Layout-Aware Zero-Shot Learning for Visual Document Matching

Qualificação de Mestrado

Lucas de Almeida Bandeira Macedo

Universidade de Brasília
Departamento de Ciência da Computação

Orientador: Prof. Dr. Pedro Garcia Freitas
Coorientador: Prof. Dr. Bruno Luiggi Macchiavello Espinoza

Outubro de 2025

Capítulos

Introdução

Contexto - Documentos e Compliance

- Ambiente Bancário
- Documentos físicos
- Imagens de documentos
- Exemplos:

Contexto - Fluxo de Compliance

Contexto - Classificação de Imagem

- Assegurar que o documento está correto
- Documentos não-digitais
- Evita fraudes

Contexto - Classificação de Imagem

- Assegurar que o documento está correto
- Documentos não-digitais
- Evita fraudes

Contexto - Classificação

Classificação Tradicional:

- Categorização em classes predefinidas
- Cross-Entropy Loss

Desempenho Atual:

- Bakkali et al. (2021): 97.70% de acurácia no RVL-CDIP

Contexto - Classificação

Classificação Tradicional:

- Categorização em classes predefinidas
- Cross-Entropy Loss

Desempenho Atual:

- Bakkali et al. (2021): 97.70% de acurácia no RVL-CDIP

O Problema:

- Novos layouts de documentos
- Classes completamente novas
- Necessidade de retreinamento
- Semanas/meses de engenharia de dados e treinamento

Zero-Shot Learning

Permite que o modelo reconheça elementos de classes nunca vistas no treinamento

Sobre a Pesquisa

Desafios

- Falta de dataset especializado
  - Imagens de Documento
  - Generalização
  - Divisão treino e teste zero-shot
- Ausência de metodologia estado-da-arte
  - Paradigma ZSL
  - Capacidade de classificar

Contribuições

1. **Novo dataset LA-CDIP**
   - Classificação ZSL
   - Derivado do RVL-CDIP
2. **Abordagem de Visual Document Matching (VDM)**
   - Similaridade de documentos
   - Metric Learning
   - Generalização Zero-Shot
3. **Avaliação sistemática**
   - Benchmark extensivo
   - Comparação com LLM

Metodologia

Dataset - Motivação

Datasets Disponíveis

- PubLayNet, DocLayNet
- DocVQA
- CORD, SROIE
- RVL-CDIP

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 has been a very active research area in recent years. However, documents often change in terms of format and their visual patterns, which may inquire 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 were not 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/AIMSHub/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).

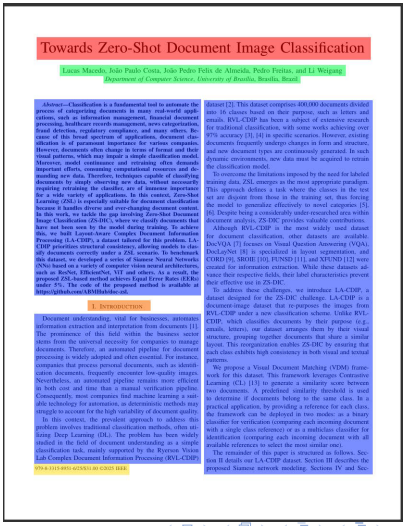
The remainder of this paper is structured as follows. Section II details our LA-CDIP dataset. Section III describes the proposed Siamese network modeling. Sections IV and Sec-

979-8-3315-8093-6/25/$31.00 ©2025 IEEE

Dataset - Motivação

Datasets Disponíveis

- PubLayNet, DocLayNet
- DocVQA
- CORD, SROIE
- RVL-CDIP


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 has been a long-standing research topic in computer vision. 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 introduce LA-CDIP, a layout-aware document dataset based on RVL-CDIP, where visually similar classes have been merged and have now to be 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 3%. The code of the proposed method is available at https://github.com/AIMSlab-unb/.

