مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال دهم، شماره دوم، تابستان ۱۴۰۲

توصیف تصاویر مبتنی بر شبکه عمیق رمزگذار-رمزگشا و سازوکار توجه بر توجه

زهرا فامیل ستاری، حسن ختن لو و الهام علیقارداش

چکیده

توصیف تصویر یک زمینه تحقیقاتی بین رشته ای در بینایی ماشین و پردازش زبان طبیعی است. بسیاری از روش‌های پیشنهاد شده برای تولید توصیف تصویر از چارچوب رمزگذار - رمزگشا پیروی کرده‌اند. به این ترتیب هر کلمه بر اساس ویژگی‌های تصویر و کلمات تولید شده قبلی تولید می‌شود. اخیراً سازوکار توجه، که میتواند با ایجاد نقشه فضایی، مناطق مرتبط تصویر با هر کلمه را برجسته کند، به طور گسترده در تحقیقات استفاده شده است. در این مقاله، ما یک روش جدید را پیشنهاد کرده‌ایم که چارچوب رمزگذار-رمزگشا را با سازوکار توجه و سازوکار توجه بر توجه ادغام کرده است. بخش رمزگذار مدل شامل چند بخش ResNet، Attention-LSTM، Multi Head Attention و Attention on Attention است. از ResNet برای استخراج ویژگی های کلی تصویر استفاده شده است. ایه‌ی Language-LSTMمسئولیت رمزگشایی را بر عهده دارد. سازوکار توجه از شواهد محلی برای افزایش نمایش ویژگی‌ها و استدلال در تولید توصیفات تصویری بهره برده و سازوکار توجه بر توجه می‌تواند روابط اشیای داخل تصاویر را به‌خوبی درک کند. این روش پیشنهادی توانسته است بر روی تصاویر مجموعه‌های داده Flickr8k و MSCOCOتوصیف‌های بهتری را نسبت به روشهای موفق موجود ارائه دهد. همچنین بر اساس معیارهای ارزیابی METEOR، ROUGEعملکرد توصیف تصویر را بهبود داده است.

کلید واژه ها

توصیف تصویر، رمزگذار-رمزگشا سازوکار توجه، سازوکار توجه بر توجه، پردازش زبان‌ها طبیعی

۱- مقدمه

توصیف تصویر، پیوندی بین پردازش زبان طبیعی و بینایی ماشین ایجاد می‌کند. سیستم های توصیف تصویر را می‌توان برای کارهای مختلف استفاده کرد؛ به عنوان مثال، این سیستم ها می توانند افراد نابینا را قادر به دریافت اطلاعات بصری در مورد محیط اطراف خود کنند. هنگامی که این سیستم ها بر روی تصاویر پزشکی آموزش داده می‌شوند، می توانند برای ارائه اطلاعات مفید به پزشکان و کمک به روش‌های تشخیص استفاده شوند[۱]. سایر کاربردهای توصیف تصویر، شامل جستجو و بازیابی جملات تصویری و همچنین آوردن هوش بصری برای ربات ها است. در تولید خودکار توصیف متنی برای تصاویر در مقایسه با طبقه بندی تصویر، تشخیص اشیا و بسیاری دیگر از وظایف بینایی کامپیوتر، با چالش‌های بیشتری مواجه است. این مسئله به این دلیل است که برای تولید توصیفی که از لحاظ محتوایی غنی و از لحاظ نحوی صحیح باشد، نیاز به ادراککلی و همه جانبه از موجودات و اشیای داخل تصویر و روابط بین آنهاست[۲]. همانطورکه ذکر شد، آموزش یک مدل کارآمد توصیف تصاویر مستلزم درک جامعی از موجودیت های اساسی در تصویر و همچنین روابط آن‌ها بوده و بسیار چالش برانگیز است. در این زمینه موفقیت‌های زیادی در

این مقاله در آبان‌ماه ۱۴۰۱ دریافت، در بهمن‌ماه بازنگری و سپس پذیرفته شد.

آزمایشگاه هوش و بینایی ربات، گروه مهندسی کامپیوتر، دانشگاه بوعلی سینا

رایانامه:khotanlou@basu.ac.ir

نویسنده مسئول : حسن ختن‌لو

dorl.net/dor/20.1001.1.23831197.1402.10.2.6.6

سال‌های اخیر به دست آمده، اما هنوز مشکلات زیادی وجود دارد. یکی از چالش های مهم این است که بیشتر تحقیقات بر روی شبکه های بازیابی تولید جمله تمرکز می‌کنند. آن‌ها تأثیر ویژگی‌های مشتق شده از تصویر را نادیده گرفته‌اند[۳]. چالش دیگر در مورد معیارهای ارزیابی پیشنهادی محققان است، که نمی‌توانند عملکرد مدل‌ها را در مقایسه با انسان به دقت اندازه‌گیری کنند[٤].

توصیه‌برچسب تصویر، با هدف تخصیص مجموعه‌ای برچسب‌های مرتبط برای تصاویراز جمله کارهایی بود که در این زمینه انجام شده است. مدل‌های مبتنی بر چسب بر این تمرکز می‌کند که چقدر برچسب‌های پیشنهادی محتوای تصویر را توصیف می‌کند. از جمله روش‌هایی که در این زمینه انجام شده است، روش‌هایی مبتنی بر خوشه بندی است[۵, ۶]. اما هدف از این مقاله تولید توصیفاتی است که تصویر را شرح دهد.

تولید الگو و پرکردن اسلات [۷-۹] و بازیابی عنوان [۱۰-۱۳] از اولین تکنیک‌هایی هستند که برای تولید توصیف تصویر به صورت خودکار استفاده شده است. سپس، نتایج بهتری با استفاده از شبکه‌های عصبی عمیق در مقایسه با روش‌های مبتنی بر الگو و بازیابی حاصل شده است. روش‌های مبتنی بر شبکه‌های عصبی عمیق[۱٤-۱٦] به دلیل آموزش با تصاویر متعدد، بدون نیاز به طراحی ویژگی‌ها توسط انسان، دقت بالایی برای تصاویر ناشناخته دارند. یکی از رایج ترین معماری‌های عمیق مورد استفاده در تولید توصیف تصویر به صورت خودکار، معماری رمزگذار - رمزگشا[۱] است. مدل رمزگذار - رمزگشا معمولاً برای ارائه توصیف متنی از تصویر استفاده می‌شود. اطلاعات در سطح پیکسل به رمزگشایشبکه عصبی کانولوشنی[۲](CNN) داده می شود تا با متراکم کردن آن را رمزگذاری کند. سپس، رمزگشا برای ترجمه این اطلاعات به زبان های طبیعی استفاده می شود[۲]. مدل‌های دیگر نیز بر پایه معماری رمزگذار — رمزگشا . شبکه‌های بازگشتی برای استخراج ویژگی‌های تصویر و تولید عنوان پیشنهاد شدند[۳, ۱۷].

با توجه به شکاف زیاد بین بینایی و روش‌های پردازش زبان، اکثر روش‌ها مشکل عدم تطبیق معنایی کامل بین تصاویر و توصیف‌های تولید شده را دارند. برای حل این مشکل، روش‌های معنایی معرفی شدند. روش‌هایی مانند تطبیق معنایی هدایت شده با شباهت چند سطحی که می‌توانست شباهت‌های معنایی محلی و سراسری را برای یادگیری همبستگی پنهان بین تصاویر و توصیفات تولید شده ترکیب کند[۱۸]، از روش‌های برخورد این چالش بود.

