Identificacão de Momentos Relevantes em Vídeos do YouTube: Uma Abordagem Multimodal com Arquitetura Transformer e Aprendizado Auto-Supervisionado Utilizando Incômodo Psicoacústico

Gustavo Ribeiro Alves Rodrigues
Universidade Federal de Minas Gerais
Belo Horizonte - MG, Brasil
gustavo.rodrigues@dcc.ufmg.br

***Abstract*—Com o aumento da popularidade das plataformas de mídia de vídeo, tornou-se comum a criação de conteúdos a partir de fragmentos de vídeos mais longos, destacando momentos de grande interesse para os usuários. Esse movimento evidencia a necessidade do desenvolvimento de novas ferramentas que auxiliem os usuários a produzir, consumir e armazenar conteúdos de vídeo de forma eficiente. Este trabalho propõe um estudo de caso que aplica um modelo projetado para a tarefa de sumarização na geração de cortes de vídeo, mantendo informações relevantes para o usuário. O estudo contribui de forma teórica, analisando o impacto de mudanças na tarefa, no dataset e nas configurações de modelos preestabelecidos na literatura, e de forma prática, auxiliando no desenvolvimento de ferramentas que supram a necessidade do consumo eficiente.**

***Index Terms*—video summarization, youtube heatmap estimator**

I. Introdução

Documentos de vídeo são atualmente uma das formas de mídia mais populares na internet, principalmente devido à ascensão das redes sociais como o Youtube e seus videocasts e plataformas de livestream como a Twitch, que utilizam vídeos como sua principal forma de conteúdo. Esses conteúdos costumam possuir longa duração e que faz com que usuários puxem partes do vídeo ou ainda criem cortes, seja para consumi-los de forma mais eficiente seja para criar um novo tipo de conteúdo. Independente da motivação para a busca de criação de conteúdos mais curtos, surge um novo desafio, desenvolver uma forma de realizar a redução de vídeos de longa duração em vídeos curtos de forma automática e com qualidade. Desenvolver formas de realizar essa tarefa pode auxiliar na eficiência não só do armazenamento mas também na forma de consumo de toda essa informação disponível na internet. Devido a isso, existem duas tarefas que podem ser bastante úteis para a resolução desse desafio,a sumarização automática de vídeos e a detecção automática de momentos de interesse são duas abordagens valiosas para enfrentar esse desafio. A sumarização de vídeos visa criar uma versão condensada de um vídeo longo, mantendo as informações essenciais e garantindo que o resumo represente de forma eficaz o conteúdo global do vídeo original. Em contraste, a predição de pontos de interesse busca identificar e extrair fragmentos específicos que atendam aos interesses do usuário, sem necessariamente refletir o conteúdo total do vídeo. Assim, enquanto a sumarização foca em preservar a integridade do conteúdo, a predição de pontos de interesse é mais voltada para atender preferências individuais.

Há diversas abordagens na literatura de como realizar essas tarefas, modelos mais antigos utilizam uma abordagem adhoc e inteligência artificial clássica, enquanto abordagens mais atuais utilizam aprendizado profundo supervisionados baseados em tarefas de processamento de linguagem natural e visão computacional. Contudo, é possível observar uma lacuna na literatura quando o assunto é a utilização de modelos multimodais, especificamente audiovisuais e de aprendizado auto supervisionado. Além disso, os datasets mais utilizados na literatura TVSum e SumMe, referência na tarefa de sumarização, possuem uma quantidade restrita de vídeos e avaliações.

Dessa forma, este trabalho propõe um estudo de caso de um modelo audiovisual utilizando aprendizado supervisionado e auto-supervisionado, baseado no desconforto psicoacústico porposto por Junior [10]. O objetivo é expandir o trabalho realizado, introduzindo alterações e conduzindo novos experimentos com em um conjunto de dados mais amplo. Será feita uma comparação com modelos existentes na literatura para avaliar o desempenho. Finalmente, o modelo resultante será aplicado na tarefa de predição de frames de interesse em vídeos do YouTube, com o intuito de encontrar um preditor robusto que reflita os interesses dos usuários.