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)

978-1-3503-8855-6/23/$31.00 ©2023 IEEE

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.

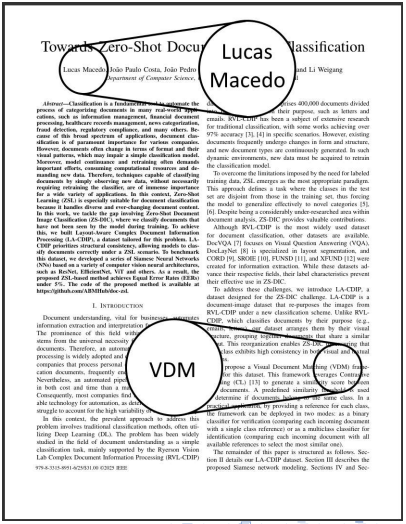
We propose a Visual Document Matching (VDM) framework for this dataset. This framework leverages Contrastive Learning (CL) [13] to generate a similarity score between two documents. A predefined similarity threshold is used to determine if documents belong to the same class. In a practical application, 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-

Dataset - Motivação

Datasets Disponíveis

- PubLayNet, DocLayNet
- DocVQA
- CORD, SROIE
- RVL-CDIP



Dataset - Motivação

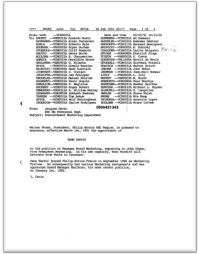
Datasets Disponíveis

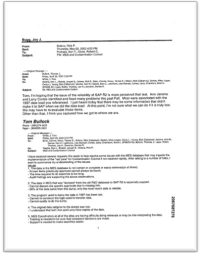
- PubLayNet, DocLayNet
- DocVQA
- CORD, SROIE
- RVL-CDIP

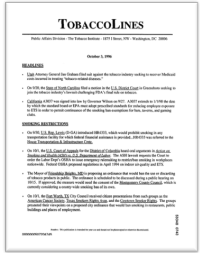
RVL-CDIP

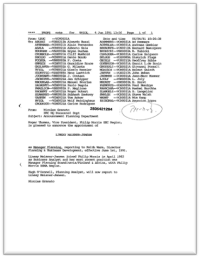
- 400.000 documentos
- 16 classes
- email, formulário, carta...
- Separado por função

