دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

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

Automatic Image Captioning Using Deep Convolutional and Recurrent Neural Networks

استاد راهنما

دکتر صفابخش

پژوهش‌گر

احمد اسدی

۹۴۱۳۱۰۹۱

اردیبهشت‌ماه ۱۳۹۵

فهرست مطالب

۱ فصل اول مقدمات **۱**

۱.۱ مقدمه ۱

۲.۱ تعریف مساله ۲

۲ فصل دوم درک صحنه **۳**

۱.۲ درک صحنه ۳

۲.۲ روش‌های مختلف موجود ۳

۳.۲ روش‌های مبتنی بر مدل‌های گرافی احتمالی ۴

۱.۳.۲ استفاده از مدل میدان تصادفی مارکف ۴

۲.۳.۲ استفاده از مدل میدان تصادفی شرطی ۷

۳.۳.۲ استفاده از سایر مدل‌های گرافی احتمالی ۹

۴.۲ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق ۱۸

۱.۴.۲ اختصاص معنا به قطعه‌های مختلف تصویر ۱۸

۲.۴.۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دوطرفه جملات و تصاویر ۱۹

۳.۴.۲ هم‌ترازسازی اطلاعات بصری و معنایی به منظور تولید خودکار شرح بر تصاویر . . ۲۴

۱ فصل اول

مقدمات

۱.۱ مقدمه

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. دقت 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 این‌صورت 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 مدلِ 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 عموما 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] به‌وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

[۱] Query
[۲] Holistic Image Caption
[۳] Natural Language Sentences
[۴] Machine 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 مبنا، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی گرفته 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 پروژه 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 طوری‌که توصیف جسم مستقر در نقطه توجه و اجسام مرتبط با آن در جملات تولیدی، وجود داشته باشد.

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

شرح کوتاه تولید شده در این پروژه، به معنی تولید جملاتی است که مستقیما 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 دقت شرح دهند.

---

[۵] Source Language

[۶] Feature Vector

[۷] Destination Language

[۸] Attention Points

۲ فصل دوم

درک 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. عموما 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 عمده زیر تقسیم‌بندی نمود:

۱. استفاده از مدل‌های گرافی احتمالی[۹]

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[۱۰] در این روش‌ها با استفاده از مدل‌های گرافی احتمالی انجام می‌شوند. به عنوان نمونه، در مقاله [۱] یک مدل میدان

[۹] Probabilistic Graphical Models (PGMs)
[۱۰] Image Segmentation

تصادفی شرطی[11] برای تجزیه معنایی[12] تصویر ارائه شده است که با استفاده از آن می‌توان در مورد حضور یا عدم حضور اجسام مختلف به طور توام 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 [۲] 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

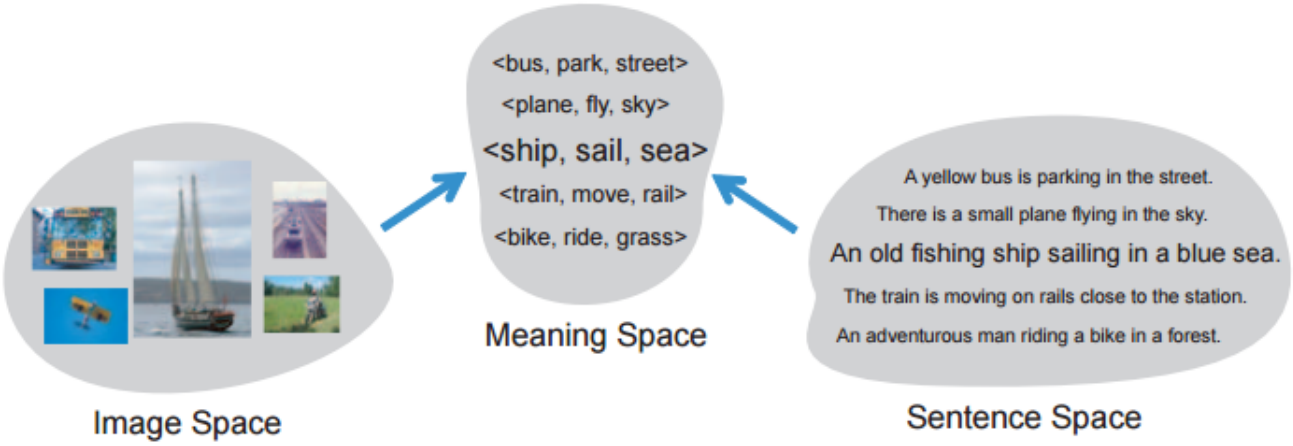
همان‌طور که قبلا ذکر شد، روش‌های مبتنی بر استفاده از مدل‌های گرافی احتمالی، از جمله پرکاربردترین روش‌ها در مرحله درک صحنه در زمینه تولید خودکار شرح بر تصاویر هستند. این روش‌ها با استفاده از نظریه گراف، آمار و احتمالات اقدام به ارائه یک توزیع احتمالی برای پارامتر مورد بررسی، با توجه به داده‌های موجود 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[13]

مقاله [۳] 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»[14] ایجاد می‌شود که بیان‌کننده اطلاعات مطلوب موجود در تصویر است. میدان[15] «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[16] می‌گویند.
شکل ۱ 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

[11] Conditional Random Field (CRF)
[12] Semantic Parsin g
[13] Markov Random Field (MRF)
[14] <Object, Activity, Scene>
[15] Field
[16] Meaning Space

فضای جملات، یک سه‌تایی ایجاد می‌شود به‌طوری‌که جملات و تصاویر متناظرشان، به یک سه‌تایی یکسان، نگاشت شوند. همان‌طور که مشخص است، با داشتن نگاشت‌هایی که خواص مذکور را داشته‌باشند، می‌توان با استفاده از سه‌تایی‌های فضای معنا، تصاویر را مدیریت کرد.



