Busca Multimodal em Vídeos: Localização de Conteúdos Baseada em Texto Descritivo

Guilherme G. Silverio[1], **Marcelo A. Mendonça**[2], **Matheus A. de Aguiar**[3]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

1. Abstract

Nos últimos anos, a busca multimodal em vídeos tem recebido grande atenção devido ao crescente volume de conteúdo audiovisual disponível online. Este projeto apresenta um sistema de recuperação de vídeos que utiliza técnicas de aprendizado de máquina para localizar e exibir vídeos com base em descrições textuais fornecidas pelo usuário. O sistema processa a descrição inserida, compara-a com um conjunto de vídeos previamente indexados e retorna os resultados mais relevantes, ordenados por grau de similaridade. A implementação é composta por um backend desenvolvido com FastAPI, responsável pelo processamento e recuperação dos vídeos, e um frontend em React, que permite ao usuário realizar buscas e visualizar os resultados de maneira interativa. Para garantir eficiência e escalabilidade, o sistema utiliza o conjunto de dados MSR-VTT para alinhamento entre texto e vídeo. A solução proposta busca melhorar a acessibilidade de vídeos por meio de descrições textuais, proporcionando uma experiência de busca mais intuitiva e precisa.

2. Introdução

Este trabalho tem como objetivo o desenvolvimento de um sistema de busca multimodal capaz de localizar segmentos de vídeos que correspondam a descrições textuais fornecidas pelo usuário. O sistema utilizará modelos de aprendizado de máquina, como o BLIP juntamente com o CLIP, para mapear texto e vídeo em um espaço compartilhado, permitindo a análise de similaridade semântica. O foco principal do trabalho é fornecer ao usuário vídeos relacionados conforme a descrição fornecida, além de apresentar também a porcentagem de similaridade correspondente entre a descrição e o conteúdo do vídeo.

3. Trabalhos Relacionados

A busca multimodal em vídeos tem se tornado um campo de grande interesse na interseção entre visão computacional e processamento de linguagem natural (NLP). Diversos trabalhos têm sido desenvolvidos para melhorar a correspondência entre descrições textuais e conteúdos audiovisuais, sendo que dois estudos fundamentais para este projeto são o MSR-VTT e o X-CLIP.

O MSR-VTT: A Large Video Description Dataset for Bridging Video and Language introduziu um dos maiores e mais amplamente utilizados datasets para pesquisa em recuperação de vídeos por meio de descrições textuais. Ele contém aproximadamente 10.000 vídeos coletados da internet, abrangendo diversas categorias e acompanhados por múltiplas descrições textuais anotadas manualmente. Esse dataset é essencial para treinar e avaliar modelos de aprendizado profundo capazes de mapear vídeos e textos em um espaço semântico compartilhado. No contexto deste trabalho, o MSR-VTT é utilizado para extrair descrições representativas dos vídeos armazenados, permitindo que a busca seja realizada de forma mais eficiente e alinhada às expectativas do usuário.

O X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval propõe um modelo avançado baseado no CLIP (Contrastive Language-Image Pretraining), adaptado para recuperação de vídeos a partir de texto. O X-CLIP incorpora aprendizado contrastivo multi-granular, permitindo capturar relações detalhadas entre segmentos de vídeo e palavras individuais, resultando em uma correspondência mais precisa entre texto e vídeo. Diferente de abordagens tradicionais, que tratam o vídeo como uma única unidade visual, o X-CLIP analisa diferentes níveis de granularidade, desde frames individuais até sequências mais longas. Esse método é relevante para este projeto, pois busca melhorar a precisão dos embeddings usados na recuperação dos vídeos, garantindo que apenas os resultados mais relevantes e contextualmente adequados sejam apresentados ao usuário.

Ambos os trabalhos fornecem a base teórica e prática para o desenvolvimento deste sistema, combinando a riqueza do dataset MSR-VTT com técnicas avançadas de modelagem de texto e vídeo, como as exploradas pelo X-CLIP. A partir dessas contribuições, este projeto visa integrar um pipeline otimizado para busca multimodal em vídeos, aprimorando a precisão e a experiência do usuário na recuperação de conteúdos audiovisuais.

