



تولید خودکار شرح بر تصاویر با استفاده از شبکه‌های عصبی کانولوشنی عمیق و بازگشته

Automatic Image Captioning Using Deep Convolutional and Recurrent Neural Networks

گزارش کارهای پیشین

استاد راهنما

دکتر صفابخش

پژوهش گر

احمد اسدی

۹۴۱۳۱۰۹۱

فروردین ماه ۱۳۹۶

چکیده

با توجه به افزایش چشمگیر تعداد تصاویر مورد استفاده کاربران در فضاهای مجازی و همین‌طور با در نظر گرفتن گرایش روزافزون کاربران به ذخیره‌سازی تصاویر در رایانه‌های شخصی، مساله مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه تصاویر موجود، به یکی از مسائل مهم و پرکاربرد در زمینه بینایی ماشین تبدیل شده است. گام اساسی در این راستا، دستیابی به سامانه‌ای است که قادر به تولید خودکار شرح برای تصاویر باشد. شرح این تصاویر که در قالب جملات زبان طبیعی ارائه می‌شود باید علاوه بر سازگاری با موضوع تصویر و توصیف صحیح صحنه، به لحاظ دستور زبان و معنا صحیح و کامل باشد.

فهرست مطالب

۱	فصل اول مقدمات
۲	۱-۱ مقدمه
۳	۲-۱ تعریف مساله
۴	۳-۱ درک صحنه
۵	۱-۳-۱ پژوهش‌های انجام شده در زمینه درک صحنه توسط مغز انسان
۷	۲-۳-۱ نتایج به دست آمده از آزمایشات
۹	۴-۱ جمع‌بندی
۱۱	فصل دوم درک صحنه
۱۲	۱-۲ درک صحنه
۱۲	۲-۲ روش‌های مختلف موجود
۱۳	۳-۲ روش‌های مبتنی بر مدل‌های گرافی احتمالی
۱۳	۱-۳-۲ استفاده از مدل میدان تصادفی مارکف
۱۶	۲-۳-۲ استفاده از مدل میدان تصادفی شرطی
۱۸	۳-۳-۲ استفاده از سایر مدل‌های گرافی احتمالی
۲۵	۴-۲ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق
۲۵	۱-۴-۲ اختصاص معنا به قطعه‌های مختلف تصویر
۲۶	۲-۴-۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دو طرفه جملات و تصاویر
۳۱	۵-۲ جمع‌بندی
۳۵	فصل سوم تولید جمله
۳۶	۱-۳ تولید جمله
۳۶	۲-۳ کاربردها
۳۷	۳-۳ روش تولید زبان طبیعی
۳۸	۴-۳ روش نزدیک‌ترین همسایه
۴۲	۵-۳ استفاده از قالب‌های آماده زبانی
۴۴	۶-۳ روش‌های مبتنی بر شبکه‌های عصبی بازگشتی
۴۶	۱-۶-۳ شبکه عصبی بازگشتی ضربی
۴۸	۷-۳ جمع‌بندی
۵۰	فصل چهارم تولید شرح متناظر صحنه با استفاده از یادگیری عمیق
۵۱	۱-۴ مقدمه
۵۲	۲-۴ تولید جمله با مفهوم مشخص

۳-۴	مدل دوطرفه نگاشت تصاویر و جملات مبتنی بر یادگیری عمیق	۵۷
۴-۳-۴	مدل زبانی مبتنی بر شبکه عصبی بازگشته	۵۷
۴-۳-۴	مدل دوطرفه نگاشت تصاویر و جملات با استفاده از شبکه عصبی بازگشته	۵۸
۴-۴	جمع‌بندی	۶۰
۵	فصل پنجم تولید شرح متناظر صحنه با استفاده از روش‌های مبتنی بر توجه بصری	۶۳
۵-۱	تولید شرح بر تصاویر با استفاده از روش‌های مبتنی بر توجه بصری	۶۴
۵-۲	روش‌های مبتنی بر توجه بصری در حوزه ترجمه ماشینی	۶۴
۵-۳	انکودر	۶۵
۵-۴	دیکودر	۶۶
۵-۵	ایده اصلی استفاده از توجه بصری	۶۶
۵-۶	دیکودر در روش مبتنی بر توجه بصری	۶۷
۵-۷	انکودر در روش مبتنی بر توجه بصری	۶۸
۵-۸	روش‌های مبتنی بر توجه بصری در حوزه تولید شرح متناظر تصویر	۶۹
۵-۹	تولید شرح متناظر تصویر با استفاده از توجه بصری و شبکه‌های عصبی	۶۹
۵-۱۰	فعالیت‌های مشابه دیگر	۷۴
۵-۱۱	جمع‌بندی	۷۷
۶	فصل ششم حافظه فعال و مقایسه آن با مدل‌های مبتنی بر توجه بصری	۸۱
۶-۱	حافظه فعال	۸۲
۶-۲	واحد بازگشته گیت‌دار	۸۲
۶-۳	شبکه GPU	۸۳
۶-۴	استفاده از حافظه فعال در ترجمه ماشینی	۸۴
۶-۵	۱-۴-۶ نسخه مارکفی شبکه GPU	۸۴
۶-۶	۲-۴-۶ نسخه توسعه‌یافته شبکه GPU	۸۵
۶-۷	بررسی عملکرد ایده حافظه فعال در حوزه ترجمه ماشینی	۸۷
۶-۸	جمع‌بندی	۸۹
۷	فصل هفتم جمع‌بندی و نتیجه‌گیری	۹۰
۷-۱	مقدمه	۹۱
۷-۲	درک صحنه	۹۲
۷-۳	تولید جمله	۹۴
۷-۴	یادگیری عمیق	۹۶
۷-۵	توجه بصری	۹۷
۷-۶	حافظه فعال	۹۹

فهرست تصاویر

۶	نمونه توصیف‌های افراد برای تصاویر	۱
۶	ساختار مطلوب اطلاعات استخراج شده از تصاویر [۱]	۲
۷	تصاویر دنیای واقعی مورد استفاده در آزمایشات [۱]	۳
۸	نمودار مقایسه‌ای عملکرد مغز در درک صحنه	۴
۹	نمونه‌ای از نتایج بهدست‌آمده از آزمایشات [۱]	۵
۱۴	نگاشت تصویر به فضای معنایی	۶
۱۵	مدل میدان تصادفی مارکف در درک صحنه	۷
۱۷	مدل سلسله‌مراتبی میدان تصادفی شرطی در درک صحنه	۸
۱۸	مدل گرافی احتمالی مورد استفاده در پژوهش [۲]	۹
۲۲	نمونه تصاویر موجود در مجموعه‌داده مورد استفاده [۲]	۱۰
۲۳	ماتریس درهم‌ریختگی مدل کامل ارائه شده در [۲]	۱۱
۲۳	نتیجه مقایسه مدل‌های مختلف در [۲]	۱۲
۲۴	نتایج نهایی بهدست آمده از مدل بر روی تصاویر. [۲]	۱۳
۲۷	طرح‌واره عملکرد روش RCNN	۱۴
۲۹	نتایج عملکرد اهداف تعریف شده در روش RCNN برای همترازسازی تصاویر و جملات	۱۵
۳۰	نتایج نهایی روش RCNN	۱۶
۳۱	نتایج حاصل از جستجوی جملات در روش RCNN	۱۷
۳۹	نتایج کیفی اختصاص جملات و تصاویر به یکدیگر با استفاده از روش نزدیک‌ترین همسایه [۳]	۱۸
۴۱	نتایج برنامه‌سازی خطی صحیح در مقایسه با جملات تولید شده انسان [۴]	۱۹
۴۱	مواردی از خروجی برنامه‌سازی خطی صحیح که نسبت به جملات انسان، برتری دارد [۴]	۲۰
۴۲	نتایج برنامه‌سازی خطی صحیح که به لحظه‌های مختلف دچار مشکل شده‌اند [۴]	۲۱
۴۴	نمونه‌های صحیح از جملات تولید شده توسط قالب‌های آماده زبانی [۵]	۲۲
۴۴	نمونه‌های اشتباه تولید شده توسط قالب‌های آماده زبانی [۵]	۲۳
۴۷	طرح‌واره شبکه عصبی بازگشتی ضربی [۶]	۲۴
۵۲	هم‌ترازسازی تصویر و جمله [۷]	۲۵
۵۳	ارتباط بین نواحی مختلف یک تصویر و عبارات جمله [۷]	۲۶
۵۴	طرح‌واره شبکه عصبی بازگشتی دوطرفه [۷]	۲۷
۵۵	انتساب نواحی مختلف تصویر به عبارات زبانی [۷]	۲۸
۵۶	طرح‌واره شبکه عصبی بازگشتی ارائه شده برا تولید جمله [۷]	۲۹
۵۶	نتایج رتبه‌بندی عبارات زبانی برای نواحی تصویر [۷]	۳۰
۵۷	نتایج تولید جمله برای تصاویر در [۷]	۳۱
۵۹	ساختار کلی شبکه ارائه شده برای نگاشت دوطرفه تصاویر و جملات در پژوهش [۸]	۳۲

۶۱	نمونه‌ای از جملات تولید شده برای تصاویر توسط مدل پیشنهاد شده در [۸]	۳۳
۶۵	ساختار کلی چارچوب کاری انکودر-دیکودر	۳۴
۶۶	ساختار دیکودر مورد استفاده در چارچوب کاری [۹]	۳۵
۶۸	ساختار کلی یک شبکه عصبی بازگشتی دوطرفه	۳۶
۷۰	یک واحد از شبکه حافظه کوتاه‌مدت بلند مورد استفاده در دیکودر پژوهش [۱۰]	۳۷
۷۳	نحوه عمل کرد الگوریتم در تغییر توجه بصری بصری و کلمه تولید شده در هر نقطه. [۱۰] . . . چند نمونه از تصاویر که در آنها توجه بصری روی یک جسم منجر به تولید کلمه دقیق متناظر شده است [۱۰].	۳۸
۷۵	نمونه‌هایی از تولید کلمات نامناسب مطابق با نقاط توجه استفاده شده در مدل [۱۰]	۴۰
۷۶	فرایند تولید شرح متناظر تصویر با استفاده از توجه بصری سخت [۱۰] . . .	۴۱
۷۷	فرایند تولید شرح متناظر تصویر با استفاده از توجه بصری نرم [۱۰] . . .	۴۲
۷۸	ساختار پشته‌ای ارائه شده در [۱۱]	۴۳
۷۹	ساختار چارچوب کاری ارائه شده در [۳۵] در حوزه تولید شرح متناظر تصویر . . .	۴۴
۸۴	ساختار گسترده شبکه عصبی GPU مطرح شده در پژوهش [۱۲]	۴۵
۸۵	ساختار نسخه ماقفی شبکه GPU ارائه شده در پژوهش [۱۳]	۴۶
۸۶	ساختار شبکه توسعه یافته GPU ارائه شده در [۱۳] . . .	۴۷
۸۸	مقایسه عمل کرد مدل‌های مختلف نسبت به طول جملات [۱۳]	۴۸

فهرست جداول

۳۹	نتایج استفاده از روش نزدیکترین همسایه در اختصاص جملات و تصاویر [۳]	۱
۴۴	نتایج معیارهای ROUGE و BLUE در حالات مختلف [۵]	۲
۵۶	نتایج معیار BLUE برای روش ارائه شده در [۷] در مقایسه با دو روش دیگر	۳
۶۰	امتیاز BLEU کسب شده توسط مدل نگاشت دوطرفه ارائه شده در مقایسه با مدل‌های دیگر [۸]	۴
۶۰	جدول نتایج بازیابی تصاویر با استفاده از جملات ورودی در مدل ارائه شده در [۸]	۵
۷۴ [۱۰]	نتایج اعمال روش [۱۰] بر روی مجموعه‌داده‌های مختلف در مقایسه با روش‌های مختلف.	۶
۸۷	جدول نتایج به‌دست آمده از مدل‌های مختلف روی مجموعه‌داده ترجمه انگلیسی به فرانسوی [۱۳]	۷

١ فصل اول

مقدمات

به دنبال پیشرفت تکنولوژی در ساخت دوربین‌های عکاسی و ورود دوربین‌های نیمه‌خودکار و خودکار به بازار، تعداد زیادی از کاربران سیستم‌های رایانه‌ای به استفاده از این تکنولوژی در ثبت تصاویر مورد علاقه خود جذب شده‌اند. دقیق و کیفیت مطلوب تصویربرداری از یک سو و سهولت استفاده از دوربین از سوی دیگر، باعث شده‌اند تعداد تصاویر ثبت شده توسط کاربران به طور روزافزون افزایش یابد؛ به‌طوری‌که امروزه اغلب کاربران، تعداد بی‌شماری از این تصاویر را در گوشی‌های تلفن همراه، تبلت‌ها و رایانه‌های شخصی خود نگه‌داری می‌کنند. از جمله مشکلاتی که در اثر ایجاد این حجم وسیع از تصاویر بوجود آمده، مشکل مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه بزرگی از تصاویر موجود، است.

برای دست‌یابی به سامانه‌ای که بتواند تعداد زیادی از تصاویر موجود را مدیریت نماید، ابتدا باید صحنه موجود در تصویر را به درستی درک کرد. درک صحیح از صحنه، عبارت است از بیان تصویر به نحوی که اطلاعات کلی موجود و هدف اصلی تصویر، واضح و مشخص باشد. این بیان می‌تواند شامل اجسام موجود در تصویر، رابطه مکانی بین اجسام، فعالیت به تصویر کشیده شده، شرایط محیطی موثر بر صحنه و مواردی از این دست باشد. از طرفی باید به نحوی محتوای تصاویر را بیان کرد که بتوان عملیات جستجو را بر اساس مدل بیان شده تصاویر انجام داد. در این صورت به‌ازای هر تصویر، یک نمونه از مدل مطابق با تصویر ایجاد و ذخیره خواهد شد. پرس‌وجوی^۱ کاربر به فضای مدل، نگاشت شده و تصویر معادل با مدل استخراج شده، به عنوان نتیجه جستجو نمایش داده می‌شود. علاوه بر این، مساله مدیریت تصاویر، به مساله مدیریت مدل‌های موجود کاهش داده می‌شود.

تولید شرح کلی بر تصاویر،^۲ بیان مناسبی از صحنه موجود در تصویر را ارائه می‌دهد. شرح تولید شده بر تصاویر، در قالب مجموعه‌ای از جملات زبان طبیعی^۳ ارائه می‌شود که عموماً بیان‌گر اجسام موجود در صحنه، ارتباطات مکانی بین اجسام و اطلاعات مشخص دیگر است که در هر پژوهش می‌تواند متفاوت باشد. بنابراین، دست‌یابی به سامانه‌ای که قادر به تولید خودکار شرح کلی بر تصاویر باشد، اساسی‌ترین گام در راستای تولید نرم‌افزارهای مدیریت تصاویر است.

یکی از اولین ایده‌های مطرح شده در این زمینه، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین^۴ به وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

^۱Query

^۲Holistic Image Caption

^۳Natural Language Sentences

^۴Machine Translation

یک جمله از زبان مبدا^۵، با روش‌های مختلف، تبدیل به یک بردار ویژگی^۶ می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل، با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد^۷ میگردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند. با توجه به فرایند مذکور، اگر به جای جمله زبان مبدا، یک تصویر به بردار ویژگی تبدیل شده و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی حاصل به جمله زبان مقصد ترجمه شود، جمله‌ای معادل با تصویر ورودی به دست خواهد آمد که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است [۱۰].

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده‌اند. علی‌رغم وجود پژوهش‌های فراوان و متفاوت، می‌توان یک بستر کلی برای تمام فعالیت‌های موجود در این زمینه ارائه داد. بر این مبنای، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی‌گرفته شده‌است، از دو بخش اساسی تشکیل می‌شود.

۱. بازنمایی تصاویر، با استفاده از بردار ویژگی

۲. تبدیل بردار ویژگی به دست‌آمده به جملات صحیح زبانی

۲-۱ تعریف مساله

در این پژوهه قصد داریم سامانه‌ای ارائه دهیم که قادر به تولید شرح کوتاه بر تصاویر باشد. دو دیدگاه اساسی در دست‌یابی به چنین سامانه‌ای مطرح است.

۱. یافتن نقاط توجه^۸ در تصاویر و تولید جملات توصیف‌کننده اجسام مستقر در این نقاط به طوری که توصیف جسم مستقر در نقطه توجه و اجسام مرتبط با آن در جملات تولیدی، وجود داشته باشد.

۲. تولید شرح جامع بر تصاویر به طوری که تمام اجسام موجود در صحنه به همراه روابط موجود بین آن‌ها توصیف شوند.

شرح کوتاه تولید شده در این پژوهه، به معنی تولید جملاتی است که مستقیماً به توصیف صحنه، اجسام موجود در صحنه و روابط بین آنها می‌پردازند. به طور کلی، دو چالش عمده در این پژوهش مورد توجه قرار خواهد گرفت:

۱. توصیف صحنه باید دقیق باشد؛ به این معنی که اجسام موجود در صحنه باید به طور دقیق از هم تفکیک شده و دسته‌بندی شوند. تصویر توصیف شده باید در قالب مناسبی بازنمایی شود که بتوان به راحتی از آن برای تولید جمله استفاده نمود.

۲. جملات تولید شده برای شرح تصویر باید به لحاظ دستور زبان، املا و معنا صحیح بوده و با تصویر مرتبط خود سازگار باشند و آن را به درستی و دقیق شرح دهند.

^۵Source Language

^۶Feature Vector

^۷Destination Language

^۸Attention Points

۳-۱ درک صحنه

مساله درک صحنه، یکی از چالش‌های بزرگ و قدیمی مطرح در زمینه بینایی ماشین است. در گذشته، هدف اغلب پژوهش‌گران از طرح این مساله، توصیف صحنه موجود در تصویر با دیدن لحظه‌ای تصویر بوده است؛ اگرچه امروزه، تعریف این مساله دچار تغییر شده است.

به طور کل نمی‌توان تعریف جامع و شاملی برای درک صحنه ارائه داد. اگرچه تعاریفی عمومی ارائه شده‌اند که کلیات این مفهوم را توضیح می‌دهند. پژوهش‌گران در این زمینه هریک سعی در ارائه تعریفی برای این مفهوم دارند که برای کاربرد مورد نظر خود کافی و مفید باشد. به عنوان مثال، یکی از جدیدترین تعاریف برای درک صحنه در پژوهش [۱۴] ارائه شده است:

* «توانایی تحلیل بصری یک صحنه برای پاسخ‌دادن به سوالاتی مانند سوالات زیر:

- چه اتفاقی در حال رخ دادن است؟

- چرا این اتفاق در حال رخ دادن است؟

- اتفاق بعدی که رخ خواهد داد، چیست؟»

به طور کل می‌توان درک صحنه را چنین معنا کرد:

* درک صحنه، فرایندی است که طی آن، یک سامانه رایانه‌ای با استفاده از الگوریتم‌های موجود، اطلاعات بصری نهفته در تصاویر را استخراج کرده و در قالب مناسبی بازنمایی^۹ کند به طوری که این اطلاعات برای توصیف صحنه کافی و مفید باشد.

با این تعریف، اگرچه مفهوم درک صحنه کمی روشن می‌شود اما نکات مبهمی مانند این که چه نوع اطلاعاتی از تصویر استخراج شود، نیاز به توضیح و تفسیر بیشتری دارند. در تمام پژوهش‌های موجود در زمینه درک صحنه، که تعدادی از آن‌ها را در فصل بعدی مورد بررسی قرار خواهیم داد، تعریف اتخاذ شده برای درک صحنه، همین تعریف است با این تفاوت که اطلاعات مورد نیاز برای استخراج، در هر پژوهش، بسته به کاربرد تعریف می‌شود. موارد مختلفی که به عنوان اطلاعات لازم برای درک و توصیف صحنه، در پژوهش‌ها به چشم می‌خورد عموماً شامل موارد زیر هستند:

۱. دسته صحنه^{۱۰} (دریا، جنگل، خیابان، کلاس درس و مواردی از این دست)

۲. دسته اجسام^{۱۱} موجود در صحنه (صندلی، مرد، گربه و مواردی از این دست)

۳. ارتباط مکانی بین اجسام موجود (بالا، کنار، پشت و مواردی از این دست)

۴. رخدادی^{۱۲} که در صحنه در حال اتفاق است (مانند نشستن، دویدن، کارکردن و مواردی از این دست)

^۹Representation

^{۱۰}Scence Class

^{۱۱}Object Class

^{۱۲}Event

۱-۳-۱ پژوهش‌های انجام‌شده در زمینه درک صحنه توسط مغز انسان

مساله درک صحنه، مانند بیشتر مسائل موجود در زمینه بینایی ماشین، الهام گرفته از نحوه رفتار انسان‌ها است. اغلب انسان‌ها با دیدن یک تصویر قادرند توصیف کامل و دقیقی از آن تصویر ارائه دهند که شامل تمام نکات لازم و ضروری نهفته در تصویر باشد. در بیشتر موارد، زمان مورد نیاز برای مغز انسان به منظور پردازش یک تصویر و توصیف آن، زمان بسیار کم و ناچیزی است. این مطلب، این ایده را در ذهن تداعی می‌کند که بخش قابل توجهی از اطلاعات مورد نیاز از هر تصویر، در اولین لحظاتی که تصویر به مغز می‌رسد (در نگاه اول) قابل استخراج است. بنابراین سامانه‌های رایانه‌ای باید قادر باشند با الگو گرفتن از مغز انسان، در کوتاه‌ترین زمان ممکن، اطلاعات کافی و مفید نهفته در تصویر را استخراج کرده و صحنه به نمایش کشیده شده در تصویر را توصیف کنند.

این فرض که مغز انسان می‌تواند در کوتاه‌ترین زمان ممکن، بیشترین حجم اطلاعات تصویر را به درستی استخراج نماید، توسط پژوهش‌گران متعددی مورد ارزیابی قرار گرفته است. از جمله اولین پژوهش‌هایی که به بررسی این فرض پرداخته‌اند می‌توان به [۱۵] و [۱۶]^{۱۳} اشاره کرد. در این پژوهش‌ها، با نشان دادن تصاویر به صورت دنباله‌ای به مجموعه‌ای از افراد، از آن‌ها خواسته شده تا بهترین و دقیق‌ترین توصیفی را که می‌توانند، برای تصاویری که دیده‌اند، بازگو کنند. در این دو پژوهش نتیجه گرفته شده است که انسان می‌تواند یک تصویر معمولی را در بازه زمانی کمتر از ۲۰۰ میلی‌ثانیه، تشخیص داده و آن را توصیف کند. اگرچه این زمان برای تشخیص و توصیف یک تصویر کافیست، زمان مورد نیاز برای به‌خاطرسباری تصویر بسیار بیشتر از این مقدار است.

در پژوهش [۱] آزمایش دیگری انجام شده که از اهمیت بسیاری برخوردار است. در پژوهش‌های قبلی، افرادی که تصاویر را توصیف می‌کردند، درباره موضوع کلی تصاویر اطلاعاتی داشتند. اما در این آزمایش، تصاویر مختلفی از دنیای واقعی که محدود به شرایط خاصی نبوده‌اند، بدون ارائه پیش‌فرض درباره موضوع، به افراد نمایش داده شده و از آن‌ها خواسته شده که تصویر را به بهترین شکل توصیف کنند. آزمایشات در این پژوهش، در دو مرحله انجام شده‌اند.

۱. توسط یک رایانه، تصاویر متعددی در بازه‌های زمانی متفاوت به افراد نمایش داده می‌شوند و پس از اتمام زمان نمایش هر تصویر، یک ماسک بصری، تصویر را می‌پوشاند. در این حالت از افراد خواسته شده است که بهترین توصیف ممکن از تصویر را تایپ کنند. شرایط محیطی آزمایشات مطابق با استانداردها رعایت شده است. هر تصویر به طور تصادفی بین ۲۷ الی ۵۰۰ میلی ثانیه روی نمایش‌گر نمایش داده شده و سپس یک ماسک روی تصویر قرار گرفته و افراد فرصت دارند تا توصیف خود را از تصویر، بنویسند.

^{۱۳}Image Series



PT = 107 ms

PT = 500 ms

This is outdoors. A black, furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball. (Subject EC)

I saw a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out. (Subject JB)

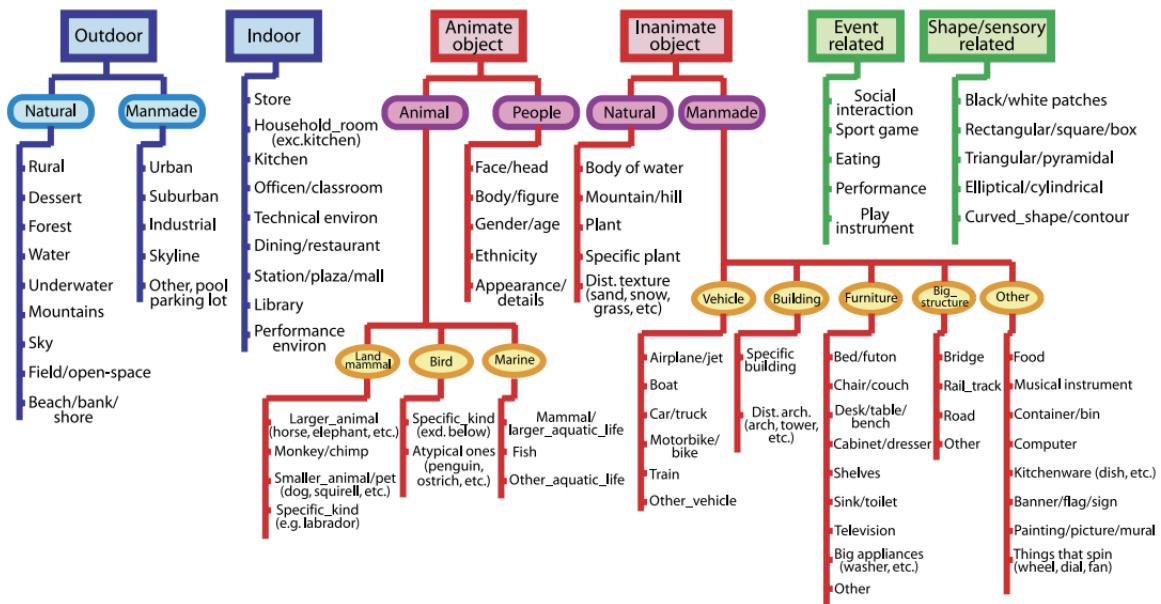


Inside a house, like a living room, with chairs and sofas and tables, no ppl. (Subject HS)

A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left. (Subject RW)

شکل ۱: نمونه توصیف‌های افراد برای تصاویر [۱]

۲. در این مرحله، آزمایش روی افراد متفاوتی انجام شده است. این گروه افراد موظفند پس از دیدن تصاویر، به بهترین شکل ممکن آن‌ها را دسته‌بندی کنند. برخلاف افراد شرکت‌کننده در آزمایش قبلی که می‌توانستند به هر شکلی اطلاعات استخراج شده را بنویسند، به افراد حاضر در این گروه یک فرم مشخص از دسته‌اطلاعات مطلوب داده شده است که افراد موظفند آن را براساس محتوای تصویری که دیده‌اند، پر کنند. شکل ۲ ساختار ۲ ساختار مطلوب پاسخ افراد را در این آزمایش نمایش می‌دهد.



شکل ۲: ساختار مطلوب اطلاعات استخراج شده از تصاویر [۱]

این ساختار با تحلیل پاسخ‌های جمع‌آوری شده از آزمایش اول استخراج شده است و شامل انواع مختلفی از اطلاعات است که افراد در آزمایش اول به آن اشاره کرده‌اند.

شکل ۳ چند نمونه از تصاویر مورد استفاده در آزمایشات این پژوهش را نمایش می‌دهد. این تصاویر از اینترنت استخراج شده‌اند. برای استخراج این تصاویر از فضای اینترنت، از یک گروه افراد شامل ۱۰ نفر که با موضوع پژوهش

آشنا نبوده‌اند خواسته‌شده تا هر یک، نام ۵ دسته صحنه مختلف را به طور تصادفی بنویسند. پس از حذف نام‌های تکراری، ۳۰ نام منحصر به فرد باقی مانده‌است. سپس تصاویر مربوط به هریک از این نام‌ها توسط موتور جستجوی گوگل استخراج شده و ۶ تصویر از صفحات اولیه نتایج به عنوان تصاویر نمونه انتخاب شده‌اند.



(آ) چند نمونه از تصاویر در محیط باز
(ب) چند نمونه از تصاویر در محیط بسته

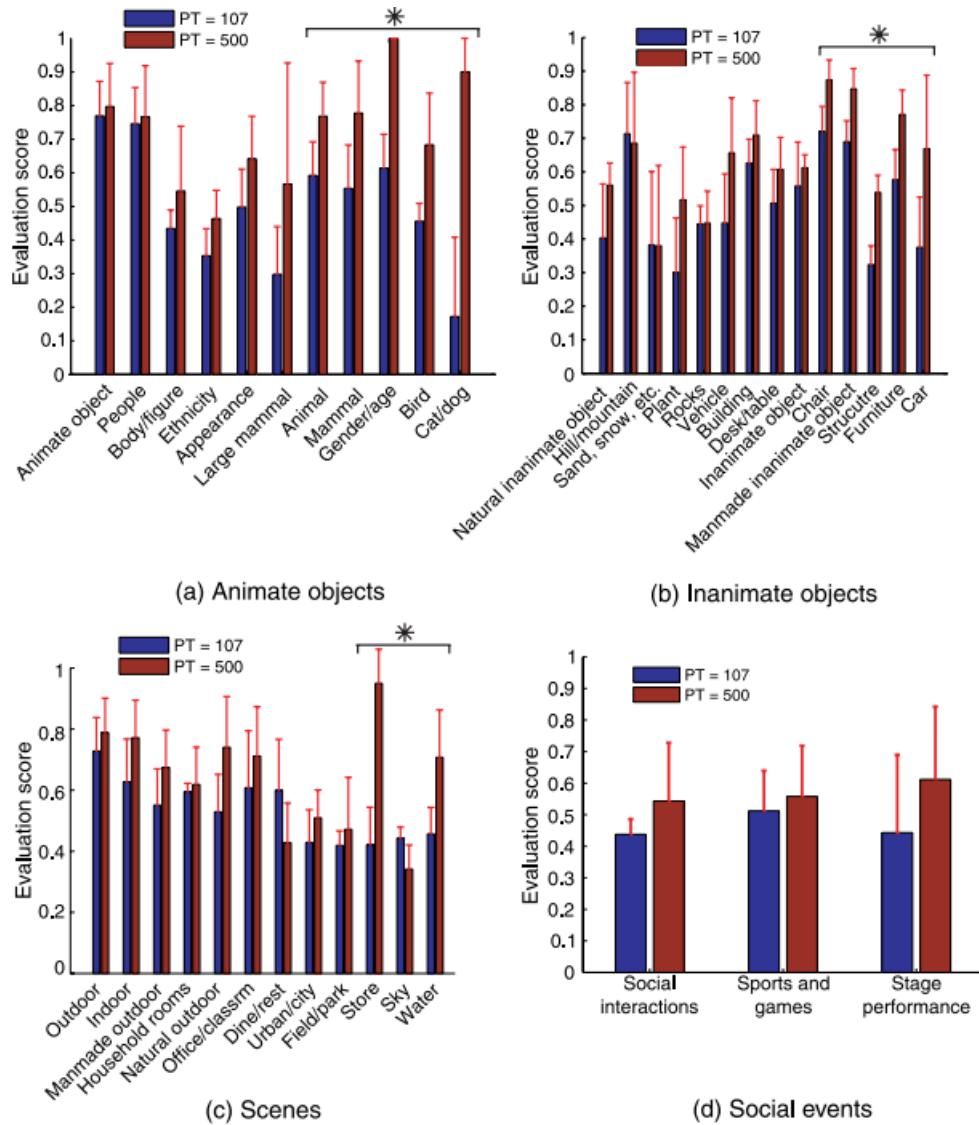
شکل ۲: تصاویر دنیای واقعی مورد استفاده در آزمایشات [۱]

ارزشمندترین نکته درباره پژوهش انجام‌شده، یافته‌های آن است. این پژوهش نکاتی را در مورد توانایی مغز انسان در توصیف صحنه روشن می‌کند که حائز اهمیت هستند. در ادامه این نتایج را بررسی خواهیم کرد.

۱-۳-۲ نتایج به دست آمده از آزمایشات

۱. حداقل زمان لازم برای مغز انسان به منظور درک صحنه، برابر با ۵۰۰ میلی ثانیه است.
۲. این مدت زمان، برای صحنه‌های ساده و بدون پیچیدگی، به حدود ۱۰۰ میلی ثانیه می‌رسد. به عنوان نمونه در شکل ۱ تصویر اول که دارای پیچیدگی‌های کمتری نسبت به تصویر دوم است در مدت زمان ۱۰۷ میلی ثانیه، به‌طور کامل توصیف شده‌است در صورتی که تصویر دوم که به نسبت، پیچیده‌تر است، مدت زمان بیشتری برای توصیف نیاز داشته‌است.
۳. با استفاده از ساختارمندسازی پاسخ‌های افراد در آزمایش دوم و اطلاعات جمع‌آوری شده در درخت پاسخ‌ها (که در شکل ۲ نمایش داده شده است) و میانگین‌گیری روی تمام تصاویر، نمودارهای مقایسه‌ای برای مدت

زمان ۱۰۷ میلی ثانیه و ۵۰۰ میلی ثانیه ایجاد شده است. شکل ۴ نمودارهای مقایسه‌ای را نمایش می‌دهد. در این نمودارها، میله‌های قرمز نشان‌دهنده نتایج برای زمان ۵۰۰ میلی ثانیه و میله‌های آبی نمایش‌دهنده نتایج برای حالت ۱۰۷ میلی ثانیه هستند. در دو نمودار اول (نمودارهای بالا سمت راست و بالا سمت چپ) تشخیص و استخراج اطلاعات مربوط به اجسام مختلف بسته به متحرک بودن^{۱۴} یا متحرک نبودن^{۱۵} آن‌ها، در نمودار سوم (نمودار پایین سمت چپ) تشخیص و استخراج اطلاعات مربوط به صحنه موجود در تصویر و در نمودار چهارم (نمودار پایین سمت راست) تشخیص و استخراج اطلاعات مربوط به رخداد موجود در تصویر، مورد بررسی قرار گرفته‌اند.



شکل ۴: نمودارهای مقایسه‌ای عملکرد مغز انسان در درک صحنه در بازه‌های زمانی ۱۰۷ و ۵۰۰ میلی ثانیه^[۱]

همان‌طور که مشخص است، مدت زمان ۱۰۷ میلی ثانیه برای مغز انسان، زمان بهینه برای توصیف صحنه است. تفاوت‌های بین نتایج در اکثر موارد، جزئی و در مقابل تفاوت زمانی موجود، بسیار کوچک هستند. به علاوه، در تمام مواردی که نیاز به اطلاعات کلی از تصویر وجود دارد، تفاوت بین دو بازه زمانی چندان

^{۱۴}Animated

^{۱۵}Inanimate

چشم‌گیر نیست، اما در مواردی که برای تشخیص نیاز به دانستن جزئیات بیشتر از تصویر وجود دارد (مانند سن، جنسیت و نوع حیوان) تفاوت بین دو زمان، قابل ملاحظه است.

همین‌طور با مقایسه تفاوت عملکرد بین حالات متحرك بودن و متتحرك نبودن اجسام، فواصل موجود در نمودارها قابل ملاحظه می‌شود. در حالت کلی، تفاوت بین عملکرد مغز در دو بازه، در حالتی که اجسام ساکن در تصویر وجود دارند به مراتب کمتر از حالتی است که اجسام موجود در تصویر، متتحرك باشند.

شکل ۵ نمونه دیگری از نتایج به دست آمده از آزمایشات را در مدت‌زمان‌های مختلف نمایش می‌دهد.

		
PT 27 ms	There was a range of dark splotches in the middle of the picture, running from most of the way on the left side, to all the way on the right side. This was surrounded primarily by a white or light gray color. (Subject: KM)	Couldn't see much; it was mostly dark w/ some square things, maybe furniture. (Subject: AM)
PT 40 ms	I saw a very bright object, shaped in a pyramidal shape. There was something black in the front, but I couldn't tell what it was. (Subject: JB)	Looked like something black in the center with four straight lines coming out of it against a white background. (Subject: AM)
PT 67 ms	Possibly outdoors. maybe a few ducks, or geese. Water in the background. (Subject: JL)	This looked like an indoor shot. Saw what looked like a large framed object (a painting?) on a white background (i.e., the wall). (Subject: RW)
PT 500 ms	It was definitely on a coast by the ocean with a large [r]ock in the foreground and at least three birds sitting on the rock. (Subject: CC)	The first thing I could recognize was a dark splotch in the middle. It may have been rectangular-shaped, with a curved top...but, that's just a guess. (Subject: KM)
	I saw the interior of a room in a house. There was a picture to the right, that was black, and possibly a table in the center. It seemed like a formal dining room. (Subject: JB)	A person, I think, sitting down or crouching. Facing the left side of the picture. We see their profile mostly. They were at a table or were some object was in front of them (to their left side in the picture). (Subject: EC)
	Some fancy 1800s living room with ornate single seaters and some portraits on the wall. (Subject: WC)	This looks like a father or somebody helping a little boy. The man had something in his hands, like a LCD screen or laptop. they looked like they were standing in a cubicle. (Subject: WC)

شکل ۵: نمونه‌ای از نتایج به دست آمده از آزمایشات [۱]

۴-۱ جمع‌بندی

با توجه به افزایش چشم‌گیر تعداد تصاویر مورد استفاده کاربران در فضاهای مجازی و همین‌طور با در نظر گرفتن گرایش روزافزون کاربران به ذخیره‌سازی تصاویر در رایانه‌های شخصی، مساله مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه تصاویر موجود، به یکی از مسائل مهم و پرکاربرد در زمینه بینایی ماشین تبدیل شده است. گام اساسی در این راستا، دست‌یابی به سامانه‌ای است که قادر به تولید خودکار شرح برای تصاویر باشد. شرح این تصاویر که در قالب جملات زبان طبیعی ارائه می‌شود باید علاوه بر سازگاری با موضوع تصویر و توصیف صحیح صحنه، به لحاظ دستور زبان و معنا صحیح و کامل باشد. فرایند تولید خودکار شرح برای تصاویر، از دو مرحله اصلی تشکیل می‌شود:

۱. نگاشت تصویر ورودی به فضای بردار ویژگی‌ها (درک صحنه)

۲. تولید جملات زبان طبیعی مبتنی بر محتوای بردار ویژگی‌ها

مساله درک صحنه، یکی از چالش برانگیزترین مسائل در زمینه بینایی ماشین است. با این وجود، تا کنون تعریف دقیق و کاملی از این مفهوم ارائه نشده است. به طور کلی می‌توان درک صحنه را فرایندی دانست که طی آن اطلاعات بصری موجود در تصویر استخراج شده و در قالب خاصی بازنمایی می‌شوند. میزان و نوع این اطلاعات را نمی‌توان به طور کلی تعریف کرد. حوزه تعریف اطلاعات و کیفیت مطلوب آن‌ها بسته به کاربرد در هر حوزه تعریف می‌شود.

در بین پژوهش‌های مربوط به تولید خودکار شرح برای تصاویر، انواع اطلاعات مطلوب، عموماً شامل موارد زیر می‌شود:

۱. دسته صحنه

۲. دسته اجسام

۳. ارتباط مکانی بین اجسام موجود

۴. رخدادی که در صحنه در حال اتفاق است

پژوهش‌گران از گذشته بر این عقیده بوده‌اند که مغز انسان در اولین لحظات مشاهده یک تصویر، قادر است اطلاعات کافی و مفید برای درک صحنه را استخراج کند. پژوهش‌های متعددی در این زمینه انجام شده‌اند که هریک به بررسی جوانب خاصی از این فرضیه پرداخته‌اند. به عنوان نمونه، پژوهش [۱۵] و [۱۶] با استفاده از دنباله‌های تصاویر، مدت زمان مورد نیاز برای مغز انسان به جهت درک صحنه را کمتر از ۲۰۰ میلی‌ثانیه تخمین زده‌اند.

در پژوهش [۱]، یک آزمایش دو مرحله‌ای برای بررسی تاثیر مدت زمان مشاهده تصاویر بر عملکرد مغز در توصیف صحنه، انجام شده است. در این آزمایش که در دو مرحله انجام شده، ابتدا گروهی از افراد با دیدن تصاویر در مدت زمان بین ۲۷ تا ۵۰۰ میلی‌ثانیه، موظف به توصیف تصویر بوده‌اند. سپس گروه دیگری از افراد با دیدن تصاویر در مدت زمان‌های مختلف، ملزم به پر کردن فرم از پیش تعیین‌شده‌ای بودند که با توجه به پاسخ‌های بهدست‌آمده از آزمایش اول، تدوین شده است.

نتایج این آزمایشات نشان می‌دهد، مدت زمان ۱۰۷ میلی‌ثانیه برای تشخیص و بخش قابل توجهی از اطلاعات موجود در تصویر کافیست؛ اگرچه، در مواردی که دق به جزئیات ضروری است (مانند تشخیص سن، جنسیت، نوع حیوان) و برای تشخیص و استخراج اطلاعات اجسام متحرک، مدت زمان ۵۰۰ میلی‌ثانیه، بهبود قابل توجهی در عملکرد مغز ایجاد می‌کند.

۲ فصل دوم

درگ صحنه

۱-۲ درک صحنه

درک صحنه یکی از چالش‌های اساسی در زمینه بینایی ماشین است که روش‌های مختلفی برای دست‌یابی به آن ارائه شده است. با وجود تعدد پژوهش‌های موجود در این مورد، ارائه تعریف جامع و شامل برای این مفهوم کاری بسیار دشوار است. عموماً این مفهوم، بسته به مورد کاربرد و هدف پژوهش، به استخراج مجموعه مشخصی از اطلاعات در مورد صحنه که برای پژوهش، کافی و مفید باشد محدود می‌شود. به همین دلیل، مجموعه اطلاعات مطلوب از تصویر که باید استخراج شود در هر پژوهش به طور خاص تعریف می‌شود.

درک صحنه در زمینه تولید خودکار شرح بر تصاویر، به طور عام شامل موارد زیر می‌شود:

۱. تشخیص اجسام موجود در صحنه و دسته‌بندی آن‌ها (مانند توپ، تلویزیون)

۲. تشخیص ارتباط مکانی بین اجسام موجود در صحنه (مانند پشت، بالا)

۳. دسته‌بندی محیط (مانند جنگل، دریا)

۴. دسته‌بندی فعالیت به تصویر کشیده شده (مانند راه‌رفتن، خوابیدن)

۲-۲ روش‌های مختلف موجود

فعالیت‌های متعددی برای تشخیص هر یک از موارد بالا انجام شده است. به طور عام می‌توان روش‌های مورد استفاده در استخراج اطلاعات مطلوب صحنه را در زمینه تولید خودکار شرح بر تصاویر به دو دسته عمدی زیر تقسیم‌بندی نمود:

۱. استفاده از مدل‌های گرافی احتمالی^{۱۶}

در این دسته از روش‌ها، با استفاده از مدل‌های گرافی احتمالی در مورد حضور یا عدم حضور اجسام مختلف در صحنه و رابطه بین اجسام موجود استنتاج نمود. همین‌طور فرایندهایی مانند قطعه‌بندی تصویر^{۱۷} در این روش‌ها با استفاده از مدل‌های گرافی احتمالی انجام می‌شوند. به عنوان نمونه، در مقاله [۱۷] یک مدل

^{۱۶}Probabilistic Graphical Models (PGMs)

^{۱۷}Image Segmentation

میدان تصادفی شرطی^{۱۸} برای تجزیه معنایی^{۱۹} تصویر ارائه شده است که با استفاده از آن می‌توان در مورد حضور یا عدم حضور اجسام مختلف به طور توان در صحنه تصمیمگیری کرد.

