UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Development of layout analysis system for historic scholar publications

Author:
Olha BAKAY

Supervisor:
Dr. Olesya MRYGLOD
Oles DOBOSEVYCH

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2019

Declaration of Authorship

I, Olha BAKAY, declare that this thesis titled, "Development of layout analysis system for historic scholar publications" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Done is better than perfect."

Unknown

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Development of layout analysis system for historic scholar publications

by Olha BAKAY

Abstract

In this work, we compare the results of different approaches for automatic document layout analysis using Convolutional Neural Networks. Although there is great progress in the Image Processing domain, there are still open problems, such as accurate detection of regions of content and classification of them into semantically similar classes. The primary purpose of work is to simplify the further processing of Ukrainian historic archives. For it, two various techniques were used. The first one is modification and re-implementation of already existing approach for document layout analysis. Another method is suggested by us and re-uses the pre-trained model on a bigger dataset. During this work, we also collected a new dataset of Ukrainian scientific publications. We evaluate these approaches on an independent test set and compare the precisions of each model.

Acknowledgements

I owe my deepest gratitude to Oles Dobosevych for continuing support and invaluable help throughout the entire project. Special thanks also to Olesya Mryglod for generating new ideas and giving constructive comments. Also, I want to thank Ukrainian Catholic University and the Faculty of Applied Sciences for making such a powerful Bachelor's Program in Computer Science, which had a significant impact on my future.

Contents

List of Figures

List of Abbreviations

NTSh	**S**hevchenko **S**cientific **S**ociety (*Naukove tovarystvo imeni Shevchenka*)
ANN	**A**rtificial **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
RLSA	**R**un **L**ength **S**moothing **A**lgorithm
YOLO	**Y**ou **O**nly **L**ook **O**nce
SSD	**S**ingle **S**hot **D**etector
IMPACT	**Imp**roving **Ac**cess to **T**ext Dataset

Chapter 1

Introduction

1.1 Motivation

1.1.1 Shevchenko Scientific Society and its history

The **Shevchenko Scientific Society** (Ukrainian: Наукове товариство імені Шевченка, НТШ, *Naukove tovarystvo imeni Shevchenka*, **NTSh**) is the Ukrainian scientific society. It was founded in 1873 as a public organisation committed to the encouragement for Ukrainian literature and Ukrainian language (*NTSh online*). However, later, in 1892, it was reorganised into the scientific society which can be referred to as the first genuinely Ukrainian Academy of Sciences. While NTSh was multidisciplinary scholar organisation, its organisational structure was defined by three topical sections: history-philosophical, philological, and mathematically-medical-natural scientific (Ярослав Грицак, 2001). NTSh was playing a dominant role in developing the Ukrainian system of scientific knowledge continuously proving the self-sufficiency and authenticity of Ukrainian national science. In particular, Ukrainian terminology and scientific language were promoted despite Ems Ukaz and Valuev Circular (*NTSh online*). The members of NTSh supported a unique self-organised activity participating in the functioning of underground university — Ukrainian University in Lviv (M.L. Dudka, 2018).

Many activities of NTSh were interrupted during the period between and including the First and the Second World Wars. In 1940, the society was dissolved by Soviet occupants and was not able to work publicly later, during Nazi occupation ("НТШ у Львові"). NTSh was revived as a union of semi-independent scientific societies in emigration. Chapters of society were established all over the world: in Paris, New York, Toronto, and on the Australian continent. Only in 1989, the NTSh was renewed in Ukraine. Currently, 23 NTSh centres are acting in Ukraine. What is more, over 1400 researchers are united in 6 sections and 35 commissions (*NTSh online*).

1.1.2 Publication activity of Shevchenko Scientific Society

Publishing is one of the essential activities of the Shevchenko Scientific Society. There are two main reasons for this. First, NTSh was dedicated to spreading Ukrainian literature and language, and publications were one of the best ways to make knowledge available to the public. Second, NTSh became a more scientific society in years, and publications still are the main form of scientific communication and the dissemination of scientific information.

Until World War II, there were a lot of new scholarly researches and notices published, amidst which launching and publishing of many new periodicals and serials publications took place. In these works, the actual information that was previously

unknown for the majority was presented. Such information included the studies dedicated to Ukrainian language, literature and science (*NTSh online*).

One of the most famous periodic scientific publications was the *Zapysky NTSh* (Notes of the Shevchenko Scientific Society). As of today, the Zapysky NTSh is the most representative body of Ukrainian science with more than 250 volumes and, moreover, it is still published (Записки НТШ 2019). The periodicals became a laboratory of scientific thought for Ukrainians from both sides of the Austro-Russian border. On its pages, young writers and scholars made their debut, which later became the embellishment of a new, modern Ukrainian literature. Except for Zapysky NTSh, there were also other valuable serial scientific and periodic publications such as the *Zbirnyk* (Collection of works), which also included scientific articles and reviews of books and were divided into different sections by topics (for example, historical and philosophical section and others) (*NTSh online*, "Periodicals and serials Shevchenko Scientific Society (1894–1939)").

The principal role of the Shevchenko Scientific Society was in the formation and affirmation of Ukrainian language and science, that was very important under the conditions of government change and influences of other cultures (such as Polish, Russian and German).

1.1.3 The heritage of Shevchenko Scientific Society and European context

The development of science is tightly connected with the development of forms of scientific communication. The knowledge, disseminated among academic peers, is automatically verified by experts and can be used as a basis for future research. "Publish or perish" (Parchomovsky, 1999) — this principle of modern science means that only published results are acknowledged, are visible, and are "real". Papers in academic periodicals remain the dominant form of presenting scientific results starting from the middle of the 17th century. For European science, it is historically connected with the foundation of the first academic periodicals in the world: "Philosophical Transactions of the Royal Society" (London, 1665)("Publishing the Phil. Trans.: the economic, social and cultural history of a learned journal, 1665–2015"." 1963) and "Journal des Sçavans" (Paris, 1665)(*Journal des Savants*). Already after a very little period after their establishment, the avalanche of academic publications testified to the exponential growth of science, see Price, 1963. Therefore, it is hard to overestimate the value of the two first journals for the history of European and World science. The editions published by NTSh play the same role for the history of Ukrainian science — as separate phenomena and as an integral part of European science. Therefore, preserving, dissemination and investigation of the heritage of NTSh is a problem of current importance.

However, there are some problems with NTSh archive publications. First, there are still many volumes of writings that are placed all over the world and are not digitised. Second, not all the publications are accessible to the public. The free access to such valuable historical data would make it possible to transmit the knowledge collected for centuries to the present world. Furthermore, it would encourage researchers to explore and discover over the tonnes of scientific data. Third, to the best of our knowledge, NTSh works have not been analysed so far.

Modern technologies and techniques of image processing enable to automate metadata collection process, organise information in a structured and convenient way for searches and analysis. So the next step will be a processing of the metadata that would significantly simplify doing researches in the future. Thereby, the main idea of this work consists in making NTSh's archives more open for future studies

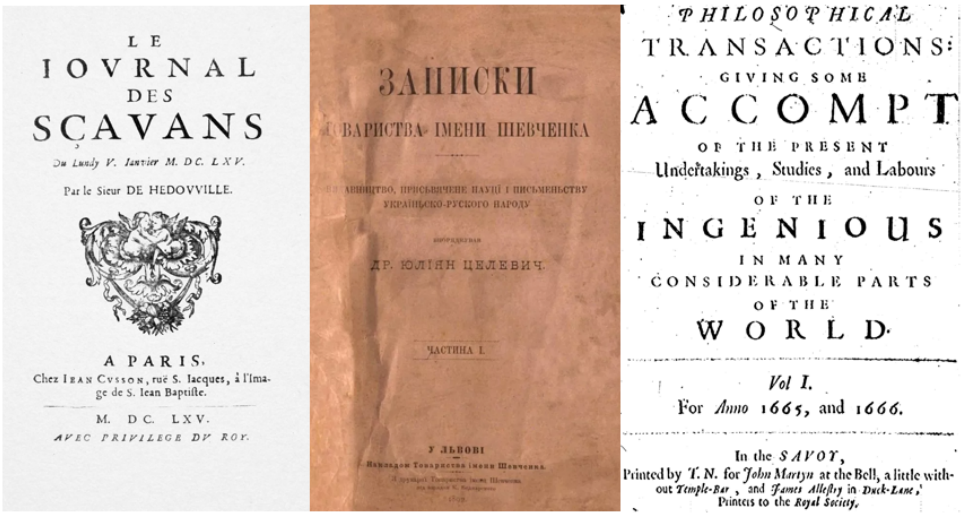


FIGURE 1.1: Title page to volume 1 of *Journal des sçavans*, *Philosophical Transactions* and *Zapysky NTSH*. *Source*: *Journal des Savants*, *Archive of NTSh*, *Phil. Trans.*

by streamlining the process of preprocessing raw data to a more convenient and searchable way.

Collections of archival documents are essential for scientists since they give us proofs of specific activities and shed a little more light on persons and societies, that were founded. Besides, they rise human's feel of identity and awareness of differences in the world's cultures. Sometimes they are even able to ensure fairness. Typically, all these historical documents were not written for personal purposes or to rewrite some parts of history so archives can be used as an objective point of view of the historical events. It can be concluded that analysing historical data makes it easier for us to understand the past and all its consequences for the next generations. Therefore, by learning all changes in evolution processes, we can get specific peak changes or find out about any evolution patterns. Researching historical data makes possible tracking of all the improvements or changes over the centuries what gives us a lot of key insights. And these insights are necessary for understanding the world nowadays.

1.1.4 Scientometrics. Bibliographic analysis. Complex network approach.

One of the effective methods to research historical data is to use specific approaches of scientometrics. Scientometrics is a discipline that studies the evolution of science through the numerous measuring of scientific information (*Scientometrics and citation index*). In other words, scientometrics does statistical researches on the structure and dynamic of scholarly information streams. A number of scientific articles published in a specific time, frequency of author citations, participation in international scientific conventions - all of these and many others are examples of indicators of scientific effectiveness which make quantitative evaluation and a comparative analysis of scientific activity and productivity on different levels (individuals, journals and institutions, countries and regions). The most popular measures are citation index, h-index and impact factor. The *citation index* is calculated by the number of references on the particular paper or author's name in others works. The *h-index*

(*Hirsch index*) is a metric which takes into account both the citation impact of the scientist's publications (amount of citations) and his/her productivity (amount of cited publications). The *impact factor* is a measure of the importance of the journal that is measured by the number of citations of its works (Durieux and Gevenois, 2010).

Modern scientometrics is mainly involved in the evaluation of science, but we are interested in its role in the analysis of history, research on development and evolution of science - this is what bibliometrics do. *Bibliometrics* is a statistical analysis of information, that was recorded in a paper form (books, articles, periodicals and others) (Rostaing, 2003). Bibliometric indicators allow to reveal an objective state of the literary editions in a certain period, help to reveal the patterns of development and on this basis determine areas of improvement of publishing (Rostaing, 2003). The methods of bibliometrics include analysis of citations, analysis of quantitative characteristics of documents, quantitative analysis of the publications of certain authors and their citations; quantitative analysis of the publications of scientists from particular countries; theoretical analysis such as studies about patterns of growth, aging and ranking of scientific documents, content analysis of scientific works, and any other issues related to the distribution of scientific documents. The use of these methods allows for tracking dynamic changes in publishing over a particular period. Moreover, based on the results, it is possible to identify whether the trends in publishing are positive or negative. This results also can cause future researches over document flow by traditional content analysis methods.

