Sumarização Automática de Vídeos: Geração de Descrições Narrativas

Ester Adaianne O. Ferreira[1], **João Pedro de Brito Tomé**[2], **Leonardo C. Filho**[3]

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

{esteradaianne, joaopedrodebritotome, leonardo.cortes}@discente.ufg.br

Abstract. *With the exponential growth of videos on digital platforms, it has become unfeasible to consume and describe all content organically. This project proposes an automated system for generating descriptive and coherent video summaries by combining techniques from computer vision and natural language processing (NLP). To support the proposal, related studies on video summarization were reviewed, and the C4 model was used as the basis for the system's architecture. Preliminary results indicate that the integration of these technologies can enable the automatic generation of accurate and contextualized descriptions, optimizing the consumption and indexing of content on a large scale.*

Resumo. *Com o aumento exponencial de vídeos em plataformas digitais, tornou-se inviável consumir e descrever todo o conteúdo de forma orgânica. Este projeto propõe um sistema automatizado para a geração de sumários descritivos e coerentes de vídeos, combinando técnicas de visão computacional e processamento de linguagem natural (PLN). Para embasar a proposta, foram revisados estudos relacionados à sumarização de vídeos, e o modelo C4 foi utilizado como base para a arquitetura do sistema. Os resultados preliminares indicam que a integração dessas tecnologias pode viabilizar a geração automática de descrições precisas e contextualizadas, otimizando o consumo e a indexação de conteúdo em larga escala.*

1. Introdução

Nos últimos anos, o volume de vídeos publicados e consumidos em plataformas digitais cresceu, impulsionado pela popularização de redes sociais, algoritmos que beneficiam vídeos curtos e interessantes e ferramentas de criação de conteúdo remuneradas [DOMO, 2024]. Com a quantidade de vídeos publicados por minuto em plataformas como o TikTok podendo chegar a mais de 15 mil e mais de 500 horas de conteúdo sendo adicionadas ao Youtube nesse mesmo período de tempo, torna-se inviável consumir e indexar ou sumarizar todo o conteúdo de maneira orgânica. [DOMO 2024, Youtube 2025]. Diante disto, a sumarização automatizada surge como um possível facilitador para o processo de analisar, indexar, resumir e sumarizar vídeos em plataformas digitais. [Tiwari and Bhatnagar 2021]

2. Trabalhos relacionados

A fim de compreender o atual estado da literatura e pesquisa sobre sumarização automática de vídeos e sumarização utilizando técnicas como processamento de linguagem natural (PLN) foram realizadas pesquisas na plataforma Google Scholar utilizando as seguintes strings de busca: ”Sumarização Automática de Vídeos: Geração

de Descrições Narrativas"e multi-video summarization"OR multi-video synopsis"OR "video-summarization"OR "video-synopsis.

A pesquisa resultou em muitos artigos e trabalhos acadêmicos, assim como alguns capítulos de livros abordando o assunto de sumarização de vídeos. Após a leitura dos artigos e trabalhos encontrados que tinham as palavras chaves mais relevantes ao tema, chegou-se à conclusão de que as pesquisas sobre sumarização de vídeos foram bastante utilizadas no contexto de sumarização de vídeos de sistemas de segurança com muitas horas de imagens ociosas ou sem grandes acontecimentos. Há também menção à pesquisa voltada para a sumarização eficiente de emoções em, por exemplo, telejornais. Encontrou-se ainda alguns artigos sobre deep learning na sumarização.

Tamires Barbieri propôs uma abordagem que leva em consideração estratégias humanas para gerar sumários multivídeos relevantes semanticamente para algoritmos de usuários de plataformas digitais. [BARBIERI 2021] Hari K.C. e Manish Pokharel fizeram uma revisão aprofundada sobre sumarização de vídeos utilizando deep learning. [Hari and Pokharel 2024]

Os trabalhos considerados mais relevantes foram devidamente documentados no repositório do github e estarão também mencionados nas referências bibliográficas do relatório.

3. Mapeamento de soluções

Para solucionar a questão da sumarização de vídeos, foram analisadas algumas possíveis soluções, e comparadas as vantagens e desvantagens entre elas. Tal análise foi feita a fim de encontrar a solução que melhor se adequasse aos desafios do projeto, como tempo para conclusão, por exemplo.

3.1. Modelos multimodais avançados

Modelos multimodais integram múltiplas fontes de informação (imagens, texto, áudio) para a geração do sumário, como o GPT e Google Cloud. Modelos multimodais que levam em consideração o contexto prometem gerar sumários mais acertados, uma vez que compreendem a dinâmica temporal e a integração entre imagem e áudio nos vídeos analisados. [Xie et al. 2022]

Algumas das vantagens desta solução são uma boa compreensão global do contexto, já que a junção da análise de várias fontes de informação permitiria a interpretação de nuances, a detecção de eventos ao longo do tempo do vídeo e a redução de ruídos no sumário. Entretanto, essa abordagem demanda grande poder computacional, com custo elevado de armazenamento, datasets grandes e com qualidade e quantidades balanceadas de diferentes formas de informação, além da complexidade de treino e implementação.

