دانشکده مهندسی

کامپیوتر

شبکه عصبی LSTM در کاربرد های داده های ویدئوی

محقق:

مهندس دانیال پهلوان مصوری

استاد راهنما

جناب دکتر مجتبی روحانی

تابستان 99

به یاد او که بالاتر از تفکر و تعقل است

فهرست مطالب

مقدمه

کلمات کلیدی :

تعریف موضوع

اهمیت و کاربردها:

روشهای مختلف حل و داده ها :

و...(پاراگراف آخر معرفی فصل بعدی هست)

همانطور که میدانید با ایجاد شبکه های اجتماعی و فضاهای اشتراک گذاشتن فایل ها موجب شده که داده های تصویری و ویدیویی زیادی در بستر شبکه داشته باشیم و حتی ممکن هست داده های تصویر و ویدیویی زیادی در بستر کامپیوتری که داریم استفاده می کنیم داشته باشیم ولی نکته مهم این داده ها این هست که برخلاف انسان ، در کامپیوترهای کاربران کمتر دیده می شود اطلاعات خاصی از این داده ها بیرون کشید و این داده ها فقط جهت نمایش به یک موجود هوشمند بنام انسان استفاده می شود ولی می توان این داده ها رو استفاده دیگر نیز کرد و می توان اطلاعات از این داده های تصویری و ویدیویی گرفت بطور مثال حالت و احساس موجود داخل تصویر چگونه هست و آیا خشمگین هست و یا خوشحال یا اینکه در این ویدیو چه اشیایی موجود هست بطور مثال یک میز قهوه ای و یک انسان و این انسان چه اقدامی دارد می کند بطور مثال درحال خواندن روزنامه و طبق رفتارهایی که از این انسان داشتیم حرکت بعدی چه خواهد بود .
حالا کاربرد این موارد بسیار زیاد هست و از جمله از آن ها می توان به استفاده دولت ها از این نتایج جهت خنثی کردن تهدید ها کرد و بطور مثال در سطح شهر دوربین های نظارتی زیادی وجود دارد و نیازمند اشخاصی هستیم که عمل مانیتورینگ انجام دهند اما انسان بخاطر محدود بودن پردازش که داره در شرایط خاص موجب میشه که نیروی انسانی زیادی تلف بشه و هرچقدر اطلاعات موجود در این ویدیو ها بیشتر باشد و یا تعداد این تولید کننده های فایل های ویدیویی بیشتر باشد موجب میشه که به نیروی انسانی زیادی نیاز داشته باشیم که در بعضی شرایط نیروی انسانی نیز برای این موارد کم میاوریم و سرعت پردازش بسیار طولانی می شود و برای اینکه این عبارت کامل جا بیفتد بیایم یک مثالی بزنیم . فک کنید در سطح شهر همانطور که قبل گفتیم دوربین های زیادی هست و جمعیت کشور هم مثل کشور چین زیاد باشد و حالا اگر انسان را برای پردازش قرار دهیم بیشتر داده های حیاتی از بین می رود چون نیروی انسانی ما نمی تواند تمام اطلاعات موجود رو در همان لحظه درک کند و نکته دوم اینکه سرعت پردازش بسیار پایین می آید بطور مثال اگر کاربر بخواهد داده های گذشته ضبط شده هم نگاه کند باید تمام فریم ها رو تماشا کند یا بیشتر آن ها را در زمانی که هر فریم طی می کند تماشا کند و این کار بسیار وقت گیر هست اما با داشتن تجهیزات که دارای چند هسته موازی هستند می شود چندین ویدیو را همزمان پردازش کنیم بدون اینکه زمانی صرف نمایش تک تک فریم ها صرف کنیم .

نکته ای که خیلی حیاتی هست و قبلا اشاره شده است حجم داده های ویدیویی در دنیای امروزی ما هست و با وجود شبکه های اجتماعی ما داده های زیادی رو در دسترس داریم و این داده ها انقدر زیاد هستند که حجمشان از ساعات عمر انسان ها نیز پیشی میگیرند و ما اینجا دیگر نمی توانیم به هیچ وجه نیروی انسانی استفاده کنیم چون میزان داده های پردازشی توسط انسان با میزان داده های تولیدی در هر روز با هم یکسان نیست و یک سر ریزی دارد و نیروی انسانی نمی تواند تمام این داده ها با فرض

اینکه مشکلی در پردازش نداشته باشیم بتواند رسیدگی کند و ما بخاطر این نیازها و نیازهای دیگه به اتوماسیون کردن کارها و استفاده از ابزارهای پردازشی میپردازیم .

بطور مثال در تحقیقی یک روش استخراج ویژگی از ویدیو داریم بر مبنای hvnLBP-TOP که در آن این ویژگی ها رو برای آنالیز احساسی مبتنی بر ویدیو استفاده میشه و در همین روش چون میزان ابعاد یا ویژگی هامون بسیار زیاد هست و توان پردازشی زیادی از ما میگیره پس مجبوریم که با روش هایی ویژگی های کم ارزش تر رو کم کنیم و به عبارتی میزان ابعاد مسئله را کاهش دهیم و می توان از روش هایی همچون تحلیل اجزای اصلی (PCA) نام برد و در کنارش نیز می تواند از lstm دو طرفه (Bi-LSTM) نیز استفاده کرد تا در حد امکان این ویژگی ها رو کاهش داد تا بتوان بهتر عملیات دسته بندی رو انجام داد و طبق این تحقیق میزان دقت در عملیات شناسایی در دیتاست MOUD به میزان ۷۱/۱ ٪ بوده و در دیتاست CMU-MOSI میزان دقت ۶۳/۹ ٪ بوده است .

در تحقیق دیگر هدف آن پیشبینی عمل بر مبنای ویدئو بوده است و چالش بزرگی که داشته این هست که برای ما انسان ها هم شاید اتفاق بیفته قضاوت اشتباه است و بطور مثال در صحنه اول شاید به اشتباه پیشبینی کنیم که جرم می خواد رخ دهد اما اگر ویدئو رو بصورت کامل ببینیم این اتفاق نیفتد و چالش بعدی تغییرات درون کلاس موجب سردگرمی پیشبینی کننده میشه و در این روش یک مدل از شبکه LSTM که دارای حافظه جداگونه هست معرفی شده بنام mem-LSTM که بتوان در لحظات اول این رسانه ویدئویی عملیات که می خواد اتفاق بیفتد پیشبینی کرد و برای پیشبینی از مثال های پیشبینی سخت استفاده شده تا کارایی الگوریتم رو بهتر بررسی بشه و در این روش از convolution neural network (CNN) در کنار LSTM استفاده شده و دلیل اینکه از LSTM با حافظه جدا یا حافظه دار استفاده شده این است که نمونه های سخت چالش بر انگیز مثال ها رو در حافظه خود داشته باشد و ازش در پیش بینی های بعدی استفاده کند و این کار باعث شده که علاوه بر اینکه در مراحل اول خوب کار کند ، در مراحلی که هیچ وجه مثالی در حافظه خود مانند این ندارد بخوبی برخورد کند در این این LSTM که وجود دارد بصورت دو طرفه می باشد که فریم های بعدی در لایه های بعدی قرار دارند و بصورت معکوس نیز به ما کمک می کند . در این تحقیق از دیتا ست UCF-101 and sport -1M استفاده شده که این نوع دیتاست طبق تحقیقاتی که داشتم دارای چالش هایی در مورد پیشبینی بوده است

در مقاله بعدی چالشی که معرفی کرده مشکلات روش سنتی LSTM هست که کار توصیف زبان طبیعی برای ویدئو رو سخت می کنه و یک معماری جدید پیشنهاد میده که از همون LSTM سنتی استفاده می کنه اما با شگردهایی مشکلات اونو در برابر دید کلی و دید جزئی به یک مسئله بهبود میده .

در مقاله ی دیگه ایده یکی جالبی زده و اینکه گفته در روش های دیگه ما تمرکزمون روی تصویر بود و روش های مختلفی برای اطلاع گرفتن از آن رو داشتیم اما اینجا صدا هم داخل کنیم میشه ویژگی های جدیدی بدست بیاوریم . سه استراتژی multimodal داریم تا بتونیم اطلاعات بهتری رو دریافت کنیم .اولین استراتژی این هست این اطلاعات رو مرتبه بندی کنیم و بر اساس مرتبه که داره روی الگوریتممون تاثیر بدیم . دومین استراتژی وابستگی کوتاه مدت صوتی تصویری هست که مثلا این صدامون نسبت به ویدئو تاثیر کمتری داشته باشه . سومین استراتژی یا محدودیت وابستگی بلند مدت حافظه هست که داده هایی که در حافظه هست که این داده بشه در بیشتر قسمت الگوریتم استفاده کرد .دیتاست هایی که مورد استفاده قرار میدیم Microsoft Research Video to Text (MSRVTT) و دیتاست Microsoft video desctiption(MSVD) می باشد .در این مقاله سعی شده که یک مقایسه ای بشه که روش با در نظر گرفتن صدا و بدون صدا در این ویدئو caption چه تاثیری داره .

در مقاله بعدی از LSTM استفاده کرده اما عنوان این الگوریتم رو lstm سلسله مراتبی گذاشته و به این معنی هست که که همانطور در کامپایلر ها ما سلسه مراتب برای تولید زبان و عبارت بعدی داریم در اینجا به عنوان phi-LSTMعنوان کرده که با مرتبه بندی این عبارات دیگه از مشکلات ترتیبی خالص رها میشیم و ادعا کرده در دیتاست های Flickr8kوFlickr30kو MS-COCO نتیجه بهتری نسبت به روش های موجود داشته است و برای داده هایی که دیده نشده نتیجه قابل قبول تری از کلمات داره .

حالا ما این انواع نگرش رو بیان کردیم و می خواهیم این نگرش ها رو ببینیم چگونه در مسئله ما پیاده سازی شده و مسئله از چه بخشی هایی تشکیل شده .

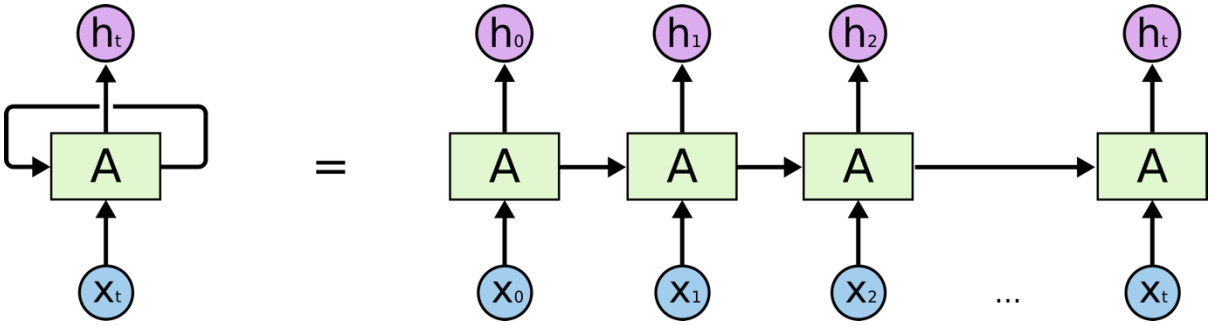
فصل دوم : تعریف دقیق مسئله و داده ها و روشهای حل

اول از هرکاری بهتر است که مسئله رو بیان کنیم و نیاز هایی که در این مسئله داریم رو بیان کنیم . در این روش ما یک ویدئو داریم که بصورت کلیپ کات شده تا حجم پردازشی کم شده باشه و اگر روی این کلیپ ها قابل قبول نتیجه داد می تونیم در بستر زمان بیشتر قرار داد مثل ایجاد یک پروتوتایپ هست و این داده ویدئویی که از دیتاست ها گرفتیم رو میخوایم با استفاده از یک روش زبان طبیعی بفهمیم و بیان کنیم تصویر چه اشیاهایی وجود دارند و اگر توانایی آن را داشتیم بتونیم عملیات که در تصویر صورت میگیره رو هم بگیم مثلا درحال روزنامه خوندن و در سطح بالاتر نیز میشه پیشبینی کنیم که در ادامه کلیپ چه اتفاقی میفته . ما در اینجا نمونه های که علمیات صورت گرفته رو بیان می کنیم و بررسی می کنیم که این عملیات ها رو چجوری و با چه روشی بررسی شده است .

ابتدا قبل از شروع کار باید بیان کنیم LSTM چیست جون می خواهیم انواع نگرش هایی که توسط این روش پایه صورت گرفته رو بیان کنیم و تعریف اولیه این الگوریتم رو در ابتدا نیاز داریم .

قبل از توضیح این شبکه باید بگین این دیدگاه شبکه از کجا اومد . همانطور در انسان میدانید ساختاری در مغز وجود داره که بطور مثال در مورد یک چیزی فکر می کند یا در حال تماشای یک ویدئو هست اطلاعات در حال بررسی در هر ثانیه ریست نمیشن و چیز جدید از اول فکر نمی کند و معنی هر اطلاعات رو از اطلاعات قبلی سرچشمه میگیرید و مثلا یک متن رو بررسی می کنید اطلاعات یک پاراگراف از کلمات داخل پاراگراف که قبل تر ازش رد شدین ارتباط داره و با خواندن کل پاراگراف معنی کامل رو متوجه میشید .

مثلا در ویدئو ما نیازمند فریم های قبلی هستیم تا متوجه بشیم چیا اتفاق افتاده و از یک فریم نمیشه استدلال ها و نتایج بزرگ گرفت . و واسه همون شبکه عصبی تعریف شده است بنام شبکه های عصبی بازگشتی یا Recurrent Neural Network که برای برطرف کردن این مشکل ساخته شده است و به عبارتی در این شبکه ها یک حلقه بازگشتی داریم که اطلاعات قبلی رو بارگذاری می کنند و این اطلاعات از بین نروند .

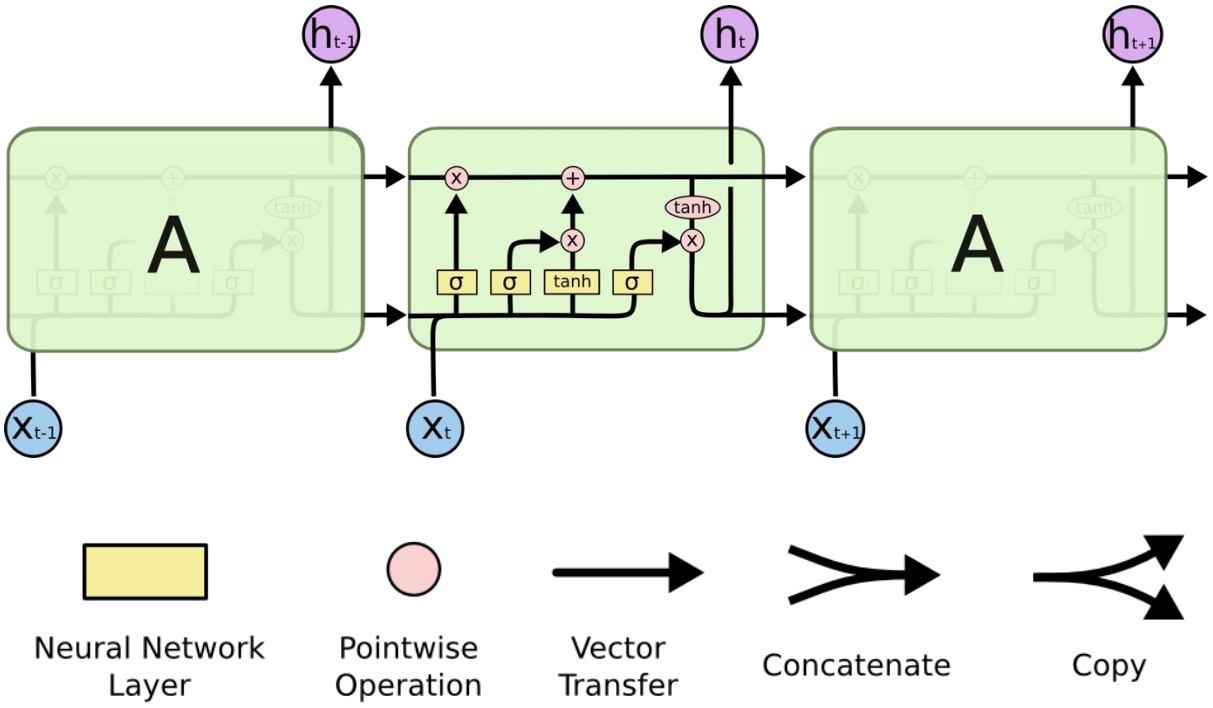


همانطور که در تصویر بالا میبینید به این صورت می توان یک عملیات فیدبک رو با شبکه عصبی ساده پیاده سازی کرد

در اینجا LSTM یک نوعی خاص از شبکه عصبی بازگشتی می باشد .

یک مشکل اساسی شبکه های عصبی بازگشتی در وابستگی بلند مدت هست به این صورت که بطور مثال ما یک تیکه از کلیپ ویدئویی داریم که به ثانیه قبل ارتباط داره اما در بعضی شرایط اون نیازمندی به فریم های های قبل تر نیز وابسته هست و این فاصله محدودیتی واسه ما ایجاد می کند و در بعضی شرایط به اطلاعات قبلی دورتر ما دسترسی نداریم و بطور خلاصه باید در اینجا دو کلمه vanishing and Exploding gradient برخورد می کنیم که این مشکل با ایجاد شبکه LSTM تا حدودی حل شده است .

شبکه LSTM مخفف کلمه Long Short Term Memory می باشد .



مقاله اول که از مجله neurocomputing Elsevier هست به ساختار ایجاد کپشن با استفاده از سلسله مراتب LSTM میپردازد .

مسئله اول که با اون روبرو هستیم این هست که عنوان بندی فایل چند رسانه ای رو بصورت کاملا مفهومی با راهکار Coase to Fine و دیدگاه سلسله مراتبی انجام بدیم . ایجاد خودکار توصیف زبان طبیعی برای ویدئو یک کار بسیار پیچیده و چالش برانگیز است. برای مقابله با موانع مدل مبتنی بر LSTM سنتی برای ویدیو captioning، ما یک معماری جدید را برای ایجاد یک ساختار شبکه جدید پیشنهاد می‌کنیم که بر ساخت یک ساختار شبکه جدید تمرکز می‌کند که می‌تواند باعث ایجاد جملات برتر نسبت به مدل پایه با LSTM شود، و مکانیسم‌های مورد توجه خاص ایجاد کند که بتواند اطلاعات بصری مفید تری را برای تولید عنوان به دست آورد. این طرح از LSTM سنتی استفاده می‌کند و از شبکه کامل convolutional با توجه به ویژگی‌های ساختار سلسله مراتبی و Coarse to fine استفاده می‌کند. مدل معرفی شده نه تنها می تواند بهتر از مدل پایه مبتنی بر مدل عمل کند، بلکه به عملکرد قابل‌مقایسه با روش‌های نوین هنری دست یابد.

مقدمه با رشد انفجاری داده‌های ویدئویی در وب، چگونه به طور یکپارچه ساختار پیچیده ویدیوهای مختلف را کنترل کنیم تا به تولید توضیح موثر برسند (Venugopalan و همکاران ۲۰۱۵ b؛ یائو و سایرین ۲۰۱۵). اگرچه برای انسان آسان است که یک ویدیو را با یک نگاه سریع توصیف کند، اما به طراحی‌های پیچیده‌تر برای کامپیوترها نیاز دارد تا همین کار را انجام دهند (Venugopalan و همکاران ۲۰۱۵ a؛ Pan و همکاران ۲۰۱۶). مدل نسل زیر باید آنچه را که در یک ویدئو ظاهر می‌شوند، ویژگی‌های این اشیا و روابط بین اشیا را مشخص کند. تمام این وظایف چالش‌های اساسی در بینایی رایانه‌ای هستند. بنابراین captioning ویدیویی، که هدف ساخت یک مدل تولید زبان است که بتواند درک معنایی را با توصیف دقیق و معنی‌دار برای ویدئوها نشان دهد، توجه قابل‌ملاحظه‌ای را به خود جلب کرده‌است. برخی از روش‌های پیشگام تلاش می‌کنند تا پیام‌های ویدئویی را از طریق مفاهیم تصویری و الگوهای محکومیت به سختی شناسایی کنند (Kojima و سایرین ۲۰۰۲). با این حال، این روش‌ها به شدت ساخته شده‌اند و جملات تولید شده کم‌تر طبیعی هستند. اخیرا، با الهام از پیشرفت‌های مدل ترجمه ماشینی عصبی (چو و همکاران ۲۰۱۴؛ Sutskever و سایرین ۲۰۱۴)، ترکیبی از شبکه‌های عصبی مصنوعی (cnns)و شبکه‌های عصبی مصنوعی (rnns)به طور گسترده اتخاذ می‌شوند که به طور قابل‌توجهی کیفیت توصیف متن ویدئو را بهبود بخشیده‌است. به طور کلی مدل‌های سی ان ان - RNN ابتدا ویدیو را به یک بردار ویژگی با طول ثابت با استفاده از سی ان ان کدگذاری می‌کند و سپس بردار ویژگی را در یک رمزگشا RNN برای ایجاد تصاویر، تغذیه می‌کند. با وجود پیشرفت موجود، مدل CNNRNN اساسی هنوز هم یک "rough" است. برای captioning ویدئو، ورودی و خروجی آن ساختارهای ترتیبی هستند که حاوی اطلاعات زمانی هستند. به خاطر عملکرد برجسته of به خصوص شبکه حافظه کوتاه‌مدت کوتاه‌مدت ((LSTM، RNN به طور طبیعی بخش مهمی در وظایف تولید زنجیره می‌شود (Venugopalan و همکاران ۲۰۱۵ a؛ ۲۰۱۵ b). با این حال، کاستی‌های آن در وظایف captioning کشف شده‌است. LSTM نیازمند خروجی قبلی به عنوان ورودی در هر لحظه است، که باعث می‌شود آموزش به شدت کند شود. واحد حافظه of پیچیده است و به دلیل مسیر طولانی انتشار، به ذخیره‌سازی مهم نیاز دارد. همه اینها آموزش of را دشوار می‌سازند. بنابراین برخی محققان تلاش کرده‌اند تا نسل متوالی را از طریق معماری مدل جدید بدون LSTM حل کنند و برخی موفقیت‌ها را در ترجمه ماشینی عصبی ایجاد کنند (Gehring و همکاران ۲۰۱۷؛ Vaswani و سایرین ۲۰۱۷).
با توجه به مشاهدات فوق، ما یک چارچوب جدید برای تولید توضیحات بهینه برای ویدئوها پیشنهاد می‌کنیم. چارچوب ما the سنتی را به تاخیر می‌اندازد و از شبکه کامل convolutional به عنوان معماری اساسی استفاده می‌کند و در عین حال توجه فراگیر و fine را که براساس ویژگی‌های ساختار کاملا convolutional طراحی شده‌اند، نشان می‌دهد. این چارچوب با هدف پرداختن به دو مساله کلیدی برای مقابله با موانع of، یعنی ایجاد

یک ساختار شبکه جدید که می‌تواند جملات برتر را نسبت به مدل‌های مبتنی بر LSTM ایجاد کند، و ایجاد مکانیزم‌های مورد توجه خاص که می‌تواند اطلاعات بصری مفید تری را برای نسل زیر فراهم کند، هدف‌گذاری می‌کند. مثال‌های captions ایجاد شده توسط مدل مبتنی بر LSTM و ما در شکل ۱ نشان‌داده شده‌اند. برای بهترین دانش ما، این یک تلاش جدید برای استفاده از شبکه کاملا convolutional خالص با توجه به captioning های ویدیویی است. مشارکت اصلی ما شامل جنبه‌های زیر است: ۱)متفاوت از مدل سنتی مبتنی بر مدل، یک معماری کاملا convolutional با دقت مناسب و موروثی، برای استفاده کامل از سطوح مختلف اطلاعات بصری درگیر در ویدئو طراحی شده‌است، و توجه موروثی برای تمرکز بهتر بر روی اطلاعات در سطح منطقه طراحی شده‌است که باید در لحظات مختلف تمرکز شود. مدل ما نه تنها می‌تواند بهتر از مدل پایه مبتنی بر مدل عمل کند، بلکه به عملکرد قابل‌مقایسه با مدل‌های the - of در MSVD دست یابد، که اثربخشی و امکان‌سنجی چارچوب پیشنهادی ما را تایید می‌کند.

-

توصیف محتوای ویدئو با زبان طبیعی در سال‌های اخیر پیشرفت خوبی داشته‌است. روش‌های موجود برای captioning ویدئو را می توان به دو دسته تقسیم کرد، یعنی، براساس الگوی یادگیری مبتنی بر الگو و مبتنی بر توالی (Kojima و همکاران ۲۰۰۲؛ Venugopalan و همکاران ۲۰۱۵ a؛ گان و سایرین ۲۰۱۷). روش مبتنی بر الگو برخی از قوانین خاص برای توصیف تولید شده را تشریح می‌کند و عنوان را به چند بخش مثل فاعل، فعل و object تقسیم می‌کند (Kojimaو همکاران ۲۰۰۲). با الگوهای از پیش تعریف‌شده، هر بخش از جمله با کلمات شناسایی‌شده از اطلاعات بصری ویدئو، و در نهایت توصیف ویدئو می‌تواند تولید شود مرتبط است. به عنوان مثال، برای توصیف فعالیت‌های انسانی با زبان طبیعی، Kojima و همکاران (۲۰۰۲)یک معماری مفهوم از اقدامات را پیشنهاد کردند و یک سلسله‌مراتب معنایی برای یادگیری روابط معنایی بین بخش‌های مختلف جملات طراحی شد. اخیرا، ژو و همکاران (۲۰۱۵)یک چارچوب یکپارچه را پیشنهاد کردند که شامل یک مدل زبانی ترکیبی، یک مدل ویدیویی عمیق و یک مدل تعبیه مشترک برای captioning ویدئو است. این روش‌ها می‌توانستند جملات مسلط را تولید کنند، اما مشکلات آشکاری داشتند. آن‌ها به شدت بر الگوهای و قواعد از پیش تعیین‌شده تکیه داشتند، که جملات تولید شده را بسیار سخت می‌کرد. روش مبتنی بر یادگیری از الگوی مبتنی بر الگو متفاوت است. به جای استفاده از قواعد از پیش تعیین‌شده، آن به طور مستقیم عنوان نهایی را با ساختار نحوی انعطاف‌پذیرتر که براساس ویدیو ورودی است، ایجاد می‌کند. دوناهیو و همکاران (۲۰۱۵)پیشنهاد شبکه‌های بلند مدت (LRCNs)را پیشنهاد کردند که از نقاط قوت of برای تشخیص بصری استفاده کرد و از LSTMبه عنوان مدل زبانی استفاده کرد. برای به دست آوردن نمایش بصری بهتر، Venugopalan و همکاران (۲۰۱۵ a)اطلاعات موقتی را با جریان نوری در نظر گرفتند و LSTM هم در کدگذار و کدگشا استفاده کردند. پن و همکاران (۲۰۱۷)ساختاری را پیشنهاد کردند که ویژگی‌های معنایی هر دو تصویر و تصویر را در نظر بگیرند، که می‌تواند اطلاعات معنایی اضافی فراهم کند. برای کسب اطلاعات مفید بیشتر، ژو و سایرین (۲۰۱۷)ویژگی‌های multimodal را در نظر گرفتند که شامل ویژگی‌های چارچوب، حرکت و صدا بود. لی و همکاران (۲۰۱۷)به طور مشترک توجه سطح منطقه‌ای و چارچوب به وظیفه of ویدیویی را به کار گرفتند تا نمایش بصری مفید و دقیق‌تر را بگیرند.

وانگ و همکاران (۲۰۱۸)یک چارچوب یادگیری تقویتی سلسله مراتبی را برای یادگیری پویایی معنایی زمانی که captioning یک ویدئو است پیشنهاد دادند. برای به دست آوردن عنوان بهتر از آنچه توسط ساختار اولیه کدگشا - رمزگشا تولید می‌شود، برخی از محققان یک ماژول اضافی به نام "Reconstructor" (وانگ و همکاران ۲۰۱۸)یا "ARNet" را اضافه کردند (Chen et al.، ۲۰۱۸). اگرچه این ماژول نام‌های متفاوتی دارند، ایده اصلی این است که برخی از مشکلات ذاتی مدل basic - رمزگشا را با انباشته کردن ساختار LSTM حل کند. چنین آثاری نقش بزرگی در فیلم‌های ویدئویی ایفا کرده‌اند و به ما الهام زیادی داده‌اند. این مزیت این است که the های ایجاد شده همگی جملات شکل خوبی هستند، که بسیار طبیعی‌تر از جملات تولید شده توسط فرمت های مبتنی بر الگو هستند.

اگر چه پیشرفت‌های زیادی در روش یادگیری توالی انجام شده‌است اما هنوز بسیاری از مشکلات در مدل‌های مبتنی بر مدل وجود دارد، که بهبود بیشتر برای فیلم‌های ویدیویی را محدود می‌سازد. برای کاهش کاستی در روش‌های مبتنی بر الگوی یادگیری مبتنی بر الگو، هم فیس بوک (Gehring و همکاران ۲۰۱۷)و گوگل (Vaswani و همکاران ۲۰۱۷)معماری‌های مدل جدید را برای حل وظایف نسل توالی بدون RNN پیشنهاد کردند.

