دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر
سمینار کارشناسی ارشد گرایش هوش مصنوعی

عنوان:

توصیف ویدئو با استفاده از شبکه‌های ژرف بازگردنده

Deep Video Captioning Using Recurrent Neural Networks

نگارش:
سید علیرضا میر محمد صادقی
۹۴۲۱۱۴۱۳

استاد راهنما:
دکتر مهدیه سلیمانی

استاد ممتحن داخلی:
دکتر شهره کسائی

بهمن ۹۵

چکیده: مسئله‌ی توصیف ویدئو [1] به دنبال ارائه‌ی خودکار توصیفی از محتوای یک داده‌ی تصویری در قالب جملات قابل فهم توسط انسان است. ورودی زمان آموزش مدل در این مسئله، یک ویدئو به همراه یک یا چندین توصیف زبان مرتبط با محتوی بصری ویدئو است. در زمان آزمون، ورودی مدل یک ویدئو بدون هیچ گونه توصیف است که هدف مدل، ایجاد توصیفی به زبان طبیعی برای ویدئوی ورودی است. لازم به ذکر است که در این مسئله، مدل تنها قابلیت ارائه‌ی توصیف‌های قابل استنباط از ویژگی‌های بصری ویدئو را دارا است و از اطلاعات موجود در صدا و یا اطلاعات پیش‌زمینه‌ای دیگر استفاده‌ای نمی‌شود. ساختار این نوشتار بدین گونه است که در بخش ۱، تعریف مسئله توصیف ویدئو ارائه می‌شود. بخش ۲، دسته‌ای از کارهای پیشین انجام شده را معرفی می‌کند. سپس در بخش‌های بعدی یک روش پیشنهادی ارائه شده و نتایج آن با روش‌های پایه مقایسه شده است و سرانجام راه‌کارهایی برای ادامه‌ی پژوهش بیان می‌شود و مطالب نوشتار جمع‌بندی می‌شوند.

واژه‌های کلیدی: شبکه‌های ژرف، شبکه‌های بازگردنده، ویدئو، توصیف ویدئو

۱ مقدمه

برای بیشتر افراد، مشاهده‌ی یک ویدئوی کوتاه و توصیف محتوی آن، کاری ساده است. اما برای رایانه‌ها، استخراج مفاهیم از پیکسل‌های ویدئو و ایجاد توصیفی به زبان‌طبیعی یک امر بسیار پیچیده است. مخلوط کردن پردازش‌زبان‌طبیعی[2] با بینایی‌ماشین [3] جهت ایجاد توصیف‌هایی به زبان انگلیسی از داده‌های تصویری، یکی از زمینه‌های فعال پژوهش در زمینه‌ی یادگیری ماشین است. راه‌حل‌های متعددی برای دامنه‌های کوچک که مجموعه‌ی کوچکی از فعالیت‌ها را شامل می‌شوند، پیشنهاد شده است، اما توصیف ویدئوها با دامنه‌باز [4] ، مانند ویدئوهای یوتیوب[5] هنوز یک مسئله‌ی باز است.

بخشی از علل عدم پیشرفت در زمینه‌ی توصیف ویدئوهای با دامنه‌باز، عدم وجود مجموعه‌دادگان جامع از ویدئو به زبان‌طبیعی است. یکی از دیگر علت‌های این واقعه، نبود مدل‌های مناسبی است که بتوانند ارتباط بین فریم‌های پشت‌سر هم ویدئو و کلمات را ضبط کنند.

پژوهش‌های گذشته، با ثابت در نظر گرفتن ساختارهای معنایی کوچک، مانند فعل، فاعل و مفعول، به عنوان یک نمایش میانی، سعی در ایجاد جملات زبان‌طبیعی برای ویدئوها داشته‌اند. اما مشخصا استفاده از این نمایش برای مجموعه‌دادهای بزرگ مناسب نیست و منجر به ایجاد جملات بسیار ساده‌تر، و گاها بی‌ربط، نسبت به محتوی ویدئو می‌شود.

در چندسال اخیر، مطالعات گسترده‌ای در حوزه یادگیری داده‌های چندحالته[6] صورت گرفته است، به طور خاص عموم این روش‌ها با استفاده از شبکه‌های عصبی ژرف پیچشی[7] و بازگردنده[8] سعی به ادغام کردن بازنمایشی از فریم‌ها و کلمات زبان‌طبیعی با هم کرده و به نتایج امیدوارکننده‌ای رسیده‌اند.

در این نوشتار بر مسئله‌ی ایجاد توصیف برای ویدئو با استفاده از شبکه‌های ژرف تمرکز می‌کنیم؛ به این معنی که داده‌هایی که مایل به ایجاد توصیف برای آن‌ها هستیم، ویدوئوها هستند. البته باید دقت داشت که این مدل‌ها، در حالتی که ویدئوی ورودی تنها متشکل از یک فریم باشد، هم‌عرض مدل‌هایی می‌شوند که روی تصاویر کار می‌کنند. ورودی مدل در زمان آموزش یک یا چند جمله و دنباله‌ای از فریم‌ها است و خروجی مدل یک یا چند جمله به زبان‌طبیعی از محتوی فریم‌ها است که با جملات ورودی مقایسه می‌شوند. در زمان آزمون، ورودی مدل دنباله‌ای از فریم‌هاست و خروجی مدل نیز یک یا چند جمله توصیف‌گر فریم‌ها است.

مباحث در این گزارش به این صورت است: در بخش ۲، صورت‌های مختلفی از مسئله‌ی توصیف ویدئو را بیان کرده و روش‌های پیشین ارائه شده برای حل آن را مرور می‌کنیم. در بخش ۳، یک روش پیشنهادی بیان می‌شود و نتایج عملی آن در بخش ۴ ارائه و با روش‌های دیگر مقایسه می‌شود. بخش ۶ به کارهای آتی، جدول زمان‌بندی پژوهش و جمع‌بندی اختصاص دارد.

۲ کارهای پیشین

در تحقیقات پیشین، از روش‌های متعددی برای توصیف ویدئوها استفاده شده است. به طور کلی می‌توان این روش‌ها را به دو دسته‌ی روش‌های سنتی و روش‌های مبتنی بر شبکه‌های ژرف تقسیم کرد. عموم روش‌های سنتی بدین صورت عمل می‌کنند که ابتدا با استفاده از مجموعه‌ای از ویژگی‌های دست‌ساز، با استفاده از الگوریتم‌های شناسایی اشیاء یا کنش‌ها، یک بازنمایش از اشیاء و کنش‌های موجود در تصویر به دست می‌آورند و سپس تلاش می‌کنند دانش به‌دست آمده از ویژگی‌های تصویری را به نحوی با دانش به دست آمده از روش‌های پردازش زبان طبیعی مخلوط کنند و به مدلی دست یابند که عملکرد مناسبی در ایجاد جملات به زبان طبیعی داشته باشد. برای مثال در [۱]، نویسندگان یک روش سه مرحله‌ای برای توصیف ویدئوها ارائه می‌دهند. در گام اول این روش، با استفاده از الگوریتم‌های بازشناسی تصویر، اشیاء و کنش‌های موجود در ویدئو استخراج می‌شوند. در ادامه و درگام دوم، با استفاده از تخمین بیشینه‌ی درست نمایی [۹] از دنیای واقعی، که با کاوش‌کردن سه‌بخشی‌های فعل، فاعل و مفعول از متون موجود در اینترنت به دست آمده است، یافته‌های گام اول را با دانش موجود از کاوش متون، مخلوط می‌کنند تا بهترین سه‌تایی را به دست آورند. در ادامه و در گام نهایی، با استفاده از مجموعه‌ای از جملات قالب (از قبل تعیین شده)، جملات نهایی ایجاد شده و بر اساس روان و منطقی بودن رده‌بندی می‌شوند و بهترین جمله انتخاب می‌شود.