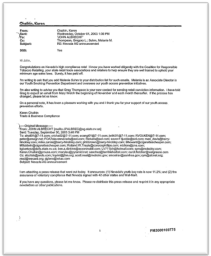
Dataset - Motivação

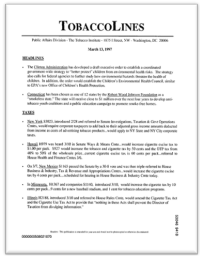




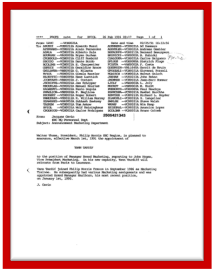


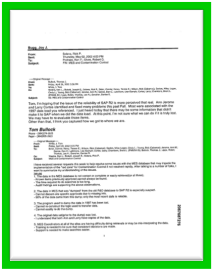


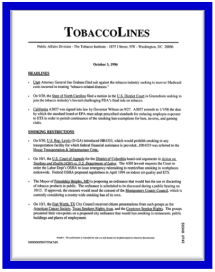


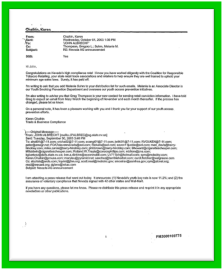


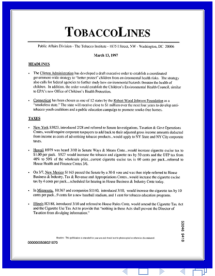
Dataset - Motivação



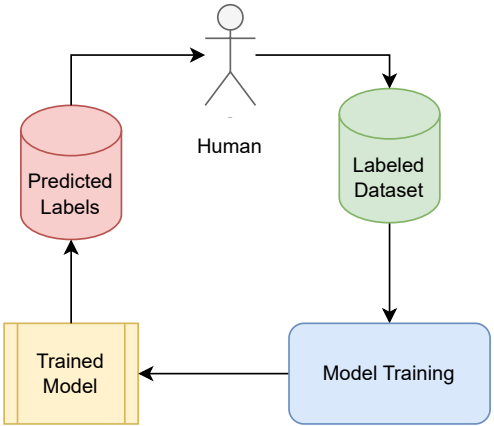




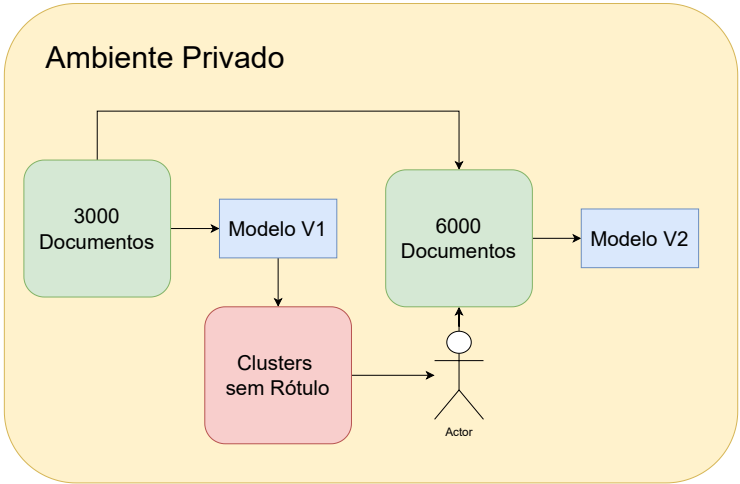




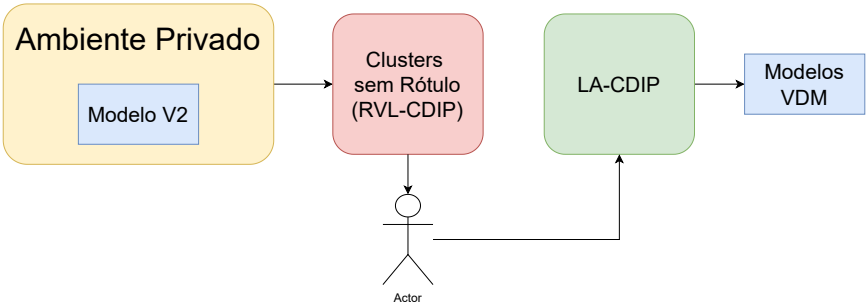
LA-CDIP Dataset - Active Learning



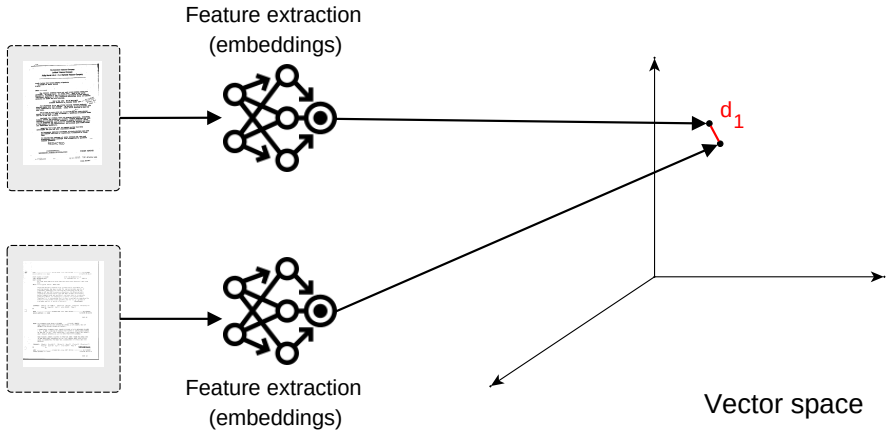
LA-CDIP Dataset - Rotulação



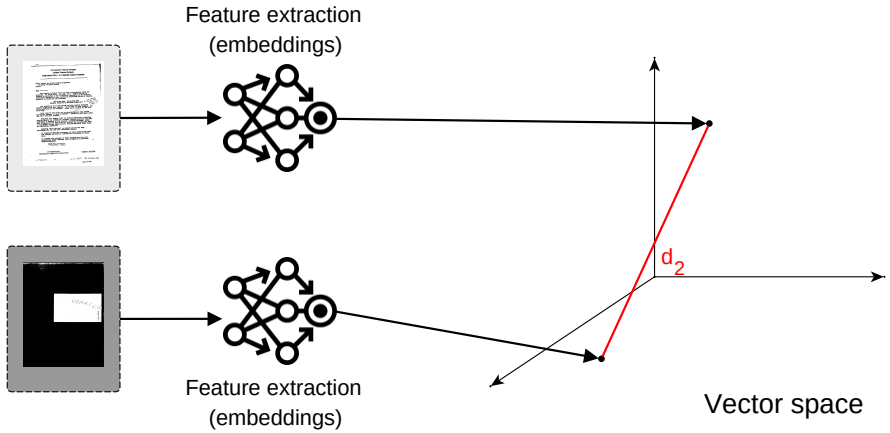
LA-CDIP Dataset - Rotulação



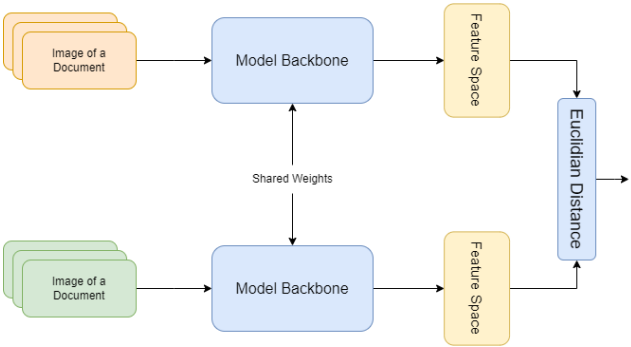
Visual Document Matching