There are many examples of how to apply scientometrics to NTSh publications and obtain some meaningful and exciting results. One of the most interesting approaches for us to use over NTSh archives is to do citation analysis. As stated by Eugene Garfield, who founded this method, *citation analysis* is a method that simplifies much work involved by a detailed review of the periodicity and behaviour of citations in a paper (Garfield and Merton, 1979). By researching NTSh periodicals, it is possible to create an entire network of co-authorship or explore the evolution of studying specific fields of sciences, the impact of foreign scholar papers on the development of Ukrainian science base, and others. One of the simplest examples is to create a network of co-authors. It can be done by finding information based on references, which are many in every periodical publication. The references usually mention somebody's publication or research article.

Simply saying, a *network* is a group of the points interconnected by lines. Many objects from different fields of our life can be formed as networks, and these networks can be viewed as a modification of graph theory. Having a lot of publication data from the scientific journals, one can create a number of complex networks, where particular authors, articles, groups of authors (and others) can serve as network's nodes and connections among these nodes will show various relations among corresponding data, including authorship, citations, using keywords or others factors (Головач, Ю and фон Фербер, К and Олємською, О and Головач, Т and Мриглод, О and Олємской, І and Пальчиков, В, 2006). By presenting the data about NTSh publications in the form of a complex network, a variety of network algorithms can be used. For example, such algorithms include those that detect the structure of the networks: define connected components (nodes) that have a greater amount of links among each other. In the case of analysing data on publications in a scientific publication, finding a structure of a network of articles or authors, combined with certain connections, will help to group them on a common scientific topic or discover the authors' collectives.

That is why to get any insights from the NTSh archives or, at least, to extract any information, there is a need to go through all papers and do document layout

analysis which includes analysing each page for presence of document regions of interest. In other words, this procedure can be called as 'data preprocessing' as it is usually done before proceeding to the research. Doubtless, it is a long and routine process. For that reason, this work is going to simplify the process of preprocessing by developing of a layout analysis system for scholar publications and in result give a starting point for the future researches.

One of the major features of similar systems is extracting metadata, such as titles, authors, references, any notes and others, for relieving establishment of the scientific literature databases. In addition to that, many other document sections such as text, years, images and captions can be helpful and useful for deeper analysis of extracted information from digitised documents (Klampfl et al., 2014). Besides, the detection of named entities and any details included in documents bodies will be a good foundation for future more in-depth analysis of documents.

7

II.

Зведенє інтеґралів еліптичних методами елементарними.

§. 1.

Получение $\int R(\xi, \sqrt{a_0+a_1\xi+a_2\xi^2+a_3\xi^3+a_4\xi^4})\,d\xi$ $a_4 \gtrless 0$ і $a_4 = 0$

в одній формі $\int R(t, \sqrt{\pm(t^2+\mu)(t^2+\lambda)})\,dt$ *).

Положім

$\eta = \sqrt{a_0+a_1\xi+a_2\xi^2+a_3\xi^3+a_4\xi^4} = \sqrt{(f-2g\xi+\xi^2)(f'-2g'\xi+\xi^2)a_4}$

і підставмо

$$\xi = \frac{p+qt}{1+t}$$

де p **і** q **в сталі довільні, то через відповідний їх добір можемо постарати ся о то, щоби сочинники при** t **були рівні зеро:**

$$f-2g\xi+\xi^2 = \frac{F-2Gt+Ht^2}{(1+t)^2}$$

$$f'-2g'\xi+\xi^2 = \frac{F'-2G't+H't^2}{(1+t)^2}$$

f, g, h, F, G, H, F', G', H', **в сталі, в котрих приходить** p **і** q.

Положім

$$G=0 = -f+g(p+q)-pq$$

$$G'=0 = -f'+g'(p+q) \quad pq$$

то знайдем потрібні нам (p, q).

$$\eta = \frac{1}{(1+t)^2}\sqrt{(F+Ht^2)(F'+H't^2)}$$

$$= \frac{\sqrt{a_4 H.H'}}{(1+t)^2}\sqrt{\pm\left(\frac{F}{H}+t^2\right)\left(\frac{F'}{H'}+t^2\right)}$$

$$= \frac{k}{(1+t)^2}\sqrt{\pm(t^2+\lambda)(t^2+\mu)} = \frac{k}{(1+t)^2}\,T$$

$$\int R(\xi\eta)\,d\xi = \int \bar{R}(t, T)\,dt.$$

Єсли $a_4 = 0$, **то поступаємо анальоґічно.**

*) Serret. — Harnack, Lehrbuch der Diff. und Integralr. T. II. стор. 48. Durége. — Theorie der ellip. Functionen.

14

Потім виступають вже коло 1235 р.: коли Данила побито в Київщині й бояреtво на якийсь час виперло єго з Галича, от у сей час галицькі бояре „и вси Болоховьсции князи съ ними“ напались на волинські волости Данила, повоювали береги р. Хомори (тече у Случ з лівого боку) й пішли до Каменця[1]), але потім повернули назад. Данилові бояре, відобравши запомогу, нагнали й забрали їх у полон. Михайло, кн. чернігівський й Ізяслав (роду з князів, мабуть, смоленських), що сидів тоді у Київі, вихагали у Данила, що-б він пустив сїх вязнів, але той на се не згодив ся[2].)

Пісьля того літопись якийсь час мовчить за Болоховців, аж вони з-являють ся знову вже пісьля находу Батия, в 1242 - 3 роках: тоді Ростислав Михайлович, борючись з Даниломъ, „собра князѣ болоховьскые и останокъ Галичанъ, приде ко бакотѣ,“ але єго відбито; Ростислав пішов за Дніпро. Тоді Данило напав ся на князїв болоховських: „гради ихъ огневи предасть и гребля ихъ роскопа... Даниль же возьма плѣнъ многъ врати ся, и поима грады ихъ: Деревичь, Губинъ и Кобудь, Городѣцъ, Божескый, Дядьковъ. Приде же Курилъ, печатникъ князя Данила, со трими тысящами нѣщець и трьми сты коньникъ, и въдасть имъ взяти Дядьковъ градъ. Оттуда же плѣнивъ землю болоховьскую и пожегъ, оставили бо ихъ Татарове, да имъ орють пшеницю и проса; Данилъ же на нѣ болшую вражьду держа, яко отъ Татаръ болшую надежду имѣаху“. При сему літописець згадує, що Данило був виручив болоховських князїв, коли вони зайшли були в волости Болеслава, князя мазовецького, й той заабрав був їх; Болеслав не хотїв попереду їх пустити, кажучи, що вони „суть особнии князи“, а не підданці Данила, але за подарунки пустив їх; се тепер літописець пригадує болоховським князям, яко кару за їх невдячність[3]).

В останній раз Болоховці з-являють ся в 50 роках XIII. в.[4]) Данило розпочав похід на землї, що безпосередне підлягали Татарам, і тоді „воевахуть людье Даниловы же и Василкови Болоховъ“, та мабуть не.

[1]) Се, мабуть, есть сучасна Каменка на Случі, на північ від р. Хомори, — див. наприклад: Барсова Географія начальной лѣтописи с. 288—290.

[2]) Ипат. с. 516; тут оповідаєть ся про се під р. 1235, на сей раз дата правдоподібна, бо се трапилось скоро після київського погрому (Михайла й Ізяслава), а той був 1235 р. — див. літопись Новгородську першу — Полное собр. лѣтоп. III. с. 50.

[3]) Ипат. с. 526—7.

[4]) Випадок сей трапив ся як видко з Галицько-волинської літописи, пісьля коронації Данила (1253 р.) й перед находом Бурундая (1258 р.); в літописи стоїть він під 1255 р. Див. за се Дашкевича Княженіе Даніила Галицкаго с. 85 і 99.

208

і щебетав що раз голоснїйше? Ти порівнував єго жартом до дука поета, що хотївби оперти ся на земних підпорах“.

З тим жайворонком може порівнати самого Залєского. Буйна, легка єго уява уносила го понад сьвіт земний и краєму ідей і мрій. Колосє народної поезиї, що виросло з реального ґрунту, з життя і чувств люду, угинало ся під ним; він инов вобивав ся в країву улуди, але рівночасно відділяв ся від дійсности. Тому єго поезиї, в більности випадків, о тілько мають виразну форму, о скілько суть вірною відбиткою дійствости, о скілько дух поета, ков той жайворонок, тріпаючи крильцями удержав ся на колосю української народної поезиї.

Олександер Колесса.

P. 1892.

FIGURE 1.2: Examples of pages in *Zbirnyk NTSh* and *Zapysky NTSh* with references and other notes that can be useful for analysis

Chapter 2

Background information

This chapter presents a short overview of the basic concepts concerning different approaches and methods that were used in the development of layout analysis system for historic scholar publications. Approaches themselves will be described in the next chapter.

As can be seen in many written works, there exist a lot of different methods to do document layout analysis. As stated in Augusto Borges Oliveira and Palhares Viana, 2017, all of them can be categorised into three groups:

- methods based on regions or blocks classification;
- methods based on pixels classification;
- methods based on connected component classification.

Block classification methods are those which divide a document page picture into some amount of blocks and classify each of them. Pixel classification methods take into account each pixel separately and use a classifier to create and mark bounding boxes of hypothetical regions. Methods based on connected component classification extract simple features from a picture and pass them to previously trained supervised learning algorithms of binary classifiers. Then by observing, combining and removing any of impurities components are finally classified.

In this work, we will compare different approaches to perform document analysis such as block-based classification method trained with pre-trained Convolutional Neural Network model with transfer learning, a run-length smoothing algorithm (RLSA) for segmentation and classification of digitised printed documents.

2.1 Artificial Neural Networks

An Artificial Neural Network (ANN) is a connected collection of simple processing elements, nodes or units. Deep Learning in an entire field that studies and uses neural networks as the main instrument. Processing capacity of a network is stored in the connections between strength units, or *weights*, received from a learning process on a set of training examples (Gurney, 2014). To put it another way, an ANN is a mathematical model that is organised in *layers*. Each layer consists of simple interconnected operating elements (also called *neurons*) that process information via the dynamic change of the state due to external inputs (Caudill, 1987). Even though the math behind the neural networks is not easy, it is still possible to obtain a general understanding of the neural networks structure and how they function.

The layer that receives raw data is called the *input layer*. The layer which gives the predictions and/or results called the *output layer*. All layers between these two layers are occupied with the actual processing and are named *hidden layers*. On input,

the layers receive the outputs from the previous layers. Each layer has an *activation function* that changes the weights of the connections by inputted data. The model in fig.2.1 consists of two *fully-connected* hidden layers where all nodes (neurons) of level n have full pairwise connections among two neighbour layers — n - 1 and n + 1 respectively.