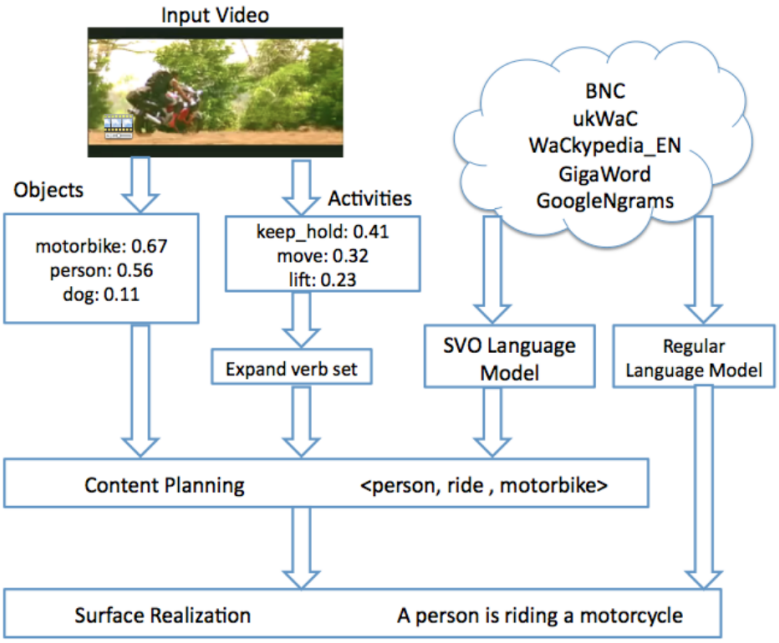
باید دقت داشت که با توجه به ابعاد بسیار بزرگ فضای ویودئو، ابتدا لازم است که این فضا خود به فضایی کوچک‌تر نگاشت شود. به همین دلیل عموم روش‌های پیشین ارائه شده، ابتدا از فریم‌های مختلف ویدئو نمونه‌برداری می‌کنند و سپس با استفاده از شبکه‌های ژرف پیچشی، سعی در شناسایی یک فضای میانی کوچک‌تر برای فریم‌ها می‌کنند (استخراج ویژگی). در نهایت نیز با توجه به هدف مسئله از یک دسته‌بند (شبکه‌ی کاملا متصل [۱۰] یا دسته‌بندهای سنتی مانند ماشین‌های بردار پشتیبان [۱۱]یا درخت‌تصمیم [۱۲]) و یا یک شبکه‌ی بازگردنده [۱۳] استفاده می‌کنند.

در این بخش، با توجه به فراگیری این چارچوب به مرور دقیق‌تر روش‌های پیشنهاد شده‌ی پیشین برای حل این مسئله می‌پردازیم.

۲.۱ ایجاد توصیف با استفاده از دانش متن‌کاوی شده

در این روش، نویسندگان پژوهش [۱] مدلی را ارائه می‌دهند که با در نظر گرفتن مخلوطی از دانش به دست آمده از حوزه‌ی پردازش‌زبان‌طبیعی و روش‌های بینایی‌رایانه‌ای، به ایجاد جملات توصیف‌کننده برای ویدئوها می‌پردازد. این مدل از دو مرحله‌ی اصلی تشکیل شده است. در مرحله‌ی اول، محتمل‌ترین فعل، فاعل و مفعول، با توجه به محتوا تصویر (روش‌های بینایی‌رایانه‌ای) و احتمال کنارهم قرارگیری این سه کلمه (روش‌های پردازش زبان طبیعی) استخراج می‌شوند. به این مرحله، مرحله‌ی برنامه‌ریزی محتوا [۱۴] نیز گفته می‌شود. سپس در مرحله‌ی دوم، با توجه به سه کلمه‌ی به دست آمده در مرحله‌ی قبل و با استفاده از یک روش ساده‌ی مبتی بر قالب، جملات توصیف‌گر ورودی ایجاد می‌شوند. این جملات در نهایت با استفاده از یک مدل زبانی احتمالاتی که روی داده‌های متنی موجود در اینترنت آموزش داده شده است، رده‌بندی می‌شوند تا بهترین جمله‌ی توصیف‌کننده‌ی یک ویدئو انتخاب شود. به این مرحله، مرحله‌ی تحقق سطح [۱۵] گفته می‌شود. در تصویر ۲.۱ شمایی از این مدل قابل مشاهده است. نویسندگان در این پژوهش، ابتدا اشیاء و کنش‌های موجود در ویدئو را با استفاده از مدل‌های بازشناسی تصویر به دست می‌آورند. با توجه به اینکه این اطلاعات عموماً دارای نوفه [۱۶] و خطای قابل توجهی هستند، گام بعدی توسط نویسندگان پیشنهاد شده است. در این گام، با در نظر گرفتن احتمال قرارگرفتن اشیاء و کنش‌های تشخیص داده‌شده در مرحله‌ی قبل، در متون موجود در منابع مختلف، محتمل‌ترین فعل، فاعل و مفعول استخراج می‌شوند. در نهایت نیز پیش‌بینی‌ها با قرار گرفتن در قالب‌های متفاوت نمره‌دهی و رتبه‌بندی می‌شوند تا بهترین جمله‌ی توصیف‌کننده انتخاب شود.

در این روش ۲.۱ برای بازشناسی اشیاء از روش ارائه شده در [۲] استفاده می‌شود. همچنین برای شناسایی کنش‌های موجود در ویدئو از مدل ارائه شده توسط [۳] استفاده می‌شود. در این روش ویژگی‌ها، بافت‌نگار بردارهای گرادیان [۱۷] و شارش‌نوری [۱۸] هستند که روی نقاط پراهمیت مکانی‌ـ‌زمانی [۱۹] محاسبه می‌شوند. در گام بعدی با اجرای خوشه‌بندی روی این ویژگی‌ها، یک نمایش کیسه‌کلمات [۲۰] برای هر خوشه از ویژگی‌ها



شکل ۲.۱: *برنامه‌ریزی محتوا و تحقق سطح* [۱]

به دست می‌آید که با تجمیع این نمایش برای تمامی فریم‌ها، بازنمایش نهایی ویدئو به صورت بافت‌نگاری از نمایش‌های کیسه کلمات به دست می‌آید. در نهایت با آموزش دادن یک مدل دسته‌بند (مانند ماشین‌های بردار پشتیبان)، عمل رخ‌داده‌شده در ویدئو شناسایی می‌شود. در مرحله‌ی برنامه‌ریزی محتوا، اطلاعات مرتبط با تصویر با استفاده از رابطه‌ی زیر، با اطلاعات به دست آمده از متون، مخلوط می‌شوند تا بهترین سه‌تایی‌های فعل، فاعل و مفعول به دست آیند.

$$visScore = p(S|vid) * p(V_{orig}|vid) * sim(V_{sim}, V_{orig}) * p(O|vid) \quad (۱)$$

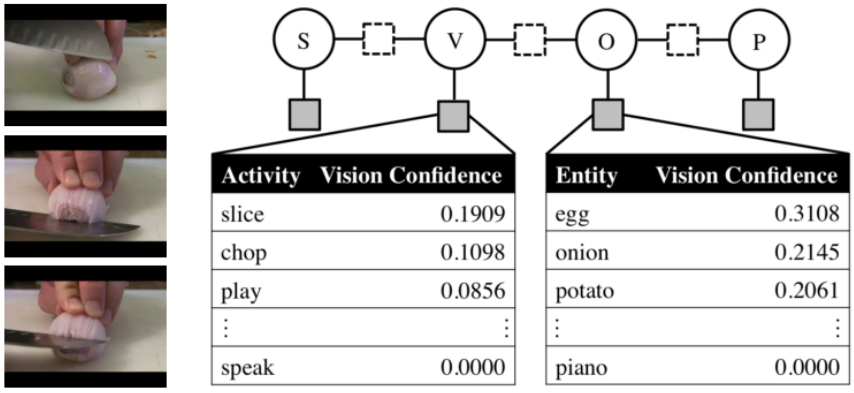
$$score = w_۱ * visScore + w_۲ * nlpScore \quad (۲)$$

در رابطه‌ی ۱ ، منظور از V_{sim} مجموعه‌ای از افعال اضافی و مشابه با افعال شناسایی شده در مرحله‌ی قبل است که با توجه به یک دسته‌بندی قبلی توسط نویسندگان، در کنار هر فعل در نظر گرفته می‌شود. (به عنوان مثال برای فعل حرکت‌کردن، افعال راه‌رفتن، دویدن و رد کردن نیز در نظر گرفته می‌شود). در مرحله‌ی بعد، سه کلمه‌ی انتخاب شده در قالب‌هایی آماده برای جملات قرار می‌گیرند، جملات ساخته‌شده توسط مدل زبانی [۴] امتیاز دهی می‌شوند و در نهایت بهترین جمله، جمله با بالاترین امتیاز، انتخاب می‌شود. نمونه‌ای از قالب‌های استفاده شده به صورت مقابل است: "A/The، Subject، Verb، Preposition (optional)، A/The، Object"

در مطالعه‌ی [۵] نویسندگان با کمی تغییر در مدل قبلی، به مدلی دست یافته‌اند که دقت بالاتری در ایجاد توصیف برای ویدئو دارد. در این مطالعه، دو تغییر اصلی نسبت به مطالعه‌ی قبلی وجود دارد. تفاوت اول اینکه در این مدل با استفاده از روش‌های شناسایی صحنه [۲۱] ، کلمه‌ای برای توصیف تصاویر پشت صحنه‌ی ویدئو تشخیص داده می‌شود و به سه‌تایی فعل، فاعل و مفعول افزوده می‌شود. در این روش برای شناسایی صحنه، ویژگی‌های LBP[۲۴]، SIFT[۲۳]، SSIM [۲۲]، HoG، GIST و ام‌های رنگ و بافت از تصویر استخراج شده‌اند. سپس با آموزش دادن

یک دسته‌بند ماشین بردار پشتیبان روی تمامی فریم‌های ویدئو و میانگین‌گیری از امتیاز هر کلاس، توزیعی روی تمامی تصاویر پشت‌صحنه در یک ویدئو به دست می‌آید.

