

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

Contexto - Documentos e Compliance

- Documentos físicos
- Imagens de documentos
- Exemplos:

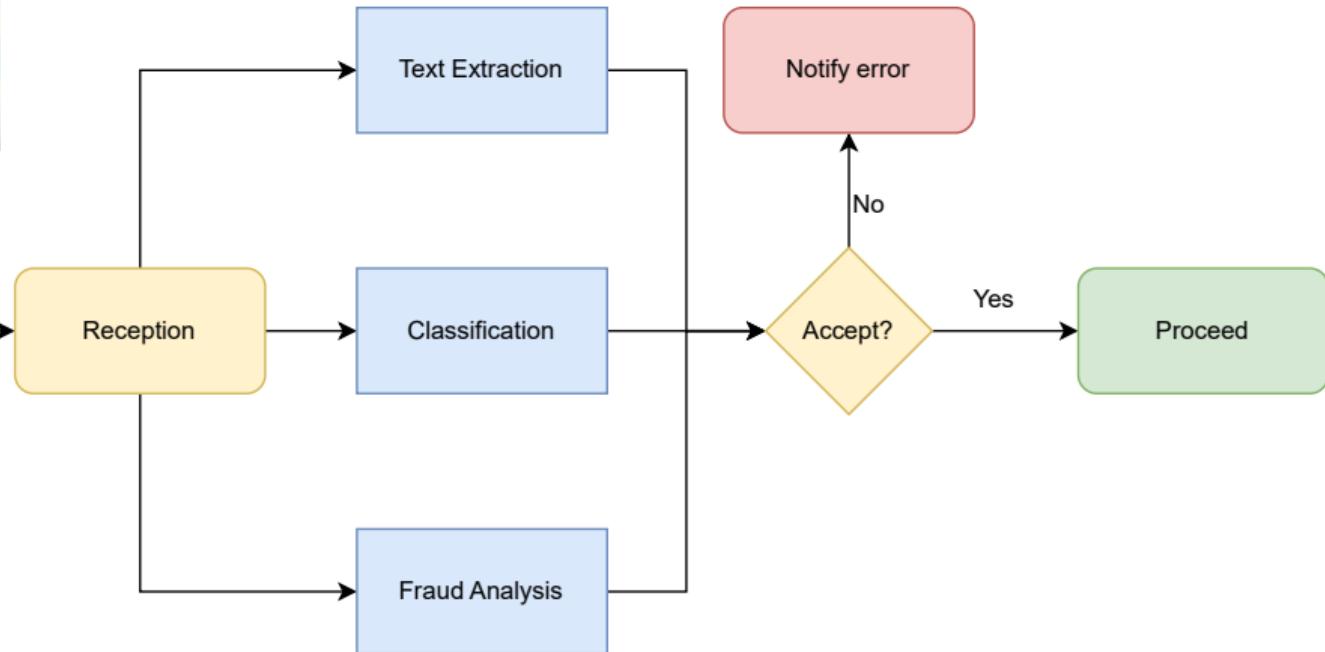


MINISTÉRIO DA FAZENDA SECRETARIA DA RECEITA FEDERAL DO BRASIL IMPOSTO SOBRE A RENDA - PESSOA FÍSICA EXERCÍCIO 2017 ANO-CALENDÁRIO 2016			
RECEBO DE ENTREGA DA DECLARAÇÃO DE AJUSTE ANUAL - OPÇÃO PELO DESCONTO SIMPLIFICADO DECLARAÇÃO ORIGINAL			
IDENTIFICAÇÃO DO DECLARANTE			
CPF do declarante 332.653.378-08	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 06199-120	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