Visual Document Matching

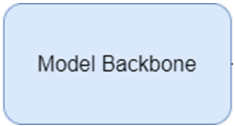


Visual Document Matching - Arquitetura



$$\mathcal{L} = \begin{cases} d(x_1, x_2)^2, & \text{if } y = 1 \\ \max(0, m - d(x_1, x_2))^2, & \text{otherwise.} \end{cases}$$

Visual Document Matching - Backbone



- AlexNet
- ResNet
- VGG
- EfficientNet
- MobileNetV3
- Vision Transformer (ViT)

Benchmarking com LLMs

Modelos Avaliados:

- LLaVA 3.2 Vision
- InternVL 2.5
- Qwen2.5-VL
- GPT-4o (2024-11-20)
- GPT-4o-mini (2024-07-18)

Avaliação:

- Zero-shot (sem fine-tuning)
- Pontuação de similaridade 0–100
- 5 níveis de categorização

Benchmarking com LLMs

Reference Image

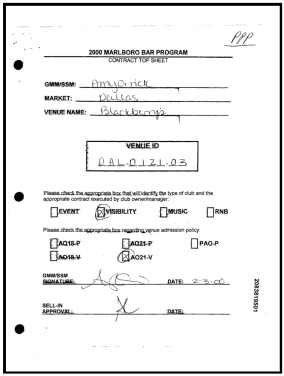
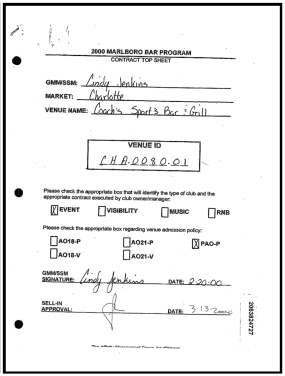


