

الحمد لله
الرحمن الرحيم!



دانشگاه آزاد اسلامی
واحد تهران جنوب
دانشکده فنی و مهندسی

پایان نامه برای دریافت درجه کارشناسی ارشد. M.Sc
مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز

عنوان :
بازیابی ریزدانه‌ای تصویر مبتنی بر محتوا

استاد راهنما :
جناب آقای دکتر کامبیز رهبر

پژوهشگر:
سید نیما سید آقا یزدی

آبان ۱۴۰۱



دانشگاه آزاد اسلامی
واحد تهران جنوب
دانشکده فنی و مهندسی

پایان نامه برای دریافت درجه کارشناسی ارشد. M.Sc
مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز

عنوان :
بازیابی ریزدانه‌ای تصویر مبتنی بر محتوا

استاد راهنما :
جناب آقای دکتر کامبیز رهبر

پژوهشگر:
سید نیما سید آقا یزدی

آبان ۱۴۰۱

تشکر و قدردانی

سپاس بیکران پروردگار یکتا را که هستی‌مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه‌چینی از علم و معرفت را روزیمان ساخت. بسیار ارزشمند بود فرصت‌هایی که توانستم از محضر استاد بزرگوارم جناب آقای دکتر کامبیز رهبر که همواره دلسوزانه و با جدیت، راهنما و راهگشای من در اکمال و اتمام پایان‌نامه بوده است، بهره گیرم و از آنان صمیمانه تشکر می‌نمایم.

همچنین از سرکار خانم ویدا صفردولابی و سرکار خانم فاطمه طاهری که مرا صمیمانه و مشفقانه در تدوین این پایان‌نامه یاری نموده، کمال تشکر و قدردانی می‌نمایم. و از خداوند متعال توفیق روزافزون را برای کلیه بزرگواران خواستارم.

تقدیم به:

پدران معنوی ام

مهندس محمدرضا کدخدازاده

و دکتر مسعود سیدآقایزدی

و خواهرم

نگار

۱	چکیده
۲	فصل اول: کلیات پژوهش
۳	۱-۱- مقدمه
۴	۱-۲- تعریف مسئله
۴	۱-۳- اهمیت و ضرورت انجام پژوهش
۴	۱-۴- کاربردها
۵	۱-۴-۱- کاربردها در پزشکی
۵	۱-۴-۲- کاربردها در صنعت فروش
۶	۱-۴-۳- کاربردها در گیاهشناسی
۶	۱-۴-۴- کاربردها در جانورشناسی
۷	۱-۴-۵- کاربردها در هنر
۷	۱-۵- فرضیه‌ها
۷	۱-۶- پرسش‌ها
۷	۱-۷- هدف و نوآوری
۷	۱-۸- متغیر پژوهش
۷	۱-۹- روش اجرای پژوهش
۹	۱-۱۰- جمع‌بندی
۱۰	فصل دوم: مروری بر پیشینه پژوهش
۱۱	۱-۲- مقدمه
۱۲	۲-۲- روش‌های بازیابی تصویر مبتنی بر متن
۱۲	۲-۲-۱- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از شبکه عصبی مصنوعی
۱۳	۲-۲-۲- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از ترکیب ویژگی
۱۳	۲-۲-۳- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از هیستوگرام

۳-۲-۳	روش‌های بازیابی تصویر مبتنی بر محتوا.....	۱۳
۳-۲-۱	روش‌های بازیابی تصویر مبتنی بر محتوا با استفاده از ترکیب ویژگی.....	۱۴
۳-۲-۲	روش‌های بازیابی تصویر مبتنی بر محتوا با استفاده از شبکه عصبی مصنوعی.....	۱۵
۳-۲-۳	روش‌های بازیابی تصویر مبتنی بر محتوا با استفاده از ترکیب ویژگی و شبکه عصبی.....	۱۶
۲-۴-۲	روش‌های بازیابی تصویر مبتنی بر طرح.....	۱۶
۱-۴-۲	روش‌های بازیابی تصویر مبتنی بر طرح با استفاده از صفرشات.....	۱۷
۲-۴-۲	روش‌های بازیابی تصویر مبتنی بر طرح با استفاده از تقویت طرح دستی.....	۱۷
۳-۴-۲	روش‌های بازیابی تصویر ریزدانه‌ای مبتنی بر طرح.....	۱۸
۵-۲	روش‌های بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا.....	۱۹
۱-۵-۲	روش‌های بازیابی ریزدانه‌ای با استفاده از ترکیب ویژگی‌ها.....	۱۹
۲-۵-۲	روش‌های بازیابی ریزدانه‌ای با استفاده از شبکه‌های عصبی.....	۲۰
۳-۵-۲	روش‌های بازیابی تصویر ریزدانه‌ای با استفاده از ترکیب ویژگی و شبکه عصبی مصنوعی.....	۲۴
۶-۲	جمع‌بندی.....	۲۶
	فصل سوم: روش پیشنهادی.....	۲۷
۱-۳	مقدمه.....	۲۸
۲-۳	روش بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا با استفاده از شبکه خودتوجهی مکانی.....	۲۹
۱-۲-۳	استخراج‌کننده ویژگی.....	۲۹
۲-۲-۳	خود توجهی مکانی.....	۳۱
۳-۲-۳	بازیابی تصویر.....	۳۲
۳-۳	راهکار پیشنهادی.....	۳۲
۴-۳	جمع‌بندی.....	۳۵
	فصل چهارم: ارزیابی و راست‌آزمایی آزمایشگاهی.....	۳۶
۱-۴	مقدمه.....	۳۷
۲-۴	معرفی پایگاه‌داده.....	۳۸

۳-۴- معیارهای ارزیابی	۳۹
۴-۴- نتایج ارزیابی	۳۹
فصل پنجم: جمع‌بندی و پیشنهادات آینده	۴۴
۲-۵- جمع بندی و نتیجه گیری	۴۵
۳-۵- پیشنهادات آینده	۴۵
منابع	۴۶

چکیده

بازیابی تصاویر، دسته‌بندی دقیق تصاویر، با استفاده از شباهت‌ها و تفاوت‌های موجود در بافت، رنگ، فرم و سایر ویژگی‌های تصویر است. بازیابی تصویر برای پرس‌وجوی مبتنی بر تصویر شامل رویکردهای متفاوتی است که می‌توان آن‌ها را در سه دسته عمده بیان نمود: بازیابی تصویر مبتنی بر طرح، بازیابی تصویر مبتنی بر محتوا و بازیابی تصویر مبتنی بر ریزدانه. در این مقاله شبکه خودتوجهی مکانی پیشنهاد شده‌است که شامل دو جزء اصلی می‌باشد. ابتدا یک شبکه عصبی کانولوشنی به‌عنوان استخراج‌کننده ویژگی پیاده‌سازی می‌شود که ویژگی‌های اولیه را از تصاویر ورودی از طریق چندین لایه کانولوشن استخراج می‌کند. سپس ماژول خودتوجهی مکانی با استفاده از مکانیسم توجه، ویژگی‌های جدید را ذخیره می‌کند. یکی از مشکلات روش استفاده از شبکه خودتوجهی مکانی آن است که تصویر ورودی، با ویژگی‌های با اهمیت کم‌تر بررسی می‌شود و ممکن است بخش‌های حاشیه‌ای در نتیجه نهایی عملکرد شبکه، تأثیرگذار باشند. در این پژوهش، روش XRAI برجسته‌سازی پیشنهاد شده‌است. این روش با امتیاز ۸۸ درصد توانسته‌است نتیجه‌ای قابل توجه نسبت به سایر روش‌های بازیابی تصویر داشته‌باشد.

کلمات کلیدی: بازیابی تصویر – بازیابی ریزدانه‌ای تصویر – بازیابی مبتنی بر محتوا

فصل اول: کلیات پژوهش

امروزه با به رسمیت شناختن تکنولوژی‌های مربوط به هوش مصنوعی و همچنین سنجش توانمندی‌های این تکنولوژی‌ها در حوزه تصویر، می‌توان بیان کرد که جستجو در میان تصاویر، به اندازه جستجو در میان متون، حائز اهمیت گشته است. از این رو، روش‌های بسیاری برای پردازش تصاویر معرفی گشته است. یکی از مهم‌ترین شاخه‌های پردازش تصویر، بازیابی تصاویر می‌باشد. بازیابی تصاویر، دسته‌بندی دقیق تصاویر، با استفاده از شباهت‌ها و تفاوت‌های موجود در بافت، رنگ، فرم و سایر ویژگی‌های تصویر است. این شاخه از علم پردازش تصویر، برای اولین بار در سال ۱۹۷۰ با رویکرد مبتنی بر متن^۱ معرفی گردید. پس از آن رویکردی متفاوت با عنوان مبتنی بر محتوا^۲ معرفی گردد که بر اساس ویژگی‌های استخراج شده از تصاویر، کار می‌کرد. این رویکرد به سرعت جایگزین رویکرد پیشین شد و در حوزه‌های پزشکی، احراز هویت، پیشگیری از وقوع جرم، امنیت محیط و... مورد استفاده قرار گرفت. در این میان چالش‌های بسیاری به هنگام استفاده از روش‌های مبتنی بر این رویکرد، پیش می‌آمد. از جمله آنکه ویژگی‌های استخراج شده با ادراک انسان فاصله معنایی بسیاری داشتند. اما با انتخاب و استخراج درست ویژگی‌های مورد محاسبه، این فاصله کمتر به چشم آمده است. به گونه‌ای که اکنون با نیاز به بررسی دقیق‌تر دسته‌بندی تصاویر، بازیابی تصاویر ریزدانه‌ای^۳ معرفی شده است که در پیدا کردن ویژگی‌های مشابه، تا حد ادراک انسان رفتار می‌کند.

¹ Text Based Image Retrieval

² Content Based Image Retrieval

³ Fine-Grained Content Based Image Retrieval

۱-۲- تعریف مسئله

با پیشرفت تکنولوژی و افزایش روزافزون داده‌های تصویری در حوزه‌های مختلف، بحث مدیریت و بازیابی تصویر، به یکی از موضوعات موردتوجه در حوزه پردازش تصویر تبدیل شده است. بازیابی تصویر همان یافتن تصاویر مشابه از مجموعه بزرگی از داده‌های تصویری است. دو رویکرد اصلی برای حل مسئله بازیابی تصویر عبارت‌اند از: مبتنی بر متن و مبتنی بر محتوا. در رویکرد مبتنی بر محتوا، هنگامی که جستجوی ما داخل یک طبقه‌بندی ریزتر انجام گیرد، بازیابی ریزدانه‌ای نامیده می‌شود. به طور مثال پیدا کردن مدل تصاویر مربوط به یک مدل ماشین، از میان یک مجموعه داده، شامل تصاویر ماشین با مدل‌های مختلف. یکی از مهم‌ترین چالش‌های بازیابی ریزدانه‌ای تصویر، روش‌های استخراج ویژگی است. این روش‌ها مبتنی بر دو رویکرد هستند: ۱- الگوریتم‌های محاسبه بردار ویژگی تصویر، ۲- استفاده از یادگیری ماشین و شبکه‌های عصبی. روش‌های حاضر دارای چند مشکل هستند و مورد اول اینکه در روش‌های مرسوم، تمرکز الگوریتم‌ها، روی مقایسه میزان تفاوت بخش‌های مختلف تصویر است. از آنجاکه می‌بایست بازیابی به صورت ریزدانه‌ای صورت گیرد، تصاویر تفاوت کمتری داشته و زمان زیادی برای محاسبه این تفاوت‌ها صرف می‌شود. مورد دوم اینکه در بعضی تصاویر، کیفیت پایین بوده و بردارهای ویژگی استخراج شده، دارای مقادیر متفاوت از مقادیر اصلی هستند. در این تحقیق سعی داریم با بررسی روش‌های استخراج ویژگی، به بهینه‌سازی این روش‌ها بپردازیم.

۱-۳- اهمیت و ضرورت انجام پژوهش

تصاویر نقش مهمی در انتقال اطلاعات دارند. با پیشرفت سریع فناوری رایانه، میزان داده‌های تصاویر دیجیتال به سرعت در حال افزایش است. نیاز اجتناب‌ناپذیری به روش‌های کارآمدی وجود دارد که می‌تواند به جستجو و بازیابی اطلاعات بصری تصاویر کمک کند.

بازیابی تصاویر از آن‌رو حائز اهمیت است که حجم زیادی از محتوای در دسترس را شامل می‌شود. باتوجه به آنکه پردازش متن گاهی با کژتابی همراه است، پیش‌بینی می‌شود جهت بررسی مطالب و دسته‌بندی آنها، با تصاویر به کاررفته در میان محتوا، چه در کتاب‌ها، فایل‌های PDF و چه در صفحات وب، بتوان به ارتباط دو یا چند مطلب پی برد. این امر وقتی مهم‌تر خواهد شد که جستجوی موردنیاز، در مسائل کمی عمیق‌تر گشته و بازیابی ریزدانه‌ای تصویر، ما را به سطوح عمیق‌تری از جستجو، هدایت کند. از سوی دیگر، پیدا کردن شباهت میان یک زیر دسته‌بندی می‌تواند ما را در ترمیم و پیش‌بینی بخش‌های آسیب‌دیده از تصاویر مشابه، راهنمایی کند.

۱-۴- کاربردها

بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا کاربردهای زیادی در صنایع مختلف اعم از تولید، فروش و ... علوم زیستی شامل پزشکی، گیاه‌شناسی، جانورشناسی و ... و هنر از جمله موارد مربوط به زیبایی‌شناسی و از همه مهم‌تر هنرهای تجسمی دارد. در هر حوزه پیدا کردن شباهت میان نمونه‌های تصویری مورد بررسی، می‌تواند وظایف مربوط به جست‌وجو را سریع‌تر و کم‌هزینه‌تر انجام دهد.

۱-۴-۱- کاربردها در پزشکی

در [۱] استفاده از جنگل‌های مسیر بهینه (بدون نظارت و با نظارت) و رویکردهای یادگیری فعال را برای بازخورد مرتبط در سیستم‌های بازیابی تصویر پزشکی مبتنی بر محتوا بررسی می‌کند. آموزنده‌ترین تصاویری که با رویکرد یادگیری فعال انتخاب می‌شوند، آن‌هایی هستند که بهترین تعادل را بین شباهت (با تصویر پرس‌وجو) و درجات خاصی از تنوع و عدم قطعیت ارائه می‌دهند. مدل یادگیری و کاربر به طور فعال در فرایند انتخاب آموزنده‌ترین تصاویر برای استفاده در آموزش، بهبود پرس‌وجو و بازگرداندن تصاویر مشابه بیشتر شرکت می‌کنند.

در [۲] یک روش ترکیبی جدید به نام تحلیل همبستگی ریزدانه^۴ برای بازیابی تصویر پزشکی پیشنهاد می‌شود. ابتدا، این مشکل را تجزیه و تحلیل می‌کند که بسیاری از مناطق محلی نامربوط در یک دسته وجود دارد. برای حل این مشکل، یک تصویر به چند نمونه ریزدانه تقسیم می‌شود. سپس، نمونه‌های ریزدانه با ویژگی‌های مشابه با همان برچسب توسط الگوریتم خوشه‌بندی k-means برچسب‌گذاری می‌شوند. در نهایت، بررسی شده است که چگونه رابطه همبستگی استخراج‌شده از نمونه‌های ریزدانه به ترکیب ویژگی‌های مختلف و به دست آوردن اطلاعات متمایزتر و کمتر اضافی برای بازیابی تصویر پزشکی کمک می‌کند.

در [۳] یک طرح بازیابی تصویر پزشکی برون‌سپاری سریع‌تر با حفظ حریم خصوصی^۵ پیشنهاد می‌شود. این طرح ابتدا یک طبقه‌بندی ساده برای تصاویر پزشکی برون‌سپاری می‌کند که دامنه بازیابی را محدود می‌کند و کارایی بازیابی را در مقایسه با طرح‌های بازیابی طبقه‌بندی‌نشده موجود بهبود می‌بخشد. دوم، یک کنترل دسترسی سبک‌وزن را برای هر کلاس با استفاده از استراتژی کنترل دسترسی مبتنی بر چندجمله‌ای پیاده‌سازی می‌کند که کنترل دسترسی دقیق را برای حفاظت از حریم خصوصی بهتر تصاویر پزشکی فراهم می‌کند. سوم تداخل اعداد تصادفی در امتیاز مربوطه را به صفر کاهش می‌دهد که دقت بازیابی را بیشتر بهبود می‌بخشد.

۱-۴-۲- کاربردها در صنعت فروش

در [۴] یک معماری انتها به انتها^۶ شامل یک شبکه مولد متقابل^۷ برای رسیدگی به تغییر دامنه در زمان آموزش و یک شبکه عصبی کانولوشنی عمیق آموزش‌دیده بر روی نمونه‌های تولید شده توسط شبکه برای یادگیری جاسازی تصاویر محصول که سلسله‌مراتبی را بین دسته‌های محصول اعمال می‌کند، پیشنهاد می‌شود. در زمان آزمایش، با استفاده از جستجوی k-امین همسایه نزدیک در برابر پایگاه داده‌ای که فقط از یک تصویر مرجع در هر محصول تشکیل شده است، شناسایی را انجام می‌دهد.