تفاوت دوم این مدل با مدل قبلی نیز، استفاده این مدل از یک گراف فاکتور [۲۵] (شکل ۲.۲) برای مخلوط کردن دانش به دست آمده از متون و ویژگی‌های تصویری است. در این روش، بعد از تعیین توابع پتانسیل، با بهره‌گیری از تخمین بیشینه‌گر احتمال پسین [۲۶] (با استفاده از الگوریتم



شکل ۲.۲: *مدل گراف فاکتور* [۵]

Max-product)، محتمل‌ترین مجموعه توامان از مقادیر را برای متغیرهای نهان مدنظر به دست می‌آوریم. توابع پتانسیل تصویری و زبانی برای این مسئله، با در نظر گرفتن $k \in \{S, V, O, P\}$ به صورت زیر تعریف شده‌اند:

$$\phi_k(t) = C_k(t) \quad , \quad \phi_{k,l}(t,s) = p(l = s|k = t) = \alpha p.(l = s|k = t) + (1 - \alpha)p_i(l = s|k = t) \qquad (۳)$$

با توجه به این‌که مشاهدات تصویری مدل، به صورت یک توزیع احتمال در خواهند آمد پتانسیل تصویری هر عنصر (فعل، فاعل، مفعول و صحنه) k، میزان اطمینانی است که دسته‌بند برای عنصر kم به کلمه‌ی tم تخصیص می‌دهد. در توابع پتانسیل زبانی، k و l دو عنصر پشت‌سر هم در دنباله‌ی $SVOP$ هستند که s و t مقادیر ممکن برای آن دو عنصر را مشخص می‌کنند.

روش ایجاد جملات در این پژوهش تقریبا تفاوتی با پژوهش قبل ندارد و از همان روش قالب محور استفاده شده است. تنها تفاوت در این گام، با توجه به افزوده شدن صحنه به سه‌تایی فعل، فاعل و مفعول، بررسی لزوم نیاز جمله به مفعول و صحنه‌ی تصویر است. در واقع با توجه به این‌که بعضی از افعال غیر متعدی هستند و یا ممکن است آوردن صحنه باعث تکرار مفاهیم شود، سه نوع متفاوت جمله از بهترین عناصر به دست آمده (SVO، SVP، SVOP) ساخته می‌شوند و توسط مدل زبانی [۴] امتیازدهی می‌شوند تا بهترین آنها به عنوان خروجی مدل انتخاب شود.

۲.۲ ایجاد توصیف با استفاده از شبکه‌های ژرف بازگردنده

یکی از بزرگترین معایب روش‌های ذکر شده در جهت ایجاد توصیف برای ویدئوها، استفاده از قالب‌های یکسان و به دست آوردن تعدادی قواعد معنایی ثابت (سه‌تایی یا چهارتایی‌های فعل، فاعل، مفعول و صحنه) برای ایجاد توصیف ویدئو به زبان‌طبیعی است. استفاده از این قالب‌های ثابت برای مجموعه کلمات بزرگتر مشکل می‌شود و همچنین جملات با قالب‌های از پیش تعیین شده و خشک، بسیار ساده‌تر از جملاتی خواهند بود که بتوانند به درستی ساختار پیچیده‌ی زبان‌طبیعی را ضبط و رعایت کنند.

در جهت رفع این مشکلات، نویسندگان [۶] با توجه به موفقیت روش‌های یادگیری ژرف در فعالیت‌های توصیف تصاویر ساده، به استفاده از

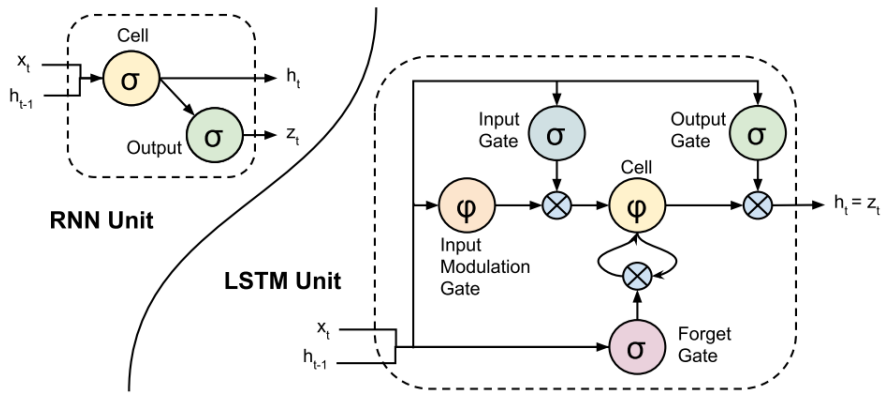
این روش‌ها روی آوردند. به طور خاص نویسندگان در این پژوهش با استفاده از ترکیبی از شبکه‌های ژرف پیچشی و نوعی خاص از شبکه‌های ژرف بازگردنده به نام LSTM[۲۷] مدلی را ارائه می‌کنند که با تغییرات کمی می‌تواند برای سه فعالیت بازشناسی کنش، توصیف تصویر و توصیف ویدئو استفاده شود.

۱.۲.۲ شبکه‌های ژرف بازگردنده

شبکه‌های بازگردنده قابلیت یادگیری پویایی‌های زمانی پیچیده را با نگاشت دنباله‌ی ورودی به دنباله‌ای از حالات‌نهان با رابطه‌ی $h_t = g(W_{xh}x_t + W_{hh}h_t + b_h)$ و در ادامه نگاشت این حالات نهان به خروجی را دارند. عملکرد این شبکه‌ها به این صورت است که هر ورودی تصویری v_t، از یک تابع غیر خطی استخراج ویژگی $\phi_v(v_t)$ که پارامتر آن V است، عبور می‌کند تا در نهایت، یک بردار ویژگی با طول ثابت، $\phi_t \in \Re^d$، از تصویر ورودی به دست آید. بعد از این‌که ورودی تصویری به یک بردار ویژگی با طول ثابت تبدیل شد، نوبت به بخش دوم مدل، یعنی شبکه‌ی بازگردنده می‌رسد. در کلی‌ترین حالت، یک مدل دنباله‌ای[۲۸] که پارامتر آن را ماتریس W مشخص می‌کند، نگاشتی از ورودی‌های x_t و حالت نهان زمان قبلی h_{t-1}، به یک خروجی در زمان فعلی z_t و حالت نهان جدید h_t به دست می‌دهد. سپس در قدم‌نهایی، برای به دست آوردن یک توزیع احتمالاتی مناسب روی کلمات خروجی محتمل در هر زمان، روی خروجی شبکه‌ی بازگردنده z_t یک تابع Softmax اعمال می‌شود.

$$P(y_t = c) = \frac{\exp(W_{zc}z_{t,c} + b_c)}{\sum_{c' \in C} \exp(W_{zc}z_{t,c} + b_c)}$$

با توجه به رابطه‌ی بالا، حدس‌های نهایی مدل (نزدیک به T، تعداد فریم‌های ویدئو)، حدس‌هایی هستند که توسط یک شبکه‌ی بسیار ژرف زده می‌شوند. ماتریس وزن‌های W نیز در این مدل، برای تمامی زمان‌ها مشترک است که این ویژگی، مدل را مجبور به یادگیری قواعدی کلی می‌کند که در میان زمان‌های متفاوت ثابت هستند که متقابلاً باعث کاهش بیش‌برارزش[۲۹] و کاهش پارامترهای شبکه خواهد شد. معماری کلی یک شبکه‌ی بازگردنده در سمت چپ شکل ۲.۳ قابل مشاهده است.



