Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

A Transformer-Based Model for Abnormal Activity Recognition

Amir Mohammad Ahmady [a], Kourosh Kiani [b,*], Razieh Rastgoo [c]

[a] *Master's student, Faculty of Electrical and Computer Science, Semnan University, Semnan, Iran*
[b] *Associate Professor, Faculty of Electrical and Computer Science, Semnan University, Semnan, Iran*
[c] *Assistant Professor, Electrical and Computer Faculty, Semnan University, Semnan, Iran*

PAPER INFO

Paper history:

Received: 07 January 2024
Revised: 02 February 2024
Accepted: 16 February 2024

Keywords:

Video processing,
Video surveillance,
Abnormal activities,
Deep learning,
Transformer network.

ABSTRACT

Given the increasing daily volume of videos generated by security cameras in personal and public spaces, monitoring the activities present in videos has become crucial. Many video surveillance systems are designed to verify performance accuracy and provide alerts during the occurrence of abnormal activities. In this regard, various intelligent models have been proposed for detecting activities in videos. Considering recent advances in artificial intelligence, particularly deep learning, this paper introduces a model based on the Transformer network. To reduce computational complexity, keypoints of the human body are utilized in this approach. Fifteen key body points are input into the Transformer model, leveraging parallel processing during training and a self-attention mechanism. This enhances the speed and accuracy of the model. Experimental results on the JHMDB public database indicate an improvement in the accuracy of detecting abnormal activities compared to baseline models.

DOI: https://doi.org/10.22075/jme.2024.32914.2604

*** Corresponding author.**
E-mail address: kourosh.kiani@semnan.ac.ir

How to cite this article:
Ahmadi, A. M., Kiani, K., & Rastgoo, R. (2024). A Transformer-based model for abnormal activity recognition in video. Journal of Modeling in Engineering, 22(76), 213-221. doi: 10.22075/jme.2024.32914.2604

مقاله پژوهشی

بکارگیری مدل مبتنی بر ترنسفورمر برای تشخیص فعالیت‌های غیرطبیعی در ویدئو

امیر محمّد احمدی[۱]، کوروش کیانی[۲،*] ، راضیه راستگو[۳]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۲/۱۰/۱۷
بازنگری مقاله: ۱۴۰۲/۱۱/۱۳
پذیرش مقاله: ۱۴۰۲/۱۱/۲۷

واژگان کلیدی:
پردازش ویدئویی،
نظارت ویدئویی، اعمال غیرطبیعی،
یادگیری عمیق،
شبکه ترنسفورمر.

چکیده

با توجه به افزایش روز افزون حجم ویدئوهای تولید شده توسط دوربین‌های امنیتی و نظارتی در مکان‌های شخصی و عمومی، نظارت بر فعالیت‌های موجود در ویدئو امری حیاتی می‌باشد. بسیاری از نظارت‌های ویدئویی برای بررسی صحت عملکرد و هشدار هنگام وقوع یا انجام اعمال غیرطبیعی می‌باشد. در این راستا، مدل‌های هوشمند مختلفی جهت تشخیص فعالیت‌های موجود در ویدئو ارائه گردیده است. با توجه به پیشرفت‌های اخیر در حوزه هوش مصنوعی و به‌خصوص یادگیری عمیق، در این مقاله، مدلی مبتنی بر شبکه ترنسفورمر ارائه می‌گردد. در این راستا، به منظور کاهش میزان محاسبات، نقاط کلیدی بدن مورد استفاده قرار می‌گیرند. تعداد ۱۵ نقطه کلیدی بدن به مدل ترنسفورمر وارد می‌گردند تا با تکیه بر پردازش موازی این شبکه در حالت آموزش و نیز مکانیسم خودتوجهی، سرعت و دقت مدل افزایش داده شود. نتایج تجربی بر روی پایگاه داده عمومی JHMDB حاکی از بهبود دقت تشخیص فعالیت‌های غیرطبیعی نسبت به مدل¬های پایه می‌باشد.