مدلی که توسط فیس بوک پیشنهاد شد، بر مبنای سی ان ان بود، در حالی که گوگل به این پیشنهاد توجه داشت. هر دو مدل به نتایج قابل‌توجهی در ترجمه ماشینی دست یافته‌اند، که پتانسیل این مدل‌ها را در دیگر وظایف مدلسازی توالی مثل تصویر یا تصویر ویدیویی نشان می‌دهد. Aneja و همکاران (۲۰۱۸)پیشنهاد کردند که از سی ان ان برای captioning تصویر استفاده کنند، و نمرات جملات تولید شده توسط مدل مبتنی بر شبکه سی ان ان قابل‌مقایسه با مدل پایه RNN بود. با الهام از چنین پیشرفتی، ما به طور خاص یک مدل captioning ویدئویی را با یک شبکه کاملا convolutional پیشنهاد می‌کنیم و مکانیسم‌های توجه جدید برای ساختار انباشت شده و محاسبه توجه سطح منطقه برای ایجاد توصیفات دقیق‌تر، ایجاد می‌نماییم.

روش بعدی :

ادغام هر دو نشانه‌های بصری و تصویری برای عنوان ویدئو گسترش‌یافته

چکیده
عنوان ویدئو به ایجاد یک حکم توصیفی برای یک کلیپ ویدئویی کوتاه خاص به طور خودکار اشاره دارد، که به تازگی به موفقیت چشمگیری دست یافته‌است. با این حال، اکثر روش‌های موجود بیشتر بر روی اطلاعات بصری تمرکز می‌کنند در حالی که نشانه‌های صوتی همگام‌سازی را نادیده می‌گیرند. ما سه استراتژی ترکیب چند multimodal برای به حداکثر رساندن مزایای اطلاعات رزونانسی صوتی - تصویری پیشنهاد می‌کنیم. اولی به بررسی تاثیر ترکیب دو حالت از پایین به مرتبه بالا می‌پردازد. دومی، وابستگی کوتاه‌مدت صوتی - کوتاه‌مدت را با به اشتراک گذاری وزن شبکه‌های front متناظر، ایجاد می‌کند. روش سوم وابستگی موقتی به مدت طولانی را از طریق به اشتراک گذاری حافظه multimodal در طول شرایط بصری و صوتی گسترش می‌دهد.
آزمایش‌ها گسترده، اثربخشی سه راهبرد ترکیبی ما بر روی دو مجموعه داده معیار، از جمله ویدیو تحقیق مایکروسافت تا متن (MSRVTT)و توصیف ویدیو مایکروسافت (MSVD)را تایید کرده‌اند.
لازم به ذکر است که وزن اشتراک گذاری می‌تواند ترکیب ویژگی visualaudio را به طور موثر هماهنگ کند و عملکرد هنر - هنری را در دو معیار BELU و meteor به دست آورد.
علاوه بر این، ما ابتدا یک چارچوب ترکیبی چند multimodal را برای رسیدگی به بخشی از موردی که در حال حاضر وجود دارد، پیشنهاد می‌کنیم. نتایج تجربی نشان می‌دهند که حتی در حالت غیبت صدا، ما هنوز هم می‌توانیم نتایج قابل‌مقایسه با کمک مدول استنباط modality صوتی بدست آوریم.

مقدمه
توصیف خودکار ویدیو با جملات طبیعی کاربردهای بالقوه بسیاری در بسیاری از زمینه‌ها مثل تعامل humanrobot، بازیابی ویدیو دارد. اخیرا، بهره‌گیری از قابلیت‌های خارق‌العاده شبکه‌های عصبی مصنوعی (CNN و Zisserman ۲۰۱۴؛ Szegedy و همکاران ۲۰۱۵؛ ۲۰۱۶)، شبکه‌های عصبی تکراری (Hochreiter و Schmidhuber ۱۹۹۷)، عنوان تصویر ویدیویی موفقیت امیدوارکننده را کسب کرده‌است. اکثر قالب‌های زیر به ترتیب می‌توانند به ترتیب به یک مرحله کدگذار و یک مرحله رمزگشا تبدیل شوند. Conditioned بر روی یک طول ثابت نمایش ویژگی‌های بصری پیشنهاد شده توسط کدگذار، رمزگشا می‌تواند یک توصیف ویدیویی مشابه را ایجاد کند. برای تولید یک نمایش طول ثابت، چندین روش پیشنهاد شده‌اند، از جمله ادغام بر روی قاب‌ها (Venugopalan و همکاران ۲۰۱۵؛ ۲۰۱۳)، sub بر تعداد ثابتی از چارچوب‌های ورودی (یائو و همکاران ۲۰۱۵)و استخراج آخرین حالت پنهان کننده کلید بصری بازگشتی (Venugopalan و همکاران ۲۰۱۵). آن روش‌های

رمزگذاری که در بالا ذکر شدند تنها براساس نشانه‌های بصری هستند. با این حال، ویدیوها شامل modality بصری و the صدا هستند. اطلاعات رزونانسی در آن‌ها برای چاپ ویدیویی برای آن‌ها ضروری است. ما معتقدیم که فقدان of اختیاری منجر به از دست دادن اطلاعات خواهد شد. به عنوان مثال، زمانی که یک فرد روی تخت دراز کشیده و یک آهنگ می‌خواند، روش‌های سنتی به عنوان "فرد روی تخت‌خواب دراز کشیده"، که ممکن است به دلیل از دست دادن اطلاعات رزونانسی و modality صوتی باشد. اگر ویژگی‌های صوتی را می توان با چارچوب (چارچوب توصیف ویدئویی)ترکیب کرد، جمله دقیق "کسی که روی تخت‌خواب دراز کشیده و آواز می‌خواند" انتظار می‌رود تولید کند. برای بهره‌برداری کامل از اطلاعات تصویری و تصویری، سه استراتژی ترکیب فضایی چند multimodal را برای به حداکثر رساندن مزایای اطلاعات رزونانسی صوتی - تصویری پیشنهاد و تحلیل می‌کنیم. اولی، تاثیر ترکیب crossmodalities را از پایین به مرتبه بالا بررسی می‌کند. دومی، وابستگی کوتاه‌مدت صوتی - کوتاه‌مدت را با به اشتراک گذاری وزن شبکه‌های front متناظر، ایجاد می‌کند. روش سوم وابستگی موقتی به مدت طولانی را از طریق به اشتراک گذاری حافظه multimodal در طول شرایط بصری و صوتی گسترش می‌دهد.

علاوه بر این، یک چارچوب پویا از ترکیب چند مولفه (fusion)نیز برای مقابله با اختلالات صوتی و مشکل موجود در طول تولید نوشته شده ارایه شده‌است. سهم مقاله ما شامل موارد زیر است:

a. سه استراتژی ترکیب چند multimodal را ارایه می‌دهیم، تا به طور موثر حالات صوتی را به عنوان برچسب ویدیویی ترکیب کنیم.

ب. ما یک ماژول استنباطی با کیفیت صدا برای رسیدگی به مشکل موجود در کیفیت صدا، از طریق ایجاد ویژگی صدا براساس ویژگی تصویری متناظر ویدیویی پیشنهاد می‌دهیم.

ج. نتایج تجربی ما براساس مایکروسافت ResearchVideo به متن (MSR - VTT)و داده‌های توصیف ویدیو مایکروسافت (MSVD)نشان می‌دهد که چارچوب‌های زمانی multimodal، منجر به نتایج بهبود یافته در caption ویدئو می‌شوند.

کاره‌ای مرتبط

عنوان ویدئو

کاره‌ای اولیه مربوط به عنوان نوشته شده را می توان به سه گروه دسته‌بندی کرد. مقوله اول روش‌های مبتنی بر الگو است. آن‌ها ابتدا ویژگی‌های معنایی پنهان در ویدئوها را شناسایی کردند و سپس یک ساختار جمله‌ها را براساس برخی از الگوهای چند جمله‌ای از پیش تعیین‌شده بدست آوردند (Krishnamoorthy و همکاران ۲۰۱۳؛ Thomason و سایرین ۲۰۱۴). سپس از مدل گرافیکی احتمالاتی برای جمع‌آوری مرتبط‌ترین محتوای ویدئو برای تولید جمله مربوطه استفاده شد. اگر چه جملات تولید شده توسط این مدل‌ها به نظر درست می‌رسند، اما فاقد غنای و انعطاف‌پذیری هستند. مقوله دوم به عنوان یک مشکل بازیابی بازی می‌کند. آن‌ها ویدیوها را با متاداده (Aradhye، Toderici، و Yagnik ۲۰۰۹)برچسب گذاری کرده و سپس ویدیو و captions را براساس این برچسب‌ها قرار دادند. اگر چه جملات تولید شده به طور طبیعی نسبت به گروه اول مقایسه شدند، اما آن‌ها به طور جدی در معرض the قرار گرفتند. طبقه‌بندی سوم به طور مستقیم نمایش دیداری را به جملات ویژه ارایه می‌کند (Venugopalan و همکاران ۲۰۱۵؛ Yao و همکاران ۲۰۱۵؛ ۲۰۱۶ و سایرین ۲۰۱۵)، که از caption تصویر الهام می‌گیرد (Vinyals و همکاران ۲۰۱۵؛ دوناهیو و سایرین ۲۰۱۵).

ما استدلال می‌کنیم که این توصیف ویدئویی تنها به اطلاعات بصری تکیه دارد در حالی که راهنمایی‌های صوتی را نادیده می‌گیرد، که عملکرد زیر را محدود خواهد کرد. برای رسیدگی به این مشکل، ما به کاوش اطلاعات صوتی در زیر عنوان ویدئو می‌پردازیم.

توالی صوتی نهفته در ویدئوها همیشه حاوی اطلاعات معنی‌دار است. اخیرا، بسیاری از محققان تلاش کرده‌اند تا اطلاعات صوتی را در کاربردهای خاص خود وارد کنند. در (Owens و همکاران ۲۰۱۶)، اونز و همکارانش صداهای محیط را به عنوان یک سیگنال نظارتی برای آموزش مدل‌های دیداری به کار گرفتند، آزمایش‌ها آن‌ها نشان داد که واحدهای شبکه آموزش‌دیده نظارت شده توسط سیگنال‌های صوتی اطلاعات معنی‌دار معنایی در مورد اشیا و صحنه‌ها داشته‌اند. رن و سایرین (رن و سایرین ۲۰۱۶)یک حافظه کوتاه‌مدت چند multimodal (LSTM)برای شناسایی سخنگو را پیشنهاد کردند که به مکان‌یابی شخصی اشاره کرد که هویت یک‌سان را با صدای در حال پیشرفت در یک ویدیو مشخص دارد. نکته اصلی آن‌ها تقسیم اوزان روی صورت و صدا برای مدلسازی وابستگی زمانی این دو روش مختلف بود.
با الهام از (رن و سایرین ۲۰۱۶)، ما پیشنهاد می‌کنیم که وابستگی موقتی در سطوح تصویری و تصویری از طریق به اشتراک گذاری وزن برای عنوان ویدئو، با هدف بررسی اینکه آیا وابستگی موقت در سطوح تصویری و تصویری می‌تواند اطلاعات رزونانسی را در بین آن‌ها ثبت کند یا خیر، ایجاد کنیم.

شبکه عصبی گسترش‌یافته Recurrent
حافظه داخلی در RNN می‌تواند اطلاعات ارزشمند را برای وظایف خاص حفظ کند. با این حال، نمی‌تواند وظایف مربوط به وابستگی زمانی بلند مدت را کنترل کند. برای افزایش توانایی حافظه of، از حافظه خارجی برای گسترش RNN در برخی کارها نظیر ماشین تورینگ Neural) NTM)، وین و Danihelka ۲۰۱۴)، شبکه حافظه (وستون، Chopra و Bordes ۲۰۱۴)استفاده شده‌است، که به سادگی RNN (ME - RNN)نامیده می‌شود. تجهیزات الکتریکی پزشکی به طور گسترده در بسیاری از کارها به کار گرفته شده‌اند.
علاوه بر رسیدگی به وظیفه‌ای که به وابستگی زمانی طولانی نیاز دارد، مانند پاسخ سوال دیداری (Xiong، Merity، و Socher ۲۰۱۶)و سیستم‌های تبادل (داج و سایرین ۲۰۱۵)، من - rnns برای مدل کردن وابستگی زمانی طولانی در وظایف مختلف به کار گرفته شده‌اند (لیو، Qiu، و هوانگ ۲۰۱۶). برای بررسی این که آیا وابستگی زمانی طولانی - تصویری می‌تواند اطلاعات رزونانسی در میان دو روش را ثبت کند، ابتدا سعی می‌کنیم یک حافظه اشتراکی صوتی - تصویری در میان روش‌های صوتی و تصویری برای caption ویدئویی بسازیم.

روش بعدی :

تولید کننده عبارت بندی عنوان عکس بصورت سلسله مراتبی شبکه lstm

به تازگی، در جایی که بیشتر آثار موجود عنوان داده متوالی خالص را به عنوان داده متوالی خالص، توصیف می‌کنند، به طور خودکار به شکل اتوماتیک عنوان توصیف محتوای یک تصویر در حال به دست آوردن تعداد زیادی از علایق تحقیقاتی بوده‌است. با این حال، زبان طبیعی دارای ساختار سلسله مراتبی زمانی با وابستگی‌های پیچیده بین هر subsequence است. در این مقاله، ما یک مدل تصویر مبتنی بر اصطلاح را با استفاده از معماری سلسله مراتبی حافظه کوتاه‌مدت (فی - LSTM)برای ایجاد توضیحات تصویری پیشنهاد می‌کنیم. در مقایسه با راه‌حل‌های con که عنوان را به شکل ترتیبی خالص تولید می‌کنند، عنوان تصویر decodes decodes - از عبارت تا جمله. آن شامل یک عبارت است برای رمزگشایی عبارت‌های اسمی طول متغیر، و یک رمزگشا برای رمز گشایی به شکل خلاصه توصیف تصویر. یک تصویر تصویر com با ترکیب عبارت‌های ایجادشده با جمله در طول مرحله استنتاج شکل می‌گیرد. علاوه بر این، مدل پیشنهادی ما یک نتیجه بهتر یا رقابتی را در مجموعه داده‌های k۸Flickr، k۳۰Flickr و MS - coco در مقایسه با مدل‌های هنری نشان می‌دهد. همچنین نشان می‌دهیم که مدل پیشنهادی ما قادر به تولید captions های جدید است (در داده‌های آموزشی مشاهده نمی‌شود)که در تمام این سه مجموعه داده‌های کلمه غنی‌تر هستند.

اگر چه مدل ترتیبی برای پردازش داده‌های sentential مناسب است، اما ساختار نحوی دیگری از زبان را در مدل‌سازی خود مورد توجه قرار نمی‌دهد. در حقیقت، زبان طبیعی، داده‌های متوالی است که دارای سلسله‌مراتب زمانی است و اطلاعات در مقیاس‌های زمانی چندگانه پخش می‌شوند [۱۶]. به انگلیسی به عنوان مثال در نظر بگیرید، پایین‌ترین سطح با کوتاه‌ترین زمان - مقیاس، شخصیت‌هایی است که پس از آن کلمات، عبارات، بندها، جملات و اسناد را دنبال می‌کنند. بنابراین، غیرقابل‌انکار است که ساختار جمله یکی از ویژگی‌های برجسته زبان است. ویکتور Yngve، یکی از نویسندگان تاثیرگذار در تئوری زبانی، در سال ۱۹۶۰ بیان کرد که "ساختار زبان شامل یک سلسله‌مراتب ساختار و یا یک سازمان تشکیل‌دهنده فوری است" [۱۷].
بنابراین، مجبور کردن یک مدل تولیدی برای آموزش دنباله‌های هموار و سپس تولید یک ساختار سطح بالا، در یک اساس گام‌به‌گام اغلب منجر به عملکرد محدود می‌شود [۱۸]. برای مثال، به طور خاص، می توان مشاهده کرد که حداقل دو سطح از ساختار در تصاویر تشریح شده انسانی در مجموعه داده‌های عمومی مانند k۸Flickr، k۳۰Flickr و MS - coco وجود دارد. در هر یک از این نوشته‌ها، چندین عبارت وجود دارد که اشیا را در یک تصویر توصیف می‌کنند. این عبارات دارای وضوح زمان - زمان برابر در سطح کلمه می‌باشند، و در طول رمزگشایی، هم بر روی هم ساختار و هم ساختار زبان کوتاه شرطی شده‌اند. بنابراین، کلمات قبلی در عنوان به جز خود عبارت که در حافظه بلند مدت کدگذاری می‌شود، در فرآیند تولید اضافی است. علاوه بر این، ساختار زیر عنوان در این عبارات بیشتر وابسته است، و بنابراین هم به تصویر و هم همه توالی‌های قبلی به عنوان زمینه‌ای برای ایجاد یک توصیف صحیح نیاز دارد. در این مقاله ما می‌خواهیم قابلیت یک مدل captioning تصویر مبتنی بر اصطلاح را بررسی کنیم که ساختار مشاهده‌شده در مدل‌سازی آن را در مقایسه با یک مدل مشابه که در دنباله‌های هموار دیده می‌شود را در بر می‌گیرد. در این راستا، ما یک مدل captioning تصویر phrasebased را با استفاده از یک معماری سلسله مراتبی سلسله مراتبی، یعنی phi - LSTM طراحی می‌کنیم که متشکل از یک decoder عبارت و یک سیگنال به اختصار (AS)برای تولید توضیحات تصویری از عبارت تا جمله می‌باشد. همانطور که در شکل ۱ نشان‌داده‌شده، با توجه به تصویری که با شبکه سی ان ان کدگذاری شد، the ابتدا برای رمزگشایی عبارت‌های اسمی (NPs)(یعنی یک موتورسوار، خیابان)که ماهیت های غالب درون تصویر را توصیف می‌کنند، استفاده می‌کنند و از کلمات به عنوان واحد اتمی استفاده می‌کنند. در عین حال، اصطلاح decoder هر یک از NP را به یک نمایش بردار ترکیبی رمز گذاری می‌کند، که به عنوان ورودی به عنوان decoder در سلسله‌مراتب بالا عمل می‌کند. به این ترتیب، NPs یک تفکیک زمانی برابر با بقیه کلمات در سطح محکومیت خواهند داشت (به عنوان مثال در). سپس، رمزگشا شکل خلاصه نوشته شده را رمزگشایی خواهد کرد، که از کلمه آخر هر یک از هر NP (یعنی motorcyclist، خیابان)و آن کلمات باقی مانده که عبارات را به هم متصل می‌کنند، ایجاد می‌شود. در نهایت، یک عنوان تصویر کامل (به عنوان مثال یک موتورسوار در خیابان)با ترکیب عبارت‌های ایجادشده با جمله به تدریج، در طول جستجوی پرتو در مرحله استنتاج شکل می‌گیرد. Empirically، مدل پیشنهادی ما نتایج بهتری را در k۸Flickr [۱۹]، k۳۰Flickr [۲۰] و MS - coco [۲۱] در مقایسه با مدل‌های هنری نشان می‌دهد.

به عنوان خلاصه، سهم ما دو برابر شده‌است:
ا. ما یک مدل سلسله مراتبی سلسله مراتبی را برای رمزگشایی عنوان تصویر از جمله به جمله، پیشنهاد می‌کنیم.
۲. ما نشان می‌دهیم که caption تصویر ایجادشده با فی - LSTM دقیق‌تر است، رمان (که در داده‌های آموزشی دیده نمی‌شود)و غنی‌تر در محتوای کلمه است.
نسخه اولیه این کار در [۲۲] ارایه شد، در حالی که کار فعلی به روش اولیه به روش‌های معنی‌دار اضافه شده‌است. اول، هدف انتخاب عبارت است با پیش‌بینی آخرین کلمه هر یک از هر یک از NP به عنوان decoder برای آموزش سادگی. دوم، ما طول نرمال سازی طول را در طول مرحله استنتاج در هر دو عبارت و سطح محکومیت، به منظور ایجاد یک عنوان طولانی‌تر، اعمال می‌کنیم. سوم، ما خروجی‌های ابزار تجزیه را با یک استراتژی اصلاح اصطلاح بهبود خواهیم داد. در نهایت، تحلیل‌های جدید و توضیحات شهودی به نتایج ما اضافه می‌شوند. ما همچنین آزمایش خود را بسط می‌دهیم تا مجموعه داده MS - coco را در نظر بگیریم (۲۱)، و نتایج خود را براساس چهار معیار ارزیابی دیگر ارزیابی کنیم (یعنی meteor [۲۳]، ROUGE [۲۴]، cider [۲۵]، cider [۲۶] و SPICE [۲۶]).

۲. کاره‌ای مرتبط
رویکردهای نسل توصیف تصویر متفاوت هستند (از نظر من)چگونه متنی که توصیف از آن مشتق شده‌است، و ۲)چگونه یک جمله ایجاد می‌شود.
۲.ا نمایش متن
برای کدگذاری اطلاعات دیداری، کاره‌ای اولیه بر روی ردیاب صوتی چندگانه و طبقه‌بندی کننده‌ها برای ثبت جنبه‌های مختلف یک تصویر مانند اشیا، ویژگی‌ها، روابط و صحنه تکیه دارند [۲۷ - ۲۷]. خروجی‌های این شناساگر و طبقه‌بندی کننده‌ها معمولا مجموعه‌ای از tuples را تشکیل می‌دهند [۳۱ - ۳۱]، که در آن توصیف بر روی آن ساخته شده‌است. چنین روشی عموما تعداد کلاس‌ها را برای هر جنبه از تصویر برطرف می‌کند. از آنجا که موفقیت بی‌سابقه سی ان ان در طبقه‌بندی تصویر و وظایف تشخیص شی، تعداد رو به رشدی از آثار از سی ان ان برای کدگذاری یک تصویر کامل استفاده می‌کند [۳، ۴، ۶، ۱۱، ۱۵، ۳۴، ۱۳، ۱۳، ۱۳، ۴۰ - ۴۲]. با در نظر گرفتن تصویر کدگذاری شده از سی ان ان و توصیف آن، بسیاری از کارها با استفاده از مدل‌های مختلف زبانی، فضای بازی چند multimodal را آموزش می‌دهند [۳ - ۷، ۹، ۱۱، ۳۴، ۳۴، ۳۴، ۳۴، ۴۳، ۴۳]. به همین ترتیب، Fang و سایرین [۱۳]، وو و سایرین [۱۴]، و شما و همکاران [۱۵] مجموعه‌ای از "ردیاب صوتی تصویری" را در داده‌های آموزشی برای کدگذاری تصویر در یک فضای معنایی، به نام ویژگی‌ها آموزش دادند. علاوه بر آن، کارهایی وجود دارند که بر رویکرد بازیابی برای ایجاد توضیحات تصویری تکیه می‌کنند. با بازیابی و رتبه‌بندی دوباره شکل تصاویر مشابه از مجموعه آموزشی [۳۴، ۳۵، ۴۰، ۴۴، ۴۵]، یک تصویر پرس و جو می‌تواند با عنوان نوشته شده انسان توصیف شود که به محتوای آن مرتبط است. با این حال، این روش قادر به توصیف یک تصویر با ترکیب نامریی اشیا نیست. بنابراین، برخی از کارها در این خط از رویکرد مجموعه‌ای از چند tuples را بازیابی می‌کند [۲۷] یا text (۳۲، ۳۳، ۴۶)تا captions جدید را شکل داده و دوباره رتبه‌بندی کند. از سوی دیگر، Mun و همکاران [۴۷] یک نقشه توجه را بر روی ویژگی تصویر پرس و جو با استفاده از the بازیابی شده به عنوان راهنما برای شکل دادن به نمایش متن ایجاد کردند.

۲.۲. نسل Description

با در نظر گرفتن زمینه‌های مختلف توصیف‌شده در بالا، چندین رویکرد برای ایجاد توضیحات تصویر توسعه داده می‌شوند، که عبارتند از: templatebased، ۲)ترکیب مبتنی بر ترکیب، و ج)مدل زبان - based.

۲.۲.۱. الگو محور

این رویکرد با استفاده از یک قالب از پیش تعریف‌شده با اسلات باز برای پر کردن از موجودیت‌های تصویر ایجاد می‌کند [۲۷، ۲۸، ۳۰، ۴۶]. آن عمدتا توسط آثاری به کار می‌رود که محتوای تصویری را به عنوان مجموعه‌ای از چند tuples نشان می‌دهد. توصیف تولید شده این روش معمولا از نظر نحوی صحیح است، اما سخت و انعطاف‌پذیر نیست. روش ترکیب‌بندی ۲.۲.۲. این روش up snippets بازیابی شده [۳۲، ۳۳] یا entities شناسایی شد [۲۹، ۳۱] تا یک توصیف تصویری را شکل دهد. برای تصمیم‌گیری درباره مجموعه‌ای از بخش‌های متنی یا entities که برای ایجاد یک عنوان کامل، دستورها آن‌ها و چسباندن کلمات بین آن‌ها استفاده می‌شود، نیاز به قوانین پیش از تعریف پیچیده دارد. توصیف تولید شده در چنین حالتی وسیع‌تر و more نسبت به رویکرد مبتنی بر الگو است، اما در زمان تست به دلیل ماهیت غیر پارامتری آن، بسیار پرهزینه است.

مدل زبانی ۲.۲.۳. - based daily

اکثر کاره‌ای اخیر به طور مشترک تصویر و زبان را به یک فضای تعبیه‌شده چند مولفه در مدل زبانی مبتنی بر شبکه عصبی برای تولید caption تصویر اختصاص می‌دهند [۳ - ۷، ۳۶]. به عنوان مثال، Kiros و همکاران [۳۶] یک مدل زبان عصبی تک پارامتری چند متغیره را پیشنهاد کردند که با ویژگی تصویر برای رمزگشایی به تصویر کشیده می‌شود. چن و همکاران [۴۸] از RNN برای ساخت یک نمایش بصری پویا از کلمات تولید شده برای کمک به پیش‌بینی کلمه بعدی در طول نسل بعدی استفاده کردند. مایو و همکاران [۳] و Karpathy & لی [۵] از RNN برای رمزگشایی عنوان طول متغیر استفاده کردند، در حالی که LSTM در [۴، ۷، ۱۳، ۳۸، ۴۹] برای رمزگشایی شرح تصویر از متن مربوطه خود استفاده شد. بافت به عنوان یک نسل می‌تواند هر یک از آن‌هایی باشد که در بخش ۲.۱ و یا ترکیبی از انواع مختلف توصیف شده‌اند. به عنوان مثال، جیا و همکاران [۳۸] از هر دو تصویر کدگذاری شده و embedding معنایی با استفاده از تجزیه و تحلیل ارتباط استاندارد مرسوم به عنوان ورودی‌های to LSTM استفاده کردند. علاوه بر این، ژو و همکاران [۹]، فو و همکاران [۱۲]، لی و همکاران [۱۱]، و Yang و همکاران [۱۱] مکانیزم مورد توجه را با decoder LSTM برای رسیدگی به بخش‌های مختلف تصویر در طول فرآیند تولید عنوان کردند. از سوی دیگر، شما و همکاران [۱۵] مکانیزم توجه را به جای فضای semantic به جای فضای multimodal در زمان ایجاد عنوان تصویر اجرا کردند.

۲.۳. ارتباط با کار ما

به طور مشابه، مدل ما با استفاده از تصویر کدگذاری شده از سی ان به عنوان متن، از LSTM برای رمز گشایی تصویر استفاده می‌کند. با این حال، به جای استفاده از کلمات tokenized به عنوان واحد اتمی برای یک LSTM متوالی، ما یک ساختار LSTM مراتبی را برای رمزگشایی توصیف تصویر از جمله به جمله، معرفی می‌کنیم. بنابراین، ورودی مدل ما در سطح جمله، توالی از کلمات و عبارات است. از نظر گسترش the از داده‌های متوالی به داده‌های ساختار یافته گراف، مدل ما کمی شبیه to [۵۰] برای تجزیه مقصود معنایی است. با این حال، از مدل نمودار LSTM [۵۰] برای به روز رسانی اطلاعات هر گره گراف براساس گره‌های همسایه آن‌ها در حالی استفاده می‌شود که ساختار هر توپولوژی گراف را حفظ می‌کند. از سوی دیگر، هدف مدل ما ساخت داده‌های ساختار یافته گراف (توصیف زبان طبیعی)از تعدادی از گره‌های unorganized (NPs)است، که در آن توپولوژی گراف در طول استنتاج نامعلوم است. همچنین، کار ما با رویکردهای مبتنی بر اصطلاح متفاوت است که از بازیابی of با استفاده از قالب یا روش ترکیب برای تولید عنوان استفاده می‌کنند (۳۲، ۳۳، ۴۶)، زیرا ما بر بازیابی تکیه نداریم. سایر رویکردهای مبتنی بر عبارت تاکید بیشتری بر یادگیری عبارت و استفاده از یک مدل زبانی ساده برای رمز گشایی جمله دارند. به عنوان مثال، Lebret و همکاران [۳۷] و Ushiku و همکاران [۳۹] انواع مختلفی از عبارت را از توصیف تصویر استخراج کردند. عبارت ربط سابق ارتباط با تصویر با نمونه‌گیری منفی را ربط می‌دهد و یک توالی از عبارات را با استفاده از یک مدل سه گرم زبانی بر روی برچسب chunking هر عبارت رمز گشایی می‌کند.