در [۵] که هدف آن ایجاد یک روش بازیابی تصویر برای مشخص کردن دسته‌بندی یک محصول است، یک مدل شبکه کانولوشنی سیامی^۸ پیشنهاد می‌شود که شامل برچسب‌های دسته و آیتم در آموزش برای تولید

^۴ fine-grained correlation analysis

^۵ Faster outsourced Medical Image Retrieval

^۶ End To End

^۷ Generative Adversarial Network

^۸ Siamese Convolutional Network

ویژگی آگاه از دسته است. این مدل با اصلاح رویه آموزشی همراه است که به طور همزمان دسته و برچسب مورد را یاد می‌گیرد. این شبکه با استفاده از یک مجموعه داده به عنوان ستون فقرات و شبکه تک‌لایه برای یادگیرنده با ویژگی متوسط پیاده‌سازی می‌شود.

۱-۴-۳- کاربردها در گیاه‌شناسی

در [۶] عملکرد یک مجموعه داده و چندین معماری شبکه عصبی کانولوشنال دیگر را هنگامی که برای ترکیبی از مجموعه داده‌ها اعمال می‌شود، ارزیابی می‌کند. برای نرمال‌سازی تأثیر عدم تعادل ناشی از ترکیب مجموعه داده‌های اصلی، از روش‌های یادگیری بیش از حد، نمونه‌برداری کم، و انتقال برای ساخت یک طبقه‌بندی‌کننده شبکه عصبی کانولوشنی سرتاسر استفاده شده است. به جای تأکید بر عملکرد بالا در هر یک از مجموعه داده‌های اصلی، تأکید بیشتری بر معیارهای مناسب برای مجموعه داده‌های نامتعادل متنوع وجود داشته است.

در [۷] بردارهای ویژگی به دست آمده از معماری‌های منفرد به هم متصل می‌شوند تا یک بردار ویژگی نهایی را تشکیل دهند. سپس ویژگی‌های استخراج شده با استفاده از طبقه‌بندی‌کننده‌های یادگیری ماشین^۹ مانند تحلیل تفکیک خطی^{۱۰}، رگرسیون لجستیک چندجمله‌ای^{۱۱}، درخت طبقه‌بندی و رگرسیون، k-نزدیک‌ترین همسایه^{۱۲}، طبقه‌بندی‌کننده جنگل تصادفی^{۱۳}، طبقه‌بندی‌کننده کیسه‌ای^{۱۴} و پرسپترون چند لایه طبقه‌بندی می‌شوند.

۱-۴-۴- کاربردها در جانورشناسی

در [۸] اثبات می‌شود که انتخاب توصیف‌گرهای عمیق مفید به خوبی به تشخیص تصویر با دانه‌ریز کمک می‌کند. به طور خاص، یک مدل جدید شبکه عصبی کانولوشنی ماسک دار^{۱۵}، بدون لایه‌های کاملاً متصل پیشنهاد شده است. بر اساس حاشیه‌نویسی‌های بخش، مدل پیشنهادی شامل یک شبکه کاملاً کانولوشنی برای مکان‌یابی قسمت‌های متمایز (مانند سر و تنه)، و مهم‌تر از آن تولید ماسک‌های جسم/قطعه وزن دار برای انتخاب توصیف‌گرهای کانولوشنی مفید و معنادار است. پس از آن، یک مدل سه‌جریانی برای جمع‌آوری توصیف‌گرهای انتخاب شده در سطح شیء و بخشی به طور همزمان ساخته می‌شود. به لطف کنارگذاشتن پارامتر لایه‌های کاملاً متصل اضافی، این شبکه ما دارای ابعاد کوچک و سرعت استنتاج کارآمد در مقایسه با سایر روش‌های ریزدانه است.

⁹ Machin Learning

¹⁰ Linear Discriminant Analysis

¹¹ Multinomial Logistic Regression

¹² K-Nearest Neighbours (KNN)

¹³ Random Forest Classifier

¹⁴ Bagging Classifier

¹⁵ Mask-Convolutional Neural Network

۱-۴-۵- کاربردها در هنر

در [۹] یک روش تخمین ریزدانه برای تخمین نمره زیبایی‌شناسی پیشنهاد می‌شود و مکانیسم‌های توجه موقعیت و کانال را برای افزایش ترکیب ویژگی‌های زیبایی‌شناسی ترکیب می‌کند. با آموزش شبکه رگرسیون جدا از شبکه طبقه‌بندی، وظیفه طبقه‌بندی را مکمل تکلیف رگرسیون می‌کند. محققان به استفاده از میانگین مربع خطا^{۱۶} به عنوان معیار ارزیابی اصلی عادت کرده‌اند، که در اندازه‌گیری خطای هر بازه ناکافی است. به منظور در نظر گرفتن کامل تصاویر، بخش‌های مختلف امتیاز زیبایی‌شناختی، به جای تمرکز بر بخش‌های نمره زیبایی‌شناختی متوسط به دلیل عدم تعادل مجموعه داده‌های زیبایی‌شناختی، یک معیار ارزیابی جدید به نام خطاهای میانگین مربع تقسیم شده^{۱۷} برای اثبات مزایا پیشنهاد می‌شود.

۱-۵- فرضیه‌ها

- فرض می‌شود تصاویر مجموعه داده مورد استفاده بدون نویز و آسیب مؤثر هستند.
- تصاویر در فرمت jpg بررسی می‌شوند.
- تصاویر دارای سه کانال رنگی قرمز، سبز و آبی هستند.

۱-۶- پرسش‌ها

آیا استفاده از رویکرد شبکه‌های عصبی به تنهایی برای استخراج ویژگی‌های متمایزکننده ریزدانه‌ای تصویر کافی است و یا تکنیک‌های بهینه کردن فیچرها لازم هست؟

۱-۷- هدف و نوآوری

بازایی دقیق تصاویر زیر طبقات یک کلاس از تصاویر. این هدف با آنالیز ریزدانه‌ای تصویر و استخراج و انتخاب ویژگی‌هایی از تصویر که تمایز دهنده تصاویر درون کلاسی باشد، انجام خواهد شد. سپس ویژگی‌هایی که می‌توانند سهم بیشتری در انجام پژوهش داشته باشند انتخاب، شده و آزمایش بر روی آن‌ها انجام خواهد گردید. همچنین روش دست‌یابی به ویژگی‌های قوی‌تر، بخشی از هدف این پژوهش می‌باشد که می‌بایست حین پژوهش بررسی شده و شناسایی ویژگی‌ها انجام گردد.

نوآوری:

۱-۸- متغیر پژوهش

الف) متغیر مستقل: مجموعه تصاویر مورد جستجو از مجموعه داده

ب) متغیر وابسته: مجموعه تصاویر بازایی شده و معیارهای ارزیابی عملکرد مدل بازایی تصویر

۱-۹- روش اجرای پژوهش

- داده‌های مورد نیاز برای راست آزمایی: پایگاه داده StanfordDogs و Cub_200_2011

¹⁶ Mean Square Errors

¹⁷ Segmented Mean Square Errors

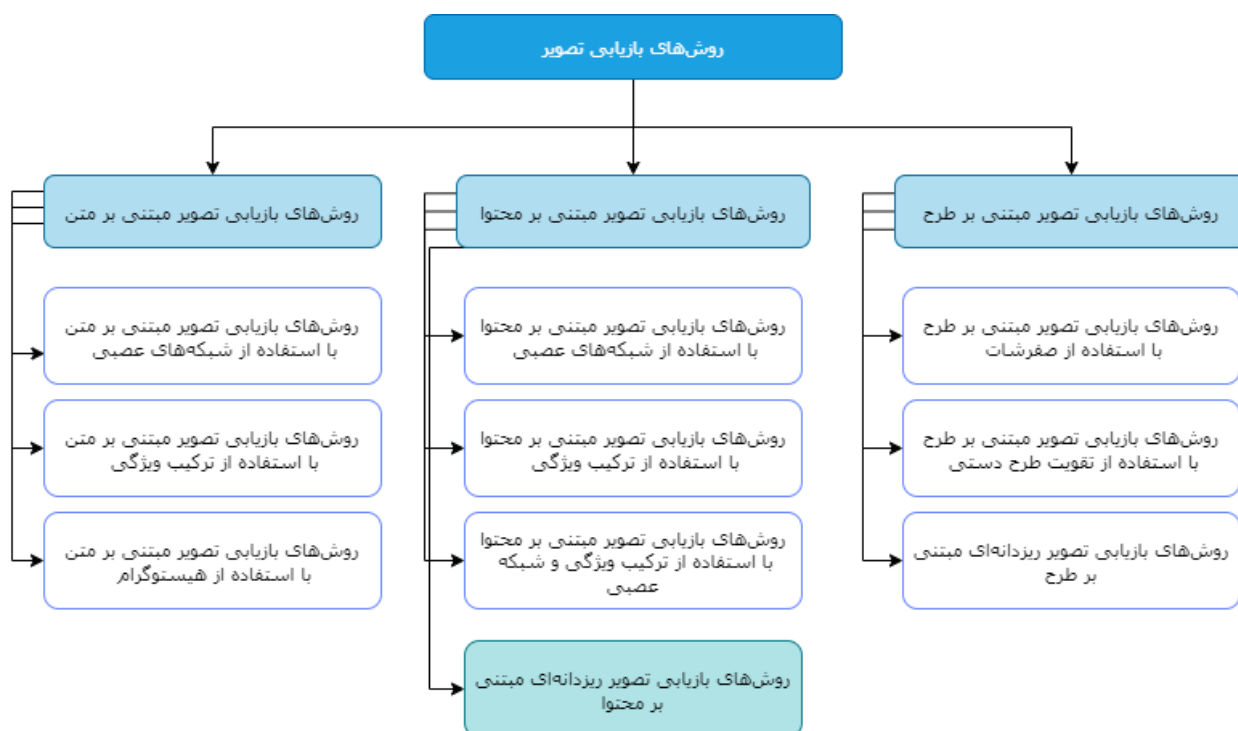
- روش تجزیه و تحلیل بر اساس یک تابع خطا ارزیابی خواهد شد.
- مدل‌های مورد استفاده در توصیف ویژگی‌های ورودی و استخراج نتایج خروجی مدل‌های بسته ریاضی و آماری خواهد بود.
- نرم‌افزار مورد استفاده در این پژوهش پایتون است.

۱-۱۰-جمع‌بندی

جستجو در میان تصاویر، به اندازه جستجو در میان متون، حائز اهمیت گشته است. از این رو، روش‌های بسیاری برای پردازش تصاویر معرفی گشته است. یکی از مهم‌ترین شاخه‌های پردازش تصویر، بازیابی تصاویر می‌باشد. این شاخه از علم پردازش تصویر، برای اولین بار در سال ۱۹۷۰ با رویکرد مبتنی بر متن معرفی گردید. پس از آن رویکردی متفاوت با عنوان مبتنی بر محتوا معرفی گردد که بر اساس ویژگی‌های استخراج شده از تصاویر، کار می‌کرد. این رویکرد به سرعت جایگزین رویکرد پیشین شد و در حوزه‌های پزشکی، گیاه‌شناسی، جانورشناسی، فروش و هنر مورد استفاده قرار گرفت. در این میان چالش‌های بسیاری به هنگام استفاده از روش‌های مبتنی بر این رویکرد، پیش می‌آمد. از جمله آنکه ویژگی‌های استخراج شده با ادراک انسان فاصله معنایی بسیاری داشتند. اما با انتخاب و استخراج درست ویژگی‌های مورد محاسبه، این فاصله کمتر به چشم آمده است. به گونه‌ای که اکنون با نیاز به بررسی دقیق‌تر دسته‌بندهای تصاویر، بازیابی تصاویر ریزدانه‌ای معرفی شده است که در پیدا کردن ویژگی‌های مشابه، تا حد ادراک انسان رفتار می‌کند.

فصل دوم: مروری بر پیشینه پژوهش

بازیابی تصویر شامل رویکردهای متفاوتی است که می‌توان آن‌ها را در سه دسته‌ی عمده‌ی بیان نمود: بازیابی تصویر مبتنی بر متن، بازیابی تصویر مبتنی بر محتوا و بازیابی تصویر مبتنی بر طرح. این دسته‌بندی‌ها هر کدام دارای زیرروش‌های مختلفی هستند که می‌توان آن‌ها را با توجه به انواع استخراج ویژگی، پردازش ویژگی و طبقه‌بندی تصاویر دسته‌بندی کرد. در ادامه به بررسی این دسته‌بندی‌ها مختصراً خواهیم پرداخت:



۲-۲- روش‌های بازیابی تصویر مبتنی بر متن^{۱۸}

تحقیقات در حوزه بازیابی تصویر از سال ۱۹۷۰ با رویکرد مبتنی بر متن آغاز گردید. در این تحقیقات، تصاویر به صورت دستی و با استفاده از توصیف‌گرهای متنی حاشیه‌نویسی می‌شدند. سپس یک سیستم مدیریت پایگاه داده، بازیابی تصویر را انجام می‌داد. در این روش کلمات کلیدی یا توضیحات برای توصیف محتوا مورد استفاده قرار می‌گرفت و مواردی همچون نام فایل، ابعاد، فرمت و محتوای تصویر را بیان می‌کردند. سپس جستجوی متنی روی پایگاه داده تصویری انجام می‌گردید و بر اساس ویژگی‌های حاشیه‌نویسی شده، تصاویر فیلتر می‌شدند و مجموعه‌ای از تصاویر که بیشترین شباهت به متن مورد جستجو داشتند، به عنوان پاسخ سیستم دریافت می‌شدند. این روش مشکلاتی را نیز شامل بود: (۱) حاشیه‌نویسی دستی برای پایگاه‌های داده بزرگ پرهزینه و بعضاً غیرممکن خواهد بود. (۲) چون حاشیه‌نویسی توسط انسان انجام می‌شد، جستجو نیز بر اساس درک انسان انجام می‌شد. (۳) این روش از زبان‌های مختلف پشتیبانی نمی‌کرد و تنها با یک زبان می‌توانست حاشیه‌نویسی صورت گیرد. (۴) خطاهای انسانی شامل ایرادات املائی و... روی پاسخ سیستم تأثیر بزرگی می‌گذاشتند.

۲-۲-۱- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از شبکه عصبی مصنوعی^{۱۹}

در [۱۰] یک روش جدید بازیابی متقابل مبتنی بر شبکه یادگیری دو جهته تصویر-متن پیشنهاد شده است. این روش یک فضای نمایش مشترک می‌سازد و به طور مستقیم شباهت داده‌های ناهمگن را اندازه‌گیری می‌کند. به طور خاص، یک شبکه نظارت چندلایه برای یادگیری ارتباط متقابل نمایش‌های تولید شده پیشنهاد شده است. علاوه بر این، یک تابع تلفات متقاطع دوطرفه برای حفظ تغییرناپذیری مودال با استراتژی یادگیری دوطرفه در فضای نمایش مشترک پیشنهاد شده است. توابع خطای سازگاری متمایز و خطای متقاطع دوطرفه در یک تابع هدف ادغام شده‌اند که هدف آن به حداقل رساندن فاصله درون کلاسی و به حداکثر رساندن فاصله بین طبقاتی است.

در [۱۱]، یک روش حمله پایان به انتهای کارآمد و ساده مبتنی بر شبکه‌های متخاصم مولد سازگار با چرخه پیشنهاد می‌شود. در مقایسه با مطالعات قبلی، این رویکرد به طور قابل توجهی هزینه برچسب‌گذاری داده‌ها را کاهش می‌دهد. علاوه بر این، این روش قابلیت حمل و نقل بالایی دارد. این می‌تواند به طرح‌های کپچا مبتنی بر متن معمولی فقط با تغییر چند پارامتر پیکربندی حمله کند که اجرای حمله را آسان‌تر می‌کند. ابتدا سینت سایبرهای کپچا^{۲۰} را بر اساس چرخه‌های شبکه مولد متقابل آموزش داده‌اند تا نمونه‌های جعلی تولید کنند. شناسه‌های اصلی مبتنی بر یک شبکه عصبی تکراری کانولوشنی با استفاده از داده‌های جعلی آموزش داده می‌شوند. متعاقباً، یک روش یادگیری انتقال فعال برای بهینه‌سازی شناسه گر اصلی با استفاده از مقادیر کمی از نمونه‌های کپچا در دنیای واقعی برچسب‌گذاری شده استفاده می‌شود.

