UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Classificação de Produtos de Supermercado: Uma Abordagem Zero-Shot

**Jonathan Freire da Silva**

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

**Jonathan Freire da Silva**

# Classificação de Produtos de Supermercado: Uma Abordagem Zero-Shot

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. João Paulo Papa

**Versão original**

**São Carlos**

**2024**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

| | |
|---|---|
| S856m | Silva, Jonathan Freire<br>Classificação de Produtos de Supermercado: Uma Abordagem Zero-Shot / Jonathan Freire da Silva ; orientador João Paulo Papa. – São Carlos, 2024.<br>50 p. : il. (algumas color.) ; 30 cm.<br><br>Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.<br><br>1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Papa, João Paulo, orient. II. Título. |

**Jonathan Freire da Silva**

# Supermarket Product Classification: A Zero-Shot Approach

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. João Paulo Papa

**Original version**

**São Carlos**

**2024**

*Dedico este trabalho aos meus amigos e à minha família, que sempre acreditaram em mim, me incentivaram e me deram todo o suporte e as condições necessárias para que eu pudesse estar aqui hoje;*

*E a todos os professores e tutores que tive ao longo da vida, desde a infância, que foram fundamentais na minha formação, transmitindo não apenas conhecimento, mas também valores essenciais para o meu crescimento.*

# AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, que sempre esteve ao meu lado, me dando força e sabedoria para superar todos os desafios ao longo dessa jornada.

Agradeço aos professores do curso, que com seu conhecimento e dedicação foram fundamentais para a minha formação, contribuindo para o meu crescimento tanto pessoal quanto profissional.

Gostaria de expressar minha gratidão aos meus colegas de pesquisa, Dr. Marcos Cleison Silva Santana e Luiz Fernando Merli de Oliveira Sementille, por sua dedicação, apoio e valiosas contribuições, que foram essenciais para o desenvolvimento deste trabalho. Seus conhecimentos e colaborações fizeram toda a diferença ao longo deste processo.

Agradeço também ao meu orientador, Prof. Dr. João Paulo Papa, por sua orientação e pelas ideais oferecidos durante a condução deste projeto.

*"Sua tarefa é descobrir o seu trabalho*
*e, então, com todo o coração, dedicar-se a ele."*
*Buda*

# RESUMO

Silva, J. F. S. **Classificação de Produtos de Supermercado: Uma Abordagem Zero-Shot**. 2024. 50p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Este trabalho propõe uma abordagem de *Zero-Shot Learning* (ZSL) para a classificação de produtos em supermercados, com o objetivo de mitigar o problema de ruptura interna de prateleiras, que ocorre quando produtos não são adequadamente repostos, resultando em prejuízos para varejistas e consumidores. A solução adotada utiliza três modelos CLIP (*Contrastive Language–Image Pretraining*), disponíveis na plataforma OpenCLIP, treinados em diferentes conjuntos de dados: LAION-2B, DFN-2B e o modelo original do OpenAI CLIP. Esses modelos foram escolhidos devido a sua capacidade de produzir *embeddings* de alta qualidade que permitem uma associação eficiente entre imagens. Além disso, foi empregado o uso da biblioteca FAISS para realizar buscas por similaridade, possibilitando a identificação rápida e eficiente de produtos com base em suas representações visuais.

A eficácia dos modelos foi avaliada em dois conjuntos de dados do setor varejista: Grozi-120 e Grozi-3.2k. Os resultados demonstraram que a abordagem ZSL, combinada com a robustez dos modelos CLIP, possibilitou uma generalização eficaz na classificação de produtos. O modelo DFN-2B se destacou, apresentando um desempenho superior, o que sugere que a qualidade dos dados utilizados no treinamento pode influenciar positivamente os resultados.

Apesar dos resultados promissores, identificou-se que ainda há espaço para melhorias, especialmente na classificação de produtos visualmente semelhantes, onde o uso de técnicas de *fine-tuning* pode melhorar a precisão. Este estudo demonstra que o ZSL, aliado aos modelos CLIP, pode ser uma solução promissora para a classificação de produtos em supermercados, eliminando a necessidade de treinamentos contínuos e ajustes frequentes nos modelos.

**Palavras-chave**: *Zero-Shot Learning*, CLIP, Classificação de Produtos, Supermercados, Visão Computacional, *Embeddings*, FAISS

## ABSTRACT

Silva, J. F. S. **Supermarket Product Classification: A Zero-Shot Approach**. 2024. 50p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This work proposes a Zero-Shot Learning (ZSL) approach for product classification in supermarkets, aiming to mitigate the issue of internal shelf outages, which occurs when products are not adequately restocked, resulting in losses for retailers and consumers. The solution adopted utilizes three CLIP models (Contrastive Language–Image Pretraining), available on the OpenCLIP platform, trained on different datasets: LAION-2B, DFN-2B, and the original OpenAI CLIP model. These models were chosen for their ability to produce high-quality embeddings that allow for efficient associations between images. Additionally, the FAISS library was employed to perform similarity searches, enabling the rapid and efficient identification of products based on their visual representations.

The effectiveness of the models was evaluated on two retail sector datasets: Grozi-120 and Grozi-3.2k. The results demonstrated that the ZSL approach, combined with the robustness of the CLIP models, enabled effective generalization in product classification. The DFN-2B model stood out, showing superior performance, which suggests that the quality of the data used in training can positively influence the results.

Despite the promising results, it was identified that there is still room for improvement, especially in the classification of visually similar products, where the use of Fine-Tuning techniques may enhance accuracy. This study demonstrates that ZSL, combined with CLIP models, can be a promising solution for product classification in supermarkets, eliminating the need for continuous training and frequent adjustments to the models.



**Keywords**: Zero-Shot Learning, CLIP, Product Classification, Supermarkets, Computer Vision, Embeddings, FAISS

## LISTA DE FIGURAS

## LISTA DE TABELAS

# LISTA DE QUADROS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| IA | Inteligência Artificial |
| AM | Aprendizado de Máquinas |
| ZSL | Zero-Shot Learning |
| ViT | Vision Transformers |
| CLIP | Contrastive Language-Image Pre-Training |
| CNN | Convolutional Neural Network |
| FAISS | Facebook AI Similarity Search |
| GPU | Graphics Processing Unit |
| CPU | Central Processing Unit |

## SUMÁRIO

# 1 INTRODUÇÃO

A rápida evolução das tecnologias de inteligência artificial (IA) e aprendizado de máquina (AM) tem impactado profundamente diversos setores, transformando a forma como os dados são analisados, decisões são tomadas e inovações são implementadas. No varejo, especialmente em supermercados, o uso dessas tecnologias tem trazido benefícios significativos, como a previsão mais precisa da demanda, otimização de estoques, personalização de ofertas para clientes, detecção de fraudes e redução de desperdícios (GUHA *et al.*, 2021)(SHANKAR, 2018). Além disso, as soluções de IA e AM têm melhorado substancialmente a experiência do cliente ao oferecer recomendações personalizadas em tempo real e automatizar processos operacionais, permitindo que as equipes se concentrem em atividades mais estratégicas. Essa integração não só aumenta a eficiência operacional, mas também proporciona um diferencial competitivo em um mercado cada vez mais dinâmico e orientado a dados.

Mesmo com esse avanços tecnológicos, um dos desafios persistentes no varejo é a ruptura interna — situação em que um produto está indisponível nas prateleiras, mesmo com estoque disponível no armazém. Esse problema acarreta perdas de vendas imediatas, prejudica a fidelidade do cliente e pode impactar negativamente a reputação da marca. Além disso, a ausência de um produto específico pode levar o consumidor a optar por uma marca ou categoria alternativa, desistir da compra ou procurar o item em outro estabelecimento, o que acarreta em perdas econômicas de curto e longo prazo para os supermercados (SON; KANG; JANG, 2019).

Os métodos tradicionais de gestão de estoque, baseados em previsões de demanda e reposição manual, são altamente suscetíveis a erros humanos e ineficiências. A aplicação de IA e AM oferece uma alternativa promissora para a automação da identificação de produtos em falta, otimização da reposição e na redução dessas perdas financeiras. A visão computacional desempenha um papel central nesse processo, automatizando a captura de imagens das prateleiras, a segmentação e classificação dos produtos, entretanto, os desafios associados à grande diversidade de produtos, semelhança visual entre itens e as constantes mudanças no portfólio de produtos tornam o problema mais complexo (WEI *et al.*, 2020).