دومی یک روش تعبیه subspace برای عبارت یادگیری و جمله تولید شده از عبارت‌های تخمینی با استفاده از یک بهینه‌سازی ترکیبی پیشنهاد داد. کار ما از نظر ا، نوع عبارت استخراج‌شده، ۲)روش یادگیری عبارت ()و ۳)روش کدگشایی محکومیت متفاوت است. اول، ما فقط با الهام از داشتن هر عبارت معادل با یک ماهیت درون تصویر، NPs را استخراج می‌کنیم.
علاوه بر این، ما هر دو عبارت خود و AS را با استفاده از LSTM که به صورت سلسله مراتبی متصل شده‌اند را آموزش می‌دهیم که در شکل ا نشان‌داده شده‌است. بنابراین، نمایش عبارت ما از the of the AS به عنوان کدگشا در سطح محکومیت یاد می‌شود.
در نهایت، یک عنوان کامل را با رمز گشایی ایجاد می‌کنیم، در حالی که به تدریج یک اسم استنباط شده را با عبارات تولید شده جایگزین می‌کنیم. اثر اخیر، اسکلت - کلید (۵ا)یک عنوان یک تصویر به تصویر را طراحی کرده‌است که در آن skel - LSTM یاد می‌گیرد تا یک جمله آخر را با کلمه آخر خود ایجاد کند، در حالی که - Attr LSTM می‌آموزد که NPs ها را رمزگشایی کند. کار آن‌ها یک مدل بالا پایین طراحی کرد، که در آن یک جمله اسکلت برای اولین بار ایجاد شد، و بعد با رمزگشایی هر یک از کلمه اسکلت برای تشکیل the ویژگی، دنبال شد.

عنوان بعدی:

پیش‌بینی اکشن از تصاویری از memorizing Hard - to – Predict

پیش‌بینی عملی بر مبنای ویدئو یک مساله مهم در زمینه بینایی رایانه‌ای با بسیاری از کاربردها نظیر جلوگیری از حوادث و فعالیت‌های جنایی است. این کار برای پیش‌بینی اقدامات در مراحل اولیه به دلیل تغییرات بزرگ بین فیلم‌های مشاهده‌شده اولیه و نمونه‌های کامل چالش برانگیز است. علاوه بر این، تغییرات درون کلاسی موجب سردرگمی در پیش‌بینی کنندگان می‌شود. در این مقاله، ما یک مدل mem - LSTM را برای پیش‌بینی اقدامات در مرحله اولیه پیشنهاد می‌کنیم، که در آن یک ماژول حافظه برای ثبت چند نمونه "سخت" و انواع مشاهدات اولیه معرفی می‌شود. روش ما از شبکه عصبی مصنوعی (سی ان ان)و حافظه کوتاه‌مدت (LSTM)برای مدل‌سازی ورودی ویدئوی مشاهده‌شده جزیی استفاده می‌کند. ما LSTM را با یک مدول حافظه افزایش می‌دهیم تا نمونه‌های ویدیویی چالش برانگیز را به یاد آوریم. با مدول حافظه، مدل mem - LSTM ما نه تنها عملکرد موثر را در مرحله اولیه به دست می‌آورد بلکه همچنین پیش‌بینی‌هایی را بدون آگاهی قبلی نسبت به نسبت مشاهده انجام می‌دهد. اطلاعات در فریم های آینده نیز با استفاده از یک لایه دو جهتی of مورد استفاده قرار می‌گیرد. آزمایش‌ها روی مجموعه داده‌های ۱۰۱ - ۱۰۱ و ورزشی - ۱ M نشان می‌دهد که روش ما بهتر از روش‌های هنری است

پیش‌بینی عملی در سال‌های اخیر به دلیل کاربرد گسترده و مهم آن در سناریوهای realworld مانند نظارت دیداری و اجتناب از تصادف، منافع فزاینده‌ای را دریافت می‌کند. متفاوت از بازشناسی در عمل، در پیش‌بینی عمل، برچسب اقدام باید قبل از اینکه کل اجرای عملیات مشاهده شود، استنباط شود. مهم‌تر اینکه، مهم است که یک الگوریتم پیش‌بینی می‌تواند پیش‌بینی‌های دقیق در مرحله شروع یک ویدیو ایجاد کند، به عنوان مثال، زمانی که تنها چند فریم از یک ویدئو دیده می‌شود. اگرچه رویکردهای شناسایی اقدام (Karpathy و سایرین ۲۰۱۴؛ تران و سایرین ۲۰۱۵)موفقیت بزرگی بدست‌آورده اند، پیش‌بینی عملی هنوز یک موضوع تحقیقاتی نسبتا جدید است و هنوز چندین مشکل دارد که باید مورد توجه قرار گیرد. در سال‌های اخیر، تلاش‌های متعددی در پیش‌بینی اقدام انجام‌گرفته است (Ryoo۲۰۱۱ a؛ Cao و همکاران ۲۰۱۳؛ Lan، Chen، Fu and۲۰۱۴؛

هنگ‌کنگ، تائو، و فو ۲۰۱۷)، اما misprediction باقی می‌ماند. یکی از دلایل اصلی این است که در برخی از اقدامات، ویژگی‌ها از چند فریم اولیه به اندازه کافی متمایز نیستند تا به خاطر شباهت بصری طبقه‌بندی شوند. بنابراین، طبقه‌بندی کننده یاد شده با استفاده از این ویژگی‌ها ممکن است قادر به یافتن مرزهای طبقه‌بندی صحیح نباشد و در نتیجه رویکردهای پیش‌بینی شکست خواهند خورد. همانطور که در شکل ۱ نشان‌داده‌شده، کارهایی مانند "بسکتبال" و "dunk بسکتبال" در مراحل اولیه خود ظاهر بسیار مشابهی دارند. شناسایی اطلاعات متمایز برای پیش‌بینی در مراحل اولیه ضروری است. کاره‌ای اخیر در (هنگ‌کنگ، تائو، و فو ۲۰۱۷)نشان می‌دهد که عملکرد پیش‌بینی معمولا زمانی پایدار می‌شود که تنها نیمی از ویدئوها دیده می‌شوند. این نشان می‌دهد که اطلاعات متمایز تشخیصی اغلب در وسط یک ویدئو دیده می‌شود و در نتیجه عملکرد پیش‌بینی را در مراحل اولیه ویدیو محدود می‌کند. کار موجود (هنگ‌کنگ، تائو، و فو ۲۰۱۷)توان تفکیک پذیری ویژگی‌ها را با انتقال اطلاعات آینده از ویدیوهای کامل به ویدیوهای جزیی افزایش می‌دهد. MTSSVM در (کنگ و فو ۲۰۱۶)تکامل عمل را مشخص می‌کند و یک مدل پیش‌بینی غیرخطی با استفاده از هسته‌ها ایجاد می‌کند. روش MSSC ارایه‌شده در (کایو و همکاران ۲۰۱۳)یک نمایش ویژگی جدید با اجرای محدودیت‌های پراکنده بر روی ویژگی‌ها را یاد می‌گیرد. با این حال، عملکرد پیش‌بینی آن‌ها در مرحله اولیه ویدئو (برای مثال، تنها ۱۰ % - ۲۰ % از فریم ها مشاهده می‌شوند)هنوز نسبتا پایین است، اگرچه قدرت تشخیص ویژگی‌ها از این قاب‌ها هم اکنون افزایش یافته‌است. به علاوه، این رویکردها عملی نیستند زیرا آن‌ها نیاز به دانستن نسبت مشاهده یک ویدئوی آزمایشی دارند.

ما mem - LSTM را برای حل مشکلات فوق معرفی می‌کنیم. ما پیشنهاد می‌کنیم که از حافظه برای ذخیره نمونه‌های آموزشی دشوار به منظور بهبود عملکرد پیش‌بینی در مرحله اولیه استفاده کنیم. ماژول حافظه مورد استفاده در این کار، پیش‌بینی‌پذیری هر نمونه آموزشی را اندازه‌گیری می‌کند، و آن نمونه‌های چالشی را ذخیره خواهد کرد. با استفاده از یک ویدئوی جزئی آزمایشی (یک نمونه پرس و جو)، حافظه، شباهت بین نمونه پرس و جو و تمام نمونه‌های ذخیره‌شده را محاسبه می‌کند، و برچسب مطلوب‌ترین را برای پرس و جو بر می‌گرداند.
همانطور که حافظه یک استخر بزرگ از نمونه‌ها را حفظ می‌کند، به ما اجازه می‌دهد تا مرزهای طبقه‌بندی پیچیده را ایجاد کنیم، که به ویژه برای ایجاد ویدیوهای جزیی در مرحله آغازین مفید هستند.
روش ما همچنین از اطلاعات آینده در یک ویدیو برای پیش‌بینی دقیق استفاده می‌کند. با استفاده از یک لایه دو جهتیof ، اطلاعات موجود در وسط و بخش‌های اواخر یک ویدیو به ویژگی‌هایی که از چارچوب‌های اول استخراج شده‌اند، منتشر می‌شود و در نتیجه قدرت تشخیصی ویژگی را بهبود می‌بخشد و عملکرد پیش‌بینی را بالا می‌برد. ما از یک چارچوب دو جریان در این مقاله استفاده می‌کنیم، که در آن جریان‌های RGB و جریان درهم ادغام می‌شوند. در هر جریان، یک شبکه عصبی convolutional برای استخراج ویژگی‌ها از قاب‌ها بکار گرفته می‌شود و سپس در هر فاصله زمانی به یک LSTM داده می‌شود. دو لایه دو جهتی دو جهتی با هم پیش رو و هم ارتباطات پشت به عقب برای مشخص کردن تکامل عملکرد زمانی و ثبت اطلاعات آینده برای پیش‌بینی به کار گرفته می‌شوند. خروجی این دو جریان به یک بردار تبدیل می‌شود و به عنوان ورودی حافظه جهانی برای مقایسه با همه نمونه‌های با پیش‌بینی سخت در نظر گرفته می‌شود. سهم اصلی این مقاله دو برابر شده‌است. ما از یک ماژول حافظه بلند مدت برای به خاطر سپردن مشاهدات دشوار استفاده می‌کنیم. این به ما اجازه می‌دهد تا مرز طبقه‌بندی پیچیده تری را بیاموزیم که به ویژه برای طبقه‌بندی ویدیوهای جزیی با اطلاعات تشخیصی ناکافی مفید است. علاوه بر این، ما از اطلاعات در قالب‌های آینده برای افزایش بیشتر عملکرد پیش‌بینی از طریق یک لایه LSTM دو جهتی استفاده می‌کنیم. این کار اساسا اطلاعات متمایزی را از بخش‌های میانی و اواخر ویدئو به بخش‌های آغازین انتقال می‌دهد. در مقایسه با روش‌های موجود) هنگ‌کنگ، کیت، و فو ۲۰۱۴؛ کایو و همکاران ۲۰۱۳؛ Ryoo ۲۰۱۱(b ، روش ما نیاز به دانستن نسبت مشاهده فیلم‌های تست ندارد و بنابراین عملی‌تر است.
تعاریف ما از تنظیمات مشکل توصیف‌شده در (هنگ‌کنگ، تائو، و فو ۲۰۱۷)پیروی می‌کنیم. برای تقلید از ورود داده‌های متوالی، یک فیلم کامل x با فریم T را به = K ۱۰ قسمت تقسیم می‌کنیم. در نتیجه هر بخش حاوی T فریم است. طول ویدئو T ممکن است برای ویدیوهای مختلف متفاوت باشد در نتیجه موجب طول متفاوتی در بخش‌های آن‌ها می‌شود. برای یک ویدیو از T طول(k) T ، } k(۱، ۰، ((Kشامل فریم هایی است که از قاب k - ۱)-(از چارچوب به هم متصل می‌شوند. یک ویدیو جزیی یا مشاهده جزیی(k) به عنوان یک subsequenceموقتی تعریف می‌شود که شامل بخش‌های آغازین k ویدئو است. سطح پیشرفت g از ویدئوی جزیی(k) x با تعداد بخش‌های موجود در ویدئوی جزیی(k) x تعریف می‌شود .g = k: نسبت مشاهده r از یک ویدئوی جزیی (k)k K: r = k K است.
به رسمیت شناختن عملی یک موضوع تحقیقاتی مهم در بینایی رایانه‌ای است، که بر ویژگی‌های استخراج‌شده از ویدئوهای عملیاتی کامل زمانی متکی است. این ویژگی‌ها مانند نقاط توجه فضا - زمان Laptev) ۲۰۰۵(و

خط سیر فشرده (وانگ و همکاران ۲۰۱۱)از ویژگی‌های spatiotemporal و ویژگی‌های ظاهری محلی تشکیل شده‌اند. در (وانگ و اشمید ۲۰۱۳)مسیر متراکم با استفاده از برآورد حرکت دوربین، لغو نویز مبتنی بر تشخیص و حضور در یک بردار فیشر بهبود پیدا کرد. مطالعات اخیر نشان داد که ویژگی‌های عملی را می توان با روش‌های یادگیری عمیق مانند شبکه‌های عصبی) convolutional سی ان ان)و شبکه‌های عصبی بازگشتی آموخت. شبکه‌های دو جریان Simonyan) و Zisserman ۲۰۱۴(بر روی قاب‌های RGB و چارچوب‌های جریان نوری ساخته شده‌اند و نتایج امیدبخش را در مجموعه داده‌های مختلف عملکردی نشان داده‌اند. در Ranzato) و همکاران ۲۰۱۴؛Srivastava ، Mansimov، و Salakhudinov ۲۰۱۵؛ سانگ و همکاران ۲۰۱۵؛ وو و سایرین ۲۰۱۵(، rnnsبرای مدلسازی همبستگی زمانی بلند مدت در ویدئو استفاده شده‌اند و نمایش ویدیو را برای طبقه‌بندی اکشن ایجاد می‌کنند. با این حال، اغلب این روش‌ها انتظار می‌رود که کنش‌ها را از طریق ویدیوهای کامل زمانی به رسمیت بشناسند. عملکرد آن‌ها در فیلم‌های اقدام ناقص به لحاظ زمانی نامعلوم است.

پیش‌بینی عملی یک وظیفه مهم دیگر برای پیش‌بینی برچسب اقدام است که در یک ویدیو تا حدی مشاهده شده‌است Ryoo .و همکاران Ryoo) ۲۰۱۱(b روش انتگرال - - - کلمات و رویکرد مبتنی بر بسته - کلمات برای پیش‌بینی عمل را پیشنهاد کردند. کایو و همکاران (کایو و همکاران ۲۰۱۳)یک فرمول احتمالاتی برای شناسایی فعالیت انسان از طریق ویدیوهای تاحدی مشاهده کردند. کنگ و همکاران (هنگ‌کنگ، کیت، و فو ۲۰۱۴)یک مدل مقیاس زمانی چندگانه را در چارچوب ماشین بردار پشتیبان برای پیش‌بینی اقدامات ناتمام ارائه کردند. از دیدگاه تعامل اجتماعی مزاحم، Lanو همکارانLan) ، Chen، و savarese ۲۰۱۴"(حرکات سلسله مراتبی" را برای پیش‌بینی عمل توسعه دادند که قادر به ثبت ساختار متداول حرکات انسان قبل از انجام عملیات است.

روش‌های یادگیری عمیق نیز در پیش‌بینی عمل نشان داده شده‌اند Ranzato .و همکاران Ranzato) و همکاران ۲۰۱۴(یک مدل تولیدی با استفاده از the برای پیش‌بینی حرکت در ویدئو معرفی کردند Srivastava .و همکاران Srivastava)، Mansimov، و Salakhudinov ۲۰۱۵(یک رویکرد یادگیری بدون نظارت را با استفاده از LSTM برای پیش‌بینی بخشی از یک عمل پیشنهاد دادند، در نتیجه آن‌ها می‌توانند با استفاده از این ویژگی‌های بخش به خوبی تعریف‌شده، به پیشنهاد پیش‌بینی عمل دستیابند. روش ما نشان نمی‌دهد که چگونه یک عمل در بخش مختلف تکامل می‌یابد؛ در عوض، ما بر به یاد آوردن مرحله اولیه عمل تمرکز می‌کنیم. علاوه بر این، اغلب روش‌های پیش‌بینی عمل نیاز به دانستن نسبت مشاهده یک ویدئوی آزمایشی برای پیش‌بینی یک پیش‌بینی دارند. روش‌های ما نیازی به دانستن نسبت مشاهده ندارند و در نتیجه می‌تواند برای پخش ویدئو استفاده شود.

حافظه کوتاه‌مدت کوتاه‌مدتLSTM) ، Schmidhuber، and Cummins۲۰۰۰(به موفقیت بزرگی در وظایف مختلف یادگیری رشته) فرناندز، گریوز، و Schmidhuber ۲۰۱۷(دست یافته‌است. یک LSTM معمولی دارای سه گیت است که شامل یک گیت ورودی در آن، یک گیت forget فوت و یک gate خروجی است. این سه دروازه اصولا واحدهای جمع جمع غیر خطی هستند. از گیت ها برای محاسبه activations از داخل و خارج از بلوک LSTMاستفاده می‌شود و فعال‌سازی سلول از طریق ضرب را مدیریت می‌کند.

هنگامی که LSTM چند لایه آموزشی را آموزش می‌دهیم، مشخص کردیم که همگرا شدن سخت است و معمولا به عملکرد آسیب می‌رساند. برای حل این مشکل، از دو استراتژی برای مقابله با آن استفاده می‌کنیم و بلوک ساختاری اصلی مورد استفاده در این تحقیق را ایجاد می‌نماییم. اول، ما LSTM را با یک ارتباط باقیمانده ترکیب می‌کنیم (او و همکاران ۲۰۱۵)تا فرآیند یادگیری را تسهیل کنند. یک اتصال باقی مانده به LSTM اضافه می‌شود تا ورودی به خروجی اضافه شود. افزودن اتصال باقیمانده به کم کردن overfitting و بهبود دقت پیش‌بینی کمک می‌کند. به طور رسمی، اضافه کردن یک اتصال باقی مانده به LSTM را می توان از طریق (شکل ۳)نشان داد:

هنگامی که LSTM چند لایه آموزشی را آموزش می‌دهیم، مشخص کردیم که همگرا شدن سخت است و معمولا به عملکرد آسیب می‌رساند. برای حل این مشکل، از دو استراتژی برای مقابله با آن استفاده می‌کنیم و

بلوک ساختاری اصلی مورد استفاده در این تحقیق را ایجاد می‌نماییم. اول، ما LSTM را با یک ارتباط باقیمانده ترکیب می‌کنیم (او و همکاران ۲۰۱۵)تا فرآیند یادگیری را تسهیل کنند. یک اتصال باقی مانده به LSTM اضافه می‌شود تا ورودی به خروجی اضافه شود. افزودن اتصال باقیمانده به کم کردن overfitting و بهبود دقت پیش‌بینی کمک می‌کند. به طور رسمی، اضافه کردن یک اتصال باقی مانده به LSTM را می توان از طریق (شکل ۲)نشان داد:

با این حال، در آزمایش ما، یک اتصال باقی مانده در جریان جریان ضعیف عمل می‌کند. عملکرد of با تنظیم‌کننده نسبت به مدل اصلی LSTM پایین‌تر است. این به دلیل ماهیت داده‌های جریان است. بر خلاف تصاویرRGB ، تصاویر جریان واریانس بسیاری بین دو فریم متوالی دارند که تنظیم‌کننده را کم‌تر مفید می‌سازد.

شکل ۳: یک ماژول حافظه، ویدیوهای جزیی و فیلم‌های کامل از همان گروه اقدام در یک محله محلی را به یاد می‌آورد، در نتیجه اجازه یک ویدیو آزمایشی در مراحل مختلف را می‌دهد تا یک برچسب اقدام مشابه را پیدا کند.

روش بعدی

تحلیل احساسات مبتنی بر ویدیو با hvnLBP - TOP و دو – LSTM

چکیده

در این مقاله، ما یک روش استخراج ویژگی جدید به نام hvnLBP - TOP برای آنالیز احساسی مبتنی بر ویدئو ارائه می‌کنیم. به علاوه، ما از تحلیل اجزای اصلی (PCA)و حافظه کوتاه‌مدت دو سویه (bi)برای کاهش ابعاد و طبقه‌بندی استفاده می‌کنیم. ما به یک صحت تشخیص میانگین of % در مجموعه داده moud و ۶۳.۹ % در مجموعه داده CMU - MOSI دست یافته‌ایم.

به طور فزاینده‌ای multimodal مقدمه با پیشرفت وب سایت‌های اشتراک ویدیو و برنامه‌های شبکه اجتماعی، آنالیز احساسی در میان پژوهشگران محبوب شده‌است. علاوه بر اطلاعات از زبان طبیعی، اطلاعات بصری دارای ویژگی‌های احساسی مهم در در ویدئو از (fer) حالات گوینده و حالات چهره است. از این رو، تحلیل احساساتی مبتنی بر ویدئو و تشخیص چهره صورت برخوردار است. بسیاری از کاره‌ای قبلی با تشخیص چهره چهره در تصاویر multimodal اهمیت زیادی در تحلیل احساسات برای توصیف ویژگی هندسی (der ۱۹۹۸ و فون Nevenهنگ‌کنگ،) مرتبط هستند. در مراحل اولیه، نشانه‌های صورت برای شناسایی بیان (Pietikainen ۲۰۰۶ و Hadid، Ahonen) LBP چهره‌ها استفاده شده‌است. بعدا ویژگی‌های بافت شامل و همکاران Mistry] می‌ستری)(hvnLBP) LBP است، و مقایسه افقی و افقی advanced دارای مشتقات LBP .استفاده شده‌است ۲۰۱۷)پیشرفته‌ترین در میان آن‌ها است. این ویژگی‌ها هر پیکسل را با پیکسل های مرزی خود مقایسه می‌کنند و هیستوگرام

الگوی از چنین نتایج مقایسه‌ای را تولید می‌کنند. در سال‌های اخیر، تکنیک‌های یادگیری ماشین مانند شبکه عصبی سطحی و همکاران ۲۰۱۵)برای استخراج ویژگی‌های احساسی مورد استفاده قرار گرفته‌اند. علاوه بر Burkert)(سطحی (سی ان ان Pietikainen ژایو و) LBP - TOP ویژگی‌های چهره، آنالیز احساسی مبتنی بر ویدئو از ویژگی‌های تصویری شامل استفاده می‌کند. بر خلاف ویژگی‌های مبتنی بر تصویر، فرصت‌های قابل‌توجهی برای بهبود دقت در اعمال ویژگی‌های (۲۰۰۷ و hand - مبتنی بر ویدئو وجود دارد. این مقاله یک ویژگی جدید برای تحلیل احساسات مبتنی بر ویدئو ارائه می‌کند. ما ویژگی را ایجاد کنیم. با استفاده از hvnLBP - TOP ترکیب کردیم تا ویژگی(TOP) رویکرد الحاق ویژگی‌ها را در سه صفحه متعامد را برای LSTM bidirectional را تا ۵۱۲ کاهش می‌دهیم. در نهایت، ما معماری of ، ما طول(PCA) تحلیل اجزای اصلی ، Rosas - پرز) moud طبقه‌بندی احساسات و رگرسیون انتخاب می‌کنیم. ما ویژگی‌های خود را در مجموعه داده‌های و همکاران ۲۰۱۶)بررسی کردیم و نتایج نشان می‌دهند CMUMOSI dataset (Zadeh و(۲۰۱۳ Morency و Mihalcea .که مدل ما دقت و کارایی بهتری نسبت به روش‌های دیگر دارد

فصل سوم: انواع روش ها

روش اول از LSTM سلسله مراتبی استفاده کرده است و حالا بررسی کنیم که این روش چیست .

در این روش گفته که ما برای اینکه یک توصیف گر تصویر داشته باشیم یک چالشی در ابتدا داریم اینکه نکات تصویری و بینایی رو باید به یک زبان تبدیل کنیم و به عبارتی ارتباط این دو رو فراهم کنیم و برای این هدف دوتا علم بینایی ماشین و پردازش و ارتباط زبان طبیعی مورد استفاده همزمان قرار میگیره .

تو سال های اخیر دوتا زیر شبکه معرفی شده از شبکه عصبی که ابتدا CNN هست که مخفف coinvulutional neural network هست که برای رمزنگاری تصاویر استفاده میشه و به بردار ویژگی ها تبدیل می کنه و دومین مورد که پیشرفت خوبی تو این سال ها داشته RNN یا recurrent neural network هست که رمزگشایی انجام میده و به توصیفات زبان طبیعی تبدیل می کنه . حالا تو شبکه بازگشتی ها یک معماری خیلی معروفی داریم بنام LSTM که مخفف از long short Term memory هست و تو این معماری بازگشتی مشکل یادآوری اطلاعات قدیمی رو حل می کند . این چارچوب lstm تو سال های

اخیر خیلی تغییر کرده و ساختارهای متفاوتی ازش معرفی شده و ساختار پایه آن توانایی قبت وابستگی بلند مدت در کنار حفظ توالی رو داره .اگر چه این ساختار ترتیب که داریم بسیار مفید هست چون داده ها رو به صورت پشت سر هم پردازش می شوند اما مشکل که ما داریم اینکه برای ساختار نحوی جملات ما باید به نکات بیشتری دقت کنیم و نمیشه پشت سر هم رو به عنوان یک جمله از نظر نحوی درست معنا کرد پس در این مقاله سعی شده از ساختار سلسله مراتبی استفاده بشه و این اطلاعات بصورت یک سلسله مراتب در تمام طول زمانی سلسله مراتبی بشن و اگه زبان انگلیسی رو مثال بزنیم پایین ترین سطح میشه کاراکترهایی که از کوتاه ترین زمان بدست میاد که از آن کلمات ، عبارات ، بندها ، جملات و اسناد را دنبال می کنند . بنابراین غیر قابل انکار هست که ساختار جمله یکی از مهمترین و برجسته ترین ویژگی های زبان هست و برای مثال victor yngve یکی از نویسندگان تاثیرگذار در تیوری زبانی در سال ۱۹۶۰ بیان می کنه که ساختار زبان از یک سلسله مراتب تشکیل شده و برای توصیف گر تصویر اگر ما یک ساختار سطح بالا رو ابتدا ایجاد کنیم عملکرد ما بسیار محدود میشه و میشه برای مثال دو سطر دو دیتاست های Flickr30k و Flickr8k و MS-coco یکسان هست پس می توان به این موضوع رسید که ما جملات از پیش آماده رو هم می تونیم استفاده کنیم و نکته مهم بعدی ساختار کلمات هست که می تواند یک کلمات توصیف گر یک جمله کامل شود . پس تا اینجا فهمیدیم جملات بصورت توالی معنی نمی شوند و نیازمند یک سلسله مراتب هستند که هرچقدر این سلسله مراتب گسترده تر باشه میزان عمق رو می توان بهتر فهمید .

ما اینجا می خواهیم ساختاری ایجاد کنیم که برخلاف مدل های که بصورت ترتیبی این عمل رو انجام میدن در این کار به صورت سلسله مراتبی صورت می گیره و نام این الگوریتم جدید رو که بر مبنای lstm می باشد رو phi-LSTM گذاشتیم

حالا ساختار کلی به این شکل هست که ما میایم تک تک کلمات رو از فریم به فریم استخراج می کنیم و با این حال با این روش cnn این تعداد کلمات زیاد می شوند و آن ها رو به عنوان ذرات اتم در نظر میگیریم و بعد این کلمات که تعدادشون زیاد هست مثلا در یک تصویری کلماتی مثل دوچرخه موتورسیکلت و غیره استخراج میشه و با استفاده از ساختار سلسله مراتبی این کلمات کدگذاری می شوند و عبارات را میسازند که این عبارات نسبت به کل کلمات بدست اومد خلاصه تر هستند و ساختار خلاصه گونه تری دارند .

پس این تحقیق از دو ساختار به صورت خلاصه تشکیل شده ابتدا ما ساختاری ایجاد می کنیم که مدلی سلسه مراتبی برای رمزگشایی عنوان تصویر یا توصیف تصویر بدهد و در قسمت دوم نشان می دهیم که توصیفات تصویر ایجاد شده با الگوریتم ما یا همان phi -LSTM از نظر دقت میزان بیشتری می باشد و به صورت یک رمان که اطلاعاتش از قبل آموزش داده نشده و اطلاعات تازه داره نمایش میدهد .

نکته ای که کمک کننده هست ودر نسخه اولیه این کار ارائه شده اما مشکل اینه که در حالت قبلی کلمات که معنا بده را پیشبینی می کرد اما در اینجا به این صورت هست که ساختار سلسله مراتبی این مفاهیم رو ایجاد می کنه و نکته دوم اینکه طول جملات نرمالیزه شده در دو حالت سطح عبارت کل و جمله و میشه کپشن های طولانی تری را تولید کرد و سوما خروحی های ابزار تجربه را با یک استراتژی اصلاح بهبود دادیم و نهایتا تحلیل های جدید و توضیحات شهودی به نتایج ما اضافه می شوند .

و ما ازمایش خود را روی دیتا ست های MS-coco انجام میدیم و نتایج خود را بر اساس چهار معیار ارزیابی بررسی می کنیم به نام های cide rouge meteor spice

در مقاله با موضوع انتقال کامل ویدیو با سنگ دانه‌های درشت با نما - به - زیبا و جالب توجه از روش زیر استفاده کرده است .

- مدل مبتنی بر RNN Revisit
از سی ان ان D ۲ / ۳ - D به عنوان کدگذار ویدئو استفاده می‌شود. هر فریم از یک ویدئو به عنوان ویژگی بعدی کدگذاری می‌شود یعنی { }، که در آن تعداد فریم ها مشخص می‌شود و فریم n - فریم ویدئو است. روش متداول برای به دست آوردن نمایش ویدیو این است که به طور متوسط از بردارهای مشخصه قاب استفاده کنیم. با این حال، استراتژی تجمع متوسط، اطلاعات موقتی را در بین قاب‌ها اعمال می‌کند. به طور موثر، LSTM را می توان به عنوان کدگذار ویدئو انتخاب کرد. حالت پنهان فعلی می‌تواند به صورت زیر به‌هنگام شود:

که در آن حالت پنهان قبلی وجود دارد و ویژگی چارچوب ورودی در مرحله زمانی فعلی است. آخرین حالت پنهان کننده LSTM می‌تواند به عنوان نمایش ویژگی جهانی در نظر گرفته شود، و سپس به رمزگشا داده می‌شود. Decoder دیگر به عنوان رمزگشا مورد استفاده قرار می‌گیرد، که توسط نمایش ویدئو راه‌اندازی می‌شود. هر کلمه در یک عنوان به یک کلمه تعبیه شده‌است. کل جمله را می توان به صورت یک توالی { } نمایش داد. در نهایت، عنوان خروجی را می توان براساس معادله زیر ایجاد کرد (۳).