اخیراً، سازوکار توجه در نقشه ویژگی CNN مورد بررسی قرار گرفته است که می‌تواند یک نقشه فضایی ایجاد کند تا اهمیت یا برجستگی خاصی به مناطق مهم تصویری مرتبط با هر کلمه تولید شده بدهد[۱۹]. بنابراین، مدل‌ها می‌توانند روابط بینویژگی‌های تصویر خاص و کلمات مربوطه در توصیف‌ها را بیاموزند[۱۷].

بیشتر روش‌های توصیف تصویر تلاش می‌کنند که مقادیر معیارهای BLEU رو افزایش دهند که شباهت بین دو جمله را حساب می‌کند. این معیار در واقع به‌عنوان میانگین هندسی حاصل ضرب دقتn-gramدر جریمه مختصر بودن براي جملات کوتاه تعریف شده است. این معیار چون به هم معنی‌ها توجه نمی‌کند، معیار خوبی برای سنجش نیست. معیار METEOR یکی دیگر از معیارهای مورد استفاده است که این معیار به‌عنوان میانگین هارمونیک دقت، recall و تطابق unigram بین جملات می‌شود. علاوه بر این، در این معیار از مترادف‌ها و ریشه کلمات نیز استفاده می‌شود. این معیار ارزیابی چند نقص BLEU را مانند ارزیابی recall و عدم تطابق صریح کلمات، برطرف می‌کند. به همین دلیل معیار METOR نسبت به معیارهای ارزیابی دیگر، معیار بهتری برای سنجش مدل پیشنهادی است.

حال با بیان ویژگی‌های مسئله توصیف متنی تصاویر، چالشها و نمونه هایی از تلاشهای انجام شده در این زمینه، به معرفی روش پیشنهادی خود برای حل این مسئله می‌پردازیم. دستاوردهای این تحقیق به طور کلی به صورت زیر خلاصه می‌شود:

- ارائه مدلی بر مبنای معماری رمزگذار — رمزگشا با تلفیق سازوکارهای توجه بر توجه در جهت بهبود عملکرد مدل.
- استفاده از سازوکار AoA[۳] و Multi-Head Attention برای نشان دادن موثرتر ویژگی‌ها و درک بهتر روابط میان اشیا.
- مقایسه عملکرد با روش‌های موجود بر روی مجموعه‌های داده Flickr8k، MSCOCO با معیارهای ارزیابی BLEU(1-4)، METEOR، ROUGE، CIDEr.

در این بخش برای ایجاد یک فضای روشن از مطالعات انجام شده در جهت حل مسئله توصیف تصاویر مرتبط با روش پیشنهادی، پژوهش‌ها در دو بخش روشهای مبتنی بر معماری رمزگذار-رمزگشا و مبتنی بر سازوکار توجه، مرور و بررسی می‌شوند.

۲- کارهای مرتبط

۲-۱- روش‌های مبتنی بر معماری رمزگذار-رمزگشا

از آنجایی که استفاده از معماری رمزگذار - رمزگشا در ترجمه ماشینی باعث بهبود شده است، به همین دلیل روش‌های توصیف تصویر نیز از این روش استفاده می‌کنند. معماری رمزگذار - رمزگشا به طور معمول دو مرحله دارد: در مرحله اول، مدل از شبکه‌های کانولوشنی (CNN) به منظور استخراج ویژگی‌ها، تشخیص اشیا و ارتباط میان آن‌ها استفاده می‌کند. در مرحله دوم، از یک مدل زبانی استفاده می‌شود که خروجی مرحله اول را به شکل کلمات تولید می‌کند[۲]. ویلیناس و همکاران[۳] مدلی به نام

[۱] Encoder-Decoder
[۲] Convolution neural network(CNN)
[۳] Attention on attention

NIC[۱]پیشنهاد کردند. این روش از CNN برای نمایش ویژگی‌های تصویر و از LSTM[۲] برای تولید توصیف استفاده می‌کند. در این روش CNN از روش جدیدی برای نرمال سازی دسته‌ای بهره می‌برد. و خروجی آخرین لایه پنهان CNNبه عنوان ورودی به رمزگذار داده می‌شود. این LSTM می‌تواند اشیایی را که قبلا با استفاده از متن توضیح داده شده‌اند را ردیابی کند. عملکرد NIC بر اساس حداکثر برآوردهای آموزشی است. از آنجایی که اطلاعات تصویر فقط در ابتدای فرآیند تغذیه می شود، ممکن است با مشکل ناپدید شدن گرادیان مواجه شوند. به این ترتیب نقش کلمات تولید شده در ابتدای جملات در تولید کلمات بعدی به تدریج ضعیف و ضعیف تر می شود. بنابراین، LSTM هنوز با چالش هایی از جمله ایجاد جملات طولانی [۲۰-۲۲]مواجه است، به همین دلیل جیا و همکاران[۱۴] توسعه‌ای از LSTM به نام gLSTM را پیشنهاد کردندکه می‌تواند جملات طولانی تولید کند. در این معماري، اطلاعات معنایی کلی با استفاده از دروازه و حالت سلول به LSTM اضافه می‌شود. وانگ و همکاران[۲۳] یک روش عمیق مبتنی بر LSTM دو طرفه پیشنهاد کردند. که این روش قادر است توصیف‌های غنی تولید کند. روش پیشنهاد شده از یک CNN و دو شبکه مجزا LSTM تشکیل شده است.

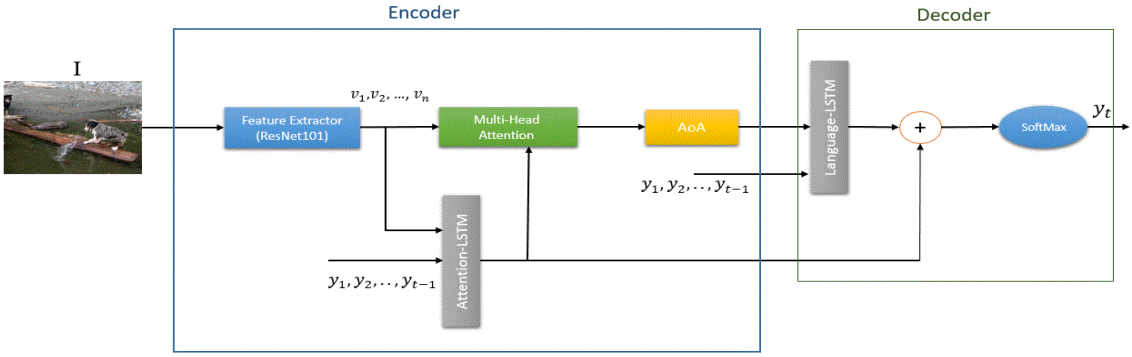
برای بهبود روش‌های مبتنی بر چارچوب رمزگذار-رمزگشا، محققان روش‌هایی مبتنی بر فضای چندوجهی پیشنهاد کردند. معماری فضای چند وجهی[۳] شامل سه بخش رمزگذار زبان و بینایی، فضای چند وجهی و رمزگشای زبان است. اولین کار در این زمینه توسط کیروس[۱۶] در سال ۲۰۱۴ انجام شده است. این روش از CNN برای استخراج ویژگی‌های تصویر و همچنین از فضای چند وجهی که تصویر و متن را نشان می‌دهد، استفاده می‌کند. چن در سال ۲۰۱۵ [۲۴] مدل توصیف تصویر را بر اساس فضای چند وجهی پیشنهاد کرد. این روش می‌توانست توصیف‌های جدیدی برای تصویر ایجاد کند، و ویژگی‌های بصری را از توصیف‌های داده شده بازیابی و همچنین نمایش بصری تصاویر را از کلمات تولید شده به صورت پویا به روزرسانی کند. این مدل برای کار با داده‌های زیاد محدودیت‌های فراوانی دارد و درکار با حافظه بلند مدت ناکارآمد است. به دلیلی وجود چالش‌های توضیح داده شده، در مرحله بعد، سازوکار توجه در این چهارچوب به کار گرفته شد که تولید توصیف برای تصویر را بهبود بخشید.