¹⁸ Text Based Image Retrieval (TBIR)

¹⁹ Text-Based Image Retrieval using Neural Networks (NN-TBIR)

²⁰ Completely Automated Public Turing test to tell Computers and Human Apart

۲-۲-۲- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از ترکیب ویژگی^{۲۱}

در [۱۲] یک روش بازیابی متقابل رسانه‌ای مبتنی بر ترکیب چند ویژگی^{۲۲} پیشنهاد می‌شود. این روش قادر به ادغام چندین ویژگی برای ارتقای درک معنایی، و اتخاذ یادگیری خصمانه برای بهبود بیشتر دقت بازنمایی زیر فضای عمومی است. سپس از شباهت در همان فضا برای مرتب‌سازی نتایج بازیابی استفاده می‌شود. در [۱۳] یک رویکرد قطعی بازیابی تصویر مبتنی بر محتوا را پیشنهاد می‌کند که ویژگی‌های بصری و متنی را برای بازیابی تصاویر مشابه ترکیب می‌کند. در مرحله اول، این روش تصویر پرس‌وجو را به‌عنوان متنی و غیر متنی طبقه‌بندی می‌کند. اگر متنی در تصویر ظاهر شود، تصویر پرس‌وجو به‌عنوان متنی طبقه‌بندی می‌شود و متن به‌عنوان کیسه کلمات متنی شناسایی و تشکیل می‌شود. اگر تصویر پرس‌وجو به‌عنوان غیر متنی طبقه‌بندی شود، ویژگی‌های برجسته بصری استخراج شده و به‌صورت کیسه کلمات بصری شکل می‌گیرد. در مرحله بعد، این روش ویژگی‌های بصری و متنی را ترکیب می‌کند و تصاویر مشابه بالا بر اساس بردار ویژگی ترکیب شده بازیابی می‌شوند. از سه حالت بازیابی پشتیبانی می‌کند: پرس‌وجو تصویر، کلمات کلیدی و ترکیبی از هر دو.

در [۱۴] یک سیستم نمایه‌سازی مبتنی بر متن را برای بازیابی و طبقه‌بندی تصاویر ماموگرافی توصیف می‌شود. این سیستم با استخراج متن (گزارش‌های ساختاریافته) و تصاویر (ماموگرام) و طبقه‌بندی در یک بخش معمولی رادیولوژی سروکار دارد. گزارش‌های ساختاریافته، حاوی متن رایگان برای تشخیص پزشکی، تجزیه و تحلیل و برچسب‌گذاری شده‌اند تا تصاویر ماموگرافی مربوطه را طبقه‌بندی کنند. فرایند بازیابی اطلاعات بر اساس برخی تکنیک‌های دست‌کاری متن، مانند تحلیل معنایی سبک، حذف کلمات توقف، و پردازش زبان طبیعی پزشکی سبک است.

۲-۲-۳- روش‌های بازیابی تصویر مبتنی بر متن با استفاده از هیستوگرام^{۲۳}

در [۱۵] یک هیستوگرام جهت‌گیری زاویه جدید به نام هیستوگرام لبه زاویه معرفی شده است. با اعمال نظریه فیثاغورث بر تصویر، ویژگی‌های بسیار مفیدی برای تطبیق، جستجو و بازیابی تصویر به‌دست‌آمده است. روش پیشنهادی نیز با روش‌های موجود مقایسه شده است و نتایج نشان می‌دهد که در مقادیر دقت و فراخوان و تعادل دقت و فراخوان از روش‌های موجود بهتر عمل می‌کند.

۲-۳- روش‌های بازیابی تصویر مبتنی بر محتوا^{۲۴}

بازیابی تصویر مبتنی بر محتوا یک رویکرد از بازیابی تصویر است که مشکلات عمده‌ی رویکرد قبلی را نداشته و بیشتر به محتوای تصاویر یک پایگاه‌داده تصویری می‌پردازد. سیستم‌های CBIR سیستم‌هایی هستند که در دو مرحله به عملیات بازیابی تصویر می‌پردازند. (۱) طبقه‌بندی تصاویر با استفاده از ویژگی‌های سطح بالا و

²¹ Text-Based Image Retrieval using Feature Fusion (FF-TBIR)

²² Multi-feature Fusion based Cross-Media Retrieval

²³ Text-Based Image Retrieval using Histogram (H-TBIR)

²⁴ Content-based Image Retrieval (CBIR)

سطح پایین تصویر (فرایند آفلاین).^{۲۵} جستجوی کاربر بر اساس یک تصویر یا متن مشخص که با استفاده از طبقه‌بندی‌های انجام‌شده در مرحله پیشین، سیستم را به سمت پاسخ کاربر هدایت می‌کند. (فرایند آنلاین). یکی از موفق‌ترین نمونه‌های سیستم بازیابی تصویر مبتنی بر محتوا، جستجوی معکوس تصاویر موتور جستجوی گوگل است. سیستم‌های بازیابی تصویر مبتنی بر محتوا جهت انواع جستجوی مختلفی پیاده‌سازی می‌شوند:

(۱) **جستجو با یک مثال:** کاربر یک تصویر نمونه را به سیستم می‌دهد و سیستم تصاویر شبیه به آن را بر اساس ضوابط سطح پایین جستجو کرده و در پاسخ ارسال می‌نماید. تصویر نمونه می‌تواند توسط کاربر تهیه شده و یا از یک مجموعه تصادفی تصاویر انتخاب گردد.

(۲) **جستجو به کمک طرح:** کاربر یک شمای کلی از آنچه به دنبال آن است را کشیده و به سیستم می‌دهد. مثلاً با چند شکل که رنگ مشخصی دارند، پیش‌طرح یک گربه را آماده کرده و به سیستم ارسال می‌کند تا مجموعه‌ای از تصاویر گربه‌ها را دریافت کند.

(۳) **جستجو با مشخص کردن ویژگی:** مثلاً کاربر در جستجوی تصاویری با رنگ آبی است.

همچنین سامانه‌های بازیابی تصویر مبتنی بر محتوا می‌توانند با نمایش نتیجه به کاربر، از وی بخواهند که نسبت به صحت جستجو موضع را اعلام کند. مثلاً «مرتبط»، «نامرتبط» و یا «خنثی». بدین‌وسیله سیستم نسبت به نتایج خودآگاه شده و می‌تواند بخش‌هایی که منجر به پاسخ‌های نامرتبط شده‌اند را اصلاح نماید.

۲-۳-۱- روش‌های بازیابی تصویر مبتنی بر محتوا با استفاده از ترکیب ویژگی^{۲۵}

در [۱۶]، هدف از پژوهش دستیابی به نمایش تصویر مؤثر برای بازیابی تصویر به روشی بدون نظارت است. برای این منظور، یک روش جمع‌آوری وزنی متقاطع^{۲۶} برای بهبود استراتژی وزنی ادغام وزنی متقاطع^{۲۷} پیشنهاد می‌شود. به طور خاص، جمع‌آوری وزنی متقاطع هر دو بخش غیر صفر و بخش سفر لایه‌های کانولوشن را وزن می‌کند، باهدف به‌دست‌آوردن نمایش تصویر قوی. به طور خاص، ویژگی‌های چند مقیاسی استخراج‌شده توسط شبکه‌های عصبی کانولوشنال را با استفاده از جمع‌آوری وزنی متقاطع پیشنهادی، با در نظر گرفتن جنبه‌های متعدد ویژگی‌های بصری گرفته‌شده توسط شبکه‌ها، جمع‌آوری می‌شود. وزن‌های مختلفی را می‌توان به ویژگی‌های استخراج‌شده توسط لایه‌های مختلف شبکه اختصاص داد. برای کاهش تلاش برای تنظیم پارامتر، یک استراتژی اولیه برای هرس فضای جستجوی وزن‌ها پیشنهاد می‌شود که با طراحی قوانین محدودیت بر اساس دانش قبلی در مورد روابط بین لایه‌های شبکه به دست می‌آید. براین‌اساس، ترکیب ویژگی‌های چندلایه وزنی برای نمایش تصویر مشابه پیشنهاد می‌شود

در [۱۷] یک روش جدید بازیابی تصویر مبتنی بر محتوا پیشنهاد می‌شود. در مرحله توصیف تصویر، این روش ابتدا توصیفگر ریزساختار سنتی را اصلاح می‌کند تا رابطه مستقیم بین ویژگی‌های شکل و بافت و بین

²⁵ Content-Based Image Retrieval using Feature Fusion (FF-TBIR)

²⁶ Fully Cross-Dimensional Weighting

²⁷ Cross-Dimensional Weighting

ویژگی‌های رنگ و بافت را به تصویر بکشد. سپس هیستوگرام الگوهای باینری محلی یکنواخت^{۲۸} تصویر را استخراج می‌کند تا اطلاعات تفاوت رنگ را به تصویر بکشد. در مرحله مقایسه تصویر، روش ما ابتدا توصیفگرهای تصاویر را با هم مقایسه می‌کند تا شباهت آنها را محاسبه کند. سپس شباهت بین هر جفت تصویر با در نظر گرفتن شباهت‌های تصاویر قابل مقایسه در مجموعه داده به روزرسانی می‌شود. بر این اساس، این روش شباهت‌های نهایی تصاویر را به دست می‌آورد.

در [۱۸] این مقاله سفری را در میان اجزای اصلی ترکیب اطلاعات ارائه می‌دهد که یک دستورالعمل برای طراحی یک سیستم بازیابی تصویر مبتنی بر محتوا، باید شامل نیازهای تقاضای کاربران باشد.

در [۱۹] یک چارچوب چند وظیفه‌ای جدید مبتنی بر جداسازی و بازسازی ویژگی^{۲۹} برای بازیابی متقابل وجهی بر اساس روش‌های رایج یادگیری مکانی پیشنهاد می‌شود که مازول جداسازی ویژگی را برای مقابله با عدم تقارن اطلاعات بین روش‌های مختلف معرفی می‌کند و تصویر را معرفی می‌کند و مازول بازسازی متن برای بهبود کیفیت مازول جداسازی ویژگی.

در [۲۰] مواد تشکیل دهنده برای تقویت رابطه بین تصاویر غذا و دستورالعمل‌ها معرفی می‌شوند، زیرا می‌توانند منطق پخت‌وپز را تا حد زیادی منعکس کنند، و یادگیری دوگانه برای ارائه یک دیدگاه مکمل با بازسازی مواد از دستورالعمل‌های تولید شده، اتخاذ شده است. به منظور بهره‌برداری کامل از مواد تشکیل دهنده برای تولید دستورالعمل‌های مؤثر، مواد تشکیل دهنده در تصاویر و نام‌های غذا با مکانیزم توجه در جریان روبه‌جلو ترکیب می‌شوند و در جریان رو به عقب، یک بازسازی کننده طراحی شده است تا مواد را از دستورالعمل‌ها بازتولید کند. علاوه بر این، یادگیری تقویتی برای هدایت بازسازی مواد تشکیل دهنده برای حفظ ویژگی‌های مؤثر در اطلاعات ترکیب شده به طور صریح استفاده می‌شود.

۲-۳-۲- روش‌های بازیابی تصویر مبتنی بر محتوا با استفاده از شبکه عصبی مصنوعی^{۳۰}

در [۲۱]، یک چارچوب بازیابی تصویر مبتنی بر محتوا برای بیماری‌های پوستی پیشنهاد می‌شود که اطلاعات چند منبعی از جمله تصاویر درموسکوپ، تصاویر بالینی و اطلاعات متا را در بر می‌گیرد. چارچوب پیشنهادی ویژگی‌های چند منبعی را در سطح شباهت متقابل ترکیب می‌کند؛ بنابراین، حل مشکلات سوگیری ابعادی شدید برای اطلاعات تصویری و غیر تصویری. سپس از تجزیه و تحلیل جامعه مبتنی بر نمودار در شبکه‌های شباهت استفاده می‌شود که در آن تصاویر مشابه به شدت به هم متصل هستند و به بازیابی تصاویر مشابه با عملکرد بهبود یافته کمک می‌کنند.

این مطالعه [۲۲] بازیابی تصویر مبتنی بر محتوا را با یک شبکه عصبی سیامی کانولوشنی پیشنهاد می‌کند. ابتدا، تکه‌های ضایعه برای ایجاد دو مجموعه‌های داده برش داده می‌شوند و جفت‌های دوتکه دلخواه یک مجموعه داده پچ-جفت را تشکیل می‌دهند. دوم، این مجموعه داده پچ-جفت برای آموزش یک شبکه استفاده می‌شود. سوم، یک پچ آزمایشی به عنوان یک پرس‌وجو در نظر گرفته می‌شود. فاصله بین این پرس‌وجو و ۲۰

²⁸ Uniform local binary patterns

²⁹ multi-task framework based on feature separation and reconstruction

³⁰ Content-Based Image Retrieval using Neural Network (NN-CBIR)

وصله در هر دو مجموعه داده با استفاده از شبکه عصبی کانولوشنی سیامی آموزش دیده محاسبه می شود. وصله های نزدیک به پرس و جو برای ارائه پیش بینی نهایی با رأی اکثریت استفاده می شود. در [۲۳] ابتدا قصد کاربر با آموزش یک شبکه عصبی کانولوشن بر اساس ویژگی های زمانی و مکانی استخراج شده از داده های ردیابی چشم وی که هنگام بازرسی ارتباط بین تصاویر مختلف جمع آوری شده است، مدل سازی می شود. با استفاده از ویژگی های بصری به عنوان پل، درجه ارتباط با تصویر جستجوی هر یک از تصاویر پایگاه داده با مدل قصد کاربر با انتقال داده های حرکت چشم از شبیه ترین تصویر از نظر بصری در میان تصاویری که به طور تکراری در بوم انباشته شده اند، محاسبه می شود. سیستم بازایی پیشنهادی به شیوه ای تکراری اجرا می شود. در هر دور تکرار، داده های حرکت چشم کاربر هنگام بازرسی سیستم جمع آوری می شود و مجموعه بوم تصاویر نیز با ضمیمه کردن آن توسط سیستم بازرسی شده توسط کاربر، به روزرسانی می شود. با مجموعه های بوم به روزرسانی شده، می توان میزان ارتباط تصاویر پایگاه داده را مجدداً محاسبه کرد و سیستم می تواند جستجوی دور جدیدی از مرتبط ترین تصاویر را آغاز کند.

در [۲۴] یک چارچوب بازایی تصویر پزشکی مبتنی بر محتوا بدون نظارت بر اساس تطبیق مکانی کلمات بصری پیشنهاد می کند. روش پیشنهادی به طور مؤثر شباهت مکانی کلمات بصری را با استفاده از یک معیار شباهت جدید به نام شاخص شباهت پرس محاسبه می کند. آزمایش ها روی سه مجموعه داده پزشکی بزرگ نتایج امیدوارکننده ای را نشان می دهد. همبستگی مبتنی بر مکان کلمات بصری به بازایی دقیق تر و کارآمدتر تصاویر پزشکی متنوع و چندوجهی نسبت به سیستم های پیشرفته بازایی تصویر پزشکی مبتنی بر محتوا کمک می کند.

در [۲۵] روشی را پیشنهاد می شود که از قدرت شبکه های عصبی کانولوشن برای پیش بینی عضویت کلاس تصویر پرس و جو برای همه کلاس های خروجی و بازایی تصاویر با استفاده از تابع فاصله تغییر یافته در فضای ویژگی موجه استفاده می کند.

۲-۳-۳- روش های بازایی تصویر مبتنی بر محتوا با استفاده از ترکیب ویژگی و شبکه عصبی^{۳۱}

در [۲۶] یک تکنیک استخراج ویژگی ترکیبی با ترکیب ویژگی های سطح بالا و ویژگی های سطح پایین برای بهبود استحکام بردار ویژگی پیشنهاد می کند. مدل پیشنهادی از مدل گوگل نت از قبل آموزش دیده به عنوان استخراج کننده ویژگی و ترکیب با ویژگی های بافت چند مقیاسی گابور استفاده می کند. بردار ویژگی نهایی برای بازایی داده های تصویر مربوطه از مجموعه داده تصویر در مقیاس بزرگ استفاده خواهد شد.

۲-۴- روش های بازایی تصویر مبتنی بر طرح^{۳۲}

بازایی تصویر مبتنی بر طرح یک دسته از روش های بازایی تصاویر است که کار آن بازایی تصویر، از یک پایگاه داده تصویر طبیعی است که با یک شکل طراحی شده توسط طرح مطابقت دارد. در حالت ایده آل، یک مدل مبتنی بر طرح باید یاد بگیرد که اجزای طرح (مثلاً پا، دم، و غیره) را با اجزای مربوطه در تصویر که ویژگی های شکل مشابهی دارند مرتبط کند. روش های ارزیابی فعلی صرفاً فقط بر ارزیابی درشت دانه تمرکز

³¹ Content-Based Image Retrieval using Feature Fusion and Neural Network (NN-FF-CBIR)

³² Sketch Based Image Retrieval (SBIR)

می‌کنند، از آن جا که تمرکز بر بازیابی تصاویری است که به همان طرح یا مشابهات آن تعلق دارند، اما لزوماً دارای ویژگی‌های شکل مشابه در طرح نیستند. در نتیجه، روش‌های موجود به‌سادگی یاد می‌گیرند که طرح‌ها را با کلاس‌هایی که در طول آموزش بررسی می‌شوند، مرتبط کنند و از این‌رو در تعمیم به کلاس‌های بررسی نشده شکست می‌خورند.

۲-۴-۱- روش‌های بازیابی تصویر مبتنی بر طرح با استفاده از صفرشات^{۳۳}

در [۲۷] یک مدل بازیابی تصویر مبتنی بر طرح چالش‌برانگیزتر با نام صفرشات را بررسی می‌کند که در آن دسته‌های آزمایشی در مرحله آموزش ظاهر نمی‌شوند. پس از درک این موضوع که طرح‌ها عمدتاً حاوی اطلاعات ساختار هستند، درحالی‌که تصاویر حاوی اطلاعات ظاهری اضافی هستند، سعی می‌شود از طریق گسستگی نامتقارن^{۳۴} به بازیابی آگاهانه از ساختار رسید. برای این منظور، روش جداسازی نامتقارن آگاه از ساختار^{۳۵} پیشنهاد می‌شود که در آن ویژگی‌های تصویر به ویژگی‌های ساختار و ویژگی‌های ظاهری تفکیک می‌شوند درحالی‌که ویژگی‌های طرح تنها به فضای ساختار، پیش‌بینی می‌شوند. از طریق جداسازی ساختار و فضای ظاهری، ترجمه دامنه دوجهته بین حوزه طرح و حوزه تصویر انجام می‌شود.

