TÓM TẮT VIDEO DỰA TRÊN BIỂU DIỄN ĐẶC TRƯNG CỦA ĐOẠN CLIP

Nguyễn Hoài Nam, Lê Quang Chiến

Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

Email: nhoainamdev@gmail.com, lqchien@husc.edu.vn

Ngày nhận bài: 26/6/2024; ngày hoàn thành phản biện: 12/7/2024; ngày duyệt đăng: 01/11/2024

TÓM TẮT

Với sự gia tăng khối lượng và đa dạng của dữ liệu video, việc tìm kiếm, trích xuất thông tin và hiểu nội dung ngày càng phức tạp và tốn thời gian. Tóm tắt video, bằng cách rút gọn video dài thành phiên bản ngắn hơn hoặc hình ảnh đại diện, nổi lên như một giải pháp tiềm năng. Kỹ thuật này có nhiều ứng dụng trong giáo dục, giải trí, an ninh, nâng cao năng suất và trải nghiệm người dùng. Các phương pháp tóm tắt truyền thống cho hiệu suất trung bình do hạn chế trong xử lý nội dung phức tạp, trong khi các kỹ thuật học sâu hiện đại đã có tiến bộ đáng kể. Bài báo này giới thiệu cách tiếp cận dựa trên biểu diễn đặc trưng của đoạn clip, khai thác thông tin không gian và thời gian qua cơ chế học tự chú ý (self-attention). Bên cạnh đó, chúng tôi đề xuất hai phương pháp tóm tắt phù hợp cho ngữ cảnh ngoại tuyến và trực tuyến dựa trên các biểu diễn đặc trưng này. Kết quả thực nghiệm cho thấy cách tiếp cận này có tiềm năng lớn cho các ứng dụng tóm tắt video thực tế.

Từ khóa: Biểu diễn đặc trưng, học sâu, self-attention, tóm tắt video.

1. MỞ ĐẦU

Tóm tắt video, hay còn gọi là Video Summarization, đã nổi lên như một giải pháp tiềm năng để khai thác tiềm năng từ dữ liệu video. Nó bao gồm việc rút gọn một video dài thành một phiên bản ngắn hơn hoặc một loạt các hình ảnh đại diện, trong khi vẫn giữ lại thông tin cốt lõi và ý chính của nội dung gốc. Mục tiêu là tạo ra các bản tóm tắt phản ánh chính xác và đầy đủ nội dung thiết yếu, giúp giảm thiểu thời gian cần thiết để người dùng nắm bắt các điểm chính của video.

Tóm tắt video được ứng dụng trong nhiều lĩnh vực khác nhau, mở ra nhiều cơ hội mới. Từ giáo dục, giải trí, an ninh đến nghiên cứu khoa học, nó có thể nâng cao hiệu suất làm việc và trải nghiệm người dùng trong các bối cảnh đa dạng. Ví dụ, một video gốc từ một sự kiện thể thao có thể được rút gọn thành một bản tóm tắt vài phút, nêu bật những khoảnh khắc quan trọng nhất như bàn thắng và các quả đá phạt đền.