Este trabalho está organizado da seguinte forma:

Referencial Teórico e Trabalhos Relacionados: Este capítulo fornece uma base teórica detalhada e revisa a literatura relevante para contextualizar o estudo e identificar os principais avanços na área.

Identify applicable funding agency here. If none, delete this.

Metodologia: Este capítulo descreve a abordagem metodológica adotada, incluindo a descrição dos métodos e técnicas utilizados, bem como o processo de implementação.

Conclusão: Neste capítulo, são discutidas as principais descobertas do trabalho, suas implicações e as possíveis direções para pesquisas futuras.

Bibliografia: O último capítulo apresenta a lista completa de referências e fontes citadas ao longo do trabalho.

II. Referencial teórico

A sumarização de vídeo e a estimativa de momentos de interesse do usuário são tarefas distintas, embora possam compartilhar um objetivo semelhante: reduzir um vídeo longo a um formato mais curto. A diferença reside no tipo de informação considerada relevante em cada tarefa. Na sumarização, o foco é preservar a maior quantidade possível de informação contida no vídeo inteiro, enquanto na predição de momentos de interesse, a prioridade é identificar os fragmentos que são especificamente interessantes para o usuário, sem necessariamente abranger toda a informação disponível. Apesar das diferenças, as abordagens para essas tarefas são bastante semelhantes e, portanto, são de grande importância para este estudo.

Devido a abrangência de artigos nessa área, foram propostos na literatura, um conjunto de protocolos para definir um ambiente comum para a avaliação e comparação dos resultados encontrados. Esses protocolos são sustentados pela utilização de dois datasets utilizados como referência para a área, sendo eles o dataset SumMe *et al.* [6] e o TVSum *et al.* [17]. Os dados são compostos por vídeos de gênero variado, sendo o SumMe composto por 25 vídeos, anotados por 15 avaliadores e o TVSum por 50 vídeos anotados por 20 avaliadores. Por consequência disso, somente serão citados nessa seção, trabalhos que utilizam esse ambiente em comum, um artigo mais completo sobre abordagens para tarefa de sumarização foi descrito por Vivekraj *et al.* [11]. A popularidade do aprendizado profundo auxiliou bastante nas abordagens atuais para a tarefa de sumarização. Grande parte dos modelos propostos para a tarefa de sumarização utilizam RNNs e CNNs. Zhang *et al.* [24] abordam a tarefa de sumarização como uma tarefa de previsão estruturada em dados sequenciais. Para isso, eles utilizam redes neurais recorrentes do tipo Long Short-Term Memory (LSTM) [8] para modelar as dependências de longo alcance presentes na tarefa de sumarização de vídeos. O trabalho realizado por Zhao *et al.* [26] inova ao propor a utilização de uma estrutura hierárquica com duas camadas de RNNs, sendo que a primeira camada codifica cenas curtas extraídas do vídeo original e a segunda camada avalia a confiança de cada sub cena para determinar se ela deve ser incluída no resumo. Essa abordagem é projetada para lidar melhor com a dependência temporal longa entre quadros em vídeos longos e reduz significativamente a carga computacional em comparação com RNNs tradicionais. Já Zhao *et al.* [25] consideram a influência da estrutura do vídeo, composto por cenas, na qualidade da sumarização. Sua abordagem foca na adaptação à estrutura hierárquica dos vídeos, integrando a segmentação de cenas diretamente no processo de sumarização. Dessa forma, eles consideram a organização natural em cenas e quadros, evitando a destruição da estrutura hierárquica e, consequentemente, melhorando a qualidade dos resumos. No estudo realizado por Mrigank *et al.* [16], os cientistas participantes formularam o problema de sumarização como uma tarefa de rotulagem de sequência. Em vez de usar modelos recorrentes como nas abordagens existentes, eles introduzem modelos de sequência totalmente convolucionais, adaptando redes de segmentação semântica, bastante utilizadas em visão computacional. Mohamed *et al.* [3] propuseram uma nova abordagem para gerar resumos automáticos de vídeos longos, baseada na análise do "actionness" (ação) nos vídeos. O estudo se baseia na ideia que quadros que contêm movimentos deliberados realizados por um agente são mais propensos a serem incluídos em resumos gerados por humanos, sugerindo que para gerar resumos automaticamente, é necessário ter um conhecimento implícito da estimativa e classificação do nível de ação nos quadros do vídeo. O trabalho realizado por Yuan *et al.* [23], propõe uma abordagem inovadora, que integra extração de características, modelagem temporal e geração de resumos em uma arquitetura end-to-end. Diferente de trabalhos anteriores, o modelo proposto por eles utiliza uma rede neural convolucional recorrente para modelar simultaneamente a estrutura espacial e temporal dos vídeos. Além disso, o artigo introduz uma nova função de perda, a Sobolev loss, que visa restringir a derivada dos dados sequenciais e explorar a estrutura temporal do vídeo. Já no artigo escrito por Junbo *et al.* [22] é proposto um modelo de memória empilhada, formada por camadas de LSTM e de memória externa onde cada camada LSTM é acompanhada de uma camada de memória. As camadas LSTM e de memória são empilhadas hierarquicamente, integrando as representações aprendidas das camadas anteriores para gerar resumos mais precisos. A abordagem se destaca por capturar dependência temporal de longo prazo para minimizar a redundância temporal.

