نگاهی بر بازیابی معنایی تصاویر در وب

محمد مهدی حاجی اسمعیلی[۱]، غلامعلی منتظر[۲]

گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع و سیستم ها، دانشگاه تربیت مدرس، تهران،
mohammadhaji@modares.ac.ir

گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع و سیستم ها، دانشگاه تربیت مدرس، تهران،
montazer@modares.ac.ir

چکیده

تعداد، تنوع و پیچیدگی محتوای تصویری در دنیای دیجیتالی به‌سرعت در حال افزایش است و این موضوع نیاز به طراحی و پیاده‌سازی سیستم‌های جویش و بازیابی محتوای تصویری را بسیار محسوس کرده است. در حال حاضر با مقیاس عظیمی از داده‌های تصویری در فضای وب روبرو هستیم که راهکارهای معمول مبتنی بر فراداده‌های دستی و انسانی پاسخگوی تنوع و تعداد بسیار زیاد آن‌ها نیست. جویشگر گوگل حدود ۶۰ هزار میلیارد صفحه‌ی قابل جستجو را نمایه‌سازی کرده و در هر ثانیه پاسخگوی بیش از ۲/۳ میلیون جستجو است. فلیکر به‌عنوان یک وبگاه اشتراک‌گذاری تصاویر در حال حاضر شامل بیش از ۱۰ میلیارد تصویر است که روزانه ۱ میلیون تصویر به آن افزوده می‌شود. وبگاه اینستاگرام در هر روز شاهد بارگذاری حدود ۸۰ میلیون تصویر است و تا به امروز حدود ۴۰ میلیارد تصویر در آن به اشتراک گذاشته شده است. اخیراً تلاش‌های بسیاری برای بازیابی تصاویر و به‌ویژه در مبحث «بازیابی‌محتوامحورتصویر» (CBIR) شکل‌گرفته است. یک سیستم بازیابی محتوامحور تصویر، قادر به جستجو و بازیابی تصاویرِ موردنظر بر اساس محتوای درونی تصویر است و نه فراداده‌هایی‌که ممکن است همراه با آن ثبت شده باشند. رویکرد این تحقیق متمرکز بر مفهومی پیچیده‌تر با نام «بازیابی معنایی» است و هدف آن بازیابی تصاویر بر اساس مفاهیم معنایی و سطح بالای موجود در آن و نه‌فقط محتوا و اجزای تشکیل‌دهنده آن است. در این مقاله در ابتدا به تعریف مسئله ی بازیابی معنایی تصاویر، اهداف و سؤالات اصلی این حوزه‌ی تحقیقاتی پرداخته و در ادامه به بررسی ضرورت و تاریخچه آن خواهیم پرداخت. در بخش سوم، پیشینه‌ی به‌کارگیری روش‌های نرم در بحث استخراج معنا و بازیابی معنایی تصاویر را در سال‌های اخیر بررسی کرده و به بیان روش‌های نوینی که دارای نتایج قابل‌قبولی بوده‌اند، خواهیم پرداخت. در ادامه و در بخش نهایی شکاف‌های مطرح در این حوزه را بیان کرده و نقاط ضعف روش‌های یاد شده را بررسی خواهیم کرد.

کلمات کلیدی

بازیابی معنایی تصویر، بازیابی محتوامحور تصویر، توصیف تصویر، روش‌های نرم رایانش، یادگیری عمیق

۱- مقدمه

برای انسان، نگاهی گذرا و کوتاه به یک تصویر یا منظره کافی است تا او را قادر به تشریح جزئیات پیچیده‌ای از آن تصویر کند درحالی‌که انجام این کار برای ماشین‌ها، مسئله‌ای حل‌نشده در طول سال‌های گذشته محسوب می‌شود [1]. برای درک تصویر، نیاز به رویکرد و ابزاری فراتر از روش‌هایی همچون «دسته‌بندی اشیا» یا «شناسایی مکان آن‌ها» است. بدون چنین ابزاری، تنها چیزی که از تصاویر به دست می‌آوریم، مجموعه‌ای از اطلاعات پردازش‌شده و خام است که هیچ‌گاه به‌خودی‌خود قادر به پر کردن خلأ بین درک معنایی انسان از یک تصویر و شناسایی سطح پایین ماشین از آن نیستند.

سه سطح مختلف برای پرس‌وجوهای محتوامحور وجود دارد [2]:

۱. سطح اول: بازیابی براساس ویژگی‌های اصلی تصویر همچون رنگ، شکل، بافت یا مکان فضایی عناصر درون تصویر. پرس‌وجوی عمومی این دسته غالباً به شکل «عکس‌هایی شبیه این عکس را پیدا کن» است.
۲. سطح دوم: بازیابی اشیایی از یک نوع خاص (که توسط کاربر مشخص می‌شود) که ممکن است با استنتاج منطقی همراه باشد. به‌عنوان‌مثال «تصویر یک زرافه را جستجو کن».
۳. سطح سوم: بازیابی بر مبنای ویژگی‌های انتزاعی سطح بالا که همراه با استنتاج بعضاً قوی در مورد چگونگی تصویر است. به‌عنوان‌مثال «تصویر کودکانی در حال بازی با یک گربه در پارک» ازجمله‌این پرس‌وجوهای سطح بالا است.

سطوح دوم و سوم در کنار هم بانام «بازیابی معنایی تصویر » شناخته می‌شوند و شکافی که بین آن دو با سطح اول به وجود می‌آید، بانام «شکاف معنایی » معرفی می‌شود. به‌عبارت‌دیگر، اختلاف عمیق بین قدرت تشریحی ضعیف ویژگی‌های سطح پایین عکس و درخواست‌های بعضاً غنی و پیچیده سطح بالای کاربر، به‌عنوان «شکاف معنایی» شناخته می‌شود. از سال ۲۰۱۲ میلادی با رشد تحقیقات در حوزه‌ی پردازش تصویر و یادگیری ماشینی، از پیچیدگی مسائل سطح دوم (که بیشتر تمرکز را بر بازیابی تصاویر حاوی یک شیء خاص قرار داده بودند)کاسته شده است [3]. لیکن مسائل سطح سوم هنوز مسئله‌ای حل نشده در این حوزه محسوب می‌شوند. این موضوع به عواملی همچون پیچیدگی استنتاج درباره‌ی روابط بین اشیا و نیاز به طراحی رابطی انسان‌پسند برای ارتباط بین درخواست‌های انسانی و خروجی ماشینی وابسته است. در شکل ۱ نمونه‌ای از تصاویر نشان داده شده که هم دارای مفاهیم سطح پایین (همچون اشیا) و هم سطح بالا (همچون رُخداد و عمل) هستند:


ج ب الف

شکل (۱) : تصاویری با مفاهیم معنایی سطح بالا

در جدول ۱ نمونه‌ای از تفاوت مفهومی بین «بازیابی محتوامحور» و «بازیابی معنایی» برای تصاویر شکل ۱ نمایش داده شده است:

جدول (۱) : تفاوت مفهومی بین بازیابی محتوامحور و بازیابی معنایی برای تصاویر شکل ۱

شکل	بازیابی محتوامحور	بازیابی معنایی
الف	مرد / فرش / کلاه	تعزیه‌خوانان روی فرش نشسته‌اند
ب	پیرمرد / جوراب / کیسه	پیرمرد جوراب‌فروش
ج	مرد / چراغ / قبر	مردی در حال دعا خواندن در قبرستان

بحث بازیابی معنایی تصاویر به چندین حوزه‌ی پردازشی و بعضاً متفاوت گسترش یافته است. تکیه بر روش‌های مجزای بینایی ماشین برای بازیابی تصاویری که همراه با مفهوم باشند کافی نیست و از طرفی وابستگی به قالب‌های از پیش تعیین شده و ایستا نیز راهکاری کاربردی برای شناسایی معنا و مفاهیم نامحدود موجود در فضای تصویر محسوب نمی‌شود [4]. روش‌های ارائه‌شده‌ای که در مرز دانش قرار دارند هنوز فاصله‌ای ۴۰ درصدی از عملکرد انسانی داشته و حتی در بهترین حالت نیز در توصیف تصاویر غیرعادی به‌ندرت دقیق عمل می‌کنند [5]. همان‌طور که درک لطیفه‌ها و ضرب‌المثل‌ها و تلاش برای ایجاد آن‌ها توسط ماشین‌ها هنوز مسئله‌ای حل‌نشده است، درک «مفاهیم ضمنی » در تصاویر نیز مسئله‌ای باز محسوب می‌شود. به‌عنوان‌مثال در حال حاضر شاید شناسایی و دسته‌بندی تصاویر، دیگر آن‌چنان کار سختی به نظر نرسد (درحالی‌که هنوز این مسئله حل‌نشده است) ولی در قدم بعدی شناسایی و درک ضمنی تصاویری همچون کاریکاتورها و تصاویر طنزمحور، مسئله‌ای است که حتی در تحقیقات امروزی نیز مطرح نشده است. به دلیل سختی بیش‌ازحد این مسئله، شاهد تحقیقات پراکنده‌ای در این حوزه‌ها در طول ۵۰ سال اخیر بوده‌ایم ([6] و [7]). درک ضمنی محتوا و مفاهیم چه در حوزه‌ی متن و چه در حوزه‌ی تصویر، نیازمند راهکاری است که در ابتدا قادر به درک کلی این حوزه‌ها باشد و در حال حاضر هیچ راهکار جامعی برای برداشتن قدم اول در این راه ارائه نشده است.

