



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Layout-Aware Zero-Shot Learning for Visual Document Matching

Lucas de Almeida Bandeira Macedo

Dissertação apresentada como requisito parcial para
qualificação do Mestrado em Informática

Orientador
Prof. Dr. Pedro Garcia Freitas

Coorientador
Prof. Dr. Bruno Luiggi Macchiavello Espinoza

Brasília
2025



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Layout-Aware Zero-Shot Learning for Visual Document Matching

Lucas de Almeida Bandeira Macedo

Dissertação apresentada como requisito parcial para
qualificação do Mestrado em Informática

Prof. Dr. Pedro Garcia Freitas (Orientador)
CiC/UnB

Prof. Dr. PLACEHOLDER Prof. Dr. PLACEHOLDER
CiC/UnB CiC/UnB

Prof.a Dr.a Cláudia Nalon
Coordenadora do Programa de Pós-graduação em Informática

Brasília, 12 de outubro de 2025

Resumo

Garantir conformidade documental requer uma identificação acurada dos documentos, que também serve o propósito de manter o dado consistente ao decorrer das esteiras de verificação. Grande parte dos estudos lidam com a identificação como uma tarefa de classificação documental, ou como uma tarefa de segmentação. Entretanto, documentos industriais estão sempre mudando sua forma, e os modelos que os classificam precisam de constantes atualizações. Nesses casos, avaliar se determinado documento está de acordo com o histórico de documentos aceitos é uma abordagem mais apropriada. Esta tese adentra no problema de comparar a aparência de dois (ou mais) documentos para determinar se eles dividem ou não a mesma disposição de informações. Portanto, esse problema é atacado com o paradigma *Zero-Shot Learning* (ZSL), que é uma abordagem poderosa para cenários onde as classes encontradas na inferência não coincidem com as classes usadas no treino. Para dar suporte ao estudo, o *Layout-Aware Complex Document Information Processing* (LA-CDIP) é introduzido, um dataset contendo 4,993 documentos, distribuídos por 144 classes, reorganizadas a partir da base de dados *Ryerson Vision Lab Complex Document Information Processing* (RVL-CDIP), realizando uma separação prioritariamente sintática, ao invés de semântica. Essa abordagem é testada usando redes siamesas e *Contrastive Learning* através de muitas arquiteturas neurais conhecidas, incluindo ResNet, EfficientNet e *Vision Transformer* (ViT). Em cenários ZSL, o método proposto atinge um *Equal Error Rate* (EER) abaixo de 5% na verificação com validação cruzada. Além disso, a abordagem *Visual Document Matching* (VDM) performa com maior precisão que *Large Language Models* (LLMs) de código aberto e rivaliza contra o modelo GPT-4o, da OpenAI, demonstrando a superioridade de uma técnica especialista sobre modelos multimodais generalistas. Essas descobertas mostram que a abordagem proposta mantém alta acurácia enquanto usa significativamente menos parâmetros que LLMs, demonstrando um uso mais prático para aplicações de conformidade documental na indústria.

Palavras-chave: Redes Neurais Artificiais, Leitura Labial Automática, Trabalho de Conclusão de Curso

Abstract

Ensuring document compliance requires accurate document identification, which plays a crucial role in maintaining consistency throughout document analysis pipelines. Several studies approach layout identification as a document image classification or segmentation task. However, due to the ever-changing nature of industry documents, a traditional classification with entropy learning is often insufficient, as models require frequent re-training. In these cases, determining whether two or more documents share the same visual layout is a more suitable approach. This paper addresses the problem of matching the visual appearance of two (or more) documents to determine whether they share the same layout. To achieve this, a Zero-Shot Learning (ZSL) approach is adopted, which is a powerful technique for scenarios where training classes do not align with those encountered during inference. Layout-Aware Complex Document Information Processing (LA-CDIP) is introduced to support the study, a dataset comprising 4,993 documents across 144 classes, which is reorganized from the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) database to emphasize visual structure over semantic content. This approach is benchmarked using a siamese network and contrastive learning framework across multiple backbone architectures, including ResNet, EfficientNet, and Vision Transformer (ViT). In zero-shot scenarios, the proposed method achieves an Equal Error Rate (EER) below 5% in 1-vs-1 verification with cross-validation. Furthermore, the Visual Document Matching (VDM) approach outperforms lighter Large Language Models (LLMs) and rivals GPT-4o, highlighting the superiority of specialized techniques over general-purpose multimodal models. These findings show that the proposed approach maintains high accuracy while using significantly fewer parameters than large multimodal models, making it more practical for real-world document compliance applications.

Keywords: Artificial Neural Networks, Automatic Lipreading, Thesis

Contents

1	Introduction	1
1.1	Contextualization	1
1.2	Problem Description	2
1.3	About this work	3
2	Theoretical Foundation	5
2.1	Traditional Classification	5
2.2	Metric Learning	6
2.3	Large Language Models	7
3	Related Work	8
3.1	Document Understanding	8
3.2	Document Understanding Databases	8
3.3	Document Processing	9
3.4	LLMs on Document Understanding	10
4	Methodology	12
4.1	Dataset Construction	12
4.2	Leveraging LLM for Benchmarking Visual Document Matching	14
4.3	Modeling	14
5	Results	17
5.1	LA-CDIP Dataset	17
5.2	Evaluation Criteria	18
5.3	LLMs	19
5.4	Partial Results	19
5.4.1	Visual Models	20
5.4.2	Large Language Models	22
5.4.3	Comparison	22

6 Conclusion	24
6.1 Timeline	24
Bibliography	26

List of Figures

2.1	Simplified architecture of a siamese network	7
4.1	Overview of the proposed LA-ZSL methodology for VDM.	12
4.2	Examples of two distinct classes under the proposed LA-CDIP dataset. Those classes originally belonged to the same class under RVL-CDIP organization.	13
4.3	Two sets of comparisons by a LLM model. The first example shows the same layout, and the second, different layouts. The scores range from 0 to 100.	15
5.1	Histogram showing the top 20 classes by their frequency. The other 124 classes are grouped in the “others” column. Note that the plot is in log scale.	17
5.2	TSNE visualization of the Test ZSL scenario. These models have been trained on the same cross-validation fold.	21
5.3	Performance vs Parameters (ZSL). The lines represent GPT-4o and GPT-4o Mini error rates, as their parameter count were not publicly disclosed. .	23

List of Tables

5.1 Comparative performance between different visual backbones and Large Language Models. Following the columns: the architecture name, the architecture edition, if exists, cross-validation over the ZSL scenario, cross-validation over the GZSL scenario, test performance on the ZSL scenario, and test performance over the GZSL scenario. Every value is a mean EER (%) value over the CV folds.	20
6.1 The timeline of the master's research. The "x" represents the current moment in the timeline.	25

Chapter 1

Introduction

1.1 Contextualization

Documents are at the core of corporate society, and they often hold immeasurable value, representing contracts, employment relationships, loans, and identities. Recently, as technology becomes more accessible, documents have become increasingly more digital, some never being printed on paper. These documents often have all important information already separated in the file's metadata and require no extra intelligence to classify or extract the information. However, many organizations still receive documents in physical form, for example, when dealing directly with individual customers, which makes full digitization impractical in many cases.

In such a scenario, the company must have a document understanding pipeline, where the document is approved regarding its form (verifying if the client sent the right type of document) and the information is extracted to support business decision-making. This pipeline has historically been executed by humans. Between the start of the digital revolution and the popularization of Optical Character Recognition (OCR) engines [1, 2], the typist profession was highly prevalent, with workers responsible for manually entering information from physical documents into digital systems. Moreover, the classification and information extraction were done by back office analysts, who spent their workforce reading and verifying if the document met the company's compliance rules. Today, with current technology, especially with the popularization of neural networks, this manual document pipeline has become unthinkable in big companies, as it is inefficient in both time and cost. This work focuses on automatic document classification.

