Técnicas de aprendizado para sumarização de vídeos

Learning techniques for video summarization

João Marcelo Tozato*, **Silvio Ricardo Rodrigues Sanches**†, **Priscila Tiemi Maeda Saito**‡

RESUMO

É perceptível uma grande disponibilidade de conteúdo em vídeo *online*. Dessa forma, estratégias de aprendizado para obter sumários relevantes de vídeos, de forma a capturar os momentos mais importantes de um dado vídeo vêm se mostrando cada vez mais necessárias. Este trabalho aborda a aplicação de técnicas de aprendizado ativo e profundo para a obtenção de sumários automáticos relevantes de vídeos. Para isto, foram realizados experimentos considerando algoritmos de clusterização, de aprendizado de máquina clássico, de aprendizado profundo, estratégias do aprendizado ativo e também comparações com sumários feitos por humanos. A metodologia proposta é validada em um conjunto de dados público referente a vídeos de diversos gêneros e alcança resultados com altas acurácias e baixas taxas de erro. Foi possível verificar as vantagens da introdução de técnicas do aprendizado ativo, de forma a selecionar os *frames* mais relevantes de um determinado vídeo de entrada e também obter sumários automáticos relevantes ao serem comparados com as anotações manuais dos vídeos.

Palavras-chave: Aprendizado Ativo. Sumarização de Vídeos. Aprendizado Profundo.

ABSTRACT

There is a noticeable amount of online video content available nowadays. Thus, learning strategies to obtain relevant summaries of videos, in order to capture the most important moments of a given video are becoming increasingly necessary. The present paper discusses the application of active and deep learning techniques to obtain relevant automatic video summaries. Therefore, we propose experiments considering clustering algorithms, classical machine learning algorithms, deep learning techniques and active learning strategies, and also comparisons with summaries made by human beings. The proposed methodology is validated on a public dataset referring to videos of various genres and achieves results with high accuracies and low error rates. We assessed and highlighted the advantages of the introduction of active learning techniques in order to select the most relevant and informative frames of a given input video. Our methods could also generate relevant automatic summaries when compared with manual human annotations of the videos.

Keywords: Active Learning. Video Summarization. Deep Learning.

1 INTRODUÇÃO

É perceptível o aumento exponencial na disponibilidade, criação e distribuição de conteúdo online no formato de vídeos no passado recente. De acordo com o YouTube, mais de 500 horas de conteúdo são disponibilizadas por minuto (STATISTA, 2021). Além disso, há um maior acesso a dispositivos capazes de realizar gravações em alta qualidade e também de disponibilizar estes conteúdos online. Tendo isto em vista, surge o seguinte problema: como obter representações relevantes e resumidas de tamanha quantidade de vídeos online?

* Engenharia de Computação, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil; tozato@alunos.utfpr.edu.br
† Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio; silviosanches@utfpr.edu.br
‡ Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio; psaito@utfpr.edu.br

Dessa forma, há a tarefa de sumarização automática de vídeos, que pode ser definida como uma forma de capturar e extrair informações essenciais, enfatizando *frames* mais representativos e informativos. Uma sumarização ideal deve levar em consideração todos os principais acontecimentos de um determinado vídeo de entrada no seu processo de seleção de *frames* mais importantes (PFEIFFER et al., 1996). Além disso, este também deve evitar a inclusão de *frames* redundantes e também de conteúdo que não apresente uma importância significante para o vídeo em questão.

