مجله ماشین‌بینایی و پردازش‌تصویر
سال ششم، شماره اول، تابستان ۱۳۹۸

مروری بر سیستم‌های برچسب زنی تصاویر

رویا راد[۱] و منصور جم‌زاد[۲]

چکیده

امروزه با رشد تکنولوژی‌های ثبت و به اشتراک گزاری تصاویر، تعداد تصاویر دیجیتال افزایش چشمگیری یافته است. مدیریت این حجم از داده‌های تصویری به سامانه‌ای کارآمد جهت مرور، دسته‌بندی، جستجو و بازیابی نیاز دارد. سامانه‌های بازیابی تصاویر در نسل‌های جدید یک عبارت معنایی را معمولاً به صورت یک یا چند کلمه کلیدی از کاربر گرفته، به دنبال بازیابی تصاویری با محتویات بصری مرتبط با آن معنا هستند. داشتن مکانیزمی که بتواند به صورت خودکار محتوای یک تصویر را مانند انسان به صورت متنی توصیف کند به کارایی این سامانه‌ها کمک زیادی می‌نماید. برچسب‌زنی خودکار تصاویر یک روش تخصصی برای بیان محتوای تصاویر به صورت کلمات کلیدی یا برچسب است. سامانه‌های برچسب‌زنی خودکار رابطه بین معنای یک متن و ویژگی‌های سطح پایین یک تصویر را با تکنیک‌های یادگیری ماشین بررسی کرده، به صورت خودکار به تصاویر چندین برچسب نسبت می‌دهند تا امکان جستجو و بازیابی مبتنی بر محتوای آن‌ها بهتر فراهم شود.

در این مقاله به بررسی مراحل مختلف پیاده‌سازی یک سامانه برچسب‌زنی خودکار خواهیم پرداخت و کارهای پیشرو را مرور کرده، مشکلات و چالش‌های موجود برای طراحی این سامانه‌ها را خواهیم دید. همچنین به معرفی چند پایگاه داده مناسب جهت بررسی و آزمودن سامانه‌های برچسب‌زنی خودکار خواهیم پرداخت.

کلیدواژه‌ها

برچسب‌زنی خودکار تصاویر، حاشیه‌نویسی، بازیابی تصویر، استخراج ویژگی

۱ مقدمه

با پیشرفت تکنولوژی‌های ثبت تصویر و در دسترس بودن آسان آنها، تعداد تصاویر موجود در فضای اینترنت افزایش قابل ملاحظه‌ای داشته است. ایجاد درک از تصاویر و امکان ساماندهی و جستجوهای موضوعی در بین آن‌ها از ایده‌آل‌هایی است که هنوز راه زیادی تا نهایی شدن در پیش دارد. در سامانه‌های بازیابی تصویر، هدف این است که از یک پایگاه داده شامل تصاویر مختلف، یک مجموعه تصویر مطلوب کاربر بازیابی شده و نمایش داده شود. این کار در طول زمان به سه شکل بازیابی مبتنی بر متن[۱] (TBIR) ، بازیابی مبتنی بر محتوا[۲] (CBIR) و بازیابی مبتنی بر معنا[۳] (SBIR) انجام شده است [۳-۱]. در نسل‌های اولیه، برای بازیابی یک تصویر تنها به متن‌های اطراف آن اتکا شده، از خصوصیات بصری تصویر استفاده چندانی نمی‌شد. به این صورت که در TBIR کاربر موضوع مورد نظر خود را به صورت یک عبارت متنی وارد کرده، این عبارت در بین متون اطراف تصاویر جستجو می‌شد. در نسل‌های دوم یا CBIR از ویژگی‌های بصری

این مقاله در اردیبهشت‌ماه۱۳۹۶دریافت، در بهمن‌ماه بازنگری و در اسفندماه همان سال پذیرفته شد.

[۱] دانشجوی دکترای گرایش هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف.
رایانامه: raad@ce.sharif.edu
[۲] دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف.
رایانامه: jamzad@sharif.edu

نویسنده مسئول: رویا راد

[1] Text Based Image Retrieval
[2] Content Based Image Retrieval
[3] Semantic Based Image Retrieval

بهره گرفته می‌شود و کاربر یک تصویر نمونه وارد کرده، به دنبال تصاویری می گردد که محتوای بصری مشابه با آن تصویر داشته باشند. در هر دو نسل سامانه‌های بازیابی فوق‌الذکر ، جستجو بر اساس مقایسه بین دو نوع همگون صورت می‌گیرد: مقایسه بین دو متن در TBIR یا مقایسه بین محتوای بصری دو تصویر در CBIR.

در نسل جدید سامانه‌های بازیابی تصاویر، SBIR، کاربر معنای مورد نظر خود را به صورت یک عبارت متنی وارد می‌کند و به دنبال یافتن تصاویری با محتویات مرتبط با آن عبارت است. بنابراین در SBIR رابطه‌ی بین معنا و محتوای بصری تصاویر بررسی می‌شود. محتوای بصری تصاویر به صورت ویژگی‌های سطح پایینی مانند رنگ و بافت استخراج می‌شوند و معنا به صورت ویژگی‌های سطح بالایی با کلماتی کلیدی یا برچسب‌ها معرفی می‌گردد. این دو دسته‌ی ویژگی سطح پایین و سطح بالا فاصله زیادی باهم دارند که کشف رابطه‌ی بین آنها کار دشواری است. به این فاصله، فاصله معنایی[1] گفته می‌شود. یک سامانه SBIR معنا را آن گونه که کاربر می‌فهمد، درک نمی‌کند و استفاده از ویژگی‌های سطح پایین نمی‌توانند به تنهایی منجر به استخراج مفاهیم سطح بالای مورد نظر انسان در جستجوها شود [۴, ۵]. به همین منظور تحقیقات گسترده‌ای به منظور کاهش فاصله معنایی با استخراج ویژگی‌های بهتر و برچسب‌زنی به تصاویر صورت گرفته است.

منطقی‌ترین روش برای جستجوی سطح بالای معانی در سامانه‌های SBIR، این است که ابتدا به تصاویر موجود در پایگاه داده برچسب‌های متنی مبتنی بر معنا نسبت داده شود و با مقایسه بین این برچسب‌ها و عبارت مورد جستجو، تصاویر مربوطه بازیابی شوند. به سامانه‌ای که به صورت خودکار حاشیه‌نویسی تصاویر را از طریق نسبت دادن کلمات کلیدی به تصاویر انجام می‌دهد، سامانه‌ی برچسب‌زنی خودکار تصاویر یا AIA[2] گفته می‌شود [۶-۱۰].

۲ آشنایی با برچسب‌زنی تصاویر

برای مدیریت کارا و بازیابی موثر مبتنی بر معنای تصاویر در یک پایگاه تصاویر، معمولاً تعدادی برچسب به هر تصویر ضمیمه می‌شود. این برچسب‌ها می‌توانند به صورت دستی یا خودکار استخراج گردند. برچسب‌زنی به صورت دستی و توسط کاربران علی‌رغم دقت بهتر آن هزینه زیادی داشته، عملاً برای تعداد بسیار زیاد تصاویر امکان پذیر نیست. همچنین این برچسب‌ها وابسته به کاربر هستند و از یکنواختی کمتری برخوردار هستند. برای بهتر شدن نتایج برچسب‌زنی دستی، پژوهش‌های خوبی در زمینه جمعیت سپاری انجام شده است [۱۱]. در این پژوهش‌ها معمولاً سعی می‌شود با طراحی بازی‌های رایانه ای جذاب و ایجاد تعامل بین چندین کاربر و یادگیری توانایی‌های هر کاربر، دقت برچسب‌ها را بالاتر ببرند.

امروزه برخی از شبکه‌های اجتماعی یا پایگاه‌های داده تصویری در کنار تصاویری که کاربران بارگذاری می‌کنند، از آنها می‌خواهند تا برچسب‌هایی را نیز ضمیمه تصاویر خود نمایند، یا برچسب‌هایی پیشنهاد می‌دهند و از کاربران می‌خواهند از بین آنها انتخاب کنند. به علت مشکلات عدم مقیاس‌پذیری در برچسب‌زنی دستی، تمایل به سمت برچسب‌زنی خودکار روز به روز بیشتر می‌شود. هدف از سامانه‌های برچسب‌زنی خودکار، تسهیل فرایند جستجو در یک پایگاه تصاویر با استفاده از برچسب‌ها است.

برچسب‌زنی می‌تواند در سطح تصویر یا در سطح ناحیه [۱۲-۱۴] انجام شود. در برچسب‌زنی در سطح تصویر برچسب‌ها به صورت کلی نسبت داده می‌شوند و مشخص نمی‌شود هر برچسب مربوط به کدام قسمت از یک تصویر است. در برچسب‌زنی در سطح ناحیه، علاوه بر برچسب‌زدن به تصاویر، ارتباط هر برچسب با نواحی آن تصویر نیز مشخص می‌شود. این امر کمک می‌کند تا در گام آموزش برای هر برچسب، تنها بر روی نواحی مربوطه تمرکز شود و با در نظر نگرفتن اشیاء نامربوط، یادگیری عمیق‌تر مفاهیم امکان پذیر گردد. در حال حاضر به علت مشکلاتی که در الگوریتم‌های ناحیه‌بندی[3] و تشخیص اشیا وجود دارد، بیشتر پژوهش‌های مربوط به حوزه‌ی برچسب‌زنی در سطح تصویر کار می‌کنند.

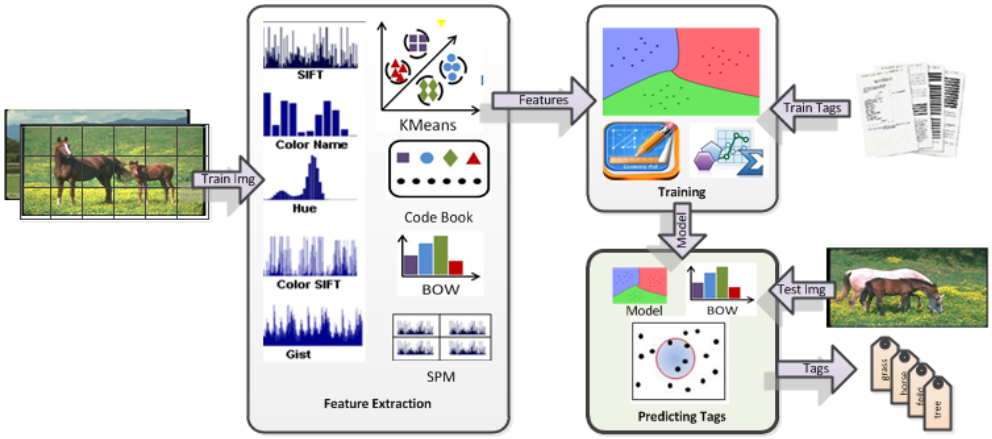
علاوه بر سامانه‌های AIA در حالت پایه که صرفاً تعدادی برچسب برای تصاویر بدون برچسب پیشنهاد می‌دهند، پژوهش‌هایی نیز جهت بهبود برچسب‌ها[4] [۱۵, ۱۶]، کامل کردن برچسب‌ها [۱۷, ۱۸]، رتبه بندی برچسب‌ها [۱۹, ۲۰]، محدوده گزینی برچسب‌ها[5] [۲۱] و برچسب زدن به صورت جمله انجام شده است. در این مقاله تمرکز تنها بر روی AIA در حالت پایه است.

۳ مراحل کار

برچسب‌زنی خودکار یکی از کاربردهای یادگیری ماشین محسوب می‌شود و از این لحاظ مانند بسیاری از کاربردهای دیگر می‌توان مراحل کار را به سه گام اصلی استخراج ویژگی، آموزش و پیش‌بینی تقسیم نمود. در ادامه به معرفی این گام‌ها خواهیم پرداخت.