۲. دیدگاه‌های مطرح

دیدگاه‌های متفاوتی نسبت به چگونگی حل این مسئله در بین محققان وجود دارد.گستره‌ی بزرگی از تلاش محققان بر روی «برچسب‌زنی تصاویر»[1] بر اساس موجودیت‌های درون آن‌ها متمرکز شده است چراکه اعتقاد دارند برچسب‌زنی تصاویر به زبان طبیعی نه‌تنها ما را قادر به جستجوی تصاویر دیگر بر مبنای محتوای آن‌ها می‌کند، بلکه مبحث بازیابی بر اساس «معنا» را نیز تا حدی پوشش می‌دهد چراکه برچسب‌ها قدرت نسبی بهتری در توصیف تصاویر و رخدادهای درون آن‌ها دارند تا ویژگی‌های سطح پایین. به‌عنوان‌مثال حتی اگر پرس‌وجوی ما یک تصویر باشد، می‌توان ابتدا برچسب‌های معنایی و موجودیتی درون آن را استخراج کرده و سپس تمامی تصاویری که ازلحاظ برچسب مشابه آن هستند را استخراج و بازیابی کرد. با اینکه این رویکرد شاهد دستاوردها و پیشرفت‌های بسیار خوبی در سال‌های اخیر بوده است [8] [9] ولی محدودیت‌های عمیقی همچون دایره‌ی لغاتِ بسته و عدم بررسی روابط بین موجودیت‌های درون تصویر، آن را تبدیل به روشی نه‌چندان مناسب برای استخراج معانی از تصاویر و بازیابی تصاویر بر اساس آن می‌کند. به‌عنوان‌مثال برای تصویری از «یک دوچرخه‌سوار که در حال پرش از روی یک تپه خاکی است»، قادر به تشخیص یک «انسان» و یک «دوچرخه» درون تصویر هستیم و می‌توانیم دو برچسب یادشده را همراه تصویر ثبت و نگهداری کنیم، ولی تکیه‌ی صرف بر این دو برچسب برای بازیابی، می‌تواند منجر به بازیابی تصاویری همچون یک «خیابان و دوچرخه‌سواران داخل آن»، «پارکینگی از دوچرخه‌ها که مردم در حال عبور از کنار آن هستند»، «یک دوچرخه فروشی»

[1] Image Labeling

و یا «مسابقه‌ی پرش از خاک‌ریز به کمک دوچرخه» شود. در بین موارد فوق، فقط یک دسته از این تصاویر، به‌عنوان یک بازیابی مناسب محسوب می‌شوند (مسابقه‌ی پرش از خاک‌ریز) و دسته‌های دیگر اشتباهاتی هستند که فقط با تکیه بر بازیابی موجودیت‌های سطح بالا (انسان و دوچرخه) رخ داده‌اند.

با توجه به موضوع یادشده، می‌توان به نقش مهم زبان طبیعی در بازیابی معنایی تصاویر پی برد چراکه بازیابی چه بر اساس یک پرس‌وجو (به زبان طبیعی) و چه بر اساس یک تصویر اولیه، باید قادر به توصیف تصویر و «روابط» درون آن باشد و آن چیزی که ما را قادر به انجام این کار می‌کند، زبان طبیعی است. این موضوع خود منجر به شکل‌گیری رویکردهایی برای توصیف تصاویر به کمک زبان طبیعی شده است. محققان این حوزه، بازیابی معنایی را مرتبط با استخراج معانی از تصاویر به کمک زبان طبیعی می‌دانند و تلاش دارند روش‌هایی برای تلفیق این دو حوزه یعنی تصویر و زبان طبیعی پیدا کنند که قادر به درک تصویر و حتی‌الامکان توصیف آن به بهترین روش به کمک زبان طبیعی باشد.

در این دیدگاه «معنا» مفاهیم سطح بالای انتزاعی محسوب می‌شود که در تصاویر می‌توان یافت، به‌عنوان‌مثال با دیدن تصویری از یک عزاداری نه‌تنها باید بتوان تشخیص داد که این جمعیت برای چه هدفی دور هم جمع شده‌اند، بلکه باید مفاهیمی همچون «غم و غصه» یا «از دست دادن یک شخص» را نیز از تصویر برداشت کرد.

۲-۱- روش‌های ایستا برای توصیف یک تصویر

مسئله‌ی توصیف یک تصویر، سال‌هاست که در حوزه‌ی بینایی ماشین و خصوصاً در حوزه‌ی تصاویر ویدئویی مطرح بوده است [10] [11] و تلاش برای حل این مسئله در طول این سال‌ها موجب شکل‌گیری روش‌هایی بسیار پیچیده و متشکل از شناسنده‌های اولیه‌ی بصری با گرامرهای زبانی ایستا شده است که با انواع و اقسام گراف‌ها و سیستم‌های منطق محور و پایگاه‌های داده‌ی «اگر-آنگاه» تلاش می‌کنند تصویری را توصیف کنند. در سال‌های اخیر، چندین روش ایستا برای توصیف تصاویر به کمک زبان‌های طبیعی همچون [12] و [13] ارائه شده‌اند که مشخصه‌ی اصلی آن‌ها غالباً تکیه بر مفاهیم بصری و قالب‌هایی است که از قبل و به‌صورت دستی توسط محقق طراحی شده‌اند. تمرکز اکثر این روش‌ها بر کاهش پیچیدگی یک تصویر و ایجاد جمله‌ای ساده برای توصیف آن است که محدودیتی نامناسب برای کار با تصاویر محسوب می‌شود. در شکل ۲ عملکرد چنین روش‌هایی برای توصیف تصاویر نشان داده شده است:

شکل (۲) : عملکرد روش‌های ایستا و قالب محور برای توصیف تصاویر (بالا) [12]، (پایین) [13]

روش استفاده شده در [12] بر مبنای استخراج مؤلفه‌های سه تایی از تصویر و پر کردن یک قالب جمله‌ای بر اساس آن‌هاست. این روش تا حدی قادر به پردازش تصاویر بوده و نه‌تنها توانسته گاو و سگ‌های درون تصاویر را تشخیص دهد بلکه از طرفی قادر به ارائه‌ی تشریحی قابل‌قبول از محتوای درون تصویر است. بااین‌حال ازلحاظ پردازش تصویر، روش فوق قوی عمل نکرده و با وجود شناسایی دو سگ، یکی از آن‌ها را پشمالو تشخیص داده است درصورتی‌که هیچ‌کدام دارای این مشخصه نیستند. همچنین قالب پیش‌فرضی که برای توصیف تصاویر به کار رفته است، به‌سرعت قابل‌تشخیص و شناسایی است که باعث می‌شود توصیف به دست آمده از تصویر بیش‌ازپیش مصنوعی به نظر آید. در روش [13] راهکاری پیچیده‌تر (ولی مشابه با [12]) و با مؤلفه‌هایی بیشتر برای تولید متن از قالب‌های پیش‌ساخته، به کار گرفته شده است. در این روش نیز شاهد عملکردی نه‌چندان قوی، هم در پردازش تصویر و هم در توصیف محتوای آن هستیم. در بین ۵ توصیفی که برای تصویر «گوسفندی در مه» ارائه شده‌اند، فقط یکی از آن‌ها به درستی قادر به تشریح تصویر است درحالی‌که بقیه‌ی توصیفات به صورتی کاملاً نادرست به توصیف تصویر پرداخته‌اند. این ضعف با اینکه در تصویر یخچال کمی بهتر شده است ولی باز هم دارای نواقصی است که با خواندن توصیفات به راحتی می‌توان متوجه آن‌ها شد. این محدودیات فقط برای تصاویر یادشده نیست و در مقاله‌های یادشده نمونه‌های بسیاری از تصاویری که به درستی توصیف شده و بخش اعظمی که به صورتی نادرست یا ناکامل توصیف شده‌اند، قابل مشاهده است.

مهم‌ترین نقطه‌ی ضعف چنین روش‌هایی، تلاش برای محدود کردن فضای توصیفی برای رسیدن به یک توصیف قابل‌قبول است. قالب‌های از پیش تعریف شده‌ای که محققان باید با دقت آن‌ها را طراحی کنند، هیچ‌گاه پاسخگوی تنوع تصاویر و ساختار جملات توصیفی آن‌ها نخواهد بود و همان‌طور که دیدیم تلاش برای این کار موجب دستیابی به جملاتی مصنوعی و غیرطبیعی می‌شود. از طرفی در بحث بازیابی، چنین رویکردی کاربر را مجبور به پیروی از قالب‌های از پیش تعریف‌شده‌ی محققان می‌کند که در سیستم‌های بازیابی، یک ویژگی نامناسب و منفی تلقی می‌شود [3].

روش‌هایی بهتر و با کارایی بیشتری همچون [14]، [15] و یا [16] نیز در طول سال‌های اخیر در این حوزه ارائه شده‌اند ولی همگی دارای نقاط ضعفی هستند که در بالا بدان‌ها اشاره کردیم.

۲-۲- روش‌های پویا برای توصیف یک تصویر

محدودیت‌های روش‌های یادشده، رویکردی را می‌طلبد که قادر به ایجاد ارتباطی عمیق بین تصویر و اشیای درون آن با متنی به زبان طبیعی باشد. این روش نه‌تنها باید قادر به شناسایی اشیا باشد بلکه باید بتواند روابط بین آن‌ها را استنتاج کرده و فعالیت‌هایی که در حال رخ دادن در تصویر است را شناسایی کند چراکه تک‌تک این ویژگی‌ها، بر معنا و مفهوم در حال رخداد درون تصویر تأثیرگذار است. از طرفی پس از شناسایی مفاهیم یادشده، این روش باید قادر به توصیف آن به زبانی طبیعی باشد که خود چالشی بزرگ محسوب می‌شود. ویژگی‌های فوق این مسئله را بسیار پیچیده‌تر از مسائلی همچون دسته‌بندی، شناسایی سبک و یا شناسایی اشیای درون تصویر و مکان آن‌ها می‌کند چراکه برخلاف این مسائل تک‌بعدی، با مسئله‌ای چندبُعدی طرف هستیم که خود ممکن است از روش‌های یادشده در حوزه‌هایی مختلف بهره گیرد.