۲. استفاده از شبکه‌های عصبی کانولوشنی عمیق در این دسته از روش‌ها، با استفاده از شبکه‌های عصبی کانولوشنی عمیق، پس از قطعه‌بندی تصاویر، اقدام به تفکیک اجسام مختلف در صحنه و برچسب‌گذاری هر جسم، بسته به یادگیری انجام شده، می‌شود. به عنوان نمونه در مقاله [۷] یک شبکه عصبی کانولوشنی عمیق معرفی شده است که قادر به برچسب‌گذاری اجسام مختلف در صحنه است. برچسب‌های مورد استفاده در این پژوهش، عبارات مختلف موجود در جملات توصیف‌گر هر تصویر در مجموعه‌دادگان هستند.

نمونه‌های متعددی از این دست پژوهش‌ها، در هر دسته، انجام شده است که در ادامه چند مورد از آن‌ها بررسی خواهد شد.

۳-۲ روش‌های مبتنی بر مدل‌های گرافی احتمالی

همان‌طور که قبلاً ذکر شد، روش‌های مبتنی بر استفاده از مدل‌های گرافی احتمالی، از جمله پرکاربردترین روش‌ها در مرحله درک صحنه در زمینه تولید خودکار شرح بر تصاویر هستند. این روش‌ها با استفاده از نظریه گراف، آمار و احتمالات اقدام به ارائه یک توزیع احتمالی برای پارامتر مورد بررسی، با توجه به داده‌های موجود در مجموعه آموزشی می‌کنند. مدل‌های استاندارد مختلفی در پژوهش‌ها مورد استفاده قرار می‌گیرند که تعدادی از آن‌ها به عنوان نمونه در این بخش مورد بررسی قرار خواهند گرفت.

۳-۲-۱ استفاده از مدل میدان تصادفی مارکف^{۲۰}[۱۸]

مقاله [۱۸] با استفاده از یک مدل ساده میدان تصادفی مارکف، فرایند درک صحنه را انجام می‌دهد و با استفاده از همین مدل، اقدام به تولید جملات توصیف‌گر تصویر می‌نماید. در این فصل به بررسی فرایند درک صحنه در این مقاله می‌پردازیم و بررسی فرایند تولید جمله را به فصل بعدی موکول می‌نماییم.

درک صحنه در این پژوهش محدود به ارتباط بین سه مفهوم در هر تصویر شده است؛ به این معنی که به ازای هر تصویر، یک سه‌تایی «جسم، فعالیت، صحنه»^{۲۱} ایجاد می‌شود که بیان‌کننده اطلاعات مطلوب موجود در تصویر است. میدان^{۲۲} «جسم»، دربر دارنده برچسب حاصل از دسته‌بندی اجسام موجود در صحنه، میدان «فعالیت»، دربر دارنده اطلاعات مربوط به فعالیت در حال انجام و میدان «صحنه» دربردارنده اطلاعات مربوط به محیط تصویر هستند. به فضای سه‌تایی‌های ایجاد شده برای اطلاعات مطلوب در درک صحنه، فضای معنا^{۲۳} می‌گویند.

شکل ۶ نمایی از نگاشت اطلاعات از فضای تصاویر و جملات به فضای معنایی، نمایش می‌دهد. همان‌طور که در شکل مشخص است، به ازای هر تصویر، یک سه‌تایی معنایی ایجاد می‌شود. همین‌طور به ازای هر جمله در

^{۱۸}Conditional Random Field (CRF)

^{۱۹}Semantic Parsin g

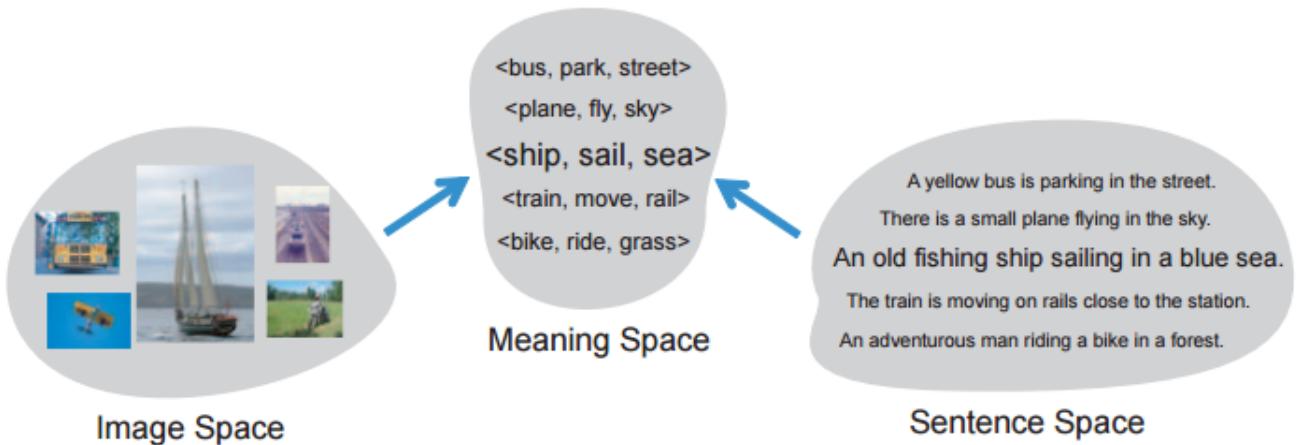
^{۲۰}Markov Random Field (MRF)

^{۲۱}<Object, Activity, Scene>

^{۲۲}Field

^{۲۳}Meaning Space

فضای جملات، یک سه‌تایی ایجاد می‌شود به‌طوری که جملات و تصاویر متناظر شان، به یک سه‌تایی یکسان، نگاشت شوند. همان‌طور که مشخص است، با داشتن نگاشتهایی که خواص مذکور را داشته باشند، می‌توان با استفاده از سه‌تایی‌های فضای معنا، تصاویر را مدیریت کرد.



شکل ۶: نگاشت تصویر به فضای معنایی. فضای معنایی شامل اطلاعات مطلوب برای استخراج در فرایند درک صحنه است. به ازای هر تصویر، یک سه‌تایی ایجاد می‌شود [۱۸].

مدل میدان تصادفی مارکف مورد استفاده در این پژوهش، یک مدل کوچک و ساده، شامل ۳ گره است. شکل ۷ طرح‌واره‌ای از مدل میدان تصادفی مارکف مورد استفاده در این پژوهش را نمایش می‌دهد. همان‌طور که در شکل مشخص است، به ازای هر کدام از میدان‌های تعریف شده در فضای معنایی، یک گره در این مدل وجود دارد. مقادیر مختلف در هر گره، برابر است با مقادیر مختلف موجود در میدان متناظر، در فضای معنا که با توجه به داده‌های مجموعه آموزشی مشخص می‌شوند. همین‌طور به ازای هر دو گره موجود در این مدل، یک یال بیان‌کننده ارتباط بین دو میدان در فضای معنایی وجود دارد.

برای استنتاج در این مدل، لازم است ابتدا فاکتورهای مورد استفاده در مدل را شناخته و مقادیر آن‌ها را مشخص نماییم. در مدل پیشنهادی، دو نوع فاکتور تعریف شده است:

۱. فاکتورهای گره

این فاکتورها، برای مشخص کردن میزان شباهت مقادیر مختلف گره با تصویر ورودی، تعریف شده‌اند. ویژگی‌های مورد استفاده برای مقداردهی این فاکتورها، شامل موارد زیر هستند:

(آ) استفاده از آشکارکننده‌های^{۲۴} فلزنسوالب^{۲۵}، به منظور محاسبه امتیاز اطمینان^{۲۶} برای هر دسته از اجسام موجود در مجموعه داده [۱۹].

پس از محاسبه امتیاز اطمینان همه دسته‌های موجود، دسته‌ای که بیشترین امتیاز را دارد می‌تواند به عنوان دسته منتخب در میدان متناظر گره، انتخاب شود. در فرایند مقداردهی این ویژگی، قبل از انجام محاسبات، اطمینان حاصل می‌شود که از هر دسته موجود، حداقل یک تصویر در مجموعه داده وجود داشته باشد.

^{۲۴}Detector

^{۲۵}Felzenszwaalb

^{۲۶}Confidence Score

(ب) استفاده از پاسخ دسته‌بندی‌کننده دیوالا^{۷۷}، ارائه شده در مقاله [۲۰]

(ج) استفاده از دسته‌بندی‌کننده مبتنی بر گیست^{۷۸}



شکل ۷: طرحواره مدل میدان تصادفی مارکف ارائه شده در پژوهش [۱۸] که شامل ۳ گره است. در این مدل، به ازای هر میدان از فضای معنا، یک گره وجود دارد و بین هر سه گره، به طور دو به دو، یک یال موجود است [۱۸].

بر اساس مقادیر محاسبه شده برای ویژگی‌های بالا و با استفاده از الگوریتم ماشین بردار پشتیبان^{۷۹}، یک دسته‌بندی برای هر گره ارائه می‌شود که بیان‌کننده دسته‌ویژگی‌های مربوط به مقادیر مختلف گره است. با استفاده از این دسته‌بندی، با ورود هر تصویر، می‌توان برای هر مقدار در هر گره، یک امتیاز شباهت محاسبه نمود. استفاده از الگوریتم یافتن نزدیک‌ترین همسایه‌های موجود برای هر تصویر ورودی، بر اساس امتیاز شباهت محاسبه شده و میانگین‌گیری روی همسایه‌های استخراج شده، معیار خوبی از تخمین مقدار هر گره، به ازای هر تصویر ورودی ایجاد می‌کند. به این ترتیب، با ورود هر تصویر می‌توان برای هر کدام از گره‌های موجود در مدل، یک مقدار محتمل مشخص نمود. سه‌تایی شامل مقادیر محتمل بدست‌آمده در هر گره، سه‌تایی متناظر تصویر ورودی در فضای معنا را مشخص می‌کند.

۲. فاکتور یال

این فاکتور، برای مشخص کردن میزان ارتباط مقادیر مختلف دو گره با یکدیگر در تصویر ورودی مورد استفاده قرار می‌گیرند.

^{۷۷}divvala

^{۷۸}Gist-based classification response

^{۷۹}Support Vector Machine (SVM)

۲-۳-۲ استفاده از مدل میدان تصادفی شرطی^{۳۰}

در این پژوهش، مساله درک صحنه در قالب یک مساله استنتاج با استفاده از مدل میدان تصادفی شرطی بیان شده است. مدل میدان تصادفی شرطی، یکی از پرکاربردترین مدل‌های گرافی احتمالی در زمینه درک صحنه است که پژوهش‌های متعددی از آن به عنوان مدل اصلی در درک صحنه استفاده کرده‌اند. به عنوان نمونه، در مقاله‌های [۲۱] و [۲۲] از مدل میدان تصادفی شرطی به منظور توصیف صحنه استفاده شده است.

پژوهش [۲۱] سعی در توصیف اجسام سه‌بعدی با استفاده از قطعه‌بندی تصاویر دو بعدی، هندسه سه‌بعدی و روابط بین صحنه و اجسام موجود، دارد. در این پژوهش، پس از استخراج ویژگی‌ها و اطلاعات بدست‌آمده از منابع مختلف، عمل استنتاج توسط یک مدل تصادفی شرطی انجام می‌شود که منجر به نگاشت تصویر ورودی به فضای معنایی می‌شود. همین‌طور در پژوهش [۲۲]، یک چارچوب کاری^{۳۱} احتمالی برای استنتاج درباره نواحی مختلف تصویر، اجسام موجود و ویژگی‌های مختلف آن‌ها مانند دسته‌بندی، موقعیت مکانی و ابعاد، مبتنی بر مدل میدان تصادفی شرطی، ارائه شده است. با توجه به وسعت و تعدد فعالیت‌های انجام شده، در این بخش، مرحله درک صحنه یک پژوهش انجام شده در زمینه تولید خودکار شرح بر تصاویر را مورد بررسی قرار می‌دهیم. لازم به ذکر است، مرحله تولید جملات توصیف‌کننده پژوهش مورد بحث، در فصل تولید جملات زبان طبیعی مورد بررسی قرار خواهد گرفت.

در پژوهش [۱۷] از مدل میدان تصادفی شرطی برای توصیف صحنه و اجسام موجود در آن استفاده شده است. میدان‌های تصادفی در این مدل، شامل متغیرهای زیر هستند:

۱. متغیرهای تصادفی بیان‌کننده برچسب دسته متناظر قطعات مختلف هر تصویر به شیوه سلسله مراتبی دارای دو سطح

۲. متغیرهای تصادفی باینری بیان‌کننده صحت دسته تشخیص داده شده برای هر جسم

شکل ۸ طرح‌واره مدل سلسله‌مراتبی ارائه شده در پژوهش [۱۷] را نمایش می‌دهد. همان‌طور که مشاهده می‌شود این مدل از دو سطح انتزاع، یکی برای برچسب قطعات مختلف تصویر و دیگری برای حضور یا عدم حضور هر دسته از اجسام در تصویر، تشکیل شده است.

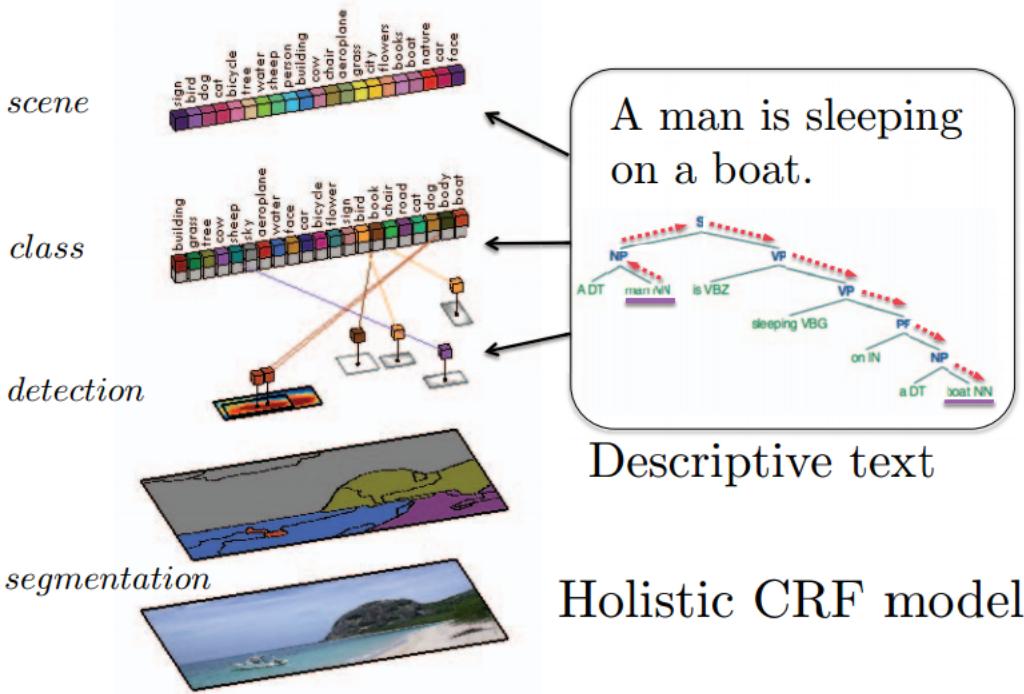
دو دسته متغیر تصادفی مختلف، که هر یک نماینده متغیرهای تصادفی موجود در یکی از این سطوح انتزاع هستند، تعریف شده‌اند؛ متغیرهای تصادفی C ، $X_i \in \{1, \dots\}$ بیان‌کننده دسته قطعه i از سطح پایین سلسله مراتب و متغیرهای تصادفی C ، $Y_j \in \{1, \dots\}$ بیان‌کننده دسته قطعه j از سطح بالای سلسله مراتب. به علاوه، دو دسته متغیر تصادفی دیگر به نام‌های b_k و z_k به ترتیب برای نمایش حضور یا عدم حضور یک تشخیص کاندید^{۳۲} و حضور یا عدم حضور جسم با دسته k در تصویر، تعریف شده‌اند. با توجه به متغیرهای تعریف شده، مدل کلی میدان تصادفی شرطی را می‌توان معادل رابطه^۱ تعریف کرد. در این رابطه $(a_\alpha)^{\Psi_\alpha^{type}}$ نماینده تابع پتانسیل تعریف شده روی متغیرهای مختلف است. با این تعریف، یافتن تخمین MAP^{۳۳}، منجر به یافتن پاسخ مورد نظر می‌شود.

^{۳۰} Conditional Random Field (CRF)

^{۳۱} Framework

^{۳۲} Candidate Detection

^{۳۳} MAP Estimation



شکل ۸: طرح‌واره مدل سلسله مراتبی مبتنی بر میدان تصادفی شرطی که بر اساس اطلاعات بصری و اطلاعات جملات توصیف‌کننده شرح محتمل تصویر را تولید می‌نماید [۱۷].

در ادامه، توابع پتانسیل مختلف که در این پژوهش تعریف شده‌اند، ارائه خواهد شد. لازم به ذکر است در تمام این موارد، برای سهولت، توابع پتانسیل به شکل لگاریتمی تعریف شده‌اند.

$$P(X, Y, b, z) = \frac{1}{Z} \prod_{type} \prod_{\alpha} \Psi_{\alpha}^{type}(a_{\alpha}) \quad (1)$$

توابع پتانسیل مختلف تعریف شده در این پژوهش عبارتند از:

۱. پتانسیل قطعه‌بندی یگانی ^{۳۴}

پتانسیل قطعه‌بندی یگانی در هر قطعه و هر ابرقطعه ^{۳۵} از تصویر، با استفاده از میانگین‌گیری روی امتیاز افزایش تکستون ^{۳۶} که در پژوهش [۲۳] ارائه شده است، انجام می‌شود.

۲. انطباق بین متغیرهای دو سطح انتزاع با یکدیگر

یک مقدار جریمه به ازای دسته‌های مخالف بین دو سطح در نظر گرفته می‌شود تا در حد امکان، دسته‌های منتخب از بین سطوح مختلف، با یکدیگر انطباق داشته باشند. پتانسیل تعریف شده در این بخش معادل رابطه ۲ تعریف می‌شود.

$$\phi_{ij}(X_i, Y_j) = \begin{cases} -\gamma & X_i \neq Y_j \\ 0 & X_i = Y_j \end{cases} \quad (2)$$

در رابطه ۲، پارامتر γ در فرآیند یادگیری که منجر به بهینه‌سازی پارامترهای مختلف مدل می‌شود، به دست می‌آید.

^{۳۴}Unary Segmentation Potential

^{۳۵}Supersegment

^{۳۶}Texton Boost

۳. پتانسیل انطباق تصویر و دسته جسم

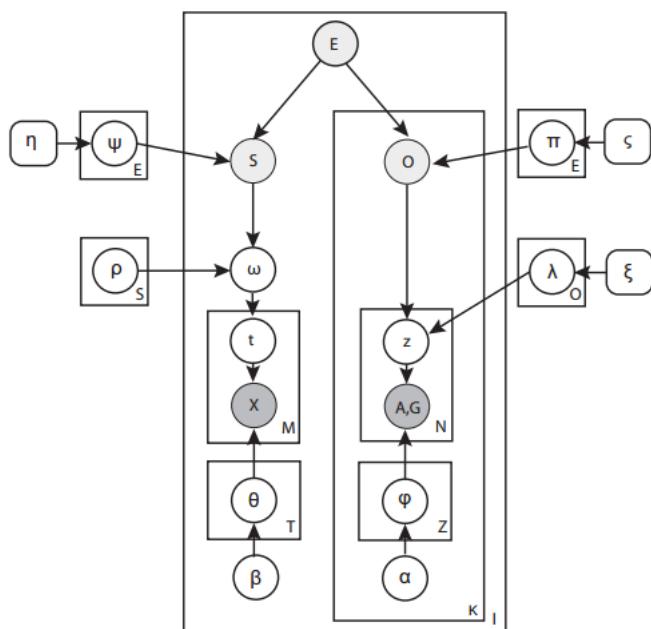
برای اندازه‌گیری میزان انطباق هر کدام از دسته‌های موجود برای اجسام با تصویر ورودی، از معیار انطباق ارائه شده در پژوهش [۲۴] توسط فلزنسوالب که به روش دی پی ام ^{۳۷} مشهور است، استفاده شده است. برای کاهش تعداد پارامترها و افزایش کارایی مدل استفاده شده، برای هر تصویر حداقل ۳ دسته جسم، به عنوان دسته‌های منتخب کاندید، در نظر گرفته می‌شوند.

۳-۳-۲ استفاده از سایر مدل‌های گرافی احتمالی

در بین پژوهش‌های موجود در زمینه درک صحنه با استفاده از روش‌های احتمالاتی، علاوه بر مدل‌های استاندارد، از مدل‌های مولد دیگر در پژوهش‌های متعددی استفاده شده است. در ادامه این بخش، به بررسی چند نمونه از این مدل‌ها خواهیم پرداخت.

۱. دسته‌بندی تصاویر بر اساس صحنه و اجسام موجود به طور توأم [۲]

مدل استفاده شده در این پژوهش، از تصاویر در سطح صحنه و سطح اجسام استفاده کرده و با یکپارچه‌سازی و تجمعی اطلاعات موجود در این دو سطح، اقدام به دسته‌بندی تصویر می‌نماید. شکل ۹ مدل استفاده شده در این پژوهش را به منظور یکپارچه‌سازی و تجمعی اطلاعات حاصل از تحلیل صحنه و تشخیص اجسام موجود در آن، ارائه می‌دهد.



شکل ۹: مدل استفاده شده به منظور تجمعی اطلاعات صحنه و اجسام موجود در آن به منظور دسته‌بندی تصاویر [۲]

یکی از اهدافی که در این پژوهش دنبال می‌شود، برحسب گذاری معنایی^{۳۸} تمام پیکسل‌های موجود در تصویر است. به همین منظور، تمام تصاویر مورد استفاده، به نواحی $10 * 10$ تقسیم شده و مورد استفاده قرار

^{۳۷}DPM

^{۳۸}Semantic Labelling

می‌گیرند. برای بررسی بهتر مدل، ابتدا متغیرهای تصادفی مورد استفاده را تعریف کرده و سپس به بررسی روند یادگیری و استنتاج مدل می‌پردازیم.

متغیر تصادفی X که حاوی اطلاعاتی مبتنی بر حضور یا عدم حضور دسته‌های مختلف صحنه است، در بخش تشخیص صحنه به کار می‌رود. اطلاعات این متغیر با استفاده از توصیف‌کننده سیفت^{۳۹} و به ازای هر ناحیه از تصویر، به دست می‌آید. برای بخش تشخیص اجسام موجود در صحنه، از دو منبع اطلاعاتی مختلف استفاده می‌شود. اطلاعات مربوط به حضور یا عدم حضور دسته‌های مختلف اجسام در متغیر تصادفی A و اطلاعات مربوط به شکل کلی آن‌ها در متغیر تصادفی G نمایش داده می‌شود.

هر گره از مدل ارائه شده، نماینده یک متغیر تصادفی است. گره‌هایی که با رنگ تیره مشخص شده‌اند، نماینده متغیرهایی هستند که در فرایند آموزش دیده می‌شوند و بقیه متغیرها، متغیرهای مخفی^{۴۰} هستند. گره‌های خاکستری روشن‌تر، متغیرهایی هستند که فقط در فرایند آموزش دیده می‌شوند در حالی که متغیرهای تیره‌تر در هر دو فرایند آموزش و آزمون مشاهده می‌شوند.

متغیر تصادفی E ، نماینده یک دسته از رخداد^{۴۱} های ممکن است. توزیع احتمال اولیه این متغیر تصادفی، یک توزیع یکنواخت فرض شده است که به هر تصویر ورودی، بر اساس همین توزیع، یک مقدار خاص از این متغیر تصادفی اختصاص داده می‌شود. با داشتن دسته رخداد موجود در تصویر، یک تصویر صحنه^{۴۲} متناظر با تصویر ورودی تولید می‌شود. با فرض وجود S دسته صحنه مختلف در مجموعه‌داده، به هر تصویر، تنها یک دسته صحنه اختصاص داده می‌شود. روند اختصاص دسته صحنه به تصویر مطابق زیر است:

* ابتدا یک دسته اولیه مطابق با توزیع احتمال شرطی $P(S|E, \psi)$ به تصویر اختصاص داده می‌شود.
 یک پارامتر چندجمله‌ای^{۴۳} حاکم بر توزیع احتمالاتی S به شرط داشتن E است. به علاوه، ψ یک ماتریس به ابعاد $S * E$ و پارامتر θ یک بردار S بعدی در نقش مقدار اولیه دیریکله^{۴۴} برای پارامتر ψ است.

* در قدم بعدی با داشتن مقدار S ، پارامترهای ω را بر اساس احتمال $P(\omega|S, \rho)$ تولید می‌کنیم. از آن جا که ω پارامتر چندجمله‌ای گره‌های مخفی t هستند، باید مجموع همه آن‌ها برابر با یک باشد. به علاوه، ρ یک ماتریس به ابعاد $S * T$ و مقدار اولیه دیریکله برای پارامتر ω است که در آن T تعداد کل t ‌ها است.

* برای تولید هر یک از M ناحیه تصویر (مقادیر متغیر تصادفی X) به شکل زیر عملی می‌کنیم:
 - یک مقدار t از توزیع احتمال $Mult(\omega)$ تولید می‌شود که مشخص‌کننده موضوعی^{۴۵} است که این ناحیه از تصویر مطابق با آن تولید شده است.

- متغیر تصادفی X از توزیع احتمالی $P(X|t, \theta)$ تولید می‌شود. θ یک ماتریس به ابعاد V_s

^{۳۹}SIFT Descriptor

^{۴۰}Latent

^{۴۱}Event

^{۴۲}Scene Image

^{۴۳}Multinomial

^{۴۴}Dirichlet prior

^{۴۵}Topic

است که در آن V_s تعداد کلمات موجود در پایگاه داده مربوط به صحنه s است. به علاوه، θ یک پارامتر چندجمله‌ای برای X است و β مقدار اولیه دیریکله برای θ .

همانند فرایندی که طی آن، تصویر صحنه به تصویر ورودی اختصاص داده می‌شود، فرایندی وجود دارد که طی آن تصویر اجسام^{۴۶} به تصویر ورودی اختصاص داده می‌شود. بر خلاف صحنه، هر تصویر می‌تواند بیش از یک جسم داشته باشد. تعداد کل اجسام موجود در یک تصویر را با K و تعداد کل دسته‌های موجود برای اجسام در مجموعه‌داده را با O نمایش می‌دهیم. فرایند زیر برای هر یک از K جسم موجود در تصویر اجرا می‌شود:

* ابتدا یک دسته جسم با توزیع احتمالی $P(O|E, \pi)$ به تصویر اختصاص داده می‌شود که در آن، π یک ماتریس به ابعاد $O * E$ و ζ یک بردار به طول O و مقدار اولیه دیریکله پارامتر π است.

* سپس با داشتن O می‌توان تمام نواحی A و G مرتبط با دسته جسم را تولید نمود. فرایند تولید این نواحی به شکل زیر است:

- متغیر تصادفی مخفی z که مشخص کننده موضوع است، از توزیع احتمالی $Mutl(\lambda, |O)$ تولید می‌شود. متغیر λ یک ماتریس به ابعاد $Z * O$ است که در آن Z تعداد کل مقادیر مختلف متغیر z است. به علاوه ζ مقدار اولیه دیریکله برای پارامتر λ است.

- نواحی مطلوب از توزیع احتمال $P(A, G|t, \phi)$ تولید می‌شوند که در آن، ϕ یک ماتریس به ابعاد $Z * V_o$ است. V_o تعداد کل کلمات موجود در مجموعه‌داده، به ازای نواحی A و G است. پارامتر α مقدار اولیه دیریکله برای پارامتر ϕ است.

با توجه به متغیرهای تصادفی توضیح داده شده در بالا، توزیع احتمالی توام کل سیستم را می‌توان مطابق با رابطه ۳ تعریف کرد.

$$\begin{aligned} P(E, S, O, X, A, G, t, z, \omega | \rho, \phi, \lambda, \psi, \pi\theta) &= P(E) \cdot P(S|E, \psi) \cdot P(\omega|S, \rho) \\ &\quad \cdot \prod_{m=1}^M P(X_m|t_m, \theta) \cdot P(t_m|\omega) \\ &\quad \cdot \prod_{k=1}^K P(O_k|E, \pi) \\ &\quad \cdot \prod_{n=1}^N P(A_n, G_n|z_n, \phi) \cdot P(z_n|\lambda, O_k) \end{aligned} \tag{۳}$$

به علاوه، با توجه به توضیحات ارائه شده در بالا، هر کدام از عبارات موجود در رابطه ۳ را می‌توان با عبارات معادل آن‌ها که در روابط ۴ تا ۱۰ آمده، جایگزین نمود.

^{۴۶}Object Image

$$P(S|E, \psi) = Mult(S|E, \psi) \quad (4)$$

$$P(\omega|S, \rho) = Dir(\omega|\rho_j), S = j \quad (5)$$

$$P(t_m|\omega) = Mult(t_m|\omega) \quad (6)$$

$$P(X_m|t_m, \theta) = P(X_m|\theta_j.), t_m = j \quad (7)$$

$$P(O_k|E, \pi) = Mult(O_k|E, \pi) \quad (8)$$

$$P(z_n|\lambda, O_k) = Mult(z_n|\lambda, O_k) \quad (9)$$

$$P(A_n, G_n|z_n, \phi) = P(A_n, G_n|\phi_j.), z_n = j \quad (10)$$

در ک صحنه در این پژوهش، محدود به استخراج سه دسته اطلاعات زیر از تصویر است:

(آ) رخدادی که در تصویر به نمایش گذاشته شده است.

(ب) صحنه‌ای که تصویر در آن ایجاد شده است.

(ج) اجمالی که در تصویر حضور دارند.

با توجه به این محدودیت و با در نظر گرفتن مدل ارائه شده، استفاده از تخمین بیشینه احتمال^{۴۷}، می‌تواند برای استخراج اطلاعات مطلوب مفید باشد. از همین رو، تخمین بیشینه احتمال، در سه سطح مختلف (هر سطح برای یک دسته از اطلاعات مطلوب) اعمال می‌شود. در سطح اجسام، احتمال رخداد تصویر ورودی به شرط اجسام موجود مطابق با رابطه ۱۱، احتمال رخداد تصویر ورودی به شرط صحنه، مطابق با رابطه ۱۲ و احتمال رخداد تصویر ورودی به شرط دسته رخداد به نمایش گذاشته شده در تصویر، مطابق با رابطه ۱۳ محاسبه می‌شوند.

$$P(I|O) = \prod_{n=1}^N \sum_j P(A_n, G_n|z_j, O) P(z_j|O) \quad (11)$$

$$P(I|S, \rho, \theta) = \int P(\omega|\rho, S) (\prod_{m=1}^M \sum_{t_m} P(t_m|\omega) P(X_m|t_m, \theta)) d\omega \quad (12)$$

$$P(I|E) \propto \sum_j P(I|O_j) P(O_j|E) P(I|S) P(S|E) \quad (13)$$

فرایند یادگیری این مدل، شامل یافتن بهترین مقادیر برای پارامترهای $\{\psi, \rho, \pi, \lambda, \theta, \beta\}$ است. این فرایند برای سه پارامتر $\{\psi, \rho, \theta\}$ با استفاده از روش انتقال پیام متغیر^{۴۸} و برای سه پارامتر $\{\pi, \lambda, \beta\}$ با استفاده از نمونه‌برداری گیبس^{۴۹} انجام می‌شود.

آزمایشات انجام شده در این پژوهش، بر روی یک مجموعه‌داده شامل تصاویر از ۸ دسته ورزشی مختلف که در هر دسته، بین ۱۳۷ تا ۲۵۰ تصویر مختلف وجود دارد، انجام شده‌اند. از جمله چالش‌های موجود در این

^{۴۷}Maximum Likelihood

^{۴۸}Variational Message Passing

^{۴۹}Gibbs Sampling

مجموعه‌داده می‌توان به وجود زمینه‌های متعدد و پیچیده در تصاویر، تنوع دسته‌های مختلف اجسام موجود، تنوع اندازه اجسام موجود از یک دسته، تنوع حالت اجسام، تنوع تعداد نمونه‌های یک جسم در یک تصویر و کوچک بودن بیش از اندازه ابعاد اجسام در تصویر اشاره کرد. شکل ۱۰ نمونه‌ای از تصاویر موجود در این مجموعه‌داده را نمایش می‌دهد.



شکل ۱۰: نمونه تصاویر موجود در مجموعه‌داده مورد استفاده [۲]

استفاده از مدل کامل ارائه شده در این پژوهش، منجر به تشخیص صحیح ۷۳.۴٪ از تصاویر شده است. شکل ۱۱ ماتریس درهم‌ریختگی^۵ مربوط به این مدل را نمایش می‌دهد. همان‌طور که در این ماتریس مشخص است، کمترین نرخ تشخیص در بین دسته‌های ورزشی موجود در این مدل، ۵۲٪ و بیشترین نرخ تشخیص

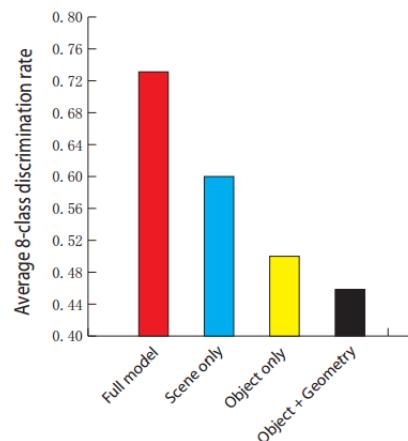
^۵. Confusion Matrix

۹۲٪ است.

	Average Perf.= 73.4%							
bocce	.52	.02	.17	.05		.25		
badminton		.92			.03		.05	
polo	.27		.62	.02		.10		
rowing	.03	.02	.03	.80		.12		
snowboarding		.18			.77		.03	.02
croquet	.27	.03	.07	.12		.52		
sailing		.13			.07		.80	
rockclimbing	.05			.02		.02		.92

شکل ۱۱: ماتریس درهم‌ریختگی مدل کامل ارائه شده برای مجموعه‌داده شامل ۸ دسته تصویر ورزشی. [۲]

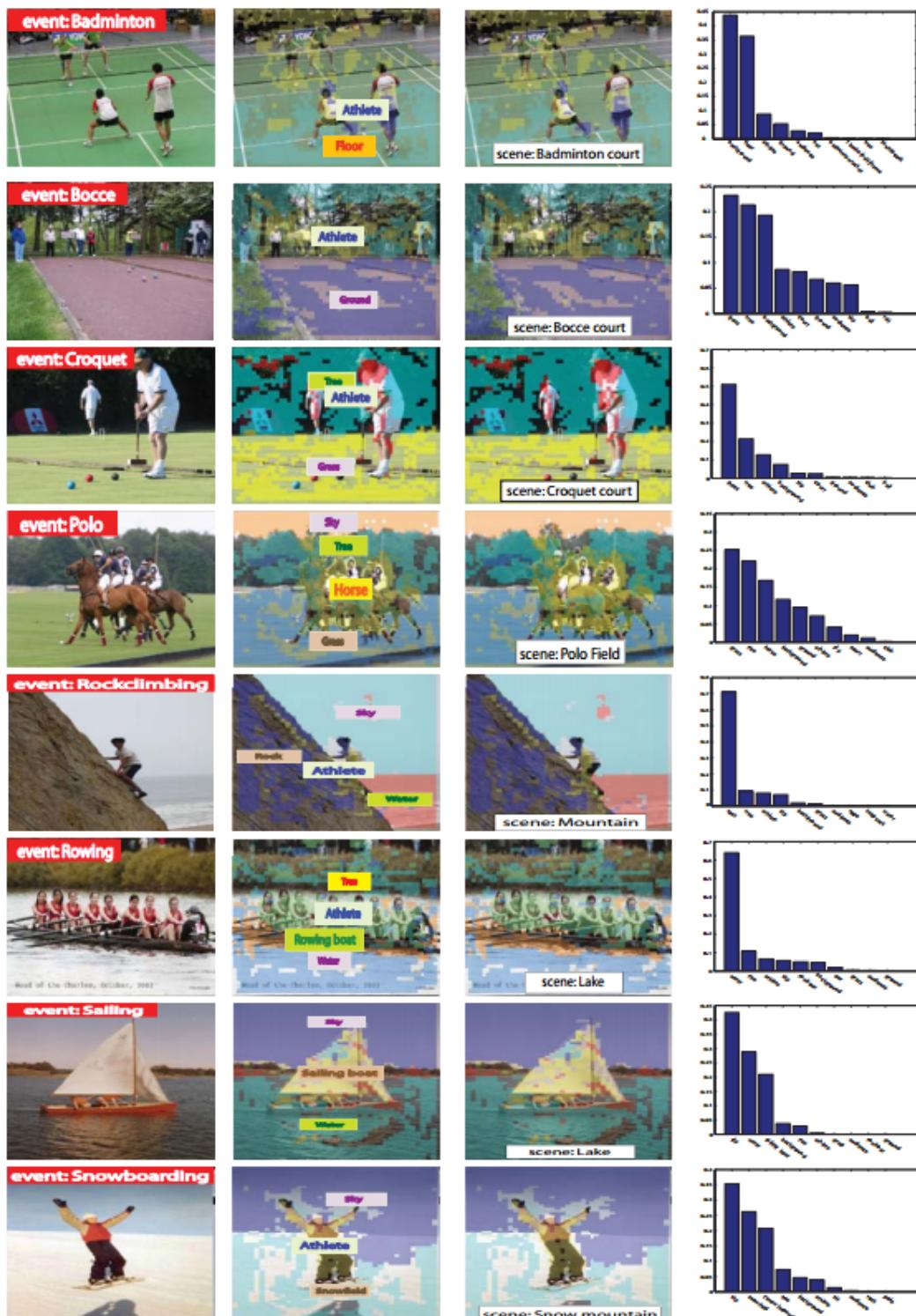
بسته به میزان استفاده از اطلاعات مختلف استخراج شده برای استنتاج، مدل‌های مختلفی به وجود می‌آیند که در شکل ۱۲ نتایج عملکرد هریک از این مدل‌ها با مدل‌های دیگر مقایسه شده است. همان‌طور که در شکل ۱۲ مشخص است، بهترین کارایی مربوط به مدل کامل است. در صورتی که در مدل، فقط از اطلاعات مربوط به صحنه استفاده شود، نتایج بدست آمده اگرچه با نتایج مدل قابل مقایسه نیست، از نتایج مدل مبتنی بر اطلاعات جسم بهتر است.



شکل ۱۲: نتیجه مقایسه مدل‌های مختلف بوجود آمده بسته به سطح اطلاعات مورد استفاده برای استنتاج. [۲]

شکل ۱۳ نتایج نهایی به‌دست آمده از مدل را نمایش می‌دهد. در این شکل، تصاویر موجود در هر سطر نماینده تصاویر موجود در یکی از دسته‌های ورزشی هستند. ستون اول برچسب به‌دست آمده از رخداد

موجود در تصویر، ستون دوم مربوط به تشخیص داده شده موجود، ستون سوم برچسب اختصاص داده شده مربوط به دسته صحنه و ستون چهارم توزیع مرتب شده اجسام به شرط رخداد را به نمایش می‌گذارند. در نمودارهای موجود در ستون چهارم، محور افقی شامل نام اجسام و محور عمودی مقدار توزیع را نمایش می‌دهد.



شکل ۱۳: نتایج نهایی به دست آمده از مدل بر روی تصاویر. [۲]

۴-۲ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق

علاوه بر فعالیت‌هایی که در زمینه تولید خودکار شرح بر تصاویر با رویکرد احتمالاتی انجام شده‌اند، تعداد زیادی از پژوهش‌گران تلاش می‌کنند تا با استفاده از روش‌های مبتنی بر شبکه‌های عصبی با این چالش روبرو شوند. در این بخش تعدادی از پژوهش‌هایی را که با استفاده از شبکه‌های عصبی سعی در درک صحنه‌های موجود در تصاویر دارند را مورد بررسی قرار می‌دهیم. شایان ذکر است، در این بخش تنها به بررسی بخشی از پژوهش‌ها که مربوط به درک صحنه است می‌پردازیم و بخش‌هایی از این پژوهش‌ها که مربوط به تولید جملات زبان طبیعی متناسب با تصویر و صحنه درک شده است را در فصل تولید جملات زبان طبیعی بررسی خواهیم نمود.

یکی از مهم‌ترین عملیات‌هایی که به نحوی در تمام پژوهش‌های قبلی وجود داشت، اختصاص یک معنا به قطعه‌های مختلف یک تصویر است. این چالش، در پژوهش‌های مرتبط با تولید خودکار شرح بر تصاویر که با استفاده از روش‌های مبتنی بر شبکه‌های عصبی به دنبال حل مشکل هستند نیز مطرح است. در ابتدا به بررسی یکی از روش‌های اختصاص معنا به هر قطعه از تصویر می‌پردازیم.

۱-۴-۲ اختصاص معنا به قطعه‌های مختلف تصویر [۲۵]

در پژوهش [۲۵] روشی ارائه شده است که با استفاده از یک شبکه عصبی کانولوشنی عمیق، علاوه بر این که می‌تواند یک تصویر را به شکل پایین به بالا، در قالب نواحی سلسله‌مراتبی قطعه‌بندی کند، قادر به استفاده به عنوان یک شبکه از پیش آموزش دیده شده در پژوهش‌های مرتبط دیگر باشد.

فرایند تشخیص اجسام در این پژوهش از سه بخش اصلی تشکیل شده است:

۱. طرح پیشنهاداتی برای نواحی به طور مستقل از دسته‌بندی^{۵۱}

۲. یک شبکه عصبی عمیق کانولوشنی که وظیفه استخراج ویژگی برای هر ناحیه را بر عهده دارد (طول بردار ویژگی استخراج شده برای تمام نواحی یکسان است).

۳. مجموعه‌ای از ماشین‌های بردار پشتیبان خطی مخصوص هر دسته

در ادامه به بررسی نحوه پیشنهاد نواحی و شبکه عصبی کانولوشنی عمیق مورد استفاده در ای پژوهش می‌پردازیم.

۱. طرح پیشنهاد نواحی

روش‌های مختلفی برای پیشنهاد نواحی ارائه شده‌اند که در اینجا از روشی موسوم به جستجوی انتخابی^{۵۲} استفاده می‌شود. نسخه‌های مختلفی از این روش ارائه شده است. نسخه ارائه شده در پژوهش [۲۶]، یکی از سریع‌ترین نسخه‌های ارائه شده است که در این بخش از همین روش استفاده می‌شود.