که در آن کلمه ورودی در مرحله زمانی فعلی تعبیه شده‌است. هنگامی که حالت پنهان در هر مرحله زمانی بدست می‌آید، کلمه متناظر در زیر را می توان تولید کرد. فقدان لگاریتم منفی یک جمله با تجمیع احتمالات لگاریتم نسبت به کلمات در جمله داده می‌شود، که به صورت زیر تعریف می‌شود:

که در آن طول جمله، نمایش ویدیو است؛ و کلمه تولید شده در مرحله زمانی فعلی است.
مدل کاملا convolutional ما با توجه
اگرچه مدل LSTM می‌تواند توصیفات زبان طبیعی برای ویدئوها ایجاد کند، اما مشکلات of هنوز وجود دارند. برای حل مشکلات، یک چارچوب جدید پیشنهاد می‌کنیم که شبکه کاملا convolutional را با هم ترکیب می‌کند (یک مدل نسل جدید برای فیلم‌های ویدیویی بدون کمک of)، یک مکانیزم توجه جدید (یک استراتژی محاسبه جدید برای وزن قاب در سطح منطقه). یک نمای کلی از مدل کاملا convolutional ما با توجه در شکل ۲ نشان‌داده شده‌است.

شبکه کاملا متصل شونده
همان طور که قبلا توضیح داده شد، ویژگی‌های یک ویدئو را می توان به صورت { } نمایش داد، و کلمه embeddings یک عنوان متناظر را می توان به شکل { } نمایش داد. ویژگی‌های تمامی فریم ها به بخش ورودی اولیه شبکه کامل convolutional و ماژول توجه داده خواهد شد. برای اینکه ورودی گشتاور اولیه حاوی اطلاعات معنایی و هم دیداری باشد، ورودی اولیه را می توان براساس معادله زیر بدست آورد (۴).

که در آن، کلمه گنجاندن کلمه در جمله در مرحله زمانی فعلی را مشخص می‌کند؛ و نشان‌دهنده ویژگی جهانی است که با در نظر گرفتن میانگین ویژگی‌های قاب‌های نمونه به دست می‌آید. بنابراین، الحاق عبارت embeddings و ویژگی‌های جهانی یک ویدیو است. همانطور که در شکل ۲ نشان‌داده‌شده، ساختار اصلی مدل ما لایه‌های انباشته از یک - دی ان ان است. اندازه هسته هر هسته convolutional وجود دارد. قسمت دوم هسته با صفر پوشیده می‌شود، چون کلمه embeddings متناظر با مراحل زمانی بعدی در مرحله زمانی فعلی در دسترس نیست. یک هسته سی ان ان می‌تواند k ویژگی ورودی را دریافت کند، یعنی، ویژگی ورودی فعلی در وسط هسته است، قسمت چپ کلمه تعبیه مراحل زمانی را دریافت می‌کند، و قسمت راست با صفر پوشش داده می‌شود. با توجه به ویژگی‌های یک شبکه سی ان ان، هر هسته لایه سطح بالا می‌تواند اطلاعات بیشتری را پردازش کند چون تعداد لایه انباشت شده افزایش می‌یابد. بنابراین ویژگی‌های خروجی آخرین لایه برای ساختار انباشته را می توان در نظر گرفت تا شامل تمام اطلاعات کلمه تعبیه ارائه‌شده در لایه اول باشد. شکل نهایی این ویدیو می‌تواند به وسیله شبکه کاملا convolutional تولید شود

به نظر جالب و جالب توجه کنید
از آنجا که ساختار مدل ما بسیار متفاوت از مدل مبتنی بر مدل است، مکانیسم‌های توجه جدید براساس ویژگی‌های ساختار ما طراحی شده‌اند. در این بخش ما دو مکانیزم مورد توجه را معرفی خواهیم کرد. اولی، توجه coarse است، که برای ارایه اطلاعات مختلف برای لایه‌های مختلف مورد استفاده قرار می‌گیرد. دومین مورد توجه موروثی است، که برای دستیابی به یک نمایش بصری دقیق‌تر مورد استفاده قرار می‌گیرد.

به دلیل ساختار انباشته یک شبکه کاملا convolutional، که در آن ورودی هر لایه از خروجی لایه قبلی می‌آید، ما می‌توانیم کل فرآیند را به عنوان یک بهینه‌سازی پیوسته از جمله ایجاد شده توسط لایه قبلی در نظر بگیریم. برای استفاده کامل از دانش فرا گرفته‌شده در هر لایه، یک توجه بسیار دقیق که با معماری سلسله مراتبی مدل ما سازگار است، برای کمک به مدل تمرکز بر فریم های برجسته و مناطق مورد استفاده قرار می‌گیرد.

هدف خشن ما فراهم آوردن اطلاعات ضروری برای لایه‌های مختلف است. آن شامل توجه موقتی و توجه به ارثی است. توجه زمانی می‌تواند اطلاعات بصری سطح فریم را فراهم کند، و توجه موروثی، اطلاعات بصری سطح ناحیه را فراهم می‌کند. با افزایش تعداد یک لایه انباشت شده، دقت بسیار زیاد می‌تواند اطلاعات بصری بیشتری را از یک ویدئو پیدا کند. هر لایه از مدل جمع‌آوری‌شده ما می‌تواند از اطلاعات دقیق تری نسبت به قبل استفاده کند تا خروجی‌های تولید شده توسط لایه قبلی را با توجه دقیق بهینه کند.

توجه موقتی به محاسبه وزن همه فریم های نمونه‌گیری شده از یک ویدئو براساس درجه اهمیت فریم های مختلف در مراحل زمانی مختلف و ارائه نتایج نهایی به لایه مربوطه هدف‌گذاری می‌کند. این روش در لایه‌های قبلی مدل استفاده می‌شود. در اینجا ما Multihead مراقبت را اتخاذ کردیم (Vaswani و همکاران، ۲۰۱۷)تا وزن‌های مورد نیاز را بدست آوریم. نتیجه توجه موقتی را می توان براساس معادله زیر محاسبه کرد (۵):

که در آن ورودی در مرحله زمانی وجود دارد؛ { } به ترتیب ویژگی‌ها و فریم های ورودی و فریم در همان فضای ویژگی هستند؛ و تعداد فریم های نمونه است. نمایش نهایی اطلاعات بصری پس از استفاده از توجه، و نشان‌دهنده وزن تمامی فریم های یک ویدئو در مرحله زمانی فعلی است. توجه چند سطحی به این مدل اجازه می‌دهد تا به طور مشترک در این اطلاعات از subspaces نمایش مختلف در موقعیت‌های مختلف شرکت داشته باشند. الحاق همه نتایج of را می توان به عنوان مجموعه‌ای از نمایش‌های مختلف یک ویدیو در نظر گرفت. همانطور که در معادله (۵)نشان‌داده‌شده، نمایش الحاق ویژگی‌هایی تولید شده توسط بسیاری از سره‌ای متفاوت است. سپس مدل ما می‌تواند تمام ویژگی‌ها را یاد بگیرد و نمایش نهایی را از الحاق به یک ویدئو به دست آورد. توجه چند سطحی می‌تواند نمایش‌های چندگانه را به دست آورد که بر روی ویژگی‌های مختلف تمرکز دارند، چون آن‌ها از اتصالات کامل استفاده می‌کنند که مستقل از هم در هر سر هستند. همانطور که تعداد یک لایه انباشت شده افزایش می‌یابد، وزن فریم های مختلف می‌تواند به طور پیوسته بهینه‌سازی شود. بنابراین، وزن نهایی فریم ها بهتر از ساختار تک لایه هستند.

توجه موقتی می‌تواند اطلاعات بصری سطح فریم را در لایه‌های قبلی فراهم کند، اما چنین توجه سطح فریم ممکن است جزئیات منطقه را نادیده بگیرد. برای تاکید بیشتر بر روی مناطق بصری دقیق‌تر محلی و کسب اطلاعات معنادار بصری برای لایه‌های بالاتر، مهم است که توجه موروثی را معطوف به تمرکز بر روی اطلاعات بصری در سطح منطقه در چند لایه بعدی کنیم.

Inherited توجه برخی پژوهشگران هر دو سطح چارچوب و توجه سطح منطقه را در captioning های ویدیویی در نظر گرفته‌اند (Li et al.، ۲۰۱۷). با این حال، آن‌ها به طور کامل از رابطه بین سطح چارچوب و اطلاعات سطح منطقه استفاده نمی‌کنند و نمی‌توانند از دانش بدست‌آمده در لحظه قبل به دلیل محدودیت‌های مدل خود بهره‌برداری کنند. برای حل مشکلات فوق، ما توجه موروثی را برای محاسبه مقادیر سطح ناحیه از اطلاعات بصری پیشنهاد می‌کنیم. نمایش‌های بصری در سطح فریم و سطح کاملا متفاوت هستند. نمایش سطح منطقه، توصیف دقیق‌تر چارچوب است و حاوی اطلاعات دقیق‌تر از نمایش سطح

چارچوب است. این بدان معنی است که نمایش سطح منطقه در زمان‌های مختلف باید ابتدا با وزن نمایش در سطح کادر مطابقت داشته باشد. توجه Inherited در لایه‌های بعدی بعد از لایه‌های با توجه زمانی مورد استفاده قرار می‌گیرد، به طوری که ما می‌توانیم از دانش بدست‌آمده در مورد چارچوب‌ها برای محاسبه وزن مناطق استفاده کنیم. نمایش و وزن مناطق در زمان‌های مختلف را می توان به صورت زیر محاسبه کرد:

که در آن ورودی در مرحله زمانی است؛ { } ویژگی‌های مناطق را مشخص می‌کند؛ تعداد فریم های نمونه‌برداری شده برای ویدئو، و وزن قاب فریم در مرحله زمانی است. نمایش نهایی هر دو وزن‌ها را در سطح کادر و آن‌هایی که در سطح منطقه محاسبه می‌شوند، در نظر می‌گیرد. بنابراین، مناطق توجه نهایی دقیق‌تر هستند، که می‌تواند اطلاعات مفید تری را برای لایه‌های بالاتر فراهم کند و سپس جملاتی تولید کند که محتوای ویدئو را دقیق‌تر توصیف می‌کنند.

تولید عنوان
تابع زیان متداول برای ویدیو captioning در بخشی از مدل مبتنی بر RNN معرفی شده‌است. با این حال، مدل کاملا convolutional ما اطلاعات بیشتری را در نظر می‌گیرد. بنابراین ما به یک تابع از دست رفتن جدید نیاز داریم که تنها نمی‌تواند مدل را هدایت کند تا captions منطقی ایجاد کند، بلکه باعث می‌شود که این مدل توجه را به ارزشمندترین مناطق معطوف کند. بنابراین مدل ما با به حداقل رساندن تابع زیان زیر آموزش‌دیده است:

که در آن بخش اول تابع لگاریتم منفی است که در معادله (۳)ذکر شد؛ و وزن‌های فریم - th در بازه زمانی مختلف است، که می‌تواند مدل را تشویق کند که بر روی همان چارچوب یا همان ناحیه از ویدیو در مراحل مختلف تمرکز نکند؛ و دو پارامتر از پیش تنظیم‌شده برای اطمینان از این که شکست احتمالی لگاریتم اتلاف به بخش عمده‌ای از افت نهایی کمک می‌کند در حالی که بخش‌های دیگر فعال هستند.

ادغام هر دو نشانه‌های بصری و تصویری برای عنوان ویدئو گسترش‌یافته

روش‌ها
در این بخش ما ابتدا چارچوب اصلی را معرفی می‌کنیم که کار ما مبتنی بر آن است. سپس، سه استراتژی ترکیب چند multimodal به ترتیب برای caption ویدئو به تصویر کشیده می‌شوند. در همین حال، چارچوب ادغام ویژگی فضایی multimodal و مولفه‌های اصلی آن نیز ارایه شده‌اند.

چارچوب عنوان اصلی عنوان
چارچوب اصلی تصویر ما از VTTS (توالی تا توالی: ویدئو تا متن)و ۳M (مدل مدل‌سازی حافظه multimodal)بسط داده شده‌است (Wang و همکاران ۲۰۱۶)، که در شکل ۱ نشان‌داده شده‌است. همانطور که در شکل ۱ نشان‌داده شده‌است، مرحله کدگذاری، ویژگی‌های بصری و مرحله رمزگشایی را رمزگذاری می‌کند. به طور خاص، ورودی‌های ویژگی دیداری توسط لایه فوقانی LSTM (سبز رنگی)ساخته می‌شوند. حافظه mulitimodal میانی (فیروزه‌ای رنگی)با استفاده از روش‌های بصری و متنی به اشتراک گذاشته می‌شود. زبان توسط حافظه multimodal bottom LSTM (قرمز رنگ)مدل‌سازی می‌شود، که بر ورودی توالی متن و اطلاعات خواندن از حافظه mulitimodal مشروط شده‌است. برچسب‌ها در شکل ۱ به ترتیب نشان‌دهنده آغاز و انت‌های دوره محکومیت هستند. اشاره می‌کند که هیچ ورودی در مرحله زمانی متناظر وجود ندارد. علاوه بر این، خطوط آبی / organge رنگی نوشته / خواندن / خواندن را به / از حافظه نشان می‌دهند.

multimodal Fusion Fusion Feature concatenating concatenating بصری و صوتی در این بخش، ما دو روش الحاقی متفاوت را پیشنهاد می‌کنیم و آن‌ها را در شکل ۲ ارایه می‌دهیم. به طور خاص، یک روش الحاقی در شکل ۲ (a)نشان‌داده شده‌است. قبل از کدگذار، جفت ویژگی‌های صوتی - تصویری از کلیپ های ویدئویی متناظر به طور مستقیم با هم الحاق می‌شوند. سپس ویژگی‌های الحاق به کدگذار LSTM ارسال می‌شوند. یک روش الحاقی دیگر در شکل ۲ (b)ارایه شده‌است. زوج‌های ویژگی‌های تصویری - تصویری ابتدا به طور جداگانه به the متناظر ارسال می‌شوند. سپس آخرین حالت‌های پنهان این دو LSTM با هم الحاق می‌شوند.

به اشتراک گذاشتن اوزان در میان قالب‌های صوتی - تصویری، اگرچه الحاق برای ترکیب ویژگی‌های تصویری و تصویری موثر است، اما نمی‌تواند اطلاعات رزونانسی را در بین آن‌ها به خوبی ثبت کند. برای رسیدگی به این مشکل، یک LSTM چند multimodal را از طریق به اشتراک گذاری وزن‌ها در حوزه‌های تصویری و تصویری برای caption ویدیویی پیشنهاد می‌کنیم. چارچوب این کدگذار LSTM multimodal در شکل ۳ (a)نشان‌داده شده‌است و به صورت زیر فرموله شده‌است:
در جاییکه آن، ft، ot و ct ورودی ورودی هستند، گیت خروجی، گیت خروجی و محتوای حافظه به روز شده را به ترتیب به ترتیب به ترتیب به ترتیب زیر ترتیب می‌دهند. s = ۰، (۱)- (۱)- (۶)مشخصه بصری LSTM (LSTM)و xs t یک ویژگی بصری استخراج‌شده توسط cnns است. s = ۱، (۱)- (۱)- (۶)، کدگذار مشخصه صوتی LSTMbased و t MFCC صوتی (mel - فرکانس cepstral)را نشان می‌دهد. به علاوه، Ws (s = ۰، ۱)ماتریس‌های وزن برای وارد کردن ویژگی‌های سمعی و بصری هستند. U ماتریس وزنی است که توسط حالت‌های مخفی of صوتی و تصویری به اشتراک گذاشته می‌شود. bs = ۰، ۱)جهت گیری مربوطه هستند.

a. حافظه خارجی حافظه خارجی بکاررفته در مقاله ما به عنوان یک ماتریس M - RK، که در آن K تعداد عناصر حافظه است، و D بعد از هر عنصر حافظه است. در هر گام زمانی t، خروجی t و سه بردار، از جمله کلیدهای ارزش کلید، بردار را پاک کرده و بردار را به ترتیب به ترتیب به ترتیب با LSTM encoders بصری و صوتی منتشر نمی‌کنند. آن‌ها را می توان با جایی محاسبه کرد که در آن s) e e e = ۰، ۱)وزن‌ها و بایاس به ترتیب برای عبارات مربوطه است.

ب. اطلاعات خواندن از حافظه خارجی ما این روش را به صورت زیر تعریف می‌کنیم:
که در آن superscript ها، دنباله‌های ورودی دیداری و صوتی را نشان می‌دهند، s = ۰، ۱)به ترتیب بردارهای خواندن برای جریان‌های تصویری یا صوتی را نشان می‌دهد، که تصمیم می‌گیرد که چه مقدار اطلاعات از حافظه خارجی خوانده خواهد شد. هر عنصر αt، k در αt را می توان از طریق محاسبه زیر بدست آورد:
که در آن g ()تابع معیار شباهت است که برای محاسبه شباهت بین هر عنصر حافظه و کلیدهای ارزشی کلیدی در زمان t استفاده می‌شود. در اینجا ما تابع متریک شباهت cosine را اعمال می‌کنیم.

3 - اطلاعات مربوط به اطلاعات بیرونی و داخلی پس از بدست آوردن اطلاعات از rt نتایج بیرونی، استراتژی ترکیب عمیقی که در مقاله پیشنهاد شده‌است (لیو، Qiu، و هوانگ ۲۰۱۶)به ترتیب برای ادغام جامع rt در حافظه‌های درونی و صوتی تصویری استفاده می‌شود. به طور دقیق، حالات hs t و سمعی و بصری، نه تنها بر روی حافظه داخلی، بلکه بر روی اطلاعات برای خواندن از حافظه خارجی، که می توان آن‌ها را از طریق آن محاسبه کرد.
که در آن Wl نشان‌دهنده ماتریس پارامتر است، که شامل گیت ترکیب می‌شود، که کنترل می‌کند که چه مقدار اطلاعات از حافظه خارجی به حافظه ادغام‌شده منتقل می‌شود و می توان از طریق آن به دست آورد.
که در آن Wp و Wq ماتریس‌های ضابطه مشابه هستند.

د. به‌هنگام‌سازی حافظه حافظه از طریق رونده‌ای زیر به روز می‌شود:
که در آن [t] e t [t] / [t] t و a• a] lt] به ترتیب حذف و اضافه کردن بردارهای ساطع شده توسط کدگذار صوتی / تصویری هستند. به‌هنگام‌سازی نهایی حافظه ترکیبی از حافظه به روز رسانی شده از جریان‌های تصویری و تصویری است. پارامتر P در مجموعه اعتبار سنجی تنظیم شده‌است.

چارچوب ارتباطی صوتی تصویری
برای اینکه هنوز از ویژگی‌های صدا استفاده کنیم، حتی هنگامی که این modality وجود ندارد، ما یک چارچوب استنباطی با کیفیت صدا(AMIN) را توسعه می‌دهیم AMIN .در شکل ۴(a) ارائه شده‌است AMIN .را می توان به صورت زیر فرموله کرد:
که در آن x ویژگی تصویر را نشان می‌دهد و y ویژگی صوت تولید شده را مشخص می‌کند. در این حالت، مجموعه پارامتری در مرحله encoding / decoding نمایش داده می‌شود و ϑ / θ مجموعه پارامتری از مرحله encodering / deconding‌است. ما از ۲ محدودیت به عنوان فقدان آموزش برای مدل AMIN استفاده می‌کنیم که به عنوان LAM ها در نظر گرفته می‌شود و فرموله می‌شود:
که در آن y ویژگی تاثیر MFCC به زمین است.

چارچوب ارتباطی صوتی تصویری
برای اینکه هنوز از ویژگی‌های صدا استفاده کنیم، حتی هنگامی که این modality وجود ندارد، ما یک چارچوب استنباطی با کیفیت صدا(AMIN) را توسعه می‌دهیم AMIN .در شکل ۴(a) ارائه شده‌است AMIN .را می توان به صورت زیر فرموله کرد:
که در آن x ویژگی تصویر را نشان می‌دهد و y ویژگی صوت تولید شده را مشخص می‌کند. در این حالت، مجموعه پارامتری در مرحله encoding / decoding نمایش داده می‌شود و ϑ / θ مجموعه پارامتری از مرحله encodering / deconding‌است. ما از ۲ محدودیت به عنوان فقدان آموزش برای مدل AMIN استفاده می‌کنیم که به عنوان LAM ها در نظر گرفته می‌شود و فرموله می‌شود:
که در آن y ویژگی تاثیر MFCC به زمین است.

چارچوب یکپارچه Fusion امکانات
هنگامی که AMIN بخوبی آموزش‌دیده است، یک چارچوب ترکیب ویژگی دینامیک می‌تواند با ترکیب AMIN با استراتژی‌های ترکیب ویژگی پیشنهادی ما بدست آید، که در شکل ۴(b) ارائه شده‌است. در مورد ویدیوهایی که هم دنباله‌های تصویری و هم تصویری هستند، می‌توانند مستقیما به چارچوب ترکیب اطلاعات چند multimodal (arrows)جامد(ارسال شوند. اگر این ویدیو فقط یک توالی بصری داشته باشد، مدل AMIN برای ایجاد ویژگی‌های صوتی براساس کلیپ ویدیویی متناظر استفاده می‌شود، پس از آن ویژگی‌های صوتی تصویری و ایجاد شده به مدل ترکیب اطلاعات چند multimodal (arrows نقطه‌چین(فرستاده می‌شوند.

آموزش و استدلال
فرض کنید که تعداد موارد زیر نوشته شده(xi ، N(yiو طول برچسبyi ، theاست که میانگین مجموعه داده مجموعه مجموعه تنظیمات را با تابع هدف ما ترکیب می‌کند.
که در آن yi X برای نشان دادن واژه ورودی بکار می‌رود، λضریب تنظیم و θ نشان می‌دهد که همه پارامترها باید در مدل بهینه‌سازی شوند. تنها به عنوان اغلب مدل‌های زبانی، یک لایه softmax برای مدل‌سازی توزیع احتمال بر روی کل دایره لغات بعدی بکار گرفته می‌شود.

که در آنWx ، Wh ، Wy، Up وbη پارامترهای مورد نیاز برای بهینه‌سازی هستند. بسته به توزیع احتمال، توالی کلمه yt می‌تواند به صورت بازگشتی نمونه‌برداری شود تا با پایان نماد در واژه‌نامه مواجه شود. با در نظر گرفتن این نسل، یک استراتژی جستجوی پرتو برای ایجاد توالی کلام انتخاب شده‌است (یو و همکاران ۲۰۱۶).

شکل caption تصویر مبتنی بر Phrase با شبکه LSTM مراتبی

۳ .معماری فیLSTM -
ایده اصلی فی LSTM - پیشنهادی رمز گشایی تصویر را از عبارت به جمله رمز گشایی می‌کند. این روش شامل یک decoder و کدگشا برای یک جمله اختصاری است. با استفاده از یک جفت sentence در مجموعه آموزشی، NPsهایی که در تصویر و ۲ (۲)با حداقل دو کلمه تشکیل شده‌اند، ابتدا با استفاده از یک الگوریتم chunkingکه در بخش ۵ توضیح داده شد، اولین chunked از جمله(S) هستند. سپس، یک AS با جایگزین کردن هر یک از هر یک از هر یک از جمله آخر کلمه chunked شکل می‌گیرد که در مثال زیر نشان‌داده شده‌است:
ما هر یک از این نوشته‌ها را در داده‌های آموزشی به یک جفت NPs به عنوان NPs تجزیه می‌کنیم، به طوری که NPs و NPs با دو decoders که به صورت سلسله مراتبی متصل شده‌اند، پردازش می‌شوند. این تجزیه ترتیب توالی را در نمودار مشروح انسان تغییر می‌دهد و در نتیجه ما توالی واقعیت زمینی متفاوتی(GTS) در طول مرحله تمرین را در مقایسه با مدل‌های RNN مرسوم خواهیم داشت. برای این منظور، the of our our، NPsاست، در حالی که the of the AS AS ها AS است.
۳.۱ decoder Phrase .
decoderدر این اثر دو نقش دارد که عبارتند از:
(۱برای رمزگشایی یک نمایش تصویر در NPs های متعدد، که ماهیت را در تصویر توصیف می‌کنند، و
۲(برای کدگذاری هر یک از NPs به یک نمایش بردار ترکیبی، که به عنوان ورودی در رمزگشا عمل می‌کند.

با در نظر گرفتن تصویری که من، یک از سی ان ان در مورد ImageNet برای رمز کردن یک تصویر در یک تصویر سه‌بعدی به کار برده می‌شود، که سپس به یک بردار - K بعدی با ماتریس تعبیه تصویر، Wipو بایاس تاثیر گذاری می‌شود. یک مدل LSTM مشابه [۴] برای رمزگشایی آن به هر یک از NPs مورد استفاده قرار می‌گیرد. برای آموزش یک مدل LSTM برای رمزگشایی کلمه i ام به طولLi ، تصویر جاسازی‌شده تصویر، و پس از آن یک نشانه آغاز به کلمهxsp ، روند ترجمه را نشان می‌دهد، و هر کلمه در NP ورودی به ترتیبی از بلوک‌های LSTM در یک روش گام‌به‌گام ورودی هستند، همانطور که در شکل ۲ نشان‌داده شده‌است. از این رو، در هر مرحله از جمله، مقادیر xi در هر مرحله از جمله، به دست می‌آیند:

که در آن، یک بردار K - D به عنوان بردار - K بعدی نمایش داده می‌شود، و به طور wi یک بردار یک - داغ است که مکان کلمه ورودی فعلی در واژه‌نامه را در زمان گام زمانی از جمله i نشان می‌دهد.
برای یک بلوک LSTM در یک گام زمانی، اجازه دهیدitp ، ftp، otp، ctpو htp نشان‌دهنده گیت ورودی، گیت خروجی، سلول حافظه و حالت پنهان در مرحله زمانی باشند. بنابراین معادلات گذار LSTM عبارت از عبارت است از:
ضرب .elementwise پارامترهایwf - Wi { LSTM ، Wo، وو، Ui، Uf، uo، uo، { uuهمگی ماتریس‌های با بعد از RK و RK هستند. در حالی که سلول حافظه واحد حافظه داخلی واحد را در ارتباط با اطلاعات پردازش‌شده در مرحله زمانی فعلی نگه می‌دارد، هر واحد گیت سازی، میزانی را که اطلاعات به روز شده، فراموش‌شده و به جلو تکثیر می‌شود را کنترل می‌کند. بنابراین حالت پنهان، یک دیدگاه جزیی از سلول حافظه واحد است.
خروجی the در هر مرحله زمانی، + RV lptp معادل احتمال شرطی یک کلمه با کلمات و تصویر قبلی، P :| wtp)ا، ا، (llاست. حقیقت زمین کلمه ورودی گام زمانی بعدی است و یک نشانه پایانی در آخرین مرحله برای

نشان دادن پایان یک .NP حالت پنهان آخرین مرحله به عنوان نمایش بردار ترکیبی از NP به کار می‌رود، که در آن
این کار به عنوان ورودی به عنوان decoder که بعد از آن شرح داده می‌شود، عمل می‌کند.