در شکل ۱ نمایی از فرایند کلی موجود در طراحی سامانه‌های برچسب‌زنی خودکار نمایش داده شده است. ابتدا ویژگی‌های تصاویر استخراج می‌شوند، سپس بر اساس این ویژگی‌ها و برچسب‌های ثبت شده برای هر تصویر، طی فرایند آموزش، یک مدل طراحی می‌شود. در مرحله پیش‌بینی برچسب یا آزمایش، ویژگی‌های تصاویر آزمایشی استخراج شده و به این مدل ارائه می‌شود تا برچسب‌هایی برای این تصاویر انتخاب گردد.

[1] Semantic gap
[2] Automatic Image Annotation
[3] Segmentation
[4] Tag refinement
[5] Tag localization



شکل ۱- نمایی از فرایند کلی موجود در سامانه‌های برچسب زنی خودکار تصاویر [۲۲].

۳-۱ استخراج ویژگی

در این مرحله بر اساس نوع الگوریتم انتخابی و خصوصیات پایگاه داده مورد استفاده، یک یا چند ویژگی انتخاب می‌شود. انتخاب ویژگی‌های مناسب، با توجه به نوع تصاویر و نوع معیار شباهت مورد استفاده، از چالش‌های اصلی در سامانه‌های بازیابی محسوب می‌شود. این ویژگی‌ها می‌توانند به صورت سراسری یا محلی انتخاب شوند.

در پردازش سراسری، ویژگی‌ها از کل تصویر استخراج می‌شوند. برای مثال میانگین شدت روشنایی پیکسل‌های تصویر، یک ویژگی سراسری محسوب می‌شود. مزیت ویژگی‌های سراسری، سرعت بالا و بار محاسباتی کمتر است. اما این ویژگی‌ها از درک جزئیات تصویر و اطلاعات مکانی آنها ناتوان هستند. در پردازش محلی، ویژگی‌ها از یک همسایگی از پیکسل‌ها استخراج می‌شوند. در بعضی از حالات تصویر به بلوک‌هایی تقسیم می‌شود، ویژگی‌های هر بلوک به صورت جداگانه استخراج شده در کنار هم بردار ویژگی مربوط به آن تصویر را تشکیل می‌دهد. با این کار در حقیقت از اطلاعات مکانی موجود تصویر نیز استفاده شده است.

پس از استخراج ویژگی‌ها، در مورد نحوه ترکیب آنها تصمیم‌گیری می‌شود. ترکیب می‌تواند به صورت هم-جوشی اولیه[1] و با الحاق بردارهای ویژگی به هم صورت گیرد یا به صورت هم‌جوشی میانی و هم‌جوشی تأخیری[2]، که در آن با هر کدام از ویژگی‌ها به صورت مجزا برخورد کرده و ترکیب را به مراحل بعدی موکول می‌نماید. در ادامه به مرور چند ویژگی پرکاربرد می‌پردازیم.

۳-۱-۱- ویژگی‌های مبتنی بر رنگ

اطلاعات رنگ به خصوص هیستوگرام رنگ به علت مقاوم بودن در مقابل چرخش و انتقال از محبوب‌ترین و پرکاربردترین ویژگی‌های مورد استفاده در بازیابی تصویر است و می‌تواند در فضاهای رنگ مختلفی از جمله RGB ,HSV, Luv, Lab, YCbCr تعریف شود. از مهم‌ترین ویژگی‌های رنگ می‌توان به هیستوگرام رنگ، ماتریس کوواریانس رنگ، ممان‌های رنگ[3]، هیستوگرام همبستگی رنگ و بردار انسجام رنگ[4] [۲۳] اشاره نمود. ممان‌های رنگ از ساده‌ترین ویژگی‌ها هستند که در کارهای زیادی مورد استفاده قرار گرفته‌اند. معروف‌ترین ممان‌ها میانگین، انحراف معیار و چولگی[5] هستند که اغلب برای هر کانال رنگ به طور جداگانه محاسبه می‌شوند. بنابراین بردار ویژگی برای آنها بسیار کوچک است. برای افزایش کارایی این ویژگی‌ها معمولا آنها را برای نواحی یا بلوک‌های مختلف استخراج می‌کنند.

هیستوگرام رنگ، توزیع رنگ را در تصویر توصیف می‌کند. این ویژگی فضای رنگ را به سبدهای[6] مختلف تقسیم کرده، تعداد پیکسل‌هایی که در هر تصویر به هر یک از این سبدها تعلق دارد را نشان می‌دهد. هیستوگرام رنگ نسبت به تغییرات چرخشی و انتقالی مقاوم است. اما از آنجا که به اطلاعات مکانی تصویر اهمیت نمی‌دهد، ممکن است دو تصویر کاملا متفاوت هیستوگرام رنگ یکسان یا خیلی شبیه بهم داشته باشند.

بردار انسجام رنگ یا CCV، اطلاعات مکانی را در هیستوگرام رنگ اضافه می‌کند. به این صورت که هر سبد از هیستوگرام رنگ را به دو بخش تقسیم می‌کند. بخش منسجم که شامل پیکسل‌های بهم متصل است و بخش غیر منسجم که شامل پیکسل‌های مجزاست. این امر دقت CCV را نسبت به هیستوگرام رنگ بیشتر نموده و در عین حال طول ویژگی آن را دو برابر می‌کند.

این ویژگی‌ها می‌توانند به صورت سراسری از کل تصویر و یا به صورت محلی و با بلوک‌بندی کردن تصویر، برای بهره گرفتن از اطلاعات محلی استخراج شوند. در نهایت ویژگی‌های استخراج

[1] Early fusion
[2] Late fusion

[3] Moments
[4] Color coherence vector
[5] Skewness
[6] Bins

شده از هر بلوک یا به صورت ساده در کنار هم چسبانده می‌شوند و یا به صورت سبد ویژگی استفاده می‌شوند که در ادامه بحث خواهد شد. در بررسی‌های صورت گرفته بین ویژگی‌های مختلف در منبع [۲۴] نشان داده شده است که ویژگی هیستوگرام رنگ به طور متوسط به عنوان بهترین ویژگی برای کاربرد بازیابی تصاویر ظاهر می‌شود.

۳-۱-۲- ویژگی‌های بافت

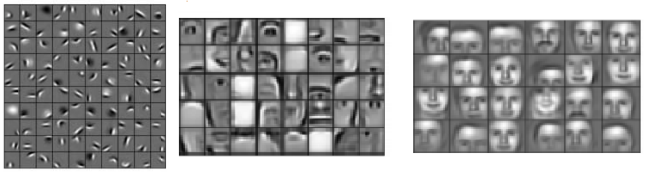
بافت تصویر اطلاعاتی در مورد آرایش و ترتیب قرار گرفتن اجزا و همچنین شدت روشنایی‌های یک تصویر ارائه می‌دهد. ویژگی بافت می‌تواند به صورت آماری یا ساختاری از تصویر استخراج شود. در رویکرد ساختاری، یک تصویر مجموعه‌ای از ریزبافت‌ها است که بصورت یک سری الگوهای منظم تکرار می‌شوند. در رویکرد آماری، که ساده‌تر و پر استفاده‌تر است، بافت تصویر به صورت کمیت‌هایی از آرایش روشنایی‌ها در یک ناحیه دیده می‌شود. رویکرد ساختاری بیشتر برای تصاویر مصنوعی و رویکرد آماری برای تصاویر طبیعی مناسب هستند. تبدیل‌های موجک و گابور [۲۵] از آنجا که مطابقت خوبی با سامانه بینایی انسان دارند، بیش از سایر ویژگی‌ها برای توصیف بافت مورد استفاده قرار گرفته‌اند. توصیف کننده هیستوگرام لبه [۲۶] نیز از ویژگی‌های پرکاربرد برای بافت است. این توصیف‌کننده تصویر را به بلوک‌هایی تقسیم کرده، هیستوگرام لبه را برای جهت‌های مختلف در هر بلوک محاسبه می‌کند. بردار ویژگی مبتنی بر بافت از طریق کنار هم قرار دادن داده‌های توصیف کننده بافت ایجاد می‌شود.

۳-۱-۳- ویژگی‌های شکل

ویژگی‌های شکل از دقت کمتری نسبت به رنگ و بافت برخوردار هستند و بیشتر در حوزه بازیابی تصاویر مصنوعی کارایی بالایی از خود نشان می‌دهند. برای مشخص کردن شکل اشیاء موجود در یک تصویر ابتدا باید محدوده اشیا شناسایی شود. برای این کار از الگوریتم‌های ناحیه‌بندی استفاده می‌شود. سپس ویژگی‌های شکل با استفاده از توصیف‌کننده‌هایی چون محیط، مساحت، قطر، نسبت طول به عرض، گِردی، ثابت‌های ممان و توصیف‌کننده‌های فوریه اندازه‌گیری می‌شود. از آنجاییکه ویژگی‌های شکل بر اثر تغییر زاویه دید یا در اثر هم‌پوشانی تغییر می‌کنند و تعیین مرز یک شی نیز با دشواری‌های زیادی همراه است، این ویژگی معمولاً برای کارهای برچسب‌زنی به تصاویر کاربرد زیادی ندارد [۷].

۳-۱-۴- ویژگی‌های ژرف [1]

در سال‌های اخیر مجموعه جدیدی از ویژگی‌ها معرفی شده است که به علت اینکه از روش‌های یادگیری ژرف استخراج می‌شود به ویژگی‌های ژرف معروف شده‌اند. این ویژگی‌ها معنای سطح بالاتری از ویژگی‌های دیگر را در بر دارند و به سیستم ادراک انسان نزدیک‌تر هستند. ویژگی‌های ژرف طی یک فرایند یادگیری ماشین با نام یادگیری ژرف [۲۷-۲۹] از تعداد بسیار زیادی نمونه و به کمک یک شبکه چند لایه‌ای استخراج می‌شوند به طوریکه خروجی هر لایه دیدی کلی‌تر و سطح بالاتر از لایه قبل ارایه می‌دهد و از پیکسل‌های تصویر به مجموعه‌ای از ریز شکل‌های تشکیل دهنده تصویر می‌رسد. دقت بالایی که ویژگی‌های مزبور در کاربردهای مختلف بینایی ماشین از خود نشان داده است بسیاری از پژوهشگران حوزه‌ی AIA را نیز به استفاده از آنها ترغیب کرده است. در شکل ۲ نمونه‌ای از ویژگی‌های ژرف را مشاهده می‌کنید.



شکل ۲- نمونه‌ای از ویژگی‌های ژرف استخراج شده از مجموعه تصاویر چهره. تصویر سمت چپ: ویژگی‌های استخراج شده از لایه‌های پایین‌تر، تصویر وسط: ویژگی‌های استخراج شده از لایه‌های میانی و تصویر سمت راست: مدل استخراج شده از لایه‌های انتهایی است[۳۰].

۳-۱-۵- سایر ویژگی‌ها

علاوه بر ویژگی‌های بصری که مستقیماً از تصویر استخراج می‌شود، ممکن است همراه با تصویر اطلاعات دیگری نیز موجود باشد که میتواند به عنوان ویژگی مورد استفاده قرار گیرد. برای مثال ممکن است همراه با تصویر، اطلاعاتی نظیر مکان جغرافیایی تصویری که گرفته شده، توضیحات[2] سایر کاربران، زمان گرفتن تصویر، متن یا ویدئوی مرتبط با آن تصویر نیز موجود باشد که از آنها نیز می‌توان برای یادگیری بهتر ارتباط بین تصاویر و برچسب‌ها بهره برد.