Após a popularização do mecanismo de atenção Ashish *et al.* [21], modelos que utilizam dessa técnica se popularizaram para tarefas de NLP e consequentemente, essa mesma técnica passou a ser utilizada para a tarefa de sumarização de vídeo.

No trabalho realizado por Yen-Ting *et al.* [13], é proposto uma abordagem inovadora para a sumarização de vídeos que integra um modelo de atenção hierárquico em dois estágios e um mecanismo de atenção de múltiplas cabeças, permitindo gerar diversos mapas de atenção, proporcionando uma análise detalhada das informações contextuais presente nos vídeos. Ping *et al.* [12] incorporam em seu modelo um mecanismo de atenção global diversificado. O SUM-GDA adapta a atenção para considerar as relações temporais entre todos os pares de quadros, permitindo uma captura global da importância de cada quadro dentro do contexto do vídeo completo. Essa abordagem aprimora a geração de resumos de vídeo, produzindo resumos mais diversificados e representativos. Nesse trabalho Litong *et al.* [4] é proposto um modelo que utiliza uma memória externa, criando um modelo de memória aumentada

capaz de prever a importância das cenas com base na visão global do vídeo. Esta abordagem adota um mecanismo de atenção global, que captura informações de todos os quadros do vídeo, proporcionando uma compreensão mais completa do conteúdo. No artigo escrito por Zhong *et al.* [9] a tarefa de sumarização é tratado como um aprendizado de sequência para sequência, onde os quadros do vídeo são a entrada e as cenas-chave selecionadas são a saída. O trabalho utiliza uma camada de encoder e decoder, onde o encoder usa um Long Short-Term Memory Bidirecional (BiLSTM) para capturar informações contextuais dos quadros de vídeo, enquanto o decoder emprega redes LSTM com mecanismos de atenção, utilizando funções-objetivo aditivas e multiplicativas, para imitar o processo humano de seleção de cenas-chave.Um dos mais importantes trabalhos, foi realizado por Evlampios*et al.* [1]. Este trabalho inova ao combinar mecanismos de atenção globais e locais, diferenciando bastante de outros trabalhos e permitindo modelar as dependências entre os quadros em diferentes níveis de granularidade. Além disso, o método incorpora um componente de codificação posicional que fornece informações sobre a posição temporal dos quadros do vídeo, resolvendo uma limitação dos mecanismos de atenção que não levam em conta a ordem de entrada dos elementos.

