Explorando o uso de VLMs para classificação Zero-Shot de Imagens

Carlos M.S. Figueiredo[1], **Tiago E. de Melo**[1]

[1]Ocean Center – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brazil

{cfigueiredo,tmelo}@uea.edu.br

Abstract. *Vision-Language Models (VLMs) have revolutionized image classification, enabling object recognition without the need for specific training. This work investigates the impact of such models with different strategies, such as direct prompts that leverage the embedded knowledge of these models, and descriptive prompts that use the reasoning ability of the models for recognizing unknown objects. We also evaluate a new strategy, where a lightweight VLM is used to textually describe objects and an LLM with greater reasoning capacity classifies the object by description. Results show the importance of the implicit knowledge of models, but that small and limited models can perform more complex tasks with more descriptive prompts. This study contributes to the understanding of the potential of zero-shot image classification with VLMs, offering valuable insights for advancements on the topic.*

Resumo. *Modelos Visão-Linguagem (VLMs) têm revolucionado a classificação de imagens, permitindo o reconhecimento de objetos sem necessidade de treinamento específico. Este artigo investiga o impacto desses modelos com diferentes estratégias, como prompts diretos que exploram a conhecimento embutido desses modelos, e prompts descritivos que usam habilidade de raciocínio dos modelos para reconhecimento de objetos desconhecidos. Avaliamos também uma estratégia alternativa, onde um VLM leve é utilizado para descrever textualmente objetos e uma LLM com maior capacidade de raciocínio classifica o objeto pela descrição. Resultamos mostram a importância do conhecimento implícito de modelos, mas que modelos pequenos e limitados podem realizar tarefas mais complexas com prompts mais descritivos. Este estudo contribui para o entendimento do potencial da classificação de imagens zero-shot com VLMs, oferecendo insights valiosos para avanços no tema.*

1. Introdução

Um dos problemas clássicos da Visão Computacional é a classificação de imagens, que consiste em atribuir rótulos ou categorias a imagens, identificando os objetos ou conceitos presentes nelas. Tradicionalmente, a classificação de imagens é realizada por meio de modelos de aprendizado de máquina profundo, principalmente redes neurais convolucionais [LeCun et al. 2015], treinados em grandes conjuntos de dados rotulados. No entanto, essa abordagem apresenta limitações, como a necessidade de grandes quantidades de dados rotulados, que podem ser custosos e demorados de obter, e a dificuldade de generalizar para novas categorias ou domínios não vistos durante o treinamento.

Nos últimos anos, os Modelos Visão-Linguagem (VLMs - *Video Language Models*) emergiram como uma nova abordagem promissora para a classificação de imagens, como evidenciado pelo grande número de publicações recentes [Radford et al. 2021]. Os VLMs são

modelos de inteligência artificial que combinam o poder da visão computacional com o Processamento de Linguagem Natural (PLN), permitindo que os modelos compreendam e relacionem informações visuais e textuais. Essa capacidade possibilita que os VLMs realizem tarefas de classificação de imagens *zero-shot*, sem a necessidade de treinamento específico, explorando o conhecimento implícito presente nos dados de treinamento multimodal por meio de *prompts* bem projetados. Embora essa abordagem tenha se mostrado viável nos Modelos de Linguagem de Grande Escala (LLMs - *Large Language Models*) [Menon and Vondrick 2022, Pratt et al. 2023], a aplicação prática das VLMs ainda é pouco explorada [Liu et al. 2024].

Ainda, existem desafios e limitações a serem superados com uso de VLMs. Grandes modelos podem ter elevado conhecimento sobre elementos visuais e capacidade de raciocínio, mas podem ser incapazes de realizar tarefas em dispositivos embarcados ou de borda. Já modelos de tamanhos mais modestos podem ser mais dependentes de uma boa estratégia de engenharia de *prompt*.