در زمینه یادگیری ماشین به هر یک از ویژگی‌هایی که از منابع مختلف اطلاعاتی یا با روش‌های مختلف استخراج می‌شود، یک وجه یا یک منظر و به سامانه‌هایی که از چندین منظر استفاده می‌کنند، سامانه‌های چندمنظری[3] یا چند وجهی[4] گفته می‌شود. در این سامانه‌ها رابطه بین منظرهای مختلف بررسی می‌شود [۳۱, ۳۲].

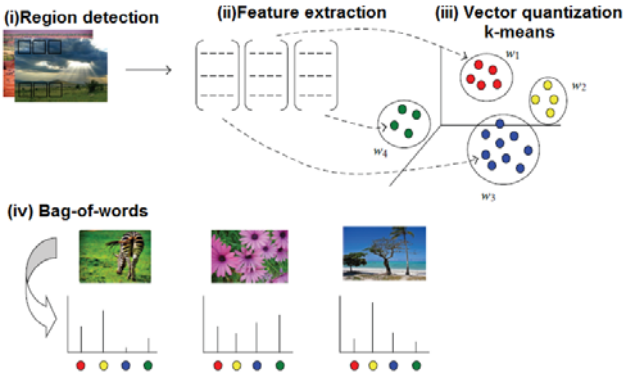
۳-۱-۶- سبد ویژگی[5] BoF-

این تکنیک که ابتدا در کاربردهای پردازش متن استفاده می‌شد، به نام سبد کلمات یا BoW[6] نیز نامیده می‌شود [۳۳, ۳۴]. در دسته‌بندی متون ابتدا کلمات پرکاربرد را کد گذاری کرده و بر اساس این کدها بردار ویژگی متن‌های مختلف را می‌سازند. در

[1] Deep features

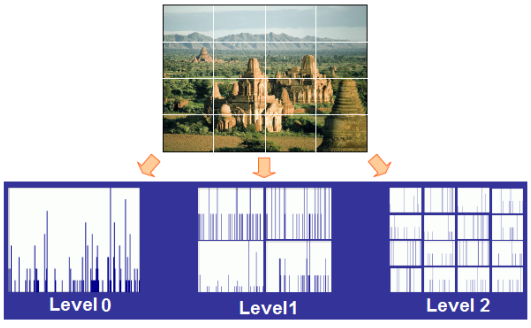
[2] Comments
[3] Multi-view
[4] Multi-modal
[5] Bag of Feature
[6] Bag of Words

پردازش تصویر نیز از همین ایده برای ساخت بردار ویژگی‌های بصری استفاده می‌شود. برای اینکار ابتدا ویژگی‌های بصری مورد نظر از تصاویر استخراج شده با یک تکنیک خوشه‌بندی مانند خوشه‌بندی k-میانگین آن‌ها را در خوشه‌های مختلف قرار می‌دهند. از مرکز هر خوشه به عنوان کد نماینده‌ی کلیه ویژگی‌های آن خوشه یاد می‌شود. سپس واژه‌نامه‌ای از این مراکز خوشه که به آن‌ها کلمه تصویری نیز گفته می‌شود، می‌سازند. در نهایت یک بردار ویژگی بر اساس تعداد واژه‌های موجود در هر تصویر تشکیل می‌شود. شکل ۳ همین مفهوم را نشان می‌دهد.



شکل ۳- استخراج ویژگی‌ها به روش سبد ویژگی - i:بلوک‌بندی تصاویر - ii:استخراج ویژگی‌ها - iii:خوشه‌بندی ویژگی‌ها - iv:تشکیل هیستوگرام برای هر تصویر بر حسب تعداد مراکز خوشه مشاهده شده[۳۵].

روش سبد ویژگی به تغییراتی مثل انتقال و دوران، مقاوم است، اما در عین حال اطلاعات محلی موجود در تصویر را نیز نادیده می گیرد. به عبارتی اگر جای قسمت‌های مختلف باهم عوض شود، نتیجه تفاوت چندانی نخواهد کرد. این مشکل را می‌توان با استفاده از تکنیک تطبیق مکانی بر روی هرم و استخراج ویژگی از چند سطح بزرگنمایی متفاوت، بهبود داد [۳۶]. ایده این کار استفاده از سطوح مختلف بزرگنمایی برای تجمیع مکانی در نواحی مختلف تصویر است. با در کنار هم قرار دادن این سطوح، هرمی از ویژگی‌های تجمیع شده به وجود می‌آید که تا حدی وابسته به مکان است. در عمل برای برقراری تعادل بین وابستگی بیش از حد و استقلال از مکان، معمولاً ویژگی‌های تصاویر در سه سطح تجمیع می‌شوند. در شکل ۴ نمایی از این تکنیک نشان داده شده است.



شکل ۴- استخراج ویژگی به روش تطبیق مکانی بر روی هرم [۳۶].

تاکنون ویژگی‌های بصری بسیار زیادی تعریف شده‌اند که می‌توان آنها را از یک تصویر استخراج کرد. اینکه کدام ویژگی مناسب‌تر است، به مسائل بسیاری بر می‌گردد. مناسب بودن یک ویژگی بستگی زیادی به مواردی چون کاربرد مساله مورد نظر، پایگاه داده مورد استفاده، میزان محاسبات و اندازه بعد ویژگی مورد انتظار، همبستگی با سایر ویژگی‌ها و نحوه ترکیب آنها دارد. در منبع [۲۴] پژوهش جامعی بر روی تعدادی از این ویژگی ها بر روی ۵ پایگاه داده صورت گرفته است.

۳-۲ مدل‌های یادگیری در برچسب‌زنی خودکار

در این مرحله یادگیری بر اساس ویژگی‌های استخراج شده صورت می‌گیرد. از منظر بازشناسی الگو، هر برچسب می‌تواند به عنوان یک دسته تلقی شود و تفاوت عمده سامانه‌های برچسب‌زنی با روش‌های دسته‌بندی، وجود چندین برچسب یا دسته برای هر تصویر است. بسیاری از مدل‌های یادگیری را می‌توان در این مرحله به کار گرفت.

به طور کلی یادگیری می‌تواند به صورت نظارتی، نیمه نظارتی یا بدون نظارت باشد. اما در برچسب‌زنی خودکار به علت لزوم وجود برچسب‌ها، معمولاً حالت بدون نظارت استفاده نمی‌شود. تا کنون پژوهش‌های زیادی در زمینه برچسب زنی خودکار و با روش‌های متفاوت صورت گرفته است. این پژوهش‌ها را می‌توان به گونه‌های مختلفی دسته‌بندی کرد. برای نمونه پژوهش [۱۰] انواع رویکردها در مدلسازی AIA را در دو گروه استقرایی[1] و هدایتی[2] دسته‌بندی کرده است. در ادامه مدلسازی AIA را در دسته‌بندی زیر مرور می‌کنیم.

۳-۲-۱- مدل‌های مولد[3]

یکی از دسته‌های مهم برای برچسب‌زنی استفاده از طراحی مدل‌های مولد است. این مدل‌ها فرض می‌کنند داده‌ها از یک توزیع تصادفی نمونه‌برداری شده‌اند و سعی می‌کنند پارامترهای توزیع را طوری تخمین بزنند که احتمال تولید آن نمونه‌ها بیشینه شود. خود این گروه را می‌توان به دو دسته تقسیم کرد که یکی رویکرد مدل مخلوط[4] و دیگری رویکرد مدل عنوان[5] را دنبال می‌کند.

در حالت مدل مخلوط هدف یافتن یک توزیع توأم بر روی نمونه‌ها و برچسب‌ها است که بتواند رابطه آنها را در مجموعه داده‌های آموزشی بهتر توصیف کند. در این مدل برای برچسب‌زنی به یک تصویر، با استخراج ویژگی‌های بصری آن تصویر، احتمال شرطی برچسب‌های مختلف به ازای آن ویژگی‌ها محاسبه می‌شود. این مدل‌ها می‌توانند به صورت تخمینی از چگالی هم‌رخدادی تصاویر و برچسب‌ها در نظر گرفته شوند.

---
[1] Inductive
[2] Transduction
[3] Generative models
[4] Mixture of models
[5] Topic models

در [۳۷] تصاویر با بیشینه کردن احتمال توأم تصویر و کلمه، برچسب زنی می‌شوند. به این صورت که همبستگی کلمه- کلمه بین تمام کلمات و همچنین همبستگی کلمه-تصویر به دست می‌آید. بر این اساس مدل تعیین می‌کند که ارزش هر جفت کلمه‌تصویر به ازای تصویر مورد آزمایش چقدر است و شرایط مورد نیاز را ارضا می‌کند یا خیر.

در مدل عنوان، تصاویر برچسب‌دار به عنوان نمونه‌هایی از مخلوط چند عنوانِ مشخص مدل می‌شوند. هر عنوان یک توزیع روی ویژگی‌های بصری و متنی تصویر است. این گروه با روش ترجمه ماشینی شروع شدند که در این حالت هر توصیف‌کننده بصری در قالب یک عنوان در نظر گرفته شده، مساله برچسب‌زنی به صورت ترجمه‌ای از چندین عنوان بصری به چندین عنوان متنی مطرح می‌شود. اغلب روش‌های مبتنی بر مدل مانند روش‌های شاخص‌گذاری معنایی مخفی[1] یا LSA ، تحلیل معنایی مخفی احتمالاتی یا PLSI و تخصیص *دیرکله* مخفی یا LDA[2] ، ابتدا برای متن کاوی معرفی شدند و پس از موفقیتی که در آن شاخه به دست آوردند، در کاربردهای پردازش تصویر نیز مورد استفاده قرار گرفتند. برای مثال در روش LDA سعی می‌شود محتوای معنایی یا GIST یک متن یا تصویر به صورت مخلوطی از عناوین خلاصه گردد. به عبارت دیگر یک مشاهده (متن یا تصویر) به صورت توزیع چندوجهی روی K عنوان، مدل می‌شود و هر کدام از این عناوین خود به صورت توزیع چندوجهی روی کلمات مدل می‌گردند[۳۸]. در [۳۹] در قالب رویکرد LDA یک رگرسیون متغیر مخفی برای یافتن همبستگی بین برچسب‌ها و ویژگی‌های بصری معرفی شده است که از طریق آن، شباهت بین دو منظر اطلاعاتی با تعدادی از عناوین مختلف محاسبه می‌گردد.

در [۴۰] و [۴۱] یک الگوریتم ترکیبی برای مساله برچسب زنی ارائه شده است که در آن ابتدا مدلی بر اساس تحلیل معنایی مخفی احتمالاتی برای تخمین احتمال پسین هر برچسب برای هر تصویر طراحی شده و بر اساس آن برچسب‌های اولیه استخراج می‌گردند. سپس یک گراف شباهت برای برچسب‌ها بر اساس میانگین گیری روی شباهت‌های بصری و متنی ساخته می‌شود و با تکنیک‌های قدم زدن تصادفی[3] نتایج برچسب‌زنی مرحله اول بهبود داده می‌شود.

یکی از روش‌های پرکاربرد در این زمینه روش تجزیه ماتریس نامنفی یا NMF[4] است که بر اساس تجزیه هر یک از ماتریس‌های ویژگی تصاویر به دو ماتریس نامنفی عمل می‌کند به طوریکه یکی از این ماتریس‌ها به عنوان بردارهای پایه یک فضای مخفی[5] و ماتریس دیگر به عنوان مختصات تصاویر در این فضای مخفی عمل می‌نمایند. در واقع روش NMF تصاویر را از فضای ویژگی‌ها به یک فضای کم‌بعد دیگر منتقل می‌کند که در این فضا الگوهای موجود در تصاویر بهتر قابل تشخیص باشند.

