



دانشگاه صنعتی امیر کبیر
(پلی تکنیک تهران)

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

Automatic Image Captioning Using Deep Convolutional and Recurrent Neural Networks

استاد راهنما

دکتر صفابخش

پژوهش گر

احمد اسدی

۹۴۱۳۱۰۹۱

اردیبهشت ماه ۱۳۹۵

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	۱	فصل اول مقدمات
۱	۱	۱.۱ مقدمه
۲	۲	۲.۱ تعریف مساله
۳	۳	۳.۱ درک صحنه
۴	۱.۳.۱	۱.۳.۱ پژوهش‌های انجام‌شده در زمینه درک صحنه توسط مغز انسان
۶	۲.۳.۱	۲.۳.۱ نتایج به‌دست آمده از آزمایشات
۸	۴.۱	۴.۱ جمع‌بندی
۱۰	۲	فصل دوم درک صحنه
۱۰	۱.۲	۱.۲ درک صحنه
۱۰	۲.۲	۲.۲ روش‌های مختلف موجود
۱۱	۳.۲	۳.۲ روش‌های مبتنی بر مدل‌های گرافی احتمالی
۱۱	۱.۳.۲	۱.۳.۲ استفاده از مدل میدان تصادفی مارکف
۱۴	۲.۳.۲	۲.۳.۲ استفاده از مدل میدان تصادفی شرطی
۱۶	۳.۳.۲	۳.۳.۲ استفاده از سایر مدل‌های گرافی احتمالی
۲۴	۴.۲	۴.۲ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق
۲۴	۱.۴.۲	۱.۴.۲ اختصاص معنا به قطعه‌های مختلف تصویر
۲۵	۲.۴.۲	۲.۴.۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دوطرفه جملات و تصاویر
۳۰	۵.۲	۵.۲ جمع‌بندی
۳۴	۳	۳ مراجع و منابع

فهرست تصاویر

۵	نمونه توصیف‌های افراد برای تصاویر	۱
۵	ساختار مطلوب اطلاعات استخراج شده از تصاویر [۱]	۲
۶	تصاویر دنیای واقعی مورد استفاده در آزمایشات [۱]	۳
۷	نمودار مقایسه‌ای عملکرد مغز در درک صحنه	۴
۸	نمونه‌ای از نتایج بهدست‌آمده از آزمایشات [۱]	۵
۱۲	نگاشت تصویر به فضای معنایی	۶
۱۳	مدل میدان تصادفی مارکف در درک صحنه	۷
۱۵	مدل سلسله‌مراتبی میدان تصادفی شرطی در درک صحنه	۸
۱۶	مدل گرافی احتمالی مورد استفاده در پژوهش [۲]	۹
۲۰	نمونه تصاویر موجود در مجموعه‌داده مورد استفاده. [۲]	۱۰
۲۱	ماتریس درهم‌ریختگی مدل کامل ارائه شده در [۲]	۱۱
۲۲	نتیجه مقایسه مدل‌های مختلف در [۲]	۱۲
۲۳	نتایج نهایی بهدست آمده از مدل بر روی تصاویر. [۲]	۱۳
۲۶	طرح‌واره عملکرد روش RCNN	۱۴
۲۸	نتایج عملکرد اهداف تعریف شده در روش RCNN برای همترازسازی تصاویر و جملات	۱۵
۲۹	نتایج نهایی روش RCNN	۱۶
۳۰	نتایج حاصل از جستجوی جملات در روش RCNN	۱۷

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؛ به طوری که 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 را به بردار ویژگی تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی را به جمله زبان مقصد ترجمه نمود، جمله‌ای معادل با تصویر ورودی به‌دست خواهد آمد. که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است.

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده است. 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 به‌دست‌آمده 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، املاء 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 [۳] 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 [۴] و [۵] اشاره کرد. در این پژوهش‌ها، با نشان دادن تصاویر به صورت دنباله‌ای^{۱۳} 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 و اگرچه این زمان برای تشخیص و توصیف یک تصویر کافیست، زمان مورد نیاز برای به‌خاطر‌سپاری تصویر بسیار بیشتر از این مقدار است.

در پژوهش [۱] آزمایش دیگری انجام شد که از اهمیت بسیاری برخوردار است. در پژوهش‌های قبلی، افراد در باره موضوع کلی تصاویر اطلاعاتی داشتند. اما در این آزمایش، تصاویر مختلفی از دنیای واقعی که محدود به شرایط خاصی نبوده‌اند، 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.

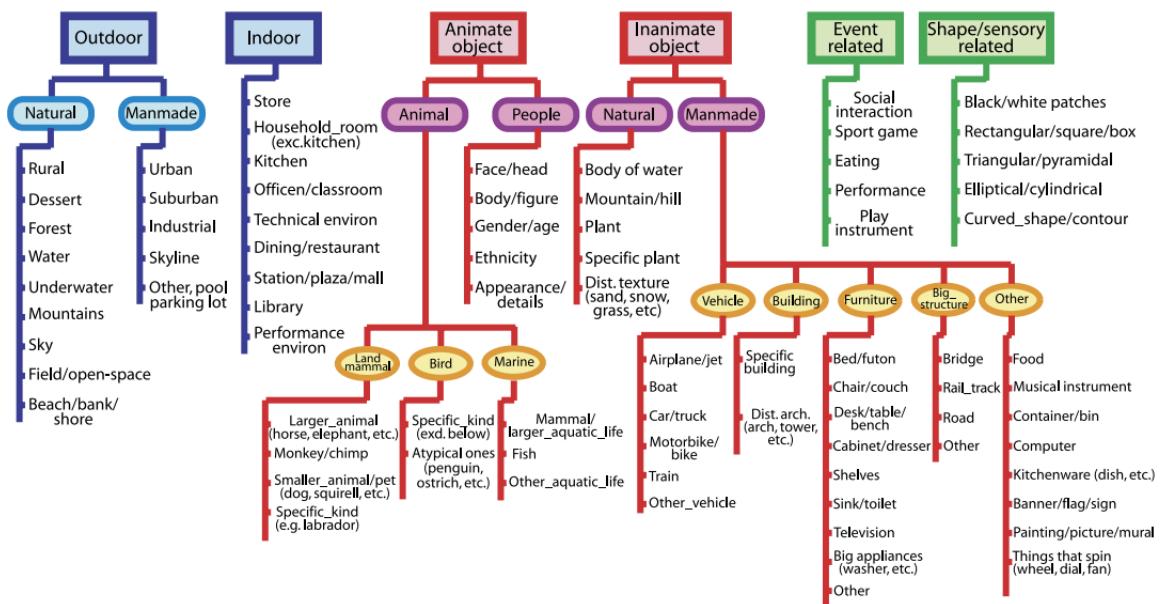
۱. توسط یک رایانه، تصاویر متعددی در بازه‌های زمانی متفاوت به افراد نمایش داده می‌شوند و پس از اتمام زمان نمایش هر تصویر، یک ماسک بصری، تصویر را می‌پوشاند. در این حالت از افراد خواسته شده بهترین توصیف ممکن از تصویر را تایپ کنند. شرایط محیطی آزمایشات مطابق با استانداردها رعایت شده است. هر تصویر به طور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