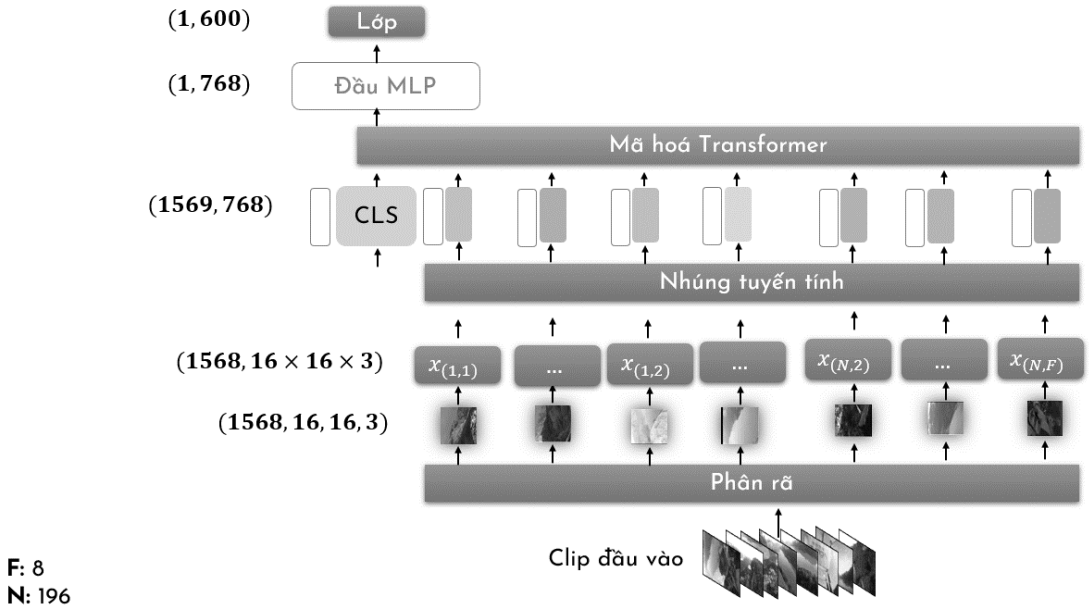
Các phương pháp tóm tắt video gần đây sử dụng nhiều kỹ thuật và mô hình khác nhau, từ các phương pháp truyền thống đến các kỹ thuật học sâu hiện đại, để đạt được hiệu suất tóm tắt tối ưu. Các phương pháp truyền thống như CSUV [1] tập trung vào các kỹ thuật tóm tắt video cơ bản như phân đoạn video và lựa chọn đoạn tóm tắt. Mặc dù những phương pháp này đã tiên phong trong lĩnh vực tóm tắt video, chúng thường có hiệu suất trung bình do hạn chế trong việc xử lý và hiểu nội dung video phức tạp. Các phương pháp học sâu đã tạo ra bước tiến lớn trong tóm tắt video, nhờ khả năng xử lý dữ liệu phức tạp và học từ lượng lớn dữ liệu. Ví dụ, VS-LMM [2] sử dụng sự tương hỗ giữa hình ảnh và ngôn ngữ để tăng cường tính mạch lạc của bản tóm tắt, dppLSTM [3] kết hợp quá trình Determinantal Point Process (DPP) với mạng Long Short-Term Memory (LSTM), re-seq2seq [4] sử dụng khung sequence-to-sequence để xử lý dữ liệu tuần tự và tạo ra các bản tóm tắt mạch lạc. Ngoài ra, Summary Transfer [5] áp dụng học chuyển giao để điều chỉnh các mô hình tóm tắt trong các lĩnh vực video khác nhau, cải thiện tính tổng quát của mô hình. Cuối cùng, DR-DSN [6] sử dụng học tăng cường sâu để tóm tắt động, tối ưu hóa sự cân bằng giữa độ dài và thông tin của bản tóm tắt.

Trong bài báo này, chúng tôi giới thiệu một cách tiếp cận dựa trên biểu diễn đặc trưng của đoạn clip. Cách biểu diễn này khai thác thông tin về không gian và thời gian thông qua cơ chế học tự chú ý (self-attention). Thông qua các biểu diễn này, chúng tôi cũng đề xuất hai phương pháp tóm tắt video lần lượt theo các ngữ cảnh ngoại tuyến và trực tuyến. Các kết quả thí nghiệm cho thấy cách tiếp cận này có tiềm năng to lớn trong việc tóm tắt video cho các ứng dụng thực tế.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Biểu diễn đặc trưng clip dựa trên Timesformer

Hình 1 biểu diễn kiến trúc tổng quan của mô hình Timesformer [7]. Kiến trúc này được cấu hình theo mô hình Timesformer-baseline được huấn luyện trước trên tập dữ liệu Kinetics-600 [8]. Theo kiến trúc, mô hình sẽ nhận đầu vào là các clip ngắn được cắt ra từ video gốc ban đầu. Mỗi clip này sau khi đi qua mô hình sẽ trích xuất ra biểu diễn đặc trưng tương ứng, đây là đặc trưng sẽ được sử dụng cho thuật toán tóm tắt video được đề xuất. Các bước chính để trích xuất đặc trưng của clip thông qua mô hình Timesformer bao gồm:



Hình 1. Minh hoạ kiến trúc tổng quan mô hình Timesformer.

