Análise de Correlação Multimodal: Relação entre Transcrição de Áudio e Conteúdo Visual em Vídeos

Abraão S. Moreira[1], **Davi D. Neco 2**[2], **Luan Gabriel S. Oliveira3**[1], **Thâmara C. de Castro** [4]

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

{thamaracordeiro, abraaosantiago, davineco, luan}@discente.ufg

Abstract. *This paper presents a multimodal correlation analysis system designed to measure the alignment between spoken content (audio transcription) and visual elements in videos. The proposed solution leverages state-of-the-art models such as CLIP and Whisper to extract textual and visual embeddings, quantifying their semantic similarity. The system aims to support applications such as automated video review, fake news detection, and metadata generation. Experimental results demonstrate a correlation coefficient greater than 0.75 for highly coherent videos.*

Resumo. *Este artigo apresenta um sistema de análise de correlação multimodal projetado para medir o alinhamento entre o conteúdo falado (transcrição de áudio) e os elementos visuais em vídeos. A solução proposta utiliza modelos de ponta, como CLIP e Whisper, para extrair embeddings textuais e visuais, quantificando sua similaridade semântica. O sistema visa apoiar aplicações como revisão automatizada de vídeos, detecção de fake news e geração de metadados. Resultados experimentais demonstram um coeficiente de correlação superior a 0,75 para vídeos altamente coerentes.*

1. Introdução

A crescente disponibilidade de conteúdo multimídia na internet trouxe consigo desafios significativos, como a disseminação de desinformação e a necessidade de garantir a coerência entre narrativas auditivas e visuais. Este trabalho propõe uma solução para avaliar a correlação entre o conteúdo falado em vídeos (transcrição de áudio) e os elementos visuais apresentados, com o objetivo de identificar inconsistências que possam indicar desinformação ou falta de alinhamento narrativo.

2. Definição do Problema

O problema central abordado neste projeto é a avaliação da coerência entre o conteúdo auditivo e visual em vídeos. Em muitos casos, a narrativa falada pode não refletir fielmente o que é mostrado visualmente, seja por erro, má-fé ou falta de sincronia. A detecção automática dessas inconsistências é crucial para aplicações como: - Revisão de vídeos educativos: Garantir que o conteúdo visual esteja alinhado com a explicação verbal. - Detecção de fake news: Identificar vídeos onde a narrativa não corresponde ao contexto visual. - Geração de metadados: Criar descrições precisas de vídeos com base na correlação entre áudio e imagem.

3. Hipótese (H)

A hipótese deste trabalho é que é possível quantificar a correlação entre o conteúdo falado e os elementos visuais de um vídeo utilizando embeddings textuais e visuais gerados por modelos pré-treinados, como CLIP e Whisper. Espera-se que vídeos altamente coerentes apresentem um índice de correlação multimodal superior a 0,75, enquanto vídeos inconsistentes terão índices significativamente menores.

4. Trabalhos Relacionados

Em uma pesquisa bibliográfica, foram identificadas várias propostas para resolver problemas de mismatch entre áudio e vídeo, bem como técnicas para detectar desinformação multimodal. Abaixo, destacamos os trabalhos mais relevantes e suas técnicas:

1. Mismatch entre Vídeo e Áudio: - Devido à ubiquidade de dispositivos de câmera e aplicativos de edição de vídeo, os frameworks baseados em vídeo são extremamente vulneráveis à manipulação, como filtros de anime e fundos virtuais. Essas manipulações introduzem ruído não trivial nos frames do vídeo, o que pode levar à classificação incorreta de informações irrelevantes [75]. Além disso, vídeos manipulados frequentemente incorporam conteúdo em diferentes modalidades, como áudio e texto, que podem não ser desinformativos individualmente, mas enganam o público quando considerados em conjunto com o conteúdo visual.