DOI: https://doi.org/10.22075/jme.2024.32914.2604

۱- مقدمه

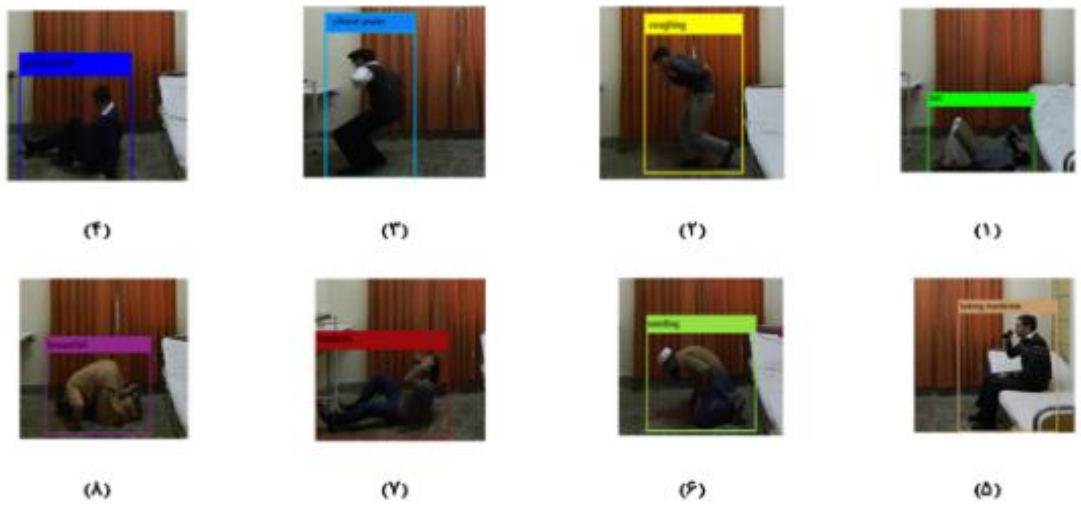
مفهوم شناسایی هوشمند فعالیت‌های غیرطبیعی بشر، امنیت داخلی و استانداردهای سیستم‌های نظارتی را بهبود و افزایش داده است [۱]. با توجه به افزایش استفاده از دوربین‌های نظارتی در محیط‌های عمومی، افزایش تعداد این دوربین‌ها و عدم امکان استفاده از نیروی انسانی برای کنترل و نظارت بر همه اطلاعات دریافتی از این دوربین‌ها، استفاده از مدل‌ها و روش‌های هوشمند برای تشخیص وقایع غیرطبیعی به صورت خودکار برای صرفه‌جویی در زمان و نیروی کار امری ضروری به نظر می‌رسد. اگرچه، چالش‌هایی در این راستا وجود دارد که از جمله مهمترین آنها می‌توان به دقت تشخیص مدل‌های هوشمند اشاره نمود.

تشخیص فعالیت‌های انسانی کاربردهای بلادرنگ مختلفی دارد. از جمله‌ی این کاربردها می‌توان به تشخیص زبان اشاره ناشنوایان [۴-۲]، نظارت بر بیمار [۱]، تشخیص فعالیت‌های غیرطبیعی [۵] اشاره نمود. در این میان، سیستم نظارت بر بیمار که در آن بیماران در میان گروهی از افراد عادی تحت نظر قرار می‌گیرند و سپس بر اساس فعالیت‌های غیرعادی آنها شناسایی می‌شوند [۵]، مورد توجه بسیاری قرار گرفته است. در شکل (۱) نمونه‌هایی از اعمال غیرطبیعی افراد بیمار شامل: ازهوش رفتن، سرفه کردن، درد قفسه سینه و از عقب افتادن و غیره نشان داده شده‌اند. در مثال قبل برای کمک به پرستاران و رسیدگی بهتر به بیماران، می‌توان از اطلاعات دوربین‌های نظارتی بیمارستان استفاده نمود و در

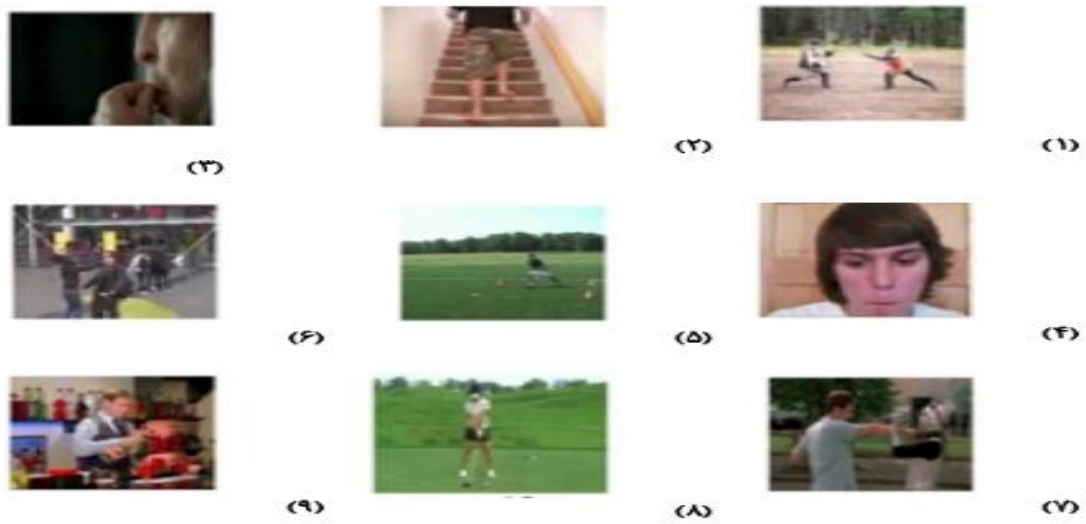
* پست الکترونیک نویسنده مسئول: kourosh.kiani@semnan.ac.ir
۱. دانشجوی کارشناسی ارشد، دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران
۲. دانشیار، دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران
۳. استادیار، دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران

استناد به این مقاله:
احمدی، امیر محمد، کیانی، کورش، و راستگو، راضیه. (۱۴۰۳). بکارگیری مدل مبتنی بر ترنسفورمر برای تشخیص فعالیت های غیرطبیعی در ویدئو. مدل سازی در مهندسی، ۲۲(۷۶)، ۲۱۳-۲۲۱. doi: 10.22075/jme.2024.32914.2604