۳.۲ .حبس ابد(AS)
رمزگشا، طراحی مشابهی را به عنوان the عبارت، به جز ورودی‌ها، خروجی‌ها وGTS ، همانطور که در شکل ۳ نشان‌داده شده‌است، دارد. ورودی کدگشا یک عنوان کامل است که تصویر را توصیف می‌کند و هر) NP به عنوان مثال مرد)و بقیه کلمات در عنوان (به عنوان مثال در)به عنوان ورودی در یک گام زمانی مجرد کدگذاری می‌شوند [t] t .بیانگر یک گام زمانی به عنوان کدگشا و N طول آن در نظر گرفتن هر یک از NP به عنوان یک واحد است، ورودی of decoder عبارت است از:
این ها مجموعه دیگری از پارامترهای trainable برای تعبیه تصویر، استفاده از کلمه اول و ماتریس تعبیه کلمه "AS" هستند، در حالی که wts یک شاخص برداری تک hot از کلمه ورودی جریان زمان است.
دو خروجی توسط مدل LSTM در هر مرحله زمانی در رمزگشا تولید می‌شوند، که عبارتند از: یک شاخص دودویی که مشخص می‌کند اگر ورودی بعدی عبارت یا عبارت باشد (یعنی علامت عبارت)و ۲)پیش‌بینی softmaxکلمه بعدی در توالی (یعنی پیش‌بینی کلمه). حقیقت زمین دومین خروجی در هر مرحله یا آخرین کلمه عبارت بعدی یا خود کلمه بعدی است که به صورت زیر فرموله شده‌است:
در کار اولیه ما [۲۲]، از یک نشانه عبارت برای نشانه عبارت استفاده کردیم، که منجر به محدودیت قادر به تشخیص مناسب بودن ورودی‌های NP مختلف در طول رمزگشایی شد. به عنوان یک جبران، یک هدف انتخاب عبارت برای حل این محدودیت معرفی شد. با این حال، این روش یک روش آموزشی پیچیده‌ای دارد، چرا که در هر مرحله زمانی که ورودی یک NP است، بیش از چندین ذره انتخاب‌شده NPs انتخاب شده‌است. برای ساده‌سازی فرآیند آموزش در اینجا، ما عبارت عبارت و عبارت انتخاب عبارت را با عبارت عبارت و پیش‌بینی softmaxکلمه آخر هر) NP یعنی معادله ۱۱)، اگر ورودی بعدی یک عبارت است)عوض می‌کنیم.
۳ .۳ آموزش مدل فیLSTM -
تابع هدف مدل ما یک تابع هزینه لگاریتم درستنمایی بیشینه است که از سرگشتگی پیش‌بینی کلمه محاسبه می‌شود و این عبارت است از: از دست رفتن پیش‌بینی نشانه. یعنی، با توجه به تصویری که من و توصیف آن Sداریم، اجازه دهید که R تعداد عبارت‌های جمله باشد، در حالی که ptp و S خروجی احتمال بلوک LSTM در مرحله زمانی و ۱ - ۱ و - t ۱ هستند. گیج بودن جمله در تصویری که من دارم
که در آن + M = N + ۱۱
= ۱R + Li) ۱ .(ما از تلفات لولا به عنوان نشانه اشاره برای طبقه‌بندی ورودی بعدی of به عبارت یا عبارت استفاده می‌کنیم. تابع هزینه طبقه‌بندی کننده این است:
که در آن پروژه، خروجی کلید مخفی بلوک LSTM در مرحله زمان می‌باشد، wps RKبه عنوان پارامترهای trainableبرای طبقه‌بندی کننده‌ها استفاده می‌شود. اگر ورودی بعدی به عنوان کدگشا یک عبارت یا در غیر این صورت باشد، lyts است. در اینجا، ktsهدف را براساس تعداد عبارات و کلمات مندرج در آن ثبت می‌کند. بنابراین، اگر = yts ۱ یا ۱(N - R) / غیر از این باشد.
در غیر این صورت. بنابراین با تعداد نمونه‌های آموزشی، تابع هدف کلی مدل ما عبارت است از:
که در آن* P = Q
= j IP .Mj این معادل با the متوسط یک کلمه با مفهوم قبلی آن‌ها و تصویر توصیف‌شده در یک عبارت تنظیم، θ۲، θλ، متوسط بر تعداد نمونه‌های آموزشی است. در اینجا θ تمام پارامترهای trainable مدل هستند.
به طور خلاصه، ساختار فی - LSTM پیشنهاد شده) برای پیش‌بینی(i کلمه بعدی داده‌شده به تمام کلمات قبلی در هر یک) NP ۲(کلمه بعدی با توجه به کلمات و عبارات قبلی (و ج)اگر ورودی بعدی یک عبارت باشد، بهینه شده‌است. این تابع عینی به مدل اجازه می‌دهد تا به انتها آموزش داده شود.

۳.۲ .حبس ابد(AS)
رمزگشا، طراحی مشابهی را به عنوان the عبارت، به جز ورودی‌ها، خروجی‌ها وGTS ، همانطور که در شکل ۳ نشان‌داده شده‌است، دارد. ورودی کدگشا یک عنوان کامل است که تصویر را توصیف می‌کند و هر) NP به

عنوان مثال مرد)و بقیه کلمات در عنوان (به عنوان مثال در)به عنوان ورودی در یک گام زمانی مجرد کدگذاری می‌شوند [t] t .بیانگر یک گام زمانی به عنوان کدگشا و N طول آن در نظر گرفتن هر یک از NP به عنوان یک واحد است، ورودی of decoder عبارت است از:
این ها مجموعه دیگری از پارامترهای trainable برای تعبیه تصویر، استفاده از کلمه اول و ماتریس تعبیه کلمه "AS" هستند، در حالی که wts یک شاخص برداری تک hot از کلمه ورودی جریان زمان است.
دو خروجی توسط مدل LSTM در هر مرحله زمانی در رمزگشا تولید می‌شوند، که عبارتند از: یک شاخص دودویی که مشخص می‌کند اگر ورودی بعدی عبارت یا عبارت باشد (یعنی علامت عبارت)و ۲)پیش‌بینی softmaxکلمه بعدی در توالی (یعنی پیش‌بینی کلمه). حقیقت زمین دومین خروجی در هر مرحله یا آخرین کلمه عبارت بعدی یا خود کلمه بعدی است که به صورت زیر فرموله شده‌است:
در کار اولیه ما [۲۳]، از یک نشانه عبارت برای نشانه عبارت استفاده کردیم، که منجر به محدودیت قادر به تشخیص مناسب بودن ورودی‌های NP مختلف در طول رمزگشایی شد. به عنوان یک جبران، یک هدف انتخاب عبارت برای حل این محدودیت معرفی شد. با این حال، این روش یک روش آموزشی پیچیده‌ای دارد، چرا که در هر مرحله زمانی که ورودی یک NP است، بیش از چندین ذره انتخاب‌شده NPs انتخاب شده‌است. برای ساده‌سازی فرآیند آموزش در اینجا، ما عبارت عبارت و عبارت انتخاب عبارت را با عبارت عبارت و پیش‌بینی softmaxکلمه آخر هر) NP یعنی معادله ۱۱)، اگر ورودی بعدی یک عبارت است)عوض می‌کنیم.
۳ .۳ آموزش مدل فیLSTM -
تابع هدف مدل ما یک تابع هزینه لگاریتم درستنمایی بیشینه است که از سرگشتگی پیش‌بینی کلمه محاسبه می‌شود و این عبارت است از: از دست رفتن پیش‌بینی نشانه. یعنی، با توجه به تصویری که من و توصیف آن Sداریم، اجازه دهید که R تعداد عبارت‌های جمله باشد، در حالی که ptp و S خروجی احتمال بلوک LSTM در مرحله زمانی و ۱ - ۱ و - ۱ t هستند. گیج بودن جمله در تصویری که من دارم
که در آن + M = N + ۱
= Li + lR) ۱.(ما از تلفات لولا به عنوان نشانه اشاره برای طبقه‌بندی ورودی بعدی of به عبارت یا عبارت استفاده می‌کنیم. تابع هزینه طبقه‌بندی کننده این است:
که در آن پروژه، خروجی کلید مخفی بلوک LSTM در مرحله زمان می‌باشد، wps RK به عنوان پارامترهای trainableبرای طبقه‌بندی کننده‌ها استفاده می‌شود. اگر ورودی بعدی به عنوان کدگشا یک عبارت یا در غیر این صورت باشد، lyts است. در اینجا، ktsهدف را براساس تعداد عبارات و کلمات مندرج در آن ثبت می‌کند. بنابراین، اگر = yts ۱ یا ۱(N - R) / غیر از این باشد.
در غیر این صورت. بنابراین با تعداد نمونه‌های آموزشی، تابع هدف کلی مدل ما عبارت است از:
که در آن* Q = P
= Mj. lP j این معادل با the متوسط یک کلمه با مفهوم قبلی آن‌ها و تصویر توصیف‌شده در یک عبارت تنظیم، ۲θ، ۸θ، متوسط بر تعداد نمونه‌های آموزشی است. در اینجا θ تمام پارامترهای trainable مدل هستند.
به طور خلاصه، ساختار فی LSTM - پیشنهاد شده) برای پیش‌بینی(i کلمه بعدی داده‌شده به تمام کلمات قبلی در هر یک) NP ۲(کلمه بعدی با توجه به کلمات و عبارات قبلی (و ج)اگر ورودی بعدی یک عبارت باشد، بهینه شده‌است. این تابع عینی به مدل اجازه می‌دهد تا به انتها آموزش داده شود.

۴ .تولید تصویر تصویر
مدل فی LSTM - عنوان تصویر را در یک روش دو مرحله‌ای ایجاد می‌کند، که در آن یک لیست از نامزدهای NPابتدا به دنبال علامت کامل، هم با استفاده از الگوریتم جستجوی پرتو، ایجاد می‌شوند. اندازه تیر برای عبارت و تولید جمله به ترتیب bp و bs هستند. تولید NPs در این کار مشابه با [۴] است، که در آن یک تصویر مشخص کدگذاری شده با شبکه سی ان ان وارد این مدل می‌شود، که به عنوان زمینه اولیه of برای تولید NPs عمل می‌کند. در هر گام زمانی، کلمات bp با بیش‌ترین احتمال نمونه‌برداری و ورودی به رمزگشا در مرحله بعدی برای استخراج کلمات بعدی است. مجموعه‌ای از بهترین توالی‌های bp که تا مرحله زمانی ایجاد می‌شوند، به عنوان کاندیداهای بالقوه برای استنباط کلمه بعدی به صورت تکراری در نظر گرفته می‌شوند، تا زمانی که همه کاندیداها نشانه پایان کلمه هستند. سپس یک امتیاز برای هر یک از هر یک از هر یک از هر یک از هر کلمه نرمال شده توسط طولNP ، از جمله نشانه پایان کلمه، محاسبه می‌شود:
سپس نامزدهای - NP ایجادشده طبق آخرین کلمه گروه‌بندی می‌شوند و کاندیداها با امتیاز کم‌تر از مقدار آستانه T دور ریخته خواهند شد. این به منظور بهبود کیفیت توصیف تصویر شکل‌گرفته است.

با این وجود، حداقل یک نامزد (از بالاترین امتیاز)صرف‌نظر از امتیاز آن برای هر گروه NP باقی خواهد ماند. پس از این، در مجموع of های کامل، از فهرست نامزدهایNP ، همانطور که در شکل ۴ نشان‌داده شده‌است، ایجاد خواهند شد. رمزگشا دو خروجی را در هر مرحله زمانی تولید می‌کند، که عبارتند از: پیش‌بینی کلمه بعدی و دوم)عبارت بعدی از ورودی بعدی. بنابراین، زمانی که مدل استنتاج می‌کند که ورودی بعدی یک عبارت است، هر یک از پاسخ‌های مربوط به واژه bs استنباط می‌شود (به عنوان مثال سگ، سگ، دو، قهوه‌ای در شکل ۴)با لیست نامزدهای NP مقایسه شده‌است. این NPs با کلمه آخرین تطبیق با کلمات) inferred به عنوان مثال یک سگ قهوه‌ای، دو سگ، دو سگ قهوه‌ای)به لیست نامزدهای پرتوی در مرحله زمانی فعلی متصل می‌شوند و جایگزین کلمات) inferred به عنوان مثال پرتوی که "سگ" به عنوان ورودی بعدی استفاده می‌کند). کلمات استنباط شده بدون هیچ جایگزین) NP به عنوان مثال، دو، قهوه‌ای)در لیست نامزدهای پرتو باقی خواهند ماند، برای مواردی که در آن رمزگشا یک NP مناسب را ایجاد نمی‌کند (به عنوان مثال کلمه یک کلمه یا یک شی کوچک). هنگامی که تمام جملات کاندید یک نشانه پایانی را استنباط می‌کنند، امتیاز هر زیر به صورت زیر محاسبه می‌شود:
و این حکم بالاترین امتیاز را به دست می‌آورد، با این وجود، با توجه به این که آن را انتخاب کرده‌اید.

۵Phrase chunking ،. محدودیت‌ها و پالایش
Phrase chunkingیک فرآیند زبان طبیعی است که یک جمله را از جمله فعل، فعل و prepositional جدا می‌کند. یک نمای کلی از ساختار توصیفات تصویر نشان می‌دهد که عناصر کلیدی که اکثریت اشکال را تشکیل می‌دهند، معمولا آن NPs هستند که ماهیت های حاکم را در یک تصویر توصیف می‌کنند. آن می‌تواند یک جسم، گروهی از اشیا یا صحنه باشد. این ورودی‌ها دارای سطح انتزاعی معادل با خروجی یک کدگذار سی ان ان هستند و با عبارت فعل و prepositional مرتبط هستند. بنابراین، NPاساسا بیش از نیمی از پیکره زبان در یک مدل زبانی را پوشش می‌دهد که برای تولید توضیح تصویری آموزش‌دیده اند. بنابراین، در این مقاله، ما یادگیری ساختار NP و جمله‌ها را تجزیه می‌کنیم به طوری که آن‌ها می‌توانند به طور مساوی پردازش شوند، در مقایسه با استخراج تمام عبارات بدون در نظر گرفتن بخشی از برچسب گفتار.(POS)
این بخش(i) الگوریتم تجزیه را توصیف می‌کند که برای به دست آوردن جفت شدگی آریلNPs ، ۲(مشکلات ناشی از محدودیت ابزار تجزیه فعلی و راه‌حل پیشنهادی ما، و ج)معیاری برای کاهش تاثیر این محدودیت‌ها در آموزش مدل captioning تصویر خود استفاده کردیم.
Phrase chunking۵.۱
برای شناسایی NPs ناشی از یک عنوان آموزشی، تجزیه وابستگی ابزار] CoreNLP Stanford ۵۲ [را اتخاذ کردیم که یک درخت رابط ساختاری را بر روی یک جمله با فراهم کردن روابط ساختاری بین کلمات تشکیل می‌دهد. اگر چه این یک جمله را مستقیما به عنوان یک تجزیه‌گر پیشنهاد می‌کند و دیگر ابزارهایchunking ، الگوی استخراج‌شده از NP انعطاف‌پذیرتر است چون ما می‌توانیم روابط ساختاری مطلوب را انتخاب کنیم.
روابطی که ما انتخاب کردیم این است:
رابطه تعریف(det) ،
*عددی(nummod) ،
تغییردهنده(adjectival (amod ،
ترکیب (ترکیب)،
تعدیل‌کننده(adverbial (advmod ، تنها زمانی انتخاب شد که معنی واژه adjective تغییر داده شود، به عنوان مثال "اتاق کم‌نور"،
تعدیل‌کننده اسمی برای "تغییر مالکیت معنوی(nmod: of nmod: poss) "، با "مورد" شامل

به طور کلی، یک تجزیه‌گر وابستگی چند triplets را استخراج می‌کند، هر یک از یک لغت استاندار، یک واژه وابسته و یک رابطه که آن‌ها را پیوند می‌دهد، به شکل رابطه (حاکم، وابسته)، از یک جمله تشکیل می‌دهد. به منظور تشکیل قسمت‌هایی از یک تجزیه‌گر وابستگی، یک گام پس پردازش ساده همانطور که در شکل ۵ نشان‌داده شده‌است، انجام می‌شود. یعنی، tripletsبا یک فرماندار یا واژه وابسته که در the کامل هم پشت سر هم هستند (به طور مثال) amod پیراهن، خاکستری)و) det پیراهن، (theبه عنوان یک ان‌پی کامل گروه‌بندی می‌شوند. همین مساله برای سه تایی متوالی صدق می‌کند (به عنوان مثال) det :مرد، (the، در حالی که واژه مستقل (به عنوان مثال "در")به عنوان یک واحد در جدول باقی می‌ماند.

۵Phrase chunking ،. محدودیت‌ها و پالایش
Phrase chunkingیک فرآیند زبان طبیعی است که یک جمله را از جمله فعل، فعل و prepositional جدا می‌کند. یک نمای کلی از ساختار توصیفات تصویر نشان می‌دهد که عناصر کلیدی که اکثریت اشکال را تشکیل می‌دهند، معمولا آن NPs هستند که ماهیت های حاکم را در یک تصویر توصیف می‌کنند. آن می‌تواند یک جسم، گروهی از اشیا یا صحنه باشد. این ورودی‌ها دارای سطح انتزاعی معادل با خروجی یک کدگذار سی ان ان هستند و با عبارت فعل و prepositional مرتبط هستند. بنابراین، NPاساسا بیش از نیمی از پیکره زبان در یک مدل زبانی را پوشش می‌دهد که برای تولید توضیح تصویری آموزش‌دیده اند. بنابراین، در این مقاله، ما یادگیری ساختار NP و جمله‌ها را تجزیه می‌کنیم به طوری که آن‌ها می‌توانند به طور مساوی پردازش شوند، در مقایسه با استخراج تمام عبارات بدون در نظر گرفتن بخشی از برچسب گفتار.(POS)
این بخش(i) الگوریتم تجزیه را توصیف می‌کند که برای به دست آوردن جفت شدگی آریلNPs ، ۲(مشکلات ناشی از محدودیت ابزار تجزیه فعلی و راه‌حل پیشنهادی ما، و ج)معیاری برای کاهش تاثیر این محدودیت‌ها در آموزش مدل captioning تصویر خود استفاده کردیم.
۵.۱Phrase chunking
برای شناسایی NPs ناشی از یک عنوان آموزشی، تجزیه وابستگی ابزار] CoreNLP Stanford ۵۲ [را اتخاذ کردیم که یک درخت رابط ساختاری را بر روی یک جمله با فراهم کردن روابط ساختاری بین کلمات تشکیل می‌دهد. اگر چه این یک جمله را مستقیما به عنوان یک تجزیه‌گر پیشنهاد می‌کند و دیگر ابزارهایchunking ، الگوی استخراج‌شده از NP انعطاف‌پذیرتر است چون ما می‌توانیم روابط ساختاری مطلوب را انتخاب کنیم.
روابطی که ما انتخاب کردیم این است:
رابطه تعریف(det) ،
*عددی(nummod) ،
تغییردهنده(amod) adjectival ،
ترکیب (ترکیب)،
تعدیل‌کننده(advmod) adverbial ، تنها زمانی انتخاب شد که معنی واژه adjective تغییر داده شود، به عنوان مثال "اتاق کم‌نور"،
تعدیل‌کننده اسمی برای "تغییر مالکیت معنوی(nmod: of nmod: poss) "، با "مورد" شامل

به طور کلی، یک تجزیه‌گر وابستگی چند triplets را استخراج می‌کند، هر یک از یک لغت استاندار، یک واژه وابسته و یک رابطه که آن‌ها را پیوند می‌دهد، به شکل رابطه (حاکم، وابسته)، از یک جمله تشکیل می‌دهد. به منظور تشکیل قسمت‌هایی از یک تجزیه‌گر وابستگی، یک گام پس پردازش ساده همانطور که در شکل ۵ نشان‌داده شده‌است، انجام می‌شود. یعنی، tripletsبا یک فرماندار یا واژه وابسته که در the کامل هم پشت سر هم هستند (به طور مثال) amod پیراهن، خاکستری)و) det پیراهن، (theبه عنوان یک ان‌پی کامل گروه‌بندی می‌شوند. همین مساله برای سه تایی متوالی صدق می‌کند (به عنوان مثال) det :مرد، (the، در حالی که واژه مستقل (به عنوان مثال "در")به عنوان یک واحد در جدول باقی می‌ماند.

هیچ راه‌حل مناسبی وجود ندارد. در نتیجه، همیشه برخی اشتباه‌ات اجتناب‌ناپذیر از خروجیparser ، صرف‌نظر از ابزار chunking استفاده‌شده وجود دارد. با استفاده از یک تجزیه‌گر وابستگی، اصطلاح chunking را با یک تجزیه‌گر پیشنهاد کرده‌ایم. حوزه‌هایparser ، تابع و گزاره یک جمله را به طور مستقیم بیان می‌کنند، و ما the NPرا در پایین‌ترین سطح قرار می‌دهیم. در این بخش، ما جفت NPs را با استفاده از the با استفاده از ۲parsers مقایسه خواهیم کرد. این NPs در ستون سمت چپ نشان داده می‌شوند در حالی که در ستون سمت راست نشان‌داده شده‌است. مثال‌هایa) ۱ - ۱(داده‌شده در زیر باS ، DP، CP، وR) DP علامت‌گذاری شده‌اند، که به ترتیب با chunking با یک تجزیه‌گر وابستگی، یک تجزیه‌گر وابستگی و یک تجزیه‌گر وابستگی تشکیل می‌شود. یک متن برجسته نشان می‌دهد که جفت شدگی آریل ها حاوی خطا هستند. یکی از اشتباه‌ات رایجی که در خروجی هر یک از the یافت می‌شود، به رسمیت شناختن اشتباه یک فعل به عنوان یک اسم تلقی می‌شود. در نتیجه، همانطور که در مثال‌های نشان‌داده‌شده، به عنوان یک شی از دست رفته شکل می‌گیرد)شکلa ۱ - ۱، ستون راست .(علاوه بر این، NPsهایی هستند که هیچ نهاد را در یک تصویر، مانند"the" ، توصیف نمی‌کنند.

از مشاهدات ما، هر دو parsers به نظر می‌رسد خروجی‌های NP مشابهی دارند. دلایلی که یک تجزیه‌گر وابستگی را انتخاب کردیم عبارتند از:
۱.نسبت به NPs با سطح سازنده بالاتر، برای انتخاب رابطه وابستگی خاص مثل :nmod" از"، نسبت به مشخص کردن سطح NP در درخت تجزیه آن، intuitiveاست.
۲.مواردی وجود دارند که در آن فعل و انفعال در گذشته بخشی از ویژگی‌های یک اسم محسوب می‌شود، و تجزیه‌گر وابستگی شانس بیشتری دارد که آن را به عنوان صفت شناسایی کند. برای مثال:
یک تعدیل‌کننده اسمی برای " "و تغییر مالکیت(nmod: poss & nmod) "در بین روابط وابستگی انتخابی ما انتخاب شده‌است. همانطور که در مثالa) ۳(نشان‌داده شده‌است، اکثر NPs chunked تحت این روابط، مطابق با یک ماهیت یا یک گروه از هویت‌های درون یک تصویر هستند، همانطور که در مثالa) ۳(نشان‌داده شده‌است. با این وجود، هنوز ابهام برای NPs chunked از nmod وجود دارد: ارتباط، در هر یک از این جمله، باید به دو NPs تبدیل شود یا بعنوان یک NP منفرد باقی بماند. مثال(b) موردی را نشان می‌دهد که در آن یک "رابطه" لازم نیست، در حالی که مثالC) ۳(یک مورد دیگر را نشان می‌دهد وقتی که الزام رابطه مبهم است.

NPsاز. Refinement ۵.۳
محدودیت‌های تجزیه‌گر تغییرات غیر ضروری را در سراسر داده‌های آموزشی ایجاد کرده‌اند که به نوبه خود اثر آموزشی مدل captioning تصویر ما را تحت‌تاثیر قرار داده‌است. به منظور کاهش اثرات تجزیه نادرست مدل ما، ما یک استراتژی پالایش بین آموزش of و رمزگشا را معرفی می‌کنیم، جایی که براساس آمار محلی داده‌های آموزشی، یک جفت NPs به روز به روزرسانی می‌شود. به عبارت دیگر، theابتدا قبل از مدل کلی با تکنولوژی توقف زودرس nique - مورد استفاده قرار گرفت nique - بر روی ارزش حیرت همه NPs اعمال شد. سپس، مدلی که با بهترین اعتبار سنجی برای تولید مجموعه‌ای از NPs مورد استفاده قرار می‌گیرد، برای تولید مجموعه‌ای از NPs مورد استفاده قرار می‌گیرد. سپس، اجزای جفت AS - NPs های آموزشی، براساس NPs تولید شده، به تدریج که کلمه اول non را به AS احیا می‌کنند، اصلاح خواهند شد و پس از آن کلمه آخر - را دنبال می‌کنند. جزییات الگوریتم پالایش پیشنهادی ما در شکل ۶، همراه با یک مثال برای درک بهتر نشان‌داده شده‌است. با در نظر گرفتن یک تصویر در داده‌های آموزشی، در مجموع دو NPs p تولید می‌شوند G .و G مجموعه کلمات اول و آخرین کلمات همه β تولید شده به ترتیب هستند، در حالی که K یک) NP از تجزیه‌گر)با کلمه W آغاز می‌شود و با یک طول of K e به پایان می‌رسد. این پالایش برای همه NPs chunked در یک جمله انجام شده‌است. مثال‌های زیر تفاوت بین جفت AS - NPs ناشی از روش پیشنهادی our de scribed قبل و بعد از استراتژی پالایش را نشان می‌دهند. مثالA) ۴(نشان می‌دهد که the بازی می‌کند، چون هیچ یک از NPs gener - ated با کلمه "کامل" شروع می‌کنند، اما برخی با کلمه "a" شروع می‌کنندb) Ex .۴(با RS ۲ تصحیح می‌شود، زیرا کلمه "ایستادن" به عنوان آخرین کلمه هر کدام از NPs تولید شده استنباط نمی‌شود. در امتحان c)۴(، عبارات یکی، نقاط جلو و انگشت اشاره او برای آن ذخیره شده‌اند، چون phrase که از تصویر به تنهایی استفاده می‌کند، با کلمه "یک"، "نقطه" و "انگشت" به کار خود پایان نمی‌دهد. این سه عبارت مطابق با هر ماهیت های غالبی در تصویر نیستند، و در نتیجه به ندرت - oc ای در میان the تصاویر مشابه وجود خواهد داشت. در حقیقت، "تنها نمی‌تواند از محتوای تصویر به تنهایی تولید شود، زیرا به سوژه خود (" دو مرد ")به عنوان متن قبلی نیاز دارد. از طرف دیگر، کلمه "دوربین" چنین استنباط می‌شود اگر چه این شی در تصویر با توجه به آمار داده‌های آموزشی، در مقایسه با "نگاه به دوربین" در سنین مختلف، در تصویر وجود ندارد. مثال d)۴(موردی را نشان می‌دهد که در آن decoder عبارت آموزش‌دیده ما به طور خودکار تصمیم می‌گیرد که کدام نهاد براساس آمار داده‌های آموزشی نگه خواهد شد.
با این استراتژی پالایش، رمزگشا به طور کامل در روش اصلاح‌شده آموزش داده خواهد شد، در حالی که کدگشا برای NPs های تصفیه‌شده هنگامی که مدل کلی آموزش‌دیده است، تنظیم شده‌است. ما بر روی کاهش تاثیر خطای ناشی از تجزیه‌گر، جفت‌های NPs به عنوان NPs ها را به گونه‌ای انتخاب می‌کنیم که به جای کار زبانی، برای تصویر captioning تصویر مناسب‌تر هستند. علاوه بر این، اشیا کم‌تر غالب که نیاز بیشتری به سابقه قبلی (حافظه بلند مدت)برای نسل خود دارند، از جمله انگشتی در امتحان - به عنوان مثالC) ۴(از طریق رمزگشا به عنوان refinement بکار گرفته می‌شود.

معماری کلی
معماری کلی در شکل ۴ نشان‌داده شده‌است.
و جریان جریان RGB روش ما را می توان به عنوان یک شبکه دو جریانی در نظر گرفت که شامل جریان برای استخراج (و همکاران ۲۰۱۵ He) می‌باشد. در هر دو جریان سیال و جریان سیال، یک شبکه باقیمانده ۱۸ لایه بر روی ویژگی‌ها برای مدل همبستگی زمانی LSTM ویژگی‌ها از هر فریم استفاده می‌شود. لایه‌های مختلف است، RGB فریم های ورودی در یک پنجره زمانی کوچک اعمال می‌شوند. معماری جریان جریان مشابه جریان بدون اتصالات باقیمانده در جریان جریان استفاده می‌کنیم. خروجی دو جریان یک الحاق LSTM اما ما تنها از است و سپس در هر فاصله زمانی به مدول حافظه پیشنهادی داده می‌شود. سپس ماژول حافظه فاصله بین نمونه پرس و جو و تمام نمونه‌های ذخیره‌شده را محاسبه می‌کند و برچسب عملیاتی نزدیک‌ترین نمونه را به نمونه پرس و جو اختصاص می‌دهد. مزیت استفاده از ماژول حافظه این است که می‌تواند مشاهدات اولیه را به خاطر داشته باشد و در نتیجه توان تعمیم را بهبود بخشد. شبکه پیشنهادی به طور خاص برای پیش‌بینی ، وMansimov، Srivastava)موجود، LSTM عمل توسعه داده می‌شود. در مقایسه با شبکه‌های روش ما(۲۰۱۶ Kautz و Molchanov۲۰۱۵؛ وو و سایرین ۲۰۱۵؛ یانگ و سایرین ۲۰۱۵؛ یانگ، Salakhudinov را با حافظه بلند مدت برای هدف حفظ نمونه‌های پیش‌بینی دشوار تقویت می‌کند. این به خصوص LSTMs زمانی مهم است که پیش‌بینی‌ها در زمانی ساخته می‌شوند که تنها چند فریم دیده شوند.
دو جهتی، LSTM ما همچنین قدرت نمایش ویدئوهای جزئی را با استفاده از اطلاعات آینده از طریق یک لایه افزایش می‌دهیم.
در مقایسه با رویکردهای پیش‌بینی اقدام موجود (هنگ‌کنگ، تائو، و فو ۲۰۱۷؛ هنگ‌کنگ، کیت، و فو ۲۰۱۴)که در عمل شناخته شود، روش ما به چنین اطلاعاتی نیاز ندارد و در نتیجه در predictors آن نسبت مشاهده باید به سناریوهای realworld عملی‌تر است.