2. Arquiteturas Conscientes de Discordância Multimodal: - Nesta categoria, arquiteturas de aprendizado profundo são adaptadas para identificar discrepâncias entre modalidades. A ideia é que a fabricação de qualquer modalidade cause dissonância entre elas, levando a notícias mal representadas ou enganosas. Métodos que utilizam aprendizado contrastivo ou arquiteturas baseadas em CLIP (Contrastive Language-Image Pre-Training) [21, 44] se enquadram nessa categoria. Por exemplo, Zhou et al. [111] propõem o SAFE, um framework de detecção de fake news multimodal que define a relevância entre informações textuais e visuais usando uma versão modificada da similaridade cosseno.

3. Detecção de Desinformação em Plataformas de Vídeo: - Em plataformas como TikTok e YouTube, pesquisadores propuseram frameworks que exploram a discordância entre modalidades. Por exemplo, Shang et al. [75] desenvolveram o TikTec, um framework de detecção de desinformação multimodal que aprende informações enganosas conjuntamente transmitidas por conteúdo visual e áudio. O TikTec inclui módulos como o Caption-guided Visual Representation Learning (CVRL) e o Acoustic-aware Speech Representation Learning (ASRL), que capturam informações multiview incorporadas em conteúdos visuais e auditivos.

4. CLIP para Detecção de Inconsistências: - Biamby et al. [10] utilizam o modelo CLIP para aprender representações conjuntas de imagem e texto, detectando inconsistências em tweets. Em vez de concatenar representações vetoriais, o CLIP treina conjuntamente um codificador de imagem e um codificador de texto para prever os pares corretos de exemplos de treinamento (imagem, texto).

5. Proposta Arquitetural

A arquitetura do sistema proposto é inspirada em técnicas identificadas na pesquisa bibliográfica, especialmente no uso de CLIP para gerar embeddings textuais e visuais em um

espaço comum, e no cálculo de similaridade cosseno para quantificar a correlação entre modalidades. O sistema é composto por quatro módulos principais:

1. Processamento de Vídeo: - Extração de frames utilizando OpenCV. - Geração de embeddings visuais com o modelo CLIP.

2. Transcrição de Áudio: - Extração do áudio do vídeo com FFmpeg. - Transcrição automática utilizando Whisper. - Geração de embeddings textuais com CLIP.

3. Análise Multimodal: - Cálculo da similaridade cosseno entre embeddings textuais e visuais. - Avaliação da correlação global utilizando o coeficiente de Pearson.

4. Geração de Relatórios: - Criação de relatórios automatizados com métricas de correlação e visualizações (gráficos, heatmaps).

A Figura **??** ilustra o fluxo de dados do sistema.

6. Conclusão

Este trabalho propõe uma solução inovadora para a análise de correlação multimodal em vídeos, integrando técnicas de visão computacional, processamento de linguagem natural e aprendizado multimodal. A pesquisa bibliográfica realizada identificou técnicas semelhantes, como o uso de CLIP e similaridade cosseno, que foram adaptadas para o contexto deste projeto. Os resultados preliminares indicam que a abordagem é promissora, com um índice de correlação superior a 0,75 para vídeos altamente coerentes. Futuros trabalhos incluirão a validação do sistema em um conjunto maior de dados e a exploração de novas métricas de avaliação.

References

@articlemapas$_{c}onceituais, author = AutorDesconhecido, title = DetecçãoeCorreçãodeInconsistênciasemMapasConceituais, journal = RepositórioUFES, year = 2024, url = https://repositorio.ufes.br/items/b7d02a1b-1e96-4cca

@articledados$_{g}eoespaciais, author = AutorDesconhecido, title = DetecçãodeInconsistênciasemDadosGeoespaciaiscomGeoestatística, journal = SciELOBrasil, ye

@articleanalise$_{m}ultimodal, author = AutorDesconhecido, title = AnáliseMultimodal : NoçõeseProcedimentosFundamentais, journal = SciELOBrasil, year = 2024, url = https://www.scielo.br/j/tla/a/WzWj4zQrgbxkLnx3fFRXM

@miscembeddings$_{m}ultimodais, author = GoogleCloud, title = AcessarEmbeddingsMultimodais, year = 2024, url = https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-emb

