

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

1 Introdução

2 Metodologia

3 Resultados

4 Conclusão

Introdução

Contexto - Documentos e Compliance

- Documentos físicos
- Imagens de documentos
- Exemplos:

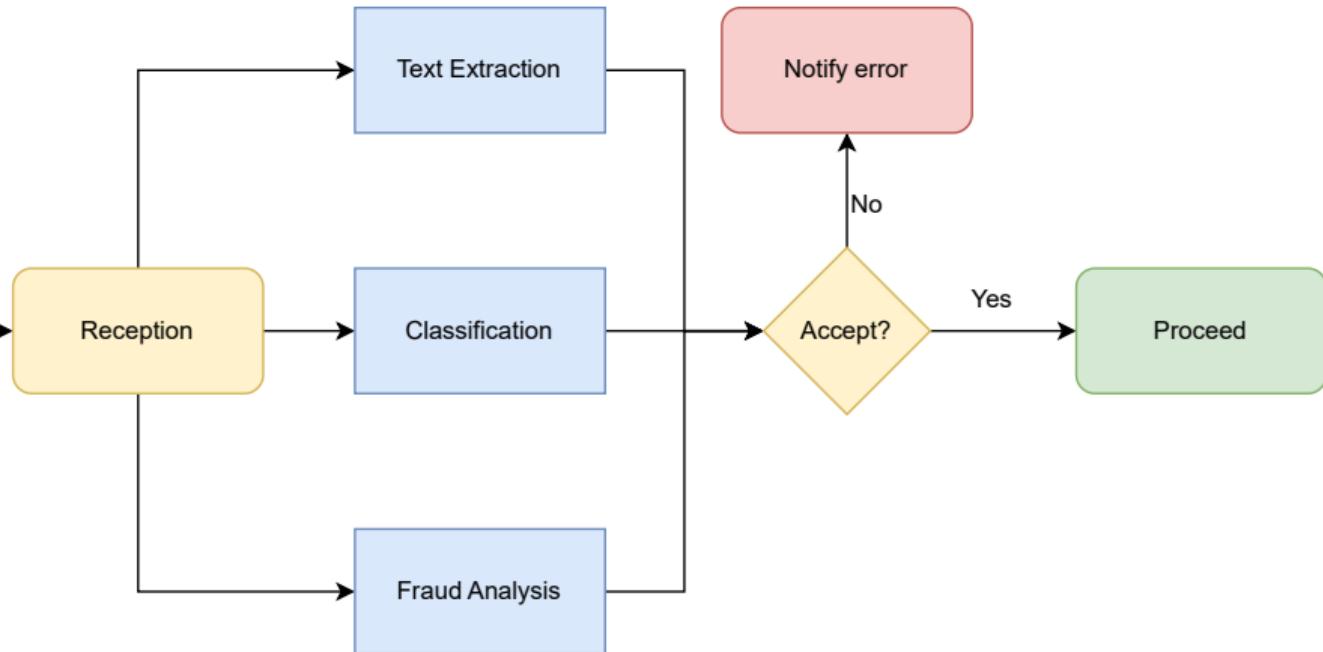


MINISTÉRIO DA FAZENDA IMPOSTO SOBRE A RENDA - PESSOA FÍSICA EXERCÍCIO 2017 ANO-CALENDÁRIO 2016			
RECEBO DE ENTREGA DA DECLARAÇÃO DE AJUSTE ANUAL - OPÇÃO PELO DESCONTO SIMPLIFICADO DECLARAÇÃO ORIGINAL			
IDENTIFICAÇÃO DO DECLARANTE			
CPF do declarante 332.653.378-08	Nome do declarante ERIKA TOMAZELLA	Telefone (11) 43791221	
Endereço RUA RUA JUSTINO ALVES BATISTA		Número 89	Complemento AP 64 BL P
Bairro/Local VILA VILANDA	CEP 06198-120	Município OGASCO	UF/ SP
(Valores em Reais)			
TOTAL RENDIMENTOS TRIBUTÁVEIS		62.260,42	
IMPOSTO DEVIDO		3.562,59	
IMPOSTO A RESTITUIR		781,16	