صورتی که سیستم هوشمند عمل غیرطبیعی را تشخیص داد، هشدار و اعلان صوتی برای رسیدگی پرستاران به بخش مربوطه فعال شود. ویدئوهای ورزشی با توجه به جذابیت و محبوبیت آن همواره مورد توجه مخاطبان هر ورزش و کارشناسان برای تجزیه و تحلیل قرار می‌گیرد. خطاهای بازیکنان و آفسایدها در بازی فوتبال نمونه‌هایی از موضوعات تحلیل ویدئویی اعمال انسان در حوزه ورزشی است که می توانند به صورت خودکار تشخیص داده شود. از دیگر کاربردهای ویدئوهای ورزشی، آسیب شناسی حرکات ورزشکاران است که رفع آن به عملکرد بهتر بازیکنان کمک شایانی خواهد کرد. بعنوان مثال در بازی گلف، حرکت ورزشکار و ژست‌های نامطلوب آنان برای ضربه زدن به توپ ضبط می‌شود و توسط مربیان و کارشناسان تحلیل می‌شود تا نقاط ضعف بازیکن شناسائی و اصلاح شود. شکل (۲) نمونه‌هایی از ویدئوهای ورزشی در مجموعه داده JHMDB [۶] را نمایش می‌دهد.



شکل ۱- فعالیت‌های مختلف غیرطبیعی تشخیص داده شده شامل: (۱) از هوش رفتن؛ (۲) سرفه کردن؛ (۳) درد قفسه سینه؛ (۴) از عقب افتادن؛ (۵) مصرف کردن دارو؛ (۶) استفراغ کردن؛ (۷) سردرد؛ (۸) به جلو افتادن [۶].



شکل ۲- نمونه‌های از ویدئوهای موجود در پایگاه داده JHMDB شامل: (۱) شمشیر بازی، (۲) بالارفتن از پله، (۳) خوردن؛ (۴) جویدن، (۵) پریدن، (۶) در آغوش گرفتن، (۷) مشت زدن، (۸) گلف بازی کردن، (۹) ریختن [۶].

تشخیص فعالیت‌های انسانی یک حوزه تحقیقاتی چالش برانگیز برای درک و تجزیه و تحلیل داده‌های ویدیویی می باشد [۷]. به طور خاص، تشخیص فعالیت‌های غیرطبیعی افراد نیز از اهمیت خاصی برخوردار می‌باشد. در این راستا، مدل‌های مختلفی برای تشخیص هوشمند این فعالیت‌ها ارائه گردیده است. با توجه به پیشرفت‌های اخیر

در حوزه هوش مصنوعی [۹-۸] و به خصوص یادگیری عمیق [۱۶-۱۰]، مدل‌های مختلفی در راستای بهبود دقت تشخیص فعالیت‌های غیرطبیعی ارائه گردیده است. در این مقاله، مدلی بر مبنای شبکه ترنسفورمر ارائه می‌گردد. شبکه ترنسفورمر، یکی از مدل‌های یادگیری عمیق می‌باشد که با تکیه بر مکانیسم خودتوجهی و نیز پردازش موازی، قادر به پردازش داده‌های ورودی با دقت و سرعت بالا می‌باشد.

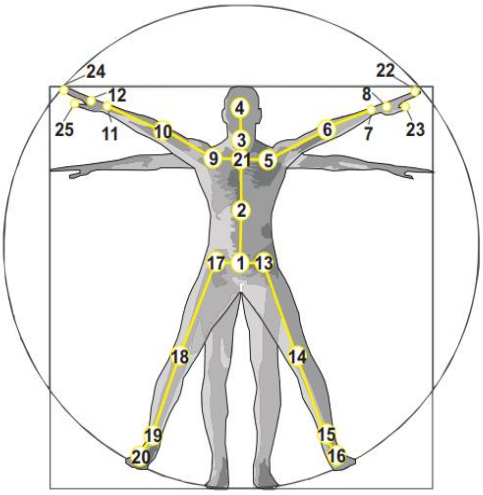
تشخیص اعمال انسان بر اساس ساختار اسکلتی به طور گسترده در کاربردهای چندرسانه‌ای مانند تعامل کامپیوتر-انسان[۲]، درک رفتار انسان[۳] و برنامه‌های کمک یار پزشکی[۴] استفاده می‌شود [۱]. استفاده از داده‌های مربوط به ساختار اسکلتی بدن انسان به دلیل کاهش حجم محاسبات، کاربردهای مختلفی در حوزه‌های مختلف دارد. در این مقاله نیز داده‌های اسکلتی بدن به‌منظور تشخیص فعالیت‌های غیرطبیعی انسان در ویدیو مورد استفاده قرار می‌گیرد. در این راستا، پس از استخراج داده‌های اسکلتی بدن انسان، این داده‌ها وارد شبکه ترنسفورمر می‌گردند تا با تکیه بر قابلیت‌های این شبکه بتوان دقت بالاتری برای تشخیص فعالیت‌های غیرطبیعی در ویدیو به دست آورد.

