مقاله پژوهشی

خلاصه‌سازی ویدئویی با روش ترکیبی گراف شبکه‌ای و خوشه‌بندی

مهسا رحیمی رسکتی، همایون مؤتمنی، ابراهیم اکبری و حسین نعمت‌زاده

چکیده: ما در دنیایی زندگی می‌کنیم که وجود دوربین‌های خانگی و قدرت رسانه باعث شده تا با حجم خیره‌کننده‌ای از داده‌های ویدئویی سر و کار داشته باشیم. مسلم است روشی که بتوان با کمک آن، این حجم بالای فیلم را با سرعت و بهینه مورد دسترسی و پردازش قرار داد، اهمیت ویژه‌ای پیدا می‌کند. با کمک خلاصه‌سازی ویدئویی این مهم حاصل شده و فیلم به یک سری فریم یا کلیپ کوتاه ولی بامعنی خلاصه می‌گردد. در این پژوهش سعی گردیده در ابتدا داده با کمک الگوریتم K-Medoids خوشه‌بندی شود؛ سپس در ادامه با کمک شبکه توجه گرافی کانولوشنالی، جداسازی زمانی و گرافی انجام گیرد و در گام بعدی با کمک روش ردکردن اتصال، نویزها و موارد تکراری حذف گردد. سرانجام با ادغام نتایج به‌دست‌آمده از دو گام متفاوت گرافی و زمانی، خلاصه‌سازی انجام گیرد. نتایج به دو صورت کیفی و کمی و بر روی سه دیتاست SumMe، TVSum و OpenCv مورد بررسی قرار گرفت. در روش کیفی به‌طور میانگین ۸۸٪ نرخ صحت در خلاصه‌سازی و ۳۱٪ میزان خطا دست یافته که به نسبت سایر روش‌ها جزء بالاترین نرخ صحت است. در ارزیابی کمی نیز روش پیشنهادی، کارایی بالاتری نسبت به روش‌های موجود دارد.

کلیدواژه: **کاوش ویدئویی، خلاصه‌سازی ویدئویی، خوشه‌بندی، K-Medoids، شبکه توجه گرافی کانولوشنالی.**

۱- مقدمه

امروزه چند رسانه، نقشی اساسی در زندگی بشر ایفا می‌کند؛ چند رسانه ترکیبی از محتویات مختلف مانند متن، تصویر، صوت، ویدئو، گرافیک و ... هستند که از طریق آنها می‌توان به هر نوع اطلاعاتی به صورت دیجیتالی دسترسی داشت. به عبارت دیگر، ارائه جذابی از داده‌های یکپارچه است [۱]. داده‌ها مجموعه‌ای از متغیرها و واحدهای اطلاعاتی هستند که در یک فرایند جمع‌آوری می‌شوند و آنها را می‌توان به هر شکل و برای هر هدفی مورد استفاده قرار داد یا به آنها دسترسی داشت. فرایندی که برای استخراج داده‌های قابل استفاده از مجموعه قابل توجهی از داده‌های خام استفاده می‌شود، داده‌کاوی نامیده می‌شود. زندگی امروزی، داده‌های زیادی را ارائه می‌دهد و یافتن اطلاعات مفید از آنها می‌تواند در جنبه‌های مختلف پزشکی، اقتصادی، آموزشی و ... به ما کمک کند. یکی از محبوب‌ترین انواع داده‌ها ویدئو است. ویدئو یک رسانه الکترونیکی است که شامل تکثیر، پخش و نمایش رسانه‌های تصویری متحرک است. ویدئوها به دلیل تأثیر قابل توجهی که بر ذهن مردم دارند، بیش از هر شکل دیگری از اطلاعات محبوب هستند. برای اکثر مردم، دیدن چیزی به جای خواندن یا گوش‌دادن محبوب‌تر و راحت‌تر است.

امروزه با رواج دوربین‌ها در همه جا، دسترسی آسان به اینترنت و پیشرفت چشمگیر در پردازش ویدئو، حجم عظیمی از داده‌های ویدئویی در دسترس است. در حالی که این داده‌ها از اهمیت زیادی برخوردار هستند و می‌توانند از بسیاری جهات کمک کنند، دسترسی به آنها چالش‌برانگیز و وقت‌گیر است. این میزان داده باعث کاهش عملکرد برنامه‌های مختلف که بر مبنای پردازش ویدئو کار می‌کنند، مانند جستجوی ویدئو، فهرست‌بندی ویدئویی، توصیه‌های ویدئویی و ... می‌شود [۲]. خلاصه‌سازی ویدئویی، روشی مؤثر برای کاهش این ایرادات و مشکلات است و کمک می‌کند تا ویدئوهای طولانی را به رویدادهای ضروری یا فریم‌های کلیدی کاهش داده که به صرفه‌جویی در زمان و انرژی بسیار کمک می‌کند. تکنیک‌های پیشرفته‌ای در این زمینه وجود دارد و پیشرفت‌های زیادی در دهه‌های گذشته حاصل شده است. با این حال، محققان هنوز در حال کار برای تولید تکنیک‌های کارآمدتر و قوی‌تر و ایجاد ویدئوهای خلاصه‌شده دقیق‌تر و مفهومی‌تر هستند [۲].

خلاصه‌سازی ویدئو به کشف معنی یا هدف واقعی یک ویدئوی طولانی در نسخه کوتاه‌تر کمک می‌کند. با انتخاب مفهومی‌ترین فریم‌های یک ویدئو که فریم‌های کلیدی نامیده می‌شود، یک نسخه انتزاعی ساخته می‌شود. خلاصه‌سازی ویدئویی در دهه ۱۹۹۰ معرفی شد و قدمت زیادی ندارد؛ اما به دلیل اهمیت آن، محققان زیادی بر روی آن کار کرده‌اند. خلاصه‌سازی ویدئو دارای سه مرحله است [۳]:

۱) اطلاعات ویدئویی ابتدا برای کشف عوامل، ساختار یا نقاط برجسته در اجزای بصری، صوتی و متنی تجزیه و تحلیل می‌شوند.
۲) فریم‌های معنی‌دار که نشان‌دهنده محتوای ویدئو هستند، انتخاب می‌شوند.
۳) تلفیق خروجی انجام می‌گیرد که شامل سازماندهی فریم‌ها/شات‌های استخراج‌شده در نسخه خلاصه‌شده ویدئو است.

خلاصه‌سازی ویدئو عمدتاً به دو نوع خلاصه‌سازی ویدئویی فریم کلیدی (یا خلاصه‌سازی ثابت ویدئویی) و خلاصه‌سازی پویای ویدئویی[1] طبقه‌بندی می‌شود. اولی با انتخاب گروهی از فریم‌های کلیدی که کل ویدئو را نشان می‌دهد خلاصه را تولید می‌کند؛ در حالی که دومی یک ویدئو را در یک نسخه کوتاه‌تر خلاصه می‌نماید.

خلاصه‌سازی پویا بازپخشی از ویدئوی اصلی است که در واقع فیلمی تشکیل‌شده از قطعات اصلی ویدئوی ابتدایی است که می‌توان آن را به دو گروه برجسته‌سازی و سکانس‌هایی از ویدئوی اصلی یا خلاصه سکانس

این مقاله در تاریخ ۳۱ خرداد ماه ۱۴۰۱ دریافت و در تاریخ ۱۶ اردیبهشت ماه ۱۴۰۲ بازنگری شد.

مهسا رحیمی رسکتی، دانشکده مهندسی کامپیوتر، دانشگاه آزاد ساری، ساری، ایران، (email: mr2.mco@gmail.com).
همایون مؤتمنی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه آزاد ساری، ساری، ایران، (email: h_motameni@yahoo.com).
ابراهیم اکبری، دانشکده مهندسی کامپیوتر، دانشگاه آزاد ساری، ساری، ایران، (email: akbari@iausari.ac.ir).
حسین نعمت‌زاده، دانشکده مهندسی کامپیوتر، دانشگاه آزاد ساری، ساری، ایران، (email: hn_61@yahoo.com).

1. Video Skimming

تقسیم کرد. برجسته‌سازی یعنی جالب‌ترین و جذاب‌ترین قسمت‌های یک ویدئو را استخراج کنیم؛ در حالی که در خلاصه سکانس، محتوا و ایده اصلی یک ویدئو به نمایش گذاشته می‌شود. در میان انواع انتزاع‌های ویدئویی، خلاصه سکانس بالاترین خلاصه معنایی را از محتوای یک ویدئو منتقل می‌کند. این خلاصه را می‌توان با شناسایی مؤلفه‌های مهم موجود در ویژگی‌های به‌دست‌آمده از ویدئو (چه تک‌وجهی یا چندوجهی) ایجاد کرد که به دلیل پویابودن این نوع خلاصه‌سازی، کمک به درک بهتری از خلاصه نهایی می‌کند و به همین دلیل اخیراً در بسیاری از تحقیقات مورد توجه محققین قرار گرفته است [۴].

