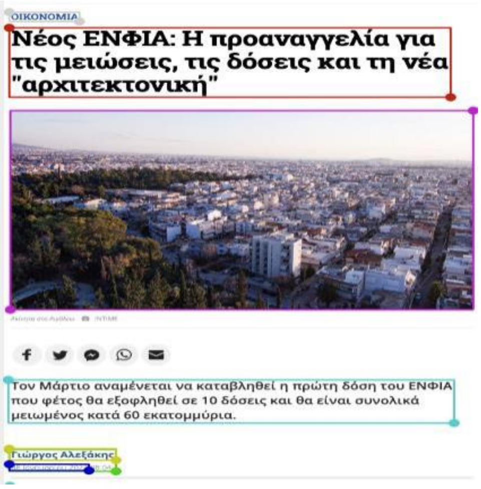
- *Keywords.* Can be at the top or the bottom segment of the page. Keywords are usually enclosed in rectangular boxes.



Custom-designed space telescope could help fulfil Stephen Hawking's dream of reaching Alpha Centauri

By Ruby Cornish

Posted Tue 16 Nov 2021 at 7:13pm, updated Wed 17 Nov 2021 at 7:57am

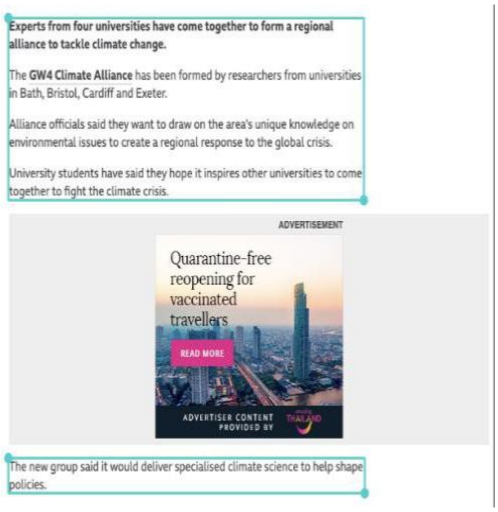


ΟΙΚΟΝΟΜΙΑ

Νέος ΕΝΦΙΑ: Η προαναγγελία για τις μειώσεις, τις δόσεις και τη νέα "αρχιτεκτονική"

Τον Μάρτιο αναμένεται να καταβληθεί η πρώτη δόση του ΕΝΦΙΑ που φέτος θα εξοφληθεί σε 10 δόσεις και θα είναι συνολικά μειωμένος κατά 60 εκατομμύρια.

Γιώργος Αλεξάκης



Experts from four universities have come together to form a regional alliance to tackle climate change.

The **GW4 Climate Alliance** has been formed by researchers from universities in Bath, Bristol, Cardiff and Exeter.

Alliance officials said they want to draw on the area's unique knowledge on environmental issues to create a regional response to the global crisis.

University students have said they hope it inspires other universities to come together to fight the climate crisis.

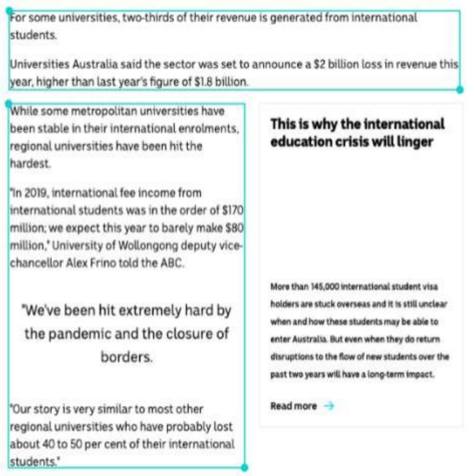
ADVERTISEMENT

Quarantine-free reopening for vaccinated travellers

READ MORE

ADVERTISER CONTENT PROVIDED BY THAILAND

The new group said it would deliver specialised climate science to help shape policies.



For some universities, two-thirds of their revenue is generated from international students.

Universities Australia said the sector was set to announce a $2 billion loss in revenue this year, higher than last year's figure of $1.8 billion.

While some metropolitan universities have been stable in their international enrolments, regional universities have been hit the hardest.

"In 2019, international fee income from international students was in the order of $170 million; we expect this year to barely make $80 million," University of Wollongong deputy vice-chancellor Alex Frino told the ABC.

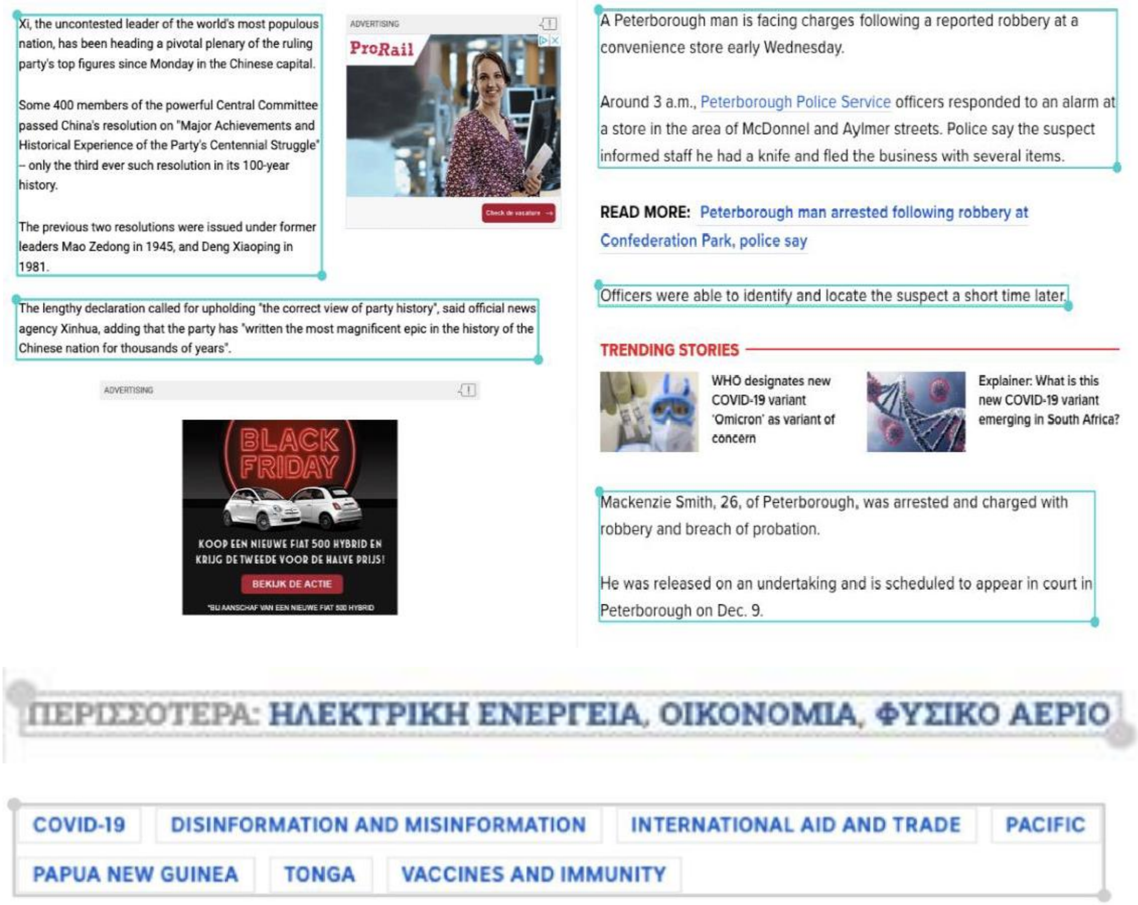
"We've been hit extremely hard by the pandemic and the closure of borders.