ساختار این مقاله به این صورت خواهد بود: در بخش بعدی، مروری کوتاه بر کارهای اخیر در حوزه‌ی تشخیص فعالیت انسان انجام می‌گردد. پس از آن، در بخش سوم، معرفی کوتاهی بر مدل ترنسفورمر ارائه می‌گردد. جزئیات مدل پیشنهادی به همراه نتایج به دست آمده نیز در بخش‌های ۴ و ۵ ارائه می‌گردند. در نهایت، در بخش ۶، جمع‌بندی و پیشنهادهای آینده مورد بحث قرار خواهند گرفت.

۲- مرور کارهای گذشته

در این بخش، مروری کوتاه بر کارهای اخیر در حوزه تشخیص فعالیت انسان ارائه می‌گردد. در مرجع [۵]، شبکه YOLO[۵] به عنوان یک مدل ستون فقرات بر مبنای شبکه عصبی کانولوشنی عصبی[۶] یا به اختصار CNN استفاده شده است. برای آموزش این مدل، مجموعه داده بزرگی از ویدئوهای بیمار، با برچسب‌گذاری هر فریم شامل اقدامات بیمار و موقعیت‌های بیمار ایجاد شده است. نتایج آزمایش مدل بر روی پایگاه داده مذکور حاکی از بهبود نتایج نسبت به مدل‌های موجود در زمینه تشخیص فعالیت‌های غیرطبیعی می‌باشد.

نمایش اسکلت یا نقاط کلیدی بدن انسان جزئیات عمیقی از وضعیت بدن انسان را به صورت فشرده ارائه می‌دهد. همین امر محققان را تشویق کرده است تا به کمک آن، کاربردهای بلادرنگ را توسعه دهند که این امر منجر به روند محاسبات سریعتر می‌گردد [۱۸-۱۷]. در این راستا، مجموعه داده‌های مختلفی ارائه گردیده است که حاوی داده‌های مربوط به نقاط کلیدی بدن می‌باشند. به عنوان نمونه، پایگاه داده NTU-RGB+D یک مجموعه داده مقیاس بزرگ برای تشخیص اعمال انسان می‌باشد که دارای چهار حالت مختلف داده می‌باشد: ویدئوی رنگی، ویدئوی نقشه عمیق[۷]، نقاط کلیدی یا داده اسکلتی سه بعدی بدن و ویدئوهای مادون قرمز. داده‌های این پایگاه داده به کمک دوربین کینتیک نسخه ۲ ضبط شده است. در شکل (۳) نقاط کلیدی بدن که با کمک سنسور کینتیک ایجاد می شود، نمایش داده شده است [۱۹].



شکل ۳- تصویری از پیکربندی ۲۵ نقطه کلیدی بدن درپایگاه داده NTU-RGB+D

علاوه براین، مجموعه داده ETRI-Activity3D نیز معرفی گردیده است که بر فعالیت‌های روزانه سالمندان تمرکز دارد [۲۰]. این پایگاه داده حاوی داده‌های واقع‌بینانه که منعکس کننده محیط کاری و موقعیت‌های خدماتی ربات است، می باشد. شکل (۴) نمونه‌ای از فریم‌های این مجموعه داده را نشان می‌دهد.

[2] Human-computer interaction
[3] Human behavior understanding
[4] Medical assistive application
[5] You Look Only Once
[6] Convolutional neural network
[7] Depth map

استناد به این مقاله:
احمدی، امیر محمد، کیانی، کورش، و راستگو، راضیه. (۱۴۰۳). بکارگیری مدل مبتنی بر ترنسفورمر برای تشخیص فعالیت های غیرطبیعی در ویدئو. مدل سازی در مهندسی، ۲۲(۷۶)، ۲۱۳-۲۲۱. doi: 10.22075/jme.2024.32914.2604

شکل ۴- نمونه‌هایی از اقدامات روزانه در پایگاه داده ETRI-Activity همراه با نقشه عمیق مربوطه و اطلاعات اسکلت به دست آمده از سنسورهای کینکت نسخه ۲