یکی از مزایای استفاده از خلاصه‌سازی پویا در مقایسه با نوع ایستای آن، استفاده از اطلاعات صوتی است؛ زیرا گاهی اوقات صوت حاوی اطلاعات مهمی است که می‌تواند در خلاصه‌سازی بهینه یاری رساند. بنابراین تفاوت اصلی بین خلاصه پویا و ایستا وجود اطلاعات حرکتی و صوتی در نوع پویاست. خروجی خلاصه‌سازی ویدئوی ایستا تنها گروهی از تصاویر است؛ در حالی که خروجی خلاصه‌سازی ویدئویی پویا حاوی محتوای داده‌های ویدئویی و صوتی است.

برخی از کلید مزایای خلاصه‌سازی پویا ویدئو عبارت هستند از

۱) انتقال طرح اصلی ویدئو در زمان کوتاه‌تر

۲) کاهش زمان انتقال برای ویدئوهای جستجوشده از طریق اینترنت

۳) بهینه‌سازی فضای ذخیره‌سازی (با ذخیره‌کردن ویدئو به شکل خلاصه‌شده آن افزایش می‌یابد).

۴) ادغام اطلاعات منتقل‌شده از طریق ویدئوهای متعدد متعلق به یک موضوع

مقایسه دو نوع خلاصه‌سازی ویدئویی نشان می‌دهد که اولی خلاصه‌ای دقیق‌تر ارائه می‌دهد؛ اما دومی به راحتی قابل درک است [۳].

خلاصه‌سازی کاربردهای فراوان دارد همچون

- فیلم‌ها
- ویدئوهای پزشکی
- ویدئوهای کاربرساخته
- ویدئوهای نظارتی
- ویدئوهای ورزشی
- و ...

در این مقاله سعی گردیده که با الهام از لی[1] و همکاران روشی بر اساس جاسازی گراف زمانی[2] ارائه شود [۵]. در این روش در ابتدا با کمک روش K-medoids ویدئو به یک سری کلاستر خلاصه و کاندیدا تقسیم می‌شود تا با این کار علاوه بر بالابردن صحت نتیجه، از حجم داده‌ها و همچنین محاسبات کاست. سپس داده‌های موجود به دو نوع ویژگی‌های زمانی و گرافی تقسیم می‌شوند. در قسمت جاسازی زمانی دقیقاً مشابه الگوریتم لی و همکاران عمل می‌گردد؛ اما در قسمت گرافی سعی شده تا از روش شبکه‌های عصبی گراف (GNN)[3] استفاده شود که به نظر می‌رسد نتایجی بهتر و با صحت بالاتری ارائه می‌دهند. ادغام محتویات[4] در گام آخر بر روی نتایج دو گام قبلی انجام می‌گیرد.

اهداف اصلی این مقاله عبارت هستند از

- انتخاب کلاسترهای منتخب جهت خلاصه‌سازی بهتر و بهینه‌تر داده ویدئویی انجام می‌گیرد.
- جاسازی زمانی و گرافی انجام گرفته تا در نهایت و با ترکیب و تلفیق آنها، خلاصه‌سازی نهایی انجام گیرد.
- نتیجه با کمک دو روش کمی و کیفی بر روی دیتاست‌های SumMe [۶]، TVSum [۷] و OpenCV [۸] انجام گرفت که بر اساس این بررسی کارایی راهکار پیشنهادی اثبات شد.

مقاله به شرح زیر سازماندهی شده است: بخش ۲ پیشینه و کارهای مرتبط را بررسی می‌کند. بخش ۳ چارچوب پیشنهادی را تشریح می‌نماید و بخش ۴ به بررسی ارزیابی روش پیشنهادی می‌پردازد. نهایتاً نتیجه‌گیری در بخش ۵ ارائه شده است.

۲- مرور ادبیاتی

نسخه کوتاه‌تر خلاصه می‌کند. مقایسه دو نوع خلاصه‌سازی ویدئویی نشان در زمینه کاوش ویدئویی و به خصوص خلاصه‌سازی ویدئویی کارهای تحقیقاتی زیادی انجام گرفته است.

راجات[5] و همکاران در سال ۲۰۱۶ به خلاصه‌سازی رویدادها در ویدئو پرداختند. در روش آنها ویدئو به تمام اجزای موجود در محتویات آن تقسیم شد (شامل متن، صدا، صورت‌ها و ...) و سپس بر این اساس خلاصه‌سازی ویدئو انجام گرفت [۹]. دیمو[6] و همکاران در سال ۲۰۱۵ با کمک روش کاربرمحور سعی در خلاصه‌سازی ویدئو داشتند. در این روش از اطلاعات سطح بالای فریم‌های حذف‌شده و اطلاعات سطح پایین ویدئو استفاده می‌شود تا خلاصه‌سازی با معنای بهتر ارائه گردد [۱۰].

اصولاً هر دو روش مبتنی بر فرهنگ لغت و خوشه‌بندی، رویکردهای بدون نظارت هستند و این مستلزم آن است که خلاصه تولیدشده آموزنده یا متنوع باشد. وانگ[7] و همکاران [۱۱] یک روش کلاسترینگ ارائه کردند که بر اساس آن روش k-mean را بهبود داده و در کنار آن با کمک ضریب سیلوئت[8] به بهینه‌ترین خلاصه‌سازی رسیدند. بیشوراجایا[9] و شارما[10] [۱۲]، روش شبکه عصبی کانولوشنال[11] را با تلفیق نتیجه نهایی با خوشه‌بندی K-means ارائه کردند و برای به‌دست‌آوردن روش بهینه‌تر در انتهای کار از ضریب سیلوئت کمک گرفتند. شاو[12] و چاودری[13] [۱۳] با ترکیب خوشه‌بندی K-means با فرمول کوله‌پشتی (IK)[14] به روش بهینه خلاصه‌سازی دست یافتند. دیگر روش‌های خوشه‌بندی، یادگیری عمیق هستند که کارایی خلاصه‌سازی را بهبود بخشیدند. معمولاً شبکه عصبی بازگشتی[15] برای رمزگشایی ارتباطات زمانی میان فریم‌های ویدئویی به کار می‌رود. ژانگ[16] و همکاران [۱۴] با کمک‌گرفتن از LSTM استخراج فریم‌های حاوی مفهوم را انتخاب نموده و از فرایند نقطه‌ای (DPP)[17] برای بهبود انتخاب فریم‌ها استفاده کردند.

1. Ping Li
2. Temporal Graph Embedding
3. Graph Neural Networks
4. Context Fusion
5. Rajat Aggarwal
6. Anastasios Dimou
7. Fengsui Wang
8. Silhouette Coefficient
9. Madhushree Basavarajaiah
10. Ananda S. Chowdhury
11. Convolutional Neural Network
12. Abhimanyu Sahu
13. Ananda S. Chowdhury
14. Integer Knapsack Type Formulation
15. Recurrent Neural Network
16. Zhang
17. Determinantal Point Process

برای خودکارکردن خلاصه‌سازی، راکن[1] و وانگ[2] [۱۵] خلاصه‌سازی ویدئو را بر اساس یادگیری تابع نگارش داده‌ای از روی داده‌های ناهمتا و با کمک محدودیت‌های موجود ایجاد کردند. برای بررسی ساختار زمانی ویدئو و اعمال تنوع محلی، لی[3] و همکاران [۱۶] مدل احتمالی بر اساس کنترل پویای فواصل زمانی ایجاد کردند و با کمک الگوریتم‌های یادگیری آن را آموزش دادند. با این کار نه تنها خلاصه‌ای از ویدئو ساخته شد که می‌توان محتوای اصلی ویدئو را تفسیر کرد. روش آنها ژاوو[4] و همکاران را را بر آن داشت [۱۷] تا روش یادگیری دوسویه را ارائه دهند. این قالب کاری با ترکیب خلاصه ایجاد شده و ساختار ویدئو، اطلاعات ویدئویی زمانی و فضایی را استخراج نماید. شاو و چاودری [۱۸] برای خلاصه‌سازی ویدئوی فردساخته از نمایش‌های مختلف گرافی داده ویدئویی استفاده کرده‌اند.

شاو و چاودری با کمک انواع نمایش‌های گرافی، خلاصه‌سازی او شخص با صحت بالاتر ارائه کردند [۱۸]. روش تشخیص مرز شات با استفاده از اطلاعات متقابل مبتنی بر گراف، ارائه و سپس یک نمودار وزنی برای هر عکس ایجاد گردید. یک فریم نماینده از هر شات با استفاده از معیار مرکزیت نمودار انتخاب گشت. روش جدیدی برای مشخص‌کردن فریم‌های ویدئویی خودمحور با استفاده از یک مدل مرکز فراگیر مبتنی بر نمودار در ادامه نشان داده شده است. در اینجا هر قاب نماینده به عنوان اتحاد یک منطقه مرکزی (گراف) و یک منطقه فراگیر (گراف) مدل‌سازی می‌شود. با بهره‌برداری از معیارهای طیفی عدم تشابه بین دو گراف (مرکز و اطراف)، مناطق بهینه مرکز و اطراف تعیین می‌شوند. نواحی بهینه برای همه فریم‌ها در یک عکس مانند قاب نماینده نگه داشته می‌شود. تفاوت‌های مرکز فراگیر در مقادیر آنتروپی و جریان نوری همراه با خصیصه PHOG ([5]HOG هرمی) از هر فریم استخراج می‌گردند. تمام فریم‌ها در یک ویدئو در نهایت با نمودار وزنی دیگری نشان داده می‌شوند که به عنوان نمودار شباهت ویدئو (VSG) نامیده می‌شود. فریم‌ها با استفاده از رویکرد مبتنی بر حداقل درخت پوشا (MST) با معیاری جدید برای لبه‌های غیرقابل قبول دسته‌بندی می‌گردند. نزدیک‌ترین فریم به مرکز هر خوشه برای ساخت خلاصه انتخاب می‌شود.