Image to Compare



Similarity Score: 98
Category: Nearly Identical

Reference Image

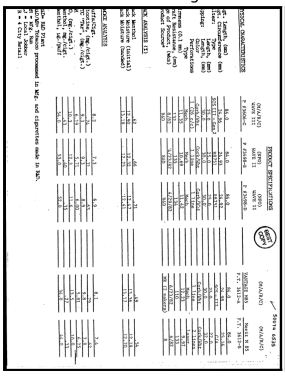
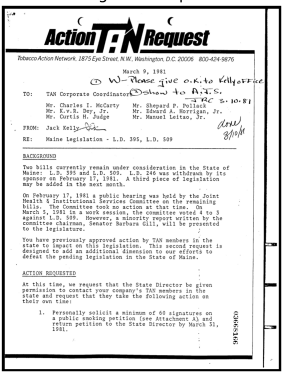


Image to Compare



Similarity Score: 15
Category: Completely Different

Resultados

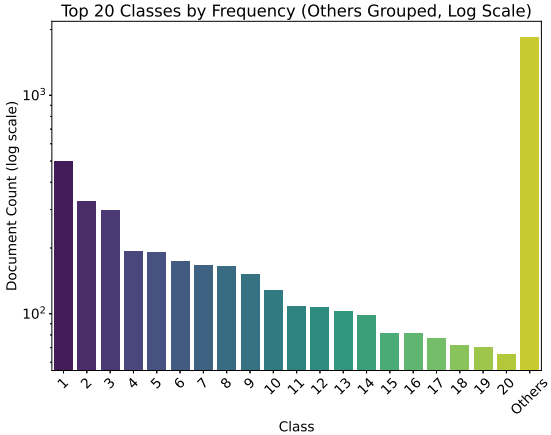
LA-CDIP Dataset - Estatísticas

Composição:

- 4.993 documentos
- 144 classes diferentes
- Min: 2 documentos/classe
- Max: 497 documentos/classe
- Mediana: 13 documentos/classe

Splits:

- ZSL: separação completa treino/teste
- GZSL: 50% overlap de classes
- 5-fold cross-validation



Métricas de Avaliação

Equal Error Rate (EER)

- Ponto onde FAR = FRR
- FAR: False Acceptance Rate
- FRR: False Rejection Rate

$$FAR(\tau) = \frac{\text{False Acceptances}}{\text{Total Negatives}} \qquad FRR(\tau) = \frac{\text{False Rejections}}{\text{Total Positives}}$$

Protocolo de Teste:

Para cada documento: 1 par similar + 1 par dissimilar

Resultados - Visão Geral

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

Resultados - Visão Geral

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

Resultados - Visão Geral

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

Resultados - Visão Geral

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

Resultados - Visão Geral

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

Resultados - Visão Geral

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

Resultados - ResNet e EfficientNet

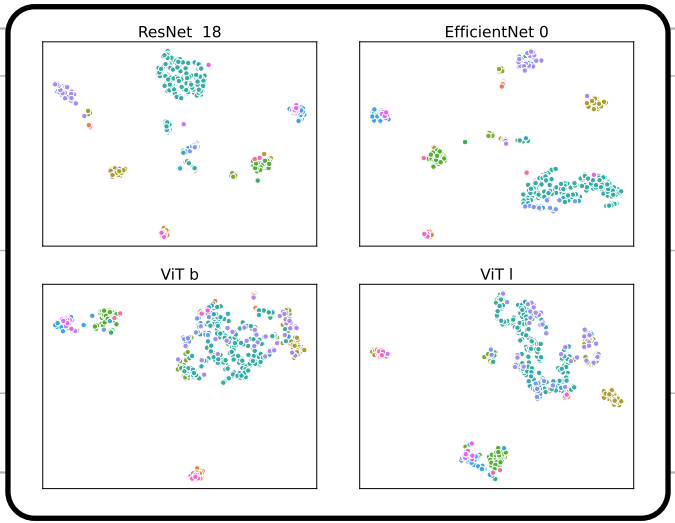
Architecture	Edition	Params	ZSL	GZSL	Test ZSL	Test GZSL
ResNet	18	11M	5.03	**1.54**	4.98	1.51
	34	21M	4.32	2.10	**4.13**	1.53
	50	23M	6.90	3.39	10.34	2.21
	101	42M	8.20	2.72	11.31	1.98
	152	58M	9.44	3.38	12.70	2.39
EfficientNet	0	4M	4.41	2.27	6.02	**0.95**
	1	6M	**3.93**	3.54	8.88	2.70
	2	7M	5.73	2.61	7.29	2.14
	3	10M	5.65	3.64	7.37	2.34

Resultados - ResNet, EfficientNet e ViT

Architecture	Edition	Params	ZSL	GZSL	Test ZSL	Test GZSL
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
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

Resultados - Visualização TSNE

Architecture	Test GZSL
ResNet	1.51
	1.53
	2.21
	1.98
	2.39
EfficientNet	**0.95**
	2.70
	2.14
	2.34
ViT	5.19
	5.26



Resultados - Large Language Models

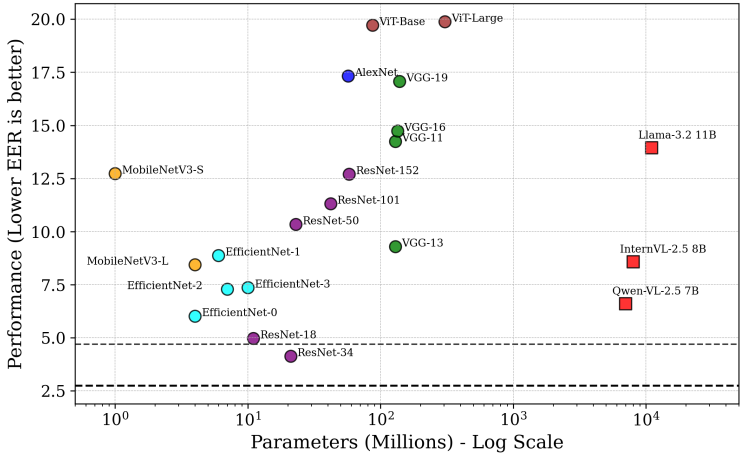
Model	Version	Params	Test ZSL	Test GZSL
Llama	3.2	11B	13.95	21.90
InternVL	2.5	8B	8.58	10.40
Qwen-VL	2.5	7B	6.61	4.20
GPT-4o mini	2024-07-18	*	4.70	4.07
GPT-4o	2024-11-20	*	**2.75**	**1.33**

* Parameter count not publicly disclosed

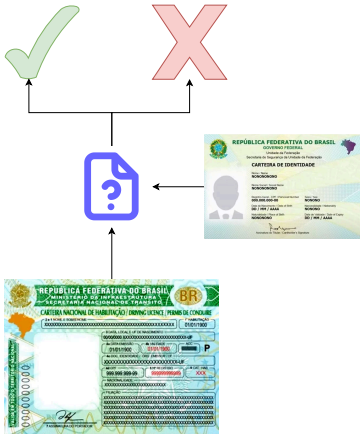
Resultados - Melhores Modelos Visuais vs LLMs

Architecture	Edition	Params	Test ZSL	Test GZSL
ResNet	18	11M	4.98	1.51
	34	21M	4.13	1.53
	50	23M	10.34	2.21
	101	42M	11.31	1.98
	152	58M	12.70	2.39
EfficientNet	0	4M	6.02	**0.95**
	1	6M	8.88	2.70
	2	7M	7.29	2.14
	3	10M	7.37	2.34
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