در پژوهش [۲۶] دو ویژگی مطرح شده است که یک جستجوی انتخابی برای ارائه نواحی معنایی تصویر باید آن‌ها را داشته باشد. ویژگی اول این است که اجسام موجود در فضا می‌توانند در هر اندازه‌ای باشند و در نتیجه نواحی ارائه شده باید بتوانند ابعاد مختلف داشته باشند. این ویژگی عموماً با روش‌های سلسله‌مراتبی

^{۵۱}Category-independent region proposals

^{۵۲}Selective Search

قابل دستیابی است. ویژگی دوم این است که نواحی مختلف باید براساس ویژگی‌های مختلفی تولید شوند. در صورتی که یک ویژگی مثل رنگ، بافت، روشنایی یا مواردی از این دست، به عنوان تنها ویژگی برای تشخیص نواحی به کار گرفته شود، الگوریتم قادر به ارائه نواحی مناسب در شرایط مختلف نخواهد بود. بنابراین ترکیب چند معیار و ویژگی باید برای تشخیص نواحی مورد استفاده قرار بگیرد.

برای دستیابی به ویژگی اول، ابتدا نواحی اولیه کوچکی روی تصویر ایجاد می‌شود. سپس با اتخاذ یک روش حریصانه و تعریف یک معیار شباخت بین نواحی همسایه، ناحیه‌هایی که شباخت زیادی با یکدیگر دارند و همسایه هستند، با هم ترکیب شده و یک ناحیه بزرگ‌تر ساخته می‌شود. به این ترتیب یک روش سلسله‌مراتبی برای ساخت نواحی با ابعاد مختلف به دست می‌آید. برای دستیابی به ویژگی دوم، از فضاهای رنگی مختلف، معیارهای شباخت مختلف و نواحی اولیه متفاوت و ترکیب پاسخ این ویژگی‌ها با هم برای ارائه نواحی و ترکیب نواحی کوچک‌تر استفاده می‌شود.

۲. شبکه عصبی کانولوشنی عمیق (استخراج ویژگی‌ها)

در این بخش از یک شبکه عصبی کانولوشنی عمیق از پیش‌آموزش دیده برای استخراج ویژگی از هر ناحیه ارائه شده در قسمت قبل، استفاده می‌شود. بردار ویژگی استخراج شده برای هر ناحیه یک بردار شامل ۴۰۹۶ مولفه است که خروجی شبکه کریشفسکی^{۵۳} آزمایش شده در چالش دسته‌بندی اجسام مسابقه ImageNet است. اطلاعات دقیق درباره این شبکه عصبی در پژوهش [۲۷] در دسترس است.

شبکه عصبی کانولوشنی عمیق ارائه شده در این پژوهش با استفاده از یک مجموعه‌داده^{۵۴} آموزش دیده شده است. از این شبکه عصبی که تحت عنوان RCNN^{۵۵} شناخته می‌شود می‌توان به عنوان یک شبکه از پیش‌آموزش دیده استفاده کرد.

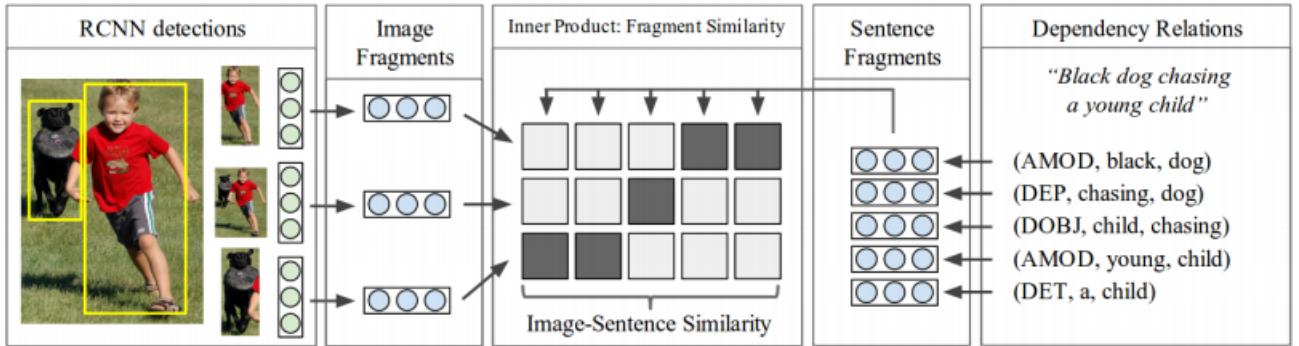
۲-۴-۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دوطرفه جملات و تصاویر [۲۸]

مدل ارائه شده در این پژوهش، مدلی است که قادر به نگاشت دوطرفه تصاویر و جملات به یکدیگر است. شکل ۱۴ طرح‌واره‌ای از این مدل را نمایش می‌دهد. ورودی مدل در سمت چپ، تصاویر و در سمت راست، جملات هستند. در این مدل، ابتدا تصاویر ورودی با استفاده از یک شبکه عصبی RCNN تبدیل به نواحی مختلف شده و برای هر ناحیه یک بردار ویژگی ۴۰۹۶ بعدی استخراج می‌شود. سپس با اعمال روش خاصی روی جملات ورودی از سمت راست (که در بخش تولید جملات زبان طبیعی به بررسی آن خواهیم پرداخت) قطعات مختلف موجود در جملات نیز استخراج شده و بین هر قطعه از جمله با تمام نواحی استخراج شده از تصویر یک معیار شباخت محاسبه می‌شود و شبیه‌ترین قطعه جمله با ناحیه مربوط به خود در تصویر، جفت می‌شوند.

^{۵۳}Krizhevsky

^{۵۴}ILSVRC 2012

^{۵۵}Regional Convolutional Neural Network



شکل ۱۴: مدل استفاده شده برای نگاشت دوطرفه تصاویر و جملات به یکدیگر با استفاده از شبکه عصبی عمیق کانولوشنی. [۲۸].

در این پژوهش پس از ناحیه‌بندی تصویر توسط شبکه RCNN، برای هر تصویر ۱۹ ناحیه استخراج می‌شود. این ۱۹ ناحیه در کنار تصویر اصلی، یک مجموعه شامل ۲۰ تصویر ایجاد می‌کنند که در پردازش‌های بعدی مورد استفاده قرار خواهند گرفت. در این مرحله باید تمام تصاویر موجود را با استفاده از یک نگاشت به فضای برداری ویژگی‌ها تبدیل نمود. برای این کار از رابطه I_b استفاده می‌شود. در این رابطه، $RCNN_{\theta_c}(I_b)$ مجموعه تمام پیکسل‌های موجود در ناحیه b ، شبکه عصبی آموزش‌دیده است که در آن θ_c مجموعه پارامترهای بهینه موجود در شبکه است. بردار حاصل ν_i برای تصویر i ، بردار نگاشت تصویر به فضای معنایی خواهد بود که محاسبه مقادیر آن مبتنی بر پیشنهاد نواحی معنایی مختلف و محاسبه ویژگی‌های مختلف روی هر ناحیه است.

$$\nu = W_m[RCNN_{\theta_c}(I_b)] + b_m \quad (14)$$

از طرفی با در نظر گرفتن بردار s_j به عنوان بردار حاصل از نگاشت جمله z به فضای معنایی و در نظر گرفتن ضرب داخلی به عنوان شباهت، $s_j^T \cdot \nu_i$ معیار شباهت بین یک تصویر و یک جمله را تعریف می‌کند. با توجه به توضیحات ارائه شده، می‌توان تابع هدف را برای شبکه کلی معادل سیستم ارائه داد. دو هدف اصلی در این شبکه قابل تعریف است:

۱. رتبه‌بندی سراسری تصاویر و جملاتی که در فرایند محاسبات شبکه عصبی بیشترین شباهت را با یکدیگر دارند باید در واقعیت هم بیشترین شباهت و ارتباط را داشته باشند.
۲. هم ترازسازی ناحیه‌ای^{۵۶} نواحی استخراج شده تصویر و عبارات استخراج شده جملات که در محاسبات شبکه عصبی بیشترین شباهت را با یکدیگر دارند، باید در واقعیت هم بیشترین شباهت و ارتباط را داشته باشند.

با توجه به مطالب گفته شده، می‌توان تابع هدف کلی را مطابق با رابطه ۱۵ تعریف کرد. در این رابطه، Θ مجموعه پارامترهای شبکه عصبی شامل $\{W_m, b_m, \theta_c, W_e, W_R\}$ است (پارامترهای W_e و W_R مربوط به بخش

^{۵۶}Fragment Alignment

تحلیل جمله هستند که در فصل مربوطه بررسی خواهند شد). C_F تابع هدف همترازسازی ناحیه‌ای، C_G تابع هدف سراسری، α و β دو ابرپارامتر^{۵۷} (با آزمون و خطا تعیین می‌شوند) و $\|\cdot\|_2^2$ یک عبارت تنظیم‌کننده^{۵۸} هستند.

$$C(\Theta) = C_F(\Theta) + \beta C_G(\Theta) + \alpha \|\Theta\|_2^2 \quad (15)$$

در ادامه به تعریف هریک از اهداف بیان شده می‌پردازیم.

۱. همترازسازی ناحیه‌ای

هدف از همترازسازی ناحیه‌ای این است که اگر عبارتی از یک جمله با یک تصویر شباهت زیادی پیدا کرد، حداقل یک ناحیه از تصویر وجود داشته باشد که نمایش‌دهنده این عبارت باشد و بقیه نواحی تصویر، ارتباط کمی با این عبارت داشته باشند. به عبارت بهتر، در صورتی که شباهت یک عبارت از یک جمله با یک تصویر از حدی بیشتر شد، شباهت حداقل یکی از نواحی موجود در تصویر با این عبارت زیاد شده و شباهت بقیه نواحی تصویر با آن کم شود. این فرض در سه حالت، رد می‌شود. اولین حالت، حالتی است که در آن ناحیه‌ای که در واقعه نمایش‌دهنده عبارت است، توسط RCNN تشخیص داده نشده باشد. دومین حالت، حالتی است که عبارت موجود به هیچ بخشی از ویژگی‌های بصری تصویر اشاره نکند و آخرین حالت، حالتی است که عبارت توصیف‌کننده، در هیچ یک از تصاویر دیگر تکرار نشده باشد در صورتی که ممکن است تصاویر دیگری هم وجود داشته باشند که شامل ویژگی‌های بصری متناظر با عبارت باشند. با توجه به شرایطی که فرض در آن‌ها نقض می‌شود، می‌توان آن را یک فرض خوب تلقی کرد که در اکثر موارد عملکرد خوبی دارد.

رابطه ۱۶ تابع هدف همترازسازی ناحیه‌ای را تعریف می‌کند. در این رابطه، y_{ij} برای تصویر i ام و جمله j ام در صورتی که با هم در مجموعه‌داده حضور داشته باشند، $+1$ و در غیر این صورت، -1 خواهد شد.

$$C_{\circ}(\Theta) = \sum_i \sum_j \max(0, 1 - y_{ij} \nu_i^T \cdot s_j) \quad (16)$$

تابع C_{\circ} تعریف شده، باعث می‌شود در حالاتی که تصویر و عبارت، در مجموعه‌داده، با یکدیگر وارد شده باشند امتیاز تابع هدف بیشتر از $+1$ شود و در غیر این صورت از -1 کمتر شود. شکل ۱۵، دو نمونه از تصاویر و جملات موجود در مجموعه‌داده را نمایش می‌دهد. C_{\circ} در سلول‌هایی که با رنگ قرمز مشخص شده‌اند، امتیاز را به سمت کمتر از -1 حرکت می‌دهد و در بقیه سلول‌ها به سمت بیشتر از $+1$.

به عبارت بهتر، C_{\circ} یک امتیاز برای مجموع تفاوت‌های نواحی مختلف از تصاویر با عبارات مختلف جملات است. به دلیل این‌که این معیار، باعث دیده نشدن موارد کمیاب می‌شود، با متغیر گرفتن پارامتر y_{ij} سعی در یافتن کمترین مقدار آن می‌کنیم. رابطه ۱۷ معیار متناظر با هدف کلی همترازسازی ناحیه‌ای را بیان می‌کند.

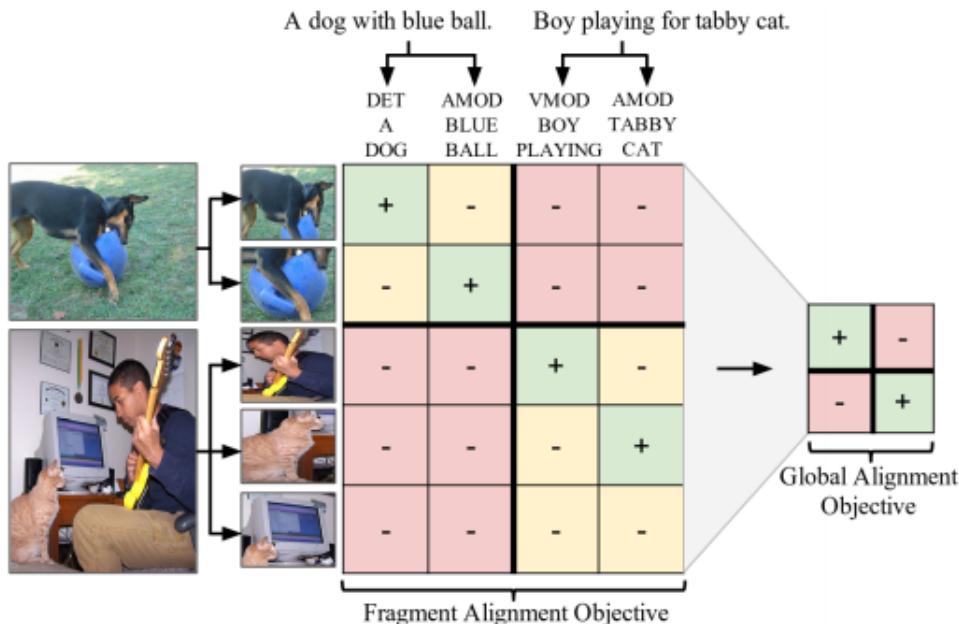
^{۵۷}Hyperparameter

^{۵۸}Regularization Term

$$C_F(\Theta) = \min_{y_{ij}} C_\circ(\Theta)$$

$$\text{s.t. } \sum_{i \in p_j} \frac{y_{ij} + 1}{2} \geq 1 \wedge \quad y_{ij} = -1, \forall i, j; m_\nu(i) \neq m_s(j) \wedge y_{ij} \in \{+1, -1\} \quad (17)$$

در این رابطه، p_j مجموعه تصاویر موجود در کیسه مثبت^{۵۹} مربوط به عبارت زام است. شایان ذکر است، تنها تصاویری که در مجموعه‌داده همراه با عبارت زام مشاهده شده‌اند در کیسه مثبت مربوط به این عبارت قرار می‌گیرند و بقیه تصاویر در کیسه منفی^{۶۰} این عبارت قرار می‌گیرند. $m_\nu(i)$ و $m_s(j)$ به ترتیب، شماره تصویر و عبارت را در مجموعه‌داده مشخص می‌کنند.



شکل ۱۵: دو نمونه از تصاویر و جملات مرتبط با آن‌ها و نتایج عملکرد اهداف تعریف شده روی آن‌ها. سطرها نمایش‌دهنده نواحی مختلف تصویر و ستون‌ها نمایش‌دهنده مفعولهای مختلف جملات هستند. سلول‌های قرمز رنگ حالتی هستند که در آن‌ها $y_{ij} = 1$ ، سلول‌های زرد نمایش‌دهنده اعضای کیسه‌های مثبت هستند که در آن‌ها $y_{ij} = -1$ است. [۲۸]

۲. رتبه‌بندی سراسری

هدف از رتبه‌بندی سراسری این است که شباهت بین یک تصویر و یک جمله، بیشینه شود اگر و تنها اگر تصویر و جمله در واقعیت نیز بیشترین شباهت را به یکدیگر داشته باشند. برای این منظور، ابتدا یک امتیاز شباهت بین یک تصویر و یک جمله تعریف می‌شود. این امتیاز مطابق با رابطه ۱۸ تعریف شده و برابر است با میانگین امتیاز شباهت دوبعدی نواحی مختلف تصویر با عبارات مختلف جمله.

$$S_{kl} = \frac{1}{|g_k|(|g_l| + n)} \sum_{i \in g_k} \sum_{j \in g_l} \max(\circ, \nu_i^T \cdot s_j) \quad (18)$$

^{۵۹}Positive Bag

^{۶۰}Negative Bag

از آن جا که برای دسته‌بندی از روش mi_SVM استفاده می‌شود، تمام امتیازها به صفر محدود می‌شوند. مقدار n که در مخرج کسر اضافه شده است، به صورت تجربی و با آزمون و خطا به دست آمده که نتایج را بهبود می‌بخشد. مقدار پیشنهاد شده در پژوهش، $n = 5$ است. تابع کلی هدف سراسری مطابق با رابطه ۱۹ تعریف می‌شود.

$$C_G(\Theta) = \sum_k (\sum_l \max(0, S_{kl} - Skk + \Delta) + \sum_l \max(0, S_{lk} - Skk + \Delta)) \quad (19)$$

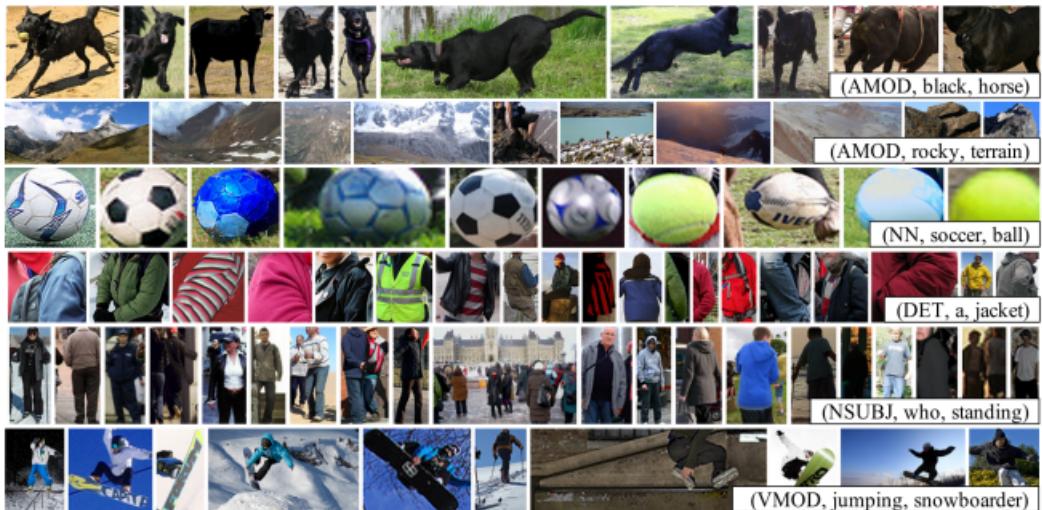
در رابطه ارائه شده، Δ یک ابرپارامتر است که با آزمون و خطا به دست می‌آید. عبارت اول درون پرانتز بیان کننده امتیاز تصویر و عبارت دوم بیان کننده امتیاز جمله هستند.

شکل ۱۶ نتایج روش پیشنهاد شده در این پژوهش را ارائه می‌دهد. همان‌طور که در شکل مشخص است، این شبکه قادر به تشخیص اجسام مختلف در تصویر و تولید یک سه‌تایی متناظر هر جسم (ناحیه معنایی) مبتنی بر جملات موجود در مجموعه‌داده مورد استفاده است.



شکل ۱۶: نتایج نهایی شبکه عصبی ارائه شده. برای هر ناحیه معنایی از تصویر، یک سه‌تایی مبتنی بر جملات موجود در مجموعه‌داده تولید شده است. همین‌طور ۵ جمله تولید شده برای هر تصویر به ترتیب امتیاز، درج شده‌اند. [۲۸]

به علاوه، با توجه به مدل ارائه شده و نگاشت دوطرفه موجود بین تصاویر و جملات، می‌توان با ورودی دادن یک جمله، تصاویر مربوط به آن جمله را استخراج نمود. شکل ۱۷ با ثابت در نظر گرفتن جملات، تصاویر مربوط به هر جمله را استخراج و نمایش داده است. هر سطر از این شکل، نمایش‌دهنده تصاویر استخراج شده مرتبط با جمله موجود در آن سطر است.



شکل ۱۷: نتایج حاصل از جستجوی جملات، با ورودی دادن یک جمله، شبکه عصبی ارائه شده در این پژوهش، قادر به استخراج تصاویر مربوط به آن جمله است. [۲۸]

روش ارائه شده در این پژوهش، به طور کامل و دقیق در پژوهش [۲۹] هم مورد استفاده قرار گرفته است، با این تفاوت که در فرایند تحلیل جمله، تغییراتی ایجاد شده است. جزئیات این روش در فصل تولید جملات زبان طبیعی مورد بررسی قرار خواهد گرفت.

۲-۵ جمع‌بندی

اولین مرحله از فرایند تولید خودکار شرح برای تصاویر، مرحله درک صحنه است. در این مرحله، تصاویر ورودی تحت عملیات مختلفی به فضای معنایی نگاشت می‌شوند. فضای معنایی در اینجا، می‌تواند فضای شامل میدان‌های اطلاعاتی از پیش تعیین شده (مانند فضای سه‌تایی‌های «جسم، رخداد، صحنه») یا فضای بردار ویژگی‌ها باشد. روش‌های مختلفی برای نگاشت تصویر ورودی به فضای معنایی ارائه شده است که به طور کلی می‌توان عموم آن‌ها را به دو بخش تقسیم کرد:

۱. روش‌های مبتنی بر مدل‌های گرافی احتمالاتی

در این روش‌ها با استفاده از مدل‌های استاندارد گرافی احتمالاتی موجود یا با ارائه یک مدل گرافی احتمالاتی، تصویر ورودی به فضای معنایی نگاشت می‌شود. در روش‌های مبتنی بر این مدل‌ها، با ارائه یک توزیع احتمال برای نقاط مختلف در فضای معنایی، محتمل‌ترین نقطه برای تصویر به عنوان نقطه نظر تصویر، انتخاب می‌شود.

(آ) مدل میدان تصادفی مارکف

یک نمونه از روش‌های مبتنی بر مدل میدان تصادفی مارکف که برای درک صحنه از آن استفاده شده است، در پژوهش [۱۸] ارائه شده است. درک صحنه در این پژوهش با ارائه یک سه‌تایی «جسم،

فعالیت، صحنه» به ازای هر تصویر، تعریف شده است. مبتنی بر همین تعریف، یک مدل میدان تصادفی مارکف شامل سه گره که دوبهدو به هم متصل هستند، تعریف شده است. هر یک از گرههای موجود در این مدل، نماینده یکی از میدان‌های سه‌گانه تعریف شده در فضای معنایی هستند. با تعریف توابع پتانسیل مختلف روی هر گره و توابع پتانسیل مختلف روی هر یال، یک تابع توزیع توام برای تمام متغیرهای تصادفی موجود در مدل ارائه شده است.

با محاسبه مقادیر پتانسیل برای تصاویر مختلف موجود در مجموعه آموزشی و با استفاده از یک ماشین بردار پشتیبان، بردارهای ویژگی شاخص برای هر گره محاسبه می‌شوند. از این بردارهای ویژگی بعده برای انطباق تصاویر با مقادیر مختلف در هر گره استفاده می‌شود.

در این پژوهش، با یافتن نزدیک‌ترین همسایه‌های یک تصویر بر حسب معیار شباهت با بردارهای ویژگی شاخص و میانگین‌گیری روی مقادیر هر گره، بهترین انطباق تصویر و نقاط فضای معنایی به دست می‌آید. به این ترتیب، برای هر تصویر ورودی، می‌توان نقطه نظری در فضای معنایی را مشخص کرد.

(ب) مدل میدان تصادفی شرطی

در پژوهش [۱۷] یک مدل میدان تصادفی شرطی سلسله‌مراتبی برای درک صحنه ارائه شده است که شامل دو سطح انتزاع است. برای گرههای موجود در هریک از سطوح انتزاع مدل، یک دسته متغیر تصادفی تعریف شده و برای کل مدل سه نوع تابع پتانسیل مختلف معرفی شده است.

اولین دسته از توابع پتانسیل معرفی شده در این بخش، توابع پتانسیل قطعه‌بندی یگانی هستند که به منظور یکپارچه‌سازی نقاط داخل یک قطعه تعریف شده‌اند. توابع پتانسیل دیگری برای انطباق بین متغیرهای تصادفی موجود در بین دو سطح انتزاع تعریف شده‌اند که در صورت مغایرت مقادیر اختصاص داده شده به متغیرهای موجود بین دو سطح، مقدار λ – و در غیر این صورت مقدار صفر دارند. این توابع در شرایطی که مقادیر متغیرهای موجود در دو سطح با هم یکسان نباشد، یک مقدار جریمه به تابع هدف اضافه می‌کنند. آخرین دسته از توابع پتانسیل مورد استفاده، برای انطباق تصویر با دسته تشخیص داده شده اجسام تعریف شده است که توسط فلزنسوالب ارائه شده و به روش دی‌پی‌ام مشهور است.

(ج) سایر مدل‌های گرافی احتمالی در پژوهش [۲]، یک مدل گرافی احتمالی مولد برای نگاشت تصویر به فضای معنایی ارائه شده است. در این مدل، از دو سطح تصویر استفاده شده است؛ تصویر سطح جسم و تصویر سطح صحنه. برای تصویر سطح صحنه، یک متغیر تصادفی، بیان‌کننده دسته صحنه و برای تصویر سطح جسم دو متغیر تصادفی، بیان‌کننده دسته و شکل جسم، ارائه شده است. روابط بین متغیرهای تصادفی در این پژوهش، براساس نحوه تولید متغیرهای تصادفی و روابط منطقی موجود بین آن‌ها طراحی شده‌اند.

تصویر ورودی در این پژوهش، ابتدا به نواحی کوچک 10×10 تقسیم می‌شود و مطابق با روش توضیح داده شده، مقدار توابع پتانسیل مختلف برای هر کدام از متغیرهای تصادفی، در هر ناحیه، محاسبه می‌شود. در این پژوهش، یک تابع احتمال شرطی برای متغیرهای تصادفی ارائه شده است که در مرحله

استنتاج، با استفاده از روش تخمین بیشترین احتمال، برچسب‌های هر تصویر مشخص می‌شوند.

۲. روش‌های مبتنی بر استفاده از شبکه‌های عصبی کانولوشنی عمیق

در این روش‌ها، با ارائه یک شبکه عصبی کانولوشنی عمیق و تعریف کردن تابع هدف برای شبکه، تابع نگاشت تصویر و فضای معنا تشکیل می‌شود. پس از ارائه توابع هدف برای هر شبکه، با بهینه‌سازی آن تابع، پارامترهای موجود در شبکه آموزش داده می‌شوند.

در پژوهش [۲۵]، روشی ارائه شده است که طی آن یک تصویر، به نواحی کوچک‌تر تقسیم می‌شود به طوری که هر ناحیه به وجودآمده، به طور یکپارچه، حاوی یک جسم باشد و هر جسم تنها در یک ناحیه قرار بگیرد. این روش موسوم به روش RCNN است. در این روش، دو ویژگی برای یک ناحیه‌بندی خوب در تصاویر ارائه شده است و پیرو این ویژگی‌ها، روشی برای طرح نواحی پیشنهادی در یک تصویر که دارای این دو ویژگی باشد، ارائه شده است.

ویژگی مطرح شده اول برای ناحیه‌بندی تصاویر این است که، ناحیه‌های ایجاد شده در هر تصویر، می‌توانند در ابعاد مختلف وجود داشته باشند زیرا اجسام موجود در تصاویر، ممکن است اندازه و تعداد متفاوتی داشته باشند. دومین ویژگی برای یک ناحیه‌بندی خوب، این است که معیار انتخاب نواحی نباید برای تمام تصاویر، یکسان در نظر گرفته شود؛ زیرا معیارهای مختلف برای ناحیه‌بندی تصاویر در شرایط مختلف، رفتارهای متفاوتی از خود نشان می‌دهند. بنابراین باید از معیارهای مختلف برای تعیین نواحی استفاده نمود.

در این پژوهش، ابتدا تصاویر مطابق با یک معیار اولیه، به مجموعه‌ای از نواحی اولیه تقسیم می‌شوند. سپس با استفاده از معیارهای مختلف مانند فضاهای رنگی مختلف، معیارهای شباht مختلف و نقاط اولیه متفاوت، با پیروی از یک روش حریصانه، نواحی کوچک‌تر که به یکدیگر شبیه‌تر هستند با هم ترکیب شده و نواحی بزرگ‌تر را می‌سازند. نواحی ایجاد شده در این روش، سپس به یک شبکه عصبی کانولوشنی عمیق داده می‌شوند و برای هر ناحیه، یک بردار ویژگی ۴۰۹۶ بعدی ایجاد می‌شود که هر ناحیه با آن بازنمایی شود.

در پژوهش [۲۸] با استفاده از روش RCNN و تعریف دو تابع هدف دیگر برای شبکه، روشی ارائه شده است که طی آن بتوان تصاویر و جملات را به طور دوطرفه به یکدیگر نگاشت کرد. توابع هدف تعریف شده در این پژوهش، دو تابع مختلف هستند. اولین تابع هدف، یک تابع هدف سراسری است. این تابع به این منظور تعریف شده است که تصاویر و جملاتی که مطابق با محاسبات شبکه عصبی ارائه شده، بیشترین شباهت را با یکدیگر دارند، در واقعیت هم شبیه‌ترین تصاویر و جملات به یکدیگر باشند. تابع هدف دوم برای این شبکه به این شکل تعریف شده است که نواحی استخراج شده از تصویر و عبارات استخراج شده از جملات که در روش ارائه شده، بیشترین شباهت را به یکدیگر دارند، در واقعیت هم بیشترین شباهت و ارتباط را با یکدیگر داشته باشند.

در این پژوهش، تصاویر ورودی با استفاده از روش RCNN به نواحی مختلف تقسیم شده و ۱۹ ناحیه با بیشترین اطمینان از بین این نواحی انتخاب می‌شود. این ۱۹ ناحیه به همراه خود تصویر به عنوان ۲۰ تصویر مختلف مورد استفاده قرار می‌گیرند. جملات ورودی با استفاده از روشی که در فصل تولید جملات زبان طبیعی توضیح داده خواهد شد، به عبارات مختلف تقسیم می‌شوند و بین هر عبارت استخراج شده و هر یک از ۲۰ تصویر موجود، یک معیار شباهت محاسبه شده و بیشترین شباهت‌ها با هم درنظر گرفته می‌شوند.

معیار شباهت مورد استفاده در این روش، ضرب داخلی بین بردارهای ویژگی عبارات و نواحی است. عبارات و نواحی که بیشترین شباهت را با یکدیگر دارند برای تولید جمله به مرحله بعد، ارسال می‌شوند.

٣ فصل سوم

تولید جمله

۱-۳ تولید جمله

در این بخش، به بررسی چالش تولید جمله و روش‌های پیشنهاد شده برای حل این چالش، در طول زمان، خواهیم پرداخت. در ابتداء ایده‌های اولیه در این مسیر را بیان نموده و به طور خلاصه مورد بحث قرار خواهیم داد و در ادامه به بررسی تفصیلی روش‌های مبتنی بر استفاده از شبکه‌های عصبی بازگشتی در تولید جملات زبان طبیعی، می‌پردازیم.

چالش تولید جمله، متوجه ساخت جملاتی به زبان طبیعی است، به طوری که از لحاظ دستور زبان، املا و معنا صحیح باشند. از طرفی با توجه به هدف اصلی ما که تولید شرح بر تصاویر است، جملات تولید شده باید علاوه بر این که شرط صحت مذکور را ارضا می‌کنند، با تصویر ورودی، صحنه توصیف شده در تصویر و رخداد به نمایش کشیده شده، هم‌خوانی داشته باشند. تضمین این هم‌خوانی از جمله معضلات دیگری است که باید برای آن چاره‌ای اندیشید.

مساله تولید خودکار جملات زبان طبیعی، یکی از مسائلی است که از دیرباز در هوش مصنوعی مطرح بوده و دارای کاربردهای فراوانی است. به عنوان یک تعریف دقیق از این مساله، می‌توان به «فرایند تولید جملات زبان طبیعی با استفاده از داده‌های غیر قابل تفسیر برای کاربران عادی در جهت افزایش قابلیت تفهیم داده به آن‌ها [۳۰]» اشاره کرد. داده‌های اولیه که جملات با استفاده از آن‌ها تولید می‌شوند، می‌توانند شامل انواع داده‌های غیر متنی از جمله نمودارها، تصاویر، اعداد و مواردی از این دست باشند.

در ادامه، ابتداء کاربردهای مساله تولید خودکار جملات زبان طبیعی^{۶۱}، خواهیم پرداخت.

۲-۳ کاربردها [۳۰]

کاربردهای بسیار زیاد و متنوعی برای تولید خودکار جملات زبان طبیعی با استفاده از داده‌های غیر متنی، ارائه شده است. در این بخش، برای روشن شدن اهمیت این مساله، تعدادی از این کاربردها را بیان خواهیم نمود.

۱. تولید خودکار شرح بر پیش‌بینی وضع آب و هوا با استفاده از نقشه‌های گرافیکی آب و هوا
۲. تولید خلاصه‌ای در باره داده‌های آماری استخراج شده از یک پایگاه داده
۳. توصیف یک زنجیره استدلالی، منتج از فرایند تصمیم‌گیری یک سیستم خبره
۴. تولید پاسخ برای پرسش‌ها در مورد یک جسم در یک سامانه مبتنی بر دانش

^{۶۱}Natural Language Generation (NLG)

موارد ذکر شده در بالا، تنها نمونه‌ای از کاربردهای وسیع این مساله در زندگی‌های روزمره را نمایش می‌دهد.

۳-۳ روش تولید زبان طبیعی

یکی از حوزه‌های پویا در زمینه هوش مصنوعی و پردازش متن، حوزه تولید زبان طبیعی است. این حوزه شامل فعالیت‌هایی است که به تولید جمله زبان طبیعی متناظر با داده‌های قابل تفسیر برای ماشین مانند پایگاه‌های دانش^{۶۲} و یا قالب‌های منطقی^{۶۳} و مواردی از این دست، می‌پردازد. روش‌های کلی که در چارچوب کاری^{۶۴} پژوهش‌های این زمینه مورد استفاده قرار می‌گیرد، به طور کلی شامل مراحل زیر است [۳۰].

۱. برنامه‌ریزی متن^{۶۵}

در این مرحله، ابتدا محتوا مورد نیاز، انتخاب می‌شود و چارچوب کلی برای کل متن، طرح ریزی می‌شود. انتخاب محتوا و طرح ریزی چارچوب کلی متن، با تکیه بر کاربرد مورد نظر در پژوهش و داده‌هایی که نیاز به تفسیر زبانی دارند، انجام می‌شود.

۲. برنامه‌ریزی جمله^{۶۶}

در این مرحله، کلمات مورد نیاز انتخاب می‌شوند، عبارات زبانی مناسب تولید می‌شوند و به طور دقیق کنارهم قرار می‌گیرند تا جملات را تشکیل دهند. انتخاب کلمات و ساخت عبارات در این بخش، بر اساس طرح کلی پی‌ریزی شده برای متن در مرحله قبل، انجام می‌شود. کنارهم قرار دادن عبارات زبانی مورد نیاز نیز، با استفاده از قواعد دستور زبان که عموماً به شکل پایگاه دانش موجود است، انجام می‌گردد.

۳. تحقق زبانی^{۶۷}

در این مرحله، که مرحله نهایی است، پردازش‌های شامل پردازش‌های نحوی^{۶۸} برای صیقل‌دادن جملات تولید شده و تصحیح نهایی آن‌ها، صورت می‌گیرد.

اولین روش‌های ارائه شده در این حوزه، عموماً محدود به کاربردهای خاص بودند. در عموم این روش‌ها، مراحل مورد نیاز برای اجرای فرایند تولید جمله به ترتیب زیر، اجرا می‌شدند.

۱. استخراج لغات متناظر با داده از طریق جدول نگاشت ثابت^{۶۹}

۲. استخراج عبارات مناسب زبانی برای تولید جمله

۳. اعمال قوانین دستور زبان برای ساخت جمله

^{۶۲}Knowledge Bases

^{۶۳}Logical Forms

^{۶۴}Framework

^{۶۵}Text Planning

^{۶۶}Sentence Planning

^{۶۷}Linguistic Realisation

^{۶۸}Syntactic

^{۶۹}Hardcoded Mapping Table

در [۳۰]، روشی برای تولید جملات زبان طبیعی، بیان کننده وضعیت حرکت قطارها در یک ترمینال مسافربری، ارائه داده است. در این پژوهش، اطلاعات مختلف موجود در پایگاه داده، به دو دسته پیام‌های «حرکت از مبدأ» و «رسیدن به مقصد» تقسیم می‌شوند و اطلاعات زمانی و شماره هر قطار در هر پیام، استخراج می‌شود. سپس با استفاده از یک جدول از پیش تعیین شده ثابت^{۷۰}، هر پیام به یک مجموعه از لغات، نگاشت می‌شود. در ادامه، عبارات زبانی مناسب جهت تولید جملات استخراج شده و با اعمال کلیشه‌های ثابت^{۷۱}، جملات نهایی تولید می‌شوند.

۴-۳ روش نزدیک‌ترین همسایه

یکی از پرکاربردترین روش‌ها در این زمینه، استفاده از روش نزدیک‌ترین همسایه است. در این روش، با استفاده از بردار ویژگی‌های به دست آمده از داده‌های مورد تفسیر، و استخراج بردار ویژگی‌های متناظر از جملات موجود در پایگاه داده، نزدیک‌ترین جمله به بردار ویژگی حاصل، به عنوان جمله توصیف‌کننده انتخاب شده و اعلام می‌شود. پژوهش‌های زیادی با استفاده از روش نزدیک‌ترین همسایه، پردازش‌های مختلفی بر روی داده‌های متنی انجام داده‌اند. به عنوان مثال در پژوهش [۳۱] با استفاده از این روش، مدلی جهت تشخیص صحت یا عدم صحت یک جمله به لحاظ دستور زبانی، ارائه شده است. در این پژوهش با استفاده از دو معیار فاصله اقلیدسی و معیار فاصله ویرایش^{۷۲}، ارائه شده است. معیار فاصله ویرایش در اصل، میزان هزینه حذف، درج یا تغییر کاراکترها را در یک دنباله کاراکتری برای رسیدن به یک جمله جدید محاسبه می‌کند. مطابق با گزارش این پژوهش، بالاترین دقیقت به دست آمده از این مدل برای تشخیص صحت یک جمله به لحاظ دستور زبانی، برابر با ۵۵٪ بوده است.

فعالیت‌های متعدد دیگری نیز با استفاده از این روش، سعی در شرح تصویر داشته‌اند. برای مثال در پژوهش [۳]، ایده اصلی در تولید شرح بر تصاویر، استفاده از نزدیک‌ترین جمله به تصویر است. روش‌های مختلف و متعدد محاسبه شباهت در این پژوهش مورد بحث و بررسی قرار گرفته‌اند که به اختصار به بیان آن‌ها خواهیم پرداخت. فرض می‌شود یک مجموعه S_{cond} شامل تمام جملات موجود در مجموعه‌داده که هر کدام شرحی بر یک تصویر هستند و یک مجموعه I_{cand} شامل تمام جملات موجود در مجموعه‌داده که هر کدام مربوط به یکی از جملات هستند، وجود دارند.

در این پژوهش، دو هدف به طور کلی مورد نظر قرار گرفته است (در هر دو تعریف، تابع (i, s, f) میزان شباهت جمله s و تصویر i را محاسبه می‌کند).

۱. پیدا کردن بهترین جمله توصیف‌کننده یک تصویر

در این مرحله، به دنبال یافتن s^* به گونه‌ای هستیم که مقدار (i, s^*, f) به ازای تصویر ورودی i بیشینه شود.

۲. پیدا کردن بهترین تصویر توصیف‌کننده یک جمله

در این مرحله به دنبال یافتن i^* به گونه‌ای هستیم که مقدار (s, i^*, f) به ازای جمله s بیشینه شود.

ایده اصلی در این پژوهش، این است که با ورود یک تصویر جدید i به سیستم، ابتدا نزدیک‌ترین تصویر موجود در مجموعه‌داده را نسبت به این تصویر پیدا کرده (i^{NN}) و سپس نزدیک‌ترین جمله به تصویر بازیابی شده (s^{NN}) را به

^{۷۰} Hardcoded Mapping Table

^{۷۱} Fixed Templates

^{۷۲} Edit Distance

عنوان جمله خروجی انتخاب می‌کنیم. برای پیدا کردن تصویر مرتبط با یک جمله نیز به همین منوال عمل می‌شود.

شکل ۱۸ نتایج اختصاص تصاویر و جملات را توسط این روش نمایش می‌دهد. در این شکل، نتایج بصری به طور کیفی و براساس میزان خطای جمله تولیدی، به چهار دسته تقسیم شده‌اند.

... describes the image without any errors (score = 4)	The selected caption describes the image with minor errors (score = 3)	... is somewhat related to the image (score = 2)	... is unrelated to the image (score = 1)
				
<i>A girl wearing a yellow shirt and sunglasses smiles.</i>		<i>A group of people walking a city street in warm weather.</i>	<i>A boy jumps into the blue pool water.</i>	<i>A dog in a grassy field, looking up.</i>
				
<i>A man riding a motor bike kicks up dirt.</i>	<i>Dogs pulling a sled in a sled race.</i>	<i>Two little girls practice martial arts.</i>	<i>A snowboarder in the air over a snowy mountain.</i>	<i>A boy in a blue life jacket jumps into the water.</i>
				
		<i>Basketball players in action.</i>		<i>A black dog with a purple collar running.</i>

شکل ۱۸: نتایج کیفی اختصاص جملات و تصاویر به یکدیگر با استفاده از روش نزدیکترین همسایه [۳]

به علاوه، جدول ۱ نتایج ارزیابی روش نزدیکترین همسایه را در اختصاص جملات و تصاویر به یکدیگر بر اساس سه معیار نمایش می‌دهد. معیار اول، میانگین امتیازی که افراد خبره به ۱۰۰۰ جملات تولید شده برای هر عکس داده‌اند را نمایش می‌دهد. این امتیازها اعداد صحیح بین ۱ تا ۴ را شامل می‌شوند و بیشینه ممکن برای این معیار، ۴ است. امتیاز بالاتر نشان‌دهنده مناسب‌تر بودن جملات تولید شده هستند. معیارهای ROUGE و BLUE معیارهایی هیتند که در پژوهش‌های ترجمه ماشین به عنوان محکی برای میزان خوب بودن ترجمه تولید شده، استفاده می‌شوند. مقادیر بالاتر در این معیارها، مناسب بودن عملکرد را نمایش می‌دهد.