لی و همکاران [۵] شبکه‌های کانولوشنی گرافی [6](GCAN) را معرفی کردند. این روش با ترکیب روش گرافی و زمانی به یک سری اطلاعات و داده‌های متقابل می‌رسد که با ترکیب آنها به شات‌های کلیدی کاندیدا می‌رسد. تعداد روش‌هایی که از گراف در خلاصه‌سازی ویدئویی استفاده کردند، بسیار کم است و در واقع استفاده از گراف باعث افزایش صحت و بالارفتن دقت می‌گردد.

لیانگ[7] و همکاران یک شبکه کانولوشنی CAAN[8] را پیشنهاد داده‌اند داده‌اند که ایده اصلی آن، ایجاد یک خلاصه‌کننده عمیق می‌باشد که روشی غیرنظارتی است [۱۹]. قالب کلی این روش پیشنهادی از یک مولد و یک جداکننده تشکیل شده است. اولی اهمیت امتیازها را برای همه فریم‌های یک ویدئو پیش‌بینی می‌کند؛ در حالی که دومی سعی می‌کند ویژگی‌های فریم‌های وزن‌دارشده با امتیاز را از ویژگی‌های فریم اصلی متمایز کند. برای ثبت رابطه زمانی عمومی و محلی فریم‌های ویدئویی، مولد از یک شبکه توالی کانولوشنال برای ساخت نمایش عمومی یک ویدئو استفاده می‌کند. برای بهینه‌سازی پارامترها، تابع هدف از سه تابع ضرر تشکیل شده که می‌تواند پیش‌بینی امتیاز اهمیت سطح فریم را به طور مشترک هدایت کند. مقایسه روش پیشنهادی با روش‌های موجود، کارآمدی روش پیشنهادی را اثبات می‌کند.

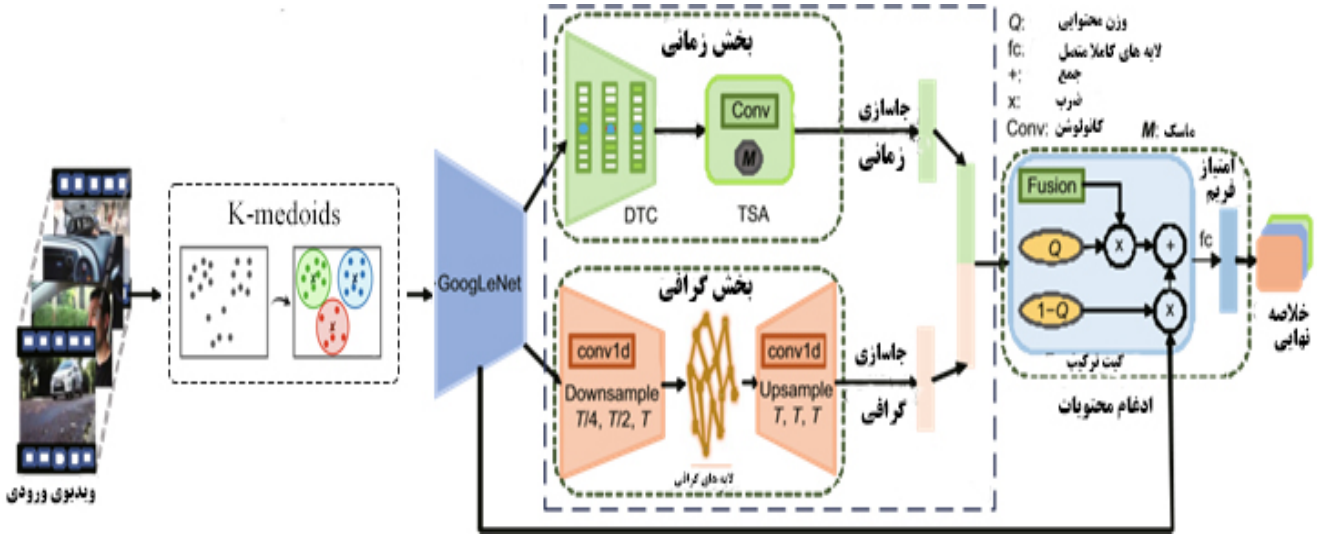
یانگ[9] و همکاران در راهکار خود از دوربین صنعتی پرسرعت برای ضبط ویدئو و از شبکه‌های عصبی کانولوشنال عمیق در زمینه اندازه‌گیری ارتعاشات بصری استفاده کرده‌اند [۲۰] و بر این اساس، دقت تشخیص هدف و سرعت ردیابی جابه‌جایی را اندازه‌گیری می‌کند. از مدل شبکه‌های عصبی کانولوشنال برای شناسایی ارتعاشات در ویدئو استفاده شده است. نشان داده شده که کارایی محاسباتی و کمیت پارامتر شبکه‌های کانولوشنال در ردیابی جابه‌جایی و حرکت در ویدئوهای بالا می‌باشد. در ابتدا شبکه عصبی کانولوشن سبک‌وزن به عنوان شبکه ستون فقرات در نظر گرفته می‌شود. در این راهکار شبکه عصبی کانولوشن استاندارد با کانولوشن قابل تفکیک عمقی و کانولوشن نقطه‌ای جایگزین می‌گردد. سپس با توجه به مزیت‌های سرعت و دقت الگوریتم‌های یادگیری عمیق برای ردیابی اشیای ویدئویی، فاصله‌های مرکز شیء و اندازه‌های مرزی را با کمک یک الگوریتم تشخیص در شبکه تخمین می‌زند. از روش شناسایی مجدد (re-ID) برای تقویت همبستگی جابه‌جایی هدف بین فریم‌های مجاور استفاده می‌شود. راهکار پیشنهادی در مقایسه با روش‌های موجود برتری خود را نشان داده است.

اسکندر و همکاران سعی کردند که یک شبکه عصبی کانولوشنال تغییریافته جدید را ارائه دهند [۲۱]؛ یعنی یک مدل نظارتی از یادگیری عمیق را برای خلاصه‌سازی ویدئوهای کریکت ارائه داده‌اند. شبکه عصبی کانولوشن کریکت پیشنهادی (C-CNN) آموزنده‌ترین ویژگی‌ها را از فریم‌های ویدئویی می‌آموزد و طبقه‌بندی باینری را به کلاس مثبت و منفی انجام می‌دهد. با کمک این روش می‌توان فریم‌های کلیدی را کل ویدئو یافته و خروجی خلاصه‌ای از کل ویدئو به دست آورد. با بررسی تجربی راهکار پیشنهادی، برتری این روش نسبت به روش‌های مشابه نشان داده شده است.

۳- راهکار پیشنهادی

در شکل ۱ مدلی کلی از راهکار پیشنهادی قابل مشاهده است. در ابتدا با کمک الگوریتم k-medoids کلاسترهایی ایجاد گردیده و با انتخاب فریم‌های کلیدی، یک سری فریم‌های نماینده انتخاب می‌شوند که برای مرحله آخر و رأی‌گیری میان فریم‌های نتیجه باقی می‌مانند. از سوی دیگر، تفاوت میان فریم‌های کلیدی سنجیده شده و بسته به یک حد آستانه، تعدادی از این فریم‌ها انتخاب شده و کلاسترهای آنها به عنوان ورودی مرحله بعدی و شروع خلاصه‌سازی جدید در نظر گرفته می‌شوند. در این گام با الهام از روش پیشنهادی لی و همکاران [۱۲] در ابتدا با کمک GoogleNet ویژگی‌های یک تصویر جداسازی می‌شوند. سپس در دو گام با کمک شاخه زمانی و گرافی یک گروه از شات‌های ویدئویی فراهم می‌آید و در انتها با کمک همین شات‌ها خلاصه دینامیکی از ویدئو به دست می‌آید که معنای ویدئوی کلی را به خوبی در خود حفظ کرده است.

1. Rochan
2. Wang
3. Y. Li
4. Zhao
5. Histogram of Oriented Gradients
6. Graph Convolutional Attention Network
7. Guoqiang Liang
8. Convolutional Attentive Adversarial Network
9. Rongliang Yang



شکل ۱: مدل سیستم پیشنهادی.

۳-۱ الگوریتم K-medoids

خوشه‌بندی برای تحلیل داده‌ها و اطلاعات به کار می‌رود. در این روش که به یادگیری غیرنظارتی هم معروف است نقاط داده را با کمک بیشینه‌کردن شباهت درون خوشه و کمینه‌کردن شباهت با نقاط داده خارج خوشه، خوشه‌بندی می‌کنند.