۲-۲- روش‌های مبتنی بر سازوکار توجه

در معماری رمزگذار-رمزگشا بدون استفاده از سازوکار توجه[۴]، یک CNN به عنوان رمزگذار برای استخراج ویژگی‌های بصری از تصویر و RNN[۵] برای تبدیل ویژگی‌های استخراج شده به کلمات استفاده شد. این روش نمی‌تواند تصویر را در طول زمان تجزیه و تحلیل کند و همچنین نمی‌تواند جنبه‌های فضایی تصویر را که مربوط به قسمت‌های مهم تصویر است را در نظر بگیرد. در روش رمزگذار - رمزگشا، توصیف با استفاده از کل صحنه تولید می‌شد. سازوکار توجه می‌تواند به صورت پویا روی عناصر مختلف یا مناطق تصویر متمرکز شود. خو وهمکاران[۲۵] در سال ۲۰۱۵، اولین روش مبتنی بر سازوکار توجه برای تولید توصیف تصویر را معرفی کردند. تفاوت اصلی روش‌های مبتنی بر توجه با روش‌های دیگر این است که می‌تواند روی بخش‌های برجسته داده‌ها تمرکز کند و همزمان کلمات مرتبط را تولید کند. این روش از دو تکنیک مختلف استفاده می‌کند: توجه سخت تصادفی[۶] و توجه نرم قطعی. اکثر رویکرهای مبتنی بر CNN از لایه‌ی بالایی ConvNet برای استخراج اطلاعات برجسته شی از تصویر استفاده می‌کنند. پادرسولی و همکاران[۲۶] سازوکارهای توجه مبتنی بر منطقه را اعمال کردند. در این روش روابط بین کلمات تولید شده و نواحی تصویر مدلسازی شد. دنگ و همکاران [۲۷] در سال ۲۰۲۰ روشی را پیشنهاد کردند که از مدل DenseNet برای استخراج ویژگی‌های کلی تصویر در مرحله رمزگذاری استفاده می‌کند. و در همان زمان، سازوکار توجه از یک دروازه نگهبان[۷] استفاده می‌کند تا تصمیم بگیرد که از کدام اطلاعات ویژگی تصویر برای تولید کلمه استفاده شود. آنها از LSTM برای بخش رمزگشایی مدل استفاده می‌کنند. این روش توانست توصیفات را تا حدودی بهبود بخشد. کائو و همکاران[۲۸] در سال ۲۰۲۰ مدل IGGAN را برای یادگیری بدون نظارت پیشنهاد کردند، که نمایش خصوصیات چند مقیاسی و روابط شی را ترکیب می‌کند. برای داشتن یک نمایش قدرتمند، تصاویر توسط ResNet با یک ماژول چند مقیاسی جدید و کانال RMCNET رمزگذاری می‌شود. روابط شی نیز توسط ماژول MAN استخراج می‌شود. شبکه IGGAN ترکیبی از شبکه‌های مولد و سازوکار توجه است.دپنگ وانگ و همکاران[۲۹] روشی مبتنی بر معماری رمزگذار — رمزگشا را طراحی کردند که شامل یک رابط هدایت کننده متن[۸]بود تا نمایش بصری را یاد بگیرد که با شناخت بصری انسان سازگار تر باشد. علاوه بر این بخش رمزگشا را به دو ماژول تقسیم می‌کند: یک تولید کننده برای تولید جمله اولیه و یک پالایش کننده برای اصلاح

[۱] Generator Neural Image Caption
[۲] Long short-term memory
[۳] Multi Modal Space
[۴] Attention Mechanism
[۵] شبکه‌های بازگشتی
[۶] Random hard attention
[۷] Sentinel gate
[۸] Text-guided relation encoder(TGER)

جمله.ماژول تولید کننده متشکل از یک LSTMاستاندارد، و یک ماژول(Gate onAttention (GOA است که سعی دارد جمله اصلی را به صورت منطقی و روان تولید کند. پالایش‌کنندهشاملیک ماژول رمزگذار توصیف، یک LSTM



شکل (۱) : ساختار کلی روش پیشنهادی

مبتنی بر توجه و یک ماژولGOAاست، که به طور مکرر جزئیات را در عنوان اصلی تغییر می‌دهد تا زیرنویس‌ها را با کمک متن راهنما غنی و دقیق تولید کند. لان هانگ و همکاران [۳۰] روشی را مبتنی بر روش توجه بر توجه(AoA) معرفی کردند که توانست توصیفات خوبی را تولید کند. در این روش AoA به قسمت رمزگذار و رمزگشای مدل افزوده شد. AoA یک گسترش کلی برای سازوکار توجه است و می‌تواند به هرکدام از آن‌ها اعمال شود.در قسمت رمزگذار، AoA به مدل سازی بهتر روابط بین اشیا مختلف در تصویر کمک می‌کند. در رمزگشا نیز AoA نتایج توجه غیر مرتبط را فیلتر می‌کند و فقط موارد مفید را نگه می‌دارد، همچنین ما روشی[۲]را پیشنهاد کرده بودیم که از یک لایه توجه محلی[۱] در چارچوب رمزگذار — رمزگشا استفاده شد، که این روش باعث شد توصیف‌های خوبی را مدل پیشنهادی تولید کند و معیارهای ارزیابی را بهبود دهد و پژوهش حاضر توسعه یافته این روش می باشد.

۳- روش پیشنهادی

در این قسمت ما روش پیشنهادی خود را معرفی کرده و جزئیات آن را بررسی می‌کنیم. ساختار کلی روش پیشنهادی در شکل (۱) مشاهده می‌شود. این ساختار از چارچوب رمزگذار - رمزگشا پیروی می‌کند و از چندین بخش تشکیل شده است. قسمت رمزگذار تصویر شامل چند ماژول استخراج کننده ویژگی (ResNet101)، Multi-Head Attention، Language-LSTM، Attention-On-Attention است. قسمت رمزگشای تصویر شامل ماژول Language-LSTM است.

همانطور که در شکل یک نشان داده شده، تصویر I به عنوان ورودی به مدل دادهمی‌شود. مدل توصیف تصویر باید جملاتی را تولید کند که از لحاظ معنایی و محتوایی صحیح باشد و می‌تواند با نمایش $S = y_1, y_2, \ldots, y_L$. که در آن y_i کلمه iام جمله S است، نشان داده شوند. همچنینLدر نمایش جملات، طول توضیحات تولید شده را نشان می‌دهد. همانطور که در معادله (۱) مشاهده می‌کنید، I به عنوان ورودی تصویر به مدل داده می‌شود و توضیحات Sدر خروجی ایجاد می‌شود.