شکل ۱: 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 ۲ 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[۱۷] فلزنسوالب[۱۸]، به منظور محاسبه امتیاز اطمینان[۱۹] برای هر دسته از اجسام موجود در مجموعه‌داده[۴].

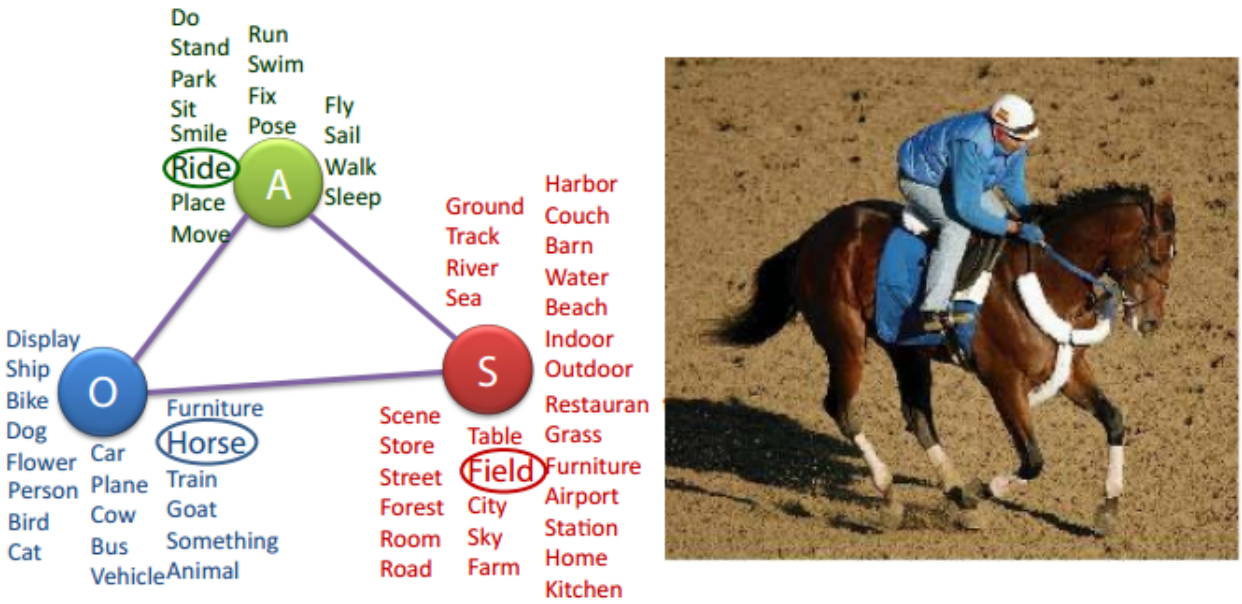
پس از محاسبه امتیاز اطمینان همه دسته‌های موجود، دسته‌ای که بیشترین امتیاز را دارد می‌تواند

[۱۷] Detector
[۱۸] Felzenszwaalb
[۱۹] 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[۲۰]، ارائه شده در مقاله[۵]

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



شکل ۲: طرح‌واره مدل میدان تصادفی مارکف ارائه شده در پژوهش [۳] 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[۳].

بر اساس مقادیر محاسبه شده برای ویژگی‌های بالا و با استفاده از الگوریتم ماشین بردار پشتیبان[۲۱]، 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 محاسبه‌شده 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. فاکتورِ یال

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

[۲۰] divvala

[۲۱] Support Vector Machine (SVM)

۲.۳.۲ استفاده از مدل میدان تصادفی شرطی[۲۲]

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 [۶] و [۷] از مدل میدان تصادفی شرطی به منظور توصیف صحنه استفاده شده است.

پژوهش [۶] سعی در توصیف اجسام سه‌بعدی با استفاده از قطعه‌بندی تصاویر دوبعدی، 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 [۷]، یک چارچوب کاری[۲۳] 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[۱] 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 داده‌شده برای هر جسم

شکل ۳ طرح‌واره مدل سلسله‌مراتبی ارائه شده در پژوهش [۱] 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.