Comparação: Visual Models vs LLMs

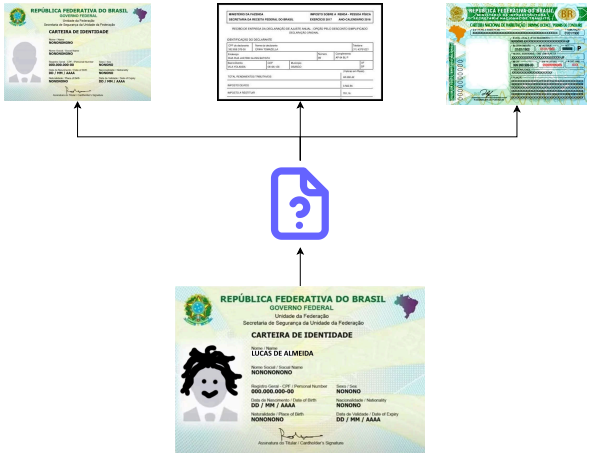


Comparação: Visual Models vs LLMs



Verificação - Classificação Binária

Comparação: Visual Models vs LLMs



Identificação - Classificação Multiclasse

Conclusão

Conclusões

1. **LA-CDIP Dataset**

   Categorizado por layout visual para Zero-Shot Learning

2. **Metodologia com VDM**

   Enforçando comparação de documentos como forma de classificação

3. **Benchmarking Sistemático**

   Comparação entre modelos visuais e LLMs

Trabalhos Futuros

Trabalhos Futuros:

- Aumentar número de amostras
- Aumentar número de classes
- Incluir fontes adicionais de documentos
- PIBIC

Publicações

Visual Document Matching for Zero-Shot Document Classification

Lucas De Almeida Bandeira Macedo ✉ [1][0009−0004−6952−4731], Joao Paulo Vieira Costa ✉ [1][0000−0001−7379−0961], Joao Pedro Felix De Almeida [1][0009−0008−3367−8033], Pedro Garcia Freitas [1][0000−0003−0866−658X], and Li Weigang [1][0000−0003−1826−1850]

University of Brasília, UnB - Brasília, Brasilia - Federal District, 70910-900

Abstract. Accurate document identification is crucial for ensuring compliance and maintaining consistency across various applications. Consequently, document classification has been extensively studied. However, the dynamic nature of documents often renders simple classification insufficient, necessitating frequent model retraining. To mitigate this constant maintenance, comparing incoming documents against known references is a more suitable approach. To address the zero-shot classification problem, we introduce Visual Document Matching (VDM). VDM focuses on verifying whether two documents share an identical visual layout structure, a particularly effective approach in scenarios where training classes do not align with those encountered during inference. Despite its significant potential, the Zero-Shot Learning (ZSL) approach remains largely underexplored in document layout understanding. To support our study, we introduce Layout-Aware Complex Document Information Processing (LA-CDIP), a dataset comprising 4,993 documents across 144 classes. We reorganized this dataset from the RVL-CDIP database to emphasize visual structure over semantic information. Our approach is benchmarked using a siamese network within a contrastive learning framework across multiple backbone architectures, including ResNet, EfficientNet, and Vision Transformer (ViT). In zero-shot scenarios, our method achieves an Equal Error Rate (EER) below 5% in 1-vs-1 verification with cross-validation. Furthermore, our VDM approach outperforms lighter Large Language Models (LLMs) and rivals GPT-4o. These findings highlight the superiority of specialized VDM techniques over general-purpose multimodal models, demonstrating high accuracy with significantly fewer parameters, making our approach more practical for real-world applications.