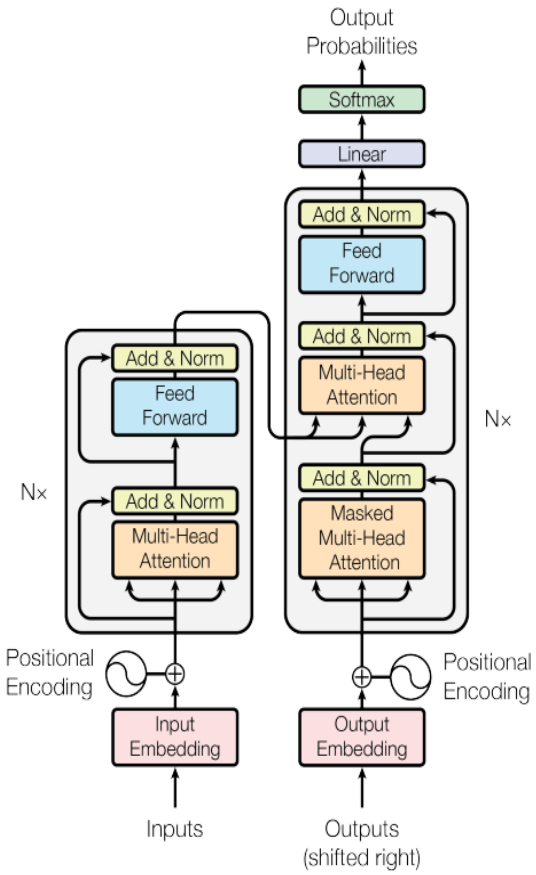
در مرجع [۲۰] یک شبکه جدید با نام شبکه کانولوشنی تطبیقی چهار جریانی[8] یا به اختصار FSA-CNN پیشنهاد شده است که دارای سه ویژگی اصلی استحکام نسبت به تغییرات مکانی-زمانی، تابع فعال‌سازی تطبیقی ورودی و گسترش رویکرد دو جریانی مرسوم می‌باشد. برتری FSA-CNN پیشنهادی با استفاده از مجموعه داده NTU RGB+D و ETRI-Activity3D مورد بحث قرار گرفته است. در مرجع [۲۱] یک مدل جدید از نقاط کلیدی یا اسکلت‌های پویا به نام شبکه‌های کانولوشن گراف فضایی یا با اختصار ST-GCN پیشنهاد گردیده است که با یادگیری خودکار الگوهای مکانی و زمانی از داده‌ها، از محدودیت‌های روش‌های قبلی فراتر می‌رود. این مدل نه تنها به قدرت بیانی بیشتر، بلکه به قابلیت تعمیم قوی‌تر منجر می‌شود. این مدل با آزمایش بر روی دو مجموعه داده مقیاس بزرگ، Kinetics و NTU-RGBD، به پیشرفت‌های قابل توجهی نسبت به روش‌های اصلی دست یافته است. هدف مدل ارائه شده در مرجع [۲۲]، پیدا کردن پنجره زمانی یا محدوده زمانی که در آن عمل غیرطبیعی اتفاق می‌افتد، می‌باشد. یادگیری داده غیرطبیعی در این مدل با مشاهده داده غیرطبیعی و همچنین داده طبیعی انجام شده است. این مدل، مسأله تشخیص فعالیت غیرطبیعی را به صورت یک مسأله رگرسیون در نظر گرفته است و هدف آن اختصاص نمره یا امتیاز بیشتر به بخش غیرطبیعی می‌باشد. در این راستا، از ویژگی‌های استخراج شده از شبکه عصبی کانولوشنی سه بعدی، به دلیل کارایی محاسباتی آن و قابلیت ثبت ظاهر و پویایی حرکت، در تشخیص فعالیت ویدیویی استفاده شده است. مرجع [۵] با تجزیه و تحلیل ویژگی‌های دنباله اسکلتی، یک شبکه دو حرکتی دوگانه[9] یا به اختصار DD-Net برای تشخیص فعالیت انسان مبتنی بر اسکلت بدن ارائه داده است. از ویژگی‌های بارز این شبکه می‌توان به پیچیدگی پایین و سرعت بالای آن اشاره نمود. در مرجع [۲۳] یک شبکه عصبی کانولوشنی سه بعدی با یک شبکه حافظه بلند کوتاه‌مدت به همراه مکانیزم توجه برای تشخیص اعمال انسان ادغام شده است. این مدل برای سناریوهای پیچیده مانند اعمالی که چندین نفر یا اشیاء در آن هستند، طراحی شده است. نتایج آزمایشات این مدل بر روی پایگاه داده‌های UCF101 و HMDB51 بهبود چشمگیر مدل پیشنهادی را نسبت به روش‌های پیشرفته موجود نشان می‌دهد. اگرچه تشخیص عمل مبتنی بر اسکلت در سال‌های اخیر به موفقیت زیادی دست یافته است [۲۴]، بسیاری از مدل‌های موجود از پیچیدگی بالا و سرعت پایین رنج می‌برند [۵]. در این مقاله، با هدف ایجاد بهبود دقت تشخیص فعالیت‌های غیرطبیعی انسان، مدلی مبتنی بر شبکه ترنسفورمر ارائه می‌گردد که جزئیات آن در بخش‌های بعدی قابل‌مشاهده می‌باشد.

۳- مدل ترنسفورمر

بسیاری از مدل‌های ارائه‌شده برای پردازش داده‌های دنباله-ای دارای ساختار کدگذار-کدگشا می‌باشند. جهت بهبود عملکرد، برخی مدل‌ها، کدگذار و کدگشا را از طریق یک مکانیزم توجه به هم متصل می‌کنند. در مرجع [۲۵] یک ساختار شبکه کارآمد به نام ترنسفورمر ارائه شده است که بر اساس مکانیزم‌های توجه ساخته شده است و کاملاً از بازگشت و کانولوشن خودداری می‌کند (شکل ۵). آزمایش‌ها در کاربردهای ترجمه ماشینی نشان می‌دهند که این مدل‌ها از نظر کیفیت برتری داشته و همچنین قابلیت موازی سازی بیشتری دارند. معماری مدل ترنسفورمر شامل ساختار کدگذار-کدگشا می‌باشد. بخش کدگذار یک دنباله ورودی را به یک دنباله از کدهای خروجی نگاشت می‌دهد. با توجه به کدهای تولید شده، کدگشا یک دنباله خروجی را در هر زمان تولید می‌کند. با تکیه بر مکانیسم خودتوجهی و پردازش موازی، مدل ترنسفورمر دارای کارآیی بالاتری نسبت به مدل‌های موجود می‌باشد.