3.2. Pré-processamento e uso subsequente de LLM's

Nesta abordagem, há um pré processamento do video em que os frames (imagens) são transcrevidos para texto utilizando visão computacional e o áudio passa pelo mesmo processo através do processamento de linguagem natural. As transcrições dos frames e do áudio são então combinadas e reduzidas a um sumário final através de uma LLM.

Algumas desvantagens dessa abordagem são que o pré processamento de frames e áudio pode ser computacionalmente custoso e vídeos longos podem levar um tempo muito

grande para análise e sumarização. No entanto, esta foi a abordagem escolhida por fornecer uma complexidade de implementação menor que modelos multimodais avançados, o que permite uma implementação mais rápida, além do uso de modelos pré-treinados.

3.3. Resultados observados

Para avaliar a eficácia das abordagens testadas, foram comparados dois sumários gerados a partir do mesmo vídeo: um utilizando um modelo baseado em LLM, feito utilizando GPT e outro gerado pelo método adotado no projeto.

O sumário criado utilizando GPT apresentou uma interpretação coesa da narrativa do vídeo, destacando a evolução do conflito fictício entre os sites de memes e o tom satírico do conteúdo. O modelo demonstrou uma capacidade de síntese eficaz, estruturando o texto de forma clara e identificando elementos-chave da discussão apresentada no áudio. No entanto, não houve menção aos elementos visuais do vídeo, limitando-se à análise textual e auditiva.

Por outro lado, o Sumário criado pelo nosso código integrou informações tanto do áudio quanto dos elementos visuais extraídos do vídeo. Isso permitiu a incorporação de detalhes adicionais, como a presença de imagens repetitivas de logotipos com a palavra ”eq”. A contextualização das imagens fortaleceu a interpretação do vídeo como um todo, proporcionando um sumário mais fiel ao conteúdo multimodal original. Entretanto, esse método exigiu um processo de pré-processamento mais extenso, incluindo análise de frames e transcrição do áudio, resultando em um maior tempo de processamento.

Com base nessa análise, foi constatado que a abordagem baseada em LLM, mesmo que mais cara para uma implementação completa, gera sumários mais ágeis e coerentes para vídeos focados na narrativa verbal, mas pode não capturar plenamente informações visuais. Já a abordagem adotada no projeto, que envolve pré-processamento detalhado, oferece um resultado mais abrangente ao incluir elementos visuais. Os sumários completos podem ser encontrados no Apêndice A.”

4. Proposta arquitetural

Para a documentação da proposta arquitetural para a solução foram utilizados os diagramas do Modelo C4. O modelo C4 foi criado para ajudar equipes de desenvolvimento de software a comunicarem arquiteturas de software.[C4 2025] Este modelo foi escolhido pois permite uma boa visão hierárquica do sistema, facilitando a compreensão da arquitetura proposta.

4.1. Diagrama de contexto

O diagrama de contexto permite uma visão ampla do software, demonstrando com que personas e outros sistemas ele interage.

Figura 1. Diagrama de contexto

4.2. Diagrama de container

O diagrama de container mostra a distribuição de responsabilidades entre os principais módulos do sistema e também como os containers se comunicam entre si.

