Towards Zero-Shot Document Image Classification

Lucas Macedo, João Paulo Costa, João Pedro Felix de Almeida, Pedro Freitas, and Li Weigang
Department of Computer Science, *University of Brasília*, Brasília, Brazil

***Abstract*—Classification is a fundamental tool to automate the process of categorizing documents in many real-world applications, such as information management, financial document processing, healthcare records management, news categorization, fraud detection, regulatory compliance, and many others. Because of this broad spectrum of applications, document classification is of paramount importance for various companies. However, documents often change in terms of format and their visual patterns, which may impair a simple classification model. Moreover, model continuance and retraining often demands important efforts, consuming computational resources and demanding new data. Therefore, techniques capable of classifying documents by simply observing new data, without necessarily requiring retraining the classifier, are of immense importance for a wide variety of applications. In this context, Zero-Shot Learning (ZSL) is especially suitable for document classification because it handles diverse and ever-changing document content. In this work, we tackle the gap involving Zero-Shot Document Image Classification (ZS-DIC), where we classify documents that have not been seen by the model during training. To achieve this, we built Layout-Aware Complex Document Information Processing (LA-CDIP), a dataset tailored for this problem. LA-CDIP prioritizes structural consistency, allowing models to classify documents correctly under a ZSL scenario. To benchmark this dataset, we developed a series of Siamese Neural Networks (NNs) based on a variety of computer vision neural architectures, such as ResNet, EfficientNet, ViT and others. As a result, the proposed ZSL-based method achieves Equal Error Rates (EERs) under 5%. The code of the proposed method is available at https://github.com/ABMHub/doc-zsl.**

I. Introduction

Document understanding, vital for businesses, automates information extraction and interpretation from documents [1]. The prominence of this field within the business sector stems from the universal necessity for companies to manage documents. Therefore, an automated pipeline for document processing is widely adopted and often essential. For instance, companies that process personal documents, such as identification documents, frequently encounter low-quality images. Nevertheless, an automated pipeline remains more efficient in both cost and time than a manual verification pipeline. Consequently, most companies find machine learning a suitable technology for automation, as deterministic methods may struggle to account for the high variability of document quality.

In this context, the prevalent approach to address this problem involves traditional classification methods, often utilizing Deep Learning (DL). The problem has been widely studied in the field of document understanding as a simple classification task, mainly supported by the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset [2]. This dataset comprises 400,000 documents divided into 16 classes based on their purpose, such as letters and emails. RVL-CDIP has been a subject of extensive research for traditional classification, with some works achieving over 97% accuracy [3], [4] in specific scenarios. However, existing documents frequently undergo changes in form and structure, and new document types are continuously generated. In such dynamic environments, new data must be acquired to retrain the classification model.

To overcome the limitations imposed by the need for labeled training data, ZSL emerges as the most appropriate paradigm. This approach defines a task where the classes in the test set are disjoint from those in the training set, thus forcing the model to generalize effectively to novel categories [5], [6]. Despite being a considerably under-researched area within document analysis, ZS-DIC provides valuable contributions.

Although RVL-CDIP is the most widely used dataset for document classification, other datasets are available. DocVQA [7] focuses on Visual Question Answering (VQA), DocLayNet [8] is specialized in layout segmentation, and CORD [9], SROIE [10], FUNSD [11], and XFUND [12] were created for information extraction. While these datasets advance their respective fields, their label characteristics prevent their effective use in ZS-DIC.

To address these challenges, we introduce LA-CDIP, a dataset designed for the ZS-DIC challenge. LA-CDIP is a document-image dataset that re-purposes the images from RVL-CDIP under a new classification scheme. Unlike RVL-CDIP, which classifies documents by their purpose (e.g., emails, letters), our dataset arranges them by their visual structure, grouping together documents that share a similar layout. This reorganization enables ZS-DIC by ensuring that each class exhibits high consistency in both visual and textual patterns.

We propose a Visual Document Matching (VDM) framework for this dataset. This framework leverages Contrastive Learning (CL) [13] to generate a similarity score between two documents. A predefined similarity threshold is used to determine if documents belong to the same class. In a practical application, by providing a reference for each class, the framework can be deployed in two modes: as a binary classifier for verification (comparing each incoming document with a single class reference) or as a multiclass classifier for identification (comparing each incoming document with all available references to select the most similar one).

The remainder of this paper is structured as follows. Section II details our LA-CDIP dataset. Section III describes the proposed Siamese network modeling. Sections IV and Sec-

979-8-3315-8951-6/25/$31.00 ©2025 IEEE