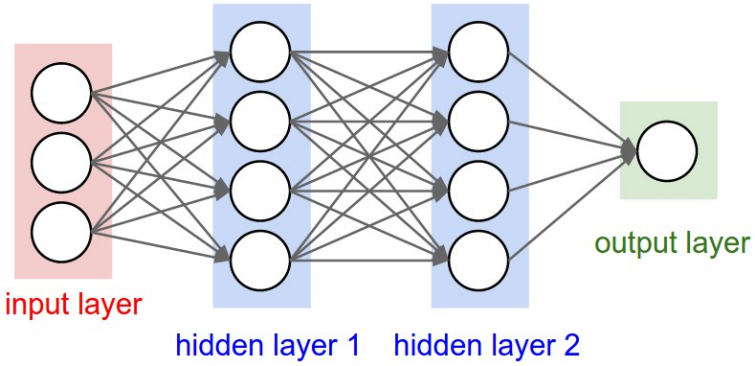


FIGURE 2.1: A neural network with three layers, three inputs, two fully-connected layers and one output layer. *Source:* Fei-Fei Li and Johnson, 2016

Another significant point neural networks are known for is their adaptiveness. It means that they can change themselves during learning from examples. While a NN is training, the weights are modified in accordance with the samples of input data. Moreover, neural networks can be used as common approaches to supervised, unsupervised and reinforcement learning problems. In most cases, the neural network needs a large number of variables and a significant amount of training data. For supervised tasks, training data must include matches of inputs and correct outputs for a specific problem. This work is an example of solving a supervised classification problem by applying a neural network to it. That is why its results will be compared to provided previously correct outputs during training and thereby model will be able to adjust its weights to find out how to perform better. At the time when the network is studied enough to provide an acceptable level of model performance, it can be used as an analytical tool for another set of data ("A Basic Introduction To Neural Networks").

2.2 Convolution Neural Network

Speaking about the image classification and neural networks it is impossible not to mention Convolution Neural Networks that were created in 1998 by Lecun et al. and now are a division of Deep Learning which is broadly used for different computer vision tasks, including document layout analysis. *Convolutional Neural Networks (CNN)* are neural networks that are designed for processing data with a known in advance homogeneous topology ("Deep Learning"). For instance, such data can be a time series or pictures. Time series can be considered as 1D-grid stored in the form of some records and measured with a fixed periodicity while an image is 2D-grid of pixels made of a photo or video frames.

The main difference between the average neural network and the convolutional network is that CNN has minimum one layer with the process of convolution which consists in using the specific linear kernel for every region of data in place of usual

matrix multiplication. There are three main types of layers for designing architectures of Convolutional Networks: *Convolutional Layer, Pooling Layer and Fully-Connected Layer* (the same one as in ordinal Neural Networks) (Fei-Fei Li and Johnson, 2016). *Convolutional layer* calculates a dot product between the part of input data (or the outputs from another layer) and the set of *kernel weights*. From the inputted kernel weights (in other words, after applying a filter) a feature map is created. This map indicates the existence of noticed features in the inputted data. *Pooling layer* by reducing the image size, received from previous layers, helps the convolutional layer to search for more features in the whole reduced image instead of concentrating on certain parts.

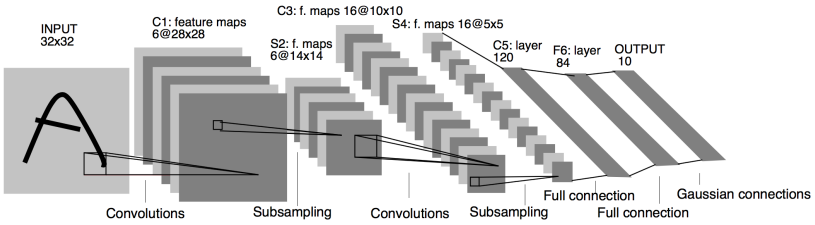


FIGURE 2.2: Architecture of LeNet-5, classical CNN with seven layers, among which there are three convolutional layers (C1, C3 and C5), two sub-sampling (pooling) layers (S2 and S4), and one fully-connected layer (F6). *Source:* LeCun et al., 1998)

The process of 2D-convolution is very helpful for problems of image recognition because it gives an opportunity to analyse only the context around a particular pixel and, thus, learn as simple as possible features (lines, angles, curves, and other.). Because if pixels are far from each other, in any case, they will be connected to next neuron, what in result will give a less positive effect. That is why the distance between pixels is an important aspect when talking about image recognition problems as it provides us with more information and makes more sense of specific fragments of image. Later, those simple features can be combined with the next layer to identify higher-level features (LeCun et al., 1998).

Currently, Convolutional Neural Networks have the best accuracies on most of the object recognition problems. Previously, CNN had a very high computational intensity. Therefore, sometimes it just limited any advantages from using it in cases when quick performance and low memory costs were necessary (Augusto Borges Oliveira and Palhares Viana, 2017). Nowadays due to the upgraded hardware CNNs can be expanded to much bigger architectures.

Deep Learning as an entire field is no longer a big black-box of algorithms. However, because of processes such as model's training (when millions of parameters are learnt), it is still not easy to understand or at least become aware of all processes and whole architecture. However, people try to improve their understanding of ANN by preparing different visualisations of neural network processing data (Zeiler and Fergus, 2014), saliency maps (*Saliency map*), and others. Even though it still not enough to have a full understanding of what CNN is and how it works, it gives pretty much good results in computer vision tasks, image segmentation in particular.

2.3 Transfer Learning

As stated in the original paper (Torrey and Shavlik, 2010), *transfer learning* is the technique of reusing a pre-trained neural network from already solved task to improve the learning process of a new related task. In other words, knowledge from a related task, for solving of which a sizeable labelled training dataset was used, is used to solve a new task, where we do not have any data. So rather than starting the training process from the very beginning, with the help of transfer learning, a training process can start with the previously-learnt patterns. It is a useful technique nowadays because there are not so many labelled datasets that would help to solve real-world problems. As described in Torrey and Shavlik, 2010, there are three most possible aspects of transfer learning improvement. They are pictured in fig.2.3 and will be explained below.

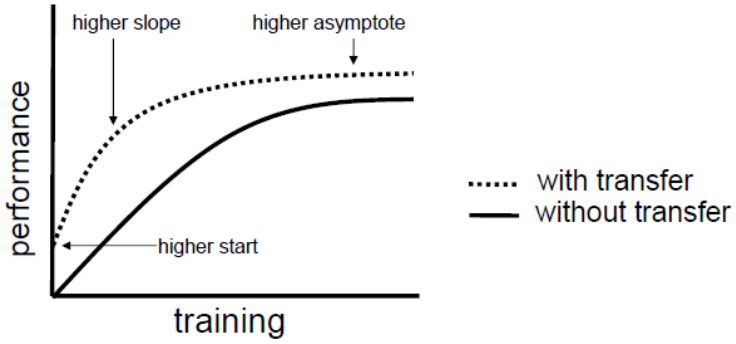


FIGURE 2.3: Transfer learning can improve the quality of the learning process in three measures. *Source:* Torrey and Shavlik, 2010

First, thanks to transfer learning, the better performance scores will have already improved the quality of learning on initial iterations for a particular task comparing to the model without any previous knowledge because of a more accurate selection of the model's learning parameters or any other transferred information. Second, the higher slope will speed up the convergence of the learning algorithm, because the amount of time needed for thorough learning is less for the transfer learning model than for a just created model. Third, the higher asymptote means that the final performance scores are better for a model trained with transfer learning instead of a model without any initial information.

Transfer learning is not the same thing as a pre-trained model with autoencoder (Schmidhuber, 2015) or a restricted Boltzmann machine (RBM) (Salakhutdinov, Mnih, and Hinton, 2007). In a standard machine learning approach, there are only a dataset and desired results, and the task is to achieve the wanted results by any means. For example, to solve a problem, a neural network could be created, which would learn some greedy algorithms and then it would become a part of an assembly of hundreds of other neural networks. However, all these actions will be dedicated only to solving one particular problem. Instead of such a way to solve the problem, transfer learning makes possible sharing of information about details of the source model to improve the performance of the current model and to reduce wasting of time for model creation. As it has been noted above, in computer vision ANNs detect some simple features (as edges, lines, others) from an inputted image in their first layers, then, in middle layers, some general shapes are detected, and

certain forms for a specific problem are identified in the last layers. So, using transfer learning technique there is only a need to re-train the last layers and use the first and the middle layers from the previous task without any modifications.

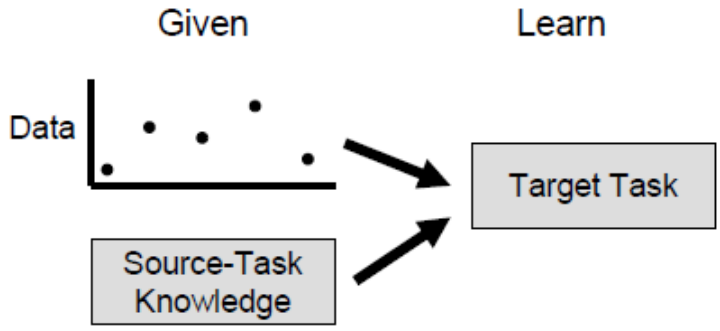


FIGURE 2.4: Transfer learning applies source-task knowledge with machine learning algorithms apart from training data. *Source:* Torrey and Shavlik, 2010

Another transfer learning feature is that gained knowledge can be transmitted only to the new model, as the old one has already solved the problem.

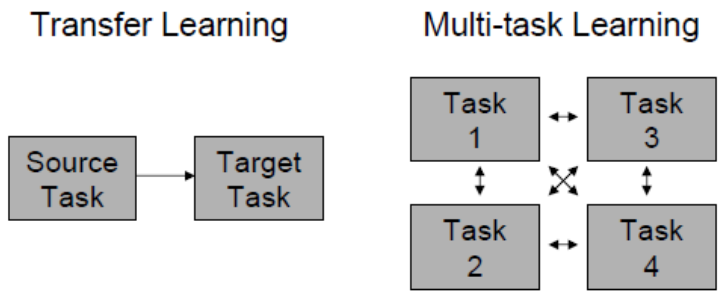


FIGURE 2.5: In transfer learning, source knowledge can be passed only in one direction from the source to the target task; in comparison, another approach, called multi-task learning, can transfer information among all the tasks. *Source:* Torrey and Shavlik, 2010

Also, transfer learning can be considered as a regularisation, because it limits the space of all hypotheses only to the valid and good ones, what is pictured in fig.2.6. To clarify and to remind, supervised learning is a process of studying with labelled examples and correct answers. It is also called learning with a teacher (Russell and Norvig, 2016). Meanwhile, learning on the examples is sometimes called *inductive learning*. That is why transfer learning can be named as *inductive transfer* (West, Ventura, and Warnick, 2007). A model that uses inductive learning algorithms should output correct results on training and testing data, as well as on real-world data. To create a model with such generalisation ability, a learning algorithm needs to have an inductive bias - a collection of assumptions about training data distribution in the real world. Then it is possible to say that transferring knowledge in the inductive learning allows the information gathered while learning the old model, to impact on results of a new model even while solving a new different task.

