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 پدر و نژاد و زیبای و ثروت است

فهرست مطالب

مقدمه

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 .

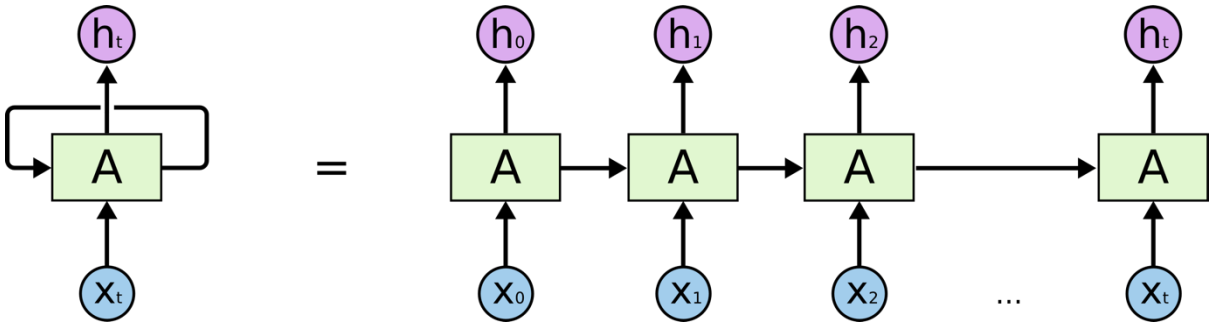
فصل دوم : تعریف دقیق مسئله و داده ها و روشهای حل

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 .

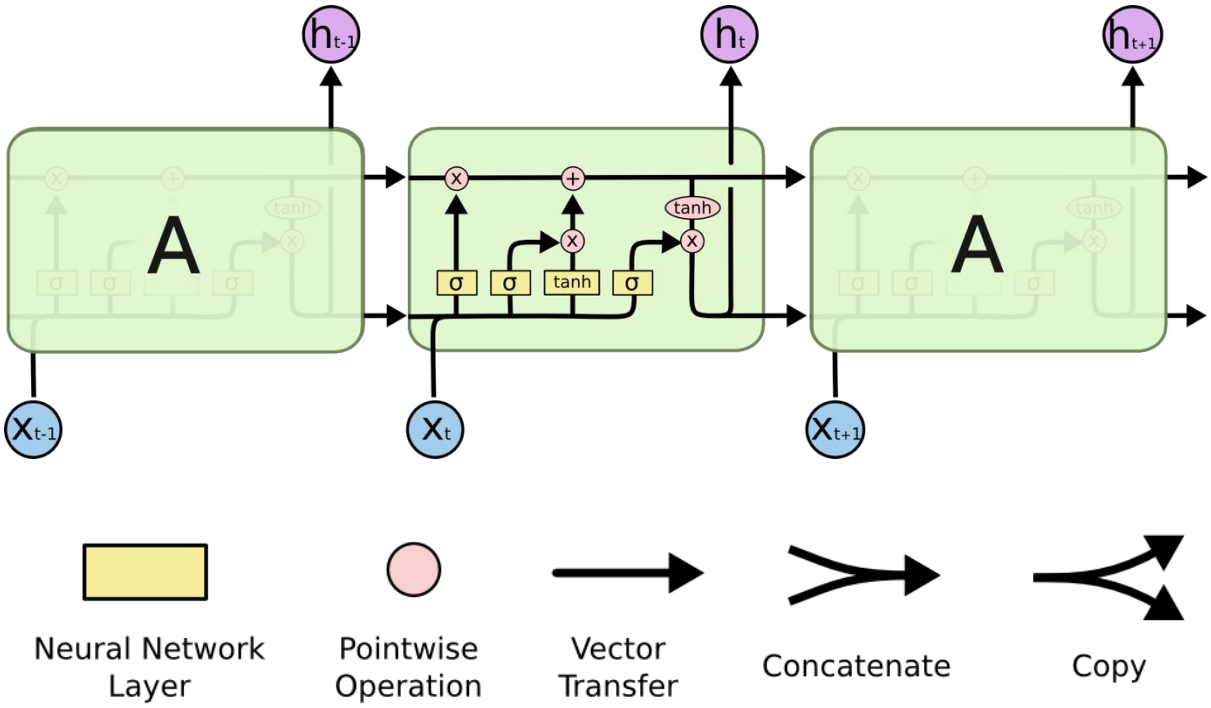


همانطور که در تصویر بالا میبینید به این صورت می توان یک عملیات فیدبک رو با شبکه عصبی ساده پیاده سازی کرد

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 .

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 تجریه 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

فصل چهارم : جمع بندی و نتیجه گیری

مراجع :