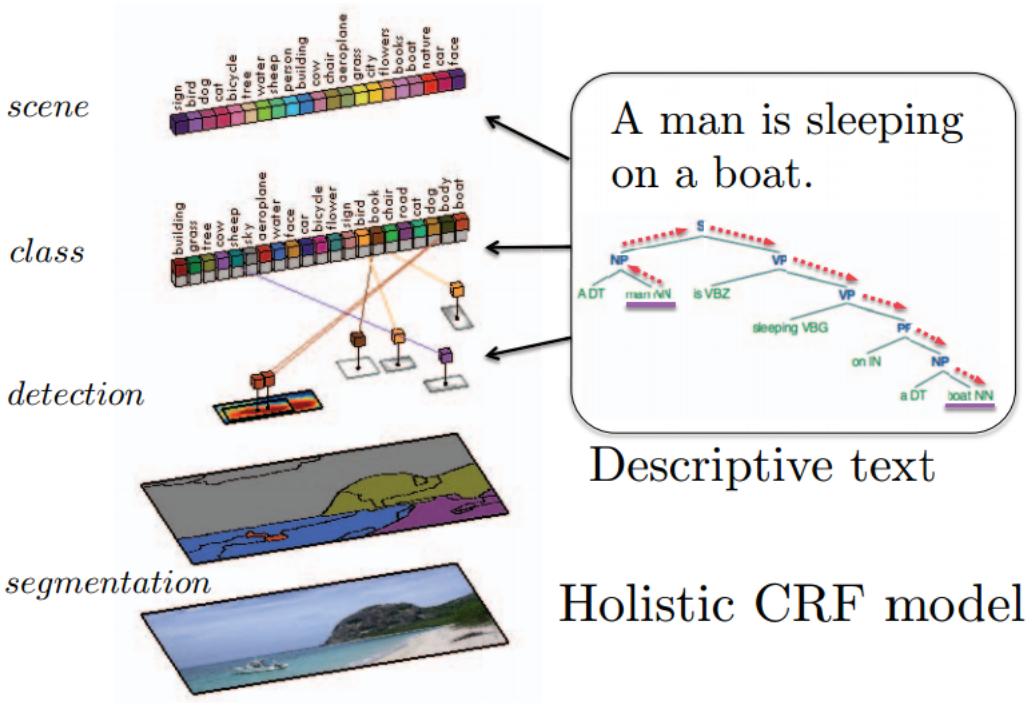
دو دسته متغیر تصادفی مختلف، که هر یک نماینده متغیرهای تصادفی موجود در یکی از این سطوح انتزاع هستند، تعریف شده‌اند؛ متغیرهای تصادفی $X_i \in ۱, \cdots, C$ بیان‌کننده دسته قطعه iام از سطح پایین سلسله مراتب و متغیرهای تصادفی $Y_j \in ۱, \cdots, C$ بیان‌کننده دسته قطعه jام از سطح بالای سلسله مراتب. به علاوه، دو دسته متغیر تصادفی دیگر به نام‌های b_l و z_k به ترتیب برای نمایش حضور یا عدم حضور یک تشخیص کاندید[۲۴] 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 رابطه۱ تعریف کرد. در این رابطه $\Psi_\alpha^{type}(a_\alpha)$ نماینده تابع پتانسیل تعریف شده روی متغیرهای مختلف است. با این تعریف، یافتن تخمین MAP[۲۵]، منجر به یافتن پاسخ مورد نظر می‌شود.

[۲۲] Conditional Random Field (CRF)

[۲۳] Framework

[۲۴] Candidate Detection

[۲۵] MAP Estimation



شکل ۳: طرح‌واره مدل سلسله مراتبی مبتنی بر میدان تصادفی شرطی که بر اساس اطلاعات بصری و اطلاعات جملات توصیف‌کننده شرح محتمل تصویر را تولید می‌نماید[۱].

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.

$$P(X,Y,b,z) = \frac{1}{Z}\prod_{type}\prod_{\alpha}\Psi_{\alpha}^{type}(a_{\alpha}) \qquad (۱)$$

توابع پتانسیل مختلف تعریف شده در این پژوهش عبارتند از:

۱. پتانسیل قطعه‌بندی یگانی[۲۶]

پتانسیل قطعه‌بندی یگانی در هر قطعه و هر ابرقطعه[۲۷] از تصویر، با استفاده از میانگین‌گیری روی امتیاز افزایش تکستون[۲۸] که در پژوهش [۸] 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 رابطه۲ تعریف می‌شود.

$$\phi_{ij}(X_i,Y_j) = \begin{cases} -\gamma & X_i \neq Y_j \\ ۰ & X_i = Y_j \end{cases} \qquad (۲)$$

[۲۶] Unary Segmentation Potential
[۲۷] Supersegment
[۲۸] Texton Boost

در رابطه۲، پارامتر γ در فرآیند یادگیری که منجر به بهینه‌سازی پارامترهای مختلف مدل می‌شود، به‌دست 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[۹] توسط فلزنسوالب که به روش دی پی ام[۲۹] مشهور است، استفاده شده است. برای کاهش تعداد پارامترها و افزایش کارایی مدل استفاده شده، برای هر تصویر حداکثر ۳ دسته جسم، به عنوان دسته‌های منتخب کاندید، در نظر گرفته می‌شوند.

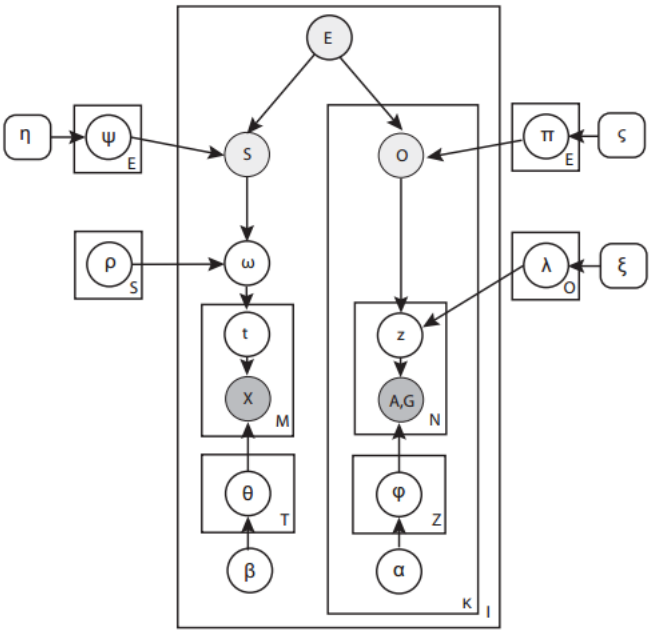
۳.۳.۲ استفاده از سایر مدل‌های گرافی احتمالی

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 توام[۱۰]

مدل استفاده شده در این پژوهش، از تصاویر در سطح صحنه و سطح اجسام استفاده کرده و با یکپارچه‌سازی و تجمیع اطلاعات موجود در این دو سطح، اقدام به دسته‌بندی تصویر می‌نماید. شکل۴ مدل استفاده شده در این پژوهش را به منظور یکپارچه‌سازی و تجمیع اطلاعات حاصل از تحلیل صحنه و تشخیص اجسام موجود در آن، ارائه می‌دهد.