در [۲۸] یک چارچوب ساده و کارآمد را پیشنهاد می‌کند که به منابع آموزشی محاسباتی بالایی نیاز ندارد و فضای تعبیه معنایی را از یک مدل بینایی به‌جای یک مدل زبان، همان‌طور که توسط مطالعات مرتبط انجام شده، یاد می‌گیرد. علاوه بر این، در مراحل آموزش و استنتاج این روش تنها از یک شبکه عصبی کانولوشنی استفاده می‌شود. در این کار، یک شبکه عصبی کانولوشنی از پیش آموزش‌دیده (یعنی ResNet۵۰) با سه هدف یادگیری پیشنهادی تنظیم شده است: خطای چهارگانه متعادل دامنه^{۳۶}، خطای طبقه‌بندی معنایی^{۳۷}، و خطای حفظ دانش معنایی^{۳۸}. با در نظر گرفتن بازیابی تصویر مبتنی بر طرح صفرشات، به‌عنوان یک مشکل تشخیص و تأیید شی، تلفات طبقه‌بندی معنایی و چهارگانه متعادل شده دامنه برای یادگیری ویژگی‌های متمایز، معنایی و نامتغیر دامنه معرفی می‌شوند. برای حفظ دانش معنایی آموخته شده با ImageNet و بهره‌برداری از آن برای مقوله‌های دیده نشده، خطای حفظ دانش معنایی آموخته شده با ImageNet و هزینه محاسباتی و افزایش دقت فرایند تقطیر دانش معنایی، قبل از آموزش، دانش معنایی حقیقت پایه به‌صورت کلاس محور تهیه می‌شود.

۲-۴-۲- روش‌های بازیابی تصویر مبتنی بر طرح با استفاده از تقویت طرح دستی^{۳۹}

در [۲۹] یک روش جدید تقویت داده‌های ویژه طرح^{۴۰} پیشنهاد می‌شود که از کمیت و کیفیت طرح‌ها به‌طور خودکار استفاده می‌کند. از جنبه کمیت، یک استراتژی تغییر شکل مبتنی بر محور منحنی^{۴۱} برای غنی‌سازی

³³ Zero-Shot Sketch based Image Retrieval (ZS-SBIR)

³⁴ Asymmetric Disentanglement

³⁵ STRucture-aware Asymmetric Disentanglement (STRAD)

³⁶ Domain-Balanced Quadruplet Loss

³⁷ Semantic Classification Loss

³⁸ Semantic Knowledge Preservation Loss

³⁹ Sketch-Based Image Retrieval using freehand Sketch Enhancement (SE-SBIR)

⁴⁰ Sketch-Specific Data Augmentation

⁴¹ Bezier Pivot Based Deformation

داده‌های آموزشی معرفی می‌شود. به‌منظور بهبود کیفیت، یک رویکرد بازسازی ضربه متوسط^{۴۲} برای تولید مجموعه‌ای از انواع جدید طرح‌ها با واریانس‌های درون کلاسی کوچک‌تر ارائه شده است. هر دوی این راه‌حل‌ها از هرگونه داده چند منبعی و نشانه‌های زمانی طرح‌ها بی‌حد هستند. علاوه بر این، نشان داده می‌شود که برخی از مدل‌های شبکه عصبی کانولوشنی عمیق اخیر که بر روی کلاس‌های عمومی تصاویر واقعی آموزش داده شده‌اند، می‌توانند انتخاب‌های بهتری نسبت به بسیاری از معماری‌های پیچیده‌ای باشند که به‌صراحت برای تشخیص طرح طراحی شده‌اند.

۲-۴-۳- روش‌های بازیابی تصویر ریزدانه‌ای مبتنی بر طرح^{۴۳}

در [۳۰] بر مشکلات اصلی بازیابی تصویر ریزدانه‌ای مبتنی بر طرح تأکید می‌شود: الف. چگونه می‌توان تفاوت بین ناهمگن رسانه‌های ناهمگن را کاهش داد؟ و ب. چگونه می‌توان تشخیص ویژگی‌های طرح را بهبود بخشید؟ به‌طور خاص، یک مدل تولید طرح برای اولین بار برای جایگزینی پیش‌پردازش معمولی لبه‌های تصویر تقریباً استخراج شده پیشنهاد می‌شود، علاوه بر این، این مدل می‌تواند معضل کمبود داده‌های طرح را کاهش دهد. سپس یک مدل جدید ریزدانه‌ای مبتنی بر طرح ایجاد می‌شود که از شبکه عصبی کانولوشنی تغییر شکل‌پذیر بهره می‌برد و درعین حال ویژگی‌های معنایی را با هم در نظر می‌گیرد. علاوه بر این، برای اولین بار یک مجموعه داده طرح-تصویر لباس ریزدانه ساخته می‌شود که دارای حاشیه‌نویسی ویژگی‌های غنی است.

در [۳۱] هدف این است که عکس موردنظر از یک طرح جزئی با کمترین تعداد ضربه ممکن بازیابی شود. این روش به‌عنوان بازیابی تصویر ریزدانه‌ای مبتنی بر طرح در پرواز نامیده می‌شود که در آن بازیابی پس از هر بار کشیدن نقاشی آغاز می‌شود. در نظر گرفته می‌شود که ارتباط معنی‌داری بین این طرح‌های ناقص در قسمت‌های طراحی-طراحی هر عکس وجود دارد. یک روش یادگیری چند دانه‌بندی-تداعی پیشنهاد می‌شود که فضای جاسازی همه طرح‌های ناقص را برای یادگیری یک فضای جاسازی مشترک کارآمد بهینه می‌کند. به‌طور خاص، بر اساس یکپارچگی طرح، یک اپیزود طرح کامل را می‌توان به چند مرحله تقسیم کرد که هر کدام مربوط به یک لایه نگاشت خطی ساده است. علاوه بر آن، این چارچوب نمایش فضای برداری طرح فعلی را برای تقریب آن با طرح‌های بعدی آن راهنمایی می‌کند. به این ترتیب، عملکرد بازیابی یک طرح با ضربه‌های کمتر می‌تواند به یک طرح با ضربه‌های بیشتر نزدیک شود.

در [۳۲] بازیابی تصویر ریزدانه‌ای مبتنی بر طرح به‌عنوان یک فرایند درشت به ریز فرموله شده است و یک مدل رتبه‌بندی متقابل آبخاری عمیق^{۴۴} پیشنهاد می‌شود که می‌تواند از تمام اطلاعات چندوجهی مفید در طرح‌ها و تصاویر حاشیه‌نویسی بهره‌برداری کند و کارایی بازیابی را بهبود بخشد هدف بر ساختن بازنمایی‌های عمیق برای طرح‌ها، تصاویر و توضیحات و یادگیری همبستگی‌های عمیق بهینه شده در چنین حوزه‌های مختلف متمرکز است؛ بنابراین برای یک طرح پرس‌وجو داده شده، تصاویر مربوطه آن با شباهت‌های ریز در سطح نمونه در یک دسته خاص می‌توانند برگردانده شوند و الزامات دقیق بازیابی سطح نمونه برآورده می‌شود.

⁴² Mean Stroke Reconstruction

⁴³ Fine-Grained Sketch-Based Image Retrieval (FG-SBIR)

⁴⁴ Deep Cascaded Cross-modal Ranking Model

در [۳۳] مجموعه‌ای از پیشرفته‌ترین روش‌های یادگیری زیرمکانی متقابل وجهی معرفی و مقایسه می‌شود و آنها را بر روی دو مجموعه داده بازیابی تصویر ریز دانه‌ای مبتنی بر طرح که اخیراً منتشر شده‌اند، محک می‌زند. از طریق بررسی کامل نتایج تجربی، نشان داده شده است که یادگیری زیرفضا می‌تواند به‌طور مؤثر شکاف دامنه طرح-عکس را مدل‌سازی کند. علاوه بر این، چند بینش کلیدی برای هدایت تحقیقات آینده ترسیم می‌شود.

۲-۵- روش‌های بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا^{۴۵}

یکی از مشکلات روش‌های امروزی در رویکرد مبتنی بر محتوا، آن است که طبقه‌بندی بر اساس تفاوت‌های بزرگ و واریانس زیاد انجام می‌گردد. به این صورت که هر نمونه از یک دسته، با هر نمونه دیگر از یک دسته‌ی دیگر دارای تفاوت زیاد محتوایی است، ولی در میان نمونه‌های یک دسته نیز تفاوت‌های محتوایی همچنان دیده می‌شود. روش ریزدانه‌ای تمرکز بر این دارد که هر دسته را به‌صورت افرازی از زیر دسته‌های مشابه‌تر توصیف می‌کند. یا حتی دسته‌بندی بر روی یک پایگاه داده تصویری انجام می‌شود که از نظر محتوایی، شباهت بالایی دارند. در این روش از متدهای مختلفی جهت استخراج ویژگی استفاده می‌شود که به‌طور کلی به دودسته تقسیم می‌شوند: (۱) استخراج ویژگی با استفاده از روش‌های سنتی و معمول و (۲) استخراج ویژگی با استفاده از به‌کارگیری شبکه‌های عصبی عمیق. در ادامه به بیان برخی از این روش‌ها، راهکارها و نگرش‌ها خواهیم پرداخت.

۲-۵-۱- روش‌های بازیابی ریزدانه‌ای با استفاده از ترکیب ویژگی‌ها^{۴۶}

در [۳۴] که روی طبقه‌بندی تصاویر گلوبول‌های سفید تمرکز دارد، یک سیستم یادگیری نیمه نظارت تهیه شده است. در این روش یک مکانیسم توجه تعاملی ریزدانه‌ای تعبیه شده که در ابتدا از تصاویر برچسب‌دار استفاده کرده و به تهیه بردارهای احتمالی حاصل از این تصویر، می‌پردازد. سپس داده‌های آموزشی بدون برچسب را با این بردارها مقایسه کرده و طبقه‌بندی می‌کند.

در [۳۵] یک ماژول مهار پیک و یک ماژول راهنمایی دانش را در یک ترانسفورماتور قرار داده که بیشتر به سرنخ‌های متمایز می‌پردازد. ابتدا ماژول مهار پیک یک سری نشانه برای هر تصویر ایجاد می‌کند و سپس ترانسفورماتور، توجه به بخش‌های متمایز را جریمه می‌کند؛ بنابراین بهره‌برداری از اطلاعات مناطق نادیده گرفته شده افزایش می‌یابد. سپس ماژول راهنمایی دانش، پاسخ به دست آمده را با مجموعه آموزشی به‌روز می‌کند.

در [۳۶] بافت کانال و اطلاعات توالی مکانی برای بازیابی مبتنی بر محتوا مورد تمرکز قرار می‌گیرند. ابتدا یک مدل عمیق جدید پیشنهاد می‌شود که هدف آن استنباط نقشه‌های توجه در امتداد بعد کانال و بعد مکانی است. با بهبود ماژول‌های توجه کانال و توجه مکانی و کاوش ترانسفورماتور، توانایی ساخت و درک مدل افزایش می‌یابد.

در [۳۷] یک مدل جدید شناسایی انسان را ارائه می‌دهد که از توزیع ماسک‌های دندان با تفاوت‌های محلی و ریزدانه‌ای بهره می‌برد. یک معماری انشعاب دوطرفه طراحی می‌شود که یکی از شاخه‌ها به‌عنوان استخراج‌کننده

⁴⁵ Fine-Grained Content Based Image Retrieval (FG-CBIR)

⁴⁶ Fine-Grained Content Based Image Retrieval using Feature Fusion (FF-FG-CBIR)

ویژگی تصویر و دیگری استخراج‌کننده ویژگی ماسک است. سپس ویژگی ماسک با ویژگی تصویر تعامل می‌یابند و وزن‌دهی صورت می‌گیرد. علاوه بر این یک مکانیسم توجه بهبودیافته برای تمرکز روی موقعیت‌های اطلاعاتی استفاده می‌شود.

در [۳۸] با اشاره به روش‌هایی که با خطای ویژگی‌های عمومی به استخراج ویژگی‌های متمایزتر کمک می‌کنند، یک تابع محاسبه خطای جدید به نام خطای متمرکز سخت ارائه می‌دهد. این تابع در استخراج ویژگی برای تمایز در تقسیم مشابه‌ترین دسته‌ها کمک می‌کند.

در [۳۹] یک معماری هرمی دو جهته موثر برای بهبود بازنمایی‌های داخلی ویژگی‌ها پیشنهاد می‌شود تا وظیفه تشخیص تصویر ریزدانه را در سناریوی یادگیری چند شات^{۴۷} انجام دهد. به طور خاص، یک هرم ویژگی چند مقیاسی و یک هرم توجه چند سطحی را در شبکه پایه مستقر می‌شود و به تدریج ویژگی‌ها را از فضا‌های دانه‌ای مختلف از طریق هر دوی آنها جمع‌آوری می‌کند. سپس یک استراتژی پالایش هدایت‌شده توجه^{۴۸} را با همکاری یک هرم توجه چند سطحی ارائه می‌شود تا عدم قطعیت ناشی از پس‌زمینه‌های مشروط به نمونه‌های محدود را کاهش دهد. علاوه بر این، روش پیشنهادی با چارچوب فرا یادگیری به‌صورت انتها به انتها بدون هیچ نظارت اضافی آموزش داده می‌شود.

در [۴۰] یک ژنراتور لنگر استخراج ویژگی محلی^{۴۹} جدید برای شبیه‌سازی اشکال ویژگی‌های نامنظم پیشنهاد می‌شود؛ بنابراین، ویژگی‌های متمایز را می‌توان به طور کامل در ویژگی‌های استخراج شده گنجاند. علاوه بر این، یک ماژول استخراج ویژگی محلی متقارن مؤثر^{۵۰} بر اساس مکانیزم توجه پیشنهاد شده است تا به طور کامل از رابطه مکانی بین ویژگی‌های محلی استخراج‌شده استفاده کند و ویژگی‌های متمایز را برجسته کند.

۲-۵-۲- روش‌های بازیابی ریزدانه‌ای با استفاده از شبکه‌های عصبی^{۵۱}

در [۴۱] که به طبقه‌بندی خودکار گیاهان پرداخته می‌شود، معماری‌های یادگیری عمیق را به دو دلیل وابستگی به مجموعه‌داده آموزشی بزرگ و عدم مقیاس‌پذیری به چالش می‌کشد. سپس از دو نمای یک برگ استفاده می‌کند تا ویژگی‌های عمومی^{۵۲} و محلی^{۵۳} تصویر را پوشش داده و با استفاده از شبکه عصبی مصنوعی کانولوشنال سیامی، وابستگی به داده آموزشی بزرگ را کاهش می‌دهد.

در [۴۲] به طبقه‌بندی گل‌های داوودی پرداخته می‌شود. برای انجام پژوهش، از یادگیری انتقالی و شبکه عصبی کانولوشن دوخطی استفاده می‌کند. از شبکه متقارن VGG۱۶ برای استخراج ویژگی بهره می‌گیرد و پس از آموزش به یک چارچوب پیشنهادی منتقل می‌کند. سپس ویژگی‌های عمومی را از دو شبکه گرفته و مورد بررسی قرار می‌دهد

⁴⁷ Few-Shot

⁴⁸ Attention-Guided Refinement Strategy

⁴⁹ Local Feature Extraction Anchor Generator

⁵⁰ Symmetrized Local Feature Extraction Module

⁵¹ Fine-Grained Image Retrieval using Neural Network (NN-FG-CBIR)

⁵² Global

⁵³ Local

در [۴۳] استفاده از معماری‌های شبکه‌های عصبی کانولوشنی عمیق در مجموعه داده‌های تصویری با بزرگ‌نمایی بالا، طبقه‌بندی دانه‌های برنج آسیب‌دیده را با دقت بالا امکان‌پذیر می‌سازد.

در [۴۴] یک شبکه بازیابی و استخراج اطلاعات متمایز به نام DRE-Net پیشنهاد می‌شود که با مشکل تشخیص تصویر با رزولوشن پایین رسیدگی می‌کند. این شبکه از دو شبکه فرعی تشکیل شده است: ۱- زیرشبکه بازیابی اطلاعات متمایز ریز^{۵۴} ۲- زیرشبکه شناسایی با رابطه معنایی خطای تقطیر^{۵۵}. ماژول اول با استفاده از ویژگی‌ها، به بازیابی جزئیات بافت حیاتی پیکسل‌ها کمک می‌کند. ماژول دوم به روابط صحیح بین هر دو پیکسل در نقشه ویژگی می‌پردازد. پس ماژول دوم می‌تواند به ماژول اول برای پیدا کردن جزئیات دقیق و قابل‌اعتماد کمک کند.

