Representation Learning for Document Image Analysis with Practical Considerations

by

Sherif ABUELWAFA

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 15, 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Mohamed Cheriet, thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Stéphane Coulombe, president of the board of examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Luc Duong, member of the jury
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Mathias Adankon, external examiner
Department of Analytics and Artificial Intelligence, National Bank of Canada

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON DECEMBER 1ST, 2021

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to my supervisor, Prof. Mohamed Cheriet, for his permanent support, inspiration, and guidance during my Ph.D. studies. I fully cherish his generosity in helping me learn a great deal of knowledge about research and academia during these years.

I also want to thank all the members of my Ph.D. committee for the honor of accepting to review my dissertation and their worthwhile feedback.

I would also like to thank all my colleagues at Synchromedia laboratory, who were always very supportive and available to discuss many ideas, and provided lots of precious insights on my work.

I owe special thanks and sincere gratitude to my family -my parents Azza and Esmat, my sister Rodayna and my fiancée Xiaobo- for their constant love, understanding, support and encouragement along the journey. Thanks for always believing in me; words cannot express the gratitude I owe you. This thesis is dedicated to you.

I would like to acknowledge the financial support for different projects discussed herein from the Natural Sciences and Engineering Research Council of Canada (NSERC).

My thanks go to everyone who has been part of this unexpected journey. I hope that you have enjoyed the ride as much as I have.

Apprentissage de représentations basé sur des considérations pratiques pour l'analyse d'images de documents

Sherif ABUELWAFA

RÉSUMÉ

Cette thèse met en place des approches d'apprentissage fiables de représentations d'images de documents qui peuvent relever les défis pratiques du monde réel auxquels est actuellement confronté le domaine de l'analyse d'images de documents. En particulier, deux défis sont relevés: effectuer une analyse efficace sur des ensembles de données à grande échelle et s'adapter à la rareté des données d'apprentissage étiquetées. Les approches proposées visent à améliorer les performances des processus d'analyse d'images de documents lorsqu'elles sont appliquées à des cas d'usages réels. À cette fin, nous abordons les défis pratiques dans deux tâches principales pour l'analyse d'images de documents, la classification et la segmentation sémantique.

Les approches actuelles de représentation de documents se concentrent généralement sur des cas d'usages basés sur l'hypothèse irréaliste selon laquelle toute représentation de documents peut être généralisée lorsqu'elle est appliquée sur des ensembles à grande échelle de données de documents. Par conséquent, nous proposons d'abord une approche de représentation de documents pour la tâche de classification de documents qui peut bien se généraliser pour de tels ensembles de données à grande échelle. Le processus de classification dans cette tâche est basé sur l'existence d'un objet visuel local distinctif (par exemple, une note de bas de page) dans l'image du document, ce qui est très pertinent pour divers cas d'usage dans le domaine de l'analyse d'image de document. L'approche proposée est appliquée à des ensembles de données qui contiennent plus de 32 millions d'images de documents et montre une performance fiable et constante dans divers ensembles de données en utilisant moins de 0,07 % des échantillons de l'ensemble de données pour l'entrainement.

De nombreuses approches récentes d'apprentissage de représentations sont basées sur l'apprentissage supervisé des caractéristiques, ce qui nécessite pour l'entrainement une grande quantité d'images de documents étiquetées pour obtenir des performances fiables. Cependant, dans les cas d'usages réels, la quantité disponible de données étiquetées est très limitée et rare, tandis qu'une grande quantité de données non étiquetées est souvent abondante. Nous proposons donc, pour la tâche de classification des documents, une approche d'apprentissage de représentations de documents capable d'apprendre des caractéristiques uniquement à partir de données non étiquetées, et sans aucune dépendance à des caractéristiques conçues manuellement. Contrairement à notre travail précédent ci-dessus, le processus de classification dans ce travail est basé sur le contexte global de l'image du document. Notre approche utilise des données non étiquetées pour apprendre une représentation qui est utilisée ultérieurement pour la classification de documents, soit avec peu de données étiquetées, soit sans données étiquetées. L'efficacité de l'approche proposée et l'amélioration des performances qui en découle sont démontrées par les résultats expérimentaux obtenus.

En considérant chaque document précédemment classifié, nous proposons enfin une approche d'apprentissage de représentations de documents pour la tâche de segmentation sémantique du document afin d'obtenir une interprétation supplémentaire du contenu de ce document et de le préparer pour d'autres tâches d'analyse. Cette approche est capable d'apprendre des caractéristiques à partir de données non étiquetées sans nécessiter de données annotées, de techniques heuristiques dépendantes des ensembles de données ou d'informations textuelles. En outre, il s'attaque au défi bien connu des similitudes inter-classes élevées entre les différentes classes sémantiques. Des expériences sur divers ensembles de données publiques démontrent l'efficacité de l'approche que nous proposons en produisant de meilleurs résultats que les approches précédentes.

Mots-clés: analyse de documents, apprentissage de représentations d'images de documents, classification de documents, segmentation sémantique de documents.

Representation Learning for Document Image Analysis with Practical Considerations

Sherif ABUELWAFA

ABSTRACT

This thesis sets up reliable document image representation learning approaches that can stand up to the practical real-world challenges currently facing the document image analysis field. Particularly, two challenges are tackled, performing efficient analysis on large-scale datasets and adapting to the scarcity of labeled training data. The proposed approaches aim to improve the performance of the document image analysis processes when applied to real-world use-cases. For this purpose, we address the practical challenges in two main tasks of document image analysis, classification and semantic segmentation.

Current document representation approaches usually focus on use-cases with an unrealistic assumption that any document representation can well generalize when applied on large-scale document datasets. Therefore, we first propose a document representation approach for the task of document classification that can generalize well for such large-scale datasets. The classification process in this task is based on the existence of a distinctive visual local object (e.g., footnote) within the document image, which is of high relevance to various use-cases in the document image analysis field. The proposed approach is applied to datasets that contain more than 32 million document images and show a consistent reliable performance across various datasets using less than 0.07% of the dataset's samples for training.

Many recent representation learning approaches are based on supervised feature learning, which requires a large amount of annotated training document images to obtain reliable performance. Meanwhile, in real-world use-cases, the available amount of labeled data is very limited and scarce, while a large amount of unlabeled data is often abundant. We, therefore, propose a document representation learning approach for the task of document classification, which is capable of learning features solely from unlabeled data, and without any dependence on hand-crafted features. Unlike our earlier work above, the classification process in this work is based on the global context of the document image. Our approach utilizes unlabeled data to learn a representation that is used later for document classification either with few labeled data or with no labeled data. The efficiency of the proposed approach and its associated performance boost is demonstrated with the obtained experimental results.

Considering each previously classified document, we finally propose a document representation learning approach for the task of document semantic segmentation to obtain an additional interpretation of that document's content and prepare it for further analysis tasks. This approach is capable of learning features from unlabeled data without requiring annotated data, dataset-dependant heuristics techniques, or textual information. In addition, it tackles the common challenge of having high inter-class similarities between different semantic classes. Experiments on various public datasets demonstrate the effectiveness of our proposed approach by yielding better results than earlier approaches.

Keywords: document analysis, document image representation learning, document classification, document semantic segmentation.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

ARI	Adjusted Rand Index
BBox	Bounding Box
BN	Batch Normalization
BoVW	Bag of Visual Words
CDBN	Convolutional Deep Belief Network
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DT	Distance Transform
ECCO	Eighteenth-Century Collections Online
FC	Fully Connected
FCN	Fully Convolutional Neural Network
FV	Fisher Vector
HOG	Histogram of Oriented Gradients
HVP	Horizontal-Vertical Partitioning
IoU	Intersection-over-Union
LDA	Linear Discriminant Analysis
NMTF	Non-negative Matrix Tri-Factorization

NN	Neural Network
OCR	Optical Character Recognition
Proj	Projection
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Units
RF	Random Forest
RNN	Recurrent Neural Network
RQ	Research questions
SC	Shape Contexts
SIFT	Scale Invariant Feature Transform
SP	Spatial Pyramid
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TML	Trainable Multiplication Layers
U-FL	Unsupervised Feature Learning
UL	Unsupervised Learning
U-PT	Unsupervised Pre-Training

INTRODUCTION

Many organizations and institutions around the globe (governments, libraries, companies) currently store millions of valuable documents that need to be well-preserved and easily accessed. Towards these goals, the digitization process has become an increasingly adopted procedure to create digital archives, in which millions of paper-based documents are transformed into document images. This process serves in guaranteeing a preserved digital content that is easily accessible for both experts and the public. In its digitized form, a document image is processed as a digital image; nevertheless, it still inherits the full characteristics of a paper-based document (e.g., layout, textual and graphical characteristics).

To fully take advantage of the obtained document images collections, many processes are needed to well-organize, analyze and interpret such collections in their new digital form. The field of document image analysis works on achieving this. Specifically, it automatically analyzes a document image based on both its overall global structure and its local textual and graphical elements to extract useful information (Baird, Bunke & Yamamoto, 2012; Nagy, 2000; Nagy, Seth & Viswanathan, 1992). Generally, the process of analyzing a document image is composed of various underlying stages such as document classification, segmentation, text recognition, etc. Each of these stages can rely on visual features, textual data or a combination of both of them to perform its process.

This thesis focuses on analyzing the document images using their visual features solely, and without any dependence on their textual data. In fact, this research aims to study two stages of document images analysis, initially a stage on the abstraction level, then an in-depth stage. In particular, in the first stage, an abstract comprehension of each document image's type (class) is performed. This can be achieved by either classifying the documents based on the availability of a specific local feature (e.g., footnote) or classifying the documents based on their global context (e.g., email, advertisement, tax form). Then, in a later stage, each document's elements

are decomposed into different semantic units. This is accomplished through semantically segmenting the content of each previously classified document into various informative regions of interest (e.g., text, table, figure). In fact, this stage is crucial for achieving more efficient and reliable document analysis. For instance, further characters recognition process can be applied to the obtained text regions using Optical Character Recognition (OCR).

The two studied document image analysis stages are investigated keeping in mind the practical challenges and faced real-world constraints. In fact, part of this research work has been deployed in two joint projects, "Digging into Data" [1] and "The Visibility of Knowledge" [2], in collaboration with humanities researchers at McGill University (Canada), Stanford University, and University of Virginia (USA). In particular, efficient computational analysis algorithms for large-scale historical collections are developed to analyze document images and surface underlying connections better. This has required efficient cooperative efforts between historical documents experts who identify the requirement and study the results, and computer science experts who design and develop relevant algorithms that can analyze the document image on the pixel level. Such collaboration has led to identifying the practical challenges facing the document image analysis field when interfacing with real-world constraints. Specifically, performing efficient analysis on large-scale datasets and the lack of annotated training samples are the two main constraints studied in this work.

Performing these many stages of document analysis in a manual setting is a very challenging procedure that is extremely expensive in terms of time and human resources. Therefore, many research works have addressed automating the analysis process with the help of different computer vision and machine learning-based algorithms. Specifically, considering the enormous increase in the machine learning algorithms efficiency and reliability in the past decade, we plan

[1] https://txtlab.org/2014/01/digging-into-data-global-currents/

[2] https://txtlab.org/2016/09/the-visibility-of-knowledge/

in this thesis to map the recent advances in the theoretical machine learning community and bridge it to address the practical challenges of the document image analysis community.

In general, for any machine learning algorithm to be efficient, a proper representation has to be obtained. One of the essential steps to do so is to find the best local features to be captured from a document image. Those features should be able to get the most out of the input data by disentangling the data's fundamental descriptive elements. This is used to provide an efficient and robust representation that can reflect the real statistics of the document image and boost the performance for the different analysis stages (e.g., classification, segmentation).

In the context of document analysis, a document image representation is mainly about obtaining an intermediate representation that provides an abstraction layer, which goes beyond the image pixels level, to solve later analysis tasks. For this representation to be efficient, it has to lead to small intraclass variations, large interclass separation using a simple decision rule. Achieving this intermediate representation can be achieved through either hand-crafted features (e.g., SIFT, SURF) or learned features. Generally, hand-crafted features need to be designed and adapted to the specific problem and domain to be addressed. In contrast, in feature learning, as features are learned from data, there is no need for domain-specific experts. Together with the very high accuracy attained by feature learning-based approaches, this advantage has attracted a lot of attention in the document analysis field.

0.1 Motivation and problem statement

Representation learning has proven to be an essential part of recent performing approaches in the field of document image analysis in general and document image classification and segmentation in particular. Yet, most of the current approaches are often based on two assumptions: i) that the trained model can generalize well when deployed on a large-scale collection of document images (i.e., the training set is a tiny portion of the dataset, where the test set is thousands

of times bigger than the training set.), ii) that the labeled training samples are available and easily accessible and obtained. Those assumptions might be reasonable and satisfactory for research scenarios but not efficient and effective for practical, real-world applications. This led to this thesis's main question, **what are efficient document representation learning approaches that do not sacrifice performance yet are capable of obtaining representations that can handle processing large-scale datasets, provide reliable generalization during the deployment phase and utilize much less labeled data or only unlabeled data during the training process?**

Despite the recent advances in document image analysis, many challenges are still yet to be solved. The nature of these challenges can be grouped into two categories. Practical challenges that relate to the real-world practical problems that currently face researchers. And technical challenges that relate more to the unique nature of document images.

0.1.1 Practical challenges

Large-scale dataset: A lot of large-scale digitization operations have a compelling need for scalable document image representations that can be reliable and efficient. Meanwhile, the design of current document representations approaches is often limited to a specific research-oriented data scale that is very small, in practice, comparing to the actual requirements of real-world applications. Therefore, performing large-scale analysis on real-world datasets containing millions of document images is an extremely difficult task that has very few precedent works.

Limited availability of labeled training data: Labeling data is an expensive process in terms of both time and labor. Yet, most of the current state-of-the-art approaches in document image analysis are based on supervised learning. In these learning approaches, large amounts of labeled data (i.e., thousands or tens of thousands of document images) need to be manually labeled to train the learning algorithms and obtain efficient performance. Generally, more labeled

training samples lead to better performance. While obtaining more labeled data is possible in some research scenarios, it poses a serious challenge for practical, real-world applications. In fact, manually annotating data is an expensive, time-consuming process since hiring experts to provide the correct labels is a costly process that takes so much time. Therefore, it is preferred to consider having less dependence on this process whenever possible or even avoid it entirely.

0.1.2 Technical challenges

High intraclass variability

In document images, the intraclass variability is so high for many classes. For instance, due to the wide variations in writing styles, fonts, and scales in a collection of manuscripts, it is challenging to learn representations that can well classify hand-written memos or semantically segment classes like paragraph, header, or section. Examples of such variations for the header class are shown in Fig. 0.1. In addition, document images incorporate high variability in the layout types for the same class, ranging from very simple ones (e.g., printed forms) to very complex ones (e.g., historical document images).

Low interclass variability

Many classes for document images have very similar visual characteristics, making it hard to learn representations that can well discriminate between those classes. For instance, considering semantic segmentation, semantic classes like paragraph, list, caption, and section are all text-based. For instance, the visual characteristics that can differentiate a paragraph from a list or a caption from a section are very subtle.

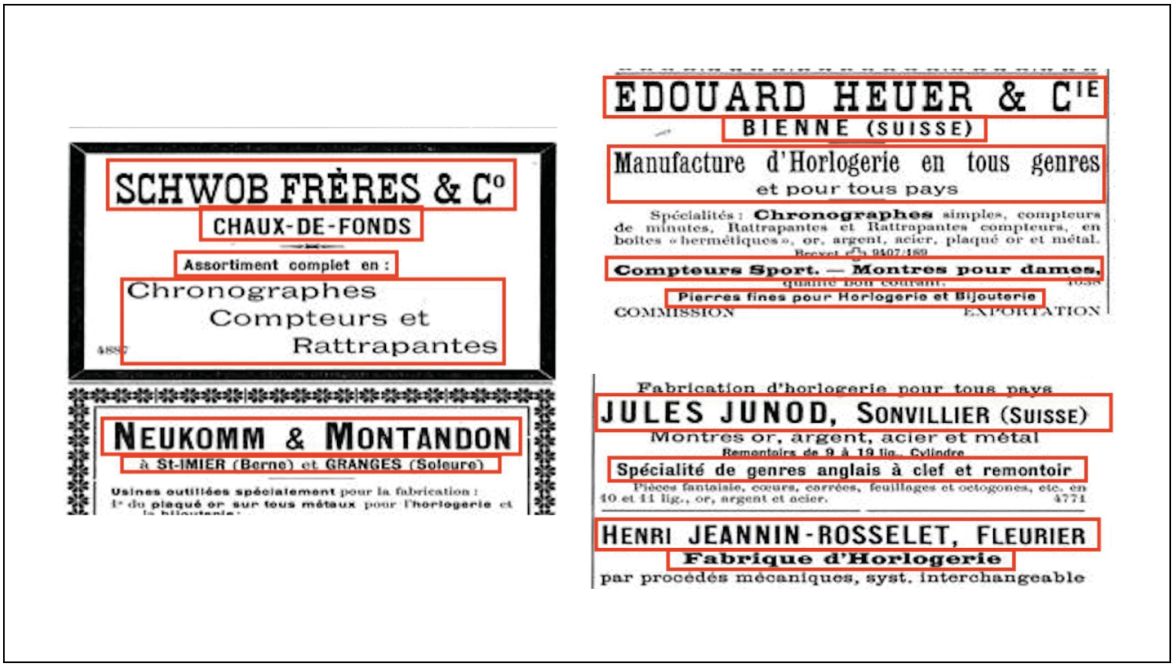


Figure 0.1 Document images with high intraclass variability of the 'header' semantic class (with red contours). Taken from the DSSE-200 dataset (Yang *et al.*, 2017)

0.2 Research questions

To address the discussed problem statement, we detail three specific research questions (RQ) in this section. Our research papers (Chapter 3-6) will offer detailed answers to these three questions.

0.2.1 Research Question (RQ1)

1. What are the practical, real-world challenges currently faced by humanities researchers when working on analyzing document image collections?

2. What kind of document image representations that can generalize well to large scale dataset (e.g., 32 million document images) using less than 0.07% of the dataset samples for training?

3. How to effectively capture specific visual features that relate to a distinctive local visual object within the document image (e.g., a footnote) and classify the document based on it?

0.2.2 Research Question (RQ2)

1. What is an efficient document representation learning approach that is capable of learning features from unlabeled data and improves the document image classification performance in both supervised and unsupervised settings without the need for any extra labeled data?

2. How to effectively capture the global context of the document image considering all its local visual features?

0.2.3 Research Question (RQ3)

1. What is an efficient document representation learning approach that is capable of learning features to semantically segment a document image using only unlabeled data during the training process?

2. How to effectively segment documents with semantic classes that contain many discontinuities and white spaces and have high inter-class similarities between them?

0.3 Contributions

Past research has introduced several document image representations for the analysis tasks. As discussed in previous sections, these representations are not optimized to provide reliable performance in the light of practical, real-world challenges. The search for more effective document analysis representations is still an active field of research. Therefore, **the purpose of this thesis is to study and introduce reliable and viable document image representations that can stand up to the practical, real-world challenges of document image analysis.** This will be achieved by investigating, designing, and developing novel document representation learning approaches that consider both practical and technical challenges to improve a spectrum of analysis tasks that range from classification to semantic segmentation. The focus of our research is on three interdependent aspects.

First, this research will focus on well-defining the practical, real-world challenges that currently face the document image analysis field in the context of an active area of research in the humanities discipline. In addition, it will study the effectiveness of utilizing recent representation learning advances in tackling the observed challenges. In this research, the document image classification task will be performed, where the existence of a distinctive visual local characteristic (e.g., footnote) within a document image will be used for the classification process. The main observed challenge is dealing with analyzing document images on a large scale (i.e., analyzing more than 32 million documents) with very limited availability of labeled training data. In light of the observed challenge, the contribution will be made by proposing a document classification approach that can generalize well for such large-scale datasets. This is to be performed while studying various representations based on a broad spectrum of features, ranging from fully hand-designed to hybrid and fully learned features. The results will show the efficiency of our proposed approach in classifying large-scale documents with reliable performance and consistency across different collections using less than 0.07% of the dataset samples for training.

Secondly, this research will focus on the problem of classifying document images considering the practical challenge of having limited -or no- labeled training data. In this research, all the local features inherited within a document image will help in capturing its global context, which will be used for the classification process. The contribution will be made by proposing a representation learning approach capable of consistently boosting the classification performance using unlabeled data to acquire knowledge that is used later to classify the documents either with few labeled data (supervised fine-tuning) or with no labeled data (clustering). It will be shown in the experimental results that our proposed approach leads to a performance boost in both cases of supervised and unsupervised classification.

Finally, this research will focus on the problem of semantically segmenting each previously classified document considering both the practical challenge of unavailability of labeled training

data during the training process and the high inter-class similarities between the different semantic classes. The contribution will be made by introducing an unsupervised end-to-end approach for semantically segmenting a document image without any dependence on labeled data, dataset-dependant heuristics techniques, or textual information. Experimental results will demonstrate the effectiveness of our proposed approach over state-of-the-art approaches.

0.4 Structure of the thesis

This thesis focuses on representation learning for document image analysis topics, its practical challenges, and proposed solutions that can tackle such challenges.

- **Chapter 1** provides a literature review on the common state-of-the-art features and representations to perform document image analysis. Moreover, it elaborates on the current literature's limitations and challenges. Additional literature review regarding the journal publications is provided in chapters 3 to 5.

- **Chapter 2** presents the general methodology of our work and defines the objectives of this thesis, which consider the state-of-the-art challenges.

- **Chapters 3 to 5** demonstrate our journal publications, including the proposed methods and obtained results in this thesis. In **chapter 3**, the practical, real-world challenges that currently face the document image analysis field are emphasized. Afterward, a document classification approach that generalizes well for large-scale datasets is proposed. In **chapter 4**, a representation learning approach that boosts the documents classification performance using solely unlabeled data is introduced. In **chapter 5**, an end-to-end representation learning approach that semantically segments document images using solely unlabeled data is proposed.

- **Chapter 6** includes the general discussion, which elaborates on the strengths and weaknesses of the proposed approaches.
- Finally, **Conclusion and Future Works** summarizes the work achieved in this thesis and offers some recommendations for future works.

CHAPTER 1

LITERATURE REVIEW

This chapter reviews the relevant literature related to features and representations commonly used in document image representation learning approaches. We first discuss the general description of document image features and their common extraction techniques. Then, we review the state-of-the-art approaches for feature learning on document image analysis with a particular focus on recent unsupervised feature learning approaches and their limitations.

1.1 Feature extraction

One of the essential steps to analyze a document image is to efficiently map the intensity values of its pixels into a relevant analysis decision. This objective can be achieved by obtaining the best features to be captured from the raw input document image; a step that helps provide efficient and robust document representation that can reflect the actual statistics of the document image and helps put a steady foundation for further analysis steps (e.g., classifying the document image or semantically segmenting it).

Formally, since the input raw document image x is usually insufficient in providing expressive information to later analysis stages, features have been introduced to obtain an intermediate representation between the raw input data and the targeted analysis step. In that case, a new document representation $\phi \in \mathbb{R}^K$ (with K features) is obtained directly from the n-dimensional raw input $x \in \mathbb{R}^n$ through a feature function $\Phi()$:

$$\phi = \Phi(x) \tag{1.1}$$

Those K features are considered to provide a better high-level representation ϕ of the document image in a way that conveys the document image's leading properties and facilitates the subsequent document analysis tasks. In fact, the most crucial part of any analysis task is to

determine the features to be utilized, where obtaining the appropriate features can make the learning process more efficient.

Most of the traditional approaches for document analysis; such as (Chen, He, Sun & Naoi, 2012) and (Kumar, Ye & Doermann, 2014), depend on carefully hand-designed features (e.g., SIFT (Lowe, 1999), Shape Contexts (SC) (Belongie, Malik & Puzicha, 2002), SURF (Bay, Tuytelaars & Van Gool, 2006) and HOG (Dalal & Triggs, 2005)). Although some of these approaches can perform unsupervised document analysis processes, the utilized features are still heavily engineered by experts based on large amounts of prior knowledge regarding the used data and desired application, which is a very complex process. For instance, some approaches considered the direction of clustering the document images based on their structure. Specifically, for each document image, (Saund, 2011) acquired horizontal and vertical line segments using line detection. Then, to obtain pairwise similarities between different document images, global histograms for the obtained segments are calculated then compared. Afterward, a similarities-based iterative greedy approach is used to perform the document clustering process. This approach is well crafted to work with document forms with a specific structure and can not easily generalize well to different types of document images. Additionally, (Kumar & Doermann, 2013) introduced the horizontal-vertical partitioning-random forest (HVP-RF) model, which is a Bag of Visual Words (BoVW) approach. Both models are based on complex pipelines that depend heavily on traditional hand-crafted features, which are labor-intensive, time-consuming, and cannot generalize well to new problems.

Due to the difficulties related to engineering hand-designed features, algorithms that can utilize data to automatically learn efficient features were needed, an objective that opened the door to feature learning (Bengio, Courville & Vincent, 2013). Such feature learning algorithms are taking advantage of the increasing amount of available data (i.e., whether it is labeled or unlabeled data) to develop document representations that can precisely express the fundamental characteristics of document images. More details regarding the concept of feature learning and the current commonly used approaches are discussed in the next section.

1.2 Feature learning

Generally, the main objective of feature learning is to utilize a set of labeled/unlabeled data to learn a representation ϕ. This objective can be achieved by parameterizing ϕ with Θ parameters,

$$\phi = \Phi(x; \Theta) \tag{1.2}$$

The Θ parameters can be perceived as a process to obtain some prior knowledge about the data. In such a case, the feature learning algorithm works on tuning the parameters Θ to guarantee a representation ϕ that makes the essential characteristics of raw input data more observable.

To train feature learning algorithms, either labeled (supervised learning) or unlabeled (unsupervised learning) data can be utilized. Very limited literature work is addressing learning document image features and representation using only unlabeled data. Nevertheless, there has been enormous work on supervised feature learning, where labeled data are essential for the training process.

1.2.1 Supervised feature learning

Most of the feature learning approaches in the literature are based on learning parameters Θ from labeled input data. For instance, (Afzal, Capobianco, Malik, Marinai, Breuel, Dengel & Liwicki, 2015; Afzal, Kölsch, Ahmed & Liwicki, 2017; Harley, Ufkes & Derpanis, 2015b) are all utilizing a huge amount of labeled data to perform reliable feature learning.

The goal of a supervised feature learning task is to use a set of given labeled training data $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$ in learning an expressive representation that can lead to a further accurate prediction (e.g., classification, segmentation) on a new input data during the inference process.

Generally, acquiring more labeled data leads to better performance, Most of the recent breakthroughs in document feature learning approaches are actually due to the availability of a

large amount of labeled training data. And since obtaining enough labeled data to perform supervised feature learning is often a challenging and expensive task due to the required time and labor for labeling, while a vast amount of unlabeled data is available and easily accessible; the unsupervised feature learning approaches are considered the optimal methods that can best achieve the objective of this research study. As a result, the rest of this chapter will mainly focus on unsupervised feature learning literature for document image analysis.

1.2.2 Unsupervised feature learning

In unsupervised feature learning approaches, only unlabeled training data x is utilized in learning features and their associated parameters Θ, which inherently embrace some characteristics of x. These approaches learn to express and represent the fundamental visual patterns of the dataset using unlabeled data and then utilize such patterns in having a high-level representation of the input document image to be used later for subsequent document analysis processes.

We will focus mainly on discussing contemporary literature approaches, including recent advances in deep learning, considering the relevance of these works to the research interests of this thesis. These approaches can be categorized into five groups, clustering-based, direct mapping (autoencoder), probabilistic, manifold learning, and self-supervised learning.

1.2.2.1 Clustering-based

In clustering-based approaches, a standard clustering algorithm, such as k-means (Jain, 2010), is used to find the centroids $\{\mu_1, ..., \mu_K\}$ of the cluster sets $C = \{c_k, k = 1, ..., K\}$ by minimizing the Euclidean distance between each obtained training sample, x, and the nearest centroid, μ_k, overall the K clusters. This is to be achieved through optimizing the following objective function $J(C)$ till a convergence is obtained:

$$J(C) = \sum_{k=1}^{K} \sum_{x \in c_k} \| x - \mu_k \|^2. \tag{1.3}$$

In (Dhillon & Modha, 2001), a spherical k-means algorithm has been introduced. A similar loss is utilized; however, cosine similarity is utilized instead of the Euclidean distance. These k-means clustering algorithms are exploited in providing a simple easy-to-implement training method for many unsupervised feature learning approaches. For instance, various 'bag of features' models (Csurka, Dance, Fan, Willamowski & Bray, 2004; Lazebnik, Schmid & Ponce, 2006) have utilized k-means with subsets of patches from the training set to construct a visual vocabulary. Additionally, a k-means dictionary learning approach (Coates & Ng, 2012) has been introduced, where a dictionary of filters can be generated.

1.2.2.2 Manifold learning

Manifold learning approaches are utilized whenever the data points are concentrated throughout a manifold (Roweis & Saul, 2000). These approaches learn to obtain a non-linear low-dimensional representation from the input data. For instance, in (Cheriet, Moghaddam, Arabnejad & Zhong, 2013), two manifold learning techniques have been exploited to learn representations for document image shape-based recognition.

Generally, most of these approaches depend on the calculation of the nearest neighbors to construct a neighborhood graph, which is a critical limitation. Specifically, when dealing with a large number of training samples, the needed calculations to obtain the neighborhood graph scale quadratically. In addition, training samples with high density throughout the manifold are needed to obtain reliable representations, which is challenging considering manifolds with high dimensions.

1.2.2.3 Probabilistic

At the probabilistic approaches, all the network variables are either visible or latent. The training process is performed by maximizing the latent variables' likelihood given the visible variables. Various probabilistic approaches have been introduced in the literature. Most of these approaches are based mainly on the famous Restricted Boltzmann Machine (RBM) model

(Hinton, Sejnowski et al., 1986), which exploits a bipartite graph between hidden h_j and visible v_i variables for the training process. Precisely, the basic block consists of two layers that are associated with a set of weights and biases, figure 1.1, in which the latent variables work on acquiring the dependencies between the visible variables. In fact, the variables of the same layer cannot be connected; therefore, it is called 'restricted'.

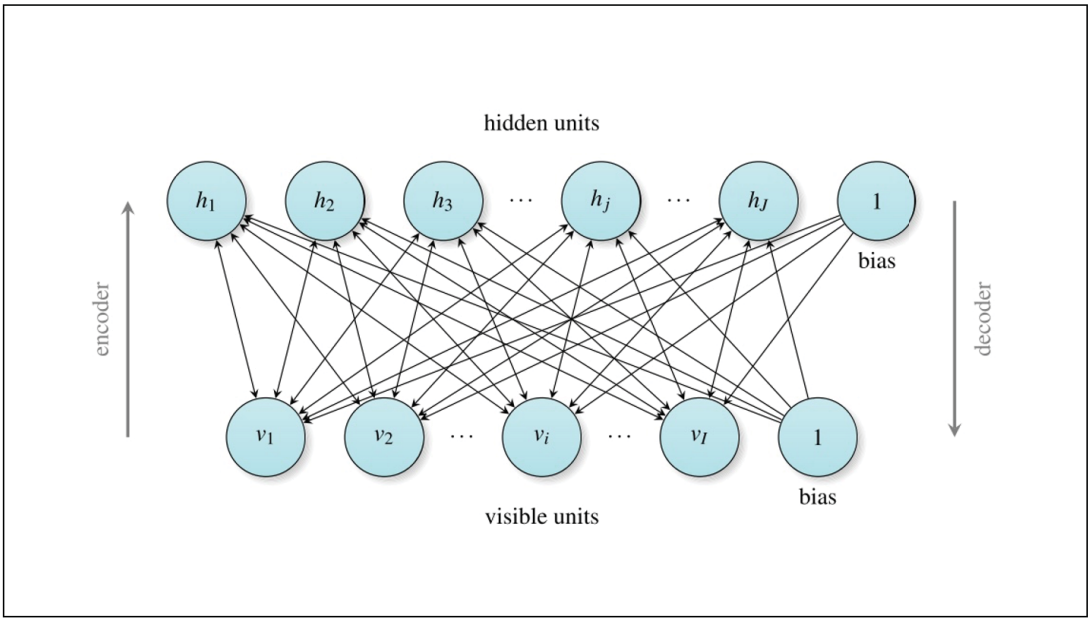


Figure 1.1 A Restricted Boltzmann Machine (RBM) model. Taken from Lopes et al. (2015, p. 158)

In (Hinton, Osindero & Teh, 2006), a Deep Belief Network (DBN) approach has been introduced to reveal the true potential of RBM. It is based on adding up many RBMs together, where the network's bottom layers are expected to detect simple low-level features from the input image. In contrast, higher layers shall unveil more complex abstraction that well-imply the actual remarks of the input image (Le Roux & Bengio, 2008). A layer-wise greedy learning algorithm is utilized for training this approach, where the inputs of a higher layer are the activations of the layer below it. As shown in figure 1.2, an RBM is being trained once per time, where its weights are being frozen once the training process is finished; then, another hidden layer is stacked into that network, and a new RBM training starts on that level.

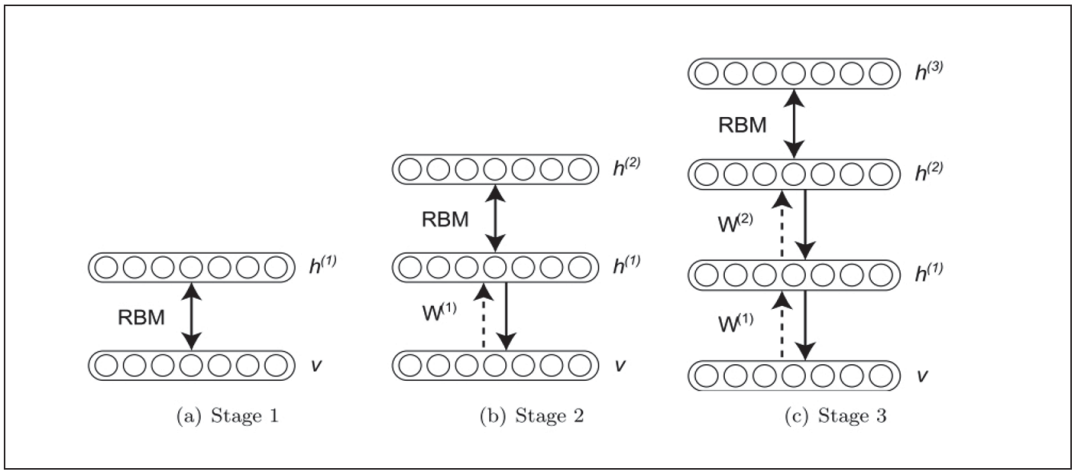


Figure 1.2 A layer-wise greedy learning process of a Deep Belief Network (DBN) with 3 hidden layers. Taken from Le Roux et al. (2008, p. 7)

Although the DBN approach is capable of acquiring many complex features through a diverse range of fields, it still suffers from scalability issues when being scaled up to full-size images with high dimensions. Therefore, a convolutional Deep Belief Network (CDBN) approach has been introduced (Lee, Grosse, Ranganath & Ng, 2009). It is a modified version of the DBN approach that incorporates the convolutional network's locality properties, weight tying, and pooling. Such addition has helped this generative model to be translation-invariant and applicable on large-size images. Nevertheless, these probabilistic approaches have not been common in recent literature. This is due to their inefficiency in learning reliable features, where intractability is common whenever multiple layers are utilized (Noroozi & Favaro, 2016).