DEVEMOS CRIAR UM DIAGRAMA AQUI

Document Classification is defined by the categorization of documents into pre-defined classes [3]. Document Image Classification (DIC) is a more specialized task in which textual information is not directly available, and the input often consists of photos

or scans of physical documents. Therefore, if the text of the document is ever required, this text should be extracted with OCR algorithms. Artificial Intelligence (AI) and Deep Learning (DL) techniques can automate document classification by leveraging traditional classification networks. The model solution may analyze just the image, which can cut the computational cost of the overall solution by not using an OCR algorithm. Sometimes, however, the image may not give enough information to make a proper decision, so models may also be modeled to classify based on text information, or even mix both in a multimodal solution. This problem is already known and well studied. Bakkali et al. [4] achieved 97.70% accuracy on the RVL-CDIP dataset [5], a popular document classification dataset.

The problem arises when previously accepted documents change in form or when entirely new document classes, not previously considered, must be handled and categorized. In such cases, models are typically retrained, as traditional classification networks may struggle to generalize to new document layouts or entirely new classes. Traditional DL classification methods require a predefined set of labels and cannot output a label outside the original training set. Therefore, introducing new labels to the model means retraining the model, changing the output layer into a wider one with more options. Very often, this also means weeks or months of data engineering, labeling, and model training. The alternative is to use Zero-Shot Learning (ZSL) techniques, which allow models to generalize to classes that were not seen during training [6].

DIAGRAMA ZSL VS TRADICIONAL AQUI

1.2 Problem Description

This work focuses on the ZSL on DIC, where the model must correctly classify a document image class that was not in the training set and therefore previously unseen by the model. ZSL techniques enable a single model to address multiple classification problems, as they are not limited by the classes used in training. Most ZSL systems work by semantically mapping the elements into a feature space, where elements from the same class are all clustered together, and different classes are placed far apart. To build a reliable ZSL system, it is necessary to identify a set of features that serve as robust representations, ensuring that documents share features if, and only if, they belong to the same class. With DL, this could also demand a dataset with a wide variety of classes, enough so that the model learns what makes two documents similar or different. These conditions allow ZSL models to group together documents, even if they were completely unseen during training. Nonetheless, not every classification dataset can be easily used in an

effective ZSL solution. The lack of a specialized dataset is one of the biggest challenges in document-image ZSL classification.

In consequence, another main challenge in image-based document ZSL classification is the lack of a state-of-the-art methodology. Due to the challenge in achieving ZSL classification with the currently available datasets, most works tackle the problem by their own methodology and are often incomparable to the ones that came before [7], unlike traditional classification and information extraction that have a widely accepted framework and baseline approaches. Many works also use pretrained Large Language Model (LLM) [8], leveraging their previous knowledge as a tool to achieve zero-shot learning with a fine-tuning framework, which limits the cost-efficiency achievable with this strategy.

But there are a few known obstacles in zero-shot document classification. First, existing datasets do not enforce disjoint class splits between training and testing, nor do they provide enough information to train an efficient ZSL model from scratch. Second, there is an absence of a well-established, state-of-the-art classification framework capable of operating under zero-shot constraints.

1.3 About this work

Following this line, this work proposes a new framework for tackling ZSL document classification: Visual Document Matching (VDM). By leveraging metric learning techniques, documents with similar patterns can be identified and grouped into the same class/group, even if their class was not seen before during training. Through this method, VDM shows itself as a zero-shot alternative for the identification of documents that are visually equivalent, as a matching problem. In other words, VDM can be handled as a binary classification problem.

This work makes several key contributions, i.e.:

1. Introduces a novel visual-only document image dataset specifically designed for the task of document ZSL classification, enabling the development and evaluation of models in this domain.
2. Proposes a VDM approach, based on image similarity, leveraging zero-shot learning techniques to generalize across unseen document layouts. This strategy enables generalization to previously unseen classes without additional intervention on the model itself.
3. Delivers a systematic evaluation of well established backbones in the context of zero-shot document layout matching, offering valuable benchmarks for future research.

This paper is structured as follows: Chapter 2 lays out the theoretical foundation of this work. Chapter 3 reviews the state of the art in zero-shot learning and visually-rich document understanding, highlighting existing methods and their limitations. Chapter 4 presents the proposed Layout-Aware Zero-Shot Learning (LA-ZSL) for VDM, detailing its architecture, components, and underlying learning paradigm. Chapter 5 describes the experimental setup, including datasets, evaluation metrics, and baseline comparisons, followed by a comprehensive analysis of the results. Finally, Chapter 6 concludes the article by summarizing the contributions, discussing potential applications, and outlining future research directions, and the timeline of this Master’s research.

Chapter 2

Theoretical Foundation

The goal of this chapter is to offer the theoretical basis for ideas that are assumed but not elaborated in Chapter 3 and Chapter 4.

2.1 Traditional Classification

It is mentioned very often in this paper that traditional classification cannot generalize over unseen classes. In the context of this work, it can be specified further: DL models, trained with a cross-entropy learning framework, cannot classify an element into a class that has not been mapped beforehand. This is because the model is trained to minimize the cross-entropy loss function over a labeled dataset, associating each input with a fixed class label. For a single instance, the cross-entropy equation is as follows:

$$CE_{loss} = - \sum_{c=1}^M y_c \log(p_c), \quad (2.1)$$

where p_c denotes the predicted probability of the input element belonging to class c , y_c is a label that is 1 if the element belongs to c , and 0 otherwise, and M is the number of classes. When a model is trained using a class-disjoint split, meaning that no class appears in both training and test sets, a cross-entropy classifier is unable to learn anything about the unseen classes. Since these classes are not represented in the training data, it is not possible to compute a meaningful cross-entropy loss for them. Therefore, to tackle ZSL, another approach is needed.

2.2 Metric Learning

Metric learning is a framework that allows ZSL. In summary, a metric-learning model $f(\cdot)$ is trained such that, over an input element, it yields a feature vector v , instead of a traditional classification. This model $f(\cdot)$ is optimized under a given metric $d(\cdot, \cdot)$, such that two feature vectors that share the same class v_1, v'_1 are closer together than two feature vectors from different classes v_1, v_2 . In other words, the model is optimized so that $d(v_1, v'_1) < d(v_1, v_2)$ and $d(v_1, v'_1) < d(v'_1, v_2)$.

A DL model following a metric learning framework is constructed as a siamese network. They are named siamese because they can be seen as bipartite architectures, with two parallel paths that converge in the end, even though both sides share the same weights. To construct a visual siamese network, an image model, $g(\cdot)$, is required to transform an input image I into a feature representation in an arbitrary n -dimensional space, $g(I)$. The $g(\cdot)$ image model is referred to as the backbone of the siamese network. This framework is illustrated in Figure 2.1 for reference.

Losses that support metric learning, such as contrastive [9] or triplet [10], follow the same strategy: cluster together elements that share a label, and split apart elements that do not. When training a siamese network, the training step should consider at least two distinct elements, and the loss function scores the step over the distance between the elements. During training, the contrastive loss considers only two elements per step, and they can either share the same class or belong to different classes. The triplet loss considers three elements per step: an anchor/reference, a positive example (same label), and a negative example (different label). In this work, the contrastive loss function is used to train the model. More details about the contrastive loss are provided in Chapter 4.3.

The metric learning nature of this method makes it highly adaptable to different applications. Since feature vectors are used, clustering is a very natural application of the method, as are verification and identification. By establishing a distance threshold, it is possible to declare that two elements are similar enough to be considered a positive case, or far enough to be a negative case. Using this strategy, and having a reference for each required class, verification can be performed by comparing a given element with a certain class, effectively employing binary classification, or identification can be performed by comparing the element with every class and taking the most similar, effectively employing multiclass classification.