در [۴۵] روش یادگیری هش با دو مشکل بررسی می‌شود: ۱- ویژگی‌های با ابعاد کم فرایند بازیابی را تسریع می‌بخشند اما به دلیل ازدست‌رفتن اطلاعات، دقت را کاهش می‌دهند. ۲- تصاویر ریزدانه منجر به ایجاد کدهای هش جستجوی یکسان در خوشه‌های مختلف در فضای پنهان پایگاه داده می‌شوند. پس این پژوهش به یک شبکه پاک‌کننده توجه مبتنی بر ثبات ویژگی^{۵۶} می‌پردازد. برای مشکل نخست، از یک ماژول پاک کردن ناحیه انتخاب‌شده^{۵۷} استفاده می‌کند که با پوشش تطبیقی برخی از مناطق تصاویر خام، شبکه را در برابر تفاوت‌های ظریف ریزدانه‌ای مقاوم می‌کند. پس کدهای هش متمایزتری در پایگاه داده هش ذخیره می‌شوند. سپس برای پایدارتر کردن رابطه بین کد هش جستجو و کد هش پایگاه داده از ماژول افزایش خطای رابطه مکانی^{۵۸} استفاده می‌کند.

در [۴۶] با معرفی یک معماری جدید به نام شبکه دروازه سلسله‌مراتبی^{۵۹} از اتصال بین دسته‌های سلسله‌مراتبی بهره‌برداری می‌کند. این شبکه از یک مکانیسم شبیه به حافظه بلندمدت-کوتاه‌مدت برای انتقال وابستگی‌ها بین طبقات سطوح مختلف در سلسله‌مراتب دسته‌بندی استفاده می‌کند؛ بنابراین، اطلاعات زمینه، در ساختار سلسله‌مراتبی، برای تقویت عملکرد تشخیص استفاده می‌گردد.

در [۴۷] نیز به طبقه‌بندی سلسله‌مراتبی پرداخته شده است. از آنجاکه وظیفه اصلی سیستم به چند وظیفه فرعی تقسیم می‌شود تا ساختار برچسب درختی ارائه گردد، این وظایف فرعی همبستگی بالایی دارند. پس کارهای فرعی می‌توانند کاندیدهایی جهت قرارگیری در دسته‌بندی سطح پایین‌تر (ریزدانه) ارائه دهند که خود آن‌ها دارای ویژگی‌هایی هستند که سطح بالاتر (درشت‌دانه) را نیز توصیف می‌کنند؛ بنابراین ما می‌توانیم یک شبکه چند وظیفه‌ای عمیق مشترک برای طبقه‌بندی سلسله‌مراتبی تصویر ارائه دهیم. این پژوهش ابتدا ماتریس رابطه بین هر دو وظیفه فرعی تعریف شده توسط ساختار سلسله‌مراتبی برچسب را استخراج کرده و سپس هر یک از وظایف فرعی از طریق ماتریس رابطه، به تمامی وظایف فرعی مرتبط بخش می‌شود. در نهایت، برای ترکیب، یک تابع همجوشی جدید بر اساس ارزیابی کار و عدم قطعیت تصمیم، طراحی می‌گردد.

⁵⁴ Fine-Grained discriminative Information Restoration: FDR

⁵⁵ Semantic Relation Distillation Loss: SRD-Loss

⁵⁶ Feature Consistency Driven Attention Erasing Network: FCAENet

⁵⁷ Selective Region Erasing Module: SREM

⁵⁸ Enhancing Space Relation Loss: ESRL

⁵⁹ Hierarchical Gate Network: HGNet

در [۴۸] با اشاره به آنکه خط‌کشی برای تشخیص ریزدانه‌ای تصویر، هم ویژگی‌های میان‌طبقاتی و هم درون‌طبقاتی را به تصویر می‌کشد، بیان می‌دارد که روش‌های موجود عموماً از داده‌های کمکی برای هدایت شبکه استفاده می‌کنند. این مورد دو اشکال دارد: ۱- استفاده از داده‌های کمکی مانند جعبه‌ی محدودکننده^{۶۰} نیازمند دانش تخصصی دارد و شامل استفاده از داده‌های گران می‌شود. ۲- استفاده از چند شبکه فرعی موجب می‌شود شبکه پیچیده‌شده و آموزش نیز مراحل متعددی داشته باشد. در ادامه این مقاله یک شبکه خودتوجهی مکانی سراسری^{۶۱} معرفی می‌کند که شامل یک ماژول خودتوجهی مکانی^{۶۲} و یک تکنیک تقطیر خودتوجهی^{۶۳} است. ماژول اول اطلاعات متنی را در ویژگی‌های محلی رمزگذاری کرده و موجب بهبود نمایش درون کلاسی می‌شود. سپس ماژول دوم اطلاعات را از ماژول اول گرفته و به یک نقشه ویژگی اولیه تقطیر می‌کند. این مورد موجب نمایش بین طبقاتی می‌شود. با محاسبه خطای طبقه‌بندی از این دو ماژول، شبکه می‌تواند تا هر دو ویژگی بین کلاسی و درون کلاسی را آموزش ببیند.

در [۴۹] به یک طرح پیشنهادی برای طبقه‌بندی ریزدانه‌ای انواع محصولات خرده‌فروشی در قفسه سوپرمارکت‌ها پرداخته می‌شود. این طرح، به طور هم‌زمان، نشانه‌های سطحی شیء^{۶۴} و نشانه‌های سطحی بخشی از تصاویر محصول^{۶۵} را ضبط می‌کند. نشانه‌های سطح شیء تصاویر محصول توسط یک شبکه جدید طبقه‌بندی بازسازی^{۶۶} تولید می‌شود. برای مدل‌سازی بدون حاشیه‌نویسی نشانه‌های سطح جزئی، قسمت‌های تبعیض‌آمیز، تصاویر محصول در اطراف نقاط کلیدی شناسایی می‌شوند. این بخش‌ها به صورت توالی‌های مرتب‌شده توسط یک حافظه کوتاه‌مدت-بلندمدت کانولوشنی کدگذاری می‌شوند و محصولات را به طور منحصربه‌فرد توصیف می‌کنند.

در [۵۰] به یکی از مشکلات بازیابی تصویر ریزدانه‌ای می‌پردازد: تنوع کم در بین کلاس‌های مختلف و درعین حال تنوع زیاد در هر کلاس. این پژوهش با بررسی خطای آنتروپی متقاطع^{۶۷} برای ایجاد ویژگی متمایز شبکه عصبی کانولوشنال بیان می‌دارد با برخی عملیات اضافی مانند نرمال‌سازی مقیاس، می‌توان بهینه‌تر عمل کرد. سپس یک نوع خطای آنتروپی متقاطع جدید را به نام خطای آنتروپی متقاطع تکه‌ای^{۶۸} معرفی می‌کند که پیاده‌سازی آن بسیار آسان‌تر است.

در [۵۱] یک شبکه ترکیبی مبتنی بر خودتوجهی^{۶۹} برای یادگیری بازنمایی‌های رایج داده‌های رسانه‌های مختلف^{۷۰} پیشنهاد می‌شود. به طور خاص، ابتدا از یک لایه خودتوجهی محلی برای یادگیری فضای توجه مشترک بین داده‌های رسانه‌های مختلف استفاده می‌شود. سپس یک روش الحاق شباهت برای درک رابطه

⁶⁰ Bounding Box

⁶¹ End-to-end Spatial Self-Attention Network: SSANet

⁶² Spatial Self-Attention: SSA

⁶³ Self-attention distillation: Self-AD

⁶⁴ Object-level

⁶⁵ Part-level

⁶⁶ Reconstruction-Classification Network: RC-Net

⁶⁷ Cross Entropy Loss

⁶⁸ Piecewise Cross Entropy loss

⁶⁹ Self-Attention Network

⁷⁰ Cross-Media

محتوایی بین ویژگی‌ها پیشنهاد می‌شود. برای بهبود بیشتر استحکام مدل، یک کدگذاری موقعیت محلی را یاد می‌گیرد تا روابط مکانی بین ویژگی‌ها را ثبت کند؛ بنابراین، رویکرد پیشنهادی می‌تواند به طور مؤثر شکاف بین توزیع‌های ویژگی‌های مختلف در وظایف بازیابی بین رسانه‌ای را کاهش دهد.

در [۵۲] یک چارچوب پاک‌کننده قطعات با نظارت خود^{۷۱} برای دسته‌بندی بصری بسیار ریز. بینش کلیدی این مدل یادگیری نمایش‌های متمایز با کدگذاری یک ماژول تحت نظارت خود است که پاک‌کردن تصادفی بخش و پیش‌بینی موقعیت متنی قطعات پاک‌شده را انجام می‌دهد. این مورد، شبکه را به سمت بهره‌برداری از ساختار ذاتی داده‌ها سوق می‌دهد، یعنی درک و تشخیص اطلاعات متنی اشیاء، در نتیجه نمایش متمایزتر در سطح بخشی را تسهیل می‌کند. این همچنین با معرفی بخش‌های آموزشی متنوع‌تر با معنای معنایی، قابلیت یادگیری مدل را افزایش می‌دهد.

در [۵۳] یک شبکه پاک‌سازی پیش‌رونده^{۷۲} پیشنهاد می‌شود. در این شبکه، یک مکانیسم پاک‌سازی چند شبکه‌ای، نمونه‌های داده را افزایش می‌دهد و به ثبت ویژگی‌های متمایز محلی کمک می‌کند، جایی که ساختار کلی تصویر به طور غیرمستقیم از طریق پاک‌سازی پیکسلی تخریب می‌شود. تجمع ویژگی‌های متقاطع با استخراج ویژگی‌های کلاس برجسته از اهمیت زیادی در بازیابی بصری ریزدانه‌ای برخوردار است. با این حال، قابلیت نمایش ویژگی متقابل لایه بر اساس یک استراتژی تجمع ساده هنوز ناکارآمد است. برای این منظور، خطای سازگاری پیشنهادی، پیوند معنایی متقابل^{۷۳} لایه‌ای را بررسی می‌کند که بلوک انگیزه لایه‌ای متقابل^{۷۴} را برای استخراج نمایش‌های ویژگی کارآمدتر از دانه‌بندی‌های مختلف راهنمایی می‌کند. همچنین آنتروپی متقاطع و آنتروپی مکمل ادغام می‌شود تا توزیع طبقات منفی را برای عملکرد بهتر طبقه‌بندی در نظر بگیرد. در [۵۴]، مدلی به نام شبکه تراز مکانی محلی^{۷۵} برای اندازه‌گیری شباهت نمونه به کلاس از طریق تراز کردن مناطق مکانی محلی به روش اسکن پیمایشی پیشنهاد شده است. به طور خاص، تراز مکانی محلی با نمونه‌برداری مداوم از وصله‌های محلی از نقشه ویژگی پرس‌وجو به دست می‌آید، جایی که هر وصله محلی به عنوان یک هسته عمل می‌کند تا شبیه‌ترین وصله‌های محلی را از نقشه‌های ویژگی پشتیبانی فیلتر کند، و شباهت‌های سطح پیچ بین پرس‌وجو را به دست آورد. نمونه و کلاس‌های پشتیبانی سپس، یک ماژول جمع‌آوری اطلاعات پیشنهاد می‌شود تا شباهت‌های سطح پیچ را در امتیاز پیش‌بینی کلاس جمع‌آوری کند، در این صورت که وصله‌های مهم برجسته می‌شوند و پس‌زمینه‌ها رقیق می‌شوند.

در [۵۵] پیشنهاد می‌شود که امتیازات فیشر^{۷۶} با تمرکز بر دو عنصر در یک شبکه عمیق آموزش‌پذیر سرتاسر جاسازی شود: تطبیق رمزگذاری با ویژگی‌های عمیق و نرمال‌سازی آمارهای مرتبه دوم استخراج‌شده^{۷۷}؛ بنابراین، از یک ماژول کدگذاری پراکنده عمیق استفاده شده که اجازه می‌دهد مرکز هر تابع گاوسی از یک

⁷¹ Self-Supervised Part Erasing Framework (SPARE)

⁷² Progressive Erasing Network

⁷³ Cross-Layer Semantic Affinity

⁷⁴ Cross-Layer Incentive

⁷⁵ Local Spatial Alignment Network

⁷⁶ Fisher Scores

⁷⁷ Normalizing The Extracted Second-Order Statistics

زیرفضای آموخته شده نمونه برداری شود و بنابراین بهتر با توزیع داده‌های ابعادی بالا تطبیق داده شود. دوم، یک ماژول نرمال‌سازی جدید را معرفی شده که یک نرمال‌سازی ماتریس جذر تقریبی را که به خوبی با نمرات فیشر سازگار است، محاسبه می‌کند. این مراحل پردازش در یک شبکه عمیق تعبیه شده‌اند به طوری که همه ماژول‌ها تنها با هدف بهبود عملکرد طبقه‌بندی با هم کار می‌کنند.

در [۵۶]، بر اساس توابع خطای ساختار عمومی^{۷۸} و استراتژی کم‌افزایی سخت^{۷۹}، خطای کمینه نرم عمومی سخت^{۸۰}، برای بهبود عملکرد بازیابی شی ریزدانه پیشنهاد می‌شود. علاوه بر این، یک پارامتر قابل یادگیری به خطای پیشنهادی وارد می‌شود که به صورت پویا توسط شبکه در طول آموزش تنظیم می‌شود. بسیاری از آزمایش‌ها نشان می‌دهند که تابع ضرر پیشنهادی برای ارتقای عملکرد بازیابی مؤثر و مفید است.

در [۵۷] یک شبکه عصبی گراف^{۸۱} به کمک عنوان برای تطبیق تصویر-متن پیشنهاد می‌شود. به طور خاص، زیرنویس‌ها از تصاویر تولید می‌شوند و از آنها به عنوان اطلاعات کمکی برای ایجاد روابط با متن استفاده می‌شود؛ بنابراین شکاف دامنه را می‌توان به طور مؤثر آزاد کرد. به منظور یادگیری روابط ریز بین تصاویر، متون و شرح‌ها، از شبکه‌های عصبی گراف برای ایجاد روابط بین داده‌های ساخت یافته استفاده می‌شود، بنابراین تطابق قوی‌تری را ممکن می‌سازد.

۲-۵-۳- روش‌های بازیابی تصویر ریزدانه‌ای با استفاده از ترکیب ویژگی و شبکه عصبی مصنوعی^{۸۲}

در [۵۸] روشی برای استفاده از یک مدل توجه چند سطحی^{۸۳} پیشنهاد می‌شود. در ابتدا سه اندازه میدان پذیرش معمولی برای نقشه‌های توجه چند سطحی انتخاب می‌شوند. سپس یادگیری چندسطحی برای استخراج ویژگی‌های متمایز از این مناطق محلی معرفی می‌گردند. این روش نگرش جدیدی در مورد چگونگی استفاده از فعال‌سازهای شبکه عصبی، برای تولید مناطق چند مقیاسی - که برای طبقه‌بندی ریزدانه‌ای مفید هستند - ارائه می‌دهد و شامل دو مرحله است: ۱- انتخاب نورون‌هایی که حداکثر فعال‌سازی را در سه نقشه ویژگی انتخاب شده دارند. این نقشه‌ها خروجی مدل‌های شبکه عصبی کانولوشنی هستند که از قبل روی تصاویر اندازه کامل، آموزش داده شده‌اند. ۲- آموزش شبکه‌های ظریف با این مناطق چند مقیاسی ایجاد شده. هر منطقه متمایز شده را می‌توان به عنوان یکی از ویژگی‌ها در نظر گرفت. سپس این نتایج برای پیش‌بینی نهایی ادغام می‌شوند.

در [۵۹] یک چارچوب سبک‌تر برای نمونه‌برداری تدریجی از قطعات متمایز، جهت یادگیری جزئیات ارائه می‌شود. در این روش ابتدا شیء از تصویر اصلی تقویت شده و سپس یک نمونه‌برداری خودتطبیقی برای شناسایی بیشتر منطقه تقویت شده انجام می‌گردد. پس این چارچوب می‌تواند از کل به شیء و از شیء به جزئیات برسد. در این میان ویژگی‌های سلسله‌مراتبی نیز سنجیده می‌شوند که هزینه‌های محاسباتی را کاهش می‌دهد.

⁷⁸ Global Structure Loss functions

⁷⁹ Hard Mining Strategy

⁸⁰ Hard Global Softmin Loss

⁸¹ Caption-Assisted Graph Neural Network

⁸² Fine-Grained Content Based Image Retrieval using Feature Fusion and Neural Network (FF-NN-FG-CBIR)

⁸³ Multi-level Attention Model

در [۶۰] یک مکانیسم انتخاب ویژگی شیء‌گرا برای ویژگی‌های کانولوشن عمیق از یک شبکه عصبی کانولوشن از پیش آموزش دیده، پیشنهاد می‌شود. نقشه‌های ویژگی کانولوشن از یک لایه عمیق بر اساس تجزیه و تحلیل پاسخ آنها به اشیاء نظارتی انتخاب می‌شوند. ویژگی‌های انتخاب شده برای نمایش ویژگی‌های معنایی اشیاء نظارتی و بخش‌های آنها با حداقل تأثیر پس‌زمینه، عملاً نیاز به روش حذف پس‌زمینه قبل از استخراج ویژگی‌ها را از بین می‌برد. فعال‌سازی‌های میانگین لایه‌ای از نقشه‌های ویژگی‌های انتخابی، توصیفگر متمایز برای هر شیء را تشکیل می‌دهند. سپس این ویژگی‌های کانولوشنی شیء‌گرا^{۸۴} با استفاده از رویکردهای هش‌سازی حساس به محلی، بر روی فضای همینگ^{۸۵} کم‌بعد پیش‌بینی می‌شوند. کدهای هش باینری فشرده به دست آمده امکان بازیابی کارآمد را در مجموعه داده‌های مقیاس بزرگ فراهم می‌کنند.