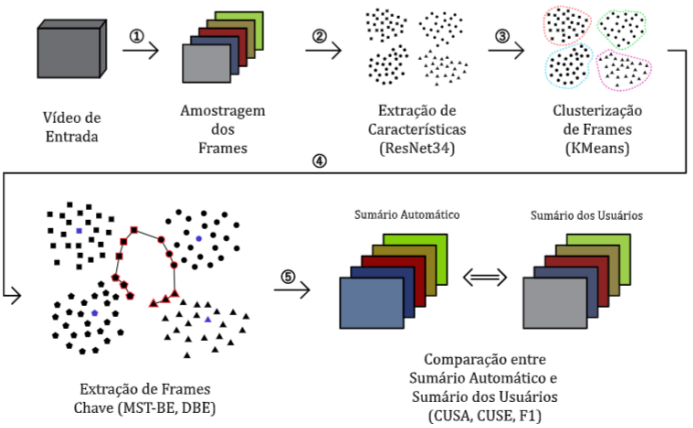
O presente projeto propõe uma nova abordagem, explorando técnicas de aprendizado ativo para sumarização de vídeos, de forma a selecionar as amostras (*frames*) mais informativos. Além disso, tendo em vista o sucesso recente do aprendizado profundo, redes neurais convolucionais residuais (HE et al., 2016) são consideradas para a extração das características dos vídeos.

Portanto, as contribuições do presente trabalho são relacionadas principalmente a extrair sumários informativos e representativos de vídeos. Para tal, é necessário estudar, desenvolver e validar técnicas relacionadas à tarefa de sumarização de vídeos. Estes sumários são obtidos por meio da utilização conjunta de estratégias de aprendizado profundo, algoritmos de clusterização e aprendizado ativo. Ademais, são conduzidos experimentos extensivos considerando abordagens distintas, de forma a explorar e determinar os *frames* mais informativos para um determinado vídeo.

2 MÉTODO (OU PROCEDIMENTOS OPERACIONAIS DA PESQUISA)

Na Figura 1 é possível verificar a metodologia proposta, que consiste na utilização de estratégias de aprendizado profundo, clusterização e aprendizado ativo para a geração de sumários automáticos de vídeos.

Figura 1 – Fluxograma da metodologia proposta.



Fonte: Autoria Própria

O primeiro passo desta é o de amostragem dos frames, que é realizado a uma taxa de 30 *frames* por segundo. Na etapa 2 da metodologia, é realizada a extração das características dos *frames* com o uso de redes neurais convolucionais e, em seguida, é realizada a clusterização deste conjunto de características (etapa 3) utilizando o algoritmo *k-means*. Tendo obtido os agrupamentos, é possível dar início à seleção dos *frames* chaves do vídeo de entrada (etapa 4), que é realizada por meio de adaptações das estratégias de aprendizado ativo MST-BE (*Minimum-Spanning Tree Boundary Edges*) e DBE (*Decreasing Boundary Edges*) (SAITO et

al., 2014). Finalmente, na etapa 5, é realizada uma comparação entre os sumários que foram gerados automaticamente e os sumários realizados por usuários, de acordo com a estratégia CUS (*Comparison of User Summaries*) (DE AVILA et al., 2011) e a métrica F_1-*Score*.

O processo de extração de características utilizado neste trabalho é baseado na utilização da arquitetura de rede neural convolucional ResNet34 (HE et al., 2016). São considerados os pesos da rede inicializados de acordo com o *dataset* ImageNet e os gradientes não são atualizados durante este processo. Dessa forma, foram extraídas as características de todos os *frames* dos vídeos do conjunto de dados considerado.

Após a obtenção e extração de características de cada um dos *frames* dos vídeos do conjunto de dados, utilizou-se o algoritmo *k-means* para realizar a clusterização dos *frames*. Entretanto, neste trabalho, este algoritmo foi utilizado como uma forma de obter as amostras raízes e amostras de fronteiras, que são fundamentais para a execução do processo de seleção de *frames* chave.