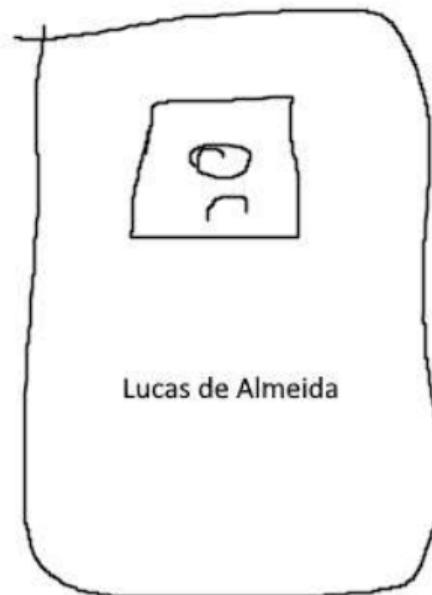


Contexto - Importância da Classificação da Imagem

- 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

LA-CDIP Dataset - Construção

HAZLETON LABORATORIES AMERICA, INC.			
PRODUCT DESIGNATION	B39		
MATERIAL SAFETY DATA SHEET			
Section 1. Source & Nomenclature Section			
EMPLOYER'S NAME	HAZLETON INC.		
ADDRESS/TELEPHONE NO.	15150 275-4455/6662		
SHIPPING ADDRESS	15150 275-4455/6662		
WHOLESALE DISTRIBUTOR	HAZLETON INC., 15150 275-4455/6662		
EMERGENCY NAME & PHONE	CHEMICAL OIL EXTRACT FORTWELL		
Section 2. Hazardous Ingredients			
BASIC MATERIAL	AMOUNT IN % OF TOTAL	REPLACEMENT NAME & DESCRIPTION	USE
			CRATE
			PERCENT
			WEIGHT
			PPM

Framework de Active Learning

Etapa 1: Clustering Preliminar

- Modelo privado de metric learning
- Hierarchical Agglomerative Clustering com estratégia de Ward
- Número dinâmico de clusters

Etapa 2: Reorganização Manual

- Limpeza dos clusters (um padrão por cluster)
- Merge de clusters com padrões similares
- Validação independente para consistência intra e inter-classe

Abordagem:

- Siamese Networks
- Metric Learning
- Mapeamento para espaço de features

Backbones Testados:

- ResNet (18, 34, 50, 101, 152)
- EfficientNet (B0-B3)
- MobileNetV3 (Small, Large)
- VGG (11, 13, 16, 19)
- Vision Transformer (Base, Large)

Princípio:

Documentos similares



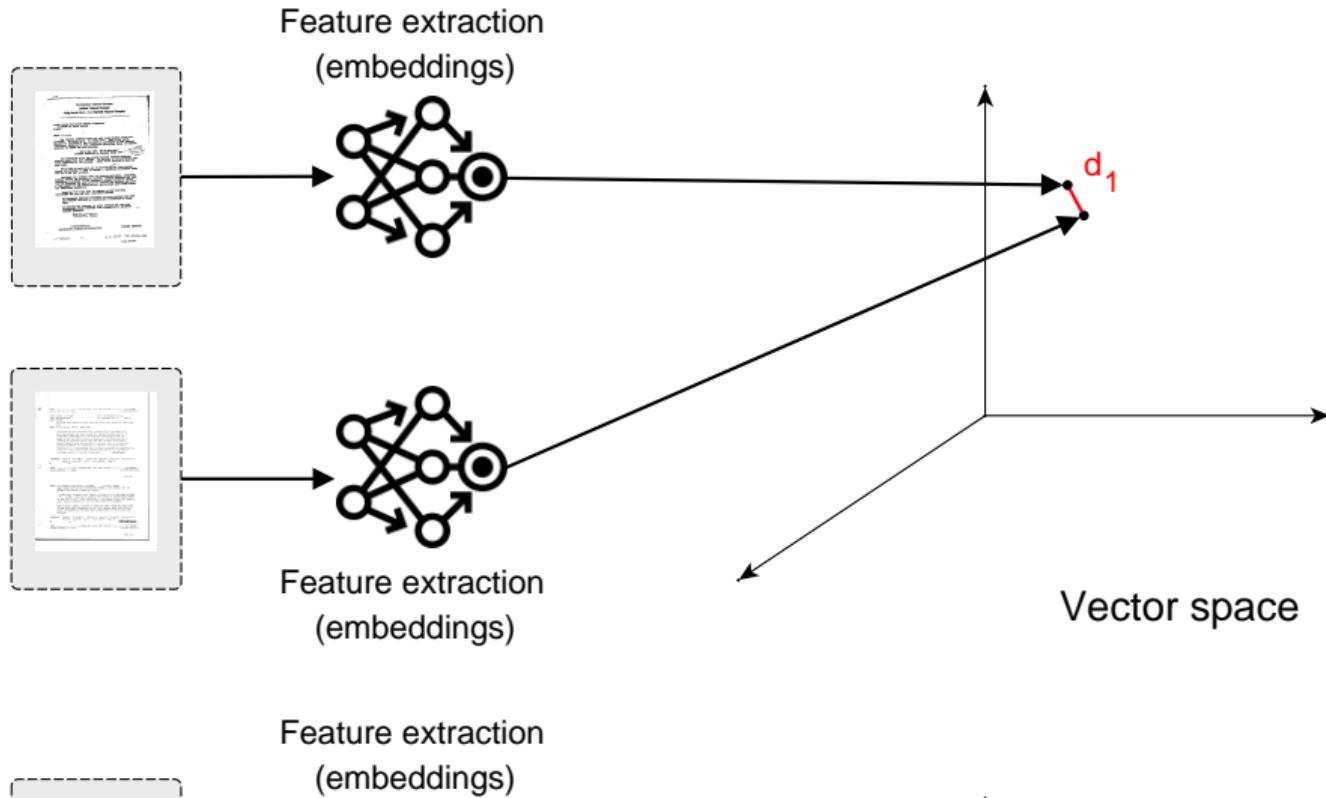
Pequena distância
no espaço de features

Documentos diferentes



Grande distância
no espaço de features

Visual Document Matching - Ilustração



Configuração

- **Input:** Matriz RGB da imagem do documento
- **Pré-processamento:**
 - Resize para (224, 224) para maioria dos modelos
 - Normalização de 0–255 para 0–1
 - Normalização com média e desvio padrão do split de treino
- **Treinamento:**
 - Formação de pares aleatórios
 - Mesma probabilidade para todas as classes
 - Contrastive Loss
 - Sem data augmentation ou pair mining

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)

Lucas de Almeida Bandeira Macedo (UnB)

Reference Image

Image to Compare

Similarity Score: 98
Category: Nearly Identical

Reference Image

Image to Compare

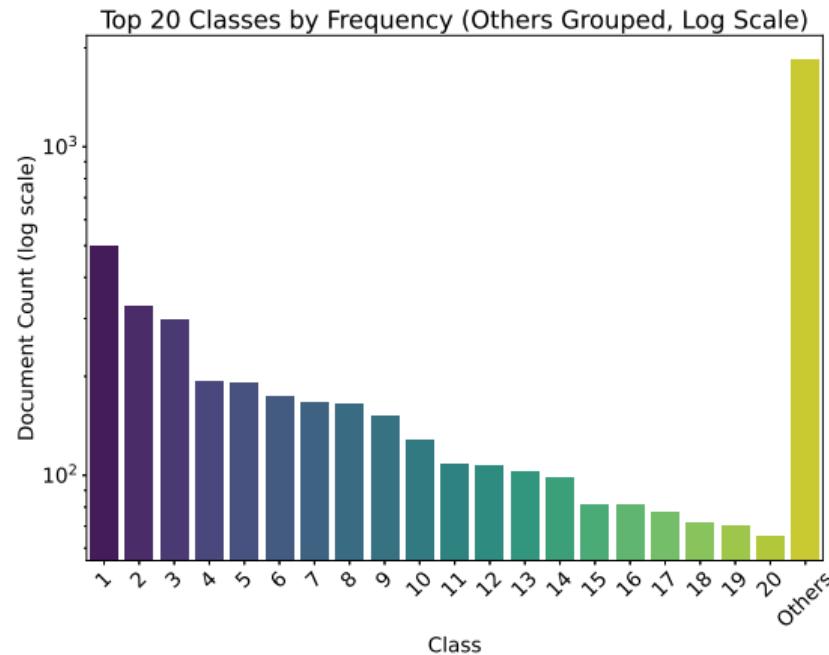
March 9, 1981
W-W- Please give o.k. to Kelly office
TO: TAN Corporate Coordinator Show to A.T.N.
Mr. Charles L. McCarty Mr. Bernard P. Pollard