1.2.2.4 Direct mapping (Autoencoder)

The direct mapping approach is mainly based on autoencoders (Bourlard & Kamp, 1988; Hinton & Zemel, 1994). An autoencoder consists of two sequential parts, an encoder that performs a parametric feature learning process using unlabeled data and a decoder that maps back the learned features to the input. Specifically, the autoencoder is a neural network consisting of one or more hidden layers with the objective of minimizing the error between the original input

and its reconstruction produced by the network. This process leads to learning a low-dimensional representation of the input document image.

Many variants of the autoencoder have been utilized in the literature for the document image analysis task (Chen, Seuret, Liwicki, Hennebert & Ingold, 2015; Chen, Seuret, Liwicki, Hennebert, Liu & Ingold, 2016b; Wei, Seuret, Liwicki, Ingold & Fu, 2017). Most of these approaches offer better efficient features when dealing with multiple layers architecture, unlike probabilistic approaches (Noroozi & Favaro, 2016). Yet, they do not provide end-to-end, fully unsupervised document analysis solutions. Specifically, after learning the representation, labeled data are still required to train the analysis process (e.g., classification, segmentation, etc.).

1.2.2.5 Self-supervised learning

Self-supervised learning approaches work to obtain an automatic supervisory signal from the available unlabeled training data, commonly by utilizing its underlying structure and its observed characteristics to predict hidden unobserved properties. It equips the training data with a free automatic labeling process, where the manual annotation step is unnecessary.

Many techniques have been introduced in the literature to achieve this objective. For instance, the relative positions between various image patches are utilized in (Doersch, Gupta & Efros, 2015; Noroozi & Favaro, 2016). While, in (Dosovitskiy, Fischer, Springenberg, Riedmiller & Brox, 2016), the model is trained using surrogate classes obtained by augmenting seed images. On the same line, image rotations are predicted in (Gidaris, Singh & Komodakis, 2018). Moreover, color histograms are predicted in (Larsson, Maire & Shakhnarovich, 2016). Finally, in (Hjelm, Fedorov, Lavoie-Marchildon, Grewal, Bachman, Trischler & Bengio, 2018; Ji, Henriques & Vedaldi, 2019), mutual information maximization is exploited. Further discussion is conducted in section 4.2.2.

Regardless of its efficiency and promising performance with natural-images, self-supervised learning approaches are not common in document image analysis' literature.

CHAPTER 2

GENERAL METHODOLOGY

In this chapter, we demonstrate the general methodology of this thesis. The focus of this thesis is on introducing efficient representation learning approaches that can be utilized in practical, real-world use-cases of the document image analysis field. These approaches shall handle two main challenges, first the large amounts of data, and second the scarcity of annotated training samples. The demonstration of that will be on two document analysis tasks, classification, and semantic segmentation. First, considering the limitation and practical challenges of the current literature, three research objectives are defined to be addressed in this thesis. Afterward, the general approach of this thesis is explained.

2.1 Research objectives

The main objective of this thesis is to introduce reliable document image representations that can stand up to the practical, real-world challenges of the document image analysis field. This main objective will be achieved with three specific objectives related to document classification and semantic segmentation tasks.

2.1.1 Objective 1: to study the practical, real-world challenges of the document image analysis field and propose a reliable document representation approach that can generalize well to large-scale datasets.

Current document representation approaches ignore the practical, real-world challenges of the document image analysis field and focus on use-cases with an unrealistic assumption that any trained model can well generalize when applied to large-scale documents collection. Therefore, our first objective is to propose a reliable document representation approach for document classification, an essential document image analysis task. This approach can generalize well to large-scale datasets (i.e., more than 32 million documents) while considering the very limited availability of the annotated samples available during the training process (i.e., less than 0.07%

of the total number of document images). In addition, a comprehensive study of various document representations is proposed. These representations are based on a broad spectrum of features that range from fully hand-designed features to hybrid and fully learned features. The proposed approach is designed and developed as the result of a close collaboration with the humanities researchers, who identify the requirements and analyze the findings. Chapter 3 will describe this collaboration in more details. This approach provides the first study that reflects the practical challenges that face the document image analysis field when interfacing with real-world constraints.

2.1.2 Objective 2: to build a reliable document representation approach that can learn features from unlabeled data for document image classification.

Many recent representation learning approaches for document image analysis are based on supervised feature learning. These approaches require a large amount of annotated training document images to obtain a reliable performance, which is practically a challenge. In real-world use-cases, the available amount of labeled data is limited and scarce, while a large amount of unlabeled data is often abundant. Our second objective is to propose a representation learning approach that is based on unsupervised feature learning. The focus here is still on the document image classification task, where the classification is based on the global context of the document image. The proposed approach to use only unlabeled data to learn a pre-trained model, which is used later to boost the performance of the document image classification in the cases of i) the unavailability of any labeled data and ii) the availability of limited labeled data. More details about the proposed approach will be presented in Chapter 4. It provides the first approach to perform an unsupervised document image classification using a representation that does not depend on any hand-crafted features or labeled data; instead, it is entirely based on feature learning using unlabeled data.

2.1.3 Objective 3: to construct a reliable document representation approach that can learn features from unlabeled data for document image semantic segmentation.

Each previously classified document image in Objective 2 can be semantically segmented to interpret that document's content further and prepare it for additional analysis tasks. Recent representation learning approaches for document image semantic segmentation are mainly supervised learning-based approaches. In these approaches, a large amount of labeled document images are needed for the training process, which is a practical challenge, as previously discussed in Objective 2. In fact, performing unsupervised document image semantic segmentation is a difficult task considering the high inter-class similarities between the semantic classes and the discontinuities and white spaces that most of them contain. Our third objective is to propose an unsupervised end-to-end approach for semantically segmenting a document image (pixel-wise) in a totally unsupervised manner. The proposed approach and its detailed experiments are demonstrated in Chapter 5. It provides the first approach to perform an unsupervised document image semantic segmentation using solely unlabeled data without depending on any textual information or dataset-dependant heuristics techniques.

2.2 General methodology

New document representation approaches have been introduced and developed in this thesis for better obtaining relevant features that can stand against the various practical and technical challenges that currently face document image analysis tasks. These approaches are associated with the objectives previously mentioned, and they can be split into three main themes: document representations for large-scale datasets, document representations for document image classification using datasets with limited to no availability of labeled training data, and document representations for document image semantic segmentation using datasets with no availability of labeled training data.

2.2.1 Document representations for large-scale dataset

The first objective proposes a reliable document representation approach that can generalize well to a large-scale dataset (32 million document images). This approach is to be applied to the document image classification task. The classification is based on the existence of a distinctive visual local characteristic (e.g., footnote) within a document image. In this approach, four models are investigated. Those models range from a conventional model based only on hand-designed features to a fully feature learning-based model. The first model is a 'rule-based' one, which is based on fully hand-designed features. It captures visual features related to some predefined rules based on the expert's prior knowledge of the footnote characteristics, such as its font size, spacing, and location on the page. The produced feature vector is then used for the classification process using a support vector machine (SVM) classifier. The second model is 'layout-based', which is a hybrid model that combines both hand-designed and learned features. This model weighs more heavily on the hand-designed features, where it depends mainly on understanding the text lines layout of the document. Similar to the previous rule-based model, it is based on the hypothesis that the footnote has some distinctive visual characteristics with respect to its size and position on the page. Initially, the obtained measures of each text line are captured by a Discrete Cosine Transform (DCT) as a signal to produce a hand-designed feature vector. Afterward, this feature vector is used as an input to an autoencoder to learn a representation -feature learning-based- used for the classification process. Although the third model is called 'CNN-based', it is still another hybrid model that combines both hand-designed and learned features with more weighting on the learned features. Initially, each document image is represented using the two top text lines and the three bottom ones. Then, a vertical projection of each of these text lines is used to create a histogram. Afterward, a concatenated version of the text lines' vertical histograms forms a hand-designed feature vector used as an input to a one-dimensional Convolutional Neural Network (CNN). In this network, a representation of the document image is learned, and the classification process is performed. The fourth model is based on a CNN architecture and relies on 'transfer learning'. Specifically, the obtained document representation is fully based on learned features without dependence on hand-designed

features. The model consists of two stages, pre-training and fine-tuning. First, an AlexNet is trained on a publicly available large dataset of natural images. The resulting learned parameters (e.g., network weights) are used to initialize the network in the second stage. In this stage, the initialized network is fine-tuned with samples of ECCO dataset to learn document representations that can lead to reliable classification performance. The final classification approach is based on an ensemble of the above four methods (i.e., established upon the majority voting system), where the experimental results show the efficiency of that approach in classifying large-scale (i.e., around 32 million) documents with a reliable performance (Chapter 3).

2.2.2 Document representations for document image classification using datasets with limited to no availability of labeled training data

The second objective proposes a reliable document representation approach to learn features from unlabeled data for document image classification. Unlike the first objective, the classification process in this objective is based on the overall global structure of the document image. The proposed approach is initially based on an unsupervised pre-training step. A convolutional neural network (CNN) is trained on an auxiliary task. Every training example is associated with a different label (exemplar) and expanded to multiple images through a data augmentation technique. Specifically, a set of randomly chosen combinations of pre-defined transformations are applied to each unlabeled training sample in the original training set to obtain a set of surrogate classes. Afterward, the learned pre-trained model, obtained in a fully unsupervised way, is utilized in both unsupervised and supervised document image classification tasks. First, when there is no accessibility to any annotated data -unsupervised classification task-, the learned pre-trained model is used to extract features and obtain representations for the unlabeled training data. These representations are to be clustered for obtaining various cluster centroids. During inference, each test sample is associated with the best-learned cluster centroid. And for evaluation, we consider the labeled data to associate each group of test samples to an actual class. Finally, when a limited amount of annotated data is available -supervised classification task-, the pre-trained model is used mainly for initialization. Then, this model is fine-tuned using the provided small annotated data and utilized for classification. The results show how consistently

efficient our approach is in boosting the classification performance at two different settings: i) as an unsupervised feature extractor to represent document images for an unsupervised classification task (i.e., clustering); and ii) in the initialization of the parameters of a supervised classification task trained with a small amount of annotated data (Chapter 4).

2.2.3 Document representations for document image semantic segmentation using datasets with no availability of labeled training data

The third objective proposes a reliable document representation approach that can learn features from unlabeled data for document image semantic segmentation. The proposed approach has two main stages, data pre-processing and training. It focuses on overcoming the challenges related to the unique properties of the document image and its semantic classes. Initially, at stage one, three steps of pre-processing are applied to each unlabeled input document image. In the first step, the distance transform (DT) is obtained for that document image, which is then concatenated with the existing RGB channels to achieve a combined representation space that utilizes distance transform and RGB information as an input to our network. Specifically, this concatenation step results in learning a novel representation that can acquire information about the spatial white spaces -horizontal and vertical- between the text lines without any labeled data. This leads to overcoming the challenge of having plenty of spatial discontinuities and white spaces in the semantic classes of document images. In addition, the learned representation is efficient in dealing with the low inter-class variability, which is a common challenge with many document images' semantic classes. In the second step, patches of the document images are obtained before pairing them in the third step to be used as an input to the upcoming training stage. In stage two, the training process is performed utilizing a network based on dilated convolutional layers. Moreover, during that process, the obtained distance transform of each image is used to automatically identify the background regions, which helps reduce the bias in the training set without the need for any labeled data. The obtained results show the efficiency of our proposed approach in performing unsupervised semantic segmentation by yielding better results than baseline approaches on various public datasets (Chapter 5).

CHAPTER 3

DETECTING FOOTNOTES IN 32 MILLION PAGES OF ECCO

Sherif Abuelwafa [1], Sara Zhalepour [1], Ehsan Arabnejad [1], Mohamed Mhiri [1], Emilienne Greenfield [2], James P. Ascher [2], Sofia Bach [2], Victoria Svaikovsky [2], Alayne Moody [2], Andrew Piper [2], Chad Wellmon [3], Mohamed Cheriet[1]

[1] Département de génie des systèmes, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] McGill University, Montreal, Canada.
[3] University of Virginia, Charlottesville, Virginia, USA.

Published in Journal of Cultural Analytics, December 3, 2018

3.1 Introduction

In "An Answer to the Question: What is Enlightenment?", the eighteenth-century German philosopher Immanuel Kant responded to a big question buried in a little footnote. But you wouldn't know it, because contemporary editions of Kant's famous essay no longer reproduce the parenthetical directive that Kant's original essay printed right under the essay's title in the December issue of the *Berlinische Monatsschrift* in 1784: "S. Decemb. 1783. S. 516." (See December 1783, p. 516). And, in fact, page 516 in the December volume of the *Berlinische Monatsschrift* 1783 has a footnote: "What is Enlightenment? This question is nearly as important as: what is truth? And should certainly be answered before one starts to enlighten! But I have yet to find it answered anywhere."

Kant's attempt to define enlightenment, then, was a reply to a specific question. The footnote to which Kant's essay refers was published in an essay written by the Berlin pastor Johann Friedrich Zöllner, who had published several sermons in the *Berlinische Monatsschrift*. Zöllner's essay, "Is it wise to no longer sanction marriage through religion?", discussed whether it would be "enlightened" to no longer require clergy to officiate weddings (Pasanek & Wellmon, 2018). Kant's essay was addressed to a larger philosophical question, but also a particular question posed on a particular page in a particular periodical by a particular contemporary. And yet

we tend to read Kant's essay, and, thus, the Enlightenment, apart from these relationships and particular moments of printed address and response.

Footnotes like the one on the bottom of page 516 in Zöllner's essay are visible traces of these mediated relationships, markers of conversations, references, connections, and the sociability of knowledge. They are also visible markers of legitimacy and authority. They demonstrate familiarity, identity, and knowledge. As Anthony Grafton writes with respect to the eighteenth-century origins of the footnote within the nascent discipline of history: "The footnote is bound up in modern life with the ideology and the technical practices of a profession." (Grafton, 1997) It is an element of the history of disciplinarity and scientific credibility.

The footnote is also one of several visual typographic practices that have shaped modern knowledge. Our effort to understand better the footnote's place within the eighteenth century is part of a larger, on-going project that seeks to understand a range of visual practices of scientific notation in the past: whether it be footnotes that communicate authority and the relationality of sources; tables that bring together disparate forms of information into geometric relations; diagrams that provide abstract representations of intellectual procedures or natural phenomena; or illustrations that provide mimetic representations of objects in the world. In each of these cases, authors, editors, and publishers used a graphic process to convey information and make truth claims, often in a way that sought to reduce complexity. And contemporary scholars continue to use these processes in order to communicate well and more efficiently with one another. Instead of reproducing the entirety of another text, we cite it. Instead of reproducing all of the underlying data of a process, we transform it into a table of relations. Diagrams abstract more detailed processes into more formal essence. Even illustrations have an indexical relationship to the larger real-world phenomena they are meant to represent.

As we explain in greater detail elsewhere, our larger project is about bringing together the intersecting strands of research from the fields of book history, the history of science, and document image analysis to better understand the analytical unit called "the page image" and its role in the history of scientific knowledge. Our aim us to take seriously the page image in a

double sense: first, as an image *of* a page, that is, to see the digitized page first and foremost as an image rather than a flawed mediation of text; and second to see the page itself as *an image*, as a visual unit rather than a primarily textual one. What have been the ways that the graphic practices of pages have underpinned the epistemic claims of scientific knowledge?

In this essay, we recount our process of using machine learning and classification algorithms to detect footnotes within the Eighteenth-Century Collections Online dataset (ECCO). ECCO represents one of the most complete digitized collections of a national publishing context within a specific historical period, consisting of over 100,000 volumes and 32 million pages published in Britain between 1700 and 1800. It has become a staple of research in the history of ideas, not just in Britain but for scholars of the Enlightenment more generally. We see the enrichment of collections like ECCO as a primary research goal for furthering historical understanding.

We discuss here the samples of training data that were collected and manually annotated, the different types of page-features that were used in the detection process, and the estimated accuracy of our predictions. The net result is metadata on the presence of footnotes within approximately thirty-two million pages of historical documents, which we share along with metadata regarding the initial training data used so that others can work with the same data. As we detail in Table 3.1, overall we are able to recall pages with footnotes with 67.8% accuracy and of those we achieve a precision of 96%. This suggests that there are a considerable number of footnotes we may be missing but that when we do detect them we do so with a very high degree of confidence. In addition to these summary statistics, we also provide users with an estimated percentage of footnoted pages per document, a table of all page IDs that have predicted footnotes on them, and finally the estimated probability of a footnote being present for all pages in ECCO. We see this as a first step in fully annotating ECCO according to our four visual categories of footnotes, tables, diagrams, and illustrations.

We want to emphasize from the outset just how challenging this process has been. "At first glance, all footnotes look very much alike," writes Grafton, "[but] even a brief exercise in comparison reveals a staggering range of divergent practices." As we quickly learned, discerning

what constitutes a footnote in the eighteenth-century is by no means a straightforward process. Training machines to identify such visual ambiguity is even more difficult. One of the issues that will need further reflection are the trade-offs between the gains of acquiring knowledge at larger scale and the challenges introduced by a fundamental uncertainty surrounding historical evidence.

Overall, we see this project contributing to a larger effort of enriching digitized collections of historical documents with more information about the documents (what is traditionally called "metadata" or data about data). We see this particular effort as contributing to knowledge about the visual qualities of page images, with a specific attention to what we are calling the history of "scientific notation." One of the major obstacles for historical understanding is the minimal amount of knowledge we have about individual documents within large document collections. We might think of this as a second-wave of digital history: the first – which is still on-going – involves the act of digitization itself. This effort is about making physical copies, which are geographically limited in their accessibility, more widely accessible to a broader reading public. The second wave, to which we see our work contributing, can be seen as the attempt to provide more knowledge about the composition of the collections to facilitate large-scale study of cultural history. ECCO metadata currently consists of features like publication date, author, title, publisher, and in some cases subject headings. There is much more that we can do to annotate collections. But to do so at large-scale requires developing algorithmic procedures for expertly labeling documents, which in turn necessitates greater collaboration between the humanities and the sciences.

However, it is important not to mistake these labels for what computer scientists call "ground truth." All knowledge is situated. This project represents the coordinated efforts of a team of 14 researchers split between the humanities and computer science, including both students and faculty, ranging in levels from BA to Masters to PhD to Professor. It entailed a lengthy collaboration to create mutual understanding and shared goals as well as a clear understanding of the cultural object of study (in this case historical practices of footnotes). The training data assembled thus represents the understanding and prior knowledge of the humanities cohort,

while the detection algorithms represent the understanding and prior knowledge of the computer science cohort. Any machine learning process inevitably encodes, explicitly and implicitly, these biases into its outputs and are thus not value free. While this may seem less controversial with a more straightforward visual object like footnotes, it is important that we continue to foreground the human learning behind machine learning. We now proceed to describe the process we used to annotate thirty-two million pages of ECCO.

3.2 What is a footnote? (Training Data)

We began our research by defining a footnote and then identifying positive and negative examples within ECCO. For our purposes we defined a footnote as: Footnotes need to be distinct, marked text at the bottom (foot) of the page that are referenced in the main part of the text.

Each of these components is important: footnotes have a distinct location; they are marked (i.e. have a distinct marker); and refer directly to a location within the main body of the text through a matching mark (Fig. 3.1). Such a definition rules out side-notations (Fig. 3.2) or unmarked commentary that may be located at the bottom of the page (Fig. 3.3). Footnotes require some rule-based distinction of being "off-set". Despite these clarifications, we encountered numerous examples of pages that looked deceptively like footnotes (Fig. 3.4, Fig. 3.5). Because footnote marks are both highly varied in the eighteenth century and also highly indistinct as images (what is the difference between a poorly printed asterisk and an ink blot from the reproduction process?) (Fig. 3.6), the footnote mark, as we quickly learned, is only weakly significant in discriminating between footnotes and commentary. For this reason, we encountered a serious limitation in our analysis that is important to signal at the outset: given the heterogeneity of footnote markers as well as their printed ambiguity (footnotes can be designated by numerous different shapes which are very hard to distinguish from other marks or blemishes on the reproduced page), our analysis does not indicate where in the body of the text the footnote is anchored. In other words, we cannot provide analysis of the footnoted word, but only an estimation of the presence of the footnote itself at the bottom of the page. Further research would be needed to reliably capture the location of the footnote mark as indicated in Figure 3.1.

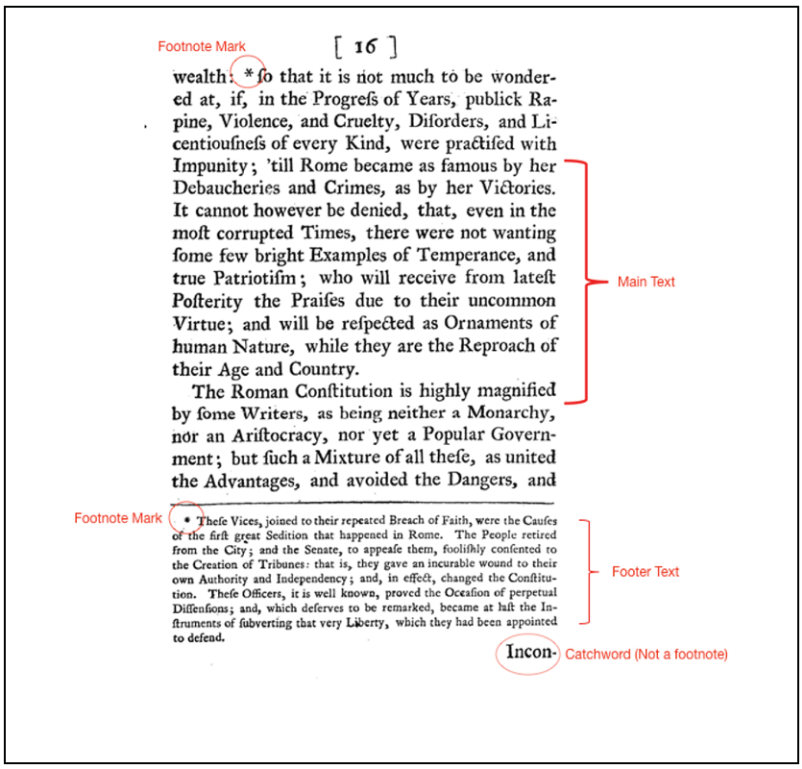
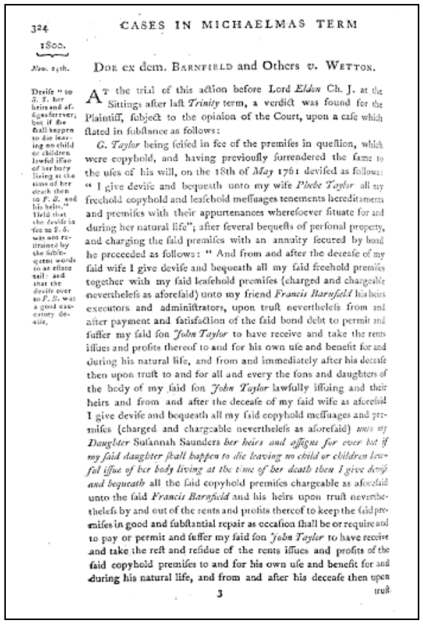


Figure 3.1 Example of a footnote. From Reflections on ancient and modern history (1742)

Based on the above definition and limitations, and with extensive discussions between students and faculty responsible for collecting training data, we manually annotated 21,939 page images for training (6,028 pages with footnotes and 15,911 pages without), and another 5,520 pages for testing (522 with and 4,998 without). All pages were randomly generated from ECCO I and II and then reviewed by a single student. Ambiguous cases were reviewed by the project investigators. As we will demonstrate, our models do not appear to show biases towards different historical timeframes within the overall dataset or between ECCO I and II, which are collected separately by Gale.

3.3 Detecting Footnotes at Large Scale (Machine Learning)

After collecting our positive and negative examples of footnoted pages, we then set out to design features and learning algorithms that could best predict the presence of footnotes on a page. We



324 CASES IN MICHAELMAS TERM

1800.

Nov. 24th.

Doe ex dem. Barnfield and Others v. Wetton.

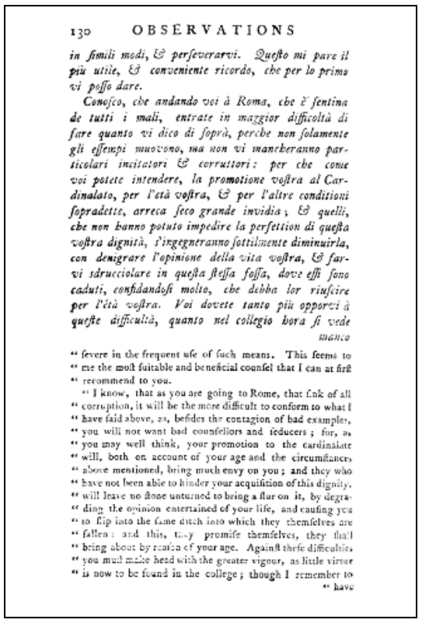
At the trial of this action before Lord *Eldon* Ch. J. at the Sittings after laſt *Trinity* term, a verdict was found for the Plaintiff, ſubject to the opinion of the Court, upon a caſe which ſtated in ſubſtance as follows:

G. Taylor being ſeiſed in fee of the premiſes in queſtion, which were copyhold, and having previouſly ſurrendered the ſame to the uſes of his will, on the 18th of *May* 1761 deviſed as follows: "I give deviſe and bequeath unto my wife *Phebe Taylor* all my freehold copyhold and leaſehold meſſuages tenements hereditaments and premiſes with their appurtenances whereſoever ſituate for and during her natural life"; after ſeveral bequeſts of perſonal property, and charging the ſaid premiſes with an annuity ſecured by bond he proceeded as follows: "And from and after the deceaſe of my ſaid wife I give deviſe and bequeath all my ſaid freehold premiſes together with my ſaid leaſehold premiſes (charged and chargeable nevertheleſs as aforeſaid) unto my friend *Francis Barnfield* his heirs executors and adminiſtrators, upon truſt nevertheleſs from and after payment and ſatisfaction of the ſaid bond debt to permit and ſuffer my ſaid ſon *John Taylor* to have receive and take the rents iſſues and profits thereof to and for his own uſe and benefit for and during his natural life, and from and immediately after his deceaſe then upon truſt to and for all and every the ſons and daughters of the body of my ſaid ſon *John Taylor* lawfully iſſuing and their heirs and from and after the deceaſe of my ſaid wife as aforeſaid I give deviſe and bequeath all my ſaid copyhold meſſuages and premiſes (charged and chargeable nevertheleſs as aforeſaid) *unto my Daughter* Suſannah Saunders *her heirs and aſſigns for ever but if my ſaid daughter ſhall happen to die leaving no child or children lawful iſſue of her body living at the time of her death then I give deviſe and bequeath* all the ſaid copyhold premiſes chargeable as aforeſaid unto the ſaid *Francis Barnfield* and his heirs upon truſt nevertheleſs by and out of the rents and profits thereof to keep the ſaid premiſes in good and ſubſtantial repair as occaſion ſhall be or require and to pay or permit and ſuffer my ſaid ſon *John Taylor* to have receive and take the reſt and reſidue of the rents iſſues and profits of the ſaid copyhold premiſes to and for his own uſe and benefit for and during his natural life, and from and after his deceaſe then upon truſt

3

Deviſe "to J. T. her heirs and aſſigns forever, but if he ſhall happen to die leaving no child or children lawful iſſue of her body living at the time of her death then to *F. B.* and his heirs." Held that the deviſe in fee to *J. T.* was not reſtrained by the ſubſequent words to an eſtate tail: and that the deviſe over to *F. B.* was a good executory deviſe.

Figure 3.2 Example of side notation. From Reports of cases argued and determined in the Courts of Common Pleas (1802)



130 OBSERVATIONS

in ſimili modi, & perſeverarvi. Queſto mi pare il più utile, & conveniente ricordo, che per lo primo vi poſſo dare.

Conoſco, che andando voi à Roma, che è ſentina de tutti i mali, entrate in maggior difficoltà di fare quanto vi dico di ſopra, perche non ſolamente gli eſſempi muovono, ma non vi mancheranno particolari incitatori & corruttori: per che come voi potete intendere, la promotione voſtra al Cardinalato, per l'età voſtra, & per l'altre conditioni ſopradette, arreca ſeco grande invidia; & quelli, che non hanno potuto impedire la perfettion di queſta voſtra dignità, s'ingegneranno ſottilmente diminuirla, con denigrare l'opinione della vita voſtra, & farvi ſdrucciolare in queſta ſteſſa foſſa, dove eſſi ſono caduti, confidandoſi molto, che debba lor riuſcire per l'età voſtra. Voi dovete tanto più opporvi à queſte difficultà, quanto nel collegio hora ſi vede manco

"ſevere in the frequent uſe of ſuch means. This ſeems to "me the moſt ſuitable and beneficial counſel that I can at firſt "recommend to you.

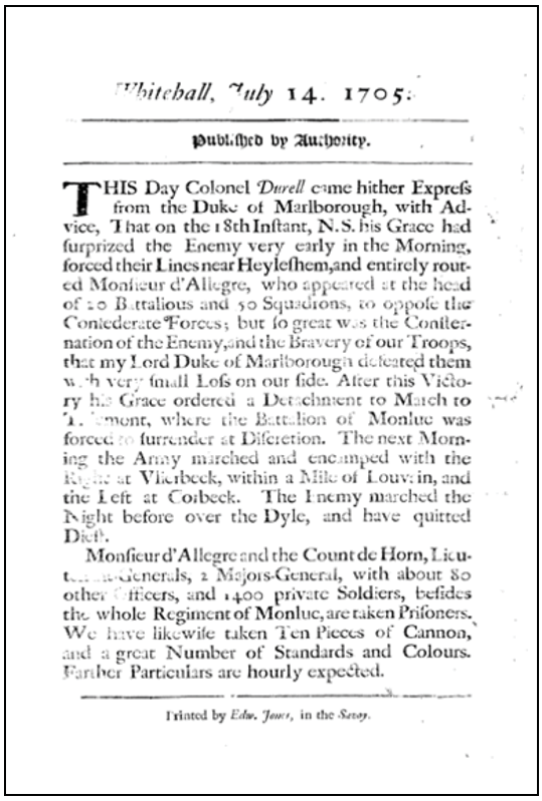
"I know, that as you are going to Rome, that ſink of all "corruption, it will be the more difficult to conform to what I "have ſaid above, as, beſides the contagion of bad examples, "you will not want bad counſellors and ſeducers; for, as "you may well think, your promotion to the cardinalate "will, both on account of your age and the circumſtances "above mentioned, bring much envy on you; and they who "have not been able to hinder your acquiſition of this dignity, "will leave no ſtone unturned to bring a ſtur on it, by degra- "ding the opinion entertained of your life, and cauſing you "to ſlip into the ſame ditch into which they themſelves are "fallen: and this, they promiſe themſelves, they ſhall "bring about by reaſon of your age. Againſt theſe difficulties "you muſt make head with the greater vigour, as little virtue "is now to be found in the college; though I remember to

"have

Figure 3.3 Example of commentary at the bottom of the page. From New observations on Italy and its inhabitants (1769)

chose to use four models which we describe here. The performance of the models is reported in Table 3.1. The designed models cover a range of machine learning approaches, beginning with a conventional model based only on hand-designed features and moving to a learning-based model that utilizes deep learning.

Our first approach is a "rule-based" model that tries to capture three overarching visual features related to the differential line-size and line-spacing of pages (thus "rule-based"). Our hypotheses for this model are that footnotes will: have a smaller font size than the main text; be located at the bottom of the page; and be indicated by significant spacing between the footnote and the main text. The advantage of this kind of approach is that the creation of custom features can target our prior knowledge of the problem (i.e. what is a footnote) and increase precision. The drawback is that the delimitation of features may not be able to capture the broader diversity of footnote behaviors in our data and thus may lower recall. This can be compensated for by more learning-based approaches where features are not pre-defined but learned from the



Whitehall, July 14. 1705.

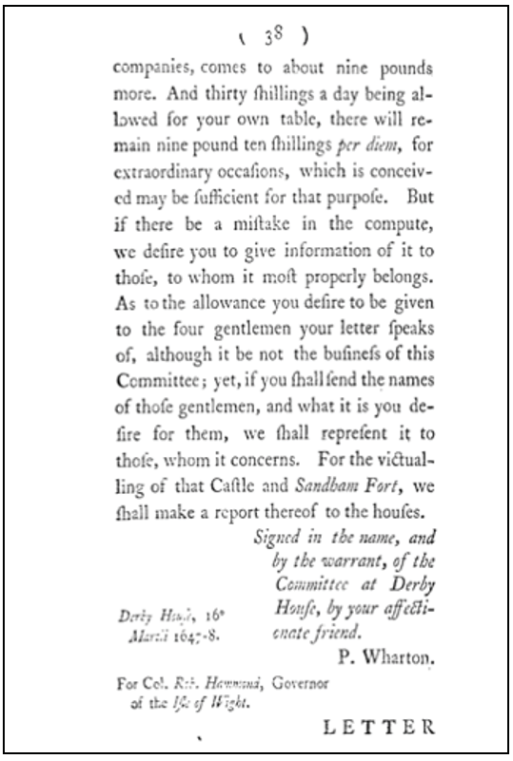
Published by Authority.

THIS Day Colonel *Durell* came hither Expreſs from the Duke of Marlborough, with Advice, That on the 18th Inſtant, N.S. his Grace had ſurprized the Enemy very early in the Morning, forced their Lines near Heyleſhem, and entirely routed Monſieur d'Allegre, who appeared at the head of 20 Battalions and 50 Squadrons, to oppoſe the Confederate Forces; but ſo great was the Conſternation of the Enemy, and the Bravery of our Troops, that my Lord Duke of Marlborough defeated them with very ſmall Loſs on our ſide. After this Victory his Grace ordered a Detachment to March to Tirlemont, where the Battalion of Monluc was forced to ſurrender at Diſcretion. The next Morning the Army marched and encamped with the Right at Vlierbeck, within a Mile of Louvain, and the Left at Corbeck. The Enemy marched the Night before over the Dyle, and have quitted Diſt.

Monſieur d'Allegre and the Count de Horn, Lieutenant-Generals, 2 Majors-General, with about 80 other Officers, and 1400 private Soldiers, beſides the whole Regiment of Monluc, are taken Priſoners. We have likewiſe taken Ten Pieces of Cannon, and a great Number of Standards and Colours. Further Particulars are hourly expected.

Printed by *Edw. Jones*, in the *Savoy*.

Figure 3.4 Example of footnote-like text in an early newspaper (1702)



(38)

companies, comes to about nine pounds more. And thirty ſhillings a day being allowed for your own table, there will remain nine pound ten ſhillings *per diem*, for extraordinary occaſions, which is conceived may be ſufficient for that purpoſe. But if there be a miſtake in the compute, we deſire you to give information of it to thoſe, to whom it moſt properly belongs. As to the allowance you deſire to be given to the four gentlemen your letter ſpeaks of, although it be not the buſineſs of this Committee; yet, if you ſhall ſend the names of thoſe gentlemen, and what it is you deſire for them, we ſhall repreſent it to thoſe, whom it concerns. For the victualling of that Caſtle and *Sandham Fort*, we ſhall make a report thereof to the houſes.