São consideradas duas principais abordagens em relação ao uso do algoritmo *k-means*, a primeira é utilizando um número fixo de *clusters* definido *a priori*, já a segunda é por meio do cálculo das distâncias euclidianas entre *frames* consecutivos de um determinado vídeo e, caso esta distância esteja acima de um limiar, que é definido experimentalmente, incrementa-se o número de *clusters* em uma unidade. O valor deste limiar, no presente trabalho, é de 0,17. A partir da clusterização dos *frames*, para extração de *frames* chave, são obtidas as amostras raízes e as amostras de fronteira. As amostras raízes são os centros de cada um dos *clusters*. As amostras de fronteira são obtidas ao analisar cada amostra do conjunto e sua respectiva *k* vizinhança, caso houver um *k* vizinho de rótulo distinto (conforme determinado pelo agrupamento), esta amostra é considerada como uma amostra de fronteira.

O primeiro método para redução e sumarização de vídeo consiste na adaptação da estratégia apresentada em Saito et al. (2012), considerando a sua etapa de redução do conjunto de dados. O conjunto resultante dessa estratégia, denominada RBS (*Root and Boundary Sampling* ou Amostragem de Raiz e Fronteira) é composto tanto pelas amostras raízes quanto pelas amostras de fronteira que foram obtidas no passo de clusterização.

Outros dois métodos adaptados e considerados para a sumarização, denominados DBE (SAITO et al., 2013) e MST-BE (SAITO et al., 2014), também consideram o conjunto de amostras raízes e de fronteira como base. No entanto, a diferença entre os métodos consiste na organização e na seleção das amostras.

No método DBE, para a organização das amostras, as arestas de fronteira são ordenadas em ordem decrescente de pesos de distâncias entre as amostras de uma dada aresta. Para a seleção das amostras, um classificador supervisionado de padrões é treinado e auxilia no processo de seleção. Nesse caso, foi utilizado o classificador SVM (*Support Vector Machines*) (WANG, 2005). Inicialmente, uma primeira instância do classificador é treinada considerando as amostras raízes obtidas pelo agrupamento inicial. Em seguida, uma aresta de fronteira por vez é avaliada, sendo selecionadas amostras (de uma dada aresta) cujos rótulos fornecidos pelo classificador sejam distintos.

Para o método MST-BE, as amostras de fronteira são organizadas computando a *minimum spanning tree* (MST) (NEŠETŘIL; MILKOVÁ; NEŠETŘILOVÁ, 2001) a partir das amostras de fronteira. Em seguida, cada aresta da MST é avaliada, sendo selecionadas as amostras de uma dada aresta, caso os rótulos das mesmas sejam distintos, de acordo com a classificação fornecida pela instância atual do classificador.

No que diz respeito à avaliação de um sumário gerado automaticamente, é preciso determinar um método objetivo para este processo avaliativo. Isto é devido ao fato de que não há uma definição clara do que pode ser considerado um sumário correto ou incorreto. Dessa forma, o método que apresenta sumários de maior fidelidade e qualidade em relação aos vídeos originais é o que leva em consideração uma anotação por parte de usuários.

Neste trabalho, é considerado o método chamado CUS (*Comparion of User Summaries* (DE AVILA et al., 2011). Esta estratégia realiza a comparação dos sumários que são gerados automaticamente com as anotações de todos os usuários.

Neste método, é realizada uma comparação entre os histogramas de cor entre os pares de *frames* selecionados pelos usuários e pelo algoritmo de sumarização, de forma que caso a distância entre os dois histogramas for menor que um determinado limiar, os *frames* são considerados correspondentes. No final deste processo, obtém-se três grupos distintos de *frames*: *frames* correspondentes (F_{corr}), *frames* não-correspondentes dos usuários ($F_{ncorr\text{-}user}$) e *frames* não-correspondentes do algoritmo de sumarização ($F_{ncorr\text{-}auto}$). A partir destes conjuntos, é possível realizar o cálculo das métricas CUS_A (taxa de acurácia), CUS_E (taxa de erro) e F_1-*Score*, que são definidas através das Eq. (1), Eq. (2) e Eq. (3). As definições de VP, FP e FN, presentes na Equação 3 são dadas pelo seguinte exemplo: em um problema de classificação binária, em que há uma classe positiva e outra negativa, Verdadeiro Positivo (VP) representa o caso em que o modelo de classificação prediz corretamente a classe positiva do problema. Falso Positivo (FP) é quando a predição do modelo é incorreta em relação à classe positiva, já Falso Negativo (FN) representa uma predição incorreta da classe negativa.