مراجع [۴۲-۴۵] از رویکرد NMF بهره گرفته‌اند. در مرجع [۴۲] به ازای هر تصویر آزمایشی نزدیکترین همسایه‌ها را یافته، با توجه به مفاهیم موجود آنها فضاهای مخفی مربوطه استخراج می‌شود، به طوریکه یکی از این فضاهای مخفی مربوط به برچسب‌ها و بقیه فضاها مربوط به ویژگی‌های بصری است. مدل ساخته شده به گونه‌ای است که شباهت بین فضاهای مخفی مختلف مربوط به یک تصویر حفظ شود و با توجه به همین شباهت بین فضای مخفیِ برچسب‌ها و فضاهای مخفی بصری، چندین برچسب پیشنهاد می‌شود. در مرجع [۴۳] برخلاف مرجع [۴۲]، برای کل تصاویر یک پایگاه تصاویر یک مدل واحد ساخته می‌شود و نیاز به ساخت مدل به ازای هر تصویر آزمایشی ندارد. این مدل نیز براساس شباهت بین فضاهای مخفی ویژگی‌های مختلف هر تصویر کار می‌کند. در مراجع [۲۲, ۴۴] علاوه بر ساخت یک مدل کلی برای تمام تصاویر، امکان ساخت فضاهای مختلف با ابعاد متفاوت وجود دارد. با میانگین گیری بین فاصله تصاویر در این فضاها یک معیار فاصله دقیقتر به دست می‌آید و بر اساس این معیار فاصله نزدیکترین همسایه‌های هر تصویر آزمایشی استخراج شده، برچسب‌هایی که در این همسایگی بیشتر تکرار شده‌اند پیشنهاد می‌شود. در مرجع [۴۵] به جای تاکید بر لزوم شباهت بین فضاهای مخفی، ویژگی‌های مختلف مربوط به هر تصویر به گروه-های مشابه تقسیم شده فضاهای مخفی در هر گروه به طور مجزا استخراج می‌شود. در طی فرایند استخراج با افزودن پارامترهایی به تابع هدف، بخشی از مفاهیم بین این فضاهای مخفی یکسان و بخشی دیگر مستقل در نظر گرفته می‌شوند.

در مرجع [۴۶] رابطه بین محتویات بصری و برچسب‌ها در یک فضای مخفی معنایی بر اساس تحلیل همبستگی کانونی هسته مدل شده و نشان داده شده است که در این فضا همسایه‌های بهتری برای هر تصویر جهت پیش‌بینی برچسب یافت می‌شود.

۳-۲-۲- مدل‌های تمایزی[6]

در مدل‌های تمایزی موضوع برچسب‌زنی خودکار تصاویر به صورت یک مساله دسته‌بندی چند برچسبی[7] [۴۷] بررسی شده و برای هر برچسب یک دسته‌بند جداگانه آموزش داده می‌شود. در حقیقت برچسب‌ها، دسته‌هایی مستقل از هم فرض می‌شوند. در گام آزمایش، برای هر تصویر با استفاده از این دسته‌بندها، به ازای برچسب‌های مختلف، تعلق تصویر به دسته مربوط به یک برچسب بررسی می‌شود. در این گروه، از روش‌های مختلفی برای یادگیری استفاده می‌گردد، مانند روش‌های ماشین بردار پشتیبان یا SVM[8]

[1] Latent Semantic Indexing
[2] Latent Dirichlet Allocation
[3] Random walking
[4] Nonnegative Matrix Factorization
[5] Latent spaces

[6] Discriminative models
[7] Multi-label classification
[8] Support Vector Machine

[۴۸, ۴۹]، یادگیری چند نمونه‌ای یا MIL[1] [۵۰-۵۲] و شبکه‌های عصبی [۵۳, ۵۴].

در مرجع [۴۸] تمرکز بر روی کاهش مشکلات برچسب‌ها شامل مشکل برچسب‌های ناکامل، برچسب‌های مبهم و برچسب‌های همپوشان است. در این پژوهش از SVM در حالت یکی علیه دیگران[2] استفاده شده و با تغییری در تابع اتلاف[3] hinge و افزودن افزودن پارامتر تحمل[4] در آن کارایی را افزایش می‌دهند. این پارامتر به صورت خودکار و با توجه به شباهت های بصری و آمار مربوط به پایگاه تصاویر تعیین می‌شود.

در مرجع [۵۱] یک مساله MIL در حالت تمایزی مطرح شده است که در این گونه مسائل که از جمله مسائل نظارتی ضعیف هستند، به جای مرتبط کردن هر تصویر با یک نمونه، نمونه‌ها در مجموعه‌هایی مرتب شده‌اند و برچسب‌ها به کل یک مجموعه نسبت داده می‌شوند. انتساب برچسب به یک مجموعه نشان دهنده این است که حداقل یکی از اعضای این مجموعه به آن برچسب مرتبط است. در این مرجع سعی شده است با تکنیک‌های نگاشت ویژگی و انتخاب ویژگی به صورت تمایزی مجموعه‌های دیده نشده برچسب‌گذاری گردد.

در [۵۰] از ترکیبی از نمایش چند نمونه‌ای در کنار تک نمونه‌ای برای برچسب‌زنی خودکار استفاده شده است. در آن پژوهش یک چارچوب یادگیری نیمه نظارتی مبتنی بر گراف طراحی شده که از این دو نمایش به طور همزمان بهره می‌گیرد و سه استراتژی برای تبدیل یک نمایش به نمایش دیگر برای مفاهیم مختلف ارائه شده است.

۳-۲-۳- جستجوگرا

از آنجا که سامانه های AIA بسیار غیرخطی هستند، یادگیری یک مدل پارامتریک ممکن است نتواند توزیع پیچیده داده‌ها را برای پیش بینی برچسب به خوبی بیان کند، روش‌های یادگیری محلی جستجوگرا که به صورت غیر پارامتریک مبتنی بر یافتن نزدیکترین همسایه عمل می کنند، در زمینه برچسب‌زنی خودکار بسیار مورد توجه قرار گرفته‌اند. در روش‌های جستجوگرا که در عین سادگی بسیار قدرتمند ظاهر شده‌اند، تمرکز بر روی یادگیری معیار شباهت یا معیار فاصله است. مثال‌هایی از این دسته پراکندگی برچسب بر روی گراف مشابهت، یا یادگیری بر اساس تکنیک‌های نزدیک‌ترین همسایه هستند. تعداد زیادی از روش‌های برترِ فعلی در زمره‌ی این دسته قرار می‌گیرند [۵۵-۶۱].

برای نمونه مرجع [۵۵] یک گراف شباهت روی تمام تصاویر می‌سازد و برچسب‌ها را روی این گراف پخش می‌کند. برای این کار از معیارهای فاصله متفاوتی استفاده کرده و با کمک تکنیک‌های یادگیری متریک[5]، یک میانگین وزنی را روی برچسب نزدیک‌ترین همسایه‌ها محاسبه نموده است.

مرجع [۵۸] از فرادادههای موجود در شبکه‌های اجتماعی به صورت غیر پارامتریک برای پیدا کردن مناسب‌تر نزدیک‌ترین همسایه‌ها بهره می‌گیرد. در این مرجع برای تعیین شباهت جهت محاسبه همسایه‌ها از معیار Jaccard استفاده شده است.

در [۵۹, ۶۰] یک الگوریتم دو مرحله‌ای بر اساس روش‌های نزدیک‌ترین همسایه طراحی شده است که در مرحله‌ی اول بر روی شباهت‌های تصویر-برچسب و در مرحله‌ی دوم بر روی شباهت-های تصویر-تصویر تمرکز می‌کند. همچنین در این پژوهش از یک چارچوب یادگیری معیار برای یادگیری وزن ویژگی‌های مختلف و معیار فاصله مناسب با هر ویژگی نیز استفاده شده است. در این پژوهش برای کاهش اثر نامتوازن بودن مجموعه تصاویر آموزشی سعی شده است برای هر تصویر مجموعه‌ای متوازن از همسایه‌ها فراهم شود. با این صورت که برای هر تصویر به ازای هر برچسب تعداد ثابت و یکسانی (۱ تا ۵) تصویر حاوی آن برچسب که به تصویر مزبور نزدیکتر بوده‌اند، انتخاب می‌شوند. سپس با میانگین گیری بردار برچسب‌ها در این مجموعه تصاویرِ همسایه، برچسب-هایی که امتیاز بیشتری دارند پیشنهاد می‌شوند. در این میانگین گیری میزان تاثیر تصاویر همسایه به نسبت نزدیکی آنها به تصویر مورد نظر است.

در مرجع [۶۱] با کمک روش خوشه‌بندی تصاویر پیش-الگوهایی[6] الگوهایی[6] برای هر دسته در دو فضای ویژگی‌های بصری و مفهومی ایجاد می‌شود. با مقایسه تصاویر آزمایشی با این پیش-الگوها در هر دو فضا برچسب‌های اولیه به دست می‌آیند که در مرحله آخر با روش‌های همجوشی برچسب‌های نهایی انتخاب می-گردند.

۳-۲-۴- روش‌های مبتنی بر یادگیری ژرف

به علت نتایج خوب حاصل از بکار گیری روش‌های مبتنی بر یادگیری ژرف در زمینه‌های مختلف، اخیراً مدل‌های زیادی بر اساس یادگیری ژرف برای سامانه‌های AIA طراحی شده است [۶۲-۶۵]. در یادگیری ژرف با الهام از ساختار عصبی مغز انسان سعی می‌شود، مفاهیم انتزاعی سطح بالاتری از داده‌ها مدل شود. معمولاً این کار از طریق یک گراف با چندین لایه پردازشی متشکل از ترکیبات خطی یا غیرخطی صورت می‌گیرد. از جمله این گراف‌ها می‌توان به روش‌هایی نظیر شبکه عصبی عمیق[7]، شبکه عصبی پیچشی[8]، شبکه باور عمیق[9] اشاره نمود. نوآوری اصلی این این روش‌ها در دو زمینه خلاصه می‌شود: استخراج ویژگی‌های

[1] Multiple-Instance Learning
[2] One vs rest
[3] Loss function
[4] Tolerance

[5] Metric Learning
[6] Prototype
[7] Deep neural network
[8] Convolutional neural Networks
[9] Deep belief network

بصری قوی‌تر و بهره‌گیری از برچسب‌های آموزشی و اطلاعات جانبی دیگر در بهبود پیش‌بینی.

در [۶۳] با استفاده از مدل پیشنهاد شده توسط کریژوسکی [۶۶] از یک شبکه پیچشی ژرف بدون هیچ گونه پیش‌آموزش استفاده شده است. از آنجایی که می‌توان مسئله برچسب‌زنی را به صورت مسئله دسته‌بندی چندبرچسبی در نظر گرفت، برای آموزش شبکه از چندین تابع اتلاف چندبرچسبی متفاوت استفاده شده است. در آخر تعداد k برچسب که بیشترین احتمال انتساب به تصویر ورودی را دارند، به عنوان برچسب‌های تصویر انتخاب شده‌اند.

مرجع [۵۸] از انواع مختلفی از فرا-داده‌ها برای یافتن نزدیکترین همسایه‌ها بهره گرفته است و با کمک یک شبکه عصبی پیچشی ژرف اطلاعات بصری بین یک تصویر و همسایه‌های آن را با هم ترکیب و برچسب‌ها را پیش‌بینی کرده است.

در [۶۴] از ویژگی‌های استخراج شده توسط یک شبکه عصبی پیچشی استفاده شده و مدلی را در چارچوب تحلیل همبستگی کانونی (CCA) برای هر دو منظر بصری و متنی طراحی کرده است. در این مرجع، چارچوب CCA در سه حالت خطی، مبتنی بر هسته و براساس خوشه‌بندی نزدیکترین همسایه به کار گرفته و مقایسه شده است.