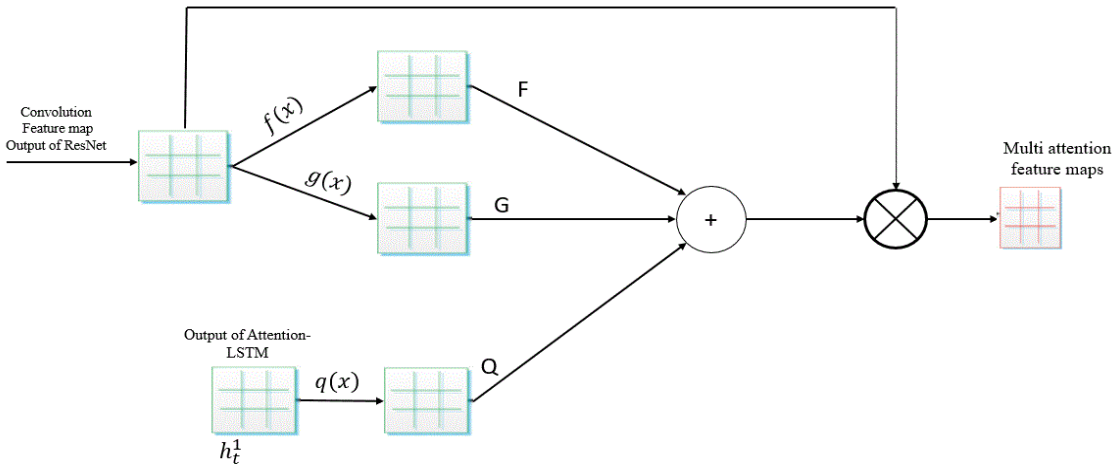
$$F : I -> S \qquad (۱)$$

روش پیشنهادی ما بر اساس پژوهش وی و همکاران[۳۱]، لان هانگ و همکاران[۳۰] و مطالعه پیشینمان[۲] است. ما در روش پیشین خود نیز از معماری رمزگذار و رمزگشا استفاده کرده ایم که شامل چندین ماژول استخراج ویژگی، Attention-LSTM، Attention-Layer، Language-LSTM است. در این معماری لغات از پیش تولید شده و ویژگی‌های استخراج شده به لایه Attention-LSTM داده می‌شود که این لایه شامل دو لایه LSTM است و سپس به یک لایه ساده توجه داده می‌شود، و سپس خروجی مدل به لایه Language-LSTMمی‌رود که شامل یک لایه LSTM است و در نهایت کلمه تولید میشود.

معماری مطالعه پیشین در روش پیشنهادی تغییر کرده است و ضمن کاهش پیچیدگی محاسباتی نسبت به کار وی و همکاران، نتایج نیز بهبود یافتند. ما در روش پیشنهادی خود بر خلاف مطالعه پایه [۲۸] از شبکه‌های مولد استفاده نکرده‌ایم. ساختار Multi-HeadAttention، Attention-LSTM نسبت به کار آنان تغییر کرده است، اما ایده ماژول AoA از مطالعه پایه گرفته شده و به مدل پیشنهادی اضافه شده است. در واقع ما در این روش، مدل قبلی پیشنهادی خود را با عوض کردن ساختار لایه‌ی سازوکار توجه و با افزودن لایه AoA بهبود داده‌ایم، ما از روش تخمین حداکثر احتمال[۲] (MLE) برای آموزش مدل استفاده کرده‌ایم. در این روش در زمان T یک توصیف واقعی تا کلمه $T-1$ به مدل داده می‌شود تا احتمال تولید y_t به عنوان عنصر واقعی جمله مدل به حداکثر برسد. برای بهینه سازی مدل از تابع تلفاتcross-entropy استفاده می‌کنیم که در معادله ۲ نمایش داده شده است.

[۱] Local attention mechanism

[۲] Maximum likelihood estimation



شکل (۲) : ساختار توجه چند وجهی پیشنهادی

$$L_{CE}(\theta) = -\sum_{t=1}^{T} \log(P_\theta(y_t^* | v, y_1^*, ..., y_{t-1}^*) \quad (۲)$$

پارامتر θ وزن های قابل یادگیری مدل را نشان می دهد. این مدل توزیعی از وجود هر کلمه را به شرط دنباله ای از پیش موجود از کلمات و ویژگی های بصری ایجاد می کند؛ زیرا هیچ توصیف واقعی از فرم موجود وجود ندارد.

برای شرح عملکرد مدل، مطابق شکل (۱)، ایجاد یک عنوان برای تصویر که از نظر نحوی و معنایی صحیح باشد را مطابق مراحل زیر دنبال می کنیم:

۳-۱- رمزگذار

قسمت‌های رمزگذار تصویر در مدل پیشنهادی شامل: استخراج کننده ویژگی‌های تصویر، Attention-LSTM، Multi-Head Attention، Attention On Attention است، که در ادامه توضیح خواهیم داد.

۳-۱-۱- استخراج کننده ویژگی

ویژگی‌های کلی تصویر، یک تصویر را به شکل جامع توصیف می‌کند. ویژگی‌های کلی تصویر محبوب هستند، زیرا نمایش بسیار فشرده‌ای از تصاویر ایجاد می‌کنند این مدل در مجموعه داده ImageNet از قبل آموزش داده شده است. ویژگی های تصویر کلی $v=v_1, v_2, v_3, ..., v_k$ جزئیات بیشتری را نسبت به تصویر ورودی پوشش می دهد. v_i یک بردار d-بعدی است که ویژگی های یک ناحیه از تصویر را نشان می دهد و k تعداد مناطق تصویر است.

$$V = CNN_{CONV}(I) \quad (۳)$$

در فرمول (۳) I تصویر ورودی است و CNN_{CONV} آخرین لایه کانولوشن ResNet101 است و در نهایت V ویژگی‌های کلی تصویر است که استخراج می‌شود.

۳-۱-۲- Attention-LSTM

این ماژول به مفاهیم مرتبط به زبان و اطلاعات متنی از پیش تولید شده توجه می‌کند. سپس به توزیع توجه در تمام قسمت‌های تصویر ادامه می‌دهد. همانطورکه در شکل (۱) نشان داده شده است، پس از ادغام ویژگی‌های سراسریV، با بردار کلمه از پیش تولید شدهy_t، $W_e{}_1$، به این لایه داده می‌شود و در ادامه چنان‌که در معادله (٤) مشاهده می‌کنید، خروجی به لایه LSTM داده می‌شود:

$$h_t^1 = LSTM(LSTM([v \cdot W_e y_{t-1}])) \quad (٤)$$

که در معادله (٤) $V = \{v_1, v_2, ...\}$، نشان دهنده ویژگی های کلی استخراج شده از تصویر است که در قسمت قبل استخراج شده است. و W بردار کلمات تعبیه شده[1] از پیش تولید شده است.

۳-۱-۳- Multi-Head Attention

این ماژول شامل یک سازوکار چند توجهی با استفاده از شواهد محلی و غیر محلی برای بازنمایی و استدلال موثرتر ویژگی در توصیف تصویر پیشنهاد شده است. همچنین در این معماری، توجه یک لایه متراکم است که ویژگی‌های سطح پایین‌تر یک منطقه محلی را دستکاری می‌کند تا یک نقشه فضایی ایجاد کند که مناطق تصویر را بر اساس اطلاعات بازخورد ارائه شده توسط Attention-LSTM برجسته می‌کند. همانطورکه در شکل (۱) مشاهده می‌کنید، خروجی لایه‌یFeature-Extractor(V)و خروجی لایه‌ی Attention-LSTM به عنوان ورودی به این لایه داده می‌شود. در شکل (۲) جزئیات این لایه را به صورت دقیق نمایش داده‌ایم. در ابتدا ویژگی‌های استخراج شده V به دو فضای F، G منتقل می‌شوند. سپس خروجی لایه ( ) Attention-LSTM به فضای جدید Q منتقل می‌شود. نمایش ریاضی این مراحل در معادلات ۵ تا ۷ نشان داده شده است.

$$f(v) = W_f V \quad (۵)$$

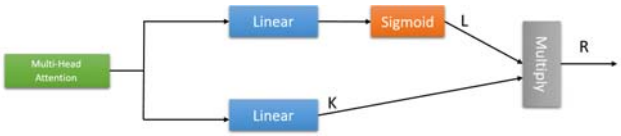
$$g(v) = W_g V \quad (٦)$$

$$q(h_t^1) = W_q h_t^1 \quad (۷)$$

و سپس جمع tanh را به صورت معادله ۸ محاسبه می‌کنیم:

[1] Embedding words

$$a_{i,t} = \tanh(f + g + q) \quad (۸)$$



شکل (۳) : ساختار پیشنهادی توجه بر توجه

$a_{i,t}$ وزن‌های استخراج شده از لایه Multi Head Attention است.