Contexto - Processamento de Documentos



Actor

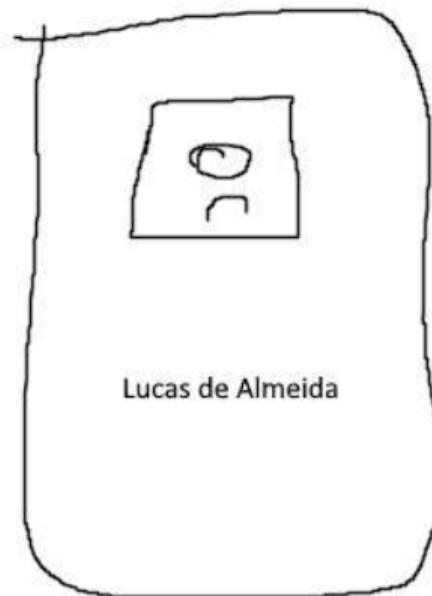


Contexto - Importância da Classificação da Imagem

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

Contexto - Importância da Classificação da Imagem

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



Classificação Tradicional:

- Categorização em classes predefinidas
- Cross-Entropy Loss

Desempenho Atual:

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

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

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

Datasets Disponíveis

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

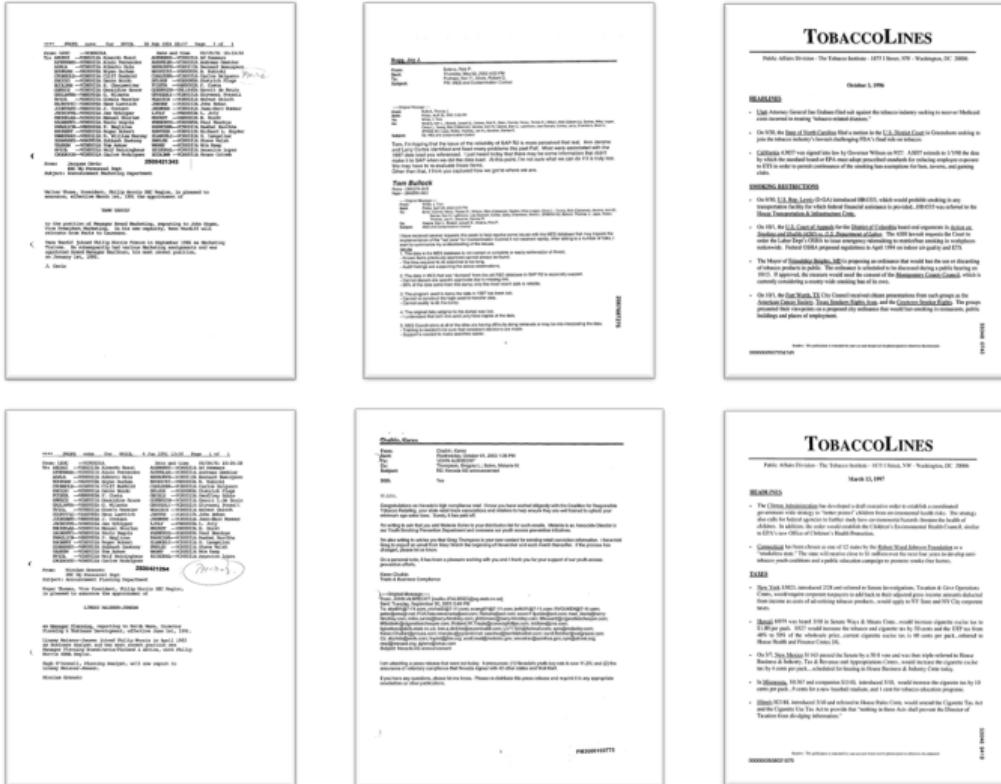
Datasets Disponíveis

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

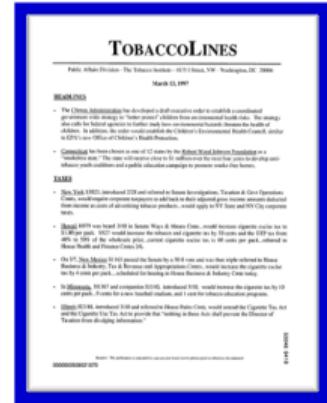
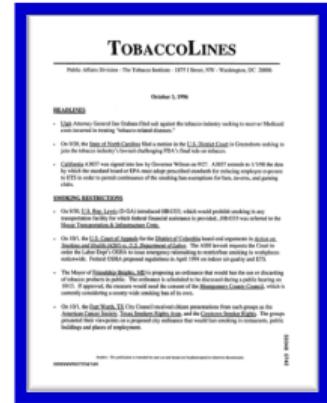
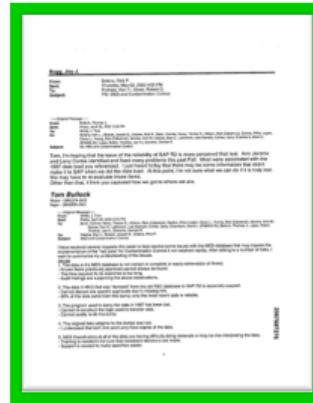
RVL-CDIP

