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

Dataset - Motivação

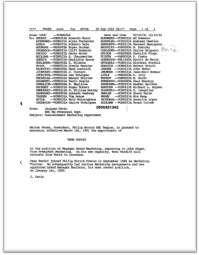
Datasets Disponíveis

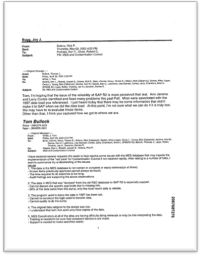
- PubLayNet, DocLayNet
- DocVQA
- CORD, SROIE
- RVL-CDIP

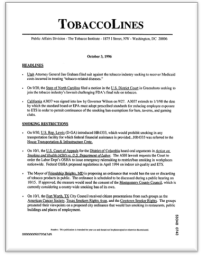
RVL-CDIP

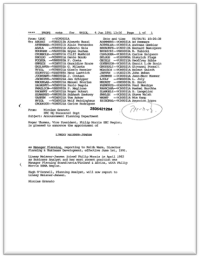
- 400.000 documentos
- 16 classes
- email, formulário, carta...
- Separado por função

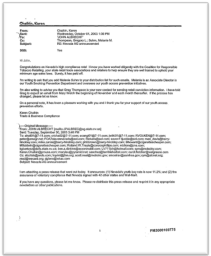
Dataset - Motivação

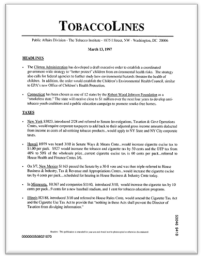




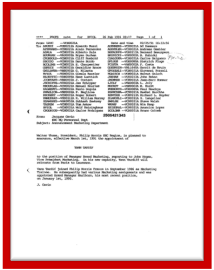


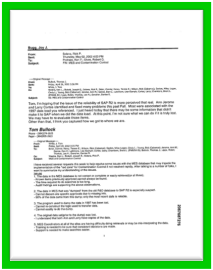


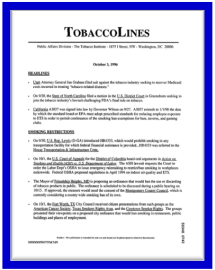


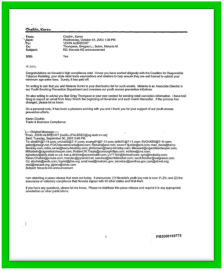


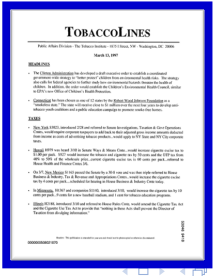
Dataset - Motivação



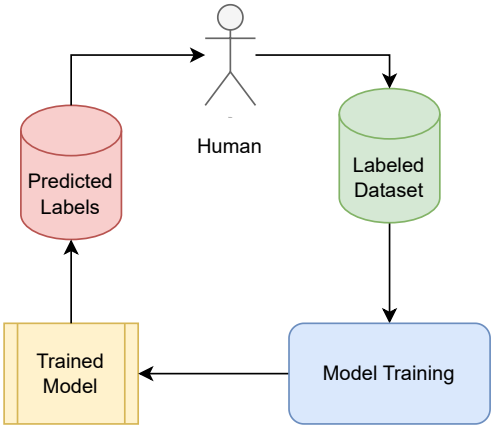