یکی از مهم‌ترین و بنیانی‌ترین مسائل این حوزه، برقراری ارتباطی معنادار بین مؤلفه‌های تشکیل‌دهنده‌ی یک تصویر و اجزای تشکیل‌دهنده‌ی متنی است که آن را توصیف می‌کند. در شکل ۳ نمونه‌ای از یک تصویر و متن توصیفی آن نمایش داده شده است:

شکل (۳) : ارتباط مؤلفه‌های یک تصویر با اجزای متن توصیفی آن [17]

در شکل فوق شاهد جمله‌ای هستیم که تصویر «سگی درون آب» را توصیف می‌کند. ابتدایی‌ترین مسئله‌ای که باید بتوان در حوزه‌ی توصیف تصاویر حل کرد، مرتبط کردن مؤلفه‌های شناسایی شده درون تصویر با اجزای درون متن است. این شکل شامل سه مؤلفه‌ی اصلی یعنی «کل تصویر»، «تصویر سگ به همراه توپ تنیس» و «تصویر توپ تنیس» است. از طرفی اجزایی از این متن باید قابل ارتباط با تصاویر ذکر شده باشند و همان‌طور که می‌بینیم قادر به جداسازی ۴ قطعه متن مجزا و متفاوت شده‌ایم که قادر به توصیف مؤلفه‌های تصویری هستند. حل چنین مسئله‌ای ما را قادر به ایجاد ارتباط عمیق‌تری بین دو فضای مجزای تصاویر و زبان طبیعی می‌کند که این موضوع خود تأثیر مستقیمی بر تولید متون جدید برای تصاویر نادیده دارد.

روش‌های متفاوت و بعضاً ناکارآمدی برای این مسئله ارائه شده است. در روش‌هایی از [18] و [19] از تحلیل همبستگی کانونی هسته برای هم‌تراز کردن قطعات متن و تصویر استفاده شده است. این روش مقیاس‌پذیری ضعیفی دارد چراکه باید به تعداد جملات و تصاویر، هسته‌های درجه‌دو ایجاد کند که این محدودیت موجب می‌شود روش‌های یادشده در محیطی غیر آزمایشگاهی، کاربردی عملیاتی نداشته باشند.

در روشی که در [12] ارائه شد، یک فضای معنای یکسان بین فضاهای تصاویر و متون ایجاد می‌شود. متأسفانه محدودیت این روش در تلاش برای دسته‌بندی و نمایش دادن تصاویر و متون به کمک ۳ مؤلفه‌ی محدود «شئ»، «فعل» و «منظره» است. رویکرد ایستای این روش همان‌طور که در قبل نیز اشاره کردیم، تلاش در محدود کردن فضای توصیفی برای رسیدن به یک توصیف قابل‌قبول دارد که موجب دستیابی به جملاتی مصنوعی و غیرطبیعی می‌شود.

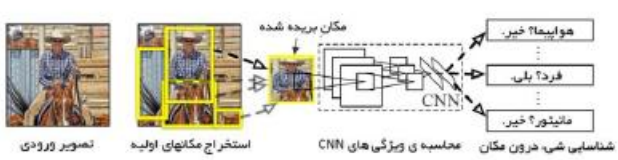
در روشی دیگر از یک «میدان تصادفی شرطی» برای استنتاج روابط فضایی در تصاویر کارتونی و ارتباط آن‌ها با توصیفات مبتنی بر زبان طبیعی استفاده شده است [7].

۳- روش‌های نوین

۳-۱- هم ترازی قطعات تصویر و متن

رویکرد [17] برای حل این مسئله، در طول سال‌های اخیر تبدیل به زیرساختی موفق در توصیف تصاویر به کمک زبان طبیعی شده است. در این راهکار، تلاش شده تا محدودیت‌های یادشده در روش‌های قبلی نه‌تنها پوشش داده شوند بلکه تبدیل به نقاط قوتی از سیستم گردند. در این روش هدف اصلی ایجاد ارتباط بین اجزای تصویر و اجزای متن توصیفی آن است و برای نیل به این هدف از تلفیق دو شبکه‌ی عصبی ژرف (در سمت تصویر) و یک «درخت تجزیه وابستگی » (در سمت متن) استفاده شده است.

برای شناسایی اجزای درون تصویر و استخراج ویژگی‌های آن‌ها، از دو شبکه‌ی ژرف استفاده شده است که شبکه‌ی اول برای شناسایی مکان اشیاء درون تصویر و شبکه‌ی دوم برای استخراج ویژگی‌های اشیاء شناسایی شده بکار می‌رود. شبکه‌ی اول شبکه‌ی ژرف R-CNN است که قادر به شناسایی مکان اشیاء در تصویر و جداسازی مکان آن‌ها نسبت به هم است. این شبکه بر روی مجموعه داده‌های ImangeNet آموزش دیده و به‌صورت دستی برای شناسایی ۲۰۰ کلاس (از ۱۰۰۰ کلاس) این مجموعه، تنظیم شده است [20]. نمونه‌ای خلاصه از معماری این شبکه در شکل ۴ نمایش داده شده است:



شکل (٤) : شبکه‌ی R-CNN برای استخراج مکان‌های اشیای درون تصویر

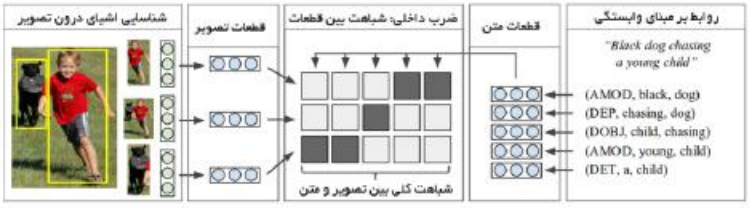
پس از شناسایی مکان هر شئ درون تصویر، پیکسل‌های این اشیاء برای استخراج ویژگی‌ها، به شبکه‌ی بعدی داده می‌شوند.در این مرحله ۱۹ قطعه‌ی تصویر بعلاوه ی قطعه‌ی کلی، تحویل شبکه‌ای هم‌ساختار با AlexNet شده

و در مرحله‌ی بعدی داده‌های ۴۰۹۶ بُعدی آخرین لایه‌ی شبکه به‌عنوان ویژگی‌های این قطعات استخراج و نگهداری می‌شوند.

در قسمت دیگر، برای محاسبه‌ی ارتباط میان متن و تصویر نیاز به انتقال متن به فضایی جدید است تا بتوان محاسبات بین ویژگی‌های استخراج‌شده‌ی تصویری و قطعات درون متن را به صورتی مناسب انجام داد. برای این کار در ابتدا متن توصیفی یک تصویر را به قطعات کوچک‌تری تقسیم می‌کنند. هدف این است که ارتباط بین قطعات متنی و تصویری مشخص و محاسبه شود و برای این کار همان‌طور که تصویر به قطعاتی کوچک‌تر تقسیم شد متن توصیفی نیز با به قطعاتی کوچک‌تر شکسته شود. در این روش برای شکستن متن به قطعات کوچک‌تر، از یک درخت تجزیه وابستگی استفاده شده و از هر ارتباط درون این درخت (یالی که بین گره‌ها برقرار است) یک قطعه متن دوکلمه‌ای استخراج شده است. به‌کارگیری درخت تجزیه وابستگی بسیار بهتر از شکستن متن بر اساس کلمه‌های مجزا و یا bigram های متصل به هم است .

در مرحله‌ی بعدی، باید کلمات را به فضایی برداری و قابل محاسبه ببرند. برای این کار در ابتدا هر کلمه‌ی درون متن را به یک بردار 1-of-k تبدیل کرده و سپس آن‌ها را با یک ضرب (W_e) به فضایی جدید می‌برند (در این تحقیق فضای جدید d=200 بُعدی است).

با داشتن قطعات تصویر و متن در فضایی قابل محاسبه، ضربی داخلی از تمامی قطعات تصویر و متن محاسبه شده و خروجی آن به‌عنوان یک شاخص امتیازدهی بین این دو زوج (قطعه‌ی تصویر و قطعه‌ی متن) در نظر گرفته می‌شود. در شکل ۲-۵ روند کار تا بدین مرحله نمایش داده شده است:



شکل (۵) : محاسبه‌ی شباهت بین اشیاء درون تصویر و قطعات متن

در مثال فوق، ۳ قطعه‌ی تصویر (کل تصویر، سگ، کودک) و ۵ قطعه‌ی متنی از زوج‌های به هم وابسته داریم. ضرب داخلی تمامی این قطعات برابر ۱۵ حالت مختلف است. نقاط پررنگ‌تر نشان‌دهنده‌ی امتیاز مثبت (پشتیبانی از شباهت بین تصویر و قطعه‌ی متن) و نقاط کم‌رنگ‌تر نشان‌دهنده امتیاز منفی (عدم هم‌ترازی بین تصویر و قطعه‌ی متن) است. به‌عنوان‌مثال در ردیف سوم (ضرب داخلی تصویر سگ و ۵ قطعه‌ی متن) شاهد دو امتیاز مثبت برای متن‌های [black,dog] و [chasing, dog] و سه امتیاز منفی برای دیگر حالات هستیم. این امتیاز، مقداری نسبی است که به‌خودی‌خود معنایی قابل‌درک ندارد ولی با ایجاد یک تابع هدف می‌توان به امتیاز به دست آمده و تغییرات آن نسبت به متن یا تصویر، معنا داد.

