Universidade de Pernambuco Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Computação Inteligente

Título: *Self-Organizing Maps* para Aprendizagem *Few-Shot* em Modelos Vision-Language

Orientador – Fernando Buarque de Lima Neto (fbln@ecomp.poli.br)
Co-orientador – Denis Mayr Lima Martins (dmlm@ecomp.poli.br)

Contexto

A aprendizagem *few-shot* (aprender a partir de poucos exemplos) tem se mostrado eficaz em tarefas unilaterais, mas seu desempenho em cenários multimodais, onde texto e imagem devem ser processados conjuntamente, ainda é limitado [1][2]. Modelos *vision-language* como CLIP (*Contrastive Language-Image Pre-training*)[3] oferecem representações robustas, porém o alinhamento entre os espaços visual e textual costuma exigir grandes volumes de dados ou *fine-tuning* extensivo.

O uso modelos de aprendizado auto-organizado, como os *Self-Organizing Maps* (SOMs) [4], pode trazer benefícios significativos por organizar *embeddings* em mapas discretos [5], facilitando a generalização a partir de poucos exemplos. Potenticais benefícios são:

- Melhor generalização em tarefas *few-shot* devido à capacidade dos SOMs de modelar topologias sem supervisionamento.
- Interpretabilidade aprimorada, uma vez que os SOMs fornecem representações espaciais organizadas de dados multi-modais.
- Adaptação dinâmica a novas classes sem retreinamento completo, explorando a plasticidade dos SOMs.

Esta proposta propõe integrar SOMs ao CLIP para melhorar o alinhamento *vision-language* em tarefas *few-shot*, como geração automática de descrições de imagens.

Problema

Em tarefas multi-modais com poucas amostras, modelos como o CLIP podem sofrer com:

- Falta de generalização devido à escassez de dados de treinamento.
- Desalinhamento entre representações visuais e linguísticas, prejudicando a precisão em few-shot learning.
- Dificuldade em adaptar-se rapidamente a novas classes sem retreinamento extenso.

Hipótese

Incorporar SOMs ao pipeline *vision-language* permitirá organizar os *embeddings* visuais e textuais em mapas topológicos que capturam relações semânticas globais, resultando em um alinhamento *cross*-modal mais robusto. Essa estrutura reduz a necessidade de fine-tuning extensivo e melhora a performance em tarefas *few-shot*.

Pergunta Principal

Como a integração de SOMs em um modelo *Vision-Language* (e.g., CLIP) pode melhorar o desempenho em tarefas multi-modais *few-shot*?

Perguntas Secundárias

- Qual é a melhor estratégia de incorporação dos SOMs (e.g., pré-treinamento, fine-tuning incremental, ou uso em tempo real) para maximizar o alinhamento *cross*-modal?
- De que maneira a topologia do mapa SOM influencia a transferência de conhecimento entre domínios visuais e textuais?

- Quais são os *trade-offs* entre desempenho, eficiência computacional e generalização ao incorporar SOMs?
- É possível adaptar dinamicamente o modelo a novas classes com poucas amostras usando SOMs?

Objetivos

Desenvolver um modelo híbrido que combine CLIP e SOMs para melhorar o alinhamento *vision-language* em tarefas few-shot, como image captioning.

Objetivos Específicos

1. Implementar uma arquitetura que integre SOMs ao modelo CLIP para refinamento de representações multi-modais.
2. Avaliar o desempenho do modelo em benchmarks *few-shot* (e.g., Flickr30k, COCO).
3. Analisar a interpretabilidade das decisões do modelo usando visualizações dos SOMs.
4. Comparar o modelo proposto com abordagens tradicionais (fine-tuning puro, meta-learning) em termos de precisão e generalização.

Produtos Esperados

1. Revisão da literatura no tema de alinhamento em modelos *vision-language*.
2. Arquitetura Modular: código aberto em PyTorch que integra CLIP e SOMs.
3. Protótipo de um sistema que facilite a exploração do espaço latente do modelo CLIP utilizando visualizações baseadas em SOMs.

Referências

[1] SILVA-RODRIGUEZ, Julio et al. A closer look at the few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 23681-23690.

[2] WANG, Ran et al. Integrated Image-Text Augmentation for Few-Shot Learning in Vision-Language Models. ACM Transactions on Intelligent Systems and Technology, v. 16, n. 2, p. 1-19, 2025.

[3] RADFORD, Alec et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PmLR, 2021. p. 8748-8763.

[4] KOHONEN, Teuvo. Essentials of the self-organizing map. Neural networks, v. 37, p. 52-65, 2013.

[5] LUO, Alan; YUAN, Kaiwen. Simple Self Organizing Map with Visual Transformer. arXiv preprint arXiv:2503.04121, 2025.