"Our story is very similar to most other regional universities who have probably lost about 40 to 50 per cent of their international students."

This is why the international education crisis will linger

More than 145,000 international student visa holders are stuck overseas and it is still unclear when and how these students may be able to enter Australia. But even when they do return disruptions to the flow of new students over the past two years will have a long-term impact.

Read more →



Xi, the uncontested leader of the world's most populous nation, has been heading a pivotal plenary of the ruling party's top figures since Monday in the Chinese capital.

Some 400 members of the powerful Central Committee passed China's resolution on "Major Achievements and Historical Experience of the Party's Centennial Struggle" -- only the third ever such resolution in its 100-year history.

The previous two resolutions were issued under former leaders Mao Zedong in 1945, and Deng Xiaoping in 1981.

The lengthy declaration called for upholding "the correct view of party history", said official news agency Xinhua, adding that the party has "written the most magnificent epic in the history of the Chinese nation for thousands of years".

ADVERTISING
ProRail
Check de vacature →
ADVERTISING
BLACK FRIDAY
KOOP EEN NIEUWE FIAT 500 HYBRID EN KRIJG DE TWEEDE VOOR DE HALVE PRIJS!
BEKIJK DE ACTIE
*BIJ AANSCHAF VAN EEN NIEUWE FIAT 500 HYBRID

A Peterborough man is facing charges following a reported robbery at a convenience store early Wednesday.

Around 3 a.m., Peterborough Police Service officers responded to an alarm at a store in the area of McDonnel and Aylmer streets. Police say the suspect informed staff he had a knife and fled the business with several items.

READ MORE: Peterborough man arrested following robbery at Confederation Park, police say

Officers were able to identify and locate the suspect a short time later.

TRENDING STORIES

WHO designates new COVID-19 variant 'Omicron' as variant of concern

Explainer: What is this new COVID-19 variant emerging in South Africa?

Mackenzie Smith, 26, of Peterborough, was arrested and charged with robbery and breach of probation.

He was released on an undertaking and is scheduled to appear in court in Peterborough on Dec. 9.

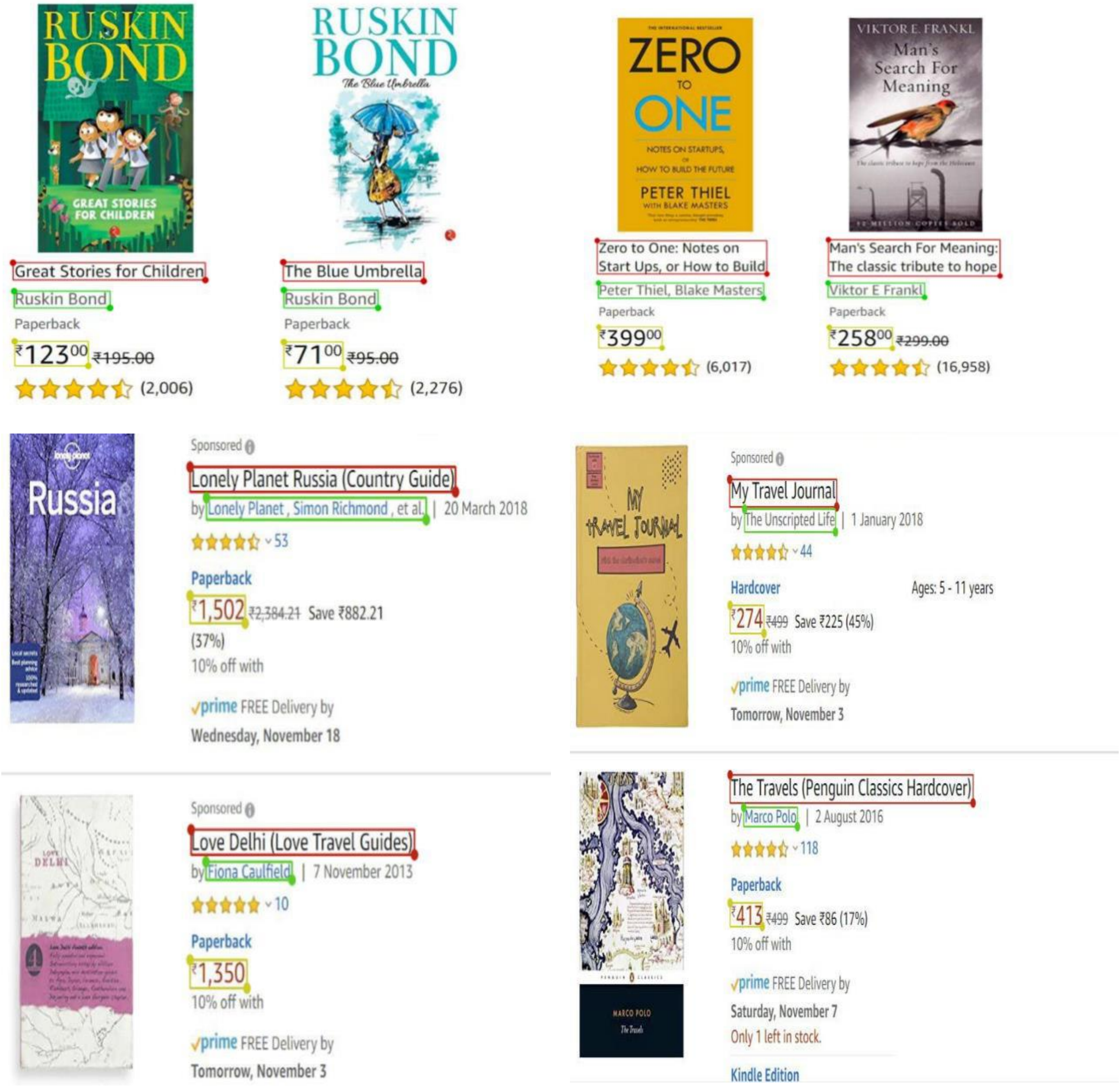
ΠΕΡΙΣΣΟΤΕΡΑ: ΗΛΕΚΤΡΙΚΗ ΕΝΕΡΓΕΙΑ, ΟΙΚΟΝΟΜΙΑ, ΦΥΣΙΚΟ ΑΕΡΙΟ

COVID-19 | DISINFORMATION AND MISINFORMATION | INTERNATIONAL AID AND TRADE | PACIFIC | PAPUA NEW GUINEA | TONGA | VACCINES AND IMMUNITY

Figure 10. Annotated examples from the news article dataset. The bounding boxes colors correspond to the following blocks: red: title, green: metadata, yellow: author, blue: date, purple: main image, cyan: text, gray: keywords.