Diante desse contexto, este artigo apresenta um estudo sobre o desempenho de diferentes modelos de VLMs, bem como de estratégias de *prompt*, para classificação *zero-shot* em um conjunto de dados de exemplo. Investigou-se o impacto de diferentes estratégias de *prompt*, como *prompt* diretos que exploram o conhecimento embutido desses modelos, e *prompt* descritivos que usam a habilidade de raciocínio dos modelos para reconhecimento de objetos desconhecidos. Avaliou-se também uma estratégia alternativa, onde um VLM leve é utilizado para descrever textualmente objetos e um LLM com maior capacidade de raciocínio classifica o objeto pela descrição.

Os resultados mostram a importância do conhecimento implícito de modelos, mas que modelos pequenos e limitados podem realizar tarefas mais complexas com prompts mais descritivos. Particularmente, grandes VLMs são mais competitivos a modelos treinados especificamente para a tarefa, mas estratégias de *prompt* fazem com que modelos muito pequenos se aproximem desse desempenho. Este estudo contribui para o entendimento do potencial e das nuances da classificação de imagens zero-shot com VLMs, abrindo caminho para novas soluções e pesquisas avançadas no tema.

O restante do artigo está organizado da seguinte forma: A seção 2 apresenta trabalhos relacionados; A seção 3 descreve dataset, modelos e prompts avaliados; Resultados são discutidos na seção 4; Conclusões e trabalhos futuros são apresentados na seção 5.

2. Trabalho relacionados

2.1. Modelos Generativos

Os Modelos de Linguagem (LLMs) são modelos de inteligência artificial treinados em grandes conjuntos de dados textuais, capazes de gerar, compreender e manipular a linguagem natural. LLMs como o GPT-3 [Brown et al. 2020] e seus sucessores demonstraram um desempenho notável em diversas tarefas, tais como tradução, resumo, geração de texto criativo.

Particularmente, LLMs mostraram-se muito úteis para tarefas de classificação *zero-shot* [Kojima et al. 2022]. O grande benefício dessa abordagem é utilizar as capacidades de aprendizado a partir de grandes bases de dados e de raciocínio para a realização de classificação de texto sem necessidade de retreino, somente utilizando técnicas de *prompt* mais acessíveis a seus usuários. Ou seja, LLMs representam a possibilidade de uso de modelos de IA mais genéricos, com menor esforço de desenvolvimento posterior, algo que tradicionalmente envolve muito tempo em projetos de treinamento de modelos específicos.

Já os grandes modelos de visão-linguagem (VLMs) representam um avanço significativo na área da IA generativa combinando o poder da visão computacional com o processamento de linguagem natural. Esses modelos são capazes de compreender e relacionar informações visuais e textuais, abrindo um leque de possibilidades para diversas aplicações, como classificação de imagens, geração de legendas, resposta a perguntas visuais e muito mais. Dentre os VLMs mais populares, destacam-se o *Contrastive Language–Image Pre-training* (CLIP) [Radford et al. 2021] e seus derivados, que utilizam uma arquitetura de aprendizado contrastivo para alinhar representações visuais e textuais em um espaço latente compartilhado. O CLIP foi treinado em um vasto conjunto de dados multimodal, permitindo que o modelo aprenda a associar imagens a descrições textuais de forma eficaz. Tal modelo é base de várias arquiteturas, como por exemplo do SmolVLM [Marafioti et al.] e o Llama Vision [Meta-AI], avaliados neste trabalho.

2.2. Classificação de Imagens

Redes Neurais Convolucionais (CNNs) têm sido amplamente utilizadas em tarefas de classificação de imagens e detecção de objetos, apresentando resultados expressivos em diversas áreas, incluindo monitoramento ambiental [Sá and Figueiredo 2022], análise de imagens médicas [Jr. et al. 2023], inspeção industrial na Indústria 4.0 [Monteiro et al. 2023] e agricultura de precisão [Mendonça and Guedes 2024]. Nessas aplicações, modelos neurais profundos exigem treinamento em grandes conjuntos de dados de imagens, cuidadosamente coletados e organizados para cada caso específico. Esse processo de treinamento demanda tempo significativo e, mesmo alterações sutis nos dados, geralmente requerem a reexecução do treinamento do modelo.