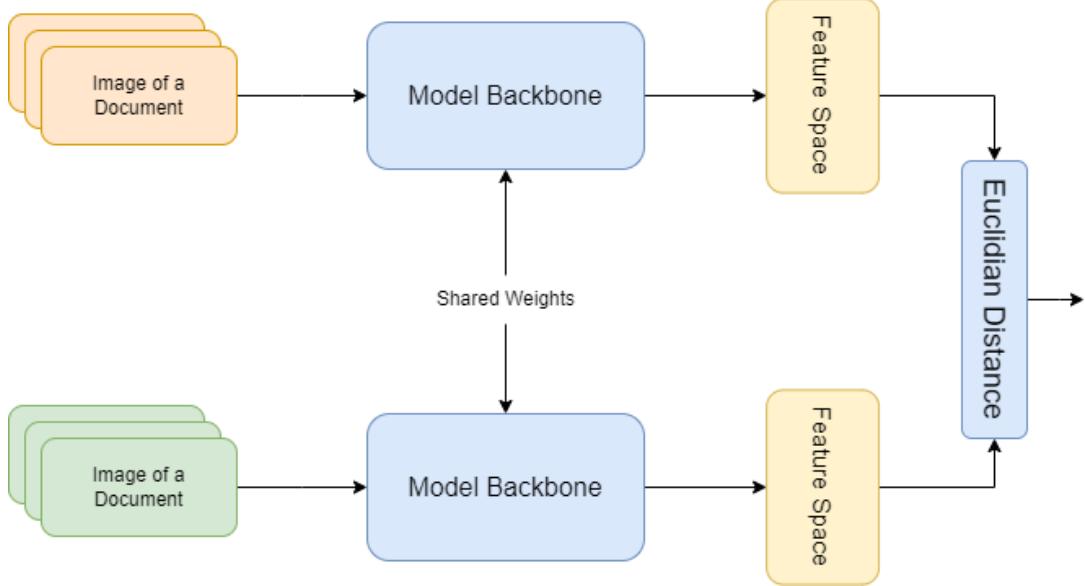


Figure 2.1: Simplified architecture of a siamese network

2.3 Large Language Models

LLMs are DL networks trained on massive corpora of textual data to, traditionally, generate human-like text, although they can be adapted to a number of different tasks. Modern LLMs, such as GPT [11] and LLaMA [12], have demonstrated remarkable performance across a wide variety of natural language processing (NLP) tasks. One of the defining characteristics of LLMs is their ability to generalize to unseen tasks without task-specific fine-tuning, a phenomenon often referred to as in-context learning. Instead of retraining the model, a user can provide a prompt containing instructions or examples, and the model adapts its behavior accordingly. This makes LLMs particularly appealing for applications where labeled data is scarce or non-existent, excelling in zero-shot and few-shot tasks.

Despite their versatility, LLMs come with challenges. Their performance can degrade in highly specialized domains or when exposed to out-of-distribution inputs. Generative LLMs in particular are very susceptible to hallucination [13], where the output answer is structurally correct but has no connection to the input text, or even has fabricated information. In a corporate scenario, integrating a generative LLM into an automation pipeline is particularly challenging. Many systems depend on structured communication, and a generative model may hallucinate on the structure itself, breaking the system even if the answer was semantically correct. Even then, they continue to shape current research and development in machine learning, and their integration into hybrid or multimodal systems is a growing area of interest.

Chapter 3

Related Work

This chapter analyzes the literature about automatic document processing, focusing on the available datasets and document classification methods. The relevance and direction of this work are justified by the limitations of each dataset and classification framework. This chapter also helps to provide the boundaries for the problem.

3.1 Document Understanding

Document Understanding is a broad field encompassing various tasks that aim to extract and interpret information from documents, which may contain text, tables, images, and complex layouts [14]. The increasing digitization of documents has led to significant advancements in document analysis techniques, particularly through deep learning and transformer-based architectures. Among the key challenges in this domain is handling the inherent complexity of scanned documents, which often exhibit noise, structural variability, and diverse formatting styles.

Several fundamental tasks define the scope of document understanding. *Document Layout Analysis* (DLA) involves detecting and categorizing different structural components of a document, such as text blocks, tables, images, and forms, to facilitate higher-level information extraction [15]. Information Extraction (IE) focuses on retrieving relevant information, such as named entities and key-value pairs, from structured and unstructured document sources [14].

3.2 Document Understanding Databases

This subsection reviews existing datasets commonly used for document classification and discusses their limitations in supporting zero-shot scenarios. By analyzing these datasets, the need for a new benchmark that explicitly enforces zero-shot constraints is highlighted.

The IIT-CDIP dataset [16], introduced in 2006, is a large-scale repository used for document classification and information retrieval. It originated from the Tobacco Documents Library, from the University of San Francisco Industry Documents Library collection, and contains millions of scanned documents. The most popular dataset when working with document-image classification is the RVL-CDIP dataset [5], which was introduced in 2015. This dataset is a labeled subset of the IIT-CDIP dataset, and organized 400,000 document images into 16 predefined categories such as letters, forms, and emails. This classification is driven by the document’s purpose and uses, which hinders the performance of a ZSL model trained from scratch, with no real-world knowledge.

More recently, the DocVQA dataset [17] was introduced, leveraging a subset of documents from the same collection. It comprises over 12,000 document images paired with 50,000 question-answer pairs, designed to evaluate models’ abilities in visual question answering on document images. DocVQA has since become a standard benchmark for assessing the performance of LLMs in document understanding tasks, particularly in multi-modal reasoning scenarios. FUNSD dataset [18], introduced in 2019, has been widely used for OCR-based semantic relation extraction from scanned forms. Similarly, SROIE [19], CORD [20] and XFUND [21] target key-value pair extraction in receipts and invoices, focusing on structured entity retrieval, such as store names, monetary values, and transaction dates. While these tasks consider the organization of visual elements within documents, they do not address VDMs as they are designed to extract specific content rather than compare the visual similarity between different documents.

While existing datasets have contributed to document understanding research, they primarily support tasks related to text extraction, classification, and structured information retrieval. These datasets do not provide a framework for evaluating VDMs, as their organization is often driven by textual or semantic content rather than visual arrangement.

3.3 Document Processing

In the domain of document analysis, various studies have addressed challenges related to document layout analysis, classification with limited data, and similarity detection. Understanding these works provides valuable insights into the current landscape and highlights the unique contributions of the research.

Veneri et al. [22] propose a method for Document Layout Analysis using variational autoencoders to detect deviations from a standard document template. Their approach is particularly suited for industrial compliance verification, where identifying visual discrepancies such as stamps, handwritten annotations, and misplaced signatures is crucial.

By learning the distribution of compliant documents, the model detects anomalies as out-of-distribution samples, making it effective for scenarios with highly imbalanced datasets. While this study shares similarities with the work in identifying visual differences across documents, it is focused on anomaly detection within a predefined template rather than assessing layout similarity between different document classes.

Zeghidi et al. [23] present CDP-Sim, a similarity metric learning approach designed to detect counterfeit Copy Detection Patternss (CDPs) using a Siamese neural network. This method effectively distinguishes original from fake CDPs by learning a similarity metric that captures subtle differences between patterns. While their work is specifically applied to counterfeit detection, the concept of learning similarity metrics through metric learning and Siamese networks is broadly applicable to various tasks requiring fine-grained visual differentiation. This principle can also be leveraged in scenarios involving document layout comparisons, where structural relationships between documents must be assessed independently of their textual content.

Sinha et al. [7] introduce CICA, a framework that enhances CLIP’s performance in zero-shot classification by improving textual-visual feature alignment through content-injected contrastive learning. This study emphasize data-efficient approaches to document classification. However, their approach is not suitable for a from-scratch training, and relies on costly pretrained models, limiting the potential to create a zero-shot cost-efficient model.

3.4 LLMs on Document Understanding

LLMs have recently gained significant attention for their capabilities in understanding and generating human-like text. Their application in document understanding as a whole has shown promising results, particularly in tasks that require comprehension of complex layouts and multimodal information.

Scius et. al. [8] investigate the application of LLMs, such as GPT-4 and RoBERTa, for zero-shot prompting and few-shot fine-tuning in document image classification. Their study demonstrates that LLMs can achieve competitive performance with minimal labeled data, challenging the traditional reliance on large annotated datasets.

With the need to process multimodal information, there has been a rapid advancement in models with visual capabilities. Llama 3.2 Vision [24], for example, was introduced as an open source alternative with multimodal reasoning capabilities. Other recent work has analyzed the relationship between performance and efficiency. In 2024, DeepSeek-VL2 [25] exemplified this trend, achieving competitive results with fewer activated parameters compared to models such as InternVL2 and Qwen2-VL.