[8] Four-stream Adaptive CNN

[9] Double-feature Double-motion Network



شکل ۵- معماری مدل ترنسفورمر [۲۴].

۴- مدل پیشنهادی مبتنی بر ترنسفورمر

بلاک دیاگرام مدل پیشنهادی در شکل (۶) قابل مشاهده می‌باشد. در ادامه این بخش، جزئیات مدل پیشنهادی ارائه می‌گردد. در این راستا، داده ورودی به شبکه، نقاط کلیدی اسکلتون بدن می‌باشد که هر اسکلتون بدن دارای ۱۵ مختصه در فضای دو بعدی می‌باشد. به منظور آماده‌سازی ویدیوهای پایگاه داده جهت ورود به مدل، نیازمند یکسان سازی تعداد فریم‌های تمام ویدیوها می‌باشیم. برای این منظور، پس از بررسی تعداد فریم‌های تمام ویدیوها و بر اساس تعداد میانگین فریم‌ها، تعداد فریم‌های هر دنباله اسکلت ورودی به تعداد ۳۲ در نظر گرفته می‌شود. بنابراین هر ویدئو دارای ابعاد (۳۲،۱۵،۲) می‌باشد که همان ماتریس فاصله‌های مجموعه مفاصل[۱۰] یا به اختصار JCD می‌باشد که بعنوان ورودی شبکه ترنسفورمر در نظر گرفته می‌شود. درواقع JCD فاصله اقلیدسی بین جفت مفصل‌های مختلف را در هر فریم محاسبه می‌کند. بنابراین ماتریس نسبت به قطر اصلی، متقارن و مقادیر قطر اصلی آن صفر می‌باشد. با توجه به تعداد اسکلتون‌ها که ۱۵ عدد می‌باشد، به ازای هر فریم $\binom{15}{2}$ یعنی ۱۰۵ ویژگی داریم. بنابراین خروجی بلوک JCD به ازای هر دنباله اسکلتون به صورت (۳۲،۱۰۵) خواهد بود. به دلیل اینکه مدل ترنسفورمر ساختار کانولوشنی یا بازگشتی ندارد، برای اینکه مدل از ترتیب دنباله استفاده کند، باید اطلاعاتی در مورد موقعیت نسبی در دنباله ویدئو به مدل تزریق شود. برای این منظور، "کدگذاری‌های موقعیتی" موجود در مدل ترنسفورمر بعنوان تعبیه‌های ورودی در بخش‌های کدگذار و کدگشا اضافه می‌گردد. پس از پردازش داده‌ها در دو بخش کدگذار و کدگشا، مدل ترنسفورمر جهت طبقه‌بندی داده‌های ورودی استفاده می‌گردد. هدف مدل، تشخیص فعالیت‌های غیرطبیعی می‌باشد که با تکیه بر کارآمدی مدل ترنسفورمر برای پردازش دقیق و موازی اطلاعات، این هدف قابل دستیابی می‌باشد. در بخش بعدی، نتایج مدل ارائه می گردد.

۵- نتایج تجربی

در این بخش، نتایج تجربی مدل پیشنهادی ارائه می‌گردد. برای این منظور، جزئیات پیاده‌سازی، پایگاه داده و نتایج مقایسه با مدل‌های موجود ارائه می‌گردد.

۵-۱- جزئیات پیاده سازی

به منظور پیاده‌ازی مدل پیشنهادی، محیط گوگل کولب به عنوان یک سرویس ابری ابری مورد استفاده قرار می‌گیرد. برای پیاده سازی مدل یادگیری عمیق ارائه شده مبتنی بر ترنسفورمر، کتابخانه تنسورفلو و کراس مورد استفاده قرار گرفته است. مدل ترنسفورمر پیشنهادی با اندازه دسته[۱۱]های مختلف (۸، ۱۶ و ۳۲) مورد آزمایش قرار گرفته است. بهینه‌ساز[۱۲] Adam با نرخ یادگیری ۰.۰۰۰۰۱ بکارگرفته شده است.

۶- پایگاه داده

پایگاه داده JHMD شامل اسکلت‌های دو بعدی از ویدیوهای رنگی سه کاناله می‌باشد که می‌تواند در موارد کلی‌تر که استنباط اطلاعات عمق ممکن است سخت یا غیرممکن باشد، استفاده گردد. این پایگاه داده می‌تواند به عنوان یک چالش در زمینه های تخمین پوز، تخمین جریان و تشخیص فعالیت انسان عمل کند [۲۶]. جدول ۱ خلاصه ای از مشخصه‌های پایگاه داده JHMDB را نشان

[10] Joint Collection Distances
[11] Batch size
[12] Optimizer

می‌دهد.