No contexto das VLMs, diversas pesquisas têm explorado o uso de VLMs em tarefas de Visão Computacional, inclusive em classificação de imagens *zero-shot* [Zhang et al. 2024]. Como exemplo, [Saha et al. 2024] explora novas arquiteturas e métodos de treinamento para melhorar a capacidade de generalização dos modelos. Já [Mirza et al. 2024] avalia o desempenho de diferentes modelos e estratégias de *prompting* em *datasets* específicos.

Embora pesquisas anteriores tenham demonstrado que *prompts* bem elaborados podem aprimorar significativamente o desempenho desses modelos, este trabalho propõe uma abordagem alternativa: a utilização de um VLM leve para gerar descrições textuais de objetos, seguido por um LLM mais robusto para realizar a classificação com base nessas descrições. Essa estratégia busca equilibrar a dependência do conhecimento implícito dos modelos com sua capacidade de raciocínio.

3. Método proposto

Esta seção apresenta a descrição do conjunto de dados, modelos e *prompts* usados para a geração de resultados.

3.1. Dataset

Para avaliação dos modelos, foi utilizado o dataset público de imagens para a tarefa de classicação disponível em `www.kaggle.com/datasets/mauriciofigueiredo/amazon-fruits-small`. O mesmo consiste de imagens de frutas amazônicas em contexto (objetos e cena), composto de 6 classes, conforme exemplificado na Figura 1. O dataset possui 120 imagens balanceadas entre as classes, sendo 75% para treino e 25% para teste.

O dataset foi escolhido por sua simplicidade, fornecendo dados sobre objetos de entendimento comum, por ter complexidade de imagens compatível a outros datasets em problemas

de classificação, mas também por possuir alguns dos objetos dificilmente encontrados em dados públicos para pré-treino de modelos genéricos. Por exemplo, as frutas pupunha e tucumã, são características da região amazônica e não são facilmente encontradas em outros datasets. Com isso, pode-se avaliar tanto o conhecimento embutido em grandes modelos pré-treinados quanto sua capacidade de racínio sobre as imagens.

Figura 1: Exemplos de imagens das classes em contexto.

3.2. Modelos e Prompts

Com o objetivo de explorar diferentes modelos e abordagens de classificação *zero-shot*, configuramos os seguintes cenários de avaliação.

3.2.1. Rede Neural Convolucional (Baseline).

Como referência de resultados para os modelos VLM, usamos a abordagem clássica para detecção de imagens partindo do treino de uma rede neural convolucional. Para isso, usamos uma transferência de aprendizado da rede InceptionResNetv2 pré-treinada no dataset ImageNet [Szegedy et al. 2016], disponível no pacote keras do tensorflow (https://keras.io). Essa rede foi treinada na partição de treino do dataset de avaliação com aumento de dados. O aumento de dados utilizado consiste do uso de adaptações aleatórias das imagens originais com deslocamento horizontal e vertical máximo de 20%, rotação máxima de 20 graus, zoom máximo de 20% e *flip* horizontal.

3.2.2. VLMs com prompts Diretos

Neste caso, avaliamos VLMs pré-treinados de diferentes capacidades e tamanhos. Usamos prompts diretos para a geração de respostas entre as classes existentes no dataset. O objetivo é avaliar o conhecimento inerente desses grandes modelos, sem necessídade de raciocínio. Os modelos testados são descritos a seguir.

SmolVLM-Instruct [Marafioti et al.]: Modelo multimodal leve e versátil que combina entradas de imagem e texto para gerar texto. Ele consegue realizar diversas tarefas, como responder perguntas sobre imagens e descrever conteúdo visual. Sua eficiência o torna ideal para uso em dispositivos móveis.