نمونه Hard Predict - to
برای مثال، پیش‌بینی اقدامات در مراحل اولیه آن‌ها ضروری است، به عنوان مثال، زمانی که تنها ۱۰ % از فریم ها مشاهده می‌شوند. اگرچه کار اخیر (هنگ‌کنگ، تائو، و فو ۲۰۱۶)قدرت نمایش ویژگی را بهبود بخشیده‌است، عملکرد پیش‌بینی در یک مرحله ابتدایی هنوز پایین است چون ویژگی‌های به‌دست‌آمده از ویدیوهای جزیی مشاهده‌شده به سختی قابل‌تشخیص هستند. ما به جای بهبود قدرت تفکیک پذیری ویژگی‌ها، پیشنهاد می‌کنیم که نمونه‌های آموزشی دشوار را در مرحله آموزش به خاطر بسپارد. همانطور که در شکل ۴ به لایه آخر شبکه پیشنهادی اضافه می‌شود که به عنوان (و همکاران ۲۰۱۷ Kaiser) نشان‌داده‌شده، حافظه حافظه جهانی عمل می‌کند. حافظه برای نگه داشتن ویدیوهای جزیی و فیلم‌های کامل از همان کلاس در یک محله محلی بهینه‌سازی می‌شود، در نتیجه اجازه می‌دهد که یک ویدیو در مرحله میانی، مرحله میانی، یا مرحله آخر تست شود تا یک فیلم مشابه با برچسب عملیاتی مشابه پیدا کند (شکل ۳). مدول حافظه شامل برای ذخیره کردن مقادیر حافظه و یک بردار V = { v } از کلیدهای حافظه، یک ماتریس K Rm * k یک ماتریس است که سن آیتم‌های ذخیره‌شده در حافظه را ردیابی می‌کند. یک حافظه به صورت m به اندازه A اضافی زیر تعریف می‌شود:

و y سطوح عملکرد مربوطه آن‌ها {(y ،z)} = { v } = V ، وk برای پیش‌بینی نمونه‌های بعدی - Kدر اینجا، است. keyvalue را نشان می‌دهد. جستجوی حافظه. اساسا، حافظه مجموعه‌ای از زوج‌های z سطوح پیشرفت و K حافظه نیاز به پرس و جوها به عنوان ورودی دارد که به دنبال مناسب‌ترین نمونه در ماتریس کلیدی را محاسبه کنیم. در آزمایش ما از d خروجی‌های آن به عنوان نتیجه باشد. سوال این است که چگونه شباهت دو معیار شباهت، محصول نقطه و هسته گاوسی برای محاسبه شباهت استفاده می‌کنیم که به وسیله آن‌ها تعریف شده‌است.

در اینجا q موج ورودی را نشان می‌دهد، و K [i] کلید i - i در حافظه است.

تلفات حافظه براساس همسایه‌های صحیح و همسایه‌های نادرست محاسبه شده‌است. فرض کنید که مقدار نزدیک‌ترین همسایه k داده می‌شود. با استفاده از یک سری از q به نمونه پرس و جو(y، z) = v تصحیح شده کوچک‌ترین شاخص است، تعریف می‌شود. همسایه‌های a ، که در آنv = V [na] ، که به عنوانq به(۰، ۱۰، n) ، نادرست نمونه‌هایی هستند که یا برچسب‌های عملیاتی نادرست را دارند: V [y] = v (y)= v (z)= v (z)= v (z) + K [q] = [q]) کوچک‌ترین شاخص هستند. از دست دادن حافظه می‌تواند به عنوان حداکثر c و b که در آن یک حاشیه مثبت است، بیان شود. این تابع زیان اساسا از محصول ζ ، که در آن(۵)، ([۰] k + [k] q - [۱] q نقطه برای محاسبه شباهت بین پرس و جو و نمونه‌های ذخیره‌شده استفاده می‌کند. ما می‌خواهیم شباهت را به کلید صحیح حداکثر کنیم و همچنین شباهت به کلیدهای نادرست را به حداقل برسانیم. به روز رسانی حافظه بسته به مقدار کلید بازگشت v متناظر با q برای توضیح این حقیقت انجام می‌شود که پرس و جوی فعلی نادرست باشد v اگر مقدار بازگشتی .n (q ،M): است q نزدیک‌ترین همسایه به Denote n .صحیح است یا نه ، q) ، ما حافظه را با نوشتن نامه پرس و جوv = [n] V(یا برچسب اکشن یا سطح پیشرفت)، به عنوان مثال، n = arg maxi a [i] به حافظه به روز خواهیم کرد. ما این مکان را در حافظه برای نوشتن پرس و جو توسط(v در حافظه q یک عدد تصادفی است که تصادفی را معرفی می‌کند، می‌یابیم. سپس پرس و جوی r ، که+ r شروع می‌شود و ما ویدیوهای جزیی را z نوشته می‌شود، در اینجا، سن یک مورد حافظه در سطح پیشرفت در مرحله اولیه خود تشویق می‌کنیم تا در حافظه ذخیره شوند. سن همه موارد از ۱ بعد از هر به‌هنگام‌سازی درست باشد (هر دو برچسب عملی و سطح پیشرفت درست v حافظه افزایش می‌یابد. اگر مقدار بازگشتی ،. و نرمال کردن آن به روزرسانی می‌شودq و K [n] هستند)، آنگاه کلید با میانگین گیری مورد

ادغام حافظه درLSTM

کاره‌ای قبلی Kaiser) و همکاران ۲۰۱۷(بعد از هر زمان - مرحله‌ای ازLSTMs ، مدول حافظه را اضافه می‌کنند. با این حال، ما استدلال کردیم که نیازی نیست هر گونه خروجی مرحله‌ای of در زمینه پیش‌بینی عمل را به یاد داشته باشیم، زیرا حرکت بین فریم های همسایه در ویدئو کاملا شبیه هم خواهد بود. در عوض، ما تنها از یک حافظه جهانی برای به خاطر سپردن همه مشاهدات اولیه در این کار استفاده می‌کنیم (شکل ۴ را ببینید). ما میانگین خروجی the را از همه timesteps میانگین می‌گیریم و این خروجی را در ماژول حافظه تغذیه می‌کنیم. با انجام این کار، ماژول حافظه می‌تواند خروجی جهانی of را به یاد داشته باشد. عملکرد حافظه جهانی شبیه طبقه‌بندی کننده است. با این حال، تفاوت اصلی این است که برچسب اقدام ارایه‌شده توسط حافظه به وسیله شباهت داده‌ها محاسبه می‌شود در حالی که برچسب ارایه‌شده توسط یک طبقه‌بندی کننده توسط مدل آموخته‌شده، محاسبه می‌شود. مشخصه‌های استخراج‌شده از ویدیوهای جزیی (به ویژه آن‌هایی که در مراحل اولیه خود هستند)در مقایسه با ویژگی‌هایی که از ویدیوهای کامل (هنگ‌کنگ، تائو، و فو ۲۰۱۷، هنگ‌کنگ، کیت، و فو ۲۰۱۴)دارای قدرت تشخیصی کمتری برخوردارند. استفاده از چنین ویژگی‌های غیرقابل‌تشخیص ممکن است قادر به یادگیری مرز طبقه‌بندی پیچیده نباشند. برعکس، محاسبه شباهت نمونه در یک استخر نمونه بزرگ با استفاده از حافظه می‌تواند یک مرز طبقه‌بندی پیچیده را یاد بگیرد و در نتیجه می‌تواند عملکرد پیش‌بینی حتی چند فریم را بهبود بخشد. ما این را در آزمایش‌ها نشان خواهیم داد.
توجه می‌تواند یک جایگزین برای مدل‌سازی همبستگی بین فریم ها در یک دوره زمانی کوچک و دسته‌های اکشن باشد. با این حال، توجه بر اطلاعات کوتاه‌مدت محلی متمرکز است، که ممکن است قادر به ارائه اطلاعات در یک حوزه بلند مدت نباشند. برای اینکه عملکرد بهتری از تشخیص اولیه داشته باشیم، مکانیسم حافظه بلند مدت در چارچوب ما معرفی می‌شود. با استفاده از این مکانیزم، این مدل می‌تواند اقدامات سخت برای پیش‌بینی اقدامات در مراحل اولیه خود را یاد بگیرد.

مدل‌سازی متن آینده

یکی از محدودیت‌های of متعارف این است که آن‌ها قادر به استفاده از بافت آینده نیستند. ما پیشنهاد می‌کنیم که از ادغام دو سویه و غنی‌سازی ویژگی‌های فعلی با انتگرال گرفتن از اطلاعات آینده استفاده کنیم. همانطور که در شکل ۵ نشان‌داده‌شده، در این کار، ما یک LSTM رو به عقب در بالای LSTM جلو بدون برهم کنش در لایه‌های پنهان اضافه می‌کنیم. دو لایه ویژگی‌های پنهان خود را به طور مستقل محاسبه می‌کنند و سپس خروجی آن‌ها در لایه خروجی در هر گام زمانی از = t ۱ تا T = T خلاصه می‌شود. مدل دو جهتی

پیشنهادی قابلیت استفاده از اطلاعات متنی قبلی و آینده برای پیش‌بینی اقدام را دارد. متفاوت از (هنگ‌کنگ، تائو، و فو ۲۰۱۷)، روش ما در استفاده از اطلاعات دو جهتی انعطاف‌پذیر است، در حالی که (هنگ‌کنگ، تائو و فو ۲۰۱۷)تنها اطلاعات مربوط به فریم های آینده را به جریان منتقل می‌کند. علاوه بر این، روش ما اطلاعات غیر ضروری را با استفاده از سلول در LSTM فیلتر می‌کند، در حالی که (هنگ‌کنگ، تائو و فو ۲۰۱۷)ممکن است نویز را در ویژگی‌های آموخته‌شده ایجاد کند.

شبکه دو جریان
ما روش را در) LRCN دوناهیو و همکاران ۲۰۱۴)دنبال کردیم تا دو شبکه جریان ایجاد کنیم که شامل جریان RGB و جریان جریان است. مدل - ResNet ۱۸ بر روی تصاویر RGB برای جریان RGB آموزش داده می‌شود. یک چارچوب جریان توسط Brox) و همکاران ۲۰۰۴(محاسبه می‌شود، و نتایج به "تصویر جریان" با مقیاس بندی و تغییر مقادیر جریان x و y به دامنه [۰، ۲۵۵]، و مرکز جریان x و y به ۱۲۸ تعیین می‌شود. پس از بدست آوردن کانال‌های x وy ، کانال سوم را با استفاده از اندازه جریان محاسبه می‌کنیم. سپس یک مدل - ResNet ۱۸ در آن تصاویر جریان آموزش داده می‌شود. ما از این مدل - ResNet ۱۸ برای استخراج ویژگی‌های جریان برای آموزش جریان جریان استفاده می‌کنیم. همانطور که پیش‌تر ذکر شد، اتصال باقی مانده و تنظیم‌کننده نتیجه جریان جریان را بهبود نمی‌بخشد و در نتیجه the دو جهتی برای آموزش تصاویر جریان استفاده می‌شود. برای ترکیب کردن نتایج هر دو جریان RGB و جریان سیال، ما خروجی جریان RGB و جریان جریان را concatenate و این نتیجه را به مدول حافظه تبدیل می‌کنیم. همچنین چندین راهبرد ترکیبی از جمله تجمع متوسط، تجمع مجموع و حداکثر تجمع روی دو ویژگی را امتحان کردیم، اما منجر به عملکرد ضعیف گردید.
این امکان وجود دارد که ویژگی‌های ویدیوهای جزیی به طور قابل‌توجهی پر سر و صدا باشند. اگر ما اجازه تعامل بین ویژگی‌های دو جریان را بدهیم، سطح نویز بیشتر افزایش خواهد یافت، و در نتیجه ممکن است حافظه را گیج کرده و باعث ایجاد جفت کلید - ارزش در مدول حافظه شود. برعکس، concatenationاجازه چنین تعاملاتی را نمی‌دهد و ویژگی‌های اصلی ویژگی‌ها را حفظ می‌کند در نتیجه منجر به عملکرد پیش‌بینی بهتری می‌شود.

تحلیل احساسات مبتنی بر ویدیو با hvnLBP - TOP و دو - LSTM

روش پیشنهادی معماری شامل ۲ ماژول اصلی است: ماژول استخراج ویژگی و مدول یادگیری متوالی، به اضافه یک ماژول پیش‌پردازش اضافی. جزییات هر ماژول به شرح زیر توضیح داده می‌شود. Preprocessing ابتدا ویدیو به قاب‌ها بریده می‌شود و به صورت تصاویر JPEG ذخیره می‌شود. دوم، ما بزرگ‌ترین چهره انسانی را با استفاده از API های تشخیص چهره شناسایی می‌کنیم. سوم، همه فریم ها در ویدئو توسط یک موقعیت مشابه کراپ می‌شوند، به طوری که ما تصویر چهره انسان را به دست می‌آوریم. در نهایت اندازه آن‌ها را در دنباله‌های تصویر ۹۸ - ۹۸ قرار می‌دهیم. استخراج مشخصه به عنوان یک توصیفگر بافت، hvnLBP شامل حالت‌های صورت ساکن صورت، اما حرکات ماهیچه‌های صورت است. از آنجا که تغییرات زمانی مشابه با بافت فضایی است، ما hvnLBP را بر روی سه صفحه متعامد، یعنی XY، XT و yt برای نشان دادن هر دو حالت صورت و هم حرکات محاسبه می‌کنیم. برای بهینه‌سازی این ویژگی، محاسبه of را تنظیم کردیم. برای تمام نتایج hvnLBP تنها ۶۵ رقم ممکن از ۰ تا ۲۵۵ وجود دارد. با توجه به این، ما یک نقشه بین این رقم در دامنه ۰ - ۶۴ ایجاد می‌کنیم. ما در هر ۵ فریم متصل به هر ۵ فریم متصل (۱ / ۶ ثانیه)همه هیستوگرام را برای هر ۵ * ۱۲ * ۵ تنظیم می‌کنیم. برای کاهش بیشتر ابعاد این ویژگی، تحلیل مولفه‌های اصلی (PCA)برای فشرده‌سازی بعد ویژگی به شکل ۵۱۲ در نظر گرفته می‌شود.

یادگیری متوالی هنگامی که ویژگی تولید می‌شود، در ماژول یادگیری متوالی زیر با لایه‌هایی از جمله: ۱)لایه Bi (Hochreiter و Schmidhuber ۱۹۹۷)؛ ۲)لایه متراکم (۲)لایه متراکم برای طبقه‌بندی (یا رگرسیون)استفاده می‌شود. تابع فعال در لایه آخر به وظیفه خاصی وابسته است: SoftMax برای کار طبقه‌بندی، و sigmoid برای کار رگرسیون.

فصل چهارم : جمع بندی و نتیجه گیری

انتقال کامل ویدیو با سنگ دانه‌های درشت با نما - به - زیبا و جالب توجه

معیارهای ارزیابی و ارزیابی
The توصیف ویدیو مایکروسافت (MSVD)محبوب‌ترین مجموعه داده برای captioning ویدئویی است (گواداراما و همکاران ۲۰۱۳). این فیلم حاوی ۱،۹۷۰ ویدیو بارگیری شده از یوتیوب با حدود ۴۰ توصیف انگلیسی برای هر ویدیو است. به طور معمول، هر کلیپ ویدئویی در مورد یک فعالیت واحد در حوزه‌های باز است. برای مقایسه عادلانه، ما از تنظیمات رایج استفاده‌شده در آزمایش‌ها خود، یعنی ۱۲۰۰ ویدئو برای آموزش، ۱۰۰ ویدیو برای اعتبار سنجی و ۶۷۰ ویدیو برای آزمایش استفاده می‌کنیم. برای ارزیابی عملکرد، ما تمام معیارهای ارزیابی عمومی از جمله bleu (Papineni و همکاران ۲۰۰۲)، cider (Vedantam و همکاران ۲۰۱۵)، meteor (Denkowski و Lavie، ۲۰۱۴)و ROUGE را در نظر می‌گیریم.

تنظیمات آزمایشی
ما ۱۵ فریم را برای هر ویدئو استفاده می‌کنیم. ۳Inception - V (Szegedy و همکاران ۲۰۱۶)و D (Ji۳C و همکاران ۲۰۱۳)برای استخراج ویژگی‌ها برای نمایش ویدیو استفاده می‌شوند. تصاویر ورودی به ۲۹۹ * ۲۹۹، و در نتیجه ابعاد ویژگی‌های چارچوب ۲،۰۴۸ هستند. ویژگی‌های منطقه‌ای قاب‌ها از لایه پایینی ۳of - V استخراج می‌شوند. در اینجا، تور آغازین بر روی ImageNet آموزش‌دیده است (دنگ و همکاران ۲۰۰۹)، و D۳C در ورزش ۱ M (Karpathy و همکاران ۲۰۱۴)قرار دارد. لایه ۳of - V قبل از اتصال کامل (۸ * ۸ * ۲،۰۴۸)برای استخراج ویژگی‌های منطقه استفاده می‌شود. هر فریم می‌تواند با ۸ * ۸ ناحیه شبکه نمایش داده شود. اندازه هسته ۱ - D از سی ان ان ۵ است و تعداد لایه‌های انباشته ۴. ما از توجه موقتی در دو لایه اول استفاده می‌کنیم و توجه را در دو لایه اخیر به ارث برده‌ایم. سر چند سر چند سر، ۸ است. بعد کلمات - تعبیه، ویژگی جهانی، ویژگی چارچوب، و ویژگی منطقه همگی ۵۱۲ مورد می‌باشند.

نتایج و تحلیل‌های آزمایش

مقایسه با رویکردهای دولت - آرت همانطور که در جدول ۱ نشان‌داده شده‌است، نتایج ما را با رویکردهای of مربوط به the مقایسه می‌کنیم. تمام روش‌های انتخاب‌شده مبتنی بر LSTM هستند. می توان مشاهده کرد که روش ما به نتایج بهتری نسبت به رویکردهای stateof - هنر در تمام معیارهای دست یافته‌است. این نشان می‌دهد که مدل مبتنی بر سی ان ان و مکانیسم‌های توجه ما می‌توانند به طور قابل‌توجهی عملکرد کل تولید را بهبود بخشند.

مقایسه با مدل پایه محور اصلی
اگرچه در جدول ۱ اثربخشی رویکرد ما را نشان می‌دهد، اما نمی‌تواند مزایای مدل کاملا convolutional را نسبت به یک مدل مبتنی بر مدل محور نشان دهد، و اینکه آیا دستاوردهای عملکرد به‌دست‌آمده از مکانیزم‌های توجه ما قابل‌توجه است یا خیر. بنابراین ما آزمایش‌های بیشتری برای اعتبار سنجی این جنبه‌ها انجام می‌دهیم. در عین حال، رویکردهای state در جدول ۱، شبکه‌های متفاوتی را به منظور دستیابی به ویژگی‌های ویدئو اتخاذ می‌کنند. برای مقایسه عادلانه، ما به طور خاص نتایج مدل مبتنی بر پایه (LSTM)، مدل CNNbased basic (FCVC)و مدل fully را با توجه (CF - CF & IA)، همانطور که در جدول ۲ نشان‌داده شده‌است، خلاصه می‌کنیم. تمام این رویکردها از ویژگی‌های of استخراج‌شده از ۳Inception - V و D۳C استفاده می‌کنند. نتایج نشان می‌دهد که مدل CNNbased ما مشخصا بهتر از مدل مبتنی بر LSTM عمل می‌کند. شایان‌ذکر است که دقت قابل‌توجه و به ارث برده می‌تواند به میزان زیادی عملکرد captioning را بهبود بخشد، که نقش‌های مهم آن‌ها را تایید می‌کند.

چرا سی ان ان، اما نه.
همانطور که قبلا ذکر شد، مدل مبتنی بر LSTM دارای اشکالاتی است در نتیجه ما از سی ان ان برای غلبه بر این موانع بهره‌برداری می‌کنیم. برای نمایش بیشتر قدرت مدل ما، برخی آزمایش‌ها برای مقایسه بهتر بین مدل مبتنی بر شبکه و LSTM انجام داده‌ایم. نتایج جدول ۱ & ۲ به وضوح برتری سی ان ان را نشان می‌دهد. می توان مشاهده کرد که نتایج یک مدل مبتنی بر شبکه سی ان ان بسیار بهتر از مدل‌های مبتنی بر مدل هستند، و مدل ما می‌تواند جملات دقیق‌تر تولید کند. با توجه به آمار جملات مرجع، می توان یافت که تعداد "# unk" در جملات تولید شده توسط مدل ما بسیار کم‌تر از مدل پایه محور اصلی است. کلمه "# unk" نشان می‌دهد که مدل نمی‌داند که کدام کلمه باید در مرحله زمانی فعلی تولید شود، و در نتیجه منجر به کاهش محکومیت و کاهش نمرات ارزیابی می‌شود. از آنجا که آموزش به سختی برای آموزش دادن دشوار است، آموزش ساختار با یک LSTM انباشت شده دشوار است و به مقدار زیادی زمان نیاز دارد. از سوی دیگر، مدل مبتنی بر شبکه سی ان ان می‌تواند لایه تولید جمله را با لایه‌ای از ساختار انباشته بهینه کند و به زمان بیشتری نسبت به مدل‌های مبتنی بر LSTM نیاز ندارد. خروجی یک شبکه سی ان ان تماما به ورودی وابسته است، که قدرت مدل را برای بهینه‌سازی جملات افزایش می‌دهد در حالی که یک مدل مبتنی بر LSTM پارامترهای زیادی برای بهینه‌سازی دارد. LSTM به حالت پنهان یک گام زمانی قبلی به عنوان ورودی مرحله زمانی فعلی نیاز دارد، که باعث می‌شود مدل زمان یادگیری مدل را آموزش دهد. سی ان ان می‌تواند سریع‌تر آموزش ببیند چون می‌تواند به صورت موازی اجرا شود. در عین حال، این مدل سریع‌تر از یک مدل مبتنیبر محصول حتی با اقدامات بیشتر ناشی از مدل انباشت شده سریع‌تر است. از آنجا که فرآیند استخراج چارچوب برای همه مدل‌ها یک‌سان است، از ویژگی‌هایی استفاده می‌کنیم که از ۳Inception - V pretrained در ImageNet و D۳C پیش از آموزش در Sports - ۱ M استخراج شده‌اند و سپس هزینه دوره آموزشی را مقایسه می‌کنیم. نتایج مربوط به این سه روش در جدول ۳ نشان‌داده شده‌است.

همانطور که انتظار داشتیم، سرعت آموزش مدل مبتنی بر شبکه سی ان ان سریع‌تر از مدل پایه محور اصلی است. مدل کاملا convolutional ما با توجه به زمان بیشتری برای آموزش نیاز دارد، زیرا مدل باید هم ویژگی‌های سطح و هم سطح ناحیه را برای ویدیو در نظر بگیرد. این امر ممکن است عملیات بیشتری را نسبت به مدل پایه محور اصلی و مدل پایه سی ان ان معرفی کند. تمام عملیات مورد استفاده در اینجا بسیار مختصر بوده و مشکلات آموزشی را افزایش نمی‌دهند. همانطور که در شکل ۳ نشان‌داده‌شده، ما چند مثال

captioning از مدل پایه LSTM، مدل پایه سی ان ان، و مدل مبتنی بر سی ان ان را با توجه به کاربرد خشن و به ارث برده‌ایم.

چگونه از توجه استفاده کنیم
همانطور که قبلا توضیح داده شد، ما بر توجه به توجه دقیق و به ارث بردیم، و دستاوردهای مثبت عملکرد را به دست آوردیم. برای بررسی بیشتر برتری مکانیسم‌های توجه ما، چندین روش مختلف را برای بهره‌برداری از مکانیزم‌های مورد توجه پیشنهاد می‌کنیم و آن‌ها را با پایه و اساس مقایسه می‌کنیم. نتایج مربوطه در جدول ۴ نشان‌داده شده‌است. FCVC مدل اصلی است، یعنی، مدل مبتنی بر شبکه سی ان ان بدون توجه. FCVC - ارزیابی تاکتیکی، مدل مبتنی بر شبکه CNN با توجه زمانی در هر لایه از این مدل است. FCVC - IA به این معنی است که ما تنها از توجه به ارث برده‌ایم. در هر لایه، ما از توجه زمانی برای محاسبه وزن فریم های مختلف استفاده می‌کنیم، و سپس توجه به وزن‌های سطح منطقه را به ارث برده‌ایم. FCVC - همه و FCVC - نهایی هر دو توجه خشن را به خود جلب می‌کنند و توجه را به ارث برده‌اند. FCVC - همه از توجه موقتی در دو لایه اول، و هر دو مورد توجه زمانی و به ارث برده در دو لایه آخر استفاده می‌کنند. FCVC - نهایی از توجه موقتی در دو لایه اول استفاده می‌کند و تنها توجه را در دو لایه آخر به ارث برده‌است. می توان مشاهده کرد که با توجه به دقت بسیار خوب و به ارث برده، ما می‌توانیم بهترین نتیجه را بدست آوریم. نتیجه of - نهایی بهتر از FCVC - تماما به این دلیل است که دومی می‌تواند تنها بر روی توجه سطح منطقه در دو لایه آخر تمرکز کند. برای کاهش دشواری یادگیری مدل، ما به طور خاص یادگیری سطح چارچوب و اوزان توجه سطح ناحیه را در لایه‌های مختلف توزیع می‌کنیم. این روش می‌تواند تعارض میان مدول‌های مختلف را کاهش داده و به دستیابی به نتایج بهتر کمک کند. بهتر است که the مدل فقط به بخش کوچکی از محاسبات واقعی اهمیت دهند. ویژگی‌های مورد استفاده در اینجا نیز از ۳Inception - V و D۳C استخراج شده‌اند.

نتیجه‌گیری
در این مقاله، ما یک شبکه کاملا convolutional را با توجه به دقت بسیار خوب و به ارث برده برای ایجاد توضیحات تصاویر ویدیویی پیشنهاد می‌کنیم. ما یک مدل مبتنی بر سی ان ان را برای جایگزینی LSTM که می‌تواند به آموزش سریع‌تر و توصیف بهتر فیلم‌ها دست یابد، می‌سازیم. ما توجه coarse و به ارث برده را توسعه می‌دهیم و امکان‌پذیری آن‌ها را با آزمایش‌ها گسترده اثبات می‌کنیم. نتایج امیدوارکننده در مجموعه داده MSVD اثربخشی مدل ما را تایید می‌کند. از آنجا که به سادگی انباشته کردن سی ان ان نمی‌تواند به طور کامل پتانسیل‌های یک مدل CNNbased را در کاره‌ای نسل توالی درک کند، ممکن است به اندازه کافی برای مجموعه داده‌های پیچیده‌تر سازگار نباشد. ما بررسی یک مدل موثرتر از شبکه CNN را بررسی خواهیم کرد، و توانایی تعمیم آن و قدرت عمومی سازی آن را برای نشان دادن به عنوان نسل در کار آینده مان افزایش خواهیم داد.

ادغام هر دو نشانه‌های بصری و تصویری برای عنوان ویدئو گسترش‌یافته

آزمایش‌ها
نمایش داده و عنوان ویدئو
برای یک ویدیو مشخص، ما ابتدا آن را با تعداد ثابتی از فریم / گیره نمونه در نظر می‌گیریم، سپس مدل سه‌بعدی از پیش آموزش‌دیده برای استخراج ویژگی‌ها استفاده می‌شود. در این میان، ویژگی‌های MFCC هر

کلیپ صوتی استخراج می‌شوند. نمایش ویژگی‌های تصویری و تصویری را می توان به صورت x } = lxs، ۲x، ...، ...، ...، { ... }، و n تعداد فریم های نمونه نمونه‌برداری / گیره نشان داد. برای اعتبار سنجی عملکرد مدل ما، از the Research مایکروسافت تا متنی Dataset (MSR - VTT)و توصیف ویدیو مایکروسافت Dataset (MSVD)استفاده می‌کنیم (چن و Dolan ۲۰۱۱). روش تقسیم آن‌ها را می توان در (ژو و همکاران ۲۰۱۶)و (یائو و سایرین ۲۰۱۵)یافت.