[۲۹] DPM



شکل ۴: مدل استفاده شده به منظور تجمیع اطلاعات صحنه و اجسام موجود در آن به منظور دسته‌بندی تصاویر[۱۰]

یکی از اهدافی که در این پژوهش دنبال می‌شود، برچسب‌گذاری معنایی [۳۰] تمام پیکسل‌های موجود در تصویر است. به همین منظور، تمام تصاویر مورد استفاده، به نواحی ۱۰ * ۱۰ 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 به‌کار می‌رود. اطلاعات این متغیر با استفاده از توصیف‌کننده سیفت [۳۱] و به ازای هر ناحیه از تصویر، به‌دست 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 [۳۲] هستند. گره‌های خاکستری روشن‌تر، متغیرهایی هستند که فقط در فرایند آموزش دیده می‌شوند در حالی‌که متغیرهای تیره‌تر در هر دو فرایند آموزش و آزمون مشاهده می‌شوند.

متغیر تصادفی E، نماینده یک دسته از رخداد [۳۳] های ممکن است. توزیع احتمال اولیه این متغیر تصادفی،

[۳۰] Semantic Labelling
[۳۱] SIFT Descriptor
[۳۲] Latent
[۳۳] Event

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 دانستن دسته رخداد موجود در تصویر، یک تصویر صحنه[۳۴] 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. ψ یک پارامتر چندجمله‌ای [۳۵] حاکم بر توزیع احتمالاتی S به شرط داشتن E است. به علاوه، ψ یک ماتریس به ابعاد $E * S$ و پارامتر η یک بردار S بعدی در نقش مقدار اولیه دیریکله [۳۶] برای پارمتر ψ است.
* در قدم بعدی با داشتن مقدار S، پارامترهای ω را بر اساس احتمال $P(\omega|S,\rho)$ تولید می‌کنیم. از 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 tها 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[۳۷] 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[۳۸] 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 $E * O$ 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:

[۳۴] Scene Image
[۳۵] Multinomial
[۳۶] Dirichlet prior
[۳۷] Topic
[۳۸] Object Image

- متغیر تصادفی مخفی z که مشخص کننده موضوع است، از توزیع احتمالی $Mutl(\lambda, |O)$ 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 $Z * V_o$ 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.

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

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.

$$P(S|E, \psi) = Mult(S|E, \psi) \quad (۴)$$

$$P(\omega|S, \rho) = Dir(\omega|\rho_{j.}), S = j \quad (۵)$$

$$P(t_m|\omega) = Mult(t_m|\omega) \quad (۶)$$

$$P(X_m|t_m, \theta) = P(X_m|\theta_{j.}), t_m = j \quad (۷)$$

$$P(O_k|E, \pi) = Mult(O_k|E, \pi) \quad (۸)$$

$$P(z_n|\lambda, O_k) = Mult(z_n|\lambda, O_k) \quad (۹)$$

$$P(A_n, G_n|z_n, \phi) = P(A_n, G_n|\phi_{j.}), z_n = j \quad (۱۰)$$

درک 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[۳۹]، می‌تواند

[۳۹] Maximum Likelihood

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 نمایش‌گذاشته شده در تصویر، مطابق با رابطه ۱۳ محاسبه می‌شوند.

$$P(I|O) = \prod_{n=1}^{N} \sum_j P(A_n, G_n|z_j, O)P(z_j|O) \quad (۱۱)$$

$$P(I|S, \rho, \theta) = \int P(\omega|\rho, S)(\prod_{m=1}^{M} \sum_{t_m} P(t_m|\omega)P(X_m|t_m, \theta))d\omega \quad (۱۲)$$

$$P(I|E) \propto \sum_j P(I|O_j)P(O_j|E)P(I|S)P(S|E) \quad (۱۳)$$

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[۴۰] و برای سه پارامتر $\{\pi, \lambda, \beta\}$ با استفاده از نمونه‌برداری گیبس[۴۱] انجام می‌شود.

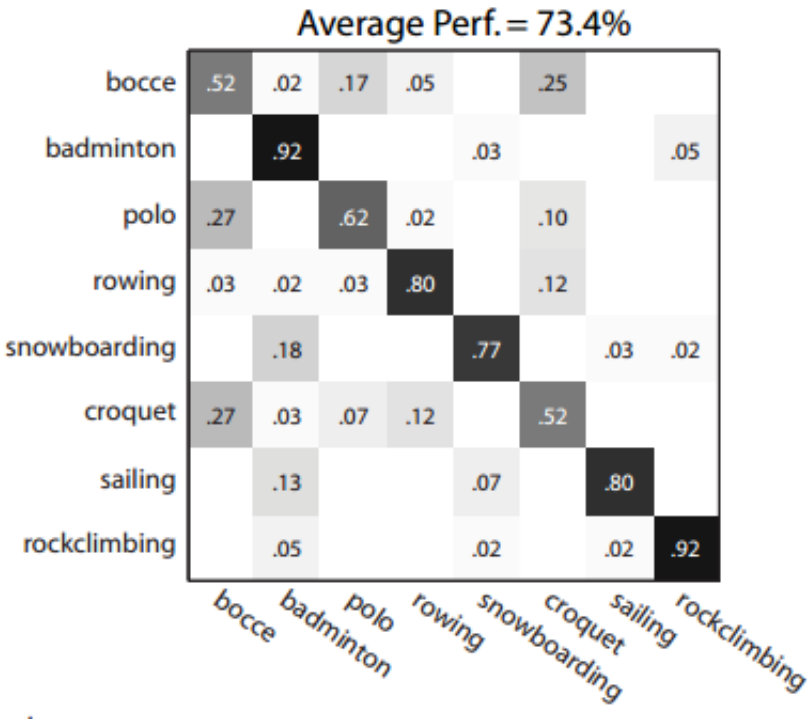
آزمایشات انجام شده در این پژوهش، بر روی یک مجموعه‌داده شامل تصاویر از ۸ دسته ورزشی مختلف که در هر دسته، بین ۱۳۷ تا ۲۵۰ تصویر مختلف وجود دارد، انجام شده‌اند. از جمله چالش‌های موجود در این مجموعه‌داده می‌توان به وجود زمینه‌های متنوع و پیچیده در تصاویر، تنوع دسته‌های مختلف اجسام موجود، تنوع اندازه اجسام 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 ۵ نمونه‌ای از تصاویر موجود در این مجموعه‌داده را نمایش می‌دهد.