⁸⁴ Object Oriented Convolutional Features

⁸⁵ Hamming

۲-۶- جمع‌بندی

بازیابی تصویر یکی از مسائل هوش مصنوعی است که در سه رویکرد مبتنی بر متن، مبتنی بر محتوا و مبتنی بر طرح، دنبال می‌گردد. هر کدام از این رویکردها در زیرروش‌های مختلفی دسته‌بندی می‌شوند که در نوع استخراج ویژگی، نوع پردازش ویژگی‌ها و طبقه‌بندی خروجی با هم تفاوت دارند. در سه رویکرد اصلی ویژگی‌های مختلف تصویر بررسی می‌گردند و هر روش با ترکیب این ویژگی‌ها به یافتن تصاویر مشابه تصویر ورودی سیستم می‌پردازد. یکی از زیرشاخه‌های مبتنی بر محتوا، بازیابی ریزدانه‌ای تصویر است که تا یک دسته پایین‌تر تشخیص را جلوتر می‌برد. هدف آن کم کردن میزان اختلاف میان ویژگی‌های هر دسته از تصاویر می‌باشد. در بازیابی ریزدانه‌ای از بردار ویژگی، شبکه‌های عصبی مصنوعی و یا ترکیب این دو استفاده می‌شود.

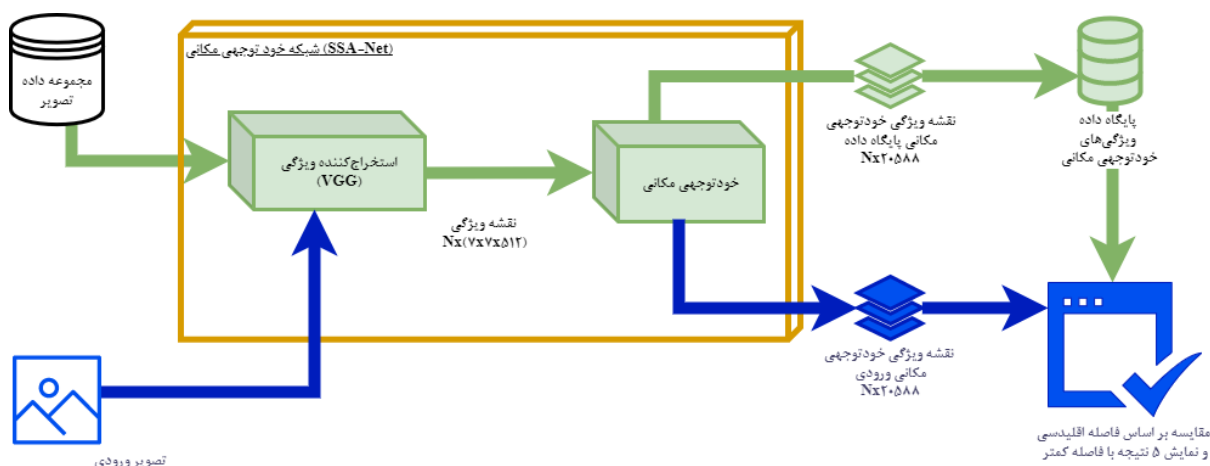
فصل سوم: روش پیشنهادی

در این فصل یک روش بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا^{۶۱} پیشنهاد می‌شود که با استفاده از یک شبکه عصبی کانولوشنی ویژگی‌های تصویر را استخراج می‌کند سپس یک ماژول خودتوجهی مکانی، ویژگی‌ها را پردازش و ذخیره می‌کند. در نهایت ماژول بازیابی تصویر، یک تصویر را از ورودی دریافت کرده، ویژگی‌های خودتوجهی مکانی آن را محاسبه می‌کند. سپس فاصله اقلیدسی این ویژگی‌ها با ویژگی‌های ذخیره شده در ماژول خودتوجهی مکانی محاسبه شده و دسته‌بندی تصاویری که کمترین فاصله اقلیدسی را با ویژگی‌های آموزش‌دیده دارند، به عنوان خروجی ارائه می‌گردند.

به عنوان راهکار پیشنهادی از روش XRAI شفاف‌سازی^{۶۲} استفاده می‌گردد که مناطق پراهمیت تصویر را شناسایی کرده و موجب می‌شود یک قاب از تصویر ورودی را استخراج کنیم. بدین صورت بخش‌هایی از تصویر که دارای اهمیت کمتری هستند مورد پردازش قرار نمی‌گیرند و عملکرد کلی شبکه بهبود می‌یابد.

۳-۲- روش بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا با استفاده از شبکه خودتوجهی مکانی

شبکه خودتوجهی مکانی^{۸۶} از دو جزء اصلی تشکیل شده است (شکل ۲ را ببینید). ابتدا یک شبکه عصبی کانولوشنی به‌عنوان استخراج‌کننده ویژگی^{۸۷} پیاده‌سازی می‌شود که ویژگی‌های اولیه را از تصاویر ورودی از طریق چندین لایه کانولوشن و ادغام استخراج می‌کند (به بخش ۳-۲-۱ مراجعه کنید)



شکل ۲. شبکه خودتوجهی مکانی

جدول ۱. اطلاعات دقیق مجموعه داده FGIR مورد استفاده برای آزمایش‌ها.

Dataset	Categories	Training	Testing
Stanford Dogs	۱۲۰	۲۰۵۷۰	۱۰
CUB_۲۰۰_۲۰۱۱	۲۰۰	۱۱۷۶۸	۲۰

در این روش از دیتاست Stanford Dogs برای بازیابی تصویر ریزدانه‌ای استفاده شده که شامل ۱۲۰ دسته‌بندی از نژادهای مختلف سگ می‌باشد. برای آموزش شبکه ۲۵۷۰ نمونه جدا شده‌اند و برای آزمایش عملکرد کل سیستم ۱۰ نمونه انتخاب شده‌اند.

همچنین برای راستی‌آزمایی بهتر از دیتاست CUB_۲۰۰_۲۰۱۱ نیز برای بازیابی تصویر ریزدانه‌ای استفاده شده که شامل ۲۰۰ دسته‌بندی مختلف پرند می‌باشد. برای آموزش شبکه ۱۱۷۶۸ نمونه جدا شده‌اند و برای آزمایش عملکرد کل سیستم ۲۰ نمونه انتخاب شده‌اند.

۳-۲-۱- استخراج‌کننده ویژگی

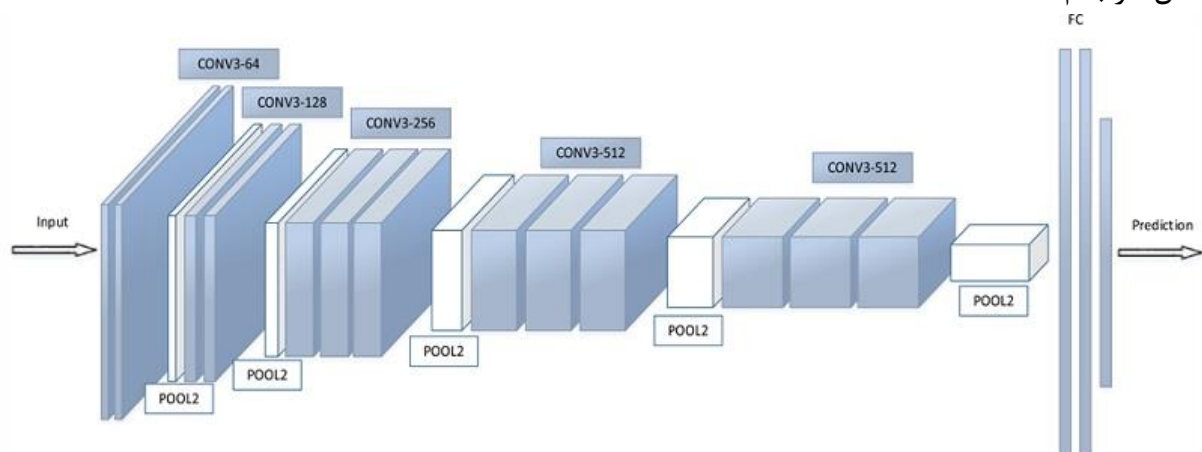
اخیراً، برای وظایف پردازش تصویر، یک رویکرد مرسوم برای استخراج ویژگی‌های اولیه، استفاده از یک شبکه عصبی کانولوشنی از قبل آموزش‌دیده به‌منظور بهره‌مندی از مقدار اولیه وزن معنادار است. چنین شبکه‌های عصبی کانولوشنی از پیش آموزش‌دیده‌ای می‌توانند ویژگی‌های سطح بالا را از تصاویر استخراج کنند. برای

⁸⁶ Spatial Self-Attention Network (SSA.Net)

⁸⁷ Feature Extractor: FE

مقایسه منصفانه با سایر روش‌های پیشرفته، از VGG-۱۶ از پیش آموزش‌دیده بر روی مجموعه داده ImageNet استفاده می‌شود.

شبکه VGG ۱۶ همان‌طور که در شکل ۳ نشان داده شده، شامل ۱۶ لایه کانولوشنی یا ۱۶ لایه پارامتری است. شبکه VGG ۱۶، شامل دو لایه کانولوشنی با ۶۴ فیلتر ۳×۳ هست که پشت سر هم قرار گرفته‌اند. سپس، یک لایه ماکس پولینگ ۲×۲ با پرش (Stride) به اندازه ۲ قرار گرفته است. این لایه ماکس پولینگ علاوه بر نمونه‌برداری، وظیفه کاهش بعد ویژگی‌ها به نصف را هم دارد. در ادامه، دو لایه کانولوشنی دیگر با ۱۲۸ فیلتر ۳×۳ و یک لایه ماکس پولینگ ۲×۲ و پرش ۲ قرار گرفته‌اند. به‌طور مشابه، سه لایه کانولوشنی با ۲۵۶ فیلتر ۳×۳ و یک لایه ماکس پولینگ ۲×۲ با پرش ۲ قرار گرفته‌اند. ۳ لایه کانولوشنی با ۵۱۲ فیلتر ۳×۳ و یک لایه ماکس پولینگ ادامه این شبکه هست که البته دو بار تکرار می‌شود. در نهایت، ویژگی‌ها تبدیل به یک بردار ویژگی می‌شوند تا در اختیار لایه‌های نورونی یا تمام‌متصل قرار گیرند. دو لایه نورونی به ابعاد ۴۰۹۶ پشت سر هم قرار گرفته‌اند. در نهایت، یک لایه نورونی به ابعاد ۱۰۰۰ که متناظر با تعداد کلاس‌های کاربرد ما هست، در نظر گرفته شده است. با توجه به اینکه پایگاه داده ImageNet شامل ۱۰۰۰ کلاس هست، در اینجا هم لایه خروجی شامل ۱۰۰۰ نورون است. در تمامی لایه‌های کانولوشنی و لایه‌های نورونی از تابع فعال‌ساز بنام RELU استفاده شده است.



شکل ۳. معماری شبکه VGG۱۶

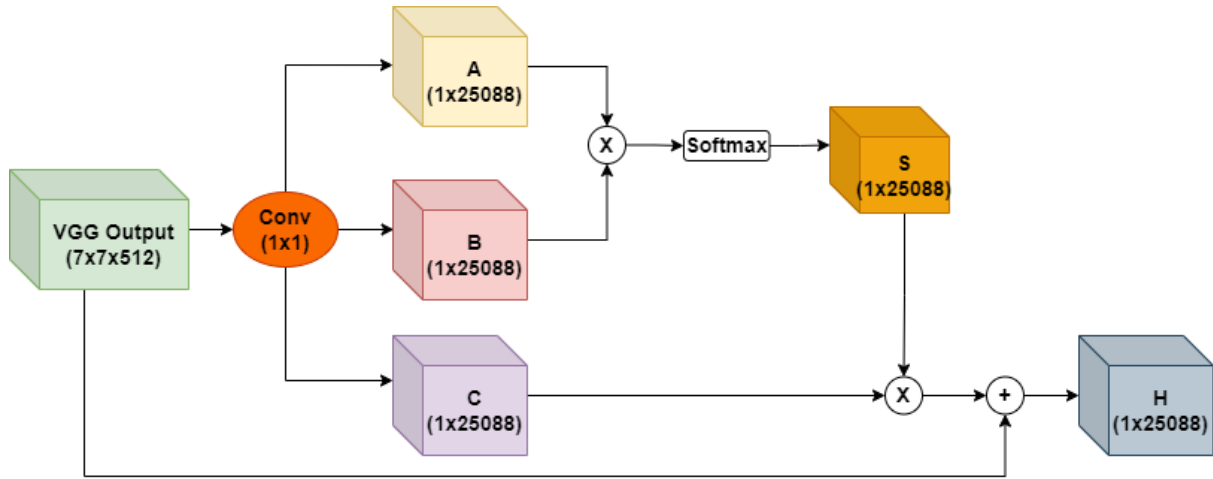
برای استخراج اولیه، سه لایه آخر که کاملاً متصل هستند حذف می‌شوند. ویژگی نقشه‌ها از تصاویر ورودی برای یک تصویر ورودی X یک مجموعه داده، خروجی نقشه ویژگی از لایه کانولوشنی نهایی گرفته می‌شود. این فرآیند به صورت زیر نشان داده شده است.

$$F = VGG(X) \quad (۱)$$

به طور خاص، استخراج‌کننده ویژگی یک تصویر ورودی X را به یک نقشه ویژگی ابعادی $F \in R^{H \times W \times K}$ نگاشت می‌کند، که در آن H ، W و K نشان دهنده ارتفاع مکانی، عرض مکانی، تعداد کانال ها/کرنل حاوی کانال هستند. این روند به ترتیب تا آخرین لایه پیش می‌رود.

۳-۲-۲- خود توجهی مکانی

ماژول خودتوجهی مکانی از مکانیزم خودتوجهی پیشنهاد شده استفاده می‌کند که توجه محلی را از طریق یک تابع softmax جمع می‌کند. این ایده گسترش می‌یابد تا به موقعیت‌های پیکسل مکانی ویژگی‌های اصلی توجه شود و از تجمیع ویژگی‌ها برای به‌دست‌آوردن نقشه‌های ویژگی خودتوجهی مکانی استفاده شود.



شکل ۳. ماژول خودتوجهی مکانی

همان‌طور که در شکل ۳ نشان داده شده است، باتوجه به نقشه‌های ویژگی اولیه $F \in R^{H \times W \times K}$ به دست آمده از استخراج‌کننده ویژگی، ابتدا سه نقشه ویژگی جدید A ، B و C با استفاده از کانولوشن 1×1 تولید می‌شود.

$\{A, B, C\} \in R^{H \times W \times K}$ همان ابعاد فضای F را داراست. سپس A و B و C را به $R^{N \times K}$ تغییر شکل می‌یابد، که در آن $N = H \times W$ تعداد پیکسل‌ها است. سپس، ضرب عناصر بین A و B ترانهاده B محاسبه می‌شود. "softmax" از نظر مکانی برای محاسبه نقشه خودتوجهی مکانی اعمال می‌شود $S \in R^{N \times N}$ که:

$$S_{ij} = \frac{\exp(A_i \otimes B_j)}{\sum_{i=1}^N \exp(A_i \otimes B_j)} \quad (2)$$

که در آن \otimes نشان‌دهنده ضرب عنصر است. S_{ij} نشان می‌دهد که چگونه شبکه تأثیر آمین موقعیت مکانی را بر موقعیت مکانی j اندازه‌گیری می‌کند. از این رو، بازنمایی ویژگی‌های مرتبط‌تر بین A و B منجر به همبستگی معنی‌دار و غنی‌تر بین آنها می‌شود و بالعکس. برای تقویت موقعیت‌های حضوری، ضرب عناصر بین $S \in R^{N \times N}$ و $C \in R^{N \times N}$ انجام می‌شود و نتایج به $R^{H \times W \times K}$ تغییر شکل داده می‌شود.