جدول ۱: نتایج استفاده از روش نزدیکترین همسایه در اختصاص جملات و تصاویر [۳]

ROUGE	BLUE	امتیاز افراد خبره
۰.۱۱	۰.۳۵	۱.۵۷

به عنوان یکی دیگر از روش‌های مورد استفاده در تولید خودکار شرح بر تصاویر، می‌توان به پژوهش ارائه شده در [۴] اشاره کرد. در این پژوهش، تصاویر موجود در مجموعه‌داده، هر کدام با تعدادی عبارت زبانی که توسط کاربران انسانی نوشته شده‌اند، توصیف شده‌اند. در این پژوهش، هدف اصلی این است که با استخراج شبیه‌ترین عبارات موجود در مجموعه‌داده به تصویر ورودی و با کنارهم قرار دادن این عبارات و ساخت جمله با استفاده از چارچوب کاری مورد استفاده در پژوهش‌های تولید زبان طبیعی، جمله متناسب، تولید و نمایش داده شود. علاوه بر استفاده از روش نزدیکترین همسایه برای انتخاب بهترین عبارات زبانی توصیف‌کننده تصویر، می‌توان از

استفاده هوشمندانه از مرحله «برنامه‌ریزی محتوا»^{۷۳} در این پژوهش، به عنوان یکی از نقاط قوت آن، یاد کرد. این کار باعث می‌شود علاوه بر تولید جملات سازگار با یکدیگر، از تولید جملات تکراری در یک شرح بر یک تصویر، خودداری شود که از نقاط قوت این روش است. این مرحله با استفاده از روش‌های بهینه‌سازی انجام می‌شود. روش برنامه‌سازی خطی صحیح^{۷۴}، به عنوان چارچوب کاری در این مرحله مورد استفاده قرار گرفته است.

در این پژوهش، ۸۹ کلاس از اجسام و ۲۶ کلاس صحنه برای تشخیص محتوا انتخاب شده و تصاویر ورودی، با استفاده از آشکارکننده‌های فلزنسوالب، به این دسته‌ها اختصاص داده می‌شوند. این کار، تخمین مناسبی از محتوای جملات ارائه می‌دهد. همین‌طور با استفاده از روش برچسب‌گذاری نقش کلمات در جمله^{۷۵}، محتوای عبارات زبانی متناظر با جملات، به طریق مشابه، دسته‌بندی می‌شود.

چهار دسته از عبارات برای هر تصویر ورودی، به این روش، استخراج می‌شود.

۱. عبارات اسمی

جستجوی عبارات اسمی موجود در مجموعه‌داده با استفاده از ویژگی بافت و رنگ به عنوان معیارهای محاسبه شbahت.

مانند:

”The brown cow”

۲. عبارات فعلی

استفاده از معیارهای مشابه با عبارات اسمی در بین عبارات فعلی موجود در مجموعه‌داده

مانند:

”boy running”

۳. عبارات اضافی نواحی و اجسام

جستجوی عبارات با استفاده از معیارهای شباهت رنگ، بافت، هیستوگرام گرادیان^{۷۶} و همین‌طور با درنظر گرفتن ویژگی‌های هندسی اجسام

مانند:

”in the sky” , ”on the road”

۴. عبارات اضافی صحنه^{۷۸}

جستجوی عبارات با استفاده از نتیجه دسته‌بندی صحنه با معیار L^2

مانند:

^{۷۳}Content Planning

^{۷۴}Integer Linear Programming (ILP)

^{۷۵}Part of Speech Tagging (POS Tagging)

^{۷۶}Region/Stuff Prepositional Phrases

^{۷۷}Histogram of Gradient (HOG)

^{۷۸}Scene Prepositional Phrases

”at the market” , ”on hot summer day” , ”in Sweden”

هر جمله شامل یک عبارت اسمی، بیان‌کننده مفعول جمله و یک یا چند نمونه از عبارت دیگر بیان‌کننده مفهوم تصویر هستند. چهار نوع عملیات انتزاعی زیر برای ساخت جمله، در نظر گرفته شده است. هر یک از اعمال زیر با استفاده از روش برنامه‌سازی خطی صحیح، انجام می‌شوند.

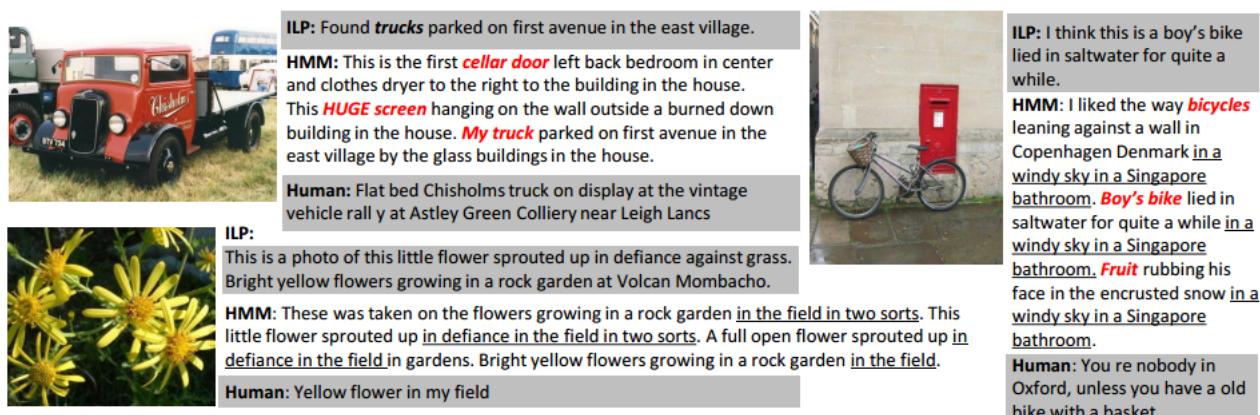
۱. انتخاب مجموعه مفعولی که نیاز به توصیف دارند (هر مفعول برای یک جمله)

۲. بازچینی و ترتیب‌دهی به مفعول‌ها

۳. انتخاب مجموعه عبارات مورد نیاز برای هر جمله

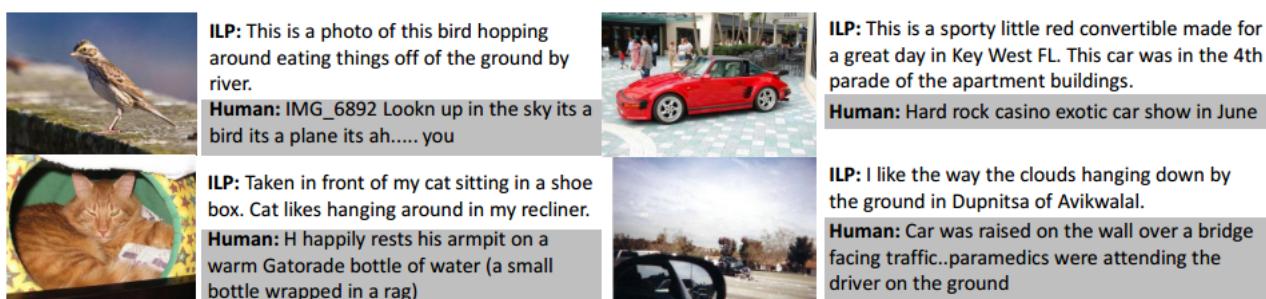
۴. بازچینی و ترتیب‌دهی عبارات در هر جمله

در شکل ۱۹، نتایج عملکرد الگوریتم را در مقایسه با جملات تولید شده توسط انسان، مشاهده می‌نمایید. مواردی که با عبارت ILP مشخص شده‌اند، خروجی‌های روش برنامه‌سازی خطی صحیح و مواردی که با عبارت HMM نمایش داده شده‌اند، جملات تولید شده توسط انسان را نمایش می‌دهند.



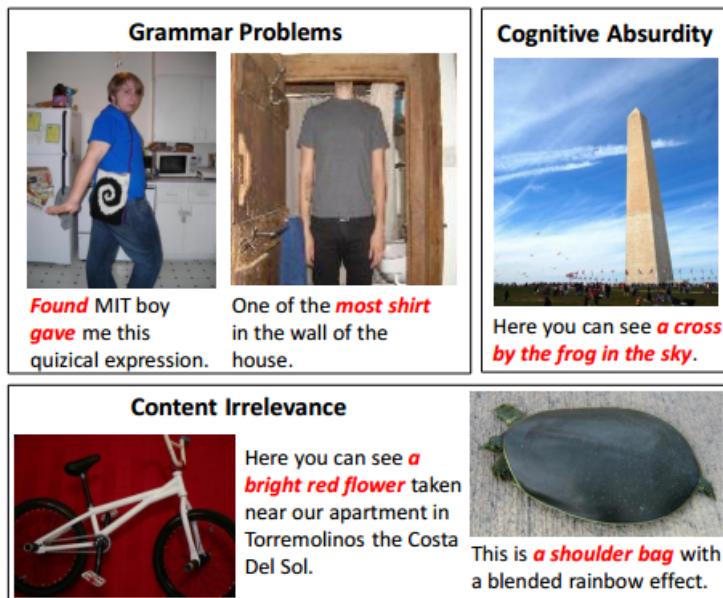
شکل ۱۹: نتایج برنامه‌سازی خطی صحیح در مقایسه با جملات تولید شده انسان [۴]

به علاوه، در شکل ۲۰، مواردی را مشاهده می‌نمایید که در آن‌ها، خروجی الگوریتم نسبت به جملات تولید شده توسط انسان، برتری دارد.



شکل ۲۰: مواردی از خروجی برنامه‌سازی خطی صحیح که نسبت به جملات انسان، برتری دارد [۴].

با وجود این که الگوریتم در برخی موارد کارایی بهتری از خود نشان داده است، جملاتی نیز وجود دارند که به لحاظ دستور زبان، عدم شناخت صحیح یا ناسازگاری محتوایی دچار مشکل شده‌اند. شکل ۲۱، نمونه‌هایی از این دست را نمایش می‌دهد.



شکل ۲۱: نتایج برنامه‌سازی خطی صحیح که به لحاظهای مختلف دچار مشکل شده‌اند [۴].

۵-۳ استفاده از قالب‌های آماده زبانی

یکی دیگر از روش‌هایی که برای تولید جمله متناظر با یک تصویر مورد استفاده قرار می‌گیرد، روش استفاده از قالب‌های آماده زبانی است. پژوهش‌هایی که از این روش برای تولید جملات زبانی استفاده کرده‌اند، غالباً پژوهش‌های وظیفه‌محور^{۷۹} هستند. حملات تولید شده در این پژوهش‌ها عموماً به طور هدفمند برای پاسخ دادن به موارد معینی تولید می‌شوند و قابلیت تعمیم‌پذیری کمتری نسبت به روش‌های دیگر دارند.

به عنوان مثال در پژوهش [۵]، یک روش مبتنی بر استفاده از قالب‌های آماده زبانی برای تولید خودکار شرح بر تصاویر ارائه شده است. جملاتی که در این پژوهش تولید می‌شوند، باید قادر به مشخص کردن اطلاعات زیر برای هر تصویر باشند:

۱. اجسامی که در تصویر مشاهده می‌شوند

۲. ویژگی‌های اجسام شامل رنگ، اندازه و موارد مشابه

۳. فاعل

۴. فعل

۵. حروف اضافه

^{۷۹}Task-Based

در این پژوهش، هدف اصلی این است که پس از بررسی تصویر با استفاده از روش‌های موجود در حاشیه‌نویسی تصویر^{۸۰}، که در آن‌ها هر تصویر با یک یا چند برچسب زبانی حاشیه نویسی می‌شود، بتوان جمله‌های بیان‌کننده موارد فوق را در تصویر با استفاده از برچسب‌های تولید شده، تولید نمود. در این پژوهش از مجموعه‌داده PASCAL^{۸۱} استفاده شده است که شامل ۱۰۰۰ تصویر و برای هر تصویر، ۵ جمله تولید شده توسط انسان است. ابتدا با استفاده از یک ابزار برچسب‌زنی نقش کلمات در جملات، تمام کلمات موجود در جملات مجموعه‌داده، برچسب زده می‌شوند. سپس برای هر تصویر، ۲ مفعول از بین ۵ جمله متناظر آن و برای هر مفعول، یک ویژگی استخراج می‌شود. به علاوه با استفاده از نرم‌افزار WordNet، مترادف‌های هر یک از کلمات استخراج شده، تا ۳ سطح، یافت می‌شوند.

در مرحله بعدی برای یافتن فاعل جمله، کافیست تعداد دفعاتی را که هر یک از ۲ مفعول استخراج شده، در بین ۵ جمله موجود، در نقش فاعل بوده‌اند شمرده و کلمه‌ای را که بیشترین تعداد تکرار به عنوان فاعل را داشته است، به عنوان فاعل جمله در نظر بگیریم. با مشخص شدن فاعل و مفعول جمله، کافیست فعل مورد نظر را با شمارش تعداد دفعات تکرار افعال مختلف در جملاتی که فاعل و مفعولشان برابر با مورد در حال بررسی است، فعل با بیشترین تکرار را انتخاب کنیم. اگر چنین فعلی یافت نشد، از فعل در قالب آماده استفاده نخواهد شد.

موارد مورد نیاز دیگر نیز به همین ترتیب استخراج می‌شوند. در انتهای، با توجه به پیدا شدن یا نشدن فعل و همین‌طور نوع فعل استخراج شده، از یکی از قالب‌های زیر برای ساخت جمله استفاده می‌شود.

۱. فعل اصلی استخراج شده است:

(معرف ۱ - ویژگی ۱ - فاعل) - فعل - حرف اضافه - (معرف ۲ - ویژگی ۲ - مفعول)

۲. فعل گذرا استخراج شده است:

(معرف ۱ - ویژگی ۱ - فاعل) - فعل - (معرف ۲ - ویژگی ۲ - مفعول)

۳. فعلی یافت نشده است:

(معرف ۱ - ویژگی ۱ - فاعل) - حرف اضافه - (معرف ۲ - ویژگی ۲ - مفعول)

شکل ۲۲ مواردی از جملات تولید شده برای تصاویر موجود در مجموعه‌داده را نمایش می‌دهد. نمونه‌هایی که در این شکل مشاهده می‌شود، نمونه‌هایی هستند که به لحاظ معنایی صحیح بوده و با تصویر مربوطه سازگاری دارند.

^{۸۰}Image Annotation

^{۸۱}<http://vision.cs.uiuc.edu/pascal-sentences/>

green, white, car, saloon	loaded, chair, shelf, plastic	train, green, yellow, building	boat, beach, water, person, sandy	white, airport, engine, building, aeroplane
A green car is parked in front of a white saloon.	A plastic chair is sitting next to a loaded shelf.	A yellow train is passing a green building.	A person is trailing the water. There is a boat in a sandy beach.	A white building is standing on an airport. There is the engine of an aeroplane.

شکل ۲۲: نمونه‌های صحیح از جملات تولید شده توسط قالب‌های آماده زبانی [۵]

به علاوه در شکل ۲۳، نمونه‌هایی از خروجی الگوریتم را در حالتی که جملات تولید شده به لحاظ شناخت صحیح فاعل، ویژگی، فعل، حروف اضافه و همین‌طور تکرار کلمات در جمله دچار مشکل شده‌اند، نمایش می‌دهد.

Incorrect subject	Incorrect attr-obj	Incorrect verb	Incorrect prep	Object Repetition
A white table is sitting on a buddha statue.	A brown person is milking a young cow.	A young person is cutting a cake.	A wooden table is sitting on a wooden chair.	There is a patio set and the wet patio in the picture.

شکل ۲۳: نمونه‌های اشتباه تولید شده توسط قالب‌های آماده زبانی [۵]

علاوه بر این، جدول ۲، نتایج معیار BLUE را در حالات مختلف (استفاده از ترکیبات چندتایی کلمات^{۸۲}) نمایش می‌دهد. همان‌طور که مشاهده می‌شود، مقادیر به دست آمده از این معیارها، قابل قبول بودن دقت جملات تولید شده توسط این روش را نمایش می‌دهند. در ستون‌هایی از جدول که از حرف s استفاده شده، تطابق بین کلمات هم‌معنی نیز در نظر گرفته شده است در بقیه ستون‌ها، کلمات دقیق با هم مقایسه شده‌اند. همان‌طور که مشخص است، استفاده از کلمات هم‌معنی نتایج بهتری را ارائه داده است.

جدول ۲: نتایج معیارهای BLUE و ROUGE در حالات مختلف [۵]

ROUGE-1-s	ROUGE-1	BLUE-3-s	BLUE-3	BLUE-2-s	BLUE-2	BLUE-1-s	BLUE-1
۰.۶۰	۰.۵۵	۰.۴۲	۰.۳۵	۰.۶۱	۰.۵۵	۰.۷۹	۰.۷۴

۶-۳ روش‌های مبتنی بر شبکه‌های عصبی بازگشتی

اخیراً، استفاده از شبکه‌های عصبی بازگشتی برای تولید جمله، توجه تعداد زیادی از پژوهش‌گران را به خود جلب کرده است. شبکه‌های عصبی بازگشتی، ضمن اثبات قدرت خود در پیش‌بینی سری‌های زمانی، در کاربردهای متعدد و متنوعی مورد استفاده قرار می‌گیرند. از جمله این کاربردها می‌توان به تولید تصاویر، تولید متن، تولید برنامه، تولید موسیقی و مواردی از این دست اشاره نمود. با توجه به ظرفیت بالا و توان یادگیری بالای این مدل‌ها،

^{۸۲}n-grams

استفاده از آن‌ها در تولید جملات زبان طبیعی مرتبط با مفهوم یا مفاهیمی خاص، نظر بسیاری را به خود جلب کرده است.

شبکه‌های عصبی بازگشتی در اوخر دهه ۱۹۹۰، ارائه شدند و فعالیت‌های محدودی بر روی مدل‌های کوچکی از آن‌ها مانند مدل Elman انجام شد. به دلیل زمان بالای مورد نیاز برای آموزش این نوع از شبکه‌های عصبی، تا سال ۲۰۱۱، عموم فعالیت‌ها در این زمینه، محدود به استفاده از مدل‌های کوچک از این شبکه‌ها بودند. در سال ۲۰۱۱، آقای هینتون^{۸۳} و همکارانش در پژوهش [۶]، با بهره‌گیری از پیشرفت‌های جدید در بهینه‌سازی روش بدون هسین^{۸۴} و اعمال این روش‌ها به فرایند تولید جمله در سطح حروف^{۸۵} قادر به آموزش یک شبکه عصبی بازگشتی در ۵ روز شدند و نتایج بهتری نسبت به مدل‌های ارائه شده تا آن زمان، حاصل کردند.

در پژوهش [۶]، علاوه بر ارائه یک روش برای آموزش شبکه‌های عصبی بازگشتی عمیق بر مبنای روش بهینه‌سازی بدون هسین، نشان داده شده است که شبکه‌های عصبی بازگشتی استاندارد، عملکرد خوبی در تولید جملات در سطح حروف از خود نشان نمی‌دهند. برای حل این مساله، نوع خاصی از این شبکه‌ها موسوم به شبکه‌های عصبی بازگشتی عمیق ضربی^{۸۶} ارائه شده‌اند.

نکات جالبی که در نتایج شبکه عصبی بازگشتی ارائه شده در این پژوهش به چشم می‌خورد، عبارتند از:

۱. تولید ساختارهای زبانی سطح بالا^{۸۷}

۲. پشتیبانی از دایره لغات بسیار وسیع

۳. یادگیری تعداد قابل توجهی از دستورات زبانی

۴. تولید تعداد زیادی اسم محتمل که در بین لغات مجموعه‌داده وجود ندارند

۵. توانایی باز و بسته کردن صحیح پرانتزها و نقل قول‌ها در فواصل طولانی بیش از ۳۰ حرف

یک شبکه عصبی بازگشتی را که دنباله زمانی $(x_1 \dots x_T)$ را به عنوان ورودی گرفته و دنباله $(h_1 \dots h_T)$ و دنباله $(o_1 \dots o_T)$ را به ترتیب به عنوان حالات مخفی و خروجی، تولید می‌کند می‌توان مطابق با رابطه ۲۰ بیان کرد. در این رابطه، W_{hx} ماتریس وزن‌های لایه ورودی به لایه مخفی، W_{hh} ماتریس وزن‌های لایه مخفی به لایه مخفی (وزن‌های بازگشتی) و W_{oh} ماتریس وزن‌های لایه مخفی به لایه خروجی است. بردارهای b ، بردارهای بایاس هستند و مقدار $W_{hh}h_{t-1}$ در نقطه $t = 1$ با یک مقدار اولیه جایگزین می‌شود.

$$h_t = \tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \\ o_t = W_{oh}h_t + b_o \quad (۲۰)$$

^{۸۳}Hinton

^{۸۴}Hessian Free (HF)

^{۸۵}Character Level Sentence Generation

^{۸۶}Multiplicative Recurrent Neural Networks (MRNN)

^{۸۷}High level linguistic structures

از آنجا که مشتقات رابطه ۲۰ قابل محاسبه هستند، به راحتی می‌توان رابطه به روزرسانی وزن‌ها را بر اساس روش نزول در امتداد گرادیان، برای شبکه عصبی بازگشتی، محاسبه کرد. با این وجود، به دلیل رابطه بین پارامترها و دینامیک شبکه، روش نزول در امتداد گرادیان کارایی خوبی در آموزش شبکه برای پیش‌بینی دنباله‌های زمانی در بازه‌های زمانی بزرگ، ارائه نمی‌دهد. به همین دلیل، تا سال ۲۰۱۱ و ارائه روش بهینه‌سازی بدون هسین در [۶] برای آموزش شبکه عصبی بازگشتی، پژوهش‌های زیادی در این زمینه انجام نمی‌شد.

برای رفع مشکل روش پسانشان خطای^{۸۸} در آموزش شبکه‌های عصبی بازگشتی، سه روش زیر پیشنهاد شده است:

۱. استفاده از شبکه‌های عصبی حالت پژواک^{۸۹}

در این شبکه‌ها، فقط وزن‌های پیش‌رو^{۹۰} به روزرسانی می‌شوند. مقادیر اولیه برای وزن‌های بازگشتی باید به طور مناسب انتخاب شوند.

۲. شبکه‌های حافظه کوتاه‌مدت بلند^{۹۱}

در این شبکه‌ها، گره‌هایی برای به خاطر سپاری ورودی‌های قدیمی تعبیه می‌شود که باعث افزایش قدرت این شبکه‌ها در پیش‌بینی بلندمدت دنباله‌های زمانی می‌شود.

۳. استفاده از بهینه‌سازی بدون هسین در آموزش شبکه‌های عصبی بازگشتی ضربی^{۹۲} این مدل، تاکنون توانسته کارایی بهتری از هر دو مدل قبلی از خود نشان دهد [۶].

در ادامه به بررسی شبکه عصبی ارائه شده در پژوهش [۶] می‌پردازیم.

۱-۶-۳ شبکه عصبی بازگشتی ضربی [۶]

ایده اصلی در طراحی این شبکه این است که به جای استفاده و آموزش یک ماتریس یکسان از وزن‌های لایه مخفی به ازای تمام ورودی‌های ممکن، به ازای هر ورودی ممکن، یک ماتریس وزن داشته باشیم. به عبارت دیگر، هر حرکتی که به عنوان ورودی به شبکه وارد شد، مشخص کننده ماتریسی باشد که به عنوان وزن‌های لایه مخفی در آموزش شبکه شرکت می‌کند. با این ایده، رابطه (۲۰) به شکل رابطه (۲۱) تبدیل می‌شود.

$$\begin{aligned} h_t &= \tanh(W_{hx}x_t + W_{hh}^{(x_t)}h_{t-1} + b_h) \\ o_t &= W_{oh}h_t + b_o \end{aligned} \quad (21)$$

همان‌طور که ملاحظه می‌شود، تنها تفاوت رابطه (۲۰) و رابطه (۲۱) در بالا نویس پارامتر W_{hh} است. پارامتر جدید را می‌توان به شکل رابطه (۲۲) تعریف کرد که در آن، $x_t^{(m)}$ مولفه m از ورودی و $W_{hh}^{(m)}$ است. لازم به

^{۸۸}Backpropagation

^{۸۹}Echo State Network

^{۹۰}Feed forward

^{۹۱}Long Short Term Memory (LSTM)

ذکر است در رابطه (۲۲) برای تعریف $W_{hh}^{(x_t)}$ از تنسور^{۹۲} استفاده شده است. در این میان، M تعداد ابعاد ورودی و ماتریس‌های $W_{hh}^{(1)}$ تا $W_{hh}^{(M)}$ ماتریس‌های مربوط به هریک از ابعاد ورودی هستند.

$$W_{hh}^{(x_t)} = \sum_{m=1}^M x_t^{(m)} W_{hh}^{(m)} \quad (22)$$

مدل ارائه شده از آنجا که به طور کامل عمومی است، در حالاتی که ابعاد ورودی و تعداد گره‌های مخفی زیاد باشد، دچار مشکل می‌شود زیرا حجم زیادی حافظه نیاز خواهد داشت. برای حل این مشکل، سعی در فاکتورگیری از پارامتر $W_{hh}^{(x_t)}$ خواهیم داشت.

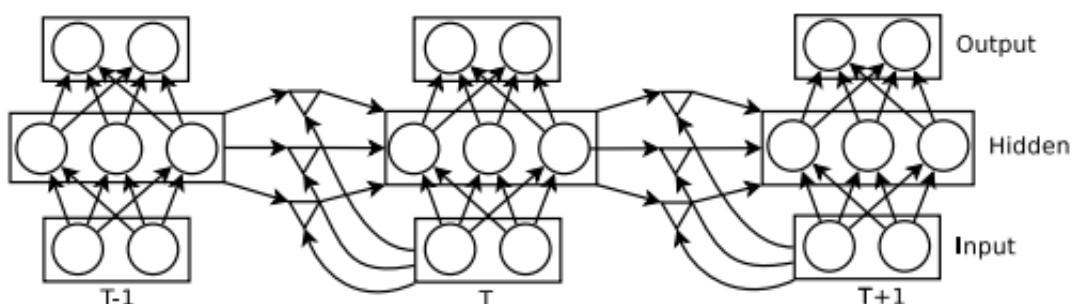
با تعریف سه ماتریس W_{hf} ، W_{fx} و W_{fh} تغییر رابطه (۲۲) به شکل رابطه (۲۳)، می‌توان این فاکتورگیری را تعریف کرد.

$$W_{hh}^{(x_t)} = W_{hf} \cdot \text{diag}(W_{fx} x_t) \cdot W_{fh} \quad (23)$$

در صورتی که ابعاد ماتریس $W_{fx} x_t$ که آن را F خواهیم نامید، بزرگ باشد، رابطه (۲۳) به همان اندازه رابطه (۲۲) پیچیده خواهد بود اما در صورتی که این ابعاد کوچک باشد، رابطه (۲۳) نسبت به رابطه (۲۲) به مراتب ساده‌تر خواهد بود. با جای‌گذاری رابطه (۲۳) در رابطه (۲۱)، شبکه عصبی بازگشتی ضربی به دست خواهد آمد. رابطه نهایی را می‌توان به شکل رابطه (۲۴) نمایش داد.

$$\begin{aligned} f_t &= \text{diag}(W_{fx} x_t) \cdot W_{fh} h_{t-1} \\ h_t &= \tanh(W_{hf} f_t + W_{hx} x_t) \\ o_t &= W_{oh} h_t + b_o \end{aligned} \quad (24)$$

شکل ۲۴ طرحواره‌ای از مدل کلی شبکه عصبی بازگشتی ضربی را که توسط رابطه (۲۴) مدل شده است، نمایش می‌دهد. علامت مثلث که در شکل نشان داده شده است، نمایش‌دهنده یک گیت^{۹۳} را نمایش می‌دهد که در آن، وزن لایه مخفی با استفاده از ورودی، انتخاب می‌شود.



شکل ۲۴: طرحواره شبکه عصبی بازگشتی ضربی [۶]

^{۹۲}Tensor

^{۹۳}Gate

نکته‌ای که هم‌چنان مبهم باقی می‌ماند، نحوه محاسبه وزن‌های مخفی است. رابطه (۲۵) نحوه محاسبه $W_{hh}^{(x_t)ij}$ را نمایش می‌دهد. در این حاصل ضرب، اگر به عنوان مثال پارامتر W_{hhif} خیلی کوچک باشد و W_{hhfj} خیلی بزرگ باشد یا بر عکس، حاصل مشتق، برای وزن‌های بسیار کوچک، بسیار بزرگ می‌شود و برای وزن‌های بسیار بزرگ، بسیار کوچک. این باعث می‌شود روش نزول در امتداد گرادیان، به حالت پایداری نرسد.

$$W_{hh}^{(x_t)ij} = \sum_f W_{hhif} W_{hhfx^t} W_{hhfj} \quad (25)$$

به دلیل همین عدم پایداری ایجاد شده، حاصل از رابطه (۲۵)، روش نزول در امتداد گرادیان، گزینه مناسبی برای آموزش این شبکه نخواهد بود. لذا لزوم استفاده از روش‌های مرتبه دوم مانند روش بدون هسین، مشهود می‌شود.

۷-۳ جمع‌بندی

چالش تولید جمله یکی از قدیمی‌ترین و پویاترین حوزه‌های فعالیتی و پژوهشی در هوش مصنوعی است که از اواسط قدن بیستم، توجه پژوهش‌گران بسیاری را به خود جلب کرده است. روش‌های مختلفی برای حل این مساله ارائه شده‌اند. از جمله این روش‌ها می‌توان به موارد زیر اشاره کرد:

۱. تولید زبان طبیعی

در این دسته از روش‌ها که از اواخر دهه بیستم تا کنون مورد استفاده قرار می‌گیرند، با طی فرایند در یک چارچوب کلی، سعی در تولید جملات مناسب دارند. این دسته از روش‌ها عموماً برای تفسیر خودکار داده‌هایی که برای کاربران انسانی غیر قابل تفسیر هستند یا تفسیر دشواری دارند، به کار می‌روند. در این روش‌ها ابتدا با استفاده از ویژگی‌های مختلفی که در داده‌های ماشینی (داده‌های قابل تفسیر برای ماشین) کلمات مناسب انتخاب شده و سپس با استفاده از کلمات منتخب، عبارات زبانی (با جایگشت دادن کلمات و حذف عبارات غیر محتمل) تولید می‌شوند. سپس با اعمال قواعد دستور زبان و چینش عبارات زبانی در کنارهم، جملات نهایی تولید می‌شوند.

۲. نزدیک‌ترین همسایه

در این دسته از روش‌ها سعی می‌شود با ورود یک تصویر و نگاشت آن به فضای ویژگی‌ها، جمله‌ای از میان تمام جملات موجود در مجموعه داده انتخاب شود که بیشترین مشابهت با بردار ویژگی تصویر را دارد. بزرگ‌ترین مشکل در این روش‌ها انتخاب معیار مناسب برای محاسبه فاصله بین یک جمله و بردار ویژگی حاصل از تصویر است. در این روش، علاوه بر این که نیاز به وجود مجموعه داده وسیع و پوشش وجود دارد، ممکن است جمله نهایی، در انتهای گویا و بیان‌کننده تمام جوانب تصویر نباشد و یا حتی با تصویر ورودی سازگاری نداشته باشد.

برای حل این مشکل، سعی شد به جای استخراج نزدیک‌ترین جمله به تصویر موجود، مشابه‌ترین عبارات زبانی را با شکستن جملات موجود به عبارات سازنده، انتخاب کرده و با بهره‌گیری از روش تولید زبان طبیعی

و یا روش‌های دیگر، چینش مناسبی از این عبارات را که در قالب یک یا چند جمله بیان شوند، تولید و به عنوان شرح بر تصویر، نمایش داد.

۳. استفاده از قالب‌های زبانی آماده

با وجود فعالیت‌های گوناگون در این زمینه و استفاده از روش‌های مختلف، همچنان تضمین صحت جمله خروجی، کار دشواری است. به همین دلیل، سعی شد با ارائه یک یا چند قالب زبانی آماده و از پیش تعیین شده برای جملات، مانند قالب‌های جملات خبری، صحت جملات نهایی را تضمین کرد. در این دسته از روش‌ها، ویژگی‌های مختلفی از تصویر استخراج می‌شود که هریک از این ویژگی‌ها یا همه آن‌ها در کنار هم قادر هستند نقش‌هایی مانند « فعل »، « فاعل »، « مفعول » و موارد مشابه را در جمله متناظر با تصویر مشخص کنند. با استخراج کلمات مناسب و شناخت نقش آن‌ها در جمله و جای‌گذاری هر یک از این کلمات در مکان مناسب نقشی خود در قالب از پیش تعیین شده، جمله متناظر با هر تصویر استخراج می‌شود.

۴. استفاده از شبکه‌های عصبی بازگشتی

اگر چه استفاده از قالب‌های آماده و از پیش تعیین شده، تا حدی مشکلات موجود را حل می‌کند اما همچنان چالش بزرگ‌تری حل نشده باقی مانده است. تولید جملات جدید، استفاده از کلمات و عبارات جدید و ابتکاری به طوری که علاوه بر تضمین رعایت دستور زبان، بتوان معنای جمله را نیز متضمن شد، چالش بزرگی است که در این مسیر کماکان وجود دارد.

استفاده از شبکه‌های عصبی بازگشتی یکی از بهترین راه‌کارهای موجود برای حل این مشکل و رویارویی با این چالش هستند. استفاده از این شبکه‌ها در اواخر قرن بیستم در بین پژوهش‌گران رواج پیدا کرد تا جایی که ناپایداری الگوریتم پساننتشار خطا در آموزش این شبکه، راه را برای پژوهش‌های بعدی بست. پس از ارائه یک روش مناسب برای بهینه‌سازی بدون هسین در سال ۲۰۱۰، روشی برای آموزش یک شبکه عصبی بازگشتی موسوم به شبکه عصبی بازگشتی ضربی بر مبنای بهینه‌سازهای بدون هسین ارائه شد و نتایج آن به طور چشم‌گیری از روش‌های موجود بیشتر بود.

ارائه شبکه عصبی بازگشتی ضربی، نقطه عطفی در مسیر علم در راستای حل چالش تولید جمله به حساب می‌آید. از حدود سال ۲۰۱۱ به بعد، استفاده از شبکه‌های عصبی بازگشتی برای تولید جمله به پویاترین و پرفعالیت‌ترین حوزه در مسائل مربوط به تولید جمله، به حساب می‌آید.

۴ فصل چهارم

تولید شرح متناظر صحنه با استفاده از
یادگیری عمیق

به طور کلی می‌توان جایگاه روش‌های مبتنی بر یادگیری عمیق را در حوزه تولید شرح متناظر تصویر، از سال ۲۰۱۴ به بعد به روشنی در میان پژوهش‌های انجام‌شده مرتبط با این موضوع دریافت. از حدود سال ۲۰۱۳ و ۲۰۱۴، روش‌های مبتنی بر یادگیری عمیق، عمل کرد بسیار مناسب‌تری نسبت به روش‌های مبتنی بر مدل‌های گرافی احتمالاتی در این زمینه از خود نشان داده‌اند که این امر موجب استقبال چشم‌گیر پژوهش‌گران از این ایده شد.

استفاده از شبکه‌های عصبی و یادگیری عمیق به طور عمده در هر دو مرحله درک صحنه و تولید جمله در بین پژوهش‌های زیادی به چشم می‌خورد. با این حال، در مرحله درک صحنه تقریباً تمام پژوهش‌های انجام شده با استفاده از یک شبکه عصبی کانولوشنی عمیق، اقدام به استخراج بردار ویژگی تصویر می‌نمایند و این بردار ویژگی را به مرحله تولید جمله ارسال می‌کنند. بر خلاف مرحله درک صحنه که ایده‌های مطرح شده در آن از تنوع کم‌تری برخوردار است، مرحله تولید جمله چالشی‌ترین بخش فرایند به‌شمار می‌رود.

در بخش‌های قبلی به طور جداگانه به بررسی روش‌های مبتنی بر یادگیری عمیق ارائه شده برای تک‌تک مراحل پرداختیم. با این حال، روش‌های ارائه شده در این بخش‌ها عموماً روش‌های پایه‌ای هستند که بهبودهای زیادی روی هر یک از آن‌ها انجام شده است. در این بخش به بیان روش‌های جدیدتر و پیچیده‌تر در این حوزه خواهیم پرداخت.

از جمله پژوهش‌هایی که با تکیه بر شبکه‌های عصبی و یادگیری عمیق اقدام به تولید شرح متناظر تصویر نموده است، می‌توان به پژوهش [۷] اشاره کرد که توسط خانم لی و همکارانش در سال ۲۰۱۵، ارائه شده و با استفاده از شبکه‌های عصبی کانولوشنی عمیق و دو نوع از شبکه‌های عصبی بازگشتی موسوم به شبکه‌های عصبی بازگشتی مالتی‌مدال و شبکه‌های عصبی بازگشتی دوطرفه، روش مناسبی برای تولید خودکار شرح بر تصاویر ارائه داده است. در این پژوهش، ابتدا با بهره‌گیری از روش شبکه عصبی کانولوشنی ناحیه‌ای، نواحی از تصویر که شامل تصویر اجسام است، استخراج شده و با استفاده از یک شبکه عصبی کریفسکی، بردار ویژگی برای هر ناحیه محاسبه می‌شود. سپس با بهره‌گیری از یک شبکه عصبی بازگشتی دوطرفه، عبارات مختلف از جمله استخراج و بردارهای ویژگی برای هر عبارت محاسبه می‌شود. سپس با استفاده از یک تابع هدف و مدل میدان تصادفی مارکف، همترازسازی بین نواحی و عبارات زبانی صورت گرفته و مدل آموزش داده می‌شود.

در ادامه با تخمین بهینه پارامترهای موجود و با استفاده از شبکه عصبی بازگشتی مالتی‌مدال، توزیع احتمال بهترین کلمه بعدی در یک جمله با داشتن کلمات قبلی و محتوای حاصل از بردار ویژگی محاسبه شده روی نواحی تصویر، محاسبه شده و بهترین کلمه بعدی تولید می‌شود. این کار تا جایی ادامه می‌یابد که شبکه، نشانه مخصوص

پایان جمله را تولید کند.

۲-۴ تولید جمله با مفهوم مشخص

استفاده از شبکه عصبی بازگشتی ضربی منجر به فراهم‌سازی بستری مناسب جهت تولید جمله در سطح حروف می‌شود. با این وجود برای تولید خودکار شرح بر تصاویر، نیازمند آن هستیم که محتوای جملات را به طور مشخص و از پیش تعیین شده داشته باشیم. به همین دلیل نیاز به ارائه روش که طی آن بتوانیم معنا و محتوای جملات تولید شده توسط شبکه عصبی بازگشتی را کنترل کنیم، مشهود می‌شود.

در پژوهش [۷] روش جدیدی برای تولید شرح خودکار بر تصاویر ارائه شده است که در مرحله تولید جمله، از نوع خاصی از شبکه‌های عصبی بازگشتی موسوم به شبکه عصبی بازگشتی دوطرفه^{۹۴}، استفاده شده است. در این روش، ابتدا از یک شبکه عصبی کانولوشنی عمیق بر روی نواحی استخراج شده از تصاویر برای استخراج ویژگی و درک صحنه استفاده شده است. از سوی دیگر، با اعمال یک شبکه عصبی بازگشتی دوطرفه بر جملات و ارائه یکتابع هدف ساختارمند، روشی برای همترازسازی جمله و اطلاعات بصری نهفته در تصویر ارائه شده است.

شکل ۲۵، نمونه‌ای از همترازسازی ارائه شده در این پژوهش برای یک تصویر را نشان می‌دهد.



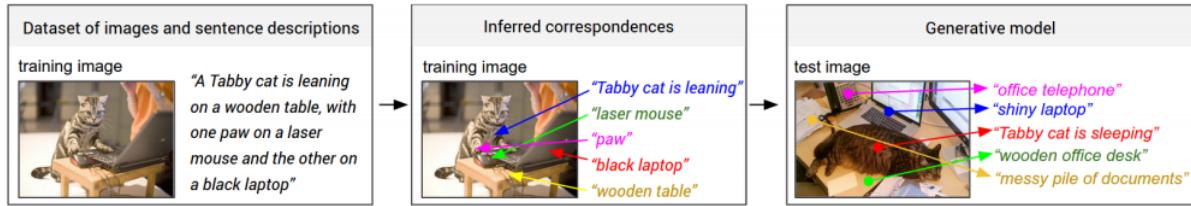
شکل ۲۵: همترازسازی تصویر و جمله [۷]

در مدل همترازسازی ارائه شده در این پژوهش، فرض بر این است که یک مجموعه‌داده شامل تعداد زیادی تصویر و جملات متناظر با هر تصویر وجود دارد. همین‌طور فرض دیگری وجود دارد مبنی بر این‌که بخش‌های مختلف هر جمله، به نواحی خاصی از تصویر اشاره می‌کنند که موقعیت این نواحی مجهول است. از طرف دیگر اشاره این بخش‌ها به نواحی مرتبط خود در تصاویر، در بین تمام مجموعه‌داده، تکرار می‌شود. به عنوان مثال، عبارات زبانی شامل کلمه «توب» در تمام تصاویر موجود در مجموعه‌داده، به نواحی از تصویر اشاره می‌کنند که دارای ویژگی‌های «توب» هستند.

در شکل ۲۶ ارتباط بین بخش‌های مختلف یک جمله و نواحی متفاوت از تصویر را مشاهده می‌نمایید. همان‌طور که در شکل مشاهده می‌شود، ابتدا برای تصاویر موجود در مجموعه‌داده و شرح متناظر با هر یک از این تصاویر،

^{۹۴}Bidirectional RNN

ارتباطات بین عبارات مختلف از جملات و نواحی تصاویر استخراج و یادگرفته می‌شود. در ادامه، با ورود یک تصویر جدید و براساس ارتباطات یادگرفته شده، شرح جدید برای تصویر تولید می‌شود.



شکل ۲۶: ارتباط بین نواحی مختلف یک تصویر و عبارات جمله^[۷]

برای تبدیل تصویر به فضای ویژگی، مطابق با آنچه در فصل درک صحنه ذکر شد، ابتدا با استفاده از روش شبکه‌های عصبی کانولوشنی ناحیه‌ای، ۱۹ ناحیه از تصویر استخراج شده و برای ۲۰ تصویر موجود، با اعمال یک بهینه‌سازی و تخمین پارامتر و اعمال یک شبکه عصبی، بردار ویژگی استخراج می‌شود. پس از استخراج بردار ویژگی از نواحی تصویر، نیازمند آن هستیم که بتوانیم از عبارات مختلف جمله، بردار ویژگی هماندازه با بردار ویژگی حاصل از تصویر، استخراج کنیم. برای این کار، در این پژوهش از شبکه عصبی بازگشتی دوطرفه استفاده شده است.

این شبکه عصبی، یک دنباله از N کلمه را به عنوان ورودی دریافت کرده و هر یک را به یک بردار در فضای h بعدی، که h اندازه بردار ویژگی حاصل از نواحی تصویر است، نگاشت می‌کند. رابطه^(۲۶)، رابطه مربوط به پارامترهای این شبکه عصبی را نمایش می‌دهد. در این رابطه، I یک بردار ستونی اندیکاتور^{۹۵} است که در اندیس کلمه t ام خود یک و در بقیه اندیس‌ها صفر دارد. W_w یک ماتریس وزن ثابت برای هر کلمه w است که برای جلوگیری از بیش‌برازش بر داده‌ها، مورد استفاده قرار می‌گیرد.