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 CV	GZSL CV	Test ZSL	Test GZSL
ResNet	18	11M	5.03	1.54	4.98	1.51
ResNet	34	21M	4.32	2.10	4.13	1.53
EfficientNet	0	4M	4.41	2.27	6.02	0.95
EfficientNet	1	6M	3.93	3.54	8.88	2.70
VGG	13	129M	7.03	4.79	9.30	3.95
ViT	Base	87M	12.43	7.97	19.72	5.19
GPT-4o	2024-11-20	*	—	—	2.75	1.33
GPT-4o mini	2024-07-18	*	—	—	4.70	4.07
Qwen-VL	2.5	7B	—	—	6.61	4.20
InternVL	2.5	8B	—	—	8.58	10.40
LLaMA	3.2	11B	—	—	13.95	21.90

Table: *

Mean EER (%) - Valores menores são melhores

Principais Observações

- **Arquiteturas menores superaram as maiores**
 - ResNet-18 e ResNet-34: melhor desempenho
 - ResNet-50, 101, 152: desempenho progressivamente pior
- **Vision Transformer (ViT): pior desempenho**
 - Overfitting nos dados de treino
 - 0% de erro de treino, alto erro de validação
 - Causa provável: dataset pequeno (4.993 documentos)
- **Melhores modelos:**
 - ResNet-34, EfficientNet-0 e EfficientNet-1
 - Balanço entre tamanho e generalização

GZSL consistentemente melhor:

- 50% das classes vistas no treino
- Maior diferença em modelos grandes
- Overfitting mitigado no cenário mais fácil

Test split ZSL mais desafiador:

- Alta variância entre splits
- Padrões completamente diferentes
- Taxas de erro mais altas

Exemplo ResNet-18:

- ZSL CV: 5.03%
- GZSL CV: 1.54%
- Test ZSL: 4.98%
- Test GZSL: 1.51%

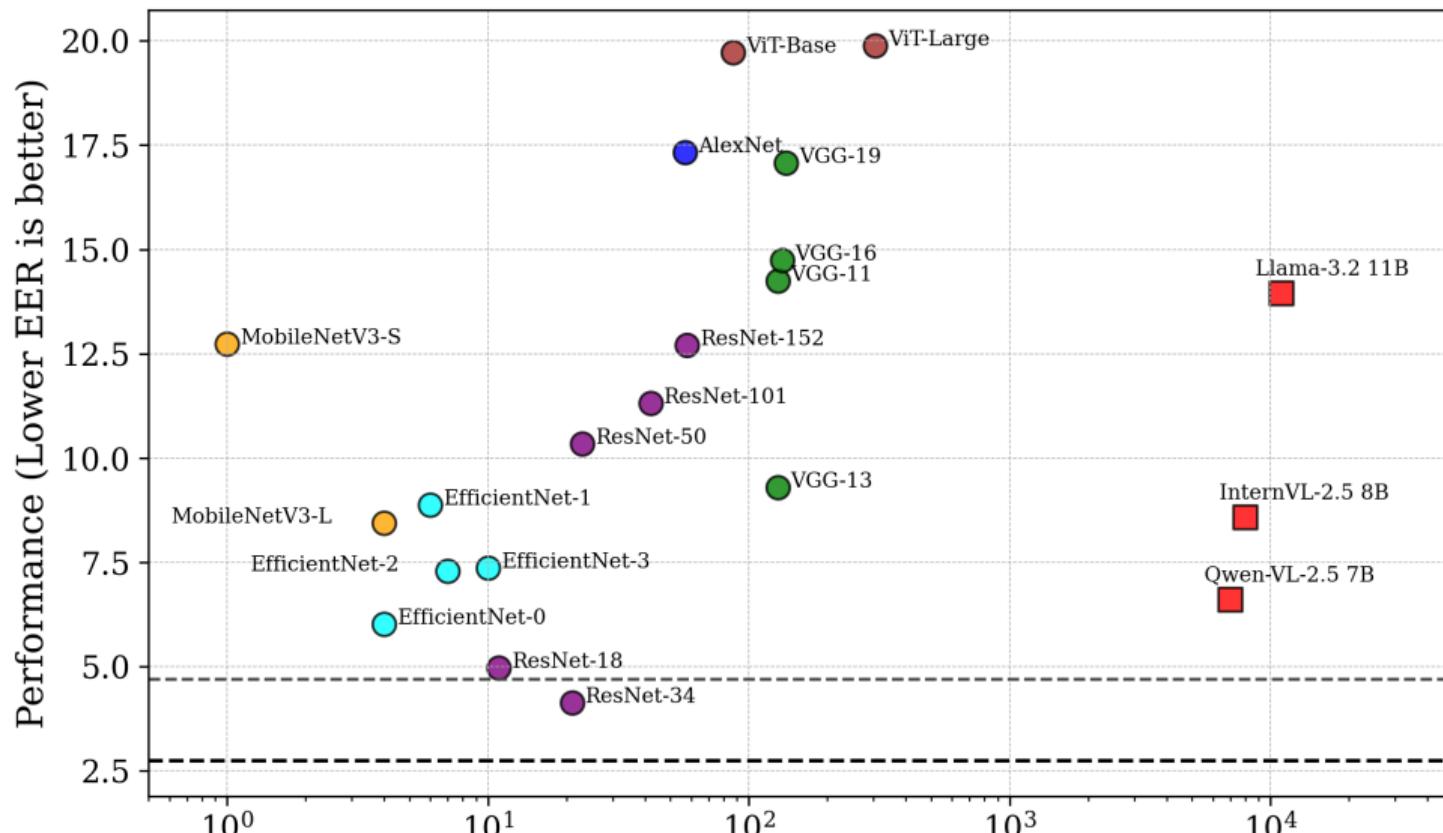
Exemplo ViT-Base:

- ZSL CV: 12.43%
- GZSL CV: 7.97%
- Test ZSL: 19.72%
- Test GZSL: 5.19%

Observações

- **GPT-4o: melhor desempenho geral**
 - ZSL: 2.75% | GZSL: 1.33%
 - Porém: centenas de vezes mais caro por inferência
- **Alternativas open-source promissoras:**
 - Qwen-VL 7B: desempenho próximo ao GPT-4o mini
 - InternVL 2.5: competitivo
- **LLaMA 3.2 Vision:**
 - Limitações no manuseio de múltiplas imagens
 - Necessário combinar imagens em input único
- **Tendências diferentes entre cenários:**
 - InternVL melhor no ZSL
 - GPT melhor no GZSL

Comparação: Visual Models vs LLMs



Sistema de Verificação (Classificação Binária)

- Dado: documento de referência + threshold
- Decisão: aceitar se distância $<$ threshold, rejeitar caso contrário

Sistema de Identificação (Classificação Multi-classe)

- Dado: referências para cada classe + threshold
- Decisão: atribuir à classe mais próxima ou rejeitar se fora do threshold

Design Choices:

- Múltiplas referências por classe (usar centroid)
- Thresholds diferentes por classe
- Armazenar embeddings ao invés de imagens
- Complexidade: $O(r)$ verificação, $O(rc)$ identificação

Principais Contribuições

① LA-CDIP Dataset

- Dataset categorizado exclusivamente por layout
- Alternativa Zero-Shot para classificação de documentos

② Benchmarking Sistemático

- Diversos backbones visuais estabelecidos
- Comparação com LLMs populares

③ Resultados Práticos

- Modelos visuais menores superam LLMs (exceto GPT-4o)
- GPT-4o melhor, mas custo-benefício desfavorável
- Alternativas open-source viáveis

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
- Data augmentation
- Permitir modelos mais complexos

Timeline da Pesquisa

Tarefa	5º	6º	7º	8º	9º–10º
Literatura	✓				
Framework	✓	✓			
Dataset Labeling		✓			
Experimentos		✓	✓		
Produção de Paper				✓	✓
Extra Labeling					✓
Qualificação					8º trim
Defesa de Mestrado					10º trim

Status atual: **8º trimestre (Qualificação)**

Obrigado!

Perguntas?

Lucas de Almeida Bandeira Macedo

lucasabmacedo@hotmail.com

Orientador: Prof. Dr. Pedro Garcia Freitas

Coorientador: Prof. Dr. Bruno Luiggi Macchiavello Espinoza