^{۱۳}Image Series

	PT = 107 ms	PT = 500 ms
	<p>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)</p>	<p>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)</p>
		<p>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)</p>

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 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 خواسته شده تا هر یک، نام ۵ دسته صحنه مختلف را به طور تصادفی بنویسند. پس از حذف نام‌های تکراری، ۲۵ الی ۳۰ نام منحصر به فرد باقی مانده است. سپس تصاویر مربوط به هریک از این نام‌ها توسط موتور جستجوی گوگل استخراج شده و ۶ الی ۲۵ 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

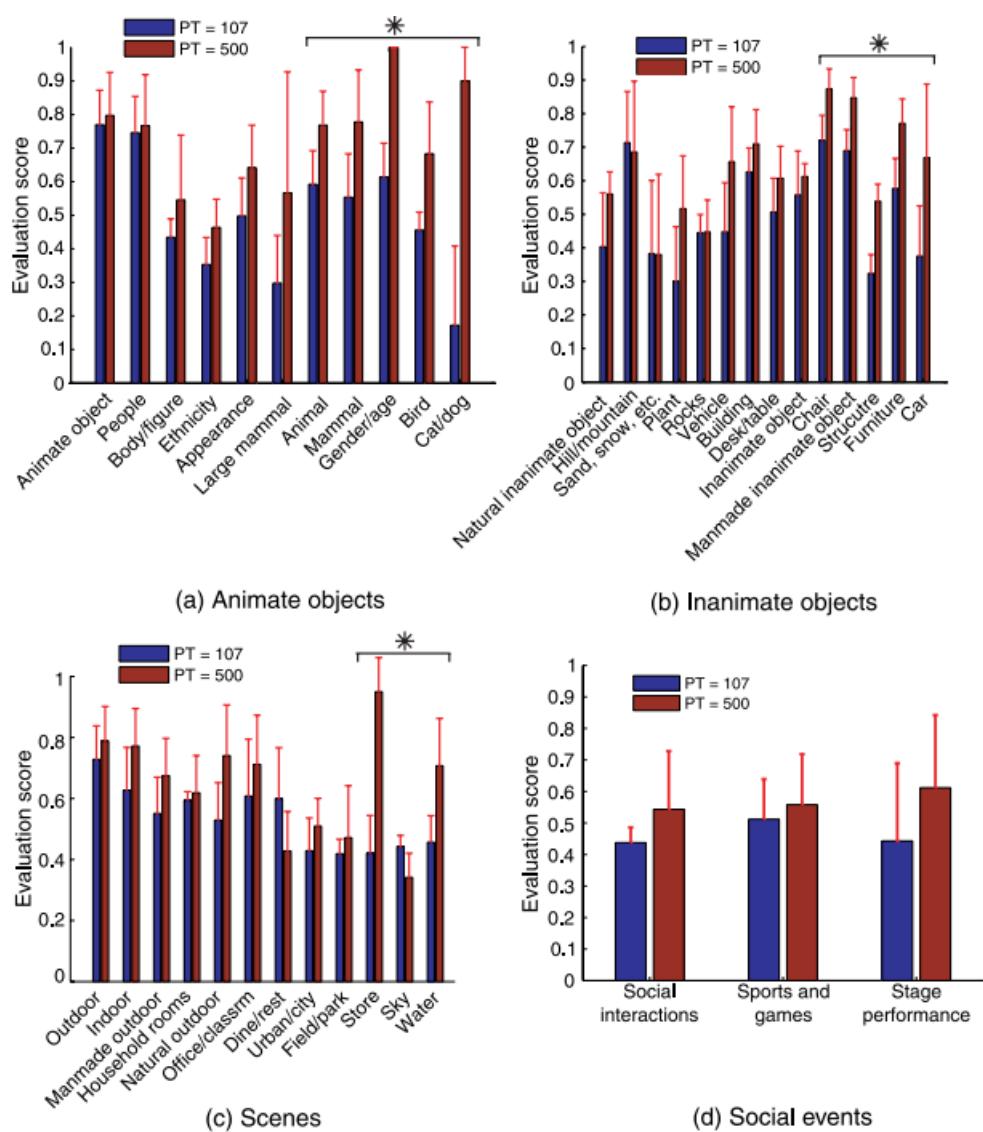
شکل ۳: تصاویر دنیای واقعی مورد استفاده در آزمایشات [۱]

ارزشمندترین نکته درباره پژوهش انجام شده، 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.

۲.۳.۱ 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، به طور کامل توصیف شده است در صورتی که تصویر دوم که به نسبت، پیچیده‌تر است، مدت زمان بیشتری برای توصیف نیاز داشته است.

