

# فهرست مطالب

|    |   |       |
|----|---|-------|
| ۱  | فصل اول مقدمات                                | ۱     |
| ۱۰ | مقدمه   | ۱.۵   |
| ۱۱ | تعریف مساله                                   | ۲.۵   |
| ۱۲ | فصل دوم درک صحنه                              | ۶     |
| ۱۲ | درک صحنه                                      | ۱.۶   |
| ۱۲ | روش‌های مختلف موجود                           | ۲.۶   |
| ۱۳ | روش‌های مبتنی بر مدل‌های گراف‌های احتمالی     | ۳.۶   |
| ۱۳ | استفاده از مدل میدان تصادفی مارکوف            | ۱.۳.۶ |
| ۱۵ | روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق | ۴.۶   |
| ۱۶ | فصل سوم تولید شرح متناظر صحنه                 | ۷     |
| ۱۰ | مقدمه   | ۱.۵   |
| ۱۱ | تعریف مساله                                   | ۲.۵   |
| ۱۲ | فصل چهارم آزمون و ارزیابی                     | ۶     |
| ۱۰ | مقدمه   | ۱.۵   |
| ۱۱ | تعریف مساله                                   | ۲.۵   |
| ۱۲ | فصل پنجم جمع‌بندی و نتیجه‌گیری                | ۶     |
| ۱۰ | مقدمه   | ۱.۵   |
| ۱۱ | تعریف مساله                                   | ۲.۵   |

# ۱ فصل اول

## مقدمات

## ۶ فصل دوم

### درک صحنه

## ۱.۶ 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 راه رفتن، خوابیدن)

## ۲.۶ 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<sup>۴۰</sup>

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<sup>۴۱</sup> در این روش‌ها با استفاده از مدل‌های گرافی احتمالی انجام می‌شوند. به عنوان نمونه، در مقاله [۱] یک مدل میدان

<sup>۴۰</sup>Probabilistic Graphical Models (PGMs)

<sup>۴۱</sup>Image Segmentation

تصادفی شرطی<sup>۴۲</sup> برای تجزیه معنایی<sup>۴۳</sup> تصویر ارائه شده است که با استفاده از آن می‌توان در مورد حضور یا عدم حضور اجسام مختلف به طور توأم 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.

## ۳.۶ 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.

### ۱.۳.۶ استفاده از مدل میدان تصادفی مارکف<sup>۴۴</sup>

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»<sup>۴۵</sup> ایجاد می‌شود که بیان‌کننده اطلاعات مطلوب موجود در تصویر است. میدان<sup>۴۶</sup> «جسم»، دربردارنده برچسب حاصل از دسته‌بندی اجسام موجود در صحنه، میدان «فعالیت»، دربردارنده 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<sup>۴۷</sup> می‌گویند.

شکل ۲ 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

<sup>۴۲</sup>Conditional Random Field (CRF)