(1) **Input clip:** Mô hình nhận đầu vào là một clip $X \in \Re^{H \times W \times 3 \times F}$, trong đó bao gồm F khung hình RGB (Red, Green, Blue) với kích thước $H\ px \times W\ px$ được lấy từ video gốc.

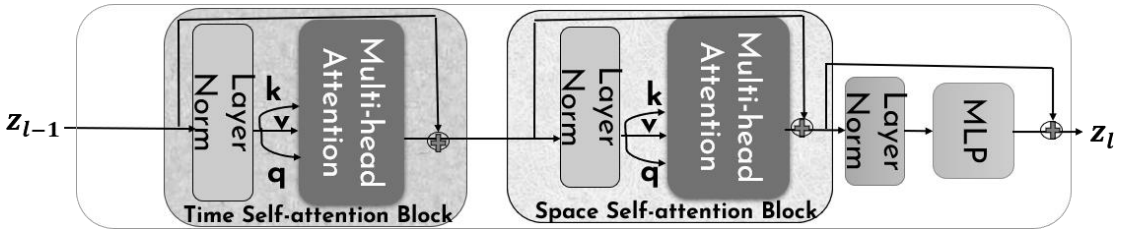
(2) **Phân rã:** Mỗi khung hình được chia thành các patch nhỏ kích thước $P\ px \times P\ px$, và các patch này không chồng chéo nhau. Những patch này sau đó được làm phẳng thành các vector $x_{(p,t)} \in \Re^{3P^2}$ với p là vị trí của patch và t là vị trí của khung hình.

(3) **Nhúng tuyến tính:** Mỗi patch $x_{(p,t)}$ được nhúng thành embedding vector $z^0_{(p,t)} \in \Re^D$ thông qua một ma trận nhúng $E \in \Re^{D \times 3P^2}$ có thể học được, kết hợp với positional embedding vector $e^{pos}_{(p,t)} \in \Re^D$. Công thức 1 biểu diễn cho phép nhúng tuyến tính này:

$$z^0_{(p,t)} = Ex_{(p,t)} + e^{pos}_{(p,t)} \quad (1)$$

(4) **Transformer Encoder:** Timesformer bao gồm L khối mã hóa. Các bước thực hiện tại mỗi khối mã hóa l (l=1..L) được minh họa như trong Hình 2. Cụ thể, mỗi khối mã hóa l, giá trị các vector query $q^{(l,a)}_{(p,t)}$ được tính cho mỗi patch từ đại diện $z^{(l-1)}_{(p,t)}$ đã mã hoá ở khối trước. Trong đó, $q^{(l,a)}_{(p,t)}$ là vector query của patch ở vị trí (p, t), tại khối mã hóa l và tại attention head a. Tiếp theo, trọng số self-attention $\alpha^{(l,a)}_{(p,t)} \in \Re^{N*F+1}$ cho mỗi query được tính thông qua dot-product giữa vector query và key, sau đó áp dụng hàm Softmax để chuẩn hóa trọng số. Cuối cùng,

nối tất cả các vector đầu ra đã mã hoá từ tất cả các attention head, kết hợp sử dụng residual connection và đưa qua một lớp Multi-Layer Perceptron (MLP) truyền kết quả $z_{(p,t)}^{(l)}$ cho khối mã hoá tiếp theo.



Hình 2. Minh hoạ kiến trúc khối mã hóa *l* sử dụng Divided Space-Time Self-Attention.