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 میلی‌ثانیه و ۵۰۰ میلی‌ثانیه ایجاد شده است. شکل ۴ نمودارهای مقایسه‌ای را نمایش می‌دهد. در این نمودارها، میله‌های قرمز نشان‌دهنده نتایج برای زمان ۵۰۰ میلی‌ثانیه و میله‌های آبی نمایش‌دهنده نتایج برای حالت ۱۰۷ میلی‌ثانیه 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 معز انسان در درک صحنه در بازه‌های زمانی ۱۰۷ و ۵۰۰ میلی‌ثانیه [۱]

^{۱۴}Animated

^{۱۵}Inanimate

همان طور که مشخص است، مدت زمان ۱۰۷ میلی ثانیه برای مغز انسان، زمان بهینه برای توصیف صحنه است. تفاوت های 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 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.

همین طور با مقایسه تفاوت عملکرد بین حالات متحرک بودن و متحرک نبودن اجسام، فواصل موجود در نمودارها قابل ملاحظه می شود. 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، متحرک باشند.

شکل ۵ نمونه دیگری از نتایج بدست آمده از آزمایشات را در مدت زمان های مختلف نمایش می دهد.

		
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 rock 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: نمونه ای از نتایج بدست 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 محتواهای 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 [۴] و [۵] 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 گروهی 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 [۶] یک مدل میدان

^{۱۶}Probabilistic Graphical Models (PGMs)

^{۱۷}Image Segmentation

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.

۱.۳.۲ استفاده از مدل میدان تصادفی مارکف^{۲۰}[۸]

مقاله [۸] 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

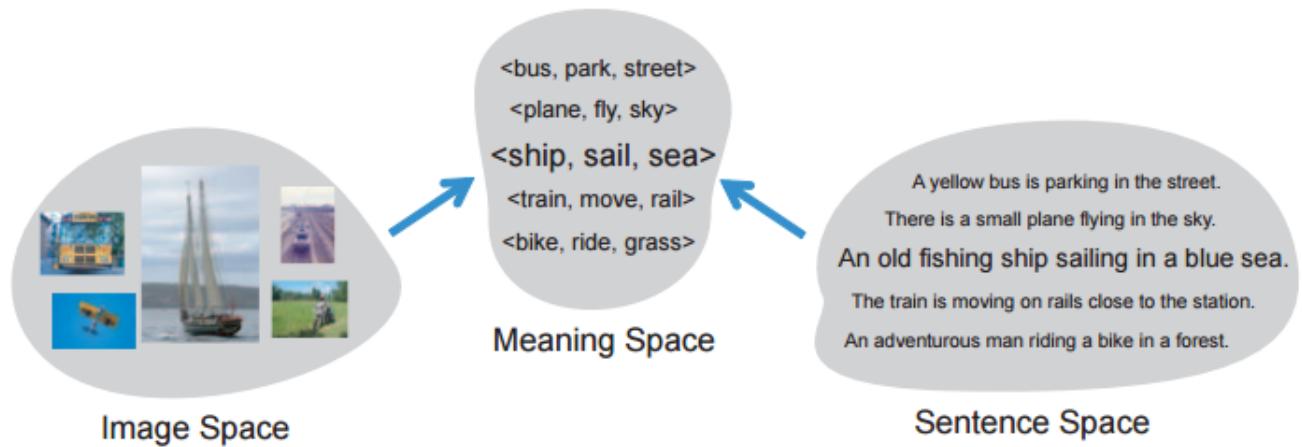
^{۲۰}Markov Random Field (MRF)

^{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 [؟]

مدل میدان تصادفی مارکف مورد استفاده در این پژوهش، یک مدل کوچک و ساده، شامل ۳ گره است. شکل ۷ طرح‌واره‌ای از 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 [۹].

پس از محاسبه امتیاز اطمینان همه دسته‌های موجود، دسته‌ای که بیشترین امتیاز را دارد می‌تواند

^{۲۴}Detector

^{۲۵}Felzenszwalb

^{۲۶}Confidence Score

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 مجموعه‌داده وجود داشته باشد.

(ب) استفاده از پاسخ دسته‌بندی‌کننده دیوالا^{۷۷}، ارائه شده در مقاله [۱۰]

(ج) استفاده از دسته‌بندی‌کننده مبتنی بر گیست[?]



شکل ۷: طرح‌واره مدل میدان تصادفی مارکف ارائه شده در پژوهش [۸] 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 [۸].

بر اساس مقادیر محاسبه شده برای ویژگی‌های بالا و با استفاده از الگوریتم ماشین بردار پشتیبان^{۷۸}، یک دسته‌بندی برای هر گره ارائه می‌شود که بیان کننده 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}Support Vector Machine (SVM)

۲.۳.۲ استفاده از مدل میدان تصادفی شرطی^{۲۹}

در این پژوهش، مساله در ک 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 در ک صحنه است که پژوهش‌های متعددی از آن به عنوان مدل اصلی در در ک صحنه استفاده کرده‌اند. به عنوان نمونه، در مقاله‌های [۱۱] و [۱۲] از مدل میدان تصادفی شرطی به منظور توصیف صحنه استفاده شده است.

پژوهش [۱۱] 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 [۱۲]، یک چارچوب کاری^{۳۰} 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 در ک 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 [۶] 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 [۶] 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.