الگوریتم k-Medoids که بهبودیافته الگوریتم k-Means است، عملکردی بسیار شبیه با آن الگوریتم دارد؛ با این تفاوت که در الگوریتم k-Medoids به‌جای استفاده از میانگین، از خود نمونه‌ها برای مرکز ثقل و نمایندگی خوشه‌ها استفاده می‌شود. با انتخاب نمونه‌های واقعی جهت نمایش یک خوشه، حساسیت روش نسبت به نمونه‌های نویز و خارج از محدوده کاهش می‌یابد. بنابراین روش k-Medoids برخلاف روش k-Means به جای اینکه مقادیر میانگین از نمونه‌ها را دریافت کند، از مرکزی‌ترین نمونه موجود در خوشه به عنوان نمایش و نماینده خوشه استفاده می‌کند.

با کمک الگوریتم K-Medoids در ابتدا ویدئو به فریم‌های کلیدی که نشان‌دهنده فعالیت خاصی هستند، تقسیم می‌شود. جهت بهبود روش K-Medoids تلاش گردیده در انتخاب k با جستجوی سریع در میان فریم‌های موجود، تعداد فریم‌هایی که از لحاظ ویژگی تصویری با یک میزان آستانه، تفاوت بیشتری دارند، شمرده شوند که این میزان برابر با تعداد K است. استفاده از این خوشه‌بندی قبل از به‌دست‌آوردن سازگاری باعث بالارفتن کارایی، دقت و حذف نویز می‌شود. پس از انتخاب فریم‌های کلیدی با کمک ضریب سیلوئت، تعدادی فریم به عنوان فریم کاندیدا انتخاب می‌گردند و خوشه‌های آنها نیز به عنوان خوشه‌های کلیدی و در واقع داده‌های گام بعدی برگزیده می‌شوند

$$a(i)=\frac{1}{|c_i|-1}\sum_{\substack{j\in c_i\\ i\neq j}} d(i,j) \qquad (1)$$

رابطه (۱) میانگین فاصله بین نقطه i و سایر نقاط داده در همان خوشه است که در آن $d(i,j)$ فاصله بین نقاط داده i و j در خوشه C_i است. (تقسیم) $|C_i|-1$ انجام می‌شود زیرا فاصله $d(i,i)$ در جمع وارد نمی‌گردد). $a(i)$ در واقع نشان می‌دهد که چه اندازه i به خوبی به خوشه خود اختصاص داده شده است (هرچه مقدار کوچک‌تر باشد، انتساب بهتر است). سپس میانگین عدم شباهت نقطه i با بعضی خوشه‌های C_k به عنوان میانگین فاصله i به تمام نقاط در C_k تعریف می‌شود (که در آن $c_k \neq c_i$) و بنابراین برای هر نقطه داده $i \in c_i$ اکنون تعریف می‌شود

$$b(i)=\min_{k\neq i}\frac{1}{|c_k|}\sum_{j\in c_k} d(i,j) \qquad (2)$$

کوچک‌ترین‌بودن (عملگر min در فرمول)، میانگین فاصله i با همه نقاط در خوشه‌های دیگر است که i عضو آنها نیست. خوشه‌ای با کوچک‌ترین میانگین نابرابری، خوشه همسایه i نامیده می‌شود زیرا این بهترین خوشه مناسب بعدی برای نقطه i است و در نهایت برای نسبت‌دادن هر گره به خوشه مناسب خود از ضریب سیلوئت استفاده می‌شود که مقدار آن برای نقطه داده i برابر است با

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)} & , \ a(i)<b(i)\\ 0 & , \ a(i)=b(i)\\ \dfrac{a(i)}{b(i)}-1 & , \ a(i)>b(i)\end{cases} \qquad (3)$$

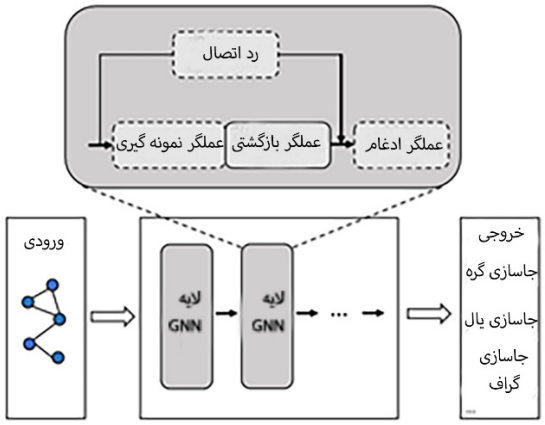
۳-۲ جداسازی زمانی و گرافی

برای جداسازی زمانی و گرافی در ابتدا از GoogleNet به عنوان شبکه کانولوشنی استفاده می‌شود. GoogleNet یک شبکه عصبی کانولوشن عمیق ۲۲لایه است که محققان Google آن را ساخته‌اند. این شبکه به دلیل وجود عمق به افزایش عملکرد چشم‌گیرش می‌انجامد. در هر سکانس ویدئویی، مجموعه $\{f_i\}_{i=1}^{T}$ شامل T فریم و f_i نشان‌دهنده iامین فریم است، به عنوان ورودی وارد GoogleNet [۲۲] شده تا ویژگی‌های فریم را استخراج کند. این ویژگی‌ها یعنی $X=[x_1,x_2,\ldots,x_T]\in\mathbb{R}^{d\times T}$ که نشانگر $x_i\in\mathbb{R}^d$امین فریم با d عنصر است. این فریم‌ها به طور جداگانه و مستقل وارد بخش زمانی و گرافی می‌شوند.

جاسازی زمانی با $B=[b_1,b_2,\ldots,b_T]$ و جاسازی گرافی با $G=[g_1,g_2,\ldots,g_T]$ مشخص می‌گردند. بنابراین در بردار و برای iامین فریم، b_i نشان‌دهنده جاسازی زمانی و g_i نشان‌دهنده جاسازی گرافی است.

۳-۲-۱ شاخه زمانی

ورودی از گام قبلی وارد این مرحله می‌شود. برای یادگیری جاسازی زمانی فریم‌ها از چندین هسته گسترش‌یافته کانولوشنال $W^r\in\mathbb{R}^{d\times w}$ با



شکل ۲: قالب کاری روش پیشنهادی.

پهنای زمانی w و نرخ کانولوشن r استفاده می‌شود تا ارتباط فضایی میان فریم‌ها را ثبت کند. خروجی جاسازی محلی در این روش برای مجموعه $\{x_i^r\}_{i=1}^{T}$ برابر است با

$$x_i^r = \sum_{j=1}^{w} (x_i + r_j \odot w_j^r) \quad (۴)$$

که $x_i^r \in \mathbb{R}^d$ و $W_i^r \in \mathbb{R}^d$ و نرخ کانولوشن r برای به‌دست‌آوردن پیچیدگی زمانی کانولوشنال (DTC)[1] به کار می‌روند. DTC میدان پذیرش نورون‌ها را بدون آنکه وضوح را کاهش دهد، بزرگ می‌کند. در اینجا $r = ۲$ و $w = ۳$ قرار داده شده است. سمبل "$\odot$" به معنی ضرب عنصرهاست. هرم کانولوشنالی زمانی از N عملگر کانولوشن به صورت موازی ایجاد شده و نرخ کانولوشن آن نیز در حال افزایش است. N امین نرخ کانولوشن $r_n = ۲^{n-1}$ برای افزایش طول بعد زمانی به کار می‌رود. خروجی کانولوشن‌های به هم متصل می‌شوند تا بردار ویژگی زمانی بهتری را ایجاد کنند؛ یعنی

$$C_i = [x_i^{r_1}, x_i^{r_2}, \dots, x_i^{r_N}] \in \mathbb{R}^{Nd} \quad (۵)$$

که منجر به جاساز زمانی $C \in \mathbb{R}^{Nd \times T}$ می‌گردد.