Signed in the name, and by the warrant, of the Committee at Derby Houſe, by your affectionate friend.

P. Wharton.

Derby Houſe, 16th March 1647-8.

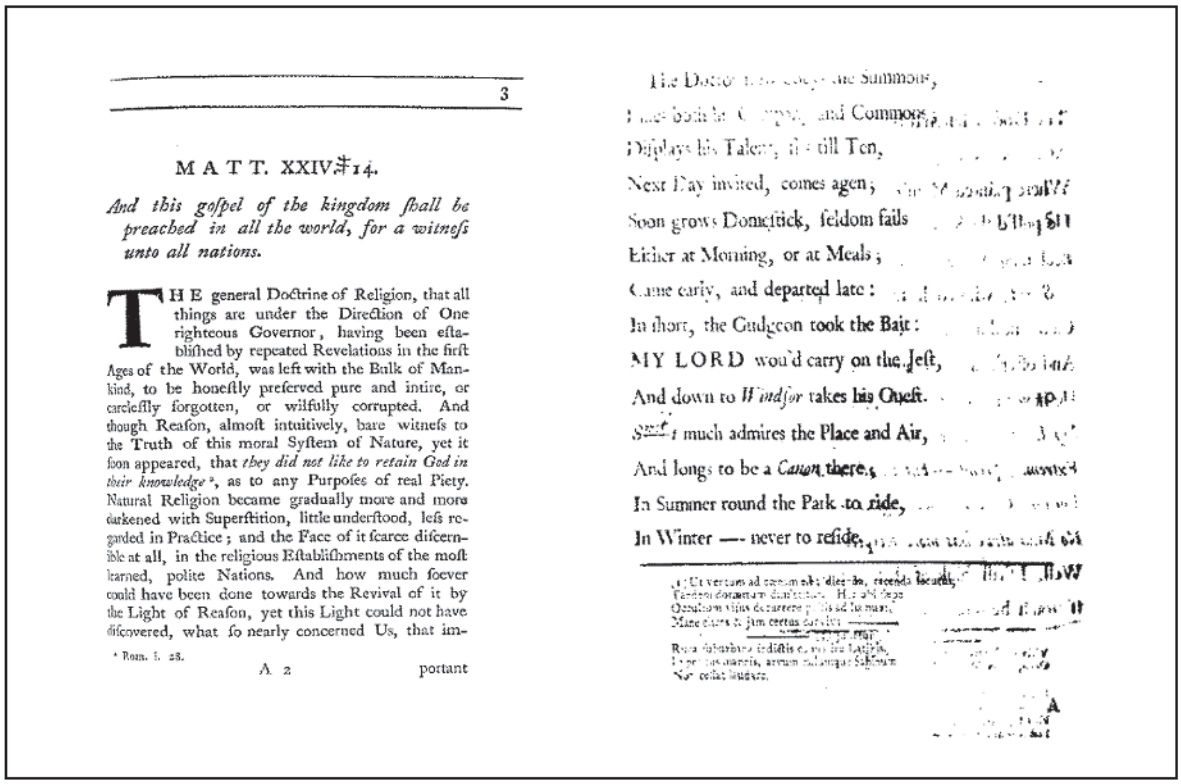
For Col. *Rich. Hammond*, Governor of the *Iſle of Wight*.

LETTER

Figure 3.5 Example of footnote-like text. From Letters between Col. Robert Hammond (1764)

training examples. However, as we show in Table 3.1, we see how overall in our results we do achieve higher precision (finding true positives) and lower recall (producing false negatives, i.e. overlooking footnotes).

In order to estimate font size (hypothesis 1), we use two methods drawn from the field of document image analysis: the bounding box method (BBox) and the horizontal projection method (Proj) (Fig. 3.7) (Dos Santos, Clemente, Ren & Cavalcanti, 2009; Likforman-Sulem, Zahour & Taconet, 2007). Bounding boxes are determined for each line by finding the rectangles containing the connected components. A connected component is defined as the continuous connection of black pixels. In theory, a connected component should correspond to individual letters, but given the imperfect reproductions of pages along with typographic irregularities introduced in historical printing practices, errors can be introduced (Fig. 3.8). These bounding



3

MATT. XXIV. ✠ 14.

And this gospel of the kingdom shall be preached in all the world, for a witness unto all nations.

THE general Doctrine of Religion, that all things are under the Direction of One righteous Governor, having been established by repeated Revelations in the first Ages of the World, was left with the Bulk of Mankind, to be honestly preserved pure and intire, or carelessly forgotten, or wilfully corrupted. And though Reason, almost intuitively, bare witness to the Truth of this moral System of Nature, yet it soon appeared, that *they did not like to retain God in their knowledge* [a], as to any Purposes of real Piety. Natural Religion became gradually more and more darkened with Superstition, little understood, less regarded in Practice; and the Face of it scarce discernible at all, in the religious Establishments of the most learned, polite Nations. And how much soever could have been done towards the Revival of it by the Light of Reason, yet this Light could not have discovered, what so nearly concerned Us, that im-

[a] Rom. i. 28.

A 2 portant

He Doctor [illegible] the Summons,
[illegible] and Commons,
Displays his Talent, till Ten,
Next Day invited, comes agen;
Soon grows Domestick, seldom fails
Either at Morning, or at Meals;
Came early, and departed late:
In short, the Gudgeon took the Bait:
MY LORD wou'd carry on the Jest,
And down to *Windsor* takes his Guest.
[illegible] much admires the Place and Air,
And longs to be a *Canon* there,
In Summer round the Park to ride,
In Winter — never to reside.

A

Figure 3.6 Examples of degraded or hard to capture footnote-marks. Part of the seventh epistle of the First book of Horace (1713) (left) and A sermon by Joseph Lord Bishop of Bristol (1739) (right)

boxes are then used to estimate the lower case letters' font sizes by finding the distance between the lower and the upper base-lines, as shown in Fig. 3.7(a). For the horizontal projection method, the horizontal intensity for each line is calculated (i.e., the pixels in the horizontal direction are summed such that there will be fewer pixels at the upper and lower levels of the line where extenders and descenders are located (capital letters or d's or y's for example)). As demonstrated in Fig. 3.7(b), the font size of a textline is estimated by calculating the distance of the inner intersected line between the derived projection and a threshold line of a value equal to 0.55.

According to our initial hypothesis, we expect that any line with a footnote would correspond to a decrease in font size when compared to the previous line. Although this would be an ideal case, such a decrease could be attributable to something other than the presence of a footnote, such as the presence of a title, figure, or tables, etc. We therefore define additional rules in order

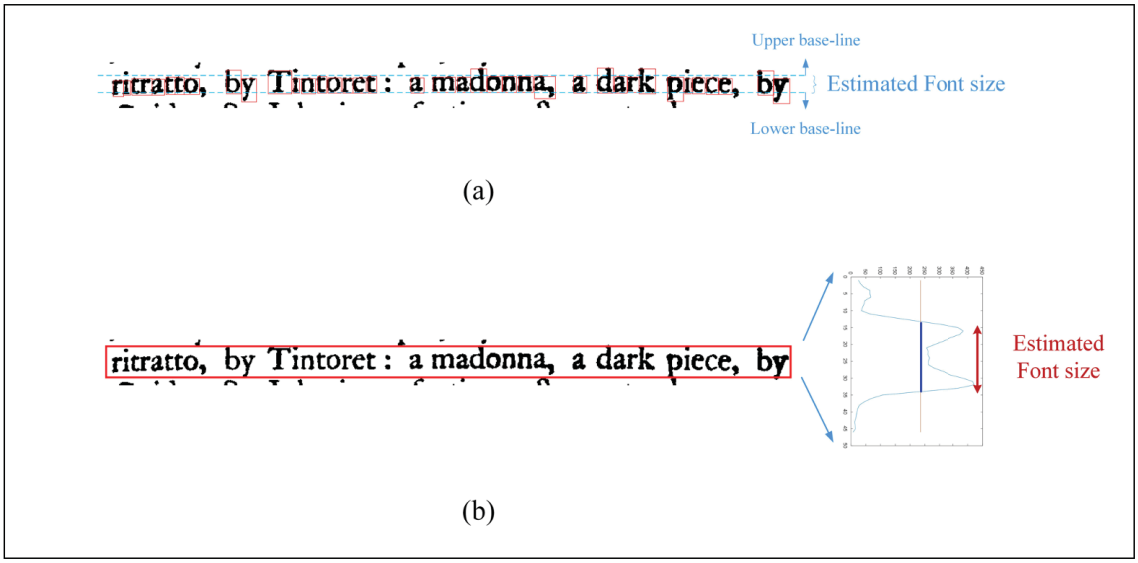


Figure 3.7 The estimated font size of (a) the bounding box based method and (b) the horizontal projection based method

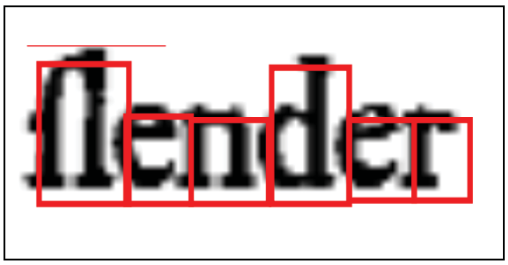


Figure 3.8 We can see in this example of the word "slender" from Hogarth's *Analysis of Beauty* a connected component that spans more than one letter due to the typeface used and the potential for bleeding between letters. Each red box represents a connected component

to improve the accuracy of our footnote detection. We convert these rules into specific features described in Appendix A (Zhalehpour, Piper, Wellmon & Cheriet, 2017).

In order to identify the footnote location (our second assumption), we define a series of further rules based on the page layout for each method (BBox and Proj). The relative position of the

estimated footnote line to all other lines on the document image is then used as a basis of further features. The third and final technique determines the spaces between the lines and uses them as a feature. More specifically, the textline below the large white space closest to the bottom of the page is considered to be a footnote candidate. The location of a possible footnote is compared to the locations estimated in the first two methods in order to partially form the final feature vector of the image. We illustrate how the three primary features of line height, line spacing, and page location perform with respect to two sample pages, one with a footnote, one without (Fig. 3.9).

Using these three primary features we develop a total of 72 features related to rule-based qualities of the page (18 BBox + 24 Proj + 30 Location and space as described in Appendix A) which are then fed into a support vector machine (SVM) classifier to detect pages with footnotes.

Our second model is a "layout-based" model that combines hand-designed and learned features, although it weighs the former more heavily. This and the next model might be considered to be hybrid models that combine custom features defined by expert knowledge and learned features defined by the machine's exposure to the training data. The layout approach primarily depends on understanding the layout of textlines on a page (Fig. 3.10). Once again, it rests on the hypothesis that footnotes will exhibit distinctive visual behavior with respect to their size and position on the page. Similar to the rule-based approach, we develop 22 custom measures for each textline based on the variables shown in Fig. 3.10 (see Appendix B for a full description of all features). Because the number of textlines varies between document images, it is necessary to extract features with a fixed length for all of the images. In order to do this, we use Discrete Cosine Transform (DCT) (Lam, 2004), where we consider the concatenation of each textline's 22 measures as a signal. This signal exhibits a repetitive behavior and thus contains frequency information such that DCT can be used to capture this information. Specifically, since most of the signal's energy (i.e., information) is concentrated in lower frequencies, and assuming that document images have at least 5 textlines, we kept only the first 300 coefficients of the DCT transform for each image.

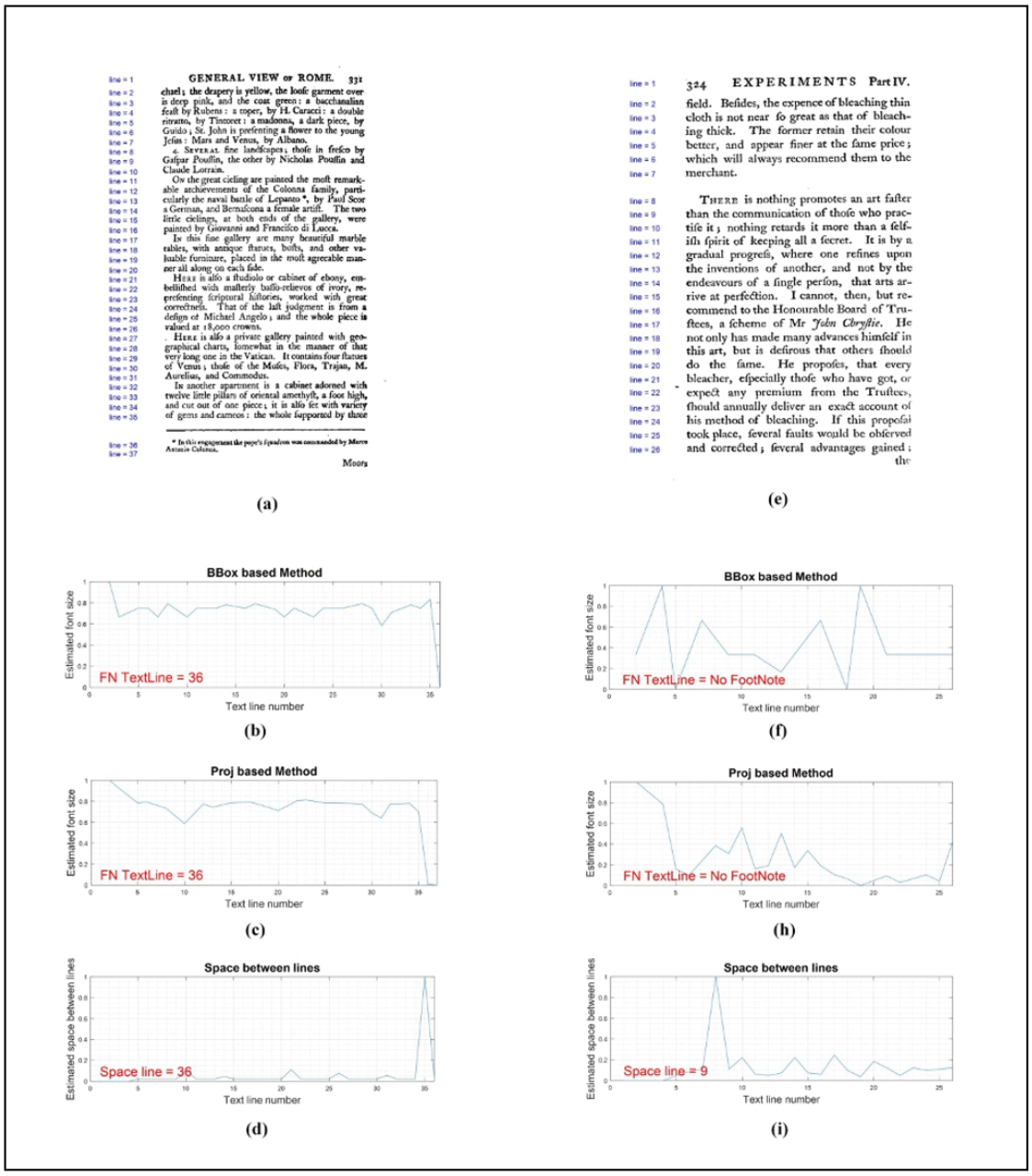


Figure 3.9 Examples of two document pages, with footnote (a) and without footnote (b), and their related results after applying the BBox-based method (b, f), the projection-based method (c, h) and estimating the lines spacing (d, i)

The final step of the "layout-based" model is classification. We use a combination of an Autoencoder overlaid with a softmax layer (Baldi, 2012). The Autoencoder creates lower dimensional representation of the provided input data in its hidden layer and then reconstructs

this data at its output layer. This representation is then fed to a softmax layer with the labels of the document images to learn the model for classifying new samples.

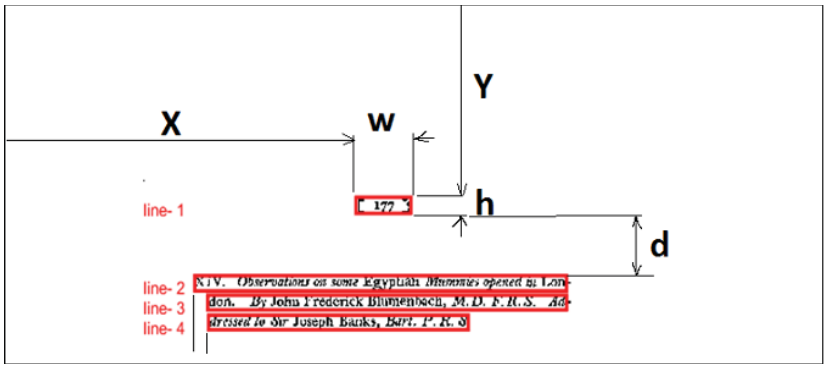


Figure 3.10 A bounding box of a textline with features X, w, Y, h, and d defined as relative positions on the page

Given the recent advances in the field of deep learning, particularly with architectures such as Convolutional Neural Networks (CNN), we used two CNN-based models for our final two approaches (Goodfellow, Bengio, Courville & Bengio, 2016). We also used two different techniques in an effort to compensate for the limited amount of labeled data available to us, because CNNs generally require large amounts of data during the training process in order to perform efficiently.

In the third approach, "CNN-based," the model is based primarily on learning the document image's features throughout the various layers of the neural network. But the model also depends on hand-designed features in order to overcome the limited amount of labeled data. Based on our hypothesis that the footnote's text and the main body's text differ in both style and font, each document image is represented using the two top textlines and the three bottom ones. (We use a projection-based segmentation method described above to detect those textlines.) Each of these textlines is represented, more precisely, as a vertical histogram (Fig. 3.11). As in the previous examples, the performance of this model will be hindered by the reliance on layout assumptions that may not always apply to our object of study. In order to capture changes in font size, here we use vertical projections of the lines, meaning the bars of the histogram represent vertical slices of the red-bounded lines. The lower height of the histogram bars represents a

lower average line-height. A concatenated version of the vertical histograms of these textlines is then used as an input to a 1-dimensional CNN (i.e., a 5000x1 histogram) (Mhiri, Abuelwafa, Desrosiers & Cheriet, 2017).

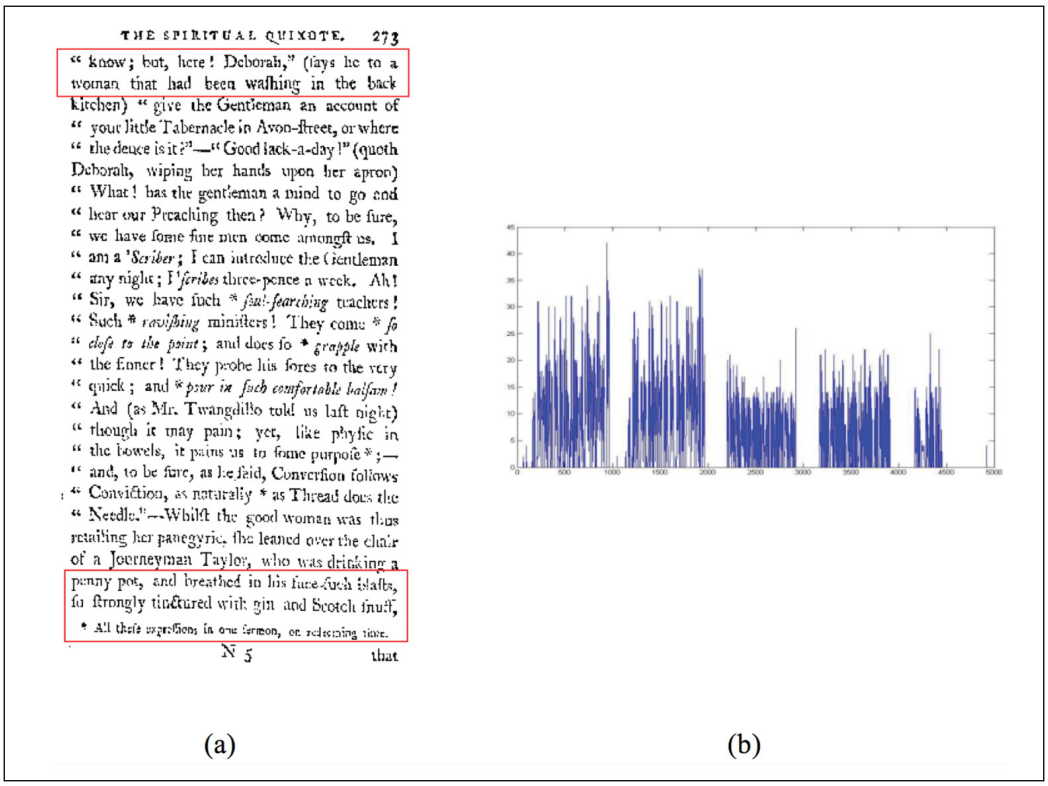


Figure 3.11 A representation of a document page (a) with its related vertical histograms (b)

Our fourth and final approach is based on transfer learning and CNN ("Transfer Learning") (Bengio *et al.*, 2013). According to this approach, the model automatically learns the features without using any hand-designed features. Transfer learning can be particularly useful given the scarcity of labeled training data in our case. The model consists of two supervised learning stages, a pre-training stage and a fine-tuning stage. In the first stage (pre-training), an AlexNet is trained on a large dataset of natural images and the resultant learned parameters (e.g., network's weights) are saved. In the second stage (fine-tuning), instead of initializing the CNN's parameters randomly, the model uses learned parameters from stage one. Then, we use the ECCO dataset to train the model to classify document images with footnotes. The novelty of

this approach is important to emphasize – footnotes are being learned first by learning features of "images" more generally and then being trained on page images more specifically. To prepare the data, we perform two pre-processing steps on the raw document images—resizing and normalization—before using them as inputs to our model. Each document image is re-sized to 227x227, and its pixel values are normalized to be in the range [0 1] (Fig. 3.12). Unlike the first three approaches, this model does not require any textline segmentation process; therefore, it avoids the segmentation errors that may result from it.

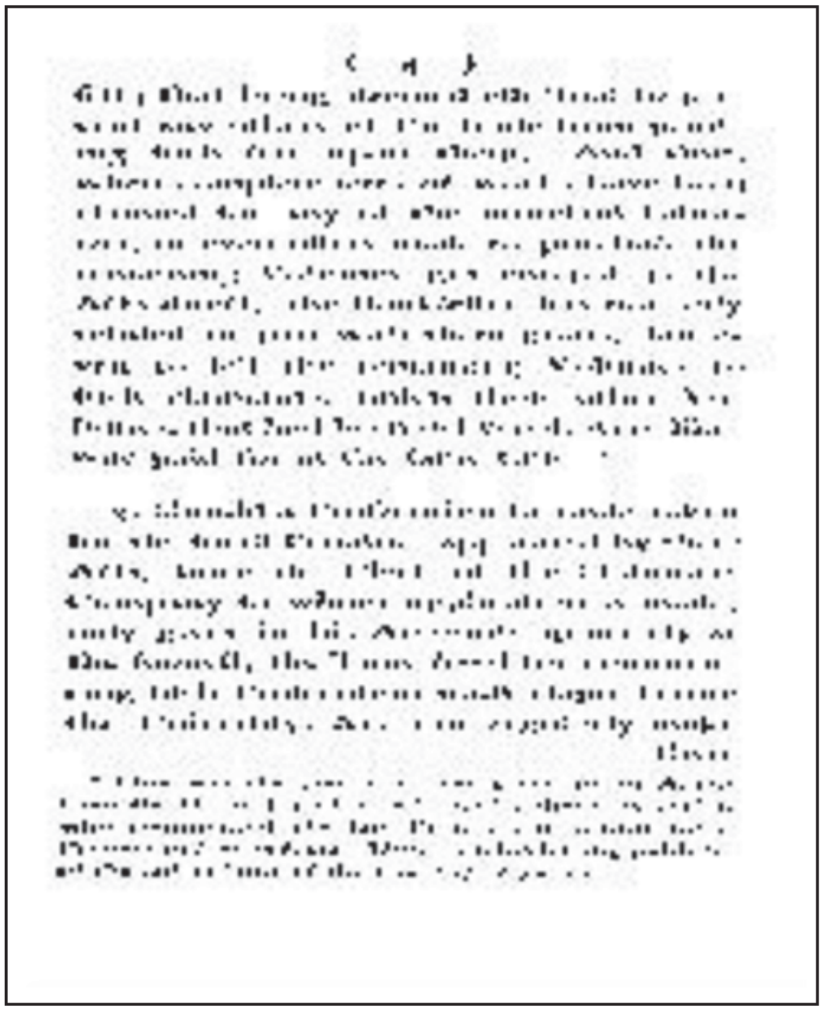


Figure 3.12 In this example, pages are binarized and then reduced in size to 227x227 pixels (or 51,529 dimensions) rendering them illegible, but ideally capturing the unique visual signature of footnotes

These then were the four models we developed to detect footnotes. As a final step, we use an ensemble detection method that combines all four classifiers. Applying this ensemble method on the test set of ECCO, we achieve 96.2% precision and 67.87% recall in our footnote detection results (Table 3.1).

Table 3.1 The individual performance of each detection approach, in addition to the final approach performance

Approaches	**Precision (%)**	**Recall (%)**	**F1 score**
Rule-based	68.24	60.8	0.643
Layout-based	60.8	69.4	0.6482
CNN-based	90.35	48.37	0.63
Transfer learning-based	74.31	41.49	0.5325
The final detection approach	**96.2**	**67.87**	**0.7959**

Applying our detection methods on the full ECCO dataset, we discovered 1,319,000 footnoted images from approximately 26,000,000 document image in ECCO I and 239,754 footnoted images from approximately 6,000,000 document image in ECCO II. We therefore estimate that roughly 4.9% of all pages in the eighteenth contained footnotes. The figures below (Figs. 3.13-3.15) provide more detailed results, including the number of document images with detected footnotes over time (publication years) as well as document images with detected footnotes according to subject classifications in both ECCO I and II. We expect in a separate piece to explore this data in more detail. We share the underlying metadata of footnote annotation to allow others to do the same.

We also provide detailed information in Tables 3.2 and 3.3 that demonstrate the consistency of our final model's performance across different time periods and subjects in both ECCO I and II. As we show, the values of the average footnote probability per page are stable (i.e., around 0.68) regardless of the year or subject of the examined document image. These tables give us confidence that our predicted levels of footnotes are not dependent on either document type or the year of publication. All of our derived data has been shared as supplementary data to this article.

Table 3.2 The average probability of footnote by years and subjects (where images are detected as footnote) at ECCO I

Year (bins)	GenRef	HistAnd Geo	Law	LitAnd Lang1	LitAnd Lang2	MedSci Tech	Reland Phil	SSAnd FineArt	Total
1695				0.57		0.64			**0.62**
1700	0.64	0.66	0.64	0.66	0.66	0.68	0.68	0.66	**0.67**
1705	0.68	0.67	0.69	0.68	0.64	0.69	0.69	0.66	**0.68**
1710	0.66	0.66	0.69	0.67	0.66	0.67	0.68	0.67	**0.68**
1715	0.68	0.67	0.68	0.67	0.68	0.68	0.68	0.68	**0.68**
1720	0.68	0.67	0.68	0.66	0.68	0.68	0.69	0.67	**0.68**
1725	0.66	0.67	0.67	0.68	0.67	0.69	0.69	0.68	**0.68**
1730	0.68	0.66	0.67	0.68	0.67	0.69	0.69	0.67	**0.68**
1735	0.67	0.67	0.66	0.67	0.67	0.68	0.69	0.67	**0.68**
1740	0.68	0.67	0.65	0.67	0.67	0.70	0.69	0.68	**0.68**
1745	0.66	0.67	0.67	0.68	0.68	0.69	0.69	0.68	**0.68**
1750	0.67	0.67	0.67	0.67	0.67	0.68	0.69	0.68	**0.68**
1755	0.67	0.68	0.68	0.68	0.68	0.68	0.69	0.68	**0.68**
1760	0.65	0.68	0.70	0.68	0.67	0.69	0.69	0.67	**0.68**
1765	0.67	0.69	0.67	0.67	0.69	0.69	0.69	0.68	**0.68**
1770	0.67	0.68	0.68	0.68	0.68	0.69	0.69	0.68	**0.68**
1775	0.66	0.68	0.68	0.68	0.68	0.72	0.68	0.68	**0.68**
1780	0.67	0.68	0.68	0.68	0.68	0.69	0.69	0.68	**0.68**
1785	0.70	0.68	0.68	0.68	0.67	0.68	0.69	0.68	**0.68**
1790	0.68	0.69	0.68	0.67	0.67	0.69	0.69	0.68	**0.68**
1795	0.66	0.68	0.69	0.67	0.67	0.68	0.69	0.68	**0.68**
1800	0.67	0.68	0.66	0.67	0.68	0.68	0.69	0.68	**0.68**
1805			0.66				0.71		**0.70**
Total	**0.67**	**0.68**	**0.68**	**0.68**	**0.68**	**0.69**	**0.69**	**0.68**	**0.68**

3.4 Appendix A - The Rule-based Footnote Detection Approach Features

The final utilized feature vector at the rule-based footnote detection approach contains 72 features, which is the combination of the features extracted at the following three techniques.

3.4.1 The Bounding Box (BBox) based Method

At this method, 18 features are being utilized based on some initial assumptions; for instance, the assumption that the font size of any footnote line is at least 0.55 smaller than the font size of the main text. More assumptions are considered and demonstrated in details in table 3.4.

Table 3.3 The average probability of footnote by years and subjects at ECCO II

Year (bins)	GenRef	HistAnd Geo	Law	LitAnd Lang	MedSci Tech	Reland Phil	SSAnd FineArt	Total
1700		0.64	0.55	0.64	0.63	0.66	0.64	**0.65**
1705	0.65	0.66		0.65	0.69	0.67	0.66	**0.66**
1710	0.59	0.66	0.68	0.67	0.68	0.67	0.64	**0.67**
1715		0.64	0.68	0.68	0.67	0.66	0.67	**0.66**
1720	0.57	0.65	0.57	0.66	0.66	0.67	0.66	**0.67**
1725	0.56	0.68	0.64	0.65	0.69	0.68	0.67	**0.68**
1730		0.65	0.67	0.66	0.67	0.66	0.67	**0.66**
1735		0.65	0.66	0.65	0.67	0.66	0.66	**0.66**
1740	0.61	0.66	0.62	0.65	0.66	0.67	0.66	**0.66**
1745	0.61	0.68	0.66	0.65	0.66	0.67	0.66	**0.67**
1750	0.72	0.67	0.66	0.66	0.68	0.68	0.66	**0.67**
1755		0.66	0.69	0.65	0.67	0.67	0.67	**0.66**
1760	0.67	0.66	0.63	0.67	0.67	0.66	0.67	**0.67**
1765	0.63	0.68	0.63	0.66	0.68	0.68	0.66	**0.67**
1770	0.67	0.69	0.66	0.66	0.67	0.67	0.68	**0.68**
1775	0.62	0.67	0.67	0.65	0.68	0.67	0.67	**0.67**
1780	0.63	0.66	0.64	0.65	0.68	0.67	0.67	**0.66**
1785	0.63	0.68	0.65	0.66	0.68	0.67	0.67	**0.68**
1790	0.64	0.67	0.68	0.66	0.67	0.67	0.67	**0.67**
1795	0.64	0.66	0.66	0.66	0.67	0.67	0.67	**0.67**
1800	0.66	0.69	0.67	0.66	0.69	0.67	0.67	**0.68**
Total	**0.65**	**0.67**	**0.67**	**0.66**	**0.67**	**0.67**	**0.67**	**0.67**

3.4.2 The Horizontal Projection (Proj) based Method

At this method, 24 features are being utilized. Table 3.5 demonstrates the extracted features in more details.

3.4.3 Location and Space based Features

Table 3.6 demonstrates the 30 extracted features in more details.

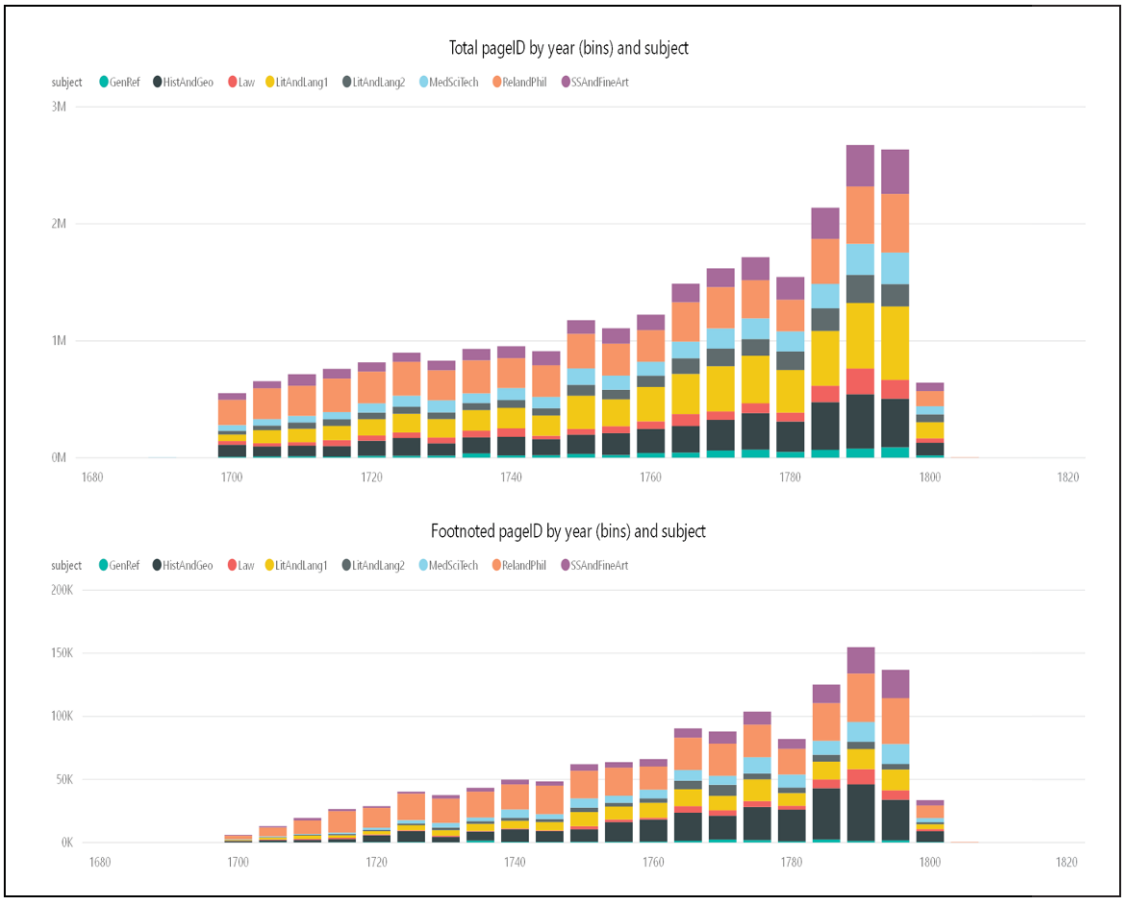


Figure 3.13 Distribution of document images (all and footnoted) in ECCO I by year using Gale's eight subject classes

3.5 Appendix B - The Layout-based Footnote Detection Approach Measures

A detailed description of the used measures at the layout-based footnote detection method is demonstrated at table 3.7.