تابع هدف نهایی، تابعی سه مؤلفه‌ای است که از یک مؤلفه برای هم‌ترازی قطعات تصویر و متن، یک مؤلفه برای هم‌ترازی کل تصویر و متن اصلی و مؤلفه‌ای نهایی برای تنظیم تشکیل شده است؛ مؤلفه‌ی هم‌ترازی قطعات متن و تصویر، تابع را به سویی هدایت می‌کند تا برای یک قطعه‌ی تصویری، حداقل یک قطعه‌ی متن توصیفی هم‌تراز با آن پیدا شود. مؤلفه‌ی هم‌ترازی کل، مطمئن می‌شود که تصویر و متن توصیفی اصلی آن، امتیازی بالاتر از تمامی توصیفات دیگر برای تصویر دارند؛ و مؤلفه‌ی تنظیم، برای جلوگیری از بیش برازش تعبیه شده است.

زیرساخت این روش طوری طراحی شده است که از یک سمت به شبکه‌های عصبی تصویری متصل شود و از سمتی دیگر به یک شبکه‌ی عصبی دو لایه که حول خروجی درخت وابستگی و کلمات تعبیه شده عمل می‌کند به‌طوری‌که خروجی شبکه، مقدار نهایی کلمه در فضای جدید خواهد بود. بدین ترتیب می‌توان تابع هدف را به کمک روشی همچون کاهش گرادیان بهینه کرد. برای بهینه‌سازی این تابع، از SGD با دسته‌های ۱۰۰ تایی و بیست چرخه تکرار برای آموزش شبکه بر روی تصاویر و متون آموزشی (شامل ۸ هزار نمونه تصویر و متن) استفاده شده است.

پس از یادگیری و بهینه‌سازی، مدل نهایی قادر است با گرفتن یک تصویر، بهترین جمله‌ی توصیفی به زبان طبیعی را از میان جمله‌هایی که قبلاً ندیده است برای این تصویر انتخاب کند و از طرفی با گرفتن یک متن، بهترین تصویر با قطعاتی هم‌تراز قطعات متن را بازیابی کند. در شکل ۶ نمونه‌ای از عملکرد این روش برای بازیابی متون برای ۳ تصویر نمایش داده شده است:

شکل (٦) : متون توصیفی بازیابی شده برای تصاویر

در تصویر سمت راست جمله به خوبی قادر به توصیف تصویر شده است. حضور «دوچرخه»، «فرد» و مهم‌تر از آن «در هوا معلق بودن» به خوبی توسط مدل شناسایی شده است بااین‌حال ضعف درخت تجزیه، خود را در اینجا نشان می‌دهد. ازآنجاکه این روش برای انتخاب قطعات متنی فقط به یال‌های درخت تجزیه اکتفا کرده است، همیشه انتظار آن می‌رود که زوج کلمات استخراج شده از این یال‌ها، واقعاً ارتباطی معنایی با هم نداشته باشند. در تصویر سمت راست، دو کلمه‌ی «Blue» و «Person» با هم از این درخت به‌عنوان یک قطعه‌ی متنی جداسازی شده‌اند. طبیعی است که در تصویر هیچ «انسانِ آبی» چه ازلحاظ رنگ لباس و چه دیگر مشخصه‌ها موجود نیست ولی همین ضعف درخت تجزیه در قطعه‌بندی کلمات موجب اشتباه مدل در انتخاب جمله‌ی توصیفی شده است : رنگ آبی برای پرچمی است که در کناره‌ی تصویر قرار دارد و نه فرد دوچرخه‌سوار.

این روش در زمان ارائه، تمامی مدل‌های دیگری را که برای حل این مسئله ارائه شده بودند، با عملکردی ۳۹ درصدی شکست داد (میزان عملکرد بر اساس معیار R@1 در بازیابی یک عنصر در نظر گرفته شده است). در بین این مدل‌ها می‌توان به [21] با عملکردی ۴/۸ درصدی و [۲۲] با عملکردی ۴/۵ درصدی برای تصاویر مجموعه‌ی Pascal1K اشاره کرد. با وجود این برتری نسبی، کارایی این روش و دقت آن، هنوز فاصله‌ی بسیار زیادی با عملکرد انسانی در توصیف تصاویر و بازیابی معنایی آن‌ها در حافظه دارد.

از طرفی نقاط ضعف آن نیز تبدیل به تنگنایی شده‌اند که مستقیماً به عملکرد آن صدمه می‌زند. علاوه بر محدودیت‌های درخت تجزیه در بحث «شمردن»،

نقاط ضعف دیگری نیز در آن وجود دارد. به‌عنوان‌مثال با داشتن جمله‌ای همچون «سگ سیاه و سفید» ممکن است درخت تجزیه آن را به دو قطعه‌ی [سگ، سیاه] و [سگ، سفید] جداسازی کند که حکایت از حضور بیش از یک سگ در تصویر دارد. همچنین برخی زوج کلمات قطعه‌بندی شده، هیچ معنای مناسبی ندارند که در بین آن‌ها می‌توان به زوج‌هایی مثل [each,other] یا [to,do] اشاره کرد. از طرفی مؤلفه‌ی تصویری نیز بدون اشکال نیست و مشکل «حذف غیرحداکثری » در شبکه‌ی R-CNN ممکن است منجر به شناسایی دوباره‌ی یک شئ در مکان‌های نزدیک به هم شود.

۳-۲- توصیف و بازیابی به کمک زبان طبیعی

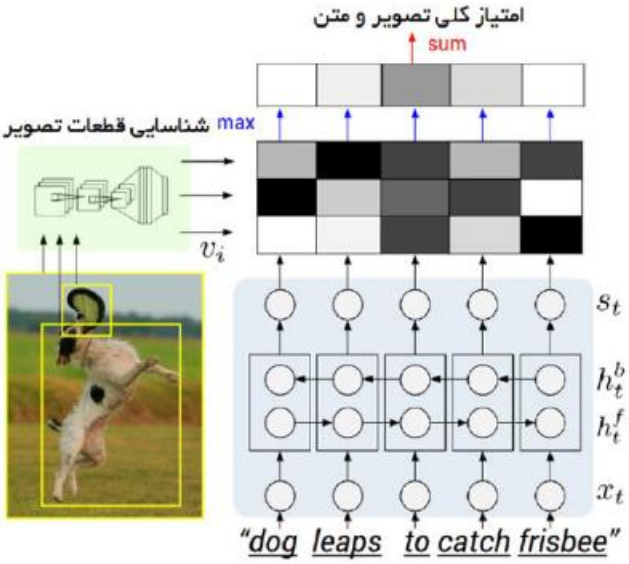
مدلی که در قسمت قبل بررسی شد [17] قابلیت‌های زایایی نداشت بلکه قادر بود بر اساس تصویر ورودی، بهترین جمله‌ی متناسب را شناسایی و بازیابی کند. در کاربردهای دنیای واقعی نمی‌توان مدلی را به مجموعه‌ای محدود و ایستا از جملات از پیش تعریف شده وابسته کرد، چراکه با فضایی نامحدود روبرو هستیم و مدلی محدود به سختی خواهد توانست ویژگی‌های متنوع آن را توصیف کند مگر با قید و بندهایی دست و پاگیر. برای نیل به این هدف نیاز به راهکاری است که قادر به توصیف تصاویر بدون تکیه بر پایگاهی از جملات از پیش نوشته باشد و این راهکار نیازمند مدل‌های تولیدی است که قادر باشند ساختار جملات در زبان طبیعی را فرا گرفته و به کمک آن، یک تصویر را توصیف کنند.

تحقیقات بسیاری برای تولید متن و جملات قابل‌قبول به زبان طبیعی انجام شده‌اند ولی در سال‌های اخیر، به لطف موفقیت روش‌های ترجمه‌ی ماشینی همچون [23] و [24] معماری‌های نوین و کارایی برای تولید دنباله‌ای از کلمات در ساختاری معنادار ارائه شده‌اند.

از مهم‌ترین نقاط ضعف روش [17] یکی مؤلفه‌ی زبانی ضعیف و دیگری نبود قابلیت تولید جملات به زبان طبیعی بود. در مقاله‌ی دیگری از همین محققان این نقاط ضعف به کمک یک شبکه‌ی عصبی بازگشتی برطرف شده است [25].