و در مرحله آخر در این لایه وزن بدست آمده $a_{i,t}$ را ضرب در ویژگی‌های استخراج شده V از ResNet می‌کنیم.

$$c_t = \sum_{k=1}^{k} a_{k,t} v_k \quad (۹)$$

۳-۱-۴- Attention on Attention[1]

AoA یک گسترش کلی برای سازوکار توجه است. و می‌تواند به هرکدام از آن‌ها اعمال شود. در قسمت رمزگذار، AoA به مدل سازی بهتر روابط بین اشیا مختلف در تصویر کمک می‌کند. همانطورکه در شکل (۳) مشاهده می‌کنید، خروجی لایه Multi-Head Attention با نام C به‌عنوان ورودی به این لایه داده می‌شود. که از لایه خطی عبور داده می‌شود، در ابتدا C از لایه خطی W_1عبور داده میشود و سپس sigmoid آن محاسبه می‌شود و همینطور C از لایه خطی W_kعبور داده می‌شود و در نهایت خروجی آن‌ها در هم ضرب می‌شود. معادلات ۱۰ تا ۱۲ توصیف مراحل گفته شده است:

$$l(c) = Sigmoid(W_l C) \quad (۱۰)$$

$$k(c) = W_k C \quad (۱۱)$$

$$R = l * k \quad (۱۲)$$

۳-۲- رمزگشا

۳-۲-۱- Language-LSTM

همانطورکه در شکل (۱) مشاهده می‌کنید، ورودی این ماژول شامل خروجی تولید شده توسط ماژول AoA و کلمات از پیش تولید شده تعبیه شده میباشد، که به شرح معادله (۱۳) است:

$$h_t^2 = LSTM([r_t, W_e y_{t-1}]) \quad (۱۳)$$

در نهایت، مجموع Attention-LSTM و Language-LSTM به یک تابع softmax که در معادله ۱۴ آمده، داده می شود تا توزیع احتمال کلمه را بدست آوریم:

$$p(y_t \mid y_{1:t-1}) = soft\max(h_t^1 + h_t^2) \quad (۱۴)$$

۴- ارزیابی

۴-۱- مجموعه داده

Flickr8k[۱۳]مجموعه داده‌ی گسترده‌ای برای توصیف تصاویر است.Flickr8k یک پایگاه داده از ۸۰۰۰ تصویر است که هر کدام دارای پنج عنوان به زبان طبیعی است که بر اساس سرویس مجموعه

شکل (٤) : نمونه‌ای از مجموعه‌داده

منابع آمازون مکانیکال ترک تولید شده‌اند. تصاویر، بیشتر در مورد انسان و حیوانات بوده و هر عنوان یا توصیف بخشی از موجودات و رویدادهای برجسته تصویر مربوطه را ارائه می‌دهد. شکل(٤) نمونه‌ای از مجموعه‌داده Flickr8k را نشان می‌دهد. مجموعه داده [۳۲] نیز مجموعه داده جامعی با ۱۲۳۲۸۷ تصویر است که در آن برای هر تصویر پنج توصیف ارائه شده است.

۴-۲- معیار ارزیابی

در حال حاضر معیارهای ارزیابی تولید توصیف تصویر همان معیارهای ارزیابی ترجمه ماشینی است. زیرا روشهای ارزیابی انسانی منسوخ شده و توصیف تصویر معیارهای ارزیابی جداگانه‌ای ندارد. مشکل معیارهای ارزیابی موجود این است که شباهت و عدم شباهت بین توصیف تولید شده و توصیف‌های مرجع را نشان می‌دهد و فقط به جنبه زبانی توجه می‌کنند، اما در حال حاضر تنها روش ارزیابی همین معیارهاست. معیارهای ارزیابی خودکار که به‌صورت گسترده استفاده می‌شود شامل: BLEU, METEOR, CIDER, ROUGE است. در این معیارها، نمرات بالاتر، جملات بهتر را نشان می‌دهد.

معیار BLEU[۳۳]، یکی از اولین معیارهای ارزیابی است که برای اندازه‌گیری شباهت بین دو جمله مورداستفاده قرار گرفته است. این معیار در ابتدا برای ترجمه ماشینی پیشنهاد شده و به‌عنوان میانگین هندسی، حاصضرب دقت n-gramدر جریمه مختصر بودن برای جملات کوتاه تعریف شده است. نحوه محاسبه دقت n-gram در فرمول(۱٤) ذکر شده است.

$$p_n = \frac{\sum_{C\in\{\text{Candidates}\}} \sum_{n-grams\in C} Count_{clip}(n-gram)}{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count(n-gram)} \quad (۱۵)$$

[1] AoA

که در اینجا $Count_{clip}$حداکثر تعداد n-gramاست که در جمله کاندید و مرجع هم زمان اتفاق می‌افتد، Count تعداد n-gramهایی است که در جمله کاندید است. برای جلوگیری از ترجمه‌های بسیار کوتاه جریمه مختصر بودن را نیز اضافه کرده‌اند.

$$BP = \begin{cases} 1 & if\ |c| > |r| \\ e^{(1-|r|/|c|)} & if\ |c| \le |r| \end{cases} \quad (۱۶)$$

در معادله ۱۶، |c| طول جمله کاندید و |r| طول جمله مرجع است. و در نهایت BLEU به صورت فرمول(۱۷) محاسبه می‌شود کهw_nدر آن ضریب وزنی است.

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} W_n \log P_n) \quad (۱۶)$$

METEOR[۳۴]معیار دیگری است که به عنوان میانگین هارمونیک دقت، recall و تطابق یک-گرام بین جملات تعریف می‌شود. همچنین در این معیار از مترادف و ریشه کلمات استفاده می‌کند. این متریک چندین نقص BLEU، مانند عدم تطابق کلمات صریح را برطرف می‌کند. آنها همچنین قادر به شناسایی شباهت معنایی بودند که توسط تطبیق مترادف مبتنی بر WordNet کنترل می شود. تعداد unigram نگاشت شده بین دو رشته m نام دارد و تعداد unigram در جمله کاندید t و همچنین تعداد unigram مرجع r نام‌گذاری شده است. دقت unigram برابر است با P=m/t و همچنین recall آن برابر است با R=m/t، سپس میانگین هارمونیک P و Rمطابق معادله ۱۷ محاسبه می‌شود.

$$F_{mean} = \frac{P \bullet R}{\alpha \bullet P(1-\alpha) \bullet R} \quad (۱۷)$$

همچین از تطابق n-gram طولانی تر برای محاسبه استفاده می‌شود. و در نهایت معیار METEOR طبق فرمول (۱۸) محاسبه می‌شود همچنین جریمه P برای تراز استفاده می‌شود.

$$M = F_{mean}(1-p) \quad (۱۸)$$

هدف معیار ارزیابی CIDER[۳۵] محاسبه تناظر بین جمله نامزد c_i و مجموعه توصیفات مرجع تصویر به صورت $S_i = S_{i1},\ldots,$ S_{im}است. برای محاسبه این معیار، همه کلمات در جملات مرجع و نامزد به ریشه نگاشت می شوند. سپس همزمانی رخداد n-gram در جملات مرجع و نامزد محاسبه می شود. در CIDER، مشابه-tf n-gram،idf که در تمام توضیحات تصویر مشترک هستند، سبک وزن هستند. در نهایت، شباهت کسینوس بین n گرم جملات کاندید و مرجع محاسبه می شود. این معیار گاهی به جزئیات بی اهمیت یک جمله وزن بیشتری می بخشد و در نتیجه ارزیابی عنوان بی اثر می‌شود[۳۶].

$$CIDEr(c_i, s_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, c_n) \quad (۱۸)$$

که $CIDEr_n$برای n-gram به طول n شباهت کسینوسی بین جمله مرجع و کاندید است.