شکل ۲.۳: *مدل LSTM و RNN* [۶]

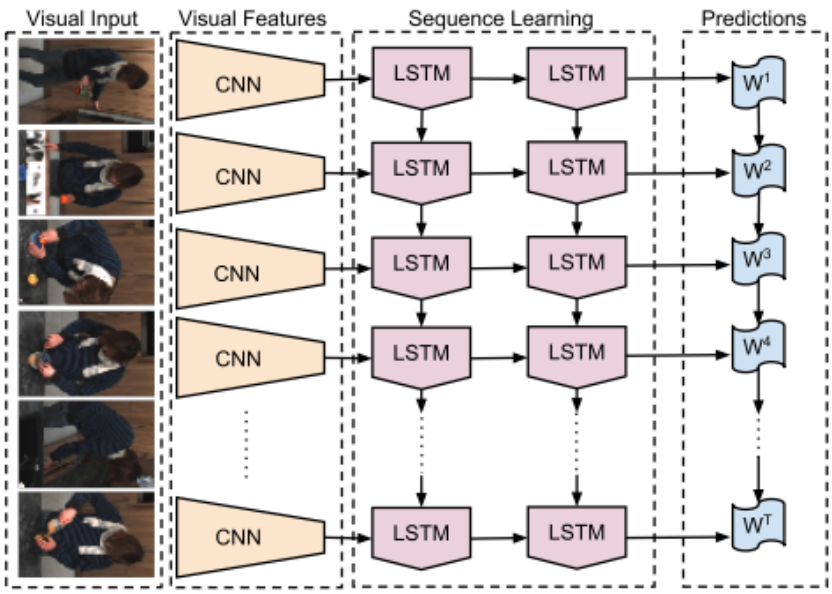
با این‌که شبکه‌های بازگردنده در زمینه‌هایی مانند بازشناسی گفتار و تولید متن موفق عمل کرده‌اند، آموزش‌دادن این دسته از مدل‌ها برای اینکه روی دنباله‌های طولانی پاسخ مناسبی دهند، کمی سخت است. در واقع در مراحل آموزش چنین مدل‌هایی با دو مشکل عمده مواجه می‌شویم، ناپدید شدن[۳۰] و یا بسیار بزرگ شدن[۳۱] مقادیر بردار گرادیان است، که به دلیل گذرانده‌شدن گرادیان‌ها از تعداد لایه‌های زیاد شبکه در زمان یادگیری است. برای حل این مشکل، مدلی بازگردنده به نام LSTM در پژوهش [۷] ارائه شده است. LSTMها با معماری خاص خود، به شبکه

اجازه می‌دهند تا در فاز آموزش، شبکه یادبگیرد که در چه زمانی حالت‌های نهان قبلی را فراموش کند و در چه زمانی حالت‌نهان فعلی را با اطلاعات جدید به روزرسانی کند. معماری یک سلول LSTM در قسمت راست شکل ۲.۳ قابل رویت است. با نمایش دادن تابع سیگموید با σ و تابع تانژانت هایپربولیک با ϕ روابط به روزرسانی سلول‌ها در LSTM به صورت زیر خواهد بود.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \qquad f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \qquad o_t = \sigma(W_{xo}x_t + w_{ho}h_{t-1} + b_o)$$
$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \qquad c_t = f_t * c_{t-1} + i_t * g_t \qquad h_t = o_t * \phi(c_t)$$

۲.۲.۲ روش ارائه شده

معماری پایه‌ی روش ارائه شده در شکل ۲.۴ قابل مشاهده است. نویسندگان در این مقاله از یک شبکه‌ی عصبی پیچشی برای استخراج ویژگی از تصاویر استفاده کرده‌اند. خروجی آخرین لایه‌های کاملا متصل شبکه‌ی پیچشی (لایه‌های fc_7 و fc_6)، به عنوان ورودی d بعدی x_t برای شبکه‌ی بازگردنده در نظر گرفته می‌شوند.



شکل ۲.۴: مدل $LRCN$ [۶]

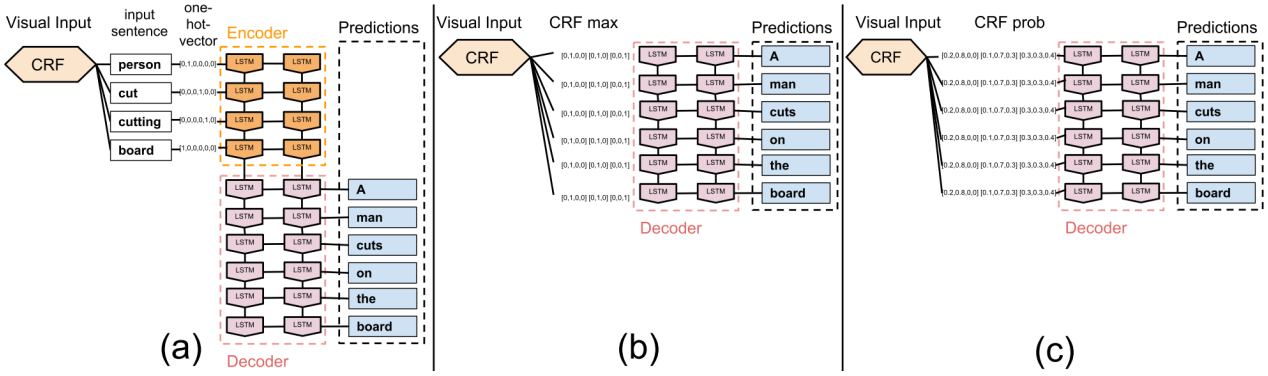
نویسندگان این پژوهش، با ارائه‌ی تعریفی از معماری پایه‌ی شبکه، اقدام به تعریف سه زیر مسئله که با معماری ارائه شده قابل حل است می‌کنند. مسائل تعریف شده شامل مسائل با ورودی دنباله‌ای و خروجی ثابت (تشخیص عمل در ویدئو)، ورودی ثابت و خروجی دنباله‌ای (توصیف تصویر) و در نهایت مسائل با ورودی و خروجی دنباله‌ای (توصیف ویدئو) هستند.

در حالت اول، از یک روش درهم‌آمیزی دیرهنگام برای به دست آوردن پاسخ نهایی شبکه از پاسخ‌های مبتنی بر زمان شبکه استفاده می‌شود. در حالت دوم که ورودی ثابت است و خروجی دنباله‌ای شکل، نویسندگان با تکرار کردن ورودی ثابت در هر زمان t مسئله را به حالت بعدی یعنی ورودی و خروجی دنباله‌ای تبدیل می‌کنند. در این حالت از یک مدل انکودر_دیکودر [۳۲] استفاده می‌شود. در این روش یک دنباله توسط انکودر به یک بردار با طول ثابت نگاشت می‌شود. سپس یک مدل دنباله‌ای دیگر که دیکودر نامیده می‌شود، بردار خروجی انکودر را به دنباله‌ای با طول متغیر

تبدیل می‌کند. در واقع در این معماری، سیستم دارای $T + T'$ گام زمانی است که در T گام ابتدای پردازش روی ورودی انجام می‌شود و از خروجی چشم‌پوشی می‌شود. سپس در T' گام بعدی، خروجی از مدل دیکودر گرفته می‌شود. با در نظر گرفتن معماری توضیح داده‌شده، پارامترهای V و W که به ترتیب مربوط به مدل تصویری و دنباله‌ای هستند، می‌توانند با بهینه‌سازی درست‌نمایی برچسب‌های خروجی y_t مشروط به ورودی و حالت‌نهان تا آن زمان، محاسبه شوند.

$$L(V, W) = -\log P_{V,W}(y_t | x_{1:t}, y_{1:t-1})$$

در این پژوهش برای ایجاد توصیف برای ویدئو از یک CRF [۳۳] استفاده شده است. دلیل این امر، نبود مجموعه‌دادگان مناسب و کافی برای آموزش مدلی بر مبنای شبکه‌های عصبی پیچشی است. CRF استفاده شده در این بخش از پژوهش با دریافت ویژگی‌های دست‌ساز از ویدئوی ورودی، سه‌تایی فعل، فاعل و مفعول را تولید می‌کند و به عنوان ورودی به انکودر وارد می‌کند. در نهایت سه معماری متفاوت (شکل ۲.۵) برای اتصال عناصر پیشنهاد شده است که با توجه به بررسی انجام شده مدل C بهترین دقت را در معیار BLEU-۴ دارا است. در این معماری انکودر حذف شده و بردار احتمال مرتبط با کلمات انتخاب شده توسط CRF در هر گام‌زمانی به عنوان ورودی دیکودر تکرار می‌شوند.



