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 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

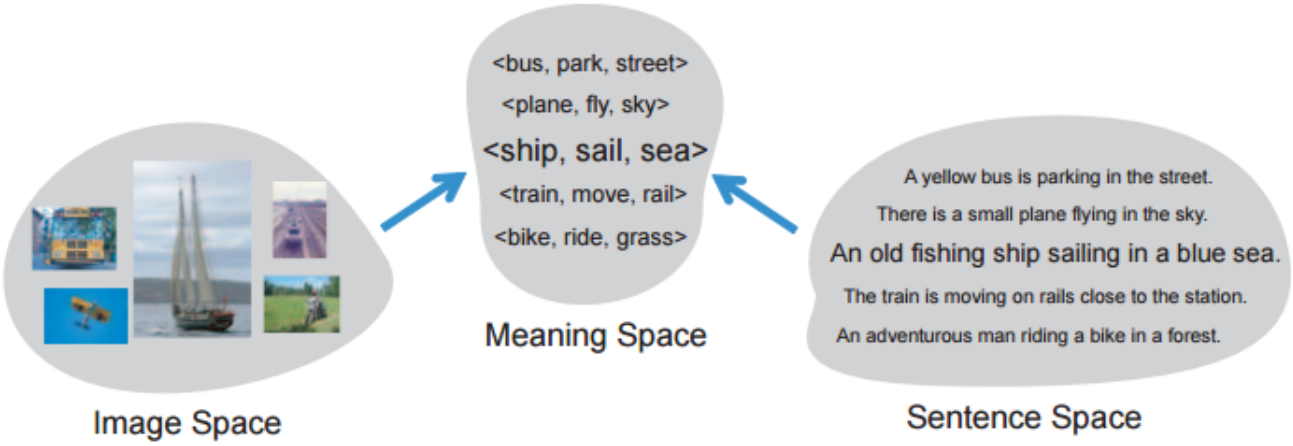
همان‌طور که قبلا ذکر شد، روش‌های مبتنی بر استفاده از مدل‌های گرافی احتمالی، از جمله پرکاربردترین روش‌ها در مرحله درک صحنه در زمینه تولید خودکار شرح بر تصاویر هستند. این روش‌ها با استفاده از نظریه گراف، آمار و احتمالات اقدام به ارائه یک توزیع احتمالی برای پارامتر مورد بررسی، با توجه به داده‌های موجود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

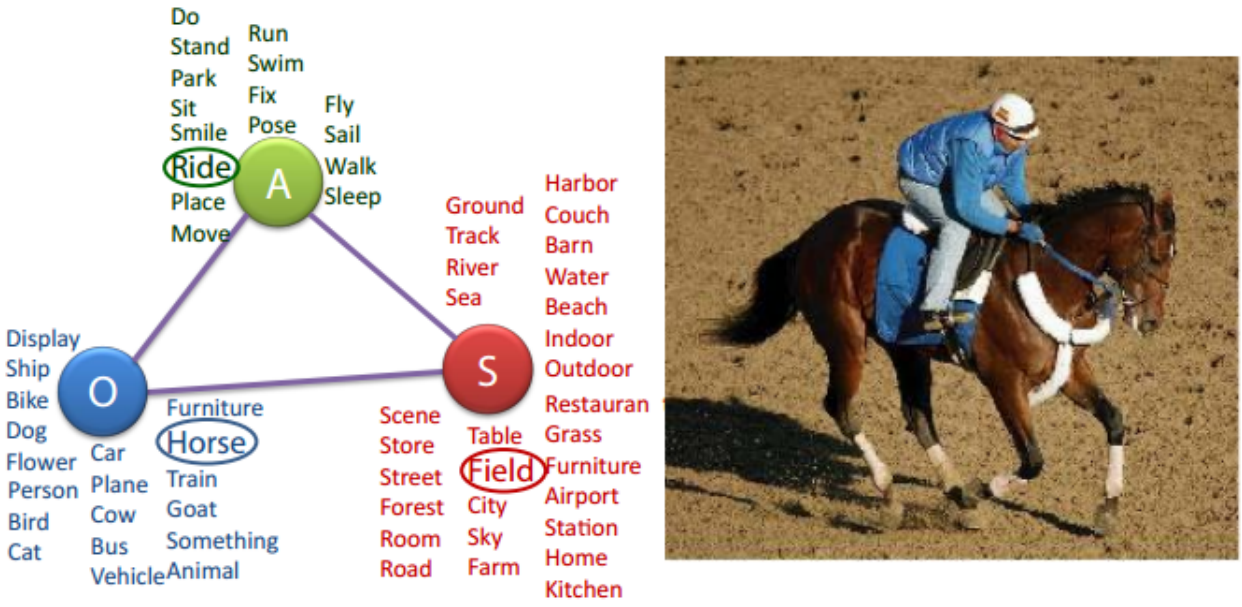
[11] Conditional Random Field (CRF)
[12] Semantic Parsin g
[13] Markov Random Field (MRF)
[14] <Object, Activity, Scene>
[15] Field
[16] Meaning Space