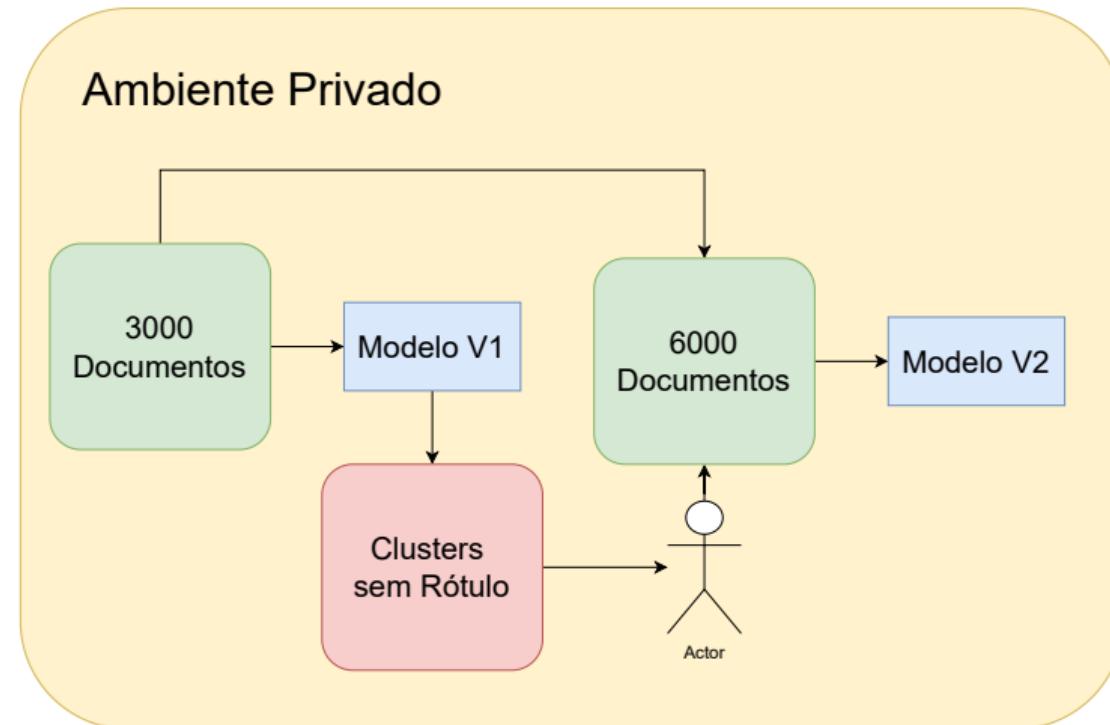
- 400.000 documentos
- 16 classes
- email, formulário, carta...
- Separado por função