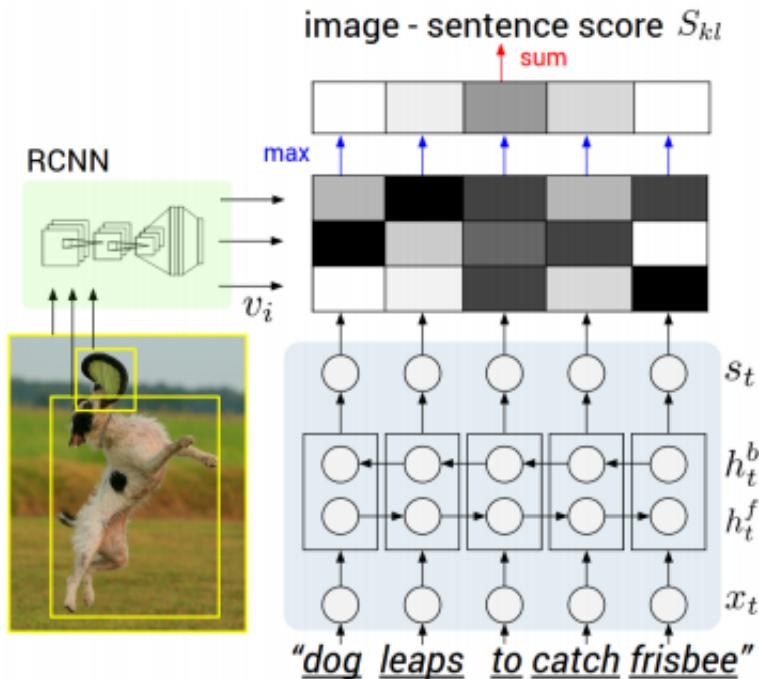
$$\begin{aligned}
 x_t &= W_w I_t \\
 e_t &= f(W_e x_t + b_e) \\
 h_t^f &= f(e_t + W_f h_{t-1}^f + b_f) \\
 h_t^b &= f(e_t + W_b h_{t-1}^b + b_b) \\
 s_t &= f(W_d(h_t^f + h_t^b) + b_d)
 \end{aligned} \tag{۲۶}$$

در این شبکه، دو جریان داده وجود دارد. جریان اول، جریان داده بین گره‌های مخفی شبکه از چپ به راست و دیگری جریان داده بین نودهای مخفی شبکه از راست به چپ است که به ترتیب با h_t^f و h_t^b نمایش داده می‌شوند. بردار نهایی s_t ، بردار حاصل از نگاشت کلمات به فضای ویژگی‌ها است که با استفاده از خود کلمه و محتوای مورد استفاده در اطراف کلمه در جمله، تولید می‌شود.

شکل ۲۷ طرح‌واره‌ای از معماری این شبکه را نمایش می‌دهد. همان‌طور که در این شکل مشخص است، در لایه

^{۹۵}Indicator

مخفی این شبکه، دو جریان داده، یکی از راست به چپ و دیگری از چپ به راست برای محاسبه تاثیر کلمات اطراف کلمه جاری بر نگاشت کلمه به فضای ویژگی‌ها، وجود دارد.



شکل ۲۷: طرح‌واره شبکه عصبی بازگشتی دوطرفه [۷]

در ادامه با بهره‌گیری از روش نگاشت دوطرفه تصاویر و جملات که در بخش درک صحنه ارائه شد، توابع هم‌ترازسازی و تابع هدف ارائه شده را مورد استفاده قرار داده و با استفاده از روش یادگیری چند نمونه‌ای، اقدام به یادگیری انتساب‌های بین نواحی مختلف تصاویر و عبارات مختلف زبانی می‌شود.

با استفاده از این روش، می‌توان برای هر ناحیه از تصویر، کلمات مناسب را تعیین کرد. اما برای تولید خودکار شرح بر تصاویر، نیاز به تولید عبارات زبانی وجود دارد. برای حل این مشکل، با در نظر گرفتن رابطه ضرب داخلی بین بردارهای ویژگی حاصل از نواحی تصویر و عبارات زبانی یک جمله به عنوان معیار شیاهت، روشی ارائه شده است که بتوان برای هر ناحیه از تصویر، عبارت زبانی مناسبی تولید کرد.

در این روش، با تعریف یک تابع انرژی و استفاده از مدل میدان تصادفی مارکف، با بهینه‌سازی تابع انرژی ارائه شده، بهترین هم‌ترازسازی برای هر یک از عبارات موجود محاسبه شده و عبارت با بهترین مقدار، انتخاب می‌شود. رابطه (۲۷)، این تابع انرژی را محاسبه می‌کند.

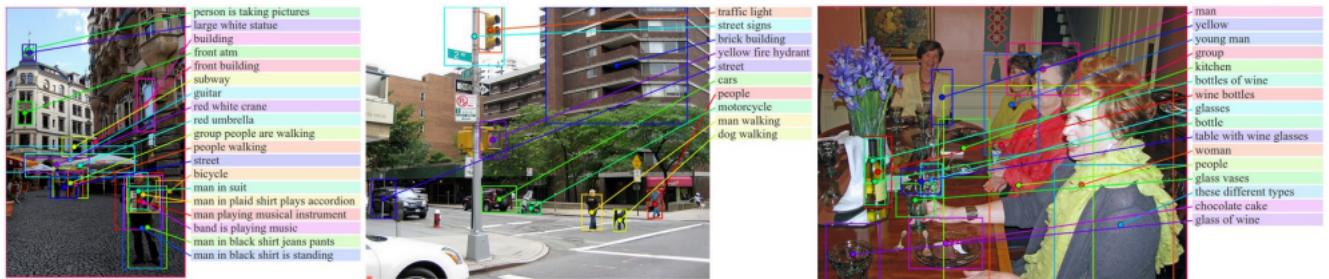
$$E(a) = \sum_{j=1 \dots N} \Psi_j^U(a_j) + \sum_{j=1 \dots M} \Psi_j^B(a_j, a_{j+1})$$

$$\Psi_j^U(a_j) = \nu_i^T \cdot s^t$$

$$\Psi_j^B(a_j, a_{j+1}) = \beta I(a_j = a_{j+1}) \quad (27)$$

می‌توان در شکل ۲۸ نتایج استفاده از این شبکه و محاسبه میزان شباهت نواحی مختلف تصویر و عبارات

مختلف از جملات را مشاهده نمود. این شکل، نتیجه آموزش اختصاص نواحی مختلف تصویر به عبارات زبانی را نمایش می‌دهد.



شکل ۲۸: انتساب نواحی مختلف تصویر به عبارات زبانی [۷]

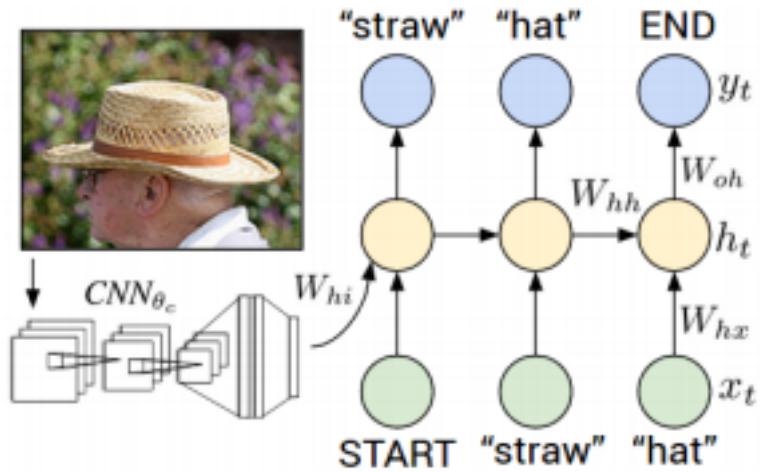
تا این مرحله، با ورود یک تصویر، عبارات زبانی متناظر، استخراج شده‌اند. هدف اصلی در این پژوهش تولید جمله برای هر تصویر است. بنابراین نیاز داریم تا با استفاده از مدل‌های ارائه شده برای تولید جمله، این کار را انجام دهیم. در پژوهش‌های زیادی، استفاده از شبکه‌های عصبی بازگشتی برای پیش‌بینی و محاسبه توزیع احتمال کلمه بعدی در یک جمله با در نظر داشتن کلمات قبلی و محتوای جمله، ارائه شده است. در این پژوهش با اعمال تغییرات کوچکی، از همین روش‌ها استفاده می‌شود.

رابطه ارائه شده برای شبکه عصبی بازگشتی که این کار را انجام می‌دهد، مطابق با رابطه (۲۸) است. در این رابطه، بردار حاصل از اعمال آخرین لایه یک شبکه عصبی کانولوشنی بر تصویر را نشان می‌دهد و بقیه پارامترها، همگی قابل آموزش هستند. بردار y_t بردار نماینده توزیع احتمالاتی تمام کلمات با در نظر گرفتن کلمات قبلی و محتوای هر ناحیه است که اندازه آن برابر است با تعداد تمام کلمات موجود در لغتنامه به علاوه یک نشانه خاص به عنوان «اتمام جمله».

$$\begin{aligned}
 b_\nu &= W_{hi}[CNN_{\theta c}(IMAGE)] \\
 h_t &= f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + I(t = 1) \cdot b_\nu) \\
 y_t &= softmax(W_{oh}h_t + b_o)
 \end{aligned} \tag{28}$$

شکل ۲۹، طرح‌واره‌ای از شبکه عصبی بازگشتی مالتی‌مدال^{۹۶} ارائه شده در این پژوهش برای تولید جمله را نمایش می‌دهد.

^{۹۶}Multimodal



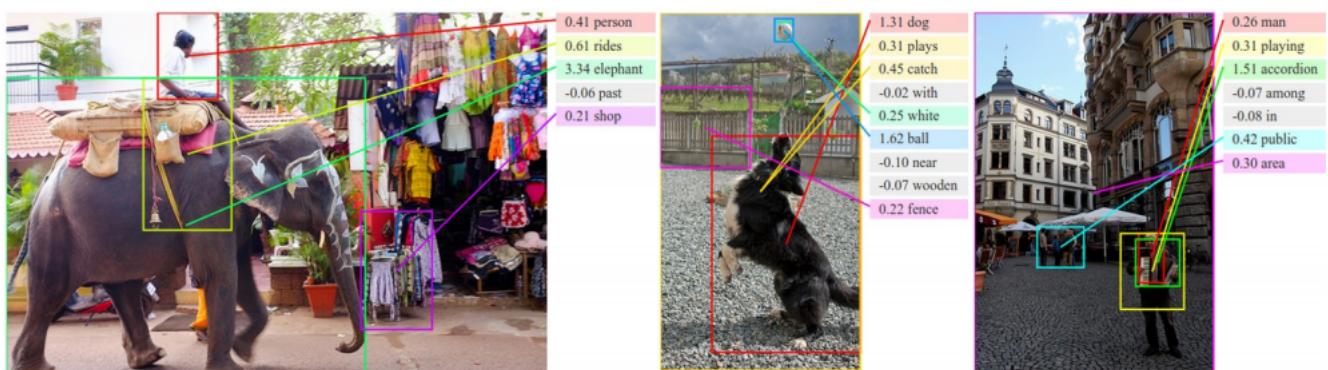
شکل ۲۹: طرح واره شبکه عصبی بازگشتی ارائه شده برای تولید جمله [۷]

جدول ۳ نتایج معيار BLUE را برای روش ارائه شده بر روی سه مجموعه داده مختلف، نمایش می‌دهد. همان‌طور که در این جدول مشخص است، نتایج به دست آمده از این روش در مقایسه با دو روش دیگر، نتایج به نسبت بهتری بوده است.

جدول ۳: نتایج معيار BLUE برای روش ارائه شده در [۷] در مقایسه با دو روش دیگر

B-4**	B-3**	B-2**	B-1**	B-3*	B-2*	B-1*	B-3	B-2	B-1	نام روش
۱۰۰	۱۶۶	۲۸.۱	۴۵.۰	—	—	—	—	—	—	نزدیک‌ترین همسایه روش [۳۲]
—	—	—	—	۲۰	۲۴	۵۵	۲۳	۲۸	۵۸	روش [۷]
۲۳۰	۳۲.۱	۴۵.۰	۶۲.۵	۲۴.۰	۳۶.۹	۵۷.۳	۲۴.۵	۳۸.۳	۵۷.۹	روش [۷]

شکل ۳۰ نتایج رتبه‌بندی جملات و عبارات برای هر ناحیه از تصویر را نمایش می‌دهد.



شکل ۳۰: نتایج رتبه‌بندی عبارات زبانی برای نواحی تصویر [۷]

به علاوه، در شکل ۳۱ نتایج تولید شرح برای تصاویر، توسط شبکه عصبی بازگشتی ارائه شده در این پژوهش، به تصویر کشیده شده است.



شکل ۳۱: نتایج تولید جمله برای تصاویر در [۷]

۳-۴ مدل دوطرفه نگاشت تصاویر و جملات مبتنی بر یادگیری عمیق

یکی از مشکلات عمدۀ در روش‌های مبتنی بر یادگیری عمیق، وجود حافظه مناسب برای به خاطر سپاری رخدادهای گذشته است. در شبکه‌های عصبی پیش‌رو عمیق که دارای l لایه هستند، ظرفیت حداکثر حافظه موجود برای رخدادهای گذشته $1 - l$ است و شبکه قادر است تنها $1 - l$ رخداد گذشته را به خاطر بسپارد. شبکه‌های عصبی بازگشتی، تا حد خوبی این مشکل را برطرف می‌نمایند. به همین دلیل، استفاده از این دسته از شبکه‌ها در بخش تولید جمله، منجر به ایجاد نتایج بهتر می‌شود.

با این حال، شبکه‌های عصبی بازگشتی نیز در مواردی که طول جمله زیاد باشد، قادر به به خاطر سپاری مناسب رخدادهای گذشته نیستند. برای رفع این مشکل، معمولاً از واحدهای گیت در شبکه‌های عصبی حافظه کوتاه‌مدت بلند استفاده می‌شود. در پژوهش [۳۳] که توسط خانم مایکولوف^{۹۷} در سال ۲۰۱۰ ارائه شده است، شبکه عصبی ای ارائه شده است که بدون استفاده از واحدهای گیت، قادر به حفظ رخدادهای گذشته دور است. پژوهش [۸] که در سال ۲۰۱۵ توسط آقای زیتنیک و همکارانش ارائه شده است، با استفاده از شبکه عصبی ارائه شده توسط خانم مایکولوف، مدلی دوطرفه برای نگاشت تصاویر و جملات به یکدیگر ارائه شده است که با داشتن تصویر قادر به تولید شرح متناظر و با داشتن شرح، قادر به بازسازی تصویر مربوطه است.

در ادامه، ابتدا مدل مطرح شده توسط خانم مایکولوف را به طور مختصر شرح داده و سپس به بررسی مدل ارائه شده توسط آقای زیتنیک می‌پردازیم.

۱-۳-۴ مدل زبانی مبتنی بر شبکه عصبی بازگشتی

در این قسمت به بررسی مدل زبانی ارائه شده توسط خانم مایکولوف در پژوهش [۳۳] می‌پردازیم. مدل ارائه شده در این پژوهش، یک مدل بسیار ساده از یک شبکه عصبی بازگشتی است. در لایه ورودی شبکه، کلمات موجود در جمله به ترتیب وارد می‌شوند. برای افزایش سرعت عملیات، به جای خود کلمات از نشان^{۹۸} در نظر گرفته شده برای کلمه استفاده می‌شود. برای محاسبه خروجی شبکه می‌توان از روابط (۳۳) تا (۲۹) استفاده نمود که در آن‌ها، کلمه t موجود در جمله، $(1 - s(t))$ بردار حالت شبکه در زمان $t - 1$ ، $u_{j,i}$ وزن مربوط به اتصال ورودی

^{۹۷}Mikolov

^{۹۸}Token

واحد به بردار حالت شبکه، v_{kj} بردار وزن مربوط به بردار حالت شبکه و خروجی آن و y_k خروجی مرحله k ام مدل را نمایش می‌دهند.

$$x(t) = W(t) + s(t - 1) \quad (29)$$

$$s_j(t) = f(\sum_i x_i(t) u_{ji}) \quad (30)$$

$$y_k(t) = g(\sum_j s_j(t) v_{kj}) \quad (31)$$

$$f(z) = \frac{1}{1 + e^{-z}} \quad (32)$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \quad (33)$$

۲-۳-۴ مدل دوطرفه نگاشت تصاویر و جملات با استفاده از شبکه عصبی بازگشتی

در مدل ارائه شده در پژوهش [۸] که توسط آقای زیتنیک در سال ۲۰۱۵ ارائه شد، با تغییر مدل زبانی ارائه شده توسط خانم مایکولوف و تبدیل آن به یک مدل دوطرفه، روشی برای نگاشت دوطرفه تصاویر و جملات به یکدیگر ارائه شده است. در این بخش به بررسی این مدل و نحوه عمل کرد آن به طور اجمالی، خواهیم پرداخت.

در این پژوهش، دو متغیر جدید به مدل زبانی مطرح شده اضافه شده‌اند. متغیر V که بیان گر بردار ویژگی تصویر است و برای منوط کردن معنای جمله به ویژگی‌های تصویر مورد اسفاده قرار می‌گیرد و متغیر U که یک متغیر مخفی است و بیان گر تفسیر بصری آخرين کلمه مشاهده شده یا تولید شده است.

برای تولید یک مدل دوطرفه، کافیست بتوانیم احتمال رخداد جمله به شرط داشتن تصویر و همین‌طور احتمال رخداد تصویر به شرط جمله را محاسبه نماییم. همین‌طور این کار را می‌توان با بخش‌هایی از تصویر و کلمات جمله انجام داد؛ به این معنی که با مدل کردن احتمال رخداد بخش‌هایی از تصویر به شرط داشتن کلمه‌ای از جمله و همین‌طور احتمال رخداد کلمه‌ای در جمله با داشتن بخشی از تصویر به طور همزمان، یک نگاشت دوطرفه بین تصاویر و جملات مرتبط با آن‌ها ایجاد نماییم.

این کار را می‌توان مطابق با رابطه (۳۴) انجام داد. این رابطه، محاسبه‌کننده میزان درست‌نمایی کلمه w_t و بردار ویژگی V به شرط داشتن کلمات قبلی W_{t-1} و تفسیر بصری هر کدام از آن‌ها U_{t-1} است.

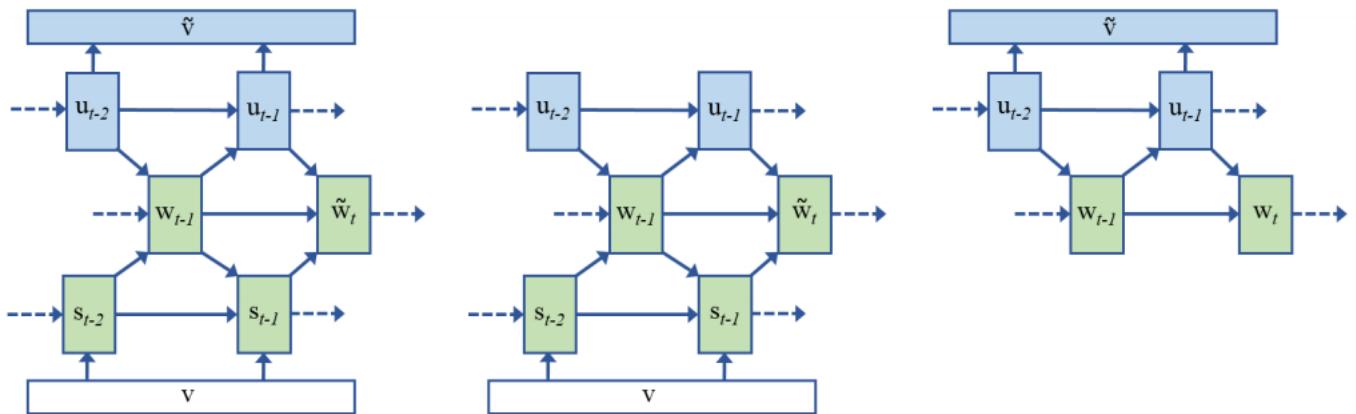
$$P(w_t, V | W_{t-1}, U_{t-1}) = P(w_t | V, W_{t-1}, U_{t-1}) P(V | W_{t-1}, U_{t-1}) \quad (34)$$

همان‌طور که در رابطه (۳۴) مشخص است، می‌توان این رابطه را به شکل حاصل‌ضرب دو عبارت نوشت که هریک از آن‌ها قابلیت مدل‌شدن توسط یک شبکه عصبی بازگشتی را دارند. از طرفی متغیرهای مورد استفاده در هر دو عبارت یکسان است و فقط جهت محاسبات متفاوت است. این نکته باعث می‌شود بتوانیم از یک شبکه عصبی بازگشتی به شکل دوطرفه برای مدل‌سازی کامل رابطه درست‌نمایی توام استفاده نماییم.

شکل ۳۲ ساختار کلی شبکه ارائه شده در این پژوهش را نمایش می‌دهد. در این تصویر، شکل سمت چپ نمایش‌دهنده مدل به طور کامل است و شکل‌های وسط و سمت راست به ترتیب نمایش‌دهنده بخش‌هایی از مدل هستند که برای تولید جمله با داشتن تصویر و تولید تصویر با داشتن جمله مورد استفاده قرار می‌گیرند.

شکل ۳۲ ساختار مدل زبانی ارائه شده توسط خانم مایکولوف را نمایش می‌دهد که متغیرهای V و W به آن اضافه

شده‌اند. اضافه کردن یک لایه V به مدل زبانی، که در شکل با رنگ سفید مشخص شده است، این امکان را می‌دهد که اطلاعات مختلفی را بتوان در مدل زبانی در نظر گرفت. این اطلاعات می‌توانند اطلاعات مربوط به نقش کلمات در جمله، مدل عنوان^{۹۹} و مواردی از این دست باشد. در این پژوهش از بردار ویژگی تصویر که مشخص‌کننده معنای تصویر است برای این قسمت استفاده شده است. این کار باعث می‌شود، معنای جمله تولید شده به محتوای تصویر منوط شود و این ضمانتی است که جمله تولید شده، توصیف‌کننده تصویر باشد.



شکل ۳۲: ساختار کلی شبکه ارائه شده برای نگاشت دوطرفه تصاویر و جملات در پژوهش [۸]

مدل دوطرفه ارائه شده، روی سه مجموعه‌داده Flickr30K، Flickr8k و MS COCO آزمایش شده است. برای بررسی کیفیت عمل کرد مدل، باید در دو آزمایش مجزا، کیفیت نگاشت تصاویر به جملات و همین‌طور کیفیت نگاشت جملات به تصاویر توسط مدل، مورد بررسی قرار گیرند. در این قسمت قصد داریم با گزارش نتایج آزمایشات در قالب جداول و تصاویر، به بررسی عمل کرد مدل بپردازیم.

در جدول ۴، مدل $RNN + IF^{100}$ یک شبکه عصبی بازگشتی است که ویژگی‌های استخراج شده از تصویر نیز به عنوان ورودی به آن داده شده است. مدل $RNN + FT^{101}$ شبکه عصبی با ورودی بردار ویژگی تصویر است که در آن خطای ایجاد شده از خروجی شبکه بازگشتی، به شبکه کانولوشنی نیز منتقل می‌شود و وزن‌های دوشبكه بازگشتی و کانولوشنی با هم به روزرسانی می‌شوند.

علاوه بر جدول فوق که نتایج عمل کرد مدل پیشنهادی را در قالب میزان امتیاز BLEU نمایش داده و با مدل‌های دیگر مقایسه می‌کند، برای بررسی کیفیت عمل کرد مدل، شکل ۳۳ جملات تولید شده مدل را با جملات نوشته شده توسط عوامل انسانی مورد مقایسه قرار می‌دهد. جملات قرمز رنگ در این تصویر، جملاتی هستند که توسط مدل ارائه شده تولید شده‌اند و جملات مشکی‌رنگ، جملاتی هستند که توسط عوامل انسانی نوشته شده‌اند. سطر آخر در این تصویر، نشان‌دهنده تعدادی از نمونه‌هایی است که در آن‌ها جملات تولید شده توسط مدل، دچار خطأ شده‌اند.

علاوه بر موارد فوق، جدول ۵ نتایج بازیابی تصاویر با وارد کردن جمله را در این مدل با مدل‌های دیگر مورد مقایسه قرار می‌دهد. در مدل‌های ارائه شده در این جدول، استفاده از عبارت T در انتهای نام مدل، بیان‌گر این نکته است که در مدل مشخص شده، جملات بر اساس درستنمایی آن‌ها با داشتن تصویر ورودی مرتب شده‌اند.

^{۹۹}Topic Model

^{۱۰۰}Image Feature

^{۱۰۱}Fine Tuned

جدول ۴: امتیاز BLEU کسب شده توسط مدل نگاشت دوطرفه ارائه شده در مقایسه با مدل های دیگر [۸].

نام مدل	Flickr8k	Flickr30k	MS COCO
RNN	۴.۵	۶.۳	۴.۷
RNN + IF	۱۱.۹	۱۱.۳	۱۶.۳
RNN + IF + FT	۱۲.۰	۱۱.۶	۱۷.۰
RNN + VGG	۱۲.۴	۱۱.۹	۱۸.۴
روش ارائه شده	۱۲.۲	۱۱.۳	۱۶.۳
+ FT	۱۲.۴	۱۱.۶	۱۶.۸
+ VGG	۱۳.۱	۱۲.۰	۱۸.۸
انسان	۲۰.۶	۱۸.۹	۱۹.۲

به علاوه، استفاده از عبارت I در نام مدل ها نمایان گر این نکته است که در این مدل ها، از خطای بازسازی تصویر نیز برای مرتب سازی جملات خروجی استفاده شده است.

جدول ۵: جدول نتایج بازیابی تصاویر با استفاده از جملات ورودی در مدل ارائه شده در [۸]

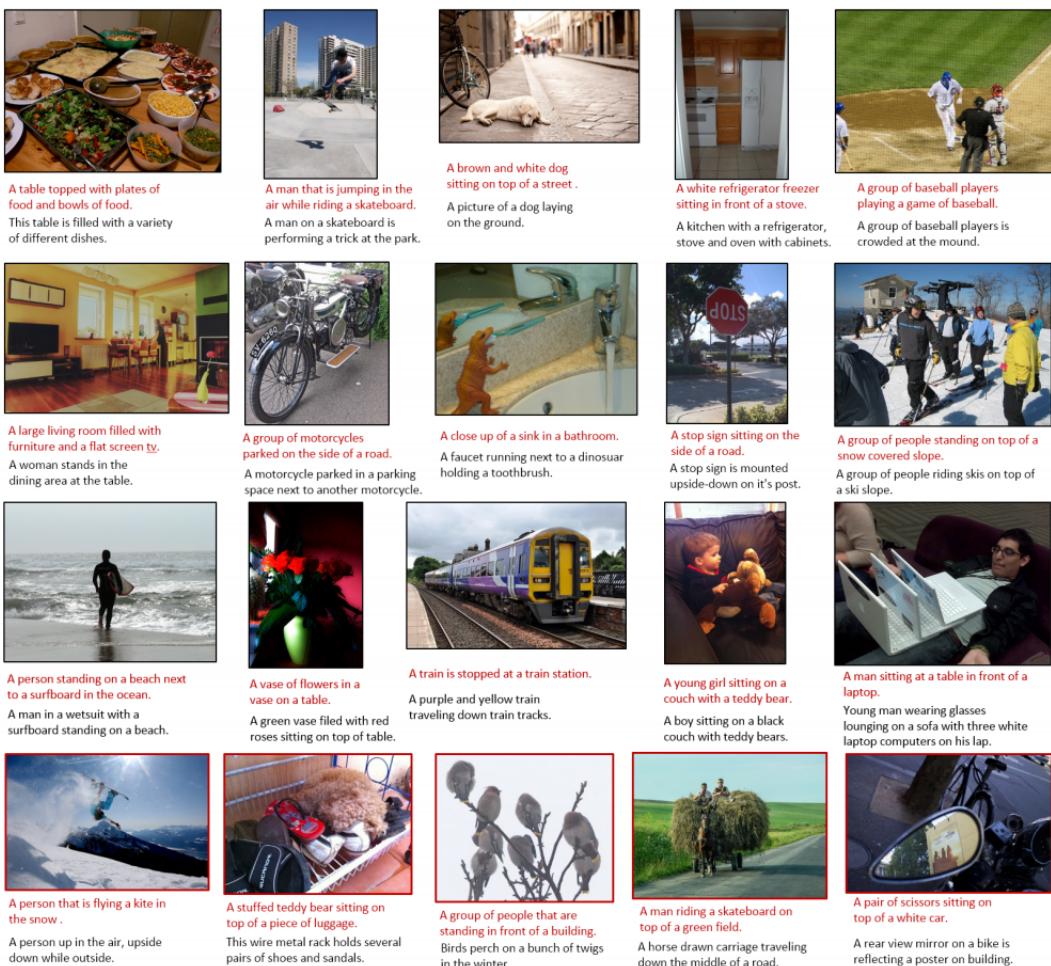
نام مدل	R@1	R@5	R@10	Med r 500
M-RNN	۱۲.۶	۳۱.۲	۴۱.۵	۱۶
RNN + VGG	۱۵.۱	۴۱.۱	۵۴.۱	۹
روش ارائه شده T	۱۷.۷	۴۴.۹	۵۷.۲	۷.۵
روش ارائه شده I + T	۱۸.۵	۴۵.۷	۵۸.۱	۷

۴-۴ جمع بندی

از اواخر سال ۱۳۹۰، روش های مبتنی یادگیری عمیق، نظر بسیاری از پژوهش گرانی را که در حوزه تولید شرح متناظر تصویر فعالیت می کردند، به خود جلب نمودند. این دسته از روش ها، به دلیل عمل کرد بهتری که از خود نشان دادند، توانستند جایگزین روش های گرافی احتمالاتی شوند.

از جمله پژوهش هایی که با استفاده از شبکه های عصبی عمیق اقدام به تولید شرح متناظر تصویر کردند، می توان به پژوهش خانم لی و همکارانش [۷] در سال ۱۵۰ در اشاره کرد. در مرحله آموزش این پژوهش، ابتدا با استفاده از روش شبکه عصبی کانولوشنی ناحیه ای که در بخش قبل، ارائه شد، نواحی تصویر که شامل تصویر یک جسم هستند، انتخاب شده و بردار ویژگی مربوط به هر کدام از این بخش ها، استخراج می شود.

پس از این مرحله، بردار ویژگی مربوط به جملات موجود در مجموعه داده، توسط یک شبکه عصبی بازگشتی دوطرفه، استخراج می شود. برای این کار، ابتدا بردار ویژگی مربوط به هر کلمه با استفاده از یک شبکه کلمه به Word To Vec، استخراج شده و به عنوان ورودی به شبکه بازگشتی دوطرفه داده می شوند. استفاده از شبکه بازگشتی دوطرفه این امکان را می دهد که تاثیر کلمات قبل و بعد از هر کلمه، در تولید بردار ویژگی جملات لحاظ



شکل ۳۳: نمونه‌ای از جملات تولید شده برای تصاویر توسط مدل پیشنهاد شده در [۸]

شود.

با بهینه‌سازی یک تابع انرژی روی این قسمت، شبکه عصبی بازگشتی دوطرفه و شبکه عصبی کانولوشنی با هم آموزش داده می‌شوند. از این طریق، بخش‌هایی از مدل که مربوط به تولید بردار ویژگی از جملات و استخراج نواحی تصاویر و بردار ویژگی مربوط به آن‌ها است، به طور کامل آموزش می‌بینند.

در ادامه فرایند آموزش شبکه، با ارائه بردار ویژگی تولید شده توسط شبکه عصبی کانولوشنی آموزش دیده در بخش قبلی به یک شبکه عصبی بازگشتی دیگر، و ارائه جملات موجود در مجموعه‌داده به آن، شبکه عصبی بازگشتی را برای تولید جمله نهایی آموزش می‌دهیم.

آزمایشات انجام شده روی این پژوهش، معیار BLEU حاصل توسط روش را روی مجموعه‌داده MS COCO در مقایسه با روش‌های دیگر ارزیابی کردند. در این آزمایشات، بهترین عمل کرد روش ارائه شده روی این مجموعه‌داده به امتیاز BLEU برابر با ۵۷.۳ رسیده است و این در حالیست که روش [۳۲] روی همان مجموعه‌داده به مقدار ۵۵.۰ رسیده است.

یکی دیگر از روش‌های ارائه شده در این بخش، روشی است که در سال ۲۰۱۵ ارائه شده است. در این روش، یک شبکه عصبی بازگشتی دوطرفه برای نگاشت جملات و تصاویر به یک دیگر استفاده شده است. مدل ارائه شده، قادر است با گرفتن تصویر به عنوان ورودی، شرح متناظر آن را در قالب یک جمله تولید و با گرفتن یک جمله به عنوان ورودی، تصویر مربوط به آن را با بازیابی نماید.

در این روش با در نظر گرفتن واحد عصبی ارائه شده در پژوهش [۳۳] و اضافه کردن دو متغیر دیگر به آن، مدل نهایی تولید شده است. متغیرهای اضافه شده به این مدل، شامل متغیری برای بردار ویژگی تصویر و متغیر دیگر برای تفسیر بصری آخرين کلمه دیده شده، است.

شبکه عصبی ارائه شده در این پژوهش، توزیع احتمال توام تصاویر و جملات را مدل‌سازی می‌نماید. در صورتی که جمله به عنوان ورودی داده شده باشد، توزیع احتمال تصویر به شرط جمله قابل محاسبه و تصویر مربوطه قابل بازیابی است. در صورتی که تصویر به عنوان ورودی داده شده باشد، توزیع احتمال جمله به شرط تصویر قابل محاسبه است.

نتایج ارائه شده در این پژوهش، با روش‌های دیگر مقایسه شد. برای تولید جمله به شرط داشتن تصویر، میزان امتیاز BLEU حاصل توسط مدل در بهترین حالت برای مجموعه‌داده Flickr8k مقدار ۱۳.۱، برای مجموعه‌داده Flickr30k مقدار ۱۲.۰ و برای مجموعه‌داده MS COCO مقدار ۱۸.۸ بوده است. این در حالیست که نتایج حاصل برای مدل RNN + VGG به ترتیب برابر با ۱۲.۴، ۱۱.۹ و ۱۸.۴ بوده و مقادیر به دست آمده برای جملاتی که توسط عوامل انسانی تولید شده‌اند به ترتیب برابر با ۲۰.۶، ۱۸.۹ و ۱۹.۲ بوده است. نتایج نشان می‌دهد، روش ارائه شده در حوزه تولید شرح متناظر تصاویر از روش‌های استاندارد دیگر بهتر بوده اما هنوز به جملات تولید شده توسط انسان نمی‌رسد.

همین‌طور برای بازیابی تصاویر با داشتن جمله ورودی، نتایج حاصل توسط مدل برای مجموعه‌داده Flickr30k به ترتیب برای معیارهای R@1، R@5 و R@10 مقدار ۵۸.۱، ۴۵.۷ و ۱۸.۵ است. این در حالیست که نتایج حاصل توسط مدل RNN + VGG به ترتیب برابر با ۵۴.۱، ۴۱.۱، ۱۵.۱ و ۹ است.

۵ فصل پنجم

تولید شرح متناظر صحنه با استفاده از
روش‌های مبتنی بر توجه بصری

۱-۵ تولید شرح بر تصاویر با استفاده از روش‌های مبتنی بر توجه بصری

ایده اصلی روش‌های مبتنی بر توجه بصری از پژوهش‌های موجود در زمینه ترجمه ماشینی گرفته شده است. این دسته از پژوهش‌ها مدلی ارائه می‌دهند که با استفاده از آن بتوان هر کلمه از جملات تولیدی را با تمرکز بر یک یا بخشی از کلمات موجود در جمله مبدأ، تولید کرد. به طور مشابه، در حوزه تولید خودکار شرح بر تصاویر، از این دسته از پژوهش‌ها به منظور حصول مدلی استفاده می‌شود که قادر باشد هر یک از کلمات موجود در جمله را با استفاده از بخشی از تصویر ورودی، تولید نماید.

در این فصل، ابتدا ایده اصلی ترجمه مبتنی بر توجه بصری را در حوزه ترجمه ماشینی ارائه خواهیم کرد و سپس کاربردهای این ایده را در حوزه تولید خودکار شرح بر تصاویر مورد بررسی قرار می‌دهیم.

۲-۵ روش‌های مبتنی بر توجه بصری در حوزه ترجمه ماشینی

همان‌طور که گفته شد، تمام روش‌های قبلی را می‌توان به دو مرحله زیر تقسیم کرد.

۱. نگاشت نمونه‌ها از فضای تصاویر به فضای ویژگی‌ها

۲. نگاشت نمونه‌ها از فضای ویژگی‌ها به فضای جملات

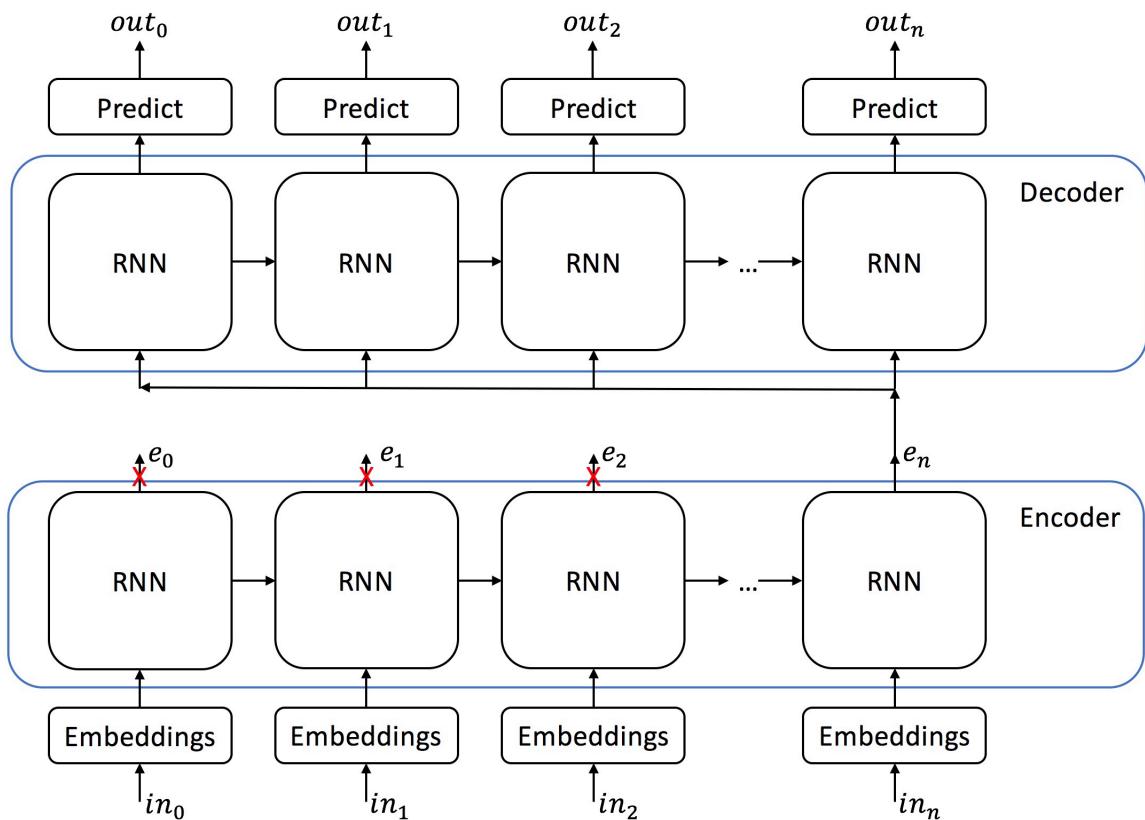
در حوزه ترجمه ماشینی، به تابع نگاشت مرحله اول، انکودر^{۱۰۲} و به تابع نگاشت مرحله دوم، دیکودر^{۱۰۳} گفته می‌شود. در این بخش، ما از این عبارات برای ارجاع به مراحل اول و دوم الگوریتم استفاده می‌نماییم. چارچوب کاری انکودر-دیکودر، در تعداد زیادی از پژوهش‌های حوزه ترجمه ماشینی به عنوان چارچوب کاری اصلی مورد استفاده قرار گرفته است. تمام روش‌های قبلی که در فصول قبل ذکر شد نیز از همین چارچوب کاری به عنوان چارچوب اصلی بهره برده‌اند. به عنوان مثال در روش‌های مبتنی بر یادگیری عمیق برای تولید خودکار شرح بر تصاویر از شبکه‌های عصبی کانولوشنی به طور معمول به عنوان انکودر و از شبکه‌های عصبی بازگشتی به عنوان دیکودر استفاده می‌شود.

شكل ؟؟ نشان‌دهنده ساختار کلی چارچوب کاری انکودر-دیکودر است.

در ادامه به بررسی بخش‌های مختلف این چارچوب کاری می‌پردازیم و سپس ایده اصلی روش‌های مبتنی بر توجه بصری را که توسط آقای بنجیو در پژوهش [۹] در سال ۲۰۱۴ ارائه شده است، مورد بررسی قرار خواهیم داد.

^{۱۰۲}Encoder

^{۱۰۳}Decoder



شکل ۳۴: ساختار کلی چارچوب کاری انکودر-دیکودر

۱-۲-۵ انکودر

انکودر در این چارچوب کاری، با گرفتن یک جمله به عنوان ورودی، بردار ویژگی متناظر جمله مبدا را تولید می‌کند. جمله ورودی با دنباله‌ای از کلمات مدل می‌شود. همین‌طور هر کلمه را با یک بردار n بعدی، که n تعداد کلمات موجود در دیکشنری است، مدل می‌شود. به این ترتیب، هر جمله ورودی، یک بردار با طول متغیر است که هر مولفه آن خودش برداری به ابعاد n است. از طرفی بردار خروجی، که همان بردار ویژگی‌ها است، یک بردار با طول ثابت و قراردادی خواهد بود.

عموماً در کاربردهای ترجمه ماشینی در هر دو بخش انکودر و دیکودر از شبکه‌های عصبی بازگشتی استفاده می‌شود. در شبکه‌های عصبی بازگشتی، خروجی هر مرحله تابعی از ورودی آن مرحله و حالت شبکه در مرحله جاری است. با فرض این‌که h_t حالت شبکه در زمان t را نمایش دهد می‌توان رابطه تولید خروجی توسط شبکه عصبی بازگشتی را مطابق با (۳۵) تعریف نمود.

$$h_t = f(X_t, h_{t-1})$$

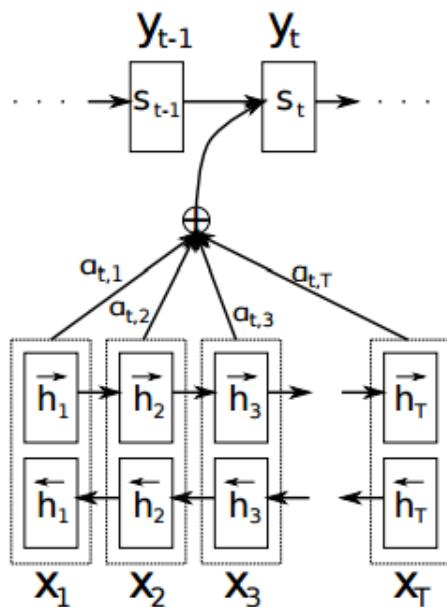
$$C = q(h_1, \dots, h_L) \quad (35)$$

به طور معمول از شبکه LSTM به عنوان تابع f استفاده می‌شود و همین‌طور به جای استفاده از تابع q حالت نهایی شبکه به عنوان بردار ویژگی مورد استفاده قرار می‌گیرد [۹].