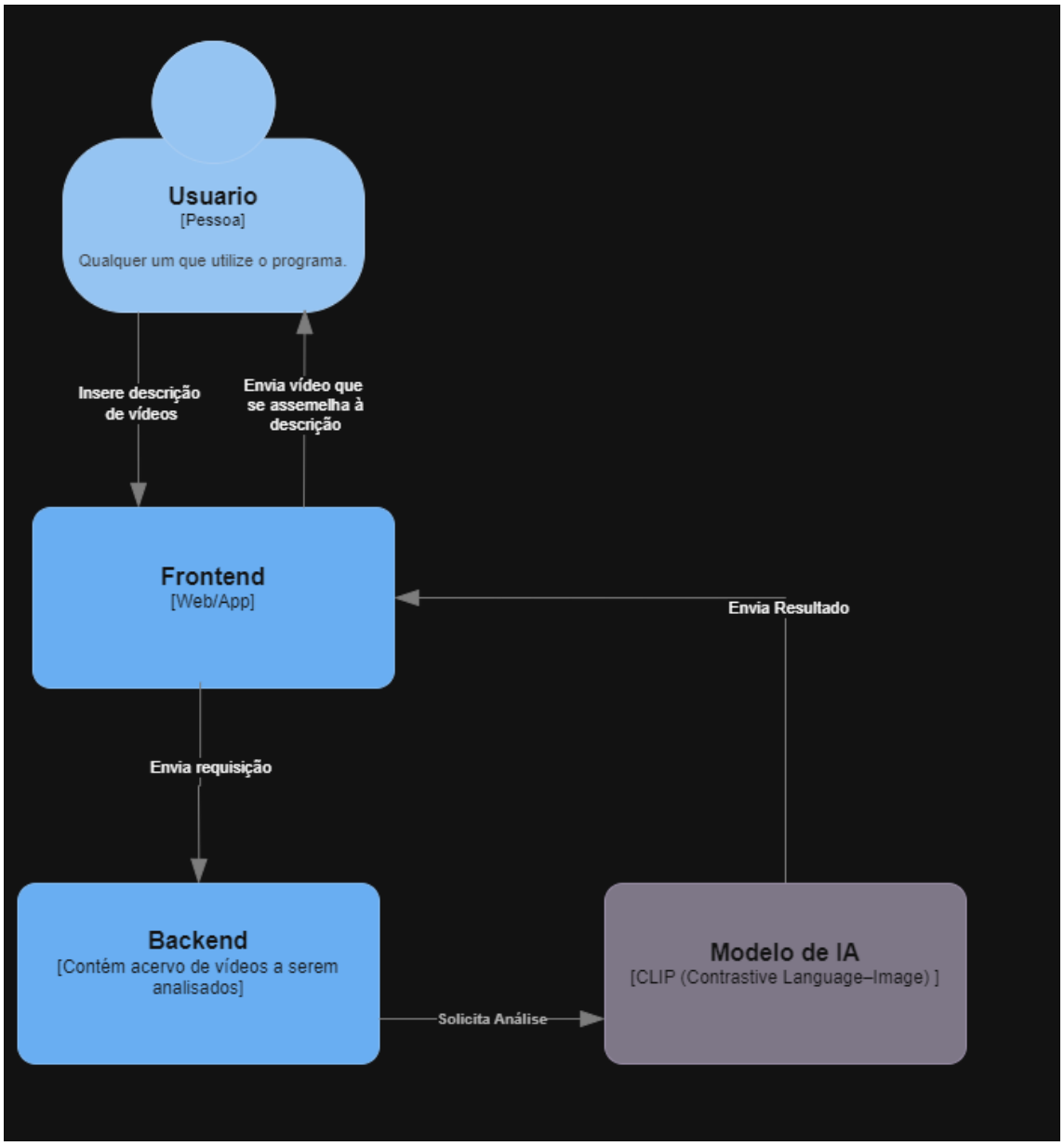
4. Dataset Escolhido

MSR-VTT (Microsoft Research Video to Text)