۳-۲-۲ شاخه گرافی

با توجه به اینکه در روش ارائه‌شده توسط لی و همکاران از رویکرد فضایی پایه استفاده شده که نمی‌تواند برای گراف‌های با مقیاس بالا به خوبی عمل کند [۱۲]، به همین خاطر در اینجا از شبکه عصبی گراف استفاده شده که در شکل ۲ آمده است. ورودی در اینجا گراف سیگنال x ($x \in \mathbb{R}^{N \times F}$ که $\mathbb{R}^{N \times F}$ فضای $N \times F$ بعدی اقلیدسی می‌باشد) است که در ابتدا و با کمک انتقال گراف فوریه $\mathcal{F}(x)$ به دامنه طیف منتقل می‌گردد و در ادامه عملیات کانولوشنالی اتفاق می‌افتد. بعد از کانولوشنالی، سیگنال نتیجه دوباره با کمک فوریه معکوس $\mathcal{F}(x)^{-1}$ به دست می‌آید. این انتقال‌ها به صورت زیر حاصل می‌شوند

$$\mathcal{F}(x) = U_x^T \quad (۶)$$

$$\mathcal{F}(x)^{-1} = U_x \quad (۷)$$

که U ماتریس بردارهای ویژه از گراف نرمال‌شده لاپلاس است. گراف لاپلاس $L = U \Lambda U^T$ بوده که Λ ماتریس مثلثی مقادیر ویژه است. کانولوشنالی گراف لاپلاس نرمال‌شده با کمک فرمول زیر به دست می‌آید

$$g \times x = \mathcal{F}^{-1}(\mathcal{F}(g) \odot \mathcal{F}(x)) = U_g^T g \odot U_x^T \quad (۸)$$

که U_g^T فیلتر دامنه طیف است. فیلتر با کمک ماتریس مثلثی ساده‌شده g_w ساده می‌گردد

$$g_w \times x = U_{g_w} U_x^T \quad (۹)$$

۳-۲-۲-۱ عملگر بازگشتی

در یک گراف، هر گره با ویژگی‌هایش مشخص می‌شود. در این گام از شبکه عصبی گراف یا GNN استفاده می‌شود که این مدل می‌تواند برای تجزیه و تحلیل نمودارها استفاده گردد. گراف‌ها ساختارهای داده‌ای قوی هستند که شامل روابط بین اشیا می‌باشند و GNNها اجازه می‌دهند که این روابط را به روش‌های جدیدی کشف کنند. مثلاً می‌توان از GNN برای شناسایی افرادی که احتمالاً محصولی را در رسانه‌های اجتماعی توصیه می‌کنند، استفاده کرد. در اینجا هدف GNN یادگیری شرایطی است که در آن $h_y \in \mathbb{R}^s$ و شامل اطلاعات خود و همسایه‌ها برای هر گره است. وضعیتی که h_y در آن جاساز شده، بردار s بعدی گره v برای ایجاد خروجی O_v است. گام‌های محاسبه این مقادیر برابر است با

$$h_v = f(x_v, x_{co[v]}, h_{N_v}, x_{N_v}) \quad (۱۰)$$

$$o_v = g(h_v, x_v) \quad (۱۱)$$

که x_v، $x_{co[v]}$، h_{N_v} و x_{N_v} ویژگی‌های V، ویژگی یال‌ها، وضعیت و ویژگی گره‌های همسایه v است. F تابع انتقال محلی است که میان تمام گره‌ها به اشتراک استفاده شده و وضعیت هر گره را بر اساس ورودی همسایه‌هایش مشخص می‌کند. g نیز تابع خروجی محلی است که نحوه ایجاد خروجی را بیان می‌کند.

۳-۲-۲-۲ ماژول ردکردن اتصال

مسلم است که در شبکه‌های عصبی، هرچه میزان لایه‌ها بیشتر باشد به نتایج بهتری می‌رسیم؛ اما از سوی دیگر، داده‌های بیشتر باعث ایجاد روابط بالا و پیش‌آمدن نویز یا پیچیدگی می‌گردند. یک روش برای حل این مسأله، روش ردکردن اتصال است تا بعضی از اتصالات موجود در شبکه (که اهمیت کمتری دارند) نادیده گرفته شوند. به عنوان یک روش می‌توان از روش رحیمی و همکاران [۲۳] بهره برد که در آن، خروجی یک لایه با ورودی آن جمع می‌گردد و از این رقم به‌دست‌آمده برای مقادیر گام بعدی استفاده می‌شود.

۳-۲-۲-۳ ماژول ادغام[2]

از روش ادغام ساده گره استفاده شده است. در این مدل عملگرهای مربوط به گره یعنی بیشینه، کمینه و مجموع به کار می‌روند تا ارائه‌ای جامع از گره داشته باشند. با کمک این عملگرها نتایج حاصل از گام‌های پیش با هم ادغام شده و خروجی حاصل بیانگر نتیجه نهایی است.

۳-۲-۲-۴ ادغام محتوا

در آخرین گام باید میان دو جاسازی زمانی $B \in \mathbb{R}^{Nd \times T}$ و گرافی $G \in \mathbb{R}^{d \times T}$ به‌دست‌آمده ادغام انجام گیرد تا شات‌های کاندیدا انتخاب گردیده و خلاصه‌سازی انجام گیرد. دو نوع اطلاعات در متغیر $Z = [B, G] \in \mathbb{R}^{(N+1)d \times T}$ با هم ادغام می‌شوند. این گام حاوی لایه خطی با تابع فعال‌سازی، لایه خطی با تابع سیگموئید[3] و گیت ترکیب[4] می‌باشد. فرمول ریاضی این دو لایه به شرح زیر است

1. Dilated Temporal Convolution
2. Pooling Module
3. Sigmoid Function
4. Fusion Gate

$$z' = ReLu(w_z z) \in \mathbb{R}^{d\times T} \quad (۱۲)$$

$$Q = sigmoid(w_g z) \in \mathbb{R}^{d\times T} \quad (۱۳)$$

که $W_g = \mathbb{R}^{(n+1)d\times T}$ و $W_z = \mathbb{R}^{d\times(n+1)d}$ وزن‌های ماتریس Q و Z' خروجی‌های دو لایه خطی است. برای حفظ معنی بین فریم‌ها، دو ماتریس ویژگی فریم $X^{(d\times T)}$ و ساختار گرافی Z' با کمک گیت ترکیب با هم ترکیب می‌شوند تا $Z_f \in \mathbb{R}^{(d\times T)}$ حاصل گردد

$$Z_f = Z' \odot Q + x \odot (1-Q) \quad (۱۴)$$

با کمک اعمال مقدار حاصل‌شده از فرمول بالا بر روی گره‌ها، مقداری نامنفی حاصل می‌گردد که اهمیت هر گره را نشان می‌دهد و در این گام با کمک الگوریتم تقسیم‌بندی زمانی هسته (KTS)[1] برای انتخاب شات‌های کلیدی به کار می‌رود.

٤- ارزیابی راهکار پیشنهادی

در این بخش به معرفی تنظیمات کلی انجام آزمایش بر روی نتایج حاصل از راهکار پیشنهادی پرداخته شده و نتایج مورد بررسی قرار می‌گیرند.

٤-١ دیتاست

کارایی راهکار پیشنهادی بر روی سه دیتاست SumMe، TvSum و OpenCv مورد ارزیابی قرار گرفت. SumMe شامل ۲۵ ویدئو با طول یک تا شش دقیقه بوده که رویدادهای مختلف ورزشی، تعطیلات و ... را پوشش می‌دهد. TVSum نیز شامل ۵۰ ویدئوی ویرایش‌شده است که در ۱۰ گروه مختلف با طول ۱/۵ تا ۱۱ دقیقه هستند. این دو دیتاست توسط کاربرها نشانه‌گذاری شده‌اند که این نشانه‌ها میزان اهمیت فریم‌ها را برای هر ویدئو نشان می‌دهند. دیتاست سوم نیز OpenCv است که از ۵۰ ویدئو با موضوعات مختلف و بیشتر در جنبه مستند ساخته شده است. طول این ویدئوها حداکثر تا ۱۰ دقیقه هستند که برای ارزیابی در بخش کمی مورد استفاده قرار گرفتند.

٤-٢ ارزیابی کیفی

در امر خلاصه‌سازی ویدئویی هنوز استانداردی برای سنجش و ارزیابی نتیجه به صورت کیفی وجود ندارد. به دلایل مختلفی چون سلیقه افراد، تخصصی‌بودن خلاصه در رشته‌های متفاوت و ... نمی‌توان نظر درستی در مورد خوبی یا بدی یک خلاصه داد و در نتیجه نمی‌توان به راحتی عمل ارزیابی را به انجام رساند.

در اینجا سعی شده تا یکی از راهکارهای کیفی ارزیابی خلاصه‌سازی به نام کاس (CUS)[2] استفاده شود که در واقع تغییر در روش F-measure و تبدیل آن به روشی کاربرمحور است [۲۴]. این روش نتیجه خلاصه‌سازی خودکار را با نظر کاربران می‌سنجد و ارزیابی مناسبی با نظر مستقیم کاربران انجام می‌دهد. تمامی ویدئوها با فرمت MPEG-۱ (۳۰ فریم در ثانیه و ۳۵۲×۲۴۰ پیکسل) هستند. ویدئوهای انتخاب‌شده بین چندین ژانر (مستند، آموزشی، زودگذر، تاریخی، سخنرانی) توزیع شده و مدت آنها از ۱ تا ۴ دقیقه متغیر است.

خلاصه‌های کاربران توسط ۵۰ کاربر ایجاد شده که هر کدام از آنها با ۵ ویدئو سر و کار داشتند؛ یعنی هر ویدئو دارای ۵ خلاصه ویدئو است که توسط ۵ کاربر مختلف ایجاد شده است. به عبارت دیگر، ۲۵۰ خلاصه ویدئویی به صورت دستی ایجاد شده‌اند.