در مرجع [۶۲] یک مدل یادگیری ژرف چند مقیاسه چند منظوره[1] برای استخراج ویژگی‌های توصیفی از تصاویر طراحی شده است. این ویژگی‌ها از سطوح مختلف لایه‌ها استخراج شده و با طی یک فرایند هم‌جوشی باهم ترکیب می‌شوند. سپس ویژگی‌های بصری استخراج شده با ویژگی‌های متنی برگرفته شده از برچسب‌های تصاویر آموزشی توسط یک زیر شبکه چند لایه‌ای پرسپترون[2] اصلاح می‌شوند. نهایتاً مساله برچسب‌زنی به دو مساله‌ی دسته‌بندی چندکلاسه و پیش بینی مقدار برچسب‌ها[3] تجزیه می‌شود که برای حل این دو مساله نیز از مدل یادگیری ژرف استفاده شده است.

در [۶۵] یک هسته چندگانه ژرف به صورت بازگشتی به عنوان ترکیب چند لایه‌ای توابع غیرخطی که هر کدام از آنها نیز ترکیبی از چند هسته ابتدایی یا میانی می‌باشد، تعریف شده است. برای یافتن ضرایب این شبکه‌ها از چهار حالت نظارتی، غیرنظارتی، نیمه نظارتی مبتنی بر هسته و نیمه نظارتی مبتنی بر لاپلاسین استفاده شده است.

در مرجع [۶۷] به منظور طراحی سامانه AIA، روش تولید فرضیه-های اشیا به صورت غیرنظارتی با روش یادگیری ژرف ترکیب شده است. برای هر تصویر فرضیه‌هایی مبنی بر وجود برچسب‌ها تولید می‌شود و ویژگی‌های تصویر برای هر فرضیه با کمک مدل شبکه عصبی ژرف استخراج می‌شود. با ترکیب ویژگی‌های تمام فرضیه-ها، ویژگی کل تصویر به دست می‌آید. سپس برای هر برچسب احتمال اینکه آن برچسب با تصویر همبسته باشد محاسبه می‌گردد.

در جدول ۱ خلاصه‌ای از پژوهش‌های مرور شده برای هر کدام از روش‌های یادگیریِ مدل آورده شده است.

۳-۳ پیش‌بینی برچسب

در این مرحله بر اساس یادگیری صورت گرفته در مرحله قبل، برای تصاویر بدون برچسب و مشاهده نشده، برچسب‌هایی پیشنهاد می‌شود. مکانیزم پیشنهاد برچسب بر اساس اعمال تصاویر به مدلِ ساخته شده در مرحله آموزش یا بر اساس تکنیک‌های جستجو و انتخاب معیار شباهت صورت می‌گیرد. تعداد برچسب برای هر تصویر می‌تواند به صورت مقدار ثابتی که معمولاً میانگین تعداد برچسب‌ها در تصاویر آموزشی است، تعیین شود یا بر اساس یک مقدار آستانه برچسب‌های با درجه اطمینان بیشتر معرفی شوند. روش دوم برای مدل‌هایی که به ازای هر برچسب یک احتمال وقوع یا درجه اطمینان نسبت می‌دهد، مناسب است.

در مرحله پیش بینی برچسب معمولاً از معیارهای مختلف شباهت استفاده می‌شود.

۴ پایگاه‌های تصاویر مربوط به AIA

پایگاه‌های تصاویر زیادی برای بازیابی تصاویر ایجاد شده است. برخی از این پایگاه داده ها مشکلاتی دارند که انها را مناسب استفاده در پژوهش‌های حوزه AIA نمی‌نماید. برای مثال تصاویر آنها برچسب گذاری نشده‌اند و فقط دسته‌ی تصاویر مشخص شده است، تصاویر با استفاده از متن‌های طولانی حاشیه نویسی شده‌اند، و یا تصاویر آنها بیشتر برای کاربردهای خاصی مانند تشخیص شی[4] یا شناسایی صحنه[5] مناسب هستند. از این بین، پایگاه‌های تصاویری نیز وجود دارند که به صورت چند برچسبی حاشیه‌نویسی شده‌اند و برای کارهای AIA قابل استفاده می باشند. چند نمونه‌ی پر استفاده‌تر از این پایگاه‌ها در جدول ۲ آورده شده است.

برای بسیاری از این پایگاه‌های تصاویر ویژگی‌های مختلفی به صورت آماده و از پیش استخراج شده وجود دارد که برای کاربردهای AIA قابل استفاده است. برای نمونه ۱۵ ویژگی مختلف و پرکاربرد برای پایگاه‌های Corel 5K, ESP-Game, IAPR TC-12 و Flickr MIR در مرجع [۵۷] به صورت رایگان قابل دسترس است[6].

همچنین برای برخی از این پایگاه‌ها، اطلاعات ناحیه‌بندی شده تصاویر نیز وجود دارد.

۵ روش‌های ارزیابی

پس از ساخته شدن مدل بر اساس تصاویر آموزشی و برچسب‌های آنها، عملکرد آن مدل بر روی داده‌های آزمایشی ارزیابی می‌شود.

[1] Multi-Modal Multi-Scale
[2] Perception
[3] Label quantity prediction
[4] Object detection
[5] Scene recognition
[6] http://lear.inrialpes.fr/people/guillaumin/data.php

به این صورت که تصاویر آزمایشی بدون برچسب به سامانه ارائه شده و برچسب‌هایی به آن نسبت داده می‌شود. با مقایسه برچسب-های پیشنهاد شده از سوی مدل و برچسب‌های مرتبط با تصاویر که در پایگاه داده موجود است، کارایی سامانه با معیارهای مختلف ارزیابی می‌شود. از جمله پارامترهای پر استفاده برای ارزیابی سامانه‌های AIA می‌توان به معیارهای دقت[1]، بازخوانی[2] و F1-score اشاره نمود [۶۸].

با در نظر گرفتن پارامترهای زیر هر یک از معیارهای فوق در ادامه تعریف می‌شوند:

tp: تعداد تصاویری که برای آن‌ها، خروجی سامانه و پایگاه داده هر دو حضور یک برچسب را تایید کنند.

fp: تعداد تصاویری که در آن سامانه به حضور برچسب و پایگاه داده به عدم حضور آن رای دهد.

fn: تعداد تصاویری که در آن سامانه به عدم حضور و پایگاه داده به حضور آن برچسب گواهی دهد.

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F1 - score = 2\frac{precision \cdot recall}{precision + recall}$$

دو معیار دقت و بازخوانی به نوعی مکمل یکدیگر هستند و معمولاً با زیاد شدن یکی، دیگری تا حدودی کاهش میابد. معیار F1-score معیارهای دقت و بازخوانی را ترکیب کرده و یک میانگین سازگار از آن‌ها را محاسبه می‌کند. معیارهای فوق برای هر برچسب به طور جداگانه محاسبه و در نهایت بین کل برچسب‌ها میانگین-گیری می‌شود.

همچنین معیار N+ نیز از جمله معیارهای پراستفاده در ارزیابی سامانه‌های AIA است. این معیار که به صورت تعداد برچسب‌هایی که بازخوانی غیر صفر دارند تعریف می‌شود، در حقیقت چگونگی برخورد سامانه با مشکل نمونه‌های مثبت نامتوازن را نشان می‌دهد.

۶ مشکلات سامانه‌های برچسب‌زنی خودکار

موضوع برچسب‌زنی خودکار به تصاویر موضوع پیچیده‌ای محسوب می‌شود و با چالش‌های زیادی روبرو است. پایین بودن دقت‌های به دست آمده برای این سامانه‌ها در مقایسه با روش‌های ساده دسته‌بندی نیز همین حقیقت را نشان می‌دهد. در این بخش به بررسی چند مشکل عمده که مانعی برای بالا رفتن دقت این سامانه‌ها به حساب می‌آید، خواهیم پرداخت.

۶-۱ کامل نبودن برچسب‌ها

سامانه‌های نظارتی یا نیمه نظارتی معمولاً از تصاویری که به صورت دستی برچسب‌زنی شده‌اند، برای آموزش خود استفاده می-کنند. در برچسب‌زنی دستی در کنار اشکالاتی مانند زمان‌بر بودن، هزینه زیاد و سلایق مختلف کاربران، یک مشکل عمده وجود دارد. تصاویر با تمام برچسب‌های درست برچسب‌گذاری نمی‌گردند. برای بسیاری از تصاویر همه برچسب های نسبت داده شده به آنها درست نیستند. به عبارت دیگر برچسب‌ها ناکامل هستند. این مشکل ممکن است ناشی از وجود برچسب‌های هم معنی یا با اشتراک در معنی، مانند گل و شکوفه باشد، یا به علت اشتباه و فراموشی کاربر در برچسب زدن تمام موارد مربوط به یک تصویر رخ دهد و یا حتی ممکن است ناشی از کوچک بودن یا بی اهمیت بودن بعضی از اشیا در مقابل بقیه آنها از دید شخصی که برچسب می‌زند و یا شلوغ بودن صحنه باشد. درست نبودن برچسب‌ها به هر علتی که رخ داده باشد، بسیار مشکل‌ساز است. چراکه در بسیاری از موارد سامانه برچسب زنی (با توجه به آنچه در مرحله آموزش آموخته است) برچسبی را به صورت صحیح به تصویری نسبت می‌دهد. اما به علت کامل نبودن برچسب‌های از پیش ثبت شده برای آن تصویر که آموزش بر اساس آنها انجام شده است، وقتی که یک شخص متخصص برچسب های پیشنهادی سیستم را مورد ارزیابی قرار میدهد، برخی از این برچسب را بعنوان برچسب غلط ارزیابی میکند و به سیستم بازخورد منفی می دهد.

نمونه‌ای از ناکامل بودن برچسب‌ها در جدول ۳ آورده شده است.

این مشکل خصوصاً برای سامانه‌هایی مانند SVM که در آن‌ها تمام نمونه‌هایی که با یک دسته برچسب نخورده‌اند و به عنوان نمونه‌های منفی محسوب می‌شوند، بیشتر نمود پیدا می‌کند [۴۸]. در مقابل آن استفاده از سامانه‌هایی مانند نزدیک‌ترین همسایه که بر مبنای شباهت کار می‌کنند، پاسخ بهتری می‌دهند.

۶-۲ متوازن نبودن دسته‌ها

بسیاری از پایگاه‌های تصاویر که برای آموزش سامانه‌های برچسب‌زنی خودکار به کار می‌روند، دچار عدم توازن شدیدی در تعداد تصاویر مربوط به برچسب‌های مختلف هستند. این مساله نه تنها در پایگاه‌های تصاویر موجود، بلکه در دنیای واقعی و در فضای اینترنت نیز وجود دارد که در آنها برای بعضی از موضوعات خاص تعداد کمی نمونه یافت می‌شود.

در مرجع [۵۹] در همین زمینه یک بررسی بر روی سه پایگاه تصاویر متداول Corel 5K، ESP-Game و IAPR TC-12 صورت گرفته است. جدول ۴ نتیجه این بررسی را برای معیارهای میانگین، میانه و بیشینه نشان می‌دهد. تفاوت زیادی که بین میانه و میانگین وجود دارد گواهی بر همین عدم توازن برچسب‌ها است. برای مثال در پایگاه داده IAPR TC-12 تعداد تصاویر برای هر برچسب در بیشترین حالت ۴۹۹۹ و در حالت میانه ۱۵۳ می‌باشد.

