دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال سیزدهم، شماره‌های۴۹و۵۰، پاییز و زمستان ۱۴۰۰
صص: ۱۰۳_۱۱۴

یک معماری دومسیره کارآمد مبتنی بر شبکه عصبی عمیق برای بازشناسی دروازه در ویدئوی بازی فوتبال

امیرحسین زنگنه* مهدی چم پور ** کامران لایقی***

*دانشجوی دکتری، گروه مهندسی کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

**استادیار، دانشکده مهندسی کامپیوتر ، دانشگاه صنعتی قوچان، قوچان، ایران.

*** استادیار، گروه مهندسی کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

تاریخ دریافت: ۱۳۹۹/۱۱/۰۶ تاریخ پذیرش: ۱۴۰۰/۰۶/۰۳

نوع مقاله: پژوهشی

چکیده

در این مقاله یک روش خودکار با استفاده از یک مدل معماری دومسیره یادگیری عمیق برای مساله تحلیل تصاویر ویدئویی ورزش فوتبال، با تاکید بر شناسایی دروازه به عنوان یکی از مهمترین عناصر رویداد گُل که مهمترین رویداد بازی فوتبال می‌باشد، ارائه کرده‌ایم. معماری پیشنهادی، شکل توسعه یافته مدل **VGG** سیزده لایه می‌باشد که طی آن یک مدل معماری دو مسیره تعریف شده است. در مدل معماری پیشنهادی برای بازشناسی دروازه در مسیر اول، مدل با مجموعه داده آموزشی، آموزش داده می‌شود. اما در مسیر دوم، مجموعه داده‌های آموزشی ابتدا توسط یک سیستم غربال‌کننده مورد بررسی قرار گرفته و بهترین تصاویر که شامل ویژگی‌های متفاوتی با ویژگی-های انتخاب شده توسط مسیر اول هستند، انتخاب می‌شوند. به عبارت دیگر در مسیر دوم، ویژگی‌هایی از شبکه‌ای مشابه مسیر اول، ولی پس از عبور از سیستم غربالگر تولید می‌شود.سپس بردارهای ویژگی تولید شده در دو مسیر با یکدیگر ادغام شده و یک بردار ویژگی سراسری حاصل می‌شود و بدین ترتیب فضاهای متفاوتی از مساله بازشناسی دروازه تحت پوشش قرار گرفته است. ارزیابی‌های متنوعی بر روی روش ارائه شده انجام شده است. نتایج ارزیابی‌ها، حاکی از بهبود دقت بازشناسی دروازه به‌وسیله مدل معماری دومسیره ارائه شده نسبت به مدل پایه می‌باشد. همچنین مقایسه روش پیشنهادی با نتایج موجود نشان می‌دهد دقت روش پیشنهادی، بهتر از نتایج منتشر شده است.

واژگان کلیدی: معماری یادگیری عمیق دو مسیره، ترکیب ویژگی‌ها، شبکه عصبی عمیق VGG، ویژگی‌های کلاسیک، معماری مشترک

۱. مقدمه

با توجه به گسترش روز افزون تجهیزات دیجیتالی ضبط و ذخیره-سازی ویدیو، امروزه کاربران بسیاری در سراسر دنیا به ناچار با حجم بسیار گسترده‌ای از داده‌های ویدیویی در ارتباط می‌باشند. برخی از این ویدیوها مربوط به حوزه سرگرمی و پرکردن اوقات فراغت کاربران و برخی دیگر نیز مرتبط با حوزه‌های نظارتی و امنیتی می‌باشند. چنانکه استفاده از سیستم‌های نظارت ویدیویی در بیشتر سازمان‌ها، ادارات، کارخانجات و محیط‌های کاری موجب مراقبت و کنترل دقیق محیط، کاهش تخلفات، افزایش توانایی در آشکارسازی سریع حوادث و نظم‌دهی محیط کاری شده است. افزایش سامانه‌های نظارتی

نویسنده مسئول: مهدی چم پور jampour@qiet.ac.ir

موجب شده تا حجم عظیمی از داده‌های چند رسانه‌ای تولید شود که در گام اول ذخیره‌سازی این حجم گسترده از داده‌های ویدیویی نیازمند استفاده از رسانه‌های ذخیره‌ساز فراوان می‌باشد و در گام دوم مدیریت محتوایی آن‌ها بصورت دستی، نیازمند تعداد بسیار زیادی کاربر انسانی و صرف مدت زمان طولانی است که در عمل امکان‌پذیر نبوده و با خطای زیادی روبرو می‌باشد.

بخش دیگری از ویدیوهای در دسترس کاربران ویدیوهای مختلف، مربوط به حوزه سرگرمی می‌باشد. کاربران با توجه به گرایش و علاقه مندی‌های شخصی، برنامه‌های ویدیوهای مختلف را تهیه می‌کنند یا سایر ویدیوهای موجود در فضای مجازی را پیگیری و مشاهده و در برخی موارد نیز اقدام به ذخیره‌سازی آن‌ها می‌کنند. در این میان لزوم خلاصه‌سازی فیلم‌های ویدیویی و خصوصا ویدیوهای ورزشی با توجه به اینکه معمولا دارای مدت زمان طولانی می‌باشند، ملموس‌تر است.

ورزش فوتبال یکی از ورزش‌های محبوب در جهان است که نه تنها هواداران، بلکه محققان زیادی از مناطق مختلف را در سراسر جهان به خود جلب کرده است. از جمله ویدیوهای ورزشی محبوب، ویدیوهای ورزش فوتبال است که به دلیل علاقه‌مندی طیف گسترده ای از مردم جهان به این ورزش از اهمیت بسیار بالایی برخوردار می باشد. مسئله زمان طولانی بازی فوتبال موجب شده که علاوه بر حجم زیاد مورد نیاز برای ذخیره‌سازی، در اغلب موارد همه مردم فرصت تماشای ۹۰ دقیقه فوتبال را نداشته باشند و البته علاقه‌مند هستند دستکم لحظات مهم و هیجان انگیز بازی را مشاهده کنند. در نتیجه با توجه به زمانگیر بودن فیلم‌های ویدیوی و ویدیوهای ورزشی و محدود بودن زمان بسیاری از علاقه‌مندان و طرفداران این ویدیوها، جهت انتقال محتوای ویدیو به بینندگان، خلاصه‌سازی ویدیو انجام می‌شود. به این ترتیب بینندگان می‌توانند بدون نیاز به مشاهده کل ویدیو، بخش‌های مهم و اصلی آن رویداد را مشاهده و درک کنند.

خلاصه سازی مطلوب یک ویدئو ورزشی فرآیند ساده ای نیست و نیازمند پردازش انسانی می‌باشد. در روش‌های سنتی یک کاربر سراسر یک ویدیو را مورد نظارت قرار داده و بخش‌های مهم آن را برچسب‌گذاری می‌کند که کاری زمانگیر و دشوار می‌باشد. برای این منظور، نیازمند تشخیص دقیق رویدادهای مهم و حساس بازی فوتبال هستیم که مبهم و هنوز به عنوان یک سوال باز مطرح می‌باشد. با این حال در این مقاله، ما دقیقا درصدد تعریف یک رویداد حساس در بازی فوتبال نیستیم بلکه برای شناسایی و تعیین رویدادها در بازی فوتبال، براساس یک پژوهش میدانی عمل کرده‌ایم. ما به منظور شناسایی مهمترین رویدادهای بازی پرطرفدار فوتبال، پرسشنامه‌ای شامل ۷ رویداد طراحی کردیم و انتخاب مهمترین رویدادها را به شرکت‌کنندگان در نظرسنجی واگذار کردیم. فرم نظرسنجی مذکور شامل رویدادهای گل، کرنر، کارت قرمز، کارت زرد، ضربه آزاد، پنالتی و برخورد توپ با تیرک دروازه می‌باشد.