In 2025, InternVL 2.5 was introduced as an evolution of its predecessor, InternVL 2.0, incorporating enhancements that resulted in performance gains, setting a new standard for open source multimodal models [26]. Moreover, its performance is competitive with leading proprietary models, such as OpenAI’s GPT-4o[27, 28]. In 2025, Qwen2.5-VL was introduced as an evolution of the Qwen2-VL series, surpassing its predecessor. According to its evaluation [29], Qwen2.5-VL excels in OCR-related tasks, chart interpretation, and document understanding. The leadership in state-of-the-art performance for these benchmarks is now primarily held by GPT-4o, InternVL 2.5, and Qwen2.5-VL, the latter two representing the strongest open-source alternatives.

Chapter 4

Methodology

Developing an effective solution for VDMs requires addressing key challenges related to dataset availability, document structure variability, and generalization to unseen layouts. This section provides a detailed description of these components and the experimental procedures used to evaluate their effectiveness. Figure 4.1 illustrates the proposed methodology, outlining the sequential steps from dataset preparation to result analysis.

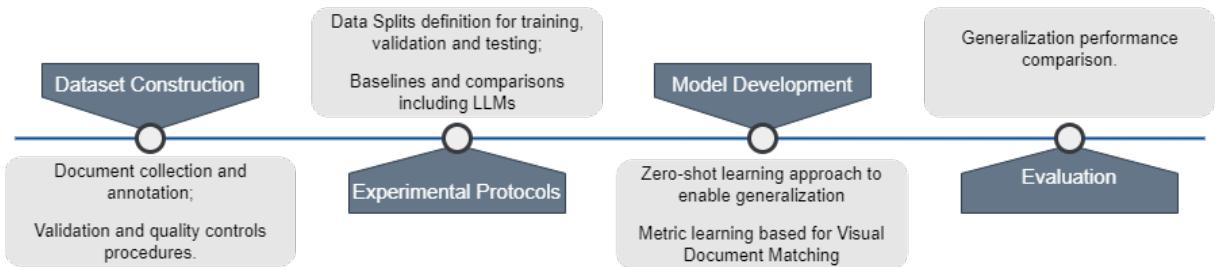


Figure 4.1: Overview of the proposed LA-ZSL methodology for VDM.

4.1 Dataset Construction

The proposed dataset, Layout-Aware Complex Document Information Processing (LA-CDIP), is built as a reorganization of the RVL-CDIP [5] dataset. While RVL-CDIP's classes are divided by their meaning and written content of the documents, the dataset is divided by the visual information and layout. This is illustrated in Figure 4.2, where both groups of documents are classified as the same class under RVL-CDIP, both being forms, but are re-arranged to be different classes under OL-CDIP, as they are visually distinct.

The construction of the database followed a light active learning framework, following two stages: preliminary clustering, followed by a manual reorganization. The clustering

Figure 4.2: Examples of two distinct classes under the proposed LA-CDIP dataset. Those classes originally belonged to the same class under RVL-CDIP organization.

is made with a private metric learning model, trained with a private document dataset, with both the model and the dataset following the same methodology as this thesis. To generate the clusters, the trained model is used to generate a feature vector for every document in the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset. Next, Hierarchical Agglomerative Clustering clustering algorithm, with Ward[30] strategy, is employed. This algorithm allows the creation of a dynamic number of clusters, opposed to a fixed k amount of clusters in algorithms, for example, required by k-Means. This is important due to the ZSL nature of the problem, as the resulting dataset needs to achieve layout and class diversity, in other words, achieve a large number of classes. This clustering served the purpose of accelerating the manual labeling process.

Through the manual labeling process, the clustering served exclusively as suggestions, as the cluster themselves were not homogeneous, neither were unique, and therefore could not be taken as ground truth. The manual labeling process was separated into two steps: first, given a cluster to analyze, clean the cluster so there is only one document pattern in the cluster. And for the second step, verify if this cluster shares a document pattern with another cluster. If so, merge both clusters into one. To ensure quality on the proposed dataset, a second independent validation of the dataset is conducted, reviewing both intra-class and inter-class consistency and fixing occasional human errors.

4.2 Leveraging LLM for Benchmarking Visual Document Matching

A structured prompt that guides LLM to compare two document images is used to benchmark VDM, evaluating visual similarity. Models were asked to assign a similarity score from 0 to 100, categorized into five levels: Nearly Identical, Highly Similar, Moderately Similar, Weak Similarity, and Completely Different. This evaluation was carried out using Google Colab [31], where the size of the model and the computational resources were limited. Figure 4.3 illustrates the input-output structure of the document comparison framework. The left and right images represent the document pairs analyzed by LLMs, which generate a similarity score and a categorical classification.

4.3 Modeling

Traditional classification methods, such as entropy learning, locks the generality of the model to the set of classes it has been trained on. Therefore, the use of Zero Shot Learning techniques are essential to be able to classify document layouts that have not been seen on the training set. A wide variety of backbones are used to experiment on the proposed dataset, but they all follow the same metric learning architecture with Siamese Networks.

The dataset is benchmarked by choosing traditional, well established vision Neural Networks as the backbones. They are ResNet [32], MobileNetV3 [33], EfficientNet [34], VGG [35], Vision Transformer (ViT) [36]. To adapt the architectures to a siamese network, only the last linear layer of each model—the classification layer—is modified into a new linear layer with an arbitrary size n , suitable for the problem. In summary, the model learns to draw a representation of a input document in a feature space with n dimensions, such that documents that share the same class are represented clustered together, and documents that have different classes are far apart.

The trained models are exclusively visual, therefore, their input data is the (R, G, B) document image matrix. First, the input image is resized into a shape compatible with each neural architecture. For most models, this shape is a $(224, 224)$ height and width. EfficientNet is the only architecture that does not follow this rule, as the different model versions increase their input size at the same time they increase the network depth and width. Then, every value of the image matrix is scaled from $0 - 255$ to $0 - 1$, and normalized with the mean and standard deviation of the current training split. During a training epoch, for each document in the dataset, a random document is chosen to close a pair. Every class has the same odds of being chosen to mitigate overfit in predominant classes. No data augmentation and no pair mining are used in these experiments.

Reference Image

PPP

2000 MARLBORO BAR PROGRAM
CONTRACT TOP SHEET

GMM/SSM: Amy Drick

MARKET: Dallas

VENUE NAME: Blackberry's

VENUE ID

DAL-0121-03

Please check the appropriate box that will identify the type of club and the appropriate contract executed by club owner/manager:

EVENT VISIBILITY MUSIC RNB

Please check the appropriate box regarding venue admission policy:

AO18-P AO21-P PAO-P
 AO18-V AO21-V

GMM/SSM
SIGNATURE: J.D. DATE: 2-3-00

SELL-IN
APPROVAL: X DATE: _____

200319501

Image to Compare

2000 MARLBORO BAR PROGRAM
CONTRACT TOP SHEET

GMN/MSM: Lindy Jenkins

MARKET: Charlotte

VENUE NAME: Coach's Sports Bar & Grill

VENUE ID

C H A - 0 0 8 . 0 . 0 1

Please check the appropriate box that will identify the type of club and the appropriate contract executed by club owner/manager:

EVENT VISIBILITY MUSIC RNB

Please check the appropriate box regarding venue admission policy:

<input type="checkbox"/> AO18-P	<input type="checkbox"/> AO21-P	<input checked="" type="checkbox"/> PAO-P
<input type="checkbox"/> AO18-V	<input type="checkbox"/> AO21-V	

GMN/MSM
 SIGNATURE: Lindy Jenkins DATE: 2-20-00

SELL-IN
 APPROVAL: J DATE: 3-13-2000

2000 MARLBORO BAR PROGRAM
 CONTRACT TOP SHEET

Similarity Score: 98
Category: Nearly Identical

Reference Image

Image to Compare

Action TAN Request

Similarity Score: 15
Category: Completely Different

Figure 4.3: Two sets of comparisons by a LLM model. The first example shows the same layout, and the second, different layouts. The scores range from 0 to 100.