[۴۰] Variational Message Passing

[۴۱] Gibbs Sampling

شکل ۵: نمونه تصاویر موجود در مجموعه‌داده مورد استفاده. [۱۰]

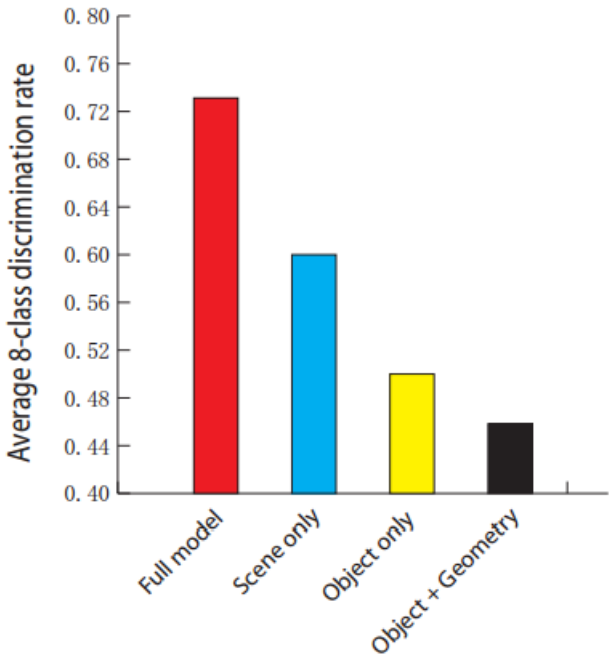
استفاده از مدل کامل ارائه شده در این پژوهش، منجر به تشخیص صحیح ٪۷۳.۴ از تصاویر شده است. شکل ۶ ماتریس درهم‌ریختگی [۴۲] مربوط به این مدل را نمایش می‌دهد. همان‌طور که در این ماتریس مشخص است، کمترین نرخ تشخیص در بین دسته‌های ورزشی موجود در این مدل، ٪۵۲ و بیشترین نرخ تشخیص ٪۹۲ است.



شکل ۶: ماتریس درهم‌ریختگی مدل کامل ارائه شده برای مجموعه‌داده شامل ۸ دسته تصویر ورزشی. [۱۰]

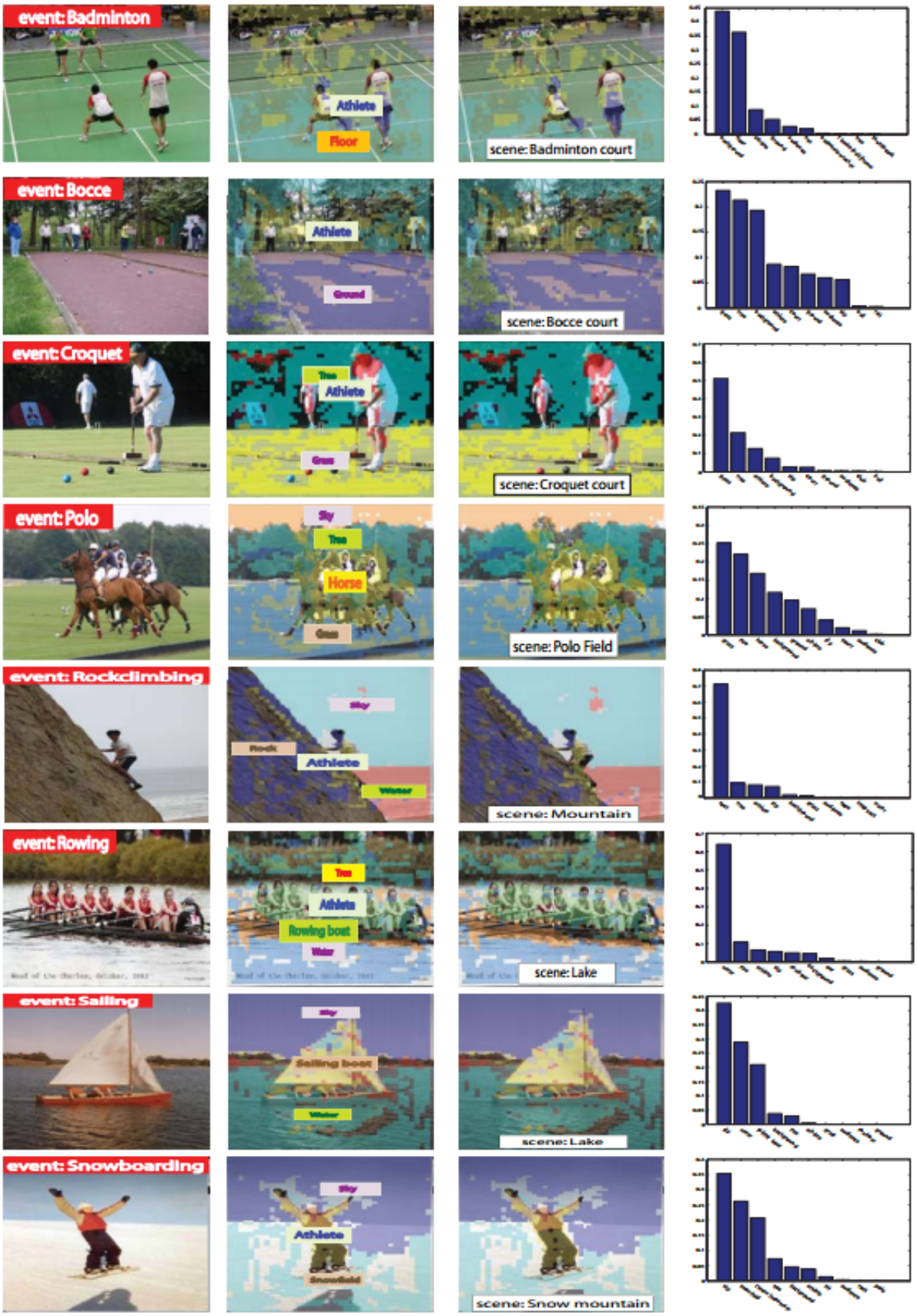
بسته به میزان استفاده از اطلاعات مختلف استخراج شده برای استنتاج، مدل‌های مختلفی به‌وجود می‌آیند که در شکل ۷ نتایج عملکرد هریک از این مدل‌ها با مدل‌های دیگر مقایسه شده است. همان‌طور که در شکل ۷ مشخص است، بهترین کارایی مربوط به مدل کامل است. در صورتی‌که در مدل، فقط از اطلاعات مربوط به صحنه استفاده شود، نتایج بدست‌آمده اگرچه با نتایج مدل کامل قابل مقایسه نیست، از نتایج مدل مبتنی بر اطلاعات جسم بهتر است.