در نهایت، یک مکانیزم تجمیع ویژگی برای بررسی تأثیر مناطق خودتوجهی مکانی در همه موقعیت‌ها در نقشه ویژگی اصلی از طریق معادلات پیاده‌سازی می‌شود:

$$H_j = \sum_{i=1}^N (s_{ij} C_i) \oplus F_j \quad (3)$$

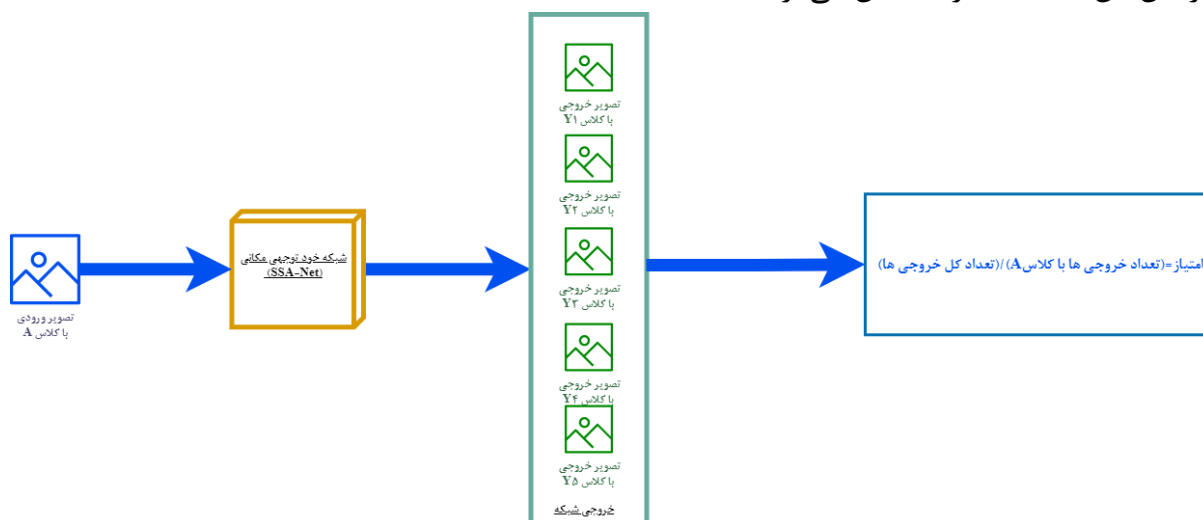
می‌توان از معادله (۳) استنباط کرد که ویژگی‌های به‌دست‌آمده توسط H_j نشان‌دهنده یک تجمع کلی از نمای زمینه‌ای بر اساس نقشه‌های خودتوجهی مکانی است. مجموعه این ویژگی‌ها به عنوان یک پایگاه داده ذخیره می‌شوند.

۳-۲-۳- بازیابی تصویر

در این بخش یک تصویر به عنوان ورودی به شبکه داده می‌شود و طبق معادلات (۱)، (۲) و (۳) نقشه ویژگی‌های خودتوجهی مکانی آن به دست می‌آید. سپس با نقشه‌های ویژگی ذخیره شده در قسمت ۳-۲-۲ و با استفاده از معادله زیر مقایسه می‌شوند.

$$D(X, Y) = \sqrt{\sum_{i=1}^{N=20588} (X_i - Y_i)^2} \quad (4)$$

که در آن X نقشه ویژگی خودتوجهی مکانی تصویر ورودی و Y نقشه ویژگی خودتوجهی مکانی هر تصویر از پایگاه داده است. سپس فاصله‌های به‌دست‌آمده، که هر کدام نگاشتی به تصویری از پایگاه داده دارند، به صورت نزولی مرتب شده و ۵ نتیجه برتر بازیابی می‌شود. خروجی سیستم بر اساس کلاسی که بیشترین احتمال را در بین این ۵ نتیجه دارد، تعیین می‌گردد.

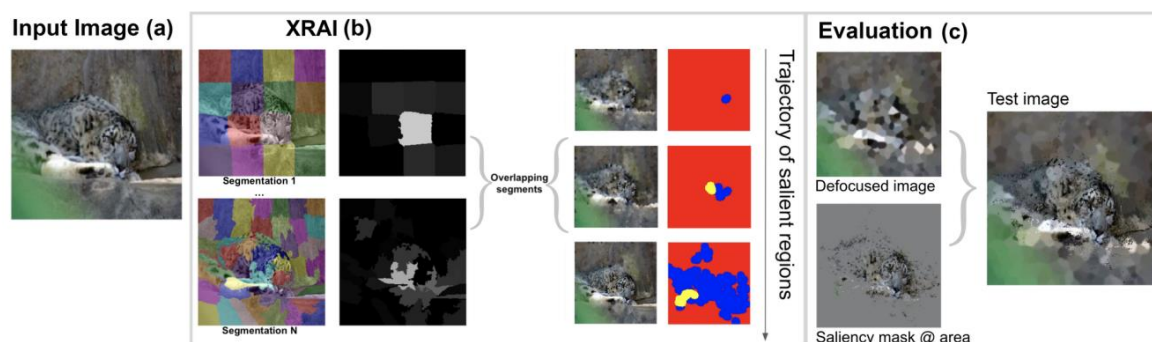


۳-۳- راهکار پیشنهادی

به عنوان نوآوری در این پژوهش، از روش XRAI شفاف‌سازی^{۸۸} استفاده شده است. این روش از گراف Felzenswalb برای تقسیم‌بندی بهره می‌گیرد. روش‌های تقسیم‌بندی معمولاً دارای چندین مجموعه از پارامترها هستند که تعداد و شکل بخش‌ها را تغییر می‌دهند. از آنجا که امکان‌پذیر نیست نتایج انتساب به مجموعه خاصی از پارامترهای فوق یا کیفیت روش تقسیم‌بندی بستگی داشته باشد، تصویر چندین بار با

استفاده از مجموعه پارامترهای مختلف قطعه‌بندی می‌شود. به طور خاص، از یک پارامتر مقیاس در مجموعه [۵۰، ۱۰۰، ۲۵۰، ۵۰۰، ۱۲۰۰] استفاده شده و بخش‌های کوچکتر از ۲۰ پیکسل نادیده گرفته می‌شود (پارامتر مقیاس عمدتاً بر اندازه بخش‌ها تأثیر می‌گذارد). برای یک پارامتر واحد، اتحاد بخش‌ها کل تصویر را محاسبه می‌کند.

بنابراین، اتحاد همه بخش‌ها مساحتی برابر با شش برابر مساحت تصویر را به دست می‌دهد و در نتیجه بخش‌های جداگانه به طور قابل توجهی همپوشانی دارند. مرزهای بخش معمولاً با لبه‌های تصویر همسو می‌شوند. برای استخراج نقشه‌های برجسته، مطلوب است که بخش‌ها شامل لبه‌ها باشند، زیرا اسناد در دو طرف یک لبه نازک اغلب به یکدیگر مرتبط هستند. برای این منظور، ماسک‌های بخش را ۵ پیکسل گشاد می‌شود تا مجموعه نهایی قطعات به دست آید.



شکل ۵. عملکرد الگوریتم XRAI روی تصویر

برای تعیین ویژگی بخش‌های مختلف تصویر، XRAI از گرادیان‌های مجتمع با خطوط پایه سیاه و سفید استفاده می‌کند. انگیزه این انتخاب به شرح زیر است: با استفاده از تکنیک گرادیان‌های یکپارچه، استفاده از یک تصویر سیاه به عنوان خط پایه، انتساب پیکسل‌های ورودی تیره را کاهش می‌دهد. برای مثال، پیکسل‌های تیره روی سوسک در شکل ۵ نسبت داده نشده‌اند، اگرچه ممکن است از پیکسل‌های روشن‌تر مهم‌تر باشند. در واقع، مقدار RGB (۰، ۰، ۰) دقیقاً مقدار صفر را دریافت می‌کند. این از فرمول گرادیان‌های یکپارچه مشخص است:

$$IG_i(x) = (x - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha * (x - x'))}{\partial x_i} d\alpha \propto \quad (5)$$

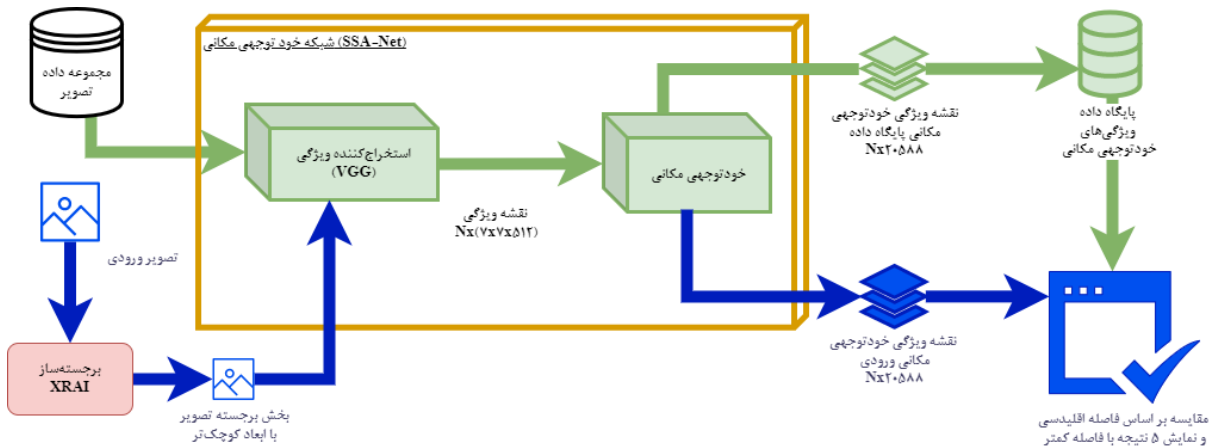
که در آن $(x - x')$ برابر با فاصله بین پیکسل i ورودی و پیکسل پایه مربوطه می‌باشد.

Algorithm 1 XRAI

- 1: Given image I , model f and attribution method g
- 2: Over-segment I to segments $s \in S$
- 3: Get attribution map $A = g(f, I)$
- 4: Let saliency mask $M = \mathbf{0}$, trajectory $T = []$
- 5: **while** $S \neq \emptyset$ and $area(M) < area(I)$ **do**
- 6: **for** $s \in S$ **do**
- 7: Compute gain²: $g_s = \sum_{i \in s \setminus M} \frac{A_i}{area(s \setminus M)}$
- 8: **end for**
- 9: $\hat{s} = \arg \max_s g_s$
- 10: $S = S \setminus \hat{s}$
- 11: $M = M \cup \hat{s}$
- 12: Add M to list T
- 13: **end while**
- 14: **return** T

شکل ۴. الگوریتم برجسته‌سازی XRAI

بدین ترتیب با اضافه شدن راهکار پیشنهادی، شکل نهایی شبکه خودتوجهی مکانی با استفاده از مکانیسم برجسته‌سازی مشابه شکل زیر خواهد بود:



شکل ۵. شبکه خودتوجهی مکانی با استفاده از مکانیسم برجسته‌سازی XRAI

۳-۴- جمع‌بندی

شبکه خودتوجهی مکانی بر اساس سه جزء اصلی ساخته شده است. یک ماژول جهت استخراج ویژگی، یک ماژول جهت خودتوجهی مکانی و یک ماژول جهت بازیابی تصویر. ماژول اول اطلاعات اصلی موردنیاز را استخراج کرده و به ماژول دوم می‌دهد. در ماژول دوم با محاسبات روی بردار کانوالو شده‌ی تصویر، میزان توجه به‌دست‌آمده و در ماژول آخر با دریافت تصویر ورودی، میزان فاصله با تصاویر مجموعه‌داده بررسی شده و دسته‌بندی بهترین جواب‌ها به عنوان نتیجه بازگردانده می‌شوند.

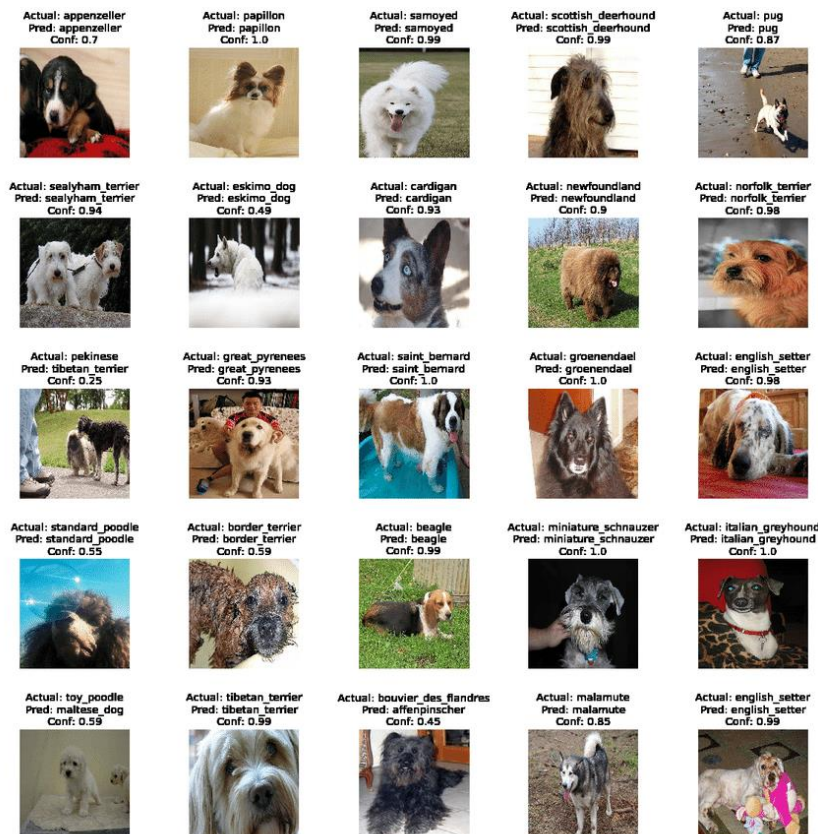
فصل چهارم: ارزیابی و راست آزمایی آزمایشگاهی

۴-۱- مقدمه

در این فصل به بررسی ارزیابی آزمایشگاهی شبکه خودتوجهی مکانی پرداخته شده است. در تکمیل بهبود استفاده از شبکه خودتوجهی مکانی، از روش شفاف‌سازی XRAI استفاده شده است. مجموعه داده استفاده شده در ۴-۲ معرفی شده‌اند و سپس نتایج رویکرد روش پایه با نتایج به دست آمده از روش پیشنهادی مقایسه شده‌اند. در این آزمایش از یک کامپیوتر شخصی مدل Macbook Air با پردازنده M1 و حافظه اصلی ۸ گیگابایت استفاده شده است. محیط توسعه و زبان مورد استفاده python است.

۴-۲- معرفی پایگاه داده

مجموعه داده Stanford Dogs^{۶۳} شامل تصاویری از ۱۲۰ نژاد سگ از سراسر جهان است. این مجموعه داده با استفاده از تصاویر و حاشیه نویسی از ImageNet برای طبقه بندی تصاویر ریز دانه ساخته شده است. این پایگاه داده، در ابتدا برای دسته بندی تصاویر دانه ریز جمع آوری شد، یک مشکل چالش برانگیز وجود داشت، زیرا برخی از نژادهای سگ ویژگی های تقریباً یکسانی دارند یا از نظر رنگ و سن متفاوت هستند. در آزمایش پیش رو، این پایگاه داده از آن جهت که ساختار سلسله مراتبی داشته و تا حد بالایی رفتارهای ریزدانه ای ارائه می دهد، انتخاب شده است.



شکل ۵. نمایی از مجموعه داده Stanford Dogs

همچنین برای راستی آزمایی بهتر روش پیشنهادی، مجموعه داده CUB_{۲۰۰-۲۰۱۱} نیز مورد بررسی قرار گرفته است. مجموعه داده (CUB-۲۰۰-۲۰۱۱) Caltech-UCSD Birds-۲۰۰-۲۰۱۱ پرکاربردترین مجموعه داده برای کار دسته بندی بصری ریز دانه است. این شامل ۱۱۷۸۸ تصویر از ۲۰۰ زیرمجموعه متعلق به پرندگان، ۵۹۹۴ تصویر برای آموزش و ۵۷۹۴ تصویر برای آزمایش است.



شکل ۶. نمایی از مجموعه داده CUB_۲۰۰_۲۰۱۱

۴-۳- معیارهای ارزیابی

معیار این آزمایش فاصله اقلیدسی بین نقشه ویژگی خودتوجهی مکانی تصویر ورودی و نقشه ویژگی خودتوجهی مکانی تصاویر پایگاه داده، می‌باشد.

$$D(X, Y) = \sqrt{\sum_{i=1}^{N=20588} (X_i - Y_i)^2} \quad (۶)$$

که در آن X نقشه ویژگی خودتوجهی مکانی تصویر ورودی و Y نقشه ویژگی خودتوجهی مکانی هر تصویر از پایگاه داده است. سپس فاصله‌های به دست آمده، که هر کدام نگاشتی به تصویری از پایگاه داده دارند، به صورت نزولی مرتب شده و ۵ نتیجه برتر ارزیابی می‌شود. خروجی سیستم بر اساس کلاسی که بیشترین احتمال را در بین این ۵ نتیجه دارد، تعیین می‌گردد. امتیاز هر ارزیابی تصویر بر اساس فرمول زیر تهیه می‌گردد.

$$Score = \frac{\text{Count Of True Labels}}{\text{Count Of All Results}} \quad (۷)$$

به عنوان مثال برای تصویر I با کلاس A که نتایج آن شامل $Results=[A,A,B,A,A]$ می‌باشند، داریم:

$$Score(I) = \frac{\text{Count(Results wiht } A)}{\text{Count(Results)}} \quad (۸)$$

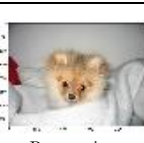
سپس عملکرد نهایی سیستم بر اساس فرمول زیر محاسبه خواهد شد:

$$Score = \frac{\sum_{i=1}^N Score(I_i)}{N} \quad (۹)$$





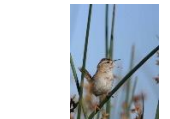




























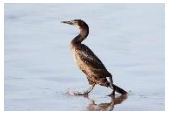






















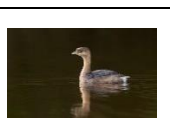
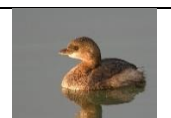


که در آن N تعداد نمونه‌های تستی و I هر کدام از تصاویر تستی می‌باشد.