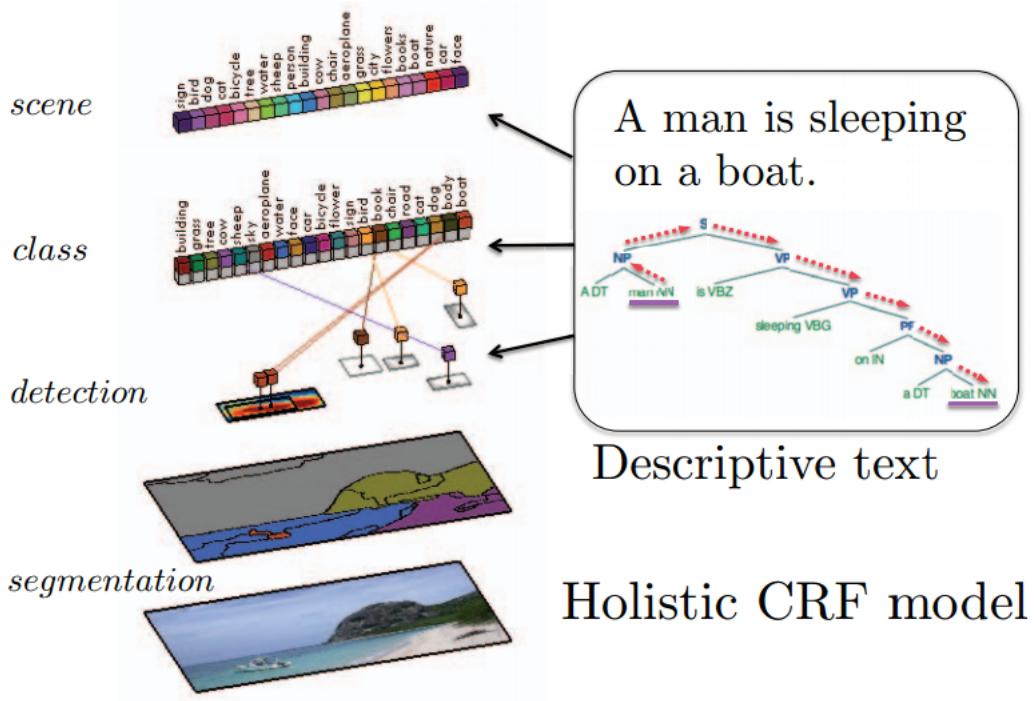
دو دسته متغیر تصادفی مختلف، که هر یک نماینده متغیرهای تصادفی موجود در یکی از این سطوح انتزاع هستند، تعریف شده‌اند؛ متغیرهای تصادفی C ، $X_i \in \{1, \dots\}$ بیان‌کننده دسته قطعه i از سطح پایین سلسله مراتب و متغیرهای تصادفی C ، $Y_j \in \{1, \dots\}$ بیان‌کننده دسته قطعه j از سطح بالای سلسله مراتب. به علاوه، دو دسته متغیر تصادفی دیگر به نام‌های b_k و z_k به ترتیب برای نمایش حضور یا عدم حضور یک تشخیص کاندید^{۳۱} 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^{۳۲}، منجر به یافتن پاسخ مورد نظر می‌شود.

^{۲۹}Conditional Random Field (CRF)

^{۳۰}Framework

^{۳۱}Candidate Detection

^{۳۲}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 [۶].

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^{۳۳}

پتانسیل قطعه‌بندی یگانی در هر قطعه و هر ابرقطعه^{۳۴} از تصویر، با استفاده از میانگین‌گیری روی امتیاز افزایش تکستون^{۳۵} که در پژوهش [۱۳] 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.

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

^{۳۳}Unary Segmentation Potential

^{۳۴}Supersegment

^{۳۵}Texton Boost

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

۳. پتانسیل انطباق تصویر و دسته جسم

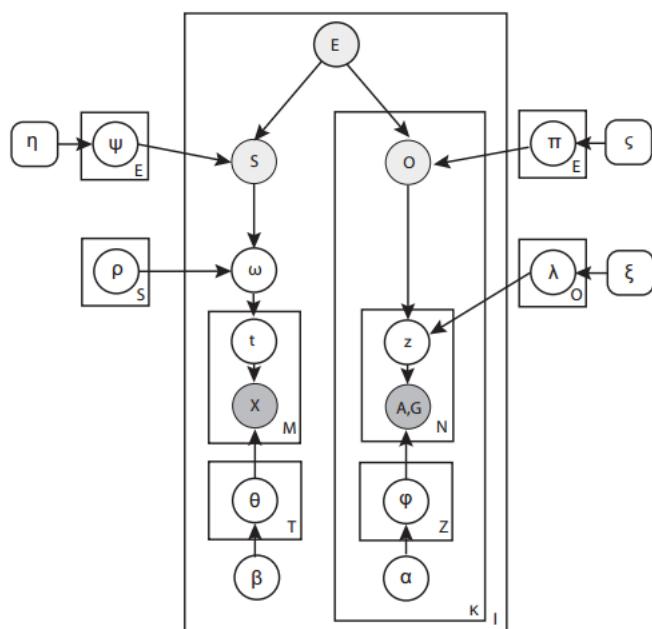
برای اندازه‌گیری میزان انطباق هر کدام از دسته‌های موجود برای اجسام با تصویر ورودی، از معیار انطباق ارائه شده در پژوهش [۱۴] توسط فلزنسوالب که به روش دی پی ام^{۳۶} مشهور است، استفاده شده است. برای کاهش تعداد پارامترها و افزایش کارایی مدل استفاده شده، برای هر تصویر حداقل ۳ دسته جسم، به عنوان دسته‌های منتخب 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 تجمع اطلاعات موجود در این دو سطح، اقدام به دسته‌بندی تصویر می‌نماید. شکل ۹ مدل استفاده شده در این پژوهش را به منظور یکپارچه‌سازی و تجمع اطلاعات حاصل از تحلیل صحنه و تشخیص اجسام موجود در آن، ارائه می‌دهد.



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

^{۳۶}DPM

یکی از اهدافی که در این پژوهش دنبال می‌شود، برچسب‌گذاری معنایی^{۳۷} تمام پیکسل‌های موجود در تصویر است. به همین منظور، تمام تصاویر مورد استفاده، به نواحی $10 * 10$ تقسیم شده و مورد استفاده قرار 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^{۳۸} 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^{۳۹} 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^{۴۰} 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^{۴۱} 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^{۴۲} حاکم بر توزیع احتمالاتی S به شرط داشتن E است. به علاوه، ψ یک ماتریس به ابعاد $S * E$ و پارامتر η یک بردار S بعدی در نقش مقدار اولیه دیریکله^{۴۳} برای پارامتر ψ است.