جدول ۱- مشخصات مجموعه داده JHMDB

مشخصه	مقدار
تعداد کل نمونه‌ها	۶۰۹
تعداد کلاس‌ها	۲۱
حالت داده	ویدئو مختصات ۲بعدی اسکلتون

۷- تجزیه و تحلیل نتایج

نتایج به دست آمده از مدل پیشنهادی بر روی پایگاه داده JHMDB با مقادیر مختلف اندازه دسته در جدول ۲ قابل مشاهده می‌باشد. مدل پیشنهادی در بهترین حالت به دقت تشخیص ۸۱.۷ با اندازه دسته ۱۶ رسیده است. علاوه بر این، نتایج مدل پیشنهادی با مدل پایه DD-Net در جدول ۳ قابل مشاهده می‌باشد که حاکی از بهبود ۵ درصدی نسبت به این مدل بر روی پایگاه داده JHMDB می‌باشد. همچنین، مدل پیشنهادی دارای پارامترهای کمتری نسبت به مدل پایه می‌باشد.

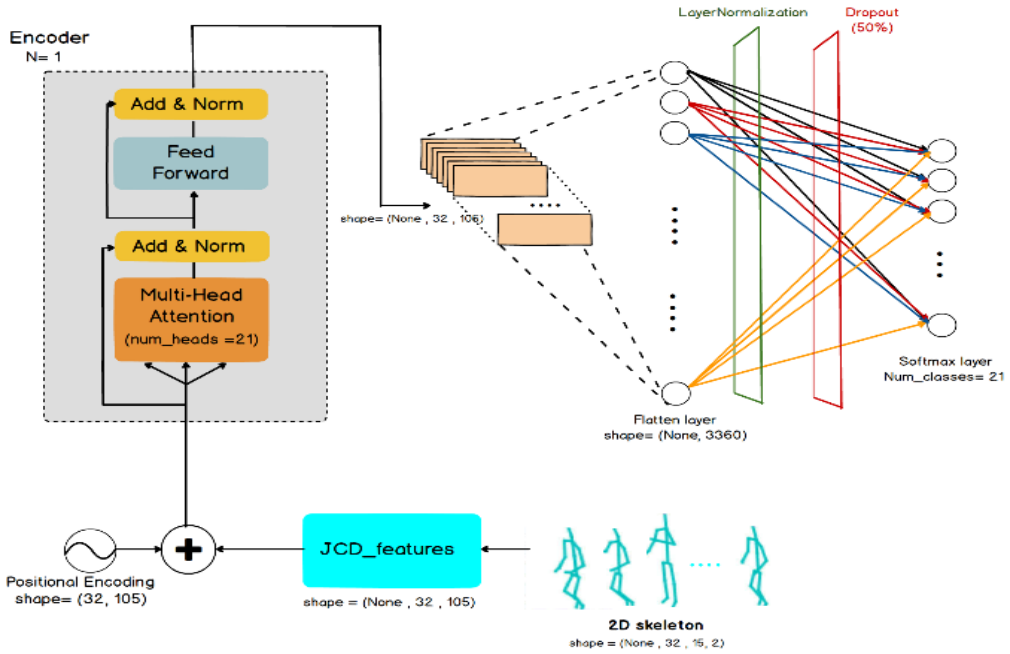
جدول ۲- نتایج مدل پیشنهادی با اندازه دسته‌های مختلف

مدل	اندازه دسته	دقت آزمون (درصد)
مدل پیشنهادی	**۱۶**	**۸۱.۷**
مدل پیشنهادی	۳۲	۸۱.۱۴
مدل پیشنهادی	۸	۸۱.۱۴

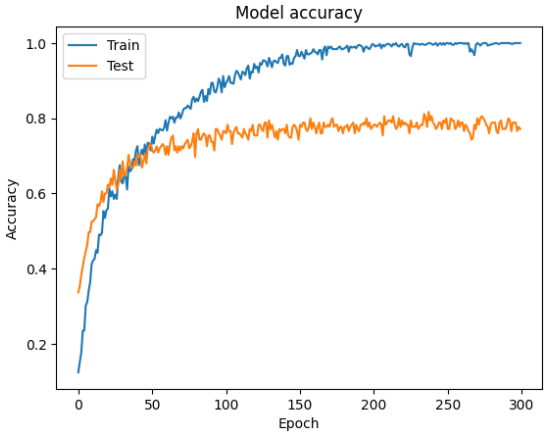
جدول ۳- مقایسه نتایج مدل پیشنهادی با مدل پایه

مدل	اندازه دسته	پارامتر	دقت آزمون (درصد)
مدل پیشنهادی	۱۶	۱.۸۲ مگابایت	۷۷.۲۰
مدل پیشنهادی	۱۶	**۱.۰۷ میلیون**	**۸۱.۷**

علاوه بر این، نمودار روند آموزش شبکه در شکل (۷) نشان داده شده است. همانگونه که این شکل نشان می‌دهد، مدل پیشنهادی پس از طی مراحل آموزش، به وضعیت پایداری می رسد.