یکی از مشخصه‌های شبکه‌های عصبی غیربازگشتی (که شامل تمامی شبکه‌های مطرح شده در این گزارش است) قید و بندها و محدودیت‌های آن‌ها در مورد داده‌ی ورودی است. این شبکه‌ها یک بردار ورودی با اندازه‌ی ثابت را گرفته (مثلاً تصویر) و برداری خروجی با اندازه‌ی ثابت تحویل کاربر می‌دهند (مثلاً کلاس‌های دسته‌بندی). از طرفی هر داده‌ی ورودی، کاملاً مستقل از داده‌های قبل و بعد خود است. این موضوع در زمانی‌که در حال کار با تصاویری برای دسته‌بندی آن‌ها هستیم، مشکلی ایجاد نمی‌کند، ولی اگر این دسته‌بندی یا شناسایی به نحوی وابسته با داده‌های قبلی یا بعدی بود، این شبکه‌ها کارایی خود را از دست می‌دهند. به‌عنوان‌مثال اگر دسته‌ی تصاویری را به‌صورت پشت سر هم از یک فایل ویدئویی استخراج کنیم و بخواهیم عملی را که در هر تصویر رخ می‌دهد دسته‌بندی کنیم، هر دنباله‌ی تصویر بر دسته‌بندی دنباله‌ی بعدی خود تأثیر خواهد گذاشت. مثلاً اگر تصویری ویدئویی از حرکات یک دشمن در بازی ویدئویی داشته باشیم و بخواهیم پیش‌بینی کنیم که آیا در ثانیه‌های بعدی این دشمن آماده‌ی حمله می‌شود و در صورت حمله کدام روش را انتخاب می‌کند، تک‌تک تصاویری که از این شخصیت در ثانیه‌های قبلی دیده شده است، تأثیر مستقیمی بر پیش‌بینی ما در مورد حرکت بعدی آن خواهد داشت.

مهم‌ترین مشخصه‌ی شبکه‌های بازگشتی این است که به ما امکان کار با دنباله‌هایی از بردارها را می‌دهند: بردارهایی در ورودی، در خروجی و یا هر دو. یکی از کاربردهای این شبکه‌ها در تولید دنباله‌ای از نویسه‌ها، کلمات و جملات است که وابسته به شرایط و وضعیتی هستند که دنباله‌ی ورودی دارد. کلمات و جملاتی که در زبان‌های طبیعی به کار گرفته می‌شوند، یک دنباله محسوب می‌شوند و قرار گرفتن هر کلمه پس از کلمه‌ی فعلی، وابستگی شدیدی نه‌تنها به کلمه‌ی فعلی بلکه به کلمات قبل از آن نیز دارد. به‌عنوان‌مثال با داشتن دنباله‌ای همچون [H E L] انتظار بیشتری می‌رود که حرف بعدی این دنباله یکی از حروف [D L M P] باشد و درصورتی‌که دنباله‌ای از کلمات [Please Hel] را داشته باشیم، انتظار بیشتری داریم که حرف بعدی این دنباله [P] باشد تا دیگر حروف الفبا. پیش‌بینی این حروف که وابسته به دنباله‌ای قبل از خود هستند، از طریق شبکه‌های بازگشتی ممکن است. روشی که در [25] برای حل مسئله‌ی تولید جملات به زبان طبیعی ارائه شده است، همچون روش قبلی از تلفیق دو مؤلفه‌ی تصویری و متنی استفاده می‌کند با این تفاوت که این بار به جای به‌کارگیری درخت تجزیه برای تبدیل فضای متون، از یک شبکه‌ی بازگشتی دو طرفه برای یادگیری ساختار جملات و بردن آن‌ها به فضای برداری بهره می‌برد. به‌کارگیری این روش که راهکاری تکمیلی بر تحقیق قبلی این محققان [17] است نه‌تنها نقاط ضعف آن را برطرف می‌کند بلکه ازلحاظ سرعت پاسخگویی و دقت عملکرد نیز به آن برتری دارد. ساختار امتیازدهی بین قطعات تصویر و قطعات متن همچون قبل بر اساس ضرب داخلی بردارهای آن‌ها و جمع کل امتیازهاست. در شکل ۷ معماری این روش در محاسبه‌ی امتیاز بین قطعات تصاویر و متن نمایش داده شده است:



شکل (۷) : محاسبه‌ی امتیاز بین اشیاء درون تصویر و قطعات متن

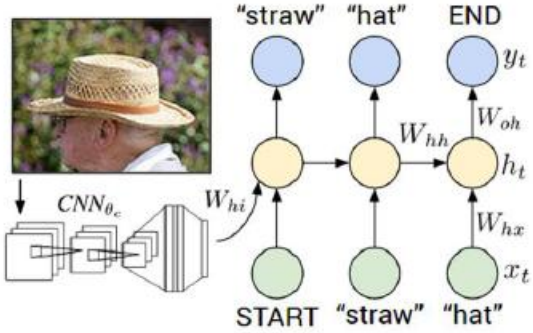
در این روش هر تصویر به همراه متن توصیفی وارد شبکه‌های عصبی (برای تصویر) و بازگشتی (برای متن) شده و امتیاز کلی بین آن‌ها محاسبه می‌شود. چون اولویت بر روی توصیف تصویر است، خطای بین توصیف اصلی و دیگر

توصیفات به‌طوری محاسبه می‌شود که به‌اندازه‌ی یک مرز مشخص بین تصویر اصلی و توصیفش با دیگر توصیفات فاصله بیفتد.

از طرفی به کمک یک میدان تصادفی مارکُفی قطعاتی از متن که ارتباط بیشتری با قطعات هر تصویر دارند، جداسازی شده و به آن اختصاص پیدا می‌کنند. علت استفاده از میدان‌های مارکُفی، در این مسئله است که تخصیص هر کلمه به ناحیه‌ای از تصویر که بیشترین امتیاز را در آن به دست آورده است، راهکاری نامناسب برای توزیع کلمات بر روی قطعات تصویر است چرا که موجب می‌شود کلمات بصورتی نامرتب و بعضاً نادرست به قطعات متفاوتی تخصیص داده شوند که در ادامه به عملکرد سیستم برای توصیف درست تصویر صدمه می‌زند. برای حل این مشکل، هم‌ترازی‌های واقعی بین قطعات تصویر و متن به عنوان متغیرهایی در یک میدان مارکُفی در نظر گرفته می‌شوند بطوریکه روابط دودویی بین کلمات همسایه، مشوّق یک هم‌ترازی برای قطعه‌ای یکسان از تصویر می‌شود. در این روش با تعریف یک تابع انرژی و کمینه کردن آن، به بهترین هم‌ترازی ممکن بین قطعات متن و تصویر می‌رسیم. خروجی این مرحله مجموعه‌ای از قطعات تصویر است که با قطعاتی از متن توصیف شده‌اند.

در این مرحله عملاً هم شبکه‌ی بازگشتی را طوری آموزش داده‌ایم تا بهترین جمله‌ها را برای تصویر آموزشی برگزیند و هم بتواند قطعات مناسبی از این جملات برای توصیف قطعات درون تصویر انتخاب کند.

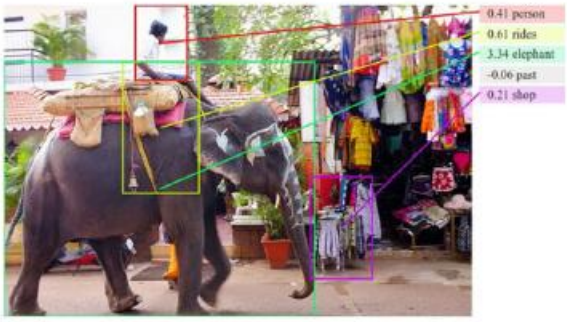
پس از اتمام آموزش، می‌توان از این معماری برای توصیف تصاویری که به‌عنوان ورودی به آن داده می‌شود استفاده کرد. ایده‌ی اصلی در این مرحله، پاس دادن اطلاعات هر قطعه تصویر به شبکه‌ی بازگشتی به‌عنوان یک ورودی به لایه‌ی پنهان است. در اینصورت شبکه در اولین تلاشش برای توصیف تصویر، به اطلاعات آن به‌عنوان یک ورودی به لایه‌ی پنهان دسترسی دارد و قادر خواهد بود دنباله‌ی کلمات را مستقیماً وابسته به این ورودی تصویری ایجاد کند. در شکل ۸ نمونه‌ای از این توصیف برای تصویری از یک «کلاه حصیری» نمایش داده شده است:



شکل (۸) : توصیف تصویری از یک «کلاه حصیری» به کمک شبکه‌ی بازگشتی

در این شکل شبکه‌ی بازگشتی با ورودی خاصی به نام START کار را شروع می‌کند. این ورودی عملاً نشانه‌ای برای شبکه است تا در اولین قدم نگاهی به تصویر داشته باشد. مشخصه‌های استخراج‌شده‌ی تصویر به‌عنوان یک ورودی از طریق W_{hi} وارد لایه‌ی پنهان شبکه‌ی بازگشتی می‌شود. درصورتی‌که شبکه به درستی آموزش دیده باشد، این ورودی تصویری موجب خروجی "straw" خواهد شد و این خروجی به‌عنوان یک ورودی جدید دوباره تحویل شبکه شده و این روند تا رسیدن به خروجی END ادامه پیدا می‌کند.