Um dos principais desafios para a tarefa de sumarização é a dificuldade para encontrar um conjunto de dados anotados para a realização da tarefa. Devido a isso, modelos que utilizam aprendizado não supervisionado começaram a surgir na literatura como uma forma de modelagem para o problema da baixa quantidade de dados que podem ser utilizados. O artigo escrito por Behrooz *et al.* [14] propõe uma estrutura generativa adversarial, composta por duas redes, uma sumarizadora e discriminadora. A sumarizadora é uma rede autoencoder de memória de longo prazo (LSTM) destinada, primeiro, a selecionar quadros de vídeo e, em seguida, decodificar a sumarização obtida para reconstruir o vídeo de entrada. A rede discriminadora é outra rede LSTM destinada a distinguir entre o vídeo original e sua reconstrução pela sumarizadora. A LSTM sumarizadora é treinada como uma adversária do discriminador, ou seja, treinada para confundir maximamente o discriminador, sendo um dos primeiros modelos utilizando GAN para a tarefa de sumarização. Nesse trabalho, Xufeng *et al.* [7] propõem um método inovador utilizando redes neurais adversariais condicionais com um mecanismo de atenção. O método envolve um gerador que produz características de quadros ponderadas e prevê pontuações de importância para cada quadro, enquanto um discriminador distingue entre essas características ponderadas e as características brutas dos quadros. Além disso, o modelo de autoatenção é aplicado buscando capturar dependências temporais de longo alcance em toda a sequência de vídeo, superando as limitações de unidades recorrentes como as LSTMs.

Os modelos anteriores negligenciam uma característica que está presente em boa parte das mídias de vídeos disponíveis na internet, a multimodalidade. Vídeos podem conter tanto informações audiovisuais quanto textuais. Trabalhos que utilizam essa característica possuem o potencial de performar melhor do que modelos que não a utilizam. O trabalho realizado por Medhine *et al.* [15] Propõe um framework único para sumarização de vídeos que abrange tanto a sumarização genérica quanto a sumarização focada em consultas. CLIP-It, o modelo proposto no trabalho, utiliza um transformer multimodal guiado por linguagem que aprende a avaliar a importância dos quadros de um vídeo com base na correlação entre a informação contida na imagem e a consulta definida pelo usuário. Além disso, como as legendas são geradas automaticamente, esse modelo pode ser utilizado no aprendizado não supervisionado, utilizando como critério de relevância, frames que possuem maior correlação com a descrição textual fornecida.

Junior [10], esse trabalho propõe um método inovador de sumarização de vídeos que utiliza informações multimodais, especificamente sinais audiovisuais, para melhorar a qualidade dos resumos gerados. Diferente da maioria dos métodos atuais que focam apenas em dados visuais, o método apresentado integra informações tanto visuais quanto auditivas em uma arquitetura baseada em transformer. O modelo proposto utiliza também um codificador posicional, capaz de capturar informações de localidade. Além disso, a abordagem inclui um novo esquema de treinamento não supervisionado usando pseudo-rótulos derivados de características psicoacústicas dos vídeos.

Um fato importante é que os datasets utilizados SumMe e TVSum, possuem características específicas para a tarefa de sumarização. Isso se deve ao fato da natureza da formulação do dataset, o que de fato poderia ser um problema para o objetivo deste estudo. Dessa forma, para testar o modelo para a tarefa de estimar momentos de interesse de usuário, será utilizado o dataset Mr.HiSum *et al.* [18]. Diferente dos conjuntos de dados referência, esse dataset é formado por mais de 30mil vídeos e rótulos confiáveis, que foram agregados a partir de mais de 50.000 usuários por vídeo. Esses rótulos representam muito melhor o interesse do usuário devido a natureza da coleta, já que, os rótulos representam proporcionalmente o número de reproduções daquele fragmento de vídeo. A proposta inclui também a validação empírica da confiabilidade dos rótulos como medida de importância dos quadros, através de transferências entre conjuntos de dados e estudos com usuários.