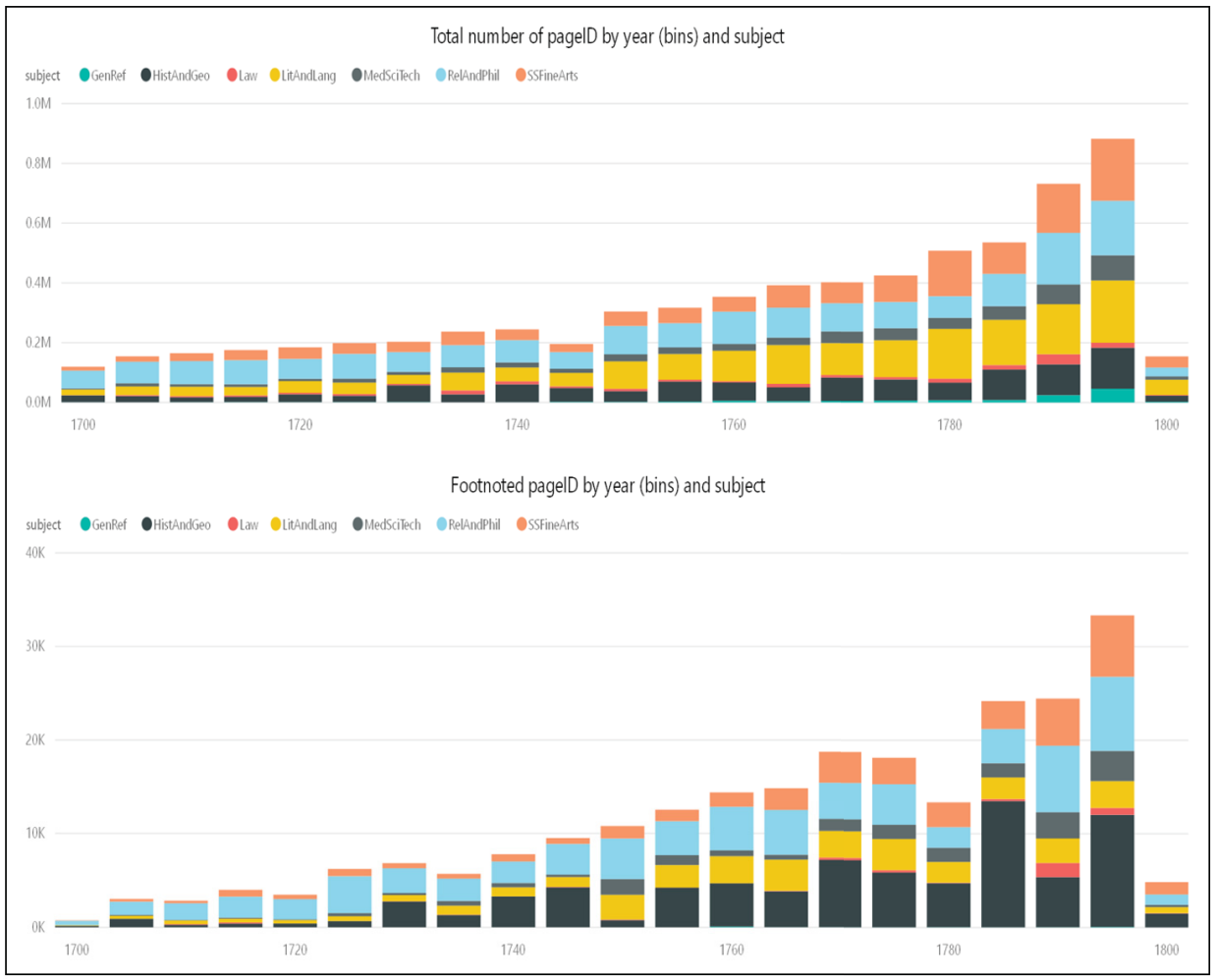


Figure 3.14 Distribution of document images (all and footnoted) in ECCO II by year using Gale's eight subject classes

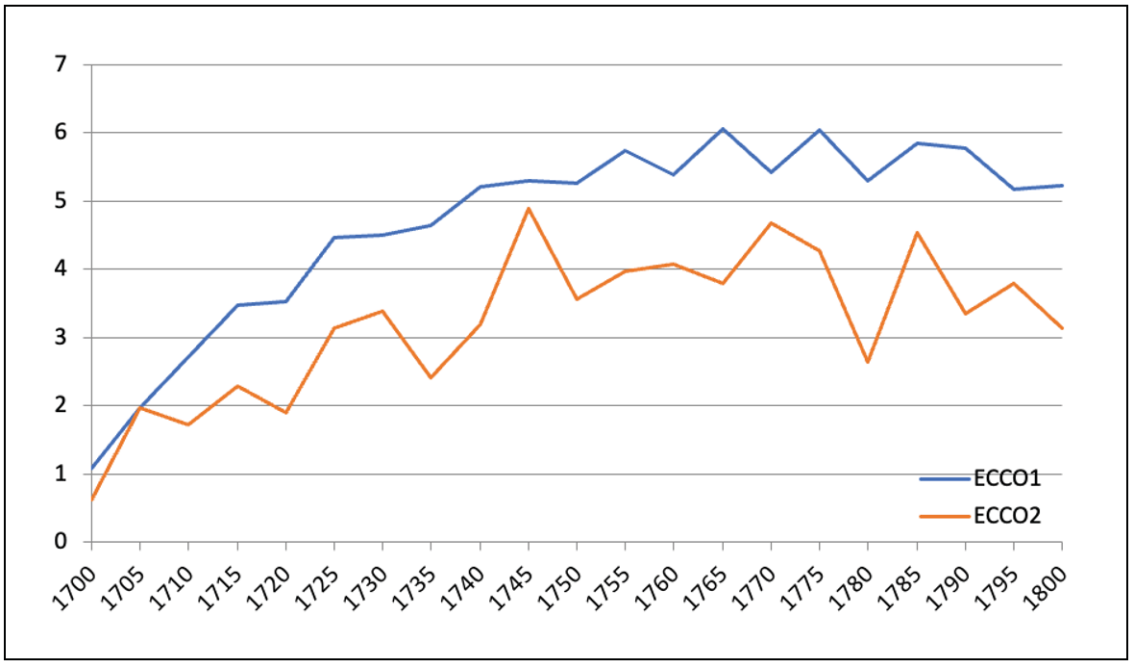


Figure 3.15 The percentage (%) of the detected footnote document images to the total document images at both ECCO I and ECCO II

Table 3.4 The Bounding Box (BBox) based method assumptions

Feature	Condition
1	1 if there is no drop more than 0.55
2	1 if there are 1+ drops of more than 0.55
3	1 if the last two lines' heights are less than 0.1
4	1 if the last line's height is less than 0.1
5	1 if the last two lines' heights are less than 0.1 and there is a footnote
6	1 if the line before last line's height is less than 0.1
7	1 if the footnote is not in the 4th line
8	1 if there are 2+ drops more than 0.55
9	1 if the footnote is not in the 4th, 5th and 6th lines
10	1 if there are 2+ drops less than 0.55 or the footnote is not in the 4th, 5th and 6th lines
11	1 if there is a drop greater than 0.15
12	1 if footnote line is in the 6th line or later
13	1 if the height of the footnote line is 0.55 greater than the line before the last line
14	1 if there is a drop of greater than 0.35 between the lines before and after the footnote line
15	1 if there is a line except the last line selected as the footnote line and there is a drop of greater than 0.35 between the lines before and after it
16	1 if there is a difference less than 0.17 between the lines before and after the footnote line
17	1 if there is a line except the last line selected as the footnote line and there is a difference less than 0.17 between the lines before and after it
18	1 if it the page has more than 3 lines

Table 3.5 The horizontal projection (Proj) based method extracted features

Feature	Condition
1	1 if there are more than 3 lines in the page
2	1 if there is no possible footnote
3	1 if there is more than one possible footnote (drops with the amount of 0.55 or more)
4	1 if the footnote line is in the first 3 lines or there are more than 3 possible footnotes
5	1 if there are more than 3 possible footnotes or there are lines shorter than 0.13 but not footnote lines
6	1 if the footnote line is in the first 3 lines or there are lines shorter than 0.13 but not footnote lines
7	1 if the footnote line is in the first 3 lines
8	1 if there are more than 3 possible footnotes
9	1 if there are lines shorter than 0.13 but not footnote lines
10	1 if the footnote line is in the first 3 lines, there are 3+ possible footnotes or there are lines shorter than 0.13, but not footnote lines
11	1 if the last line or the line before it has a height less than 0.1
12	1 if the last line or the line before it has height less than 0.1 and there is a footnote
13	1 if the last line has a height less than 0.1
14	1 if the line before the last line has a height less than 0.1
15	1 if there still exists a footnote line
16	1 if the height of the last line is less than 0.4
17	1 if the last line has a height less than 0.1 and there exist a footnote line
18	1 if the line before the last line has a height less than 0.1 and there exist a footnote line
19	1 if the height of the last line is less than 0.4 and there is a footnote line and the last line or the line before has a height less than 0.1
20	1 if the greatest height drop is equal or greater than 0.4 and there is at least a 0.25 drop between the line before and after footnote
21	1 if there is at least a 0.25 drop between the line before and after the footnote and the footnote line's height is less than 0.4
22	1 if the greatest height drop is equal or greater than 0.4 and the height of the last line is less than 0.4
23	1 if the height of the last line is less than 0.4 and the greatest height drop is equal or greater than 0.4 and there is at least a 0.25 drop between the line before and after footnote
24	1 if the height of the footnote line is 0.4 below the highest height of all the other lines except the first 3 and last lines

Table 3.6 Location and space based features

Feature	Condition
1	1 if there is more than 10 lines in the page
2	1 if there is a space peak in the 2nd 1/4th of the page and there is more than 10 lines in the page
3	1 if there is a space peak in the 3rd 1/4th of the page and there is more than 10 lines in the page
4	1 if there is a space peak in the 4th 1/4th of the page and there is more than 10 lines in the page
5	1 if there is only one peak in the page, select its location: (Peak location/ # of lines) and there is more than 10 lines in the page
6	1 if there is more than one peak in the page, select the last one's location: (Peak location/ # of lines) and there is more than 10 lines in the page
7-8	1 if there is a footnote in the last 1/4th of the page: (FN location/ # of lines) and there is more than 10 lines in the page
9-10	1 if there is a footnote in the page: (FN location/ # of lines) and there is more than 10 lines in the page
11-16	Check if feature 5 appears anywhere around feature 7(Proj) using a threshold from ±0.02 by a 0.02 step and up to ±0.14
17-22	Check if feature 6 appears anywhere around feature 8(Proj) using a threshold from ±0.02 by a 0.02 step and up to ±0.14
23-26	Check if feature 5 appears anywhere around feature 7(BBox) using a threshold from ±0.02 by a 0.02 step and up to ±0.14
27-30	Check if feature 6 appears anywhere around feature 8(BBox) using a threshold from ±0.02 by a 0.02 step and up to ±0.14

Table 3.7 The layout-based footnote detection method used measures

Measure	Description
1	Number of objects (characters - connected components) ∈ current textline
2-3	$h_{current\ textline} - h_{previous\ textline}$, $h_{current\ textline} - h_{next\ textline}$
4-5	$w_{current\ textline} - w_{previous\ textline}$, $w_{current\ textline} - w_{next\ textline}$
6-7	$x_{current\ textline} - x_{previous\ textline}$, $x_{current\ textline} - x_{next\ textline}$
8	$(x+w)_{current\ textline} - (x+w)_{previous\ textline}$
9	$(x+w)_{current\ textline} - (x+w)_{next\ textline}$
10	$(y+h)_{current\ textline} - (y)_{next\ textline}$
11	$(y)_{current\ textline} - (y+h)_{previous\ textline}$
12	$(y)_{current\ textline} - (y+h)_{first\ textline}$
13	$(y+h)_{current\ textline} - (y)_{last\ textline}$
14-16	$w_{current\ textline}$, $h_{current\ textline}$, $x_{current\ textline}$
17	$(x+w)_{current\ textline}$
18	Number of foreground pixel to the number of all of pixels
19	$feature18_{current\ textline} - feature18_{current\ textline}$
20	$feature18_{current\ textline} - feature18_{next\ textline}$
21	Number of foreground pixels in right half of the textline to the number of foreground pixels in left part of the textline
22	Average ratio of black and white pixels for each row of textline image

CHAPTER 4

UNSUPERVISED EXEMPLAR-BASED LEARNING FOR IMPROVED DOCUMENT IMAGE CLASSIFICATION

Sherif Abuelwafa [1], Marco Pedersoli [1], Mohamed Cheriet[1]

[1] Département de génie des systèmes, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Published in IEEE Access
Volume 7, September 2019, Pages 133738 - 133748

Abstract

Many recent state-of-the-art approaches for document image classification are based on supervised feature learning that requires a large amount of labeled training data. In real-world problem of document image classification, the available amount of labeled data is limited and scarce while a large amount of unlabeled data is often available at almost no cost. In this paper, we present an approach for learning visual features for document analysis in an unsupervised way, which improves the document image classification performance without increasing the amount of annotated data. The proposed approach trains a neural network model on an auxiliary task in which every training example is associated with a different label (exemplar) and expanded to multiple images through a data augmentation technique. Thus, the learned model, which is trained in an unsupervised way, is used to boost the document classification performance. In fact, this learned model has proved to be consistently efficient in two different settings: i) as an unsupervised feature extractor to represent document images for an unsupervised classification task (i.e., clustering); and ii) in the parameters initialization of a supervised classification task trained with a small amount of annotated data. We perform experiments on the Tobacco-3482 dataset and demonstrate the capability of our approach to improve i) the unsupervised classification accuracy up to 2.4%; and ii) the supervised classification accuracy by 1.5% without any extra data or by 5% when using 3000 additional not annotated samples.

Keywords: document image classification, document analysis, document image representation.

4.1 Introduction

Document image classification is a crucial step in the process of document analysis. Finding the document category is essential to later analysis steps, such as text recognition and document retrieval (Dengel & Dubiel, 1995). The current state-of-the-art approaches for document image classification depend on either carefully hand-crafted features (Bukhari & Dengel, 2015; Chen *et al.*, 2012; Kumar & Doermann, 2013) or feature learning (Afzal *et al.*, 2015,1; Das, Roy & Bhattacharya, 2018; Harley, Ufkes & Derpanis, 2015a; Kang, Kumar, Ye, Li & Doermann, 2014). Engineering features is a complex process that requires special expertise for designing and adapting the features to the desired domain and makes it hard to generalize to new tasks (Abuelwafa, Mhiri, Hedjam, Zhalehpour, Piper, Wellmon & Cheriet, 2017; Goodfellow *et al.*, 2016). Recently, approaches that directly learn features from data have received more interest and it is also the approach that we use. Among the feature learning approaches, methods based on Convolutional Neural Networks (CNNs), in which features are learned by the convolutional layers (Afzal *et al.*, 2015,1; Das *et al.*, 2018; Harley *et al.*, 2015a; Kang *et al.*, 2014), achieved state of the art performance.

In terms of supervision, most of the successful feature learning approaches in the domain of document image classification are based on a supervised pre-training paradigm. Using fully supervised feature learning is often an efficient solution that provides very good results as long as enough labeled training data can be provided. This is not often the case in document classification, because the process of manually annotating data is slow and expensive in terms of both, the needed time and expertise. This results in a limited amount of labeled data that can actually be used in the feature learning process. On the other hand, a large amount of related unlabeled data is widely available (e.g., HathiTrust digital library[3] that contains millions of digitized document images).

Thus, semi-supervised and unsupervised approaches seem to be a good solution to improve the classification results without increasing the amount of annotated data. For instance,

[3] https://www.hathitrust.org/

unsupervised feature learning approaches at the pre-training stage (Bengio *et al.*, 2013; Erhan, Bengio, Courville, Manzagol, Vincent & Bengio, 2010) can provide substantial classification improvements (LeCun, Bengio & Hinton, 2015) without additional data annotation. In these approaches, structural and spatial-related features are learned using only unlabeled data. Then, these learned features are used at a later fine-tuning stage, improving the supervised classification performance.

In this paper, we propose to first learn a neural network model during a pre-training phase on a set of data without annotation, thus in an unsupervised manner. This is performed with an exemplar learning in which a neural network is trained to accomplish the auxiliary task of classifying each sample in a data-augmented version of the original dataset. Then, we tackle the problem of document image classification, in which the pre-trained model is used in two different ways: i) in an unsupervised manner, by clustering on features extracted with the pre-trained model ii) initializing a supervised training with the pre-trained network weights. In both cases, the pre-trained model consistently improves the classification performance, over the baseline approaches on the respective tasks, without the need to use any additional labeled data. Note that for unsupervised classification, the reported results are with respect to a baseline that does not utilize the learned featured of our pre-trained network; in addition to, other methods based on a more complex clustering algorithm and hand-crafted features. For the supervised case, the baseline to compare with is a trained model without our pre-trained initialization of the weights.

4.1.1 Contributions of this paper

Our paper provides the following contributions:

- We propose a unified unsupervised pre-training based framework that is simple, yet capable of consistently boosting the performance of both unsupervised and supervised classification.
- To the best of our knowledge, our approach is the first to perform an unsupervised document image classification using a representation that is entirely based on feature learning using

unlabeled data, and does not depend on any hand-crafted features. In the experimental results, we show that our approach outperforms the previous baseline approaches.

- We demonstrate and experimentally validate that by incorporating a small fraction of unlabeled data from a related-dataset we can easily gain up to 2.4% boost in the unsupervised classification performance and over 5% boost in the supervised classification performance.

The organization of this paper is as follows; section 5.2 provides a comprehensive review study on the related work. In section 5.3, the proposed approach is introduced in details. The experimental setup is presented at section 5.4 and the results with their related analysis are discussed at section 4.5. Finally, the paper is concluded at section 6.3.

4.2 Related Work

4.2.1 Document image classification

The problem of document image classification has been tackled in the literature through many approaches that differentiate based on i) the chosen features and ii) the utilized learning mechanism (Chen & Blostein, 2007).

Considering the chosen features, recent approaches in the literature are either content (text) based (Tang, He, Baggenstoss & Kay, 2016), visual appearance based or a combination of both (Noce, Gallo, Zamberletti & Calefati, 2016). The content-based approaches are typically restricted to documents with text and depend mainly on Optical Character Recognition (OCR) methods, which may output text with errors that can affect the classification performance (Kumar, 2013). To avoid this, our proposed work is based instead on the visual appearance characteristics of the document image and does not rely on OCR.

Conventionally, visual appearance based approaches utilizes hand-crafted features (Bukhari & Dengel, 2015). For instance, Scale Invariant Feature Transform (SIFT) (Lowe, 2004) is exploited by (Chen *et al.*, 2012) and Speeded Up Robust Features (SURF) (Bay *et al.*, 2006) is used by

(Kumar *et al.*, 2014). However, lately, visual appearance approaches that are based on feature learning (Kang *et al.*, 2014) have attracted considerable attention.

Since this work is mainly focused on the pre-training stage of the classification process, a review on the related visual appearance based works is discussed in further detail below.

The simplest and most used pre-training approach that has been utilized extensively in recent years is supervised pre-training, in which a big and fully labeled dataset is used to perform a pre-training process (Goodfellow *et al.*, 2016). For instance, (Afzal *et al.*, 2015,1; Das *et al.*, 2018; Harley *et al.*, 2015a) are all incorporating a supervised pre-training process. In this process, annotated samples are used to train a network in a supervised manner, then that pre-trained network's learned parameters are used to initialize a fine-tuning network and perform the process of document classification. Usually a huge amount of annotated data is exploited in this process; for example, around 1 million labeled images of ImageNet (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein et al., 2015) are used in (Afzal *et al.*, 2015; Das *et al.*, 2018; Harley *et al.*, 2015a) and 320,000 labeled images of RVL-CDIP (Harley *et al.*, 2015a) are used in (Afzal *et al.*, 2017). Das, et al.(Das *et al.*, 2018) extended that approach using an ensemble of region-based classifiers (i.e., a strategy that has been introduced by (Roy, Das & Bhattacharya, 2016)). However, the method is still limited to a specific set of documents (e.g., forms, memo), and cannot be applied easily to other document types because it depends on the spatial features of the documents and requires a manual readjustment to the learning algorithm for any new document type. Similarly, (Harley *et al.*, 2015a) evaluated enforcing learning region-specific features and concluded that it is not effective in case of enough training data.

In addition to supervised classification, some related works in literature have explored classifying the document images in an unsupervised manner, hence considered as 'unsupervised classification' (i.e., more details about the unsupervised classification process are discussed at section 4.3.3). For instance, (Kumar & Doermann, 2013) introduced the horizontal vertical partitioning-random forest (HVP-RF) model, which trains a random forest classifier to learn structural patterns from

SURF features (Bay *et al.*, 2006) codebook. This model has a complex pipeline that depends heavily on traditional hand-crafted features; in contrast, our approach achieves better results using a pipeline that is based entirely on unsupervised feature learning. Moreover, the CONFIRM algorithm (Tensmeyer & Martinez, 2019) uses page elements such as OCR transcriptions and rule lines to obtain collection-dependent features. Using rule lines makes this approach limited and more specific to tables. Additionally, depending on OCR is not ideal as discussed earlier in this section.

4.2.2 Unsupervised feature learning

Unsupervised feature learning often works on modeling the distribution of the training data to learn the common invariant features in it. For instance, Deep Belief Networks (DBNs) (Hinton *et al.*, 2006) learn features by yielding the parameters that maximize the latent variables likelihood given the observed ones. The main drawback of this technique is its inefficiency due to the intractability of the estimation of the latent variables likelihood. On the other hand, in direct mapping techniques, features are learned by minimizing the error between an input sample and the reconstructed output or some variants of it (e.g., stacked denoising auto-encoders (Vincent, Larochelle, Lajoie, Bengio & Manzagol, 2010), k-sparse auto-encoder (Makhzani & Frey, 2013) and variational auto-encoder (Kingma & Welling, 2013)). Another interesting approach for improving the classification accuracy of documents is to perform an unsupervised pre-training. On the contrary to the supervised approach, the unsupervised pre-training depends only on unlabeled data. This means fast and cheap access to the available data since the labeling process has been bypassed. Even if very appealing, the impact of unsupervised pre-training on the final classification performance is still limited and not performing as effective as supervised pre-training.

A special case of unsupervised pre-training, is *self-supervised pre-training*. In that case, the learning task exploits the structure of the training data, such that data annotations are already available or come for free. In this way, normal supervised learning techniques can be used on those pseudo-annotations. For instance, the spatial information of neighboring patches is used

to automatically label the input data through either context prediction (Doersch *et al.*, 2015) or solving jigsaw puzzles (Noroozi & Favaro, 2016). Additionally, (Gidaris *et al.*, 2018) applies four different rotations to each unlabeled sample and trains a network to recognize the correct one. On the same line, an exemplar-based learning with CNN is introduced by Dosovitskiy et al. (Dosovitskiy *et al.*, 2016). In this approach, data-augmentation is applied to each unlabeled sample to create a set of surrogate classes and a network is trained to discriminate between them. Due to its simplicity and closeness to the classification tasks, the Exemplar-CNN based learning has inspired the pre-training part of our framework. However, various changes in the architecture have been introduced for better adaptation to the problem of structural document classification.

4.3 The proposed methodology

Our proposed framework is based on an unsupervised pre-training step in which a convolutional neural network (CNN) model is learned using only unlabeled data. This is followed by two different document image classification approaches: an unsupervised classification on the learned representation and a supervised classification initialized with the pre-trained model.

More insights on the different learning stages and other related steps are detailed in the following subsections.

4.3.1 Pre-processing

As shown in table 4.1, our network is based on the AlexNet architecture (Krizhevsky, Sutskever & Hinton, 2012). Thus, in order to match the network input size, all the utilized input document images, at the stages of unsupervised pre-training and classification, are resized to 227x227 pixel resolution. To provide an efficient processing performance, the resizing process keeps the fundamental structural features of the document, while reducing other less critical information for our model (e.g., the exact shape of characters and words). After resizing the image, a binarization process is performed: the image pixels values are rounded to either 0 or 1.

4.3.2 Unsupervised pre-training stage

The main objective of this stage is to train a CNN model using a set of unlabeled data. As shown in Fig. 4.1, the training procedure is composed of two steps: first, the generation of augmented data and surrogate classes; and then the actual training of the neural network to classify these generated classes. The two steps are detailed in the following paragraphs.

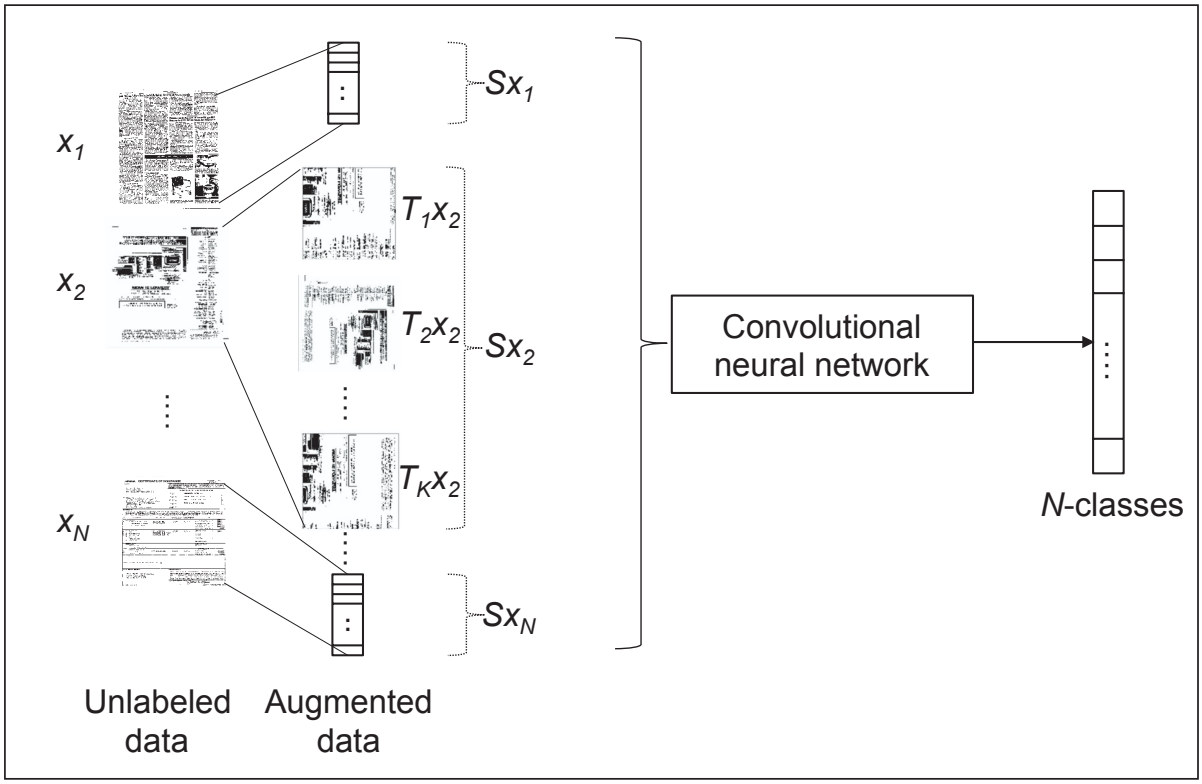


Figure 4.1 Proposed unsupervised pre-training stage

4.3.2.1 Generate augmented data and surrogate classes

Inspired by data augmentation (Wang & Perez, 2017) and similarly to (Dosovitskiy *et al.*, 2016), we generate a set of transformations of our original document images such that the augmented data are still valid and realistic document representations. We consider an initial training set $\mathcal{X}$ containing N unlabeled document images. A set of randomly chosen combinations of pre-defined transformations $\{T_1, ...T_K\}$ is applied to each image $x_i \in \mathcal{X}$, which produces K

Table 4.1 The architecture of the CNN model used in our experiments

Layer (type)	Output shape	Filter
input (InputLayer)	1 x 227 x 227	-
conv_1 (Conv2D)	96 x 55 x 55	11 x 11
max_pooling_1 (MaxPooling2D)	96 x 27 x 27	3 x 3
conv_2 (Conv2D)	256 x 27 x 27	5 x 5
max_pooling_2 (MaxPooling2D)	256 x 13 x 13	3 x 3
conv_3 (Conv2D)	384 x 13 x 13	3 x 3
conv_4 (Conv2D)	384 x 13 x 13	3 x 3
conv_5 (Conv2D)	256 x 13 x 13	3 x 3
max_pooling_3 (MaxPooling2D)	256 x 6 x 6	3 x 3
flatten (Flatten)	9216	-
dense_1 (Fully-connected)	4096	-
dense_2 (Fully-connected)	4096	-
dense_3 (Fully-connected)	N*	-

* Number of surrogate classes.

augmented versions of this image. Specifically, each augmented image $T_k x_i$ is the result of incrementally applying (with 50% probabilities) three basic transformations. To guarantee robust, descriptive and generic learned features, the following basic transformations that relate to some core characteristics of the document images have been used: rotation by angles 90 or -90 degrees, zooming-in by a uniformly sampled factor between 1 and 1.15, and horizontal flipping. Algorithm 1 provides more details on how the augmentation process is carried out. Each unlabeled image, x_i, is now considered a surrogate class S_{x_i}, and its corresponding generated transformations $\{T_1 x_i, ...T_K x_i\}$ are samples of that class with a surrogate label $i \in N$. Fig. 4.2 shows some generated samples of a surrogate class. We will show that the numbers of surrogate classes N and samples per surrogate class K have a critical impact on the classification performance; more insights are discussed in subsection 4.5.2.1.

4.3.2.2 Train the network

An exemplar learning process is accomplished using the obtained set of N surrogate classes and their $N * K$ samples. Specifically, a neural network is trained to associate each sample $T_k x_i$ to its

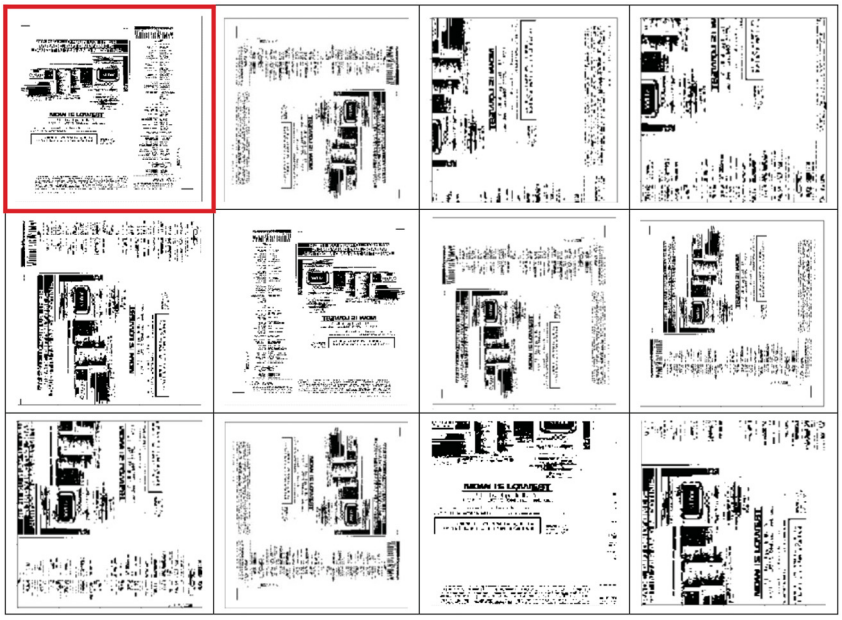


Figure 4.2 Applying T transformations (i.e., rotation by angles ±90°, zooming-in by a uniformly sampled factor between 1 and 1.15, and horizontal flipping) to an unlabeled document image x_i from the Tobacco-3482 dataset to generate $\{T_1 x_i, ...T_K x_i\}$ samples of a surrogate class S_{x_i}. The seed image x_i is at the top left corner

related surrogate class S_{x_i} by minimizing the augmented samples cross-entropy loss:

$$\begin{aligned} L(X) &= \sum_{i=1}^{N} \sum_{k=1}^{K} l(T_k x_i, i), \\ l(x, i) &= -log(p(y = i; x)), \end{aligned} \tag{4.1}$$

where $p(y = i; x)$ is the probability of sample x to belong to class i and $p(.)$ is the softmax output of our network. After training, the obtained network parameters θ are considered to be invariant to the transformations used during the augmentation process.

The used network, as reported in table 4.1, contains eight layers (i.e., five convolutional and three fully connected layers) with around 56 million parameters. A zero padding is included to all the convolutional layers except the last one. In addition, the last fully-connected layer is coupled with an N-way softmax that provides an estimate of each class's conditional probability.

Algorithm 4.1 Generate surrogate classes: for each image x_i, we generate a transformation as a random composition of rotation of θ degrees (R_θ), zoom-in by a factor z (Z_z) and horizontal flip (F)

for ***each*** $x_i \in \mathcal{X}$ **do**
  for $k = 1 \ \ to \ \ K$ **do**
    $T_k = I$ ▷ I: identity transform
    $rotate \sim Bernoulli(0.5)$
    if $rotate$ **then**
      $\theta \leftarrow$ either $-90°$ or $90°$
      $T_k = T_k \circ R_\theta$
    end
    $zoom\text{-}in \sim Bernoulli(0.5)$
    if $zoom\text{-}in$ **then**
      $z \sim U(1, 1.15)$
      $T_k = T_k \circ Z_z$
    end
    $flip \sim Bernoulli(0.5)$
    if $flip$ **then**
      $T_k = T_k \circ F$
    end
  end
  $S_{x_i} = \{T_1 x_i, \ldots, T_K x_i\}$
end

4.3.3 Unsupervised classification stage

As illustrated in Fig. 4.3, the unsupervised classification is actually a clustering process in its core. During training, we divide the training data into clusters and then associate each cluster to the best class in the test data. Thus, we separate the data into M classes in an unsupervised manner, but then for the evaluation, we consider the labeled data to associate each group to an actual class. This is a common way to evaluate unsupervised learning for a classification task (Kumar & Doermann, 2013); more insights are discussed at the clustering step.