The models are trained with a supervised contrastive learning framework, and use the Contrastive Loss [9] as the loss function. Its objective is to minimize the distance between two elements of the same class, but increase the distance otherwise. Given a sample pair $\{(a, b)\}$ of feature spaces, and y as a label with value 1 if a and b share the same class, 0 otherwise. Therefore:

$$Loss = y * d(a, b)^2 + (1 + (-1 * y)) * \text{abs}(m - d(a, b))^2, \quad (4.1)$$

where d represents a metric function and m a hyperparameter defining the lower bound distance between samples of different classes. In this context of this paper, $m=0.5$ and d is the Euclidean distance between two points $P = (x_1, x_2, \dots, x_n)$ and $Q = (y_1, y_2, \dots, y_n)$ in an n -dimensional space defined mathematically as:

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}. \quad (4.2)$$

Chapter 5

Results

5.1 LA-CDIP Dataset

The proposed dataset is composed of 4993 documents, divided in 144 different classes. Each class has at least two documents, the biggest class has 497 documents, and the median size of the classes is 13, exposing an extra challenge on dealing with a unbalanced dataset. Figure 5.1 shows the class distribution.

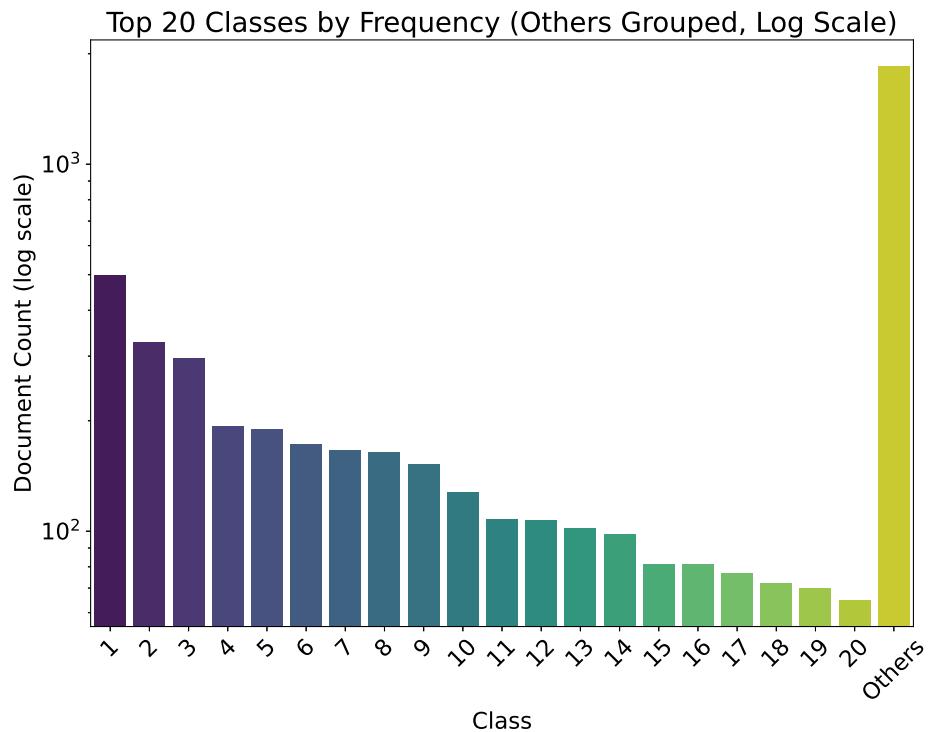


Figure 5.1: Histogram showing the top 20 classes by their frequency. The other 124 classes are grouped in the “others” column. Note that the plot is in log scale.

To maintain train-test data consistency, the data is split following the ZSL and Generalized Zero-Shot Learning (GZSL) protocols proposed by Xian et al [37]. Each protocol is a different method to split train and test data: ZSL is the complete separation of train and test classes - no overlap between the splits; while the GZSL follows a more realistic scenario by partially overlapping train and test classes, having half the test or validation set be of seen classes, and half of unseen classes. In both scenarios, the data is split into train and test data, and the train data is further divided by a 5-fold cross-validation [38].

The data for the test scenario and the data for each split are chosen randomly, and the same for every experiment. The test scenario is picked as a sixth split for the cross-validation, so it follows the same rules for the other 5 splits. For the ZSL scenario, a sixth of the classes are randomly chosen for each split, with no overlapping. Naturally, this scenario creates splits with variable sizes, as the classes themselves have variable sizes. For the GZSL scenario, half the classes in the whole dataset are chosen as classes that do not overlap between splits, and other half are diluted between the splits. While constructing a split, first, the non overlapping classes are chosen for the split, then the remaining slots are filled with the overlapped classes, achieving splits with a constant size for this scenario.

This dataset can be used for a traditional visual classification problem, but since the problem is tackled with metric learning, additional considerations must be taken into account. While using a Siamese network, running inference on a single point of data returns no information, as the architecture requires some sort of comparison between different data points. Randomly selecting pairs every test can fluctuate the results, therefore, to maintain the test consistency, a test protocol is used that assures the same pairs are tested every time. For every document in the set, two other documents are chosen: one that shares the same class, and one that does not. This document is chosen randomly by its class, therefore, every class has the same probability of being picked. A protocol exists for the test set and for every cross-validation split.

5.2 Evaluation Criteria

Since a metric learning model yields a distance between two data points (in this case, two documents), a threshold needs to be defined to declare whether the two data points belong to the same class or not, to calculate how accurate the model is. For this, Equal Error Rate (EER) is used as the metric for the experiments [39]. EER is the point at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. The EER is calculated by setting the FAR equal to the FRR and finding the corresponding threshold at this point [40]. In other words, the EER is the value of error rate when the

threshold value τ_{EER} gives

$$FAR(\tau_{EER}) = FRR(\tau_{EER}), \quad (5.1)$$

where

$$FAR(\tau) = \frac{\text{Number of false acceptances at threshold } \tau}{\text{Total number of negative samples}} \quad (5.2)$$

is the probability that a negative sample is incorrectly classified as positive [41]. On the other hand,

$$FRR(\tau) = \frac{\text{Number of false rejections at threshold } \tau}{\text{Total number of positive samples}} \quad (5.3)$$

is the probability that a positive sample (e.g., a genuine user) is incorrectly classified as negative.

Ten models are trained for each chosen backbone (see Chapter 4.3): a 5-fold cross-validation for each of the 2 scenarios, ZSL and GZSL. As shown in Section 4.1, each split follows a fixed validation protocol to ensure consistency. The trained models from each split are also evaluated on the independent test set, to compare them with the LLMs. The mean EER is reported for every cross-validation scenario, as well as the mean EER on the independent test set.

5.3 LLMs

For the LLMs analysis, the chosen models are LLaVA 3.2 Vision, InternVL 2.5, Qwen2.5-VL, GPT-4o (2024-11-20), and GPT-4o-mini (2024-07-18). They were chosen given their performance in document understanding benchmarks [29, 24], which demonstrated capabilities in tasks involving visual-text reasoning, OCR, and layout-based document analysis. These models were evaluated in a zero-shot manner, without any fine-tuning or additional training.

5.4 Partial Results

The results of the research are presented in this chapter. This chapter's discussions revolves around Table 5.1, where the results of every trained model are presented. The table shows the mean EER value of the cross-validation in both scenarios, as well as the mean performance of each fold on the independent test set. While the ZSL and the GZSL scenarios are very influential on the visual models performance, this difference is not relevant on the LLM test, as they have not been fine-tuned for the task. While they may have been originally trained with some documents of the RVL-CDIP database,

both scenarios are effectively ZSL. The Chapter 5.4.1 discusses the results of the trained visual models and compare the effects of different backbones in the result. Chapter 5.4.2, compares the results returned by the multimodal LLMs chosen for this work. Finally, Chapter 5.4.3, compares both approaches.

Table 5.1: Comparative performance between different visual backbones and Large Language Models. Following the columns: the architecture name, the architecture edition, if exists, cross-validation over the ZSL scenario, cross-validation over the GZSL scenario, test performance on the ZSL scenario, and test performance over the GZSL scenario. Every value is a mean EER (%) value over the CV folds.