A. Psychoacoustic Annoyance (PA)

Diversas pesquisas na área da saúde discutem o efeito do som no comportamento. Independentemente de ser ruído ou música. O artigo escrito por Zwicker *et al.* [27], propõem uma métrica capaz de calcular o grau de incômodo auditivo percebido. O incômodo psicoacústico (PA) mede o desconforto subjetivo que uma pessoa experimenta devido à exposição a certos sons. É diferente do nível de pressão sonora (SPL), que é uma medida física da amplitude das ondas sonoras. Enquanto o SPL é uma medida objetiva da intensidade de um som, o PA foca na percepção subjetiva do incômodo, levando em consideração vários fatores psicoacústicos. No modelo de Zwicke [27], o incômodo acústico de um som está relacionado aos seguintes índices psicoacústicos: Flutuação (F)

e Rugosidade (R): medem a modulação de um sinal ao longo do tempo. Loudness (N): baseado em estudos com humanos, mede o volume percebido. Nitidez (S): calculada por uma soma ponderada dos níveis de loudness em diferentes bandas de frequência. A fórmula para calcular o PA é dada por:

$$PA = N_5(1 + \sqrt{\omega_S^2 + \omega_{FR}^2})$$

$$\omega_s = \begin{cases} (S - 1.75) \cdot 0.25 \cdot lg(N_5 + 10) \ Se \ S > 1.75 \\ 0 \ Se \ S \leq 1.75 \end{cases}$$

$$\omega_{FR} = \frac{2.18}{N_5^{0.4}}(0.4F + 0.6R)$$

onde N é a loudness, N5 é o percentil 95 de loudness, S é a nitidez, F e R são flutuação e rugosidade, respectivamente.

Assim como no trabalho de referência, este estudo propõe utilizar o incômodo psicoacústico como ferramenta para o treinamento não supervisionado. Isso se deve ao fato da natureza multisensorial humana, onde nossa atenção pode ser guiada não somente pelo que observamos (imagem) mas também pelo que ouvimos (som).

III. Metodologia

Esse estudo consiste em estender os trabalhos realizados no trabalho realizado por Junior [10]. Dessa forma,é necessário analisar tanto o modelo e suas possíveis variações, quanto a metodologia aplicada.

A. Modelo referencia

O modelo utilizado neste estudo foi proposto por Junior [10], esse modelo recebe uma sequência de características multimodal (audiovisual) extraídas de quadros de vídeo e áudio. Para projetar essas características em um espaço de menor dimensão, utilizam-se cabeçotes MLP específicos para cada modalidade. A sequência multimodal resultante é então processada por um codificador Transformer, que aplica mecanismos de atenção para gerar representações contextuais das características integradas. Os vetores de características que passaram pelo codificador transformer possuem informações contextuais sobre toda a sequência para ambas as modalidades.Dessa forma o modelo base utiliza uma rede de regressão de pontuação que toma as características com atenção e gera uma pontuação de importância para cada uma. As pontuações finais são obtidas agregando as pontuações específicas de cada modalidade, as modalidades são então agregadas utilizando a média das pontuações para cada segmento utilizado como entrada.

$$AggFunc = (y_{aud} + y_{img})/2$$

A figura 2 representa a arquitetura que será utilizada no estudo.

Além disso, o modelo PGL-SUM [1] foi implementado como modelo de controle. Dessa forma, é foi possível verificar se os resultados encontrados após a implementação estão de acordo com o descrito na literatura. Isso permite que a implementação possa ser avaliada, garantindo a veracidade dos resultados encontrados.