معیار ارزیابی ROUGE[۳۷]در ابتدا برای ارزیابی سیستمهای خلاصه سازی پیشنهاد شده است. و این ارزیابی از طریق مقایسه n-gramهای همپوشانی، دنباله‌های کلمه و جفت کلمات انجام می‌شود. در این کار ما از L-Rouge استفاده می‌کنیم که طولانی‌ترین دنباله‌های مشترک بین جفت جمله را اندازه گیری می‌کند[۳۶].

۴-۳- آماده سازی داده

به دلیل ناکافی بودن داده‌ها در مجموعه داده Flickr8k، ما به تقویت داده[۱] در این مجموعه نیاز داشتیم. از این رو با روشهای مختلف، ۲۰۰۰ داده به مجموعه‌داده Flickr8kاضافه کردیم. ما داده‌ها را به روش‌های مختلفی تقویت کردیم. از جمله: افزودن نویز گاوسی[۲] به تصاویر، چرخاندن تصاویر از چپ به راست[۳]، افزودن روشنایی[۴] به تصاویر و آینه کردن تصاویر.

۴-۴- جزئیات پیاده سازی

برای پیاده سازی مدل پیشنهادی، از زبان برنامه نویسی پایتون نسخه ۳,۹ در چارچوب کتابخانه Kerasاستفاده کردیم. در این تحقیق از روشی برای تفکیک داده های ورودی به داده های آموزشی و داده های آزمون استفاده شد. که ۸۰ درصد داده های ورودی برای آموزش مدل و ۲۰ درصد برای آزمون مدل تقسیم شدند. همچنین یک کامپیوتر با یک پردازنده i7 -۶۷۰۰ Coreشرکت اینتل با ۲۴ گیگابایت حافظه رم و یک پردازنده گرافیکی GTX NVIDIA ۳۰۷۰ مورداستفاده قرار گرفت. در روش پیشنهادی، تعداد حالت‌های پنهان LSTMو تعبیه ۱۲۸ است. همچنین از مدل از پیش آموزش دیده شده ResNet-101در شبکه تصویر استفاده کردیم. تعداد پارامترهای مدل ما برابر با ۲۴۲۴۶۱۴۶ است. ما مدل خود را با بهینه ساز Adamآموزش داده‌ایم. مدل ما بر مجموعه‌داده Flickr8kدر ۱۰۰ دوره آموزش دید و همچنین این آموزش حدود ۹۸ ساعت طول کشید. همچنین آموزش مدل بر مجموعه داده MSCOCOدر ۱۱۰ دوره آموزش حدود ۱۸۰ ساعت طول کشید.برای محاسبه‌ی معیارهای ارزیابی از کتابخانه‌ی استانداردی pycocoevalcapاستفاده کرده‌ایم.

۴-۵- نتایج تجربی

ما در این روش سعی بر تولید توصیفاتی داشتیم که از لحاظ محتوایی غنی و از لحاظ نحوی صحیح باشد. همانطور که قبلا ذکر شد، یکی از بزرگترین مشکلات موجود در توصیف خودکار تصویر عدم وجود معیار ارزیابی مناسب است. معیارهای ارزیابی موجود و همچنین روش‌های پیشین سعی بر این دارند که توصیفی تولید کنند که بسیار شبیه به جمله مرجع باشد. اما ما در این روش سعی داشتیم روشی را پیشنهاد کنیم، که طی آن جملاتی تولید شوند که ضمن اینکه دقیقا جمله مرجع نباشند اما به خوبی تصویر

[۱] Data augmentation
[۲] Gussian augmentation
[۳] Filip augmentation
[۴] Brightness augmentation

را بیان و توصیف کند و از لحاظ نحوی صحیح باشد. به همین دلیل معیارهای ارزیابی موجود معیار مناسبی برای سنجش کار نبود. به دلیل اینکه معیار مناسب دیگری وجود نداشت معیارهای METEOR و ROUGE که نسبتا معیارهای مناسب تری برای سنجش هستند برای ارزیابی انتخاب شدند. به خصوص معیار METEOR که از ریشه کلمات و هم معنی آن برای مقایسه استفاده می‌کند معیار مناسبی برای سنجش عملکرد مدل ما بود. در شکل (۵) نمونه‌ای از توصیف تولید شده بر مجموعه داده Flickr8k را مشاهده می‌کنید. همانطور که مشاهده می‌شود، این توصیف توانسته است، به خوبی تصویر را توصیف کند و همچنین از لحاظ نحوی صحیح است.

A man in a wetsuit rides his bicycle on a beach

شکل (۵) : نمونه‌ای از توصیف تولید شده توسط مدل پیشنهادی

٤-٦- تجزیه و تحلیل کمی

در جدول (۱) مقایسه کار انجام شده، با روش قبلی پیشنهادی که روش پایه ما هست را نشان می‌دهد.در روش پیشین خود نیز از معماری رمزگذار و رمزگشا استفاده کرده ایم که شامل چندین ماژول استخراج کننده ویژگی، Attention-LSTM، Attention-Layer، Language-LSTM است. در روش پیشین [۲] نیز مدل پیشنهادی قبلی ما توصیفاتی تولید می‌کرد که از لحاظ نحوی و معنایی صحیح بود و توانسته بود در معیارهای ارزیابی مقادیر خوبی کسب کند، ما با بهبود لایه‌ی توجه و افزودن لایه‌ی توجه بر توجه به این مدل، همانطور که در جدول (۱) مشاهده می‌کنید، توانستیم معیارهای ارزیابی را بهبود دهیم.

در شکل (٦) نیز مقایسه توصیفات تولید شده توسط دو مدل را مشاهده می‌کنید. همانطور که می‌بینید، مدل جدید پیشنهادی ما توصیفات بهتری را تولید کرده است. اما با وجود اینکه توصیفات خیلی بهتر شده‌اند از دید معیارهای ارزیابی استفاده شده، تغییر چشم گیری نداده است. این مسئله تایید دیگری بر عدم ارزیابی کامل و دقیق توسط معیارهای موجود است. و همانطور که مشاهده می‌کنید، هر سه توصیفی که در شکل(٦) تولید شده است ، از لحاظ معنایی و نحوی صحیح است، و به خوبی تصویر را توصیف می‌کند. همانطور که مشاهده می‌کنید این جملات دقیق شبیه به جمله مرجع نیست، به همین دلیل باعث شده است معیارهای BLEU که هدف از آن محاسبه‌ی شباهت n-gram است، کاهش بیابد. به طور مثال همانطور که در تصویر مشاهده می‌کنید در جمله مرجع کلمه woods وجود ندارد، اما روش ما به دلیل استفاده از مکانیزم توجه توانسته است woods را در تصویر تشخیص دهد. و در تصویر دوم در شکل(٦)، کلمه is running به جای کلمه jumps تشخیص داده شده است، این باعث می‌شود معیار BLEU کاهش پیدا کند در حالیکه جملات تولید شده از لحاظ معنایی و نحوی صحیح است.

بنابراین همانطور که مشاهده می‌کنید در جداول مقایسه روش پیشنهادی با روش‌های پیشین در معیارهای BLUEافزایش نداشته‌ایم و در معیار METEORو ROUGE بهبود خوبی داشته‌ایم.