O treinamento de modelos tradicionais de aprendizado de máquina para classificar produtos de supermercado requer grandes volumes de dados de treinamento, o que é difícil, custoso e demorado. Além disso, esses modelos enfrentam dificuldades de generalização quando novos produtos são introduzidos, o que exige processos frequentes de *fine-tuning*, aumentando os custos computacionais e o risco de *overfitting* dos modelos. Além disso, durante esses processos de *fine tuning*, os novos produtos permanecem não reconhecidos, criando lacunas temporárias na capacidade de identificação do sistema (SRIVASTAVA,

2024).

Uma solução promissora para superar essas limitações é o *Zero-Shot Learning* (ZSL), técnica em que o modelo pode identificar e classificar itens não vistos durante o seu treinamento. Isso é particularmente útil em cenários como os de supermercados, onde novos produtos são introduzidos continuamente, exigindo que o modelo tenha alta capacidade de generalização. O ZSL reduz a necessidade de novos dados de treinamento e elimina a dependência de *fine-tuning* constante, oferecendo uma solução mais ágil e escalável para esse cenário.

Neste trabalho, propõe-se a utilização do ZSL com o modelo CLIP *(Contrastive Language–Image Pretraining)* (RADFORD *et al.*, 2021) obtidos através do OpenCLIP (ILHARCO *et al.*, 2021). O CLIP se destaca por associar imagens e textos sem a necessidade de exemplos rotulados específicos para cada tarefa, enquanto o OpenCLIP oferece uma implementação aberta e otimizada, com uma alta quantidade de modelos pré-treinados em diferentes *datasets*. A combinação de ZSL com CLIP proporciona uma abordagem promissora para a classificação de produtos em supermercados, mitigando os desafios impostos pela diversidade e constante atualização do catálogo de produtos.

## 1.1 Hipótese de Pesquisa

A hipótese central deste trabalho é que a aplicação de técnicas de ZSL, utilizando modelos baseados na arquitetura CLIP, pode auxiliar significativamente o processo e a eficiência na classificação de produtos em supermercados. Ao eliminar a necessidade de *fine-tuning* constante, o uso de ZSL permite uma redução dos custos computacionais e oferece uma resposta mais rápida e eficaz às constantes mudanças no catálogo de produtos. Dessa forma, a aplicação dessas técnicas pode auxiliar no problema da ruptura interna, melhorando a disponibilidade dos produtos nas prateleiras e minimizando os impactos negativos na experiência do cliente e nos resultados financeiros dos supermercados.

## 1.2 Objetivos Gerais e Específicos

O objetivo geral deste trabalho é avaliar a viabilidade do uso de ZSL, com base no modelo CLIP, para a classificação de produtos em supermercados, visando uma solução eficiente e adaptável que não dependa de treinamento ou *fine tuning* constante.

Os objetivos específicos incluem: Implementar o modelo CLIP utilizando versões pré-treinadas fornecidas pelo OpenCLIP, para realizar a classificação de produtos em supermercados com a técnica ZSL. Avaliar o desempenho dos modelos em termos de acurácia, comparando diferentes variações do modelo CLIP. Explorar técnicas complementares, como o uso de bancos de dados de vetores e busca por similaridade com *embeddings*, para melhorar a eficiência e a capacidade de generalização do modelo. Aplicar técnicas de *data*

*augmentation* para melhorar a capacidade do modelo de lidar com diferentes variações visuais dos produtos, simulando cenários do mundo real. Comparar os resultados obtidos e identificar as principais vantagens e limitações da abordagem ZSL em relação a métodos tradicionais de classificação de produtos. Propor direções para pesquisas futuras, sugerindo como o ZSL pode ser adaptado a outros contextos do varejo e explorando novas técnicas de visão computacional e AM.

## 1.3 Organização do Documento

Este trabalho está estruturado da seguinte maneira:

- Capítulo 2 - Revisão da Literatura: Apresenta uma revisão de literatura referente ao uso de IA e visão computacional no varejo, discutindo as principais técnicas e modelos aplicados à detecção e classificação de produtos.
- Capítulo 3 - Desenvolvimento: Detalha o processo de implementação dos modelos CLIP, descrevendo os conjuntos de dados utilizados, os parâmetros adotados, as técnicas de pré-processamento das imagens e os métodos de avaliação empregados.
- Capítulo 4 - Resultados e Discussões: Apresenta e analisa os resultados obtidos nos experimentos, comparando o desempenho dos modelos e discutindo os principais desafios e limitações encontrados durante o desenvolvimento do trabalho.
- Capítulo 5 - Conclusão: Resume as principais conclusões do estudo, destacando a eficácia da abordagem ZSL para a classificação de produtos em supermercados, e propõe direções para futuras pesquisas, incluindo sugestões para melhorar a generalização dos modelos e explorar novas técnicas.

# 2 FUNDAMENTAÇÃO TEÓRICA

Este trabalho explora a aplicação de *Zero-Shot Learning* (ZSL) na classificação de produtos de supermercados, uma técnica que permite ao sistema classificar itens sem um treinamento específico anterior sobre esses produtos. Essa abordagem busca replicar a habilidade humana de generalizar e adaptar-se a novos contextos, sendo especialmente relevante em cenários onde novos produtos são constantemente introduzidos, apresentam alta similaridade com outros itens ou quando há escassez de dados anotados para o treinamento de modelos AM.

## 2.1 Reconhecimento de produtos no contexto do varejo

A classificação de produtos em ambientes de varejo tem sido amplamente investigada, com diversos autores abordando diferentes enfoques e técnicas, por exemplo, (WEI *et al.*, 2020) discutem os desafios e estratégias para o reconhecimento de produtos utilizando técnicas de visão computacional, com foco em abordagens de aprendizado profundo aplicadas à identificação de itens em prateleiras. Esses autores destacam a complexidade da variabilidade intra-classe, isto é, as diferenças entre exemplares de um mesmo produto, exemplificada na figura 1, e a dificuldade de se obter conjuntos de dados anotados adequados para o treinamento de modelos de AM. Além disso, eles ressaltam a necessidade de constante atualização dos modelos devido à introdução frequente de novos produtos.


Figura 1 – Exemplo de variações intra-classe. Fonte: Elaborada pelo Autor

(MERLER; GALLEGUILLOS; BELONGIE, 2007) propuseram uma abordagem que utiliza dados coletados *in vitro* para reconhecer produtos em ambientes reais. Essa pesquisa destaca a importância de se considerar a diversidade de condições ambientais ao classificar produtos de supermercado, já que essas variações podem impactar diretamente a precisão do reconhecimento.

De maneira similar, (ACHAKIR; MOHTARAM; ESCARTIN, 2023) desenvolveram uma solução automatizada baseada em IA para a detecção de produtos fora de estoque em ambientes de varejo. Embora o foco principal da pesquisa esteja na detecção de itens

ausentes, os métodos e técnicas utilizados podem ser adaptados para melhorar a precisão da classificação de produtos em supermercados. Este estudo demonstra a importância de soluções inteligentes e adaptáveis para lidar com os desafios do ambiente de varejo, onde a eficiência na gestão de estoque e a precisão na classificação são essenciais.

Outra abordagem relevante é a proposta por (SAKAI; KANEKO; SHIRAISHI, 2023), que oferece uma solução significativa ao permitir o reconhecimento de produtos de varejo com base em imagens limpas de sites de comércio eletrônico, utilizando apenas um ou poucos exemplos do mesmo produto, essa abordagem, conhecida como *one-shot learning*, visa facilitar a integração entre o mundo *online* e as lojas físicas. Esta abordagem é muito semelhante ao problema tratado neste trabalho, onde se dispõe de apenas uma imagem de exemplo do produto, que será usada como base para a classificação.