طی این پژوهش میدانی، پرسشنامه را بین ۲۰۰ نفر در محدوده‌های سنی مختلف پخش کرده و از مخاطبان درخواست کردیم که به ۷

جدول ۱. نتایج نظرسنجی در مورد مهمترین رویداد بازی فوتبال

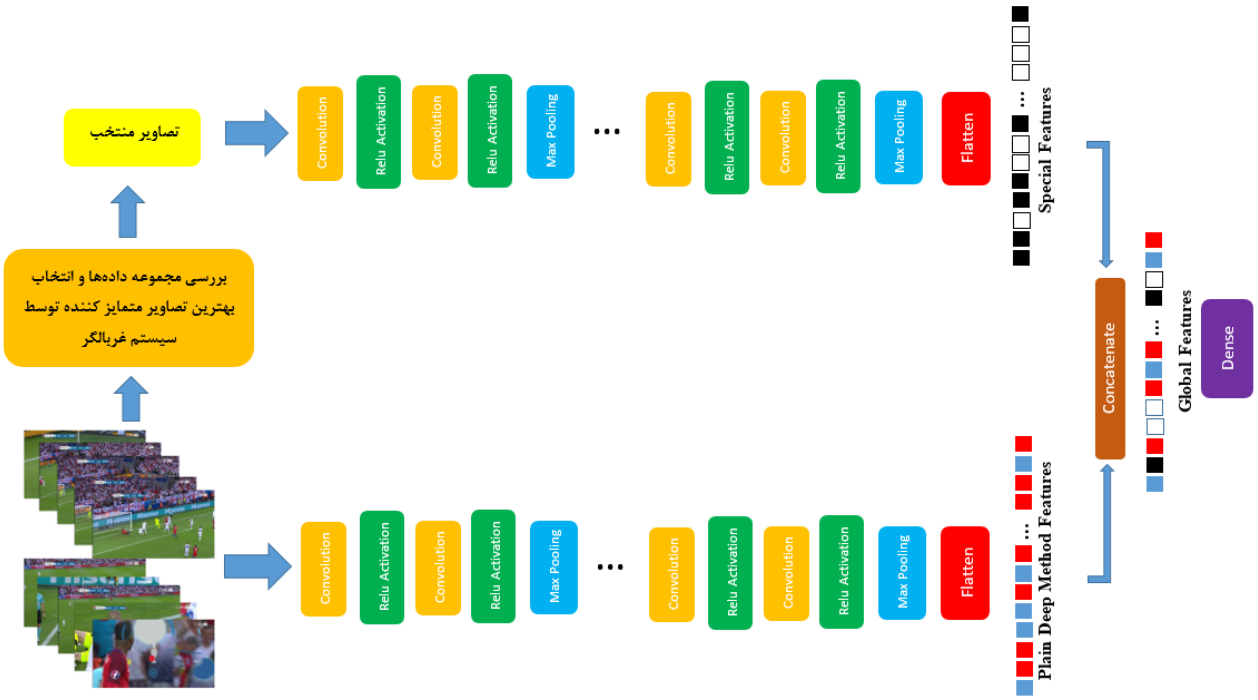
ردیف	نوع رویداد	میانگین امتیاز دریافتی
۱	گل	۶,۰۵
۲	کرنر	۳,۹۴
۳	کارت قرمز	۴,۱۶
۴	کارت زرد	۲,۸۸
۵	ضربه آزاد	۴,۶۶
۶	پنالتی	۵,۶۰
۷	برخورد توپ با تیرک دروازه	۴,۴۹

رویداد مهم مطرح شده در پرسشنامه از عدد یک (کم اهمیت‌ترین) تا عدد ۷ (پر اهمیت‌ترین) یک امتیاز را اختصاص دهند. نتایج نظرسنجی در جدول ۱ ارائه شده است. همانطور که انتظار داشتیم براساس میانگین نظر شرکت‌کنندگان در نظرسنجی مهمترین رویداد، رویداد گل و پس از آن رویداد پنالتی می‌باشد.

براساس نتایج حاصل از نظر سنجی، مواردی که در حین تماشای یک بازی فوتبال توجه مردم را به خود جلب می‌کنند شامل شوت، پنالتی، کارت زرد، کارت قرمز، خطاها، ضربات آزاد، کرنر و گل می‌باشند که آنها را به عنوان رویدادهای مهم وحساس در بازی فوتبال تعریف می‌کنیم. لازم به ذکر است که براساس نتایج حاصله، رویداد «گُل» مهمترین و حساس‌ترین رویداد در بازی فوتبال به شمار می‌آید.

از سوی دیگر، کشف و شناسایی رویدادها و وقایع پیچیده در ویدیوها عملیاتی چالش برانگیز و پیچیده است که توجه محققان زیادی را در جامعه بینایی رایانه به خود جلب کرده است. در مقایسه با تشخیص مفهوم‌های مجزا[1]، که عمدتاً بر شناسایی اشیاء خاص و صحنه در تصاویر ثابت یا کلیپ‌های کوتاه ویدیویی شامل حرکات ساده متمرکز است، تشخیص رویداد چندرسانه‌ای با فیلم‌های پیچیده‌تری مرتبط می‌باشد که شامل تعامل انسان با اشیاء مختلف در صحنه‌های متفاوت است و پردازش آن‌ها معمولا چند دقیقه یا حتی چند ساعت زمان نیاز دارد.بنابراین، یک رویداد، یک انتزاع معنایی از توالی سطح بالاتر نسبت به یک یا چند مفهوم است. به عنوان مثال، رویداد گُل را می‌توان ترکیب چندین مفهوم مانند اشیاء (بازیکنان، توپ، دروازه و تور)، صحنه (زمین چمن که مسابقه روی آن انجام می‌شود)، اقدامات

[1] Atomic concept



شکل ۱. معماری دومسیره پیشنهادی جهت ترکیب ویژگی‌ها استخراج شده توسط شبکه عصبی آموزش داده شده با مجموعه دادهای آموزشی و شبکه عصبی آموزش داده شده با دادهای منتخب مبتنی بر سیستم غربالگر.

(حرکت بازیکنان، موقعیت توپ، شادی کردن بازیکنان) و مفاهیم صوتی (صدای گزارشگر و تماشاچیان، صدای بازیکنان، تشویق تماشاچیان) و غیره تعریف نمود.

با توجه به مشکلات موجود لزوم خلاصه‌سازی خودکار ویدیوها کاملا محسوس است. در خلاصه‌سازی خودکار ویدیو با معرفی رویدادهای مهم و حساس، سیستم قادر است پس از دریافت یک ویدیو در ورودی با حذف افزونگی‌های بصری و فریم‌های تکراری، یک کلیپ خلاصه شده از ویدیو که در برگیرنده محتوای ویدیوی اصلی می‌باشد را در اختیار کاربر قرار دهد.

در این مقاله ما یک روش خودکار برای تحلیل تصاویر ویدئویی ورزش فوتبال ارائه می‌کنیم که با تاکید بر شناسایی یکی از عوامل قابل توجه در تشخیص رویدادهای مهم و حساس بازی فوتبال، یک روش خودکار برای تشخیص و خلاصه‌سازی لحظات مهم بازی تلقی می‌شود. همانطور که در شکل ۱ نشان داده شده در این مقاله، با ترکیب ویژگی‌های استخراج شده توسط یک مدل معماری دومسیره یادگیری عمیق، روشی خودکار معرفی می‌کنیم که در آن دروازه به عنوان یک عامل تفکیک کننده شناسایی می‌شود. اگر چه این عامل به تنهایی نشان دهنده لحظات همیشه حساس نیست ولی گامی موثر بسوی تفکیک لحظات مهم بازی فوتبال و خلاصه‌سازی آن به شمار می آید.

ادامه این مقاله به شرح زیر سازماندهی شده است: در بخش ۲، کارهای انجام شده در زمینه خلاصه‌سازی ویدیویی فوتبال مورد بررسی قرار می‌گیرد، سپس در بخش ۳، روش پیشنهادی به تفصیل شرح داده می‌شود؛ در بخش ۴ نتایج تجربی ارائه شده است و در نهایت، نتیجه‌گیری در بخش ۵ ذکر شده است.

۲. کارهای مرتبط

تشخیص خودکار رخدادها و تفسیر معنایی صحنه‌ها، یک کار چالش برانگیز در خلاصه‌سازی ویدیو بازی فوتبال است. این کار می‌تواند با استخراج ویژگی‌ها در سطوح معنایی مختلف انجام شود. ویژگی‌های سطح پایین

تصویر مانند رنگ، شکل و بافت، توپ، دهانه دروازه، و همچنین ویژگی‌های ویدئویی مانند فریم‌های متوالی و عکس‌ها برای به دست آوردن ویژگی‌های سطح بالا از قبیل شناسایی وضعیت ویدیو مانند حالت پخش مجدد بازی و حالت وقفه ایجاد شده در بازی استفاده می‌شوند.

کارهایی برای خلاصه سازی ویدیو انجام شده که ما آن‌ها را به دو دسته کلی تقسیم کرده‌ایم: ۱) روش‌هایی که برای خلاصه‌سازی از ویژگی‌های مختلف ویدیو مانند ویژگی‌های دیداری، شنیداری یا متن‌های مرتبط با ویدیو استفاده می‌کنند. ۲) روش‌هایی که منحصرا ویژگی‌های دیداری موجود در فریم‌های ویدیو را برای شناسایی رویداد، مورد استفاده قرار می‌دهند.