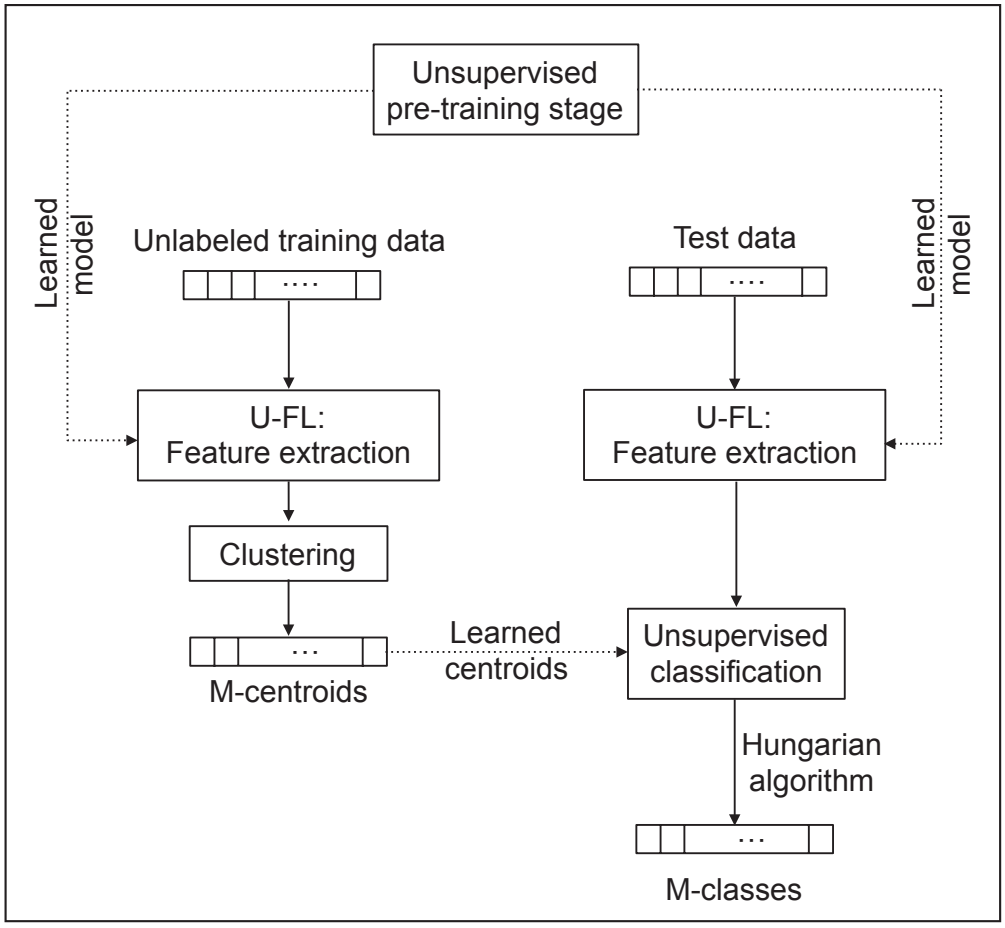


Figure 4.3 Proposed unsupervised classification stage

4.3.3.1 Feature extraction

In the scenario of the unavailability of any annotated data, the derived pre-trained model is used to extract features. In this case, we consider the learned neural network as a function $f : \mathbb{R}^A \rightarrow \mathbb{R}^E$, which maps each image $x_i \in X$ from its original space $\mathbb{R}^A$ to the representation space $\mathbb{R}^E$. The choice of representation and its related feature vector length E is studied in more detail in subsection 4.5.1.1.

4.3.3.2 Clustering

Each obtained representation, $f(x_i)$, from the previous step is used as an input to a clustering algorithm. Since the main focus of this work is on the representation learning part, two off-the-shelf standard clustering algorithms, k-means (Jain, 2010) and spherical k-means (Dhillon & Modha, 2001), are utilized. In k-means, the centriods $\{\mu_1, ..., \mu_M\}$ of the cluster sets $C = \{c_m, m = 1, ..., M\}$ are found through minimizing the Euclidean distance between each obtained document image representation, $f(x_i)$, and the nearest centroid, μ_m, over all the M clusters using the following objective function $J(C)$:

$$J(C) = \sum_{m=1}^{M} \sum_{f(x_i) \in c_m} \| f(x_i) - \mu_m \|^2. \tag{4.2}$$

The spherical k-means algorithm is also based on a similar loss, but the cosine similarity is used instead of the Euclidean distance.

Once the cluster centroids $\{\mu_1, ..., \mu_M\}$ are obtained during the training process, each test sample is then assigned to its nearest centroid in the unsupervised classification process. Afterwards, each cluster of test samples is assigned to an actual class (i.e., from the test set true labels) in an optimal way using the Hungarian algorithm (Kuhn, 1955). This algorithm considers a matching matrix of the predicted cluster labels and true labels and returns the indices of the best matching pairs.

4.3.4 Supervised classification stage

If a limited amount of annotated data is available, the learned parameters θ of the same network architecture of the unsupervised pre-training are used as an initialization to improve the supervised classification performance. As illustrated in Fig. 4.4, this neural network is then fine-tuned on the provided small annotated data with cross-entropy loss function and an M-way softmax classification layer. Notice that M is now the real number of classes of the task.

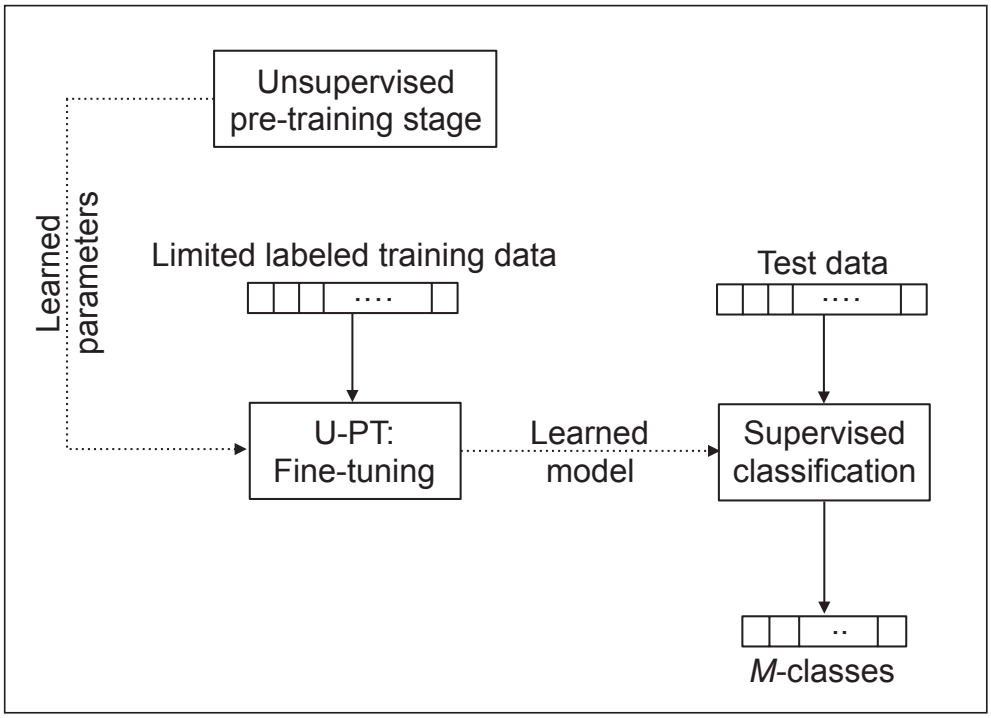


Figure 4.4 Proposed supervised classification stage

4.4 Experimental setup

In this section, the used datasets and the implementation details for the different experiments are explained.

4.4.1 Datasets

During the unsupervised pre-training stage, two datasets have been utilized with our proposed framework. In both datasets, only the document images are used. The first dataset is Tobacco-3482[4] (Kumar & Doermann, 2013), which contains 3,482 document images and 10 document classes. The second dataset is RVL-CDIP dataset[5] (Harley *et al.*, 2015a). This dataset originally contains 400,000 document images and 16 document classes, but only a small subset of those images (i.e., up to 5000) has been used throughout the pre-training stage. This is because more

[4] https://lampsrv02.umiacs.umd.edu/projdb/project.php?id=72

[5] http://scs.ryerson.ca/ aharley/rvl-cdip/

images did not further improve the results and made the training longer, as demonstrated at section 4.5.2.1.

At the later stage of both unsupervised and supervised classification, only the Tobacco-3482 dataset has been utilized. At the unsupervised classification stage, the document images are used solely without their associated labels; while at the supervised classification stage, the document images of the Tobacco-3482 dataset and their related labels have been utilized.

At the unsupervised pre-training stage and when using the Tobacco-3482 dataset, we performed the process ten times, one for each partition using 1,000 samples of the related training set. This is to guarantee that all the test samples, at the later classification stages, are completely unseen and have not been used previously during pre-training. On the other hand, when using RVL-CDIP dataset for pre-training, we performed the process only once since all the used samples are considered unseen for the testing process.

To evaluate our document image classification approach at either the unsupervised classification or the supervised classification stages, we follow the same evaluation protocol presented in the literature (Afzal *et al.*, 2017; Harley *et al.*, 2015a; Kang *et al.*, 2014) to guarantee fair comparisons. Initially, the Tobacco-3482 dataset is divided into 1,000 samples for training and the rest of the samples (2,482) for testing. Since the samples in the original dataset are unevenly distributed between its 10 classes, we make sure that the training set contains exactly 100 samples per class. Then, the training set is divided into 800 samples for training and 200 for validation, where each class is represented with 80 images for training and 20 images for validation. To guarantee a reliable estimation of the proposed approach performance, we report the median classification accuracy of ten randomly-created partitions of the dataset.

4.4.2 Implementation details

All the provided results are based on implementations carried out on an Nvidia GeForce GTX 960 GPU using Theano (Team, Al-Rfou, Alain, Almahairi, Angermueller, Bahdanau, Ballas, Bastien, Bayer, Belikov et al., 2016) and Keras API [6].

For the pre-training stage, the Adam optimization algorithm (Kingma & Ba, 2014) has been used to train our models with a learning rate of $1e-4$ for 120 epochs; while at the supervised classification stage, the same algorithm has been utilized with $1e-6$ learning rate for 1100 epochs.

During the pre-training stage, the unlabeled training data has been subdivided into batches of 5 samples, where for each epoch, the run-time was around 2 seconds per batch. While during the supervised classification stage, the run-time was around 8 seconds per epoch using 800 samples for training and 200 samples for validation.

At the unsupervised classification stage, the number of times the clustering algorithm will be run with randomly initialized centroids (*'n_init'*) and the maximum number of iterations for each run (*'max_iter'*) are set to 50 and 300, respectively, in case of k-means; while they are set to 150 and 300 in case of spherical k-means. In addition, the 'linear_assignment' function provided by scikit-learn library (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg et al., 2011) is used to implement the Hungarian algorithm.

4.5 Results and discussion

4.5.1 Unsupervised feature learning

In this subsection, we discuss in details the unsupervised classification performance and the effect of the learned representation on it.

[6] https://github.com/keras-team/keras

4.5.1.1 Selection of the learned representation

To study the effect of the learned representation on the unsupervised classification performance, various experiments have been performed using a partition of the Tobacco-3482 dataset. To evaluate the unsupervised classification performance (i.e. clustering is well-matched with the test set's true labels), we follow the literature (Kumar & Doermann, 2013) in computing the purity (Manning, Raghavan & Schütze, 2010) and the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985).

Specifically, at the feature extraction stage, we study the correlation between the different characteristics of the learned document representations and the unsupervised classification (clustering) performance. The representation characteristics mentioned here refer to the location of the layer to extract the features from and its associated feature vector length E, table 4.1 provides more details about the different types of layers and their associated locations in the neural network and related output shapes.

Table 4.2 shows the performance of different learned representations with various locations and dimensionality that ranges from $E = 4,096$ to $E = 43,264$. Although the *flatten* representation has a larger feature vector ($E = 9,216$) than the *dense_2* representation ($E = 4,096$), the former performs better than the latter. This is due to the fact that the *flatten* representation preserves the spatial locality information of its obtained features unlike *dense_2*. On the other hand, since the number of the unlabeled training samples is limited ($N = 1000$), and considering the curse of dimensionality, it is understandable that both high-dimensional representations *conv_5 (*$E = 43,264$*)* and *flatten+dense_1 (*$E = 13,312$*)* obtain a poor performance despite preserving full/some spatial locality information about their features.

Our best results are obtained when using the *flatten* representation for both clustering algorithms, k-means and spherical k-means.

Table 4.2 The unsupervised classification (clustering) ARI and purity when utilizing various learned representations (i.e., on a partition of the Tobacco-3482 dataset)

Representation	**Feature vector length (E)**	**k-means**		**Spherical k-means**	
		ARI	**Purity**	**ARI**	**Purity**
conv_5	43,264	0.1387	0.4057	0.2321	0.4899
flatten + dense_1	13,312	0.2094	0.4694	0.2742	0.5254
flatten	9,216	0.2726	0.5242	**0.2759**	**0.5294**
dense_2	4,096	0.2141	0.4895	0.2194	0.5153

4.5.1.2 Unsupervised classification (Clustering) results

Table 4.3 reports the unsupervised classification results using our proposed unsupervised feature learning (U-FL) based representations, which show an improvement in the performance compared to the best four performing representations in the literature (Kumar & Doermann, 2013). These codewords based representations are either global-based (G-BOW) or partitioning-based that use either spatial-pyramid (SP) or horizontal vertical partitioning (HVP) to capture the spatial dependencies. Afterward either Euclidean distance (E) or random forest (RF) is used to compute similarities. For our proposed approach, we compare two configurations: 'without additional data (w/o add. data)' refers to using 1000 training samples (unlabeled) of Tobacco-3482 at pre-training, while 'with additional data (w/ add. data)' denotes utilizing 3000 unlabeled samples from RVL-CDIP dataset. In our experiments, the best configuration seems to be k-means with additional data although the difference with respect to the other configurations of our algorithm is relatively small.

Compare with previous approaches, our proposed representation outperforms the HVP-RF representation (Kumar & Doermann, 2013) by 4 points, in both ARI and purity, without the need of any additional data (U-FL (w/o add. data) -spherical k-means-) and 5 points using additional data (i.e., 3000 unlabeled samples from RVL-CDIP dataset) (U-FL (w/ add. data) -k-means-).

Table 4.3 The unsupervised classification (clustering) ARI and purity results of our learned representation and the state-of-the-art representations

Representation	ARI	Purity
G-BOW-RF (Kumar & Doermann, 2013)	0.21	0.48
SP-RF (Kumar & Doermann, 2013)	0.22	0.46
HVP-E (Kumar & Doermann, 2013)	0.18	0.46
HVP-RF (Kumar & Doermann, 2013)	0.24	0.49
Proposed U-FL (w/o add. data) -k-means-	0.27	0.52
Proposed U-FL (w/ add. data) -k-means-	**0.29**	**0.54**
Proposed U-FL (w/o add. data) -spherical k-means-	0.28	0.53
Proposed U-FL (w/ add. data) -spherical k-means-	0.27	0.52

4.5.2 Unsupervised pre-training

This subsection studies the supervised classification performance and its correlation with the pre-training parameters.

4.5.2.1 Selection of the pre-training parameters

We study the importance of the number of surrogate classes N and the number of samples per surrogate class K on the supervised classification task using a partition of the Tobacco-3482 dataset for evaluation.

First, we study the correlation between the supervised classification performance and the used number of samples per surrogate class K. To do so, we examine the classification performance with various K values using the Tobacco-3482 dataset at the unsupervised pre-training stage with 1000 surrogate classes ($N = 1000$). Fig. 4.5 shows that increasing the number of samples per surrogate class K results in an improvement in the accuracy that saturates as the number of samples becomes larger.

Then, to examine the correlation between the supervised classification performance and the number of used surrogate classes N (i.e., and consequently the total number of pre-training samples), we apply our proposed approach with various N values. This is performed using the

RVL-CDIP dataset with 40 samples per surrogate class ($K = 40$). Note that, in this experiment, the RVL-CDIP dataset is used instead of the Tobacco-3482 (only at the unsupervised pre-training stage), where it offers studying the performance with surrogate classes N values that are beyond 1000 (i.e., the Tobacco-3482 dataset is limited to 1000 training samples). On the other hand, the Tobacco-3482 is still used at the supervised classification stage.

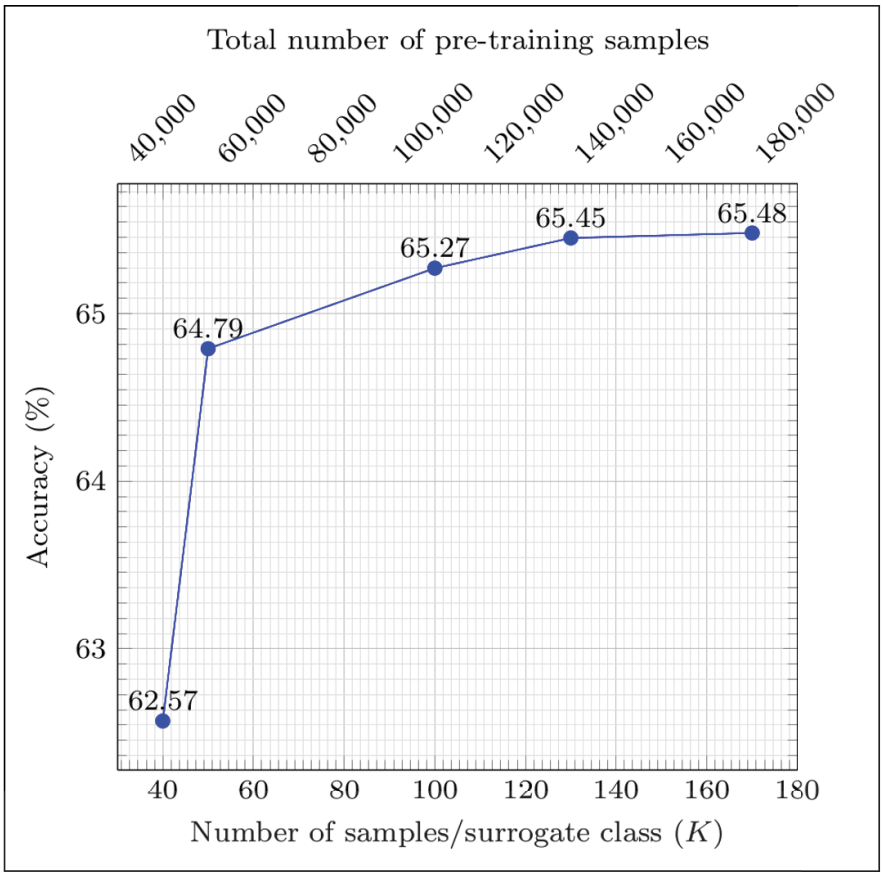


Figure 4.5 The supervised classification accuracy, on a partition of the Tobacco-3482 dataset, with different numbers of used samples/surrogate class (K) and fixed 1000 surrogate classes (N)

Fig. 4.6 shows that the accuracy generally improves when increasing the number of used surrogate classes N with a clear saturation after a certain point. For instance, in the last point of Fig. 4.6, although the number of surrogate classes N has been increased by 2000 (i.e., 66%), the classification performance has not improved significantly, only by 0.08%. This is expected

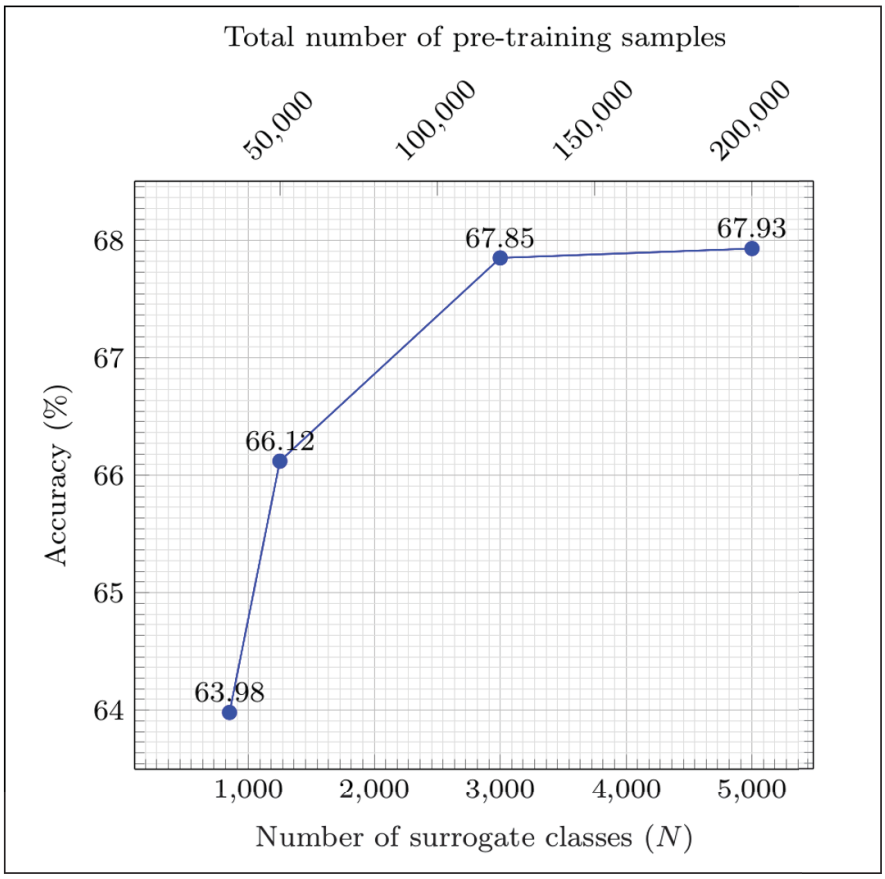


Figure 4.6 The supervised classification accuracy, on a partition of the Tobacco-3482 dataset, with different numbers of utilized surrogate classes (N) and a fixed 40 samples per surrogate class (K)

since utilizing more surrogate classes can lead to considering too similar images as different classes, which leads to harder pre-training discrimination and less effective learned parameters (Dosovitskiy *et al.*, 2016).

4.5.2.2 Supervised classification results

Table 4.4 demonstrates the supervised classification median and mean accuracy on the Tobacco-3482 dataset, where the parameters initialization is with either i) no pre-training, ii) our proposed unsupervised pre-training (U-PT) based learned parameters θ without any additional data (w/o add. data) (i.e., based on the training data of the Tobacco-3482 dataset using 1000 surrogate

classes N with 100 samples/class K (100K samples)), or iii) our proposed unsupervised pre-training (U-PT) based learned parameters θ with additional related unlabeled data (w/ add. data) (i.e., based on a small portion of the training data of the RVL-CDIP dataset using 3000 surrogate classes N with 40 samples/class K (120K samples)). Note that the used N and K values are based on a trade-off between the accuracy and the computational cost of the algorithm (i.e., values where the accuracy starts to saturate while the computation is still moderate).

Table 4.4 The supervised classification median and mean accuracy, on the Tobacco-3482 dataset, with different parameters initialization methods

Parameters initialization method	**Median accuracy (%)**	**Mean accuracy±std (%)**
No U-PT	63.38	62.74±0.017
Proposed U-PT (w/o add. data)	65.01	65.13±0.012
Proposed U-PT (w/ add. data)	68.86	68.95±0.012

The obtained results show that incorporating our proposed unsupervised pre-training (U-PT) based learned parameters θ can efficiently and consistently lead to a boost in the supervised classification accuracy over the performance of the method when trained from scratch. The improvement is over 1.5% without the need of any extra data and using only an unlabeled version of the same training data to be used at the supervised classification stage. Additionally, our approach is capable of boosting the classification accuracy to over 5% when substituting the previously used data with unlabeled data from a related dataset (e.g., RVL-CDIP dataset).

4.5.3 Discussion

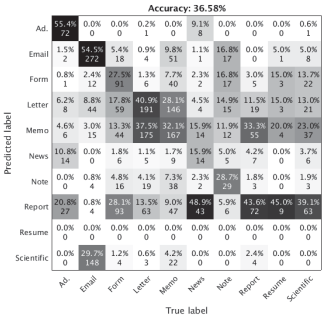
To illustrate the performance of both unsupervised and supervised classification on the same metric space, the accuracy of the unsupervised classification process is calculated through efficiently utilizing the Hungarian algorithm (Kuhn, 1955) to find the optimal assignment between each cluster of document images and its corresponding class in the ground truth (true label).

Fig. 4.7 and table 4.5 demonstrate the impact of utilizing the learned pre-trained model on the document image classification performance with both of its unsupervised and supervised

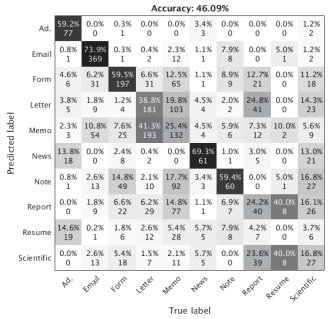
settings, specifically: i) on the unsupervised classification accuracy using the model's learned representations ii) on the supervised classification accuracy using the model's learned parameters θ. In both cases, the results are compared to their relevant baselines.

Fig. 4.7 reports the confusion matrices of tests performed on one partition of the Tobacco-3482 dataset. We observe that incorporating our proposed unsupervised feature learning (U-FL) based representations with the unsupervised classification leads to a better class grouping. This is except for some classes which have low inter-class layout variations with each other (i.e., the high layout similarities between the classes of report, resume and scientific). Similarly, for the supervised classification, our proposed unsupervised pre-training (U-PT) based learned parameters θ yields better grouping results in many classes comparing to training the network from scratch.

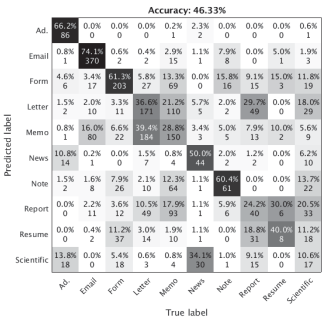
Table 4.5 compares the performance of our methods for supervised and unsupervised classification and other approaches. In order to have a fair comparison, all methods are trained (either supervised or unsupervised) on 1000 samples of the Tobacco-3482 dataset. We can separate the methods in unsupervised (upper part of the table) and supervised (lower part of the table). All the supervised methods outperform the unsupervised ones. This is expected as in the unsupervised case, classes are grouped based only on clustering approaches and no labels are used. Among the unsupervised methods, we can see that the features extracted from our network architecture without any pre-training (No U-FL) perform quite poorly. However, when we use the features from our pre-trained network (U-FL), the results are much better. This shows that our unsupervised pre-training approach is very effective in learning good features. Additionally, our methods obtain better results than(Kumar *et al.*, 2014), which is based on a random forest and hand-crafted features that are selected for the specific task. For the supervised classification (lower part of table 4.5), we can see a similar pattern in which using unsupervised pre-training (U-PT) helps to improve the performance, going from 63.38% to 68.86% for the pre-training with 3000 images. In fact, our unsupervised pre-training (U-PT) gets closer to the performance of a model pre-trained with over one million labelled data (ImageNet). Finally, we see that in the case of having access to a large amount of similar labelled data (e.g., utilizing 320,000 annotated



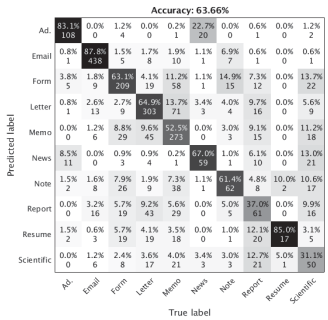
(a)) Unsup. classification: No U-FL



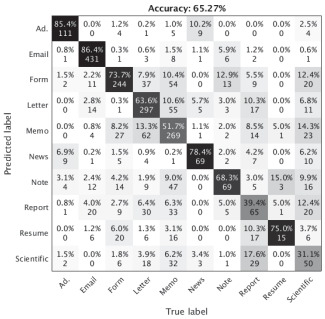
(b)) Unsup. classification: Proposed U-FL (w/o add. data)



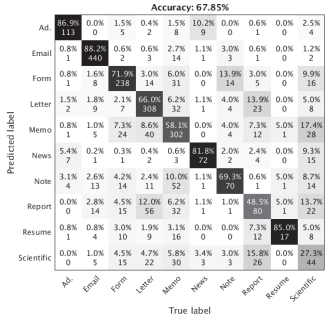
(c)) Unsup. classification: Proposed U-FL (w/ add. data)



(d)) Supervised classification: No U-PT



(e)) Supervised classification: Proposed U-PT (w/o add. data)



(f)) Supervised classification: Proposed U-PT (w/ add. data)

Figure 4.7 Confusion matrices for different models on one partition of the Tobacco-3482 dataset. (a) Unsupervised classification using features from a randomly initialized network. (b) Unsupervised classification using features from a network pre-trained on 1000 non-annotated samples. (c) Unsupervised classification using features from a network pre-trained on 3000 non-annotated samples. (d) Supervised classification without any pre-training. (e) Supervised classification with unsupervised pre-training on 1000 non-annotated samples. (f) Supervised classification with unsupervised pre-training on 3000 non-annotated samples

document images in(Afzal *et al.*, 2017)), results can be further boosted up to 90%. Overall, we can see that with limited training data (1000 training samples) and without a proper pre-training (No U-FL and No U-PT), CNN-based methods perform quite poorly. However, incorporating our proposed unsupervised pre-training enables these methods to be trained more effectively and leads to better results without the need of extra annotated data.

Table 4.5 Classification median accuracy (i.e., on the Tobacco-3482 dataset-ten partitions-) for unsupervised and supervised methods with different pre-training approaches

	Method	**Pre-training**		**Median accuracy (%)**
		Unsup.	**Supervised**	
Unsup.	No U-FL	-	-	36.76
	HVP-RF (Kumar *et al.*, 2014)	-	-	43.80
	Proposed U-FL (w/o add. data)	1000	-	45.26
	Proposed U-FL (w/ add. data)	3000	-	46.25
Supervised	No U-PT	-	-	63.38
	Proposed U-PT (w/o add. data)	1000	-	65.01
	Proposed U-PT (w/ add. data)	3000	-	68.86
	S-PT (w/ ImageNet)	-	$\sim 1,000,0000$	72.89
	S-PT (w/ doc. images) (Afzal *et al.*, 2017)	-	$320,0000$	90.04

4.6 Conclusion

Contrary to conventional document image classification methods that use either hand-crafted features or supervised pre-training approaches, we propose a visual features learning approach that is based on unsupervised pre-training. The proposed approach uses only unlabeled data to learn a pre-trained model, which is used later for unsupervised and supervised classification. Our approach improves the performance of the document image classification problem in the cases of i) the unavailability of any labeled data, ii) the availability of limited labeled data and iii) the availability of additional unlabeled data. Our experimental results corroborate the capability of our approach to improve the accuracy of CNN-based classification methods. Although other supervised pre-training approaches may provide more improvement in the classification performance, our approach has a crucial advantage of not requiring any additional manually annotated data.

Acknowledgment

The authors thank the NSERC of Canada (Grants no. RGPIN 2014-04649 and RGPIN 2018-04825) for their financial support.

CHAPTER 5

UNSUPERVISED LEARNING FOR DOCUMENT IMAGE SEMANTIC SEGMENTATION

Sherif Abuelwafa[1], Ehsan Arabnejad[1], Marco Pedersoli[1], Mohamed Cheriet[1]

[1] Département de génie des systèmes, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Submitted to Elsevier Pattern Recognition, September 2021

Abstract

Semantic segmentation of document images is an essential, yet challenging, step in performing document image analysis. The current state-of-the-art for this task depends mainly on supervised learning-based approaches, in which a large amount of labeled document images are needed for the training process. In this work, we propose an unsupervised end-to-end approach for semantically segmenting document images without requiring labeled data, dataset-dependant heuristics techniques or textual information. The proposed approach successfully overcomes various challenges that relate to the unique characteristics of the document image and the nature of its semantic classes. For instance, it learns a robust representation of the document image thanks to our introduced representation space that combines both distance transform and RGB information of each image. This novel representation is capable of learning discriminative features of the document's semantic classes, regardless of the inherent discontinuities and white spaces in those classes and the high inter-class similarities between them. In addition, dilated convolutional layers in a CNN-based approach are utilized to well-capture the features of both small-sized and large-sized semantic classes. Moreover, a novel technique is introduced to automatically identify the image's background regions to reduce the bias in the training set without the need for any annotations. We demonstrate that our proposed approach is efficient in performing unsupervised semantic segmentation by yielding better results than baseline approaches. Experiments have been performed on various public datasets to demonstrate the robustness of our approach.

Keywords: semantic segmentation of document images, document analysis, semantic page segmentation, document image representation.

5.1 Introduction

Document image segmentation is a core step in the document image analysis pipeline since it provides a detailed understanding of the document image by segmenting it into various similar logical regions (e.g., figure, text, table), which are usually involved in further processing stages (i.e., performing Optical Character Recognition (OCR) on the obtained text regions).

On traditional document image segmentation approaches, pixels are grouped based only on their visual appearance (i.e., appearance-based classification where the text pixels shall be distinguished from other regions' pixels like figures and tables). This is commonly referred to as 'page segmentation'. On the other hand, in recent semantic segmentation approaches, the classification of pixels is actually both visual appearance and semantic-based. Specifically, it is based on the underlying text content and semantic, in addition to its visual characteristics. For instance, instead of classifying all the text pixels as one region, they are classified into sub-regions that are semantically relevant (e.g., paragraph, list, caption, etc.). In fact, the problem of document image semantic segmentation can be perceived as a pixel classification problem, where a label is assigned to each pixel in the document image.

Utilizing semantic segmentation-based approaches for document image analysis helps in understanding the content and the structure of the document image by segmenting it into homogeneous regions that are semantically alike (e.g., background, figure, table, paragraph).

Most of the recent advances in the document semantic segmentation field are based on fully supervised learning approaches, which are profoundly dependent on the availability of a large amount of labeled document images. And since segmenting a document image semantically involves a pixel-wise ground truthing process, a costly annotation step is needed to prepare the training data (Ji *et al.*, 2019; Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár & Zitnick, 2014). This has led to a scarcity in the availability of labeled training document images.

While using synthetic generated data, as in (Yang *et al.*, 2017), can arguably be applied to supervised learning techniques without the need for an expensive data annotation process, these data do not follow the same distribution as real data, and therefore, performance is lower because of the domain shift. Inspired by the success of unsupervised semantic segmentation in other domains (e.g., natural images (Ji *et al.*, 2019)), we propose performing pixel-wise semantic segmentation for document images depending fully on unsupervised learning and without requiring annotated data, textual information (e.g., text embedding maps) or dataset-dependant heuristic techniques. Generally, doing so on document images is an extremely challenging problem, and a relatively unexplored area.

In fact, many bottlenecks in performing unsupervised semantic segmentation on document images are actually due to the fundamental properties of the document image and the nature of its semantic classes. For instance, in a document image, each class to be segmented often contains plenty of spatial discontinuities and white spaces. Specifically, each segmented class (e.g., paragraph, list, table) can contain both text and lots of white pixels that shall be considered as a part of that class, even without sharing the same characteristics with those adjacent text pixels. This is unlike the segmented classes in a natural image (e.g., a human face, a car, a cat), where each object to be segmented is a naturally connected pixels region and barely contains white pixels. Moreover, many of the semantic classes in document images (e.g., paragraph, list, caption, section) are actually all text-based and often the visual differences between them are subtle, which makes those semantic classes ambiguous to the learned model. Additionally, almost all the document images dataset are imbalanced at pixel level and biased in nature, where the majority of pixels are assigned to a handful number of dominating semantic classes (e.g., background) that are way more represented than the other ones.

In the case of supervised learning based approaches, labeled data and textual information can provide a great help in overcoming the document images challenges mentioned earlier. On one hand, the labeled training data can help in training the model to learn features that can well differentiate between semantic classes with high inter-class similarities. Additionally, using textual data (e.g., OCR information) can help in reducing the confusion between such

classes (Enendu, 2019). On the other hand, having information about the number of samples per class can help in achieving a weighted loss function that can lead to an efficient learned model. However, in unsupervised learning, neither the labeled data nor the textual information are available; therefore, none of the above options are feasible.

In this work, we propose a novel end-to-end approach for segmenting a document image semantically, and in an unsupervised manner, using solely unlabelled training data and without the need to use any additional textual information or dataset-dependant heuristic procedures. To achieve this, our approach focuses mainly on overcoming the previously discussed challenges associated with performing an unsupervised semantic segmentation on document images. First, we introduce a combined representation space that utilizes both distance transform in addition to RGB information of the document image as an input to our network. This results in learning a novel representation that can acquire information about the spatial white spaces -horizontal and vertical- between the text lines, which leads to overcoming the white spaces issue without the need to any labeled data. In addition, it contributes to having a better semantic segmentation of different textual semantic classes that are visually similar. Moreover, dilated convolutional layers are utilized for combining features from multiple spatial scales to efficiently learn features that are related to both the small-sized and the large-sized classes. Furthermore, to encounter the biased nature of the document images dataset, the obtained distance transform of each image is used to automatically identify the background regions, which helps in reducing the bias in the training set without the need for any labeled data.

