Comentário crítico: *Learning Transferable Visual Models From Natural Language Supervision*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo introduz o CLIP, um modelo de *embedding* multimodal cujo objetivo é alinhar o espaço latente de imagens e textos. Os autores demonstram que o CLIP representou um marco no aprendizado multimodal por ser capaz de realizar esse alinhamento em larga escala, evidenciando que, mesmo sem rótulos tradicionais, é possível obter representações eficazes.

Para viabilizar o treinamento, foi construído um conjunto de dados de grande escala, denominado *WebImageText*, composto por 400 milhões de pares imagem-texto coletados da internet, inspirado no *WebText* utilizado no GPT-2.

O principal ponto do treinamento está na introdução do aprendizado contrastivo, cujo objetivo é maximizar a similaridade entre pares corretos de imagem e texto e minimizar a similaridade entre pares incorretos. Dessa forma, os *embeddings* de imagem e texto tendem a se alinhar e ocupar posições próximas no espaço latente. Além disso, os autores reutilizaram arquiteturas consolidadas — ResNet para imagens e Transformers para texto — mas treinaram os modelos a partir do zero, de modo a adequá-los ao cenário multimodal.

Um aspecto crítico discutido no artigo é o uso de dados massivos coletados da web, que introduz ruído e pode gerar sobreposição com benchmarks. Essa característica dificulta a avaliação precisa da real capacidade de generalização do modelo e levanta preocupações sobre qualidade e curadoria dos dados.

O CLIP inaugura uma nova etapa no aprendizado multimodal, combinando escala de dados, arquiteturas robustas e aprendizado contrastivo para alcançar resultados expressivos em diversas tarefas de visão computacional. No entanto, sua dependência de grandes volumes de dados coletados da web ressalta limitações quanto à qualidade, vieses e confiabilidade dessas informações. Assim, embora represente um avanço técnico notável, o modelo também evidencia a necessidade de maior reflexão sobre transparência, curadoria de dados e implicações éticas no desenvolvimento de sistemas de inteligência artificial em larga escala.