The labels for the books dataset are title, author, and price. Annotated examples from the books dataset are presented in Figure 11. Here are some metadata and rationale for labeling each attribute:

- *Title.* Usually in black color and positioned below the books' cover image.
- *Author*. One or more authors of the book can be selected. Usually in blue or gray color and positioned between the title and the price.
- *Price.* Usually in black or red color and positioned below the author. One very visually discriminative feature of the price is the Indian rupee symbol inside the block.



Great Stories for Children
Ruskin Bond
Paperback
₹123⁰⁰ ~~₹195.00~~
(2,006)

The Blue Umbrella
Ruskin Bond
Paperback
₹71⁰⁰ ~~₹95.00~~
(2,276)

Zero to One: Notes on Start Ups, or How to Build
Peter Thiel, Blake Masters
Paperback
₹399⁰⁰
(6,017)

Man's Search For Meaning: The classic tribute to hope
Viktor E Frankl
Paperback
₹258⁰⁰ ~~₹299.00~~
(16,958)

Sponsored
Lonely Planet Russia (Country Guide)
by Lonely Planet , Simon Richmond , et al. | 20 March 2018
53
Paperback
₹1,502 ~~₹2,384.21~~ Save ₹882.21 (37%)
10% off with
prime FREE Delivery by Wednesday, November 18

Sponsored
My Travel Journal
by The Unscripted Life | 1 January 2018
44
Hardcover
Ages: 5 - 11 years
₹274 ~~₹499~~ Save ₹225 (45%)
10% off with
prime FREE Delivery by Tomorrow, November 3

Sponsored
Love Delhi (Love Travel Guides)
by Fiona Caulfield | 7 November 2013
10
Paperback
₹1,350
10% off with
prime FREE Delivery by Tomorrow, November 3

The Travels (Penguin Classics Hardcover)
by Marco Polo | 2 August 2016
118
Paperback
₹413 ~~₹499~~ Save ₹86 (17%)
10% off with
prime FREE Delivery by Saturday, November 7
Only 1 left in stock.
Kindle Edition

Figure 11. Annotated examples from books dataset. The bounding boxes colors correspond to the following blocks: red: title, green: author, yellow: price.

3.3 Proposed Approach

3.3.1 Data Augmentation

Data augmentation is used to expand the dataset so that the model can be more robust to image variations. For this purpose, photometric and geometric distortions are commonly used by researchers. We use photometric distortion by altering the saturation, value, and hue of the images. In regards to geometric distortions, we add random scaling and translating. Mosaic data augmentation is also included with the YOLOv5 implementation and combines four training images into one in various ratios enabling the model to learn to identify objects at smaller scales (Appendix A).

3.3.2 YOLOv5 Models

YOLOv5 offers 5 models which are pretrained on the COCO [40] dataset. The difference between those models is the trade-off between the detection accuracy and the detection speed. The most lightweight model, YOLOv5n6, is really fast but not very accurate. On the other hand, the biggest model, YOLOv5x6 is the most accurate model but comes at the cost of slow inference speed. In Figure 12 the detection speed and accuracy of each model on the COCO dataset is presented.

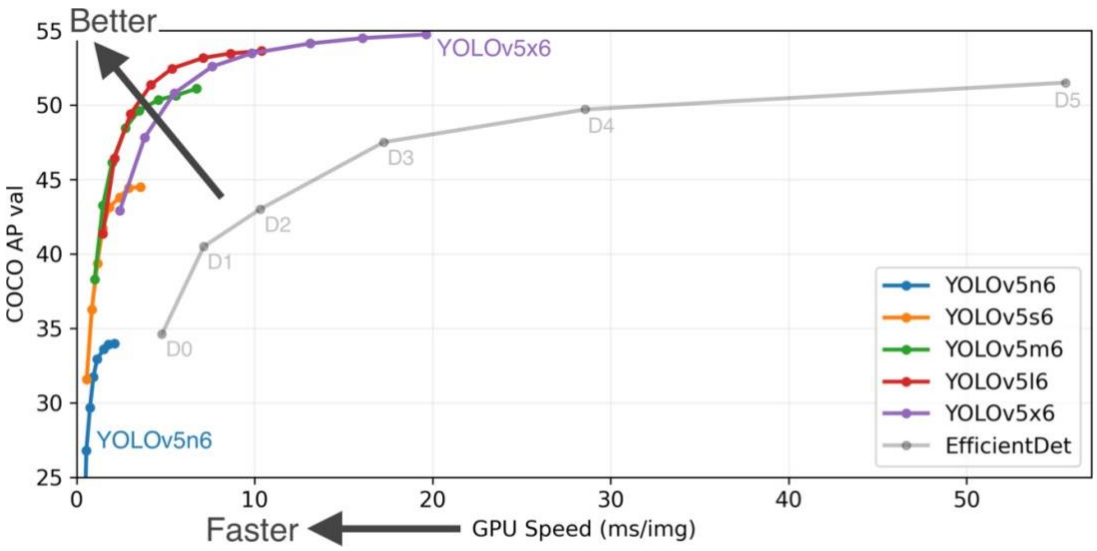


Figure 12. YOLOv5 models. Detection speed and detection accuracy trade-off [26].

3.3.3 YOLOv5 Architecture

The general architecture of YOLO models consists of three components:

- *Backbone component*. Usually, a pretrained CNN is used for extracting features from the image.
- *Neck component.* The feature maps which are extracted by the backbone are aggregated and utilized by the neck. A neck, in most cases, consists of several bottom-up and top-down paths. This component aids in the transmission of small-object information and prevents it from being lost to higher levels of abstraction. This is accomplished by upsampling the resolution of the feature maps, allowing distinct layers from the backbone to be aggregated and influence the detection phase [41].
- *Head component.* Predicts bounding box coordinates and class probabilities.