در ابتدا از ۵۰ کاربر درخواست می‌شود تا فیلم را دیده و بعد به صورت دستی طبق میل خود خلاصه‌ای ایستا از آن فیلم تهیه کنند. برای راحتی کار، فریم‌های نمونه در اختیارشان قرار می‌گیرد تا از میان آنها خلاصه‌ها را پیدا کنند. کاربر در انتخاب فریم‌ها و تعداد آنها کاملاً مختار و آزاد است. در گام دوم، این فریم‌های انتخاب‌شده با فریم‌های خلاصه‌سازی خودکار مورد مقایسه قرار می‌گیرند. در گام سوم کیفیت این فریم‌ها با کمک دو مقیاس CUS_A و CUS_E سنجیده می‌شود

$$CUS_A = \frac{n_{mAS}}{n_{Us}} \quad (۱۵)$$

$$CUS_E = \frac{n_{\bar{m}AS}}{n_{Us}} \quad (۱۶)$$

که n_{mAS} برابر است با تعداد فریم‌های کلیدی مشابه بین خلاصه خودکار (AS) و خلاصه کاربران که دقیقاً برعکس $n_{\bar{m}AS}$ یعنی تعداد فریم‌های کلیدی که در این دو با هم برابر نیستند می‌باشد. n_{Us} نیز تعداد فریم‌های موجود در خلاصه کاربر (Us) است.

٤-٣ ارزیابی کمی

مقیاس F-measure به میزانی گسترده در مقالات مختلف جهت ارزیابی کارایی مورد استفاده قرار می‌گیرد [۱۲]، [۱۴]، [۱۷] و [۲۵]. برای تمام دیتاست‌ها، نشانه‌گذاری‌ها از سطح فریم به سطح شات تغییر سطح می‌دهند و شات‌های کلیدی را برای خلاصه‌هایی که زمانی کمتر از ۱۵٪ ویدئوی اصلی دارند، انتخاب می‌کنند. برای محاسبه F-measure در ابتدا مقادیر صحت[3] و پوشش[4] محاسبه می‌گردند. در این حالت صحت و پوشش با فرمول‌های زیر به دست می‌آیند

$$Precision = \frac{\left|S_{gen} \cup S_{human}\right|}{\left|S_{human}\right|} \quad (۱۷)$$

$$Recall = \frac{\left|S_{gen} \cup S_{human}\right|}{\left|S_{gen}\right|} \quad (۱۸)$$

که در روابط بالا S_{gen} خلاصه‌های ایجادشده توسط الگوریتم و S_{human} خلاصه ایجادشده توسط بشر است. بنابراین F-measure از رابطه زیر به دست می‌آید

$$F-measure = \frac{2\times precision \times recall}{precision + recall} \times 100\% \quad (۱۹)$$

٤-٤ تنظیمات ارزیابی

برای ارزیابی روش کمی نیاز به یک سری تنظیمات پیش‌پردازشی است و بر اساس آنچه ژانگ و همکاران ارائه کردند، این ارزیابی در ۳ گام انجام می‌پذیرد: مرکزی (C)[5]، تفضیلی (A)[6] و انتقالی (T)[7] که تنظیمات آن در در جدول ۱ آمده است. برای آموزش روش پیشنهادی از ۸۰٪ داده‌ها استفاده شده و مجموعه آزمون برابر با ۲۰٪ باقیمانده است.

1. Kernel Temporal Segmentation
2. Comparison of User Summaries
3. Precision
4. Recall
5. Canonical
6. Augmented
7. Transfer

جدول ۱: تنظیمات ارزیابی.

آزمون	آموزش	تنظیمات	دیتاست
۲۰٪ از SumMe	۸۰٪ از SumMe	C	
۲۰٪ از SumMe	OVP + YouTube + TVSum + ۸۰٪ SumME	A	SumMe
SumMe	OVP + YouTube + TVSum	T	
۲۰٪ از TVSum	۸۰٪ از TVSum	C	
۲۰٪ از TVSum	OVP + YouTube + SumME + ۸۰٪ TVSum	A	TVSum
TVSum	OVP + YouTube + SumME	T	

جدول ۲: مقادیر ارزیابی روش‌های موجود در برابر راهکار پیشنهادی [۱۹].

راهکار پیشنهادی	VSN [۲۶]	$VSUMM_1$ [۲۷]	$VSUMM_2$ [۲۷]	STIMO [۲۸]	DT [۲۹]	OV [۳۰]	
۰٫۸۹	۰٫۸۰	۰٫۸۵	۰٫۷۰	۰٫۷۲	۰٫۵۳	۰٫۷۰	CUS_A
۰٫۲۸	۰٫۲۶	۰٫۳۸	۰٫۲۷	۰٫۵۸	۰٫۲۹	۰٫۵۷	CUS_E

جدول ۳: تفاوت میان CUS_A در سطح اطمینان ۹۸٪ روش‌ها با روش پیشنهادی.

تفاوت‌ها	بازه اطمینان (۹۵٪)	
	مینیمم	ماکسیمم
روش پیشنهادی– $VSUMM_1$	۰٫۰۱	۰٫۰۳
روش پیشنهادی– $VSUMM_2$	۰٫۱۶	۰٫۱۸
روش پیشنهادی– OV	۰٫۱۲	۰٫۲۱
روش پیشنهادی– DT	۰٫۳	۰٫۳۸
روش پیشنهادی– STIMO	۰٫۱۳	۰٫۱۸

جدول ۴: تفاوت میان CUS_E در سطح اطمینان ۹۸٪ روش‌ها با روش پیشنهادی.

تفاوت‌ها	بازه اطمینان (۹۸٪)	
	مینیمم	ماکسیمم
روش پیشنهادی– $VSUMM_1$	−۰٫۱۵	−۰٫۰۱
روش پیشنهادی– $VSUMM_2$	−۰٫۰۱	۰٫۱۲
روش پیشنهادی– OV	−۰٫۴۲	−۰٫۱
روش پیشنهادی– DT	−۰٫۰۳	۰٫۰۷
روش پیشنهادی– STIMO	−۰٫۴	−۰٫۱۸

٤-۵ جزئیات پیاده‌سازی

ویژگی‌ها از روی تابعی با ۱۰۲۴ بعد در GoogleNet روی ImageNet آموزش داده شده‌اند. راهکار پیشنهادی بر روی ماشینی با پردازشگر Intel(R) Core(TM) i۷–۱۰۷۵۰H CPU@۲٫۶۰GHz GHz ۲٫۵۹ و NVIDIA GeForce GTX ۱۶۵۰ Ti پیاده‌سازی شده است.

۵- نتایج و تحلیل داده‌ها

این بخش به بررسی روش‌های ارزیابی راهکار پیشنهادی به دو روش کمی و کیفی می‌پردازد.

۵-۱ ارزیابی کیفی

نتایج حاصل از راهکار پیشنهادی توسط روش کاس بر روی ۵۰ ویدئو از سایت OpenCV مورد بررسی قرار گرفت. نرخ خطا و نرخ صحت به‌دست‌آمده در ابتدا با ۵ روش دیگر که آنها نیز بر روی ویدئوهای OpenCV اعمال شده‌اند، مقایسه می‌گردد که نتیجه این مقایسه در جدول ۲ مشخص است.

مقایسه دوبه‌دوی راهکارهای پیشنهادی با هر یک از ۵ روش با بازه اطمینان ۹۸٪ انجام گرفته که به نوعی برتری راهکار پیشنهادی را نشان می‌دهد. طبق جدول ۲، روش پیشنهادی دارای بالاترین میزان نرخ صحت نتایج است و نرخ خطای نسبتاً پایینی نیز دارد که این، بیانگر برتری روش می‌باشد. در ادامه در جداول ۳ و ۴، مقایسه دوبه‌دوی نتایج حاصل از راهکار پیشنهادی که توسط روش کاس به دست آمده است، با هر یک از ۵ روش و در بازه اطمینان ۹۸٪ انجام گرفته و به نوعی برتری راهکار پیشنهادی را نشان می‌دهند. در میان روش‌ها $VSUMM_1$ دومین رتبه را بعد از راهکار پیشنهادی دارد؛ اما در عین حال نرخ خطای آن بالاتر است. VSN نرخ خطای پایین‌تری نسبت به روش پیشنهادی دارد و توانسته از تخمین اشتباه جلوگیری کند؛ اما با وجود این نرخ صحت آن به میزان ۰٫۹ پایین‌تر از روش پیشنهادی است. با توجه به موارد قیدشده، راهکار پیشنهادی توانسته که بالاترین نرخ صحت و یکی از پایین‌ترین نرخ خطاها را کسب کند. طبق قاعده در سطوح اطمینان ۹۸٪ چنانچه ارزش مقادیر برابر با صفر باشد، نشان‌دهنده عدم کیفیت است. در هیچ کدام از مقایسه‌های انجام‌شده، نتیجه صفر حاصل نشده که نشانگر کیفیت بالای روش پیشنهادی است. از سوی دیگر همان طور که مشخص است نرخ صحت الگوریتم پیشنهادی از تمام روش‌ها بالاتر بهتر می‌باشد. در مورد نرخ خطا همان طور که نتایج نیز نشان می‌دهند، نرخ خطای روش پیشنهادی از اکثر روش‌ها پایین‌تر بوده و بهتر عمل می‌کند؛ اما نسبت به دو روش دیگر $VSUMM_2$ و VSN ضعیف‌تر است. با وجود این، تفاوت میان نرخ خطای روش پیشنهادی و این دو روش کم بوده و نرخ صحت بالاتری نیز نسبت به این دو دارد.