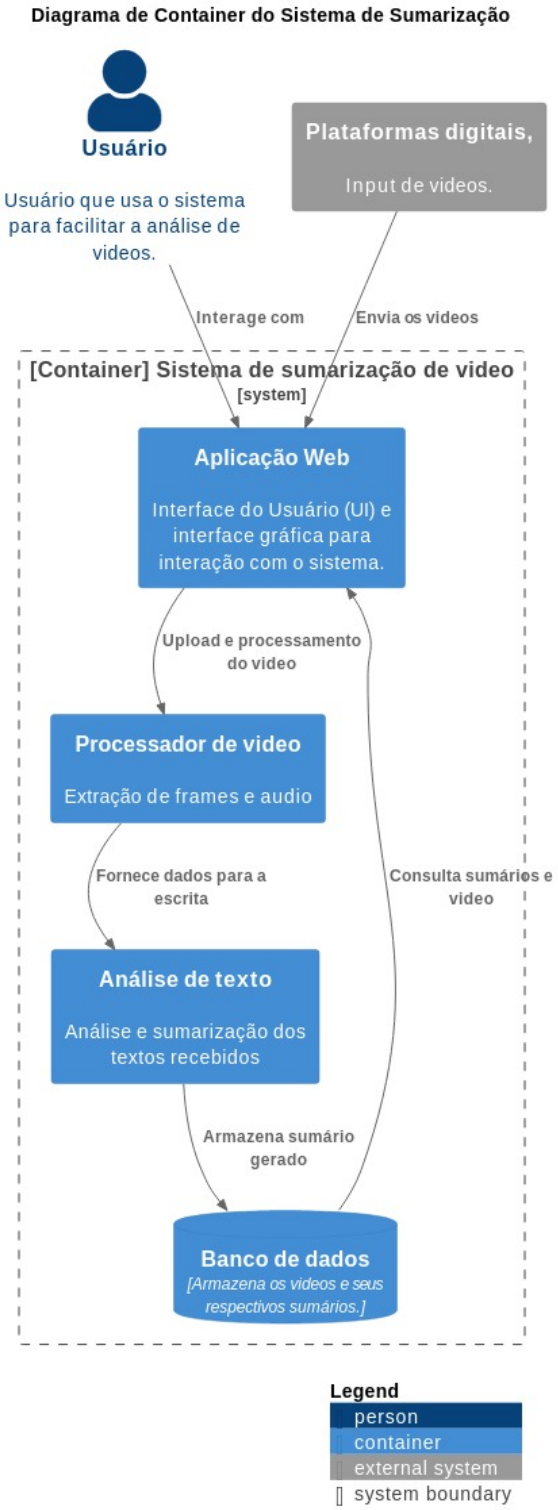

Figura 2. Diagrama de container

4.3. Diagrama de componentes

Mostra de que componentes é feito um container e alguns detalhes relevantes da implementação.

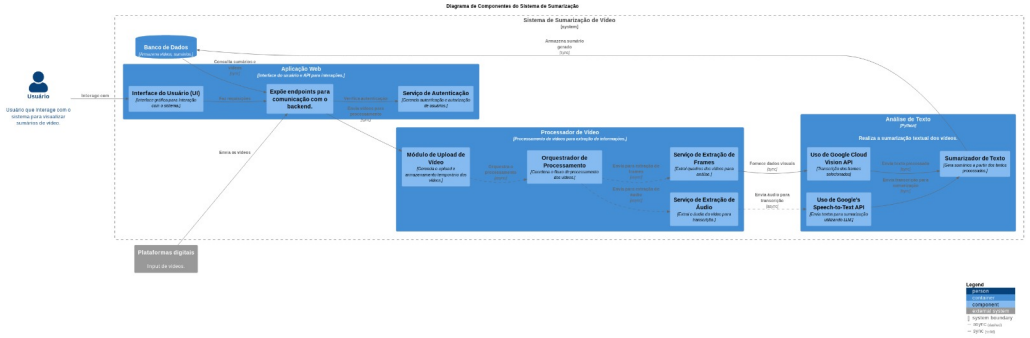

Figura 3. Diagrama de componentes

5. Solução implementada

A solução implementada foi desenvolvida principalmente utilizando a linguagem de programação Python, através da ferramenta Google Collab. Para processamento dos vídeos, foram utilizadas API's para a descrição dos frames e transcrição destes e dos áudios, como BLIP e Whisper.

O vídeo a ser sumarizado é recebido pela aplicação, e para evitar tempo excessivo no processamento foi definido um limite de 100Mb por vídeo. Na etapa de pré processamento, são então extraídas do vídeo informações como tempo de duração, áudio e frames. A extração do áudio em mp3 é realizada pela API ffmpeg, e há configurações sobre a duração dos arquivos de áudio. A extração dos frames é realizada por segundo, ou seja, um número definido de frames por segundo, e tal configuração também pode ser editada.

Para evitar custos computacionais elevados, foi implementada uma fórmula que limita o número de frames extraídos de um vídeo com base na sua duração. Quando o vídeo ultrapassa um tempo máximo predefinido, a extração de frames deixa de ser uma taxa fixa por segundos e passa a ser por uma quantidade fixa de frames por todo o vídeo.

Com os frames extraídos, eles são processados com o uso da biblioteca Pillow e salvos em uma pasta própria. Foi definida uma qualidade máxima de 720p a fim de reduzir o uso de recursos computacionais, mas essa configuração é passível de alteração. O áudio é transcrito pela API Whisper enquanto a API BLIP realiza a descrição textual dos frames, em inglês.

Há duas opções de sumarização: levando em conta frames vizinhos, ou não. Para fins de melhor contextualização, é recomendado que seja feita a sumarização entre frames vizinhos. Neste caso, ollama cria um sumário entre frames vizinhos e é feito um novo vetor para a lista de descrições de frames, o que auxilia na criação de contexto, permitindo uma melhor análise temporal. Caso a sumarização seja por frames vizinhos, o vetor com as novas descrições é usado para gerar uma descrição resumida de todas elas. Caso não, são utilizadas as descrições do processamento inicial.