برپاسازی آزمایش در طول آموزش مدل، شروع و برچسب‌های نهایی به ترتیب به هر جمله اضافه می‌شوند. کلماتی که در دایره لغات وجود نداشته باشند با نشانه UNK جایگزین می‌شوند. علاوه بر این، ماسک‌ها به طور جداگانه برای آموزش دسته‌ای بهتر به جملات، بصری و شنوایی اضافه می‌شوند.
پارامترها به صورت زیر تنظیم می‌شوند، اندازه جستجوی تیر، بعد تعبیه کلمه و بعد حالت پنهان LSTM به ترتیب ۵، ۴۶۸ و ۵۱۲ هستند. به ترتیب اندازه حافظه - بینایی و بینایی - متن به ترتیب ۶۴ * ۱۲۸ و ۱۲۸ * ۵۱۲ می‌باشد. برای اجتناب از overfitting، ترک‌تحصیل "(Srivastava و سایرین ۲۰۱۴)" با ۰.۵ نرخ در هر دو لایه اتصال به طور کامل متصل و لایه‌های خروجی "LSTM" استفاده می‌شود، اما نه در گذاره‌ای دوره‌ای میانی. به علاوه، شیب به محدوده [-۱۰، ۱۰] جهت جلوگیری از انفجار گرادیان کاهش یافت. الگوریتم بهینه‌سازی که برای چارچوب‌های deep استفاده می‌شود، ADADELTA (Zeiler ۲۰۱۲)است.
در مورد کیفیت شنیداری، شبکه مکمل، شامل ۳ لایه کاملا متصل برای کدگذار و رمزگشا می‌شود.
واحدهای SI برای کدگذار لایه‌های پنهان ۱۰۲۴، ۵۱۲ و ۲۵۶ نیز به طور جداگانه و ۲۵۶، ۵۱۲، ۱۰۲۴ برای کدگشا برای لایه‌های پنهان هستند.
ارزیابی مدل‌های تلفیقی multimodal ویژگی
ارزیابی عملکرد مدل‌های مختلف multimodal ویژگی
برای اعتبار سنجی کارایی of صوتی به عنوان چارچوب (framework)، ما چندین مدل ترکیب ویژگی را توسعه می‌دهیم و آن‌ها را به صورت زیر ارایه می‌دهیم: concatenating: V - CatL - A / V - A ویژگی‌های تصویری و تصویری قبل / بعد از کدگذار LSTM. VShaWei - A / V - A - A: به اشتراک گذاری weights / memeory در طول مراحل تصویری و صوتی در طول مرحله کدگذاری.
ما مدل‌های ترکیب ویژگی‌های خود را با چندین مدل توصیف ویدئویی مقایسه می‌کنیم، شامل ۳M (Wang و همکاران ۲۰۱۶)، مدل بصری (مدل basic اصلی ما)، مدل صدا (به جای ویژگی‌های بصری). نتایج مقایسه براساس ویژگی‌های بصری D۳C در جدول ۱ نشان‌داده شده‌است.
جدول ۱ نشان می‌دهد که عملکرد مدل‌های ترکیب بصری و صوتی ما، از جمله "V - CatL - A"، " - V - CatHA A" و "V - ShaWei - A" به طور یک‌سان بهتر از مدل‌هایی هستند که فقط بر ویژگی‌های بصری یا تصویری مشروط می‌باشند. علاوه بر این عملکرد "V - cath - A" بهتر از "V - CatL - A" است، که نشان می‌دهد که ویژگی‌های سمعی و بصری در لایه بالاتر کارآمدتر از آن در لایه پایین است.
علاوه بر این، نتایج "VShaMem - A" و "V - ShaWei - A" از مدل‌های "V - CatL - A" و "V - A - A" برتر هستند، که اشاره می‌کند وابستگی موقتی در حوزه‌های صوتی و تصویری می‌تواند باعث افزایش کارایی در caption تصویری شود. علاوه بر این، عملکرد مدل "V - ShaWei - A" از مدل "V - ShaMemA" پیشی می‌گیرد که نشان می‌دهد وابستگی زمانی کوتاه کارآمدتر است.
شاید به این دلیل باشد که وابستگی زمانی کوتاه می‌تواند اطلاعات رزونانسی را در میان شرایط بصری و صوتی موثرتر ثبت کند.
بهترین مدل ما می‌تواند بهبود زیادی را نسبت به ۳M توسط ۳۸.۳ - ۳۵.۱ / ۱ % در BLUE @ ۴ و با افزایش ۱.۵ درصد در امتیاز meteor براساس ویژگی D۳C ایجاد کند

ارزیابی of های Generated مدل‌های مختلف multimodal امکانات تلفیقی شکل ۵، برخی از جملات تولید شده توسط مدل‌های مختلف و واقعیت زمینی humanannotated براساس مجموعه تست VTT - of را نشان می‌دهد. ما می‌توانیم ببینیم که مدل صوتی همیشه جملات غلطی تولید می‌کند، که ممکن است به این دلیل باشد که فقدان of بصری منجر به فقدان اطلاعات جدی می‌شود. از سوی دیگر، VShaWei - یک مدل می‌تواند جملات و اهداف مربوط به اهداف، فعالیت‌ها و اهداف مرتبط را ایجاد کند. در مورد اولین ویدیو، جمله ایجاد شده توسط مدل بصری بیشتر بر روی نشانه‌های بصری تمرکز دارد در حالی که اطلاعات شنیداری را نادیده

می‌گیرد. در نتیجه، آن محتوای اشتباه ("به مرد" در مقابل اخبار)تولید می‌کند. V - کت - مدلی یک شی دقیق "یک مرد" را ایجاد می‌کند و "صحبت می‌کند" در حالی که محتوای "اخبار" را lossing می‌کند. این به این خاطر است که به طور مستقیم، الحاق به ویژگی‌های صوتی و تصویری ممکن است منجر به فروپاشی اطلاعات شود. هر دو A - ShaMem - V و A - ShaWei - V می‌توانند جملات مرتبط تری با کمک نشانه‌های صوتی ایجاد کنند. با توجه به مدل A - ShaMem - V، بیشتر بر روی خلاصه دوره طولانی‌تر اطلاعات تمرکز می‌کند، که طنین در میان شرایط بصری و صوتی را محو می‌کند و یک کلمه انتزاعی دیگر را ارائه می‌دهد. با توجه به مدل وی A - ShaWei -، توجه بیشتری به این رویداد در یک ساختار دقیق‌تر دارد که موضوعات واقعی را نشان می‌دهد، و نشان می‌دهد که می‌تواند اطلاعات رزونانسی در میان دو روش را به طور موثر ثبت کند. در مورد ویدئوی دوم، همه مدل‌ها می‌توانند "شنا" و "هدف" را در آب ایجاد کنند. در حالی که تنها V - ShaWei - مدل یک جسم دقیق ("ماهی" در مقابل "انسان" و "انسان" تولید می‌کند. دلیل این است که A - ShaWei - V می‌تواند هم حرکت و هم جسم حساس به صدا (اطلاعات رزونانسی در میان حالت‌های بصری و صوتی)، غیر از شی استاتیک که شبیه یک انسان است را ثبت کند. در مورد ویدئوی سوم، تنها V - ShaWei - یک مدل عمل مرتبط تری ایجاد می‌کند (نمایش "در برابر بازی")، که نشان می‌دهد که مدل می‌تواند ماهیت یک عمل را ثبت کند. با توجه به این ویدیو، V - کت - A و A - ShaWei - V می‌تواند اقدامات مرتبط تری ("ضربه زدن به یک دیوار"، "استفاده از یک تلفن")با کمک اطلاعات صوتی ایجاد کند. با این حال، V - ShaMem - یک مدل بیشتر بر روی رویداد جهانی تمرکز می‌کند و یک حکم "دروغ بر روی تخت‌خواب" را ایجاد می‌کند. علاوه بر این، مدل بصری توجه بیشتری به اطلاعات بصری می‌کند و همچنین توصیف "دروغ بر روی تخت‌خواب" را تولید می‌کند. در مورد پنجمین ویدیو، رویداد در این ویدیو بیشتر مربوط به اطلاعات دیداری است. در نتیجه، تصویری، A - VShaMem و A - ShaWei - V همه اعمال دقیق‌تر ("رقص" در مقابل "بازی" و "آواز")تولید می‌کنند. علاوه بر این، A - ShaMem - V و A - A - V تعداد دقیق اشیا ("گروهی از" در مقابل یک دختر "،" یک شخصیت کارتونی "و" کسی ")را ارایه می‌دهد، که نشان می‌دهد وابستگی موقتی در شرایط بصری و تصویری برای شناسایی هدف مفید است. علاوه بر این، A - ShaWei - V شی دقیق‌تر ("شخصیت‌های کارتونی" در مقابل "مردم" را ارایه می‌دهد، که اعتبار زمانی کوتاه تری در ثبت اطلاعات رزونانسی

ارزیابی تلفیقی پویای multimodal مشخصه

ارزیابی of صوتی تکمیلی براساس وزارت توسعه اجتماعی

برای تایید اینکه آیا the مکمل صوتی اثرات قابل‌مقایسه با مدل اصلی دارد، ما مدل‌هایی را با مدل V - ShaMem - GA (مشابه با استفاده از ویژگی‌های صوتی تولید شده به جای ویژگی‌های صوتی اصلی)، V - ShaMem - صفر (مشابه با استفاده از صفر برای جایگزین کردن ویژگی‌های صوتی)و مدل بصری براساس مجموعه داده VTT - MSR مقایسه می‌کنیم. VShaWei - GA، V - ShaCatH - GA، V - ShaWei -، V - ShaCatHzero معانی مشابهی را با عبارات مشابه به اشتراک می‌گذارند. نتایج مقایسه در جدول ۲ نشان‌داده شده‌است. مدل‌هایی با ویژگی‌های سمعی و بصری ایجاد شده (V - cath - GA، V - ShaMemGA و V - ShaWei - GA)، از مدل‌های متناظر با ویژگی‌های صوتی تصویری و تصویری صفر (V - صفر، V - صفر و V - صفر - صفر)و مدل بصری برتری دارند، که نشان می‌دهد ویژگی‌های صوتی مکمل اطلاعات مفیدی را منتقل می‌کنند.

ارزیابی of صوتی تکمیلی براساس MSVD

برای بررسی بیشتر اثربخشی ویژگی‌های صوتی مکمل، برچسب ویدئویی را براساس مجموعه داده MSVD که هیچ فرمان صوتی ندارد ارزیابی می‌کنیم. ویژگی‌های صوتی در ابتدا به وسیله شبکه استنتاج modality صدا (AMIN)تولید می‌شوند، و سپس این ویژگی‌ها با اطلاعات دیداری از طریق چارچوب‌های ترکیبی multimodal featuree برای نوشتن ویدئو ترکیب می‌شوند. برای اعتبار سنجی اینکه آیا ویژگی‌های صوتی مکمل حاوی اطلاعات مفید هستند یا خیر، ما مدل‌هایی را با مدل‌های V - ShaWeiGA، V - ShaMem - GA با مدل بصری مقایسه می‌کنیم. علاوه بر این، برای تایید این که آیا pretraining مبتنی بر مجموعه داده مجموعه داده بزرگ، عملکرد را افزایش خواهد داد یا نه، به ترتیب VShaWei - GA - Pre - GA و مدل‌های Pre - GA)، به جز این که قبل از آموزش در مجموعه داده V، مدل‌هایی هستند که اولین pretrained مبتنی بر VTT توسعه اجتماعی هستند. نتایج مقایسه در جدول ۳ ارائه شده‌است. در میان جدول ۳، عملکرد مدل‌های V - ShaWei - GA بهتر از مدل بصری (۳M)، مدل بصری هنری)است، که دوباره تایید می‌کند که مشخصه‌های صوتی مکمل اطلاعات معنی‌دار را برای برچسب ویدئویی حمل می‌کنند. به علاوه، مدل‌هایی که pretraining بهترین عملکرد را

بدست می‌آورند، که دانش بدست‌آمده از دیگر مجموعه داده‌ها را نشان می‌دهد، می‌تواند وظیفه خاص ما را افزایش دهد.
ارزیابی حکم‌های Generated چارچوب یکپارچه Fusion استاندارد multimodal مشخصه
شکل ۶ برخی جملات تولید شده توسط GA را نشان می‌دهد (مدل‌هایی که تنها ویژگی‌های صوتی تولید شده)، ۳M (وانگ و همکاران ۲۰۱۶)، مدل‌های V - GA و حقیقت بر مبنای انسانی براساس مجموعه تست MSVD ارایه می‌دهد. در مورد اولین ویدیو، جمله ایجاد شده توسط مدل بصری بیشتر بر نشانه‌های بصری تمرکز دارد. در نتیجه، محتوای اشتباه "پیانو" ایجاد می‌کند، که به این دلیل است که شی پشت این پسر مانند یک پیانو است و فضای زیادی را در تصویر می‌کشد. V - ShaMem - GA که مجهز به فرمان صوتی تولید شده می‌باشد، هدف بیشتر مربوط به "ویولن" را می‌گیرد، که بعدا تایید می‌کند که روش صوتی مکمل مفید است و مدل V - ShaWei - GA می‌تواند نشانه‌های بصری و صوتی مرتبط را ثبت کند. مدل GA "گیتار" بیشتر از "پیانو" را در مقایسه با واژه دقیق "ویولن" ایجاد می‌کند، که اثربخشی فرمان‌های صوتی تولید شده را تایید می‌کند. با توجه به دومین ویدیو، مدل V - GA قادر به تولید عمل دقیق‌تر ("ریختن سس به یک قابلمه" در مقابل پختن چیزی ")است، که نشان می‌دهد که مدل VShaWei - GA قادر به ثبت اطلاعات رزونانسی در شرایط بصری و صوتی به طور موثر می‌باشد. مشابه مدل V - ShaWei - GA، مدل GA عمل دقیقی را ایجاد می‌کند که نشان می‌دهد مشخصه‌های صوتی تولید شده معنی‌دار است. در مورد ویدئوی سوم، V - ShaWei - GA و GA می‌تواند شی مرتبط تری را ایجاد کند ("دختر در برابر مرد")

ارزیابی ویژگی‌های صوتی تکمیلی براساس MSVD برای تایید بیشتر موثر بودن ویژگی‌های صوتی مکمل، به عنوان مثال زیر را براساس مجموعه داده‌های MSVD ارزیابی می‌کنیم که هیچ فرمان صوتی ندارد. ویژگی‌های صوتی در ابتدا به وسیله شبکه استنتاج modality صدا (AMIN)تولید می‌شوند، و سپس این ویژگی‌ها با اطلاعات دیداری از طریق چارچوب‌های ترکیبی multimodal featuree برای نوشتن ویدئو ترکیب می‌شوند. برای اعتبار سنجی اینکه آیا ویژگی‌های صوتی مکمل حاوی اطلاعات مفید هستند یا خیر، ما مدل‌هایی را با مدل‌های V - ShaWeiGA، V - ShaMem - GA با مدل بصری مقایسه می‌کنیم.

علاوه بر این، برای تایید این که آیا pretraining مبتنی بر مجموعه داده مجموعه داده بزرگ، عملکرد را افزایش خواهد داد یا نه، به ترتیب VShaWei - GA - ShaMem - GA و مدل‌های Pre - GA)، به جز این که قبل از آموزش در مجموعه داده MSVD، مدل‌هایی هستند که اولین pretrained مبتنی بر VTT توسعه اجتماعی هستند. نتایج مقایسه در جدول ۳ ارائه شده‌است. در میان جدول ۳، عملکرد مدل‌های V - ShaWei - GA بهتر از مدل بصری (۳M)، مدل بصری هنری)است، که دوباره تایید می‌کند که مشخصه‌های صوتی مکمل اطلاعات معنی‌دار را برای برچسب ویدئویی حمل می‌کنند. به علاوه، مدل‌هایی که pretraining بهترین عملکرد را بدست می‌آورند، که دانش بدست‌آمده از دیگر مجموعه داده‌ها را نشان می‌دهد، می‌تواند وظیفه خاص ما را افزایش دهد. ارزیابی حکم‌های Generated چارچوب یکپارچه Fusion استاندارد multimodal شکل ۶، برخی جملات تولید شده توسط GA را نشان می‌دهد (مدل‌هایی که تنها ویژگی‌های صوتی تولید شده)، ۳M (وانگ و همکاران ۲۰۱۶)، V - ShaWei - GA و واقعیت زمینی مشروح بر مبنای مجموعه تست of را نشان می‌دهد. در مورد اولین ویدیو، جمله ایجاد شده توسط مدل بصری بیشتر بر نشانه‌های بصری تمرکز دارد. در نتیجه، محتوای اشتباه "پیانو" ایجاد می‌کند، که به این دلیل است که شی پشت این پسر مانند یک پیانو است و فضای زیادی را در تصویر می‌کشد. V - ShaMem - GA که مجهز به فرمان صوتی تولید شده می‌باشد، هدف بیشتر مربوط به "ویولن" را می‌گیرد، که بعدا تایید می‌کند که روش صوتی مکمل مفید است و مدل V - ShaWei - GA می‌تواند نشانه‌های بصری و صوتی مرتبط را ثبت کند.

مدل GA "گیتار" بیشتر از "پیانو" را در مقایسه با واژه دقیق "ویولن" ایجاد می‌کند، که اثربخشی فرمان‌های صوتی تولید شده را تایید می‌کند. با توجه به دومین ویدیو، مدل V - GA قادر به تولید عمل دقیق‌تر ("ریختن سس به یک قابلمه" در مقابل پختن چیزی ")است، که نشان می‌دهد که مدل VShaWei - GA قادر به ثبت اطلاعات رزونانسی در شرایط بصری و صوتی به طور موثر می‌باشد. مشابه مدل V - ShaWei - GA، مدل GA عمل دقیقی را ایجاد می‌کند که نشان می‌دهد مشخصه‌های صوتی تولید شده معنی‌دار است. در مورد ویدئوی سوم، V - ShaWei - GA و GA می‌تواند شی مرتبط تری را ایجاد کند ("دختر در

نتیجه‌گیری در این مقاله، ما سه استراتژی ترکیب چند multimodal برای ادغام اطلاعات صوتی به مدل‌هایی برای اضافه کردن اطلاعات صوتی را پیشنهاد می‌کنیم. هر یک از این سه استراتژی می‌توانند به طور یکنواخت عملکرد of ویدئو را افزایش دهند، که نشان‌دهنده valuableness های صوتی نهفته در ویدئوها است. علاوه بر این، مدل‌های تلفیقی از طریق تقسیم وزن‌ها در حوزه‌های تصویری و تصویری می‌توانند به خوبی اطلاعات را در میان خود مدل کرده و بهترین نتایج را بدست

آورند. علاوه بر این، براساس مدل ترکیب مدل multimodal، یک چارچوب ترکیبی چند multimodal را پیشنهاد می‌کنیم که برای رسیدگی به مسایل مربوط به کیفیت صدا، چارچوب ترکیبی چند multimodal را ارایه می‌دهد.

آن می‌تواند ویژگی‌های صوتی نویدبخش براساس ویژگی‌های بصری متناظر را هنگامی که the صوتی گم شده‌است، ایجاد کند.

شکل caption تصویر مبتنی بر Phrase با شبکه LSTM مراتبی

۶ .آزمایش

۶.۱ datasets .

مدل phi - LSTM در سه مجموعه داده‌های محک آزمایش شد [۱۹] ، Flickr۳۰ [۲۰k] ، و MS - coco [۲۱]. این مجموعه داده‌ها به ترتیب شامل ۸۰۰۰، ۳۱۰۰۰ و ۱۲۳،۲۸۷ هستند، که هر کدام با حداقل پنج شرح تصویر تهیه‌شده توسط انسان از منبع یابی جمعیت تهیه می‌شوند. ما مجموعه داده در دسترس را که در [۵] مورد استفاده قرار می‌گیرد، دنبال می‌کنیم. یعنی اعتبار سنجی و آزمایش هر یک شامل ۱۰۰۰ تصویر برای مجموعه داده‌هایFlickr۳۰ Flickr۳۰k & ۳۰k و ۵۰۰۰ تصویر برای مجموعه داده MS - coco است. بقیه تصاویر برای آموزش استفاده می‌شوند.

ارزیابی معیارهای ارزیابی

ما پنج معیار خودکار را بکار می‌گیریم، از جمله Evaluation ارزیابی]bilingual (bleu)[۲۴ ، متریک ارزیابی تصویر مبتنی بر اجماع]meteor)[۲۵ ، ارزیابی مبتنی بر اجماع]SPICE)[۲۶ [و ارزیابی کیفیت تصویر ایجاد شده]SPICE)[۲۶ .[متریک bleu دقت سازگاری - n گرم را بین یک عنوان تولید شده و تمام جملات مرجع اندازه‌گیری می‌کند، در حالی که متریک متریک به جای دقت محاسبات را اندازه‌گیری می‌کند. در اینجا ما فقط ROUGE - Lرا گزارش کردیم که از طولانی‌ترین توالی مشترک به جای - n گرم استفاده می‌کند meteor .
captionتولید شده و رشته مرجع را با نگاشت هر unigram با استفاده از سه ماژول متفاوت، که "دقیق"، "stem porter"و "WordNet synonymy" هستند، مقایسه می‌کند. امتیاز نهایی - F میانگین محاسبه‌شده از تعداد نگاشت unigram است. متریک cider میانگین cosine هر یک از - N گرم بین عنوان و مرجع تولید شده را با هم ترکیب می‌کند. وزن کم‌تر به - n گرم می‌دهد که معمولا در تمام اشکال مرجع در مجموعه داده‌ها رخ می‌دهد. در آخر، گروه SPICE caption تصویر و منابع آن را به یک گراف صحنه تقسیم می‌کند تا tuples برای هر پیشنهاد معنایی بسازد. سپس، نمره F را در ترکیب هر چند tuples منطقی محاسبه می‌کند.

۶.۳ .جزئیات آزمایشی

گذشته از مدل پیشنهادی فیLSTM - ، ما یک آزمایش بر روی یک مدل پایه انجام دادیم که به عنوان توالی از کلمات عنوان تصویر را پردازش می‌کند. این اساسا یک reimplementation کاری است که در [۴] توضیح داده شد، اما بدون استفاده از مدل‌های طراحی‌شده چندگانه و استفاده از VGGnet [۵۴] به جای GoogleLeNet [۵۵] برای کدگذاری تصویر برای مقایسه عادلانه با مدل ما. تمام تنظیمات تجربی در مدل پایه و سیستم ما یک‌سان هستند مگر این که در غیر این صورت بیان شود. در طول مرحله آموزش، ما از caption خام بدون هیچ preprocessing به عنوان ورودی یک تجزیه‌گر زبان به منظور بدست آوردن یک جفت NPs مناسب‌تر استفاده می‌کنیم. سپس، همه کلمات در جفت NPs به حالت پایین‌تر تبدیل می‌شوند، با حذف punctuations، و کلمه‌ای که کم‌تر از ۵ بار در داده‌های آموزشی دور ریخته می‌شود، به طوری که the های تصویر ما با آن شکل هماهنگ هستند [۵]. برای جلوگیری از انفجار گرادیان به دلیل) to overlength نسبت به طول متوسط همه داده‌های آموزشی)، ما جمله را در جدول ۱ مشخص کردیم. برایNPs overlength ، ما چند کلمه اول را به جای چند کلمه آخرtruncate ، چون بخش دوم NPs معمولا دارای محتوای معنایی بیشتری هستند.

طول جفت شدگی آریل ها، آن‌هایی هستند که بعد از پالایش توصیف‌شده در بخش ۵.۳. مورد بررسی قرار می‌گیرند، به طوری که تعداد of که تحت‌تاثیر قرار می‌گیرند، کم‌تر از ۰.۵ % کل داده‌های آموزشی هستند.

کدگذار سی ان ان مورد استفاده در این مقاله - ۱۶ the ۵۴] pretrained [بر ImageNet است، اما بدون تنظیم پارامترهای شبکه سی ان ان، برای مقایسه عادلانه با نسخه اولیه این کار [۲۲]. ما همچنین نتایج کمی بدست‌آمده با استفاده از ویژگی‌pool ۵ - of ۱۵۲] ۵۶ [را به عنوان کدگذار تصویر برای مراجع در نظر گرفتیم . The LSTMبا اندازه پنهان = K ۲۵۶Flickr) ۸(و = k ۱۲K Flickr۵) ۳۰(& MS - coco به‌k کار گرفته می‌شود. مدل ما باRMSprop ، با استفاده از minibatch ۳۰۰Flickr) ۸(k، ۵۰۰Flickr) ۳۰(و ۷۰۰k)MS - coco) برای هر تکرار بهینه‌سازی شده‌است.
نرخ یادگیری به ۰.۰۰۱ و تنظیمات ترک‌تحصیل برای اجتناب از overfitting تنظیم شده‌است. در طول مرحله آزمایش نشان دادیم که مدل پیشنهادی ما عنوان بهتری با اندازه تیر بزرگ تولید می‌کند، همانطور که در شکل ۷ نشان‌داده‌شده، در حالی که عملکرد مدل پایه زمانی که اندازه تیر بزرگ استفاده می‌شود افت می‌کند.
با توجه به Vinyals و همکاران، یک مدل خوب آموزش‌دیده باید نتیجه بهتری با اندازه پرتوی بزرگ‌تر داشته باشد و بهترین عملکرد را با یک اندازه پرتو نسبتا کوچک نشان دهد (۵۷). با این وجود، ما مدل خود را با استفاده از اندازه تیر = bp ۳۰ و = bs ۲۰ مقایسه کردیم، و مدل پایه با اندازه تیر = b) ۳ و = b ۲۰(تست شد.

مدل ما عنوان را در یک روش دو مرحله‌ای، از ان‌پی کامل به عنوان توصیف کامل در بخش ۴ تولید می‌کند. ما چند نمونه از NPs تولید شده در شکل ۸ را نشان می‌دهیم. برای انتخاب یک مقدار مناسب از[T] T ، ما تغییرات چند متریک و یگانگی جمله را در the های ایجاد شده با استفاده از یک مقدار متغیر در هر مجموعه داده‌ها بررسی می‌کنیم. نتیجه آزمایش مجموعه داده MS - coco در شکل ۹ نشان‌داده شده‌است. مشاهده شده‌است که تمام متریکbleu) n ، cider، meteorو(ROUGE - L به تدریج با آستانه افزایش می‌یابد و در مجموعه داده‌هایFlickr ۸ kوFlickr ۳۰ kبه یک بهینه در- T = ۱.۶ برای مجموعه داده MS - coco می‌رسد.
افزایش بیشتر T تاثیر متفاوتی بر متریک های n مختلف دارد، که در آن bleu و cider کاهش می‌یابند در حالی که meteor و ROUGE - L به طور نامنظم در نوسان هستند. علاوه بر آن، منحصر به فرد بودن جمله به طور مداوم با افزایش T به عنوان نتیجه‌ای از انتخاب کم‌تر نامزدهای NP کاهش می‌یابد. ما همچنین توجه داریم که تغییرات زیادی در معیار SPICE وجود ندارد، که در آن امتیاز در محدوده - ۰.۱۶۵ of در عرض متغیر T نوسان می‌کند. این نشان می‌دهد که مقدار آستانه نه تنها بر روی کلمات تاثیر می‌گذارد و در پیش‌بینی اشیا، ویژگی‌ها و روابط مناسب کمک زیادی نمی‌کند.
مقایسه ۶.۴. با مدل‌هایstate
جداول ۲ و ۳ و ۴ عملکرد مدل ما را در مقایسه با مدل‌های state نشان می‌دهند، در حالی که جدول ۵ عملکرد مدل ما را در مقایسه با مدل پایه مورد ارزیابی با سرور تست آنلاین MS - coco گزارش می‌دهدn - B .، MT، RG، CDو SP به ترتیب برایn ، meteor، ROUGE - L، ciderو SPICE قرار دارند. * نشان می‌دهد که نتایج با ensemblingمدل آموزش‌دیده چندگانه، در حالی که(w) بدست می‌آید(r .و .w) l(r .به ترتیب به ترتیب به ترتیب و بدون عبارت refinement اشاره دارد. (مدل کامل ما است که با - ResNet ۱۵۲ به عنوان کدگذار تصویر آموزش داده می‌شود.
به دنبال آن بگردید. در مقایسه با روش‌هایی که تنها از شبکه سی ان ان به عنوان کدگذار استفاده می‌کنند، مدل ما بهتر یا قابل‌مقایسه با سایر مدل‌های دیگر از مدل‌های هنری است، از جمله مدل‌های مبتنی بر عبارت ارایه‌شده توسط Lebret و همکاران [۳۷] و Ushiku و همکاران [۳۹]. توجه داشته باشید که مدل فعلی ما امتیاز - bleu ۱ و - bleu ۲ را دارد، اما امتیاز - bleu ۳ و - bleu ۴ در مقایسه با نتایج اولیه ما منتشر شده در [۲۲]. این به این دلیل است که نظم کمتری از معیارهای bleu به سمت جمله کوتاه گرایش دارد [۳۸]، ۵، اما ما طول نرمال سازی طول را در الگوریتم جستجوی) beam ۱۵(و (۱۶)اضافه کرده‌ایم تا عنوان طولانی‌تر را ایجاد کنیم. در نتیجه، ما قادر به افزایش میانگین طول شکل ایجاد شده توسط تقریبا سه کلمه، به عنوان مثال از ۶.۸ کلمه (همانطور که در جدول ۲ آمده‌است)تا ۹.۷۲ کلمات برای مجموعه داده‌Flickr ۸ kرا افزایش دهیم. عنوان بیشتر برای مقایسه بهتر با مدل‌های دیگر مطلوب است. جداول ۲ - ۴ همچنین اثربخشی الگوریتم refinement NPرا نشان می‌دهند، چون تقریبا bleu ۱ در همه سه مجموعه داده زمانی که استراتژی پالایش بکار گرفته می‌شود، وجود دارد.
اگرچه ما نتایج بدست‌آمده با - ResNet ۱۵۲ به عنوان کدگذار تصویر برای مرجع آینده را گزارش کردیم، نتایج - VGG ۱۶هنوز در مقایسه با اکثر آثار موجود در جدول مقایسه عادلانه تری دارند. از آنجا که هدف از کار ما بررسی قابلیت یک مدل captioning تصویر مبتنی بر عبارت است، در مقایسه با مدل مشابهی که در دنباله‌های هموار آموزش‌دیده است، ما مکانیزم توجه را پیاده‌سازی نمی‌کنیم و یا اطلاعات اضافی برای مدل خود فراهم نمی‌کنیم، زیرا خارج از محدوده این مقاله است. با این وجود، ما استدلال می‌کنیم که مدل ما با

مدل توجه نرم قابل‌مقایسه است [۹]، که نیازمند محاسبه بیشتر اهمیت نسبی هر مکان در نقشه‌های ویژگی در هر گام زمانی است.

مدل ما عنوان را در یک روش دو مرحله‌ای، از ان‌پی کامل به عنوان توصیف کامل در بخش ۴ تولید می‌کند. ما چند نمونه از NPs تولید شده در شکل ۸ را نشان می‌دهیم. برای انتخاب یک مقدار مناسب از[T] T ، ما تغییرات چند متریک و یگانگی جمله را در the های ایجاد شده با استفاده از یک مقدار متغیر در هر مجموعه داده‌ها بررسی می‌کنیم. نتیجه آزمایش مجموعه داده MS - coco در شکل ۹ نشان‌داده شده‌است. مشاهده شده‌است که تمام متریکbleu) n ، cider، meteorوROUGE - L) به تدریج با آستانه افزایش می‌یابد و در مجموعه داده‌هایFlickr ۸ kوFlickr ۳۰ kبه یک بهینه در- T = ۱.۶ برای مجموعه داده MS - coco می‌رسد. افزایش بیشتر T تاثیر متفاوتی بر متریک های n مختلف دارد، که در آن bleu و cider کاهش می‌یابند در حالی که meteor و ROUGE - L به طور نامنظم در نوسان هستند. علاوه بر آن، منحصر به فرد بودن جمله به طور مداوم با افزایش T به عنوان نتیجه‌ای از انتخاب کم‌تر نامزدهای NP کاهش می‌یابد. ما همچنین توجه داریم که تغییرات زیادی در معیار SPICE وجود ندارد، که در آن امتیاز در محدوده - of ۰.۱۶۵ در عرض متغیر T نوسان می‌کند. این نشان می‌دهد که مقدار آستانه نه تنها بر روی کلمات تاثیر می‌گذارد و در پیش‌بینی اشیا، ویژگی‌ها و روابط مناسب کمک زیادی نمی‌کند.