In summary, our paper provides the following contributions:

- We propose an end-to-end approach for semantically segmenting a document image (pixel-wise) in a total unsupervised manner.
- We introduce a novel representation based on a combined representation space (RGB and distance-transform) to effectively segment the document image even with semantic classes that contain many discontinuities and white spaces, and have high inter-class similarities between them. In addition, dilated convolutional layers in a CNN-based approach are

exploited for addressing the need to capture the features of both small-sized and large-sized semantic classes.

- We propose a novel automatic technique to identify the image's background regions for reducing the training set bias without the need to any annotated data.
- To the best of our knowledge, this is the first work to perform an unsupervised document image segmentation using solely unlabeled data, and without depending on any textual information or dataset-dependant heuristics techniques. Experimental results demonstrate that our approach outperforms the baseline approaches using publicly available datasets.

The rest of this paper is organized as follows. Section 5.2 gives a comprehensive overview of the related work. In section 5.3, our proposed approach is presented with details. Afterwards, in section 5.4, we evaluate our approach and discuss further analysis with demonstrated results. Finally, section 5.5 concludes the paper.

5.2 Related Work

Many approaches for document image segmentation have been studied in the literature. These approaches can be grouped into two groups, page segmentation and, recently, semantic segmentation.

5.2.1 Page Segmentation

Most of the conventional page segmentation approaches are based on dividing the document image into several patches, then carefully-tuned hand-designed features are used to train a classifier and assign logical classes to those patches. These approaches rely often on document layout assumptions and heuristic rules; therefore, they cannot generalize well to different types of documents and their varying visual characteristics. For instance, image patches with associated textural features are utilized in both (Vil'kin, Safonov & Egorova, 2013) and (Oyedotun & Khashman, 2016). Specifically, the document image is divided into various patches, where the texture features are extracted and used for the patches classification process. The

classes of the obtained patches are often corrected based on the analysis of neighboring regions, like in (Vil'kin *et al.*, 2013). And since using image patches leads often to rough and inaccurate segmentation boundaries, some approaches have considered the page segmentation problem as a pixel classification task, such as (Chen, Wei, Hennebert, Ingold & Liwicki, 2014). Yet, this approach depends mainly on hand-designed features to perform segmentation. Specifically, the texture features combined with other color and coordinates features have been utilized.

Recent advances in feature learning have led to exploring many related approaches for document image page segmentation. Some of these approaches are based on supervised feature learning. For instance, in (Chen, Seuret, Hennebert & Ingold, 2017), a simple network of one convolutional layer has been utilized to learn features from image patches. In contrast, deeper networks, such as residual neural network (ResNet), are used in (Pondenkandath, Seuret, Ingold, Afzal & Liwicki, 2017). Additionally, (Jobin & Jawahar, 2017) has depended on Fisher vector encoded convolutional neural networks (FV-CNN) and fully connected convolutional neural networks (FC-CNN) for extracting features, then an Support Vector Machines (SVM) is used for segmentation. Besides, (Alberti, Seuret, Pondenkandath, Ingold & Liwicki, 2017b) has utilized a Linear Discriminant Analysis (LDA) for initializing the weights of a neural network and used that network later for segmenting a document image. Subsequently, (Wick & Puppe, 2018) and (Barakat & El-Sana, 2018) have depended mainly on a Fully Convolutional Neural Network (FCN) to perform document image segmentation.

On the other hand, other approaches are based on unsupervised feature learning. For instance, (Chen *et al.*, 2015) used a convolutional autoencoder to learn features, then for segmentation, an SVM is used. An extension to that approach has been introduced in (Chen *et al.*, 2016b), where a Conditional Random Field (CRF) model is applied after learning the local features using the stacked autoencoders. Although the above techniques have attempted performing unsupervised feature learning on document images, they have not tackled the challenge of performing an unsupervised segmentation. Specifically, those document segmentation approaches are not actually fully unsupervised, since the autoencoder is only used for feature learning, and then labeled data are needed to train a classifier in a supervised manner.

In (Wei *et al.*, 2017), a mixed approach that depends on both hand-designed and learned features has been introduced. Specifically, this approach investigated selecting features that has been already learned and fine-tuned. The work highlighted that most of the features learned by the autoencoder are redundant and not effective for page segmentation tasks.

In (Studer, Alberti, Pondenkandath, Goktepe, Kolonko, Fischer, Liwicki & Ingold, 2019), cross-domain transfer learning has been studied. In particular, VGG and ResNet based neural networks have been utilized, with weights that are initialized using ImageNet pre-trained parameters, for the task of document image segmentation. The obtained results show that cross-domain transfer learning has little, and sometimes harmful, effect on the desired task.

5.2.2 Semantic Segmentation

Considering the advances in supervised semantic segmentation techniques (i.e., which are mainly designed for natural images), recent works have investigated utilizing such techniques for extracting the semantic structure of document images based on a supervised learning manner. For instance, in (He, Cohen, Price, Kifer & Giles, 2017), a multi-scale fully convolutional neural network (FCN) is used for the task of semantic page segmentation. In that approach, the learning process is achieved based on supervised learning; and a conditional random field (CRF) is utilized to improve the obtained results. In addition, in (Lee, Hayashi, Ohyama & Uchida), a combination of a U-Net and a trainable multiplication layers (TMLs) is utilized for obtaining the co-occurrence from features maps to gain better semantic segmentation performance. Furthermore, in (Sarkar, Aggarwal, Jain, Gupta & Krishnamurthy, 2019), a CNN-RNN based network pipeline is introduced to perform a hierarchical structure segmentation on high resolution 'form' images. Moreover, (Yang *et al.*, 2017) utilized a multimodel fully convolutional neural networks (MFCN) for segmenting the document images based on both appearance and semantics bases. The work is mainly based on supervised learning and needs around 135,000 annotated document images -synthetically generated- to train the model. On the other hand, it explored the notion of utilizing unsupervised learning for improving the performance of the supervised task (i.e., improvements that range from 0.6% to 2.6%). In fact, the unsupervised learning part is only complementary to

the supervised learning pipeline and cannot be implemented to be used in a standalone manner. Additionally, that approach depends on both a text embedding map during the training process and the PDF file information to perform post-processing.

As presented above, many approaches in the literature have been proposed to utilize supervised learning to tackle the problem of document images semantic segmentation. However, performing semantic segmentation based on unsupervised learning is relatively an unexplored area of research. In this work, we investigate performing pixel-wise semantic segmentation for document images through solely unsupervised learning, and without the need to any labeled training data, dataset-dependant heuristics techniques or textual information.

5.3 The proposed methodology

Our proposed approach has two main stages, as shown in Fig. 5.1: data pre-processing and training. The training stage roughly utilizes an adapted version of the clustering objective set by (Ji *et al.*, 2019); nevertheless, many major changes and contributions have been introduced to achieve a better reliable adaptation for the document image segmentation task. In fact, these improvements and contributions have helped in overcoming two major challenges that are related to the unique characteristics of document images and their layout. More details are discussed below.

Small-sized and large-sized semantic classes:
First, we have observed the need to capture both local low-level and global high-level spatial features simultaneously during the learning process. Since document images have a unique nature of including small-sized semantic classes (e.g., section, caption) that exist in relatively small portions of the document image, comparing to other semantic classes (e.g., figure) that occupy larger portions of the image, low-level spatial features are normally needed in order to capture these small-sized classes and their related visual distinctive remarks. For instance, columns (a) and (b) in Fig. 5.3 demonstrate the small portion that the 'section' class -in yellow- is occupying comparing to other semantic classes (e.g., figure -in green- or paragraph -in purple-).

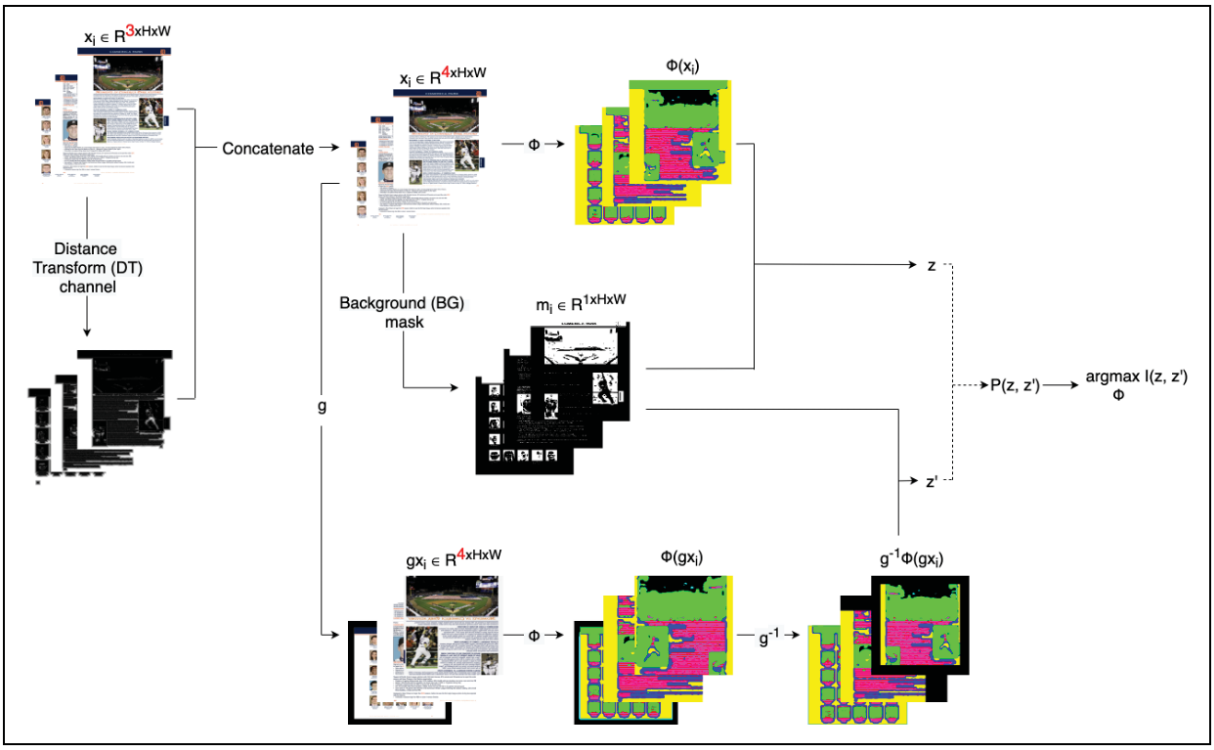


Figure 5.1 The proposed approach flowchart. The convolution process is expressed with dashed lines

Performing a pooling process, and its associated down-sampling process, urges the need to perform a counter up-sampling process (e.g., interpolation) to return back the the actual size of the input image. This process leads subsequently to introducing distortions and experiencing a loss in critical local spatial features that are essential for distinguishing between similar small-sized semantic classes (e.g., caption and section) and segmenting them. Therefore, in order to maintain the obtained local low-level spatial features across the different layers of the proposed network and avoid any loss in them, our proposed network architecture does not consider any pooling layers. More details on the effect of utilizing pooling (e.g., max-pooling) on the learned representation and the unsupervised segmentation performance are detailed in section 5.4.4.2.

Beside avoiding utilizing any pooling process, small-sized receptive fields are needed in order to capture the local features of the small-sized semantic classes. To obtain these small-sized receptive fields, convolutional filter with small sizes should be used. In fact, using these filters

leads normally to better segmentation of the pixels related to the small-sized semantic-based classes, alongside a sharper-edged segmentation masks. However, this introduces some extra undesired oversegmentation in other large-sized semantic-based classes that occupy large portions of the document image (e.g., figure, table, paragraph) since the receptive field in that case is not large enough to be well-aware of the global correlations in the semantic-context at the spatial-level. An issue that can be avoided by utilizing larger filter sizes to obtain larger receptive fields. More details on the effect of utilizing different receptive field sizes on the learned representations are detailed in section 5.4.4.3. The obtained observations have led to a conclusion that both small and big receptive fields are needed to be considered in order to efficiently detect both small-sized and large-sized classes. To achieve this, and inspired by (Yu & Koltun, 2015), dilated convolutional layers are utilized. They provide a technique to combine features from multiple spatial scales without the need to utilize other techniques (e.g., multi-scale pyramids) that relies often on spatial subsampling. In fact, dilated convolutional layers offer a way to acquire both the local and the global knowledge using the same size of the convolutional filter, which can spread out its weight values more distant in the space to learn more about the global context. This allows having small receptive fields that grow can very large (i.e., exponentially), while the number of parameters grow linearly. To adjust how further the receptive field can grow, a dilation factor d is used.

Dataset biased-nature:

Our second observation, regarding the challenges unique to the document images, is the necessity to reduce the dataset bias. In fact, due to the nature of the textual data distribution in document images, there are always dominating semantic classes that occupy large areas of the document, where lots of pixels are assigned to those classes. Specifically, most of the pixels in the document images belong to the background and only fraction of the total document's pixels contain foreground information and other semantic classes. This results in a persisting problem of imbalanced and biased dataset, where some classes are much more represented than others. In order to overcome such challenge, it is common, during the training process, to utilize loss function with weighting factors that variate for each class based on the number of pixels which

represent that class in the dataset (Zhang, Du, Yoshida & Yang, 2019). This helps in imposing different importance level to each class based on its occurrence. Actually, this technique can work well for semantic segmentation problems based on supervised learning, where the ground truth labels for the utilized training samples and the distribution of each class are available. But this is not the case in our proposed approach that is fully-based on unsupervised learning with no in-advance knowledge about the dataset classes distribution. Therefore, instead of having a weighted mutual information based loss function under unsupervised learning, we aim instead to reduce the bias in the utilized training set in an entirely unsupervised manner. This is achieved through the distance transform information, as explained in 5.3.2.1, to automatically create a background mask to minimize the effect of the dominating background pixels. More details on how to obtain the mask, and its effect on re-balancing the dataset is discussed in further details in 5.3.3.1.

5.3.1 Network Architecture

The proposed network consists of one backbone and two output heads (i.e., a main one and an auxiliary one). The backbone is formed by four dilated convolution blocks. For each block, five dilated convolution layers with dilation factors $d = \{1, 2, 4, 8, 16\}$ are applied (i.e., each with padding $p = \{1, 2, 4, 8, 16\}$, respectively). For each dilated convolution layer, a batch normalization (BN) (Ioffe & Szegedy, 2015) is adopted; then an activation process using rectified linear units (ReLU) (Nair & Hinton, 2010) is applied. At the end of each dilated convolutional block, the produced features maps of each dilation factor are concatenated to form the input of the next block. A summarization of our proposed network architecture is provided in table 5.1, where a convolutional kernel size of 3×3 is utilized across the network, except for the last output layer. At that layer, a kernel size of 1×1 is utilized with no padding to perform the softmax activation on each pixel. The main output head layer is utilized in both training and testing processes, where it outputs C_{GT} predictions that matches the number of classes in the ground truth. On the other hand, the auxiliary over-clustering head layer is used only during

the training process to improve the learning process. It outputs C_{OC} predictions that refers to a number of classes that is larger than the ground truth classes' number (i.e., $C_{OC} > C_{GT}$).

Table 5.1 The network architecture of the proposed UL model

Layer	**Dilation factor (d)**	**Output shape**
input -InputLayer-		4 x 600 x 600
→ d_conv2D_1_1 (w/ padding)	1	13 x 600 x 600
→ d_conv2D_1_2 (w/ padding)	2	13 x 600 x 600
→ d_conv2D_1_4 (w/ padding)	4	13 x 600 x 600
→ d_conv2D_1_8 (w/ padding)	8	13 x 600 x 600
→ d_conv2D_1_16 (w/ padding)	16	12 x 600 x 600
conv_block_1 (concat.)		= 64 x 600 x 600
→ d_conv2D_2_1 (w/ padding)	1	26 x 600 x 600
→ d_conv2D_2_2 (w/ padding)	2	26 x 600 x 600
→ d_conv2D_2_4 (w/ padding)	4	26 x 600 x 600
→ d_conv2D_2_8 (w/ padding)	8	26 x 600 x 600
→ d_conv2D_2_16 (w/ padding)	16	24 x 600 x 600
conv_block_2 (concat.)		= 128 x 600 x 600
→ d_conv2D_3_1 (w/ padding)	1	26 x 600 x 600
→ d_conv2D_3_2 (w/ padding)	2	26 x 600 x 600
→ d_conv2D_3_4 (w/ padding)	4	26 x 600 x 600
→ d_conv2D_3_8 (w/ padding)	8	26 x 600 x 600
→ d_conv2D_3_16 (w/ padding)	16	24 x 600 x 600
conv_block_3 (concat.)		= 128 x 600 x 600
→ d_conv2D_4_1 (w/ padding)	1	26 x 600 x 600
→ d_conv2D_4_2 (w/ padding)	2	26 x 600 x 600
→ d_conv2D_4_4 (w/ padding)	4	26 x 600 x 600
→ d_conv2D_4_8 (w/ padding)	8	26 x 600 x 600
→ d_conv2D_4_16 (w/ padding)	16	24 x 600 x 600
conv_block_4 (concat.)		= 128 x 600 x 600
main output head		C_{GT}* x 600 x 600
auxiliary over-clustering head [at training only]		C_{OC}** x 600 x 600

* Number of ground truth semantic classes.
** Number of over-clustering semantic classes.

5.3.2 Data pre-processing

For a dataset of document images $\mathcal{X}$ that contains N unlabeled samples, three stages of pre-processing are applied to each input document image $x_i \in \mathbb{R}^{3\times H\times W}$.

5.3.2.1 Stage 1: Obtaining the distance transform (DT)

In order to obtain the distance transform (DT) of x_i, a binarization process is performed on it first. Specifically, a phase-based binarization algorithm (Nafchi, Moghaddam & Cheriet, 2014) is utilized. This algorithm is based on phase-congruency and uses the phase of wavelet transform to accurately determine the pixels of the text and the edges of its strokes. In addition, it is capable of dealing with dark backgrounds, which is a common feature in some historical manuscripts. Afterwards, the obtained binarized image, Fig. 5.4-b, is used to calculate the distance transform (DT) of x_i, Fig. 5.4-c, based on the chessboard distance metric (Cantrell, 2000). Particularly, each pixel in the acquired distance transform (DT), $x_{i_{dt}}$, represents the chessboard distance between that pixel $x_{i_{dt}} = (h_{dt}, v_{dt})$ and its nearest boundary pixel $x_{i_b} = (h_b, v_b)$, where $h \in \{1, \ldots, H\}$ and $v \in \{1, \ldots, W\}$. Formally, the chessboard-based distance transform for each pixel $x_{i_{dt}}$ is calculated through,

$$x_{i_{dt}} = max(|h_b - h_{dt}|, |v_b - v_{dt}|). \tag{5.1}$$

Afterwards, an upper bound condition is set, in which any pixel distance value $x_{i_{dt}}$ above a certain threshold β will be set to β

$$x_{i_{dt}} = min(x_{i_{dt}}, \beta). \tag{5.2}$$

This process results in obtaining a distance transform (DT) channel to be concatenated with the other RGB channels to achieve an image $x_i \in \mathbb{R}^{4 \times H \times W}$ that contains four channels R, G, B and DT. Note that all the four channel are normalized, where all the pixels values in RGB channels are divided by 255, and in the DT channel by β.

More details on the effect of utilizing different DT metrics and the choice of the upper bound β, among several conditions, and their effects on the performance are discussed in more details at subsection 5.4.4.1.

5.3.2.2 Stage 2: Obtaining the image patches

Since the process of semantic segmentation is pixel-based, it is computationally expensive. Therefore, the training process is to be operated on pairs of image patches, and not on whole images. Specifically, for each image $x_i \in \mathbb{R}^{4\times H\times W}$, many patches are being obtained, where every image patch x_u is centered at a pixel location $u \in \omega$, where $\omega = \{1, \ldots, H\}x\{1, \ldots, W\}$.

5.3.2.3 Stage 3: Pairing the image patches

As discussed in details in the upcoming subsection, the training process of our proposed approach depends mainly on a pair of two images as its input. Those image patches pairs are generated using i) a pre-defined patch spatial shift t (i.e., where image patches that are close together are perceptually grouped together) and ii) a pre-defined geometric transformation g to aid in learning representations that are invariant to such transformation. Ideally, g and t shall be applied on each individual patches x_u at location u to obtain the relevant image patch pair $(x_u, g(x_{u+t}))$, where $g(x_{u+t})$ is the neighbour patch at location $u + t$ after applying the transformation g. But, practically, it is way more efficient to apply g on the entire image x, which inherently applies g on all the image's patches at the same time, in parallel. Specifically, g consists of a random scaling with a factor that ranges from 0.4 to 1.6 and a random horizontal flipping. It is applied to each image $x_i \in \mathcal{X}$ to produce a transformed version of that image gx_i. Additionally, the used spatial displacement t is set as follows, $t \in T = \{5, \ldots, 10\}$ pixels.

The above three pre-processing stages lead to image pairs $(x_{i_u}, g(x_{i_{u+t}})), i = \{1, \ldots, N\}$ with each $x_i \in \mathbb{R}^{4\times H\times W}$. Such pairs are used as an input to the training stage. More details on obtaining the pair of patches in practice, during the training stage, is discussed in more detail in the upcoming subsection.

5.3.3 Training process

5.3.3.1 Background pixels masks

In order to improve the learning during the training process, the distance transform (DT) channel of each image in the image pair, $(x_u, g(x_{u+t}))$, is utilized to automatically obtain a mask of the background pixels in that image. This mask is obtained in a totally unsupervised manner and can be applied to any document images dataset since the domination of the background pixels is a common challenge across all datasets of document images.

In natural images, the distance transform (DT) of each pixel, obtained in Eq.(5.1), represents how far that pixel is from the closest boundary. In document images, such boundary can be considered as the nearest textual data. For instance, the farther the pixel from the text boundary, the larger the distance is; similarly, the nearer, the smaller the distance is. Therefore, that distance transform information, $0 \leq x_{i_{dt}} \leq \beta$, can be useful in obtaining the likeliness of each pixel being near or far from the text boundary; therefore, belonging to the background of the document image.

Formally, to convert the distance transform information of x_i into a an applicable pixel-level background mask $m_i \in \mathbb{R}^{1\times H\times W}$, each pixel in the mask, m_{i_p}, to be calculated through,

$$m_{i_p} = e^{\frac{-x_{i_{dt}}}{a}}, \tag{5.3}$$

where $-x_{i_{dt}}$ is the negated value of the DT of each pixel and a is a constant factor that controls the strictness of considering a pixel as a background or not. As a result, for each image pair $(x_u, g(x_{u+t}))$, a pair of corresponding background masks is generated $(m_u, g(m_{u+t}))$. Fig. 5.2-c shows examples of automatically obtained background masks for full images m_i.

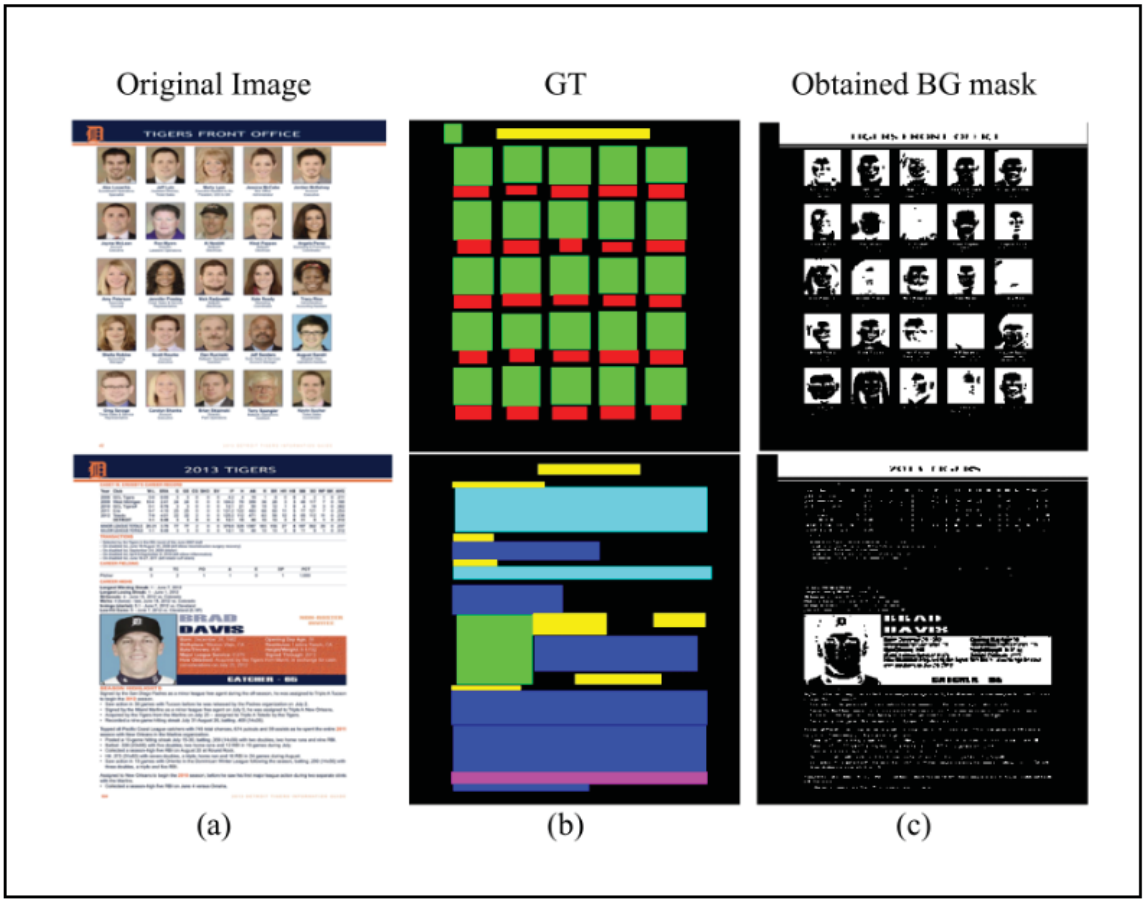


Figure 5.2 Examples of automatically obtained background (BG) masks for full images m_i using a=10. Note the major similarities between the BG pixels in (b) (i.e., the black pixels in the GT) and the obtained BG pixels in (c) (i.e., the black pixels in m_i)

5.3.3.2 Objective function

Going through our proposed network, the representations $(\phi(x_u), \phi(g(x_{u+t})))$ of each images pair, $(x_u, g(x_{u+t}))$, are obtained initially. In fact, these representations represent the predicted labels, where $\phi \in \mathbb{R}^{C_{GT} \times H \times W}$. Moreover, these representations are processed with the background masks $(m_u, g(m_{u+t}))$ to produce an updated representations that are utilized as an input to our mutual information (MI) based objective function. More details are discussed below.

Mutual information is a metric on how much information is shared between two instances. Considering that $\phi(x_u) = \phi_u(x)$, let the first instance of our mutual information formula

$z \sim \phi_u(x_i)$ embodies the representation of each original image x_i at the patch centered at the pixel u. And in order to well-correlate z with the correspondent representation of the transformed version of the original image, $\phi_{u+t}(gx_i)$, this version needs to be rolled-back to its original geometric state before the transformation, where $z' = [g^{-1}\phi(gx_i)]_{u+t}$.

Meanwhile, the masks $(m_u, g(m_{u+t}))$, obtained in Eq.(5.3), are applied to each channel of the corresponding representations $(\phi_u(x), [g^{-1}\phi(gx)]_{u+t})$, where an element-wise multiplication is utilized. This produces newly updated representations $(\Phi_u(x), [g^{-1}\Phi(gx)]_{u+t})$ that weight all the masked background pixels. Specifically, since the masks $(m_u, g(m_{u+t}))$ have high values (close to 1) on the text pixels and the areas around the text, and low values farther from text (i.e. background), applying the element-wise multiplication attenuates the background pixels effect during the process of calculating the mutual information and performing the optimization process.

Formally, inspired by (Ji *et al.*, 2019), the objective function of our proposed approach is based on maximizing the mutual information I estimate between z and z':

$$\max_{\Phi} I(z, z') = \max_{\Phi} I(\Phi_u(x_i), [g^{-1}\Phi(gx_i)]_{u+t}) \tag{5.4}$$

Considering that the weighted representation $\Phi(x_i)$ can be interpreted as the label of an image x_i, the main goal of Eq.(5.4) is to maximize I between every individual patch prediction $\Phi_u(x_i)$ and the prediction of the neighbouring patch $[g^{-1}\Phi(gx_i)]_{u+t}$ to conserve what is common between them, and neglecting any other details that are specific to only one of them. This mutual information I can be calculated through:

$$I(z, z') = P(z, z').\ln \frac{P(z,z')}{P(z).P(z')} \tag{5.5}$$

Besides, the joint (co-occurrence) probability $P(z, z')$ can be calculated by:

$$\begin{aligned} P(z, z') &= \sum_{t \in T} P(z, z'|t).P(t) \\ &= \frac{1}{n|G||\omega||T|} \sum_{i=1}^{N} \sum_{g \in G} \sum_{t \in T} \sum_{u \in \omega} \Phi_u(x_i).[g^{-1}\Phi(gx_i)]_{u+t}^{\intercal}, \end{aligned} \tag{5.6}$$

which is mainly a common co-occurrence probability estimate for each paired data (z, z'), in which z relates to z' by a spatial displacement t. In fact, the sequential sums in the above equation can be performed through the convolution process, where the processing of all pixels u and their related displacement t (i.e., pair of patches representations) can be performed in parallel.

Generally, and thanks to the novel representations $(\Phi_u(x), [g^{-1}\Phi(gx)]_{u+t})$ with weighted background pixels, the process of maximizing the mutual information I in our objective function, Eq.(5.4), leads to learning features that cope well with the document images unique characteristics. Specifically, it encourages the network to equally learn features of various semantic classes, while neglecting the dominating effect of the background samples in the training set space.

5.4 Experimental results and discussion

5.4.1 Datasets

In order to demonstrate the generalization capabilities of our proposed approach, two publicly available datasets have been used. The first dataset is DSSE-200 dataset (Yang *et al.*, 2017). It contains 200 document images and seven semantic classes (i.e., background, table, figure, paragraph, section, list and caption). The second dataset is DIVA-HisDB dataset (Simistira, Seuret, Eichenberger, Garz, Liwicki & Ingold, 2016). It is composed of three medieval manuscripts (i.e., CB55, CSG18, CSG863), which contains 150 document images that are annotated pixel-wise (i.e., 50 images per manuscript), and four semantic classes (i.e., background, main text body, comments and decoration figures).

5.4.2 Evaluation

To evaluate our proposed unsupervised document image semantic segmentation approach, we follow the same standard evaluation protocols presented in the literature (Ji *et al.*, 2019; Long, Shelhamer & Darrell, 2015). Specifically, our performance evaluation is based on three metrics, per-pixel accuracy, mean intersection-over-union (IoU) and weighted IoU.

For calculating the per-pixel accuracy, the following formula is used:

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP}, \tag{5.7}$$

where TP, TN, FN and FP denote the per-pixel true positives, true negatives, false negatives and false positives, respectively. Additionally, the IoU for each class in the dataset is calculated by:

$$IoU = \frac{TP}{TP+FP+FN}. \tag{5.8}$$

To calculate the mean IoU, the average of all the $IoUs$ (i.e., for all the classes) is calculated. In fact, for the mean IoU metric, a uniform distribution of the number of pixels per class is assumed, which is not the case in the document image datasets (more details are in section 5.3). While the actual distribution of the number of pixels per class is considered for calculating the weighted IoU metric, which makes it a more suitable metric to evaluate the performance for the document image datasets.

For the calculation of the weighted IoU, the IoU of each class (i.e., which is obtained in Eq.(5.8)) is weighted by the number of pixels in that class. Afterwards, the average of all the obtained weighted $IoUs$ (for all the classes) is computed.

Note that all samples of the dataset are utilized during both the training and testing processes. During the training process, our model is trained on unlabeled samples, without the need to any annotation. On the other hand, during the testing process, the samples and their ground-truth annotations are utilized in the evaluation process of the model. Specifically, a linear assignment

algorithm (i.e., the Hungarian algorithm (Kuhn, 1955)) is utilized to find the optimal mapping between each learned cluster and a ground-truth class.

5.4.3 Implementation details

All the performed training and testing processes have been carried out on a GeForce RTX 2080 GPU, with models implemented using PyTorch [7] library. To train our models, the Adam optimization algorithm (Kingma & Ba, 2014) has been utilized with a learning rate of $1e-5$ and a batch size of 2 image pairs. The training time is around 10 minutes per epoch for 600 epochs.

During the training phase only, an auxiliary over-clustering output head is utilized, in addition to the main output head, with C_{OC} that is 2-3 times larger than C_{GT}. For instance, in case of DSSE-200 dataset, $C_{OC} = 15$ and $C_{GT} = 7$; while with DIVA-HisDB dataset, $C_{OC} = 10$ and $C_{GT} = 4$. Besides, the mask factor $a = 10$ is utilized since it leads to an $m_{i_p} = 0$ when the pixel is far from the text (background) and $m_{i_p} = 1$ when the pixel distance is 0 (text/image). Moreover, an upper bound threshold $\beta = 15$ is set.

5.4.4 Results

A detailed analysis on utilizing and tuning various parameters and their effect on obtaining reliable results has been performed. In the following subsections, the related experiments and their acquired observations with analysis are discussed in more details. Note that the training process in all the following experiments have been performed from scratch, without any pre-training and using an input size of 600 x 600. In addition, the reported results are based on the DSSE-200 dataset.

5.4.4.1 Effect of the representation space

First, experiments have been performed to obtain the best representation space that can lead to reliable unsupervised semantic segmentation results. Specifically, we have studied various

[7] http://pytorch.org/

representation spaces that consider either RGB -3 channels-, a distance transform (DT) -1 channel- or a combination of both RGB and distance transform (DT) -4 channels-.

The visual results shown in Fig. 5.3-c and Fig. 5.3-d demonstrate that using RGB or DT solely can barely work for distinguishing between the background and the foreground of the document image, but it is completely ineffective for performing unsupervised semantic document segmentation. Specifically, the learned features from the RGB representation space or the DT representation space are not distinctive enough to lead to a reliable separation between the different categories of foreground (e.g., text, figure, etc.) or the different semantic classes of text (e.g., caption, list, table and section). On the other hand, as shown in Fig. 5.3-e, combining the distance transform (DT) with the RGB representation space (RGBDT) has led to obtaining much better discriminative features that led to a noticeable improvement in segmenting different semantic classes in an unsupervised manner.

Yet, as demonstrated in the DT and RGBDT results in Fig. 5.3, some extra contours are always introduced around the main text body of the document image whenever a distance transform is used. Checking the contours-related pixels in the distance transform channel, Fig. 5.4-c, it is found that the distance values of those pixels are different than the distance values for the background pixels. This has prevented the contours-related pixels from being well-perceived as background by the learned model.