[1] Precision
[2] Recall

در این پایگاه تصویر حدود ۷۵% تصاویر، فرکانسی کمتر از میانگین فرکانس برچسب‌های آن پایگاه داده دارند.

جدول ۳-نمونه‌هایی از برچسب‌های ناکامل. دو تصویر ابتدایی از پایگاه تصویر Corel 5K و دو تصویر انتهایی از پایگاه تصویر ESP-Game را نشان می‌دهد [۲۲].

تصویر	برچسب‌های ثبت شده	سایر برچسب-های مرتبط
	bear, reflection, water, black	lake, grass, grizzly, brown
	jet, mountain, plane	tree, sky, cloud, flight, blue
	room, white, lamp, blue, tv, picture, chair, window, floor, table, apartment	flower, curtain, hotel, photo, sofa
	bald, map, green, man	shose, grass, tree, , field

برای درک تاثیری که این مشکل بر روی کارایی سامانه‌های برچسب‌زنی می‌گذارد، کارایی سامانه خط پایه[1] معرفی شده در [۵۶] به ازای برچسب‌های با فرکانس متفاوت بررسی شده است. برای برچسب‌هایی با فرکانس کمتر از ۲۰% میانگین، معیار F1-score به نتیجه ضعیف ۱۹/۷% می‌رسد، در حالیکه برای ۲۰% برچسب‌ها با بیشترین فرکانس این معیار نتیجه بالایی معادل ۵۰/۶% به دست می‌آورد.

در شکل ۵ تعداد تصاویر را به ازای هر برچسب در پایگاه‌های تصاویر Corel 5K، ESP-Game و IAPR TC-12 مشاهده می‌کنید.

۶-۳ انتخاب ویژگی مناسب

یکی از چالش‌های مهم در طراحی سامانه‌های برچسب زنی خودکار انتخاب ویژگی مناسب است. انتخاب ویژگی تاثیر بسیار زیادی بر روی کارایی سامانه خواهد داشت. همان گونه که در بخش ۳-۱ گفته شد، برای این سامانه‌ها معمولاً ترکیبی از ویژگی-های مختلف مبتنی بر رنگ و بافت استفاده می‌شود. بسیاری از این ویژگی‌ها به عنوان یک ویژگی منفرد دقت خوبی دارند، اما ممکن است به علت همبستگی با برخی ویژگی‌های دیگر در یک مجموعه باعث کاهش کارایی شوند. استفاده از تکنیک‌های انتخاب ویژگی یا وزن دادن به ویژگی می تواند موثر واقع شود.

۶-۴ پیچیدگی زیاد مدل‌ها

برخی از روش‌های برچسب‌زنی از مدل‌های پیچیده ریاضی با تعداد زیادی پارامتر استفاده می‌کنند. این مدل‌های پیچیده معمولا برای تعداد زیاد معنا دقت خوبی ارائه نمی‌دهند. برای آموزش در مدل-های پیچیده به مجموعه بسیار بزرگی از تصاویر آموزشی نیاز هست. از طرفی چنین سامانه‌هایی با خطر بیش برازش[2] مواجه خواهند بود که در آن یک مدل با دقت بالا بر روی مجموعه آموزشی منطبق می‌شود به طوریکه برای تصاویری از خارج از آن مجموعه نمی‌تواند پاسخ خوبی بیابد. در حقیقت در این حالت عملاً یادگیری به درستی صورت نگرفته است.

۶-۵ پیچیدگی برچسب‌ها

تنوع و پیچیدگی برچسب‌ها یا کلماتی که به هر تصویر نسبت داده می‌شود از دیگر چالش‌های پیش روی سامانه‌های AIA است. کلماتی که نمایانگر یک شی یا مفهوم بصری مشخص باشند، در صورتیکه نمونه‌های آموزشی مناسبی داشته باشند، به راحتی آموخته می‌شوند اما یادگیری برچسب‌هایی با مفاهیم انسانی و سطح بالا مانند "تنهایی"، "محبت" یا "تنفر" که نشانه‌های بصری مشخصی ندارند بسیار دشوار است. همچنین اغلب در بین کلمات مورد استفاده در یک پایگاه، رابطه معنایی وجود دارد. برای مثال رابطه کلی-جزئی بودن در کلمات "حیوان" و "گربه" وجود دارد. به طور کلی بررسی رابطه بین کلمات و همبستگی آنها می تواند برای ارائه پیشنهادهای بهتر موثر باشد. برای مثال وجود برچسب "ابر" احتمال وجود برچسب "آسمان" را بیشتر می‌کند. همچنین برخی کلمات ظاهر نوشتاری یکسانی دارند، در حالیکه معنای متفاوتی را نشان می‌دهند. مثلاً کلمه "شیر" در فارسی که به سه معنای متفاوت به کار می‌رود. برای یادگیری چنین مفاهیمی لازم است ابتدا خوشه-بندی بر روی نمونه‌های آموزشی صورت گیرد و هر خوشه به طور مجزا آموزش داده شود.

[1] Baseline

[2] Overfitting

۶-۶ برچسب‌زنی در دنیای واقعی

در سال‌های اخیر تعداد تصاویر تولید شده به طور روزانه به طرز چشم‌گیری رو به افزایش است. اگر چه تلاش‌هایی برای برچسب زنی خودکار تصاویر در دنیای واقعی و خارج از یک پایگاه داده مشخص، با برچسب‌ها و تصاویر محدود صورت گرفته است، اما هنوز راه درازی برای رسیدن به دقت قابل قبول در برچسب‌گذاری در پیش داریم.

جدول ۴-مقایسه تعداد برچسب‌ها به ازای هر تصویر و تعداد تصاویر به ازای هر برچسب به صورت میانگین، میانه و بیشینه [۵۹] .

خصوصیت	Corel 5K	ESP-Game	IAPR TC-12
میانگین تصویر به ازای برچسب	۵۸/۶	۳۲۶/۷	۳۴۷/۷
میانه تصویر به ازای برچسب	۲۲	۱۷۲	۱۵۳
بیشینه تصویر به ازای برچسب	۱۰۰۴	۴۵۵۳	۴۹۹۹
میانگین برچسب به ازای تصویر	۳/۴	۴/۷	۵/۷
میانه برچسب به ازای تصویر	۴	۵	۵
بیشینه برچسب به ازای تصویر	۵	۱۵	۲۳

کار با برچسب‌های بسیار زیاد (به اندازه اسامی موجود در یک فرهنگ لغت) با ابهامات و اشتراکات بسیار و نویز بالا، به یک مرحله پیش پردازش پیچیده جهت حذف برچسب‌های اضافی، گروه‌بندی کردن برچسب‌های هم‌معنی و رفع ابهام از برچسب‌هایی با معانی مختلف نیاز دارد. همچنین در مواجه با تعداد بسیار زیاد تصاویر برای هر برچسب، سامانه‌هایی با قابلیت‌های انتخاب مجموعه مناسب با تنوع و پوشش قابل قبول برای آموزش، توانایی شاخص‌گذاری بالا و استفاده از تکنیک‌های توزیع شده برای افزایش توان پردازشی مورد نیاز است.

۶-۷ مشکلات دیگر

تصاویر طبیعی پیچیدگی‌های زیادی دارند. وجود پس‌زمینه‌های شلوغ، تعداد اشیاء زیاد، تفاوت فاصله اشیاء از دوربین، تفاوت زاویه دوربین برای اشیاء مختلف، هم‌پوشانی اشیا و تغییرات روشنایی از جمله مواردی هستند که کار برچسب‌زنی خودکار را دشوار می‌سازند. این مسائل گاهی سبب می‌شود دو تصویر از دو دسته متفاوت توصیف مشابهی پیدا کنند، به عبارت دیگر با برچسب‌های یکسان توصیف شوند و یا دو تصویر از یک دسته، بسیار متفاوت تشخیص داده شده و برچسب های کاملاً متفاوتی به آنها نسبت داده شود.

۷ جمع‌بندی

در این مقاله سامانه‌های برچسب‌زنی خودکار تصاویر مرور شده است. در طراحی یک سامانه‌ی برچسب‌زنی خودکار مانند بسیاری از کاربردهای یادگیری ماشین، سه مرحله وجود دارد. در مرحله اول استخراج ویژگی بر اساس داده‌ها و شرایط مساله صورت می‌گیرد. این ویژگی‌ها معمولاً به صورت ترکیبی از انواع ویژگی‌های رنگ و بافت انتخاب می‌شوند. برخی از این ویژگی‌ها به علت اینکه خصوصیات مکانی تصویر را نادیده می‌گیرند، بهتر است به صورت سبد ویژگی استخراج شوند. پس از استخراج ویژگی، با استفاده از برچسب‌های مجموعه تصاویر آموزشی مدلی برای یادگیری ساخته می‌شود که می‌تواند مدل مولد، مدل تمایزی، جستجوگرا یا به صورت یادگیری ژرف باشد. در مدل مولد سعی می‌شود نوع توزیع ویژگی‌ها تشخیص داده شده، پارامترهای آن توزیع تخمین زده شوند. در مدل تمایزی برای هر برچسب یک دسته‌بند مجزا آموزش داده می‌شود که تعلق یا عدم تعلق آن برچسب به هر تصویر را پیش‌بینی کند. در مدل‌های جستجوگرا برچسب‌های یک تصویر بر اساس برچسب‌های تصاویر مشابه که در همسایگی آن تصویر هستند انتخاب می‌شوند و مدل‌های مبتنی بر جستجوی ژرف به صورت شبکه‌هایی چند لایه برای استخراج ویژگی‌ها و نمایش مفاهیم سطح بالاتر از تصاویر طراحی می‌شوند. پس از ساخت مدل و آموزش آن، در فاز پیش بینی برچسب با ارائه ویژگی‌های تصاویر آزمایشی به مدل مزبور برچسب هایی متناسب با هر تصویر پیش‌بینی می‌شود. برای ارزیابی سامانه‌های برچسب‌زنی معمولاً از معیارهای دقت، بازخوانی، F1-score و N+ استفاده می‌شود. از جمله مشکلات موجود برای طراحی سامانه‌های برچسب‌زنی تصاویر می‌توان به کامل نبودن برچسب‌های تصاویر آموزشی، عدم توازن در تعداد تصاویر مربوط به هر برچسب در مجموعه تصاویر آموزشی، انتخاب ویژگی‌های مناسب برای مدل و وجود خطای بسیار در برچسب‌های نسبت داده شده به تصاویر خامی که در دنیای واقعی و خارج از پایگاه‌های تصاویر محدود طراحی شده برای این سامانه‌ها وجود دارد اشاره نمود.

جدول ۱ - جمع‌بندی روش‌های مرور شده برحسب نوع مدل به کار گرفته شده، شماره مرجع، نام نویسنده و سال نشر مقاله.