Dataset - Motivação

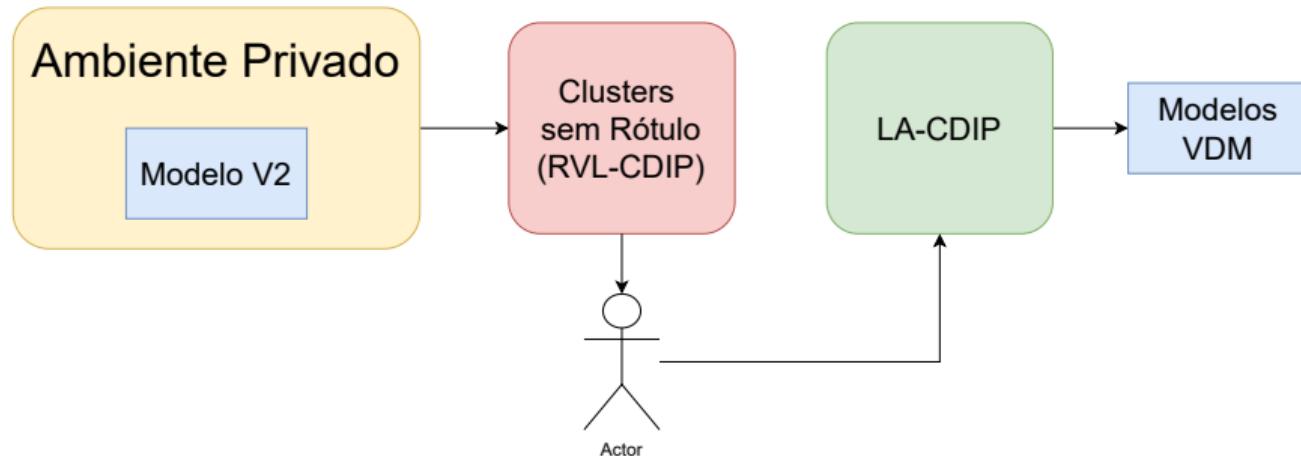


Dataset - Motivação

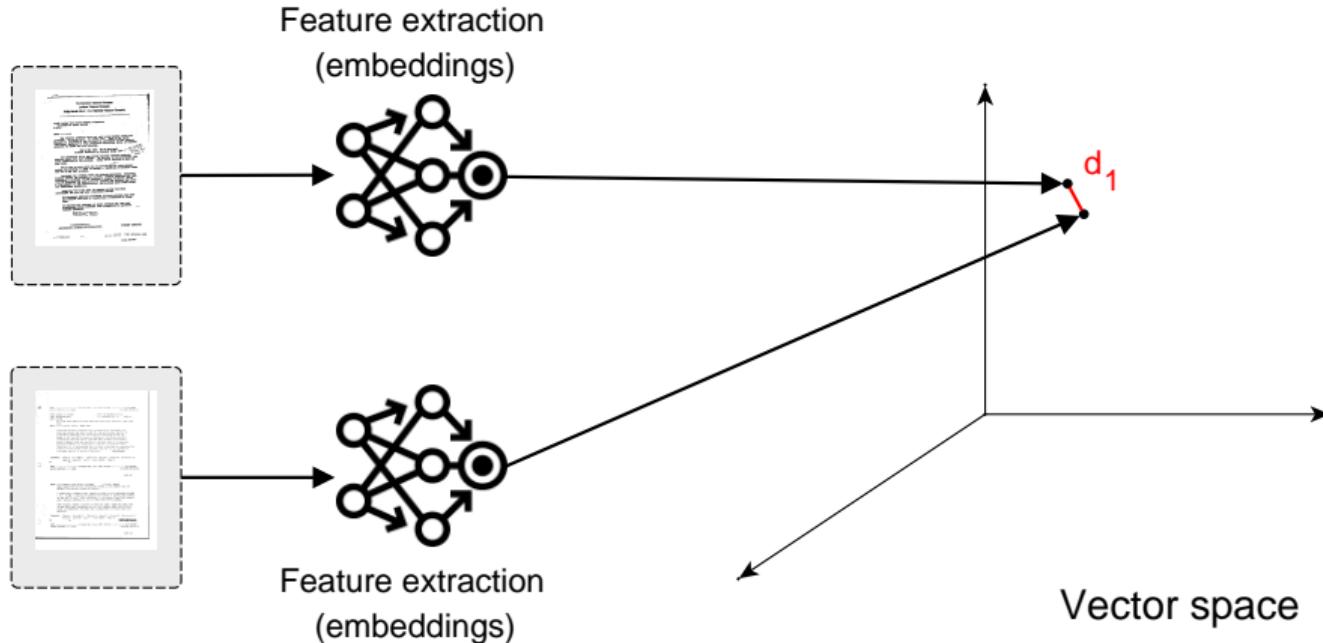




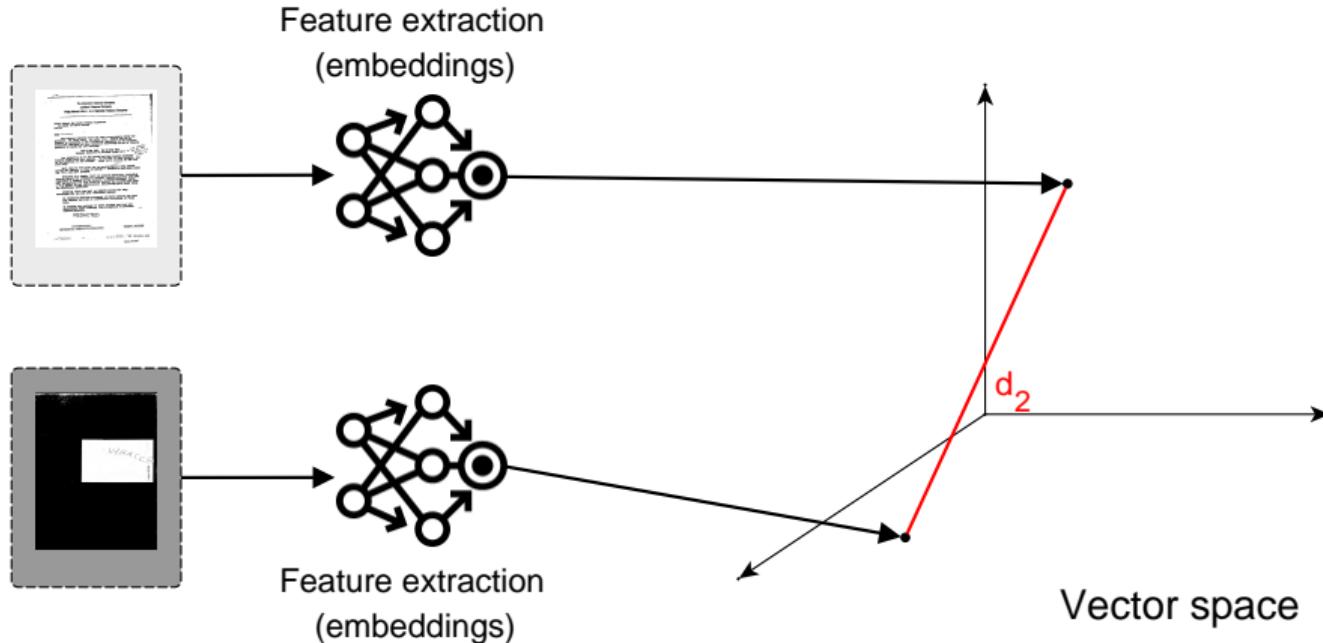
LA-CDIP Dataset - Processo de Rotulação

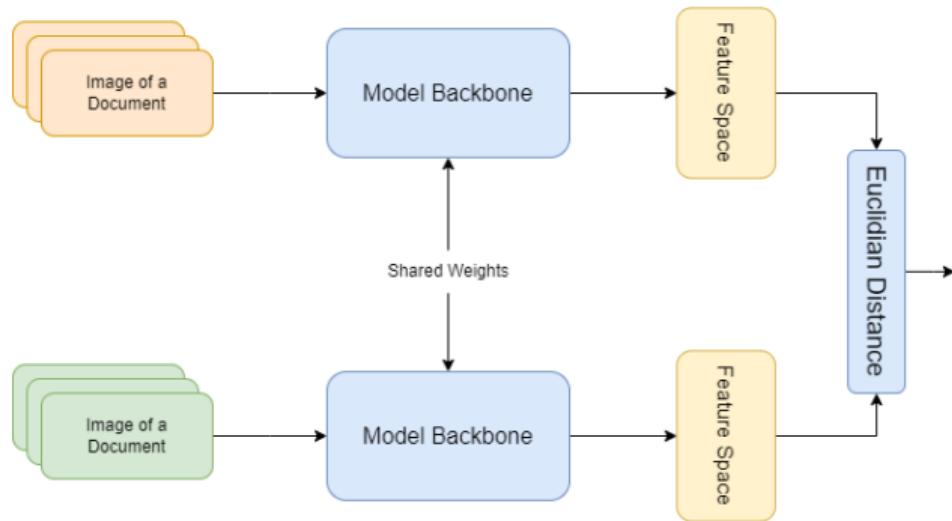


Visual Document Matching - Ilustração



Visual Document Matching - Ilustração





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

Model Backbone

Model Backbone



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

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 - Exemplos

Reference Image

PPP

2000 MARLBORO BAR PROGRAM

CONTRACT TOP SHEET

GMM/SSM: Amy Drnick

MARKET: DALLAS

VENUE NAME: Blackberry

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:

<input checked="" type="checkbox"/> EVENT	<input checked="" type="checkbox"/> VISIBILITY	<input type="checkbox"/> MUSIC	<input type="checkbox"/> RNB
---	--	--------------------------------	------------------------------

Please check the appropriate box regarding venue admission policy:

<input checked="" type="checkbox"/> AC1B-P	<input checked="" type="checkbox"/> AD21-P	<input type="checkbox"/> PAO-P
<input checked="" type="checkbox"/> AD1V-V	<input checked="" type="checkbox"/> AD21-V	