شکل ۲.۵: *نمونه‌های معماری پایه برای توصیف ویدئو*[۶]

پژوهش [۸] که بعد از این پژوهش انجام شد، دو مشکل اساسی آن را در زمینه‌ی توصیف ویدئو برطرف می‌کند. در واقع در این پژوهش با استفاده از شبکه‌ی پیچشی به جای CRF باعث می‌شود مدل قابلیت آموزش به صورت انتهابه‌انتها [۳۴] را به دست آورد و همچنین نیاز به تعیین ویژگی‌ها و نقش‌های معنایی به صورت دستی (که برای مجموعه‌لغات بزرگ بسیار مشکل است) از بین رود. در این مدل به آموزش مدلی پرداخته شده‌است که با توجه به ویژگی‌های تصویری V و کلمات قبلی تولید شده توسط مدل، $S_{t-1}, \cdots, S_1$ ، احتمال درست‌نمایی جمله S را بیشینه کند.

$$\theta^* = \arg\max_{\theta} \sum_{V,S} \log p(S|V;\theta) \rightarrow \log p(S|V) = \sum_{t=0}^{N} \log p(S_{w_t}|V, S_{w_1}, \cdots, S_{w_{t-1}}) \qquad (۴)$$

در این مدل از یک شبکه‌ی پیچشی که در کتابخانه‌ی Caffe [۹] موجود است (شبکه‌ی عصبی AlexNet [۱۰]) استفاده شده است که روی ۲.۱ میلیون تصویر از مجموعه‌داده‌ی ILSVRC-۲۰۱۲ [۱۱] پیش‌آموزش داده شده است. از هر ۱۰ فریم ویدئو یک فریم نمونه‌گیری شده است و از خروجی لایه‌ی آخر کاملا متصل شبکه پیچشی (fc_7) به عنوان بردار ۴۰۹۶ بعدی نمایش‌گر ویدئو استفاده شده است. ورودی انکودر برداری است که از کنار هم گذاشتن بردار ویژگی‌های ویدئو و همچنین بردار ویژگی کلمه‌ی ایجاد شده‌ی قبلی به دست می‌آید. ورودی دیکودر نیز خروجی انکودر در هر گام زمانی است. در نهایت با اعمال تابع Softmax روی خروجی دیکودر، کلمه با بالاترین احتمال را به عنوان خروجی انتخاب می‌کنیم. با اعمال تغییرات ذکر شده، بهبود ۲ درصد در معیار BLEU-۴ نسبت به پژوهش قبلی مشاهده می‌شود.

۲.۳ بهبود توصیف متنی بر شبکه‌های بازگردنده با استفاده از دانش متن‌کاوی شده

در ادامه‌ی پژوهش قبلی، در پژوهش [۱۲] نویسندگان مدل زبانی ارائه شده را به شکلی تغییر می‌دهند تا از دانش داده‌کاوی شده از متون موجود در اینترنت نیز بهره ببرد. برای مخلوط کردن دانش موجود از متون با مدل‌زبانی ارائه شده در پژوهش قبلی، از سه روش ادغام سریع [۳۵] ، ادغام دیرهنگام [۳۶] و ادغام ژرف [۳۷] استفاده شده‌است. در روش ادغام سریع، بخش‌هایی از شبکه که زبان‌طبیعی را مدل می‌کنند، روی متون موجود، پیش‌آموزش داده می‌شوند و در نهایت روی متون مرتبط با مجموعه‌ی داده، تنظیم دقیق [۳۸] می‌شوند. دو روش دیگر از یک LSTM مجزا، که روی متون موجود اینترنت، از پیش آموزش داده شده‌اند استفاده می‌کنند. همچنین در این پژوهش به جای نمایش کلمات به صورت One-Hot (نحوه‌ای از نمایش که هر کلمه به صورت برداری به طول تمامی مجموعه لغات نمایش داده می‌شود که تنها در نمایه‌ی مرتبط با کلمه‌ی فعلی مقدار آن برابر با یک و در دیگر نقاط مقدار آن صفر است)، از روش‌های بازنمایش آماری کلمات از پیش آموزش‌دیده شده، استفاده می‌شود. مزیت استفاده از این دست نمایش‌ها برای کلمات، کاهش قابل توجه ابعاد داده‌های زبانی و همچنین استفاده از مفاهیم موجود در کلمات است.

۱.۲.۳ ادغام دیرهنگام

استفاده‌ی روش ادغام دیرهنگام از شبکه‌های پیش‌آموزش دیده، مشابه استفاده‌ی روش‌های ترجمه‌ی ماشینی زبان، در دیکودر است. در هر مرحله از ایجاد جمله، مدل توصیف ویدئو یک توزیع روی کلمات ایجاد می‌کند. خروجی نهایی با در نظرگرفتن میانگین وزن‌دار مجموع امتیازات داده‌شده توسط مدل‌زبانی و مدل موجود توصیف ویدئو ایجاد می‌شود. به طور دقیق‌تر، با در نظر گرفتن y_t به عنوان خروجی مدل در زمان t، P_{VM} و P_{LM} به عنوان توزیع‌هایی که به ترتیب مدل‌های توصیف ویدئو و زبان تولید می‌کنند، برای تمامی لغات خروجی می‌توان تابع امتیاز زیر را در نظر گرفت:

$$p(y_t = y') = \alpha P_{VM}(y_t = y') + (1 - \alpha)P_{LM}(y_t = y')$$

۲.۲.۳ ادغام ژرف

در روش ادغام ژرف، مدل‌زبانی جداگانه به صورت ژرف‌تری با مدل ترکیب می‌شود (برخلاف ادغام دیرهنگام که تنها در انتهای مدل از مدل‌زبانی دارای دانش استفاده شده بود). نحوه‌ی انجام این‌کار بدین صورت است که حالت‌نهان مدل زبانی مجزا (h_t^{LM}) LSTM در کنار حالت‌نهان مدل توصیف ویدئو موجود h_t^{VM} قرار داده می‌شود و از بردار حاصله‌ی جدید برای پیش‌بینی کلمه‌ی فعلی استفاده می‌شود. احتمال انتخاب یک کلمه‌ی خاص در زمان t در این مدل متناسب با رابطه‌ی زیر است:

$$p(y_t|y_{<t}, v) \propto \exp(Wf(h_t^{VM}, h_t^{LM}) + b)$$

در این رابطه، b و v به ترتیب بایاس و بردار ویژگی‌های تصویر را نمایش می‌دهند. برای جلوگیری از بیش‌برازش مدل به وزن‌های پیش‌آموزش داده شده‌ی مدل زبانی، وزن‌های این مدل در مرحله‌ی آموزش ثابت نگه‌داشته می‌شوند و تنها وزن‌های شبکه‌ی توصیف ویدئو موجود به روزرسانی می‌شوند تا از وزن‌های مدل‌زبانی نیز در به دست آمدن وزن‌های نهایی استفاده شده باشد.