Para treinar o modelo base para o método de sumarização, minimizamos o Erro Quadrático Médio as pontuações estimadas e as pontuações anotadas.

$$RMSE = \sqrt{\sum_{i=1}^{T} \frac{(y_i' - y_i)^2}{T}}$$

Finalmente, as pontuações de áudio e vídeo são combinadas para criar a pontuação final. Durante o processo, o conjunto de frames é dividido em segmentos usando a Segmentação Temporal Kernel e em seguida, os segmentos são escolhidos com base no algoritmo da Mochila 0/1, onde a pontuação média é utilizada como ”valor” e a quantidade de quadros como ”peso”.

Para que o modelo estime as pontuações para cada frame do vídeo, é necessário que o modelo seja alimentado por features de áudio e vídeo que podem ser extraídas de qualquer modelo existente na literatura. Dessa forma, assim como nos trabalhos de Junior [10] e PGL [1], foram utilizados features de vídeo extraídas do modelo CLIP ViT B/32 e da camada pool5 do modelo GoogleNet [20], ambos treinados para o dataset ImageNet [2]. Além disso, devido ao caráter audiovisual do modelo, o áudio de cada vídeo foi normalizado para uma taxa de amostragem de 44.1Khz e as features foram extraídas para cada segundo utilizando o modelo EsresNext [5] pré-treinado no dataset Audioset. Por fim, após os segmentos serem agregados utilizando o algoritmo KTS, as melhores sequências são selecionadas utilizando o algoritmo da mochila 0/1, que é empregado para selecionar os 15% melhores frames dos vídeos.

A utilização desse modelo se deve ao fato que o modelo foi pensado para utilizar o incômodo psicoacústico como pseudo-rótulo. Além disso, por conta da natureza da atenção humana ser multissensorial, a utilização de features de áudio parecem bastante promissoras para alcançar os objetivos pretendidos.

A escolha do conjunto de dados para esse estudo é de extrema importância, devido a isso, foi optado por utilizar os datasets SumMe e TVSum para alimentar o modelo e realizar os testes de implementação e avaliação de resultados. Essa decisão foi sustentada pelo fato da necessidade de se manter em uma plataforma comum com resultados já documentados na literatura e a utilização dos mesmos nos trabalhos de referência. Além disso, o conjunto de dados Mr.HISum [19], composto por vídeos extraídos do YouTube, também será utilizado na próxima etapa do projeto para a tarefa de estimativa de pontos de interesse. Esse dataset foi escolhido pois possui mais de 30mil vídeos, onde as avaliações de interesse de cada vídeo é o número de vezes que um determinado fragmento do vídeo foi reproduzido. A natureza desse dataset se assemelha com o objetivo final deste estudo.

O incômodo psicoacústico proposto por Zwicker [referencia] foi utilizado como pseudo label para o treinamento auto supervisionado dos modelos. Essa decisão foi tomada pois foi observado uma correlação entre o incômodo psicoacústico e

as partes mais reproduzidas do subconjunto de 3 mil vídeos presentes no dataset MrHiSum, replicando os resultados observados no trabalho referência [referencia], onde também foi observado uma correlação entre o incômodo psicoacústico e as anotações realizadas pelos usuários nos datasets de referência SumMe e TVSum.

Para a aquisição do incômodo psicoacústico, foi utilizado um código baseado no modelo proposto por Zwicker [referencia] do incômodo psicoacústico, e para isso foi necessário que o passasse por um tratamento de reamostragem, normalizando o dataset em 16 Khz de taxa de amostragem. Dessa forma, para cada áudio, foi gerado um vetor, onde cada elemento representa 1 segundo de vídeo e possui como valor o incômodo psicoacústico estimado pelo modelo.