[۳۷]: Liu، ۲۰۱۳	احتمال توام تصویر کلمه را بیشینه می کند.
[۳۸]: Rasiwasia، ۲۰۱۰	توزیع چند وجهی تصویر روی k عنوان و توزیع چند وجهی عناوین روی برچسب‌ها محاسبه می‌شود.
[۳۹]: Putthividhy، ۲۰۱۰	همبستگی بین برچسب‌ها و ویژگی‌های بصری به کمک روش LDA محاسبه

		می‌شود.
مدل مولد	[۴۰]: Tian، ۲۰۱۴	به روش PLSI احتمال پسین[1] هر برچسب محاسبه می‌شود.
	[۴۲]: Kalayeh، ۲۰۱۴	به روش NMF برای هر تصویر فضاهای مخفی مختلف به گونه‌ای که نمایش آنها به هم شباهت داشته باشند استخراج می‌گردد.
	[۴۳]: Rad، ۲۰۱۵	یک مدل کلی برای تمام تصاویر به روش NMF و بر اساس شباهت بین فضاهای مخفی ساخته می‌شود.
	[۴۴]: Rad، ۲۰۱۷	اجازه می‌دهد فضاهای مخفی استخراج شده به روش NMF ابعاد متفاوتی با توجه به بعد ویژگی مربوطه داشته باشد.
	[۴۵]: Rad، ۲۰۱۷	برای استخراج فضاهای مخفی به روش NMF برخی عامل‌های مخفی را شبیه و برخی از آنها را یکسان در نظر می‌گیرد.
	[۴۶]: Ballan، ۲۰۱۴	بر اساس تحلیل همبستگی کانونی هسته رابطه بین ویژگی‌های بصری و متنی را مدل می‌کند.
مدل تمایزی	[۴۷]: Xu، ۲۰۱۵	برچسب زنی را به صورت مساله دسته‌بندی چند برچسبی بررسی کرده برای هر برچسب یک دسته بند آموزش می‌دهد.
	[۴۸]: Verma، ۲۰۱۳	بر اساس SVM با تابع اتلاف hinge و به روش یکی بر علیه دیگران کار می‌کند.
	[۴۹]: Zhou، ۲۰۱۵	برای هر یک از هزاران مفهوم مختلف یک تشخیص دهنده مفهوم به روش SVM در سطح تصویر و در سطح ناحیه طراحی می‌کند.
	[۵۰]: Jinhui، ۲۰۱۰	به صورت یادگیری چند نمونه‌ای و به روش مبتنی بر گراف کار می‌کند.
	[51]:Richang، ۲۰۱۴	از یک روش یادگیری چند نمونه‌ای به همراه راهکارهای انتخاب ویژگی بهره می‌گیرد.
	[۵۲]:Ding، ۲۰۱۶	برای حل مساله یادگیری چند نمونه‌ای، برچسب‌زنی روش‌های مبتنی برگراف و مفاهیم مخفی را ترکیب می‌کند.
	[۵۳]: Savita، ۲۰۱۳	یک شبکه عصبی پرسپترون چند لایه‌ای با ویژگی‌های مخصوص طراحی می‌کند.
	[۵۴]: Shin، ۲۰۱۶	از یک روش شبکه‌های عصبی بازگشتی RNN بر پایه ویژگی‌های CNN ژرف استفاده می‌کند.
جستجوگرا	[۵۵]: Guillaumin، ۲۰۰۹	یک گراف شباهت کل تصاویر بر اساس معیارهای فاصله متفاوت و یادگیری متریک ساخته، بین برچسب‌های همسایه میانگین گیری می‌کند.
	[۵۸]: Johnson، ۲۰۱۵	از فراداده‌های موجود در شبکه‌های اجتماعی برای افزایش دقت محاسبه نزدیکترین همسایه استفاده می‌کند.
	[۵۹]: Verma، ۲۰۱۲	برای هر تصویر مجموعه متوازنی از همسایه‌ها پیدا می‌کند و با یادگیری متریک برچسب‌ها را نسبت می‌دهد.
	[۶۰]: Verma، ۲۰۱۷	مانند مرجع ۵۹.
	[۶۱]: Bahrololoum، ۲۰۱۷	تصاویر را خوشه‌بندی کرده و برای هر دسته الگوهایی ایجاد می‌کند.
	[۵۸]: Johnson، ۲۰۱۵	اطلاعات دیداری بین یک تصویر و همسایه‌های آن را با کمک یک شبکه عصبی پیچشی ژرف ترکیب می‌کند.
	[۶۲]: Niu، ۲۰۱۷	ویژگی‌های سطوح مختلف از یک شبکه یادگیری ژرف با هم ترکیب شده همراه با برچسب‌های آموزشی در یک شبکه پرسپترون اصلاح می‌شوند.
	[۶۳]: Gong، ۲۰۱۳	از یک شبکه پیچشی ژرف که برای آموزش آن از چندین تابع اتلاف چندبرچسبی استفاده شده بهره می‌گیرد.
	[۶۴]: Murthy، ۲۰۱۵	مدلی را به روش CCA بر اساس ویژگی‌های استخراج شده توسط یک شبکه

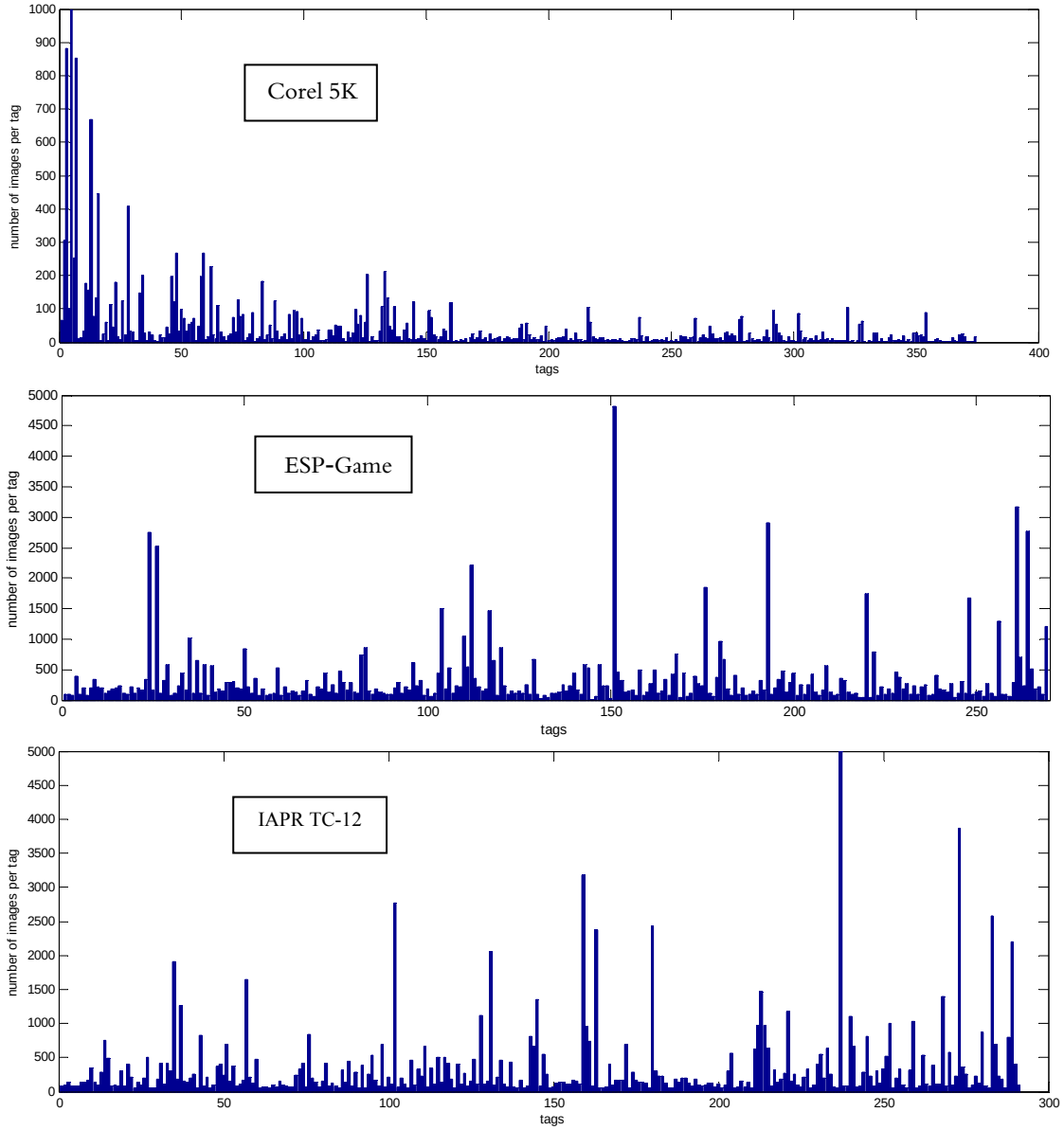
[1] Posterior

یادگیری ژرف	عصبی پیچشی برای دو منظر بصری و متنی طراحی می‌کند.
[۶۵]: Jiu، ۲۰۱۷	یک هسته چندگانه ژرف براساس ترکیب چند لایه‌ای توابع غیرخطی که هر کدام از آنها نیز ترکیبی از چند هسته ابتدایی یا میانی می‌باشد، تعریف می‌شود.
[۵۸]: Johnson، ۲۰۱۵	اطلاعات بصری بین یک تصویر و همسایه‌های آن را با کمک یک شبکه عصبی پیچشی ژرف ترکیب می‌کند.
[۶۷]: sang، ۲۰۱۷	برای هر تصویر فرضیه‌هایی مبنی بر وجود برچسب‌ها تولید می‌شود و ویژگی‌های تصویر برای هر فرضیه با کمک مدل شبکه عصبی ژرف استخراج می‌شود.

جدول ۲- برخی از پایگاه تصاویر پر استفاده در حوزه AIA به همراه خصوصیات آنها.

پایگاه تصویر	تعداد تصاویر	تعداد مفاهیم (دسته‌ها)	تعداد کل برچسبها	میانگین برچسب برای تصویر	تعداد تصاویر آموزشی	تعداد تصاویر آزمایشی
MIR-Flickr[1]	۲۵۰۰۰	۳۸	۱۳۸۶	۹	۱۲۵۰۰	۱۲۵۰۰
NUS-Wide[2]	۲۶۹۶۴۸	۸۱	۵۰۱۸	-	۱۶۱۷۸۹	۱۰۷۸۵۹
NUS-WIDE-LITE[2]	۵۵۶۱۵	۸۱	۵۰۱۸	-	۲۷۸۰۷	۲۷۸۰۷
Corel 5K	۵۰۰۰	۱۰۰	۲۶۰	۳/۴	۴۵۰۰	۵۰۰
Corel 60K[3]	۶۰۰۰۰	۵۹۹	۴۱۷	-	-	-
ESP Game[4]	۲۰۷۷۰	-	۲۶۸	۴/۷	۱۸۶۸۹	۲۰۸۱
IAPR TC-12[5]	۱۹۶۲۷	۴۱	۲۹۱	۵/۷	۱۷۶۶۵	۱۹۶۲
Pascal VOC 2,07[6]	۹۹۶۳	۲۰	۸۰۴	-	۲۵۱۰+۲۵۰۱	۴۹۵۲
Caltech 256[7]	۳۰۶۰۸	۲۵۶	۲۵۶	۱	-	-
Caltech-101[8]	۸۷۶۵	۱۰۱	۱۰۱	۱	-	-
Lable Me[9]	۴۱۷۲۴	۱۸۳	-	۳,۳	-	-
ImageNet[10]	۱۴۱۹۷۱۲۲	۲۱۸۴۱	۲۱۸۴۱	-	-	-

[1] http://press.liacs.nl/mirflickr/
[2] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[3] http://wang.ist.psu.edu/docs/related.shtml
[4] http://hunch.net/~jl/
[5] http://imageclef.org/photodata
[6] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/
[7] http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[8] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[9] http://labelme2.csail.mit.edu/Release3.0/browserTools/php/dataset.php
[10] http://image-net.org/download



شکل ۵- تعداد تصاویر برای هر برچسب. بالا: Corel 5K، وسط: ESP-Game، پایین: IAPR TC-12. محور افقی برچسب‌ها و محور عمودی تعداد تصویر به ازای هر برچسب را نشان می‌دهد [۲۲].

مراجع

[1] T. Dharani, I.L. Aroquiaraj, A survey on content based image retrieval, Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on, IEEE2013, pp. 485-490.

[2] P. Shrivas, U.K. Lilhore, N. Agarwal, A Survey on Image Retrieval by Different Features and Techniques, (2017).