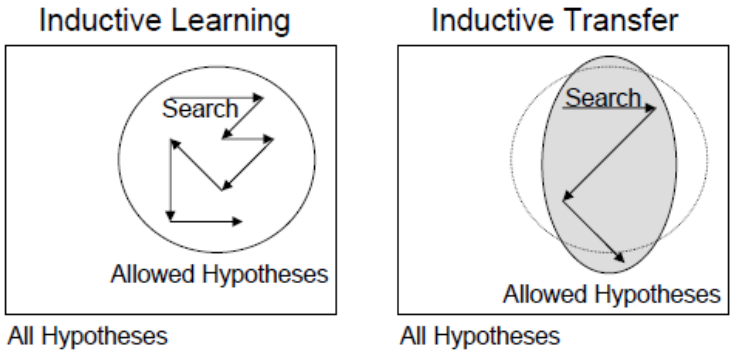


FIGURE 2.6: Inductive learning can be considered as a directed search through a specific hypothesis space (Mitchell, 1997). Inductive transfer uses source-knowledge to regulate inductive bias, that can modify the hypothesis space *Source:* Torrey and Shavlik, 2010

Another key thing to remember about transfer learning is that it can also be adverse. If transfer learning is reducing model performance, then a negative transfer is happening. One of the critical issues in the development of transfer learning methods is to generate positive transfer between respectively selected tasks preventing from negative transfer between less related tasks.

In this work, transfer learning was used because of the absence of enough amount of labelled training data for a neural network to perform well on analysing pages of NTSh's publications.

Chapter 3

Related Works

This chapter presents an overview of methods that are currently used for document layout analysis (including block segmentation and classification) and other methods, that we adopted for this certain work. We will start with the classic techniques and proceed to neural network approaches.

3.1 Run Length Smoothing Algorithm

Run Length Smoothing Algorithm (*RLSA*) is a technique used for block segmentation in Document Image Processing domain. This method consists of one operation - dividing a page into blocks in such a way that one block covers only one type of data (text, graphics and others). Besides, RLSA can be applied to an image in a row-by-row or column-by-column approach.

First of all, a document page is digitised and converted to a binary image. A binary image is a black-and-white image. For simplicity suppose that in a binary image, white pixels are stored as zeroes, and black are ones, respectively. As can be seen, such an image is simply a binary sequence. The RLSA has rules for transforming input binary image (binary sequence):

- 1's in the input are never changed in the output sequence;
- 0's in the input are changed to 1's in output if a number of 0's in a row is less or equal to the previously established limit.

For example, let x be input sequence,

$$x = 0\,0\,0\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,1\,0\,1\,1\,0\,1\,1\,1\,0\,0\,0\,0$$

with a limit equals to 5. The output y will be equal to

$$y = 1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1.$$

The second rule is called smoothing rule because it merges two subsequences to the one if there is not a big distance between them. When talking about images, there should be various thresholds (values of C that determine the number of pixels that will be connected) for row-by-row and column-by-column approaches, because distances among document objects are highly different vertically and horizontally. If thresholds are chosen correctly, then page blocks of common data will be detected. Black regions, which consist of 1's, will be blocks of segmentation.

Generally, run length smoothing algorithm consists of next steps:

1. Applying horizontal smoothing to the document page with some threshold C_h;

2. Applying vertical smoothing to the document page with some threshold C_v;
3. Applying logical operation AND to results of first and second steps;
4. Applying horizontal smoothing to the result of the third step with a smaller threshold C_a.

However, there are many different implementations of RLSA that has lower computer performance, and instead of four steps performs only three or even two. In the three-step approach, the first two steps are switched, so that vertical smoothing is applied first. The third step is horizontal smoothing performed by using algebra theory of $A \cap B = A \backslash \overline{B}$, such that A and B are sets. So, the three-step approach looks like this (Shih, 2010):

1. Applying vertical smoothing to the document page with some threshold C_v;
2. If the amount of 0's in the horizontal direction of the original document image is larger than C_h, then the corresponding pixels in an output image after step 1) are changed to 0's, if not, stay the same;
3. Applying additional horizontal smoothing to the document page with some comparatively little threshold C_a.

Besides, by combining the second and third steps of the previous three-step approach, we may obtain a two-step approach. Also, there is a popular implementation of RLSA on the Internet (*pythonRLSA package*), but it does not perform smoothing operation and modifies the entire input image.

3.2 Fast CNN-based document layout analysis

The process of analysing document structure can be divided into two steps: block segmentation and text discrimination (classifying blocks by their features in classes like text, graphics and others). Some approaches do these two steps simultaneously, and some do successively such that firstly they divide an input image into segments and then classify them. Because of the success of neural networks in classification problems, since Krizhevsky, Sutskever, and Hinton, 2012 was published, neural networks are now used in solving problems of document analysis domain. Augusto Borges Oliveira and Palhares Viana proposed a three-step approach in their paper (Augusto Borges Oliveira and Palhares Viana, 2017):

1. preprocessing a document page and dividing it into its blocks of content;
2. calculating vectors, which will be a sum of horizontal and vertical projections of block content on axes;
3. after training a CNN which had vectors from step 2) as an input, detecting classes to which the contents of blocks belong.

The main result of "Fast CNN-based document layout analysis" paper is a technique, which can classify segments based on vertical and horizontal projections. This techniques performs with the same accuracy as classifying the entire block, but works faster and do not need a significant amount of training data.

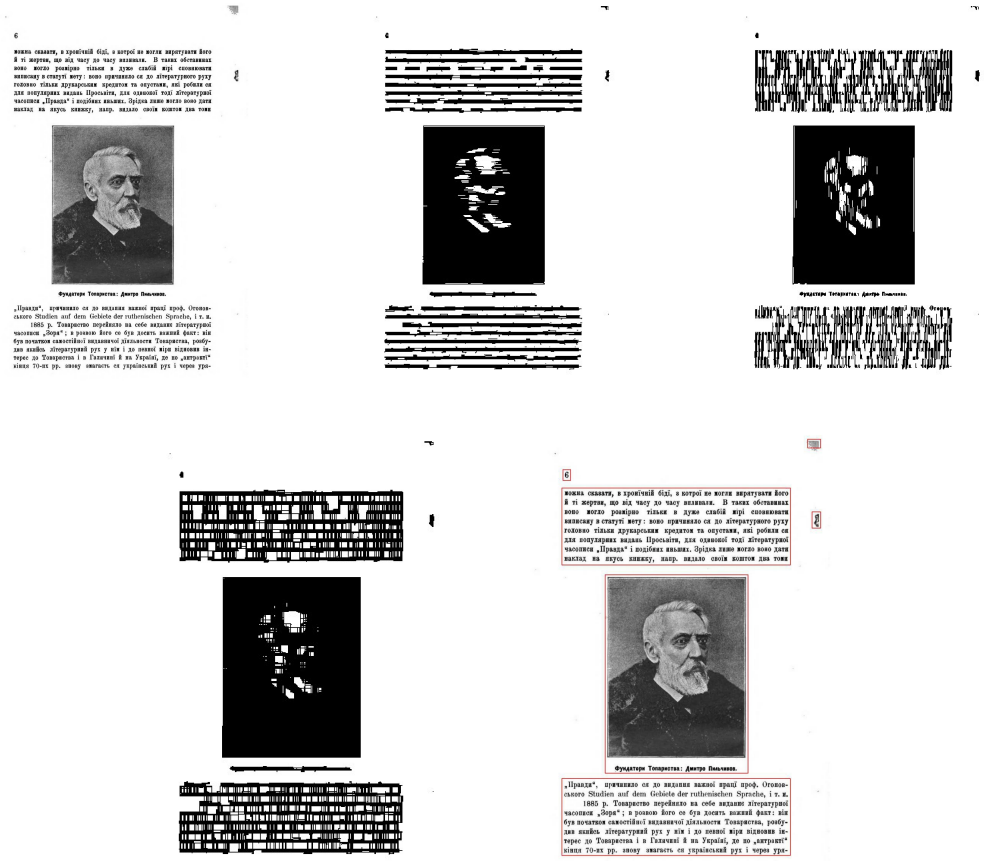


FIGURE 3.1: (From the left to right) First image is a mixed example of a document page with text and image, which was already converted to binary; Second and third images are example of applying RLSA in the horizontal and vertical directions; Third image is result of applying logical AND and the fourth image is result of block segmentation. *Source*: Own implementation of RLSA on a page from the Zapysky NTSh

3.2.1 Selecting blocks with content from a document page

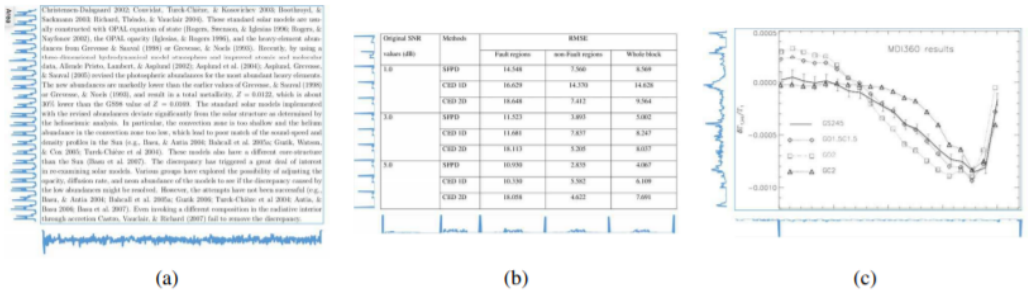
Before classifying text blocks on a document page, the entire page needs to be segmented into smaller regions of interest. For this, the next steps need to be done (see fig.3.2 below):

1. Convert an input page to a binary form;
2. RLSA applies to the result of step 1) with horizontal and vertical directions and then obtained binary images are summed with logical operator AND;
3. A 3×3 dilation operation applies twice over the result of step 2) and converts all pixels in a square 3×3 to the white ones if there is at least one white pixel in that square. This step will combine all parts of one region, in other words, it will create blobs;
4. Received blobs are denoted as rectangles, and these rectangles are our wanted blocks.

FIGURE 3.2: The process of segmenting a page into blocks of content. a) Converted to grayscale mode. b) Result image after applying RLSA (Wong, Casey, and Wahl, 1982). c) Result image after applying twice dilation by a 3×3 mask. d) Resulting blocks of content. *Source*: Augusto Borges Oliveira and Palhares Viana, 2017

3.2.2 Fast 1D CNN based classification

Received blocks are now needed to be classified. For this input images are resized to 100×100 and their vertical and horizontal projections are calculated. It is easy to notice that the results of different classes projections (text, image, table) that were considered in the article have very different characteristics:



(a) (b) (c)

FIGURE 3.3: Examples of three classes ((a) text, (b) table and (c) image) of blocks of content and their corresponding vertical and horizontal projections. *Source*: Augusto Borges Oliveira and Palhares Viana, 2017

For classifying blocks of content one-dimensional CNN architecture was used which receives as input data horizontal and vertical projections of images as two one-dimensional arrays. Primarily, each of projections is independently gone through a convolutional path that consists of a series of three one-dimensional layers with 50 filters with size 3×1, MaxPooling layer with a kernel size of 2 pixels and a 0.1 dropout and ReLu (Nair and Hinton, 2010) activation function. After that, two paths are connected into one structure, and that structure is gone through a fully-connected layer with 50 inputs nodes and three outputs nodes with 0.1 dropouts and softmax activation function for classification into three classes. Fig.3.4 below shows the architecture of proposed 1D CNN based classification approach.