To avoid this, a condition has been introduced during the calculation of the distance transform in an attempt to unify the distance values of both the pixels of the background and the document's main body extra contour. Utilizing the chessboard distance transform, Eq.(5.1), two upper bound (UB) conditions have been investigated: a gradual condition and a strict condition. First, a gradual upper bound condition based on the following formula has been investigated.

$$x_{i_{dt}}^{new} = log(1 + \alpha x_{i_{dt}}^{2}), \tag{5.9}$$

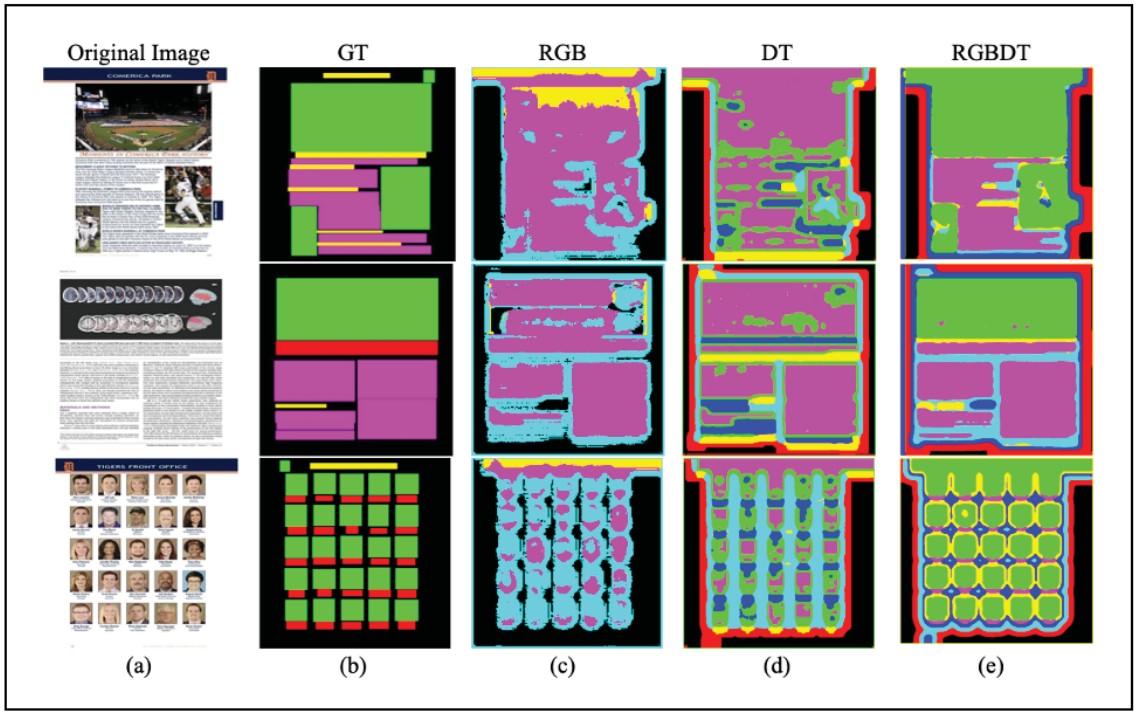


Figure 5.3 The effect of utilizing different types of representation spaces on the unsupervised segmentation performance

where α is a parameter to control the upper bound value that a pixel distance value $x_{i_{dt}}$ will saturate at. Fig. 5.4-d demonstrates the result of utilizing this condition with $\alpha = 1$, where the extra contour pixels are gradually getting closer to the distance values of the background pixels.

On the other hand, a strict upper bound condition can be set where any distance value $x_{i_{dt}}$ above a certain threshold β will be set to β

$$x_{i_{dt}} = min(x_{i_{dt}}, \beta), \tag{5.10}$$

Fig. 5.4-e shows that using a strict upper bound (e.g., $\beta = 15$) condition led to a more representative distance transform. In this case, all the pixels around the document's main body have the same distance value of the document's background. Fig. 5.5-d shows the effectiveness of utilizing the strict upper bound condition in limiting the document's main body extra contours

comparing to utilizing the same distance transform without any upper bound condition, Fig. 5.5-b, or with a gradual upper bound condition, Fig. 5.5-c.

Based on the previously discussed visual qualitative results and the quantitative results presented in Table 5.2, the combined representation space of RGB and distance transform (RGBDT) with a strict upper bound (UB) condition has shown its reliability; therefore, it is the representation space setting to be used for the rest of this paper.

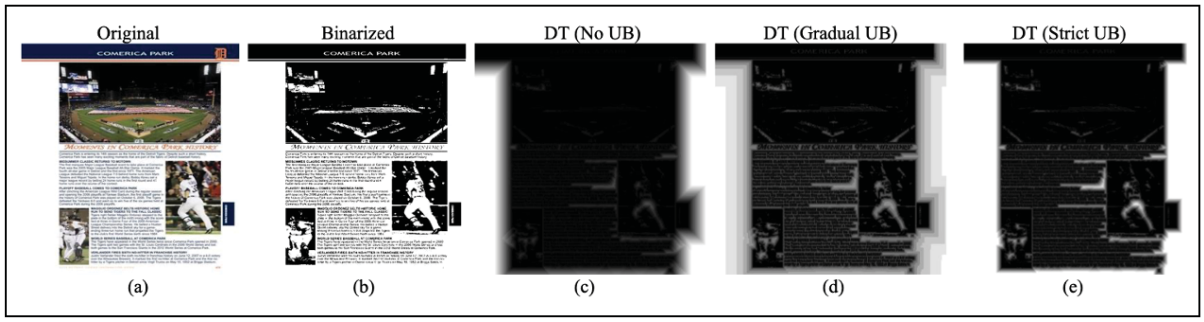


Figure 5.4 The effect of utilizing different upper bound (UB) conditions on obtaining the distance transform (DT) from the binarized image

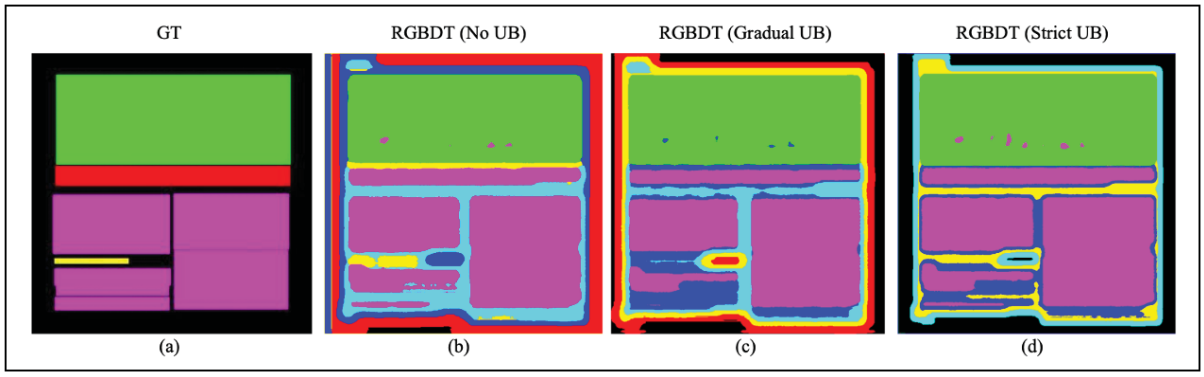


Figure 5.5 The effect of utilizing different upper bound (UB) conditions on limiting the document's main body extra contours

5.4.4.2 Effect of max-pooling

In order to test the effect of utilizing max-pooling on the unsupervised segmentation performance, we have conducted three experiments with identical parameters and alike network architectures

Table 5.2 The performance in terms of mean IoU (%) on the DSSE-200 dataset using different input spaces

Input space	Mean IoU
RGB	20.24
DT	21.13
RGBDT (No UB)	23.35
RGBDT (Gradual UB)	25.08
RGBDT (Strict UB)	**26.34**

except that the max-pooling layer(s) (i.e., and its corresponding interpolation process) have been removed gradually from one experiment to another. Table 5.3 presents the network architectures of the three experiments with their corresponding mean IoU performance.

Fig. 5.6-b and Fig. 5.6-c demonstrate the segmentation results of the network architectures that include three max-pooling layers and one max-pooling layer, respectively. On the other hand, Fig. 5.6-d shows the result of the network architecture that does not consider any max-pooling layer(s).

Although the same convolutional filter size has been utilized across the three experiments, a large difference in the visual results has been obtained. In fact, the obtained visual results, in Fig. 5.6, emphasizes the critical role that the max-pooling layer(s) plays in learning or missing some critical features. Specifically, although using max-pooling has led to an apparent better visual segmentation results and higher mean IoU, it has overlooked the small-sized semantic classes (e.g., the class 'section' -in yellow- at Fig. 5.6-a). Meanwhile, removing the max-pooling layer(s) has helped in capturing more low-level spatial features that aid in locating and segmenting such small-sized semantic classes. As a drawback of avoiding the max-pooling layer(s), the obtained representation has lost part of its sensitivity to the global semantic correlations at the spatial-level, which led to an undesired oversegmentation in large-sized semantic classes (e.g., the oversegmentation in the 'figure' class -in green- at Fig. 5.6-d). To get over this issue and other related issues, more investigations have been conducted on utilizing various receptive field sizes.

Table 5.3 The performance in terms of mean IoU (%) on the DSSE-200 dataset using network architectures with different number of max-pooling layer(s). (64,. . . ,512) represents the number of filters for each layer, where 'M' is a max-pooling layer

	Network architecture	**Mean IoU**
w/ 3 max-pooling	(64, 'M', 128,'M', 256, 'M', 512)	24.36
w/ 1 max-pooling	(64, 128,'M', 256, 256, 512, 512)	24.74
w/o max-pooling	(64, 128, 256, 256, 512, 512)	22.55

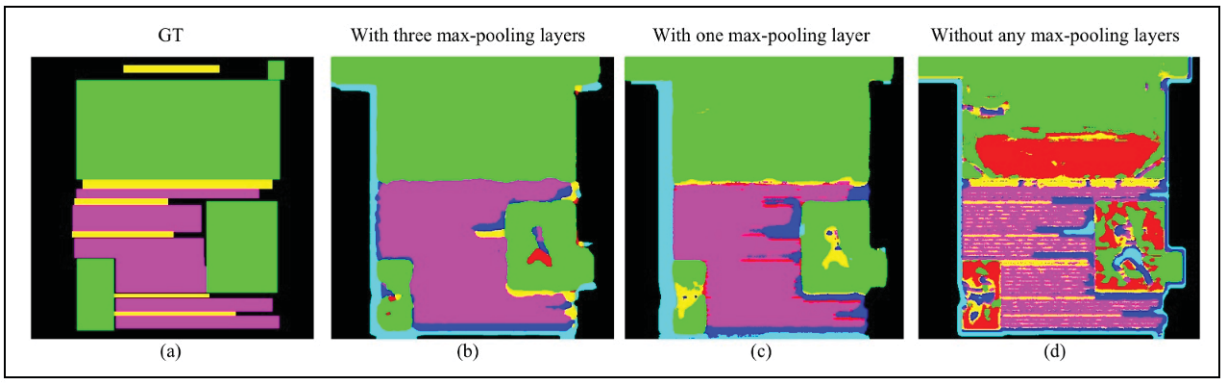


Figure 5.6 The effect of utilizing max-pooling on the unsupervised segmentation performance

5.4.4.3 Effect of receptive field size

To demonstrate the impact of the CNN's receptive field size on the performance of the unsupervised segmentation task and eliminating the undesirable oversegmentation in large-sized semantic classes, four experiments have been conducted using different convolutional filter sizes. Specifically, the network architecture in section 5.3.1 has been utilized across all the experiments, with an exception that a different filter size has been considered for each of these experiments (e.g., 3x3, 5x5, 7x7, 11x11). In this network architecture, the pooling process is avoided, four dilated convolution blocks are utilized and the selected filter size is utilized across all the layers of the network.

Results at Fig. 5.7 demonstrate that using large filter sizes (e.g., 11x11 and 7x7) are good for large-sized semantic classes (e.g., the 'figure' class -in green-), but they led to missing

small-sized semantic classes (e.g., the 'section' class -in yellow-) and obtaining less sharp segmented bounding boxes. In contrast, utilizing a filter size of 3x3 works well for small-sized semantic classes (e.g., the 'section' class -in yellow-), where sharper segmented bounding boxes are obtained. Moreover, the undesired oversegmentation in large-sized semantic classes have been eliminated largely (i.e., comparing to Fig. 5.6-d).

Based on the previous qualitative findings and the quantitative results in Table 5.4, a network architecture that avoids the pooling process, considers dilated convolution blocks and utilize a small convolutional filter size (e.g., 3x3) is considered to work efficiently and reliably with the task of unsupervised document images semantic segmentation.

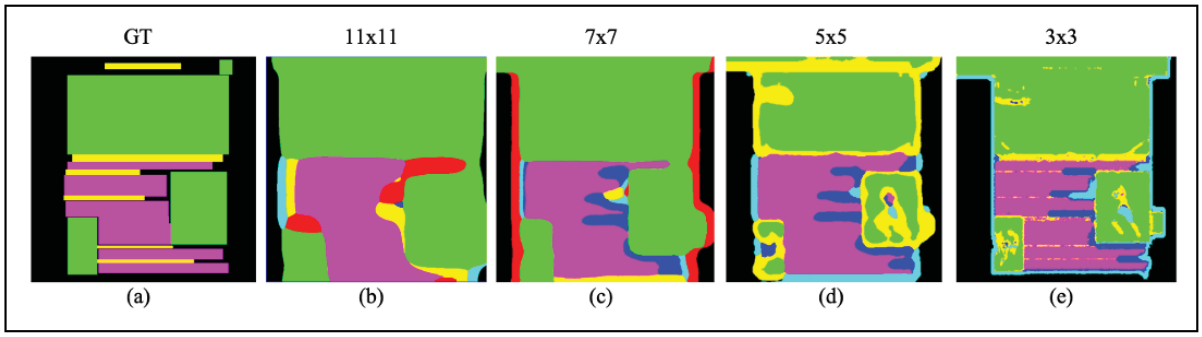


Figure 5.7 The effect of utilizing different convolutional filter sizes on the unsupervised segmentation performance

Table 5.4 The performance in terms of mean IoU (%) on the DSSE-200 dataset using various convolutional filter sizes and dilated convolutional blocks

Filter size	**Mean IoU**
11x11	17.51
7x7	20.66
5x5	23.90
3x3	24.74

5.4.5 Discussion

Since performing document images semantic segmentation based on unsupervised learning is a relatively unexplored area of research, it was challenging to obtain state-of-the-art results that are suitable to be compared with the results of our proposed unsupervised learning (UL) based approach; therefore, we had to establish baselines, from scratch, to fairly evaluate our approach. Specifically, our proposed approach is compared with two baseline algorithms. First, k-means (Lloyd, 1982), which is a simple clustering algorithm that is widely used in the literature. Additionally, Non-negative Matrix Tri-Factorization (NMTF) (Yoo & Choi, 2010), which is a more complex clustering algorithm that provides further representative cluster centers in comparison with k-means. More details on how the datasets have been prepared to be fed to the baseline algorithms; in addition to their implementations are discussed in 5.6.

Table 5.5 compares the semantic segmentation performance of our proposed approach with that of the two baseline algorithms. The performance is reported based on four datasets and in terms of three standard metrics; accuracy, mean IoU and weighted IoU. Results show our approach to outperform the two baseline algorithms for all the datasets. Furthermore, our proposed approach is compared to the vanilla IIC approach (Ji *et al.*, 2019) on the DSSE-200 dataset. The results demonstrate that considering the adapted objective function of (Ji *et al.*, 2019), in the proposed approach, alongside our introduced document-related contributions has led to more than 8% improvement in the performance using the weighted IoU metric. In contrast, applying the supervised learning based MFCN approach (Yang *et al.*, 2017) on the DSSE-200 dataset can boost the weighted IoU performance to up to 28% higher than the proposed approach, which is expected considering the unsupervised nature of our work. Specifically, on one hand, our proposed approach does not need any labeled samples during the training process. On the other hand, (Yang *et al.*, 2017) has to access a large amount of labeled training samples (i.e., around 135,000 annotated document images) to achieve such performance boost.

Discussing the baseline algorithms, we note that the way in which those algorithms and our approach function are different. In fact, NMTF and k-means are based on information of

Table 5.5 The performance (%) on four different datasets obtained by our proposed approach compared to various approaches

Approach		**Accuracy**	**Mean IoU**	**Weighted IoU**
k-means	(DSSE)	49.6	18.9	41.2
NMTF	(DSSE)	53.4	20.3	42.5
IIC (Ji *et al.*, 2019)	(DSSE)	51.5	21.6	44.1
UL (**ours**)	(DSSE)	**61.5**	**27.4**	**52.3**
k-means	(CB55)	42.0	18.2	36.0
NMTF	(CB55)	39.6	18.4	33.2
UL (**ours**)	(CB55)	**67.8**	**34.6**	**61.3**
k-means	(CSG18)	38.8	16.1	33.7
NMTF	(CSG18)	35.1	15.6	29.0
UL (**ours**)	(CSG18)	**63.0**	**27.0**	**57.8**
k-means	(CSG863)	45.1	19.8	39.5
NMTF	(CSG863)	45.6	20.7	39.0
UL (**ours**)	(CSG863)	**66.4**	**29.2**	**62.3**

individual pixels. While, on the other hand, our proposed approach is based on information of a region of pixels. Specifically, k-means and NMTF use the similarity between individual pixels to separate them into different groups. In contrast, our approach uses local regional information, combined in different ways, for clustering and hence segmentation.

Fig. 5.8 shows three different document image samples from the DSSE dataset (Yang *et al.*, 2017) and their corresponding results using our proposed approach and the baseline methods. Starting from the far left column in Fig. 5.8, the first column shows the ground truth of the samples, where it is clear how the white pixels between the text are considered as text (i.e., the same goes for other regions of the foreground). In the second column, the results of the k-means method are shown, in which the segmentation is very local and even small regions of text (or strokes) are separated into different classes and clearly there is no separation between different regions containing text (i.e., such as list, caption, etc.). Additionally, the third column shows the results of the NMTF method, where the segmentation is still pixel-based and very local. Although NMTF could use different features and their combination, but the effect is still insignificant. Lastly, the forth column shows the results of our proposed approach, in which the

effect of the receptive field size and dilated convolution blocks are clearly demonstrated, where the learned representation can well-capture regional spatial information to separate different semantic classes.

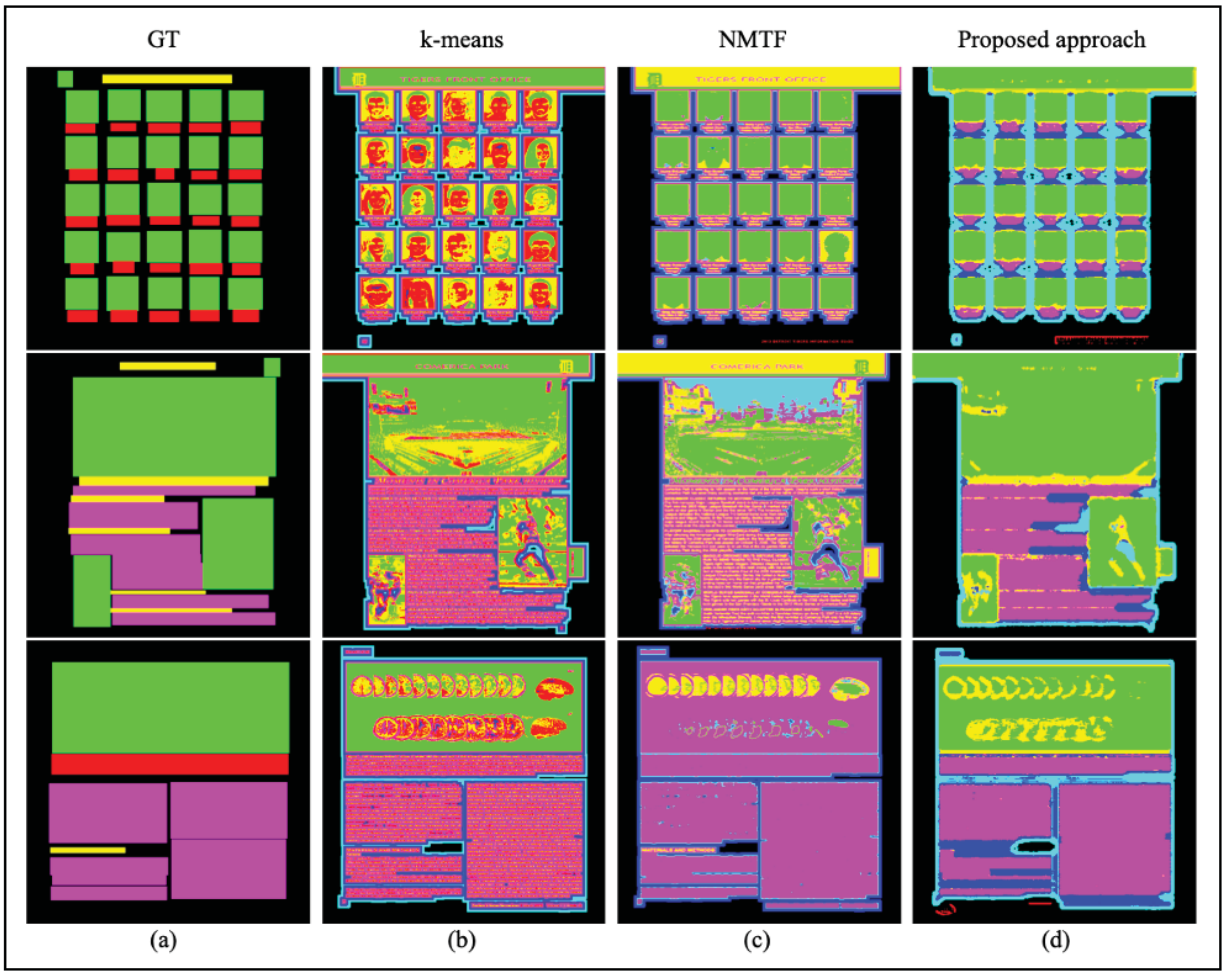


Figure 5.8 Comparing the results of k-means (b), NMTF (c) and our proposed approach (d)

In fact, we can see that for the small 'figures' -in green- (e.g., the top image in column (d)), our proposed approach's segmentation is producing uniform regions without undesired oversegmentation for those areas, unlike k-means and NMTF. The same goes for large 'figures' (e.g., the middle and bottom images in column (d)), except that a bit of oversegmentation appears in regions that contains very large white pixels. Yet, the oversegmentation is way

less outstanding, unlike k-means and NMTF (e.g., the results at the middle row) where the oversegmentation of figures is totally visible.

In the middle row, one of the 'section' heading regions -in yellow- is segmented clearly by our proposed method. This demonstrates that with enough surrounding information, our approach can capture semantic information without the need for annotations. Moreover, the 'caption' class -in red- is one of the rare and hard classes to be captured by an unsupervised learning approach. This is considering that this class shares a lot of visual characteristics with the 'text' class -in purple-, which makes those two classes barely distinguishable, even from a human perspective (e.g., the 'caption' class at the third row sample).

5.5 Conclusion

We proposed an end-to-end approach for semantically segmenting document images without the need to any labeled data, textual information or dataset-dependant heuristics techniques. The proposed approach utilizes a combined representation space, a novel automatic dataset re-balancing technique and dilated convolutional layers to obtain a novel learned representation that can overcome the challenges related to the unique characteristics of the document image and its semantic classes. Our results demonstrate that our approach is robust, where it outperforms the baseline approaches on various public datasets. Although other supervised-learning based approaches might provide better semantic segmentation performance, our proposed approach fully opens up novel research verticals, in which the widely available unlabeled data can be utilized in performing document image semantic segmentation. In future work, we will include investigating the effect of increasing the training set size on the inference performance of our approach. Additionally, further investigations can be performed on the generalization capacity of the proposed approach with large datasets that contains millions of document images.

Acknowledgments

The authors thank the NSERC of Canada (Grant no. RGPIN 2019-05230) for their financial support.

5.6 Appendix - More details on the used baselines (k-means and NMTF)

In appendix 5.6, we provide more insights on how the datasets have been prepared to be fed to the baselines algorithms. In addition, we summarize the baseline algorithms implementation in order to obtain relevant results.

In order to prepare the datasets to be fed to these algorithms, all the images' channels (i.e., RGB and DT) are converted to a matrix with the size of $e_i \times P$, where e_i is the number of pixels in the image and P is its dimensionality. Then, all of the corresponding matrices are concatenated into a one matrix Y of size $E \times P$, where $E = \{e_1 + e_2 + \dots\}$.

In k-means, the goal is to find the clusters by minimizing the distances between the samples and the cluster centers. The main parameter to be set is the number of clusters, which is set equal to the number of classes in differnet datasets, C_{GT}.

On the other hand, Non-negative matrix factorization (NMF) (Lee & Seung, 2000) was originally proposed for low rank approximation of matrix by decomposing it to the product of two matrices with lower ranks. NMF can also be considered as a soft-clustering algorithm which is compared to k-means (i.e., which assigns a sample to a specific cluster), where it assigns samples to different clusters with coefficients that determine the degree of association. NMF is a versatile algorithm and its objective function can be modified in different ways to obtain desired properties. One of the variants of NMF is Non-negative matrix tri-Factorization (NMTF) (Yoo & Choi, 2010), in which the data matrix is decomposed to the product of 3 matrices $Y = WSG^T$.

NMTF shows the properties of co-clustering (i.e., clustering the rows and columns of the data matrix simultaneously) that produce better clustering results. The objective function of NMTF is:

$$\begin{aligned} &\min_{W,S,G} \left\|Y - WSG^T\right\|_F^2, \\ s.t. \quad &W \geq 0,\ S \geq 0,\ G \geq 0 \\ &W^T W = I,\ G^T G = I, \end{aligned} \tag{5.11}$$

where Y is a $E \times P$ matrix, and W, S, G are matrices of sizes $E \times k$, $k \times r$ and $P \times r$, respectively. In fact, the two orthogonality constraints are forcing the algorithm to produce a k-means-like results. The parameter k is the number of the desired clusters for the rows of the data matrix Y. Additionally, the parameter r is the number of clusters for the columns of data matrix Y. Both k and r are set to the number of the desired clusters for each dataset, which is equal to the number of classes C_{GT}. In order to convert the results of NMTF (soft-clustering) to cluster assignment, we utilize a scheme based on applying k-means on the coefficients obtained by NMTF.

CHAPTER 6

GENERAL DISCUSSION

This thesis has addressed several problems related to document image representations for document image analysis. The introduction and literature reviewed in Chapter 1 showed limitations of current document representations and their relevant features in tackling several practical and technical challenges. Specifically, the following question was investigated: what are efficient document representation approaches capable of obtaining representations that can handle processing large-scale datasets, provide reliable generalization during the deployment phase, and utilize much less labeled data or only unlabeled data during the training process? Considering these challenges and limitations, three research objectives have been established in Chapter 2 2 and led to proposing three novel document image representation approaches that can stand up to the practical, real-world challenges of the document image analysis field. These approaches, their relevant contributions and evaluations are discussed in Chapter 3, Chapter 4 and Chapter 5. In the following sections, our proposed approaches and contributions are discussed, focusing on their strength and limitations, considering the general advances they made in the document image analysis' state of the art.

6.1 Efficient document image representations for large scale dataset

In Chapter 3, the first objective was covered where we studied the practical, real-world challenges of the document image analysis field and proposed a reliable document representation approach that can generalize well to large scale datasets. For this purpose, we developed a classification approach based on an ensemble of four methods that rely on a broad spectrum of features that range from fully hand-designed features to hybrid and fully learned features. The methods are rule-based, layout-based, CNN-based, and transfer learning-based.

The proposed approach is the first study that reflects the practical challenges facing the document image analysis field when interfacing with real-world constraints. The obtained results demonstrate the performance consistency of our proposed approach across different time periods

and subjects within the utilized large-scale dataset (32 million document images). Despite the capability of our proposed approach in generalizing well to such a large-scale dataset, it still has some limitations. For instance, three methods -in the final approach ensemble- are depending on hand-designed features. Moreover, the utilized feature learning-based methods rely on supervised learning, where access to annotated training samples is essential. Both of those limitations have been studied and tackled in the following two objectives of the thesis. For the sake of completeness, our article -in Appendix I- proposes an additional document image classification approach that depends on fully-learned features utilizing limited amount of labeled data.

6.2 Efficient document image representation learning for classifying datasets with limited to no availability of labeled training data

The work in Chapter 4 offers three modifications over the previous contribution. First, it is fully based on feature learning, without any dependence on hand-designed features. Second, the proposed document representation learning approach does not require any annotated training document images during the pre-training step. It is based on unsupervised feature learning, where only unlabeled data is used. Finally, the performed classification process is based on the global context of the document image, unlike the previous work in which the classification process depends on the existence of a distinctive visual local characteristic (e.g., footnote) within the document image.

In this contribution, we developed an unsupervised pre-training-based framework that is simple, very effective in learning good features, and capable of consistently boosting both unsupervised and supervised classification performance. The approach is the first to perform an unsupervised document image classification using a representation that is entirely based on feature learning using unlabeled data and does not depend on any hand-crafted features. Although some other supervised pre-training approaches may provide more improvement in the classification performance compared to our proposed unsupervised pre-training approach, our approach has a crucial advantage of not requiring any additional manually annotated training samples. This is

unlike such supervised learning-based approaches that require an immense amount of annotated data (i.e., tens or hundreds of thousands) to reach such performance.

6.3 Efficient document image representation learning for semantically segmenting datasets with no availability of labeled training data

For each previously classified document image by the proposed approach in Chapter 4, further interpretation of that document's content is performed through semantic segmentation. Aligning with the previous work on avoiding utilizing any annotated document images during the training process, the work in Chapter 5 proposed an end-to-end approach for semantically segmenting a document image (pixel-wise) in a totally unsupervised manner.

The proposed approach is the first to conduct an unsupervised document image segmentation process using solely unlabeled data without any dependence on any textual information or dataset-dependant heuristics techniques. Better semantic segmentation performance might be obtained through other supervised-learning-based approaches. Nevertheless, a large amount of labeled training samples (i.e., hundreds of thousands) are needed to be accessed to achieve such a performance boost.

CONCLUSION AND RECOMMENDATIONS

In this thesis, we have presented original contributions to the state of the art of document image representation for the document image analysis field. Efficient document image representations shall be able to deal with real-world practical challenges rather than being tuned according to unrealistic assumptions optimized for specific research scenarios. Two main practical challenges have been studied in this thesis, generalizing well to large-scale datasets and dealing with limited to no availability of labeled training data.

First, document representations shall be capable of generalizing well to large-scale datasets that contain millions of document images. The contribution of this thesis shows directions for designing reliable document representations that consider the scale of large datasets and their associated generalization challenge at the core of the design process. In addition, the contribution emphasizes the efficiency of combining various representations to build a reliable tool that can empower experts in different fields with the possibility to analyze large-scale datasets effectively.

Additionally, obtaining document image representations that can deal with limited to no availability of labeled training data is of very high interest since, in real-world use-cases, the access to labeled data is restricted, while unlabeled data is abundant. The contributions of this thesis open up novel research verticals for utilizing millions of widely available unlabeled data in performing document image analysis. Specifically, two novel approaches for document image representation learning have been introduced in a particular sequence to highlight a proper framework for document image analysis. At first, when no labeled training data, or very few, are available for the document classification task, while many unlabeled data is accessible, our framework depends on unlabeled data to obtain a reliable classification process. Then, when each document image is classified, and no labeled training data is available, our framework relies on unlabeled data to perform further analysis within that document image by performing semantic segmentation.

Summary of contributions

In this thesis, the following novel contributions have been introduced:

- A document representation approach that can generalize well for a very large-scale dataset (32 million documents) has been introduced. This is the first study that considers the practical challenges which face the document image analysis field when interfacing with real-world constraints.

- An unsupervised document representation learning approach for document image classification is proposed. This is the first approach to perform an unsupervised document image classification using a representation that is entirely based on feature learning using unlabeled data and does not depend on any hand-crafted features.

- An unsupervised document representation learning approach for document image semantic segmentation is proposed. To the best of our knowledge, it is the first work to perform an unsupervised document image segmentation using solely unlabeled data and without depending on any textual information or dataset-dependant heuristics techniques.

Limitations and future work

Although the proposed document representation approach in Chapter 3 is capable of generalizing well to large-scale datasets, it still has some weaknesses. In the final classification approach ensemble, three out of the four explored methods are based on hand-designed features. Additionally, the fourth feature learning-based method in the ensemble is based on supervised learning, in which annotated training samples availability is crucial. Even though these weaknesses have been addressed partly by exploring unsupervised document representation learning approaches in the rest of the thesis, further investigations can still be conducted on exploring the performance of the proposed approach when extended to semantic segmentation tasks.

The document representation learning approaches proposed in Chapters 4 and 5 are mainly trained in an unsupervised manner, which is critical for eliminating the need for labeled data during the training process. The obtained learned representations have proved their efficiency and robustness. However, the performance of the proposed approaches is still limited comparing to other supervised-learning based approaches. Nonetheless, an immense amount of labeled training samples are required for such a performance boost to materialize. This performance limitation can be addressed by investigating recent semi-supervised learning techniques, where few labeled data are utilized during the training process alongside the unlabeled data. Moreover, further investigations can be performed on the generalization capacity of the proposed approaches with large-scale datasets that contain millions of document images.

Generally, the obtained representations in this work have demonstrated their effectiveness for document classification and semantic segmentation tasks, which cover only part of the document analysis pipeline. A further possible extension of this work would be to investigate the efficiency of the introduced concepts when transferred to additional analysis tasks, such as word segmentation and character recognition.

7.1 Articles in peer reviewed journals

1. Sherif Abuelwafa, Sara Zhalepour, Ehsan Arabnejad, Mohamed Mhiri, Emili- enne Greenfield, James P. Ascher, Sofia Bach, Victoria Svaikovsky, Alayne Moody, Andrew Piper, Chad Wellmon, and Mohamed Cheriet. “Detecting Footnotes in 32 million pages of ECCO.” Journal of Cultural Analytics. (December 3, 2018).

2. Sherif Abuelwafa, Marco Pedersoli, and Mohamed Cheriet. "Unsupervised exemplar-based learning for improved document image classification." IEEE Access 7 (2019): 133738-133748.

3. Sherif Abuelwafa, Ehsan Arabnejad, Marco Pedersoli, and Mohamed Cheriet. "Unsupervised learning for Document Image Semantic Segmentation." Submitted to Elsevier Pattern Recognition (September 2021).

4. Mohamed Mhiri, Sherif Abuelwafa, Christian Desrosiers, and Mohamed Cheriet. "Hierarchical representation learning using spherical k-means for segmentation-free word spotting." Pattern recognition letters 101 (2018): pp. 52-59.

7.2 Articles in peer reviewed conference proceedings

1. Sherif Abuelwafa, Mohamed Mhiri, Rachid Hedjam, Sara Zhalehpour, Andrew Piper, Chad Wellmon, and Mohamed Cheriet. "Feature learning for footnote-based document image classification." In International Conference Image Analysis and Recognition, Springer, Cham, (July, 2017): 643-650.

2. Mohamed Mhiri, Sherif Abuelwafa, Christian Desrosiers, and Mohamed Cheriet. "Footnote-based document image classification using 1D convolutional neural networks and histograms." In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, (2017): pp. 1-5.

APPENDIX I

FEATURE LEARNING FOR FOOTNOTE-BASED DOCUMENT IMAGE CLASSIFICATION

Sherif Abuelwafa [1], Mohamed Mhiri [1], Rachid Hedjam [1], Sara Zhalehpour [1], Andrew Piper [2], Chad Wellmon [3], Mohamed Cheriet[1]

[1] Département de génie des systèmes, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] McGill University, Montreal, Canada.
[3] University of Virginia, Charlottesville, Virginia, USA.