Llama-3.2-vision [Meta-AI]: Modelo aberto da Meta otimizado para reconhecimento visual, raciocínio sobre imagens, legendagem e resposta a perguntas gerais sobre uma imagem. Dentro os modelos abertos, este promete o melhor desempenho perante outros disponíveis. Embora o mesmo possua diferentes tamanho e capacidades, usamos sua versão com 11 bilhões de parãmetro, mantendo maior proximidade de recursos computacionais com o SmolVLM.

Gemini-Pro Vision 1.5 [Gemini-Team 2024]: Modelo multimodal da Google com reconhecida capacidade de entendimento multimodal. Modelo fechado, muito grande e para tarefas complexas, mas que pode ser acessado via API para uso em dispositivos diversos.

Para todos esses modelos o prompt dado em inglês, devido ao enfoque dos mesmos na língua internacional, como pergunta direta foi: ”What is this fruit? Respond in one word.The options are [açaí, cupuaçu, graviola, guarana, pupunha, tucumã]”.

3.2.3. VLMs com prompts descritivos.

Assumindo que o resultado dos VLMs podem ser limitados pelo desconhecimento das classes pelos seus nomes introduzidos no prompt direto, testamos a capacidade de raciocínio do modelo caso introduzirmos a descrição dos objetos a serem detectados. As descrições dos objetos foram feitas de forma simples, partindo do princípio de uma comando que seria dado por uma pessoa comum. Usamos o modelo SmolVLM, por ser o mais econômico e com resultados limitados da primeira estratégia. O prompt utilizado foi:

”Considering that tucumã is a round fruit with a green skin and yellow pulp. Considering that pupunha is a fruit with a shiny yellowish or reddish skin. Considering that cupuaçu is a fruit with a velvety brown skin and white pulp. Considering that guaraná is a fruit with red skin but usually shows its large white seed with a black dot. Considering that graviola is a green fruit with a spiny appearance. Considering that açaí is a fruit that grows in bunches, is round, and has a purple color. Answer in one word. The answer options are: [açaí, cupuaçu, graviola, guarana, pupunha, tucumã]”.

3.2.4. Descrições VLMs e Raciocínio LLMs.

Ao observar que a descrição dos objetos pode ajudar a qualidade das respostas dos VLMs, mas ainda apresentando dificuldades, surgem as hipóteses de que ou a descrição do objeto no prompt não está alinhada com a base de conhecimento do modelo pré-treinado ou o mesmo não tem uma capacidade de raciocínio conjunto imagem-texto. Para avaliar este caso, foi criado uma arquitetura híbrida VLM+LLM, onde o VLM somente descreve as características físicas dos objetos da imagem e um LLM com maior capacidade de raciocínio as relaciona com as descrições dos objetvos. No caso, usamos o SmolVLM descrito anteriormente com o prompt:

”Describe the physical characteristics of this fruit”.

Em seguida, usamos o LLama-3.2 [Grattafiori 2024], leve, com 3 bilhões de parâmetros, com o prompt:

”What fruit has similar characteristics to the description? DESCRIPTION. Considering that tucumã is a round fruit with a green skin and yellow pulp. Considering that pupunha is a fruit with a shiny yellowish or reddish skin. Considering that cupuaçu is a fruit with a velvety brown skin and white pulp. Considering that guaraná is a fruit with red skin but usually shows its large white seed with a black dot. Considering that graviola is a green fruit with a spiny appearance. Considering that açaí is a fruit that grows in bunches, is round, and has a purple color. Answer in one word. The answer options are: [açaí, cupuaçu, graviola, guarana, pupunha, tucumã]”.

4. Resultados

De acordo com os modelos e estruturas de prompts descritos na seção anterior, apresentamos os resultados obtidos executando os modelos nos dados de testes do dataset de exemplo. A Tabela 1 apresenta as acurácias de todos os cenários testados, bem como apresenta o tamanho

dos modelos testados, em termos de quantidade de parâmetros, para fins de comparação. As subseções a seguir, detalham cada cenário por meio de suas matrizes de confusão.