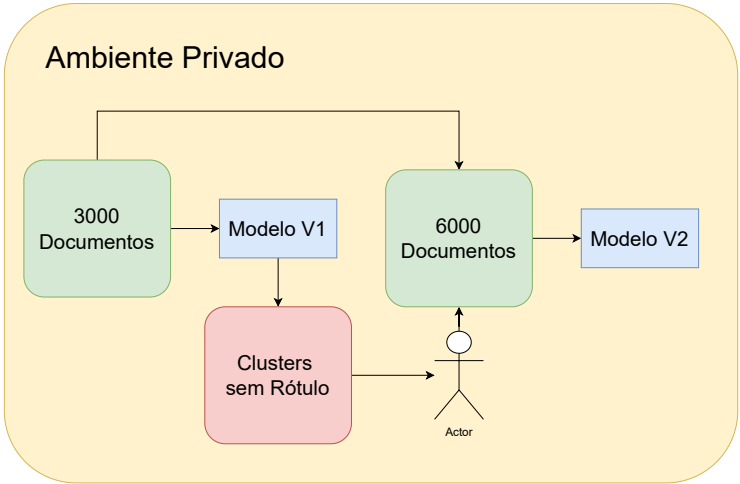




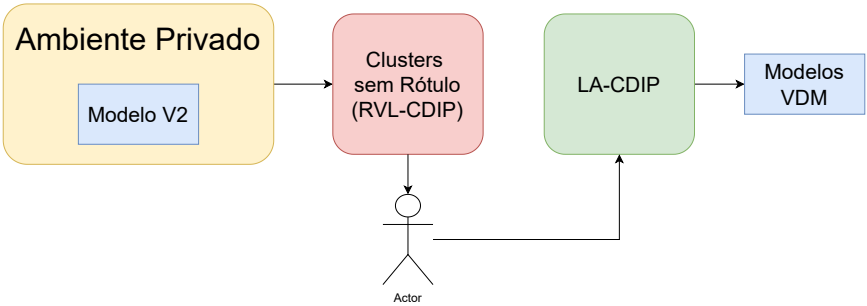
LA-CDIP Dataset - Active Learning



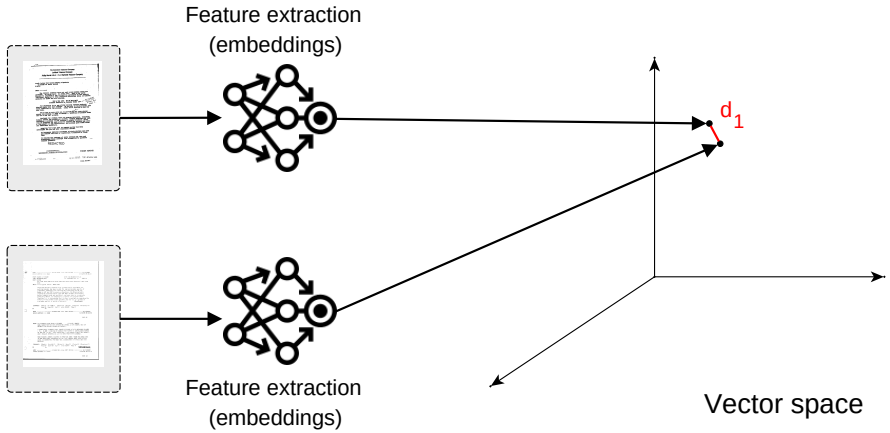
LA-CDIP Dataset - Rotulação



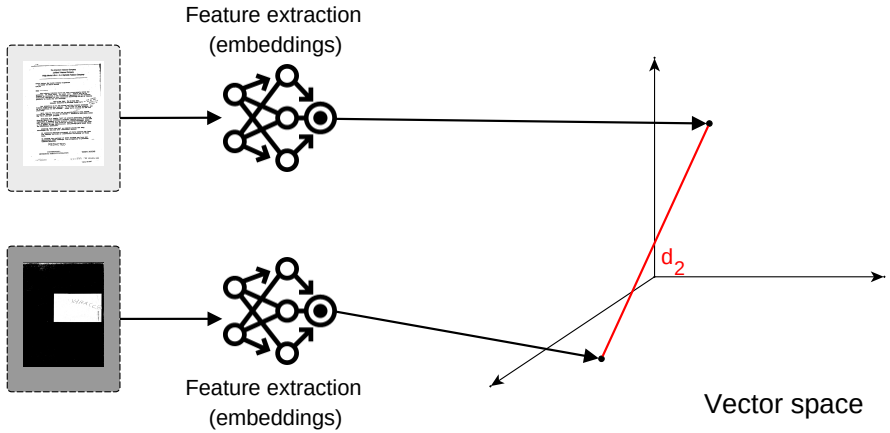
LA-CDIP Dataset - Rotulação



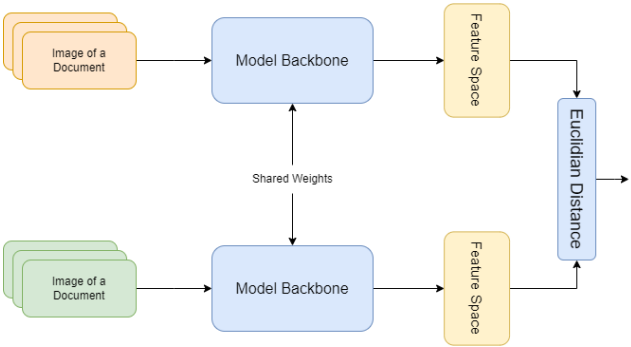
Visual Document Matching



Visual Document Matching

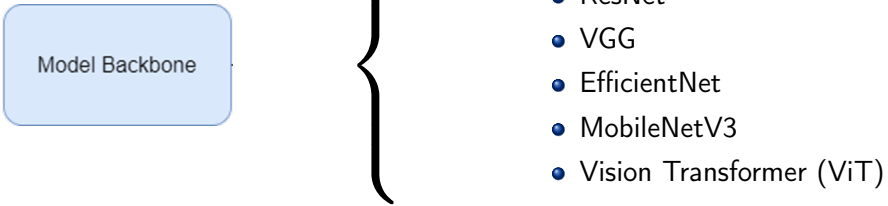


Visual Document Matching - Arquitetura



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

Visual Document Matching - Backbone



Benchmarking com LLMs

Modelos Avaliados:

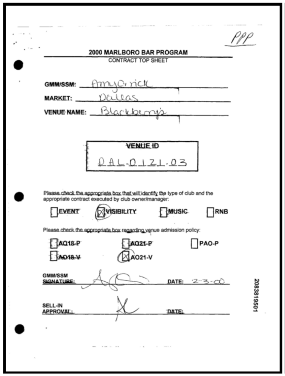
- LLaVA 3.2 Vision
- InternVL 2.5
- Qwen2.5-VL
- GPT-4o (2024-11-20)
- GPT-4o-mini (2024-07-18)

Avaliação:

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