در جدول (۲)، مقایسه بین مطالعات موفق انجام شده و روش ما روی مجموعه داده Flickr8k نشان داده شده است. مدل پیشنهادی ما توانسته‌است مقادیر خوبی را در معیارهای ارزیابی بدست آورد. و همچنین در معیار METEOR که قبلا اهمیت آن شرح داده شد، پیشرفت خوبی اتفاق افتاده است. همانطور که گفته شد، یکی از بزرگترین مشکلات در زمینه تولید توصیف تصاویر، خوب نبودن معیارهای ارزیابی است، بنابراین علاوه بر مشاهده نتایج معیارهای ارزیابی با تجزیه و تحلیل توضیحات تولید شده، می‌توان فهمید که توضیحات خوبی توسط مدل پیشنهادی ما تولید شده است. در جدول(۳) نیز مقایسه روش پیشنهادی با روش‌های پیشین بر مجموعه داده MSCOCO انجام شده است، همانطور که در جدول مشاهده می‌کنید در این مجموعه‌داده نیز مدل پیشنهادی توانسته است مقادیر خوبی را در معیارهای ارزیابی بدست آورد و همچنین معیار ROUGEنیز پیشرفت خوبی داشته است.

شکل(۷) نمونه ای از درک صحیح تصویر و اجزای آن توسط مدل ما است. در شکل(۸) مقایسه توصیف تولید شده توسط روش پیشنهادی و توسط یکی از روش‌های پیشین انجام شده است. هر دو مدل توانسته‌اند جملات خوبی را تولید کند، اما مدل پیشنهادی MAGANسعی بر این داشته که توصیفی تولید کند که بیشتر شبیه به توصیق مرجع باشد، اما روش ما توصیفی تولید کرده است که هم از لحاظ محتوایی صحیح هست و هم دقیقا جمله مرجع نیست.به این ترتیب در روش ما ضمن تولید خروجی مناسب برای این تصویر از دایره واژگان گسترده‌تری استفاده شده که مطلوبیت و کیفیت مناسبی را از دید ناظر انسانی به خروجی می‌دهد.

جدول (۱) : مقایسه روش پیشنهادی جدید با روش پیشنهادی قبلی

روش	Flickr8k						
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDER
روش پیشین پیشنهادی[۲]	٦۰,٤	٤۱,۳	۲۷,۷	۱۸,۷	۲٤,٤	٥۱,٦	٤۱,٥
روش پیشنهادی	٦۱,٦	٤۲,۳	۲۷,۸	۱۸,۹	۲٥,٤	٥۲,٦	٤۲,٥

New-Method: a man is riding a bike on a trail through the woods.
Previous-Method: a person is riding a bike on a dirt bike.
Ground-Truth: a person riding a bike doing a very high jump.
Ground -Truth: a person is in the air with a trick bike

New-Method: a black dog is running through a field of snow.
Previous-Method: a black dog is running through the now.
Ground -Truth: a black dog jumps in the snow.
Ground -Truth: a black dog bounds through a path in the snow.

New-Method: a basketball player dribbles the ball.
Previous-Method: a basketball player in a white uniform is playing with a ball in the background.
Ground -Truth: a basketball player dribbles the ball.
Ground -Truth: a basketball player in the white strip is dribbling the ball on the court.

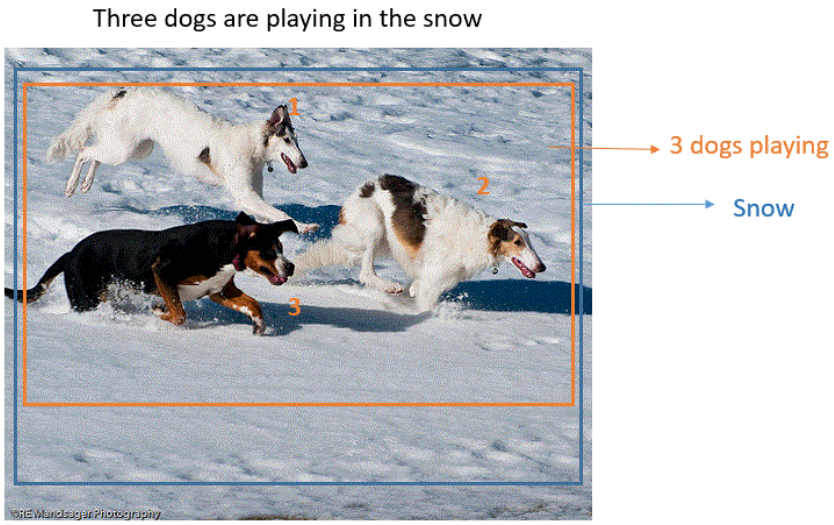
شکل (٦) : مقایسه توصیفات تولید شده

جدول (۲) : مقایسه روش پیشنهادی جدید با روش‌های قبلی بر روی مجموعه داده Flickr8k

روش	Flickr8k						
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDER
خو و همکاران[۲٥]	٦۷	٤٥,۷	۳۱,٤	۲۱,۳	۲۰,۳	-	-
فو و همکاران[۳۸]	٦۳,۹	٤٥,۹	۳۱,۹	۲۱,۷	۲۰,٤	٤۷	٥۳,۸
لی و همکارن[۳۹]	٥۷,۲	۳۷,۹	۲۳,۹	۱٤,۸	۱٦,٦	٤۱,۹	۳٦,۲
زو و همکاران[٤۰]	٦۳,۲	٤٤,۸	۳۱,۳	۲۱,٥	۲۰,٤	-	-
ونگ و همکاران[۲۳]	٦۸,۹	٤۷,٤	۳۳,٤	۲۳,۸	۲۱,٤	٤۸,۸	٥۹,٤
چنگ و همکاران[٤۱]	٦۳,۹	٤٦,۲	۳۲,٥	۲۲,۲	-	-	-
دینگ و همکاران[٤۲]	٦۷,۲	٤٥,۱	۳۰,٥	۲۱,٥	-	-	-
شیائو و همکاران[٤۳]	٦٤,۸	٤٦,۷	۳۲,۳	۲۱,۸	۲۰,٥	-	-
روش پیشین پیشنهادی[۲]	٦۰,٤	٤۱,۳	۲۷,۷	۱۸,۷	۲٤,٤	٥۱,٦	٤۱,٥
روش ما	٦۱,٦	٤۲,۳	۲۷,۸	۱۸,۹	**۲٥,٤**	**٥۲,٦**	٤۲,٥

جدول (۳) : مقایسه روش پیشنهادی جدید با روش‌های قبلی بر روی مجموعه داده MSCOCO

روش	MSCOCO						
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDER
یو و همکاران[٤٤]	٧٣,١	٥٦,٥	٤٢,٤	٣١,٦	٢٥,٠	٥٣,٣	٩٤,٣
گو و همکاران[٤٥]	٦٧,١	٤٧,٨	٣٢,٣	٢١,٥	٢٠,٩	٤٧,٢	٦٩,٥
شانکو و همکاران[٢٨]	–	–	–	٢١,٩	٢١,١	٤٧,٢	٦٤
لیان و همکاران[٤٦]	٦١,٧	٤٢,٨	٢٨,٦	١٩,٣	٢٠,٢	٤٥,٠	٦١,٨
فنگ و همکاران[٤٧]	٥٨,٩	٤٠,٣	٢٧,٠	١٨,٦	١٧,٩	٤٣,١	٥٤,٩
لو وهمکاران[١٩]	٧٤,٦	٥٨,٢	٤٤,٣	٣٣,٥	٢٦,٤	٥٥,٠	١٠٦,١
چن و همکاران[٤٨]	٧٢,٥	٥٥,٦	٤١,٤	٣٠,٦	٢٤,٦	٥٢,٨	٩١,١
روش ما	٧٠,٦	٤٦,١	٢٩,١	٢١,٢	**٢٨,٤**	**٥٤,٦**	٤٩,٥



شکل (۷) : توصیف تولید شده توسط مدل پیشنهادی

Ground-Truth: Dog looking out the window of a car in review mirror

MAGAN[31]: A dog looking out a car window in mirror

Our-Method: a reflection of person holding a dog in the back of the car

شکل (۸) : مقایسه توصیف تولید شده با یکی از روش های پیشین

۵- نتیجه‌گیری

با توجه به مزایای تولید توصیفات متنی تصویر توسط هوش مصنوعی، روش‌های بسیاری تا به امروز ارائه شده است. ما در این روش تلاش کردیم توصیفات تولید شده را بهبود بخشیم. همچنین مدل ما از منظرمعیارهای METEOR و ROUGEدارای بهبود در عملکرد بود.