۵-۲ ارزیابی کمی

همان طور که در جدول ۵ آمده است، روش پیشنهادی در اکثر حالات دارای مقادیر بالاتر F-measure نسبت به روش‌های موجود است.

روش vsLSTM و dppLSTM هر دو خلاصه‌سازی با کارایی کمی دارند؛ چرا که برای مدل‌کردن روابط زمانی متغیرهای خود از LSTM کمک گرفتند و به دلیل محدودیت‌هایی که روش LSTM دارد نتوانستند به خوبی ارتباط میان فریم‌ها را در ویدئوهای با زمان و طول بالا نشان دهند. در صورتی که روش پیشنهادی با کمک‌گرفتن از TSA توانسته که به راحتی این مشکل را حل کند و روابط معنایی میان فریم و شات‌ها را نیز حفظ نماید. روش SUM-GAN نیز در روند خلاصه‌سازی خود از LSTM بهره برده و دقیقاً مانند روش‌های قیدشده در مورد ویدئوهای طولانی‌مدت دچار ایراد است. روش DR-DSN با به‌کارگیری روش‌های یادگیری و روش‌های rewardدهنده توانسته که بر تنوع ویدئوها فائق

جدول ۵: مقایسه کارایی روش پیشنهادی با روش‌های مختلف.

method	F-measure					
	SumMe			TVSum		
	C	A	T	C	A	T
vsLSTM [۱۴]	۳۷٫۶	۴۱٫۶	۴۰٫۷	۵۴٫۲	۵۷٫۹	۵۶٫۹
dppLSTM [۱۴]	۳۸٫۶	۴۲٫۹	۴۱٫۸	۵۴٫۷	۵۹٫۶	۵۸٫۷
$SUM\text{-}GAN_{sup}$ [۳۱]	۴۱٫۷	۴۳٫۶		۵۶٫۳	۶۱٫۲	
$DR\text{-}DSN_{sup}$ [۳۲]	۴۲٫۱	۴۳٫۹	۴۲٫۶	۵۸٫۱	۵۹٫۸	۵۸٫۹
HAS-RNN [۳۲]		۴۴٫۱			۵۹٫۸	
DyseqDPP [۱۶]	۴۴٫۳			۵۸٫۴		
$SASUM_{sup}$ [۳۳]	۴۵٫۳			۵۸٫۲		
SUM-FCN [۱۵]	۴۷٫۵	۵۱٫۱	۴۴٫۱	۵۶٫۸	۹۲٫۲	۵۸٫۲
$UnpairedVSN_{psup}$ [۳۴]	۴۸٫۰					
$CSNet_{sup}$ [۳۵]	۴۸٫۶	۴۸٫۷	۴۴٫۱	۵۸٫۵	۵۷٫۱	۵۷٫۴
A-AVS [۲۵]	۴۳٫۹	۴۴٫۶		۵۹٫۴	۶۰٫۸	
M-AVS [۲۵]	۴۴٫۴	۴۶٫۱		۶۱٫۰	۶۱٫۸	
$PCDL_{sup}$ [۱۷]	۴۳٫۷	۴۴٫۱		۵۹٫۲	۶۱٫۳	
$GCAN_{sup}$ [۱۲]	۵۳٫۰	۵۴٫۲	۴۶٫۸	۶۰٫۷	۶۱٫۱	۵۸٫۷
روش پیشنهادی	۵۴٫۰	۵۴٫۲	۵۰٫۰	۶۰٫۰	۸۹٫۹	۶۵٫۶

آمده و خلاصه‌سازی خوبی داشته باشد و مشکلات موجود در روش SUM_GAN را رفع کند؛ اما در کنار آن، مشکل پرهزینه‌بودن روش و قوانین مربوط به آموزش را دارد. HSARNN از LSTM دوسویه بهره می‌برد ولی باز هم به دلیل محدودیت‌های این روش توانسته است که روابط عمومی بین شات‌های ویدئوهای طولانی را به خوبی نشان دهد. DYSeqDPP روش قبلی خود یعنی SeqDPP را بهبود بخشیده و با کمک الگوریتم آموزشی به نتایج خوبی رسیده است؛ اما در عوض روش آموزش این الگوریتم ساده نیست و باعث ایجاد نقطه ضعف در این روش شده است. SASUM یا نیاز به تعریفگرهای متنی از ویدئو دارد که معمولاً دسترسی به آنها سخت یا غیرممکن است و یا نیاز به توصیفگرهای ویدئویی دارد که ساخت آنها پرهزینه بوده و اغلب برای خلاصه‌سازی به خوبی عمل نمی‌کنند.

این رویکرد با معرفی شبکه‌های کانولوشنالی، کارایی را در روش افزایشی بر روی SumMe بهبود بخشیده است؛ اما در عین حال نشانه‌های عمومی زمانی ویدئوها را در نظر نگرفته که باعث کاهش کیفیت نتایج در این روش شده است. CSNet از یک شبکه دوجریانه برای استفاده از هر دو ویژگی‌های عمومی و محلی فریم‌ها استفاده می‌کند؛ اما نتوانسته است که ساختار گرافی داده‌ها را در نظر بگیرد. AVS از دو تابع افزایشی[1] (A-AVS) و ضربی[2] (M-AVS) استفاده کرده و LSTM دومسیره را به کار برده تا به کارایی بالاتری در TVSum برسد؛ اما با این کار ساختار حقیقی نمونه فریم‌ها را از یاد برده و از خلاصه‌سازی مفهومی و معنایی فاصله گرفته است. GCAN نیز نسبت به سایر روش‌های مشابه برتری داشته که این نشانگر تأثیر استفاده از هر دو جاسازی گرافی و زمانی است. اما این روش کمی برای ویدئوهای طولانی‌مدت پرهزینه می‌باشد و به همین خاطر، روش پیشنهادی در ابتدا با روش خوشه‌بندی توانست که فریم‌های مهم‌تر را جداسازی کند و از حجم فریم‌ها و پیچیدگی‌های آتی بکاهد. در کنار آن با کمک تغییر روش گرافی و بهبود آن، نتایج بامعنی‌تری ایجاد نموده و صحت نتیجه را افزایش دهد.

۶- نتیجه‌گیری

با توجه به اهمیت داده‌های ویدئویی و عملیات پردازش این نوع داده، این مقاله به خلاصه‌سازی ویدئویی پرداخته است. برای رسیدن به این هدف، در ابتدا به وسیله الگوریتم خوشه‌بندی K-medoid و سپس کمک‌گرفتن از دو بخش جاسازی ساختاری زمانی و گرافی، خلاصه‌سازی روی شات‌های کلیدی برگزیده از مرحله قبل انجام می‌گردد. برای ارزیابی راهکار پیشنهادی سعی شده تا نتایج حاصل با کمک دو روش کیفی و کمی بر روی سه دیتاست مورد بررسی قرار گیرد. برای ارزیابی کیفی با الگوریتم کاس و تغییر در آن، می‌توان به ارزیابی درستی از نتیجه خلاصه‌سازی ویدئویی رسید. با این روش می‌توان نتایج خلاصه‌سازی خودکار را به صورت مستقل و جدا با نظر تک‌تک کاربران سنجید و در مورد کیفیت ویدئوها و نزدیکی آنها به ادراک انسانی تصمیم گرفت.

این راهکار بر روی ۵۰ ویدئو تکرار شد و میانگین ۸۹٪ نرخ صحت را در خلاصه‌سازی دارد و میزان خطای آن ۲۸٪ است که به نسبت سایر روش‌ها جزء بالاترین نرخ صحت‌ها می‌باشد و نرخ خطای آن نیز نسبت به اغلب روش‌ها پایین بوده و نسبت به سایر روش‌ها چندان بالا نیست.

نتیجه بررسی با روش کمی که در سه شرایط مختلف و بر روی دو دیتاست انجام گرفت، منجر به دست‌یافتن به F-measure با درصد بالا شده که نسبت به روش‌های مشابه، کارایی بالاتری از خود نشان داده است. برای کارهای آتی پیشنهاد می‌شود که با تغییر الگوریتم خوشه‌بندی گام اول سعی درد بهبود نتیجه نهایی داشت. همچنین با وجود اینکه قسمت مربوط به بخش گرافی یا شبکه عصبی گرافی نتایج مطلوبی داشته است اما می‌توان با بهینه‌ترکردن بخش زمانی و تغییر آن با الگوریتم‌های مشابه موجود، نتیجه نهایی را بهبود بخشید.

1. Additive
2. Multiplicative

مراجع

[1] A. Messina and M. Montagnuolo, "Fuzzy mining of multimedia genre applied to television archives," in *Proc. IEEE Int.Conf. on Multimedia and Expo*, pp. 117-120, Hannover, Germany, 23 Jun.-26 Apr. 2008.

[2] A. Bora and S. Sharma, "A review on video summarization approcahes: recent advances and directions," in *Proc. Int. Conf. on Advances in Computing, Communication Control and Networking, ICACCCN'18*, pp. 601-606, Greater Noida, India, 12-13 Oct. 2018.