Foram realizadas algumas extrações de PA para um subconjunto de 3 mil vídeos de dados contidos no dataset Mr.HiSum [19], com o objetivo de avaliar a possibilidade de utilizar o incomodo psicoacústico como pseudo rótulos para um futuro treinamento auto supervisionado, as figuras a seguir, representam os resultados encontrados

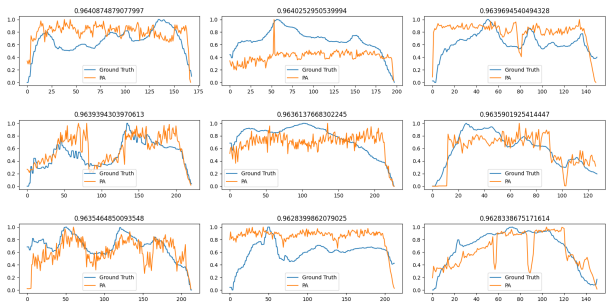

Fig. 1. 100:109 melhores semelhanças de cossenos entre o PA e o número de reproduções.

Durante o desenvolvimento desse projeto, experimentos parciais foram realizados focando na tarefa de sumarização de vídeo com o objetivo de alinhar o modelo antes de prosseguir com os próximos passos e avaliar os impactos das mudanças nas features de alimentação no desempenho do modelo.

A métrica F1-Score foi utilizada como ferramenta para quantificar a qualidade do modelo e do código. Isso se deve à sua ampla aceitação na literatura para problemas de sumarização, onde o número de frames importantes e não importantes é geralmente desbalanceado. Além de ser frequentemente utilizada nos estudos de comparação para a tarefa de sumarização, essa métrica é particularmente adequada para este tipo de problema, pois equilibra a precisão e a revocação, proporcionando uma avaliação mais robusta do desempenho do modelo quando há um desbalanceamento significativo entre as classes que faz parte da natureza do problema de sumarização. Por último, com o objetivo de melhorar a convergência do modelo e controlar o estado inicial do ambiente de testes, foi utilizado a distribuição xavier uniforme com ganho $\sqrt{(2)}$ para a inicialização dos pesos.

A Tabela 1, resume o desempenho do modelo testado bem como os impactos das mudanças nas features de alimentação e no aprendizado supervisionado e não supervisionado em comparação com os resultados dos modelos de referência.

Modelo	**F1-Score (%)**
CLIP-It! Não supervisionado	52,5
PVT Não supervisionado original	52,6
CLIP-It!	54,2
PGL Original	55,6
PVT-Clip Original	56,7
PGL-Não supervisionado (nosso)	41,54
PVT-Clip (nosso)	45,97
PVT-Google (nosso)	59,42
PGL (nosso)	63,06

TABLE I

TABELA 1 - ESTUDO COMPARATIVO DO MODELO IMPLEMENTADO UTILIZANDO A MÉTRICA F1-SCORE PARA O DATASET SUMME EM COMPARAÇÃO COM OS MELHORES RESULTADOS DOS MODELOS DOCUMENTADOS NA LITERATURA

Para os experimentos acima, foi utilizado o melhor resultado, utilizando os seguintes hiperparâmetros.

Taxa de aprendizado	5^{-5}
Dropout	1^{-5}
Batch-size	1
Epochs	200
Heads	8
Segs	4

Podemos observar que a utilização do PA para o dataset Mr.HiSum [19] pode ser bastante útil para alcançar bons resultados para a tarefa de predição de momentos de interesse em aprendizado auto supervisionado, isso pois graficamente pudemos observar uma grande relação entre o incômodo psicoacústico e os pontos mais reproduzidos dos vídeos amostrados.

Entretanto, mesmo com resultados promissores com o PA, os resultados preliminares dos experimentos com o modelo PGL mostraram que a implementação ainda não está alinhada com os resultados encontrados na referência, evidenciando que ainda existem pontos a serem corrigidos.

IV. CONCLUSÃO E PRÓXIMOS PASSOS

Este trabalho preliminar teve como objetivo experimentar a implementação e suas possíveis variações e avaliar a viabilidade da utilização do incômodo psicoacústico e o modelo proposto no trabalho base [10] que será utilizado como modelo para a predição automática de momentos de interesse em vídeos do Youtube.