<sup>۴۳</sup>Semantic Parsin g

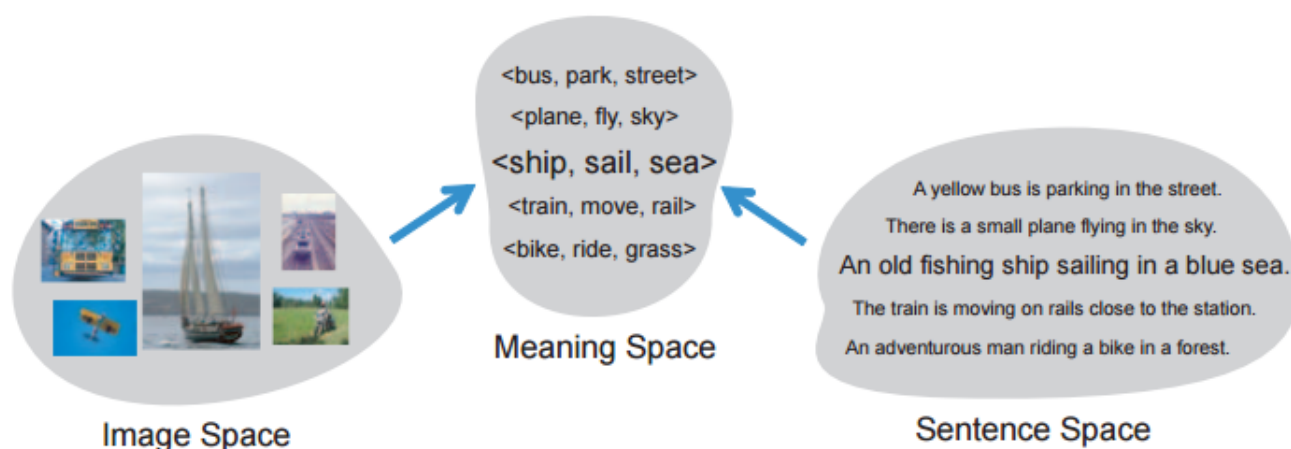
<sup>۴۴</sup>Markov Random Field (MRF)

<sup>۴۵</sup><Object, Activity, Scene>

<sup>۴۶</sup>Field

<sup>۴۷</sup>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 ۲: 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 [۴].

مدل میدان تصادفی مارکف مورد استفاده در این پژوهش، یک مدل کوچک و ساده، شامل ۳ گره است. شکل ۳ 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. فاکتور گره

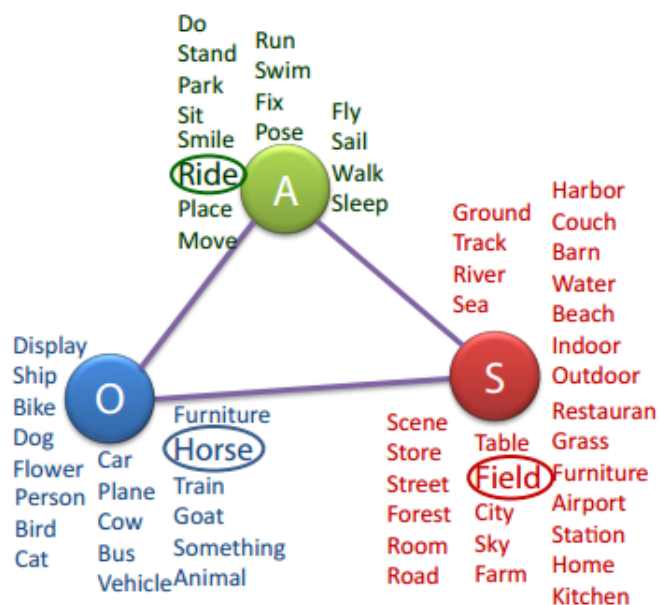
این فاکتور، برای مشخص کردن میزان شباهت مقادیر مختلف گره با تصویر ورودی، تعریف شده‌اند. سه ویژگی‌های مورد استفاده برای مقداردهی این فاکتورها، شامل موارد زیر هستند:

(آ) استفاده از آشکارکننده‌های<sup>۴۸</sup> ارائه شده در مقاله [۴]، به منظور محاسبه امتیاز اطمینان<sup>۴۹</sup> برای هر دسته از اجسام موجود در مجموعه داده.

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

<sup>۴۸</sup>Detector

<sup>۴۹</sup>Confidence Score



شکل ۳: 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].

انجام محاسبات، اطمینان حاصل می‌شود که از هر دسته موجود، حداقل یک تصویر در مجموعه داده وجود داشته باشد.

## ۲. فاکتور یال

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

## ۴.۶ روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی عمیق

## ۷ فصل سوم

تولید شرح متناظر صحنه



## ۶ فصل چهارم

### آزمون و ارزیابی

## ۶ فصل پنجم

### جمع‌بندی و نتیجه‌گیری

- [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.