GMM/SSM: [Signature] **DATE:** 2-3-00

SELL-IN APPROVAL: X **DATE:** _____

Similarity Score: 98
Category: Nearly Identical

Image to Compare

2000 MARLBORO BAR PROGRAM							
CONTRACT TOP SHEET							
GM/WSM:	<u>Lindy Jenkins</u>						
MARKET:	<u>Charlotte</u>						
VENUE NAME:	<u>Coach's Sports Bar & Grill</u>						
VENUE ID							
<u>L.H.B.0080-01</u>							
<p>Please check the appropriate box that will identify the type of club and the appropriate contract executed by club owner/manager:</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 25%;"><input checked="" type="checkbox"/> EVENT</td> <td style="width: 25%;"><input type="checkbox"/> VISIBILITY</td> <td style="width: 25%;"><input type="checkbox"/> MUSIC</td> <td style="width: 25%;"><input type="checkbox"/> RNB</td> </tr> </table>		<input checked="" type="checkbox"/> EVENT	<input type="checkbox"/> VISIBILITY	<input type="checkbox"/> MUSIC	<input type="checkbox"/> RNB		
<input checked="" type="checkbox"/> EVENT	<input type="checkbox"/> VISIBILITY	<input type="checkbox"/> MUSIC	<input type="checkbox"/> RNB				
<p>Please check the appropriate box regarding venue admission policy:</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 25%;"><input type="checkbox"/> AO18-P</td> <td style="width: 25%;"><input type="checkbox"/> AO21-P</td> <td style="width: 25%;"><input checked="" type="checkbox"/> PAO-P</td> </tr> <tr> <td style="width: 25%;"><input type="checkbox"/> AO18-V</td> <td style="width: 25%;"><input type="checkbox"/> AO21-V</td> <td style="width: 25%;"></td> </tr> </table>		<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	
<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						
GM/WSM SIGNATURE:	<u>Lindy Jenkins</u>						
	DATE: <u>2-20-00</u>						
SELL-IN APPROVAL:	<u>J</u>						
	DATE: <u>3-13-2000</u>						
The Marlboro International Program for Restaurants							

Reference Image

Similarity Score: 15
Category: Completely Different

Image to Compare

Action T.N. Request

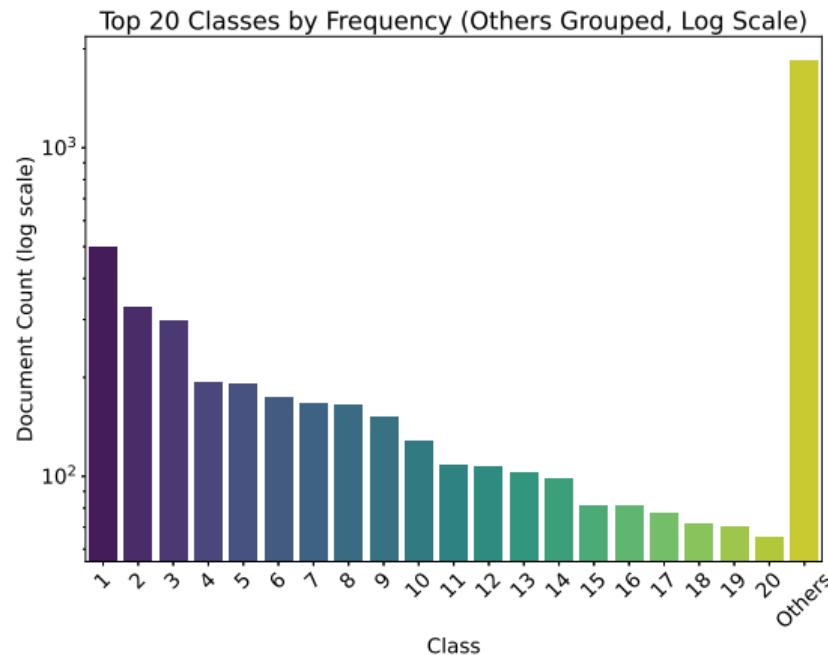
Resultados

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



Equal Error Rate (EER)

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

$$FAR(\tau) = \frac{\text{False Acceptances}}{\text{Total Negatives}}$$

$$FRR(\tau) = \frac{\text{False Rejections}}{\text{Total Positives}}$$

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 - 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