دیکودر به منظور نگاشت فضای ویژگی‌ها به فضای جملات مورد استفاده قرار می‌گیرد. خروجی انکودر، ورودی دیکودر است. با این فرض، ورودی دیکودر یک بردار ویژگی با طول ثابت است و خروجی آن که یک جمله به زبان مقصد است، همانند جمله مبدا، یک بردار با طول متفاوت شامل بردارهای بازنمایی کلمات است. دیکودر در اصل در هر مرحله، به دنبال یافتن کلمه‌ای است که با داشتن کلمات تولید شده قبلی و بردار ویژگی موجود، محتمل‌ترین کلمه نسبت به بقیه کلمات موجود در دیکشنری باشد.تابع احتمال مربوطه را می‌توان به فرم (۳۶) تعریف نمود.

$$p(y_t|C, y_1, y_2, \dots, y_{t-1}) = g(y_t, s_t, C) \quad (36)$$

در رابطه (۳۶)، C نشان‌دهنده بردار ویژگی، y_i نشان‌دهنده لغت i ام تولید شده از زبان مقصد و بردار s_t نشان‌دهنده حالت شبکه بازگشتی مورد استفاده به عنوان دیکودر است.
شکل ۳۵ ساختار کلی دیکودر را نمایش می‌دهد.



شکل ۳۵: ساختار دیکودر مورد استفاده در چارچوب کاری [۹]

۳-۲-۵ ایده اصلی استفاده از توجه بصری

همان‌طور که بیان شد، در چارچوب کاری انکودر-دیکودر، ابتدا جمله ورودی که شامل تعداد نامعلوم کلمه است به یک بردار با طول متفاوت مدل می‌شود. بردار تولید شده توسط یک انکودر به یک بردار با طول ثابت، که همان بردار ویژگی‌ها است، نگاشت شده و در نهایت بردار ویژگی تولید شده توسط یک انکودر به یک بردار با طول متفاوت که نماینده جمله زبان مقصد است، نگاشت می‌شود.
فرآیند مذکور یک محدودیت جدی دارد و آن این است که انکودر باید بتواند تمام اطلاعات مورد نیاز برای تولید

جمله را در یک بردار با طول ثابت بگنجاند و دیکودر باید بتواند تمام اطلاعات مورد نیاز خود را از همین بردار با موجود با طول ثابت، استخراج کند. این محدودیت باعث می‌شود قدرت کد کردن اطلاعات در بردار ویژگی کاهش یابد. برای حل این مشکل از ایده نقاط توجه استفاده می‌نماییم.

در این دسته از روش‌ها به جای این‌که انکودر فقط یک بردار ویژگی تولید کند، بردارهای ویژگی مختلفی ایجاد می‌کند که هر بردار با مرکز بر روی یک یا بخشی از جمله مبدا تولید شده است. به این ترتیب، هر بردار تولید شده شامل اطلاعات معنایی یک یا بخشی از جمله مبدا می‌باشد. به این طریق، دیکودر می‌تواند با انتخاب بین بردارهای معنایی تولید شده در هر مرحله، کلمه تولیدی را با مرکز بر روی معنای یک کلمه و کلمات مجاور آن در جمله مبدا، تولید کند.

در ادامه به بررسی تغییراتی که باید در انکودر و دیکودر اتفاق بیفتد تا بتوان به جای یک بردار ویژگی مجموعه‌ای از بردارهای ویژگی با مرکز محلی ایجاد نمود و از آن‌ها برای تولید جمله استفاده نمود را مورد بررسی قرار می‌دهیم. برای سهولت فهم تغییرات، ابتدا تغییرات دیکودر را مطرح نموده و سپس به بررسی تغییرات انکودر خواهیم پرداخت.

۴-۲-۵ دیکودر در روش مبتنی بر توجه بصری

فرض می‌کنیم به جای تنها یک بردار ویژگی، L بردار ویژگی از ورودی استخراج شده باشد. آن‌ها را در یک ماتریس به شکل $C = [c_1, \dots, c_L]^T$ بازنمایی می‌نماییم. فرض می‌کنیم بردار ویژگی c_i به دنباله حاشیه‌نویسی‌های $h = [h_1, \dots, h_L]^T$ وابسته است. حاشیه‌نویسی h_i خود یک متغیر تصادفی به شکل برداری است که دارای دو ویژگی بسیار مهم می‌باشد.

۱. حاوی اطلاعات استخراج شده از تمام جمله است

۲. مرکز استخراج اطلاعات بر روی کلمه نام و کلمات اطراف آن بوده است.

با تعریف این دو ویژگی، حاشیه‌نویسی‌ها را می‌توان همان بردار ویژگی جمله تصور کرد با این شرط که علاوه بر این که معنای کل جمله را کد کرده‌اند، مرکز بیشتری بر معنای کلمه نام و کلمات مجاور آن دارند. به عبارت بهتر هر حاشیه‌نویسی علاوه بر این‌که معنای کلی جمله را کد می‌کند، حاوی معنای محلی مربوط به کلمات هم هست. با تعریف حاشیه‌نویسی به شکل فوق و با تکیه بر فرض‌های انجام شده، می‌توانیم مدل احتمالاتی ارائه شده را به شکل (۳۷) تغییر دهیم.

$$p(y_i|y_1, \dots, y_{i-1}, X) = g(y_{i-1}, S_i, c_i) \quad (37)$$

که در آن:

$$c_i = \sum_{j=1}^L \alpha_{ij} h_j \quad (38)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^L \exp(e_{ik})} \quad (39)$$

$$e_{ij} = f(s_{i-1}, h_j) \quad (40)$$

^{۱۰۴}Annotation

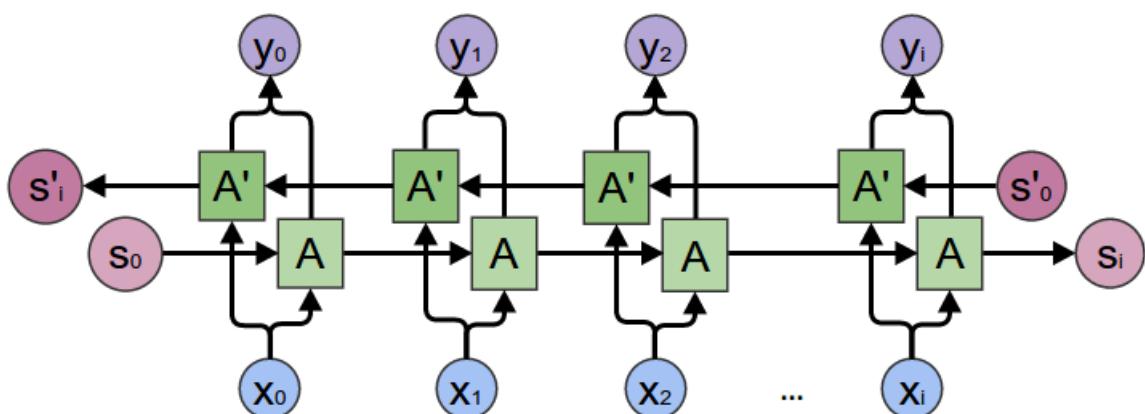
متغیر تصادفی e_{ij} که در رابطه (۴۰) تعریف شده است نمایان‌گر میزان شباهت کلمه i ام در جمله خروجی به کلمه j ام در جمله ورودی است. وظیفه این متغیر، هم‌ترازسازی^{۱۰۵} ورودی و خروجی است. α_{ij} یک نرمال‌سازی روی امتیازهای محاسبه شده انجام می‌دهد. از این متغیر نرمال شده به عنوان وزن حاشیه‌نویسی‌ها استفاده می‌شود. مطابق با رابطه (۳۸) بردار ویژگی مورد استفاده برای تولید کلمه در جمله مقصود، از طریق یک میانگین‌گیری بر اساس وزن معنایی کلمات تولید می‌شود.

در رابطه (۴۰) s_{i-1} بردار حالت شبکه دیکودر در زمان $1 - i$ و f یکتابع امتیاز است. تابع امتیاز مورد استفاده در این رابطه را می‌توان با یک شبکه عصبی پیش‌رو^{۱۰۶} مدل‌سازی کرد. در صورت استفاده از شبکه عصبی پیش‌رو برای مدل‌سازی تابع شباهت، در صورتی که از هم‌ترازسازی نرم^{۱۰۷} استفاده شود، تابع هدف مشتق‌پذیر شده و می‌توانیم از الگوریتم پساننتشار خطاب برای آموزش استفاده نماییم.

۵-۲-۵ انکودر در روش مبتنی بر توجه بصری

برای طراحی انکودر در این بخش، باید مکانیزمی ارائه شود که قادر باشد حاشیه‌نویسی‌های h_1 تا h_L را طوری تولید کند که دو شرط مطرح شده در بخش قبلی را ارضاء نمایند. به عبارت دیگر باید بردارهای ویژگی‌ای استخراج نماییم که علاوه بر این‌که حاوی معنای کل جمله باشند، هر یک از آن‌ها بر روی معنای یک کلمه و کلمات اطراف آن تمرکز بیشتری نسبت به سایر بردارها داشته باشند تا بتوانیم علاوه بر مدل‌سازی معنای کلی جمله، از معنای محلی کلمات هم استفاده نماییم.

به این منظور از یک شبکه عصبی بازگشتی دوطرفه در مدل‌سازی انکودر استفاده می‌نماییم. شکل ۳۶ ساختار کلی یک شبکه عصبی بازگشتی دوطرفه را نمایش می‌دهد.



شکل ۳۶: ساختار کلی یک شبکه عصبی بازگشتی دوطرفه

همان‌طور که در شکل ۳۶ مشخص است، یک شبکه عصبی بازگشتی دوطرفه شامل دو شبکه پیش‌رو در خلاف جهت یکدیگر است. حالت‌های مخفی شبکه پیش‌رو را به راست را با $\rightarrow h_i$ و حالت‌های مخفی شبکه پیش‌رو را به چپ را با $\leftarrow h_i$ نمایش می‌دهیم. همان‌طور که در شکل پیداست، خروجی‌های شبکه در این ساختار هم به حالت‌های سمت راست و کلمات سمت راست در جمله و هم به حالات و کلمات سمت چپ وابسته هستند. پس همین

^{۱۰۵} Alignment

^{۱۰۶} Feed Forward Neural Network

^{۱۰۷} Soft Alignment

خروجی‌ها را می‌توان به عنوان حاشیه‌نویسی‌هایی که هر دو ویژگی را دارند مورد استفاده قرار داد. یکی از راه‌های ساده برای ایجاد حاشیه‌نویسی با استفاده از حالات شبکه‌های پیش‌رو و رو به راست و رو به چپ این است که مطابق با رابطه (۴۱) با پشت سر هم قرار دادن حالات شبکه، حاشیه‌نویسی مورد نیاز را تولید نماییم.

$$h_j = [h_j^{\rightarrow T}, h_j^{\leftarrow T}]^T \quad (41)$$

۳-۵ روش‌های مبتنی بر توجه بصری در حوزه تولید شرح متناظر تصویر

در بخش قبل به بیان ایده اصلی روش‌های مبتنی بر توجه بصری در حوزه ترجمه ماشینی پرداختیم. ساختار کلی انکودرها و دیکودرها در این قالب و همین‌طور نحوه تولید بردارهای ویژگی مختلف از جمله مبدا و استفاده از این بردارها در تولید جمله مقصد را مورد بررسی قرار دادیم. در این بخش به بررسی پژوهش‌های خواهیم پرداخت که از این ایده در حوزه تولید شرح متناظر تصویر بهره جسته‌اند.

یکی از برجسته‌ترین و مورد توجه‌ترین پژوهش‌ها از این دست، پژوهشی است که آقای بنجیو و همکارانش در سال ۲۰۱۵ ارائه داده‌اند [۱۰]. در این بخش به بررسی این پژوهش خواهیم پرداخت.

۱-۳-۵ تولید شرح متناظر تصویر با استفاده از توجه بصری و شبکه‌های عصبی [۱۰]

در این پژوهش که در سال ۲۰۱۵ توسط آقای بنجیو و همکارانش ارائه شده است از ایده استفاده از توجه در حوزه ترجمه ماشینی استفاده شده است تا شرح متناظر تصاویر با دقت بیشتری تولید شود. چارچوب کاری انکودر-دیکودر مانند آن‌چه در بخش قبلی مطرح شد در این پژوهش مورد استفاده قرار گرفته است. انکودر ارائه شده در این پژوهش، یک شبکه عصبی کانولوشنی است که قادر به تولید L بردار ویژگی مختلف است. به هر یک از این بردارهای ویژگی یک حاشیه‌نویسی^{۱۰۸} تصویر گفته می‌شود. بردارهای حاشیه‌نویسی، همان‌طور که در بخش قبل ذکر شد، باید دارای دو شرط زیر باشند:

۱. حاوی معنای تصویر به طور کلی باشند.

۲. تمرکز بیشتری روی یکی از بخش‌های تصویر داشته باشند.

برای این‌که بتوانیم دو شرط فوق را در بردارهای حاشیه‌نویسی تولید شده از انکودر بگنجانیم از خروجی لایه ما قبل آخر شبکه عصبی کانولوشنی به عنوان بردارهای حاشیه‌نویسی استفاده می‌کنیم. هر بردار حاشیه‌نویسی یک بردار D بعدی است که مربوط به یک بخش از تصویر می‌شود و آن را با a_i نمایش می‌دهیم. بنابر این داریم:

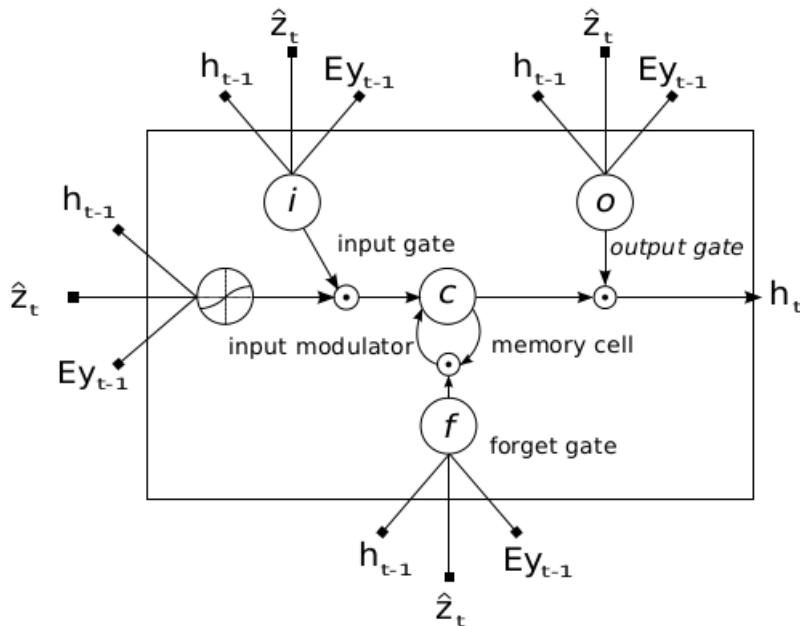
$$a = \{a_1, a_2, \dots, a_L\}, a_i \in R^D \quad (42)$$

در این پژوهش از یک شبکه حافظه کوتاه‌مدت بلند به عنوان دیکودر استفاده شده است. این شبکه با دریافت مجموعه بردارهای حاشیه‌نویسی a ، جمله‌ای به زبان انگلیسی تولید می‌کند که شامل دنباله‌ای از C کلمه است. هر کلمه با یک بردار K بعدی نمایش داده می‌شود که K تعداد کلمات موجود در دیکشنری است. در هر یک از بردارهای بازنمایی کلمات فقط یک مولفه یک است و ماقبی مولفه‌ها صفر هستند. مولفه‌ای که برابر با یک است

^{۱۰۸}Annotation

نمایش دهنده اندیس کلمه در دیکشنری است.

شکل ۳۷ یک سلول از شبکه حافظه کوتاهمدت بلند مورد استفاده در این پژوهش به عنوان دیکودر را نمایش می‌دهد. روابط مربوط به یادگیری این شبکه را می‌توان مطابق با روابط (۴۳) تا ۴۸ نمایش داد. در همه روابط،



شکل ۳۷: یک واحد از شبکه حافظه کوتاهمدت بلند مورد استفاده در دیکودر پژوهش [۱۰]

تابع T یک تابع نگاشت خطی به شکل $T : R^{D+m+n} * R^n$ است که پارامترهای آن آموزش داده شده‌اند. متغیر i_t ورودی، f_t خروجی سلول فراموشی، c_t حافظه، o_t خروجی و h_t حالت مخفی شبکه را نمایش می‌دهند.

$$i_t = \sigma(T(Ey_{t-1}, h_{t-1}, \hat{z}_{t-1})) \quad (43)$$

$$f_t = \sigma(T(Ey_{t-1}, h_{t-1}, \hat{z}_{t-1})) \quad (44)$$

$$o_t = \sigma(T(Ey_{t-1}, h_{t-1}, \hat{z}_{t-1})) \quad (45)$$

$$g_t = \sigma(T(Ey_{t-1}, h_{t-1}, \hat{z}_{t-1})) \quad (46)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (47)$$

$$h_t = o_t \odot \tanh(c_t) \quad (48)$$

بردار \hat{z}_{t-1} بردار معنای تصویر را نمایش می‌دهد که با استفاده از بردارهای حاسیه‌نویسی تولید شده در انکودر تولید می‌شود. ماتریس E ، ماتریس جانمایی 109 به ابعاد $K * m$ است. تابع σ تابع فعالیت سیگموئیدی و \odot حاصل ضرب مولفه‌های نظیر به نظیر بردارها را نمایش می‌دهند.

فرایند آموزش دیکودر کاملاً مطابق با فرایند آموزش معمول شبکه حافظه کوتاهمدت بلند است. تنها تفاوت در این پژوهش وجود و نحوه محاسبه بردار معنای \hat{z}_{t-1} است که توجه بصری را تعریف می‌کند. برای محاسبه این متغیر

¹⁰⁹Embedding Matrix

تابع ϕ را تعریف می‌نماییم. این تابع در هر لحظه از زمان با استفاده از مجموعه بردارهای حاشیه‌نویسی a برداری تولید می‌کند که به عنوان بردار ویژگی استخراج شده از تصویر در هر لحظه مورد استفاده قرار می‌گیرد.

تابع phi می‌تواند به دو شکل بردار ویژگی را تولید نماید. روش اول این است که ابتدا با تولید وزن‌های مثبت α_i برای هر ناحیه از تصویر با بردار حاشیه‌نویسی a_i یک احتمال برای میزان مناسب بودن ناحیه i از تصویر برای استفاده در تولید کلمه در زمان t تعریف شود. سپس بردار حاشیه‌نویسی با بیشترین احتمال برای تولید کلمه انتخاب شده و به مراحل بعدی ارسال شود. این روش را تحت عنوان روش توجه سخت^{۱۰} نام‌گذاری می‌نماییم. روش دوم برای تولید بردار ویژگی تصویر در هر لحظه توسط تابع ϕ این است که اعداد مثبت تولید شده α_i را به طور مستقیم به عنوان معیاری جهت سنجش میزان مناسب بودن نسبی نواحی نسبت به یکدیگر مورد استفاده قرار دهیم و با استفاده از یک میانگین‌گیری وزن‌دار بر حسب همین وزن‌های مثبت از بردارهای حاشیه‌نویسی اقدام به تولید بردار ویژگی تصویر نماییم. به این روش، روش توجه نرم^{۱۱} می‌گوییم.

به جهت سهولت در امر رابطه‌بندی توجه بصری در فرایند آموزش انکودر و دیکودر، متغیر تصادفی $s_{t,i}$ را معرفی می‌نماییم که نشان‌دهنده این است که آیا در زمان t از بردار ویژگی مربوط به ناحیه i ام تصویر، برای تولید کلمه استفاده می‌شود یا خیر. اگر در زمان t از بردار ویژگی ناحیه i ام، که همان بردار حاشیه‌نویسی با اندیس i است، به منظور تولید کلمه استفاده شود، مقدار متغیر $s_{t,i}$ برابر با یک و در غیر این صورت برابر با صفر قرار می‌گیرد.

با استفاده از متغیر تصادفی تعریف شده می‌توان به راحتی روابط مربوط به مدل‌سازی توجه بصری نرم و سخت را به شرح زیر تشکیل داد. نکته آخر این‌که در پژوهش مورد بررسی، به منظور تولید احتمال کلمه بعدی با توجه به کلمات تولید شده قبلی و بردار ویژگی استخراج شده از تصویر از رابطه (۴۹) استفاده شده است.

$$p(Y_t|a, Y_1^{t-1}) \propto \exp(L_o(EY_{t-1} + L_h h_t + L_z \hat{z}_t)) \quad (49)$$

در رابطه (۴۸) ماتریس‌های شبکه هستند که باید آموزش داده شوند.

۱. توجه بصری سخت

با فرض یک توزیع Multinoulli مطابق رابطه (۵۰) می‌توان متغیر \hat{z}_t را به عنوان بردار ویژگی استخراج شده نهایی با توجه به بردارهای حاشیه‌نویسی a و توجه بصری بصری، به شکل رابطه (۵۱) محاسبه نمود.

$$p(s_{t,i} = 1 | s_{j < t}, a) = \alpha_{t,i} \quad (50)$$

$$\hat{z}_t = \sum_t s_{t,i} a_i \quad (51)$$

برای آموزش وزن‌های شبکه یک تابع هدف به نام L_s مطابق با رابطه (۵۲) مطرح می‌شود که یک کران پایین از بیشینه درستنما می‌باشد $p(Y|a)$ است که در آن Y دنباله کلمات تولید شده نهایی و a بردارهای حاشیه‌نویسی تولید شده از روی تصویر را نمایش می‌دهند.

^{۱۰}Hard Attention

^{۱۱}Soft Attention

$$\log p(Y|a) = \log \sum_s p(s|a)p(Y|s, a) \geq \sum_s p(s|a) \log p(Y|s, a) = L_s \quad (52)$$

با ارائه تابع هدف L_s مطابق با رابطه (52) و بهینه‌سازی آن می‌توان رابطه بهروزرسانی وزن‌ها در فرایند آموزش را محاسبه نمود. رابطه (53) محاسبات مربوطه را نمایش می‌دهد.

$$\frac{\partial L_s}{\partial W} = \sum_s p(s|a) \left[\frac{\partial \log p(Y|s, a)}{\partial W} + \log p(Y|s, a) \frac{\partial \log p(s|a)}{\partial W} \right] \quad (53)$$

به جای متغیر s_t در رابطه (53) می‌توان با استفاده از روش نمونه‌برداری مونت‌کارلو^{۱۲} نمونه‌های تصادفی \tilde{s}_t تولید کرد و سپس با استفاده از رابطه (54) تابع هدف را بهینه نمود.

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^N \left[\frac{\partial \log p(Y|\tilde{s}^n, a)}{\partial W} + \log p(Y|\tilde{s}^n, a) \frac{\partial \log p(\tilde{s}^n|a)}{\partial W} \right] \quad (54)$$

۲. توجه بصری نرم

همان‌طور که گفته شد، تولید بردار ویژگی را می‌توان با میانگین‌گیری وزن‌دار روی بردارهای حاشیه‌نویسی انجام داد. در شرایطی که وزن‌های تخصیص داده شده به بردارهای حاشیه‌نویسی برابر صفر نباشند، توجه بصری نرم، فرایندی خواهد بود شامل تولید بردار ویژگی با استفاده از تمام حاشیه‌نویسی‌های موجود و با تمرکز روی تعدادی از حاشیه‌نویسی‌ها که ضریب بیشتری دارند. از آنجا که این شیوه محاسبه بردار ویژگی شامل یک بردار میانگین‌گیری وزن‌دار است، تمام تابع هدف مشتق‌پذیر شده و امکان استفاده از روش پس‌انتشار خطابه برای یادگیری وزن‌ها فراهم می‌شود. در این روش به طور کلی می‌توان بردار ویژگی \hat{z}_t را مطابق با رابطه (55) محاسبه نمود.

$$E_{p(s_t|a)}[\hat{z}_t] = \sum_{i=1}^L \alpha_{t,i} a_i \quad (55)$$

مطابق با رابطه (46) حالت مخفی شبکه یک ترکیب خطی از بردار ویژگی استخراج شده از تصویر به همراه یک غیرخطی‌سازی با استفاده از تابع \tanh است. برای تقریب مرتبه اول حالت مخفی شبکه می‌توان از امید ریاضی بردار ویژگی \hat{z}_t در رابطه (46) استفاده کرد. با در نظر گرفتن رابطه (48) می‌توان متغیر n_t را به شکل (55) تعریف نمود. با این تعریف، متغیر $n_{t,i}$ مشخص کننده متغیر

^{۱۲}Monte Carlo Sampling

n_t است در شرایطی که $a_i = \hat{z}_t$ باشد. با استفاده از متغیر تعریف شده، میانگین هندسی وزن دار نرمال شده 113 را برای تولید کلمه k مطابق با رابطه (۵۶) تعریف می نماییم.

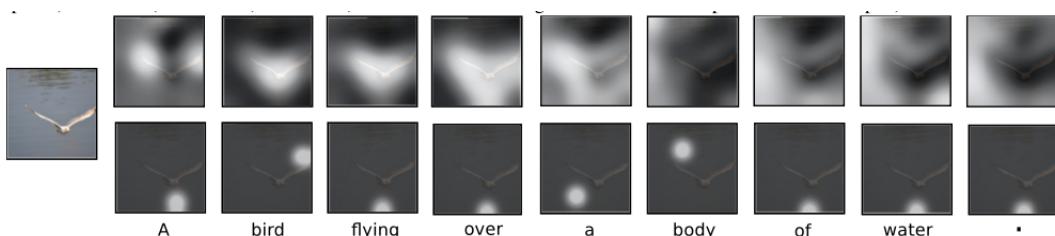
$$NWGM[P(y_t = k|a)] = \frac{\prod_i \exp(n_{t,k,i})^{p(s_{t,i}=1|a)}}{\sum_j \prod_i \exp(n_{t,j,i})^{p(s_{t,i}=1|a)}} = \frac{\exp(E_{p(s_{t,i}|a)}[n_{t,k}])}{\sum_j \exp(E_{p(s_{t,i}|a)}[n_{t,j}])} \quad (56)$$

از آنجا که $E[n_t] = L_o(EY_{t-1} + L_hE[h_t] + L_zE[\hat{z}_t])$ می تواند به خوبی توسط بردار ویژگی \hat{z}_t تخمین زده شود. این بدین معناست که میانگین هندسی وزن دار نرمال شده لایه نهایی شبکه می تواند با اعمال تابع $soft max$ به امید ریاضی ترکیبات خطی لایه های پایین تر محاسبه شود.

آزمایشات انجام شده در این پژوهش روی سه مجموعه داده Microsoft COCO و Flickr30k و Flickr8k اجرا شده است که به ترتیب شامل ۳۰۰۰۰، ۸۰۰۰ و ۸۲۷۸۳ تصویر با شرح تولید شده توسط عوامل انسانی هستند. دو مجموعه داده اول برای هر تصویر، ۵ شرح مختلف و مجموعه داده سوم در برخی تصاویر بیش از ۵ شرح را شامل می شوند. در تمام پژوهش ها به منظور یکسان سازی آزمایشات و نتایج، از ۵ شرح برای هر تصویر استفاده شده است. هر دو نوع محاسبه توجه بصری در این پژوهش مورد آزمایش قرار گرفته اند و نتایج هریک به طور جداگانه بیان شده است. در این پژوهش از معیارهای BLEU و METEOR به منظور ارزیابی مدل استفاده شده است. همان طور که در جدول ۶ مشخص است، در هر سه مجموعه داده، پژوهش [۱۰] بهترین عملکرد را نسبت به روش های دیگر از خود نشان داده است. استفاده از توجه بصری نرم در مجموعه داده های MS COCO Flickr30k و Flickr8k عمل کرد بهتری نسبت به روش های دیگر از لحاظ معیار METEOR از خود نشان داده است. همین طور استفاده از توجه بصری سخت، بهترین عملکرد را در معیار BLEU از خود نشان داده است.

یکی از فعالیت های مفید برای بررسی نحوه عملکرد مدل که در این پژوهش مورد استفاده قرار گرفته است، بصری کردن فرایند تولید کلمه توسط مدل است. در این پژوهش، توجه بصری روی تصویر در هر مرحله به همراه کلمه تولید شده در هر مرحله مشخص شده اند که در درک نحوه عمل کرد مدل و همین طور پیدا کردن دلایل ایجاد کلمات غیر مرتبط بسیار کمک کننده هستند.

شکل ۳۸ توجه بصری در هر زمان را برای تولید هر کلمه برای یک تصویر نمونه نمایش می دهد. ردیف بالا نمایش دهنده عمل کرد روش با استفاده از توجه بصری نرم و ردیف پایین نمایش دهنده عمل کرد روش با استفاده از توجه بصری سخت است. در این نمونه خاص، نتیجه تولید جمله برای هر دو روش یکسان بوده است.



شکل ۳۸: نحوه عمل کرد الگوریتم در تغییر توجه بصری با توجه به کلمه تولید شده در هر نقطه. [۱۰]

^{۱۱۳}Normalized Weighted Geometric Mean (NWGM)

جدول ۶: نتایج اعمال روش [۱۰] بر روی مجموعه‌داده‌های مختلف در مقایسه با روش‌های مختلف.

METEOR	BLEU-4	BLEU-3	BLEU-2	BLEU-1	نام مدل	مجموعه‌داده
–	–	۲۷.۰	۴۱.۰	۶۳.۰	Google NIC	Flickr8k
۱۷.۳۱	۱۷.۷	۲۷.۷	۴۲.۴	۶۵.۶	Log Bilinear	Flickr8k
۱۸.۹۳	۱۹.۵	۲۹.۹	۴۴.۸	۶۷.۰	Soft Attention	Flickr8k
۲۰.۳۰	۲۱.۳	۳۱.۴	۴۵.۷	۶۷.۰	Hard Attention	Flickr8k
–	۱۸.۳	۲۷.۷	۴۲.۳	۶۶.۳	Google NIC	Flickr30k
۱۶.۸۸	۱۷.۱	۲۵.۴	۳۸.۰	۶۰.۰	Log Bilinear	Flickr30k
۱۸.۴۹	۱۹.۱	۲۸.۸	۴۳.۴	۶۶.۷	Soft Attention	Flickr30k
۱۸.۴۶	۱۹.۹	۲۹.۶	۴۳.۹	۶۶.۹	Hard Attention	Flickr30k
۲۰.۴۱	–	–	–	–	CMU/MS Research	MS COCO
۲۰.۷۱	–	–	–	–	MS Research	MS COCO
–	۲۰.۳	۳۰.۴	۴۵.۱	۶۴.۲	BRNN	MS COCO
–	۲۴.۶	۳۲.۹	۴۶.۱	۶۶.۶	Google NIC	MS COCO
۲۰.۰۳	۲۴.۳	۳۴.۴	۴۸.۹	۷۰.۸	Log Bilinear	MS COCO
۲۳.۹۰	۲۴.۳	۳۴.۴	۴۹.۲	۷۰.۷	Soft Attention	MS COCO
۲۳.۰۴	۲۵.۰	۳۵.۷	۵۰.۴	۷۱.۸	Hard Attention	MS COCO

در تمام تصاویر، محدوده‌های روش‌نتر، محدوده‌هایی هستند که در آن‌ها ضریب میانگین‌گیری بیشتر بوده و توجه بیشتری در محاسبات روی آن‌ها متمرکز شده است. شکل ۳۹ چند نمونه از تصاویر را نمایش می‌دهد که در آن‌ها توجه بصری روی یک جسم منجر به تولید کلمه دقیق متناظر آن جسم شده است. کلمه تولید شده در شرح نهایی تولید شده برای تصویر در زیر هر تصویر نمایش داده شده است.

به علاوه، شکل ۴۰ نمایش‌دهنده شرایطی است که در آن کلمه تولید شده متناظر توجه بصری بصری نیست. با استفاده از بصری‌سازی محل توجه بصری و کلمه تولید شده در هر مرحله، می‌توان به راحتی مشاهده کرد که کلمه تولید شده متناظر کدام نقطه از تصویر، نامناسب است.

علاوه بر موارد فوق، نمونه‌ای از بررسی تمام مراحل تولید شرح متناظر صحنه برای یک تصویر را در حالت‌های استفاده از توجه بصری سخت در شکل ۴۱ و توجه بصری نرم در شکل ۴۲ قابل مشاهده است. هر کلمه تولید شده در هر مرحله در کنار میزان فعال‌سازی شبکه مربوط به آن کلمه نمایش داده شده است.

۴-۵ فعالیت‌های مشابه دیگر

استفاده از توجه بصری در حوزه تولید شرح متناظر تصویر، از سال ۲۰۱۵، توجه بسیاری از پژوهش‌گران را به خود جلب نموده است و پژوهش‌های زیادی با استفاده از این ایده سعی در تولید جمله برای تصاویر، ویدئوها، صوت و انواع ورودی‌های مشابه نموده‌اند. همین‌طور در حوزه ترجمه ماشینی، نسخه‌های متفاوت و متنوعی از این ایده



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.

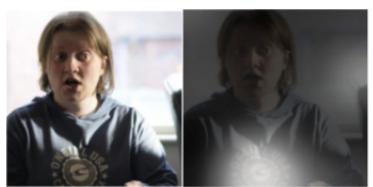


A giraffe standing in a forest with trees in the background.

شکل ۳۹: چند نمونه از تصاویر که در آن ها توجه بصری روی یک جسم منجر به تولید کلمه دقیق متناظر شده است [۱۰].



A large white bird standing in a forest.



A woman holding a clock in her hand.



A man wearing a hat and a hat on a skateboard.



A person is standing on a beach with a surfboard.



A woman is sitting at a table with a large pizza.



A man is talking on his cell phone while another man watches.

شکل ۴۰: نمونه هایی از تولید کلمات نامناسب مطابق با نقاط توجه استفاده شده در مدل [۱۰]

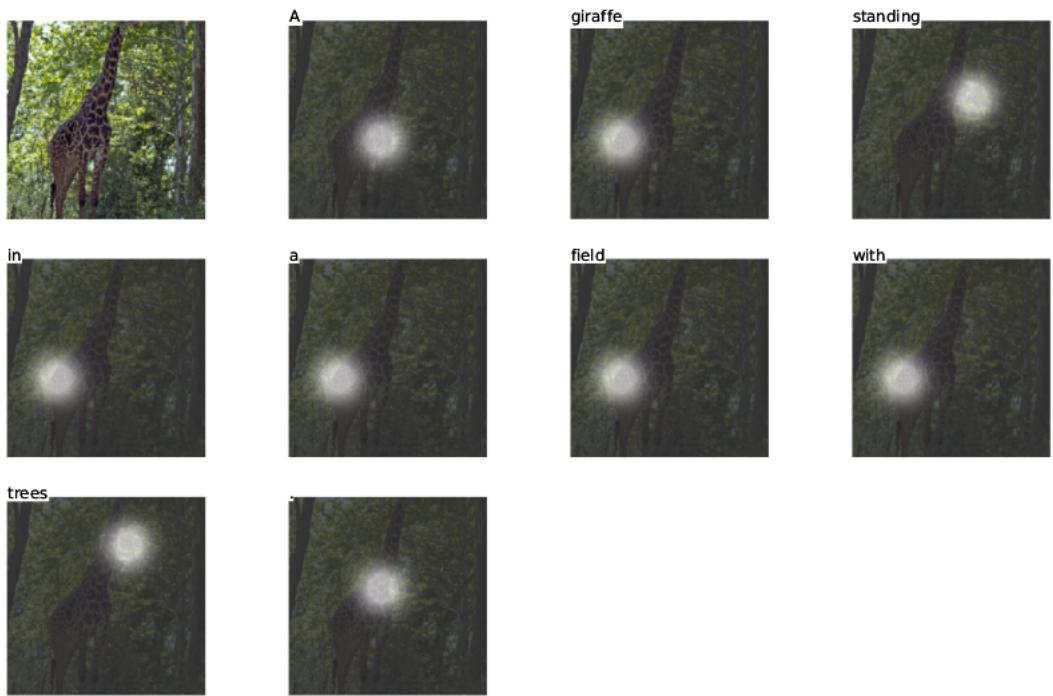
برای دست یابی به ترجمه های بهتر ارائه شده اند.

یکی از پژوهش هایی که در این زمینه برای بهبود عمل کرد ترجمه ماشینی با استفاده از نقطه توجه ارائه شده است، پژوهشی است که آقای منینگ و همکارانش در سال ۲۰۱۵ ارائه دادند [۱۱]. در این پژوهش، که بر روی مجموعه داده WMT که شامل جملات انگلیسی و معادل آلمانی آنها است اجرا شده، از یک ساختار پشتی ای مطابق شکل ۴۳ استفاده شده است. در این ساختار، برای آموزش، جمله انگلیسی و معادل آلمانی آن به یک دیگر الصاق شده و ساختار انکوادر-دیکوادری با هم آموزش می بینند. سپس با دریافت ورودی جمله انگلیسی یا آلمانی، معادل آنها تولید می شود.

در پژوهش ارائه شده نیز مانند پژوهشی که آقای بنجیو در سال ۲۰۱۵ در حوزه ترجمه ماشینی انجام دادند، دو نوع توجه محاسبه و مورد آزمایش قرار گرفته است. توجه اول که معادل توجه نرم است، تحت عنوان توجه سراسری ^{۱۱۴} و توجه دوم که معادل توجه سخت است، تحت عنوان توجه ناحیه ای ^{۱۱۵} مطرح شده اند. در این آزمایش هم مشابه نتایج پژوهش [۱۰] توجه ناحیه ای، در بسیاری موارد عمل کرد بهتری نسبت به توجه سراسری از خود

^{۱۱۴}Global Attention

^{۱۱۵}Local Attention

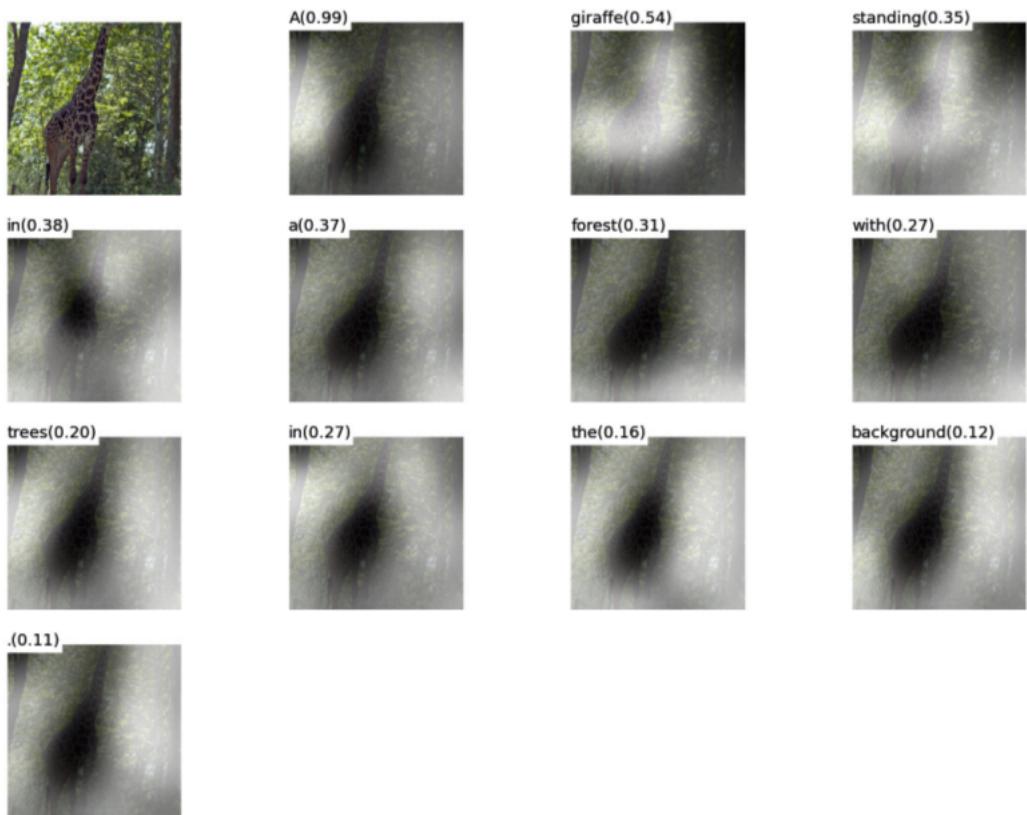


(a) A giraffe standing in a field with trees.

شکل ۴۱: فرایند تولید شرح متناظر تصویر با استفاده از توجه بصری سخت [۱۰]

نشان داده است.

پژوهش مشابهی در حوزه پرسش و پاسخ بصری توسط ایده مشابه پژوهش [۱۱] در سال ۲۰۱۶ توسط آقای ینگ و همکارانش در [۳۴] ارائه شده است. در این پژوهش، بردار ویژگی تصویر به شرح متناظر آن که در مجموعه داده موجود است، الصاق شده و ساختار انکودر-دیکودری به شکل مشابهی آموزش می‌بینند. سپس با ورود یک تصویر جدید، بردار ویژگی آن به ساختار داده شده و جمله مرتبط با تصویر جدید توسط ساختار تولید می‌گردد. آقای بنجیو در پژوهش [۳۵] در سال ۲۰۱۵، چارچوب کاری ای را مبتنی بر استفاده از نقطه توجه ارائه کردند که قابل استفاده در حوزه‌های ترجمه ماشینی، تولید شرح متناظر تصویر، توصیف ویدئو و گفتار است. در این پژوهش، علاوه بر ارائه یک روش برای محاسبه نقطه توجه بصری و استخراج بردار ویژگی با استفاده از بردارهای حاشیه‌نویسی، صحبت‌هایی در مورد امکان انتقال یادگیری در چارچوب ارائه شده انجام شده است. در این پژوهش در مورد هر یک از چهار حوزه‌ای که ذکر شد، صحبت شده و نحوه استفاده از چارچوب کاری در هر یک از این حوزه‌ها تبیین شده است. همین‌طور در این پژوهش اثبات شده است که استفاده از مکانیزم نقطه توجه، این امکان را به مدل می‌دهد که به طور بدون نظارت، رابطه همترازی بین بخش‌های ورودی و خروجی را یاد بگیرد تا بتوان از این ویژگی در انتقال یادگیری استفاده نمود. نتایج استفاده از این چارچوب کاری در حوزه‌های مختلف بهتر از روش‌های دیگر گزارش شده است. شکل ۴۴ ساختار چارچوب را در حوزه تولید شرح متناظر تصویر نمایش می‌دهد.



(b) A giraffe standing in a forest with trees in the background.