Besides, authors of this approach created an additional dataset, on which 1D CNN performed with 96.75% accuracy level. In comparison to this approach, the

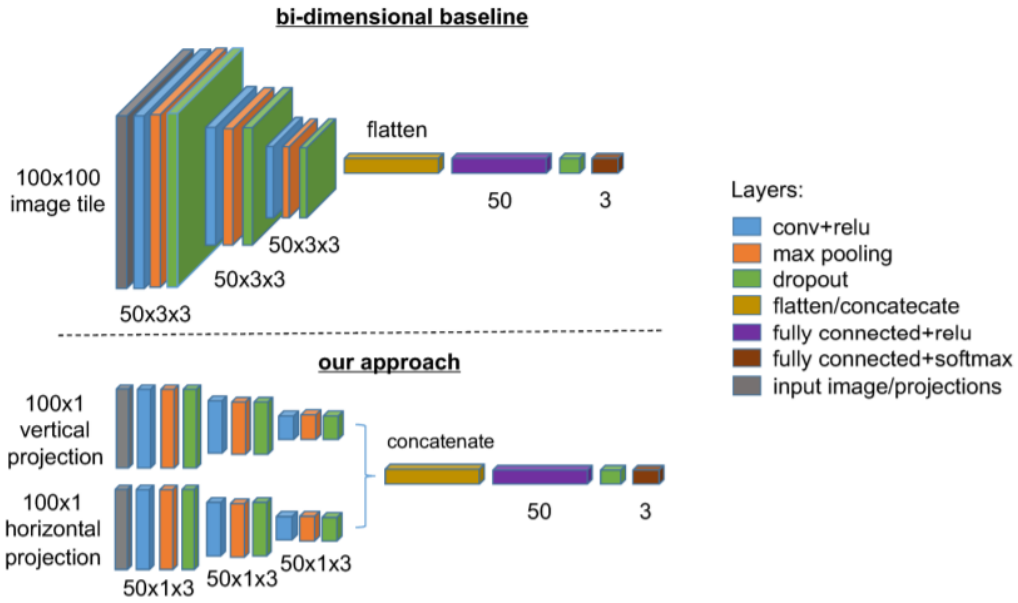


FIGURE 3.4: Comparison of used bi-dimensional baseline and the proposed one-dimensional approach. *Source*: Augusto Borges Oliveira and Palhares Viana, 2017

bi-dimensional CNN model performs with 97.19% accuracy and as input receives the entire image of a page (see the fig.3.4). Authors believe that such difference is minor. Moreover, the processing time for one picture using 1D CNN is 0.783 ± 0.078 secs, and for 2D CNN such time is 6.1 ± 0.223 secs (Augusto Borges Oliveira and Palhares Viana, 2017). These calculations were done on NVidia Tesla K80 GPU and the dataset on which model was trained collected by authors. As can be seen, the one-dimensional approach is faster in about 6.1 times. Also, the performance of the proposed method was compared with other state-of-the-art techniques, but it was not very representative. The reason for this is that datasets on which models were trained is different and sometimes with limited access.

Even though, while working on the one-dimensional approach, the authors determined some cases when the model would most likely make a mistake. As can be seen in fig.3.5, the most often errors were the following: formulas that were classified as an image but labelled as text (fig.3.5 a); problems with segmenting blocks of different data classes (fig.3.5 b); mistakes which were done while manually marking the data (fig.3.5 c) (Augusto Borges Oliveira and Palhares Viana, 2017).

3.3 You only look once algorithm

Object Detection is the subfield of Computer Vision and currently is the most well-studied domain that is widely used in real life, from video surveillance to self-driving cars. Object Detection solves the problem of recognition of the objects on a given picture and also localisation of the detected object on an image.

Using classifiers like VGGNet (Simonyan and Zisserman, 2014) or Inception (Szegedy et al., 2015) was an old approach to object detection. By sliding a window over a

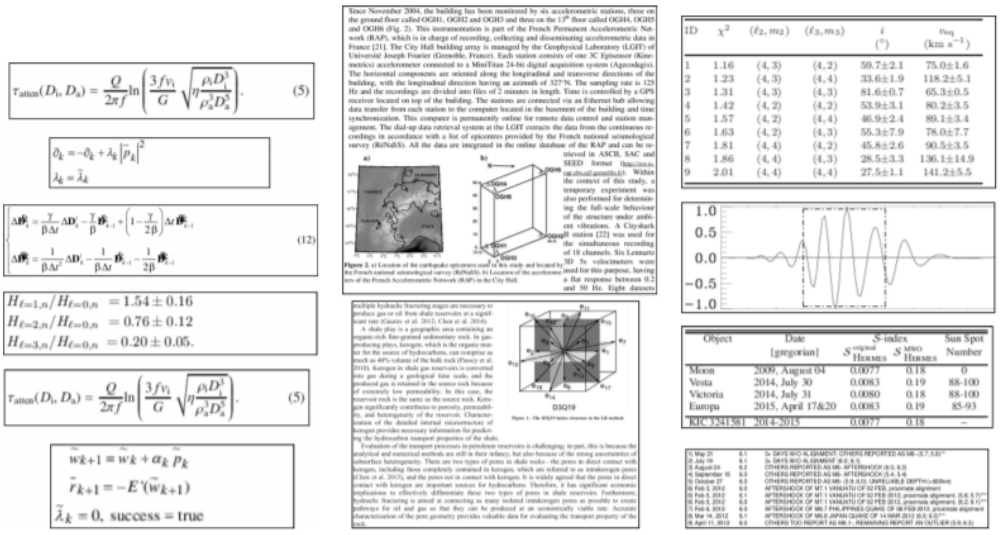


Figure 6. The most common errors found in our methodology are related to: (a) formulas annotated as text and classified as image; (b) problems in blocks segmentation annotated as text and classified as image; and (c) mistakes in manual annotation: the graph as annotated as text and classified as image, the tables were annotated as text and classified as tables.

FIGURE 3.5: The most often mistakes that were found in the model's results. *Source*: Augusto Borges Oliveira and Palhares Viana, 2017

picture, the classifier predicts what is inside a particular window. With such an approach, the classifier will go through every pixel of an image for a few times and will make hundreds of predictions, but will output only the ones with the most significant probability. However, it is a very slow approach. Another possible method is to use a technique called region proposals with a classifier. This technique is about predicting regions of a picture where possibly can be placed interesting information. Then apply classifier only on these regions. Region proposals will be quicker than sliding window, but still, both approaches are slow, because classifier must be run many times.

"You only look once" is the opposite technique. YOLO, as stated in its name, looks at an image exactly one time (Redmon et al., 2016). One of the problems in first YOLO architecture was fully connected layers at the end of the neural network. Later, it was proven on the real-world data that fully-connected layers decrease the performance of a model because of long training time, and they create constraints for input and output data. By YOLO algorithm, an input image is divided into an $N \times N$ grid. Then over each cell, some number of bounding boxes are created. A *bounding box* is a rectangle placed over a cell with a centre in it. Each bounding box has five elements: width, height, offsets to the corresponding cell, and a box confidence score. The *confidence score* for each bounding box shows with which confidence bounding box contains an object. Confidence score knows nothing about what type of object is placed in the box; it merely shows whether shapes of the bounding box are good enough. Besides, each cell makes only one prediction about the class of the object, which is placed inside the bounding box by making probability distribution for all other possible classes. The confidence score for a bounding box multiplies by the class probability and gives a final score. The final score provides us with the confidence that a particular bounding box contains a specific type of object. Most of the bounding boxes will have small confidence level, and because of this, they will not

be shown in the result. *Non-maximal suppression (NMS)* is a technique that compares each bounding box by its score and nullifies any of the overlapping boxes. In other words, this technique chooses the best prediction (Hosang, Benenson, and Schiele, 2017).

There are three versions of YOLO, and each is an improvement of the previous one. Also, there are many implementations based on the improvements of the original object detection YOLO algorithm, such as Tiny YOLO, Fast YOLO, and Single Shot Detector (Liu et al., 2016). YOLO and all other future improved architectures show their performance well not only on the problem for which they were initially created but also for more specific tasks (for instance text recognition in the wild, facial recognition). In cases, when an only small amount of data is available, we pre-train YOLO model on bigger datasets and then train via transfer learning on the related problem.

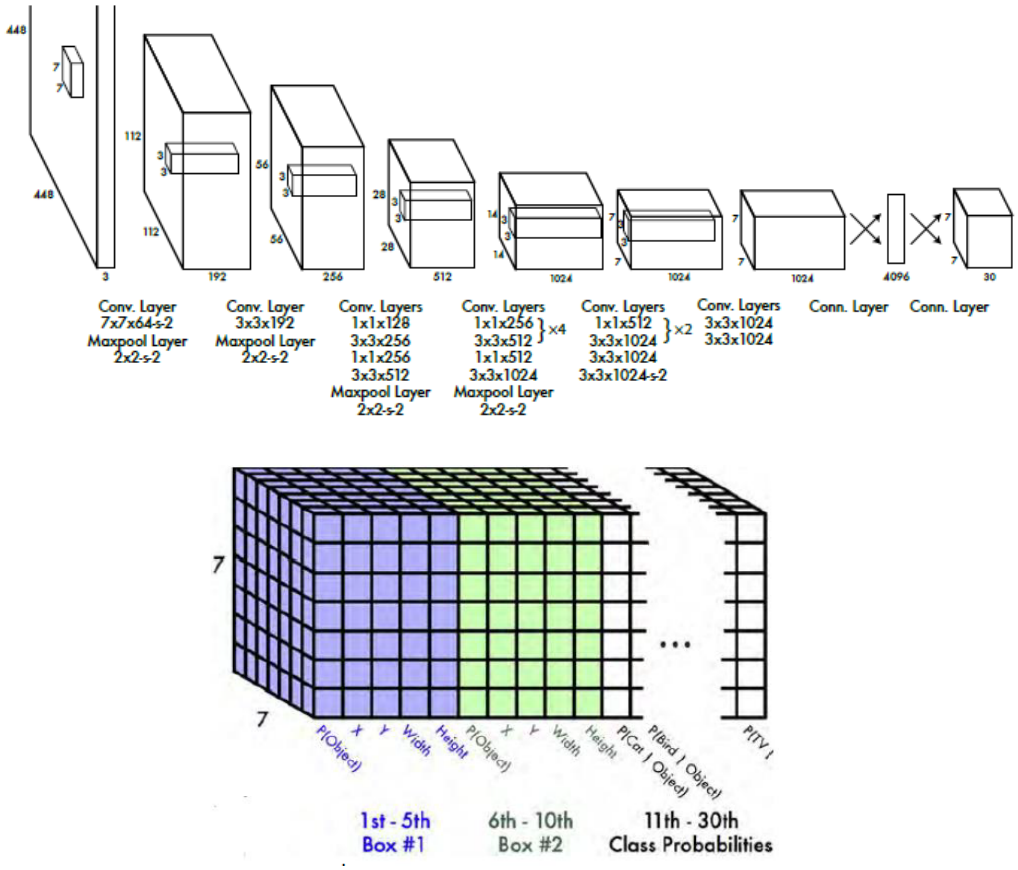
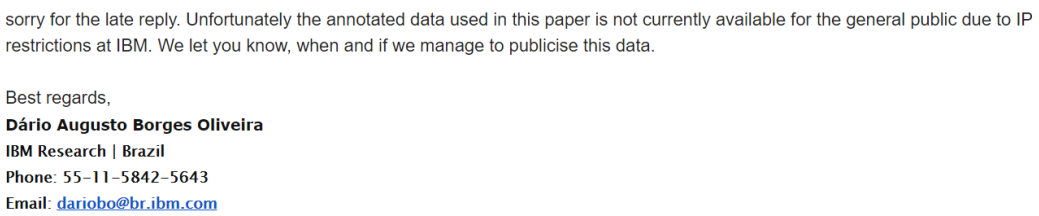