۲_۱ رو ش‌هایی مبتنی بر ویژگی‌های ویدیو

ویژگی‌های صوتی شامل تشویق تماشاگران و هیجان مفسران ورزشی استخراج شده، و همزمان نشانه‌های(ویژگی‌های) بصری تشخیص داده شده‌اند. بعد از استخراج مفهوم معنایی و توجه به توالی معنایی

رویدادهایی که با هم مرتبط هستند، مانند ورود توپ به دروازه و هلهله تماشاچیان، قوانین موجود برای شناسایی رویداد به کار گرفته می‌شوند[۱]. در کاری مشابه برای تجزیه و تحلیل محتوی ویدیو اقدام به استخراج ویژگی‌های سطح پایین و سطح میانی از کانال‌های صدا / تصویری کردند[۲].

روشی برای آنالیز معنایی ویدیو و خلاصه‌سازی ویدیو با شناسایی مفاهیم با استفاده از یک شبکه بیزی معرفی شده است که در آن، رویدادهای برجسته بازی با استفاده از ویژگی‌های صوتی با استفاده از قوانین تولید شده و دانش این حوزه از کلیپ‌های ویدیو، شناسایی می‌شوند[۳]. مجموعه‌ای از کلیپ‌های برجسته که شامل رویدادهای حساس بازی هستند، برچسب‌گذاری شده و در یک چکیده ویدئویی برای کاربردهای مختلف مانند مرور رویدادهای مهم، شاخص‌گذاری و بازیابی ویدیو بکار برده می‌شوند.

استخراج ویژگی‌های صوتی (صدای سوت داور) و تصویری ویدیو برای شناسایی وقفه‌های ایجاد شده در بازی، مورد استفاده قرار گرفته‌اند. [۴]. برای مثال در بازی فوتبال زمانی که سوت داور شنیده می‌شود به این معنی است که یک خطا اتفاق افتاده یا توپ خارج از میدان بوده و در نتیجه یک وقفه در بازی رخ داده است. از جمله مزایای کار یاد شده، عمومی بودن و کاربردی بودن آن برای همه بازی‌هایی است که دارای ساختار بازی/ وقفه می‌باشند.

همچنین [۹]–[۵] نیز از ویژگی صوتی به عنوان یکی از مهمترین ویژگی‌ها برای شناسایی رویدادهای حساس استفاده کردند. به عنوان مثال، برای شناسایی رویداد گل در بازی فوتبال از تغییرات صدای گزارشگر و تغییرات صدای تماشاچیان استفاده کردند. به عقیده آن‌ها افزایش شدید انرژی صوتی نشان دهنده رویدادی خاص در بازی می‌باشد [۱۱] ,[۱۰].

کارهای انجام شده دسته اول که از ویژگی‌های مختلف برای شناسایی رویداد استفاده می‌کنند با محدیدویت‌های از جمله: ۱- افزایش تعداد سنسورها و تجهیزات سخت افزاری به منظور ضبط صوت، ۲- محدودیت در فاصله ضبط داده‌ها، با استفاده از دوربین می‌توان رویدادها را از فاصله دور ثبت و ضبط نمود در حالیکه اگر بخواهیم همان ویدیو را با صدا تهیه کنیم با محدودیت فاصله روبرو خواهیم بود. ۳- حذف نویز و صداهای اضافی موجود در ویدیو که توسط تماشاچیان تولید می‌شود و می‌تواند موجب خطا در عملکرد سیستم شود. به عنوان مثال روش‌هایی که با شناسایی صدای سوت داور اقدام به شناسایی رویداد می‌کنند در مواردی که تماشاچیان اقدام به سوت زدن در حین بازی می‌کنند با خطا روبرو می‌شوند. ۴- در ویدیوهایی که در ورزشگاه‌های سرپوشیده تهیه می‌شوند، صدای تماشاچیان صدای غالب بوده و عملا صدای سوت داور و بازیکنان توسط روش‌های مالتی‌مُدال قابل استفاده نیستند. ۵- روش‌های مالتی‌مدال فقط روی ویدیوهایی که تحت شرایط خاصی تهیه شده‌اند، قابلیت استفاده را دارا بوده و عمومی نیستند.

۲_۲ روش‌های مبتنی بر ویژگی‌های فریم

شرکت‌های پخش ویدیویی از تکرارهای[۲] صحنه‌های هیجان انگیز و مهم استفاده می‌کنند تا روی رویدادهای خاص بازی با جزئیات کامل تأکید کرده و آنها را برای بینندگان خود نمایش دهند. صحنه تکرار به طور عمده شامل نمایش حرکت آهسته یک رویداد جالب و گاهی اوقات لوگو بازی (علامت ویژه مسابقه یا علامت تجاری اسپانسر برای برخی از فریم‌ها) است، که در آغاز و پایان صحنه تکرار استفاده می‌شود. استفاده از ویژگی تکرار رویدادهای حساس بازی نیز در برخی از کارهای مشابه برای خلاصه‌سازی ویدیو مورد استفاده قرار می‌گیرد.

برای شناسایی رویدادهای حساس بازی اقدام به شناسایی لوگو بازی کرده‌اند[۱۲]. به نظر آن‌ها هنگامی رویداد گل شناسایی می‌شود که یک وقفه در مسابقه تشخیص داده می‌شود یا برخی علایم از تشویق بازیکنان مشاهده می‌شوند و یا پخش مجدد بازی از زوایای مختلف که توسط دوربین های مختلف بدست آمده‌اند، نمایش داده می‌شوند. هنگامی که لوگوی مسابقات در ویدیو پخش می‌شود اقدام به تشخیص صحنه تکرار می‌کنند و سپس برای خلاصه سازی ویدیو با استفاده از شناسایی صحنه‌ی تکرار، شناسایی مبتنی بر قاعده گل و تشخیص حمله، اقدام می‌کنند. این تشخیص از طریق تشخیص مرز براساس دهانه‌ی دروازه، طبقه‌بندی عکس، تشخیص صحنه‌ی تکرار، و تشخیص بورد ثبت امتیازات امکان‌پذیر است.

برای شناسایی رویدادهای حساس بازی فوتبال اقدام به تشخیص صحنه‌های پخش مجدد در ویدیو کردند[۱۳]. به نظر آن‌ها صحنه‌های پخش مجدد حاوی رویدادهای مهم بازی می‌باشند. برای شناسایی صحنه‌های پخش مجدد نیز اقدام به شناسایی لوگوی بازی در فریم‌های ویدیو کرده‌اند. آن‌ها به محض تشخیص لوگوی مسابقات در یک فریم، به فریم‌های قبلی برگشته و این کار را تا رسیدن به فریمی که حاوی یک تصویر از نمای دور[۳]، است ادامه می‌دهند. مجموعه فریم‌های بین تصویر نمای دور و لوگوی مسابقات به عنوان رویداد مهم بازی خلاصه می‌شوند.

برای شناسایی رویدادهای حساس بازی اقدام به تشخیص صحنه‌های پخش مجدد ویدیو کرده‌اند[۱۴]. به عقیده آن‌ها لوگوی مسابقات قبل

[۲] Replay

[۳] Long View Shot

شکل ۲. معماری اولیه شبکه VGG و لایه‌های استفاده شده در آن

و بعد صحنه‌های پخش مجدد ویدیو قرار دارند. آن‌ها برای تشخیص وجود یا عدم وجود لوگوی مسابقات در فریم‌های ویدیو از شبکه عصبی کانولوشن استفاده کرده و براساس رویدادهای حساس شناسایی شده اقدام به خلاصه سازی ویدیو کرده‌اند.

در روش‌های مبتنی بر شناسایی لوگوی [۱۷]–[۱۵] مسابقات ۱- باید لوگوی مسابقات برای سیستم تعریف شود ۲- سیستم فقط به تصویر لوگو مسابقه حساس بوده و هیچ دانشی در مورد نوع رویداد اتفاق افتاده نداشته و در نتیجه امکان خلاصه سازی ویدیو براساس نوع رویداد در این روش وجود ندارد و ۳- این روش عمومی نبوده و فقط برای ویدیوهایی طراحی شده که توسط یک کاربر انسانی از قبل مورد بررسی قرار گرفته باشد که سلیقه و انتخاب کاربر شرکت‌های پخش ویدیویی در آن دخیل است.