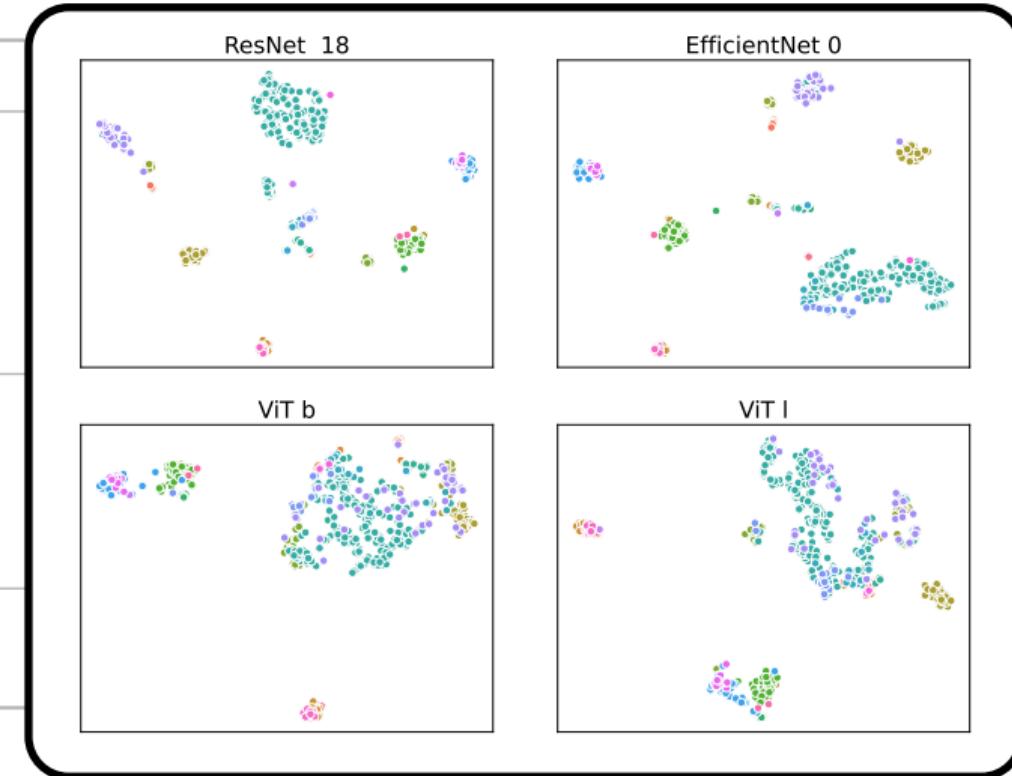
Resultados - Visualização TSNE

Architecture

ResNet

EfficientNet

ViT



Test GZSL

1.51
1.53
2.21
1.98
2.39

0.95
2.70
2.14
2.34

5.19
5.26

Resultados - Large Language Models

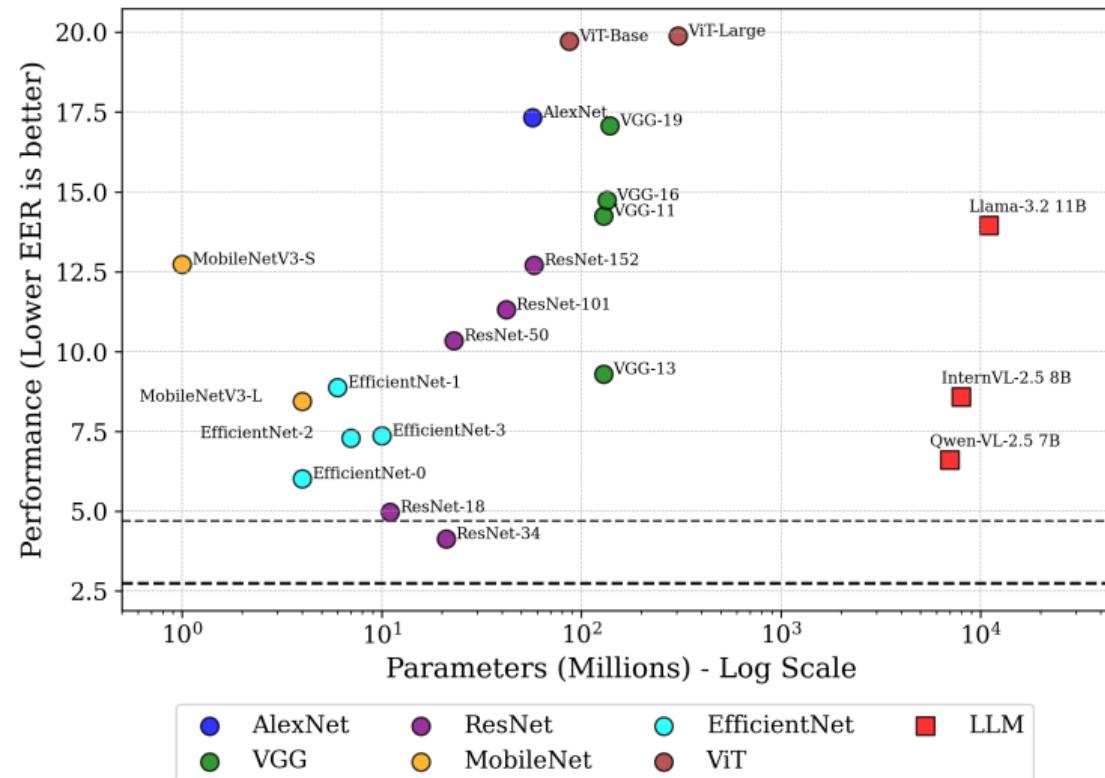
Model	Version	Params	Test ZSL	Test GZSL
Llama	3.2	11B	13.95	21.90
InternVL	2.5	8B	8.58	10.40
Qwen-VL	2.5	7B	6.61	4.20
GPT-4o mini	2024-07-18	*	4.70	4.07
GPT-4o	2024-11-20	*	2.75	1.33