## 2.2 *Datasets*

Considerando os desafios mencionados anteriormente, muitos *datasets* foram desenvolvidos para auxiliar no enfrentamento dessas dificuldades. Esses conjuntos de dados são fundamentais para a pesquisa e o desenvolvimento de soluções de AM, fornecendo uma base de dados realista e diversificada para testar e avaliar as abordagens propostas. Além disso, os *datasets* permitem que pesquisadores e desenvolvedores avaliem a eficácia de suas soluções em diferentes cenários e condições, o que é crucial para garantir que as metodologias sejam robustas e eficazes nos diversos contextos aplicados.

### 2.2.1 Grozi-120

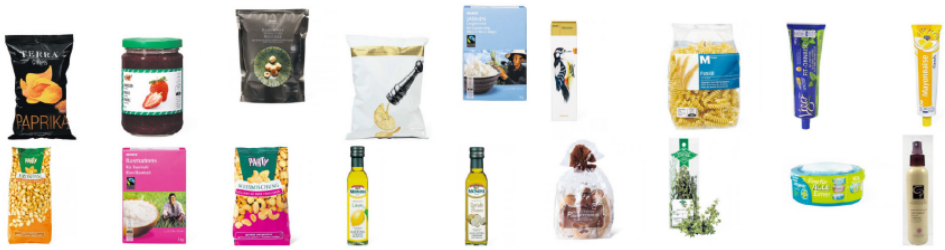
O *dataset* Grozi-120 (MERLER; GALLEGUILLOS; BELONGIE, 2007) é uma coleção de imagens de produtos de supermercado, contendo um total de 120 classes diferentes de produtos, e para cada produto, são disponibilizadas duas gravações distintas: imagens *in vitro*, extraídas da internet, e imagens *in situ*, capturadas a partir de vídeos filmados dentro de uma loja de supermercado. Essa dualidade de fontes de imagem foi projetada para facilitar o estudo das diferenças entre condições ideais e reais de captura de imagem e seu impacto no desempenho de algoritmos de reconhecimento de objetos.


Figura 2 – Exemplo de imagens *inSitu* do *dataset* Grozi-120. Fonte: Adaptada de (MERLER; GALLEGUILLOS; BELONGIE, 2007)


Figura 3 – Exemplo de imagens *inVitru* do *dataset* Grozi-120. Fonte: Adaptada de (MERLER; GALLEGUILLOS; BELONGIE, 2007)

### 2.2.2 Grozi-3.2k

O Grozi-3.2k(GEORGE; FLOERKEMEIER, 2014) é um *dataset* maior desenvolvido para avançar a pesquisa em recuperação de imagens e reconhecimento de produtos em ambientes de varejo e supermercados. Focado em classificação, o *dataset* contém 8.350 imagens de produtos e 680 imagens de teste, todas anotadas, capturadas tanto em condições ideais quanto em cenários do mundo real. Um aspecto inovador deste trabalho é que as imagens usadas para treinamento e as usadas para teste vêm de condições diferentes, simulando situações mais realistas, onde as imagens de treinamento foram capturadas em condições ideais, utilizando apenas uma imagem por tipo de produto, enquanto o conjunto de avaliação foi coletado em cenários do mundo real, utilizando um telefone celular em condições completamente diferentes. Este *dataset* visa estimular novas pesquisas em classificação de produtos de supermercados e varejo em geral.



(a) Sample training images from 'Food', 'Plants', and 'Healthcare' categories

(b) Sample testing images from our collected dataset

Figura 4 – Imagens de exemplo do *dataset* Grozi-3.2k. Fonte: (GEORGE; FLOERKEMEIER, 2014)

## 2.3 Classificação *Zero-Shot Learning*

*Zero-Shot Learning* (ZSL) é uma abordagem da inteligência artificial que visa solucionar problemas de classificação quando há falta de dados anotados para treinamento ou quando esses dados são escassos. Diferente do aprendizado supervisionado, que depende de grandes *datasets* rotulados, e do aprendizado não supervisionado, que se baseia em padrões entre exemplos não rotulados, o ZSL vai além, reconhecendo classes inéditas para o sistema durante o treinamento (M; VEDHAMANI; B, 2023; DINU; LAZARIDOU; BARONI, 2015).

Enquanto o aprendizado semi-supervisionado combina dados rotulados e não rotulados para melhorar o desempenho de reconhecimento das classes conhecidas, o ZSL é mais radical, exigindo a capacidade de generalizar para classes completamente desconhecidas (SUN; GU; SUN, 2021). Essa capacidade reflete o comportamento humano, onde, por exemplo, uma pessoa pode identificar um novo tipo de refrigerante mesmo sem ter tido contato prévio com o produto, baseando-se apenas em seu conhecimento sobre refrigerantes em geral, como tamanho, formato e cor. A similaridade entre essa habilidade humana e o ZSL reside na transferência de conhecimento entre classes conhecidas e desconhecidas, por meio de características compartilhadas (SUN; GU; SUN, 2021).

Essas características podem ser capturadas por modelos de aprendizado de máquina por meio de vetores de *embeddings* (ZHANG; XIANG; GONG, 2017; LAMPERT; NICKISCH; HARMELING, 2009). Em modelos ZSL, os *embeddings* desempenham um papel crucial na generalização para classes não vistas durante o treinamento. Segundo (PALATUCCI *et al.*, 2009), os *embeddings* permitem ao modelo explorar as semelhanças entre as características das classes conhecidas e das amostras, de forma que o modelo pode inferir a classe correta para novas amostras, mesmo sem ter sido exposto a elas.

No contexto de ZSL, quando um novo produto é introduzido para classificação, o modelo pode utilizar os *embeddings* das classes conhecidas para inferir a categoria correta do novo produto com base em características compartilhadas, como cor, tamanho e formato. Dessa forma, ZSL e *embeddings* são fundamentais para a classificação de produtos em um ambiente de varejo em constante mudança, permitindo que o sistema se adapte à crescente diversidade de produtos sem a necessidade de grandes conjuntos de dados rotulados e extensivos processos de retreinamento.

## 2.4 Visão computacional

A visão computacional tem como objetivo replicar a capacidade de percepção visual humana em máquinas, utilizando algoritmos e modelos matemáticos para interpretar e entender o conteúdo visual de imagens e vídeos. Essa área é parte fundamental da IA, com aplicações que incluem o reconhecimento de objetos, a segmentação de imagens,

a detecção de faces, a análise de vídeos e a classificação de imagens (MARENGONI; STRINGHINI, 2010). A visão computacional tem sido amplamente utilizada em sistemas de reconhecimento de produtos no varejo, onde é necessário processar grandes volumes de imagens para identificar produtos de forma rápida e precisa.

Um dos principais desafios enfrentados pela visão computacional no contexto do varejo é a necessidade de adaptar-se à variabilidade dos dados visuais, como a presença de diferentes condições de iluminação, ângulos de captura e até mesmo variações na aparência dos produtos ao longo do tempo. Para superar essas dificuldades, uma técnica amplamente utilizada são as redes CNN (*Convolutional Neural Networks*)(O'SHEA; NASH, 2015), que têm a capacidade de extrair automaticamente características relevantes de imagens, gerando representações numéricas, os *embeddings*.

Durante o treinamento de modelos de visão computacional, os *embeddings* são ajustados para capturar a semântica das imagens, ou seja, suas características mais importantes em relação às classes conhecidas, por exemplo, em um modelo treinado para reconhecer frutas, os *embeddings* de diferentes tipos de frutas serão posicionados em um espaço vetorial de modo a refletir suas semelhanças e diferenças. Em abordagens de *Zero-Shot Learning*, os *embeddings* não são ajustados apenas para as classes vistas, mas também projetados para permitir a generalização para novas classes, utilizando características compartilhadas, como atributos visuais comuns ou descrições textuais (ZHANG; XIANG; GONG, 2017).

## 2.5 Modelos baseados em *transformers*

Nos últimos anos, modelos de visão computacional baseados em *Transformers*(VASWANI *et al.*, 2023), conhecidos como *Vision Transformers* (ViT) (DOSOVITSKIY *et al.*, 2015), têm ganhado destaque devido à sua capacidade de extrair características tanto de dados visuais quanto textuais. Esses modelos são projetados para lidar com a complexidade e a grande variabilidade dos dados visuais, oferecendo uma abordagem alternativa e em certos casos mais precisas em relação aos métodos tradicionais baseados em CNNs.