The YOLOv5s6 model architecture, as well as its modules, are presented in Figure 13. These modules are:

- The *focus module* is the initial layer of the backbone component, and it is intended to reduce the models' calculations. The focus layer acts basically as a SpaceToDepth transformation layer that rearranges blocks of spatial data into depth. The input image is sliced into four segments. Then, the four slices are concatenated in-depth and sent off to the CBL module.
- The *CBL Module* includes a convolution layer, a batch normalization (BN) layer, and a leaky ReLU activation function.
- The *CSP1_X module* is the backbone component's third layer. The CSP1_X and CSP2_X modules are based on the CSPNet [42]. These modules divide the input feature mapping into two parts, and then combine them, reducing the number of calculations. CBL modules and X residual modules (Resunits) are contained in the CSP1_X to extract more effectively deeper features of the image. The value of X corresponds to the number of Resunits.

- The *Res unit module* consists of mainly two CBL modules, and produces its output by adding the output of those modules with the original input
- The *CSP2_X module* is the initial layer of the neck component. The key distinction between it and the CSP1_X block is that the value of X in the CSP2_X block denotes the number of CBL modules.
- The *Spatial Pyramid Pooling (SPP) module* is the backbone component's ninth layer, and it turns a variable-dimension feature map into a fixed-dimension feature map and is intended to improve the receptive field of the model and short out important features from the backbone [24]. The module firstly passes the input through the CBL module. Then, the max-pooling operation is applied vie three parallel maximum pool layers. Then, the output after those operations is concatenated with the feature map outputted after the CBL module. Finally, the output of the concatenation is passed through a CBL module.

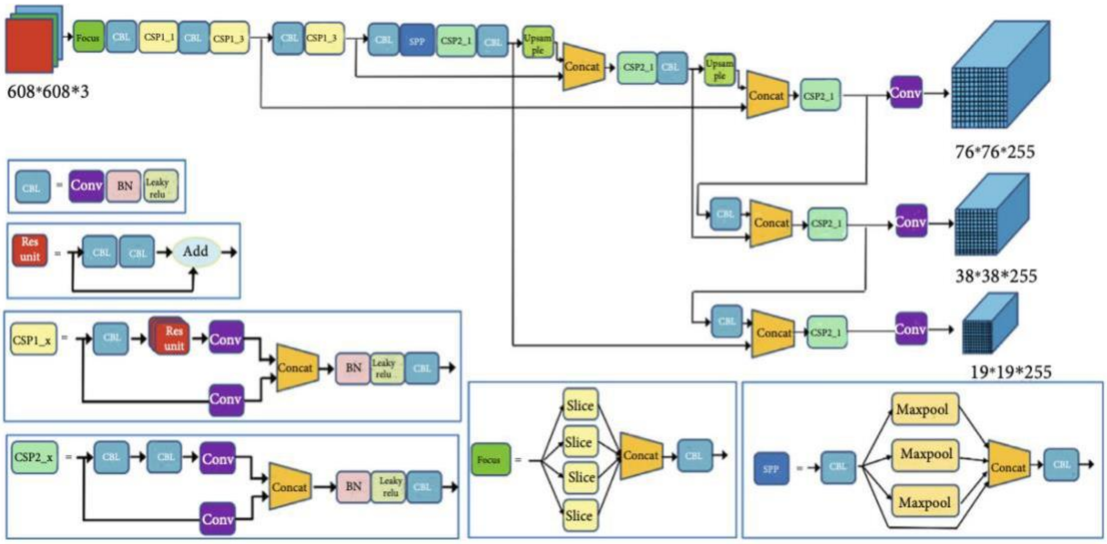


Figure 13. YOLOv5s6 architecture [43].

3.3.4 Training Procedure

It's important to note that the screenshots from the news articles had quite big and variable dimensions. The height ranged from 944 to 17724 pixels and the width was less variable ranging from 800 to 1283. In comparison, the screenshots from the books dataset had lower dimensions and variability. The height ranged from 235 to 1080 pixels and the width was also less variable ranging from 800 to 1283. Basic statistics about the news and the books datasets are shown in Appendix B.

Theoretically, the best detection accuracy would be obtained by using the biggest model and the highest resolution possible for training. However, we were restricted by the GPUs' memory (NVIDIA GTX 1080Ti 11 GB). We used the official pretrained yolov5s6 model on the COCO dataset. The training environment consisted of python 3.9.9 and PyTorch 1.10 [44]. PyTorch is an open-source machine learning library that enables efficient tensor operations via the GPU, supports automatic differentiation, and has been utilized for the implementation of the YOLOv5 models.

We tested all model sizes but with capped resolutions (as much as the GPU's memory allowed). The best results were obtained using the yolov5s6 model with 3520 x 3520 pixels resolution for the news dataset and the yolov5s6 model with 2048 x 2048 pixels resolution (which captures all resolutions) for the books dataset. Stochastic gradient descent is used as the optimizer with an initial learning rate of 0.01, 0.937 momentum, and weight decay of 0.0005. We used the official pretrained yolov5s6 model on the COCO dataset. The epochs, batch size, and patience are set to 5000, 2, 500 for the news dataset and 1000, 5, 100 for the books dataset respectively. In each case, the best model based on validation loss is saved during training. The loss consists of three components: bounding box regression loss, confidence loss, and classification loss.

Wandb.ai [45] is used to track, compare and visualize the experiments. Wandb.ai offers a web interface where one can explore metadata e.g., GPU power/memory usage, hyperparameters of the models, key metrics such as mAP, training/validation losses, and more.

Since the images were mostly rectangular, rectangular training is used. That means that the bigger side of the image will be resized to the image_size that is set and the smaller side will shrink according to the aspect ratio of the image and will get padded to become a squared image of image_size x image_size dimensions. The news dataset was split into two datasets. One dataset contains 687 images from 28 domains and the other 41 images from 3 domains (to test one-shot-learning). For the rest of this thesis, these datasets are referred to as news dataset and English news TF dataset respectively.

4 Results and Discussion

4.1 News Dataset

The model yielded 0.87 mAP with stratified (based on domain) 5-fold cross-validation on the news dataset over all classes. In Table 1, the mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 (0.5 IoU threshold) are presented. The title, main απimage, text, and keywords have a very high mAP (more or equal to 0.91). However, the metadata, author, and date have a lower mAP (less or equal to 0.82). This can be attributed to the fact that the metadata section which usually includes the date and the author blocks is usually quite smaller in comparison to the other classes. Moreover, resizing the input image to a smaller resolution could exacerbate the problem further. A higher resolution could possibly increase the mAP for those classes. The standard deviation for the mAP for all classes except for the author and date is less or equal to 0.02, so the model is quite stable.