* در قدم بعدی با داشتن مقدار S ، پارامترهای ω را بر اساس احتمال $(\rho|S, \omega)$ تولید می‌کنیم. آن‌جا 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:

^{۳۷}Semantic Labelling

^{۳۸}SIFT Descriptor

^{۳۹}Latent

^{۴۰}Event

^{۴۱}Scene Image

^{۴۲}Multinomial

^{۴۳}Dirichlet prior

- یک مقدار t از توزیع احتمال $Mult(\omega)$ تولید می‌شود که مشخص‌کننده موضوعی^{۴۴} 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 $T * V_s$ 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^{۴۵} 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 $Mult(\lambda, |O)$ تولید می‌شود. متغیر λ یک ماتریس به ابعاد $O * Z$ است که در آن Z تعداد کل مقادیر مختلف متغیر z است. به علاوه ξ 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 $V_o * Z * V_o$ 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.

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

^{۴۴}Topic

^{۴۵}Object Image

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.

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

$$P(\omega|S, \rho) = Dir(\omega|\rho_{j.}), S = j \quad (5)$$

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

$$P(X_m|t_m, \theta) = P(X_m|\theta_{j.}), t_m = j \quad (7)$$

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

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

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

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) اجسامی که در تصویر حضور دارند.

با توجه به این محدودیت و با در نظر گرفتن مدل ارائه شده، استفاده از تخمین بیشینه احتمال^{۴۶}، 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 $\{\beta, \psi, \rho, \pi, \lambda, \theta\}$ است. این فرایند برای سه پارامتر $\{\beta, \psi, \rho, \theta\}$ با استفاده از روش انتقال پیام متغیر^{۴۷} و برای سه پارامتر $\{\pi, \lambda, \beta\}$ با استفاده از نمونه‌برداری گیبس^{۴۸} انجام می‌شود.

^{۴۶}Maximum Likelihood

^{۴۷}Variational Message Passing

^{۴۸}Gibbs Sampling

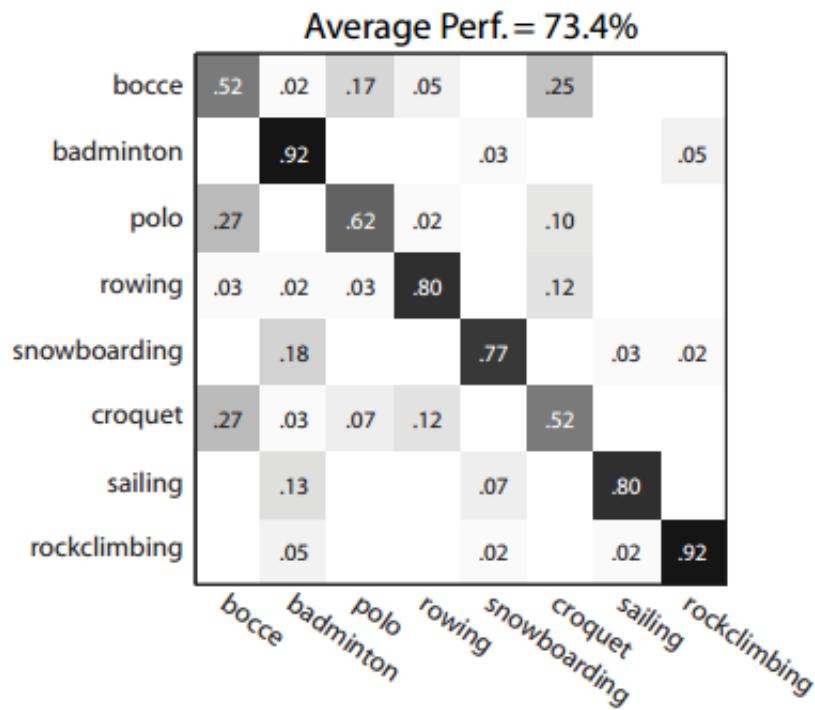
آزمایشات انجام شده در این پژوهش، بر روی یک مجموعه‌داده شامل تصاویر از ۸ دسته ورزشی مختلف که در هر دسته، بین ۱۳۷ تا ۲۵۰ تصویر مختلف وجود دارد، انجام شده‌اند. از جمله چالش‌های موجود در این مجموعه‌داده می‌توان به وجود زمینه‌های متنوع و پیچیده در تصاویر، تنوع دسته‌های مختلف اجسام موجود، تنوع اندازه اجسام 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

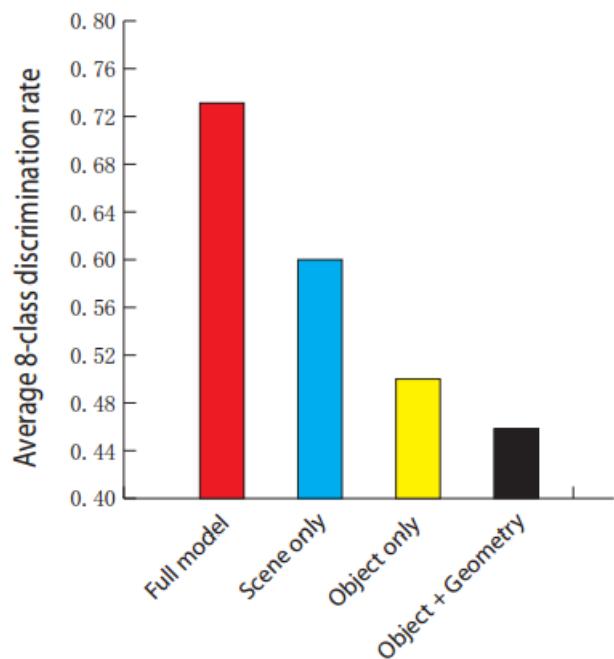
۱۱ ماتریس درهم‌ریختگی^{۴۹} مربوط به این مدل را نمایش می‌دهد. همان‌طور که در این ماتریس مشخص است، کمترین نرخ تشخیص در بین دسته‌های ورزشی موجود در این مدل، ۵۲٪ و بیشترین نرخ تشخیص ۹۲٪ است.