یادگیری عمیق به عنوان یکی از تکنیک‌های یادگیری ماشین، از پیشرفت‌های فناوری واحدهای پردازش گرافیکی[۴] استفاده کرده است، و این امر به نوبه خود استفاده گسترده از آن را فراهم آورده است. کریمی و همکاران [۱۸] از یادگیری عمیق برای شناسایی رویدادهای ورزش فوتبال با تاکید بر استخراج رویداد کارت زرد و قرمز استفاده کرده‌اند. آن‌ها ابتدا تصاویر ورزش فوتبال را از سایر تصاویر تفکیک کرده و در مرحله بعد اقدام به شناسایی رویداد می‌کنند.

تکنیک‌های یادگیری عمیق به نتایج بسیار خوبی در بسیاری از مسائل مهم در مقایسه با روش‌های سنتی دست یافته‌اند. شبکه‌های عصبی پیچشی[۵] یکی از مدل‌های یادگیری عمیق با لایه‌های متعدد می‌باشند که شامل سطوح چندگانه هستند. در مقایسه با شبکه‌های کاملا متصل، شبکه‌های عصبی پیچشی دارای قابلیت تعمیم بالاتری هستند. این امر آن‌ها را برای کاربردهای مختلف از جمله تشخیص اشیا، و دسته‌بندی تصاویر مناسب می‌کند [۲۱]–[۱۹]. با توجه به اینکه در این مقاله هدف ما بازشناسی دروازه به عنوان عامل شناسایی لحظات حساس می‌باشد ما از یک شبکه عصبی پیچشی استفاده کرده‌ایم.

۳. روش پیشنهادی

در این بخش با توجه به ضرورت توسعه روش‌های خودکار و کارآمد برای خلاصه‌سازی رویدادهای مهم ویدیو، به معرفی روش پیشنهادی می‌پردازیم. ما از یک مدل پایه یادگیری عمیق برای استخراج ویژگی‌ها استفاده می‌کنیم، اما پیش‌تر نشان داده شده که با استفاده از شبکه‌های عصبی عمیق همچنان ممکن است برخی ویژگی‌های مفید برای دسته‌بندی کشف و استخراج نشوند [۲۲] در نتیجه روش‌های ترکیبی می‌توانند برای این منظور کارآمد باشند. ما در این مقاله، با ارائه یک معماری دومسیره، در یک مسیر به استخراج ویژگی‌های مبتنی بر شبکه عمیق پرداخته و در مسیر دوم به کمک یک سیستم غربال‌گر ابتکاری به استخراج ویژگی‌های مکمل می‌پردازیم که در بخش تجربیات نشان داده شده است ترکیب این دو مسیر، توصیف مطلوبتری از تصاویر به منظور تفکیک‌پذیری ایجاد می‌کند و سبب بهبود کارآیی سیستم در شناسایی هدف می‌شود. در ادامه ما ابتدا معماری پایه VGG که در این مقاله مورد استفاده قرار گرفته شده است و انگیزه استفاده از آن را شرح می‌دهیم، سپس در زیربخش بعدی سیستم غربال‌گر ابتکاری‌مان را معرفی کرده و در زیربخش آخر، مدل ترکیبی پیشنهادی را ارائه و تشریح می‌کنیم.

۳_۱ معماری مدل پایه

ما در این مقاله از مدل پایه شبکه عصبی عمیق VGG-۱۳ برای بازشناسی تصاویر حاوی دروازه استفاده می‌کنیم. مدل VGG یک معماری شبکه عصبی پیچشی است که توسط سایمون و زیسرمن در سال ۲۰۱۴ پیشنهاد شد. این شبکه نشان داد که می‌توان با افزایش عمق شبکه، دقت دسته‌بندی را بهبود بخشید. انگیزه ما، در به کارگیری شبکه VGG عمق شبکه بوده است. چنانکه، عمق مناسب شبکه یادگیری عمیق در عملکرد آن بسیار موثر می باشد. معماری VGG با هدف تعامل بین عمق مطلوب شبکه و از سوی دیگر کاهش تعداد پارامترها در شبکه طراحی شده است چنانکه در همه لایه‌ها از فیلتر پیچشی (کانولوشن) ۳×۳ با طول گام ۱ و همچنین یک حداکثر تجمع[۶] ۲×۲ استفاده شده است. تابع فعال‌سازی که شبکه VGG

[۴] Graphical Processing Units

[۵] Convolutional Neural Networks

[۶] Max pooling

با آن کار می‌کند یکسوساز خطی (ریلو[۷]) می‌باشد و چنانکه در شکل ۲ نشان داده شده است در لایه آخر از تابع سیگموئید[۸] استفاده می‌شود.

روش‌های مبتنی بر یادگیری عمیق، کارآیی بسیار مطلوبی در استخراج ویژگی‌ها دارند، اما تضمین‌کننده استخراج همه ویژگی‌ها، یا به عبارتی بهترین ویژگی‌ها نیستند [۲۳]. در نتیجه با اطمینان می‌توان گفت که استخراج خودکار ویژگی‌های تصویر توسط یک مدل یادگیری عمیق ساده برای کاربردهایی مانند بازشناسی دروازه در تصاویر بازی فوتبال اگر چه مفید است اما کامل نبوده و ترکیب ویژگی‌های مختلف استخراج شده که هرکدام قادر به پوشش بخشی از فضای مساله می‌باشند سبب بهبود دقت در عملکرد سیستم خواهند شد. لذا در ادامه به معرفی مسیر دوم معماری پیشنهادی، به عنوان مکملی برای استخراج ویژگی مطلوب می‌پردازیم.

۳_۲ سیستم غربالگر تصاویر

ما در مسیر دوم به ارائه یک مدل شبکه عصبی عمیق می‌پردازیم که بطور سری در ادامه سیستم غربالگر ابتکاری‌مان قرار دارد چنانکه ویژگی‌هایی را از تصاویر استخراج می‌کند که در بخش نتایج نشان می‌دهیم ترکیب آنها با ویژگی‌های شبکه عصبی عمیق در مسیر اول، سبب بهبود کارآیی سیستم بازشناسی می‌شود. هدف از زیرسیستم غربالگر در مسیر دوم، کمک به شناسایی و جداسازی تصاویر آموزشی برتر می‌باشد که منظور از واژه برتر در اینجا، اشاره به تصاویری است که توسط ماشین از قابلیت تفکیک‌پذیری بالاتری برخوردارند. سیستم غربالگر پیشنهادی از یادگیرنده‌های ضعیف[۹] بر اساس ضریب همبستگی و توابع چگالی احتمال استفاده می‌کند. برای این منظور ابتدا فرآیند شناسایی بهترین تصاویر توسط سیستم غربالگر را به عنوان یک منبع اولیه از ویژگی‌های بسیار مطلوب در شناسایی دروازه شرح می‌دهیم.

ما ابتدا ضریب همبستگی هر تصویر از تصاویر شامل دروازه (مثبت) در مجموعه تصاویر آموزشی را نسبت به سایر تصاویر آموزشی، اعم از حاوی دروازه (مثبت) یا تصاویر غیردروازه (منفی) را محاسبه می‌کنیم. سپس بر اساس ضرایب به دست آمده، توابع چگالی احتمال آنها را نسبت به همه تصاویر آموزشی حاوی دروازه و غیردروازه محاسبه می‌کنیم. تصاویری که بتوانند تمایز بهتری بین تصاویر شامل دروازه و بدون دوازه ایجاد کنند، به عنوان تصاویر منتخب انتخاب می‌شوند. به عبارت دیگر، به صورت یک رابطه ۱ به n، توابع چگالی احتمال هر تصویر بر اساس میزان ضریب همبستگی آن نسبت به سایر تصاویر محاسبه می‌شود. به این ترتیب به هر تصویر از مجموعه تصاویر آموزشی به دید یک یادگیرنده ضعیف اما سراسری نگاه می‌کنیم. بدیهی است، تصویری که بتواند بیشترین همبستگی با سایر تصاویر مثبت و عدم همبستگی با تصاویر منفی را کسب کند از ویژگی‌های مطلوبتری برخوردار است که می‌تواند در نهایت منجر به توصیف بهتر تصاویر دیده نشده گردد.