FIGURE 3.6: The architecture of YOLO v1. *Source*: Redmon et al., 2016

Chapter 4

Datasets

This chapter presents an overview of datasets that were used in our experiments (which are described in chap.5). Some of the datasets were publicly available while others we collected by ourselves. Even though there many different datasets that are using in training neural networks for problems related to text recognition domain, only a small amount of them are publicly available. For instance, authors of previously described approach Dario Augusto Borges Oliveira and Matheus Palhares Viana noted in their work that "The comparison of document image analysis methods using the same datasets is not simple because some are paid (UW-III), some are unavailable in their home web sites (ICDAR-2009), or do not have the same kind of documents (academic papers) we built in out database (MediaTeam)." (Augusto Borges Oliveira and Palhares Viana, 2017). However, at the same time, when we requested access to their dataset a little less than two months later, we received the next answer (fig.4.1):


sorry for the late reply. Unfortunately the annotated data used in this paper is not currently available for the general public due to IP restrictions at IBM. We let you know, when and if we manage to publicise this data.

Best regards,
Dário Augusto Borges Oliveira
IBM Research | Brazil
Phone: 55-11-5842-5643
Email: dariobo@br.ibm.com

FIGURE 4.1: Answer from creators of the needed dataset. *Source*: personal email

Next, we will describe the datasets to which we got accessed during the work on this article, as well as datasets collected by ourselves.

4.1 Improving Access to Text Dataset

Improving Access to Text (IMPACT) - is the research dedicated to collecting images from libraries, that are participating in IMPACT (Papadopoulos et al., 2013). This research is a long-term project because more images are added to collections, more variations and conditions are identified and stored. The main goal of IMPACT is to provide the most variety of examples of conditions for any further subprojects. Conditions are any image objects, image structure, the language of the content on an image, fonts, and others. Also, there are many modifications of IMPACT, including a part of the IMPACT Centre of Competence in Digitisation called IMPACT Digitisation (*IMPACT Center of Competence*). The IMPACT Digitisation Image Repository has more than half a million images of typical text-based pages collected from the

Type	Number of documents
Book Page	335,640
Newspaper Rage	142,748
Legal Document Page	80,289
Journal Page	19,573
Other Document Page	18,957
Unclassified Page	5,423
Total Pages	**602,630**

TABLE 4.1: Distribution of document types. *Source:* Papadopoulos et al., 2013

biggest European libraries. The dataset contains even digitised and processed pages of works dated about the 1500th year. Also, in this dataset included material from books, brochures, newspapers and other typewritten works. All of these make the IMPACT Digitisation indispensable source of information for future researches on Image Processing, Optical Character Recognition (OCR) and enriching the language.

A cautiously chosen subset of such images has been improved with corresponding ground truth. Ground truth in image processing is the verification that confirms the specific properties of digital images are appropriate. For instance, it could be done via producing a human transcription of a digital image by an accurate recording of each symbol and word on the image. This ground truth verification will help to evaluate the accuracy of automated image processing.

The IMPACT Dataset of Historical Document Images collected by C.Papadopoulos, S.Pletschacher, C.Clausner, A.Antonacopoulos has about 602,630 images of different document types in a variety of languages (including, English, Polish, Old Church Slavonic, Russian and others). These images of pages were provided from libraries in the United Kingdom, Spain, France, Germany, Czech Republic, Slovenia, Poland and other countries (Papadopoulos et al., 2013). The images are very diverse in their source of origin (see table 4.1).

Each object of an image was labelled and classes of all labels are shown on the table 4.2. As can be seen, there are eight main categories, two of which also have subcategories. To give more clarity to these categories:

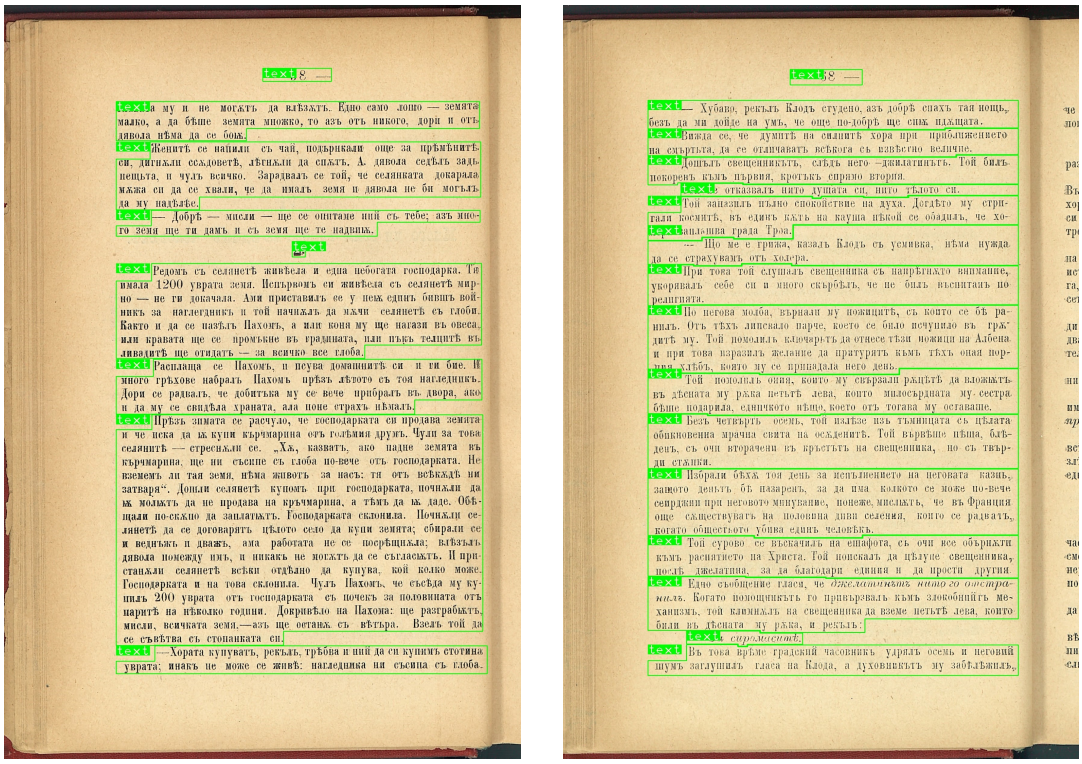
- a heading is words written at the top of a text as a title;
- the paragraph is a part of a text, that has begun on a new line and at least one sentence;
- the footer is a part of the text that printed at the bottom of each page such as a page number, title;
- footnote is a note or reference with additional information connected to the text above and is written at the bottom of a page.

Although these aspects provide the unique and varied dataset, they also bring many problems. For example:

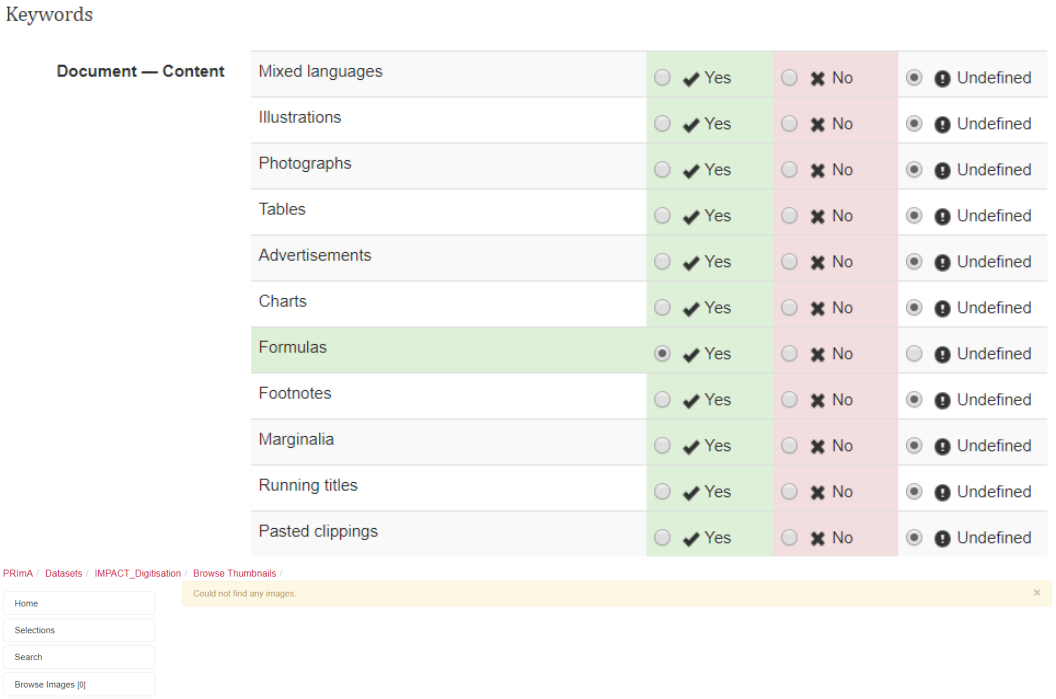
1. Because of many different sources of data, images are labelled differently. In the example below, the left image does not include spaces before and after the text to the labels, but the right image does.