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

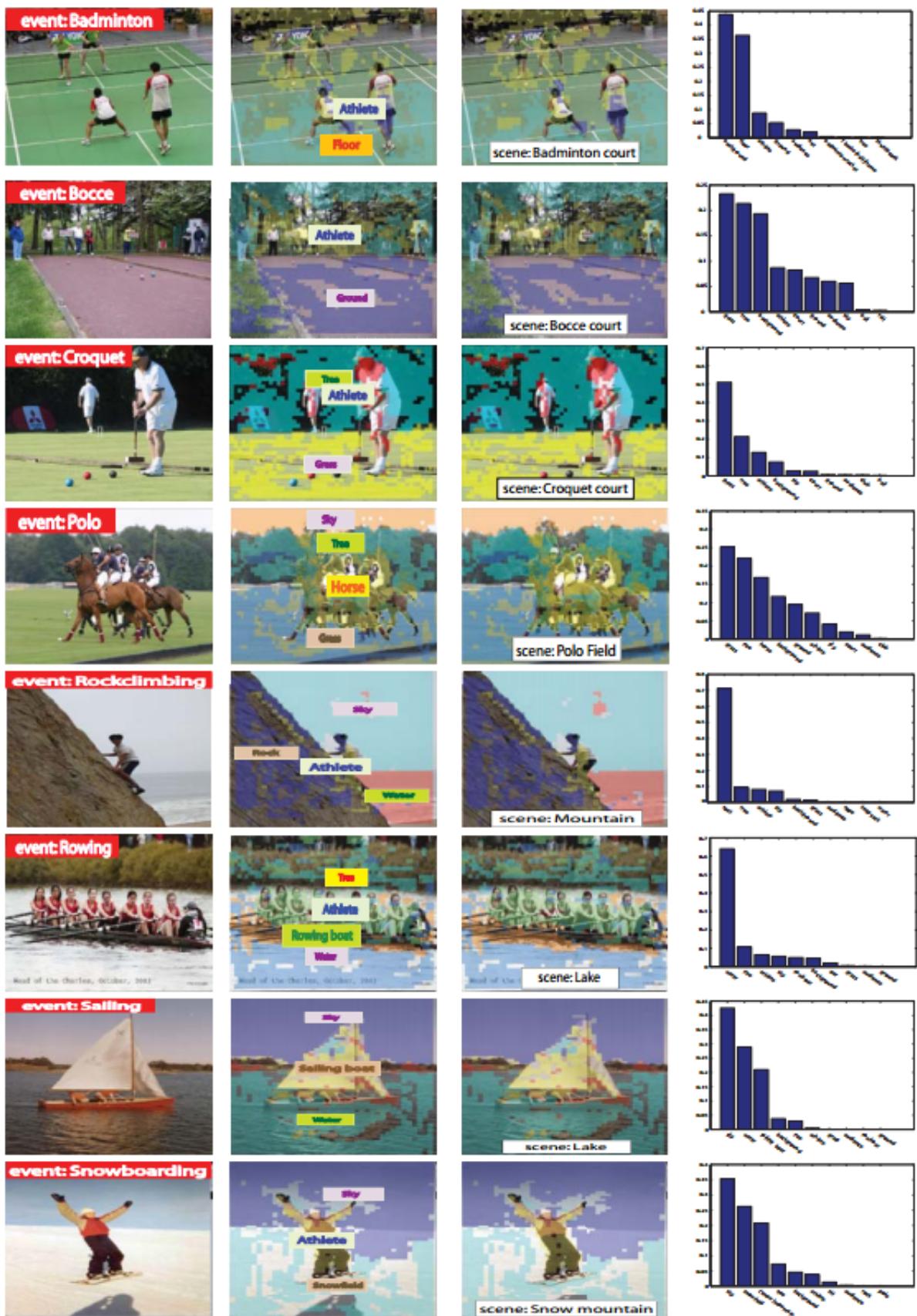
بسته به میزان استفاده از اطلاعات مختلف استخراج شده برای استنتاج، مدل‌های مختلفی به وجود می‌آیند که در شکل ۱۲ نتایج عملکرد هریک از این مدل‌ها با مدل‌های دیگر مقایسه شده است. همان‌طور که در شکل ۱۲ مشخص است، بهترین کارایی مربوط به مدل کامل است. در صورتی که در مدل، فقط از اطلاعات مربوط به صحنه استفاده شود،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

^{۴۹}Confusion Matrix



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

شکل ۱۳ نتایج نهایی به دست 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 به دست آمده از رخداد موجود در تصویر، ستون دوم برچسب‌های تشخیص داده شده مربوط به اجسام 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 [۱۷]

در پژوهش [۱۷] 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^{۵۰}.

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^{۵۱} استفاده می‌شود. نسخه‌های مختلفی از این روش ارائه شده است. نسخه ارائه شده در پژوهش [۱۸]، یکی از سریع‌ترین نسخه‌های ارائه شده است که در این بخش از همین روش استفاده می‌شود.

در پژوهش [۱۸] 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 فضای می‌توانند در هر اندازه‌ای باشند و در نتیجه نواحی ارائه شده باید بتوانند ابعاد مختلف داشته باشند. این ویژگی عموماً با روش‌های سلسله‌مراتبی

^{۵۰} Category-independent region proposals

^{۵۱} 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^{۵۲} آزمایش شده در چالش دسته‌بندی اجسام مسابقه ImageNet است. اطلاعات دقیق درباره این شبکه عصبی در پژوهش [۱۹] در دسترس است.

