Anais 2025
27ª Semana de Pesquisa da Universidade Tiradentes
"Ciência, Água e Vida: Inovação para transformar o clima do amanhã"
03 a 07 de novembro de 2025
ISSN: 1807-2518

O Delta Semântico: Validando a Generalização de VLMs em Domínios Sintéticos

Ana Carolina Andrade Passos[1,2] (PIBIC/CNPq);
Victor Flávio de Andrade Araújo[1,2,3,4] (Orientador)
ana.apassos@souunit.com.br;

[1]UNIT - Universidade Tiradentes/Ciência da Computação/Aracaju/SE.
[2]GPTICS - Grupo de Pesquisa Interdisciplinar em Tecnologia, Computação e Sociedade
[3]INCT-SANI - Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva
[4]INCT-SIM-AI - National Institute of Science and Technology in Simulation and Monitoring for Individual Assistance in Extreme Climate Events

10300007 - Ciência da Computação; 10301003 – Teoria da Computação

RESUMO

Modelos de Visão-Linguagem (VLMs), como o CLIP ViT-B/32 (RADFORD et al., 2021), demonstraram capacidade de realizar classificação *zero-shot* através do seu alinhamento entre imagens e descrições textuais, sendo aplicados em diversas tarefas de visão computacional (YOKOYAMA et al., 2025). Com a intensificação de eventos climáticos extremos, como inundações urbanas, mecanismos como o CLIP oferecem potencial para análise automatizada de imagens em sistemas de monitoramento e resposta a desastres (JIANG et al., 2021). Contudo, o treinamento desses modelos é predominantemente baseado em coleções de fotos da Internet, o que levanta dúvidas sobre sua robustez e capacidade de generalização. Dessa forma, existe uma lacuna de domínio significativa entre imagens reais e dados gerados sinteticamente, formando um obstáculo para aplicações que dependem de simulações (TOBIN et al., 2017), como o estudo e a simulação virtual de desastres climáticos, incluindo inundações. Diante disso, este estudo tem como objetivo investigar diretamente essa lacuna, avaliando a capacidade de um VLM de generalizar seu entendimento semântico de conceitos complexos, especificamente inundações urbanas, ao ser testado em um conjunto de dados sintéticos gerado na plataforma Unity. Para tal, aplicamos uma metodologia de delta semântico, adaptada de estudos de percepção social (HAUSLADEN et al., 2025), a qual mede a compreensão de atributos específicos de forma mais robusta do que a classificação binária. O delta é calculado subtraindo a similaridade de cosseno de um prompt neutro, como "A foto de uma rua", da similaridade de um prompt adjetivado, como "A foto de uma rua inundada". Utilizamos um conjunto de dados composto por 880 imagens, o qual possui 440 de contextos de "Inundação" e 440 de "Sem Inundação", testando-as contra um conjunto de adjetivos que representam esses cenários, como "parcialmente inundada" e "seco e limpo". Um delta positivo indica que o adjetivo aumenta a similaridade, demonstrando reconhecimento do atributo pelo modelo. A análise quantitativa do delta validou a capacidade de generalização do VLM para o domínio sintético, visto que o modelo demonstrou reconhecer os atributos semânticos em todas as condições principais. Para imagens da categoria "Inundação", a aplicação dos adjetivos correspondentes resultou em delta médio positivo de +2.16 (escala de 0-100), indicando que o atributo enriqueceu a descrição. Inversamente, para imagens "Sem Inundação", a aplicação dos mesmos adjetivos de inundação produziu um delta médio negativo de -2.48, confirmando que o modelo penalizou corretamente a descrição semanticamente incorreta. Os casos de controle de imagens e *prompts* (Inundação/Sem Inundação e Sem Inundação/Inundação), também mostraram valores delta próximos de zero ou negativos, reforçando a capacidade de distinção do modelo. Em síntese, este estudo demonstra que o CLIP ViT-B/32 mantém capacidade de reconhecimento semântico em domínios sintéticos, com deltas que indicam distinção consistente entre contextos de inundação e não-

inundação. Embora a ausência de comparação com imagens reais limite a quantificação da lacuna de domínio, os resultados sugerem potencial promissor para o desenvolvimento de sistemas de resposta a desastres climáticos baseados em simulações.

PALAVRAS-CHAVE: CLIP, domínio sintético, classificação zero-shot

ABSTRACT

Vision-Language Models (VLMs), such as CLIP ViT-B/32 (RADFORD et al., 2021), have demonstrated the ability to perform zero-shot classification through their alignment between images and textual descriptions, being applied across various computer vision tasks (YOKOYAMA et al., 2025). With the intensification of extreme weather events, such as urban flooding, mechanisms like CLIP offer potential for automated image analysis in disaster monitoring and response systems (JIANG et al., 2021). However, the training of these models is predominantly based on collections of photographs from the Internet, which raises questions about their robustness and generalization capacity. Consequently, there exists a significant domain gap between real images and synthetically generated data, forming an obstacle for applications that depend on simulations (TOBIN et al., 2017), such as the study and virtual simulation of climate disasters, including floods. Considering this, the present study aims to directly investigate this gap by evaluating the ability of a VLM to generalize its semantic understanding of complex concepts, specifically urban flooding, when tested on a synthetic dataset generated in the Unity platform. To this end, we apply a semantic delta methodology, adapted from social perception studies (HAUSLADEN et al., 2025), which measures the comprehension of specific attributes in a more robust manner than binary classification. The delta is calculated by subtracting the cosine similarity of a neutral prompt, such as "A photo of a street," from the similarity of an adjectival prompt, such as "A photo of a flooded street." We utilized a dataset composed of 880 images, which includes 440 from "Flooding" contexts and 440 from "No Flooding" contexts, testing them against a set of adjectives representing these scenarios, such as "partially flooded" and "dry and clean." A positive delta indicates that the adjective increases similarity, demonstrating the model's recognition of the attribute. Quantitative analysis of the delta validated the VLM's generalization capacity to the synthetic domain, as the model demonstrated recognition of semantic attributes across all primary conditions. For images in the "Flooding" category, the application of corresponding adjectives resulted in a positive mean delta of +2.16 (on a 0-100 scale), indicating that the attribute enriched the description. Conversely, for "No Flooding" images, the application of the same flooding adjectives produced a negative mean delta of -2.48, confirming that the model correctly penalized the semantically incorrect description. Control cases of images and prompts (Flooding/No Flooding and No Flooding/Flooding) also showed delta values close to zero or negative, reinforcing the model's discriminative capacity. In summary, this study demonstrates that CLIP ViT-B/32 maintains semantic recognition capability in synthetic domains, with deltas indicating consistent distinction between flooding and non-flooding contexts. Although the absence of comparison with real images limits the quantification of the domain gap, the results suggest promising potential for the development of simulation-based climate disaster response systems.

KEYWORDS: CLIP, synthetic domain, zero-shot classification

REFERÊNCIAS/REFERENCES:

- *Yokoyama, Kaname, Chihiro Nakatani, and Norimichi Ukita. "Dynamic Group Detection using VLM-augmented Temporal Groupness Graph." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.*

- *Hausladen, Carina I., et al. "Social perception of faces in a vision-language model." Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. 2025.*

- *Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.*

- *Tobin, Josh, et al. "Domain randomization for transferring deep neural networks from simulation to the real world." 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017.*

- *Jiang, Xin, et al. "Rapid and large-scale mapping of flood inundation via integrating spaceborne synthetic aperture radar imagery with unsupervised deep learning." ISPRS journal of photogrammetry and remote sensing 178 (2021): 36-50.*