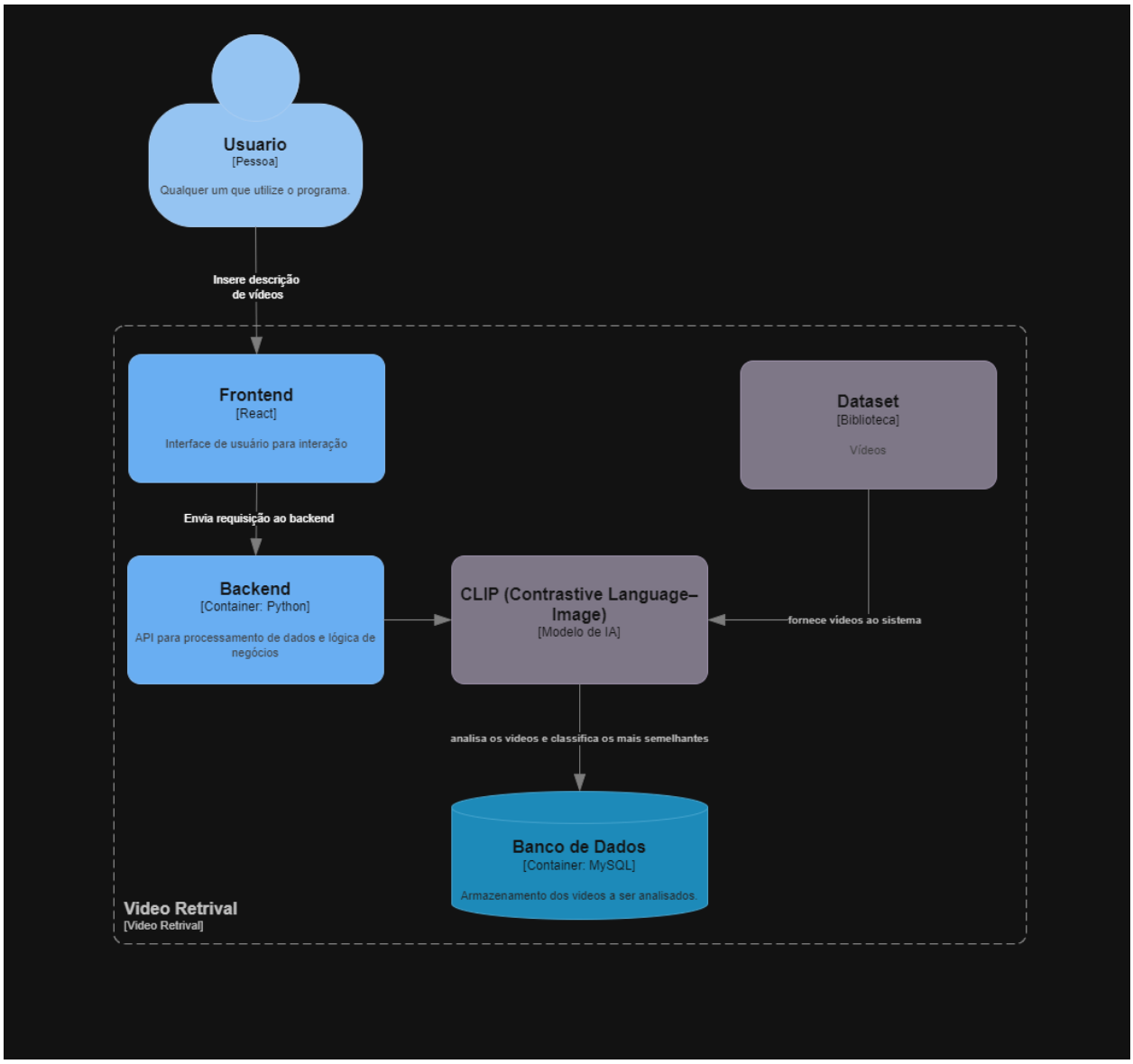
Link dataset

5. Proposta Arquitetural

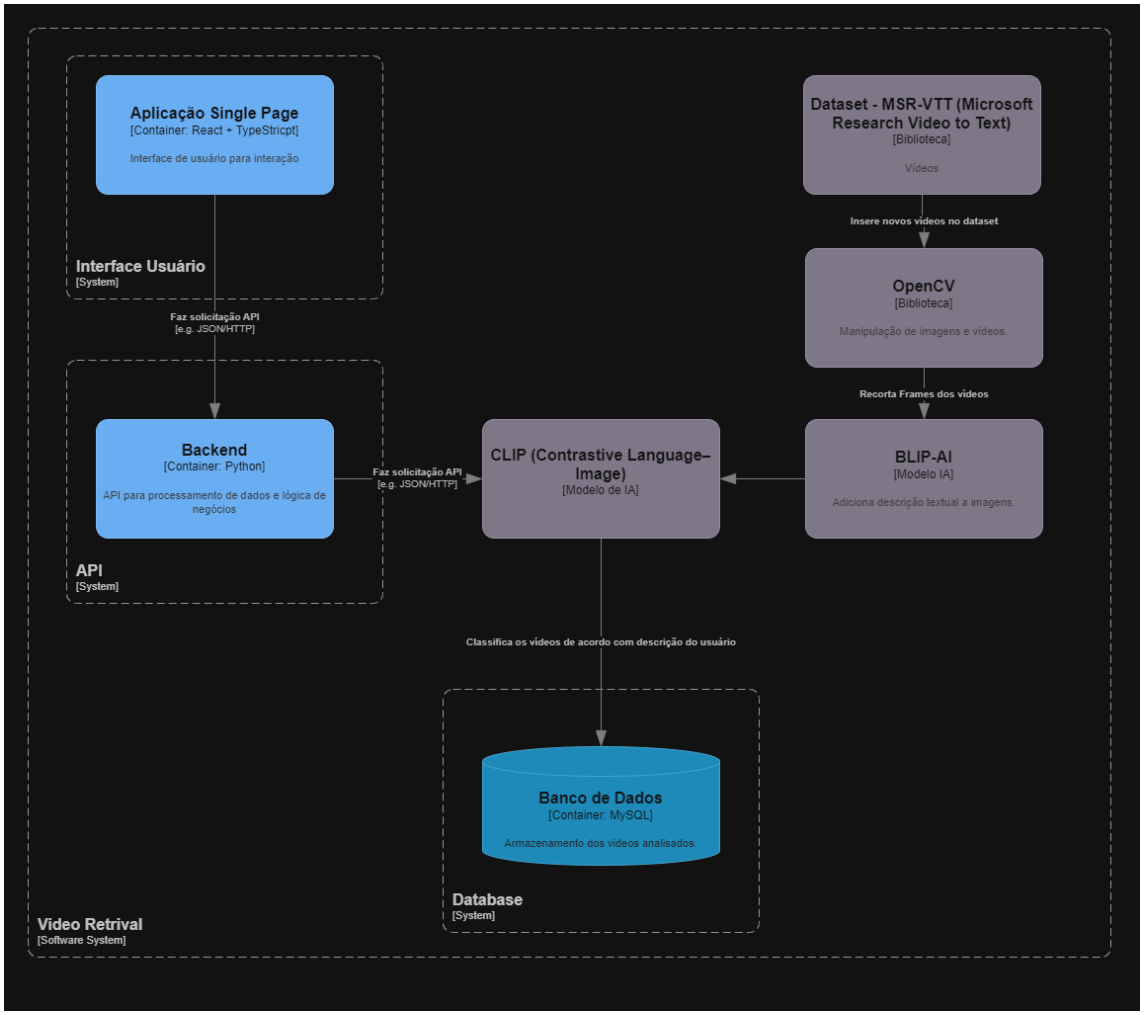
5.1. Diagrama de Contexto



5.2. Diagrama de Container



5.3. Diagrama de Componente



6. Solução Proposta

Nossa solução sobre a proposta se deu a partir da criação de uma API sendo consumida por um frontend, onde disponibilizaria ao usuário um input para inserir a descrição dos vídeos que ele gostaria que fossem trazidos a ele.

Usamos o dataset MSR-VTT para ”povoar” nossa base de dados. Ele é um dataset multimodal, porém não conseguimos encontrar o arquivo que continha as descrições dos vídeos, então tivemos que fazer a descrição dos vídeos usando uma ferramenta chamada BLIP (Bootstrapping Language-Image Pre-training). Ela gera descrições a partir de imagens, portanto, tivemos que usar o OpenCV para extrair os frames dos vídeos, para que o BLIP utilizasse esses frames e então gerasse as descrições.

Com as descrições realizadas, usamos o CLIP para gerar os embeddings visuais e textuais, comparando com as descrições criadas pelo BLIP e assim gerando a similaridade da descrição feita pelo usuário com a descrição de cada vídeo, levando ao usuário os vídeos relacionados juntamente com a porcentagem dessa similaridade.

7. Conclusão

A implementação da nossa API demonstrou resultados satisfatórios ao buscar vídeos semelhantes de acordo com a descrição fornecida pelo usuário. No entanto, alguns pontos importantes foram identificados durante o processo que podem ser aprimorados para melhorar a precisão e a relevância dos resultados entregues.