Os modelos ViT introduziram uma nova arquitetura que se mostrou extremamente eficaz em diversas tarefas de visão computacional, como classificação de imagens, segmentação semântica e detecção de objetos, a figura 5 mostra como é a arquitetura do modelo ViT. Diversos modelos baseados na arquitetura do ViT foram desenvolvidas para as diversas tarefas de visão computacional, dentre esses modelos podemos destacar o CLIP(RADFORD *et al.*, 2021).

**CLIP** (*Contrastive Language–Image Pretraining*) desenvolvido pela OpenAI, combina textos e imagens durante o treinamento, permitindo que o modelo compreenda contextos visuais e linguísticos de forma conjunta. Este modelo é composto por duas redes

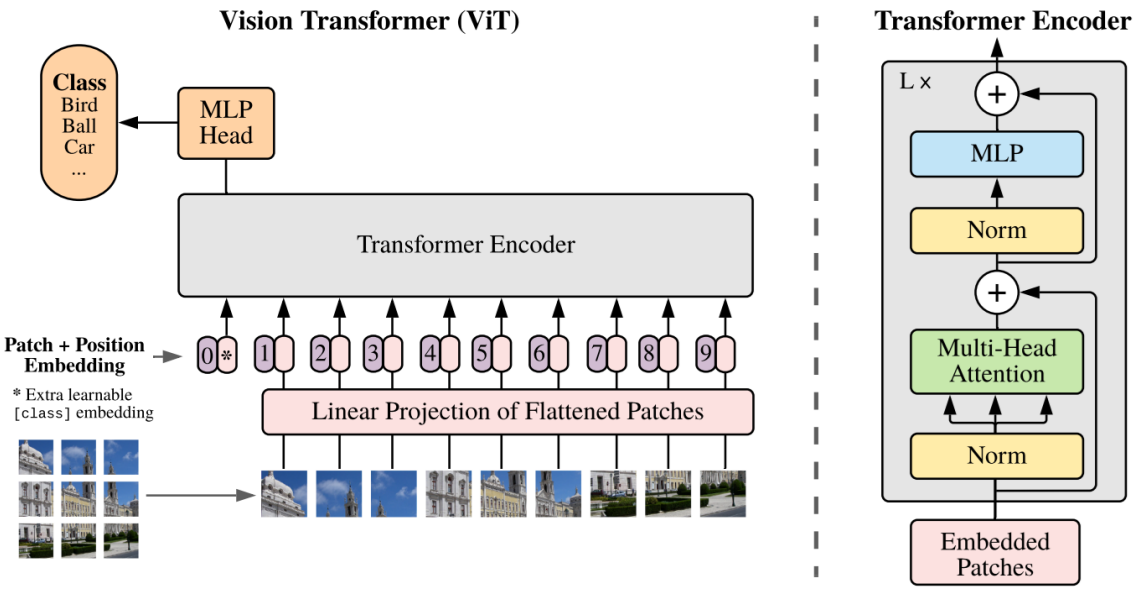



Figura 5 – Arquitetura de ViT. Fonte: (DOSOVITSKIY *et al.*, 2015)

neurais distintas: uma dedicada ao processamento de texto, utilizando um modelo baseado em *Transformers*, e outra dedicada ao processamento de imagens, empregando uma rede ResNet(HE *et al.*, 2015) ou uma ViT.(PINECONE, 2023) O grande diferencial do CLIP é sua capacidade de aprender e unificar representações a partir de textos e imagens, isso permite que o modelo compreenda e relacione conceitos visuais e textuais sem a necessidade de dados de treinamento específicos para cada tarefa, o que possibilita um desempenho impressionante em uma ampla variedade de tarefas de visão computacional e processamento de linguagem natural, incluindo classificação de imagens, busca por similaridade e geração de legendas para imagens, com base em uma única arquitetura.(RADFORD *et al.*, 2021)

A Figura 6 ilustra essa arquitetura, onde as representações de imagem e texto são unificadas em um espaço comum. Essa abordagem permite que o CLIP generalize bem para tarefas não vistas anteriormente, como classificar uma nova imagem com base em uma descrição textual, ou até mesmo identificar imagens que sejam semanticamente similares sem que tenham sido explicitamente rotuladas.

## 2.6 Busca por similaridade

Em um contexto onde a classificação de produtos se baseia em *embeddings*, a busca por similaridade se torna uma etapa crucial e altamente sensível para o resultado final. A busca por similaridade envolve comparar um novo *embedding* com um banco dados de *embeddings* existentes para encontrar os mais semelhantes. Essa técnica é fundamental e é umas das bases para o ZSL, onde a classificação é feita com base na similaridade entre as características de novas amostras e as características de classes conhecidas. Dessa forma, o

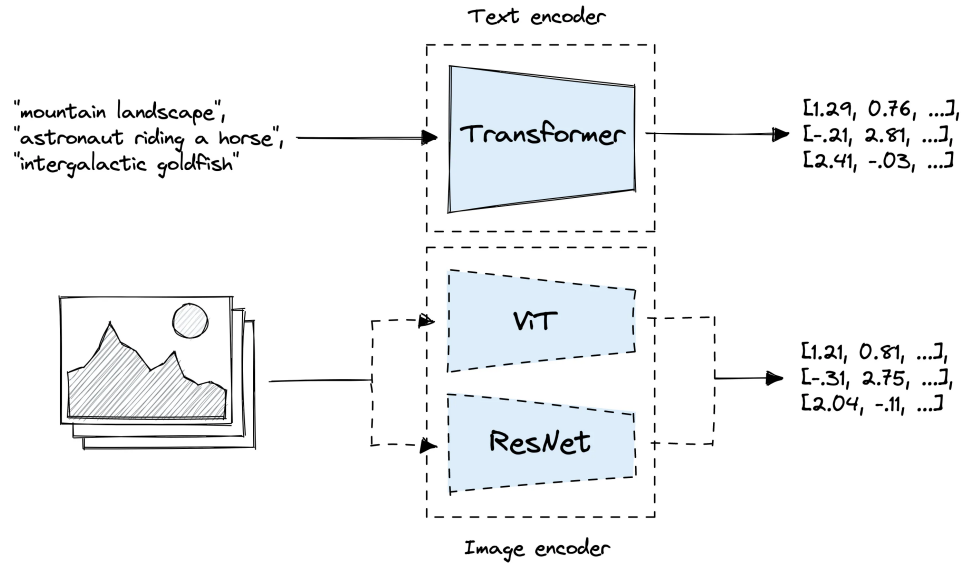



Figura 6 – Arquitetura do CLIP. Fonte: (PINECONE, 2023)

ZSL permite a classificação de produtos sem a necessidade de exemplos de treinamento específicos para cada classe, utilizando apenas as relações de similaridade no espaço dos *embeddings*.

### 2.6.1 FAISS

Desenvolvida pelo *Meta AI Research* o **FAISS**(*Facebook AI Similarity Search*)(DOUZE *et al.*, 2024) constitui uma biblioteca desenvolvida, com o propósito de realizar buscas eficientes e rápidas em grandes coleções de vetores de *embeddings*. FAISS é otimizada para executar operações de busca por similaridade de forma extremamente eficiente, tirando proveito tanto da CPU quanto da GPU para acelerar significativamente o processamento de busca(JOHNSON; DOUZE; JÉGOU, 2019). Esta biblioteca foi projetada para lidar com coleções massivas de vetores, o que a torna uma escolha amplamente popular em projetos que envolvem grandes volumes de dados. Essa otimização e flexibilidade permitem que o FAISS seja aplicada em uma variedade de contextos, onde a necessidade de buscas rápidas e precisas em dados de alta dimensionalidade é crítica.