$$CUS_A = \frac{F_{corr}}{F_{corr}+F_{ncorr-user}} \quad (1)$$

$$CUS_E = \frac{F_{ncorr-auto}}{F_{corr}+F_{ncorr-user}} \quad (2)$$

$$F_1 - Score = \frac{VP}{VP+\frac{1}{2}(FP+FN)} \quad (3)$$

2.1 Descrição do Conjunto de dados

O conjunto de dados considerado é uma subseção do *Open Video Project*, contendo 50 vídeos de gêneros distintos. Todos estes vídeos foram amostrados em 30 *frames* por segundo. Além disso, possuem durações entre 1 e 4 minutos com uma resolução de 352 x 240 *pixels*. Devido à natureza do problema de sumarização de vídeos, não é viável a realização de anotações a nível de *frames*, ou seja, é preciso determinar alguma outra forma de gerar rótulos para comparação com os algoritmos de sumarização. Para isso, cada um dos vídeos deste conjunto de dados apresenta 5 sumários, que foram criados por usuários distintos. Dessa forma, é possível realizar comparações objetivas entre os sumários automáticos e sumários gerados por usuários utilizando as métricas descritas anteriormente.

Tabela 1 – Nome, descrição e número de clusters das estratégias de aprendizado propostas.

Abordagem	**Descrição**	***k* Clusters**
RBS_*k*AUTO	Somente redução, sem anotação	Automático
DBE_*k*AUTO	Redução e seleção, anotação manual das raízes	Automático
DBE_*k*10	Redução e seleção, anotação automática das raízes	10
DBE_*k*20	Redução e seleção, anotação automática das raízes	20
DBE_*k*30	Redução e seleção, anotação automática das raízes	30
MST-BE_*k*AUTO	Redução e seleção, anotação manual das raízes	Automático
MST-SE_*k*10	Redução e seleção, anotação automática das raízes	10
MST-SE_*k*20	Redução e seleção, anotação automática das raízes	20
MST-SE_*k*30	Redução e seleção, anotação automática das raízes	30

Fonte: Autoria Própria

2.2 Cenários

Na Tabela 1 é possível verificar os cenários que foram propostos para a realização dos experimentos. Esta também explicita quais as etapas dos algoritmos de aprendizado ativo são consideradas (redução e seleção), além da presença ou não da anotação manual das amostras raízes dos *clusters*. Nesta também é possível verificar a quantidade de clusters (*k*) do algoritmo *k-means* foram utilizados para as abordagens propostas, ou se este número foi determinado automaticamente através da estratégia descrita na Subseção de Clusterização de *frames*.

3 RESULTADOS

A Tabela 2 mostra os resultados que foram obtidos pelas abordagens propostas, de acordo com as métricas CUS_A, CUS_E e F_1-*Score*. Os melhores resultados obtidos pelas abordagens considerando cada uma das métricas estão destacados em negrito. É possível observar que a abordagem RBS_*k*AUTO foi a que apresentou uma maior acurácia (CUS_A), dentre todas as estratégias consideradas. Além disso, pode-se afirmar que esta abordagem foi uma das que mais se aproximou aos sumários realizados pelos usuários.

Tabela 2 – Resultados obtidos por cada abordagem considerando a média das métricas CUS_A, CUS_E e F_1-Score no *dataset Open* Video.