۳.۲.۳ نمایش توزیعی کلمات

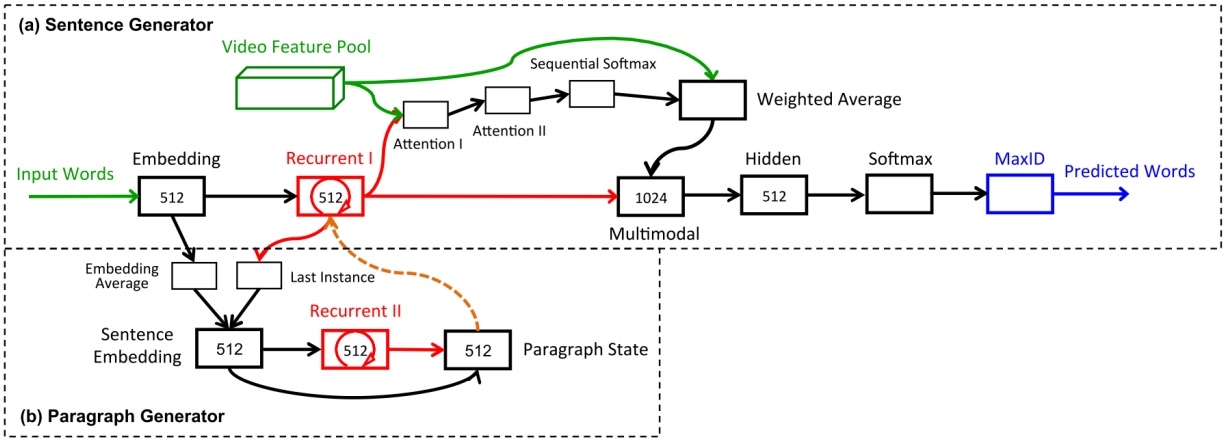
شبکه‌ی توصیف ویدئو موجود، مانند بسیاری از مدل‌های توصیف ویدئو و تصویر دیگر، از نمایش One-Hot برای کلمات استفاده می‌کند. در این حالت در زمان آموزش، بخش مرتبط با زبان شبکه، با استفاده از داده‌های آموزش، یک نمایش با ابعاد کمتر (۵۰۰ بعدی) از کلمات را یاد می‌گیرد. اما این یادگیری از داده‌هایی محدود و دارای نوفه موجود در مجموعه‌دادگان رخ می‌دهد که بهینه نیست. روش‌های زیادی هستند که برای

حل این مشکل، یک مدل همه‌منظوره روی متون بسیار زیاد آموزش می‌دهند تا بتوان از آنها در مدل‌های دیگر استفاده کرد. در این پژوهش از نمایش بردارهای GloVe [۱۳] استفاده شده است که کلمات را در قالب یک بردار ۳۰۰ بعدی نمایش می‌دهد. در نهایت هم با توجه به غیر محدب بودن تابع هزینه‌ی شبکه، از میانگین خروجی چندین نمونه‌ی متفاوت [۳۹] از مدل، به عنوان خروجی نهایی استفاده شده است. اعمال تغییرات آورده شده در این پژوهش، منجر به افزایش دقت ۵ درصدی در معیار BLEU-۴ و ۴.۲ درصدی در معیار Meteor در مجموعه داده‌گان یوتیوب شده است.

۲.۴ ایجاد توصیف با استفاده از شبکه‌های ژرف بازگردنده چند سطحی

از نقاط ضعف روش‌های پیشین می‌توان به دو مورد اصلی اشاره کرد. مورد اول این‌که این روش‌ها عموما برای بهبود جملات روی مدل زبانی شبکه تمرکز می‌کردند و به ویژگی‌های تصویری استخراج شده توسط شبکه‌ی پیچشی بسنده می‌کردند. مورد دوم این‌که در این روش‌ها، تنها یک جمله به عنوان خروجی مدل ایجاد می‌شد که گاهاً جهت توضیح یک ویدئوی پرمحتوا، به هیچ عنوان کافی نیست. در پژوهش [۱۴] نویسندگان با ارائه‌ی یک معماری پیچیده‌تر، سعی در رفع مشکلات اشاره شده دارند.

در این پژوهش، یک شبکه‌ی عصبی ژرف بازگردنده چندسطحی برای توصیف ویدئوهای طولانی با چندین پاراگراف ارائه شده است. با توجه به اینکه در هر پاراگراف، جملات باید مرتبط با یکدیگر ایجاد شوند، شبکه‌ی چندسطحی از دو ایجاد کننده [۴۰] ، یک ایجادکننده‌ی جمله و یک ایجادکننده‌ی پاراگراف، تشکیل می‌شود. در سطح پایین، وظیفه‌ی ایجادکننده‌ی جمله تولید جملاتی است که یک بازه‌ی زمانی کوتاه و مشخص در ویدئو را توصیف می‌کنند. در این پژوهش از مکانیزم توجه [۴۱] در هر دو حوزه‌ی زمان [۴۲] و مکان [۴۳]، جهت انتخاب دقیق‌تر المان‌های تصویری مرتبط با جمله‌ی در حال تولید، استفاده شده است. در سطح بالاتر، نمایشی از جملات ایجاد شده را به عنوان ورودی دریافت می‌کند و وضعیت پاراگراف با به عنوان خروجی ایجاد می‌کند، این خروجی (وضعیت پاراگراف) به عنوان حالت اولیه به شبکه‌ی ایجادکننده‌ی جملات داده می‌شود. هرکدام از این شبکه‌ها، یک شبکه‌ی بازگردنده خاص به نام GRU [۱۵] هستند که مدل ساده شده‌ی LSTMها هستند که در قسمت قبل معرفی شدند. معماری کلی شبکه‌ی نهایی در شکل ۲.۶ قابل مشاهده است.



شکل ۲.۶: *شبکه‌ی بازگردنده چند سطحی برای توصیف ویدئو*

۱.۲.۴ ایجادکننده‌ی جمله

این شبکه در هر گام زمانی با ورود یک بردار One-Hot از کلمه، با استفاده از یک جدول تعبیه [۴۴] ، بردار ورودی را به یک بردار فشرده‌تر (۵۱۲ بعد) تبدیل می‌کند. سپس مانند روش‌های پیشین، بردار فشرده‌ی به دست آمده وارد لایه‌ی بازگردنده ۱، شبکه‌ی انکودر می‌شود. تفاوت این شبکه با شبکه‌ی مدل‌های قبلی در استفاده از GRU به جای LSTM است. همچنین به عنوان تابع فعال‌سازی از تابع ReLu [۴۵] به جای تابع سیگموید استفاده شده است. خروجی شبکه‌ی ۱، به لایه‌های توجه [۴۶] داده می‌شوند تا ضرایب توجه برای ویژگی‌های تصویر به دست آیند. ضرایب توجه، با در نظر گرفتن یک توزیع احتمال روی ویژگی‌های هر فریم در هر زمان به دست می‌آید. اگر ویژگی‌های موجود در ویدئو رو با $v_{KM}, \cdots, v_1$ نشان‌دهیم که در آن M طول ویدئو و K تعداد وصله‌ها [۴۷] روی هر فریم است. هدف، به دست آوردن مجموعه‌ای از ضرایب $\beta^t_{KM}, \cdots, \beta^t_1$ در هر زمان، به طوری که $\sum_{m=1}^{KM} \beta^t_i = 1$ است. برای این‌کار ابتدا امتیاز توجه برای هر فریم m محاسبه می‌شود.

$$q^t_m = W^T \phi(W_q v_m + U_q h^{t-1} + b_q)$$

که در آن، ϕ تابع $stanh$ در نظر گرفته شده است. سپس با گرفتن Softmax از امتیازهای توجه به دست آمده، امتیاز توجه برای هر ویژگی به دست می‌آید:

$$\beta^t_m = \frac{\exp(q^t_m)}{\sum_{m'=1}^{KM} \exp(q^t_{m'})}$$

در نهایت بردار ویژگی‌نهایی در لایه‌ی میانگین وزن‌دار، با ضرب ضرایب به دست آمده از لایه‌ی توجه در بردار ویژگی‌های تصویر به دست می‌آید: $u^t = \sum_{m=1}^{KM}$. در ادامه خروجی لایه‌ی میانگین وزن‌دار ویژگی‌های تصویر (۱۰۲۴ بعدی)، به همراه خروجی لایه‌ی بازگردنده ۱، وارد لایه‌ی چندحالته می‌شوند. در این لایه، عناصر تصویر و متن بایکدیگر مخلوط می‌شوند. برای حالتی که ویژگی‌های تصویری دو کاناله باشند (u_a و u_o)، خروجی این لایه (m^t) برابر با $\phi(W_{m,o} u^t_o + W_{m,a} u^t_a + U_m h^t + b_m)$ است.

۲.۲.۴ ایجادکننده‌ی پاراگراف

ایجادکننده‌ی جمله در مرحله‌ی قبل، عملیات مرتبط با یک‌جمله را انجام می‌دهد. حالت اولیه‌ی لایه‌ی بازگردنده ۱ برای اولین جمله برابر با صفر در نظر گرفته می‌شود، اما بعد از ایجاد اولین جمله، حالت اولیه این لایه، با توجه به معنی مفهومی جملات قبل، توسط ایجادکننده‌ی پاراگراف تنظیم می‌شود. در مراحل ایجاد یک جمله، میانگین تعبیه‌شده‌های تمامی کلمات جمله توسط لایه‌ی میانگین تعبیه [۴۸] نگه‌داشته می‌شود که در کنار آخرین وضعیت لایه‌ی بازگردنده ۱ به عنوان حالت جمله، به عنوان ورودی لایه‌ی تعبیه‌ی جمله [۴۹] در نظر گرفته می‌شوند. خروجی این لایه، لایه‌ی بازگردنده ۲ است. این لایه زمانی لایه‌ی بعدی، که حالت پاراگراف است، را به روزرسانی می‌کند که نشانگر انتهای جمله [۵۰] توسط ایجاد کننده‌ی جمله تولید شده باشد.