Published in the Proceeding of International Conference Image Analysis and Recognition (ICIAR), July 2017, Pages 643-650

Abstract

Classifying document images is a challenging problem that is confronted by many obstacles; specifically, the pivotal need of hand-designed features and the scarcity of labeled data. In this paper, a new approach for classifying document images, based on the availability of footnotes in them, is presented. Our proposed approach depends mainly on a Deep Belief Network (DBN) that consists of two phases, unsupervised pre-training and supervised fine-tuning. The main advantage of using this approach is its capability to automatically engineer the best features to be extracted from a raw document image for the sake of generating an efficient representation of it. This feature learning approach takes advantage of the vast amount of available unlabeled data and employs it with the limited number of labeled data. The obtained results show that the proposed approach provides an effective document images classification framework with a highly reliable performance.

Keywords: unsupervised feature learning, hierarchical representation learning, document image classification.

1. Introduction

Protecting humanity's cultural heritage is highly needed to understand our past and prepare for the future; therefore, digitizing historical manuscripts and printed books becomes an essential approach to guarantee a well preserved history and a widely accessible content to thousands of researchers around the globe. A great example that shows the adaptation of this global movement is the Eighteenth Century Collections Online (ECCO)[8] that contains around 200,000 volumes (32 million image pages) of online archive related to the eighteenth-century printed books.

In fact, introducing an efficient data-driven approach that can understand such huge amount of widely available historical data will play a vital role in unveiling the secrets of such precious content. Generally, understanding a historical document image involves a wide spectrum of subprocesses and tasks that range from layout analysis and document image classification to optical character recognition (OCR). In this work, the main focus will be on the task of classifying document images based on the presence of footnotes in them. Considering the valuable information that is usually contained in a footnote and the strong ties it creates with other documents; footnotes are believed to be reflecting how ideas can be circulated and exchanged between various manuscripts and books throughout centuries and civilizations (Grafton, 1997; Pasanek & Wellmon, 2015). Therefore, obtaining document images with footnotes has been raised to form the main focus of this research paper.

The performance of the document images classification process dependents highly on the used representation of the document image, where learning how to map the intensity values of the document images' pixels into a relevant decision (i.e., the presence of a footnote in the image or not) is critical. In order to obtain an expressive representation of the document image, the best features have to be captured from it. Such features help in getting better high-level representations of the raw data in a way that explicates the document image's main properties and facilitates the subsequent classification process.

[8] http://find.galegroup.com/ecco/

Most of the traditional systems for document image classification depend on carefully hand-designed features (e.g., SIFT (Lowe, 1999), SURF (Bay *et al.*, 2006) and HOG (Dalal & Triggs, 2005)). Those features are being engineered by experts relying on their prior knowledge regarding the used data and the desired application, which is a very complex process. Actually, such hand-designed features are labor-intensive, time consuming and cannot generalize well to new problems. Due to the difficulties related to engineering hand-designed features, we focus in this work on feature learning approaches (Bengio *et al.*, 2013) that can take advantage of the increasing amount of available informative data to automatically learn even better feature representations than the hand-designed ones. Utilizing feature learning in the field of document image classification has been adopted in some recent works, such as the work of (Kang *et al.*, 2014). But those works have relied heavily on using only labeled data for training their feature learning algorithms.

In order to train feature learning algorithms either labeled (supervised learning) or unlabeled (unsupervised learning) data can be used. Apparently, acquiring more data leads to better performance regardless of the used learning approach (Banko & Brill, 2001), where most of the recent breakthroughs in the results of machine learning approaches are actually due to the availability of a large amount of training data. And since obtaining enough labeled data to perform supervised feature learning is often a very difficult and expensive task due to the required time and labor for labeling, while a vast amount of unlabeled data is available and easily accessible; we are investigating in this research paper the capacity of incorporating an unsupervised feature learning algorithm alongside a supervised one to achieve an optimal approach for document image classification.

In the light of the previously mentioned challenges of hand-designed features and the scarcity of labeled data, the main contribution of this paper is in proposing and evaluating a feature learning approach for document image classification that is capable of the following; generating the best representation of an input raw image through automatically engineer the best features using a trainable feature extractor instead of using hand-crafted features. This is performed while

depending mainly on the largely available unlabeled images besides the restrictedly available labeled images.

The rest of the paper is organized as follows. The proposed approach is outlined in section 2. In section 3, the used dataset and the obtained experimental results are reviewed alongside some observations and related discussion. Finally, section 4 demonstrates the conclusion and the future work.

2. The Proposed Approach

Since a footnote is the center of interest in our document classification process; and through studying the document images that contain footnotes, we observed a compelling common feature. We found that document images with footnotes usually contain two different font sizes between their main text and their footnotes text, a property that turned out to be at the core of our hypothesis.

To train a classifier to differentiate efficiently between document images with footnotes and document images with no footnotes, we need to provide it with many positive and negative examples of document images that contain footnotes and do not contain footnotes, respectively. And since we have a very limited amount of labeled data with ground-truth; while, on the other hand, unlabeled data are widely available, it will be effective to utilize an efficient approach that can leverage such wide availability of unlabeled data. As a result of the above factors, our document classification algorithm is based mainly on teaming up an unsupervised pre-training phase with a supervised fine-tuning phase. In fact, this setup has proved to be very efficient in the case of scarcity of labeled data (Glorot, Bordes & Bengio, 2011). According to (Bengio *et al.*, 2013; Erhan *et al.*, 2010), exploiting the process of unsupervised pre-training in initializing a later supervised classification process can be certainly helpful for this classification process.

Generally, our proposed document classification approach consists of 4 stages, Fig. I-1. This approach depends at its core on a feature learning model that is based on Deep belief network (DBN) architecture (Hinton *et al.*, 2006; Lee, Ekanadham & Ng, 2007) and composed of two

main phases (unsupervised pre-training and a supervised fine-tuning). The following subsections will provide more insights about each stage of them.

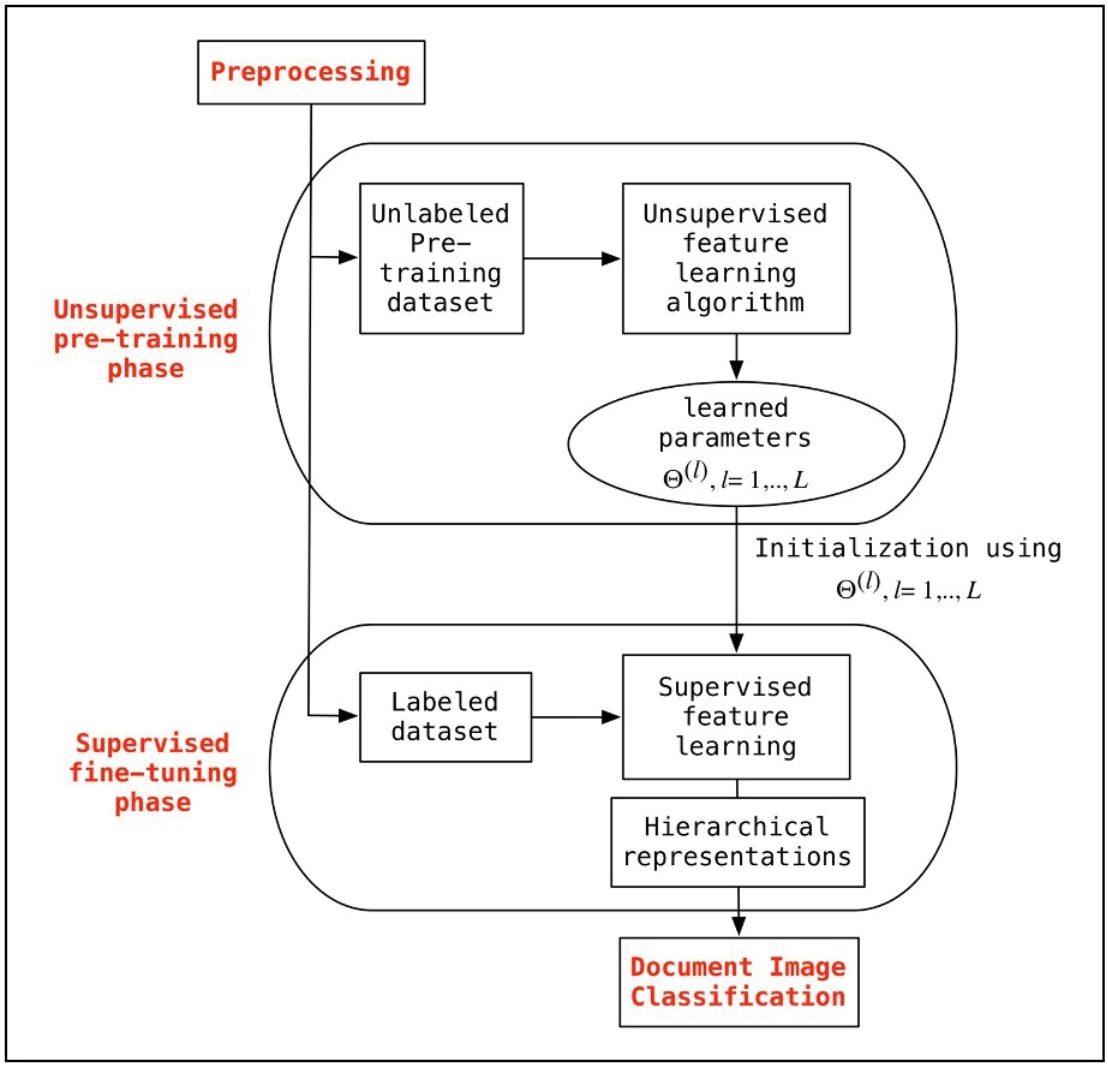


Figure-A I-1 The proposed approach pipeline

2.1 Pre-processing

Considering our observations and hypothesis, each document image is represented by a concatenated image of its two top text-lines and two bottom text-lines, Fig. I-2. In order to obtain these text-lines, a projection-based text-line segmentation method is used (Dos Santos

et al., 2009). Afterward, each concatenated image is negated, normalized then resized to 45x500 for a faster performance.

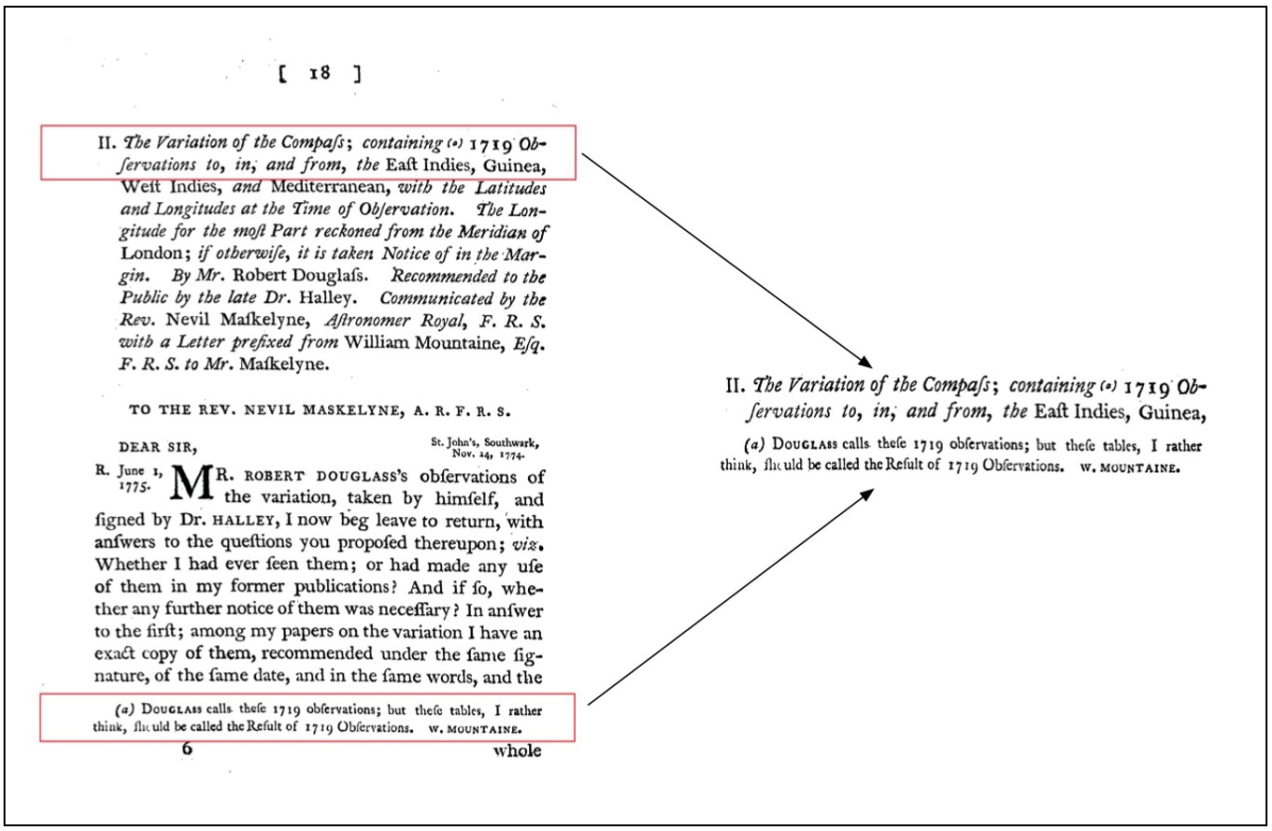


Figure-A I-2 An example from ECCO dataset for a concatenated image of two top text-lines and two bottom text-lines

2.2 Unsupervised pre-training phase

In this phase, an unlabeled pre-training dataset that contains a large amount of document images is utilized as an input to our unsupervised feature learning algorithm, which is based on DBN architecture. DBN is a generative model in which the dependencies between the nodes in one layer is being statistically encoded in the layer above it by using Restricted Boltzmann machines (RBMs). To train our model, a layer-wise greedy learning algorithm is exploited, where the inputs of a higher layer are the calculated activations of the layer below it. Specifically, an RBM is being trained once per time, where the obtained parameters θ^l are being frozen once the

training process is finished; then, another RBM layer is stacked into that network, and a new training process starts on that level. This process is repeated until the last layer is trained.

For each RBM layer, the following energy function $E(v, h)$ is defined to express the negative log likelihood (cost function) of this layer (Lee *et al.*, 2007)[9]:

$$-\log P(v,h) = E(v,h) = \frac{1}{2}\sum_i v_i^2 - \sum_{i,j} v_i w_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i \qquad \text{(A I-1)}$$

where c_i is the bias of the visible node v_i, b_j is the bias of the hidden node h_j and w_{ij} is the weight between v_i and h_j. After the training process, a set of parameters $\theta^l = (w_{ij}, b_j, c_i)$ is being learned.

Training a DBN with L layers results in a set of L learned parameters $\theta^l, l = 1, \ldots, L$, which shall contain implicitly some information about the characteristics of used document images.

2.3 Supervised fine-tuning training phase

A labeled fine-tuning dataset is utilized as an input to a supervised feature learning algorithm. This algorithm can be perceived as a simple Multi-Layer Perceptron (MLP) with the same architecture as the utilized DBN and initialized using the set of learned parameters θ^l obtained at the pre-training phase. In particular, after training our DBN at the previous stage, the parameters of each layer θ^l are used in the initialization process of the same corresponding parameters at our neural network in the current phase. A fine-tune process to the previously learned parameters θ^l is conducted and results in learning high-level hierarchical representations of each document image.

2.4 Classification

A logistic regression classifier is added to our fine-tuning network, as an output layer, in order to classify the used document images into two classes. The fine-tuning dataset is utilized

[9] considering having real values at the visible nodes (input document image).

in the classification training and testing processes using the document images' high-level representations obtained from the fine-tuning stage.

3. Experimental Results and Discussion

Many experiments have been conducted in order to assess our proposed approach. Besides using f-measure as an evaluation metric, the following two steps are utilized. First, a cross-validation technique with 10-folds has been used to evaluate the final classification performance. Additionally, in order to investigate the effect of images' layout complexity, we re-conducted the experiments using a relaxed version of our dataset. In fact, the obtained results show that our approach is notably effective in classifying images based on the presence of footnotes in them even with images with complex layouts.

3.1 Datasets

The images used in our experiments are part of the ECCO dataset used for "The Visibility of Knowledge"[10] project. We utilize two subgroups of this dataset within our proposed approach; an unlabeled dataset that is utilized at the pre-training phase ; in addition, a labeled dataset is exploited at the fine-tuning phase (i.e., this includes the processes of training, validation and testing).

3.1.1 Pre-training dataset

An unlabeled dataset that contains 6895 document images is utilized to learn features in an unsupervised-manner in the pre-training stage of our proposed approach.

3.1.2 Fine-tuning dataset

For fine-tuning and training our proposed approach, a labeled dataset that contains the ground-truth of document images classes is utilized. This dataset includes 4322 labeled samples of

[10] https://txtlab.org/2016/09/the-visibility-of-knowledge/

ECCO dataset (2138 images contain footnotes and 2184 images without footnotes). In order to study the effect of the images' layout complexity on the results, about 1000 images with complex structures have been removed. This has led to a relaxed version of the dataset with 2894 images. The relaxed dataset only contains images with one column and does not contain figures, tables or formulas. Fig. I-3 shows some examples of images with complex layouts.

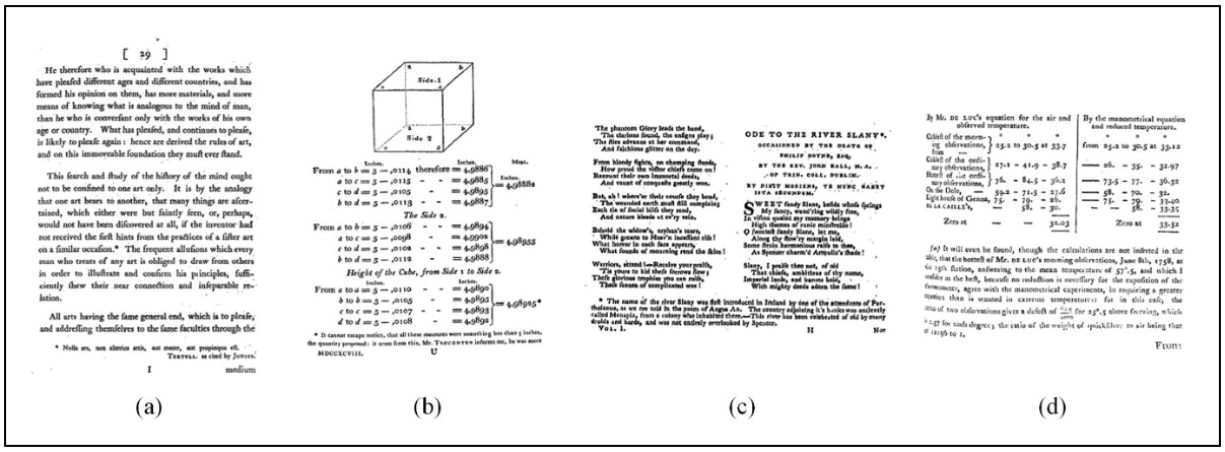


Figure-A I-3 Examples of document images with complex layouts: (a) a simple page (b) a page with formulas and figures (c) a page with two columns (d) a page with tables

3.2 Experimental Setup

In implementing our experiment, we used Python and Theano (Bergstra, Breuleux, Bastien, Lamblin, Pascanu, Desjardins, Turian, Warde-Farley & Bengio, 2010). Our DBN network composed of 2 layers, each layer contains 1000 hidden units. Specifically, the final architecture can be described as 45x500 - 1000 - 1000 - 2. In this case, 2 represents the number of classes (i.e., 0: an image with no footnote and 1: an image with a footnote). The utilized learning rates for pre-training and fine-tuning are 0.01 and 0.05, respectively. In addition, a batch size with a value 10 is set while considering 20 epochs for pre-training and 400 epochs for fine-tuning.

Furthermore, we used a 10-fold cross-validation setup to conduct our experiments. In each cycle, 8 folds are assigned to training phase, 1 fold is used at the validation phase to tune the

model's hyper-parameters and the final fold is utilized as a test set to calculate the generalization performance of our proposed model when applied to unseen test images.

3.3 Results

As demonstrated in Table I-1, our approach has an overall f-measure of 81.37% using the original dataset; a value that has been increased by 4.46% after relaxing the problem. Investigating more, we can observe a big difference between the precision and the recall values using the original dataset; while, on the other hand, the difference between these values is so slight when it comes to the relaxed set. This clearly indicates that the original set contains images with footnotes that are hard to be detected; therefore, relaxing the problem and filtering out the dataset of its complex images has noticeably contributed in improving the value of precision (around 7%). These observations implies the critical role of images with complex structures in affecting the classification overall performance, and raise the need to tackle them.

Table-A I-1 Experimental results using both the fine-tuning original dataset and its relaxed version. Values are in percent

	Precision	**Recall**	**F-measure**
Orginal Set	78.75	84.37	81.37
Relaxed Set	85.63	86.16	85.83

4. Conclusion and Future Work

We have proposed a document image classification framework that is significantly suitable for classification problems associated with a limited availability of labeled data. The proposed approach aims to take advantage of the largely available unlabeled data through incorporating a DBN-based unsupervised feature learning procedure. Our cross-validation-based experimental results demonstrated empirically that our approach can attain an efficient generalization performance on classifying document images based on the availability of footnotes in them. Although this framework is capable of acquiring many tangled features, it finds challenges in dealing with document images with complex structures. The upcoming step towards a more

efficient classification model that can tackle these challenges is to exploit more pre-training data and investigate the criticality of the unlabeled data in reinforcing the overall classification performance of our approach.

Acknowledgments.

This publication was made possible by a grant from SSHRC Canada for ”The Visibility of Knowledge” project. The statements made herein are solely the responsibility of the authors.

BIBLIOGRAPHY

Abuelwafa, S., Mhiri, M., Hedjam, R., Zhalehpour, S., Piper, A., Wellmon, C. & Cheriet, M. (2017). Feature learning for footnote-based document image classification. *International Conference Image Analysis and Recognition*, pp. 643–650.

Afzal, M. Z., Capobianco, S., Malik, M. I., Marinai, S., Breuel, T. M., Dengel, A. & Liwicki, M. (2015). DeepDocClassifier: Document classification with deep convolutional neural network. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1111–1115.

Afzal, M. Z., Kölsch, A., Ahmed, S. & Liwicki, M. (2017). Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification. *arXiv preprint arXiv:1704.03557*.

Alberti, M., Bouillon, M., Ingold, R. & Liwicki, M. (2017a). Open evaluation tool for layout analysis of document images. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 4, 43–47.

Alberti, M., Seuret, M., Pondenkandath, V., Ingold, R. & Liwicki, M. (2017b). Historical document image segmentation with LDA-initialized deep neural networks. *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, pp. 95–100.

Baird, H. S., Bunke, H. & Yamamoto, K. (2012). *Structured document image analysis*. Springer Science & Business Media.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49.

Banko, M. & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pp. 26–33.

Barakat, B. K. & El-Sana, J. (2018). Binarization free layout analysis for arabic historical documents using fully convolutional networks. *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp. 151–155.

Bay, H., Tuytelaars, T. & Van Gool, L. (2006). Surf: Speeded up robust features. *European conference on computer vision*, pp. 404–417.

Belongie, S., Malik, J. & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4), 509–522.

Bengio, Y., Courville, A. & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. & Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. *Proceedings of the Python for scientific computing conference (SciPy)*, 4(3), 1–7.

Bojanowski, P. & Joulin, A. (2017). Unsupervised learning by predicting noise. *arXiv preprint arXiv:1704.05310*.

Bourlard, H. & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4), 291–294.

Bukhari, S. S. & Dengel, A. (2015). Visual appearance based document classification methods: Performance evaluation and benchmarking. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 981–985.

Cantrell, C. D. (2000). *Modern mathematical methods for physicists and engineers*. Cambridge University Press.

Chen, K., Wei, H., Hennebert, J., Ingold, R. & Liwicki, M. (2014). Page segmentation for historical handwritten document images using color and texture features. *2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 488–493.

Chen, K., Seuret, M., Liwicki, M., Hennebert, J. & Ingold, R. (2015). Page segmentation of historical document images with convolutional autoencoders. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1011–1015.

Chen, K., Liu, C.-L., Seuret, M., Liwicki, M., Hennebert, J. & Ingold, R. (2016a). Page segmentation for historical document images based on superpixel classification with unsupervised feature learning. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp. 299–304.

Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Liu, C.-L. & Ingold, R. (2016b). Page segmentation for historical handwritten document images using conditional random fields. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 90–95.

Chen, K., Seuret, M., Hennebert, J. & Ingold, R. (2017). Convolutional neural networks for page segmentation of historical document images. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, 965–970.

Chen, N. & Blostein, D. (2007). A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1), 1–16.

Chen, S., He, Y., Sun, J. & Naoi, S. (2012). Structured document classification by matching local salient features. *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 653–656.

Cheriet, M., Moghaddam, R. F., Arabnejad, E. & Zhong, G. (2013). Manifold learning for the shape-based recognition of historical Arabic documents. In *Handbook of Statistics* (vol. 31, pp. 471–491). Elsevier.

Coates, A. & Ng, A. Y. (2012). Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade* (pp. 561–580). Springer.

Csurka, G., Dance, C., Fan, L., Willamowski, J. & Bray, C. (2004). Visual categorization with bags of keypoints. *Workshop on statistical learning in computer vision, ECCV*, 1(1-22), 1–2.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, 886–893.

Das, A., Roy, S. & Bhattacharya, U. (2018). Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks. *arXiv preprint arXiv:1801.09321*.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Dengel, A. & Dubiel, F. (1995). Clustering and classification of document structure-a machine learning approach. *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 2, 587–591.

Dhillon, I. S. & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2), 143–175.

Dimmick, D., Garris, M. & Wilson, C. (1991). NIST structured forms reference set of binary images (sfrs). *NIST Special Database*, 2.

Doersch, C. & Zisserman, A. (2017). Multi-task self-supervised visual learning. *The IEEE International Conference on Computer Vision (ICCV)*.

Doersch, C., Gupta, A. & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430.

Dos Santos, R. P., Clemente, G. S., Ren, T. I. & Cavalcanti, G. D. (2009). Text line segmentation based on morphology and histogram projection. *2009 10th International Conference on Document Analysis and Recognition*, pp. 651–655.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M. & Brox, T. (2016). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9), 1734–1747.

Enendu, S. (2019). *Predicting Semantic Labels of Text Regions in Heterogeneous Document Images*. (Master's thesis, University of Twente).

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P. & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb), 625–660.

Gidaris, S., Singh, P. & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Glorot, X., Bordes, A. & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323.

Golub, G. H. & Van Loan, C. F. (2012). *Matrix computations*. JHU press.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016). *Deep learning*. MIT press Cambridge.

Grafton, A. (1997). *The footnote: A curious history*. Harvard University Press.

Harley, A. W., Ufkes, A. & Derpanis, K. G. (2015a). Evaluation of deep convolutional nets for document image classification and retrieval. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 991–995.

Harley, A. W., Ufkes, A. & Derpanis, K. G. (2015b). Evaluation of deep convolutional nets for document image classification and retrieval. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 991–995.

He, D., Cohen, S., Price, B., Kifer, D. & Giles, C. L. (2017). Multi-scale multi-task fcn for semantic page segmentation and table detection. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, 254–261.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hinton, G. E. & Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. *Advances in neural information processing systems*, 6, 3–10.

Hinton, G. E., Sejnowski, T. J. et al. (1986). Learning and relearning in Boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317), 2.

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527–1554.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A. & Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193–218.

Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning*, pp. 448–456.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651–666.

Ji, X., Henriques, J. F. & Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874.

Jobin, K. & Jawahar, C. (2017). Document image segmentation using deep features. *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pp. 372–382.

Kang, L., Kumar, J., Ye, P., Li, Y. & Doermann, D. (2014). Convolutional neural networks for document image classification. *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 3168–3172.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kölsch, A., Afzal, M. Z., Ebbecke, M. & Liwicki, M. (2017). Real-Time Document Image Classification using Deep CNN and Extreme Learning Machines. *arXiv preprint arXiv:1711.05862*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pp. 1097–1105.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), 83–97.

Kumar, J. (2013). *Efficient Machine Learning Methods for Document Image Analysis*. (Ph.D. thesis, Fac. Graduate School, Univ. Maryland, College Park, MD, USA).

Kumar, J. & Doermann, D. (2013). Unsupervised classification of structurally similar document images. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pp. 1225–1229.

Kumar, J., Ye, P. & Doermann, D. (2014). Structural similarity for document image classification and retrieval. *Pattern Recognition Letters*, 43, 119–126.

Lam, E. Y. (2004). Analysis of the DCT coefficient distributions for document coding. *IEEE Signal Processing Letters*, 11(2), 97–100.

Larsson, G., Maire, M. & Shakhnarovich, G. (2016). Learning representations for automatic colorization. *European conference on computer vision*, pp. 577–593.

Lazebnik, S., Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2, 2169–2178.

Le Roux, N. & Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural computation*, 20(6), 1631–1649.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Lee, D. D. & Seung, H. S. (2000). Algorithms for Non-Negative Matrix Factorization. *Proceedings of the 13th International Conference on Neural Information Processing Systems*, (NIPS'00), 535–541.

Lee, H., Ekanadham, C. & Ng, A. (2007). Sparse deep belief net model for visual area V2. *Advances in neural information processing systems*, 20, 873–880.

Lee, H., Grosse, R., Ranganath, R. & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616.

Lee, J., Hayashi, H., Ohyama, W. & Uchida, S. Page Segmentation using a Convolutional Neural Network with Trainable Co-occurrence Features.

Li, X.-H., Yin, F., Xue, T., Liu, L., Ogier, J.-M. & Liu, C.-L. (2019). Instance Aware Document Image Segmentation using Label Pyramid Networks and Deep Watershed Transformation. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 514–519.

Li, Y., Zou, Y. & Ma, J. (2018). DeepLayout: A Semantic Segmentation Approach to Page Layout Analysis. *International Conference on Intelligent Computing*, pp. 266–277.

Likforman-Sulem, L., Zahour, A. & Taconet, B. (2007). Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4), 123–138.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, pp. 740–755.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2), 129–137.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Lopes, N. & Ribeiro, B. (2015). *Machine Learning for Adaptive Many-Core Machines-A Practical Approach*. Springer.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision*, 2, 1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Mairal, J., Ponce, J., Sapiro, G., Zisserman, A. & Bach, F. R. (2009). Supervised dictionary learning. *Advances in neural information processing systems*, pp. 1033–1040.

Makhzani, A. & Frey, B. (2013). K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Manning, C., Raghavan, P. & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100–103.

Mettam, G. R. & Adams, L. B. (1999). How to prepare an electronic version of your article. In Jones, B. S. & Smith, R. Z. (Eds.), *Introduction to the Electronic Age* (pp. 281-304). New York, NY: E-Publishing Inc.

Mhiri, M., Abuelwafa, S., Desrosiers, C. & Cheriet, M. (2017). Footnote-based document image classification using 1D convolutional neural networks and histograms. *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–5.

Nafchi, H. Z., Moghaddam, R. F. & Cheriet, M. (2014). Phase-based binarization of ancient document images: Model and applications. *IEEE transactions on image processing*, 23(7), 2916–2930.

Nagy, G. (2000). Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 38–62.

Nagy, G., Seth, S. & Viswanathan, M. (1992). A prototype document image analysis system for technical journals. *Computer*, 25(7), 10–22.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *ICML*.

Noce, L., Gallo, I., Zamberletti, A. & Calefati, A. (2016). Embedded Textual Content for Document Image Classification with Convolutional Neural Networks. *Proceedings of the 2016 ACM Symposium on Document Engineering*, pp. 165–173.

Noroozi, M. & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. *European Conference on Computer Vision*, pp. 69–84.

Oliveira, S. A., Seguin, B. & Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 7–12.

Oyedotun, O. K. & Khashman, A. (2016). Document segmentation using textural features summarization and feedforward neural network. *Applied Intelligence*, 45(1), 198–212.

Pasanek, B. & Wellmon, C. (2015). The enlightenment index. *The Eighteenth Century*, 56(3), 359–382.

Pasanek, B. & Wellmon, C. (2018). Enlightenment, Some Assembly Required. *The Eighteenth Centuries: Global Networks of Enlightenment*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine Learning in Python.

Pondenkandath, V., Seuret, M., Ingold, R., Afzal, M. Z. & Liwicki, M. (2017). Exploiting state-of-the-art deep learning methods for document image analysis. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 5, 30–35.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500), 2323–2326.

Roy, S., Das, A. & Bhattacharya, U. (2016). Generalized stacking of layerwise-trained Deep Convolutional Neural Networks for document image classification. *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 1273–1278.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Sarkar, M., Aggarwal, M., Jain, A., Gupta, H. & Krishnamurthy, B. (2019). Document Structure Extraction for Forms using Very High Resolution Semantic Segmentation. *arXiv preprint arXiv:1911.12170*.

Saund, E. (2011). A graph lattice approach to maintaining dense collections of subgraphs as image features. *2011 International Conference on Document Analysis and Recognition*, pp. 1069–1074.

Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M. & Ingold, R. (2016). Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 471–476.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Strunk Jr., W. & White, E. B. (1979). *The Elements of Style* (ed. 3rd). New York, NY: Macmillan.

Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M. & Ingold, R. (2019). A Comprehensive Study of ImageNet Pre-Training for Historical Document Image Analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 720-725.

Tang, B., He, H., Baggenstoss, P. M. & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1602–1606.

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A. et al. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.

Tensmeyer, C. & Martinez, T. (2019). CONFIRM–Clustering of noisy form images using robust matching. *Pattern Recognition*, 87, 1–16.

Titterington, D. M. (2005). *Statistical analysis of finite mixture distributions*. (Ph.D. thesis, Institute of Philosophy).

van der Geer, J., Hanraads, J. A. J. & Lupton, R. A. (2000). The art of writing a scientific article. *J. Sci. Commun.*, 163, 51-59.

Vil'kin, A., Safonov, I. & Egorova, M. (2013). Algorithm for segmentation of documents based on texture features. *Pattern recognition and image analysis*, 23(1), 153–159.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec), 3371–3408.

Wang, J. & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*.

Wei, H., Seuret, M., Liwicki, M., Ingold, R. & Fu, P. (2017). Selecting fine-tuned features for layout analysis of historical documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, 281–286.

Wick, C. & Puppe, F. (2018). Fully convolutional neural networks for page segmentation of historical document images. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 287–292.

Xu, Y., He, W., Yin, F. & Liu, C.-L. (2017). Page segmentation for historical handwritten documents using fully convolutional networks. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, 541–546.

Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D. & Lee Giles, C. (2017). Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5315–5324.

Yoo, J. & Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5), 559–570.

Yu, F. & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhalehpour, S., Piper, A., Wellmon, C. & Cheriet, M. (2017). Footnote-based document image classification. *International Conference Image Analysis and Recognition*, pp. 634–642.

Zhang, W., Du, Y., Yoshida, T. & Yang, Y. (2019). DeepRec: A deep neural network approach to recommendation with item embedding and weighted loss function. *Information Sciences*, 470, 121–140.

Zhong, Y., Karu, K. & Jain, A. K. (1995). Locating text in complex color images. *Pattern recognition*, 28(10), 1523–1535.