Tabela 1: Resultados dos experimentos de classificação de imagens.

Modelo	Acurácia (%)	Tamanho (GB)
Baseline	86,67	56M
SmolVLM prompt direto	20	2B
Llama-Vision prompt direto	73,33	11B
Gemini Pro-Vsion prompt direto	76,67	120B (estimado)
SmolVLM prompt descritivo	33,33	2B
Llama-Vision prompt descritivo	43,33	11B
SmolVLM + LLama3.2	63,33	2B + 3B

4.1. VLMs com prompts diretos

A Figura 2 mostra a matriz confusão de todos os modelos avaliados. Como pode-se observar, o modelo convolucional treinado especificamente na partição de testes do dataset tem um desempenho superior na acurácia, de 86,67%, e um bom equilíbrio entre as classes, com exceção da classe tucumã, mostrando também uma certa complexidade inerente do dataset. O modelo SmolVLM apresentou o pior desempenho (20%), com respostas enviesadas a uma única classe, produzindo ainda uma classificação fora das classes possíveis, indicada na matriz como *none*. Particularmente, o modelo respondeu uma imagem de açaí como blueberry, que de fato se parecem, mas não era uma opção passada no prompt.

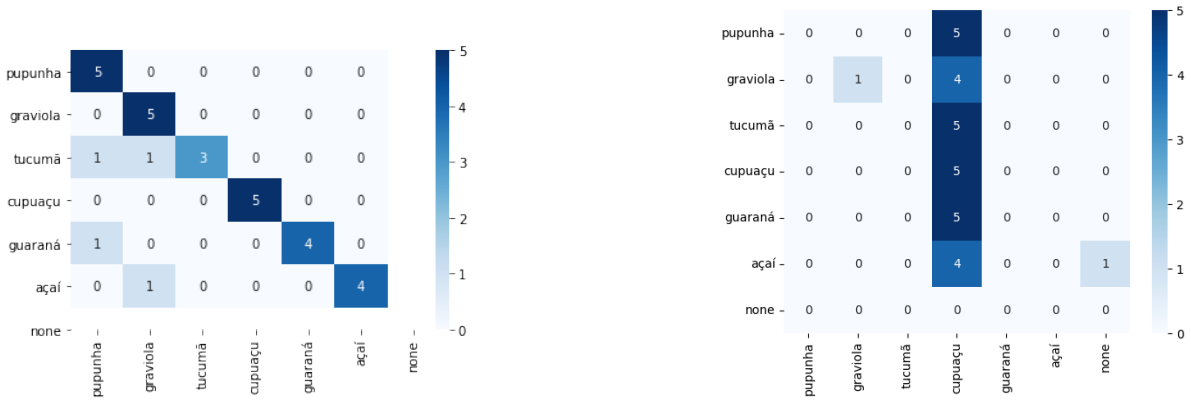
Já os modelos Llamma-Vision e Gemini-Pro-Vision ficaram mais próximos do baseline em termos de acurácia, com 73,33 e 76,67%, respectivamente. Ambos reconheceram a maioria das classes, mas o Llama-Vision parece desconhecer uma das classes do dataset, o tucumã.

Esses resultados mostram que os modelos maiores possuem um conhecimento inerente importante sobre as características das imagens, bem como de seus nomes presentes no prompt direto. Ainda, percebe-se a viabilidade de classificação Zero-Shot mesmo em um dataset com elementos muito específicos de uma região, dificilmente encontrados em datasets específicos de objetos. Ou seja, há a indicação de uso de dados de imagens diversos e em grandíssima escala no treino desses modelos.