Apesar dos resultados da relação entre o PA e os momentos mais reproduzidos dos vídeos para o dataset Mr.HiSum [19] parecerem promissores, os resultados são limitados a amostragem de 10% dos vídeos que estão disponíveis no dataset. Além disso, podemos perceber que o modelo ainda não está alinhado com os resultados documentados na literatura. Isso se deve ao fato que este trabalho ainda está em processo de desenvolvimento.

Por isso, fica definido como próximos passos do projeto o alinhamento do código com os resultados encontrados na literatura, removendo qualquer tipo de ruído nos resultados dos experimentos, a realização da extração de features e incômodo

psicoacústico para o dataset Mr.HiSum e o estudo de caso da implementação do modelo proposto para a tarefa de predição automática de momentos de interesse.

Estes resultados preliminares encontrados enfatizam a importância de uma abordagem cuidadosa na implementação de modelos de aprendizado. Além disso, auxilia no desenvolvimento de uma base sólida na investigação da utilização do incômodo psicoacústico como pseudo rótulo para um aprendizado auto supervisionado.

REFERENCES

[1] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 226–234, December 2021.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] M. Elfeki and A. Borji. Video summarization via actionness ranking, 2019.

[4] L. Feng, Z. Li, Z. Kuang, and W. Zhang. Extractive video summarizer with memory augmented neural networks. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 976–983, New York, NY, USA, 2018. Association for Computing Machinery.

[5] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio, 2021.

[6] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *European Conference on Computer Vision*, 2014.

[7] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2296–2304, New York, NY, USA, 2019. Association for Computing Machinery.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Z. Ji, K. Xiong, Y. Pang, and X. Li. Video summarization with attention-based encoder-decoder networks, 2018.

[10] E. R. A. Junior. An audiovisual approach for video summarization using psychoacoustic features. Master's thesis, Universidade Federal de Minas Gerais, 2023.

[11] V. V. K., D. Sen, and B. Raman. Video skimming: Taxonomy and comprehensive survey. *ACM Computing Surveys*, 52(5):1–38, Sept. 2019.

[12] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao. Exploring global diverse attention via pairwise temporal relation for video summarization, 2020.

[13] Y.-T. Liu, Y.-J. Li, F.-E. Yang, S.-F. Chen, and Y.-C. F. Wang. Learning hierarchical self-attention for video summarization. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3377–3381, 2019.

[14] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2982–2991, 2017.

[15] M. Narasimhan, A. Rohrbach, and T. Darrell. Clip-it! language-guided video summarization, 2021.

[16] M. Rochan, L. Ye, and Y. Wang. Video summarization using fully convolutional sequence networks, 2018.

[17] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015.

[18] J. Sul, J. Han, and J. Lee. Mr. hisum: A large-scale dataset for video highlight detection and summarization. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[19] J. Sul, J. Han, and J. Lee. Mr. hisum: a large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems*, 36, 2024.

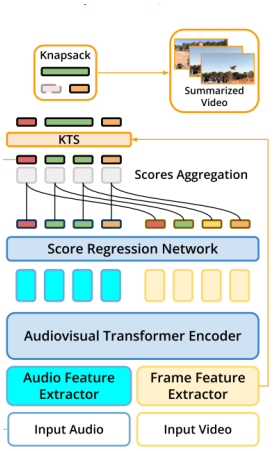

Fig. 2. Representação da arquitetura do modelo PVT, figura 4.1 - Página 34 [10]

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[22] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan. Stacked memory network for video summarization. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 836–844, New York, NY, USA, 2019. Association for Computing Machinery.

[23] Y. Yuan, H. Li, and Q. Wang. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, 7:64676–64685, 2019.

[24] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory, 2016.

[25] B. Zhao, X. Li, and X. Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7405–7414, 2018.

[26] B. Zhao, X. Li, and X. Lu. Hierarchical recurrent neural network for video summarization, 2019.

[27] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.

V. APÊNDICE