Benchmarking com LLMs

Reference Image



2000 MARLBORO BAR PROGRAM
CONTRACT TOP SHEET

GM/MSSM: Dimmick
MARKET: Dallas
VENUE NAME: Blackhawks

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:

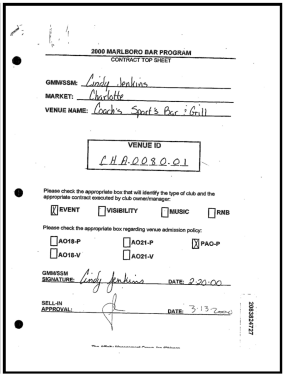
EVENT [X] VISIBILITY MUSIC RNB

Please check the appropriate box regarding venue admission policy:

AO18-P AO21-P PAO-P
AO18-V [X] AO21-V

GM/MSSM SIGNATURE: DATE: 2-3-00
SELL-IN APPROVAL: DATE:

Image to Compare



2000 MARLBORO BAR PROGRAM
CONTRACT TOP SHEET

GM/MSSM: Cindy Jenkins
MARKET: Charlotte
VENUE NAME: Coach's Sports Bar 'n Grill

VENUE ID
CHA-0080-01

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

[X] EVENT VISIBILITY MUSIC RNB

Please check the appropriate box regarding venue admission policy:

AO18-P AO21-P [X] PAO-P
AO18-V AO21-V

GM/MSSM SIGNATURE: Cindy Jenkins DATE: 2-20-00
SELL-IN APPROVAL: DATE: 3-13-2000

Similarity Score: 98
Category: Nearly Identical

Reference Image

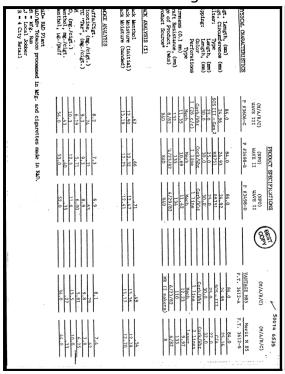
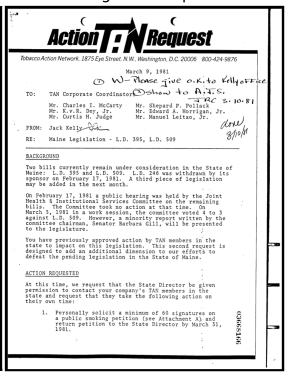


Image to Compare



Action TAN Request

Tobacco Action Network, 1875 Eye Street, N.W., Washington, D.C. 20006 800-424-5876

March 9, 1981

TO: TAN Corporate Coordinators

Mr. Charles I. McCarty
Mr. E.V.R. Dey, Jr.
Mr. Curtis H. Judge
Mr. Shepard P. Pollack
Mr. Edward A. Horrigan, Jr.
Mr. Manuel Latter, Jr.

FROM: Jack Kelly

RE: Maine Legislation - L.D. 395, L.D. 589

BACKGROUND

Two bills currently remain under consideration in the State of Maine: L.D. 395 and L.D. 589. L.D. 244 was withdrawn by its sponsor on February 17, 1981. A third piece of legislation may be added in the next month.

On February 17, 1981 a public hearing was held by the Joint Health & Institutional Services Committee on the remaining two bills. The committee took action on these bills on March 5, 1981 in a vote which resulted in a recommendation of 8 to 3 against L.D. 589. However, a minority report written by the committee chairman, Senator Barbara Gill, will be presented to the legislature.

You have previously approved action by TAN members in the state to impact on this legislation. This second request is designed to add an additional dimension to our efforts to defeat the pending legislation in the State of Maine.

ACTION REQUESTED

At this time, we request that the State Director be given permission to contact your company's TAN members in the state and request that they take the following action on their own time:

1. Personally solicit a minimum of 60 signatures on a public smoking petition (see Attachment A) and return petition to the State Director by March 31, 1981.

Similarity Score: 15
Category: Completely Different

Resultados

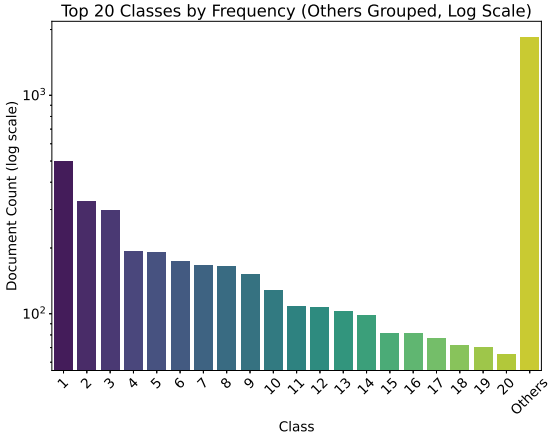
LA-CDIP Dataset - Estatísticas

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 - ResNet e EfficientNet

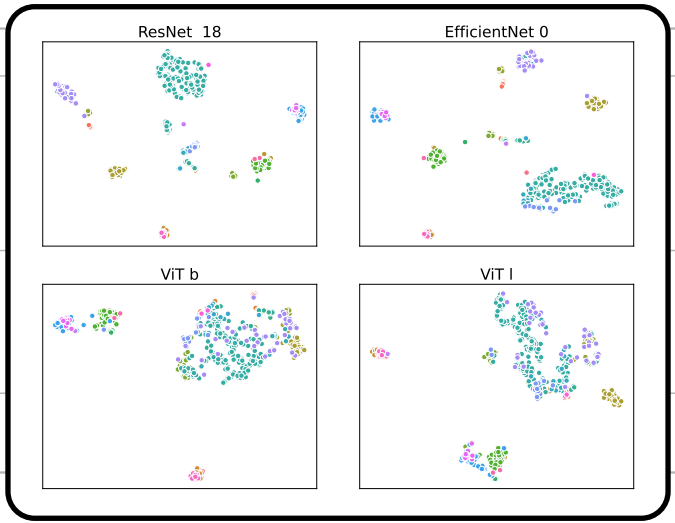
Architecture	Edition	Params	ZSL	GZSL	Test ZSL	Test GZSL
ResNet	18	11M	5.03	**1.54**	4.98	1.51
	34	21M	4.32	2.10	**4.13**	1.53
	50	23M	6.90	3.39	10.34	2.21
	101	42M	8.20	2.72	11.31	1.98
	152	58M	9.44	3.38	12.70	2.39
EfficientNet	0	4M	4.41	2.27	6.02	**0.95**
	1	6M	**3.93**	3.54	8.88	2.70
	2	7M	5.73	2.61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