مقایسه ۶.۴. با مدل‌هایstate

جداول ۲ و ۳ و ۴ عملکرد مدل ما را در مقایسه با مدل‌های state نشان می‌دهند، در حالی که جدول ۵ عملکرد مدل ما را در مقایسه با مدل پایه مورد ارزیابی با سرور تست آنلاین MS - coco گزارش می‌دهدn - B .، MT، RG، CDو SP به ترتیب برایn ، meteor، ROUGE - L، ciderو SPICE قرار دارند. * نشان می‌دهد که نتایج با ensemblingمدل آموزش‌دیده چندگانه، در حالی که(w) بدست می‌آید(r .و .w) ا(r .به ترتیب به ترتیب به ترتیب و بدون عبارت refinement اشاره دارد. (مدل کامل ما است که با - ResNet ۱۵۲ به عنوان کدگذار تصویر آموزش داده می‌شود.

به دنبال آن بگردید. در مقایسه با روش‌هایی که تنها از شبکه سی ان به عنوان کدگذار استفاده می‌کنند، مدل ما بهتر یا قابل‌مقایسه با سایر مدل‌های دیگر از مدل‌های هنری است، از جمله مدل‌های مبتنی بر عبارت ارایه‌شده توسط Lebret و همکاران [۳۷] و Ushiku و همکاران [۳۹]. توجه داشته باشید که مدل فعلی ما امتیاز - bleu ۱ و - bleu ۲ را دارد، اما امتیاز - bleu ۳ و - bleu ۴ در مقایسه با نتایج اولیه ما منتشر شده در [۲۲]. این به این دلیل است که نظم کمتری از معیارهای bleu به سمت جمله کوتاه گرایش دارد [۳۸]، ۵، اما ما طول نرمال سازی طول را در الگوریتم جستجوی) beam ۱۵(و (۱۶)اضافه کرده‌ایم تا عنوان طولانی‌تر را ایجاد کنیم. در نتیجه، ما قادر به افزایش میانگین طول شکل ایجاد شده توسط تقریبا سه کلمه، به عنوان مثال از ۶.۸ کلمه (همانطور که در جدول ۲ آمده‌است)تا ۹.۷۲ کلمات برای مجموعه دادهFlickr ۸ kرا افزایش دهیم. عنوان بیشتر برای مقایسه بهتر با مدل‌های دیگر مطلوب است. جداول ۲ - ۴ همچنین اثربخشی الگوریتم refinement NPرا نشان می‌دهند، چون تقریبا ۱ bleu در همه سه مجموعه داده زمانی که استراتژی پالایش بکار گرفته می‌شود، وجود دارد.

اگرچه ما نتایج بدست‌آمده با - ResNet ۱۵۲ به عنوان کدگذار تصویر برای مرجع آینده را گزارش کردیم، نتایج - VGG۱۶ هنوز در مقایسه با اکثر آثار موجود در جدول مقایسه عادلانه تری دارند. از آنجا که هدف از کار ما بررسی قابلیت یک مدل captioning تصویر مبتنی بر عبارت است، در مقایسه با مدل مشابهی که در دنباله‌های هموار آموزش‌دیده است، ما مکانیزم توجه را پیاده‌سازی نمی‌کنیم و یا اطلاعات اضافی برای مدل خود فراهم نمی‌کنیم، زیرا خارج از محدوده این مقاله است. با این وجود، ما استدلال می‌کنیم که مدل ما با مدل توجه نرم قابل‌مقایسه است [۹]، که نیازمند محاسبه بیشتر اهمیت نسبی هر مکان در نقشه‌های ویژگی در هر گام زمانی است.

۷.۳ .محدودیت‌های مدل مشاهده‌شده با آنالیز کیفی

برای به دست آوردن دیدگاه بیشتر در مورد این که چگونه تعداد وقوع هر کلمه در مجموعه آموزشی بر پیش‌بینی کلمه "هنگام تولید" تاثیر می‌گذارد، حداقل پنج کلمه برتر را ثبت می‌کنیم که از هر دو مدل در جدول ۸ استنباط می‌شود. سپس، ما به صورت دستی تصاویر را بررسی می‌کنیم که حاوی این کلمات هستند، و کلماتی را که به درستی در توصیف تصویر مربوطه شان استفاده می‌شوند را برجسته می‌کنیم. در شکل ۱۰ به عنوان مثال، یک جفت تصویر caption - برخی از کلمات به درستی استنباط شده در شکل ۱۰ نشان‌داده شده‌است. از جدول ۸ می‌توانیم ببینیم که مدل مبتنی بر phrase به طور کلی قادر به استنباط درست کلماتی

است که کم‌تر دیده می‌شوند در مقایسه با مبنا. تنها استثنا در مجموعه دادهFlickr ۸ kوجود دارد، که در آن خط مبنا به طور صحیح کلمه "اسنوبورد" را تشخیص می‌دهد که تنها ۴۴ بار دیده می‌شود. عنوان مربوطه در تصویر اول در شکل ۱۰ نشان‌داده شده‌است، و متوجه شدیم که مدل ما "snowboarder" را برای آن تصویر استنتاج کرده‌است، که به طور طبیعی باعث می‌شود نسل اجرایی "زیاده‌روی" کند. علاوه بر این، ما پنج کلمه برتر را ثبت کردیم، که اغلب آن‌ها در تصاویر ایجاد شده در مدل ما و اساس در جدول ۹ غایب هستند. از این جدول، مشاهده می‌کنیم که کلماتی که در آن مدل ما قادر به استنباط است در حالی که اساس نمی‌تواند وجود داشته باشد: "یکFlickr) "۸(k، "یکFlickr) "۸(k، "یک" و "سه .(MS - coco) "در مورد کلمه"a" ، این به این دلیل است که مجموعه تست of دارای ویژگی‌های بیشتری با شروع با حروف صدادار در مقایسه با اشیا، مانند "یک پیراهن نارنجی" و "بازار بیرونی" است که در شکل ۱۱ نشان‌داده شده‌است. چنین دنباله‌ای معمولا امتیاز پایینی دارد تا زمانی که کلمه شی پیش‌بینی شود (یعنی امتیاز توالی یک فضای باز "بسیار پایین‌تر از" بازار "قبل از واژه سوم" بازار " پیش‌بینی می‌شود). با توجه به ردیف دوم و حذف روند جستجوی تیر در هر مرحله زمانی، توالی با امتیاز کم‌تر در مرحله زمانی اولیه، به راحتی از بین می‌رود، به خصوص آن‌هایی که توالی قبلی قبلی دارند. بنابراین، ایجاد عنوان در یک روش مبتنی بر کلام، از چنین مشکلی اجتناب می‌کند، زیرا نتیجه توالی جمله کوتاه، تاثیر کمتری از کلمات قبلی در طول جستجوی پرتو دارد.
در مورد کلمه "آنجا"، این مساله را می توان از مدل ما استنباط کرد، زیرا رمزگشایی به صورت جداگانه را تجزیه و تحلیل می‌کنیم، که به طور طبیعی پیش‌بینی کلمه "را" وظیفه of عبارت " می‌سازد. بدون کلمه "a" به عنوان رقیب، واژه "آنجا" به احتمال بیشتری به عنوان اولین واژه در یک عنوان پیش‌بینی می‌شود. در مورد کلمه "سه" با کلمه "دو" به عنوان رقیب استفاده می‌شود. اینها دلایلی هستند که مدل ما قادر به ایجاد تصاویر captions منحصر به فرد در مقایسه با مبنا است. از سوی دیگر، مدل پایه شانس بهتری برای پیش‌بینی واژه "بالا" و "ناشی از" با تاثیر بیشتر کلمات قبلی دارد. واژه‌های دیگری که نمی توان از هر دو مدل استنباط کرد معمولا کلماتی جایگزین دارند که امتیاز بالاتری دارند. برای مثال، "پسر / دختر" و "بعدی"، گزینه بهتری برای "فرزند" و "توسط" است. علاوه بر این، هر دوی این مدل‌ها قادر به استنباط "همزمان" نیستند "و"، که اغلب برای توصیف فعالیت‌های چندگانه انجام‌شده توسط افراد

۸ .نتیجه‌گیری

این مقاله یک مدل مفهومی مبتنی بر عبارت) فی(LSTM - را برای تولید عنوان تصویر در یک روش سلسله مراتبی ارایه داد، که در آن NPs که اشیا برجسته را در یک تصویر توصیف می‌کنند، ابتدا تولید می‌شوند، قبل از این که یک عنوان کامل از NPs تشکیل شود. هر یک NP تولید شده به صورت یک بردار ترکیبی کدگذاری می‌شود که به عنوان ورودی یک گام زمانی در سطح جمله عمل می‌کند. چنین طراحی اجازه می‌دهد تا NPs در یک مقیاس زمانی ثابت، رمزگشایی شوند، در حالی که تنوع تفکیک زمان - مقیاس در سطح محکومیت را کاهش می‌دهد. نتایج تجربی نشان می‌دهند که caption تصویر تولید شده در چنین حالتی، در مقایسه با یک مدل ترتیبی خالص با استفاده از کلمات به عنوان واحد اتمی، دقیق‌تر است. علاوه بر این، فرآیند رمزگشایی سلسله مراتبی امکان ایجاد تصاویر جدید را با کلمات مختلف فراهم می‌کند که باید تولید شوند. کار آینده ما بر طراحی یک مدل دو جهتی مبتنی بر اصطلاح برای captioning تصویر متمرکز خواهد بود.

پیش‌بینی اکشن از تصاویری از memorizing Hard - to – Predict

datasets

ما از - the ۱۰۱) سومرو، ضمیر، و شاه ۲۰۱۲)و مجموعه داده‌های ورزشی - ۱ Karpathy) M و همکاران ۲۰۱۴(استفاده می‌کنیم تا روش ما را ارزیابی کنیم. مجموعه داده ۱۰۱ - ۱۰۱ شامل ویدیوهایی است که در ۱۰۱ اقدامات انسانی، مانند "تبلیغ بیسبال" و "گیتار" پخش شده‌اند. این مجموعه داده‌ها شامل ۱، ۱۳۳، ۱۵۸ ویدیو است که به ۴۸۷ ویدئو پخش شده‌اند. در این کار، ما از (هنگ‌کنگ، تائو، و فو ۲۰۱۷)پیروی می‌کنیم و تنها در بخشی از مجموعه داده - M ۱ M برای انجام یک مقایسه عادلانه تست می‌کنیم. ما از ۵۰ کلاس اول در مجموعه داده - M ۱ M استفاده می‌کنیم و فیلم‌های ۹۲۲۳ نمونه به دنبال (هنگ‌کنگ، تائو، و فو ۲۰۱۷)را انتخاب می‌کنیم.

جزییات اجرا
ما - the ۱۸) او و همکاران ۲۰۱۵)و - VGG ۱۹ برای داده‌های جریان نوری و نوری جداگانه استفاده کردیم - The . ۱۸ در مجموعه آموزشی - ImageNet ILSVRC ۲۰۱۲ آموزش‌دیده است (دنگ و همکاران ۲۰۰۹). سپس - the ۱۸ روی مجموعه دادهUCF ۱۰۱ تنظیم می‌شود و می‌تواند به ۶۹.۱ % دقت برسد - Pre - .۱۸ در ImageNet می‌تواند به طور قابل‌توجهی مشکل overfitting مدل را کاهش داده و دقت % from به ۶۹.۱ % را افزایش دهد. ما جریان نوری را از صفر به سی ان آموزش می‌دهیم. در تنظیماتLSTM RGB ، ویدئو خام به یک توالی تصویر تبدیل می‌شود. یک شبکه - ResNet ۱۸ تنظیم‌شده برای استخراج ویژگی‌ها از فریم‌های ویدئوی مورد استفاده قرار می‌گیرد. ورودی to ویژگی‌های استخراج‌شده از هر فریم ویدیویی است. اندازه ورودی و خروجی of ۵۱۲ مورد است. مدل دو جهتی LSTM شامل یک لایه LSTM جلو و یک لایه LSTM عقب است، که هر لایه LSTM اتصال باقی مانده و تنظیم‌کننده را دارد. هر گام زمانیof ، یک لایه خطی (اندازه: ۵۱۲، تعداد کلاس)و یک لایه Logsoftmax به عنوان طبقه بند را دنبال می‌کند. در تنظیمات نمودار جریان ما، ویژگی‌های جریان از تصاویر جریان با استفاده از شبکه سی ان به عنوان کانال RGB استخراج می‌شوند. اندازه ورودی ofبرابر با ۴۰۹۶ و اندازه خروجی of ۵۱۲ می‌باشد. ما تنها از ۱ لایه در بالای مشخصه‌های جریان استفاده می‌کنیم. هر مرحله از LSTM نیز یک لایه خطی و یک لایه Logsoftmax را به عنوان طبقه‌بندی کننده دنبال می‌کند. تنظیمات آموزش مفصل مانند LSTM RGB است. حافظه در روش ما را می توان به عنوان یک طبقه‌بندی کننده دید.
خروجی‌های هر دو پروفایل های RGB و LSTM جریان از تمامی مراحل زمانی به طور متوسط و سپس به مدول حافظه داده می‌شوند. مقدار خروجی حافظه برچسب کلاس است. در طول آموزش و آزمایش، اندازه حافظه تا ۵۰۰۰ تنظیم شده‌است، Kبرای ۱۶ تنظیم شده‌است و اندازه کلید ۱۰۲۴) معادل مجموع of RGB و (Flow Analysisاست. تمامی این شبکه‌ها با استفاده از الگوریتم گرادیان تصادفی(SGD) که بر روی یک تیتان Xsingle اجرا می‌شود، آموزش داده می‌شوند.

نتایج
عملکرد پیش‌بینی ما روش خود را با روش پویای کیف - کلمات(DBoW) ، b) bagو همکاران ۲۰۱۵)، MTSSVM .Chen et al)، ۲۰۱۲)،) DeepSCNهنگ‌کنگ، تائو، و فو ۲۰۱۷)بر روی مجموعه داده - UCF ۱۰۱ و مجموعه داده‌های ورزشی ۱ M مقایسه می‌کنیم. روش‌های RGB و روش‌های RGB - BiLSTM به عنوان روش‌های پایه استفاده می‌شوند. توجه داشته باشید که تمام این روش‌های مقایسه باید نسبت مشاهده برای پیش‌بینی دقیق را بدانند، در حالی که روش ما به آن نیازی ندارد - UCF .۱۰۱ مجموعه داده‌ها. نتایج ارایه‌شده در شکل ۶ (a)نشان می‌دهد که روش ما به طور پیوسته به نتایج برتری نسبت به روش‌های پیش‌بینی رفتار حالت هنر در مجموعه داده ۱۰۱ - ۱۰۱ دست می‌یابد. روش ما به ترتیب برابر با ۶ % و ۳.۳۳ % در نسبت‌های مشاهده ۰.۱ و ۰.۲ می‌باشد. این نشان می‌دهد که ماژول حافظه در شبکه‌های ما نمونه‌های پیش‌بینی hardto را حفظ می‌کند و در نتیجه می‌تواند به افزایش کارایی پیش‌بینی کمک کند. همانطور که ویدیو در مجموعه داده ۱۰۱ - ۱۰۱ مشاهده می‌شود، تشخیص نمونه‌های بیشتری در شبکه ما دشوار نیست، و بنابراین فاصله عملکرد بین روش ما و DeepSCN به حدود ۲ % در مشاهدات ما کاهش می‌یابد. تمام این روش‌ها با ویژگی‌هایC ۳ Dتغذیه می‌شوند. با این حال، تمام این روش‌ها یک مدل برای هر نسبت مشاهده ایجاد می‌کنند و به طور خاص نمونه‌های دشوار را حفظ نمی‌کنند. علاوه بر این، این روش‌ها باید نسبت یک نمونه آزمایش را بدانند و بنابراین عملی نیست. روش ما نسبت بهIBoW ، DBoWو MSSC با ۱۴.۷۲ %، ۱۴.۷۲ % و ۱۶.۹۷ %، در نسبت مشاهده ۱، بهتر عمل می‌کند. روش ما همچنین بهتر از SVM + SVM با هسته chi بهتر عمل می‌کند. روشC ۳D، پیچ‌ها ۳ Dرا بر روی ویدئو برای بازشناسی عملیات انجام می‌دهد، اما برای پیش‌بینی عمل بهینه نشده است. با مشاهده نسبت ۰.۱، روش ما از روشC ۳ Dبا ۱۱.۰۲ % بهتر عمل می‌کند، نشان می‌دهد که مزایای ماژول حافظه و لایه‌های bi دو جهتی را در روش ما برای پیش‌بینی عمل نشان می‌دهد. روش ما به طور پیوسته بهتر از RGBو - LSTM RGB است که نشان می‌دهد سود استفاده از جریان جریان در مرحله پیش‌بینی را نشان می‌دهد. در نسبت مشاهده ۰.۱، فاصله عملکرد بین روش ما و دو جریان(RGB) RGB وRGB) (Bi RGB حدود ۱ % است. با این حال، به عنوان فریم های بیشتری مشاهده می‌شود، این فاصله به ترتیب ۱۹.۰۶ % برای LSTM RGBو ۱۵.۷۵ % برای - LSTM RGB را نشان می‌دهد که اهمیت استفاده از جریان جریان در وظیفه پیش‌بینی را نشان می‌دهد.
این نتیجه همچنین نشان می‌دهد که اطلاعات جریان ممکن است در فریم اول ویدیو بسیار متمایز نباشند. اکثر ویدئوها تنها یک حرکت غیر اطلاع‌رسانی در فریم اول دارند و در نتیجه اطلاعات جریان استخراج‌شده از این فریم ها پر سر و صدا هستند. مجموعه داده - M. ۱ روش ما همواره بهتر از تمام روش‌های مقایسه عمل

می‌کند. در مقایسه با روش‌های موجود پیش‌بینی اقدام موجودDeepSCN ، MTSSVM، MSSC، IBoWو DBoW، روش ما به نتایج برتری نسبت به تمام نسبت‌های مشاهده دست می‌یابد. با اینکه DeepSCN اطلاعات متمایزی را از ویدیوهای کامل به ویدیوهای جزیی منتقل می‌کند، ممکن است این روش انتقال داده شود که منجر به مساله انتقال منفی می‌شود. علاوه بر این، نگاشت انتقال لزوما یک خط خطی است، و بنابراین ممکن است اساسا قدرت ویژگی‌های چند فریم اول را بهبود ندهد.
در مقایسه، روش ما از ماژول حافظه برای کشف نمونه‌های سخت برای پیش‌بینی استفاده می‌کند، که به ما این امکان را می‌دهد تا مرز طبقه‌بندی پیچیده تری را یاد بگیریم. علاوه بر این، LSTMدو جهتی در روش ما از نگاشت غیر خطی برای انتقال اطلاعات از فریم های آینده به چارچوب‌های فعلی استفاده می‌کند در نتیجه قدرت تفکیک پذیری ویژگی‌ها را بهبود می‌بخشد. بهبود روش ما نسبت به MTSSVM و MSSC در نسبت مشاهده ۰.۱ درصد و ۱۰.۹ % است، که قابل‌توجه است. بزرگ‌ترین شکاف عملکرد بین روش ما و IBoW در نسبت مشاهده ۱.۰ % است؛ و % it بین روش ما و DBoW در نسبت مشاهده رخ داده‌است. روش ما همچنین بهبود قابل‌توجهی را بر روی LSTM RGB و - Bi RGB نشان می‌دهد که مزایای استفاده از جریان جریان را نشان می‌دهد. جدول ۱ عملکرد پیش‌بینی آنی را نشان می‌دهد که در آن تنها ۱۰ % فریم ها به همه روش‌ها مشاهده می‌شوند. روش ما به ۵۱.۰۲ % روی مجموعه داده - UCF ۱۰۱ و ۵۷.۶۰ % بر روی مجموعه داده - M ۱ M می‌رسد، که بالاتر از تمام روش‌های مقایسه است. با تشکر از ماژول حافظه، روش ما زمانی عمل می‌کند که فریم های کمتری وجود داشته باشند، حتی اگر تشخیص آن‌ها دشوار باشد. توجه داشته باشید که روش ما نیاز به نسبت‌های مشاهده در تست ندارد، در حالی که روش‌های موجود مانندDeepSCN ، MTSSVM، MSSC، IBoWو DBoW همگی به این اطلاعات نیاز دارند. این نشان می‌دهد که روش ما عملا عملی است

یون. این نشان می‌دهد که روش ما عملا عملی است. جدول ۱ یک مشاهده جالب را نشان می‌دهد، که بهبود روش ما بر روی روش دوم در مورد ورزش - M ۱ است که کمی بیشتر از یک مجموعه داده - the ۱۰۱ است. یک دلیل ممکن این است که اطلاعات پس‌زمینه در مجموعه داده - M ۱ در چارچوب‌های اول اطلاعات بیشتری نسبت به مجموعه داده ۱۰۱ - ۱۰۱ وجود دارد، و در نتیجه می‌تواند برای ایجاد فعالیت‌های مختلف ورزشی مورد استفاده قرار گیرد. بنابراین، حافظه مجموعه داده - M ۱ M به طور موثرتر از اولی روی مجموعه داده ۱۰۱ - ۱۰۱ کار می‌کند چون انتظار می‌رود که ویژگی‌ها را از فریم های اولیه به یاد داشته باشد.

مقایسه با انواع مختلف mem - LSTM را با انواع مختلف مقایسه می‌کنیم تا اثربخشی هر ماژول را در روش ما نشان دهیم. این متغیرها عبارتند از - ResNet :۱۸، LSTM، LSTM، and Bi، LSTM flow، RGB Bi + LSTM .Flowهمه این متغیرها فقط یک جریان RGB هستند. مقایسه نتایج با mem - LSTM در نسبت مشاهده ۰.۵ و ۱.۰ در جدول ۲ نشان‌داده شده‌است. روش ما بهبود قابل‌توجهی را نسبت به تمام روش‌های مختلف در نسبت‌های مشاهده ۰.۵ و ۱.۰. به دست می‌آورد که مزیت استفاده از ماژول حافظه را می توان در تفاوت عملکرد روش ما و روش‌ها بدون حافظه مشاهده کرد. این بهبود به ترتیب ۵.۲۱ % و ۴.۶۹ % در نسبت مشاهده ۰.۵ و ۱.۰ است. این نشان می‌دهد که حافظه در روش ما هنوز می‌تواند برخی از نمونه‌های چالش برانگیز را به خاطر بسپارد تا در مرحله میانی یا دیر ویدیو دسته‌بندی شود، در نتیجه مرزه‌ای طبقه‌بندی پیچیده را ایجاد کرده و عملکرد را بهبود بخشد. مزیت استفاده از اتصالات باقیمانده در LSTM را می توان از بهبود of مانده بر روی LSTMوانیل مشاهده کرد. اضافه کردن اتصالات باقیمانده در حدود ۱ % بهبود را افزایش می‌دهد - Bi .با ۲.۱۹ % و ۳.۲۷ % بهتر عمل می‌کند و اهمیت استفاده از یک لایه LSTM backward بر بالای the وانیل را نشان می‌دهد. لایه LSTM اطلاعات متمایزی را از قالب‌های آینده به چارچوب‌های فعلی منتقل می‌کند.
حتی اگر ویژگی‌های مشاهدات فعلی به اندازه کافی متمایز نباشند، اطلاعات منتقل‌شده می‌تواند ویژگی‌ها را غنی کرده و در نتیجه عملکرد پیش‌بینی را افزایش دهد. ما همچنین عملکرد جریان جریان در این آزمایش را مورد آزمایش قرار می‌دهیم. یک LSTM وانیل روی یک سری از مشخصه‌های جریان استخراج‌شده با استفاده از - ۱۸ResNet اجرا می‌شود. نتایج نشان می‌دهند که جریان جریان عملکرد مشابهی را با جریان RGB (vanilla LSTM)به دست می‌آورد. ما همچنین عملکرد of در جریان‌های جریان را مورد آزمایش قرار می‌دهیم. با این حال، عملکرد آن نسبتا پایین است، احتمالا به این دلیل که نویز ناشی از جریان نوری از فریم های آینده به چارچوب‌های فعلی منتقل می‌شود.

نتیجه‌گیری

در این مقاله، ما یک شبکه پیش‌بینی اقدام جدید با هدف بهینه‌سازی عملکرد در مرحله آغاز ویدئو ارائه کرده‌ایم. ما یک ماژول حافظه را معرفی می‌کنیم که نمونه‌های دشوار را به خاطر دارد. با به حداقل رساندن اتلاف حافظه، شبکه کلیدهای حافظه را با جایگزین کردن کلید با پرس و جو و یا به طور میانگین اصلی و پرس و جو به روز می‌کند. این موضوع اساسا به ما این امکان را می‌دهد تا مرزهای طبقه‌بندی پیچیده را بیاموزیم که به ویژه برای متمایز ساختن ویژگی‌ها از چند فریم بسیار مفید است. برای افزایش بیشتر قدرت تفکیک پذیری ویژگی‌ها، ما پیشنهاد می‌کنیم که از اطلاعات مربوط به آینده توسط یک LSTM دو جهتی استفاده کنیم. ما همچنین اتصالات باقیمانده را به the اضافه می‌کنیم و لایه‌های پنهان را با افزودن یک محدودیت به شبکه LSTMاضافه می‌کنیم. آزمایش‌ها گسترده در مجموعه داده‌های ۱۰۱ - ۱۰۱ و ورزشی - M ۱ نشان می‌دهد که روش ما در مقایسه با روش‌های پیشرفته، به ویژه هنگامی که تنها چند فریم اولیه مشاهده می‌شوند، به کارایی بالای پیش‌بینی دست می‌یابد. با بسیاری از رویکردهای پیش‌بینی عملیات موجود، روش ما نیاز به دانستن نسبت‌های مشاهده نمونه‌های تست ندارد. این ویژگی جذاب، رویکرد ما را عملی می‌کند.

تحلیل احساسات مبتنی بر ویدیو با hvnLBP - TOP و دو – LSTM

آزمایش‌ها برای آزمایش تاثیر ویژگی پیشنهادی، چندین آزمایش بر روی دو مجموعه داده طراحی کردیم: داده‌های moud (پرز - Rosas ،Mihalcea ،و Morency۲۰۱۳)و مجموعه داده‌های CMUMOSI (زاده و همکاران ۲۰۱۶). مجموعه داده moud شامل ۴۹۸ کلمه ویدئوی بازبینی محصول در زبان اسپانیایی است. هر ویدیو شامل چندین گونه مرور محصول با برچسب‌های احساسی مثبت، منفی یا خنثی است، و ما فقط ۴۵۰ utterances برچسب دار مثبت یا منفی برای استفاده تجربی داریم. مجموعه داده CMU - MOSI شامل بخش‌های ۲۱۹۹ از ۹۳ ویدیو نظرات فیلم در یوتیوب است. هر بخش در محدوده (۳ و ۳)برچسب زده می‌شود. در میان همه بخش‌های ۲۱۹۹، فقط ۱۱۶ مورد از آن‌ها طول بیش از ۱۰ ثانیه طول کشیده. بنابراین در آزمایش ما، بخش‌های طولانی‌تر از ۱۰ ثانیه با بریدن دم دراز مدت ۱۰ ثانیه کوتاه می‌شوند. در آزمایش moud، طبقه‌بندی احساسات دوگانه را انجام می‌دهیم و از دقت و نمره ۱F برای ارزیابی عملکرد استفاده می‌کنیم. ما به طور تصادفی ۶۴ % را انتخاب کردیم: ۱۶ %: ۲۰ % برای آموزش، اعتبار سنجی و مجموعه تست. به طور خاص، از ۲۸۸ نمونه برای آموزش، ۷۲ برای اعتبار سنجی و ۹۰ مورد برای تست استفاده کردیم. در آزمایش MOSI، طبقه‌بندی احساسات دوگانه، طبقه‌بندی احساسات پنج طبقه و رگرسیون احساسی را انجام می‌دهیم. ما از هر دو دقت و امتیاز ۱F برای طبقه‌بندی باینری، دقت طبقه‌بندی کلاس پنج کلاس و خطای مطلق میانگین (MAE)برای تکالیف رگرسیون استفاده می‌کنیم. میزان آموزش، اعتبار سنجی و مجموعه تست ۱۳۷۶: ۴۷۹: ۴۷۹ قرار دارد.

ما عملکرد مدل خود را با مدل‌های state بعدی برای تحلیل احساسات ویدئویی مقایسه می‌کنیم: AU + SVM (پرز - Rosas و همکاران ۲۰۱۳)یک مدل SVM را بر روی ویژگی‌های واحد اکشن آموزش می‌دهد و ویژگی - theart را برای تحلیل احساسات ویدئویی در مجموعه داده moud نگه می‌دارد. V - tfn (زاده و همکاران ۲۰۱۷)the بصری مدل تحلیل احساسات چند - در مجموعه داده of - CMU است.

نتیجه‌گیری این مقاله یک ویژگی بیان چهره جدید برای تحلیل احساسات ویدئویی را پیشنهاد می‌دهد. ما از یک معماری یادگیری ماشینی برای تایید ویژگی پیشنهادی خود استفاده می‌کنیم. نتایج آزمایش نشان‌دهنده کارایی این ویژگی است. تقدیر و تشکر: پروژه (گرانت :No۶۱۶۷۳۲۳۵)با حمایت بنیاد ملی علوم طبیعی چین پشتیبانی می‌کند.

مراجع :

همون منابع تو pdf های نامبرده