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.



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].

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 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[؟]

[۱۷] Detector
[۱۸] Felzenszwaalb
[۱۹] Confidence Score
[۲۰] divvala

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.

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

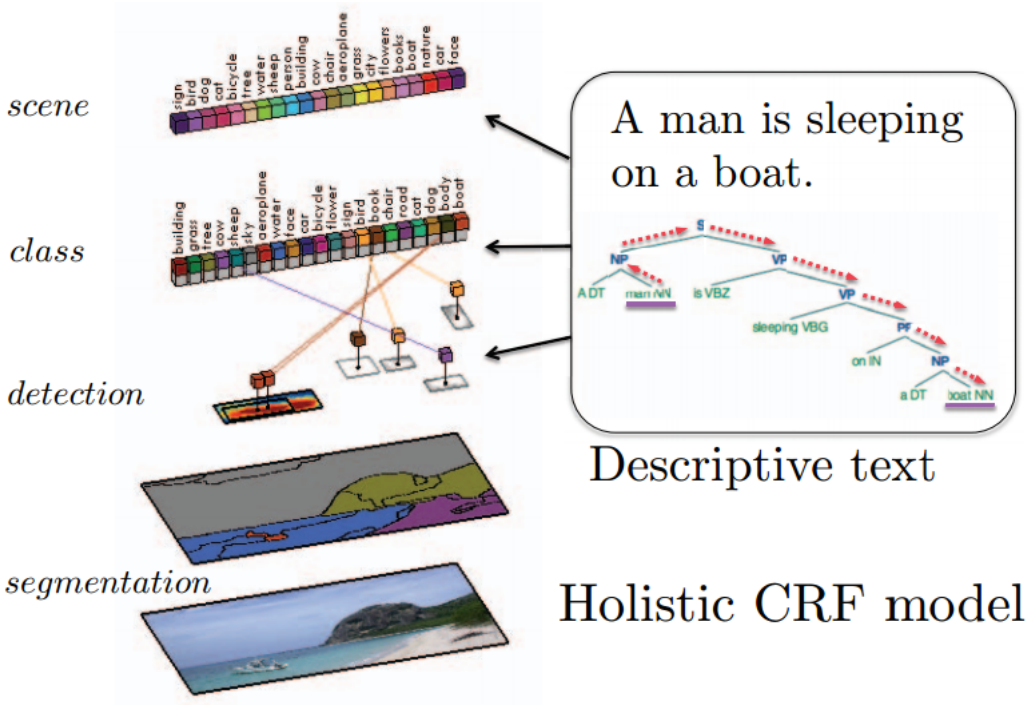
[۲۱] Support Vector Machine (SVM)

[۲۲] Conditional Random Field (CRF)

[۲۳] Framework

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.



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].

۴.۲ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق

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.