Trong bài báo gốc [7], nhiều phương pháp tính toán self-attention khác nhau đã được thử nghiệm. Các kết quả phân tích chỉ ra rằng, phương pháp Divided Space-Time Self-Attention (T+S) cho thấy sự hiệu quả trong việc học cấu trúc thời gian - không gian trong video. Vì vậy, trong nghiên cứu này, chúng tôi áp dụng phương pháp T+S trong mỗi khối mã hóa để trích xuất đặc trưng của mỗi clip.

(5) **Clip Embedding**: Biểu diễn đặc trưng clip y được lấy từ khối mã hoá cuối cùng L sau khi áp dụng hàm chuẩn hóa *LN()* theo Công thức 2.

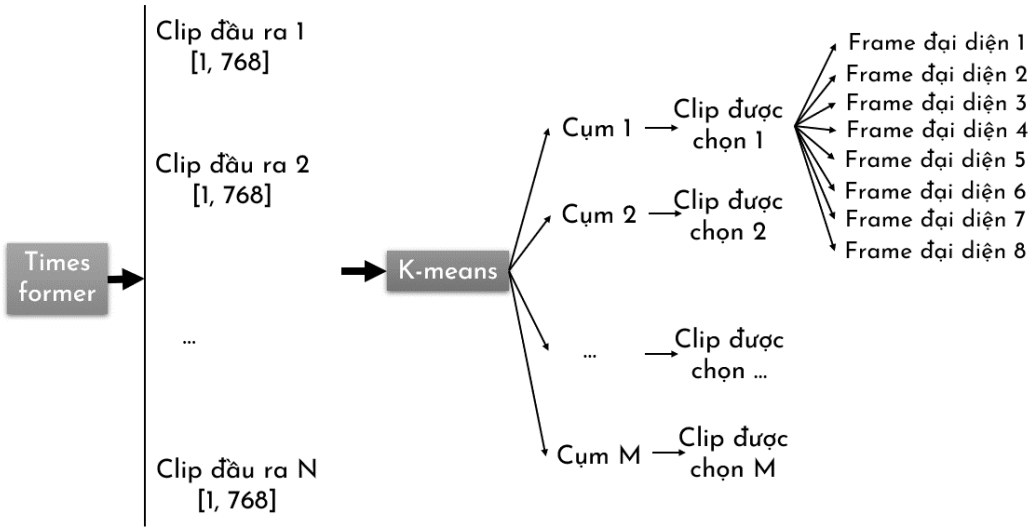
$$y = LN\left(z_{(0,0)}^{(L)}\right) \in \Re^D \quad (2)$$

2.2. Thuật toán tóm tắt video

Dựa vào mô hình Timesformer, mỗi clip đầu vào sẽ được trích xuất thành một vector 768-dims. Các vector này, sau đó, sẽ được sử dụng để thực hiện tóm tắt video. Trong bài báo này, chúng tôi đề xuất hai thuật toán tóm tắt video tương ứng với hai cách tiếp cận riêng cho bài toán này: (1) Tóm tắt video ngoại tuyến; và (2) Tóm tắt video trực tuyến.