Region type/subtype	**Number**
Text	**573,725**
Heading	42,345
Paragraph	388,636
Drop capital	6,211
Caption	294
Header	35,023
Footer	409
Footnote	2,897
Footnote continued	187
Signature mark	10,642
Catch word	20,678
TOC-entry	6,217
Page number	37,727
Marginalia	11,091
Credit	11,307
Graphic	**10,151**
Logo	4
Stamp	937
Handwritten annotation	2,343
Punch hole	419
Signature	15
Other	6,135
Image	**1,312**
Line Drawing	**8**
Separator	**30,998**
Table	**1,558**
Chart	**5**
Maths	**355**

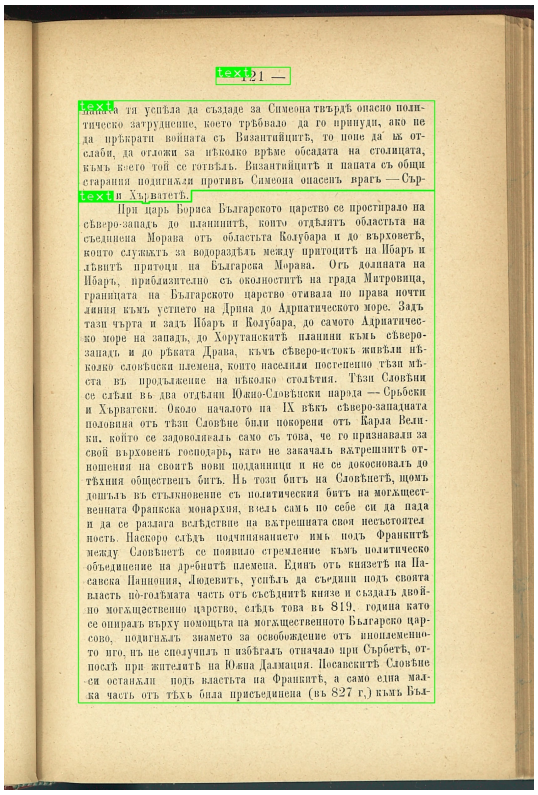
TABLE 4.2: The total number of labelled types and subtypes. *Source:* Papadopoulos et al., 2013



2. Some of the classes are absent when searching on a site.



3. Strange labelling of some paragraphs, in cases, when they include blank regions that they do not belong to.



4.2 Zapysky Dataset

Within this project, we collected a dataset of images from the Zapysky NTSh. At first, we were marking photos with such labels: photo, title, subtitle, text, page_num, footer, separator, reference, author, watermark, caption, table. We used a program called labelling for it. However, at the time when we reached out about IMPACT dataset, we decided to change a little the way of marking data. From that moment labelling method is sticking to the IMPACT data designation and also there is a new label called paragraph which contains information about the author. For a detailed view of labels, see fig.4.3.

4.3 Zbirnyk Dataset

The Zbirnyk dataset is similar to Zapysky dataset but received as a result of marking images from the Zbirnyk NTSh. When we started working on the Zbirnyk dataset we had already known about the IMPACT dataset and decision was made to increase the number of mathematical formulas in the dataset. The reason for this is that we wanted to try how our approaches would work with formulas, but the IMPACT has only 355 math formulas, what is quite a small number for the dataset of size 602,630 labels. So pages of publications exactly on mathematical topics were collected for the dataset. Also, this time marking was done in the Supervisely and labels are the same as in the IMPACT dataset. For a detailed view of labels, see fig.4.4.

IMPACT label	Our Label	Number of labels
Region type/subtype		**628**
Text		**286**
Heading	title	18
Heading	subtitle	28
Paragraph	text	124
Paragraph	author	2
Caption	caption	7
Footer	footer	6
Footnote	reference	8
Page number	page_num	93
Graphic		**312**
Other	watermark	312
Image	**photo**	**10**
Separator	**separator**	**19**
Table	**table**	**1**

TABLE 4.3: The total number of labels in the Zapysky dataset.

IMPACT label	1120-2163-1-PB [1]	1108-2162-1-PB [2]
Region type/subtype	**601**	**544**
Text	**375**	**352**
Heading	8	12
Paragraph	258	232
Caption	34	39
Header	30	28
Footnote	9	8
Signature mark	4	4
Page number	32	29
Graphic	**2**	**2**
Other	2	2
Separator	**38**	**32**
Maths	**186**	**158**

TABLE 4.4: The total number of labels in the Zbirnyk dataset.

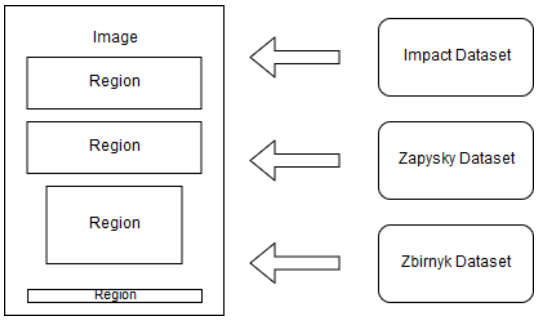
[1] "1120-2163-1-PB " - Збірник НТШ. Том 1. Докази істованя інтеґралів рівнань ріжничкових Володимир Левицкий

[2] "1108-2162-1-PB" - Збірник НТШ. Том 1. Про переступ чисел е і рі Володимир Левицкий

Chapter 5

Implementation Details

To conduct all experiments similarly, without making modifications every time, was decided to create class Image, that would have all information about regions of an input image (each region is an object of class Region). Class Image also has functions for scaling up its instances, make corrections to it and export input image along with its regions, that are also modified appropriately. Along with this, for each dataset were implemented connectors from data class to Image class.



Also, modules for RLSA algorithm, converting PDF files to JPEG and converting instances of Image class to TFRecords were created.

Chapter 6

Experiments

This chapter presents a summary of conducted experiments to see which approach works better on the task of document layout analysis. Before performing anything, we programmatically do data preprocessing of each page scan to remove frames, backgrounds. As in this work, we compare results from different approaches, and firstly, we reproduced the technique described in Augusto Borges Oliveira and Palhares Viana, 2017 on IMPACT dataset and later on the dataset collected by ourselves. Then we check how YOLO works on detecting and classifying block on a document page.

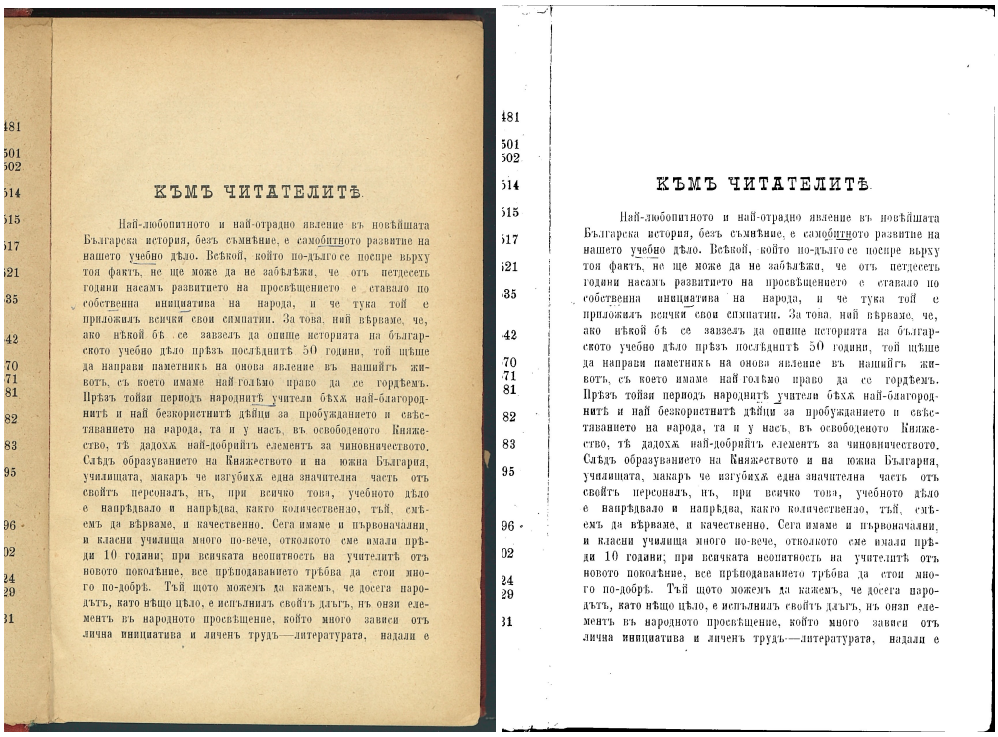
6.1 Preprocessing

Sometimes digitised images have undesirable areas along the edges because of bad scanning. In order to remove those areas and bring all the images to one type, some corrections were implemented. This process consists of the next stages:

1. Background removal:
   (a) The conversion from an RGB image to grey;
   (b) Use of Gaussian smooth to blur an image;
   (c) Use of a binary threshold to check the intensity of every pixel in comparison to a threshold and assign to it 1 or 0 accordingly;
2. Black frame removal (see fig.6.1):
   (a) Find contours of the most significant white area in an image (contours are specified as a curve that connects all continuous dots (along the perimeter) having the same colour or intensity);
   (b) Cut image — we believe that the black frame is only when the largest white area covers more than 90%of the entire image in height and width. Limits were set experimentally;
3. Padding Removal (in other words, white frame removal) (see fig.6.2):
   (a) Removal white frame (all of the white pixels) around the text;

6.2 Fast 1D CNN experiments

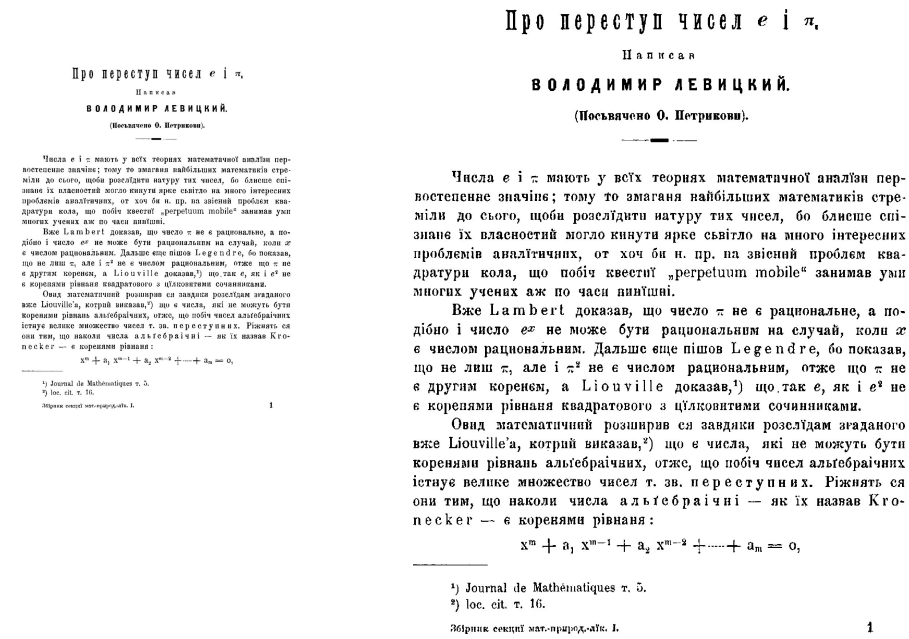
One of the experiments that were done during this work is checking the possibility to use Fast one-dimensional CNN based approach on a more significant amount



КЪМЪ ЧИТАТЕЛИТѢ.