Abordagem	CUS_A	CUS_E	F_1-Score
RBS_*k*AUTO	**0,93**	0,21	0,23
DBE_*k*AUTO	0,22	0,24	0,16
DBE_*k*10	0,27	0,22	0,21
DBE_*k*20	0,51	0,15	0,24
DBE_*k*30	0,63	0,22	0,21
MST-BE_*k*AUTO	0,38	0,18	**0,29**
MST-SE_*k*10	0,36	0,19	0,24
MST-SE_*k*20	0,61	**0,12**	0,23
MST-SE_*k*30	0,76	0,18	0,20

Fonte: Autoria Própria (2021)

No que diz respeito à métrica F_1-*Score*, a estratégia MST-BE_*k*AUTO foi a que apresentou um maior valor para a mesma. Entretanto, não é possível realizar comparações entre as abordagens propostas e as da literatura em relação a esta métrica específica, pois estes valores não são fornecidos nos trabalhos originais. A abordagem que apresentou a menor taxa de erro foi a MST-BE_*k*20, que pode ser atribuído ao fato desta produzir sumários contendo poucos *frames* chave.

4 CONCLUSÃO

Neste trabalho foi proposta uma metodologia baseada em estratégias de aprendizado ativo, algoritmos de clusterização e aprendizado profundo para abordar o problema de sumarização de vídeos. Foram conduzidos experimentos extensivos considerando diferentes abordagens, incluindo adaptações de estratégias de aprendizado ativo propostas na literatura e variações no método de obtenção de agrupamentos. Além disso,

utilizou-se uma metodologia objetiva para a avaliação dos sumários obtidos pelos algoritmos, de forma a quantificar os resultados em relação a sumários realizados por seres humanos, o que reduz a subjetividade neste processo.

Foi possível observar as vantagens do uso de técnicas de aprendizado profundo, que foram utilizadas de forma a obter características representativas do conjunto de dados de entrada. Explorou-se também o potencial de estratégias decorrentes do aprendizado ativo no processo de seleção dos *frames* chave que compõem o sumário automático final, de forma a incorporar as amostras mais informativas neste conjunto.

AGRADECIMENTOS

O presente estudo foi realizado com o apoio das Instituições: Universidade Tecnológica Federal do Paraná, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Processo 431668/2016-7. Agradeço ao Programa de Bolsas de Iniciação Científica PIBIC 2020/2021 por ter fornecido o auxílio financeiro para o desenvolvimento do presente trabalho.

REFERÊNCIAS

DE AVILA, Sandra Eliza Fontes et al. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. **Pattern Recognition Letters**, Elsevier, v. 32, n. 1, p. 56–68, 2011.

HE, Kaiming et al. **Deep residual learning for image recognition**. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

NEŠETŘIL, Jaroslav; MILKOVÁ, Eva; NEŠETŘILOVÁ, Helena. Otakar Boruvka on minimum spanning tree problem Translation of both the 1926 papers, comments, history. **Discrete mathematics**, Elsevier, v. 233, n. 1-3, p. 3–36, 2001.

PFEIFFER, Silvia et al. Abstracting digital movies automatically. **Journal of Visual Communication and Image Representation**, Elsevier, v. 7, n. 4, p. 345–353, 1996.

SAITO, Priscila Tiemi Maeda et al. A data reduction and organization approach for efficient image annotation. In: **Proceedings of the 28th annual ACM symposium on applied computing**. 2013. p. 53-57.

SAITO, Priscila Tiemi Maeda et al. An active learning paradigm based on a priori data reduction and organization. **Expert Systems with Applications**, Elsevier, v. 41, n. 14, p. 6086–6097, 2014.

SAITO, Priscila Tiemi Maeda et al. Improving active learning with sharp data reduction. **WSCG'2012**, 2012.

SAITO, Priscila Tiemi Maeda. **Active learning with applications to the diagnosis of parasites**. 2014. 72 f. Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Computação, Campinas, SP. Disponível em: <http://www.repositorio.unicamp.br/handle/REPOSIP/275528>

STATISTA. **Hours of video uploaded to YouTube every minute as of February 2020**. 2021. Disponível em: <https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/>. Acesso em: 18 jun. 2021.

WANG, Lipo (Ed.). **Support vector machines: theory and applications**. Springer Science & Business Media, 2005.