[3] S. Gandhani, N. Singhal, Content based image retrieval: survey and comparison of CBIR system based on combined features, International Journal of Signal Processing, Image Processing and Pattern Recognition, 8 (2015) 155-162.

[4] A. Kumar, J. Kim, W. Cai, M. Fulham, D. Feng, Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data, Journal of digital imaging, 26 (2013) 1025-1039.

[5] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22 (2000) 1349-1380.

[6] A.-M. Tousch, S. Herbin, J.-Y. Audibert, Semantic hierarchies for image annotation: A survey, Pattern Recognition, 45 (2012) 333-345.

[7] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recognition, 45 (2012) 346-362.

[8] F. Wang, A survey on automatic image annotation and trends of the new age, Procedia Engineering, 23 (2011) 434-438.

[9] S. Kadam, S. Bajpai, P. Yelmar, Annotation: an investigative survey of annotation types and systems, Proceedings of the International Conference on Advances in Engineering and Technology2014, pp. 102-105.

[10] X. Li, T. Uricchio, L. Ballan, M. Bertini, C.G. Snoek, A.D. Bimbo, Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval, ACM Computing Surveys (CSUR), 49 (2016) 14.

[11] A. Doan, R. Ramakrishnan, A.Y. Halevy, Crowdsourcing systems on the world-wide web, Communications of the ACM, 54 (2011) 86-96.

[12] C. Yang, M. Dong, J. Hua, Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning, Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE2006, pp. 2057-2063.

[13] H. Frigui, J. Caudill, Region based image annotation, Image Processing, 2006 IEEE International Conference on, IEEE2006, pp. 953-956.

[14] Y. Wang, T. Mei, S. Gong, X.-S. Hua, Combining global, regional and contextual features for automatic image annotation, Pattern Recognition, 42 (2009) 259-266.

[15] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, R. Jain, Tri-clustered tensor completion for social-aware image tag refinement, IEEE transactions on pattern analysis and machine intelligence, 39 (2017) 1662-1674.

[16] J. Wang, J. Zhou, H. Xu, T. Mei, X.-S. Hua, S. Li, Image tag refinement by regularized latent Dirichlet allocation, Computer Vision and Image Understanding, 124 (2014) 61-70.

[17] Z. Lin, G. Ding, M. Hu, J. Wang, X. Ye, Image tag completion via image-specific and tag-specific linear sparse reconstructions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2013, pp. 1618-1625.

[18] Z. Feng, S. Feng, R. Jin, A.K. Jain, Image tag completion by noisy matrix recovery, European Conference on Computer Vision, Springer2014, pp. 424-438.

[19] Y. He, C. Kang, J. Wang, S. Xiang, C. Pan, Image tag-ranking via pairwise supervision based semi-supervised model, Neurocomputing, 167 (2015) 614-624.

[20] J.-W. Jeong, H.-K. Hong, D.-H. Lee, i-TagRanker: an efficient tag ranking system for image sharing and retrieval using the semantic relationships between tags, Multimedia Tools and Applications, 62 (2013) 451-478.

[21] A.R. Zamir, M. Shah, Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs, IEEE transactions on pattern analysis and machine intelligence, 36 (2014) 1546-1558.

[۲۲] ر. راد, برچسب‌زنی خودکار تصاویر بر مبنای تجزیه ماتریس نامنفی به صورت چند منظری، دانشکده کامپیوتر، دانشگاه صنعتی شریف، تهران، ۱۳۹۶.

[23] G. Pass, R. Zabih, Histogram refinement for content-based image retrieval, Applications of Computer Vision,. WACV'96., Proceedings 3rd IEEE Workshop on, 1996 IEEE1996, pp. 96-102.

[24] T. Deselaers, D. Keysers, H. Ney, Features for image retrieval: an experimental comparison, Information Retrieval, 11 (2008) 77-107.

[25] H.G. Feichtinger, T. Strohmer, Gabor analysis and algorithms: Theory and applications, Springer1998.

[26] C.S. Won, Feature extraction and evaluation using edge histogram descriptor in mpeg-7, Advances in Multimedia Information Processing-PCM 2004, Springer2005, pp. 583-590.

[27] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 521 (2015) 436-444.

[28] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Transactions on Signal and Information Processing, 3 (2014).

[29] L. Deng, D. Yu, Deep learning: methods and applications, Foundations and Trends® in Signal Processing, 7 (2014) 197-387.

[30] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, Proceedings of the 26th annual international conference on machine learning, ACM2009, pp. 609-616.

[31] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634, (2013).

[32] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE transactions on pattern analysis and machine intelligence, 37 (2015) 2531-2544.

[33] M. Ivasic-Kos, I. Ipsic, S. Ribaric, A knowledge-based multi-layered image annotation system, Expert systems with applications, 42 (2015) 9539-9553.

[34] R. Shekhar, C. Jawahar, Word image retrieval using bag of visual words, Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, IEEE2012, pp. 297-301.

[35] C.-F. Tsai, Bag-of-words representation in image annotation: A review, ISRN Artificial Intelligence, 2012.

[36] Y. Jia, C. Huang, T. Darrell, Beyond spatial pyramids: Receptive field learning for pooled image features, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE2012, pp. 3370-3377.

[37] M. Li, J. Lui, B. Wang, Z. Li, W.-Y. Ma, Dual cross-media relevance model for image annotation, Google Patents2013.

[38] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, Proceedings of the international conference on Multimedia, ACM2010, pp. 251-260.

[39] D. Putthividhy, H.T. Attias, S.S. Nagarajan, Topic regression multi-modal latent dirichlet allocation for image annotation, Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE2010, pp. 3408-3415.

[40] D. Tian, X. Zhao, Z. Shi, An Efficient Refining Image Annotation Technique by Combining Probabilistic Latent Semantic Analysis and Random Walk Model, Intelligent Automation & Soft Computing, (2014), 1-11.

[41] D. Tian, X. Zhao, Z. Shi, Refining image annotation by integrating PLSA with random walk model, Advances in Multimedia Modeling, Springer2013, pp. 13-23.

[42] M.M. Kalayeh, H. Idrees, M. Shah, NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE2014, pp. 184-191.

[43] R. Rad, M. Jamzad, Automatic image annotation by a loosely joint non-negative matrix factorisation, IET Computer Vision, 9 (2015) 806-813.

[44] R. Rad, M. Jamzad, Image annotation using multi-view non-negative matrix factorization with different number of basis vectors, Journal of Visual Communication and Image Representation, 46 (2017) 1-12.

[45] R. Rad, M. Jamzad, A multi-view-group non-negative matrix factorization approach for automatic image annotation Multimedia tools and applications, (2017).

[46] L. Ballan, T. Uricchio, L. Seidenari, A. Del Bimbo, A cross-media model for automatic image annotation, Proceedings of International Conference on Multimedia Retrieval, ACM2014, pp. 73.

[47] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering, 26 (2014) 1819-1837.

[48] Y. Verma, C. Jawahar, Exploring SVM for Image Annotation in Presence of Confusing Labels, Proceedings of the 24th British Machine Vision Conference2013.

[49] B. Zhou, V. Jagadeesh, R. Piramuthu, Conceptlearner: Discovering visual concepts from weakly labeled image collections, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2015, pp. 1492-1500.

[50] T. Jinhui, L. Haojie, G.J. Qi, T.S. Chua, Image Annotation by Graph-Based Inference With Integrated Multiple/Single Instance Representations, Multimedia, IEEE Transactions on, 12 (2010) 131-141.

[51] H. Richang, W. Meng, G. Yue, T. Dacheng, L. Xuelong, W. Xindong, Image Annotation by Multiple-Instance Learning With Discriminative Feature Mapping and Selection, Cybernetics, IEEE Transactions on, 44 (2014), 669-680.

[52] X. Ding, B. Li, W. Xiong, W. Guo, W. Hu, B. Wang, Multi-instance multi-label learning combining hierarchical context and its application to image annotation, IEEE Transactions on Multimedia, 18 (2016) 1616-1627.

[53] P. Savita, D. Patel, A. Sinhal, A Neural Network Approach to Improve the Efficiency of Image Annotation, International Journal of Engineering Research and Technology, ESRSA Publications2013.

[54] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R.M. Summers, Learning to read chest X-rays: recurrent neural cascade model for automated image annotation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 2497-2506.

[55] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, Computer Vision, 2009 IEEE 12th International Conference on, IEEE2009, pp. 309-316.

[56] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, Computer Vision—ECCV 2008, Springer2008, pp. 316-329.

[57] L. Wu, E. Chen, Q. Liu, L. Xu, T. Bao, L. Zhang, Leveraging tagging for neighborhood-aware probabilistic matrix factorization, Proceedings of the 21st ACM international conference on Information and knowledge management, ACM2012, pp. 1854-1858.

[58] J. Johnson, L. Ballan, L. Fei-Fei, Love thy neighbors: Image annotation by exploiting image metadata, Proceedings of the IEEE International Conference on Computer Vision2015, pp. 4624-4632.

[59] Y. Verma, C. Jawahar, Image annotation using metric learning in semantic neighbourhoods,

Computer Vision–ECCV 2012, Springer2012, pp. 836-849.

[60] Y. Verma, C. Jawahar, Image annotation by propagating labels from semantic neighbourhoods, International Journal of Computer Vision, 121 (2017) 126-148.

[61] A. Bahrololoum, H. Nezamabadi-pour, A multi-expert based framework for automatic image annotation, Pattern Recognition, 61 (2017) 169-184.

[62] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, S.-F. Chang, Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation, arXiv preprint arXiv:1709.01220, (2017).

[63] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, arXiv preprint arXiv:1312.4894, (2013).

[64] V.N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using deep learning representations, Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM2015, pp. 603-606.

[65] M. Jiu, H. Sahbi, Nonlinear Deep Kernel Learning for Image Annotation, IEEE Transactions on Image Processing, 26 (2017) 1820-1832.

[66] R. Salakhutdinov, G. Hinton, Deep boltzmann machines, Artificial Intelligence and Statistics2009, pp. 448-455.

[67] M. Fang, S.-h. LV, K.-x. ZHENG, J. Chi, C. Fei, Y. Ke, D. Yong, Image Annotation by Object Hypotheses-oriented Deep Neural Networks, DEStech Transactions on Computer Science and Engineering, (2017).

[68] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal of Machine Learning Technologies, 2 (2011) 37-63.

[69] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE transactions on pattern analysis and machine intelligence, 29, (2007), 394-410.

رویا راد مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۷۹ از دانشگاه صنعتی امیرکبیر و مدرک کارشناسی ارشد و دکترای خود را در دانشگاه صنعتی شریف و در گرایش هوش مصنوعی در سالهای ۱۳۸۱ و ۱۳۹۶ دریافت کرد. وی از سال ۱۳۸۴ به عضویت هیئت علمی دانشگاه آزاد واحد پرند در آمده است.

منصور جم‌زاد مدرک کارشناسی ارشد علوم کامپیوتر از دانشگاه مک گیل، کانادا و دکترا در رشته مهندسی کامپیوتر از دانشگاه واسدا، توکیو، ژاپن. از سال ۱۳۷۴ بعنوان عضو هیت علمی در دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف مشغول بکار می‌باشد. دروس اصلی که تدریس نموده پردازش تصویر و بینایی ماشین است. زمینه‌های اصلی تحقیقاتی مورد علاقه ایشان برچسب‌گذاری تصاویر، بازیابی تصویر مبتنی بر محتوا، نشانه‌گذاری در تصاویر، پنهان‌نگاری، تشخیص تومورهای سرطان در تصاویر، ردگیری و کاربردهای صنعتی بینایی‌ماشین است.