	Classes	Precision	Recall	map0.5
0	all	0.92 ± 0.01	0.83 ± 0.02	0.87 ± 0.01
1	title	0.93 ± 0.07	0.93 ± 0.02	0.94 ± 0.01
2	metadata	0.92 ± 0.03	0.73 ± 0.02	0.78 ± 0.02
3	author	0.87 ± 0.03	0.64 ± 0.06	0.69 ± 0.05
4	date	0.92 ± 0.04	0.74 ± 0.05	0.82 ± 0.04
5	main_image	0.95 ± 0.02	0.93 ± 0.02	0.96 ± 0.02
6	text	0.92 ± 0.02	0.90 ± 0.01	0.91 ± 0.02
7	keywords	0.92 ± 0.02	0.93 ± 0.01	0.94 ± 0.01

Table 1. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the news dataset.

Test-time augmentation (TTA) is also explored. Similarly, as the data augmentation that is applied during the training phase, TTA can be used during test time. TTA performs augmentations during test time, so instead of just using the original image for inference, random augmentations such as random scaling and flipping are performed and then the average of these predictions is taken into account for the final prediction. TTA results in significant improvements in mAP over all classes as illustrated in Table 2. The mean mAP for all classes is 0.90 compared to 0.87 without TTA. The order of the mAP values for all classes is similar to the ones without TTA. The standard deviation for the mAP for all classes except for the author and date is less or equal to 0.02 as well.

	Classes	Precision	Recall	map0.5
0	all	0.93 ± 0.01	0.86 ± 0.01	0.90 ± 0.01
1	title	0.96 ± 0.02	0.97 ± 0.01	0.97 ± 0.01
2	metadata	0.95 ± 0.02	0.76 ± 0.02	0.82 ± 0.02
3	author	0.89 ± 0.05	0.67 ± 0.05	0.75 ± 0.04
4	date	0.94 ± 0.04	0.79 ± 0.03	0.87 ± 0.03
5	main_image	0.96 ± 0.02	0.98 ± 0.01	0.99 ± 0.01
6	text	0.92 ± 0.03	0.93 ± 0.01	0.94 ± 0.01
7	keywords	0.91 ± 0.02	0.94 ± 0.01	0.95 ± 0.02

Table 2. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the news dataset with TTA.

4.2 One-shot Learning on English News TF and Greek dataset

One-shot learning capabilities of the model are also explored by using transfer learning to fine-tune the model to unseen news websites in English but also in another language (Greek). In both cases, we use one different image from each domain of the dataset for training to test how different input images affect the performance of the model and check the variability of the results. We use one of the pretrained models (on the news dataset) and fine-tune it with just one image per domain. This experiment is conducted 5 times for each dataset.

In Table 3, the mean and standard deviation for all classes and for the metrics of precision, recall, and map0.5 are presented for the English news TF dataset. The mAP (0.75) is quite lower than the mAP of the original news dataset (0.87). The title, main image, and text have still high mAP values but metadata, author, date, and keywords have quite lower values. This can be attributed to the fact that there may be high variability of this dataset in certain classes compared to the original news dataset and likewise the high mAP values of the title, main image and text can be explained by the fact that they are most commonly visually consistent across domains. Moreover, the standard deviation for the mAP for most of the classes is quite high. This is to be expected since only one image for each domain is used for training each time and it may not be able to capture the same degree of variation that is required for all the webpages of that domain.

	Classes	Precision	Recall	map0.5
0	all	0.80 ± 0.12	0.73 ± 0.08	0.75 ± 0.08
1	title	0.98 ± 0.01	1.00 ± 0.00	0.99 ± 0.00
2	metadata	0.54 ± 0.29	0.36 ± 0.03	0.34 ± 0.04
3	author	0.68 ± 0.21	0.46 ± 0.25	0.59 ± 0.21
4	date	0.67 ± 0.20	0.71 ± 0.34	0.69 ± 0.31
5	main_image	0.97 ± 0.01	0.90 ± 0.07	0.93 ± 0.06
6	text	0.88 ± 0.07	0.95 ± 0.03	0.97 ± 0.01
7	keywords	0.87 ± 0.10	0.69 ± 0.15	0.76 ± 0.12

Table 3. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the English news TF dataset.

TTA is also applied to the English news TF dataset. Using the same methodology as described above, the results are presented in Table 4. TTA leads to higher mAP over all classes; 0.79 compared to 0.75 for the vanilla case. The order of the mAP values is similar to the ones without TTA and the standard deviation for the mAP for most of the classes is reduced.

	Classes	Precision	Recall	map0.5
0	all	0.83 ± 0.11	0.76 ± 0.08	0.79 ± 0.07
1	title	0.98 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
2	metadata	0.61 ± 0.28	0.38 ± 0.00	0.35 ± 0.03
3	author	0.75 ± 0.21	0.51 ± 0.37	0.70 ± 0.21
4	date	0.69 ± 0.17	0.71 ± 0.31	0.70 ± 0.29
5	main_image	0.98 ± 0.01	0.96 ± 0.03	0.98 ± 0.02
6	text	0.91 ± 0.05	1.00 ± 0.00	0.99 ± 0.00
7	keywords	0.87 ± 0.11	0.77 ± 0.13	0.82 ± 0.09

Table 4. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the English news TF dataset with TTA.

In Table 5, the mean and standard deviation for all classes and for the metrics of precision, recall, and map0.5 are presented for the Greek dataset. The mAP (0.89) is a bit higher compared to the mAP of the original news dataset (0.87). The title, metadata, main image, and text have still high mAP values but, author, and date have lower values. This high mAP across all classes can be attributed to the fact that there may be a high degree of similarity of this dataset with the original news dataset. Moreover, the standard deviation for the mAP for some of the classes is quite high e.g., author. This is to be expected since only one image is used for training as explained in the English news dataset above.

	Classes	Precision	Recall	map0.5
0	all	0.87 ± 0.04	0.87 ± 0.05	0.89 ± 0.04
1	title	0.89 ± 0.04	0.92 ± 0.04	0.96 ± 0.01
2	metadata	0.92 ± 0.05	0.89 ± 0.09	0.96 ± 0.02
3	author	0.86 ± 0.10	0.80 ± 0.21	0.82 ± 0.17
4	date	0.70 ± 0.07	0.71 ± 0.10	0.71 ± 0.11
5	main_image	0.97 ± 0.01	1.00 ± 0.00	0.99 ± 0.00
6	text	0.84 ± 0.06	1.00 ± 0.01	0.98 ± 0.01
7	keywords	0.92 ± 0.05	0.76 ± 0.05	0.84 ± 0.07

Table 5. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the Greek news dataset.

TTA is also applied to the Greek dataset. Using the same methodology as described above, the results are presented in Table 6. TTA leads to a slighter higher mAP over all classes, 0.90 compared to 0.89 for the vanilla case. The order of the mAP values is similar to the ones without TTA and the standard deviation for the mAP for all of the classes is reduced.