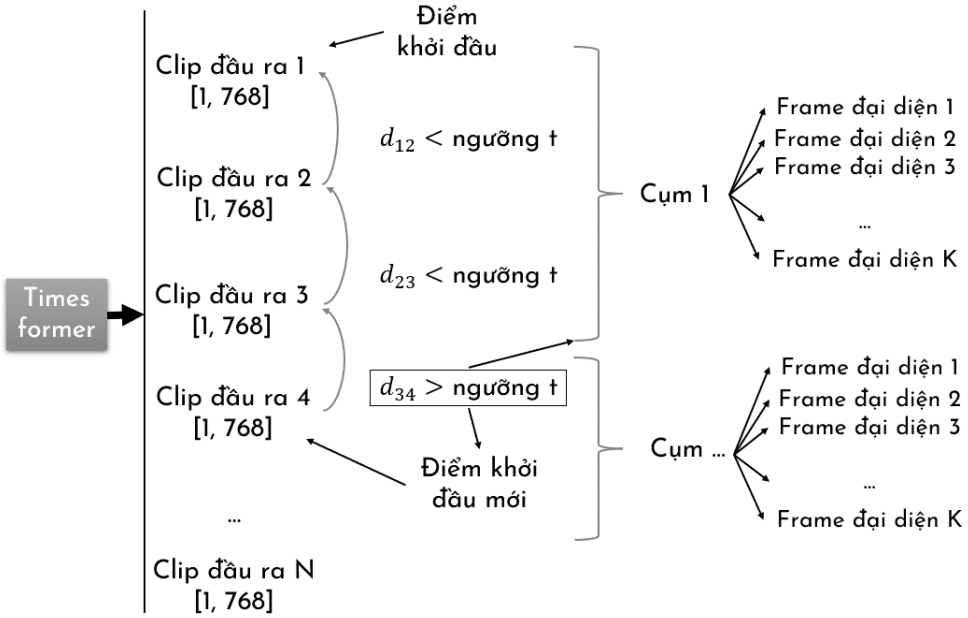
2.2.1. Thuật toán tóm tắt video ngoại tuyến

Bước đầu tiên của thuật toán này sử dụng phương pháp phân cụm K-means với K là số lượng cụm bằng 15 phần trăm tổng số biểu diễn đặc trưng (các vector 768-dims) được trích xuất. Sau khi thực hiện phân cụm, ta sẽ có được K cụm với mỗi cụm bao gồm các biểu diễn đặc trưng của các clip tương đồng với nhau. Bước tiếp theo thực hiện tính khoảng cách giữa mỗi tâm cụm và danh sách biểu diễn đặc trưng để tìm ra clip tương ứng có khoảng cách gần nhất với mỗi tâm cụm. Kết quả ta có được danh sách K clip được chọn. Cuối cùng, dựa trên danh sách K clip này, chúng ta sẽ chọn ra các khung hình (frame) đại diện. Thuật toán được biểu diễn như trong Hình 3 dưới đây.



Hình 3. Minh họa thuật toán tóm tắt video ngoại tuyến dựa trên phân cụm K-means

2.2.2. Thuật toán tóm tắt video trực tuyến



Hình 4. Minh họa thuật toán tóm tắt video trực tuyến

Thuật toán tóm tắt video trực tuyến được biểu diễn như trong Hình 4. Thiết lập clip đầu tiên là điểm khởi đầu, thuật toán bắt đầu duyệt từ clip thứ 2. Tại mỗi clip đang xét, thuật toán thực hiện so sánh sự khác nhau giữa clip hiện tại và với clip kế trước. Nếu

như độ khác nhau vượt qua ngưỡng t, thì tiến hành lưu lại các clip ở điểm khởi đầu tới clip kế trước của clip hiện tại tạo thành một đoạn video chứa các clip tương đồng. Sau đó, lấy ngẫu nhiên K khung hình đại diện trong mỗi đoạn video này. Thuật toán sẽ lặp lại từ đầu với thiết lập điểm khởi đầu mới là clip hiện tại cho tới khi vượt qua clip cuối cùng.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả

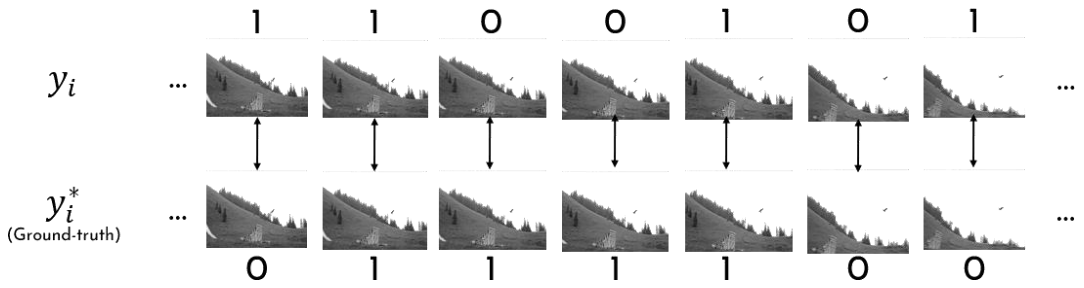
3.1.1. Cơ sở dữ liệu

Chúng tôi tiến hành thí nghiệm trên tập dữ liệu SumMe [9], tập dữ liệu này bao gồm 25 video có thời lượng tối đa 6 phút, với 25 chủ đề khác nhau, bao gồm thể thao, sự kiện và ngày lễ (xem một số khung hình mẫu trong Hình 5). SumMe có các tập chú thích được thu thập từ 15 đến 18 người khác nhau cho mỗi video. Mỗi bản tóm tắt có độ dài từ 5 đến 15 phần trăm thời lượng của video gốc.

Hình 5. Một số khung hình mẫu của video Jumps.mp4 trong tập dữ liệu SumMe.

3.1.2. Phương pháp đánh giá

Để thực hiện việc đánh giá tính hiệu quả của hệ thống, chúng tôi đánh giá dựa trên thước đo F1-score. F1-score được tính toán dựa trên kết quả tóm tắt được tạo ra từ hệ thống tóm tắt video của chúng tôi $y_i \in \{0,1\}$ với phần chú thích có sẵn trong tập dữ liệu SumMe $y_i^* \in \{0,1\}$. Trong đó, $y_i = 1$ nếu khung hình i được chọn nằm trong kết quả tóm tắt, ngược lại là 0; tương tự với y_i nhưng với chú thích có sẵn trong SumMe (xem Hình 6).



Hình 6. Minh họa cách đánh giá cho bài toán tóm tắt video với tập dữ liệu SumMe.

3.1.3 Kết quả thí nghiệm

Trong kết quả thí nghiệm, F1-score sẽ được tính riêng cho từng video tóm tắt. F1-score của tập dữ liệu nhận được bằng cách tính trung bình hoặc chọn mức tối đa cho mỗi video. Các kết quả thí nghiệm theo hai thuật toán tóm tắt video ngoại tuyến và trực tuyến được báo cáo ở Bảng 1 dưới đây.

Bảng 1. Bảng kết quả thí nghiệm theo hai thuật toán trên tập dữ liệu SumMe

TT	Tên video	FPS	N-frames	F1-score (ngoại tuyến)	F1-score (trực tuyến)
1	Air Force One	25	4494	21.08	16.94
2	Base jumping	30	4729	21.25	15.75
3	Bearpark climbing	25	3341	18.92	17.02
4	Bike Polo	30	3064	23.91	15.94
5	Bus in Rock Tunnel	30	5133	17.21	17.28
6	car over camera	30	4382	29.22	14.96
7	Car rail crossing	30	5075	15.99	17.02
8	Cockpit Landing	30	9046	16.07	16.22
9	Cooking	15	1287	29.27	17.23
10	Eiffel Tower	25	4971	20.07	15.88
11	Excavators river crossing	25	9721	16.78	15.68
12	Fire Domino	30	1612	28.51	17.87
13	Jumps	25	950	17.39	18.06
14	Kids playing in leaves	30	3187	18.00	17.96
15	Notre Dame	24	4608	17.45	17.38
16	Paintball	24	6096	18.26	15.57
17	paluma jump	30	2574	23.37	16.43
18	playing ball	30	3119	22.37	16.22
19	Playing on water slide	30	3065	21.57	18.48
20	Saving dolphins	30	6683	17.23	16.17
21	Scuba	30	2221	23.73	16.59
22	St Maarten Landing	25	1751	29.83	16.99

23	Statue of Liberty	25	3863	22.40	17.44
24	Uncut Evening Flight	30	9672	15.98	15.55
25	Valparaiso Downhill	30	5178	20.67	15.72
	Trung bình			21.06	16.66

Bên cạnh đó, chúng tôi cũng báo cáo kết quả thí nghiệm trong so sánh với một số nghiên cứu trước đây để có được cái nhìn tổng quan hơn, kết quả thể hiện ở Bảng 2.