در این مقاله ما روشی مبتنی بر رمزگذار - رمزگشا پیشنهاد دادیم. و همچنین از لایه توجه چندگانه و توجه بر توجه استفاده کردیم. لایه توجه چندگانه باعث می‌شود که مدل به قسمت‌های مهم تصویر بیشتر توجه کند و همچنین نمایش بهتری از تصویر نشان دهد. همچنین لایه توجه بر توجه باعث شد که مدل ما بتواند روابط بین اشیا را به خوبی درک کند. همچنان حوزه توصیف تصویر به صورت خودکار جای پیشرفت بسیاری دارد. همچنین چالش معیارهای ارزیابی همانطورکه در مقدمه ذکر شد جای کار بسیار داردو هنوز محققان در این حوزه با این مشکل رو به رو هستند.

مراجع

[۱] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proceedings of the second workshop on shortcomings in vision and language*, 2019, pp. 26-36 .

[۲] Z. F. Sattari, H. Khotanlou, and E. Alighardash, "Improving Image Captioning with Local Attention Mechanism," in *2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, 2022: IEEE, pp. 1-5 .

[۳] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164 .

[۴] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407-2415 .

[۵] M. Zamiri and H. S. Yazdi, "Image annotation based on multi-view robust spectral clustering," *Journal of Visual Communication and Image Representation,* vol. 74, p. 103003, 2021.

[۶] M. Zamiri, T. Bahraini, and H. S. Yazdi, "MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging," *Expert Systems with Applications,* vol. 173, p. 114657, 2021.

[۷] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi , "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 220-228 .

[۸] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, no. 12, pp. 2891-2903, 2013.

[۹] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*, 2010: Springer ,pp. 15-29 .

[۱۰] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2596-2604 .

[۱۱] V. Ordonez, G. Kulkarni, and T. Berg" ,Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems,* vol. 24, 2011.

[۱۲] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *European conference on computer vision*, 2014: Springer, pp. 529-545 .

[۱۳] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research,* vol. 47, pp. 853-899, 2013.

[۱۴] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128-313 .۷

[۱۵] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems,* vol. 27, 2014.

[۱۶] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539,* 2014.

[۱۷] Q. Wu, C. Shen, P. Wang, A. Dick, and A. Van Den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 6, pp. 1367-1381, 2017.

[۱۸] J. Li, N. Xu, W. Nie, and S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching," *Visual Informatics,* vol. 5, no. 4, pp. 41.۲۰۲۱ ,۴۸-

[۱۹] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375-383 .

[۲۰] M .Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR),* vol. 51, no. 6, pp. 1-36, 2019.

[۲۱] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[۲۲] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259,* 2014.

[۲۳] C. Wang and X. Gu, "Image captioning with adaptive incremental global context attention," *Applied Intelligence,* pp. 1-23, 2021.

[۲۴] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422-2431 .

[۲۵] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015: PMLR, pp. 2048-2057 .

[۲۶] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1242-1250 .

[۲۷] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, "Image captioning using DenseNet network and adaptive attention," *Signal Processing: Image Communication,* vol. 85, p. 115836, 2020.

[۲۸] S. Cao, G. An, Z. Zheng, and Q. Ruan, "Interactions guided generative adversarial network for unsupervisedimage captioning," *Neurocomputing,* vol. 417, pp. 419-431, 2020.

[۲۹] D. Wang, Z. Hu, Y. Zhou, R. Hong, and M. Wang, "A Text-Guided Generation and Refinement Model for

Image Captioning," *IEEE Transactions on Multimedia,* 2022.

[۳۰] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634-4643 .

[۳۱] Y. Wei, L. Wang, H. Cao, M. Shao, and C. Wu, "Multi-attention generative adversarial network for image captioning," *Neurocomputing,* vol. 387, pp. 91-99, 2020.

[۳۲] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014: Springer, pp. 740-755 .

[۳۳] K. Papineni, S. Roukos, T. Ward, andW.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318 .

[۳۴] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MTevaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72 .

[۳۵] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575 .

[۳۶] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," *arXiv preprint arXiv:1612.07600,* 2016.

[۳۷] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81 .

[۳۸] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligningwhere to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, no. 12, pp. 2321-2334, 2016.

[۳۹] L. Li, S. Tang, Y. Zhang, L. Deng, and Q .Tian, "GLA: Global–local attention for image description," *IEEE Transactions on Multimedia,* vol. 20, no. 3, pp. 726-737, 2017.

[۴۰] X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, and X. Niu, "Image captioning with triple-attention and stack parallel LSTM," *Neurocomputing,* vol. 319, pp. 55-65, 2018.

[۴۱] Y. Cheng, F. Huang, L. Zhou, C. Jin, Y. Zhang, and T. Zhang, "A hierarchical multimodal attention-based neural network for image captioning," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 889-892 .

[۴۲] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing,* vol. 398, pp. 520-530, 2020.

[۴۳] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," *IEEE Transactions on Multimedia,* vol. 21, no. 11, pp. 2942-2956, 2019.

[۴۴] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention ",in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651-4659 .

[۴۵] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10323-10332 .

[۴۶] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7414-7424 .

[۴۷] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125-4134 .

[۴۸] L. Chen *et al.*, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659-5667 .

زهرا فامیل ستاری در سال ۱۳۹۸ مدرک کارشناسی خود را در رشته مهندسی کامپیوتر از دانشگاه بوعلی سینا دریافت نمود و در سال ۱۴۰۰ مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه بوعلی سینا اخذ نمود. حوزه‌های پژوهشی مورد علاقه ایشان یادگیری عمیق، بینایی ماشین و پردازش زبان طبیعی است.

حسن ختن‌لو استاد گروه مهندسی کامپیوتر دانشگاه بوعلی سینا، در سال ۱۳۸۷ دکترای مهندسی کامپیوتر را از دانشگاه پیر و ماری کوری اخذ و تا به حال به عنوان عضو هیات علمی گروه مهندسی کامپیوتر، مشغول به فعالیت‌می باشد و زمینه‌های تحقیقاتی مورد علاقه ایشان، پردازش تصویر و ویدیو، پردازش تصاویر پزشکی، شناسایی الگو و یادگیری ماشین است.

الهام علیقارداش مربی گروه مهندسی کامپیوتر دانشگاه سید جمال الدین اسدآبادی، در سال ۱۳۹۱مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر- گرایش هوش مصنوعی از دانشگاه بوعلی سینا اخذ نمود. وی در حال حاضر دانشجوی دکتری مهندسی کامپیوتر در دانشگاه بوعلی سینا می باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان بازشناسی الگو، پردازش تصویر، بینایی ماشین، یادگیری ماشین یادگیری چندوجهی و یادگیری عمیق است.