Architecture	Edition	Params	ZSL	GZSL	Test ZSL	Test GZSL
AlexNet		57M	8.92	5.45	17.33	6.31
VGG	11	129M	7.47	5.01	14.24	3.95
	13	129M	7.03	4.79	9.30	3.95
	16	134M	8.29	5.23	14.74	4.82
	19	139M	7.30	4.57	17.08	3.90
ResNet	18	11M	5.03	1.54	4.98	1.51
	34	21M	4.32	2.10	4.13	1.53
	50	23M	6.90	3.39	10.34	2.21
	101	42M	8.20	2.72	11.31	1.98
	152	58M	9.44	3.38	12.70	2.39
MobileNetV3	Small	1M	7.98	5.06	12.74	5.26
	Large	4M	8.16	4.27	8.45	4.43
EfficientNet	0	4M	4.41	2.27	6.02	0.95
	1	6M	3.93	3.54	8.88	2.70
	2	7M	5.73	2.61	7.29	2.14
	3	10M	5.65	3.64	7.37	2.34
ViT	Base	87M	12.43	7.97	19.72	5.19
	Large	305M	13.16	7.57	19.88	5.26
Llama	3.2	11B	—	—	13.95	21.90
InternVL	2.5	8B	—	—	8.58	10.40
Qwen-VL	2.5	7B	—	—	6.61	4.20
GPT 4o mini	2024-07-18	*	—	—	4.70	4.07
GPT 4o	2024-11-20	*	—	—	2.75	1.33

* The parameter count of GPT-4o has not been publicly disclosed.

5.4.1 Visual Models

Among visual models, smaller and more cost-efficient architectures outperformed larger ones. This trend is evident in different ResNet variants: ResNet-18 and ResNet-34 showed

strong performance, while larger versions (ResNet-50, ResNet-101, and ResNet-152) performed progressively worse. Additionally, the ViT, despite excelling on datasets like ImageNet [42], was the worst-performing model in this task. Figure 5.2 shows that, while none of the class separations are perfect, ResNet-18 and EfficientNet-b0 have a better inter-class separation, if compared against ViT-b and ViT-l. These larger models overfitted to the training data, with some reaching 0% train error in certain epochs while maintaining a high validation error. The most likely cause of this behavior is the dataset size—LA-CDIP contains only 4,993 documents, with approximately two-thirds used for training in each fold. However, techniques such as tuple mining and data augmentation could help mitigate this effect by increasing the diversity of training samples and improving feature learning. The overall best models are the ResNet-34, and EfficientNet-0 and 1. They offer a balance in size, achieving a good generalization of the problem while also learning the intricacies the task demands.

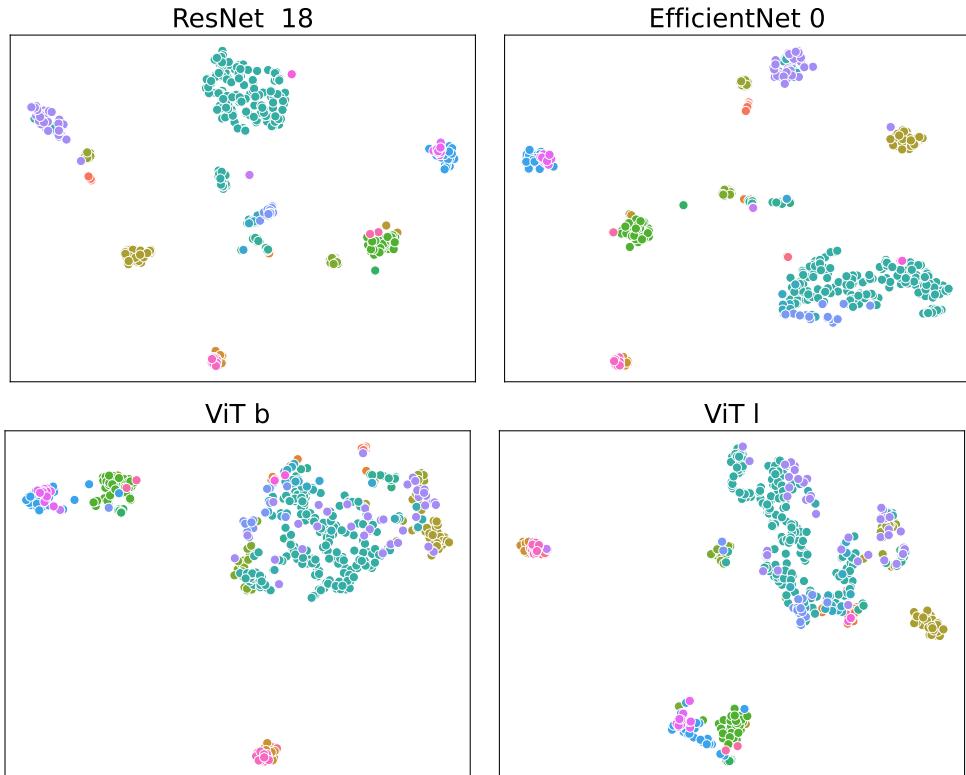


Figure 5.2: TSNE visualization of the Test ZSL scenario. These models have been trained on the same cross-validation fold.

Comparing the ZSL and the GZSL results, the GZSL consistently gives better results. This is expected, since half of the validation set on each fold is composed of classes seen in training. The biggest discrepancies between the two modalities can be seen in the largest models, where the overfitting damage is mitigated by the easier scenario. Even so, the

best performance on the GZSL scenario is usually connected with a good performance on the ZSL scenario.

In the test scenarios, the reported values represent the mean performance across all folds, evaluated on the same set. Since this is a Zero-shot data split, there is inherently a high variance over the different splits, since they test on entirely different patterns each fold. By chance, the ZSL test split is more challenging than average, resulting in higher error rates for most models relative to the mean cross-validation error rate. The test GZSL scenario follows the opposite reasoning.

5.4.2 Large Language Models

All models were tested under similar conditions, however, further performance improvements could be achieved through refined prompt engineering. Therefore, the results presented should not be considered as definitive comparisons of the models capabilities but rather a baseline for assessing the viability of the proposed method. Despite these limitations, LLM results align with existing benchmarks for document understanding tasks for these models.

As shown in Table 5.1, GPT-4o exhibited the best overall performance, achieving the lowest error rates in both the ZSL and GZSL scenarios. The GPT-4o-mini variant performed similarly but showed slightly lower accuracy, reflecting its trade-off between efficiency and capability.

Additionally, Llama 3.2 Vision faced challenges in handling multiple document inputs, requiring images to be combined into a single input before processing. This limitation impacted its ability to directly compare layouts across distinct documents, differentiating its approach from the other evaluated models.

InternVL 2.5 and Qwen2.5-VL emerged as open-source alternatives for ZSL VDM. Notably, the QwenVL 7B model reached performance levels close to those of GPT-4o mini, further highlighting the efficiency of these compact architectures.

Since the LLMs have not been fine-tuned with the training data, the only difference between the ZSL and GZSL scenarios is the number of classes to compare: ZSL has exactly $\frac{1}{6}$ of the classes and GZSL can have every class, making GZSL a more diverse test. Surprisingly, the two models performed with different trends between both scenarios: InternVL was better at the ZSL split, and GPT better at the GZSL split.

5.4.3 Comparison

Given that GZSL is an advantageous scenario for vision models—since they have been trained on half of the classes—while LLMs have not been fine-tuned for this task, both

settings effectively serve as pure zero-shot evaluations. Therefore, the comparison was conducted in the ZSL scenario.

When compared to most LLMs, LA-ZSL enabled visual models to handle ZSL VDM while maintaining a significantly lower parameter count. This relationship between performance and model efficiency is illustrated in Figure 5.3.