[۴۲] Confusion Matrix



شکل ۷: نتیجه مقایسه مدل‌های مختلف به‌وجود آمده بسته به سطح اطلاعات مورد استفاده برای استنتاج. [۱۰]

شکل ۸ 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 برچسب‌های تشخیص داده شده مربوط به اجسام موجود، ستون سوم برچسب اختصاص داده‌شده 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 ۸: نتایج نهایی به‌دست آمده از مدل بر روی تصاویر. [۱۰]

۲. حاشیه‌نویسی تصویر[۴۳] با استفاده از قطعه‌بندی و دسته‌بندی صحنه و اجسام موجود[۱۱]

۳. درک صحنه بر اساس نواحی مختلف تصویر، اجسام موجود و روابط سه‌بعدی بین آن‌ها[۱۲]

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 ازتصویر می‌پردازیم.

۱.۴.۲ اختصاص معنا به قطعه‌های مختلف تصویر[۱۳]

در پژوهش [۱۳] 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 دیده‌شده 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[۴۴]

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[۴۵] استفاده می‌شود. نسخه‌های مختلفی از این روش ارائه شده است. نسخه ارائه شده در پژوهش [۱۴]، یکی

[۴۳] Image Annotation
[۴۴] Category-independent region proposals
[۴۵] Selective Search

از سریع‌ترین نسخه‌های ارائه شده است که در این بخش از همین روش استفاده می‌شود.
در پژوهش [۱۴] 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 عموما با روش‌های سلسله‌مراتبی قابل دست‌یابی است. ویژگی دوم این است که نواحی مختلف باید براساس ویژگی‌های مختلفی تولید شوند. در صورتی‌که 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 شباهت بین نواحی همسایه، ناحیه‌هایی که شباهت 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 به‌دست می‌آید.برای دست‌یابی به ویژگی دوم، از فضاهای رنگی مختلف، معیارهای شباهت مختلف و نواحی اولیه متفاوت و ترکیب پاسخ این ویژگی‌ها با هم برای ارائه نواحی و ترکیب نواحی کوچک‌تر استفاده می‌شود.

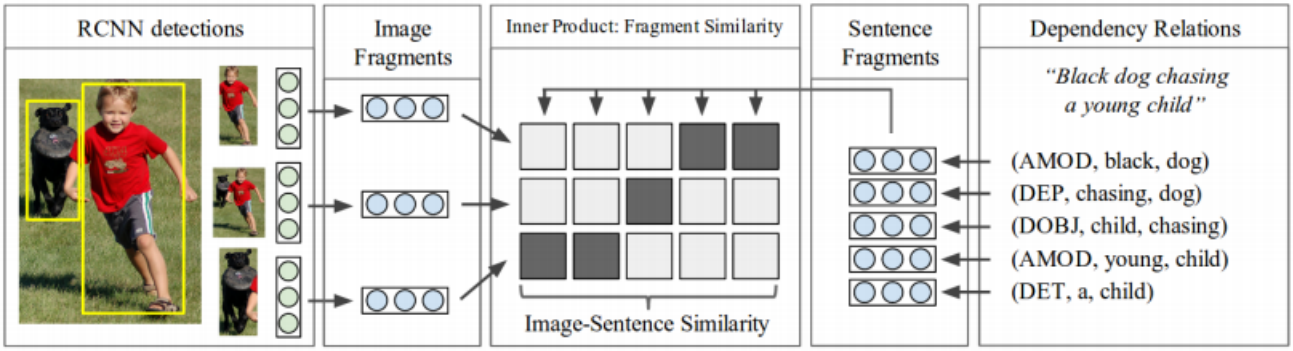
۲. شبکه عصبی کانولوشنی عمیق (استخراج ویژگی‌ها)
در این بخش از یک شبکه عصبی کانولوشنی عمیق از پیش‌آموزش‌دیده برای استخراج ویژگی از هر ناحیه ارائه شده در قسمت قبل، استفاده می‌شود. بردار ویژگی استخراج شده برای هر ناحیه یک بردار شامل ۴۰۹۶ مولفه است که خروجی شبکه کریشفسکی[۴۶] آزمایش شده در چالش دسته‌بندی اجسام مسابقه ImageNet است. اطلاعات دقیق درباره این شبکه عصبی در پژوهش [۱۵] در دسترس است.