Keywords: Zero-shot learning · Metric Learning · Document Understanding · Document Image Classification · Visual Document Matching

1 Introduction

Documents are central to many organizations, acting as official records for contracts, identities, and evidence, and playing a crucial role in legal, financial, and

ICDAR

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

SIBGRAPI

Visual Document Matching for Zero-Shot Document Classification

Lucas De Almeida Bandeira Macedo ✉[1][0009-0004-6952-4731], Joao Paulo Vieira Costa ✉[1][0000-0001-7379-0961], Joao Pedro Felix De Almeida[1][0009-0008-3367-8033], Pedro Garcia Freitas[1][0000-0003-0866-658X], and Li Weigang[1][0000-0003-1826-1850]

University of Brasília, UnB - Brasília, Brasilia - Federal District, 70910-900

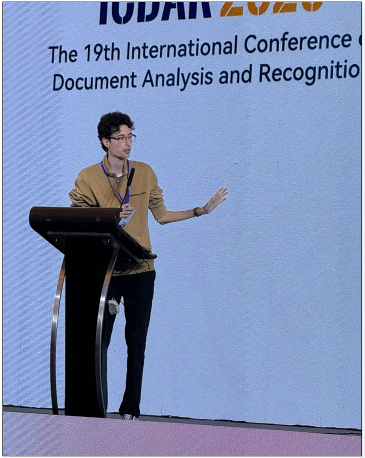
Abstract. Accurate document identification is crucial for ensuring compliance and maintaining consistency across various applications. Consequently, document classification has been extensively studied. However, the dynamic nature of documents often renders simple classification insufficient, necessitating frequent model retraining. To mitigate this constant maintenance, comparing incoming documents against known references is a more suitable approach. To address the zero-shot classification problem, we introduce Visual Document Matching (VDM). VDM focuses on verifying whether two documents share an identical visual layout structure, a particularly effective approach in scenarios where training classes do not align with those encountered during inference. Despite its significant potential, the Zero-Shot Learning (ZSL) approach remains largely underexplored in document layout understanding. To support our study, we introduce Layout-Aware Complex Document Information Processing (LA-CDIP), a dataset comprising 4,993 documents across 144 classes. We reorganized this dataset from the RVL-CDIP database to emphasize visual structure over semantic information. Our approach is benchmarked using a siamese network within a contrastive learning framework across multiple backbone architectures, including ResNet, EfficientNet, and Vision Transformer (ViT). In zero-shot scenarios, our method achieves an Equal Error Rate (EER) below 5% in 1-vs-1 verification with cross-validation. Furthermore, our VDM approach outperforms lighter Large Language Models (LLMs) and rivals GPT-4o. These findings highlight the superiority of specialized VDM techniques over general-purpose multimodal models, demonstrating high accuracy with significantly fewer parameters, making our approach more practical for real-world applications.

Keywords: Zero-shot learning · Metric Learning · Document Understanding · Document Image Classification · Visual Document Matching

1 Introduction

Documents are central to many organizations, acting as official records for contracts, identities, and evidence, and playing a crucial role in legal, financial, and

ICDAR



Apresentação

Descrição da Tarefa	Trimestre									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Disciplina: IA 1										
Disciplina: Seminário										
Disciplina: Projeto e Análise de Algoritmos										
Definição de Escopo										
Revisão Bibliográfica										
Experimentos com Modelos Pré-treinados										
Disciplina: Fundamentos de Sistemas Computacionais										
Disciplina: Estágio Docência										
Construção do Framework										
Rotulação do Dataset										
Experimentos de Arquitetura										
Produção do Artigo										
Rotulação Adicional do Dataset										
Apresentação do Artigo										
Qualificação										×
Defesa da Dissertação										

Obrigado!

Lucas de Almeida Bandeira Macedo
lucasabmacedo@hotmail.com

Orientador: Prof. Dr. Pedro Garcia Freitas
Coorientador: Prof. Dr. Bruno Luiggi Macchiavello Espinoza