A biblioteca FAISS oferece inúmeras vantagens que a tornam uma ferramenta poderosa para a busca em vetores de alta dimensão. Primeiramente, ela é capaz de realizar buscas de forma extremamente rápida, mesmo em cenários onde o número de vetores é excepcionalmente grande, o que é obtido por meio de algoritmos avançados de indexação e técnicas de quantização que visam reduzir a complexidade das buscas(DOUZE *et al.*, 2024). Além disso, a flexibilidade da FAISS se manifesta no suporte a diversas métricas de similaridade, como a distância euclidiana e a similaridade de cosseno, permitindo que a biblioteca seja adaptada a diferentes necessidades de aplicação. Em termos de processamento e escalabilidade, FAISS pode ser implementada em sistemas que utilizam tanto

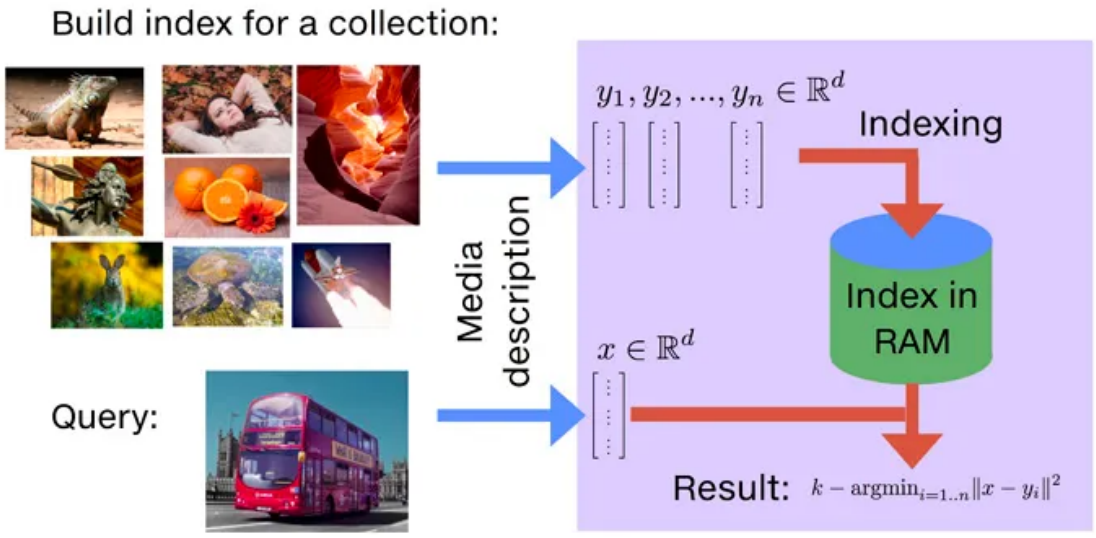



Figura 7 – Busca por similaridade com FAISS. Fonte: (JOHNSON; DOUZE; JéGOU, 2017)

CPU quanto GPU(JOHNSON; DOUZE; JÉGOU, 2019), o que possibilita o processamento paralelo e acelera de forma significativa as operações de busca. Por fim, a capacidade de FAISS de lidar com grandes volumes de dados se revela particularmente útil em um ambiente como o de varejo, onde há um número vasto de produtos e a quantidade de *embeddings* pode ser extremamente alta.(JOHNSON; DOUZE; JéGOU, 2017)

# 3 DESENVOLVIMENTO

O problema de ruptura interna pode ser desmembrado em vários desafios computacionais, como a detecção e localização de produtos, identificação e mensuração de espaços vazios, classificação e contagem de produtos e seus subtipos, e agrupamento de imagens semelhantes. O *pipeline* típico para o reconhecimento de produtos consiste em três grandes etapas, representadas na Figura 8. Primeiramente, é necessário detectar os produtos (i) e suas regiões na imagem. Em seguida, realiza-se o recorte (ii) do produto da imagem, para então classificar (iii) cada recorte.

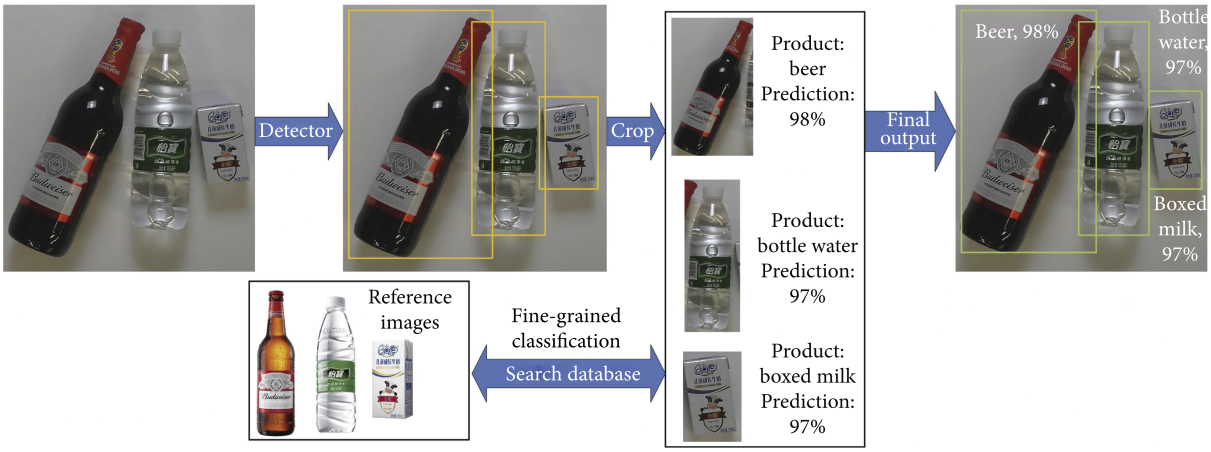



Figura 8 – Pipeline de reconhecimento de produtos. Fonte: (WEI *et al.*, 2020)

Para realizar a classificação utilizando técnicas de ZSL, é necessário seguir uma série de etapas fundamentais, que garantem a correta aplicação do método, essas etapas são detalhadas a seguir:

## 3.1 Base de dados das imagens de referência

Para iniciar o processo de classificação, é necessário gerar uma base de imagens de referência, construída a partir da coleta de imagens de cada produto disponível no catálogo. Essas imagens servirão como base para a classificação dos produtos e para a avaliação dos modelos que serão testados posteriormente. Após a coleta, as imagens passam por um processo de padronização, incluindo o redimensionamento para uma resolução uniforme, a centralização para remover bordas externas desnecessárias, e a normalização, ajustando os valores de pixel para um intervalo padrão. Esse processo facilita o processamento subsequente e garante consistência nos *embeddings* gerados, resultando em melhor desempenho dos modelos e, consequentemente, em resultados mais precisos.

## 3.2 *Data augmentation*

Para aumentar a diversidade e robustez das imagens presentes nos *datasets*, aplicaremos técnicas de *data augmentation*. *Data augmentation* envolve a aplicação de várias transformações às imagens existentes, criando novas versões das mesmas. O objetivo é simular a variedade de condições que podem ocorrer em um ambiente real, ajudando o modelo a generalizar melhor e a ser mais robusto diante dessas diferentes condições. As transformações que serão aplicadas incluem:

- Rotação: As imagens serão rotacionadas em ângulos aleatórios de até 15º, para simular diferentes ângulos dos produtos nas prateleiras.
- Mudança de perspectiva: Ajustes na perspectiva das imagens serão realizados para simular diferentes perspectivas de visualização, como vistas de cima ou de lado.
- Alteração de resolução e cores: Modificações na resolução e nos canais de cores da imagem, visando simular diferentes condições de qualidade, iluminação e pequenas variações na aparência dos produtos.
- Recortes: Partes das imagens serão recortadas aleatoriamente, simulando imprecisões e variações nos recortes dos produtos detectados.

Essas novas imagens permitirão que a base de dados de referência contenha o mesmo produto sob diversas condições, melhorando sua capacidade de generalização e obtendo um cenário mais fidedigno ao que será encontrado no dia a dia, conforme exemplificado na figura 9.