شبکه عصبی کانولوشنی عمیق ارائه شده در این پژوهش با استفاده از یک مجموعه‌داده[۴۷] آموزش دیده شده است. از این شبکه عصبی که تحت عنوان RCNN[۴۸] شناخته می‌شود می‌توان به عنوان یک شبکه از پیش‌آموزش‌دیده استفاده کرد.

۲.۴.۲ ناحیه‌بندی عمیق تصاویر به منظور نگاشت دوطرفه جملات و تصاویر[۱۶]

مدل ارائه شده در این پژوهش، مدلی است که قادر به نگاشت دوطرفه تصاویر و جملات به یک‌دیگر است. شکل۹ 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 شباهت محاسبه می‌شود و شبیه‌ترین قطعه جمله با ناحیه مربوط به خود در تصویر، جفت می‌شوند.

[۴۶] Krizhevsky
[۴۷] ILSVRC 2012
[۴۸] Regional Convolutional Neural Network



شکل ۹: مدل استفاده شده برای نگاشت دوطرفه تصاویر و جملات به یک‌دیگر با استفاده از شکبه عصبی عمیق کانولوشنی.[۱۶]

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 رابطه۱۴ 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 ν_i برای تصویر iام، 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.

$$\nu = W_m[RCNN_{\theta_c}(I_b)] + b_m \qquad (۱۴)$$

از طرفی با در نظر گرفتن بردار s_j به عنوان بردار حاصل از نگاشت جمله jام به فضای معنایی و در نظر گرفتن ضرب داخلی به عنوان شباهت، $\nu_i^T \cdot s_j$ 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[۴۹]

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.

[۴۹] Fragment Alignment

با توجه به مطالب گفته شده، می‌توان تابع هدف کلی را مطابق با رابطه۱۵ 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[۵۰] (با آزمون و خطا تعیین می‌شوند) و $||\Theta||_۲^۲$ یک عبارت تنظیم‌کننده[۵۱] هستند.

$$C(\Theta) = C_F(\Theta) + \beta C_G(\Theta) + \alpha||\Theta||_۲^۲ \quad (۱۵)$$

در ادامه به تعریف هریک از اهداف بیان‌شده 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 صورتی‌که 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 واقه 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 صورتی‌که 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.
رابطه۱۷ تابع هدف هم‌ترازسازی ناحیه‌ای را تعریف می‌کند. در این رابطه، y_{ij} برای تصویر iام و جمله jام در صورتی‌که با هم در مجموعه‌داده حضور داشته باشند، +۱ و در غیر این صورت، -۱ خواهد شد.

$$C_۰(\Theta) = \sum_i \sum_j max(۰, ۱ - y_{ij} \nu_i^T \cdot s_j) \quad (۱۶)$$

تابع $C_۰$ تعریف شده، باعث می‌شود در حالاتی که تصویر و عبارت، در مجموعه‌داده، با یک‌دیگر وارد شده باشند امتیاز تابع هدف بیشتر از +۱ شود و در غیر این‌صورت از -۱ کمتر شود. شکل۱۰، دو نمونه از تصاویر و جملات موجود در مجموعه‌داده را نمایش می‌دهد. $C_۰$ در سلول‌هایی که با رنگ قرمز مشخص شده‌اند، امتیاز را به سمت کمتر از -۱ حرکت می‌دهد و در بقیه سلول‌ها به سمت بیشتر از +۱.

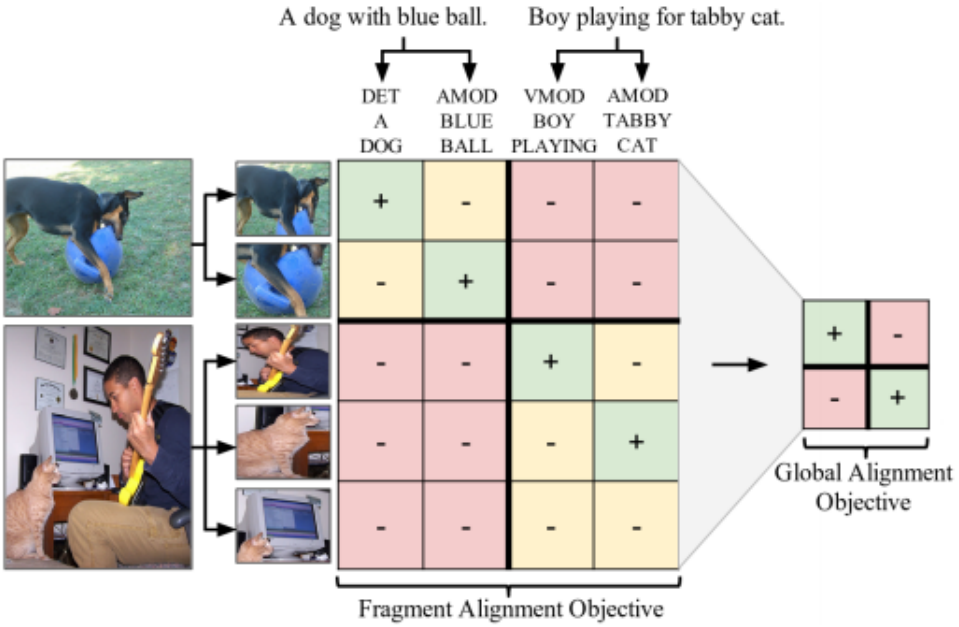
به عبارت بهتر، $C_۰$ یک امتیاز برای مجموع تفاوت‌های نواحی مختلف از تصاویر با عبارات مختلف جملات است. به دلیل این‌که این معیار، باعث دیده نشدن موارد کم‌یاب می‌شود، با متغیر گرفتن پارامتر y_{ij} سعی