شبکه عصبی کانولوشنی عمیق ارائه شده در این پژوهش با استفاده از یک مجموعه‌داده^{۵۳} آموزش دیده شده است. از این شبکه عصبی که تحت عنوان RCNN^{۵۴} شناخته می‌شود می‌توان به عنوان یک شبکه از پیش‌آموزش دیده استفاده کرد.

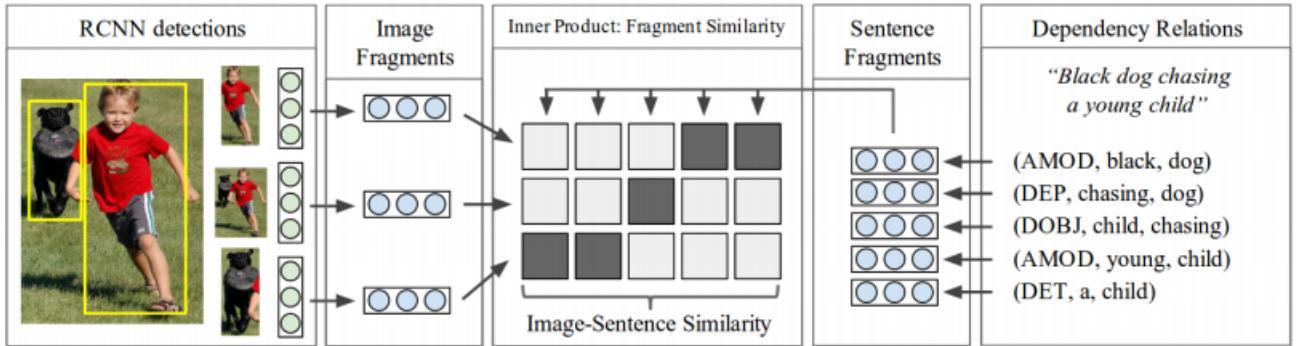
۲.۴.۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دو طرفه جملات و تصاویر [۲۰]

مدل ارائه شده در این پژوهش، مدلی است که قادر به نگاشت دو طرفه 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.

^{۵۲}Krizhevsky

^{۵۳}ILSVRC 2012

^{۵۴}Regional Convolutional Neural Network



شکل ۱۴: مدل استفاده شده برای نگاشت دو طرفه تصاویر و جملات به یکدیگر با استفاده از شبکه عصبی عمیق کانولوشنی. [۲۰]

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، I_b مجموعه تمام پیکسل‌های موجود در ناحیه b ، $RCNN_{\theta_c}$ شبکه عصبی آموزش‌دیده است که در آن θ_c مجموعه پارامترهای بهینه موجود در شبکه است. بردار حاصل v_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.

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

از طرفی با در نظر گرفتن بردار s_j به عنوان بردار حاصل از نگاشت جمله زام به فضای معنایی و در نظر گرفتن ضرب داخلی به عنوان شباهت، $s_j \cdot v_i^T$ 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. هم‌ترازسازی ناحیه‌ای^{۵۵} 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.

^{۵۵}Fragment Alignment

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 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^{۵۶} (با آزمون و خطا تعیین می‌شوند) و $\|\cdot\|_2^2$ یک عبارت تنظیم‌کننده^{۵۷} 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 کم‌باب می‌شود، با متغیر گرفتن پارامتر z_{ij} سعی

^{۵۶}Hyperparameter

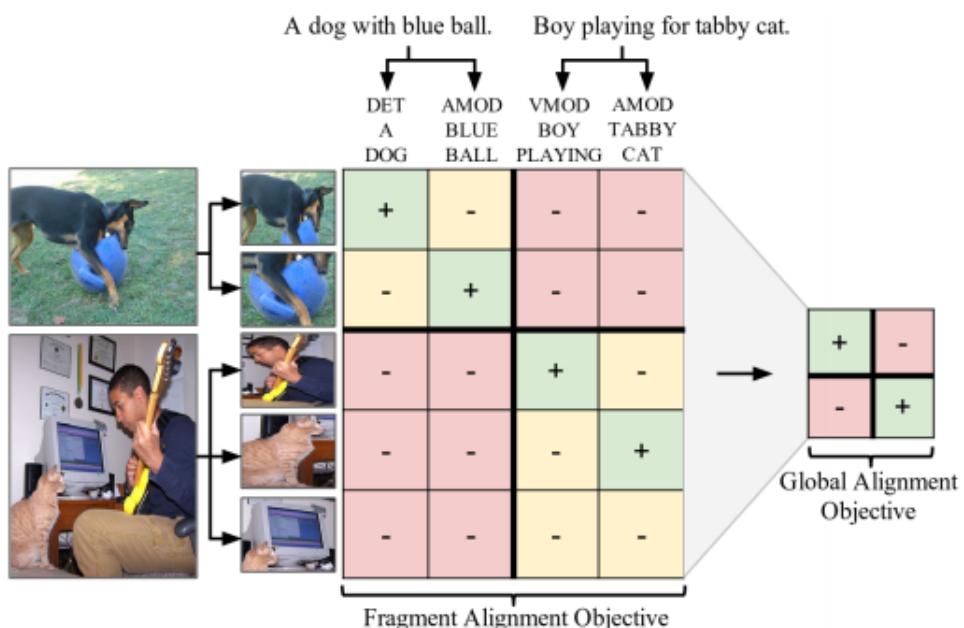
^{۵۷}Regularization Term

در یافتن کمترین مقدار آن می‌کنیم. رابطه ۱۷ معیار متناظر با هدف کلی همترازسازی ناحیه‌ای را بیان می‌کند.