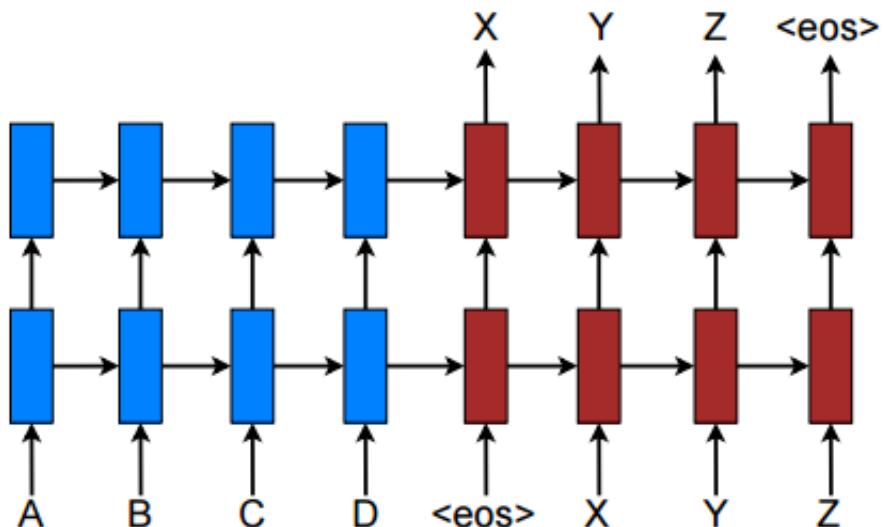
شکل ۴۲: فرایند تولید شرح متناظر تصویر با استفاده از توجه بصری نرم [۱۰]

۵-۵ جمع‌بندی

چارچوب کاری انکودر-دیکودر یکی از اصلی‌ترین چارچوب‌های کاری در حوزه ترجمه ماشینی و پیرو آن تولید شرح متناظر تصویر به شمار می‌رود. انکودر در این چارچوب کاری وظیفه نگاشت ورودی به فضای معنا و دیکودر وظیفه نگاشت فضای معنا به فضای خروجی را بر عهده دارد. در حوزه ترجمه ماشینی معمولاً از یک شبکه عصبی حافظه کوتاه‌مدت بلند به عنوان دیکودر استفاده می‌شود. این شبکه عصبی با دریافت کلمات جمله ورودی به ترتیب، بردار حالت مخفی خود را به روزرسانی می‌نماید. در نهایت می‌توان از این بردار به عنوان بردار حاصل نگاشت جمله ورودی به فضای معنا استفاده نمود.

دیکودر در این چارچوب کاری با دریافت بردار ویژگی تولید شده توسط دیکودر، عمل تولید خروجی را بر عهده داشت. در حوزه ترجمه ماشینی معمولاً یک شبکه عصبی بازگشتی برای دیکودر می‌تواند مورد استفاده قرار بگیرد. به طور معمول، بردار ویژگی تولید شده توسط انکودر، به عنوان یک ورودی به دیکودر داده می‌شود و دیکودر در هر مرحله با تولید یک کلمه به عنوان خروجی، بردار حالت مخفی خود را به روزرسانی نموده و با استفاده از بردار حالت مخفی جدید، اقدام به تولید کلمه جدید می‌نماید.

یکی از محدودیت‌های جدی فرایند مذکور این است که بردار ویژگی فقط یک بردار با طول ثابت است و اولاً انکودر باید بتواند تمام اطلاعات قابل استخراج را تنها در این بردار جاسازی نماید و ثانیاً دیکودر باید بتواند تمام اطلاعات مورد نیاز خود برای تولید کلمه و جمله را فقط از همین یک بردار استخراج نماید. این مشکل، پژوهش‌گران را بر



شکل ۴۳: ساختار پشتهای ارائه شده در [۱۱]

آن داشت تا بردار ویژگی را از یک بردار با طول ثابت به یک دنباله بردار با طول ثابت و تعداد متغیر تغییر دهنده. به بردارهای ویژگی تولید شده در حالت جدید، حاشیه‌نویسی می‌گویند. این حاشیه‌نویسی‌ها باید دارای دو شرط زیر باشند:

۱. دربرگیرنده تمام معنای ورودی باشند.

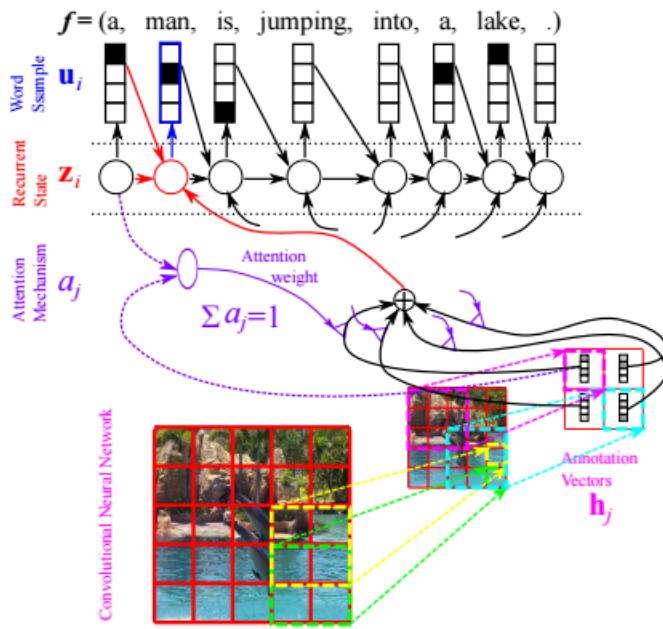
۲. تمرکز بیشتری روی معنای یک بخش مشخص از ورودی داشته باشند.

با در نظر گرفتن این ویژگی‌ها، دیکودر قادر خواهد بود تا هنگام تولید هر کلمه، روی معنای یک بخش از جمله تمرکز بیشتری داشته باشد و فقط از آن بخش برای تولید کلمه استفاده نماید. به این شکل، کلمات تولید شده شباهت بیشتری به ورودی خواهند داشت و ترجمه‌های بهتری حاصل خواهد شد.

در سال ۲۰۱۵، آقای بنجیو و همکارانش در پژوهش [۱۰] روشی ارائه دادند که در آن برای اولین بار از ایده استفاده از نقطه توجه در حوزه ترجمه ماشینی برای تولید شرح متناظر تصویر استفاده نمودند. در این پژوهش، از یک شبکه عصبی کانولوشنی به عنوان انکودر استفاده شده است. خروجی شبکه از لایه ماقبل آخر گرفته شده که منجر به ایجاد تعداد زیادی بردار ویژگی از تصویر می‌شود که هر کدام از این بردارهای ویژگی، از یک ناحیه از تصویر ایجاد شده‌اند و تمرکز بیشتری روی آن ناحیه داشته‌اند.

بدین ترتیب با استفاده از یک شبکه عصبی بازگشتی به عنوان دیکودر و استفاده از بردارهای حاشیه‌نویسی ایجاد شده توسط انکودر می‌توان به راحتی عملیات تولید شرح متناظر تصویر را انجام داد. تنها نکته‌ای که باید مشخص شود، چگونگی استفاده از بردارهای حاشیه‌نویسی است. در این پژوهش دو روش مختلف برای استفاده از بردارهای حاشیه‌نویسی مطرح شده است.

روش اول موسوم به روش توجه سخت، روشی است که در آن فقط یک بردار حاشیه‌نویسی انتخاب شده و از آن برای تولید جمله استفاده می‌شود. در این روش به هر یک از بردارهای حاشیه‌نویسی توسط یک مدل که قبلاً آموزش دیده است، یک وزن اختصاص می‌دهیم و سپس با توجه به وزن‌های تخصیص داده شده به هر بردار حاشیه‌نویسی،



شکل ۴۴: ساختار چارچوب کاری ارائه شده در [۳۵] در حوزه تولید شرح متناظر تصویر

یکی از آن‌ها را به عنوان بردار ویژگی تصویر انتخاب کرده و از آن در مراحل بعدی استفاده می‌کنیم. روش دوم موسوم به روش توجه نرم، روشی است که در آن یک بردار ویژگی کلی از روی بردارهای حاشیه‌نویسی تولید شده و از آن بردار در مراحل بعدی استفاده می‌شود. برای تولید این بردار نیز مانند روش توجه سخت، ابتدا توسط یک مدل که از پیش‌آموزش دیده است، به هر یک از بردارهای حاشیه‌نویسی یک وزن اختصاص می‌دهیم. سپس می‌توان با محاسبه امید ریاضی بردارهای حاشیه‌نویسی با توجه به وزن هر یک از آن‌ها بردار ویژگی نهایی را برای تصویر تولید و از آن برای تولید جمله استفاده کرد.

آزمایشات انجام شده روی این مدل نشان می‌دهد، معیار BLEU-1 حاصل از این روش با استفاده از توجه سخت معمولاً از مدل توجه نرم بیشتر بوده است. مطابق با نتایج گزارش شده در این پژوهش، میزان امتیاز BLEU-1 حاصل توسط توجه سخت روی مجموعه‌داده‌های Flickr30k، MS COCO و Flickr8k به ترتیب برابر با ۷۱.۸، ۶۶.۹ و ۶۷.۰ است. این در حالیست که امتیاز حاصل توسط توجه نرم روی همین مجموعه‌های داده، به ترتیب برابر با ۷۰.۸، ۶۶.۷ و ۶۷.۰ و امتیاز کسب شده توسط مدل Log Bilinear در بهترین حالت، به ترتیب برابر با ۷۰.۸، ۶۵.۶ و ۶۰.۰ بوده است.

مطابق با آزمایشات انجام شده، استفاده از توجه نرم، معیار METEOR را نسبت به استفاده از توجه سخت افزایش می‌دهد. طبق نتایج گزارش شده در پژوهش، امتیاز METEOR حاصل از توجه نرم به ترتیب روی مجموعه‌داده‌های MS COCO و Flickr30k، Flickr8k با ۱۸.۹۳، ۱۸.۴۹ و ۲۳.۹۰ بوده است. این در حالیست که امتیاز Log Bilinear کسب شده توجه سخت به ترتیب برابر با ۲۰.۳۰، ۱۸.۴۶ و ۲۳.۰۴ و امتیاز کسب شده توسط روش در بهترین حالت به ترتیب برابر است با ۱۷.۳۱، ۱۶.۸۸ و ۲۰.۰۳. این موضوع نشان می‌دهد با وجود این که جملات تولید شده توسط روش توجه سخت با در نظر گرفتن جملات موجود در مجموعه‌داده از امتیاز بالاتری نسبت به جملات تولید شده توسط توجه نرم برخوردارند؛ استفاده از توجه نرم، منجر به تولید جملات قابل قبول‌تری توسط انسان می‌شود.

پژوهش‌های مختلفی از این ایده در حوزه‌های مختلف استفاده نموده‌اند که گزارش مختصری از تعدادی از این پژوهش‌ها ارائه شده است.

۶ فصل ششم

حافظه فعال و مقایسه آن با مدل‌های
مبتنی بر توجه بصری

۱-۶ حافظه فعال

در بخش‌های قبل در رابطه با روش‌های مختلف تولید شرح متناظر تصویر صحبت کردیم. در این بخش قصد داریم، جدیدترین روش را که در حوزه ترجمه ماشینی مطرح شده و مورد استفاده قرار می‌گیرد بیان کرده و کاربرد آن را در حوزه تولید خودکار شرح متناظر تصویر مورد بررسی قرار دهیم. روش مورد بررسی، در حوزه تولید شرح متناظر تصویر به روش حافظه فعال^{۱۱۶} موسوم است.

همان‌طور که در بخش‌های قبلی ذکر شد، روش نقطه توجه با تمرکز بر روی یک بخش از تصویر و تولید کلمه مربوط به آن بخش سعی در تولید جمله می‌نماید. در این بخش برخلاف نقطه توجه، در هر مرحله با در نظر گرفتن کل تصویر اقدام به تولید کلمه می‌نماییم.

از آن‌جا که در این ایده، واحدهای بازگشتی گیت‌دار^{۱۱۷} پایه معماری شبکه را تشکیل می‌دهند، در این فصل ابتدا به بیان مختصر ساختار این واحدها پرداخته، سپس با استفاده از این واحدها، اقدام به ساخت نسخه کانولوشنی^{۱۱۸} آن خواهیم نمود. در نهایت با بررسی یک نمونه از پژوهش‌ها در حوزه تولید شرح متناظر تصویر، کارکرد ایده را در این حوزه مورد بررسی قرار داده و مقایسه‌ای از نحوه عملکرد این ایده در مقابل استفاده از روش‌های مبتنی بر نقطه توجه انجام خواهیم داد.

۲-۶ واحد بازگشتی گیت‌دار

ساختار واحدهای بازگشتی گیت‌دار شباht زیادی با ساختار شبکه حافظه کوتاه‌مدت بلند دارد. تنها تفاوت این واحدها در اندازه بردار ورودی و حالت است. در این واحدها، ابعاد ورودی و ابعاد بردار حالت با هم برابر است که منجر به افزایش تعمیم‌پذیری مدل می‌شود. رابطه این واحدها را می‌توان مطابق با روابط (۵۷) تا (۵۹) مدل‌سازی نمود. در این روابط، متغیرهای W , W' , W'' , U , U' و U'' ماتریس‌های وزن و بردارهای B , B' و B'' بردارهای بایاس هستند. تمام عبارات به شکل Wx بیان‌کننده حاصل ضرب ماتریس در بردار و عبارات به شکل $s \odot r$ بیان‌کننده حاصل ضرب درایه‌های نظیر به نظیر بردارها هستند. از آنجا که درایه‌های بردارهای r و u همگی در بازه $[0, 1]$ هستند، به این بردارها، گیت گفته می‌شود.

^{۱۱۶}Active Memory

^{۱۱۷}Gated Recurrent Unit (GRU)

^{۱۱۸}Convolutional Gated Recurrent Unit (CGRU)

$$GRU(x, s) = u \odot s + (1 - u) \odot \tanh(Wx + U(r \odot s) + B) \quad (57)$$

$$u = \sigma(W'x + U's + B') \quad (58)$$

$$r = \sigma(W''x + U''s + B'') \quad (59)$$

از واحدهای بازگشتی گیتدار به عنوان واحدهای یک شبکه بزرگ به این شکل استفاده می‌شود که ابتدا با اعمال ورودی و بردار حالت به واحد اول، خروجی و بردار حالت جدید محاسبه می‌شود، سپس با اعمال خروجی محاسبه شده و بردار حالت جدید به واحد در مرحله بعدی، خروجی و بردار حالت به روزرسانی می‌شوند. با توجه به نحوه استفاده از این واحد در شبکه، می‌توان به جای اعمال ورودی‌ها به طور جداگانه، تمام آن‌ها را با هم در یک ماتریس سه‌بعدی در حالت اولیه شبکه s قرار داد و محاسبات را انجام داد. با این روش، محاسبات مطابق با روابط (۶۰) تا (۶۲) تغییر می‌یابد که در آن‌ها، $s * U$ بیان‌کننده کانوالو کردن با انک فیلتر ^{۱۱۹} U با ماتریس s است. به واحد بازگشتی جدید که با استفاده از کانولوشن دسته فیلتر محاسبه می‌شود، واحد بازگشتی کانولوشنی گیتدار گفته می‌شود.

$$CGRU(x, s) = u \odot s + (1 - u) \odot \tanh(U * (r \odot s) + B) \quad (60)$$

$$u = \sigma(U' * s + B') \quad (61)$$

$$r = \sigma(U'' * s + B'') \quad (62)$$

یک دسته فیلتر، یک تنسور ۴ بعدی به ابعاد $[k_w, k_h, m, m]$ است که شامل $k_w \cdot k_h \cdot m^2$ پارامتر است و در آن k_w و k_h به ترتیب عرض و ارتفاع فیلتر هستند. این دسته فیلتر با ماتریس s به ابعاد $[w, m, h]$ کانوالو می‌شود که منجر به تولید یک ماتریس با ابعاد مشابه مطابق با رابطه (۶۳) می‌شود.

$$U * s_{[x,y,i]} = \sum_{u=-[k_w/2]}^{[k_w/2]} \sum_{v=-[k_h/2]}^{[k_h/2]} \sum_{c=1}^m s[x+u, y+v, c] \cdot U[u, v, c, i] \quad (63)$$

عمل کرد واحد CGRU کاملا مشابه عمل کرد لایه کانولوشن در یک شبکه کانولوشنی است. با توجه به نحوه عمل کرد این واحد، تشکیل ساختار یک شبکه l لایه‌ای به سادگی انجام می‌شود.

۳-۶ شبکه GPU

شبکه GPU که در این بخش مورد بررسی قرار می‌گیرد مطابق با ساختار ارائه شده در پژوهش [۱۲] است که به منظور یادگیری الگوریتم ضرب اعداد بزرگ مورد استفاده قرار گرفته است. با در نظر گرفتن ساختار و روابط مربوط به واحد CGRU، تعریف ساختار شبکه GPU به سهولت انجام می‌شود.

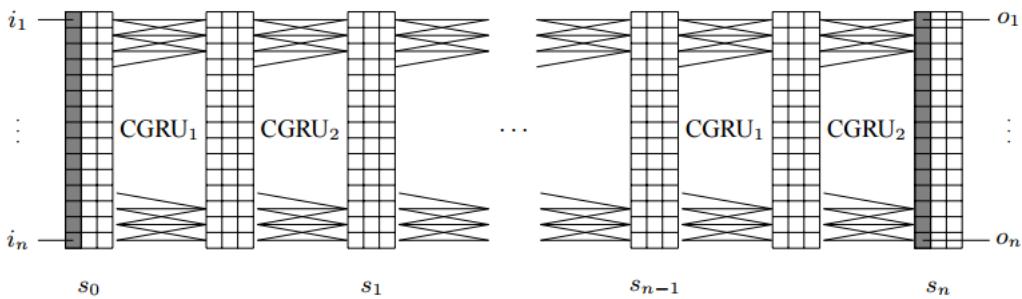
در مرحله اول، دنباله (i_1, \dots, i_n) به عنوان ورودی‌های شبکه، در ماتریس حالت اولیه s قرار داده می‌شوند.

^{۱۱۹}Kernel Bank

سپس با اعمال CGRU_l واحد مختلف به صورت یک دنباله به بردار حالت اولیه، خروجی نهایی را مطابق با رابطه (۶۷) محاسبه می‌شود.

$$s_{t+1} = \text{CGRU}_l(CGRU_{l-1} \cdots CGPU_1(s_t) \cdots) \text{ and } s_{fin} = s_n \quad (64)$$

ساختار گسترده شبکه عصبی مدل شده توسط رابطه (۶۴) را که دارای ۲ لایه و $w = 3$ است را می‌توان در شکل ۴۵ مشاهده نمود.



شکل ۴۵: ساختار گسترده شبکه عصبی GPU مطرح شده در پژوهش [۱۲]

برای تولید خروجی نهایی شبکه، ابتدا تمام آیتمها را در ستون اول s_{fin} ضرب نموده و برای بدست آوردن مقدار لاجستیک^{۱۲۰} آن، در ماتریس O ضرب می‌نماییم. با فرض $[O_{s_{fin}}]_{\circ, k, :} = l_k$ (سینتکس زبان پایتون)، مولفه با بیشترین مقدار $o_k = argmax(l_k)$ را به عنوان خروجی شبکه معروفی می‌نماییم. به همین منظور در طول فرایند آموزش، از یک لایه soft max در لایه آخر به عنوان خروجی استفاده می‌نماییم و از منفی لگاریتم احتمال به عنوان تابع خطا استفاده می‌نماییم.

۴-۶ استفاده از حافظه فعال در ترجمه ماشینی

مطابق با پژوهش [۱۳] که آقای کایزر و همکارانش در سال ۲۰۱۷ انجام دادند، استفاده از شبکه GPU استاندارد، به شکلی که در این گزارش تا اینجا ذکر شد، در حوزه ترجمه ماشینی منجر به نتایج قابل توجهی نمی‌شود. مطابق این پژوهش، معیار سرگشتگی^{۱۲۱} نسخه استاندارد این شبکه در حوزه ترجمه ماشینی نمی‌تواند به کمتر از ۳۰ برسد. این در حالی است که مدل‌های دیگر در این حوزه به سرگشتگی کمتر از ۴ دست یافته‌اند. از طرف دیگر معیار BLEU این مدل به سختی به حدود ۵ می‌رسد در حالی که مدل‌های دیگر می‌توانند به مقدار بالاتر از ۲۰ در این معیار برسند. سوال اساسی این است که کدام قسمت از مدل باعث ایجاد این میزان کاهش دقت در نتایج می‌شود؟

۴-۶-۱ نسخه مارکفی شبکه GPU

عملکرد نامناسب این مدل در حوزه ترجمه ماشینی در مرحله اول مربوط به استقلال مولفه‌های خروجی نسبت به یکدیگر است. همان‌طور که در شکل ۴۵ قابل مشاهده است، تمام o_i ها نسبت به یکدیگر به شرط s_{fin} مستقل

^{۱۲۰}Logistic

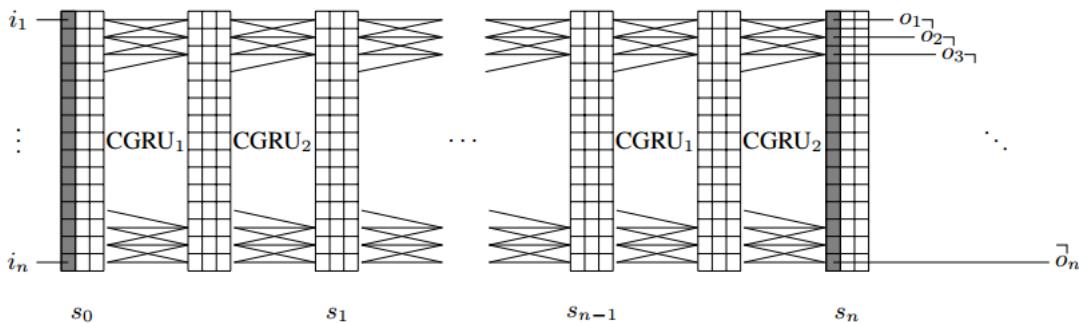
^{۱۲۱}Perplexity

هستند؛ در حالی که در کاربرد ترجمه ماشینی و تولید جمله، لازم است کلمات جمله به یکدیگر وابسته باشند. نکته قابل توجه این است که برای ایجاد این وابستگی فقط کافیست نحوه تولید خروجی را تغییر دهیم و نیازی به ایجاد تغییرات اساسی در مدل نیست.

به منظور ایجاد این وابستگی در لایه خروجی می‌توان رابطه خروجی را مطابق با رابطه (۶۵) تغییر داد.

$$l_k = O_{concat}(s_{fin}[\circ, k, :], E' o_{k-1}) \quad (65)$$

همان‌طور که در رابطه (۶۵) مشخص است، تنها تغییر ایجاد شده در نحوه تولید خروجی این است که ابتدا هر مولفه از ستون اول s_{fin} را با بردار شامل تاثیر مولفه مرحله قبلی الحق^{۱۲۲} کرده و سپس در ماتریس O ضرب می‌نماییم. ماتریس E در این رابطه یک ماتریس جاسازی است. سپس خروجی نهایی را به شکل $o_k = argmax(l_k)$ محاسبه می‌نماییم. به این نسخه از شبکه، شبکه مارکفی GPU^{۱۲۳} گفته می‌شود. شکل ۴۶ نمایش‌دهنده ساختار نسخه مارکفی این شبکه است.



شکل ۴۶: ساختار نسخه مارکفی شبکه GPU ارائه شده در پژوهش [۱۲]

۲-۴-۶ نسخه توسعه یافته شبکه GPU

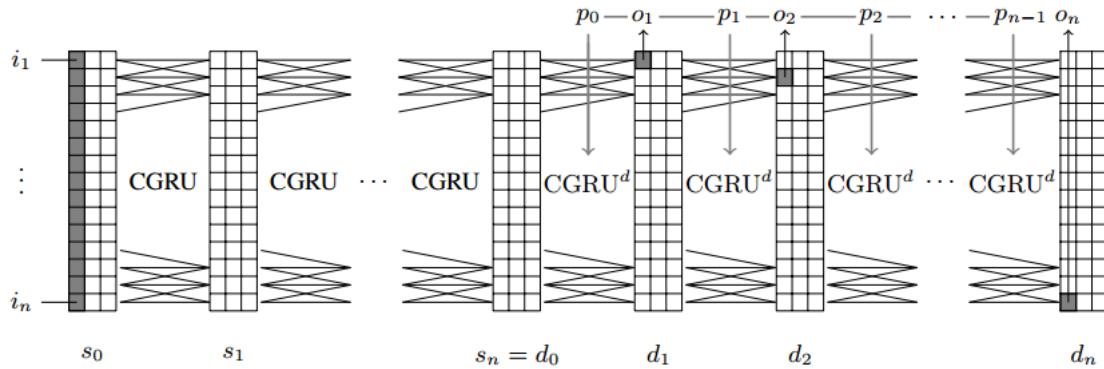
با وجود این که نسخه مارکفی این شبکه بهبود قابل توجهی در معیار سرگشتگی مدل ایجاد می‌کند، اما نتایج حاصل شده از آن هنوز قابل قبول نیست. این نسخه به سرگشتگی حدود ۱۲ و امتیاز BLEU مشابه نسخه استاندارد دست می‌یابد که با وجود بهبود ایجاد شده نسبت به نسخه استاندارد، کماکان قابل مقایسه با مدل‌های موجود دیگر در حوزه ترجمه ماشینی نیست.

علت وقوع این ناکارامدی در مدل مارکفی ارائه شده این است که وابستگی‌های مارکفی که در مدل بین مولفه‌های خروجی در نظر گرفته شده است، به اندازه کافی شدت ارتباط واقعی موجود بین این مولفه‌ها را مدل نمی‌کند. برای افزایش میزان وابستگی این مولفه‌ها، نسخه توسعه یافته شبکه ارائه شده است که در آن به جای یک لایه خروجی، از یک دیکوڈر حافظه فعال^{۱۲۴} استفاده شده است. شکل ۴۷ ساختار نسخه توسعه یافته شبکه GPU را نمایش می‌دهد.

^{۱۲۲}Concatenation

^{۱۲۳}Moarkovian Neural GPU

^{۱۲۴}Active Memory Decoder



شکل ۴۷: ساختار شبکه توسعه یافته GPU ارائه شده در [۱۳]

همان طور که در شکل پیداست، دیکودر حافظه فعال درست از لایه s_n شروع می‌شود. لایه d_n را در نظر می‌گیریم. در دیکودر حافظه فعال، از یک نوار تنسور^{۱۲۵} اضافه برای خروجی استفاده می‌شود که آن را با p نمایش می‌دهیم. ساختار نوار تنسور خروجی کاملاً مشابه ساختار بردار d است. در بخش دیکودر، فرایند محاسبه خروجی از p شروع می‌شود. ابتدا $p = \text{CGRU}_l^d(\text{CGRU}_{l-1}^d(\dots \text{CGRU}_1^d(d_t, p_t) \dots, p_t), p_t)$ می‌شود.

$$d_{t+1} = \text{CGRU}_l^d(\text{CGRU}_{l-1}^d(\dots \text{CGRU}_1^d(d_t, p_t) \dots, p_t), p_t) \quad (66)$$

شایان توجه است که رابطه (۶۶) همان رابطه (۶۴) است که یک ورودی p_t به واحدهای CGRU استفاده شده در آن اضافه شده است. واحدهای CGRU^d مورد استفاده در این رابطه را می‌توانیم به شکل رابطه (۶۷) مدل‌سازی نمود.

$$\begin{aligned} \text{CGRU}^d(s, p) &= u \odot s + (1 - u) \odot \tanh(U * (r \odot s) + W * p + B) \int_1^{100} x dx \\ u &= \sigma(U' * s + W' * p + B') \\ r &= \sigma(U'' * s + W'' * p + B'') \end{aligned} \quad (67)$$

برای محاسبه خروجی k ام در دیکودر، بردار k ام در اولین ستون d_k را در ماتریس O به شکل $[:, k]$ ضرب می‌نماییم و سپس مولفه با بیشترین مقدار را به شکل $o_k = \text{argmax}(l_k)$ به عنوان خروجی، معین می‌کنیم. خروجی o_k دوباره با یک جاسازی دیگر به صورت یک بازنمایی متراکم به p به شکل رابطه (۶۸) اضافه می‌شود.

$$p_{k+1} = p_k, p_k[:, k, :] \leftarrow E' o_k \quad (68)$$

با این کار، تاثیر تمام خروجی‌ها، به صورت تجمعی، روی نوار تنسور اضافه شده به مدل، تجمیع می‌شود و در هر مرحله برای تولید خروجی، تمام مولفه‌های قبلی تاثیرگذار می‌شوند.

^{۱۲۵}Tape Tensor

۵-۶ بررسی عملکرد ایده حافظه فعال در حوزه ترجمه ماشینی

در این قسمت نتایج مربوط به آزمایشات انچام شده در پژوهش [۱۳] را مورد بررسی قرار می‌دهیم. از آنجا که تمام روابط ارائه شده تا اینجا مشتق‌پذیر هستند، می‌توان از هر الگوریتمی که مبتنی بر نزول تصادفی در امتداد گرادیان^{۱۲۶} باشد، به منظور آموزش شبکه استفاده نمود. در تمام آزمایشات انچام شده در این پژوهش از روش بهینه‌سازی آدام^{۱۲۷} که در پژوهش [۳۶] ارائه شده، بهره گرفته شده است. تمام شبکه‌های مورد استفاده در این پژوهش که مبتنی بر ایده حافظه فعال طراحی شده‌اند دارای ۲ لایه هستند. پارامترهای $w = 512$ و $m = 32$ در تمام آزمایشات مقدار ثابت دارند.

مجموعه داده مورد استفاده در این پژوهش، همان مجموعه‌داده مورد استفاده در پژوهش [۹] است که توسط آقای بنجیو در سال ۲۰۱۴ ارائه شده و برای اولین بار، ایده استفاده از نقطه توجه در حوزه ترجمه ماشینی را مطرح نموده است. این مجموعه‌داده که به WMT موسوم است، برای ترجمه جملات از زبان انگلیسی به فرانسوی جمع‌آوری شده است و دارای ۳۲۰۰۰ کلمه در دیکشنری خود است.

مدل مبتنی بر نقطه توجه استفاده شده به عنوان معیار مقایسه در این آزمایش، مدلی است که توسط آقای هینتون و همکارانش در سال ۲۰۱۵ در پژوهش [۳۷] ارائه شده است. تنها تغییری که در این مدل ایجاد شده است، این است که به جای واحدهای LSTM، از واحدهای GRU استفاده شده است که باعث کاهش تعداد پارامترهای مدل از ۱۲۰ میلیون به حدود ۱۱۰ میلیون می‌شود. این تغییر به این دلیل ایجاد شده است که مقایسه ایده حافظه فعال با ایده نقطه توجه را قابل توجیه نماید.

در مدل توسعه‌یافته ارائه شده در این بخش، یکی از مهم‌ترین پارامترهایی که قبل از انچام آزمایش باید مقدار بهینه آن به خوبی تعیین شود، اندازه خروجی مورد انتظار شبکه است که مستقیماً مشخص‌کننده تعداد لایه‌های شبکه می‌شود. برای انتخاب مقدار بهینه این پارامتر از یک جستجوی حریصانه، استفاده شده است. جستجوی مذکور به این شکل صورت گرفته است که مقادیر مختلف برای این پارامتر قرار داده شده است و مقداری که کمترین میزان سرگشتنگی را نتیجه داده به عنوان مقدار بهینه انتخاب شده است.

جدول ۷ نمایش‌دهنده نتایج به دست‌آمده از ۴ مدل مختلف روی مجموعه‌داده مذکور است.

جدول ۷: جدول نتایج به دست‌آمده از ۴ مدل مختلف روی مجموعه‌داده ترجمه انگلیسی به فرانسوی [۱۳]

نام مدل	سرگشتنگی	معیار BLEU
شبکه GPU	۳۰.۱	< ۵
نسخه مارکفی	۱۱.۸	< ۵
نسخه توسعه‌یافته	۳.۳	۲۹.۶
نقطه توجه با واحدهای GRU	۳.۴	۲۶.۴

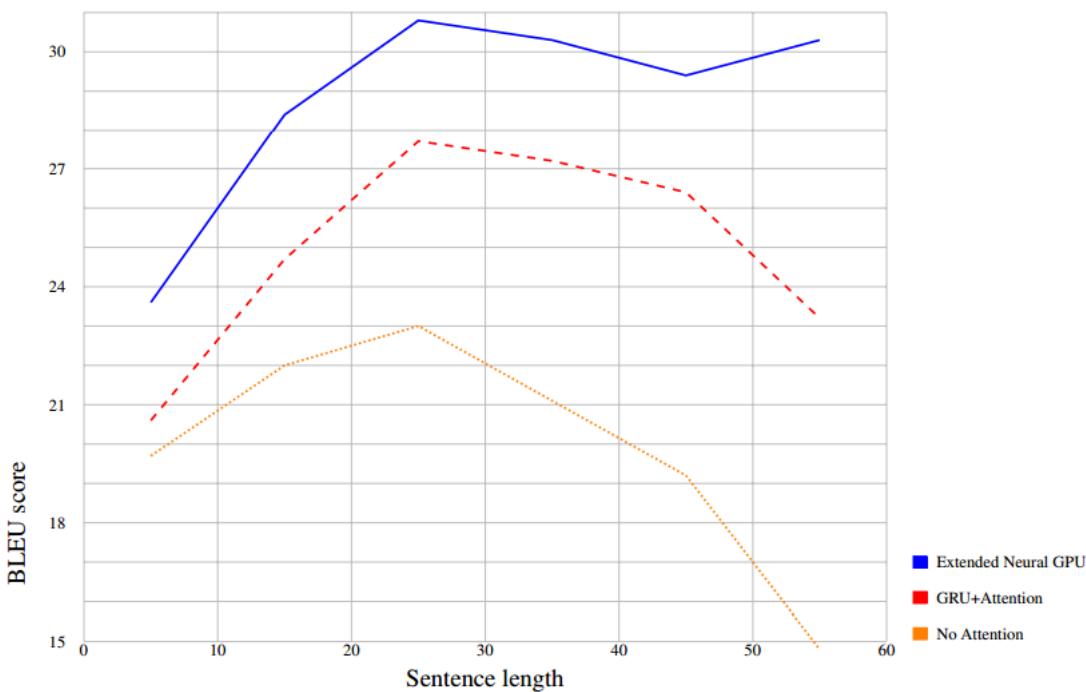
نتایج ارائه شده در جدول ۷ نشان می‌دهد که مدل حافظه فعال می‌تواند در حوزه ترجمه ماشینی به مدل‌های مبتنی بر توجه بسیار نزدیک شود و حتی بتواند از با تعداد پارامترهای کمتر، نتایج بهتری را نسبت به آن‌ها از خود

^{۱۲۶}Stochastic Gradient Descent

^{۱۲۷}Adam Optimizer

نشان دهد. از طرفی نتایج پژوهش‌های دیگر نشان می‌دهد روش‌های مبتنی بر نقطه توجه، با داشتن نمونه‌های کوتاه‌تر می‌توانند تعمیم‌پذیری بیشتری نسبت به مدل‌های حافظه فعال از خود نشان دهند.

به علاوه، با نگاه دقیق‌تر به فرایند مدل‌های مبتنی بر توجه در می‌یابیم که یکی از مشکلاتی که این نوع از مدل‌ها به خوبی آن‌ها را حل می‌کنند، متفاوت بودن طول جملات تولید شده است. به همین منظور آزمایش دیگری انجام شد تا عمل کرد مدل‌های مبتنی بر حافظه فعال و نقطه توجه را نسبت به طول جملات با یکدیگر مورد بررسی قرار دهد. شکل ۴۸ نشان‌دهنده عمل کرد مدل‌های مختلف روی جملات با طول‌های متفاوت است. در این نمودار، محور افقی مشخص کننده طول جملات و محور عمودی میزان معیار BLEU است. مطابق با نتایج نمایش شده در این شکل، مدل حافظه فعال حساسیت کمتری در برابر طول جملات نسبت به مدل‌های دیگر از خود نشان می‌دهد.



شکل ۴۸: مقایسه عمل کرد مدل‌های مختلف نسبت به طول جملات [۱۳]

شایان ذکر است، نکته کلیدی مورد استفاده در ایده حافظه فعال که منجر به برتری عمل کرد نسخه توسعه یافته آن نسبت به مدل‌های مبتنی بر نقطه توجه شده است، در نظر گرفتن وابستگی بیشتر بین مولفه‌های خروجی است. در اکثر مدل‌های قبلی که ارائه شد، این وابستگی بین مولفه‌های خروجی در سطح ساختار واحدهای GRU یا LSTM دیده شده است. این در حالیست که در ایده حافظه فعال، از این سطح عبور کرده و آن را در سطح معماری کل شبکه مورد بررسی قرار داده‌ایم.

یکی دیگر از نکات بسیار مهم در رابطه با مقایسه مدل‌های مبتنی بر نقطه توجه و مدل‌های حافظه فعال این است که با در نظر گرفتن این که نسخه توسعه یافته ایده حافظه فعال می‌تواند با تعداد پارامترهای کمتر به عمل کرد مدل‌های مبتنی بر نقطه توجه برسد و حتی گوی سبقت را از آن‌ها برباید، آیا می‌توان در همه حوزه‌های دیگر، حافظه فعال را به طور کامل جایگزین نقطه توجه کرد؟ پاسخ این است که حافظه فعال همواره می‌تواند جایگزین توجه نرم شود زیرا بار محاسباتی توجه نرم به مراتب بیشتر از حافظه فعال است و جایگزینی آن تقریباً در همه

موارد می‌تواند مفید فایده باشد.

با این وجود در رابطه با مدل‌های مبتنی بر نقطه توجه با توجه سخت، نمی‌توان به راحتی در رابطه با جایگزینی آن‌ها با حافظه فعال نظر داد. زیرا ممکن است در حوزه‌هایی غیر از ترجمه ماشین، تمرکز بر روی یک بخش از ورودی در مراحل مختلف، سودمند باشد. با تمام این تفاسیر، ایده‌های حافظه فعال و نقطه توجه، در تضاد با یکدیگر نیستند و می‌توانند در مدل‌هایی به صورت ترکیبی با یکدیگر مورد استفاده قرار بگیرند.

۶-۶ جمع‌بندی

در این بخش، یکی از جدیدترین روش‌هایی را که در حوزه ترجمه ماشینی مورد استفاده قرار می‌گیرد و در پژوهش [۱۳] که در سال ۲۰۱۷ ارائه شده است، عمل کرد بهتری نسبت به روش‌های مبتنی بر نقطه توجه از خود نشان داده است را، که به نام حافظه فعال شناخته می‌شود، مورد بررسی قرار دادیم. در ابتدا واحد بازگشتی گیت‌دار و واحد بازگشتی گیت‌دار کانولوشنی معرفی شدند که روابط و ساختاری مشابه شبکه حافظه کوتاه‌مدت بلند دارند. سپس با استفاده از این واحدها، اقدام به ساخت ساختار شبکه GPU نمودیم.

مطابق با پژوهش [۱۳]، ساختار شبکه GPU به همان شکل که ارائه شد، توان رقابت با مدل‌های مشابه دیگر را در حوزه تولید شرح متناظر تصویر، ندارد. به همین دلیل نسخه‌های مارکفی و توسعه‌یافته این شبکه را ارائه نمودیم که عمل کرده‌ای بهتری از خود نشان دادند. این شبکه در حوزه یادگیری الگوریتم، به ویژه در حوزه یادگیری الگوریتم‌های ساده مانند ضرب و جمع اعداد بسیار بزرگ، عمل کرد بسیار خوبی از خود نشان داده است.

ایده حافظه فعال بر خلاف ایده روش‌های مبتنی بر توجه، بر این است که در هر مرحله از تولید خروجی، از تمام حافظه موجود استفاده نماییم. نکته‌ای که باعث کاهش چشم‌گیر عمل کرد این مدل در حوزه ترجمه ماشینی می‌شود، عدم وجود وابستگی کافی بین مولفه‌های خروجی شبکه است. با اعمال وابستگی‌های مارکفی، نسخه مارکفی این شبکه قابل حصول است که علاوه بر بهبود عمل کرد نسخه استاندارد، توانایی رقابت با مدل‌های مبتنی بر توجه را ندارد. در نسخه توسعه‌یافته این شبکه، وابستگی بین مولفه‌های خروجی، شدیدتر از وابستگی‌های مارکفی در نظر گرفته شده است که باعث افزایش چشم‌گیر کارایی مدل و رقابت‌پذیری مدل با نسخه‌های مبتنی بر توجه شده است.

مطابق با نتایج گزارش شده در پژوهش [۱۳]، نسخه توسعه‌یافته شبکه قادر به دست‌یابی به امتیاز BLEU برابر با ۲۹.۶ روی مجموعه‌داده جملات معادل زبان انگلیسی و فرانسوی شده است. این در حالیست که مدل مبتنی بر نقطه توجه امتیاز BLEU برابر با ۲۶.۴ را کسب کرده است. این بهبود عمل کرد از در نظر گرفتن وابستگی بیشتر بین مولفه‌های خروجی حاصل شده است.

مدل حافظه فعال می‌تواند با تعداد پارامترهای به مراتب کمتر نسبت به مدل‌های مبتنی بر نقطه توجه، عمل ترجمه ماشینی را انجام دهد. با این وجود هنوز مدل‌های مبتنی بر نقطه توجه که از توجه سخت استفاده می‌کنند عمل کرده‌ای بهتری نسبت به مدل حافظه فعال از خود نشان می‌دهند.

۷ فصل هفتم

جمع‌بندی و نتیجه‌گیری

در این مستند، گزارش مختصری درباره مساله تولید خودکار شرح بر تصاویر و روش‌های پیشنهادی برای حل چالش‌های موجود در این مسیر را ارائه دادیم. مساله تولید خودکار شرح بر تصاویر، به معنای تولید جملات زبان طبیعی برای هر تصویر است به طوری که این جملات شامل سه شرط زیر باشند:

۱. صحنه، اجسام موجود در تصویر، رابطه مکانی اجسام و اطلاعاتی از این دست که به درک تصویر کمک می‌کنند باید در جملات تولید شده وجود داشته باشند و دقیق و کامل باشند.
۲. جملات تولید شده، خود، باید به لحاظ معنایی، دستور زبان و املایی صحیح بوده و نقصی نداشته باشند.
۳. جملات تولید شده باید با تصاویر مرتبط با خود، سازگاری داشته باشند.

ایده‌های اولیه در این مسیر از پژوهش‌های موجود در زمینه ترجمه ماشین ایجاد شده است که در آن‌ها، ابتدا یک جمله ورودی از یک زبان مبدا، با استفاده از روش‌های مختلفی به یک بردار ویژگی تبدیل می‌شود و سپس در مرحله دوم، بردار ویژگی حاصل، با استفاده از روش‌های خاص دیگری به جملات زبان طبیعی به زبان مقصد، تبدیل می‌شوند. حال اگر به جای جمله از زبان مبدا، یک تصویر به این سامانه وارد شود و با روشی بتوان این تصویر را به همان بردار ویژگی نگاشت کرد، جمله نهایی معادل معنایی تصویر ورودی خواهد شد. با استفاده از این فرایند می‌توان به طور خودکار برای تصاویر، شرح مناسبی به زبان طبیعی تولید کرد.

در این مسیر، دو چالش عمده در پیش رو وجود دارد که باید مرتفع شوند:

۱. چالش درک صحنه
فرایند استخراج اطلاعات بصری نهفته در تصویر و بازنمایی مناسب این اطلاعات را به گونه‌ای که بتواند برای پردازش‌های بعدی مناسب باشد، فرایند درک صحنه می‌نامند. روش‌های مختلف و متعددی برای حل این چالش، تاکنون مطرح شده‌اند. در این مساله، باید بتوان تصاویر ورودی را به نحوی موثر و مفید به فضای ویژگی‌ها نگاشت کرد به طوری که بازنمایی حاصل، بتواند در مرحله تولید جمله، منجر به تولید جملات معنادار و مناسب شود.