ما از توابع چگالی احتمال برای نمایش میزان تفکیک‌پذیری هر یک از یادگیرنده‌های ضعیف استفاده کرده‌ایم؛ به عنوان مثال شکل ۳ (سمت راست) مربوط به یک یادگیرنده مطلوب است که بخوبی تصاویر دارای دروازه و غیردروازه را تفکیک کرده است. در عوض، (سمت چپ) توابع چگالی احتمال یک یادگیرنده ضعیف ناموفق را نشان می‌دهد که همپوشانی دو تابع بیانگر عدم توانایی در تفکیک مطلوب تصاویر دارای دروازه و غیردروازه می‌باشد. ما به منظور محاسبه خودکار میزان همپوشانی توابع چگالی احتمال، روابط (۱) تا (۷) را بسط داده‌ایم چنانکه این روابط به ما کمک می‌کنند به مقداری عددی به منظور تصمیم‌گیری در خصوص مقدار همپوشانی توابع چگالی احتمال، و به طور کلی انتخاب یا عدم انتخاب یادگیرنده‌های ضعیف اقدام کنیم. بدیهی است هر چه مقدار همپوشانی که از رابطه (۷) به دست می‌آید کمتر باشد میزان تفکیک‌پذیری توسط یادگیرنده ضعیف بهتر بوده و بنابراین مطلوب انتخاب ما است. بنابراین بر اساس توزیع نرمال دو تابع چگالی احتمال داریم:

$$\mu_{\phi(X_i)} = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \qquad (1)$$

$$\mu_{\phi(Y_j)} = \frac{1}{n}\sum_{j=1}^{m}\phi(Y_j) \qquad (2)$$

که در آن X_i مجموعه داده مثبت، Y_j مجموعه داده منفی، $\phi(X_i)$ تابع چگالی احتمال داده مثبت، $\phi(Y_j)$ تابع چگالی احتمال داده منفی، $\mu_{\phi(X_i)}$ میانگین تابع چگالی احتمال داده‌های مثبت و $\mu_{\phi(Y_j)}$ تابع چگالی احتمال داده‌های منفی می‌باشند. برای محاسبه فاصله ۲ تابع چگالی احتمال داریم:

$$D = |P-Q|^2 \qquad (3)$$

$$= \left|\mu_{\phi(X_i)} - \mu_{\phi(Y_j)}\right|^2 \qquad (4)$$

$$= \left|\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(Y_j)\right|^2 \qquad (5)$$

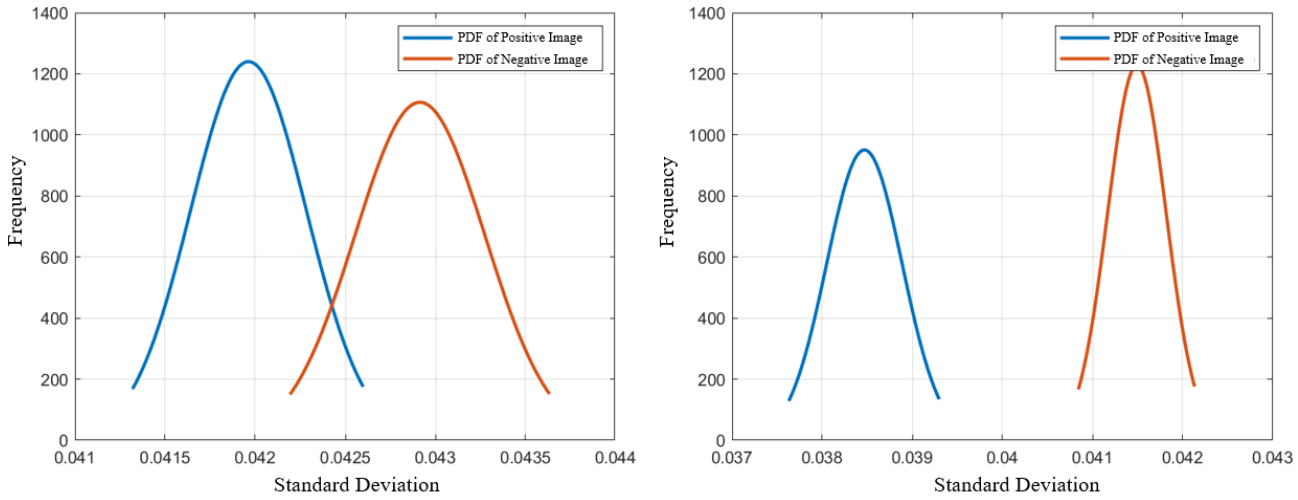
که با بسط آن خواهیم داشت:

$$= \left(\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(Y_j)\right)^T \times \left(\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(Y_j)\right) \qquad (6)$$

[۷] ReLU

[۸] Sigmoid

[۹] Weak lerner



شکل ۳. تابع چگالی احتمال یک تصویر مثبت (سمت راست) و تابع چگالی احتمال یک تصویر منفی (سمت چپ).

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi(X_i)^T \phi(X_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \phi(Y_i)^T \phi(Y_j) - \frac{2}{n\,m} \sum_{i=1}^{n} \sum_{j=1}^{m} \phi(X_i)^T \phi(Y_j) \quad (7)$$

حاصل رابطه (۷) مقداری عددی در بازه صفر تا یک خواهد بود که هر چه مقدار آن کمتر باشد میزان همپوشانی دو تابع چگالی احتمال کمتر خواهد بود که تفکیک پذیری بیشتر مجموعه هدف و غیرهدف را بیان می‌کند. ما از آن به عنوان عاملی برای انتخاب ۲۰۰ تصویر با بیشترین معیار جهت تفکیک تصاویر مثبت (دروازه) و منفی (غیردروازه) به منظور آموزش شبکه در مسیر دوم استفاده می‌کنیم. در زیربخش بعد، معماری دو مسیره پیشنهادی و نهایتا ترکیب ویژگی‌ها برای استخراج ویژگی‌های سراسری را معرفی می‌کنیم.

۳_۳ معماری مدل شبکه مشترک پیشنهادی

در مدل معماری پیشنهادی برای بازشناسی دروازه در مسیر اول، مدل با مجموعه داده آموزشی، آموزش داده می‌شود. اما در مسیر دوم، مجموعه داده‌های آموزشی ابتدا توسط یک سیستم غربال‌کننده که در بخش ۲-۳ معرفی شد مورد بررسی قرار گرفته و بهترین تصاویر که شامل ویژگی‌های متفاوتی با ویژگی‌های انتخاب شده توسط مسیر اول هستند، انتخاب می‌شوند. به عبارتی یک غربال اولیه بر روی تصاویر آموزشی اعمال می‌کنیم و شبکه مسیر دوم با مجموعه غربال شده آموزش داده می‌شود که سبب تاکید بیشتر بر محتوای هدف، تقویت و بهبود تفکیک‌پذیری مدل پیشنهادی می‌گردد. در معماری نسخه پایه شبکه VGG، ورودی شبکه شامل مجموعه‌ای از قطعه‌های تصویر است اما در این مقاله به منظور بازنمایی بهتر ویژگی‌های تصویر، در حقیقت با تعریف مسیر دوم، به تقویت داده‌های سراسری پرداخته‌ایم. به این ترتیب، در معماری دو مسیره پیشنهادی، تصویر ورودی که اندازه آن ۲۰۰ × ۲۰۰ است به دو مسیر وارد می شود و در مسیر اول از لایه‌های پیچشی عبور می‌کند که عمق آن‌ها از ۳۲ در لایه اول تا ۲۵۶ در لایه چهارم افزایش پیدا می‌کند. سپس لایه‌های پیچشی با سه لایه اتصال کامل دنبال می‌شوند. در مسیر دوم، ویژگی‌هایی از شبکه‌ای مشابه ولی پس از عبور از سیستم غربالگر تولید می‌شود.

بعد از لایه Flatten هر مسیر دارای یک بردار ویژگی ۸۰۰۰۰ بُعدی می‌باشد که با هم ادغام شده و یک بردار ۱۶۰۰۰۰ بُعدی حاصل می‌شود. بردار بدست آمده در حقیقت یک بردار ویژگی سراسری می‌باشد که از ترکیب ویژگی‌های دو مسیر بدست می‌آید. که این بردار ویژگی سراسری، ورودی لایه Dense را تشکیل می‌دهد.

در بخش بعدی، با تحلیل مدل پایه شبکه عمیق VGG و روش پیشنهادی، بر روی پایگاه داده تصاویر، با توجه به اینکه در دو مسیر معماری پیشنهادی فضاهای متفاوتی از مساله تحت پوشش قرار گرفته است انتظار داریم که بردار ویژگی مشترک ایجاد شده از کارآیی مطلوبتری در مقایسه با بردار ویژگی مدل پایه برخوردار باشد.

۴. نتایج تجربی و آزمایش‌ها

در این بخش به ارزیابی و تحلیل روش پیشنهادی می‌پردازیم. در ابتدا مشخصات پایگاه داده تصاویر مورد استفاده را معرفی کرده و

جدول ۲. پارامترهای مربوط به پیاده‌سازی شبکه عمیق پیشنهادی

Parameter	Value
Optimizer	Adam
Loss function	Binary cross-entropy
Performance metric	Accuracy
Total Classes	2 (Gate and Non-Gate)
Batch Size	32
Epoch	50

سپس روش پایه و روش پیشنهادی را مورد ارزیابی قرار می‌دهیم.