	Classes	Precision	Recall	map0.5
0	all	0.88 ± 0.02	0.89 ± 0.03	0.90 ± 0.02
1	title	0.90 ± 0.04	0.91 ± 0.04	0.96 ± 0.02
2	metadata	0.91 ± 0.02	0.95 ± 0.07	0.94 ± 0.01
3	author	0.89 ± 0.06	0.82 ± 0.16	0.85 ± 0.13
4	date	0.77 ± 0.06	0.73 ± 0.13	0.74 ± 0.11
5	main_image	0.97 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
6	text	0.84 ± 0.03	1.00 ± 0.00	0.98 ± 0.00
7	keywords	0.84 ± 0.03	0.80 ± 0.04	0.86 ± 0.05

Table 6. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the Greek news dataset with TTA.

4.3 State-of-the-art Comparison on the Books Dataset

The model yielded 0.95 mAP with stratified (based on domain) 5-fold cross-validation on the books dataset over all classes. In Table 7, the mean and standard deviation for all classes and for the metrics of precision, recall, and map0.5 (0.5 IoU threshold) are presented. The mean mAP for all classes is 0.95. All the classes; title, author, and price have a very high mAP (greater or equal to 0.91). The standard deviation for the mAP for all classes except the title is less or equal to 0.02.

	Classes	Precision	Recall	map0.5
0	all	0.97 ± 0.01	0.93 ± 0.02	0.95 ± 0.02
1	title	0.99 ± 0.01	0.84 ± 0.05	0.91 ± 0.05
2	author	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02
3	price	0.96 ± 0.03	0.98 ± 0.02	0.97 ± 0.02

Table 7. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the books dataset.

TTA was also applied to the books dataset. Using the same methodology as described above, the results are presented in Table 8. TTA leads to similar mAP order for all classes with a slight increase in the title class. No significant change for standard deviation as well.

	Classes	Precision	Recall	map0.5
0	all	0.97 ± 0.01	0.94 ± 0.03	0.95 ± 0.03
1	title	1.00 ± 0.01	0.86 ± 0.07	0.93 ± 0.05
2	author	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.04
3	price	0.96 ± 0.02	0.98 ± 0.02	0.97 ± 0.03

Table 8. Mean and standard deviation for all classes for the metrics of precision, recall, and map0.5 for the books dataset.

Our approach yields better precision, recall, and mAP compared to the state-of-the-art approach as shown in Table 9. This can be attributed to the deeper model architecture, the novel data augmentation methods e.g., mosaic, and all the further improvements that were discussed in Chapter 2, in the evolution from YOLOv2 to YOLOv5.

Methods	Precision	Recall	map0.5
State of the art	0.97	0.48	0.74
Proposed Approach	0.97 ± 0.01	0.93 ± 0.02	0.95 ± 0.02
Proposed Approach + TTA	0.97 ± 0.01	0.94 ± 0.03	0.95 ± 0.03

Table 9. Mean and standard deviation comparison of our approach with a state-of-the-art approach for all classes for the metrics of precision, recall, and map0.5 for the books dataset.

For the books dataset, in both cases (with or without TTA), the results are generally better than the ones from the news dataset. This is to be expected since with the book dataset the full resolution of the images is used during training and so, small objects can be more easily detected. Also, the news datasets include articles from multiple domains whereas the books dataset contains book product details only from one domain.

4.4 Detection Speed and Accuracy Trade-off

In all cases, TTA leads to various degrees of improvement in regards to the mAP. However, it’s important to note that there is a trade-off between detection speed and accuracy when using TTA. TTA leads in most cases to half or less FPS for all datasets as shown in Table 10.

Dataset	TTA	FPS	mAP (over all classes)
news	no	34.92 ± 3.51	0.87 ± 0.01
	yes	17.16 ± 0.78	0.90 ± 0.01
Greek news	no	22.52 ± 0.72	0.89 ± 0.04
	yes	9.20 ± 0.12	0.90 ± 0.02
English news TF	no	17.34 ± 0.25	0.75 ± 0.08
	yes	11.94 ± 0.16	0.79 ± 0.07
books	no	18.78 ± 5.72	0.95 ± 0.02
	yes	5.72 ± 1.53	0.95 ± 0.03

Table 10. FPS (mean and standard deviation) and mAP over all classes for each dataset with and without TTA.

5 Conclusion and Future Work

In this thesis, we reformulated the problem of WIE as an object detection task. For this purpose, a dataset with news articles was collected and annotated. A state-of-the-art detector, YOLOv5, was used to extract specific attributes such as the news articles' title, metadata, author, date, main image, text, and keywords. The model yielded 90% mAP (over all classes) in stratified (based on website domain) 5-fold cross-validation. One-shot learning capabilities of the model were also explored by using transfer learning to fine-tune the model to unseen news websites in English but also in another language (Greek) achieving 79% mAP and 90% mAP respectively. A books dataset, with classes title, author, date was used to compare our approach with a state-of-the-art approach where a previous version of YOLO (version 2) was utilized. The mAP of our approach yielded 95% mAP compared to the state-of-art approach which yielded 74% mAP. In all cases, TTA leads to various degrees of improvement in respect to the mAP but with a detection speed/accuracy trade-off. Our approach works not only in multiple websites within the same domain (e.g., news articles) but can also work for other domains (e.g., book listings). Regarding future directions, testing this method to extract specified fields from other domains e.g., job listings would be interesting. We were also restricted by the GPU's memory so experiments with higher resolution images and bigger models could possibly yield better results.

References

[1] B. Y. Lin, Y. Sheng, N. Vo, and S. Tata, “FreeDOM: A Transferable Neural Architecture for Structured Information Extraction on Web Documents,” Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul. 2020, doi: 10.1145/3394486.3403153.

[2] V. Govindasamy and G. Suresh, “Complex Event Processing Based User-Friendly Web Data Extraction,” International Journal of Emerging Engineering Research and Technology, vol. 2, no. 3, pp. 186–193, 2014.

[3] A. Suresh Babu, I. Sadia Naureen, M. Rao, and G. Sirisha, “WEB INFORMATION EXTRACTION USING DEPTA,” International Journal of Computer Science Engineering and Information Technology Research, vol. 2, no. 2, 2012.

[4] T. Weninger, R. Palacios, V. Crescenzi, T. Gottron, and P. Merialdo, “Web Content Extraction -a Meta-Analysis of its Past and Thoughts on its Future.”

[5] S. K. Patnaik, C. N. Babu, and M. Bhave, “Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks,” Big Data Mining and Analytics, vol. 4, no. 4, pp. 279–297, Dec. 2021, doi: 10.26599/bdma.2021.9020012.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5).”

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[8] L. Liu et al., “Deep learning for generic object detection: A survey,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 261–318, 2020.

[9] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective Search for Object Recognition,” International Journal of Computer Vision, vol. 104, no. 2, pp. 154–171, Apr. 2013, doi: 10.1007/s11263-013-0620-5.