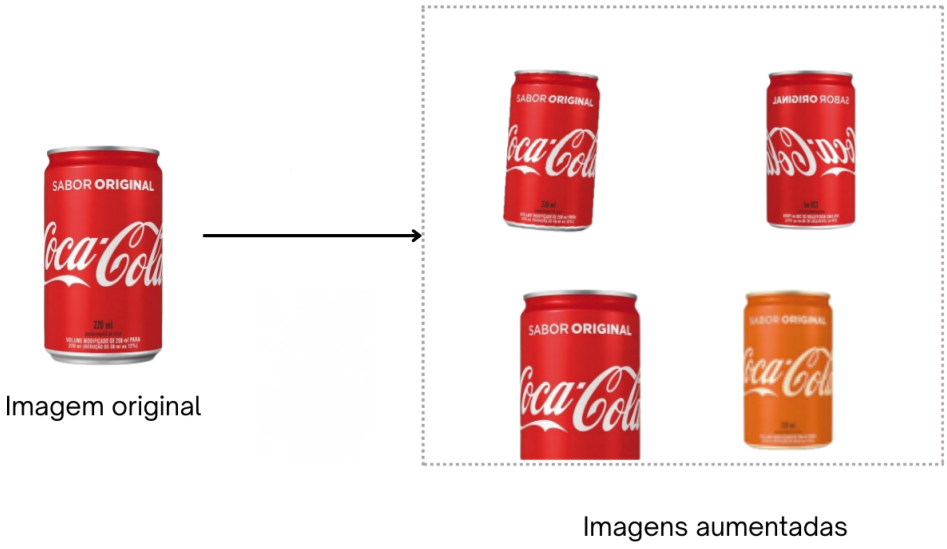



Figura 9 – Exemplo de imagens aumentadas. Fonte: Elaborada pelo Autor

## 3.3 Geração de *embeddings* de imagens

Cada imagem, incluindo as aumentadas, será processada pelos modelos para gerar seus respectivos *embeddings*. Neste trabalho, utilizaremos e testaremos os modelos ViT-L pré-treinados, disponíveis no repositório do OpenCLIP(ILHARCO *et al.*, 2021). Especificamente utilizaremos a arquitetura ViT-L/14, utilizaremos três modelos, um treinado no *dataset* LAION-2B(SCHUHMANN *et al.*, 2022), outro no DFN-2B(FANG *et al.*, 2023) e o CLIP original (RADFORD *et al.*, 2021) criado pela OpenAI. Os *embeddings* gerados por esses modelos serão armazenados em um banco de dados vetorial utilizando o FAISS, que será configurado para suportar consultas rápidas, permitindo a recuperação rápida dos K *embeddings* mais similares presentes na base de dados de referência. Para determinar a similaridade, será utilizada a distância Euclidiana, dada por $d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$.

## 3.4 Classificação dos produtos

A classificação dos produtos é a etapa crucial deste projeto, sendo responsável por identificar e categorizar os produtos com base em suas imagens recortadas após a detecção. O processo começa com a padronização da imagem, de forma semelhante à das imagens de referência, após esse pré-processamento, a imagem é processada pelo modelo para gerar os *embeddings* do produto a ser classificado. Esses *embeddings* são então comparados com os armazenados na base de dados de referência utilizando o FAISS, que recupera os K itens mais similares ao da imagem do produto em questão. Com base nos K itens mais similares retornados pelo FAISS, determinamos a categoria ou identidade do produto.

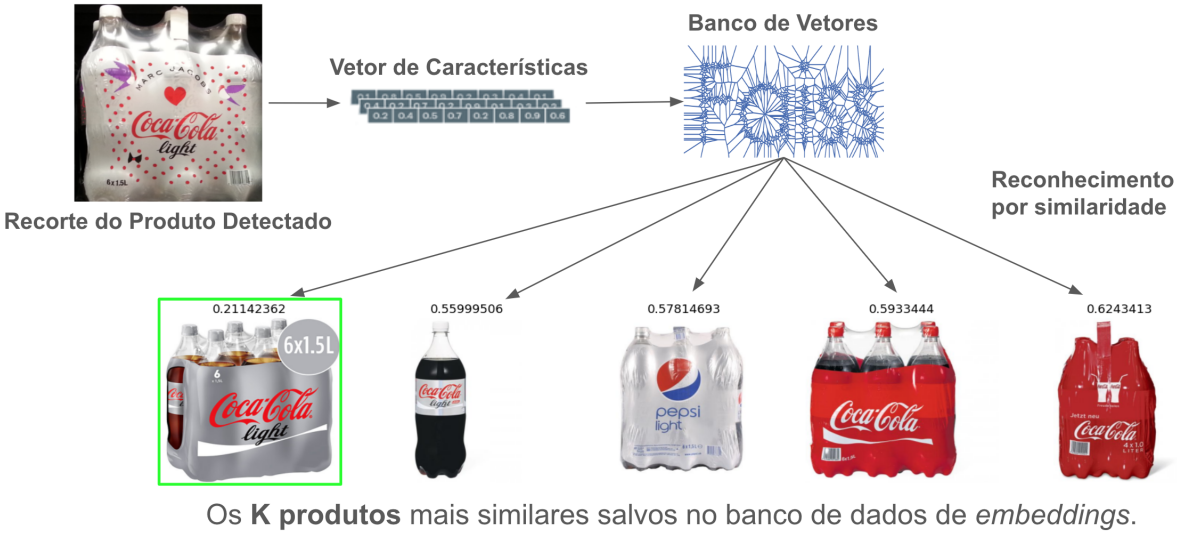



Figura 10 – Etapas para classificação. Fonte: Elaborada pelo Autor

## 3.5 Avaliação e validação

Para garantir que o método desenvolvido seja preciso, robusto e eficaz em um ambiente de produção, será necessário avaliar, validar e mensurar os resultados obtidos por cada modelo empregado. As métricas utilizadas para avaliar o desempenho dos modelos serão:

- *Accuracy*: Mede a proporção de acertos nas previsões sobre o número total de previsões realizadas. É calculada como:

$$Accuracy = \frac{\text{Predições corretas}}{\text{Número Total de predições}}$$

- *Top-k Accuracy*: Mede a proporção de casos em que a classificação correta do produto está entre as K previsões mais prováveis do modelo. Esta métrica é particularmente útil onde múltiplas sugestões podem ser consideradas.

$$\textit{Top-k Accuracy} = \frac{\text{Predições corretas no top k}}{\text{Número total de predições}}$$

Com base nessas métricas, os experimentos de classificação serão realizados utilizando os diferentes modelos e *datasets*, comparando seu desempenho em termos *Accuracy* e *Top-k Accuracy*. Os resultados serão analisados para avaliar o desempenho geral dos modelos e identificar possíveis melhorias. Além disso, a análise permitirá verificar a eficácia dos modelos em condições variadas de captura de imagem (como ângulos, iluminação e qualidade), assegurando assim a eficiência do sistema para aplicações em tempo real. A partir dos resultados obtidos, poderão ser identificadas áreas de melhoria e possíveis otimizações nos modelos e no pipeline de classificação empregado.

# 4 RESULTADOS E DISCUSSÕES

Nesta seção, apresentamos e discutimos os resultados obtidos na tarefa de classificação de produtos utilizando a abordagem ZSL com os diferentes modelos baseados na arquitetura CLIP. As métricas de desempenho consideradas incluem *Accuracy* e *Top-10 Accuracy*, que são comumente utilizadas para avaliar a eficácia de sistemas de classificação de imagem. Os resultados foram obtidos a partir de experimentos realizados com dois conjuntos de dados distintos: Grozi-120 e Grozi 3.2k, conforme demonstrado na tabela 1 e nos gráficos das figuras 11 e 12. Foram utilizados três modelos CLIP para testarmos a classificação dos produtos: CLIP OpenAI, CLIP DFN-2B e CLIP LAION-2B, cada um com características e conjuntos de treinamento distintos.

| ***Dataset*** | **Modelo** | ***Accuracy*** | ***Top-10 Accuracy*** |
|---|---|---|---|
| **Grozi-120** | **Clip OpenAI** | 0,5914 | 0,6906 |
| | **Clip DFN-2B** | 0,6323 | 0,7344 |
| | **Clip LAION-2B** | 0,5955 | 0,6784 |
| **Grozi 3.2k** | **OpenAI Clip** | 0,5325 | 0,7609 |
| | **Clip DFN-2B** | 0,5576 | 0,8152 |
| | **OpenClip LAION-2B** | 0,5378 | 0,7695 |