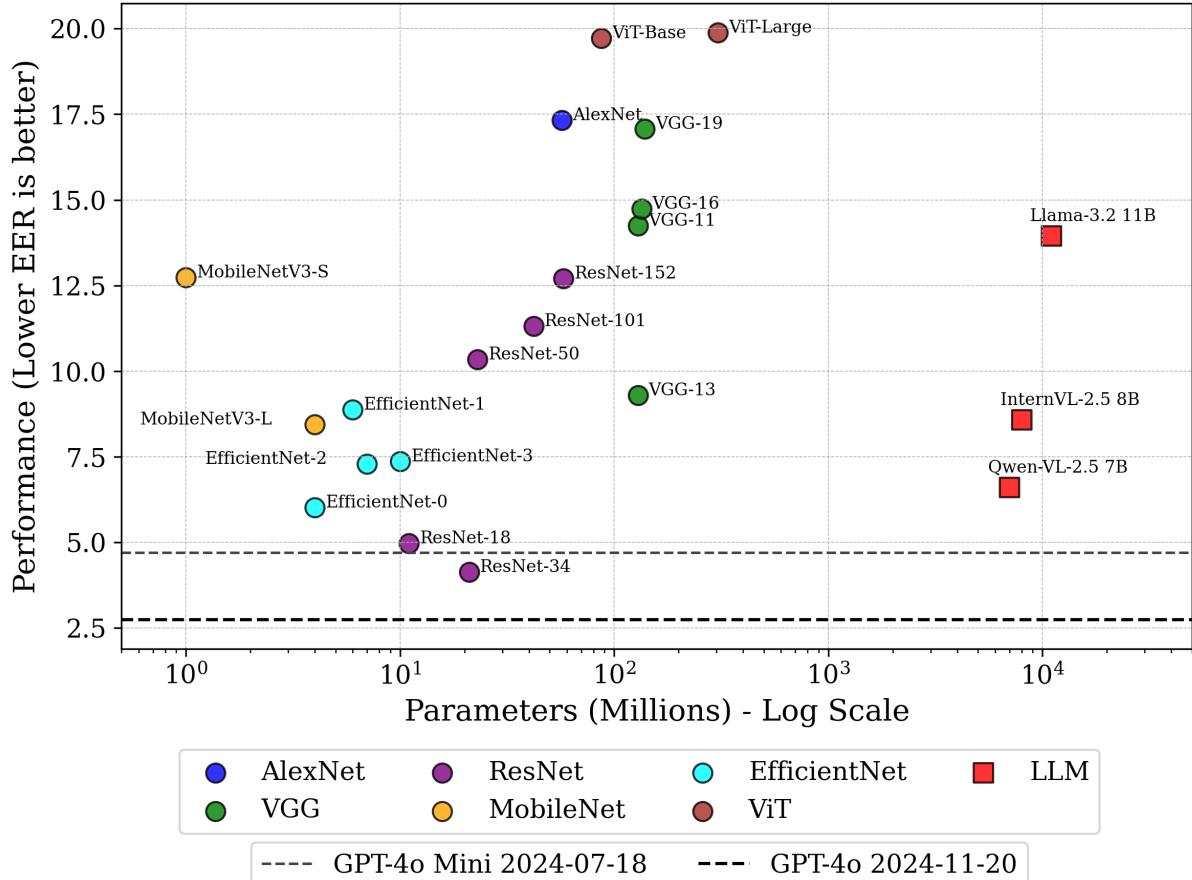


Figure 5.3: Performance vs Parameters (ZSL). The lines represent GPT-4o and GPT-4o Mini error rates, as their parameter count were not publicly disclosed.

The results highlight a trade-off between model complexity and effectiveness. Smaller visual backbones, such as ResNet-18 and EfficientNet-0, achieve competitive performance while using significantly fewer parameters than large multimodal models. Notably, ResNet-18 and ResNet-34 exhibit lower error rates than several larger architectures, reinforcing the efficiency of lightweight vision models in document matching tasks.

Chapter 6

Conclusion

This work introduces LA-CDIP, a document image dataset categorized exclusively by the layout information of the document image. This dataset enables VDM research, providing a Zero-Shot alternative to the Document Image Classification problem. The dataset is benchmarked by training siamese networks employing the LA-ZSL approach on a diverse catalog of well-established visual backbones and compare their results with popular Large Language Models that require no additional training. Generally LLMs underperform compared to visual models, except with GPT-4o, which achieves a slight performance gain over visual models. Still, it is hundreds of times more expensive per inference, making its advantage less practical.

There are some known limitations of the work and plans to tackle them. First, as mentioned, the complexity of an model architecture is currently an obstacle, as the dataset is relatively small. It is an ongoing work to solve this problem by increase its size, in number of samples per class and number of classes. Increasing the number of document sources by gathering documents from sources other than the RVL-CDIP dataset, also hold value as research, as it increases the generalization of the trained models, and is already mapped as future work. This should allow effectiveness in the bigger models. In the topic of the training pipeline, employing data augmentation techniques should relieve the necessity to expand the dataset, allowing the utilization of greater and more complex models.

6.1 Timeline

This section presents the timeline of this master’s research. The production of this document happened in the 8th semester. Therefore, everything that came before represents the past: classes, steps of the research and experiments. Everything that comes after is the intended work and next steps, culminating on the thesis defense, in the 10th semester.

Task Description	Trimester									
	1 ^o	2 ^o	3 ^o	4 ^o	5 ^o	6 ^o	7 ^o	8 ^o	9 ^o	10 ^o
Class: Artificial Intelligence 1										
Class: Seminar										
Class: Algorithm Design and Complexity										
Scope Definition										
Literature Review										
Pretrained Image Models Experiments										
Class: Fundamentals of Computer Systems										
Class: Internship										
Framework Construction										
Dataset Labeling										
Architecture Experiments										
Paper Production										
Qualification									x	
Extra Dataset Labeling										
Training Framework Enhancement										
Multimodal Training										
Master's Thesis Defense										x

Table 6.1: The timeline of the master's research. The "x" represents the current moment in the timeline.

Bibliography

- [1] Kay, Anthony: *Tesseract: an open-source optical character recognition engine*. Linux Journal, 2007(159):2, 2007. 1
- [2] Mindee: *doctr: Document text recognition*. <https://github.com/mindee/doctr>, 2021. 1
- [3] Liu, Li, Zhiyu Wang, Taorong Qiu, Qiu Chen, Yue Lu, and Ching Y. Suen: *Document image classification: Progress over two decades*. Neurocomputing, 453:223–240, 2021. <https://doi.org/10.1016/j.neucom.2021.04.114>. 1
- [4] Bakkali, Souhail, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol: *EAML: ensemble self-attention-based mutual learning network for document image classification*. Int. J. Doc. Anal. Recognit., 24(3):251–268, September 2021, ISSN 1433-2833. <https://doi.org/10.1007/s10032-021-00378-0>, visited on 2025-06-03. 2
- [5] Harley, Adam W., Alex Ufkes, and Konstantinos G. Derpanis: *Evaluation of deep convolutional nets for document image classification and retrieval*. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 991–995. IEEE Computer Society, 2015. <https://doi.org/10.1109/ICDAR.2015.7333910>. 2, 9, 12
- [6] Xian, Yongqin, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata: *Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly*. IEEE Trans. Pattern Anal. Mach. Intell., 41(9):2251–2265, 2019. <https://doi.org/10.1109/TPAMI.2018.2857768>. 2
- [7] Sinha, Sankalp, Muhammad Saif Ullah Khan, Talha Uddin Sheikh, Didier Stricker, and Muhammad Zeshan Afzal: *CICA: content-injected contrastive alignment for zero-shot document image classification*. In Smith, Elisa H. Barney, Marcus Liwicki, and Liangrui Peng (editors): *Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, Greece, August 30 - September 4, 2024, Proceedings, Part IV*, volume 14807 of *Lecture Notes in Computer Science*, pages 124–141. Springer, 2024. https://doi.org/10.1007/978-3-031-70546-5_8. 3, 10
- [8] Scius-Bertrand, Anna, Michael Jungo, Lars Vögtlin, Jean-Marc Spat, and Andreas Fischer: *Zero-shot prompting and few-shot fine-tuning: Revisiting document image classification using large language models*. In Antonacopoulos, Apostolos, Subhasis Chaudhuri, Rama Chellappa, Cheng-Lin Liu, Saumik Bhattacharya, and Uma-pada Pal (editors): *Pattern Recognition - 27th International Conference, ICPR*

2024, Kolkata, India, December 1-5, 2024, Proceedings, Part XIX, volume 15319 of *Lecture Notes in Computer Science*, pages 152–166. Springer, 2024. https://doi.org/10.1007/978-3-031-78495-8_10. 3, 10