Bảng 2. Bảng so sánh kết quả thí nghiệm với các nghiên cứu trước đây trên tập dữ liệu SumMe

TT	Phương pháp	F-avg	F-max
1	CSUV [1]	23.1	-
2	VS-LMM [2]	-	40.3
3	dppLSTM [3]	-	43.2
4	VS-DSF [10]	18.3	-
5	Summary Transfer [5]	-	41.2
6	DR-DSN [6]	-	41.3
7	re-seq2seq [4]	-	45.1
8	SASUM [11]	-	45.3
9	Timesformer (Thuật toán ngoại tuyến)	21.06	29.83
10	Timesformer (Thuật toán trực tuyến)	16.66	18.48

** F-avg là thước đo F1-score trung bình; F-max là thước đo F1-score cao nhất.*

3.2. Thảo luận

Trong thí nghiệm này, ban đầu chúng tôi đã thiết lập tần suất lấy mẫu là 1/1, tức là giữ toàn bộ video gốc để thực hiện việc tóm tắt video. Như vậy, việc đánh giá sẽ đảm bảo được tính chính xác vì số lượng khung hình đại diện sẽ là toàn bộ khung hình của video gốc. Tuy nhiên, vấn đề đã nảy sinh khi chúng tôi tiến hành tóm tắt với một số video có độ dài hơn 8 phút, phần cứng đã không đáp ứng được các video này. Để giải quyết, chúng tôi tiến hành tăng tần suất lấy mẫu lên, việc làm này có ưu điểm sẽ giảm gánh nặng bộ nhớ, tuy nhiên, hành động này vô tình lược bỏ một số lượng lớn khung hình quan trọng, dẫn đến kết quả đánh giá trở nên kém đi.

Thông qua thí nghiệm, chúng ta thấy được hướng tiếp cận khác biệt với bài tóm tắt video của hai thuật thuật toán. Cụ thể, đối với thuật toán dựa trên phân cụm K-means, ta cần nắm rõ số lượng cụm chính xác là bao nhiêu. Để có được điều này, đương

nhiên phải có tổng số khung hình của video gốc, vì vậy, phương pháp này sẽ hoạt động hiệu quả với dạng video đã được lưu trữ, có sẵn như một bộ phim, một video trên Youtube. Thêm nữa, với độ phức tạp tính toán của K-means, khiến cho việc tóm tắt diễn ra không được quá nhanh, sẽ không phù hợp với hướng tiếp cận tóm tắt video trực tuyến. Đối với hướng tiếp cận này, dữ liệu video sẽ được cập nhật liên tục cho đến khi dừng, ví dụ như một buổi livestream. Lúc này, thuật toán tóm tắt video trực tuyến tiến hành xét các clip mới vào với clip cuối cùng trong bản tóm tắt, nếu giống, thì bỏ qua clip mới này, ngược lại, tiến hành lưu vào clip vào bản tóm tắt. Như vậy, thuật toán này sẽ bỏ qua việc lưu trữ các clip tương đồng với clip trước đó cho tới khi video kết thúc, giúp giảm bớt gánh nặng lưu trữ đáng kể.

4. KẾT LUẬN

Bài báo này đã giới thiệu một phương pháp tóm tắt video mới dựa trên biểu diễn đặc trưng clip thông qua cơ chế self-attention của mô hình Timesformer. Phương pháp này có thể tạo ra các bản tóm tắt video đầy đủ thông tin và có thể được áp dụng cho cả hai ngữ cảnh ngoại tuyến và trực tuyến, mang lại tính linh hoạt cao. Kết quả thí nghiệm cho thấy cách tiếp cận này có tiềm năng to lớn trong việc tóm tắt video cho nhiều lĩnh vực khác nhau, giúp người dùng tiết kiệm thời gian và nâng cao hiệu suất công việc.