در شکل ۹ نمونه‌ای از عملکرد این روش قبل از به‌کارگیری میدان‌های تصادفی مارکُفی برای هم‌تراز کردن تصاویر و قطعات درون آن‌ها، نمایش داده شده است:



شکل (۹) : هم‌ترازی بین متن و قطعات درون تصویر

برخلاف آنچه که در [17] یا [22] دیدیم، وضعیت هر قطعه متن و هر کلمه‌ی بازیابی شده، تأثیر مستقیمی بر انتخاب دیگر کلمات دارد. به‌عنوان‌مثال در تصویر چپ یک فیل در حال گذر کردن از «کنار» یک مغازه است و شبکه‌ی بازگشتی قادر به انتخاب کلمه‌ی "past" برای این اتفاق شده است و نه کلماتی جایگزین همچون "towards" یا "from" که ازلحاظ گرامری در این جمله صدق می‌کنند. در شکل ۱۰ نیز عملکرد روش یادشده در اتصال اطلاعات به هم برای تولید یک توصیف کلی از تصویر، نمایش داده شده است:

شکل (۱۰) : توصیف تصاویر به زبان طبیعی [25]

روش یادشده به خوبی قادر به توصیف تصاویر به زبان طبیعی است، بااین‌حال نباید از این حقیقت غافل شد که بخش اعظمی از توصیفات در داده‌های آموزشی وجود داشته‌اند. به‌عنوان‌مثال در مجموعه داده‌ی آموزشی Flickr30K که شامل ۳۰ هزار تصویر به همراه توصیف‌های آن‌هاست، جمله‌ی "two young girls are playing with lego" وجود دارد و شبکه‌ی بازگشتی این جمله را مستقیماً برای تصویری تستی (در شکل نیامده است) انتخاب کرده، بدین معنی که هیچ فرایند تولیدی صورت نگرفته است. این موضوع شاید در توصیف تصویری از دو دختر که در حال بازی با لگو هستند، مفید به نظر بیاید ولی نشان از ترجیح شبکه به حفظ جملات تا یادگیری دقیق ساختار آن‌ها دارد. البته این موضوع در تمامی تصاویر صادق نیست، به‌عنوان‌مثال جمله‌ی "man in black shirt is playing guitar" در هیچ کجای داده‌های آموزشی وجود ندارد و شبکه خود قادر به تولید این جمله بر اساس محتوای تصویر شده است. البته دو جمله‌ی "man

"in black shirt" و "is playing guitar" به‌طور میانگین ۴۰ بار در داده‌های آموزشی وجود دارند ولی این موضوع که شبکه قادر بوده به کمک این دو قطعه، یک متن توصیفی مناسب تولید کند، نشان از حرکت به سمت این هدف دارد.

نقاط ضعف این راهکار را می‌توان در ساختار معماری‌های شبکه‌ی عصبی آن‌ها جستجو کرد. شبکه عصبی R-CNN فقط قادر به قبول تصاویری با اندازه‌ی ثابت است بدین معنی که برای استفاده از این روش باید تمامی تصاویر ورودی را به‌اندازه‌ی قابل‌قبول توسط شبکه ببریم که این خود تأثیر مستقیمی بر کارایی مدل در پیش‌بینی‌های زمان‌واقعی و زنده دارد. همچنین با توجه به قدیمی شدن شبکه‌ی AlexNet برای استخراج ویژگی‌های تصویر، استفاده از آن موجب استخراج ویژگی‌هایی با دقت کمتر می‌شود. از طرفی تنها راه دسترسی شبکه‌ی بازگشتی به داده‌های تصویری، اضافه کردن این اطلاعات از طریق یک بایاس اولیه به لایه‌ی نهان شبکه است؛ نشان داده شده که این راهکار قدرت تشریحی ضعیف‌تری نسبت به روش‌های دیگر در ارسال وضعیت به شبکه‌های بازگشتی دارد [26]. درنهایت، این روش از یک معماری دو مدله استفاده می‌کند که به صورتی مجزا از هم عمل می‌کنند. حل این مسئله به کمک ساختاری یکپارچه و تک مدله، هنوز یک مسئله‌ی باز و حل‌نشده است.

در جدول ۲-۱، خلاصه‌ای از امتیازات روش‌های متفاوت برای حل این مسئله آمده است (امتیازات B-x مربوط به شاخص BLEU در حوزه‌ی ترجمه‌ی ماشینی است):

جدول (۲) : عملکرد روش‌های مختلف در توصیف تصاویر FlickrXK

	Flickr8K				Flickr30K			
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
Nearest Neighbor	—	—	—	—	—	—	—	—
Mao et al.	58	28	23	—	55	24	20	—
Google NIC	63	41	27	—	66.3	42.3	27.7	18.3
LRCN	—	—	—	—	58.8	39.1	25.1	16.5
MS Research	—	—	—	—	—	—	—	—
Chen and Zitnick	—	—	—	14.1	—	—	—	12.6
Karpathy et al.	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7

۳-۳- دیگر روش‌ها

در حال حاضر دو رویکرد اصلی در حوزه‌ی روش‌های غیرایستا وجود دارد. در رویکرد اول، در ابتدا از شبکه‌ای ژرف برای پیش‌بینی یک سبد کلمه‌ی تصویری استفاده می‌شود. سپس و در مرحله‌ی بعد از یک مدل زبانی بیشینه آنتروپی (ME LM) برای تولید یک متن که کمینه‌ای از قطعات تصویر را پوشش می‌دهد، استفاده می‌شود. این رویکرد همانند آن چیزی است که در قسمت قبل و در [17] آن را بررسی کردیم. رویکرد دوم از خروجی شبکه‌ای ژرف (برای تصویر) به‌عنوان ورودی شبکه‌ای بازگشتی (برای متن) استفاده می‌شود که به آن «شبکه‌ی عصبی چندحالته» نیز گفته می‌شود که آن را نیز در قسمت قبل و در [25] و [4] بررسی کردیم.

به لطف مجموعه داده‌های حجیمی همچون Flickr30K یا Microsoft COCO و با رشد تحقیقات در حوزه‌ی شبکه‌های عصبی و تلفیق آن‌ها با هم، تحقیقات قابل‌توجهی از سال ۲۰۱۳ در این حوزه شکل گرفته است. اکثر روش‌ها و تحقیقات این حوزه، در چالش Microsoft COCO [5] شرکت کرده‌اند و می‌توان دید که در حال حاضر ۱۵ تحقیق با کارایی‌های بعضاً مشابه در این زمینه ارائه شده‌اند. در تحقیقی از [27] نقاط کلی قوت و ضعف این دو رویکرد مقایسه شده‌اند. روش‌های مبتنی بر شبکه‌های بازگشتی عملکرد بالایی در تولید جملات و توصیف تصاویر با تکیه بر داده‌های آموزشی دارند. در این روش‌ها، مدل زبانی ارتباط مستقیمی با طراحی شبکه‌ی بازگشتی دارد و نه طراحی مؤلفه‌ای مبتنی بر مشخصه‌های موجود در زبان‌های طبیعی. هر چقدر داده‌های بیشتری برای آموزش شبکه در نظر گرفته شده و حافظه‌ی بیشتری برای به خاطر سپردن کلمات قبلی به آن تخصیص داده شود، مدل زبانی به دست آمده در این شبکه‌ها با دقت بیشتری عمل می‌کند [4]. بااین‌حال این دقت بالا در شبکه‌های بازگشتی در بسیاری از اوقات به خاطر به‌کارگیری مستقیم جملات توصیفی در داده‌های آموزشی است و نه تولید جملات جدید. از طرفی با اینکه دقت روش‌های مبتنی بر مدل‌های زبانی از شبکه‌های بازگشتی کمتر است، ولی در مواردی این روش‌ها کارایی بهتری در تولید جملات از به هم چسباندن تکه‌های متنی از خود نشان می‌دهند [27]. ازلحاظ سرعت بازیابی و پاسخ‌دهی، روش‌هایی که کاملاً به شبکه‌های عصبی متکی هستند، بیشترین سرعت را از خود نشان می‌دهند که این موضوع مرهون زمان زیادی است که برای آموزش این شبکه‌ها صرف شده است. از طرفی در روش‌های مبتنی بر نزدیک‌ترین همسایه [28] و روش‌هایی که نیاز به دسترسی به پایگاه داده دارند [29] شاهد تأخیر در پردازش و ایجاد خروجی هستیم که استفاده از آن‌ها را چه در موتورهای جستجو و سیستم‌های بازیابی تصویر و چه در کاربردهای زمان‌واقعی با مشکل مواجه می‌کند.

به‌غیراز رویکرد مبتنی بر شبکه‌های عصبی ژرف و شبکه‌های بازگشتی، راهکارهای بعضاً متفاوتی با این دو رویکرد نیز در تحقیقات اخیر اتخاذ شده‌اند. در روشی از [30] تلاش شده معماری ارائه شده توسط دو مقاله‌ی فوق را بهبود بخشند و با اضافه کردن مؤلفه‌ای برای تشخیص اهمیت و برجستگی نقاط خاصی از تصویر، دقت توصیف را افزایش دهند. این روش نیز قادر به برتری بر Google NIC نبوده و با وجود تلاش برای بهبود آن، در رتبه‌ی سوم قرار گرفته است. بااین‌حال نکته‌ی قابل‌توجه در این روش، پیدا کردن اجزای مهم در تصویر است حتی بدون اینکه توسط شبکه‌ی ژرف شناسایی شده باشند. این موضوع باعث افزایش قابلیت مدل در توصیف تصاویری می‌شود که مؤلفه‌ی تصویری قادر به شناسایی بخشی از اجزای آن نیست. نمونه‌ای از عملکرد این مدل در شکل ۱۱ نمایش داده شده است:

شکل (۱۱) : عملکرد روش [31] در توصیف تصاویر

دو تصویر بالا و وسط به صورتی درست توصیف شده‌اند و در آن‌ها قسمتی از تصویر که بیشترین برجستگی و اهمیت را به خود اختصاص داده نمایش داده شده است. تصویر پایین، تصویری است که به صورتی نادرست توصیف شده است و یکی از نکات مثبت تحقیق یادشده در این است که می‌تواند به کمک استفاده از میزان برجستگی، نشان دهد کدام قسمت از تصویر موجب توصیف نادرست آن شده است. در مثال فوق، شناسایی نادرستِ طراحی دایره‌ای و ساعت‌واری که بر روی لباس زن وجود دارد، موجب اشتباه در توصیف تصویر شده است. بااین‌وجود، این قابلیت آن‌چنان‌که باید و شاید برتری خاصی برای این تحقیق محسوب نمی‌شود چراکه بسیاری از روش‌های دیگر نیز تلاش داشته‌اند مفهوم برجستگی و هم‌ترازی بین قطعات متن و تصویر را بر اساس ساختار و معماری خود مدل کنند که در قسمت قبل و در بررسی روش [17] به آن اشاره کردیم.