[۵۰] Hyperparameter
[۵۱] Regularization Term

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

$$C_F(\Theta) = min_{y_{ij}} C_\circ(\Theta)$$
$$s.t. \sum_{i \in p_j} \frac{y_{ij} + ۱}{۲} \geq ۱ \wedge \qquad y_{ij} = -۱, \forall i, j; m_\nu(i) \neq m_s(j) \wedge y_{ij} \in \{+۱, -۱\} \qquad (۱۷)$$

در این رابطه، p_j مجموعه تصاویر موجود در کیسه مثبت[۵۲] مربوط به عبارت jام است. شایان ذکر است، تنها تصاویری که در مجموعه‌داده همراه با عبارت jام مشاهده شده‌اند در کیسه مثبت مربوط به این عبارت قرار می‌گیرند و بقیه تصاویر در کیسه منفی[۵۳] این عبارت قرار می‌گیرند. $m_\nu(i)$ و $m_s(j)$ به ترتیب، شماره تصویر و عبارت را در مجموعه‌داده مشخص می‌کنند.



شکل ۱۰: دو نمونه از تصاویر و جملات مرتبط با آن‌ها و نتایج عملکرد اهداف تعریف‌شده روی آن‌ها. سطرها نمایش‌دهنده نواحی مختلف تصویر و ستون‌ها نمایش دهنده قطعه‌های مختلف جملات هستند. سلول‌های قرمز رنگ حالاتی هستند که در آن‌ها $y_{ij} = -۱$، سلول‌های زرد نمایش‌دهنده اعضای کیسه‌های مثبت هستند که در آن‌ها $y_{ij} = -۱$ است. [۱۶]

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

[۵۲] Positive Bag
[۵۳] Negative Bag

شباهت بین یک تصویر و یک جمله تعریف می‌شود. این امتیاز مطابق با رابطه۱۸ تعریف شده و برابر است با میانگین امتیاز شباهت دوبه‌دوی نواحی مختلف تصویر با عبارات مختلف جمله.

$$S_{kl} = \frac{۱}{|g_k|(|g_l| + n)} \sum_{i \in g_k} \sum_{j \in g_l} max(۰, \nu_i^T \cdot s_j) \qquad (۱۸)$$

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، $n = ۵$ است. تابع کلی هدف سراسری مطابق با رابطه ۱۹ تعریف می‌شود.

$$C_G(\Theta) = \sum_k (\sum_l max(۰, S_{kl} - Skk + \Delta) + \sum_l max(۰, S_{lk} - Skk + \Delta)) \qquad (۱۹)$$

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.

شکل۱۱ 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 ۱۱: 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.[۱۶]

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. شکل۱۲ 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 ۱۲: 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.[۱۶]

۳.۴.۲ هم‌ترازسازی[۵۴] اطلاعات بصری و معنایی به منظور تولید خودکار شرح بر تصاویر[۱۷]

در پژوهش [۱۷] عملیات تولید خودکار شرح برای تصاویر به‌طور کل با استفاده از شبکه‌های عصبی انجام شده‌است. همان‌طور که گفته شد، در این بخش به بررسی درک صحنه در این پژوهش می‌پردازیم.

درک صحنه در این پژوهش، با استفاده از به‌کارگیری یک شبکه عصبی کانولوشنی عمیق انجام شده است. این عملیات با انتساب قطعات کوچک جملات به بخش‌های تصویر صورت می‌گیرد. برای این منظور، برای هر قطعه از یک تصویر یک عبارت زبانی توصیف‌کننده قطعه تولید می‌شود. سپس در مرحله دوم با داشتن عبارات توصیف‌کننده قطعات مختلف یک تصویر، عملیات ساخت جمله انجام می‌شود.

ورودی مدل در این قسمت، مجموعه‌ای از تصاویر و شرح تولید شده توسط عوامل انسانی است که در مجموعه‌داده وجود دارد. نکته‌ای که در این بخش حائز اهمیت است، این است که در صورتی‌که یک تصویر به تعدادی از کاربران نمایش داده‌شود و از کاربران خواسته شود که بهترین شرح برای تصویر را بنویسند (بدون این‌که کاربران با یک‌دیگر در ارتباط باشند یا از شرح تولید شده دیگران خبر داشته باشند) بخش‌های یکسانی در شرح‌های تولید شده کاربران وجود خواهد داشت که به نواحی خاصی از تصویر مربوط هستند که موقعیت آن‌ها برای ما نامعلوم است. برای مثال اگر تصویر از یک ریل راه‌آهن و یک قطار باشد، در شرح‌های تولید شده توسط کاربران، عباراتی بیان‌کننده این دو مفهوم وجود خواهند داشت که تکرار آن‌ها در بین شرح‌های موجود از عبارات دیگر بیشتر است. همین موضوع، پایه‌ای برای هم‌ترازسازی در این بخش است. یافتن رابطه مخفی بین عبارات در جملات و نواحی مختلف تصویر منجر به ارائه مدلی برای انتساب این عبارات و نواحی تصویر به یک‌دیگر می‌شود.

[۵۴] Alignment

[۱۶]

مراجع

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Ladickỳ, 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.