- [9] Chopra, Sumit, Raia Hadsell, and Yann LeCun: *Learning a similarity metric discriminatively, with application to face verification*. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society, 2005. <https://doi.org/10.1109/CVPR.2005.202>. 6, 16
- [10] Hoffer, Elad and Nir Ailon: *Deep metric learning using Triplet network*, December 2018. <http://arxiv.org/abs/1412.6622>, visited on 2025-06-09, arXiv:1412.6622 [cs]. 6
- [11] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei: *Language Models are Few-Shot Learners*, July 2020. <http://arxiv.org/abs/2005.14165>, visited on 2025-06-08, arXiv:2005.14165 [cs].
7
- [12] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample: *LLaMA: Open and Efficient Foundation Language Models*, February 2023. <http://arxiv.org/abs/2302.13971>, visited on 2025-06-08, arXiv:2302.13971 [cs].
7
- [13] Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung: *Survey of Hallucination in Natural Language Generation*. ACM Comput. Surv., 55(12):248:1–248:38, March 2023, ISSN 0360-0300. <https://doi.org/10.1145/3571730>, visited on 2025-06-08. 7
- [14] Abdallah, Abdelrahman, Daniel Eberharter, Zoe Pfister, and Adam Jatowt: *A survey of recent approaches to form understanding in scanned documents*. Artificial Intelligence Review, 57(12), 2024. <https://doi.org/10.1007/s10462-024-11000-0>. 8
- [15] Zhong, Xu, Jianbin Tang, and Antonio Jimeno-Yepes: *Publaynet: Largest dataset ever for document layout analysis*. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE, 2019. <https://doi.org/10.1109/ICDAR.2019.00166>. 8
- [16] Lewis, David D., Gady Agam, Shlomo Argamon, Ophir Frieder, David A. Grossman, and Jefferson Heard: *Building a test collection for complex document information processing*. In Efthimiadis, Efthimis N., Susan T. Dumais, David Hawking, and Kalervo Järvelin (editors): *SIGIR 2006: Proceedings of the 29th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, pages 665–666. ACM, 2006. <https://doi.org/10.1145/1148170.1148307>. 9

- [17] Mathew, Minesh, Dimosthenis Karatzas, and C. V. Jawahar: *Docvqa: A dataset for VQA on document images*. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE, 2021. <https://doi.org/10.1109/WACV48630.2021.00225>. 9
- [18] Jaume, Guillaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran: *FUNSD: A dataset for form understanding in noisy scanned documents*. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6, 2019. 9
- [19] Huang, Zheng, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar: *ICDAR2019 competition on scanned receipt OCR and information extraction*. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, September 2019. <https://doi.org/10.1109/ICDAR.2019.00244>. 9
- [20] Park, Seunghyun, Seung Shin, Byeongchang Kim, Junbum Cha, and Hwalsuk Lee: *CORD: a consolidated receipt dataset for post-ocr parsing*. In *Document Intelligence Workshop at NeurIPS 2019*, 2019. <https://arxiv.org/abs/1908.07414>. 9
- [21] Xu, Yiheng, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei: *Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding*. CoRR, abs/2104.08836, 2021. <https://arxiv.org/abs/2104.08836>. 9
- [22] Youssef, Ali, Gabriele Valvano, and Giacomo Veneri: *Document layout analysis with variational autoencoders: An industrial application*. In Ceci, Michelangelo, Sergio Flesca, Elio Masciari, Giuseppe Manco, and Zbigniew W. Ras (editors): *Foundations of Intelligent Systems - 26th International Symposium, ISMIS 2022, Cosenza, Italy, October 3-5, 2022, Proceedings*, volume 13515 of *Lecture Notes in Computer Science*, pages 477–486. Springer, 2022. https://doi.org/10.1007/978-3-031-16564-1_46. 9
- [23] Zeghidi, Hédi, Carlos Fernando Crispim Junior, and Iuliia Tkachenko: *Cdp-sim: Similarity metric learning to identify the fake copy detection patterns*. In *IEEE International Workshop on Information Forensics and Security, WIFS 2023, Nürnberg, Germany, December 4-7, 2023*, pages 1–6. IEEE, 2023. <https://doi.org/10.1109/WIFS58808.2023.10374744>. 10
- [24] AI, Meta: *Llama 3.2: From cloud to edge, now with vision*. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024. 10, 19
- [25] Wu, Zhiyu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu,

- Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan: *Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding*. CoRR, abs/2412.10302, 2024. <https://doi.org/10.48550/arXiv.2412.10302>. 10
- [26] Chen, Zhe, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huirong Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang: *Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling*. CoRR, abs/2412.05271, 2024. <https://doi.org/10.48550/arXiv.2412.05271>. 11
- [27] OpenAI: *Hello gpt-4o*. <https://openai.com/index/hello-gpt-4o/>, 2024. 11
- [28] OpenAI: *Gpt-4o mini: advancing cost-efficient intelligence*. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. 11
- [29] Bai, Shuai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al.: *Qwen2.5-vl technical report*. arXiv preprint arXiv:2502.13923, 2025. <https://arxiv.org/abs/2502.13923>. 11, 19
- [30] Jr., Joe H. Ward: *Hierarchical grouping to optimize an objective function*. Journal of the American Statistical Association, 58(301):236–244, 1963. <https://doi.org/10.1080/01621459.1963.10500845>. 13
- [31] Research, Google: *Google colaboratory*. <https://colab.research.google.com/>, 2024. 14
- [32] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun: *Deep residual learning for image recognition*. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. <https://doi.org/10.1109/CVPR.2016.90>. 14
- [33] Howard, Andrew, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu: *Searching for mobilenetv3*. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1314–1324. IEEE, 2019. <https://doi.org/10.1109/ICCV.2019.00140>. 14
- [34] Tan, Mingxing and Quoc V. Le: *Efficientnet: Rethinking model scaling for convolutional neural networks*. In Chaudhuri, Kamalika and Ruslan Salakhutdinov (editors): *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. <http://proceedings.mlr.press/v97/tan19a.html>. 14

- [35] Simonyan, Karen and Andrew Zisserman: *Very deep convolutional networks for large-scale image recognition*. In Bengio, Yoshua and Yann LeCun (editors): *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. <http://arxiv.org/abs/1409.1556>. 14
- [36] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby: *An image is worth 16x16 words: Transformers for image recognition at scale*. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. <https://openreview.net/forum?id=YicbFdNTTy>. 14
- [37] Xian, Yongqin, Tobias Lorenz, Bernt Schiele, and Zeynep Akata: *Feature generating networks for zero-shot learning*. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5542–5551. Computer Vision Foundation / IEEE Computer Society, 2018. http://openaccess.thecvf.com/content_cvpr_2018/html/Xian_Feature_Generating_Networks_CVPR_2018_paper.html. 18
- [38] Wong, Tzu-Tsung and Po-Yang Yeh: *Reliable accuracy estimates from k-fold cross validation*. *IEEE Trans. Knowl. Data Eng.*, 32(8):1586–1594, 2020. <https://doi.org/10.1109/TKDE.2019.2912815>. 18
- [39] Tolosana, Ruben, Ruben Vera-Rodriguez, Carlos Gonzalez-Garcia, Julian Fierrez, Santiago Rengifo, Aythami Morales, Javier Ortega-Garcia, Juan Carlos Ruiz-Garcia, Sergio Romero-Tapiador, Jiajia Jiang, et al.: *Icdar 2021 competition on on-line signature verification*. In *Document Analysis and Recognition-ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 723–737. Springer, 2021. 18
- [40] Hofbauer, Heinz and Andreas Uhl: *Calculating a boundary for the significance from the equal-error rate*. In *2016 International Conference on Biometrics (ICB)*, pages 1–4, 2016. 18
- [41] Agrawal, Pinki, Ravikant Kapoor, and Sanjay Agrawal: *A hybrid partial fingerprint matching algorithm for estimation of equal error rate*. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pages 1295–1299, 2014. 19
- [42] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei: *Imagenet: A large-scale hierarchical image database*. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. <https://doi.org/10.1109/CVPR.2009.5206848>. 21