Tabela 1 – Resultados dos modelos nos *datasets* Grozi-120 e Grozi 3.2k

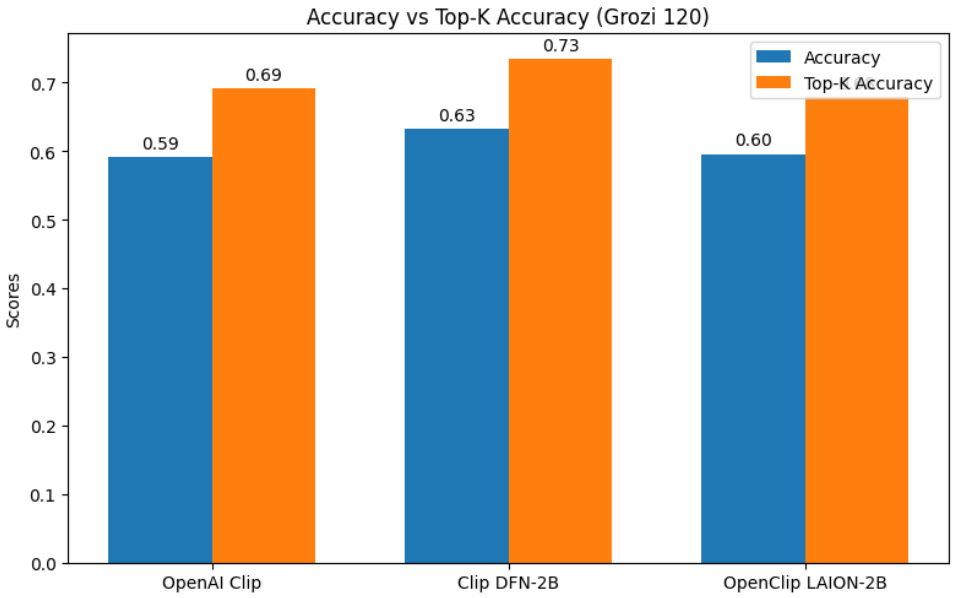



Figura 11 – Resultados no *dataset* Grozi 120. Fonte: Elaborada pelo autor

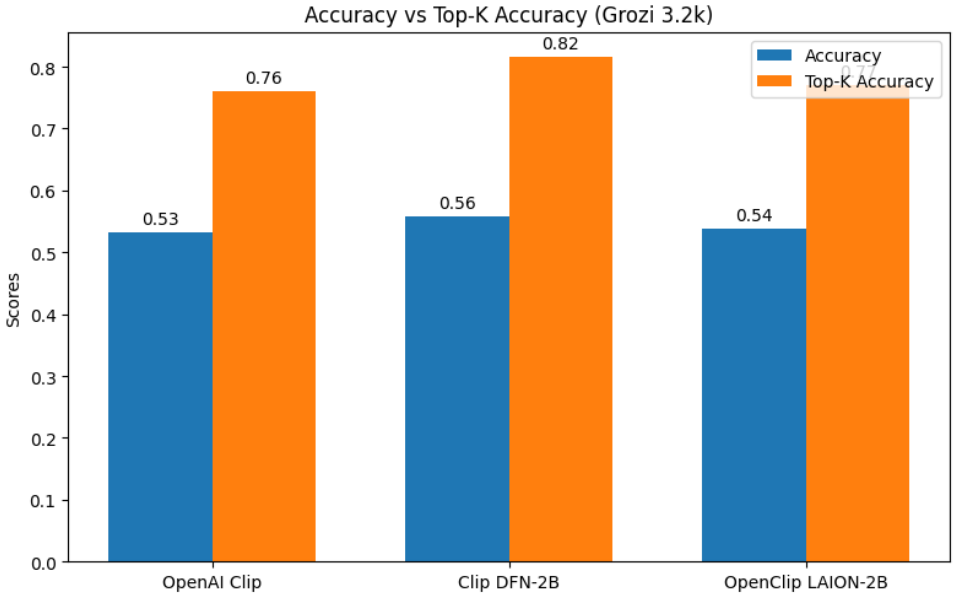



Figura 12 – Resultados no *dataset* Grozi 3.2k. Fonte: Elaborada pelo autor

Os resultados mostram que os três modelos testados - CLIP OpenAI, CLIP DFN-2B e CLIP LAION-2B - apresentaram desempenhos bastante próximos, com variações relativamente pequenas entre si, tanto no *dataset* Grozi-120 quanto no Grozi 3.2k. A métrica de *Accuracy*, que mede a proporção de classificações corretas, variou entre 59,14% e 63,23% no conjunto Grozi-120, e entre 53,25% e 55,76% no Grozi 3.2k. Já a *Top-10 Accuracy*, que indica a frequência com que a classe correta estava entre as 10 primeiras sugestões, foi ligeiramente mais alta, variando de 67,84% a 73,44% no Grozi-120, e de 76,09% a 81,52% no Grozi 3.2k.

Essa proximidade nos resultados sugere que a arquitetura CLIP é bem robusta e ajustada para tarefas de classificação de itens não vistos durante o treinamento, mostrando uma capacidade de generalização consistente, mesmo quando aplicada a diferentes bases de dados. No entanto, é importante destacar que a principal variação entre os resultados pode ser atribuída à diferença nos conjuntos de dados utilizados para o treinamento dos modelos. Por exemplo, o CLIP DFN-2B apresentou uma leve superioridade em todas as métricas, especialmente no Grozi 3.2k, o que pode indicar que o *dataset* usado para o seu treinamento foi mais adequado para as características dos produtos presentes nesses *datasets* testados.

Além disso, os resultados indicam que a abordagem ZSL pode ser aplicada de forma eficaz em domínios como a classificação de produtos em supermercados, mas ainda há desafios a serem superados. A abordagem ZSL, ao tentar classificar produtos sem um treinamento prévio explícito para todas as classes, enfrenta dificuldades ao lidar com a grande diversidade de classes e com as variações nas condições de captura das imagens, como ângulo de visão, iluminação e qualidade da imagem. Essas variáveis podem influenciar

diretamente na precisão do modelo, levando a erros de classificação.

## 4.1 Análise das Métricas

Ao analisar mais detalhadamente as métricas, a *Accuracy* relativamente baixa em alguns experimentos revela limitações da abordagem ZSL para determinadas classes de produtos, especialmente em cenários onde uma única predição correta é necessária. Em ambientes reais, como sistemas de inventário automatizado, essa limitação pode ser significativa, sugerindo que técnicas adicionais de refinamento de predição, como a utilização de algoritmos de votação ou agregação de resultados, seriam benéficas para melhorar o desempenho final.

Por outro lado, a *Top-10 Accuracy* elevada demonstra que os modelos são capazes de fornecer boas sugestões em cenários práticos. Em aplicações onde múltiplas opções podem ser apresentadas ao usuário, como sistemas de recomendação ou assistentes de compra, essa métrica se torna particularmente útil. Isso sugere que, embora os modelos possam não acertar a classe correta de imediato, eles frequentemente oferecem opções viáveis dentro das 10 primeiras sugestões, o que é vantajoso em sistemas que permitem a intervenção humana na escolha final.

Um ponto crucial a ser considerado é o tamanho e a diversidade dos *datasets* utilizados. O Grozi-120 e o Grozi 3.2k, embora sejam adequados para um estudo inicial, são pequenos em comparação com *datasets* de outros domínios, como o ImageNet, que contém milhões de imagens. Essa limitação de tamanho pode ter prejudicado a capacidade dos modelos de aprender padrões mais complexos, restringindo sua habilidade de generalizar para novos cenários. O uso de *datasets* maiores e mais diversificados poderia potencialmente melhorar a performance dos modelos em tarefas de classificação de produtos, fornecendo uma base mais robusta para o aprendizado.

Outro aspecto relevante é a qualidade dos dados de treinamento. O modelo CLIP DFN-2B, que obteve o melhor desempenho, parece ter se beneficiado de um processo de treinamento mais robusto e de uma curadoria de dados mais rigorosa. Isso ressalta a importância não apenas de escolher uma arquitetura de modelo adequada, mas também de garantir que os dados usados no treinamento sejam de alta qualidade e representem bem o domínio de aplicação. Conjuntos de dados bem curados, com maior diversidade de imagens e melhor filtragem, são fundamentais para aumentar a eficácia dos modelos ZSL.