۳.۲.۴ آموزش

با در نظر گرفتن $s_{1:t-1}$ به عنوان جملات قبلی موجود در پاراگراف و $w^n_{1:t-1}$ به عنوان کلمات قبلی موجود در جمله‌ی nم، تابع درست‌نمایی ایجاد یک کلمه به صورت $p(w^n_t | s_{1:n-1}, w^n_{1:t-1}, V)$ خواهد بود. با توجه به این تعریف، تابع هزینه ایجاد یک پاراگراف برابر است با:

$$L(s_{1:N}|V) = -\frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \log P(w^n_t | s_{1:n-1}, w^n_{1:t-1}, V)}{\sum_{n=1}^{N} T_n}$$

در این رابطه، T_n و N به ترتیب تعداد کلمات موجود در جلمه‌ی nم و تعداد جملات موجود در پاراگراف است. با در نظر گرفتن پاراگراف‌های متعدد برای هر داده‌ی آموزش، تابع هزینه‌ی نهایی به صورت زیر به دست می‌آید:

$$L = -\frac{\sum_{y=۱}^{Y} \sum_{n=۱}^{N_y} \sum_{t=۱}^{T_n^y} \log P(w_t^{n,y} | s_{۱:n-۱}^y, w_{۱:t-۱}^{n,y}, V_y)}{\sum_{y=۱}^{Y} \sum_{n=۱}^{N_y} T_n^y}$$

معماری ارائه شده در این پژوهش منجر به بهبود دقت در معیارهای *BLEU* به میزان ۵ درصد و در معیار *METEOR* به میزان ۲ درصد می‌شود.

۲.۵ سایر روش‌ها

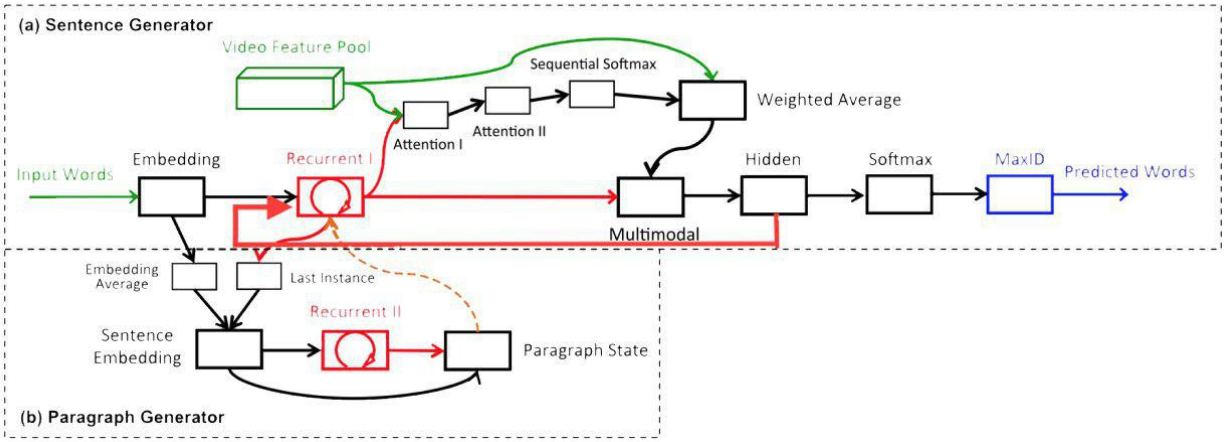
روش‌های متعدد دیگری برای توصیف ویدئو ارائه شده‌اند که فصل مشترک عموم این روش‌ها، استفاده از شبکه‌های پیچشی برای استخراج ویژگی و استفاده از شبکه‌های بازگردنده برای مدل‌کردن زبان طبیعی و ایجاد جملات است. در [۱۶] نویسندگان، برخلاف پژوهش ارائه شده در [۶] از یک معماری انکودر–دیکودر ساده برای مدل زبانی استفاده می‌کنند، اما با تغییر مدل مربوط به ویژگی‌های تصویری و افزودن مکانیزم توجه و تعریف نوعی خاص از شبکه‌های بازگردنده، به دقتی بهتر از پژوهش ارائه شده در بخش ۲.۱ و تقریبا معادل با دقت پژوهش ارائه شده در بخش ۲.۲ می‌رسند.

در پژوهش [۱۷] نویسندگان، با استفاده از دانش موجود در مدل زبانی پیش‌آموزش داده‌شده روی متون خارج از دامنه، راه‌حلی برای توصیف اشیا جدید که در زمان آزمون دیده‌نشده‌اند ارائه می‌کنند. در این پژوهش با استفاده از ماژول‌های موجود، یک لایه‌ی جدید چندحالته برای ترکیب دانش زبان‌طبیعی موجود و ویژگی‌های تصویری و انتقال دانش [۵۱] ارائه شده است. در این مدل، دو شبکه‌ی مجزا روی مجموعه‌دادگان مجزا آموزش داده می‌شوند. برای مدل‌کردن زبان از یک شبکه‌ی بازگردنده و برای استخراج ویژگی‌های تصویری از یک شبکه‌ی پیچشی استفاده می‌شود. هر دو شبکه‌ی ذکر شده روی مجموعه‌دادگان جدا، شامل اشیایی که در مجموعه‌دادگان اصلی حضور ندارند، پیش‌آموزش داده شده‌اند. در نهایت تمامی شبکه روی مجموعه‌ی دادگان اصلی آموزش داده می‌شود. در این شبکه وظیفه‌ی لایه‌ی چندحالته مخلوط کردن ویژگی‌های تصویری و زبانی و به دست آوردن یک بازنمایش مناسب از اشیایی که در تصویر هستند و در مجموعه‌ی دادگان حضور ندارند است. برای این‌کار یک تابع هزینه برای این لایه در نظر گرفته شده است که پیش‌بینی مدل‌های مجزا و مدل توصیف تصویر اصلی مدل را با یکدیگر مخلوط می‌کند.

۳ روش ارائه شده

یکی از بهترین روش‌های ارائه شده برای توصیف ویدئو، پژوهش [۱۴] است که در بخش ۲.۴ به آن اشاره شد. در این روش از سلسله‌مراتبی از شبکه‌های بازگردنده در کنار توجه و یادگیری چندحالته استفاده می‌شود. یکی از نقاط ضعف این روش اما استفاده‌ی حداقلی از ویژگی‌های تصویری ویدئو است. ویدئوها به عنوان یک منبع تصویری زمانی–مکانی، حاوی اطلاعات بسیاری هستند که استفاده‌ی مناسب از آنها می‌تواند به افزایش دقت قابل توجه مدل منجر شود. در این روش، از خروجی لایه‌ی بازگردنده‌ی ایجادکننده جمله، برای به دست آوردن ویژگی‌های مهم تصویر (توجه روی ویژگی‌های تصویری) استفاده می‌شود و سپس با به دست آمدن ویژگی‌های تصویری مرتبط با جمله‌ی درحال ایجاد، این ویژگی‌ها با دانش‌زبانی مدل در لایه‌ی چندحالته ترکیب می‌شوند تا کلمه‌ی بعدی خروجی را ایجاد کنند. روش پیشنهادی ما در این گزارش، بر مبنای پژوهش [۱۴] است. برای رفع مشکل اشاره شده، ما یک چرخه‌ی بازخورد از خروجی مدل با توجه به ویژگی‌های تصویری به شبکه‌ی ایجاد کننده‌ی جمله وارد می‌کنیم. بدین صورت سعی می‌کنیم تا مدل در زمان آموزش، یادگیرد تا از ویژگی‌های تصویری به دست آمده در ایجاد کلمات بعدی استفاده کند (تنها از کلمات ایجاد شده تا به حال برای ایجاد کلمه‌ی بعدی استفاده نکنیم). در واقع روش‌پیشنهادی اولیه‌ی ما بر پایه‌ی افزایش در هم تنیدگی ویژگی‌های تصویری و متنی بنا شده است. برای بررسی این مدعا، ساختار ارائه شده در [۱۴] را بدین نحو تغییر می‌دهیم که خروجی لایه‌ی چندحالته به