شکل ۶- مدل پیشنهادی



شکل ۷- نمودار دقت-خطا برای مدل پیشنهادی

۸- نتیجه‌گیری و کارهای آینده

در این مقاله، مدلی مبتنی بر شبکه ترنسفورمر جهت تشخیص فعالیت‌های غیرطبیعی انسان در ویدیو ارائه گردیده است. برای این منظور، داده‌های اسکلت بدن به عنوان ورودی مدل در نظر گرفته می‌شوند. علاوه بر این، به منظور یکسان‌سازی تعداد فریم‌های ویدیو، پیش‌پردازشی بر روی داده‌ها انجام می‌گیرد تا بر اساس اطلاعات مربوط به حداقل، حداکثر و پراکندگی تعداد فریم‌ها، تعداد بهینه فریم‌ها تعیین گردد. پس از آن، مدل ترنسفورمر جهت غنی-سازی داده‌های اسکلتون مورد استفاده قرار می‌گیرد. تشخیص فعالیت‌های غیرطبیعی نیز با استفاده از طبقه‌بند نهایی که متشکل از یک لایه تمام‌متصل می‌باشد، انجام می‌گردد. نتایج تجربی بر روی پایگاه داده JHMDB حاکی از بهبود نتایج نسبت به مدل‌های موجود می‌باشد. در کارهای آتی، ترکیب مدل ترنسفورمر و شبکه‌های گرافی به منظور غنی‌سازی بیشتر داده‌ها و نیز استفاده از اطلاعات فضایی داده‌ها مورد بررسی قرار خواهد گرفت.

مراجع

[1] C. Dhiman, and D.K. Vishwakarma. "A review of state-of-the-art techniques for abnormal human activity." *Engineering Applications of Artificial Intelligence* 77, (2019): 21-45.

[2] R. Rastgoo, K. Kiani, and S. Escalera. "ZS-SLR: Zero-Shot Sign Language Recognition from RGB-D Videos." *arXiv:2108.10059*, (2021).

[3] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou. "Multi-modal zero-shot sign language recognition". *arXiv:2109.00796*, (2021).

[4] R. Rastgoo, K. Kiani, and S. Escalera. "Word separation in continuous sign language using isolated signs and post-processing." *arXiv:2204.00923*, 2022.

[5] M.A. Gul, M.H. Yousaf, S. Nawaz, Z.U. Rehman, and H.W. Kim. "Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture." *Electronics* 12, no. 9 (2020): 1993.

[6] "JHMDB: Joint-annotated Human Motion Data Base". https://ps.is.mpg.de/code/jhmdb-joint-annotated-human-motion-data-base. Access Date: Feb. 2024.

[7] M. Jain, H. Jégou, and P. Bouthemy. "Improved Motion Description for Action Classification." *Frontiers in ICT* 2, no. 28 (2015).

[8] R. Rastgoo, and V. Sattari Naeini. "A neuro-fuzzy QoS-aware routing protocol for smart grids." *22nd Iranian Conference on Electrical Engineering (ICEE)*, pp. 1080-1084, 2014.

[9] R. Rastgoo, and V. Sattari Naeini. "Tuning parameters of the QoS-aware routing protocol for smart grids using genetic algorithm." *Applied Artificial Intelligence* 30, no. 1 (2016): 52-76.

[10] N. Majidi, K. Kiani, and R. Rastgoo. "A deep model for super-resolution enhancement from a single image." *Journal of AI and Data Mining* 8, no. 4, (2020): 451-460.

[11] K. Kiani, R. Hematpour, and R. Rastgoo. "Automatic grayscale image colorization using a deep hybrid model." *Journal of AI and Data Mining* 9, no. 3 (2021): 321-328.

[12] R. Rastgoo, and V. Sattari-Naeini. "Gsomcr: Multi-constraint genetic-optimized qos-aware routing protocol for smart grids." *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 42, (2018): 185-194.

[13] R. Rastgoo, and K. Kiani. "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning." *Journal of Modeling in Engineering* 17, no. 58 (2019): 103-111.

[14] S. Zarbafi, K. Kiani, and R. Rastgoo. "Spoken Persian digits recognition using deep learning." *Journal of Modeling in Engineering* 21, (2023): 163-172.

[15] F. Alinezhad, K. Kiani, and R. Rastgoo. "A Deep Learning-based Model for Gender Recognition in Mobile Devices." *Journal of AI and Data Mining* 11, (2023): 229-236.

[16] F. Bagherzadeh, and R. Rastgoo. "Deepfake image detection using a deep hybrid convolutional neural network." *Journal of Modeling in Engineering* 75, no. 21 (2023): 19-28.

[17] F. Yang, Y. Wu, S. Sakti, and S. Nakamura. "Make Skeleton-based Action Recognition Model Smaller, Faster and Better." *Proceedings of the ACM Multimedia Asia*, 2019.

[18] M.G. Morshed, T. Sultana, A. Alam and Y.K. Lee. "Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities." *Sensors* 23, no. 4 (2023): 2182.

[19] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.Y. Duan, and A.C. Kot. "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, no. 10 (2020): 2684-2701.

[20] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim. "ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly." *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[21] S. Yan, Y. Xiong, and D. Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition." *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] W. Sultani, C. Chen and M. Shah. " Real-World Anomaly Detection in Surveillance Videos." *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[23] E.M. Saoudi, J. Jaafari, and S.J. Andaloussi. "Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN." *Scientific African* 21, (2023).

[24] R. Rastgoo, K. Kiani, and S. Escalera. "Hand sign language recognition using multi-view hand skeleton." *Expert Systems with Applications* 150, (2020): 113336.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 2023.

[26] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M.J. Black. "Towards Understanding Action Recognition." *IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013.