# 5 CONCLUSÕES

A classificação de produtos utilizando técnicas de ZSL demonstrou ser uma abordagem promissora, especialmente em cenários onde a diversidade de classes é alta e a necessidade de generalização é crítica. Os resultados obtidos com os modelos baseados na arquitetura CLIP mostram que, mesmo sem um treinamento específico no domínio de produtos de supermercado, os modelos conseguiram realizar a tarefa de classificação com um grau razoável de precisão. Um aspecto importante desta abordagem é que os testes foram realizados utilizando ZSL, sem a necessidade de treinamento ou *fine-tuning* dos modelos. Isso representou uma vantagem significativa, pois eliminou a necessidade de gastar com recursos computacionais intensivos que seriam exigidos para treinar ou ajustar os modelos, tornando o processo mais eficiente e acessível.

Entretanto, a acurácia atingida pelos modelos, embora significativa, ainda deixa margem para aprimoramentos. Uma das principais limitações observadas foi a dificuldade dos modelos em diferenciar produtos que possuem aparências muito semelhantes ou que variam em pequenos detalhes, como variações de embalagens e marcas. Esse desafio evidencia a necessidade de técnicas mais avançadas e especializadas para lidar com essas nuances.

Uma alternativa para superar essas limitações seria a aplicação de técnicas de *Fine-Tuning*. O *Fine-Tuning* permite ajustar modelos pré-treinados em grandes *datasets*, como o LAION-2B ou DFN-2B, utilizando *datasets* específicos do contexto desejado, neste caso o de supermercados. Ao adaptar os modelos com dados que refletem mais fielmente o contexto de aplicação, é possível aumentar significativamente a acurácia e a robustez das classificações. Com um *Fine-Tuning* especifico para supermercados, o modelo não apenas ajustaria seus pesos para reconhecer melhor as características específicas dos produtos, mas também melhoraria sua capacidade de diferenciar entre classes semelhantes.

Além do *fine-tuning*, o alto *Top-10 Accuracy* sugere que técnicas como votação ou agregação de resultados poderiam ser exploradas para determinar a classificação exata, melhorando ainda mais a precisão e a confiabilidade das predições. Outras técnicas, como *data augmentation* personalizada para este contexto, também podem ser exploradas para aumentar ainda mais a precisão.

Portanto, embora a abordagem ZSL tenha se mostrado eficaz, o caminho para uma solução de classificação de produtos altamente precisa passa pela adaptação dos modelos ao contexto específico de aplicação. O uso de *fine-tuning* com *datasets* no contexto de supermercados, juntamente com outras técnicas de aprimoramento, representa uma direção promissora para futuras pesquisas e desenvolvimentos nesta área. Ao focar em tais melhorias,

podemos desenvolver sistemas mais robustos e precisos, capazes de operar de forma eficaz em ambientes reais de produção, como em supermercados e outros estabelecimentos comerciais.

# REFERÊNCIAS


ACHAKIR, F.; MOHTARAM, N.; ESCARTIN, A. An automated ai-based solution for out-of-stock detection in retail environments. *In*: **2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)**. [*S.l.: s.n.*], 2023. p. 1–6.

DINU, G.; LAZARIDOU, A.; BARONI, M. **Improving zero-shot learning by mitigating the hubness problem**. 2015.

DOSOVITSKIY, A. *et al.* **Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks**. 2015.

DOUZE, M. *et al.* The faiss library. 2024.

FANG, A. *et al.* **Data Filtering Networks**. 2023. Disponível em: <https://arxiv.org/abs/2309.17425>.

GEORGE, M.; FLOERKEMEIER, C. Recognizing products: A per-exemplar multi-label image classification approach. *In*: FLEET, D. *et al.* (ed.). **Computer Vision – ECCV 2014**. Cham: Springer International Publishing, 2014. p. 440–455. ISBN 978-3-319-10605-2.

GUHA, A. *et al.* How artificial intelligence will affect the future of retailing. **Journal of Retailing**, v. 97, n. 1, p. 28–41, 2021. ISSN 0022-4359. Re-Strategizing Retailing in a Technology Based Era. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022435921000051>.

HE, K. *et al.* **Deep Residual Learning for Image Recognition**. 2015.

ILHARCO, G. *et al.* **OpenCLIP**. Zenodo, 2021. If you use this software, please cite it as below. Disponível em: <https://doi.org/10.5281/zenodo.5143773>.

JOHNSON, J.; DOUZE, M.; JÉGOU, H. Billion-scale similarity search with GPUs. **IEEE Transactions on Big Data**, IEEE, v. 7, n. 3, p. 535–547, 2019.

JOHNSON, J.; DOUZE, M.; JéGOU, H. **Faiss: A Library for Efficient Similarity Search**. 2017. Acesso em: 11 jun. 2024. Disponível em: <https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/>.

LAMPERT, C. H.; NICKISCH, H.; HARMELING, S. Learning to detect unseen object classes by between-class attribute transfer. *In*: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2009. p. 951–958.

M, M. P.; VEDHAMANI, A.; B, S. K. Zero-shot learning for text classification: Extending classifiability beyond conventional techniques. *In*: **2023 IEEE Region 10 Symposium (TENSYMP)**. [*S.l.: s.n.*], 2023. p. 1–6.

MARENGONI, M.; STRINGHINI, S. Tutorial: Introdução à visão computacional usando opencv. **Revista de Informática Teórica e Aplicada**, v. 16, n. 1, p. 125–160, Mar. 2010. Disponível em: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v16_n1_p125>.

MERLER, M.; GALLEGUILLOS, C.; BELONGIE, S. Recognizing groceries in situ using in vitro training data. *In*: **2007 IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2007. p. 1–8.

O'SHEA, K.; NASH, R. **An Introduction to Convolutional Neural Networks**. 2015. Disponível em: <https://arxiv.org/abs/1511.08458>.

PALATUCCI, M. *et al.* Zero-shot learning with semantic output codes. *In*: . [*S.l.: s.n.*], 2009. v. 22, p. 1410–1418.

PINECONE. Multi-modal ml with openai's clip. **Pinecone Learning Series**, 2023. Acesso em: 12 jun. 2024. Disponível em: <https://www.pinecone.io/learn/series/image-search/clip/>.

RADFORD, A. *et al.* **Learning Transferable Visual Models From Natural Language Supervision**. 2021.

SAKAI, R.; KANEKO, T.; SHIRAISHI, S. Framework for fine-grained recognition of retail products from a single exemplar. *In*: **2023 15th International Conference on Knowledge and Smart Technology (KST)**. [*S.l.: s.n.*], 2023. p. 1–6.

SCHUHMANN, C. *et al.* **LAION-5B: An open large-scale dataset for training next generation image-text models**. 2022. Disponível em: <https://arxiv.org/abs/2210.08402>.

SHANKAR, V. How artificial intelligence (ai) is reshaping retailing. **Journal of Retailing**, v. 94, n. 4, p. vi–xi, 2018. ISSN 0022-4359. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022435918300769>.

SON, J.; KANG, J. H.; JANG, S. The effects of out-of-stock, return, and cancellation amounts on the order amounts of an online retailer. **Journal of Retailing and Consumer Services**, v. 51, p. 421–427, 2019. ISSN 0969-6989. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0969698918305186>.

SRIVASTAVA, M. M. **RetailKLIP : Finetuning OpenCLIP backbone using metric learning on a single GPU for Zero-shot retail product image classification**. 2024.

SUN, X.; GU, J.; SUN, H. Research progress of zero-shot learning. **Applied Intelligence**, v. 51, n. 6, p. 3600–3614, 2021. ISSN 1573-7497. Disponível em: <https://doi.org/10.1007/s10489-020-02075-7>.

VASWANI, A. *et al.* **Attention Is All You Need**. 2023.

WEI, Y. *et al.* Deep learning for retail product recognition: Challenges and techniques. **Computational Intelligence and Neuroscience**, v. 2020, p. e8875910, nov 2020. ISSN 1687-5265. Acesso em: 14 de nov. 2023. Disponível em: <https://doi.org/10.1155/2020/8875910>.

ZHANG, L.; XIANG, T.; GONG, S. Learning a deep embedding model for zero-shot learning. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2017.