Tuy nhiên, phương pháp dựa trên mô hình Timesformer yêu cầu phần cứng mạnh mẽ để xử lý dữ liệu video dài. Điều này dẫn đến một số hạn chế khi áp dụng vào các video dài trong thực tế, đòi hỏi sự tối ưu hóa thêm về mặt phần cứng và thuật toán. Mặc dù kết quả hiện tại chỉ ở mức tiềm năng, việc tiếp tục tinh chỉnh mô hình và tìm kiếm các phương pháp chọn lọc hiệu quả hơn là cần thiết để cải thiện kết quả tóm tắt video.

TÀI LIỆU THAM KHẢO

[1]. M. Gygli, H. Grabner, H. Riemenschneider, and L. van Gool. Creating summaries from user videos. In European Conference on Computer Vision (ECCV), pages 505–520, 2014.

[2]. M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3090–3098, 2015.

[3]. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In European Conference on Computer Vision (ECCV), pages 766–782, May 2016.

[4]. K. Zhang, K. Grauman, and F. Sha. Retrospective Encoders for Video Summarization. In European Conference on Computer Vision (ECCV), pages 383–399, 2018.

[5]. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1059–1067, 2016.

[6]. K. Zhou, Y. Qiao, and T. Xiang. Deep reinforcement learning for unsupervised video summarization with diversity representativeness reward. 2018.

[7]. Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? 2021.

[8]. João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017.

[9]. Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc van Gool. Creating summaries from user videos. 2014.

[10]. M. Otani, Y. Nakashima, E. Rahtu, J. Heikkila, and ¨ N. Yokoya. Video summarization using deep semantic features. In Asian Conference on Computer Vision (ACCV), volume 10115, pages 361–377, 2016.

[11]. H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao. Video Summarization via Semantic Attended Networks. In AAAI Conference on Artificial Intelligence, pages 216–223, 2018.

CLIP FEATURE REPRESENTATION-BASED VIDEO SUMMARIZATION

Nguyen Hoai Nam, Le Quang Chien

Faculty of Information Technology, University of Sciences, Hue University

Email: nhoainamdev@gmail.com, lqchien@husc.edu.vn

ABSTRACT

With video data's increasing volume and diversity, tasks such as searching, extracting information, and understanding content have become more complex and time-consuming. By condensing lengthy videos into shorter versions or representative images, video summarization has emerged as a potential solution. This technique has numerous applications in education, entertainment, and security, enhancing productivity and user experience. Traditional summarization methods yield average performance due to limitations in handling complex content, whereas modern deep-learning techniques have shown significant advancements. This paper introduces an approach based on clip feature representation, leveraging spatiotemporal information through a self-attention mechanism. Additionally, we propose two summarization methods suitable for offline and online contexts based on these feature representations. Experimental results demonstrate that this approach holds significant potential for practical video summarization applications.

Keywords: Deep-learning, feature representation, self-attention, video summarization.

Nguyễn Hoài Nam sinh ngày 22/07/2002 tại Thừa Thiên Huế. Ông tốt nghiệp thủ khoa đầu ra cử nhân ngành Công nghệ thông tin, chuyên ngành Khoa học máy tính tại trường Đại học Khoa học, ĐH Huế năm 2024. Hiện nay, ông đang công tác tại công ty TNHH Phần mềm FPT Quy Nhơn (QAI).

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, Thị giác máy tính.

Lê Quang Chiến sinh ngày 15/09/1983 tại Thừa Thiên Huế. Năm 2005, ông tốt nghiệp cử nhân chuyên ngành Tin học tại trường Đại học Khoa học, Đại học Huế. Năm 2007, ông nhận bằng thạc sĩ chuyên ngành khoa học máy tính tại trường Đại học Khoa học, Đại học Huế. Năm 2016, ông nhận học vị tiến sĩ chuyên ngành Tin học tại trường SOKENDAI (The Graduate University for Advanced Studies), Nhật Bản. Hiện nay, ông đang công tác tại khoa Công nghệ Thông tin, trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Xử lý và nhận dạng ảnh, xử lý video, học máy, thị giác máy tính.