4.2. VLMs com prompts descritivos

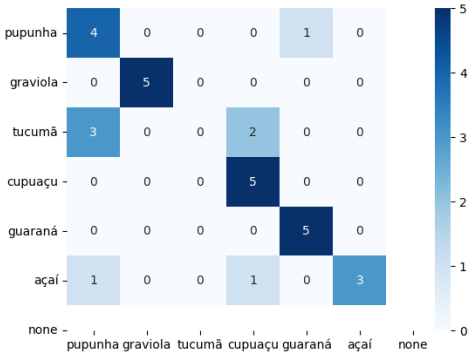
Como observado no primeiro experimento, há a hipótese de algumas classes serem desconhecidas para um determinado VLM, fato esse observado principalmente com o SmolVLM. Dessa forma, o prompt descritivo pode ajudar na classificação zero-shot aproveitando-se da capacidade de raciocínio dos modelos. Para efeito de comparação, mostramos na Figura 3 os resultados do SmolVLM e Llama-Vision, apenas.

Para o SmolVLM, a estratégia elevou os acertos de 20 para 33,33%, comprovando que sua limitação estaria na base de conhecimento pré-treinada, Isso faz sentido pelo tamanho muito reduzido do mesmo. No entanto, percebe-se que o desempenho ainda é muito inferior em relação ao baseline. Já o Llama-Vision, piorou sua métrica geral, de 73,33 para 43,33%, ainda não reconhecendo a classe tucumã, mas aumentando as confusões com as demais classes. Esse fato pode indicar que o conhecimento implícito de um modelo pode ser mais relevante do que sua capacidade de raciocínio para uma tarefa, embora possamos questionar que prompts descritivos melhores pudessem elevar a capacidade dos modelos ao custo de maior esforço humano para tal.

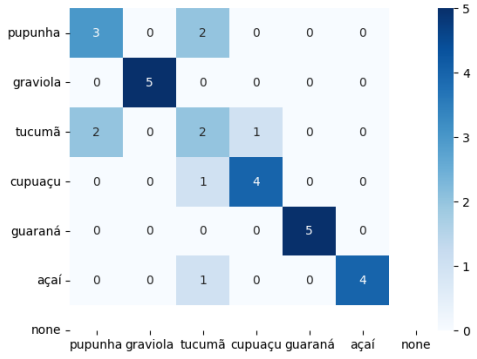


(a) Baseline

(b) SmolVLM

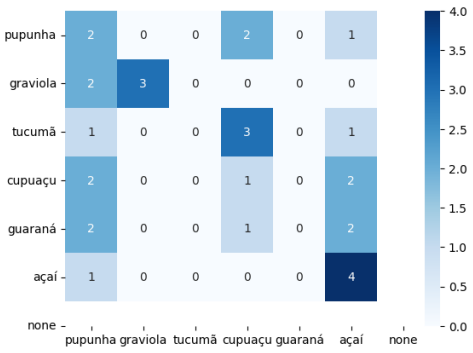


(c) Llamma Vision

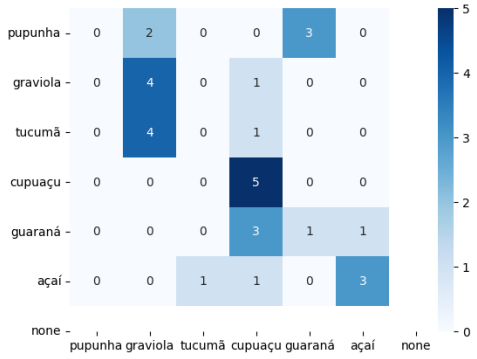


(d) Gemini Pro Vision

Figura 2: Resultados dos modelos com prompts diretos.



(a) SmolVLM



(b) Llamma Vision

Figura 3: Resultados dos modelos com prompts descritivos.

4.3. Descrições VLMs e raciocínio LLMs

Diante da hipótese levantada no experimento anterior, de que um esforço maior para descrever os objetos por meio de seus prompts poderia ajudar a capacidade de um VLM, avaliamos o SmolVLM apenas como um gerador de descrições textuais, submetendo posteriormente essa descrição a um LLM para a classificação zero-shot. Escolhemos o SmolVLM para esta situação devido à necessidade de uso de dois modelos, aumentando o custo de predição. Dessa forma, uma aplicação já poderia ter um LLM para tarefas de texto, então acrescentaríamos um

VLM pequeno no pipeline para acrescentar apenas a habilidade de trabalhar com imagens.