در رویکردی دیگر که به عملکرد قابل‌قبولی در بین روش‌های توصیف تصویر رسیده است، صرفاً از گستردگی داده‌های آموزشی Microsoft COCO (که شامل ۱۰۰ هزار تصویر و توصیف است) برای انتخاب نزدیک‌ترین توصیف به تصویر استفاده شده است [28]. این رویکرد که از روش‌های نزدیک‌ترین همسایه برای این کار استفاده می‌کند، فرض را بر این قرار داده که در صورت گستردگی مجموعه داده‌ی آموزشی، باید بتوان بدون نیاز به تولید متن، از جملاتی که در داده‌های آموزشی قرار دارند، برای توصیف تصاویر استفاده کرد. این روش برخلاف دیگر روش‌ها از معماری‌های پیچیده و زیرساخت‌های گسترده برای هم‌ترازی متن و تصویر استفاده نکرده است و بااین‌حال قادر به دستیابی به امتیازی قابل‌قبول نیز بوده است. بااین‌حال این موضوع به معنی قدرت روش یاد شده در تولید متن نیست بلکه این عملکرد با تکیه‌ی صرف به حجم عظیم داده‌ها و «حفظ» کردن جملات توصیفی ممکن شده است و طبیعی است که می‌توان ضعف این روش را در مواجه شدن با تصاویری خارج از دامنه‌ی معنایی تصاویر Microsoft COCO انتظار داشت.

در تحقیقی از [27] عملکرد تقریباً یکسان شبکه‌های بازگشتی و روش مبتنی بر نزدیک‌ترین همسایه بررسی شده است. در این تحقیق نشان داده شده که با وجود برتری روش‌های مبتنی بر شبکه‌های بازگشتی، عملکرد آن‌ها در توصیف تصاویر و بازیابی آن‌ها، بسیار شبیه به روش‌های مبتنی بر نزدیک‌ترین همسایه است چرا که این شبکه‌ها بیش از اینکه تلاش کنند جمله‌ای جدید و نادیده را تولید کنند، تمایل به استفاده‌ی دوباره از جملات موجود در داده‌های آموزشی دارند و این رویکرد تفاوت چندانی با تکیه بر نزدیک‌ترین همسایه برای پیدا کردن نزدیک‌ترین جمله‌ی توصیفی از داده‌های آموزشی برای تصویر ندارد. در این تحقیق تلاش شده با یک مدل زبانی مناسب، جملاتی نادیده که در داده‌های آموزشی وجود نداشتند برای توصیف تصاویر تولید شود. با وجود ادعای تحقیق یاد شده در برتری مدل زبانی نسبت به شبکه‌های بازگشتی، این روش در عمل از شبکه‌ی بازگشتی [23] برای یادگیری ارتباط بین متن و تصویر استفاده می‌کند که شبکه‌ای عام‌منظوره بوده و در اصل برای مقاصد ترجمه‌ی ماشینی طراحی شده است. در روشی شبیه به این روش که از شبکه‌ی بازگشتی قدرتمندتر (و پیچیده‌تری) استفاده می‌کند [4]، برتری مدل زبانی یاد شده، از بین می‌رود.

در بین روش‌های ارائه شده در MS COCO 2015 یک روش با وجود سادگی به دقت بیشتری نسبت به دقت انسانی برای شاخص M4 (میانگینِ مقدارِ جزئیات در هر توصیف) رسیده که در هیچ یک از روش‌های دیگر رخ نداده است [29]. در این روش همانند روش‌های قبلی از یک شبکه‌ی ژرف برای استخراج ویژگی‌های تصویر استفاده شده و سپس از فاصله‌ی کُسینوسی برای بازیابی نزدیکترین تصاویر به این تصویر از یک بانک اطلاعاتی استفاده می‌شود. ۱۰ تصویر بازیابی شده هر کدام دارای ۵ توصیف مختص به خود هستند که در نهایت برابر ۵۰ توصیف بازیابی شده می‌شوند. در ادامه شروع به حذف جملاتی می‌کنیم که دارای کلماتی با بیشترین تعداد استفاده در بین ۵۰ جمله نیستند. این فرایند با انتخاب کلمه‌ای با بیشترین تعداد رخداد در جمله شروع می‌شود و اگر کلمه‌ی یاد شده در دیگر جملات حضور نداشته باشد، جمله رد می‌شود. برای جلوگیری از تأثیر کلمات پراستفاده بر این فرایند، از ۱۰۰ کلمه‌ای که طبق Google n-grams بیشترین استفاده را در زبان انگلیسی دارند [32]، چشم‌پوشی می‌شود. با بررسی این روش متوجه می‌شویم که علت برتری آن نسبت به شاخص انسانی نه به خاطر دقت بالای آن در تولید جملات و درک تصاویر بلکه به خاطر استفاده‌ی مستقیم از جملات توصیفی موجود در داده‌های آموزشی است. این به‌کارگیری داده‌های آموزشی تا حدی مستقیم است که حتی همچون روش‌هایی مثل [27] و [28] تلاشی برای تولید متن جدید نیز صورت نگرفته است و محققان فقط به تنوع داده‌های آموزشی و توصیفات آن‌ها امید بسته‌اند. طبیعی است چنین رویکردی نه تنها مشکلات روش‌های مبتنی بر نزدیک‌ترین همسایه را دارد، بلکه زمان و حجم محاسبات زمان‌واقعی آن نیز به خاطر سربارِی که در هر بار دسترسی به پایگاه داده‌ی تصویری و محاسبه‌ی فاصله‌ی کُسینوسی با ۱۰ تصویر دارد، بالاتر از دیگر روش‌های یاد شده در این گزارش است.

٤- شکاف‌های تحقیقاتی

این محدودیت‌ها را می‌توان در موارد زیر دسته‌بندی کرد:

الف. ضعف کلی مدل‌ها: با اینکه در حوزه‌ی شناسایی اشیاء درون تصاویر به دقتی نزدیک به انسان [8] و در مواردی حتی بهتر از آن [33] رسیده‌ایم، ولی

در حال حاضر مدل‌های ارائه شده برای توصیف تصاویر و بازیابی معنایی آن‌ها، تفاوتی حدود ۳۵ تا ۴۰ درصد نسبت به دقت انسانی دارند. در اکثر روش‌های ارائه شده که ساختاری دو مؤلفه‌ای دارند، نیاز به تلفیق دو دامنه‌ی تصویری و متنی است که هر روشی رویکرد خود را برای تلفیق این دو فضا اتخاذ کرده است و نقطه‌ی ضعف این مدل‌ها را می‌توان در این روش تلفیقی، طراحی مدل زبانی و خروجی تصویر (از شبکه‌ی ژرف) جستجو کرد. به‌طورکلی می‌توان گفت که قدرت این مدل‌ها، ارتباط مستقیمی با داده‌های آموزشی محدودی دارد که بر روی آن‌ها آموزش می‌بینند. بدین ترتیب طبیعی است که انتظار داشت خروجی یک مدل بر روی داده‌های Pascal1K (با هزار داده)، Flickr8K (با ۸ هزار داده) و Flickr30K (با ۳۰ هزار داده) متفاوت باشد و هرچقدر تعداد و کیفیت این داده‌های آموزشی بالاتر رود، قابلیت‌های تولیدی مدل‌ها نیز افزایش یابند. از طرفی مدل‌های ارائه شده هنوز دارای نواقص مختلفی هستند و در بسیاری از موارد ارجحیت را بر حفظ کردن داده‌های آموزشی می‌گذارند تا تولید متنی جدید و با قابلیت توصیف بالا. به‌عنوان‌مثال حتی در Google NIC (رتبه‌ی برتر COCO 2015) نیز گزارش شده است که حدود ۸۰ درصد توصیفات بر روی تصاویر انتخابی، در داده‌های آموزشی به‌طور کامل و یکپارچه وجود داشته‌اند که عملاً کار مدل در اینجا بیشتر شبیه به پیدا کردن نزدیک‌ترین جمله‌ی توصیفی است تا تولید متنی جدید و مختص به تصویر (Vinyals et al. 2015).

ب. ضعف مدل‌های تصویری: مدل‌های تصویری فعلی، بر مبنای داده‌های ImageNet آموزش داده شده‌اند و این موضوع خود به معنای ضعف تمامی آن‌ها در توصیف تصاویری خارج از دامنه‌ی معنایی ImageNet است و به‌خاطر این موضوع شاهد خروجی‌هایی هستیم که قادر نیستند بیش از ۱۰۰۰ کلاس را دسته‌بندی کند. همچنین لازم به ذکر است که دسته تصاویر ImageNet برای آزمودن قابلیت روش‌های شناسایی اشیاء انتخاب شده‌اند و به‌عنوان‌مثال در آن شاهد دسته‌هایی از انواع و اقسام نژاد حیواناتی همچون سگ و گربه هستیم و این رویکرد ImageNet موجب ضعف مدل‌های شکل گرفته بر مبنای آن می‌شود. از طرفی ساختار ایستای شبکه‌های ژرفی که در این مدل‌ها به کار رفته‌اند، موجب نیاز به انجام پیش پردازش‌هایی بر روی داده‌های تصویری می‌شود که خود زمان پاسخ‌دهی سیستم را بالا می‌برد. همچنین مشکلاتی که در شناسایی اشیاء در این شبکه‌ها وجود دارد (همچون ضعف در حذف حداکثری) در بسیاری از اوقات موجب شناسایی نادرست و یا شناسایی چندین باره‌ی اشیاء می‌شود که این اشتباه تأثیر خود را به صورتی زنجیره‌وار بر فرایند توصیف تصویر و بازیابی آن می‌گذارد.