[3] M. K. Mahesh and K. Pai, "A survey on video summarization techniques," in *Proc. Innovations in Power and Advanced Computing Technologies, i-PACT'19*, 5 pp., Vellore, India, 22-23 Mar. 2019.

[4] V. K. Vivekraj, D. Sen, and B. Raman, "Video skimming: taxonomy and comprehensive survey," *ACM Computing Surveys*, vol. 52, no. 5, Article ID: 106, 38 pp., Sept. 2019.

[5] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Computer Vision and Pattern Recognition*, vol. 111, Article ID: 107677, Mar. 2020.

[6] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," In: D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, (eds) *Computer Vision-ECCV'14*, Lecture Notes in Computer Science, vol 8695. Springer, pp. 505-520, 2014.

[7] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: summarizing web videos using titles," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'15*, pp. 5179-5187, Boston, MA, USA, 7-12 Jun. 2015.

[8] G. Bradski, A. Keahler, and V. Pisarevsky, "Learning-based computer vision with Intel's open source computer vision library," *Intel. Technology J.*, vol. 9, no. 2, pp. 119-130, May 2005.

[9] D. Zhao, J. Xiu, Y. Bai, and Z. Yang, "An improved item-based movie recommendation algorithm," in *Proc. 4th Int. Conf. on Cloud Computing and Intelligence Systems, CCI'16*, pp. 278-281, Beijing, China, 17-19 Aug. 2016.

[10] A. Dimou, D. Matsiki, A. Axenopoulos, and P. Daras, "A user-centric approach for event-driven summarization of surveillance videos," in *Proc. 6th Int. Conf. on Imaging for Crime Prevention and Detection, ICDP'15*, 6 pp., London, UK, 15-17 Jul. 2015.

[11] H. Zeng, *et al.*, "EmotionCues: emotion-oriented visual summarization of classroom videos," *IEEE Trans. on Visualization and Computer Graphics*, vol. 27, no. 7, pp. 3168-3181, Jul. 2021.

[12] P. Li, C. Tang, and X. Xu, "Video summarization with a graph convolutional attention network," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 6, pp. 902-913, 2021.

[13] S. S. de Almeida, *et al.*, "Speeding up a video summarization approach using GPUs and multicore CPUs," *Procedia Computer Science*, vol. 29, pp. 159-171, 2014.

[14] K. Zhang, W. L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," In: B. Leibe, J. Matas, N. Sebe, and M. Welling, (eds) *Computer Vision-ECCV'16*, Lecture Notes in Computer Scienc, vol 9911. Springer, pp. 766-782, 2016.

[15] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," In: V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, (eds) *Computer Vision-ECCV'18,* Lecture Notes in Computer Science, vol 11216. Springer, pp. 358-374, 2018.

[16] Y. Li, L. Wang, T. Yang, and B. Gong, "How local is the local diversity? reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization," In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) *Computer Vision-ECCV'18,* Lecture Notes in Computer Science, vol 11216. Springer, pp. 156-174, 2018.

[17] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3989-4000, Oct. 2020.

[18] B. U. Kota, A. Stone, K. Davila, S. Setlur, and V. Govindaraju, "Automated whiteboard lecture video summarization by content region detection and representation," in *Proc. 25th Int. Conf. on Pattern Recognition, ICPR'21*, pp. 10704-10711, Milan, Italy, 10-15 Jan. 2021.

[19] G. Liang, Y. Lv, S. Li, S. Zhang, and Y. Zhang, "Video summarization with a convolutional attentive adversarial network," *Pattern Recognition*, vol. 131, Article ID: 108840, Nov. 2022.

[20] R. Yang, S. Wang, X. Wu, T. Liu, and X. Liu, "Using lightweight convolutional neural network to track vibration displacement in rotating body video," *Mechanical Systems and Signal Processing*, vol. 177, Article ID: 109137, Sept. 2022.

[21] S. Sikandar, R. Mahmum, and N. Akbar, "Cricket videos summary generation using a novel convolutional neural network," in *Mohammad Ali Jinnah University Int. Conf. on Computing, MAJICC'22,* 7 pp., Karachi, Pakistan, 27-28 Oct. 2022.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'15*, 9 pp., Boston, MA, USA, 7-12 Jun. 2015.

[23] A. Rahimi, T. Cohn, and T. Baldwin, "Semi-supervised user geolocation via graph convolutional networks," in *Proc of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 2009-2019, Melbourne, Australia, Jul. 2018.

[24] A. P. Ta, M. Ben, and G. Gravier, "Improving cluster selection and event modeling in unsupervised mining for automatic audiovisual video structuring," In: K. Schoeffmann, B, Merialdo, A. G, Hauptmann, and C. W. Ngo, Andreopoulos, Y., Breiteneder, C. (eds) *Advances in Multimedia Modeling. MMM 2012. Lecture Notes in Computer Science,* vol 7131. Springer, Berlin, pp. 529-540, 2012.

[25] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709-1717, Jun. 2019.

[26] X. Li, Q. Li, D. Yin, L. Zhang, and D. Peng, "Unsupervised video summarization based on an encoder-decoder architecture," *J. of Physics: 5th Int. Conf. on Advanced Algorithms and Control Engineering, ICAACE'22,* vol. 2258, Article ID: 012067, Sanya, China, 20-22 Jan, 2022.

[27] S. E. F. de Avila, *et al.*, "VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56-68, Jan. 2011.

[28] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 529-540, Jan. 2009.

[29] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International J. on Digital Libraries*, vol. 6, no. 2, pp. 219-232, 2006.

[30] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *Proc. of the 6th ACM Int. Conf. on Multimedia*, pp. 211-218, Bristol, UK, 13-16 Sept. 1998.

[31] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2982-2991, Honolulu, HI, USA, 21-26 Jul. 2017.

[32] K. Y. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. on Artificial Intelligence*, pp. 7582-7589, New Orleans, LA, USA, 2-7 Feb. 2018.

[33] H. W. Wei, *et al.*, "Video summarization via semantic attended networks," in *Proc. AAAI Conf. on Artificial Intelligence*, pp. 216-223, New Orleans, LA, USA, 2-7 Feb. 2018.

[34] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proc IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 7894-7903, Long Beach, CA, USA, 15-20 Jun. 2019.

[35] Y. Jung, D. Cho, D. Kim, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proc AAAI Conf. on Artificial Intelligence*, pp. 8537-8544, Honolulu, HI, USA, 27 Jun.-1 Feb. 2019.

مهسا رحیمی رسکتی در سال ۱۳۸۷ مدرک کارشناسی مهندسی نرم‌افزار خود را از دانشگاه پیام نور و در سال ۱۳۹۲ مدرک کارشناسی ارشد مهندسی نرم‌افزار خود را از دانشگاه آزاد اسلامی قزوین و در سال ۱۴۰۰ مدرک دکتری مهندسی نرم‌افزار خود را از دانشگاه آزاد اسلامی ساری دریافت نمود. وی از سال ۱۳۸۹ تا کنون در دانشکده مهندسی کامپیوتر دانشگاه پیام نور و فرهنگیان مشغول به تدریس می‌باشد. زمینه‌های علمی مورد علاقه ایشان عبارتند از: داده کاوی، مهندسی نرم‌افزار و خلاصه‌سازی.

همایون موتمنی کارشناسی مهندسی کامپیوتر–نرم‌افزار را در سال ۱۳۷۴ از دانشگاه شهید بهشتی،کارشناسی ارشد مهندسی کامپیوتر– هوش را در سال ۱۳۷۷ ودکترای مهندسی کامپیوتر– نرم‌افزار را در سال ۱۳۸۶ از دانشگاه علوم وتحقیقات تهران اخذ نموده است وی از سال ۱۳۷۷ به‌عنوان عضو هیات علمی دانشگاه آزاد اسلامی بوده و هم‌اکنون استاد تمام دانشگاه آزاد اسلامی در رشته مهندسی کامپیوتر می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: مهندسی نرم‌افزار، ارزیابی کارایی، محاسبات تکاملی و سیستم‌های فازی.

ابراهیم اکبری مدرک دکتری خود را در رشته علوم کامپیوتر از دانشگاه فناوری مالزی، در سال ۱۳۹۴ دریافت نمود. وی اکنون استادیار گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد ساری است. تحقیقات او در زمینه تجزیه و تحلیل داده‌ها، الگوریتم‌ها و کاربردهای داده کاوی، یادگیری ماشین و تشخیص الگو است.

حسین نعمت‌زاده فارغ التحصیل دکترای علوم کامپیوتر (نرم افزار) از دانشگاه یو تی ام مالزی است. وی از سال ۱۳۹۱ عضو هیات علمی و استادیار گروه کامپیوتردانشگاه آزاد اسلامی واحد ساری بوده است. ایشان همچنین از سال ۱۴۰۰ در دانشگاه مالاگا اسپانیا به عنوان محقق در زمینه علم داده در حال فعالیت می باشند. زمینه تحقیقاتی ایشان به طور کلی حوزه علم داده و به طور خاص انتخاب ویژگی و هوش مصنوعی قابل توضیح است.