Най-любопитното и най-отрадно явление въ новѣйшата Българска история, безъ съмнѣние, е самобитното развитие на нашето учебно дѣло. Всѣкой, който по-дълго се поспре върху тоя фактъ, не ще може да не забѣлѣжи, че отъ петдесетъ години насамъ развитието на просвѣщението е ставало по собствена инициатива на народа, и че тука той е приложилъ всички свои симпатии. За това, ний вѣрваме, че, ако нѣкой бѣ се завзелъ да опише историята на българското учебно дѣло прѣзъ послѣднитѣ 50 години, той щѣше да направи паметникъ на онова явление въ нашийтъ животъ, съ което имаме най-голѣмо право да се гордѣемъ. Прѣзъ тойзи периодъ народнитѣ учители бѣхж най-благороднитѣ и най безкористнитѣ дѣйци за пробужданието и свѣстяванието на народа, та и у насъ, въ освободеното Княжество, тѣ дадохж най-добрийтъ елементъ за чиновничеството. Слѣдъ образуванието на Княжеството и на южна България, училищата, макаръ че изгубихж една значителна часть отъ свойтъ персоналъ, нъ, при всичко това, учебното дѣло е напрѣдвало и напрѣдва, както количествено, тъй, смѣемъ да вѣрваме, и качествено. Сега имаме и първоначални, и класни училища много по-вече, отколкото сме имали прѣди 10 години; при всичката неопитность на учителитѣ отъ новото поколѣние, все прѣподаванието трѣбва да стои много по-добрѣ. Тъй щото можемъ да кажемъ, че досега народътъ, като нѣщо цѣло, е испълнилъ свойтъ дългъ, въ онзи елементъ въ народното просвѣщение, който много зависи отъ лична инициатива и личенъ трудъ—литературата, надали е

FIGURE 6.1: Results of the two first steps performed on a page from the Zapysky NTSh.



Про перестуи чисел *e* і π.

Написав

ВОЛОДИМИР ЛЕВИЦКИЙ.

(Посьвячено О. Петрикови).

Числа *e* і π мають у всїх теориях математичної аналїзи первостепенне значінє; тому то змагання найбільших математиків стреміли до сього, щоби розслїдити натуру тих чисел, бо ближе спізнанє їх власностий могло кинути ярке свїтло на много інтересних проблємів аналїтичних, от хоч би н. пр. на звісний проблєм квадратури кола, що побіч квестиї „perpetuum mobile“ занимав уми многих ученых аж по часи нинїшні.

Вже Lambert доказав, що число π не є рациональне, а подібно і число e^x не може бути рациональним на случай, коли x є числом рациональним. Дальше еще пішов Legendre, бо показав, що не лиш π, але і π^2 не є числом рациональним, отже що π не є другим коренем, а Liouville доказав,[1]) що так *e*, як і e^2 не є коренями рівнаня квадратового з цїлковитими сочинниками.

Овид математичний розширив ся завдяки розслїдам згаданого вже Liouville'a, котрий виказав,[2]) що є числа, які не можуть бути коренями рівнань альґебраїчних, отже, що побіч чисел альґебраїчних істнує велике множество чисел т. зв. переступних. Ріжнять ся они тим, що наколи числа альґебраїчні — як їх назвав Kronecker — є коренями рівнаня:

$$x^m + a_1 x^{m-1} + a_2 x^{m-2} + \cdots + a_m = 0,$$

[1]) Journal de Mathématiques т. 5.
[2]) loc. cit. т. 16.

Збірник секциї мат.-природ.-лїк. I. 1

FIGURE 6.2: Result of the third step performed on a page from the Zapysky NTSh.

and more diverse classes. In particular, we added mathematical formulas as an additional class. Below in tab.6.1 and tab.6.2 you can see the results. The results were

TABLE 6.1: Train results (1120-2163-1-PB)

	Math	**Separator**	**Paragraph**
Math	179	0	6
Separator	0	38	0
Paragraph	12	5	358

TABLE 6.2: Test results (1108-2162-1-PB)

	Math	**Separator**	**Paragraph**
Math	148	1	5
Separator	0	32	0
Paragraph	16	5	331

obtained by training on 1120-2163-1-PB (part of the dataset) through 100 epoch with SGD optimiser and parameters lr=0.001 and momentum=0.9. We chose the other three classes than in the original Fast CNN-based document layout analysis, but the result is comparable. In Fast CNN-based document layout analysis, we have an accuracy of 96.75%, and here we have 94.98% on the test data.

6.3 YOLO

Most approaches for document layout analysis use two steps, but we decided to reduce it to one step using more modern architectures. That is why we decided to use SSD_Inception model (a descendant of YOLO) with pre-trained weights ImageNet (Deng et al., 2009). Below, in tab.6.3, you can see the results. The performance of transfer learning, even for the case of one class classification is quite pure. The results show that the algorithm is bad even on segmenting a page, not even talking about classification.

TABLE 6.3: Results of using transfer learning with YOLO

	Train	**Test**
mAP@0.5IOU	0,625412	0,105045

Chapter 7

Conclusions

This work is about segmentation and classification blocks of a document page scan. The primary motivation for this work is to prepare data for further analysis of Shevchenko Scientific Society publications. Because we believe that NTSh's works have invaluable information that should be researched and processed. That is why the main goal was to develop the system which would be able to analyse document structure. However, this is not an innovative task, and there are already existing solutions for it. Firstly to test already existing approaches and our suggestion, we used publicly available IMPACT dataset but later decided to collect a new one from NTSh's publications. This decision was made because of marking specifics in this dataset and not enough amount of some classes. For instance, the IMPACT dataset has only 385 formulas (which is a small amount according to the entire dataset size) and also for some reason scans with these formulas cannot be downloaded from the web-page. That is why, precisely the publications with many mathematical formulas were chosen to increase the number of class labels. The approach that was published on the International Conference on Computer Vision 2017 (Augusto Borges Oliveira and Palhares Viana, 2017) was modified and reimplemented for using in this work and gave pretty good results. Also, we suggested and evaluated our approach of pretrained SSD model, which gave bad results in the end.

Bibliography

"A Basic Introduction To Neural Networks". http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html.

Archive of NTSh. URL: http://chtyvo.org.ua/authors/Naukove_tovarystvo_imeni_Shevchenka/.

Augusto Borges Oliveira, Dario and Matheus Palhares Viana (2017). "Fast CNN-based document layout analysis". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1173–1180.

Caudill, Maureen (1987). "Neural networks primer, part I". In: *AI expert* 2.12, pp. 46–52.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Durieux, Valérie and Pierre Alain Gevenois (2010). "Bibliometric indicators: quality measurements of scientific publication". In: *Radiology* 255.2, pp. 342–351.

Fei-Fei Li, Andrej Karpathy and Justin Johnson (2016). *CS231n: Convolutional Neural Networks for Visual Recognition*. URL: http://cs231n.stanford.edu/.

Garfield, Eugene and Robert King Merton (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. Wiley New York.

Gurney, Kevin (2014). *An introduction to neural networks*. CRC press.

Hosang, Jan, Rodrigo Benenson, and Bernt Schiele (2017). "Learning non-maximum suppression". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4507–4515.

IMPACT Center of Competence. https://www.digitisation.eu/. Managed by Fundación Biblioteca Virtual Miguel de Cervantes.

Journal des Savants. http://www.persee.fr/collection/jds.

Klampfl, Stefan et al. (2014). "Unsupervised document structure analysis of digital scientific articles". In: *International journal on digital libraries* 14.3-4, pp. 83–99.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Liu, Wei et al. (2016). "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer, pp. 21–37.

Mitchell, Tom M. (1997). *Machine Learning*. Publisher: McGraw-Hill Science/Engineering/Math.

M.L. Dudka, Yu.V. Holovatch (2018). "Clandestine Ukrainian university in Lviv". In: *Leopoli Scientific Collection*. (in Ukrainian). URL: http://www.icmp.lviv.ua/sites/default/files/preprints/pdf/1802U.pdf.

Nair, Vinod and Geoffrey E Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

NTSh online. (In Ukrainian). URL: https://ntsh.org.

Papadopoulos, Christos et al. (2013). "The IMPACT dataset of historical document images". In: *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*. ACM, pp. 123–130.

Parchomovsky, Gideon (1999). "Publish or perish". In: *Mich. L. Rev.* 98, p. 926.

Phil. Trans. URL: https://gallica.bnf.fr/ark:/12148/bpt6k55806g/f1.image.

Price, Derek J De Solla (1963). "Little science, big science". In:

"Publishing the Phil. Trans.: the economic, social and cultural history of a learned journal, 1665–2015"." (1963). In: URL: https://arts.st-andrews.ac.uk/philosophicaltransactions/brief-history-of-phil-trans/.

pythonRLSA package. URL: https://pypi.org/project/pythonRLSA/.

Redmon, Joseph et al. (2016). "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Rostaing, Hervé (2003). "Basic principles of bibliometrics. Application to Research Development". In: *The competitive intelligence and industrial vision in the 21st century*.

Russell, Stuart J and Peter Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton (2007). "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 791–798.

Saliency map. https://en.wikipedia.org/wiki/Saliency_map.

Savenko, Victor. "Periodicals and serials Shevchenko Scientific Society (1894–1939)". In:

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117.

Scientometrics and citation index. http://lib.med.edu.ua/home/medicni-vidanna-atestovani-vak-ukraieni/naukometria-ta-indeks-cit. (in Ukrainian).

Shelpuk, Sergiy. "Deep Learning". Notes from the serie of lectures at the Ukrainian Catholic University.

Shih, Frank Y (2010). *Image processing and pattern recognition: fundamentals and techniques*. John Wiley & Sons.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Головач, Ю and фон Фербер, К and Олємською, О and Головач, Т and Мриглод, О and Олємской, І and Пальчиков, В (2006). "Складні мережі". In: Журнал фізичних досліджень 10. (in Ukrainian), pp. 247–291.

Записки НТШ (2019). (In Ukrainian). URL: https://uk.wikipedia.org/wiki/ÐŮÐřÐ¿ÐýÑĄÐžÐý_ÐİÐřÑČÐžÐ¿ÐšÐ¿ÐşÐ¿_ÑĆÐ¿ÐšÐřÑĂÐýÑĄÑĆÐšÐř_ÑŰÐijÐ-Ð¡ÑŰ_ÐĺÐ-ÐšÑĞÐ-Ð¡ÐžÐř.

Купчинський, ОА. "НТШ у Львові". In: Енциклопедія історії України 7. (In Ukrainian).

Ярослав Грицак (2001). "Наукове товариство ім. Т. Шевченка". In: Довідник з історії України (А–Я). (in Ukrainian). URL: http://map.lviv.ua/statti/grycak.html.

Torrey, Lisa and Jude Shavlik (2010). "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, pp. 242–264.

West, Jeremy, Dan Ventura, and Sean Warnick (2007). "Spring research presentation: A theoretical foundation for inductive transfer". In: *Brigham Young University, College of Physical and Mathematical Sciences* 1, p. 32.

Wong, Kwan Y., Richard G. Casey, and Friedrich M. Wahl (1982). "Document analysis system". In: *IBM journal of research and development* 26.6, pp. 647–656.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833.