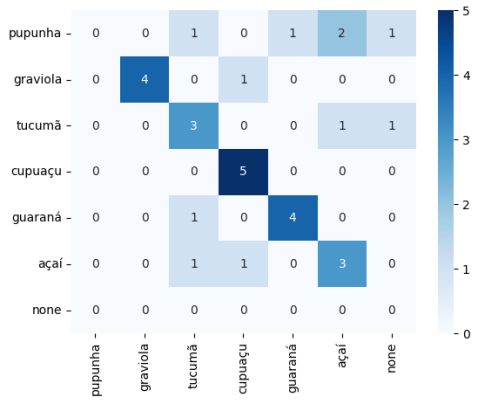


Figura 4: Resultados dos modelos com SmolVLM e Llama3.2.

Observa-se na Fig. 4 que o desempenho melhorou em comparação ao experimento anterior, onde a acurácia aumentou de 33,33 para 63,33%. Isso significa que o modelo pequeno SmolVLM foi incapaz de reconhecer as classes pelo nome, mas foi capaz de gerar descrições mais genéricas das mesmas, pelas suas características físicas. Algumas descrições obtidas do SmolVLM são apresentadas na tabela 2. Podemos observar que, somente pela descrição, há chance de confusão de Pupunha com Tucumã, por exemplo, explicando um pior resultado para essas classes e apontando esse problema como uma limitação da abordagem.

Tabela 2: Exemplos de descrições do SmolVLM.

Classe	**Descrição**
Pupunha	*It is round and orange in color, with a smooth, shiny surface. It is about the size of a small orange. The fruit is covered in small, sharp thorns.*
Graviola	*The fruit is large, green, and has a bumpy texture. It is split into two halves, revealing a white interior with brown seeds.*
Tucumã	*The fruit is round and has a smooth, shiny surface. It is orange in color, with a slightly lighter shade on the top. The fruit is about the size of a golf ball.*
Cupuaçu	*The fruit is large, round, and brown. It has a thick, rough, brown skin. The inside of the fruit is white and has a large, round, white mass in the center.*
Guaraná	*The fruit is red and round, with a small black dot in the center of each fruit. The fruit is surrounded by green leaves.*
Açaí	*The fruit is small, round, and has a dark purple color. It is likely a type of berry or plum.*

4.3.1. Discussão dos resultados

Observa-se que os maiores modelos (Gemini Pro Vision, com estimados 120 bilhões de parâmetros, e Llama-Vision, com 11 bilhões de parâmetros), construídos sobre uma base de conhecimento maior, têm os resultados mais competivos com o baseline (aprox. 10%). Por outro lado, o modelo mais econômico, SmolVLM (2 bilhões de parâmetros), teve desempenho bem inferior independente da estratégia de prompt. É natural assumir que um modelo maior tem mais conhecimento inerente, sendo essa a alternativa prioritária para modelos Zero-Shot de alto desempenho. No entanto, esses modelos podem não ser viáveis a dispositivos embarcados e de computação na borda.

No entanto, os resultados mostram a possibilidade de ter modelos menores, mas exigindo maior exploração das suas características de raciocínio por meio de prompts melhor elaborados. A ideia ilustrada neste trabalho, tanto com um VLM com prompt descritivo quanto da solução VLM+LLM, é a de explorar a capacidade das VLMs em entender características físicas genéricas dos objetos de interesse (formas, cores, tamanhos etc.) Com soluções VLMs puras, mostra-se a necessidade de maior esforço do desenvolvedor para elaboração de prompts. Na solução VLM+LLM, mostra-se que descrições genéricas obtidas por uma VLM poderiam alimentar algum pós-processamento mais leve para a classificação das imagens. Vale notar que a solução VLM+LLM tem quantidades de parâmetros somados ainda menor que o Llama-Vision (5 bilhões vs. 11 bilhões de parâmetros), podendo ser alternativa interessante com restrições computacionais intermediárias.