۲. چالش تولید جمله
تولید جملات به زبان طبیعی که علاوه بر صحت معنایی، دستور زبانی و املایی، قادر به توصیف و تفسیر اطلاعات غیر قابل تفسیر برای کاربران انسانی هستند، از جمله مهم‌ترین و پویاترین حوزه‌های پژوهشی در زمینه هوش مصنوعی است و توجه پژوهش‌گران بسیاری را به خود جلب کرده است. در این مساله، باید

بتوان بردار ویژگی حاصل از تصویر را که در مرحله درک صحنه تولید شده است، به نحوی کارا و موثر به یک جمله در زبان طبیعی نگاشت کرد.

۲-۷ درک صحنه

اولین مرحله از فرایند تولید خودکار شرح برای تصاویر، مرحله درک صحنه است. در این مرحله، تصاویر ورودی تحت عملیات مختلفی به فضای معنایی نگاشت می‌شوند. فضای معنایی در اینجا، می‌تواند فضای شامل میدان‌های اطلاعاتی از پیش تعیین شده (مانند فضای سه‌تایی‌های «جسم، رخداد، صحنه») یا فضای بردار ویژگی‌ها باشد. روش‌های مختلفی برای نگاشت تصویر ورودی به فضای معنایی ارائه شده است که به طور کلی می‌توان عموم آن‌ها را به دو بخش تقسیم کرد:

۱. روش‌های مبتنی بر مدل‌های گرافی احتمالاتی

در این روش‌ها با استفاده از مدل‌های استاندارد گرافی احتمالاتی موجود یا با ارائه یک مدل گرافی احتمالاتی، تصویر ورودی به فضای معنایی نگاشت می‌شود. در روش‌های مبتنی بر این مدل‌ها، با ارائه یک توزیع احتمال برای نقاط مختلف در فضای معنایی، محتمل‌ترین نقطه برای تصویر به عنوان نقطه نظر تصویر، انتخاب می‌شود.

(آ) مدل میدان تصادفی مارکف

یک نمونه از روش‌های مبتنی بر مدل میدان تصادفی مارکف که برای درک صحنه از آن استفاده شده است، در پژوهش [۱۸] ارائه شده است. درک صحنه در این پژوهش با ارائه یک سه‌تایی «جسم، فعالیت، صحنه» به‌ازای هر تصویر، تعریف شده است. مبتنی بر همین تعریف، یک مدل میدان تصادفی مارکف شامل سه گره که دوبه‌دو به هم متصل هستند، تعریف شده است. هر یک از گره‌های موجود در این مدل، نماینده یکی از میدان‌های سه‌گانه تعریف شده در فضای معنایی هستند. با تعریف توابع پتانسیل مختلف روی هر گره و توابع پتانسیل مختلف روی هر یال، یکتابع توزیع توام برای تمام متغیرهای تصادفی موجود در مدل ارائه شده است.

با محاسبه مقادیر پتانسیل برای تصاویر مختلف موجود در مجموعه آموزشی و با استفاده از یک ماشین بردار پشتیبان، بردارهای ویژگی شاخص برای هر گره محاسبه می‌شوند. از این بردارهای ویژگی بعده برای انطباق تصاویر با مقادیر مختلف در هر گره استفاده می‌شود.

در این پژوهش، با یافتن نزدیک‌ترین همسایه‌های یک تصویر بر حسب معیار شباهت با بردارهای ویژگی شاخص و میانگین‌گیری روی مقادیر هر گره، بهترین انطباق تصویر و نقاط فضای معنایی به دست می‌آید. به این ترتیب، برای هر تصویر ورودی، می‌توان نقطه نظر در فضای معنایی را مشخص کرد.

(ب) مدل میدان تصادفی شرطی

در پژوهش [۱۷] یک مدل میدان تصادفی شرطی سلسله‌مراتبی برای درک صحنه ارائه شده است که

شامل دو سطح انتزاع است. برای گرههای موجود در هریک از سطوح انتزاع مدل، یک دسته متغیر تصادفی تعریف شده و برای کل مدل سه نوع تابع پتانسیل مختلف معرفی شده است.

اولین دسته از توابع پتانسیل معرفی شده در این بخش، توابع پتانسیل قطعه‌بندی یگانی هستند که به منظور یکپارچه‌سازی نقاط داخل یک قطعه تعریف شده‌اند. توابع پتانسیل دیگری برای انطباق بین متغیرهای تصادفی موجود در بین دو سطح انتزاع تعریف شده‌اند که در صورت مغایرت مقادیر اختصاص داده شده به متغیرهای موجود بین دو سطح، مقدار λ – و در غیر این صورت مقدار صفر دارند. این توابع در شرایطی که مقادیر متغیرهای موجود در دو سطح با هم یکسان نباشد، یک مقدار جریمه به تابع هدف اضافه می‌کنند. آخرین دسته از توابع پتانسیل مورد استفاده، برای انطباق تصویر با دسته تشخیص داده شده اجسام تعریف شده است که توسط فلزنسوالب ارائه شده و به روش دی پی ام مشهور است.

(ج) سایر مدل‌های گرافی احتمالی در پژوهش [۲۲]، یک مدل گرافی احتمالی مولد برای نگاشت تصویر به فضای معنایی ارائه شده است. در این مدل، از دو سطح تصویر استفاده شده است؛ تصویر سطح جسم و تصویر سطح صحنه. برای تصویر سطح صحنه، یک متغیر تصادفی، بیان‌کننده دسته صحنه و برای تصویر سطح جسم دو متغیر تصادفی، بیان‌کننده دسته و شکل جسم، ارائه شده است. روابط بین متغیرهای تصادفی در این پژوهش، براساس نحوه تولید متغیرهای تصادفی و روابط منطقی موجود بین آن‌ها طراحی شده‌اند.

تصویر ورودی در این پژوهش، ابتدا به نواحی کوچک $10 * 10$ تقسیم می‌شود و مطابق با روش توضیح داده شده، مقدار توابع پتانسیل مختلف برای هر کدام از متغیرهای تصادفی، در هر ناحیه، محاسبه می‌شود. در این پژوهش، یک تابع احتمال شرطی برای متغیرهای تصادفی ارائه شده است که در مرحله استنتاج، با استفاده از روش تخمین بیشترین احتمال، برچسب‌های هر تصویر مشخص می‌شوند.

۲. روش‌های مبتنی بر استفاده از شبکه‌های عصبی کانولوشنی عمیق
در این روش‌ها، با ارائه یک شبکه عصبی کانولوشنی عمیق و تعریف کردن تابع هدف برای شبکه، تابع نگاشت تصویر و فضای معنا تشکیل می‌شود. پس از ارائه تابع هدف برای هر شبکه، با بهینه‌سازی آن تابع، پارامترهای موجود در شبکه آموزش داده می‌شوند.

در پژوهش [۲۵]، روشی ارائه شده است که طی آن یک تصویر، به نواحی کوچک‌تر تقسیم می‌شود به طوری که هر ناحیه به وجود آمده، به طور یکپارچه، حاوی یک جسم باشد و هر جسم تنها در یک ناحیه قرار بگیرد. این روش موسوم به روش RCNN است. در این روش، دو ویژگی برای یک ناحیه‌بندی خوب در تصاویر ارائه شده است و پیرو این ویژگی‌ها، روشی برای طرح نواحی پیشنهادی در یک تصویر که دارای این دو ویژگی باشد، ارائه شده است.

ویژگی مطرح شده اول برای ناحیه‌بندی تصاویر این است که، ناحیه‌های ایجاد شده در هر تصویر، می‌توانند در ابعاد مختلف وجود داشته باشند زیرا اجسام موجود در تصاویر، ممکن است اندازه و تعداد متفاوتی داشته باشند. دومین ویژگی برای یک ناحیه‌بندی خوب، این است که معیار انتخاب نواحی نباید برای تمام تصاویر، یکسان در نظر گرفته شود؛ زیرا معیارهای مختلف برای ناحیه‌بندی تصاویر در شرایط مختلف، رفتارهای

متفاوتی از خود نشان می‌دهند. بنابراین باید از معیارهای مختلف برای تعیین نواحی استفاده نمود. در این پژوهش، ابتدا تصاویر مطابق با یک معیار اولیه، به مجموعه‌ای از نواحی اولیه تقسیم می‌شوند. سپس با استفاده از معیارهای مختلف مانند فضاهای رنگی مختلف، معیارهای شباهت مختلف و نقاط اولیه متفاوت، با پیروی از یک روش حریصانه، نواحی کوچکتر که به یکدیگر شبیه‌تر هستند با هم ترکیب شده و نواحی بزرگتر را می‌سازند. نواحی ایجاد شده در این روش، سپس به یک شبکه عصبی کانولوشنی عمیق داده می‌شوند و برای هر ناحیه، یک بردار ویژگی ۴۰۹۶ بعدی ایجاد می‌شود که هر ناحیه با آن بازنمایی شود. در پژوهش [۲۸] با استفاده از روش RCNN و تعریف دوتابع هدف دیگر برای شبکه، روشی ارائه شده است که طی آن بتوان تصاویر و جملات را به طور دوطرفه به یکدیگر نگاشت کرد. توابع هدف تعریف شده در این پژوهش، دوتابع مختلف هستند. اولین تابع هدف، یک تابع هدف سراسری است. این تابع به این منظور تعریف شده است که تصاویر و جملاتی که مطابق با محاسبات شبکه عصبی ارائه شده، بیشترین شباهت را با یکدیگر دارند، در واقعیت هم شبیه‌ترین تصاویر و جملات به یکدیگر باشند. تابع هدف دوم برای این شبکه به این شکل تعریف شده است که نواحی استخراج شده از تصویر و عبارات استخراج شده از جملات که در روش ارائه شده، بیشترین شباهت را به یکدیگر دارند، در واقعیت هم بیشترین شباهت و ارتباط را با یکدیگر داشته باشند.

در این پژوهش، تصاویر ورودی با استفاده از روش RCNN به نواحی مختلف تقسیم شده و ۱۹ ناحیه با بیشترین اطمینان از بین این نواحی انتخاب می‌شود. این ۱۹ ناحیه به همراه خود تصویر به عنوان ۲۰ تصویر مختلف مورد استفاده قرار می‌گیرند. جملات ورودی با استفاده از روشی که در فصل تولید جملات زبان طبیعی توضیح داده خواهد شد، به عبارات مختلف تقسیم می‌شوند و بین هر عبارت استخراج شده و هر یک از ۲۰ تصویر موجود، یک معیار شباهت محاسبه شده و بیشترین شباهتها با هم درنظر گرفته می‌شوند. معیار شباهت مورد استفاده در این روش، ضرب داخلی بین بردارهای ویژگی عبارات و نواحی است. عبارات و نواحی که بیشترین شباهت را با یکدیگر دارند برای تولید جمله به مرحله بعد، ارسال می‌شوند.

۳-۷ تولید جمله

چالش تولید جمله یکی از قدیمی‌ترین و پویاترین حوزه‌های فعالیتی و پژوهشی در هوش مصنوعی است که از اواسط قدن بیستم، توجه پژوهش‌گران بسیاری را به خود جلب کرده است. روش‌های مختلفی برای حل این مساله ارائه شده‌اند. از جمله این روش‌ها می‌توان به موارد زیر اشاره کرد:

۱. تولید زبان طبیعی

در این دسته از روش‌ها که از اواخر دهه بیستم تا کنون مورد استفاده قرار می‌گیرند، با طی فرایند در یک چارچوب کلی، سعی در تولید جملات مناسب دارند. این دسته از روش‌ها عموماً برای تفسیر خودکار داده‌هایی که برای کاربران انسانی غیر قابل تفسیر هستند یا تفسیر دشواری دارند، به کار می‌روند. در این روش‌ها ابتدا با استفاده از ویژگی‌های مختلفی که در داده‌های ماشینی (داده‌های قابل تفسیر برای ماشین) کلمات مناسب انتخاب شده و سپس با استفاده از کلمات منتخب، عبارات زبانی (با جایگشت دادن کلمات و حذف عبارات غیر محتمل) تولید می‌شوند. سپس با اعمال قواعد دستور زبان و چینش عبارات زبانی در

کارهم، جملات نهایی تولید می‌شوند.

۲. نزدیکترین همسایه

در این دسته از روش‌ها سعی می‌شود با ورود یک تصویر و نگاشت آن به فضای ویژگی‌ها، جمله‌ای از میان تمام جملات موجود در مجموعه‌داده انتخاب شود که بیشترین مشابهت با بردار ویژگی تصویر را دارد. بزرگترین مشکل در این روش‌ها انتخاب معیار مناسب برای محاسبه فاصله بین یک جمله و بردار ویژگی حاصل از تصویر است. در این روش، علاوه بر این که نیاز به وجود مجموعه‌داده وسیع و پوشای وجود دارد، ممکن است جمله نهایی، در انتهای گویا و بیان‌کننده تمام جوانب تصویر نباشد و یا حتی با تصویر ورودی سازگاری نداشته باشد.

برای حل این مشکل، سعی شد به جای استخراج نزدیکترین جمله به تصویر موجود، مشابه‌ترین عبارات زبانی را با شکستن جملات موجود به عبارات سازنده، انتخاب کرده و با بهره‌گیری از روش تولید زبان طبیعی و یا روش‌های دیگر، چینش مناسبی از این عبارات را که در قالب یک یا چند جمله بیان شوند، تولید و به عنوان شرح بر تصویر، نمایش داد.

۳. استفاده از قالب‌های زبانی آماده

با وجود فعالیت‌های گوناگون در این زمینه و استفاده از روش‌های مختلف، همچنان تضمین صحت جمله خروجی، کار دشواری است. به همین دلیل، سعی شد با ارائه یک یا چند قالب زبانی آماده و از پیش تعیین شده برای جملات، مانند قالب‌های جملات خبری، صحت جملات نهایی را تضمین کرد. در این دسته از روش‌ها، ویژگی‌های مختلفی از تصویر استخراج می‌شود که هریک از این ویژگی‌ها یا همه آن‌ها در کنار هم قادر هستند نقش‌هایی مانند « فعل »، « فاعل »، « مفعول » و موارد مشابه را در جمله متناظر با تصویر مشخص کنند. با استخراج کلمات مناسب و شناخت نقش آن‌ها در جمله و جای‌گذاری هر یک از این کلمات در مکان مناسب نقشی خود در قالب از پیش تعیین شده، جمله متناظر با هر تصویر استخراج می‌شود.

۴. استفاده از شبکه‌های عصبی بازگشتی

اگر چه استفاده از قالب‌های آماده و از پیش تعیین شده، تا حدی مشکلات موجود را حل می‌کند اما همچنان چالش بزرگ‌تری حل نشده باقی مانده است. تولید جملات جدید، استفاده از کلمات و عبارات جدید و ابتکاری به طوری که علاوه بر تضمین رعایت دستور زبان، بتوان معنای جمله را نیز متضمن شد، چالش بزرگی است که در این مسیر کماکان وجود دارد.

استفاده از شبکه‌های عصبی بازگشتی یکی از بهترین راه‌کارهای موجود برای حل این مشکل و رویارویی با این چالش هستند. استفاده از این شبکه‌ها در اوخر قرن بیستم در بین پژوهش‌گران رواج پیدا کرد تا جایی که ناپایداری الگوریتم پسانشان خطا در آموزش این شبکه، راه را برای پژوهش‌های بعدی بست. پس از ارائه یک روش مناسب برای بهینه‌سازی بدون هسین در سال ۲۰۱۰، روشی برای آموزش یک شبکه عصبی بازگشتی موسوم به شبکه عصبی بازگشتی ضربی بر مبنای بهینه‌سازهای بدون هسین ارائه شد و نتایج آن به طور چشم‌گیری از روش‌های موجود بیشتر بود.

ارائه شبکه عصبی بازگشتی ضربی، نقطه عطفی در مسیر علم در راستای حل چالش تولید جمله به حساب

می‌آید. از حدود سال ۲۰۱۱ به بعد، استفاده از شبکه‌های عصبی بازگشتی برای تولید جمله به پویاترین و پرفعالیت‌ترین حوزه در مسائل مربوط به تولید جمله، به حساب می‌آید.

خانم لی و همکارانش در سال ۲۰۱۵، در پژوهش [۷]، با استفاده از شبکه‌های عصبی کانولوشنی عمیق و دو نوع از شبکه‌های عصبی بازگشتی موسوم به شبکه‌های عصبی بازگشتی مالتی‌مودال و شبکه‌های عصبی بازگشتی دوطرفه، روش مناسبی برای تولید خودکار شرح بر تصاویر ارائه داده است.

در این پژوهش، ابتدا با بهره‌گیری از روش شبکه عصبی کانولوشنی ناحیه‌ای، نواحی از تصویر که شامل تصویر اجسام است، استخراج شده و با استفاده از یک شبکه عصبی کریشفسکی، بردار ویژگی برای هر ناحیه محاسبه می‌شود. سپس با بهره‌گیری از یک شبکه عصبی بازگشتی دوطرفه، عبارات مختلف از جمله استخراج و بردارهای ویژگی برای هر عبارت محاسبه می‌شود. سپس با استفاده از یک تابع هدف و مدل میدان تصادفی مارکف، همترازسازی بین نواحی و عبارات زبانی صورت گرفته و مدل آموزش داده می‌شود.

در ادامه با تخمین بهینه پارامترهای موجود و با استفاده از شبکه عصبی بازگشتی مالتی‌مودال، توزیع احتمال بهترین کلمه بعدی در یک جمله با داشتن کلمات قبلی و محتوای حاصل از بردار ویژگی محاسبه شده روی نواحی تصویر، محاسبه شده و بهترین کلمه بعدی تولید می‌شود. این کار تا جایی ادامه می‌یابد که شبکه، نشانه مخصوص پایان جمله را تولید کند.

۴-۷ یادگیری عمیق

از اواخر سال ۲۰۱۳، روش‌های مبتنی یادگیری عمیق، نظر بسیاری از پژوهش‌گرانی را که در حوزه تولید شرح متناظر تصویر فعالیت می‌کردند، به خود جلب نمودند. این دسته از روش‌ها، به دلیل عمل کرد بهتری که از خود نشان دادند، توانستند جایگزین روش‌های گرافی احتمالاتی شوند.

از جمله پژوهش‌هایی که با استفاده از شبکه‌های عصبی عمیق اقدام به تولید شرح متناظر تصویر کردند، می‌توان به پژوهش خانم لی و همکارانش [۷] در سال ۲۰۱۵ اشاره کرد. در مرحله آموزش این پژوهش، ابتدا با استفاده از روش شبکه عصبی کانولوشنی ناحیه‌ای که در بخش قبل، ارائه شد، نواحی تصویر که شامل تصویر یک جسم هستند، انتخاب شده و بردار ویژگی مربوط به هر کدام از این بخش‌ها، استخراج می‌شود.

پس از این مرحله، بردار ویژگی مربوط به جملات موجود در مجموعه‌داده، توسط یک شبکه عصبی بازگشتی دوطرفه، استخراج می‌شود. برای این کار، ابتدا بردار ویژگی مربوط به هر کلمه با استفاده از یک شبکه کلمه به بردار Word To Vec، استخراج شده و به عنوان ورودی به شبکه بازگشتی دوطرفه داده می‌شوند. استفاده از شبکه بازگشتی دوطرفه این امکان را می‌دهد که تاثیر کلمات قبل و بعد از هر کلمه، در تولید بردار ویژگی جملات لحاظ شود.

با بهینه‌سازی یک تابع انرژی روی این قسمت، شبکه عصبی بازگشتی دوطرفه و شبکه عصبی کانولوشنی با هم آموزش داده می‌شوند. از این طریق، بخش‌هایی از مدل که مربوط به تولید بردار ویژگی از جملات و استخراج نواحی تصاویر و بردار ویژگی مربوط به آن‌ها است، به طور کامل آموزش می‌بینند.

در ادامه فرایند آموزش شبکه، با ارائه بردار ویژگی تولید شده توسط شبکه عصبی کانولوشنی آموزش دیده در بخش قبلی به یک شبکه عصبی بازگشتی دیگر، و ارائه جملات موجود در مجموعه‌داده به آن، شبکه عصبی بازگشتی را

برای تولید جمله نهایی آموزش می‌دهیم.

آزمایشات انجام شده روی این پژوهش، معیار BLEU حاصل توسط روش را روی مجموعه‌داده MS COCO در مقایسه با روش‌های دیگر ارزیابی کرده‌اند. در این آزمایشات، بهترین عمل کرد روش ارائه شده روی این مجموعه‌داده به امتیاز BLEU برابر با ۵۷.۳ رسیده است و این در حالیست که روش [۳۲] روی همان مجموعه‌داده به مقدار ۵۵.۰ رسیده است.

یکی دیگر از روش‌های ارائه شده در این بخش، روشی است که در پژوهش [۸] در سال ۲۰۱۵ ارائه شده است. در این روش، یک شبکه عصبی بازگشتی دوطرفه برای نگاشت جملات و تصاویر به یکدیگر استفاده شده است. مدل ارائه شده، قادر است با گرفتن تصویر به عنوان ورودی، شرح متناظر آن را در قالب یک جمله تولید و با گرفتن یک جمله به عنوان ورودی، تصویر مربوط به آن را با بازیابی نماید.

در این روش با در نظر گرفتن واحد عصبی ارائه شده در پژوهش [۳۳] و اضافه کردن دو متغیر دیگر به آن، مدل نهایی تولید شده است. متغیرهای اضافه شده به این مدل، شامل متغیری برای بردار ویژگی تصویر و متغیر دیگر برای تفسیر بصری آخرین کلمه دیده شده، است.

شبکه عصبی ارائه شده در این پژوهش، توزیع احتمال توان تصاویر و جملات را مدل‌سازی می‌نماید. در صورتی که جمله به عنوان ورودی داده شده باشد، توزیع احتمال تصویر به شرط جمله قابل محاسبه و تصویر مربوطه قابل بازیابی است. در صورتی که تصویر به عنوان ورودی داده شده باشد، توزیع احتمال جمله به شرط تصویر قابل محاسبه است.

نتایج ارائه شده در این پژوهش، با روش‌های دیگر مقایسه شد. برای تولید جمله به شرط داشتن تصویر، میزان امتیاز BLEU حاصل توسط مدل در بهترین حالت برای مجموعه‌داده Flickr8k مقدار ۱۳.۱، برای مجموعه‌داده Flickr30k مقدار ۱۲.۰ و برای مجموعه‌داده MS COCO مقدار ۱۸.۸ بوده است. این در حالیست که نتایج حاصل برای مدل RNN + VGG به ترتیب برابر با ۱۲.۴، ۱۱.۹ و ۱۸.۴ بوده و مقادیر به دست آمده برای جملاتی که توسط عوامل انسانی تولید شده‌اند به ترتیب برابر با ۲۰.۶، ۱۸.۹ و ۱۹.۲ بوده است. نتایج نشان می‌دهد، روش ارائه شده در حوزه تولید شرح متناظر تصاویر از روش‌های استاندارد دیگر بهتر بوده اما هنوز به جملات تولید شده توسط انسان نمی‌رسد.

همین‌طور برای بازیابی تصاویر با داشتن جمله ورودی، نتایج حاصل توسط مدل برای مجموعه‌داده Flickr30k به ترتیب برای معیارهای $R@1$, $R@5$, $R@10$ و $R@500$ در بهترین حالت برابر با ۴۵.۷، ۱۸.۵، ۵۸.۱ و ۷ است. این در حالیست که نتایج حاصل توسط مدل RNN + VGG به ترتیب برابر با ۴۱.۱، ۱۵.۱، ۵۴.۱ و ۹ است.

۵-۷ توجه بصری

چارچوب کاری انکودر-دیکودر یکی از اصلی‌ترین چارچوب‌های کاری در حوزه ترجمه ماشینی و پیرو آن تولید شرح متناظر تصویر به شمار می‌رود. انکودر در این چارچوب کاری وظیفه نگاشت ورودی به فضای معنا و دیکودر وظیفه نگاشت فضای معنا به فضای خروجی را بر عهده دارد. در حوزه ترجمه ماشینی معمولاً از یک شبکه عصبی حافظه کوتاه‌مدت بلند به عنوان دیکودر استفاده می‌شود. این شبکه عصبی با دریافت کلمات جمله ورودی به ترتیب، بردار حالت مخفی خود را به روزرسانی می‌نماید. در نهایت می‌توان از این بردار به عنوان بردار حاصل نگاشت جمله

ورودی به فضای معنا استفاده نمود.

دیکودر در این چارچوب کاری با دریافت بردار ویژگی تولید شده توسط دیکودر، عمل تولید خروجی را بر عهده خواهد داشت. در حوزه ترجمه ماشینی معمولاً یک شبکه عصبی بازگشتی برای دیکودر می‌تواند مورد استفاده قرار بگیرد. به طور معمول، بردار ویژگی تولید شده توسط انکودر، به عنوان یک ورودی به دیکودر داده می‌شود و دیکودر در هر مرحله با تولید یک کلمه به عنوان خروجی، بردار حالت مخفی خود را به روزرسانی نموده و با استفاده از بردار حالت مخفی جدید، اقدام به تولید کلمه جدید می‌نماید.

یکی از محدودیت‌های جدی فرایند مذکور این است که بردار ویژگی فقط یک بردار با طول ثابت است و اولاً انکودر باید بتواند تمام اطلاعات قابل استخراج را تنها در این بردار جاسازی نماید و ثانیاً دیکودر باید بتواند تمام اطلاعات مورد نیاز خود برای تولید کلمه و جمله را فقط از همین یک بردار استخراج نماید. این مشکل، پژوهش‌گران را آن داشت تا بردار ویژگی را از یک بردار با طول ثابت به یک دنباله بردار با طول ثابت و تعداد متغیر تغییر دهند. به بردارهای ویژگی تولید شده در حالت جدید، حاشیه‌نویسی می‌گویند. این حاشیه‌نویسی‌ها باید دارای دو شرط زیر باشند:

۱. در برگیرنده تمام معنای ورودی باشند.

۲. تمرکز بیشتری روی معنای یک بخش مشخص از ورودی داشته باشند.

با در نظر گرفتن این ویژگی‌ها، دیکودر قادر خواهد بود تا هنگام تولید هر کلمه، روی معنای یک بخش از جمله تمرکز بیشتری داشته باشد و فقط از آن بخش برای تولید کلمه استفاده نماید. به این شکل، کلمات تولید شده شباهت بیشتری به ورودی خواهند داشت و ترجمه‌های بهتری حاصل خواهد شد.

در سال ۲۰۱۵، آقای بنجیو و همکارانش در پژوهش [۱۰] روشی ارائه دادند که در آن برای اولین بار از ایده استفاده از نقطه توجه در حوزه ترجمه ماشینی برای تولید شرح متناظر تصویر استفاده نمودند. در این پژوهش، از یک شبکه عصبی کانولوشنی به عنوان انکودر استفاده شده است. خروجی شبکه از لایه ماقبل آخر گرفته شده که منجر به ایجاد تعداد زیادی بردار ویژگی از تصویر می‌شود که هر کدام از این بردارهای ویژگی، از یک ناحیه از تصویر ایجاد شده‌اند و تمرکز بیشتری روی آن ناحیه داشته‌اند.

بدین ترتیب با استفاده از یک شبکه عصبی بازگشتی به عنوان دیکودر و استفاده از بردارهای حاشیه‌نویسی ایجاد شده توسط انکودر می‌توان به راحتی عملیات تولید شرح متناظر تصویر را انجام داد. تنها نکته‌ای که باید مشخص شود، چگونگی استفاده از بردارهای حاشیه‌نویسی است. در این پژوهش دو روش مختلف برای استفاده از بردارهای حاشیه‌نویسی مطرح شده است.

روش اول موسوم به روش توجه سخت، روشی است که در آن فقط یک بردار حاشیه‌نویسی انتخاب شده و از آن برای تولید جمله استفاده می‌شود. در این روش به هر یک از بردارهای حاشیه‌نویسی توسط یک مدل که قبل از آن آموزش دیده است، یک وزن اختصاص می‌دهیم و سپس با توجه به وزن‌های تخصیص داده شده به هر بردار حاشیه‌نویسی، یکی از آن‌ها را به عنوان بردار ویژگی تصویر انتخاب کرده و از آن در مراحل بعدی استفاده می‌کنیم.

روش دوم موسوم به روش توجه نرم، روشی است که در آن یک بردار ویژگی کلی از روی بردارهای حاشیه‌نویسی تولید شده و از آن بردار در مراحل بعدی استفاده می‌شود. برای تولید این بردار نیز مانند روش توجه سخت، ابتدا توسط یک مدل که از پیش‌آموزش دیده است، به هر یک از بردارهای حاشیه‌نویسی یک وزن اختصاص می‌دهیم.

سپس می‌توان با محاسبه امید ریاضی بردارهای حاشیه‌نويیسی با توجه به وزن هر یک از آن‌ها بردار ویژگی نهایی را برای تصویر تولید و از آن برای تولید جمله استفاده کرد.

آزمایشات انجام شده روی این مدل نشان می‌دهد، معیار BLEU-1 حاصل از این روش با استفاده از توجه سخت معمولاً از مدل توجه نرم بیشتر بوده است. مطابق با نتایج گزارش شده در این پژوهش، میزان امتیاز BLEU-1 حاصل توسط توجه سخت روی مجموعه‌داده‌های MS COCO، Flickr8k و Flickr30k به ترتیب برابر با ۷۱.۸، ۶۶.۹ و ۶۷.۰ است. این در حالیست که امتیاز حاصل توسط توجه نرم روی همین مجموعه‌های داده، به ترتیب برابر با ۷۰.۸، ۶۶.۷ و ۶۷.۰ و امتیاز کسب شده توسط مدل Log Bilinear در بهترین حالت، به ترتیب برابر با ۷۰.۸، ۶۵.۶ و ۶۵.۰ بوده است.

مطابق با آزمایشات انجام‌شده، استفاده از توجه نرم، معیار METEOR را نسبت به استفاده از توجه سخت افزایش می‌دهد. طبق نتایج گزارش شده در پژوهش، امتیاز METEOR حاصل از توجه نرم به ترتیب روی مجموعه‌داده‌های MS COCO، Flickr30k و Flickr8k برابر با ۱۸.۴۹، ۱۸.۹۳ و ۲۳.۹۰ است. این در حالیست که امتیاز کسب شده توسط روش Log Bilinear در بهترین حالت به ترتیب برابر است با ۱۸.۴۶، ۲۰.۳۰ و ۲۳.۰۴ و امتیاز کسب شده توسط روش Log Bilinear این‌که جملات تولید شده توسط روش توجه سخت با در نظر گرفتن جملات موجود در مجموعه‌داده از امتیاز بالاتری نسبت به جملات تولید شده توسط توجه نرم برخوردارند؛ استفاده از توجه نرم، منجر به تولید جملات قابل قبول‌تری توسط انسان می‌شود.

پژوهش‌های مختلفی از این ایده در حوزه‌های مختلف استفاده نموده‌اند که گزارش مختصری از تعدادی از این پژوهش‌ها ارائه شده است.

۶-۷ حافظه فعال

در این بخش، یکی از جدیدترین روش‌هایی را که در حوزه ترجمه ماشینی مورد استفاده قرار می‌گیرد و در پژوهش [۱۳] که در سال ۲۰۱۷ ارائه شده است، عمل کرد بهتری نسبت به روش‌های مبتنی بر نقطه توجه از خود نشان داده است را، که به نام حافظه فعال شناخته می‌شود، مورد بررسی قرار دادیم. در ابتدا واحد بازگشتی گیت‌دار و واحد بازگشتی گیت‌دار کانولوشنی معرفی شدند که روابط و ساختاری مشابه شبکه حافظه کوتاه‌مدت بلند دارند. سپس با استفاده از این واحدها، اقدام به ساخت ساختار شبکه GPU نمودیم.

مطابق با پژوهش [۱۳]، ساختار شبکه GPU به همان شکل که ارائه شد، توان رقابت با مدل‌های مشابه دیگر را در حوزه تولید شرح متناظر تصویر، ندارد. به همین دلیل نسخه‌های مارکفی و توسعه‌یافته این شبکه را ارائه نمودیم که عمل کرده‌های بهتری از خود نشان دادند. این شبکه در حوزه یادگیری الگوریتم، به ویژه در حوزه یادگیری الگوریتم‌های ساده مانند ضرب و جمع اعداد بسیار بزرگ، عمل کرد بسیار خوبی از خود نشان داده است.

ایده حافظه فعال بر خلاف ایده روش‌های مبتنی بر توجه، بر این است که در هر مرحله از تولید خروجی، از تمام حافظه موجود استفاده نماییم. نکته‌ای که باعث کاهش چشم‌گیر عمل کرد این مدل در حوزه ترجمه ماشینی می‌شود، عدم وجود وابستگی کافی بین مولفه‌های خروجی شبکه است. با اعمال وابستگی‌های مارکفی، نسخه مارکفی این شبکه قابل حصول است که علاوه بر بهبود عمل کرد نسخه استاندارد، توانایی رقابت با مدل‌های مبتنی

بر توجه را ندارد. در نسخه توسعه یافته این شبکه، وابستگی بین مولفه‌های خروجی، شدیدتر از وابستگی‌های مارکفی در نظر گرفته شده است که باعث افزایش چشم‌گیر کارایی مدل و رقابت‌پذیری مدل با نسخه‌های مبتنی بر توجه شده است.

مطابق با نتایج گزارش شده در پژوهش [۱۳]، نسخه توسعه یافته شبکه قادر به دستیابی به امتیاز BLEU برابر با ۲۹.۶ روی مجموعه‌داده جملات معادل زبان انگلیسی و فرانسوی شده است. این در حالیست که مدل مبتنی بر نقطه توجه امتیاز BLEU برابر با ۲۶.۴ را کسب کرده است. این بهبود عمل کرد از در نظر گرفتن وابستگی بیشتر بین مولفه‌های خروجی حاصل شده است.

مدل حافظه فعال می‌تواند با تعداد پارامترهای به مراتب کمتر نسبت به مدل‌های مبتنی بر نقطه توجه، عمل ترجمه ماشینی را انجام دهد. با این وجود هنوز مدل‌های مبتنی بر نقطه توجه که از توجه سخت استفاده می‌کنند عمل کردهای بهتری نسبت به مدل حافظه فعال از خود نشان می‌دهند.

۷-۷ نتیجه‌گیری

به دلیل پیچیدگی تحلیلی و عمل کردهای ضعیف، روش‌های مبتنی بر مدل‌های گرافی احتمالاتی، از سال ۲۰۱۳ به بعد کاربرد زیادی در بین پژوهش‌گران حوزه تولید شرح متناظر تصویر نداشتند و به طور کلی روش‌های مبتنی بر یادگیری عمیق، گوی سبقت را از روش‌های گرافی احتمالاتی ربوده‌اند. اما در میان روش‌های مبتنی بر یادگیری عمیق می‌توان پژوهش‌های موجود را به دو دسته کلی تقسیم‌بندی نمود.

۱. روش‌های استاندارد

این روش‌ها از ترکیب یک شبکه عصبی کانولوشنی با یک یا چند شبکه عصبی بازگشتی به منظور تولید شرح متناظر تصویر استفاده می‌نمایند. تعداد پژوهش‌هایی که در این دسته قرار می‌گیرند بسیار زیاد است و بخش قابل توجهی از این پژوهش‌ها با اعمال تغییرات کوچک در ساختار، تلاش برای دستیابی به دقت‌های بیشتر می‌نمایند. در بین پژوهش‌های موجود در این حوزه، می‌توان پژوهش [۷]، که توسط خانم لی ارائه شده است، را به عنوان یکی از کامل‌ترین پژوهش‌ها معرفی نمود که در هر دو بخش درک صحنه و تولید جمله، نوآوری‌های زیادی داشته است. فرایند آموزش این پژوهش کمی پیچیده است. تعداد پارامترهای مدلی که در این پژوهش مطرح شده است، حدود ۶۰ میلیون پارامتر است.

به نظر می‌رسد بتوان با ترکیب بخش‌هایی از این پژوهش‌ها با یکدیگر، روش جدیدی ارائه داد که به دقت‌های بهتری دست‌بیابد. به عنوان مثال، انتخاب چند کلیدواژه برای هر تصویر و مشروط کردن شبکه عصبی تولید‌کننده جمله، به تولید جملاتی که شامل این کلیدواژه‌ها باشند، انتظار می‌رود بتواند بهبود خوبی در نتایج حاصل ایجاد نماید.

۲. روش‌های مبتنی بر توجه بصری

این دسته از پژوهش‌ها از سال ۲۰۱۵ با پژوهش آقای بنجیو [۱۰] به شکل جدی مطرح شدند. در این دسته از پژوهش‌ها، ارائه مکانیزم مناسب برای محاسبه نقطه توجه از اهمیت بالایی برخوردار است. تعداد پارامترهای مدلی که در این پژوهش ارائه شده است حدود ۱۲۰ میلیون پارامتر است که منجر به نیاز به پردازنده گرافیکی قوی دارد. با این حال، فرایند آموزش این مدل‌ها ساده‌تر از مدل‌های قبلی است.

یکی از ایده‌های دیگر برای نوآوری در پژوهش این است که از ایده حافظه فعال در زمینه تولید شرح متناظر تصویر استفاده شود. تا کنون پژوهشی از این ایده در حوزه تولید شرح متناظر تصویر استفاده نکرده است. با توجه به این که مقاله [۱۳] در سال ۲۰۱۷ ارائه شده و مقایسه‌ای بین این روش و روش‌های مبتنی بر توجه ارائه داده است، احتمال این که پژوهش‌گرهای دیگری اقدام به استفاده از آن در حوزه ترجمه ماشینی نمایند بسیار زیاد است. با این حال هنوز نکات مبهم بسیاری در این روش برای من وجود دارد.

علاوه بر این، ترکیب ایده حافظه فعال با روش‌های مبتنی بر توجه بصری با استفاده از یک ماسک روی حافظه می‌تواند مورد استفاده قرار بگیرد که در پژوهش [۱۳] به عنوان روال‌های آتی مطرح شده است. ممکن است ارائه یک ترکیب مناسب و استفاده از آن در حوزه تولید شرح متناظر تصاویر بتواند نتایج قابل مشاهده‌ای را ارائه دهد.

مراجع

- [1] Fei-Fei, Li, Iyer, Asha, Koch, Christof, and Perona, Pietro. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10–10, 2007.
- [2] Li, Li-Jia and Fei-Fei, Li. What, where and who? classifying events by scene and object recognition. in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE, 2007.
- [3] Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [4] Kuznetsova, Polina, Ordonez, Vicente, Berg, Alexander C, Berg, Tamara L, and Choi, Yejin. Collective generation of natural image descriptions. in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers- Volume 1*, pp. 359–368. Association for Computational Linguistics, 2012.
- [5] Gupta, Ankush and Mannem, Prashanth. From image annotation to image description. in *International Conference on Neural Information Processing*, pp. 196–204. Springer, 2012.
- [6] Sutskever, Ilya, Martens, James, and Hinton, Geoffrey E. Generating text with recurrent neural networks. in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1017–1024, 2011.
- [7] Karpathy, Andrej and Fei-Fei, Li. Deep visual-semantic alignments for generating image descriptions. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.
- [8] Chen, Xinlei and Lawrence Zitnick, C. Mind’s eye: A recurrent visual representation for image caption generation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2422–2431, 2015.
- [9] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [10] Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron C, Salakhutdinov, Ruslan, Zemel, Richard S, and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. in *ICML*, vol. 14, pp. 77–81, 2015.

- [11] Luong, Minh-Thang, Pham, Hieu, and Manning, Christopher D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
- [12] Kaiser, Lukasz and Sutskever, Ilya. Neural gpus learn algorithms. *arXiv preprint arXiv:1511.08228*, 2016.
- [13] Kaiser, Lukas and Bengio, Samy. Can active memory replace attention? *arXiv preprint arXiv:1610.08313v2 [cs.LG]*, 2017.
- [14] Hoiem, Derek, Hays, James, Xiao, Jianxiong, and Khosla, Aditya. Guest editorial: Scene understanding. *International Journal of Computer Vision*, 112(2):131–132, 2015.
- [15] Potter, Mary C. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.
- [16] Potter, Mary C, Staub, Adrian, Rado, Janina, and O’Connor, Daniel H. Recognition memory for briefly presented pictures: the time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance*, 28(5):1163, 2002.
- [17] Fidler, Sanja, Sharma, Abhishek, and Urtasun, Raquel. A sentence is worth a thousand pixels. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1995–2002, 2013.
- [18] Farhadi, Ali, Hejrati, Mohsen, Sadeghi, Mohammad Amin, Young, Peter, Rashtchian, Cyrus, Hockenmaier, Julia, and Forsyth, David. Every picture tells a story: Generating sentences from images. in *Computer Vision–ECCV 2010*, pp. 15–29. Springer, 2010.
- [19] Felzenszwalb, Pedro, McAllester, David, and Ramanan, Deva. A discriminatively trained, multiscale, deformable part model. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.
- [20] Divvala, Santosh K, Hoiem, Derek, Hays, James H, Efros, Alexei A, and Hebert, Martial. An empirical study of context in object detection. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1271–1278. IEEE, 2009.
- [21] Lin, Dahua, Fidler, Sanja, and Urtasun, Raquel. Holistic scene understanding for 3d object detection with rgbd cameras. in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

- [22] Ladicky, L'ubor, Sturgess, Paul, Alahari, Karteek, Russell, Chris, and Torr, Philip HS. What, where and how many? combining object detectors and crfs. in *Computer Vision–ECCV 2010*, pp. 424–437. Springer, 2010.
- [23] Ladicky, Lubor, Russell, Chris, Kohli, Pushmeet, and Torr, Philip HS. Graph cut based inference with co-occurrence statistics. in *Computer Vision–ECCV 2010*, pp. 239–253. Springer, 2010.
- [24] Felzenszwalb, Pedro F, Girshick, Ross B, McAllester, David, and Ramanan, Deva. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.
- [25] Girshick, Ross, Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [26] Uijlings, Jasper RR, van de Sande, Koen EA, Gevers, Theo, and Smeulders, Arnold WM. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
- [27] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [28] Karpathy, Andrej, Joulin, Armand, and Li, Fei Fei F. Deep fragment embeddings for bidirectional image sentence mapping. in *Advances in neural information processing systems*, pp. 1889–1897, 2014.
- [29] Karpathy, Andrej and Fei-Fei, Li. Deep visual-semantic alignments for generating image descriptions. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [30] Reiter, Ehud and Dale, Robert. Building applied natural language generation systems. *Natural Language Engineering*, 3(01):57–87, 1997.
- [31] Lawrence, Steve, Fong, Sandiway, and Giles, C Lee. Natural language grammatical inference: a comparison of recurrent neural networks and machine learning methods. in *International Joint Conference on Artificial Intelligence*, pp. 33–47. Springer, 1995.

- [32] Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan L. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
- [33] Mikolov, Tomas, Karafiat, Martin, Burget, Lukas, Cernocky, Jan, and Khudanpur, Sanjeev. Recurrent neural network based language model. in *Interspeech*, vol. 2, p. 3, 2010.
- [34] Yang, Zichao, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Smola, Alex. Stacked attention networks for image question answering. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.
- [35] Cho, Kyunghyun, Courville, Aaron, and Bengio, Yoshua. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.
- [36] Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [37] Vinyals, Kaiser, Koo, Petrov, Sutskever, and Hinton. Grammar as a foreign language. in *Advances in Neural Information Processing Systems*, 2015.