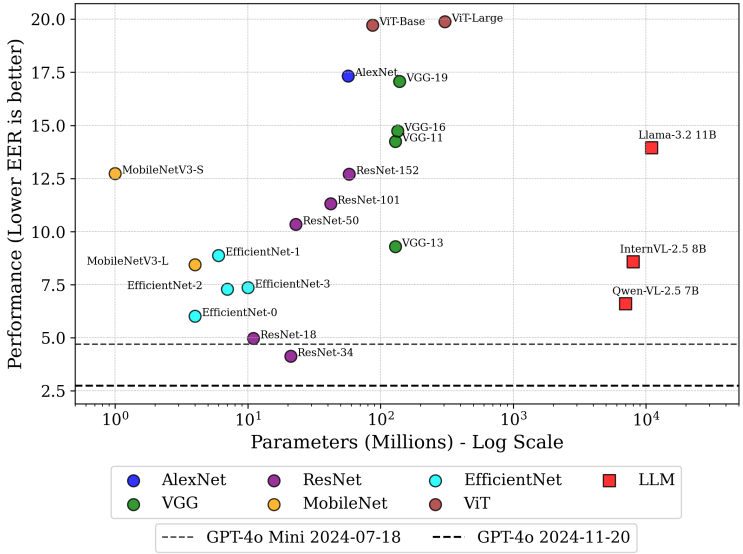
Model	Version	Params	Test ZSL	Test GZSL
Llama	3.2	11B	13.95	21.90
InternVL	2.5	8B	8.58	10.40
Qwen-VL	2.5	7B	6.61	4.20
GPT-4o mini	2024-07-18	*	4.70	4.07
GPT-4o	2024-11-20	*	**2.75**	**1.33**

* Parameter count not publicly disclosed

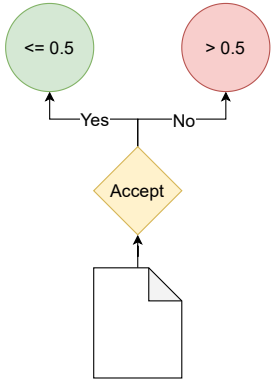
Resultados - Melhores Modelos Visuais vs LLMs

Architecture	Edition	Params	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

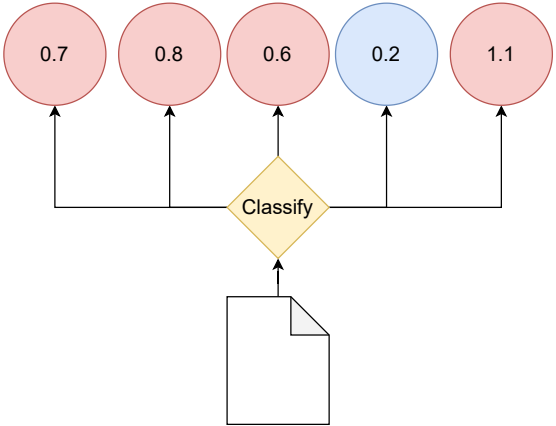
Comparação: Visual Models vs LLMs



Comparação: Visual Models vs LLMs



Sistema de Verificação



Sistema de Identificação

Conclusão

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

Limitações e Trabalhos Futuros

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