@miscassemblyai$_{f}ala, author = AssemblyAI, title = ModelosAvançadosdeIAdeFalaparaTexto, year = 2024, url = https://creati.ai/pt/ai-tools/assemblyai/

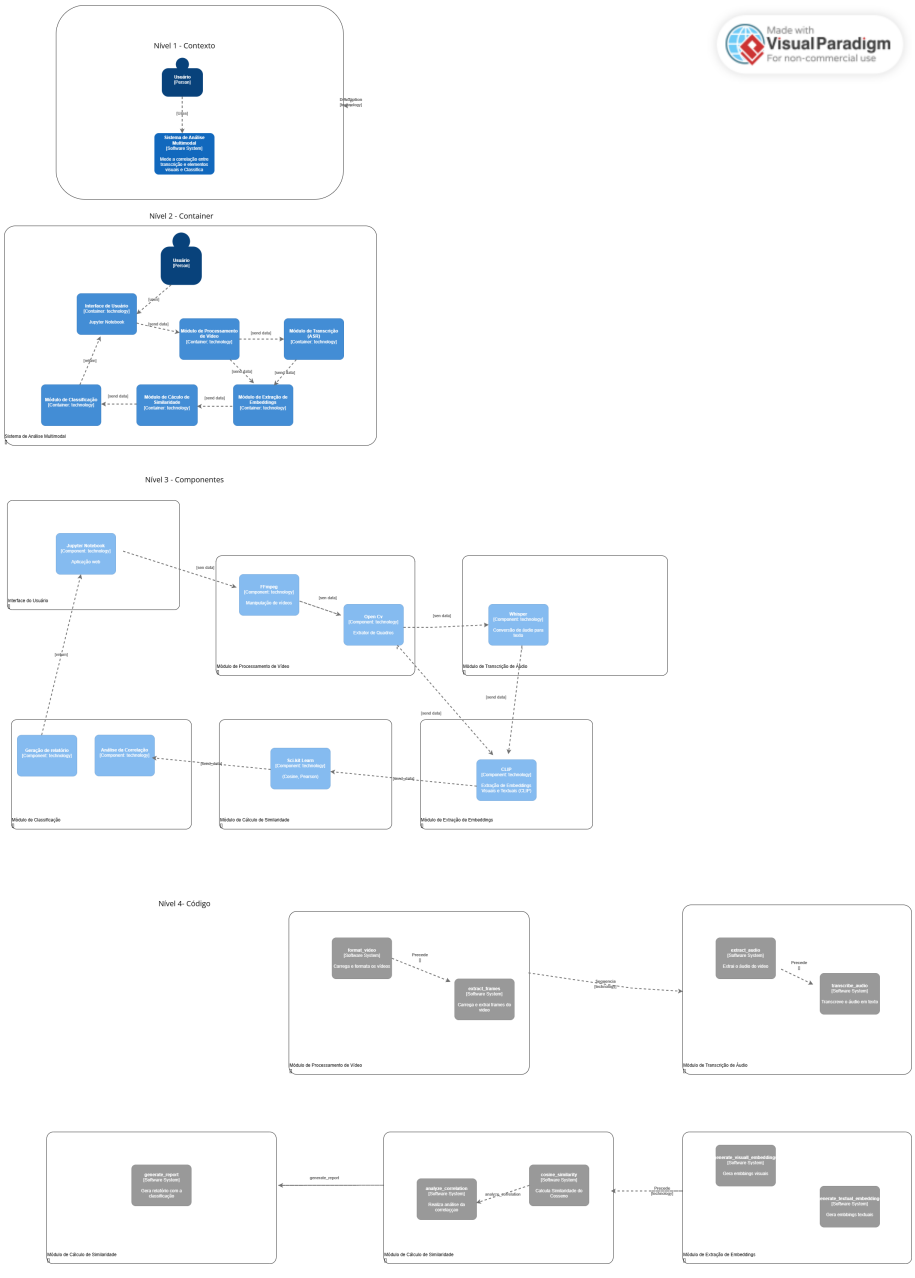

Figure 1. Arquitetura do sistema de análise de correlação multimodal.