O desempenho superior do baseline, treinado especificamente para o dataset em questão, tanto de acurácia quanto de tamanho do modelo, mostra que quando o esforço de coleta de dados e treino é viável, essa solução deve ser preferencial. Mas os resultados observados indicam que os avanços rápidos das VLMs têm potencial para diminuir essa margem.

5. Considerações Finais

Este artigo mostrou que grandes VLMs, treinados em grandes quantidades de dados, têm conhecimento inerente importante, podendo ter desempenhos competitivos a modelos especializados na tarefa de classificação de imagens. No entanto, necessitam de recursos computacionais elevados, podendo dificultar a adoção em soluções embarcadas e de borda. Para esses casos, o trabalho mostra que pequenos modelos podem se adaptar melhor ao adotarem melhores estratégias de prompt. Particularmente, mostramos uma solução interessante, composta por um VLM para gerar descrições de imagens e um LLM para classificação zero-shot das descrições, ambos pequenos. Tal solução tem potencial de ser competitiva caso soluções embarcadas já assumam a necessidade de LLMs pequenas.

Como trabalhos futuros, vislumbramos: (i) a experimentação dos modelos em datasets mais desafiadores em termos de descrição dos objetos de interesse; (ii) exploração de melhores descrições nos prompts, avaliando a capacidade de raciocínio das VLMs; (iii) a realização de estratégias de refinamento de modelos pré-treinados para especializá-los na tarefa de descrever genericamente imagens de interesse, dispensando maior conhecimento implícito de objetos, mas servido de base para outras estratégias de classificação.

Referências

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gemini-Team (2024). Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*.

Grattafiori, A. (2024). The llama 3 herd of models. *arXiv2407.21783*.

Jr., A., Filho, A., Sabino-Silva, R., and Carneiro, M. (2023). Convolutional neural networks for the molecular detection of covid-19. In *Anais da XII Brazilian Conference on Intelligent Systems*, pages 51–62, Porto Alegre, RS, Brasil. SBC.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436—-444.

Liu, S., Yu, S., Lin, Z., Pathak, D., and Ramanan, D. (2024). Language models as black-box optimizers for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12687–12697.

Marafioti, A., Noyan, M., Farré, M., Bakouch, E., and Cuenca, P. Smolvlm - small yet mighty vision language model. `https://huggingface.co/blog/smolvlm`. Acessado em Janeiro de 2025.

Mendonça, A. and Guedes, E. (2024). Classificação e detecção inteligentes de grãos para agricultura digital na cultura do milho. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

Menon, S. and Vondrick, C. (2022). Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Meta-AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. Acessado em Janeiro de 2025.

Mirza, M. J., Karlinsky, L., Lin, W., Doveh, S., Micorek, J., Kozinski, M., Kuehne, H., and Possegger, H. (2024). Meta-prompting for automating zero-shot visual recognition with llms. In *European Conference on Computer Vision*, pages 370–387. Springer.

Monteiro, G., Camelo, L., Aquino, G., Fernandes, R. d. A., Gomes, R., Printes, A., Torné, I., Silva, H., Oliveira, J., and Figueiredo, C. (2023). A comprehensive framework for industrial sticker information recognition using advanced ocr and object detection techniques. *Applied Sciences*, 13(12).

Pratt, S., Covert, I., Liu, R., and Farhadi, A. (2023). What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Saha, O., Van Horn, G., and Maji, S. (2024). Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17552.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv:1602.07261*.

Sá, T. and Figueiredo, C. (2022). Self-driving vessels: Yolov5 approach for water surface object detection. In *Anais do XIV Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 31–40, Porto Alegre, RS, Brasil. SBC.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2024). Vision-language models for vision tasks: A survey.