ورودی شبکه‌ی بازگشتی ایجادکننده‌ی جمله وارد می‌شود. این تغییر باعث افزایش ابعاد شبکه‌های ایجادکننده‌ی جمله، پاراگراف و تعبیه‌ی آخرین جمله می‌شود. این افزایش ابعاد متعاقباً منجر به افزایش پارامترهای لازم برای آموزش شبکه می‌شود که برای جلوگیری از بیش‌برازش، از روش Dropout استفاده می‌کنیم در کنار بهره‌گیری از منظم‌سازهای L۱ و L۲ استفاده می‌کنیم. در نهایت معماری پیشنهادی به صورت زیر است:



شکل ۳.۷: *نمونه‌های معماری ارائه‌شده برای توصیف ویدئو*

۴ کارهای آتی

یکی از روش‌هایی که اخیرا در زمینه‌ی پردازش ویدئو مورد استفاده قرار گرفته و منجر به بهبودهای قابل توجهی در دقت این روش‌ها شده است، استفاده از شبکه‌های بازگردنده‌ی دو طرفه [۵۲] است. در این شبکه‌ها، علاوه بر اینکه از فریم‌های ویدئو به ترتیب زمانی برای آموزش شبکه استفاده می‌شود، از فریم‌های آینده‌ی ویدئو نیز برای آموزش یک شبکه‌ی بازگردنده‌ی در جهت مخالف شبکه‌ی اولیه استفاده می‌شود. در واقع هدف این دست از روش‌ها، استفاده از رخدادهای آینده‌ی ویدئو در ایجاد توصیف برای ویدئو است. از کارهای آینده‌ی این پژوهش، بررسی نحوه‌ی ادغام این دست از شبکه‌ها در معماری مدنظر و آموزش‌دادن آنها است.

همچنین یکی از نقاط ضعف شبکه‌های بازگردنده، عدم توانایی در مدل‌کردن دنباله‌های طولانی است. علت رخداد این امر نیز وجود تنها یک بردار حالت در این شبکه‌ها است. اخیرا یک دسته از شبکه‌های عصبی به نام شبکه‌های حافظه [۵۳] [۱۸] است که هدف آنها، پاسخ‌دهی به وابستگی‌های زمانی پیچیده و با فاصله‌ی زیاد در داده‌ها است. یکی دیگر از کارهای آتی این پژوهش، بررسی به‌کارگیری این شبکه‌ها در کنار شبکه‌های بازگردنده‌ی موجود برای ایجاد جملات زبان‌طبیعی است.

در نهایت، خلاصه‌ای از مراحل و میزان پیشرفت این پژوهش در جدول ۴.۱ آمده است.

۵ جمع‌بندی

در این گزارش مسئله‌ی توصیف ویدئو، نسخ مختلفی از مسائل موجود و یک چارچوب کلی برای حل این دست از مسائل، با معرفی روش‌های ارائه شده پیشین و بررسی مزایا و معایب هرکدام، معرفی شد.با توجه به پیچیدگی ذاتی روش‌های پردازش ویدئو و نیاز به وجود سخت‌افزارهای قدرتمند، اکثر روش‌های ارائه شده که به پاسخ‌های مناسبی رسیده‌اند و در این گزارش نیز بررسی شدند، مختص یک یا دو سال اخیر هستند. در

جدول ۴.۱: *جدول زمان‌بندی*

عنوان فعالیت	مدت زمان لازم	درصد پیشرفت	زمان اتمام
مطالعه و بررسی روش‌های موجود و راه‌کارهای قابل استفاده	۳ ماه	۱۰۰	شهریور ۹۴
آزمایش روش‌های موجود مقایسه آن‌ها	۲ ماه	۱۰۰	آبان ۹۴
بررسی و یافتن کاستی‌های روش‌های موجود	۱ ماه	۷۵	آبان ۹۴
پیشنهاد و پیاده‌سازی و ارزیابی روش جدید	۴ ماه	۵	اسفند ۹۴
ارزیابی روش نهایی و مقایسه با روش‌های دیگر	۲ ماه	۰	اردیبهشت ۹۵
نگارش پایان‌نامه	۲ ماه	۰	تیر ۹۵

این گزارش سعی شد روش‌های گوناگون حل این مسئله با توجه به ابزارهای مورد استفاده در آن (شبکه‌های بازگشتی، کانولووشنال ، مکانیزم تمرکز و روش‌های سنتی) مورد تقسیم‌بندی قرار بگیرند. در بخش روش‌ارائه‌شده ..... و روشی برای بهبود توصیف ایجاد شده برای ویدئوها ارائه دادیم. در نهایت راه‌کاری آتی و جدول زمان‌بندی ادامه‌ی کار در بخش کارهای آتی ارائه شد.

مراجع

[1] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge," *NAACL HLT Workshop on Vision and Language*, pp.19–10, .2013

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, .2008 CVPR .2008 IEEE Computer Society Conference on*, pp.8–1, IEEE, .2008

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, .2008 CVPR .2008 IEEE Conference on*, pp.8–1, IEEE, .2008

[4] A. Pauls and D. Klein, "Faster and smaller n-gram language models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp.267–258, Association for Computational Linguistics, .2011

[5] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild," *Coling*, pp.1227–1218, .2014

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, U. T. Austin, U. Lowell, and U. C. Berkeley, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *Cvpr*, .2015

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol.9, no.8, pp.1780–1735, .1997

[8] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, U. C. Berkeley, M. Rohrbach, U. C. Berkeley, R. Mooney, and K. Saenko, "Translating Videos to Natural Language Using Deep Recurrent Neural Networks," *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'15)*, no.June, pp.1504–1494, .2015

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp.678–675, ACM, .2014

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp.1105–1097, .2012

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol.115, no.3, pp.252–211, .2015

[12] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving LSTM-based Video Description with Linguistic Knowledge Mined from Text," *Arxiv*, .2016

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation.," in *EMNLP*, vol.14, pp.43–1532, .2014

[14] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video Paragraph Captioning using Hierarchical Recurrent Neural Networks," *Cvpr*, .2016

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:14123555.*, .2014

[16] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," *arXiv:151103476. [cs]*, pp.1038–1029, .2015

[17] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," *Cvpr*, pp.10–1, .2016

[18] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:14103916.*, .2014

۶ واژه‌نامه

[۱] Deep Video Captioning
[۲] NaturalLanguage Processing
[۳] Computer Vision
[۴] Open-domain
[۵] Youtube
[۶] Multimodal
[۷] Convolutional Neural Networks
[۸] Recurrent Neural Networks
[۹] Maximum Likelihood Estimation
[۱۰] Fully Connected Network
[۱۱] Support Vector Machine
[۱۲] Decision Tree
[۱۳] Recurrent Neural Networks
[۱۴] Content Planning Stage
[۱۵] Surface Realization
[۱۶] Noise
[۱۷] Histogram Of Gradients
[۱۸] Histogram Of Optical Flow
[۱۹] Spatio-temporal interest points
[۲۰] Bag of Words
[۲۱] Scene Detection
[۲۲] Structural Similarity Index
[۲۳] Scale Invariant Feature Transform
[۲۴] Local Binary Patterns
[۲۵] Factor Graph
[۲۶] Maximum a posteriori
[۲۷] Long Short Term Memory
[۲۸] Sequence Model
[۲۹] Overfitting
[۳۰] Vanishing Gradients
[۳۱] Exploding Gradients
[۳۲] Encoder-Decoder
[۳۳] Conditional Random Field
[۳۴] End-to-End
[۳۵] Early Fusion
[۳۶] Late Fusion
[۳۷] Deep Fusion
[۳۸] Finetune
[۳۹] Ensemble
[۴۰] Generator
[۴۱] Attention

[۴۲] Temporal Attention
[۴۳] Spatial Attention
[۴۴] Embedding
[۴۵] Rectified Linear Unit
[۴۶] Attention Layers
[۴۷] Patch
[۴۸] Embedding Average Layer
[۴۹] Sentence Embedding
[۵۰] End Of Sentence، EOS
[۵۱] Transfer Learning
[۵۲] Bi-directional LSTM، BiLSTM
[۵۳] Memory Networks