1. Threshold de Similaridade: Ajustar o threshold de 0.5 pode ajudar a reduzir o número de vídeos irrelevantes apresentados ao usuário, garantindo que apenas vídeos com uma maior correspondência de conteúdo sejam destacados.

2. Qualidade das Descrições Geradas: Utilizar uma melhor qualidade de frames para gerar as descrições, garantindo que as descrições geradas reflitam com maior precisão o conteúdo do vídeo.

3. Eliminação de Descrições Redundantes: Implementar um comparador de descrições para eliminar descrições excessivamente semelhantes, mantendo apenas uma para representar um conjunto de vídeos com características visuais semelhantes.

4. Integração de Análise de Áudio (Transcrição de Áudio com Whisper): Integrar a análise de áudio dos vídeos na comparação das descrições, utilizando ferramentas como o Whisper para transcrever o conteúdo falado dos vídeos e comparar com as descrições textuais.

Essas melhorias podem impactar diretamente na precisão da nossa aplicação, tornando a entrega de vídeos mais relevantes mais eficiente e melhor alinhada com as expectativas do usuário.

8. References

References

[1] Liu, X., Zhang, H., Zhang, X., et al., *X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval*, disponível em: https://arxiv.org/abs/2204.01218, acesso em: Fev. 2025.

[2] Radford, A., Kim, J.W., Hallacy, C., et al., *Multimodal video retrieval with CLIP: a user study*, disponível em: https://arxiv.org/abs/2111.07881, acesso em: Fev. 2025.

[3] Xu, J., Yang, J., Huang, Z., et al., *MSR-VTT: A Large Video Description Dataset for Bridging Video and Language*, disponível em: https://arxiv.org/abs/1609.08675, acesso em: Fev. 2025.