Por fim, são enviados para ollama as informações: transcrição do áudio e descrição textual resumida dos frames. Então, a LLM cria um sumário final.

6. Trabalhos futuros

Com os resultados preliminares da solução implementada, as principais áreas em que melhorias podem ser feitas como seguimento deste trabalho identificadas foram: aprimoramento do modelo de visão computacional para melhor descrição de frames, aprimoramento do modelo de processamento de linguagem natural para melhor interpretar as descrições de frames e as correlacionar com o áudio, integração com plataformas de vídeo(youtube, tiktok, etc) e redes sociais (X, instagram, etc), análise de prioridade de imagens para melhor qualidade do sumários, o que auxiliaria na implementação de detecção de clickbait.

7. Conclusão

A sumarização automática de vídeos é uma área promissora, especialmente com o aumento exponencial da quantidade de conteúdo audiovisual nas plataformas digitais. A solução proposta, que combina técnicas de visão computacional e processamento de linguagem natural, demonstrou ser eficaz na geração de descrições narrativas automáticas de vídeos, proporcionando resumos contextualmente precisos.

Apesar de ainda haver desafios a serem superados, como o aprimoramento da precisão das descrições e a adaptação a diferentes tipos de vídeos, os resultados obtidos indicam que essa abordagem tem grande potencial para facilitar a indexação e a navegação em grandes volumes de conteúdo audiovisual.

O desenvolvimento contínuo e a expansão das capacidades do sistema são fundamentais para que ele se torne uma ferramenta eficiente e útil para plataformas digitais e usuários, promovendo uma melhor organização e consumo do conteúdo

A. Sumários Gerados

Video utilizado: BATALHA DOS MEMES

A.1. Sumário criado utilizando GPT

O vídeo apresenta uma discussão intensa e exagerada sobre qual site de memes é superior: ”Site dos Memes”ou ”Salve American Meme”. A conversa começa com um debate aparentemente comum, mas rapidamente escala para um nível absurdo, com personagens exaltados, brigas e até consequências globais. Dentre os momentos mais marcantes: Um personagem afirma que ”Site dos Memes”é melhor que ”Salve American Meme”, o que gera uma reação indignada. Outros personagens começam a gritar ao mesmo tempo, tornando a conversa caótica. Em meio à discussão, ocorre um assassinato em ”sangue frio”devido à discordância sobre a qualidade dos sites de memes. O conflito se expande a um nível global, com países como EUA e França apoiando ”Salve American Meme”, enquanto China e Rússia defendem ”Site dos Memes”. O exagero e a dramatização satírica do vídeo refletem um tom humorístico e absurdo, ridicularizando disputas exageradas sobre preferências na internet. O vídeo usa humor nonsense e sátira para mostrar como debates aparentemente banais podem se tornar discussões sem sentido e desproporcionais.

A.2. Sumário criado pelo Projeto

O vídeo apresenta imagens repetitivas de logotipos ou marcas com a palavra ”eq”em diferentes momentos e locais, possivelmente para enfatizar sua presença ou como parte da

identidade visual. Por outro lado, o áudio do vídeo é composto por discussões animadas sobre um conflito entre dois sites: ”salfa, American meme”e ”site dos menes”. Os participantes expressam que o ”site dos menes”é superior a qualquer coisa e mesmo que outros países estejam entrando em guerra por causa do ”salta America memes”, eles continuariam defendendo o ”site dos menes”. Essa discussão se tornou uma batalha entre seguidores dos dois sites, com algumas pessoas afirmando que tal debate não tem sentido pois ambos são bobos e a única coisa certa é o ”site dos menes”. A discussão se tornou tão intensa que alguém foi até mesmo assassinado apenas por falar a favor do ”salta America memes”. No final, um indivíduo anuncia que está saindo da discussão porque ela é bobo e não tem sentido.

Referências

BARBIERI, T. T. d. S. (2021). *Sumarização automática multivídeo baseada em estratégias humanas*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, University of São Paulo.

C4 (2025). C4 model. Technical report, c4model.com.

DOMO (2024). Data never sleeps 12.0. Technical report, DOMO.

Hari, K. and Pokharel, M. (2024). A review on video summarization using deep learning: Approaches, challenges and future direction. *SSRN*.

Tiwari, V. and Bhatnagar, C. (2021). A survey of recent work on video summarization: approaches and techniques. *Springer Nature*.

Xie, J., Chen, X., Lu, S.-P., and Yang, Y. (2022). A knowledge augmented and multimodal-based framework for video summarization. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Youtube (2025). Youtube in numbers. Technical report, YouTube.