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

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

در این رابطه، p_j مجموعه تصاویر موجود در کیسه مثبت^{۵۸} 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^{۵۹} این عبارت قرار می‌گیرند. ($m_\nu(i)$ و $(j) m_s(j)$ 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 نمایش دهنده قطعه‌های مختلف جملات هستند. سلول‌های قرمز رنگ حالاتی هستند که در آن‌ها $y_{ij} = 1$ ، سلول‌های زرد نمایش‌دهنده اعضای کیسه‌های مثبت هستند که در آن‌ها $y_{ij} = -1$ است. [۲۰]

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

^{۵۸}Positive Bag

^{۵۹}Negative Bag

شباخت بین یک تصویر و یک جمله تعریف می‌شود. این امتیاز مطابق با رابطه ۱۸ تعریف شده و برابر است با میانگین امتیاز شباخت دوبه‌دی نواحی مختلف تصویر با عبارات مختلف جمله.

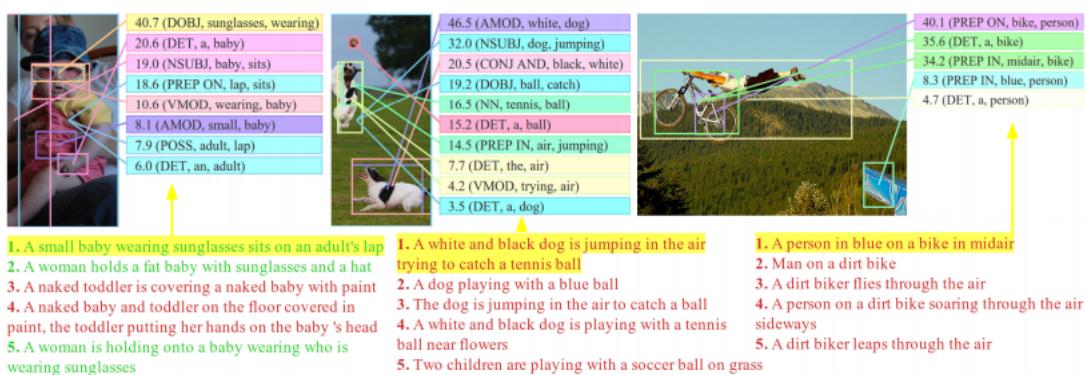
$$S_{kl} = \frac{1}{|g_k|(|g_l| + n)} \sum_{i \in g_k} \sum_{j \in g_l} \max(\circ, \nu_i^T \cdot s_j) \quad (18)$$

از آن‌جا 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 به‌دست آمده که نتایج را بهبود می‌بخشد. مقدار پیشنهاد شده در پژوهش، $n = 5$ است.تابع کلی هدف سراسری مطابق با رابطه ۱۹ تعریف می‌شود.

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

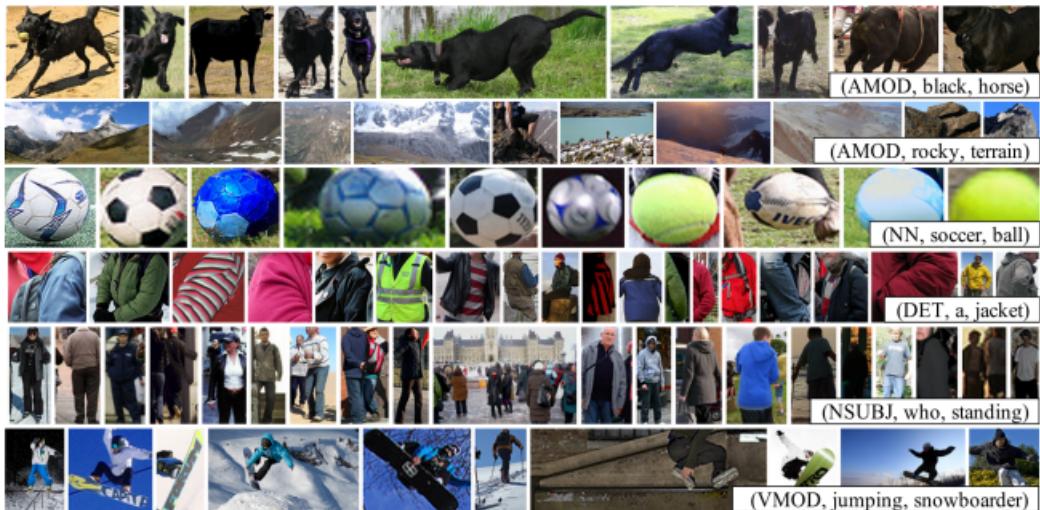
در رابطه ارائه شده، Δ یک ابرپارامتر است که با آزمون و خطا به‌دست می‌آید. عبارت اول درون پرانتز بیان‌کننده امتیاز تصویر و عبارت دوم بیان‌کننده امتیاز جمله هستند.

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. [۲۰]

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. [۲۰]

روش ارائه شده در این پژوهش، به طور کامل و دقیق در پژوهش [۲۱] 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.

۵.۲ 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 [۸] 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، 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 بعداً 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 [۶] 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 دی‌پی ام مشهور است.

(ج) سایر مدل‌های گرافی احتمالی در پژوهش [۲۱]، 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 [۱۷]، 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، روشنی 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 [۲۰] با استفاده از روش RCNN و تعریف دو تابع هدف دیگر برای شبکه، روشنی ارائه شده است که طی آن بتوان تصاویر و جملات را به طور دوطرفه به یکدیگر نگاشت کرد. تابع 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

طبيعي 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 مراجع و منابع

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] Hoiem, Derek, Hays, James, Xiao, Jianxiong, and Khosla, Aditya. Guest editorial: Scene understanding. *International Journal of Computer Vision*, 112(2):131–132, 2015.
- [4] Potter, Mary C. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.
- [5] 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.
- [6] 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.
- [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] 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.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] Ladický, Lubor, Sturgess, Paul, Alahari, Kartik, 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.
- [13] 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.
- [14] 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.
- [15] Li, Li-Jia, Socher, Richard, and Fei-Fei, Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2036–2043. IEEE, 2009.
- [16] Gould, Stephen, Fulton, Richard, and Koller, Daphne. Decomposing a scene into geometric and semantically consistent regions. in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1–8. IEEE, 2009.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.