[10] R. Girshick, Robots.ox.ac.uk. [Online]. Available: https://www.robots.ox.ac.uk/~tvg/publications/talks/fast-rcnn-slides.pdf

[11] R. Girshick, “Fast R-CNN,” arXiv.org, 2015.

[12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/tpami.2016.2577031.

[13] “Getting Started with SSD Multibox Detection - MATLAB & Simulink - MathWorks Benelux,” Mathworks.com, 2016. https://nl.mathworks.com/help/vision/ug/getting-started-with-ssd.html

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi: 10.1109/cvpr.2016.91.

[15] H. Jain and S. K. Nandy, “Incremental Training for Image Classification of Unseen Objects,” ResearchGate, Aug. 2019. https://www.researchgate.net/publication/345061606_Incremental_Training_for_Image_Classification_of_Unseen_Objects

[16] C. Szegedy et al., “Going deeper with convolutions,” 2014.

[17] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” 2016.

[18] S. Ioffe, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” 2015.

[19] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement.”

[20] “Deep Learning for Vision Systems,” Manning Publications, 2020.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection.”

[22] C.-Y. Wang, H.-Y. Liao, I-Hau. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, “CSPNET: A NEW BACKBONE THAT CAN ENHANCE LEARNING CAPABILITY OF CNN A PREPRINT,” 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” Computer Vision – ECCV 2014, pp. 346–361, 2014, doi: 10.1007/978-3-319-10578-9_23.

[25] F. Lin, Q. Wu, J. Liu, D. Wang, and X. Kong, “Path aggregation U-Net model for brain tumor segmentation,” Multimedia Tools and Applications, Mar. 2020, doi: 10.1007/s11042-020-08795-9.

[26] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, ... wanghaoyang0106. (2021). ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support (v6.0). Zenodo. https://doi.org/10.5281/zenodo.5563715

[27] A. Rosebrock, “Intersection over Union (IoU) for object detection - PyImageSearch,” PyImageSearch, Nov. 07, 2016. https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

[28] J. Hammer, J. Mchugh, and H. Garcia-Molina, “Semistructured data: The TSIMMIS experience,” ResearchGate, Sep. 15, 2002.

[29] G. Arocena and A. Mendelzon, “WebOQL: Exploiting Document Structure in Web Queries,” undefined, 2017.

[30] S. Soderland, Machine Learning, vol. 34, no. 1/3, pp. 233–272, 1999, doi: 10.1023/a:1007562322031.

[31] Ion Muslea, S. Minton, and C. A. Knoblock, “STALKER: Learning Extraction Rules for Semistructured, Web-based Information Sources *,” undefined, 2017.

[32] M. Califf and R. Mooney, “Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction,” Journal of Machine Learning Research, vol. 4, pp. 177–210, 2003, Accessed: Feb. 08, 2022.

[33] A. Schulz, J. Lassig, and M. Gaedke, “Practical Web Data Extraction: Are We There Yet? - A Short Survey,” 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Oct. 2016, doi: 10.1109/wi.2016.0096.

[34] A. W. Hogue and David Ron Karger, “Thresher: automating the unwrapping of semantic content from the World Wide Web,” ResearchGate, 2005.

[35] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, “VIPS: a Vision-based Page Segmentation Algorithm,” ResearchGate, 2003.

[36] “Automatic Web Content Extraction by Combination of Learning and Grouping | Proceedings of the 24th International Conference on World Wide Web,” ACM Other conferences, 2015. https://dl.acm.org/doi/10.1145/2736277.2741659.

[37] “NewsCatcher News API,” Newscatcherapi.com, 2022. https://newscatcherapi.com/

[38] “selenium,” PyPI, Nov. 22, 2021. https://pypi.org/project/selenium/

[39] wkentaro, “wkentaro/labelme: Image Polygonal Annotation with Python (polygon, rectangle, circle, line, point and image-level flag annotation).,” GitHub, Nov. 24, 2021. https://github.com/wkentaro/labelme

[40] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context.”

[41] A. Benjumea, Izzeddin Teeti, Fabio Cuzzolin, and A. Bradley, “YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles,” ResearchGate, Dec. 22, 2021.

[42] Q. Song et al., “Object Detection Method for Grasping Robot Based on Improved YOLOv5,” Micromachines, vol. 12, no. 11, p. 1273, Oct. 2021, doi: 10.3390/mi12111273.

[43] Q. Xu, Z. Zhu, H. Ge, Z. Zhang, and X. Zang, “Effective Face Detector Based on YOLOv5 and Superresolution Reconstruction,” Computational and Mathematical Methods in Medicine, vol. 2021, pp. 1–9, Nov. 2021, doi: 10.1155/2021/7748350.

[44] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” 2019.

[45] “Weights and Biases, Inc,” Wandb.ai, 2015. https://wandb.ai/site

Appendix A

	Augmentation	Value
0	HSV_H	0.015
1	HSV_S	0.7
2	HSV_V	0.4
3	Scale	0.5
4	Translate	0.1
5	Mosaic	1

Table 11. Data augmentation values.

Appendix B

domains	count	min height	max height	mean height	std height	min width	max width	mean width	std width
abc.net.au	71	5348	16081	9292.0	2554.0	800	800	800.0	0.0
aljazeera.com	25	6062	14286	8052.0	1653.0	800	800	800.0	0.0
aol.com	21	4409	15060	8024.0	2548.0	800	800	800.0	0.0
bbc.co.uk	44	1733	13325	7966.0	2575.0	800	800	800.0	0.0
bbc.com	33	1821	14134	8798.0	2310.0	800	800	800.0	0.0
bizjournals.com	22	3682	9000	5700.0	1453.0	800	970	846.0	77.0
cbsnews.com	20	3077	8564	5318.0	1332.0	800	800	800.0	0.0
chicagotribune.com	22	3939	8214	5305.0	1062.0	800	800	800.0	0.0
chinadaily.com.cn	27	944	8859	3212.0	1549.0	800	1010	862.0	98.0
chron.com	21	3891	13528	7064.0	2524.0	800	800	800.0	0.0
cnbc.com	20	3489	10650	5708.0	1518.0	800	800	800.0	0.0
cnn.com	21	3160	17724	9129.0	2982.0	800	800	800.0	0.0
dw.com	23	5014	9593	6751.0	1007.0	980	980	980.0	0.0
espn.com	21	1756	6289	3232.0	1224.0	800	800	800.0	0.0
fastcompany.com	20	10257	17073	11856.0	1907.0	800	800	800.0	0.0
france24.com	20	2681	9138	5371.0	1789.0	800	800	800.0	0.0
globalnews.ca	15	6612	13790	8618.0	1986.0	800	800	800.0	0.0
go.com	21	2103	6967	4180.0	1413.0	800	800	800.0	0.0
huffpost.com	22	3622	10756	7042.0	1641.0	800	800	800.0	0.0
indiatimes.com	22	4049	8808	5556.0	1288.0	800	1003	920.0	102.0
irishtimes.com	21	4808	11456	7678.0	2140.0	800	800	800.0	0.0