همچنین در انتهای این بخش مقایسه‌ای با سایرکارهای مشابه انجام پذیرفته است. پلتفرم مورد استفاده در پیاده‌سازی این تحقیق پایتون بوده و پارامترهای مربوط به شبکه عمیق پیشنهادی نیز در جدول شماره ۲ ارائه شده است.

۴_۱ پایگاه‌داده تصاویر

با توجه به تحقیقات انجام شده، در حال حاضر هیچ مجموعه داده اختصاصی برای تحلیل اشیاء موجود در زمین فوتبال به صورت دسترسی رایگان برای امور تحقیقاتی وجود ندارد. با این حال، تعدادی پایگاه داده تصاویر وجود دارد که شامل ویدئو و تصاویر متنوعی از جمله زمین فوتبال می‌باشند که آنها نیز بطور خاص الزاما شامل دروازه نیستند ضمن آنکه تعداد آنها نیز به اندازه‌ای که ما برای آموزش شبکه پیشنهادی استفاده کنیم نیست. به عنوان مثال، تنها مجموعه داده‌ی در دسترس، مربوط به ویدیوهای مربوط به پنج مسابقه فوتبال از لیگ فوتبال اسپانیا (La Liga) است [۱۴] که آن نیز به دلیل محدودیت در تعداد مسابقات و تیم‌های شرکت‌کننده، همچنین عدم تنوع در شرایط مختلف روشنایی، آب و هوا و غیره از جامعیت کافی برخوردار نیست. درنتیجه ما برای رفع محدودیت‌های یاد شده اقدام به تهیه یک مجموعه داده تصاویر از ویدئوهای فوتبال در شرایط بسیار متنوع کردیم که مشخصات آن در ادامه ذکر شده است.

۴_۲ آماده سازی پایگاه داده تصاویر فوتبال

با توجه به خلاء پایگاه داده تصاویر مرتبط با موضوع تحقیقاتی مورد نظر این مقاله، ما یک مجموعه داده به شرح اطلاعات زیر ارایه کردیم. ما ۱۰۰۰ تصویر از تعداد ۲۰ مسابقه فوتبال از سراسر جهان از جمله لیگ آسیا (شامل کشورهای ایران، ژاپن، کره و غیره)، لیگ‌های اروپایی (شامل کشورهای آلمان، اسپانیا، ایتالیا و غیره)، لیگ‌های آمریکایی (به عنوان مثال برزیل، آرژانتین و غیره) استخراج کردیم. به منظور حفظ جامعیت داده‌ها و تنوع در شرایط واقعی، بازی‌ها مربوط به ساعات مختلف و در فصل‌های متفاوت می‌باشند که از بیش برازش مدل‌های پیشنهادی مبتنی بر یادگیری ماشینی جلوگیری نماید.

به لحاظ آماری، تصاویر به گونه‌ای جمع‌آوری شده‌اند که نیمی از آنها حاوی دروازه و نیمه دیگر تصاویر دیگری از بازی فوتبال هستند یعنی ما در این پایگاه داده تصاویر، ۵۰۰ تصویر شامل دروازه داریم و ۵۰۰ تصویر دیگر از بازی فوتبال می‌باشد که در آنها دروازه مشاهده نمی‌شود. همچنین اندازه تصاویر متنوع است و هیچ محدودیتی برای آنها لحاظ نشده است در نتیجه طول و عرض تصاویر مختلف است با این حال کوچکترین تصویر در اندازه ۶۴۰ × ۲۸۸ است و بزرگترین تصویر در اندازه ۱۹۲۰× ۱۰۸۰ می‌باشد لازم به ذکر است که همه تصاویر رنگی هستند.

۴_۳ معیارهای ارزیابی روش پیشنهادی

ما به منظور ارزیابی عملکرد روش پیشنهادی از ۴ معیار ارزیابی شامل بازیابی[۱۰] (رابطه ۸)، وضوح[۱۱] (رابطه ۹)، معیار-اف[۱۲] (رابطه ۱۰) و دقت[۱۳] (رابطه ۱۱) استفاده کرده ایم. همچنین بر اساس این پارامترها مشخصه عملکرد سیستم[۱۴] را نیز محاسبه می‌کنیم. در این ارزیابی‌ها هدف پیدا کردن فریم‌هایی از یک ویدئوی بازی فوتبال است که در آنها دروازه مشاهده شود، انگیزه نویسندگان از شناسایی فریم‌های شامل دروازه آن است که برش‌هایی از ویدئوی بازی که در آنها دروازه مشاهده می‌شود احتمالا جزء بخش‌های حساس بازی است و در نتیجه برای خلاصه‌سازی ویدئوی بازی فوتبال می‌تواند به عنوان یک ویژگی سطح بالا تلقی گردد. اگر چه می‌توان دانش شناسایی برش‌های حساس بازی را هنوز هوشمندانه‌تر تعریف کرد که آن نیز می‌تواند به عنوان بخشی از کارهای آتی به شمار آید.

$$Recall = \frac{TP}{TP+FP} \qquad (۸)$$

$$Precision = \frac{TP}{TP+FN} \qquad (۹)$$

که در آن TP[۱۵] تعداد نمونه های مثبتی است که به درستی مثبت شناسایی شده‌اند، TN[۱۶] تعداد نمونه‌های منفی که به درستی منفی شناسایی شده‌اند، FP[۱۷] تعداد شناسایی‌های مثبت کاذب و FN[۱۸]تعداد شناسایی‌های منفی کاذب می‌باشند. سپس مقدار معیار-اف f- و دقت به شرح زیر تعریف می‌شوند:

$$\mathrm{f}-measure = \frac{۲*Precision*Recall}{Precision+Recall} \qquad (۱۰)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۱۱)$$

[۱۰] Recall
[۱۱] Precision
[۱۲] F-measure
[۱۳] Accuracy
[۱۴] ROC
[۱۵] True Positive
[۱۶] True Negative
[۱۷] False Positive
[۱۸] False Negative

جدول ۳. مقایسه نتایج روش پیشنهادی و روش پایه

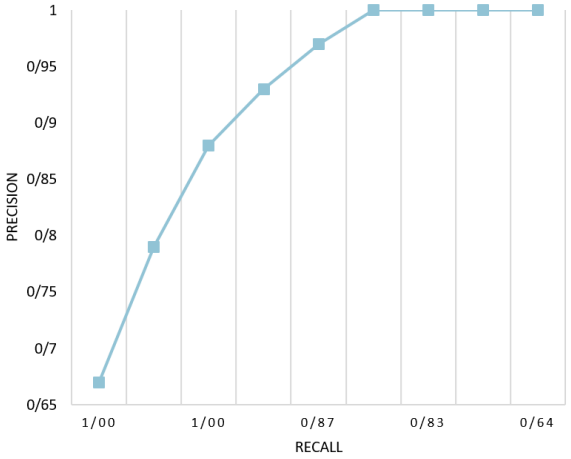
نام روش	معماری پایه VGG	روش پیشنهادی
TP	٪ ۸۹	٪ ۹۳
TN	٪ ۸۵	٪ ۹۱
FN	٪ ۱۱	٪ ۷
FP	٪ ۱۵	٪ ۹
معیار بازیابی	۰،۸۵	۰،۹۱
معیار وضوح	۰،۸۹	۰،۹۳
معیار- اف	۰،۸۶	۰،۹۱
دقت	٪ ۸۷	٪ ۹۲
زمان آموزش	۰،۴۲۶ ثانیه/عکس	۰،۵۵۶ ثانیه/عکس

زمان آموزش ثانیه به ازای هر عکس می‌باشد.