پ. ضعف مدل‌های زبانی: ضعف مدل‌های زبانی بیش از مدل‌های تصویری بر عملکرد سیستم‌های بازیابی تأثیر دارد بخصوص با در نظر گرفتن این موضوع که کاربران چنین سیستم‌هایی غالباً از طریق رابط زبان طبیعی با آن‌ها ارتباط برقرار می‌کنند تا دیگر رابط‌ها همچون رابط‌های تصویری. مسئله‌ی شکستن یک متن به قطعاتی هدفمند و هم‌ترازی این قطعات با تکه‌هایی از تصویر مسئله‌ای حل نشده در این حوزه محسوب می‌شود. درخت تجزیه‌ای که در [17] برای قطعه‌بندی متن استفاده شده بود، کارایی و درک لازم را برای شکستن یک متن به قطعاتی معنایی ندارد و در مواردی حتی قطعاتی تولید می‌کند که موجب اشتباه سیستم در هم‌ترازی متن و تصویر می‌شود. از طرفی نبود درکی معنایی از استقلال اشیاء درون تصویر و روابط فضایی آن‌ها موجب می‌شود که مدل قادر به جداسازی این مفاهیم از یکدیگر و شمردن آن‌ها نباشد و در اکثر تصاویری که نیاز به توصیف چندین عنصر متفاوت ولی یکسان (همچون دو گربه یا دو کودک) دارند، مدل یا اشیاء را یک شیء یکسان در نظر گرفته و یا آن‌ها را بصورت جمعی خطاب می‌کند (به‌عنوان‌مثال «کودکان در حال بازی هستند»). دراین‌بین طراحی شبکه‌های بازگشتی یا مدل‌زبانی مناسبی که قادر به توصیف تصاویر بصورتی مناسب باشد، هنوز مسئله‌ای باز در این حوزه است و عملکرد فعلی این مدل‌ها نسبت به عملکرد انسانی قابل قبول نبوده و نیاز به تحقیقات بیشتری دارد.

ت. در نظر نگرفتن داده‌های مکمل: اکثر روش‌های یاد شده برای توصیف تصاویر و بازیابی آن‌ها، ساختاری دو مؤلفه‌ای دارند و هسته‌ی اصلی این سیستم‌ها، روشی تلفیقی است که بین متون و تصاویر ارتباط برقرار می‌کند [27]. رویکرد غالبی که در این روش‌ها مشهود است، نگاه به بازیابی معنایی تصویر به عنوان یک مسئله‌ی هم‌ترازی متن و اشیای درون تصویر است. به‌غیر از ویژگی‌های استخراج شده از شبکه‌های ژرفی که برای شناسایی اشیاء در تصویر آموزش دیده‌اند، هیچ‌گونه اطلاعات دیگری از تصویر به مؤلفه‌ی متنی/تلفیقی ارسال نمی‌شود. اطلاعاتی همچون نوع مکان، ساختار هنری تصویربرداری، قطعه‌های معنایی اشیاء (شامل تمامی پیکسل‌های تشکیل دهنده‌ی هر شیء) یا عمق نسبی اشیای درون تصویر در مدل‌های موجود در این حوزه بکار گرفته نشده‌اند. این موضوع به ساختار شبکه‌های ژرفی برمی‌گردد که این روش‌ها از آن‌ها برای استخراج ویژگی‌های تصاویر استفاده می‌کنند. در حال حاضر طراحی شبکه‌ای که قادر به استخراج چندین ویژگی همچون نوع تصویر و عمق نسبی آن و قطعه‌بندی اشیای درون تصویر باشد، مسئله‌ای باز و حل نشده محسوب می‌شود و همچنین ازآنجاکه این روش‌ها نیاز به مدل‌هایی موفق برای استخراج ویژگی‌های تصویر دارند، به‌ناچار مجبور به به‌کارگیری شبکه‌های موفق این حوزه در سالهای اخیر شده‌اند که همگی بر روی داده‌های ImageNet آموزش دیده و برای وظایف تعیین شده در آن طراحی شده‌اند. این تکیه بر شبکه‌های فوق و تلاش برای ساده نگه‌داشتن معماری روش بازیابی معنایی، سبب شده است که داده‌های مکملی که می‌توان از تصویر برای فرایند توصیف استخراج کرد، نقشی در این روش‌ها نداشته باشند.

۵- نتیجه

در این تحقیق به تعریف مسئله‌ی بازیابی معنایی تصاویر پرداختیم. این مبحث نیازمند طرح مسئله و تشریح اهمیت کاربردی آن به عنوان یک حوزه‌ی تحقیقاتی مهم و تأثیرگذار در مباحث پردازش تصویر و بینایی ماشین است. در قسمت اول به بررسی مفاهیم و مباحث موجود در این حوزه پرداخته و نقاط قابل‌تأمل در آن را مطرح کردیم، سپس به معرفی دو نگرش اصلی در مواجهه با چنین مسائلی پرداختیم و روش‌های منتخبی از این دو نگرش را مطرح کردیم. در بخش بعدی وارد جزئیات کار این روش‌ها شده و سعی کردیم مهم‌ترین و تأثیرگذارترین تحقیق‌های چند سال اخیر در این حوزه را بررسی کنیم. در ادامه و در بخش انتهایی شکاف‌های تحقیقاتی و موارد تأثیرگذاری را که توسط این تحقیقات به‌طور عمیق کنکاش نشده بودند را حول ۴ محور اصلی بررسی کردیم.

مراجع

[1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *J. Vis.*, vol. 7, no. 1, p. 10, Jan. 2007.

[2] J. Eakins, M. Graham, and C. I. Retrieval, "University of Northumbria at Newcastle," *Content-based Image Retr.*, 1999.

[3] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 20–54, 2015.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07–12–June, no. 4, pp. 3156–3164.

[5] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[6] R. Mihalcea, "The multidisciplinary facets of research on humour," *Appl. Fuzzy Sets Theory*, pp. 412–421, 2007.

[7] A. Chandrasekaran, A. K. Vijayakumar, S. Antol, M. Bansal, D. Batra, C. L. Zitnick, and D. Parikh, "We Are Humor Beings: Understanding and Predicting Visual Humor," Dec. 2015.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[10] R. Gerber and N. H. Nagel, "Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences," *Proc. Int. Conf. Image Process.*, pp. 805–808, 1996.

[11] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. Zhu, "I2T : Image Parsing to Text Description," *Proc. IEEE*, vol. 98, no. 8, pp. 1–21, 2010.

[12] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6314 LNCS, no. PART 4, pp. 15–29, 2010.

[13] G. Kulkarni, V. Premraj, V. Ordonez, S. Member, S. Dhar, S. Li, S. Member, and Y. Choi, "BabyTalk : Understanding and Generating Simple Image Descriptions," vol. 35, no. 12, pp. 2891–2903, 2013.

[14] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, A. Mensch, A. Berg, X. Han, T. Berg, and O. Health, "Midge: Generating Image Descriptions From Computer Vision Detections," *Eacl*, pp. 747–756, 2012.

[15] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," *Acl-2010*, no. July, pp. 1250–1258, 2010.

[16] P. Kuznetsova, V. Ordonez, and A. Berg, "Collective generation of natural image descriptions," *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, no. July, pp. 359–368, 2012.

[17] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," in *Proceedings of NIPS 2014*, 2014, pp. 1–9.

[18] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 966–973, 2010.

[19] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. April, pp. 67–78, 2014.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Nips*, pp. 1–10, 2015.

[21] A. Frome, G. Corrado, and J. Shlens, "Devise: A deep visual-semantic embedding model," *Adv. Neural ...*, pp. 1–11, 2013.

[22] R. Socher, A. Karpathy, Q. V Le, C. D. Manning, and A. Y. Ng, "Grounded Compositional Semantics for Finding and Describing Images with Sentences," *Tacl*, vol. 2, no. April, pp. 207–218, 2014.

[23] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1724–1734, 2014.

[24] I. Sutskever, O. Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," *Nips*, pp. 1–9, 2014.

[25] A. Karpathy and F. F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07–12–June, pp. 3128–3137.

[26] J. Martens, "Generating Text with Recurrent Neural Networks," *Neural Networks*, vol. 131, no. 1, pp. 1017–1024, 2011.

[27] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language Models for Image Captioning : The Quirks and What Works," *Acl-2015*, no. Me Lm, pp. 100–105, 2015.

[28] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning," *arXiv Prepr.*, pp. 1–6, 2015.

[29] M. Kolář, M. Hradiš, and P. Zemčík, "Technical Report: Image Captioning with Semantically Similar Images," p. 3, 2015.

[30] K. Xu, J. L. B. R. Kiros, K. C. A. Courville, and R. S. R. S. Z. Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, Feb. 2015.

[31] J. Mao, W. Xu, Y. Yang, J. Wang, B. Research, and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," *To Appear ICLR-2015*, vol. 1090, no. 2014, pp. 1–14, 2015.

[32] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden,

"Quantitative analysis of culture using millions of digitized books.," *Science*, vol. 331, no. 6014, pp. 176–82, 2011.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015.