latimes.com	21	4273	10205	7144.0	1458.0	800	800	800.0	0.0
mercurynews.com	21	4611	9093	6611.0	1258.0	800	820	801.0	4.0
metro.co.uk	21	5419	15204	8156.0	2725.0	800	800	800.0	0.0
nbcnews.com	26	2781	7523	4804.0	1356.0	800	800	800.0	0.0
ndtv.com	23	2563	4832	3350.0	581.0	1283	1283	1283.0	0.0
netscape.com	21	1007	2694	1790.0	389.0	980	980	980.0	0.0
npr.org	22	5133	14425	7692.0	2458.0	800	800	800.0	0.0
nypost.com	15	3525	8791	5571.0	1813.0	800	800	800.0	0.0
thestar.com	11	4890	8582	6412.0	1278.0	1176	1176	1176.0	0.0
variety.com	15	5746	9435	7368.0	1083.0	913	913	913.0	0.0

Table 11. Basic statistics about the domain count, width and height of the news dataset.

domains	count	min height	max height	mean height	std height	min width	max width	mean width	std width
news247.gr	24	5350	10937	6943.0	1127.0	1000	1000	1000.0	0.0
sport24.gr	21	2744	8096	4093.0	1518.0	800	800	800.0	0.0

Table 12. Basic statistics about the domain count, width and height of the Greek dataset.

domains	count	min height	max height	mean height	std height	min width	max width	mean width	std width
amazon.in	138	235	1080	706.0	277.0	398	1920	1073.0	545.0

Table 13. Basic statistics about the domain count, width and height of the books dataset.

Appendix C

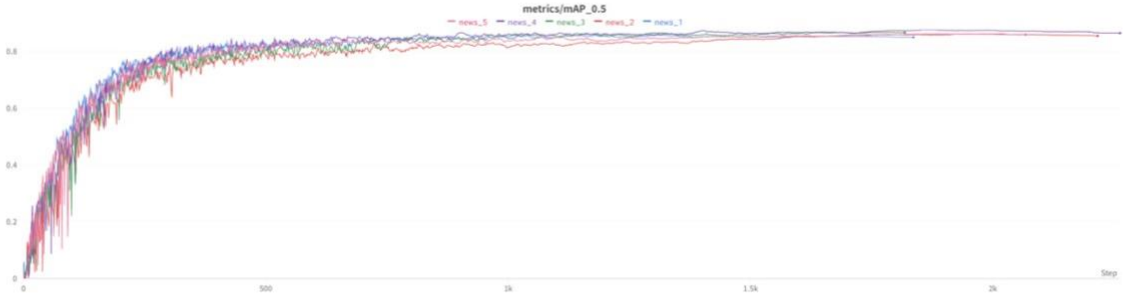


Figure 14. mAP per epoch (validation set) of training for each experiment with the news dataset.

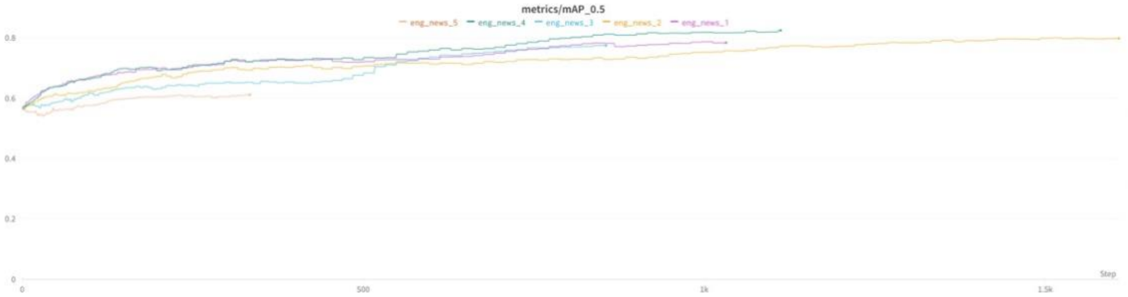


Figure 15. mAP per epoch (validation set) of training for each experiment with the English news TF dataset.

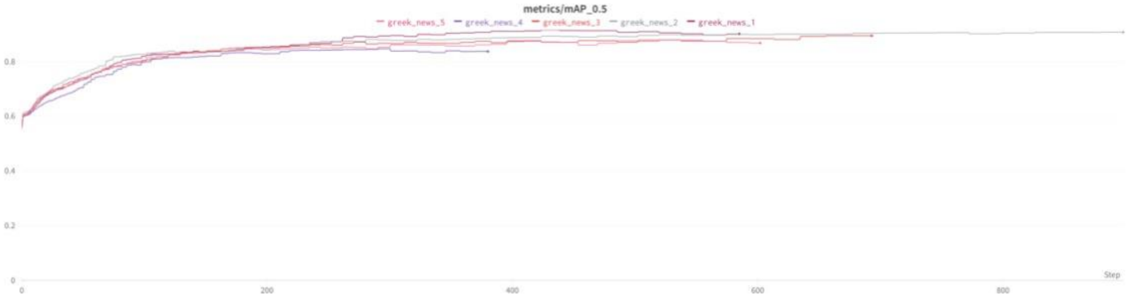


Figure 16. mAP per epoch (validation set) of training for each experiment with the Greek news dataset.

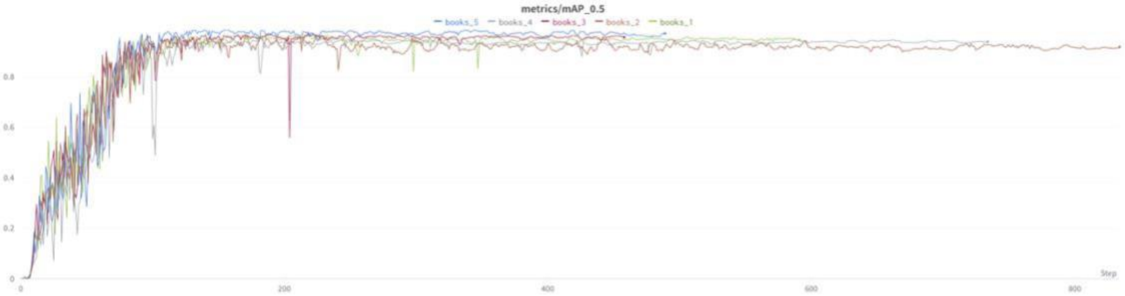


Figure 17. mAP per epoch of training for each experiment with the books dataset.