۴_۴ ارزیابی و مقایسه روش ارائه شده

در این بخش، نتایج ارزیابی روش پیشنهادی در ترکیب ویژگی‌های استخراج شده توسط مدل شبکه عصبی عمیق با ویژگی‌های سراسری استخراج شده به روش کلاسیک ارائه شده است. طی این ارزیابی با شرایط یکسان، مدل VGG-۱۳ که مدل پایه معماری ارائه شده می‌باشد و روش پیشنهادی روی پایگاه‌داده معرفی شده در بخش ۴-۲ مورد ارزیابی قرار داده‌ایم. در این ارزیابی که هدف آن شناسایی دروازه در تصاویر می‌باشد، مجموعه داده‌های تست که شامل تصاویر دربردارنده‌ی دروازه و تصاویر فاقد دروازه می‌باشند به شبکه عصبی عمیق پایه و شبکه عصبی عمیق پیشنهادی ارائه شده در بخش ۳، به عنوان ورودی داده شده و سپس نتایج روش پیشنهادی روی مجموعه تصاویر تست محاسبه و نتایج حاصل با نتایج کارهای قبلی [۲۴] مقایسه شده است. جدول ۳ این نتایج را نشان می‌دهد چنانکه مشاهده می‌شود دقت روش پیشنهادی با معماری دو مسیره نسبت به روش پایه (VGG) برای شناسایی تصاویر حاوی دروازه از بهبود قابل توجهی برخوردار شده است و از ٪۸۷ به ٪۹۲ افزایش یافته است. این نتیجه با بهبود متناسب و معناداری بر روی همه مولفه‌های حساسیت و شناسایی حاصل شده است که بیانگر بهبود کارآیی روش پیشنهادی نسبت به روش پایه می‌باشد. روش پیشنهادی علی‌رغم بهبود دقت دارای زمان آموزش بیشتری می‌باشد. همچنین بر این اساس، دو مولفه دقت و کارآیی محاسبه شده و در آن جدول ذکر شده است. ضمن آنکه بر اساس این دو مولفه، منحنی مشخصه عملکرد سیستم در شکل ۴ نمایش داده شده است.

همچنین نتایج حاصل از مقایسه روش پیشنهادی با سایر روش‌ها براساس پارامتر دقت در جدول ۴ ارائه شده است. نتایج حاصل به وضوح برتری روش پیشنهادی در مقایسه با سایر کارهای انجام شده قبلی در این زمینه را نشان می‌دهد.



شکل ۴ . منحنی مشخصه عملکرد سیستم روش پیشنهادی

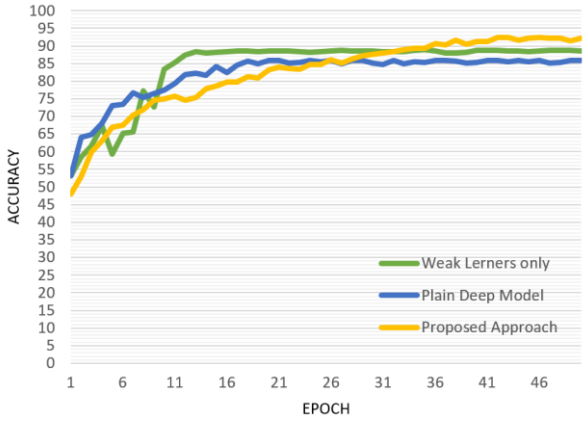
در یک مقایسه دیگر ما نتایج حاصل از روش پیشنهادی را که یک روش ترکیبی است با هر یک از نتایج شبکه عصبی پایه و روشی که

جدول ۴. مقایسه نتایج روش پیشنهادی و سایر روش‌ها

روش	دقت
معماری پایه VGG	٪ ۸۷
همه یادگیرنده‌های ضعیف [۲۴]	٪ ۵۷
یادگیرنده‌های ضعیف منتخب [۲۴]	٪ ۸۰،۵
روش پیشنهادی	٪ ۹۲

در آن صرفا از سیستم غربال‌گر استفاده کرده است مقایسه کردیم که در شکل ۵ نشان داده شده است.

چنانکه از این تصویر مشاهده می شود به ازای آموزش در مرحل میانی و پس از آن، نتایج از هر دو مدل مجزا بهتر می باشد.



شکل ۵. مقایسه نتایج روش‌های پیشنهادی با روش پایه VGG و روش سیستم غربالگر به تنهایی

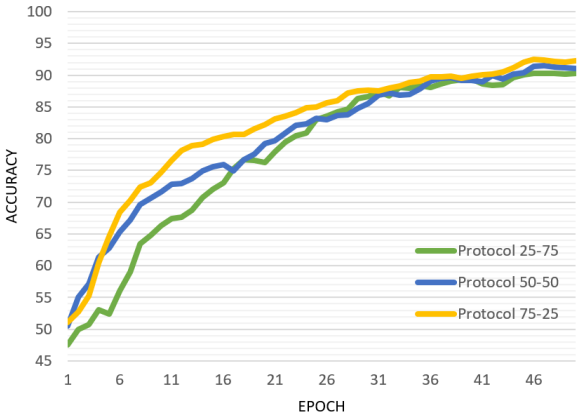
از طرف دیگر به منظور تحلیل بیشتر روش پیشنهادی با توجه به آنکه در معماری روش‌های مبتنی بر شبکه‌های عصبی عمیق نقش تعداد داده‌های آموزشی مهم است ما برای بررسی میزان تاثیر داده‌های آموزشی در میزان دقت روش پیشنهادی، از سه قرارداد تقسیم داده های آموزشی و آزمایشی استفاده می کنیم. در قرارداد

اول، ۲۵٪ کل داده‌های آموزشی را برای آموزش مدل و ۷۵٪ را برای آزمایش اســـتفاده می‌کنیم. طی قرارداد دوم، ما ۵۰٪ داده‌ها را برای آموزش و ۵۰٪ را برای آزمایش اســـتفاده می‌کنیم و بالاخره در قرارداد سوم، ۷۵٪ داده‌های آموزشی برای آموزش مدل و ۲۵٪ برای آزمایش مورد اســـتفاده قرار می‌گیرند. به این ترتیب علاقه‌مندیم نقش تعداد داده‌های آموزشی بر مدل پیشنهادی را نیز مورد ارزیابی قرار دهیم. نتایج روش پیشنهادی با تقسیم‌بندی ذکر شده بر روی پایگاه‌داده تصاویر معرفی شده در بخش ۴-۲ اعمال شده است که در جدول ۵ و نمودار گرافیکی آن طی مراحل[۱۹] مختلف در شکل ۶ قابل مشاهده می‌باشد.

نتایج ارزیابی روش پیشنهادی در دسته بندی مجموعه تصاویر آزمایشی برای تعدادی از تصاویر در شکل ۷ ارائه شده‌اند. شکل ۷-

جدول ۵. مقایسه نتایج روش پیشنهادی و روش پایه

رهیافت	دقت		
قرارداد ارزیابی	۷۵-۲۵	۵۰-۵۰	۲۵-۷۵
روش عمیق پایه	٪ ۸۷	٪ ۸۵	٪ ۶۷
روش پیشنهادی	٪ ۹۲	٪ ۹۱	٪ ۹۰



شکل ۶. ارزیابی روش پیشنهادی با سه قرارداد متنوع در استفاده از داده‌های آموزشی و آزمایشی ۲۵-۷۵، ۵۰-۵۰ و ۷۵-۲۵ درصد

الف به وضوح توانایی روش پیشنهادی در شناسایی تصاویر فاقد دروازه در مجموعه تصاویر و شکل ۷-ب توانایی روش پیشنهادی در شناسایی و دسته بندی تصاویر دارای دروازه را نشان می‌دهد. براساس نتایج ارائه شده روش پیشنهادی توانسته با درصد احتمال بالایی تصاویر هر دسته را شناسایی و تفکیک کند که گویای کارایی و دقت روش پیشنهادی می‌باشد.

برای ارزیابی روش پیشنهادی اقدام به مقایسه نتایج روش پیشنهادی با سایر کارهای انجام شده کرده‌ایم. برای این منظور برای مقایسه و ارزیابی روش پیشنهادی با نتایج کار انجام شده توسط کریمی و

ب الف

شکل ۷. ارزیابی روش پیشنهادی در دسته‌بندی مجموعه داده‌های تست به دو دسته تصاویر فاقد دروازه (الف) و دسته تصاویر دارای دروازه (ب)

همکاران، داده‌های مربوط به رویداد کارت زرد ارائه شده توسط کریمی و همکاران[۲۰] را دانلود و سپس نتایج روش پیشنهادی با کار انجام شده توسط کریمی و همکاران را در جدول شماره ۶ ارائه کرده‌ایم. براساس نتایج ارائه شده روش پیشنهادی در شناسایی رویداد کارت زرد دارای دقت بهتری می‌باشد.