* Parameter count not publicly disclosed

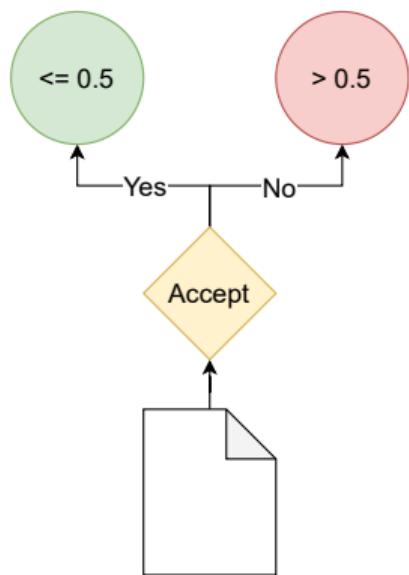
Resultados - Melhores Modelos Visuais vs LLMs

Architecture	Edition	Params	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
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

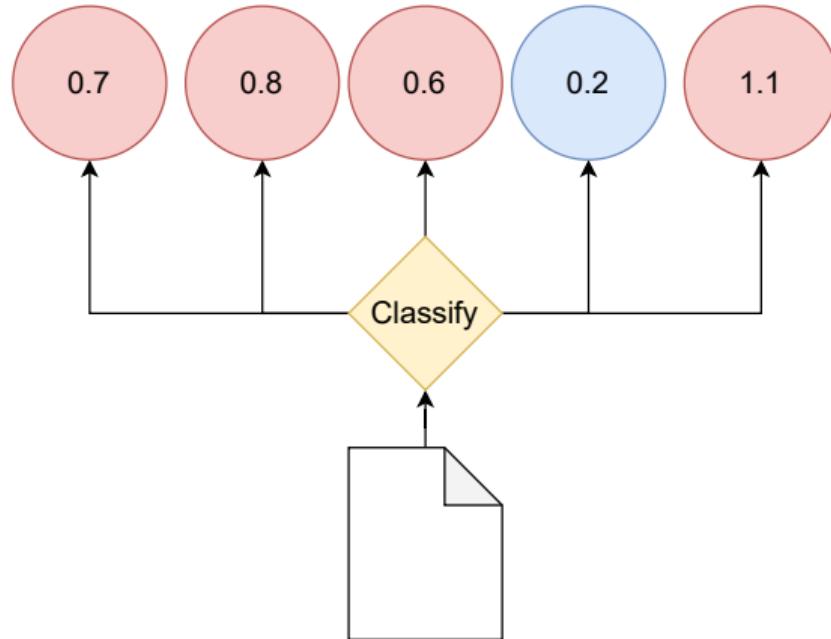
Comparação: Visual Models vs LLMs



Comparação: Visual Models vs LLMs



Sistema de Verificação



Sistema de Identificação

Conclusão

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

Limitações Conhecidas:

- Dataset relativamente pequeno
- Complexidade de arquiteturas limitada
- Apenas documentos do RVL-CDIP

Trabalhos Futuros:

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

Obrigado!

Lucas de Almeida Bandeira Macedo

lucasabmacedo@hotmail.com

Orientador: Prof. Dr. Pedro Garcia Freitas

Coorientador: Prof. Dr. Bruno Luiggi Macchiavello Espinoza