۵. نتیجه‌گیری

جدول ۶. مقایسه نتایج روش پیشنهادی در شناسایی کارت زرد و سایر روش‌ها

روش	دقت
کریمی و همکاران [۱۸]	٪ ۹۲٫۶۶
روش پیشنهادی	٪ ۹۵٫۲۶

امروزه حجم بسیار فراوانی از ویدئوهای مختلف در اختیار کاربران در سراسر جهان قرار دارد. برخی از این ویدیوها مربوط به حوزه سرگرمی و برخی دیگر نیز مرتبط با حوزه‌های نظارتی و امنیتی می‌باشند. از جمله این ویدئوها، ویدیوهای ورزشی و خصوصا ویدئوهای ورزش فوتبال است که به دلیل علاقه‌مندی طیف گسترده‌ای از مردم جهان به این ورزش دارای اهمیت بالایی است. علاوه بر موضوع علاقه‌مندی فوتبال دوستان، مسئله زمان طولانی بازی فوتبال است که در اغلب موارد همه مردم فرصت تماشای ۹۰ دقیقه بازی فوتبال را ندارد و البته علاقه‌مند هستند دستکم لحظات مهم و هیجان انگیز بازی را مشاهده کنند. به همین علت اخیرا برخی از فراهم‌کنندگان خدمات ارائه ویدئوهای ورزشی به خلاصه‌سازی بازی فوتبال پرداخته‌اند که با

[۱۹] Epoch

[۲۰] https://github.com/FootballAnalysis/footballanalysis/tree/main/Dataset/Soccer٪۲۰Event٪۲۰Dataset٪۲۰(Image)

استقبال کاربران‌شان مواجه شده است. در این مقاله ما یک روش خودکار با استفاده از یک مدل معماری دومسیره یادگیری عمیق برای تحلیل تصاویر ویدئویی ورزش فوتبال، با تاکید بر شناسایی دروازه به عنوان یکی از مهمترین عناصر رویداد گُل که مهمترین رویداد بازی فوتبال می‌باشد، ارائه کردیم.

رشد چشمگیر روش‌های مبتنی بر یادگیری عمیق توانسته نتایج قابل قبولی را در حوزه تحلیل تصویر فراهم کنند. روش‌های مبتنی بر یادگیری عمیق کارآیی بسیار مطلوبی در استخراج ویژگی‌ها دارند، اما تضمین‌کننده استخراج همه ویژگی‌ها یا به عبارتی بهترین ویژگی‌ها نیستند. در نتیجه با اطمینان می‌توان گفت که استخراج خودکار ویژگی‌های تصویر برای بازشناسی دروازه در تصاویر بازی فوتبال اگر چه بسیار مفید است اما کامل نبوده و ترکیب ویژگی‌ها سبب بهبود دقت سیستم شناسایی هدف می‌گردد. ترکیب ویژگی‌های استخراج شده با روش‌های سنتی که قادر به پوشش مسئله هدف هستند با روش‌های خودکار استخراج ویژگی عمیق پروسه‌ای هوشمندانه است که می‌تواند منجر به کسب نتایجی بهتر در مقایسه با هر یک از آن‌ها به تنهایی شود.

در این مقاله، ما یک روش خودکار هوشمند بازشناسی دروازه در جهت شناسایی لحظات حساس بازی فوتبال به منظور خلاصه‌سازی آن ارائه کردیم که در آن با ارائه یک مدل معماری دومسیره یادگیری عمیق، از هر دو مسیر ویژگی استخراج شده توسط شبکه‌های عصبی عمیق و ویژگی‌های سنتی بهره‌مند شدیم. در این مقاله نشان دادیم که ترکیب این ویژگی‌ها توصیف مطلوب‌تری به‌منظور بازشناسی هدف (دروازه) حاصل می‌کند. براساس نتایج ارائه شده، روش پیشنهادی در مقایسه با سایر روش‌ها از دقت بیشتر و خطای کمتری برخودار می‌باشد.

مراجع

[۱] P. Shi and X. Yu, "Goal event detection in soccer videos using multi-clues detection rules," in *Management and Service Science, ۲۰۰۹. MASS'۰۹. International Conference on*, ۲۰۰۹, pp. ۱–٤.

[۲] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. ۱۰, no. ۲, pp. ۲۵۲–۲۵۹, ۲۰۰۸.

[۳] M. H. Kolekar, "Bayesian belief network based broadcast sports video indexing," *Multimedia Tools and Applications*, vol. ٥٤, no. ۱, pp. ۲۷–٥٤, ۲۰۱۱.

[٤] D. W. Tjondronegoro and Y.-P. P. Chen, "Knowledge-discounted event detection in sports video," *Ieee transactions on systems, man, and cybernetics-part a: Systems and humans*, vol. ٤۰, no. ٥, pp. ۱۰۰۹–۱۰۲٤, ۲۰۱۰.

[٥] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Transactions on Multimedia*, vol. ۷, no. ٦, pp. ۱۰٦٦–۱۰۸۳, ۲۰۰٥.

[٦] B. Li, J. H. Errico, H. Pan, and I. Sezan, "Bridging the semantic gap in sports video retrieval and summarization," *Journal of Visual Communication and Image Representation*, vol. ۱٥, no. ۳, pp. ۳۹۳–٤۲٤, ۲۰۰٤.

[۷] H.-G. Kim, S. Roeber, A. Samour, and T. Sikora, "Detection of goal events in soccer videos," in *Storage and Retrieval Methods and Applications for Multimedia ۲۰۰٥*, ۲۰۰٥, vol. ٥٦۸۲, pp. ۳۱۷–۳۲٦.

[۸] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models," in *Multimedia and Expo, ۲۰۰۳. ICME'۰۳. Proceedings. ۲۰۰۳ International Conference on*, ۲۰۰۳, vol. ۳, p. III–۲۹.

[۹] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham, "Highlights for more complete sports video summarization," *IEEE multimedia*, vol. ۱۱, no. ٤, pp. ۲۲–۳۷, ۲۰۰٤.

[۱۰] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong, "Semantic event detection using conditional random fields," in *۲۰۰٦ Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'۰٦)*, ۲۰۰٦, pp. ۱۰۹–۱۰۹.

[۱۱] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic Bayesian network," *IEEE Transactions on Multimedia*, vol. ۸, no. ٤, pp. ۷٤۹–۷٦۰, ۲۰۰٦.

[۱۲] M. Y. Eldib, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar, and M. El-Saban, "Soccer video summarization using enhanced logo detection," in *Image Processing (ICIP), ۲۰۰۹ ۱٦th IEEE International Conference on*, ۲۰۰۹, pp. ٤۳٤٥–٤۳٤۸.

[۱۳] B. Fakhar, H. R. Kanan, and A. Behrad, "Event detection in soccer videos using unsupervised learning of Spatio-temporal features based on pooled spatial pyramid model," *Multimedia Tools and Applications*, pp. ۱–۳۱, ۲۰۱۹.

[۱٤] J. Yu, A. Lei, and Y. Hu, "Soccer Video Event Detection Based on Deep Learning," in *International Conference on Multimedia Modeling*, ۲۰۱۹, pp. ۳۷۷–۳۸۹.

[۱٥] Z. Dang, J. Du, Q. Huang, and S. Jiang, "Replay detection based on semi-automatic logo template sequence extraction in sports video," in *Image and Graphics, ۲۰۰۷. ICIG ۲۰۰۷. Fourth International Conference on*, ۲۰۰۷, pp. ۸۳۹–۸٤٤.

[۱٦] H. Pan, P. Van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *icassp*, ۲۰۰۱, pp. ۱٦٤۹–۱٦٥۲.

[۱۷] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. Kim, "Event detection based approach for soccer

video summarization using machine learning," *International Journal of Multimedia and Ubiquitous Engineering*, vol. ۷, no. ۲, pp. ٦۳–۸۰, ۲۰۱۲.

[۱۸] A. Karimi, R. Toosi, and M. A. Akhaee, "Soccer Event Detection Using Deep Learning," *arXiv preprint arXiv:* ۲۱۰۲,۰٤۳۳۱, ۲۰۲۱.

[۱۹] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ۲۰۱٤, pp. ٥۸۰–٥۸۷.

[۲۰] S. Gerke, K. Muller, and R. Schafer, "Soccer jersey number recognition using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, ۲۰۱٥, pp. ۱۷–۲٤.

[۲۱] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ۲۰۱٦, pp. ۷۷۰–۷۷۸.

[۲۲] S. He and L. Schomaker, "Deep adaptive learning for writer identification based on single handwritten word images," *Pattern Recognition*, vol. ۸۸, pp. ٦٤–۷٤, ۲۰۱۹.

[۲۳] S. F. Chevtchenko, R. F. Vale, V. Macario, and F. R. Cordeiro, "A convolutional neural network with feature fusion for real-time hand posture recognition," *Applied Soft Computing*, vol. ۷۳, pp. ۷٤۸–۷٦٦, ۲۰۱۸.

[۲٤] A. Zanganeh and M. Jampour, "Automatic Weak Learners Selection for Pattern Recognition and its application in Soccer Goal Recognition," ۲۰۱۹ *٤th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, ۲۰۱۹, pp. ۲٤۰-۲٤٥, doi: ۱۰,۱۱۰۹/PRIA.۲۰۱۹,۸۷۸٥۹٦٦.