۴-۴- نتایج ارزیابی

جدول ۲. نتایج آزمایشگاهی بر روی مجموعه داده StanfordDogs

Input	Outputs					Score
 Lhasa	 Lhasa	 Lhasa	 Lhasa	 Lhasa	 silky_terrier	80%
 Irish setter	 Irish setter	 Irish setter	 Irish setter	 Irish setter	 flat_coated_tretriever	80%
 Pomernian	 Pomernian	 Pomernian	 Pomernian	 Pomernian	 Pomernian	100%
 Afghan_hound	 bloodhound	 Aghan_hound	 bloodhound	 Afghan_hound	 Afghan_hound	60%
 otterhound	 otterhound	 otterhound	 otterhound	 otterhound	 otterhound	100%
 Afghan_hound	 Afghan_hound	 Afghan_hound	 Afghan_hound	 Afghan_hound	 Aghan_hound	100%
 Pekinese	 Pekinese	 Shih-Tzu	 Shih-Tzu	 Pekinese	 Pekinese	60%
 Bernese_mountain	 Bernese_mountain	 Bernese_mountain	 Bernese_mountain	 Bernese_mountain	 Bernese_mountain	100%
 chow	 chow	 chow	 chow	 chow	 chow	100%
 clumber	 clumber	 clumber	 clumber	 clumber	 clumber	100%

جدول ۳. نتایج آزمایشگاهی بر روی مجموعه داده CUB_۲۰۰_۲۰۱۱

Input	Outputs					Score
 Wren	 Warbler	 Wren	 Hummingbird	 Wren	 Wren	%۶۰
 Woodpecker	 Woodpecker	 Woodpecker	 Woodpecker	 Flicker	 Woodpecker	%۸۰
 Sparrow	 Thrasher	 Sparrow	 Sparrow	 Sparrow	 Sparrow	%۸۰
 Vireo	 Vireo	 Vireo	 Warbler	 Vireo	 Vireo	%۶۰
 Gull	 Gull	 Grebe	 Gull	 Gull	 Gull	%۱۰۰
 Grebe	 Grebe	 Grebe	 Cormorant	 Grebe	 Grebe	%۸۰
 Blackbird	 Blackbird	 Blackbird	 Blackbird	 Blackbird	 Oriole	%۸۰
 Gull	 Gull	 Gull	 Gull	 Gull	 Gull	%۱۰۰
 Gull	 Gull	 Gull	 Gull	 Gull	 Gull	%۱۰۰
 Grebe	 Grebe	 Grebe	 Grebe	 Grebe	 Grebe	%۱۰۰

Input	Outputs					Score
 Gull	 Gull	 Gull	 Gull	 Gull	 Gull	%100
 Tern	 Tern	 Tern	 Tern	 Tern	 Tern	%100
 Merganser	 Merganser	 Merganser	 Merganser	 Merganser	 Merganser	%100
 Woodpecker	 Woodpecker	 Woodpecker	 Woodpecker	 Woodpecker	 Woodpecker	%100
 Warbler	 Warbler	 Warbler	 Warbler	 Sparrow	 Oriole	%80
 Sparrow	 Sparrow	 Sparrow	 Sparrow	 Wren	 Waterthrush	%60
 Sparrow	 Sparrow	 Sparrow	 Sparrow	 Sparrow	 Sparrow	%100
 Sparrow	 Sparrow	 Sparrow	 Sparrow	 Sparrow	 Sparrow	%100
 Sparrow	 Grosbeak	 Sparrow	 Sparrow	 Flycatcher	 Sparrow	%60
 Wren	 Wren	 Wren	 Ovenbird	 Wren	 Wren	%80

جدول ۴. نمایش نتایج آزمایشگاهی و مقایسه عملکرد آن با سایر روش‌های بازیابی تصویر، بر روی مجموعه داده Stanford Dogs

Method	Score
FCAN [۶۴]	۸۴.۵٪
PDFR [۶۵]	۷۱.۹٪
PC-DenseNet-۱۶۱ [۶۶]	۸۳.۶٪
[۶۷] HDWE	۷۹.۶٪
EfficientNet-B۰ [۶۸]	۶۱.۲٪
PC [۶۹]	۶۱.۹٪
SSA [۷۰]	۸۶٪
SSA with XRAI Saliency	۸۸٪

جدول ۴ بیان‌کننده مقایسه نتایج عملکرد روش‌های مختلف در بازیابی تصویر ریزدانه‌ای بر روی مجموعه داده Stanford Dogs می‌باشد. در روش EfficientNet-B۰ از کانولوشن‌های قابل تفکیک عمیق استفاده شده و در میان سایر روش‌ها از عملکرد ضعیف‌تری برخوردار است. در روش PC از آموزش با تقسیم‌بندی‌های پیچیده، در جهت ریزدانه‌ای کردن بازیابی تصویر استفاده شده است که نسبت به روش پیشین ۰.۷ درصد بهبود داشته است. روش PDFR، از فیلترهای متمایز و آشکارسازی استفاده می‌کند که نسبت به روش قبلی ۱۰ درصد بهبود ایجاد کرده است. در روش‌های FCAN، PC-DenseNet-۱۱۶ و HDWE نیز از شبکه‌های کانولوشنی متعددی استفاده شده که نتیجه آن‌ها نسبت به موارد قبلی بین ۸ الی ۱۳ درصد افزایش امتیاز صورت گرفته است. روش SSA که در این مقاله نیز بررسی شد نسبت به سایر روش‌ها کارآمدتر است ولی روش پیشنهادی عملکردی بهتری نسبت به تمامی روش‌های ذکر شده دارد.

جدول ۵. نمایش نتایج آزمایشگاهی و مقایسه عملکرد آن با سایر روش‌های بازیابی تصویر، بر روی مجموعه داده CUB_200_2011

Method	Score
PDFR [۷۱]	٪۸۲.۶
HDWE [۷۲]	٪۸۴.۳
SSA [۷۳]	٪۸۵
SSA with XRAI Saliency	٪۸۷

جدول ۵ بیان‌کننده مقایسه نتایج عملکرد روش‌های مختلف در بازیابی تصویر ریزدانه‌ای بر روی مجموعه داده CUB_200_2011 می‌باشد. روش PDFR، از فیلترهای متمایز و آشکارسازی استفاده می‌کند و ۸۲.۶ درصد صحت عملکرد داشته است. در روش HDWE نیز از شبکه‌های کانولوشنی متعددی استفاده شده که نتیجه آن‌ها نسبت به مورد قبلی ۱.۷ درصد افزایش امتیاز صورت گرفته است. روش SSA که در این پژوهش نیز بررسی شد نسبت به سایر روش‌ها کارآمدتر است ولی روش پیشنهادی عملکردی بهتری نسبت به تمامی روش‌های ذکر شده دارد.

فصل پنجم: جمع‌بندی و پیشنهادات آینده

۵-۲- جمع بندی و نتیجه گیری

یکی از مهم‌ترین شاخه‌های پردازش تصویر، بازیابی تصاویر می‌باشد. این شاخه از علم پردازش تصویر، برای اولین بار در سال ۱۹۷۰ با رویکرد مبتنی بر متن معرفی گردید. پس از آن رویکردی متفاوت با عنوان مبتنی بر محتوا معرفی گردد که بر اساس ویژگی‌های استخراج شده از تصاویر، کار می‌کرد. این رویکرد به سرعت جایگزین رویکرد پیشین شد و در حوزه‌های پزشکی، گیاه‌شناسی، جانورشناسی، فروش و هنر مورد استفاده قرار گرفت. بازیابی تصویر در سه رویکرد مبتنی بر متن، مبتنی بر محتوا و مبتنی بر طرح، دنبال می‌گردد. هر کدام از این رویکردها در زیرروش‌های مختلفی دسته‌بندی می‌شوند که در نوع استخراج ویژگی، نوع پردازش ویژگی‌ها و طبقه‌بندی خروجی با هم تفاوت دارند. در سه رویکرد اصلی ویژگی‌های مختلف تصویر بررسی می‌گردند و هر روش با ترکیب این ویژگی‌ها به یافتن تصاویر مشابه تصویر ورودی سیستم می‌پردازد. یکی از زیرشاخه‌های مبتنی بر محتوا، بازیابی ریزدانه‌ای تصویر است که تا یک دسته پایین‌تر تشخیص را جلوتر می‌برد. هدف آن کم کردن میزان اختلاف میان ویژگی‌های هر دسته از تصاویر می‌باشد. در بازیابی ریزدانه‌ای از بردار ویژگی، شبکه‌های عصبی مصنوعی و یا ترکیب این دو استفاده می‌شود. در این میان چالش‌های بسیاری به هنگام استفاده از روش‌های مبتنی بر این رویکرد، پیش می‌آید. از جمله آنکه ویژگی‌های استخراج شده با ادراک انسان فاصله معنایی بسیاری داشتند. اما با انتخاب و استخراج درست ویژگی‌های مورد محاسبه، این فاصله کمتر به چشم آمده است. شبکه خودتوجهی مکانی بر اساس سه جزء اصلی ساخته شده است. یک ماژول جهت استخراج ویژگی، یک ماژول جهت خودتوجهی مکانی و یک ماژول جهت تقطیر خودتوجهی. ماژول اول اطلاعات اصلی موردنیاز را استخراج کرده و به ماژول دوم می‌دهد. در ماژول دوم با محاسبات روی بردار کانال‌های شده‌ی تصویر، میزان توجه به دست‌آمده و در ماژول آخر با ارائه یک تصویر به عنوان ورودی، دسته‌بندی تخمینی سیستم ارائه می‌گردد.

۵-۳- پیشنهادات آینده

به عنوان تحقیقات آینده در این روش، پیشنهاد می‌شود رویکرد بازیابی تصویر ریزدانه‌ای مبتنی بر محتوا، با حفظ هسته اصلی خودتوجهی مکانی و اضافه شدن ماژول یادگیری تقویتی ادامه یابد. برای بهتر شدن تحقیقات می‌توان از یادگیری بدون نظارت استفاده کرد و داده‌هایی که ساختار سلسله مراتبی ندارند نیز بررسی گردند. سپس نتایجی که سیستم به عنوان خروجی برگردانده بررسی شده و امتیاز بگیرند. ممکن است در مواردی منجر به ساختن خوشه‌های جدید تصاویر شود و مجموعه تصاویر بتوانند با امتیازهای متفاوت دسته‌بندی سلسله مراتبی را ایجاد کنند.

منابع

-
- [1] Bressan, R. S., Bugatti, P. H., & Saito, P. T. M. (2022). Optimum-path forest and active learning approaches for content-based medical image retrieval. In *Optimum-Path Forest* (pp. 95–107). Elsevier. <https://doi.org/10.1016/b978-0-12-822688-9.00012-8>
 - [2] Wang, X., Lan, R., Wang, H., Liu, Z., & Luo, X. (2021). Fine-grained correlation analysis for medical image retrieval. In *Computers & Electrical Engineering* (Vol. 90, p. 106992). Elsevier BV. <https://doi.org/10.1016/j.compeleceng.2021.106992>
 - [3] Duan, Y., Li, Y., Lu, L., & Ding, Y. (2022). A faster outsourced medical image retrieval scheme with privacy preservation. In *Journal of Systems Architecture* (Vol. 122, p. 102356). Elsevier BV. <https://doi.org/10.1016/j.sysarc.2021.102356>
 - [4] Tonioni, A., & Di Stefano, L. (2019). Domain invariant hierarchical embedding for grocery products recognition. In *Computer Vision and Image Understanding* (Vol. 182, pp. 81–92). Elsevier BV. <https://doi.org/10.1016/j.cviu.2019.03.005>
 - [5] Rahman, A., Winarko, E., & Mustofa, K. (2022). Product image retrieval using category-aware siamese convolutional neural network feature. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 34, Issue 6, pp. 2680–2687). Elsevier BV. <https://doi.org/10.1016/j.jksuci.2022.03.005>
 - [6] Gajjar, V. K., Nambisan, A. K., & Kosbar, K. L. (2022). Plant Identification in a Combined-Imbalanced Leaf Dataset. In *IEEE Access* (Vol. 10, pp. 37882–37891). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/access.2022.3165583>
 - [7] Sundara Sobitha Raj, A. P., & Vajravelu, S. K. (2019). DDLA: dual deep learning architecture for classification of plant species. In *IET Image Processing* (Vol. 13, Issue 12, pp. 2176–2182). Institution of Engineering and Technology (IET). <https://doi.org/10.1049/iet-ipr.2019.0346>
 - [8] Wei, X.-S., Xie, C.-W., Wu, J., & Shen, C. (2018). Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. In *Pattern Recognition* (Vol. 76, pp. 704–714). Elsevier BV. <https://doi.org/10.1016/j.patcog.2017.10.002>
 - [9] Jin, X., Deng, Q., Lou, H., Li, X., & Xiao, C. (2022). Fine-grained Regression for Image Aesthetic Scoring. In *Cognitive Robotics*. Elsevier BV. <https://doi.org/10.1016/j.cogr.2022.07.003>
 - [10] Li, Z., Lu, H., Fu, H., & Gu, G. (2022). Image-text bidirectional learning network based cross-modal retrieval. In *Neurocomputing* (Vol. 483, pp. 148–159). Elsevier BV. <https://doi.org/10.1016/j.neucom.2022.02.007>
 - [11] Li, C., Chen, X., Wang, H., Wang, P., Zhang, Y., & Wang, W. (2021). End-to-end attack on text-based CAPTCHAs based on cycle-consistent generative adversarial network. In *Neurocomputing* (Vol. 433, pp. 223–236). Elsevier BV. <https://doi.org/10.1016/j.neucom.2020.11.05>
 - [12] Jiang, Y., Du, J., Xue, Z., & Li, A. (2022). Cross-Media Retrieval of Scientific and Technological Information Based on Multi-Feature Fusion. In *Neurocomputing*. Elsevier BV. <https://doi.org/10.1016/j.neucom.2022.06.061>
 - [13] Unar, S., Wang, X., Wang, C., & Wang, Y. (2019). A decisive content based image retrieval approach for feature fusion in visual and textual images. In *Knowledge-Based Systems* (Vol. 179, pp. 8–20). Elsevier BV. <https://doi.org/10.1016/j.knosys.2019.05.001>

-
- [14]Farruggia, A., Magro, R., & Vitabile, S. (2014). A text based indexing system for mammographic image retrieval and classification. In *Future Generation Computer Systems* (Vol. 37, pp. 243–251). Elsevier BV. <https://doi.org/10.1016/j.future.2014.02.008>
- [15]Yasmin, M., Sharif, M., Irum, Isma, & Mohsin, S. (2013). Powerful Descriptor for Image Retrieval Based on Angle Edge and Histograms. In *Journal of Applied Research and Technology* (Vol. 11, Issue 5, pp. 727–732). Universidad Nacional Autonoma de Mexico. [https://doi.org/10.1016/s1665-6423\(13\)71581-5](https://doi.org/10.1016/s1665-6423(13)71581-5)
- [16]Wang, Q., Lai, J., Yang, Z., Xu, K., Kan, P., Liu, W., & Lei, L. (2019). Improving cross-dimensional weighting pooling with multi-scale feature fusion for image retrieval. In *Neurocomputing* (Vol. 363, pp. 17–26). Elsevier BV. <https://doi.org/10.1016/j.neucom.2019.08.025>
- [17]Niu, D., Zhao, X., Lin, X., & Zhang, C. (2020). A novel image retrieval method based on multi-features fusion. In *Signal Processing: Image Communication* (Vol. 87, p. 115911). Elsevier BV. <https://doi.org/10.1016/j.image.2020.115911>
- [18]Piras, L., & Giacinto, G. (2017). Information fusion in content based image retrieval: A comprehensive overview. In *Information Fusion* (Vol. 37, pp. 50–60). Elsevier BV. <https://doi.org/10.1016/j.inffus.2017.01.003>
- [19]Zhang, L., & Wu, X. (2022). Multi-task framework based on feature separation and reconstruction for cross-modal retrieval. In *Pattern Recognition* (Vol. 122, p. 108217). Elsevier BV. <https://doi.org/10.1016/j.patcog.2021.108217>
- [20]Zhang, M., Tian, G., Gao, H., Liu, S., & Zhang, Y. (2022). Multimodal feature fusion and exploitation with dual learning and reinforcement learning for recipe generation. In *Applied Soft Computing* (Vol. 126, p. 109281). Elsevier BV. <https://doi.org/10.1016/j.asoc.2022.109281>
- [21]Wang, Y., Fariah Haq, N., Cai, J., Kalia, S., Lui, H., Jane Wang, Z., & Lee, T. K. (2022). Multi-channel content based image retrieval method for skin diseases using similarity network fusion and deep community analysis. In *Biomedical Signal Processing and Control* (Vol. 78, p. 103893). Elsevier BV. <https://doi.org/10.1016/j.bspc.2022.103893>
- [22]Zhang, K., Qi, S., Cai, J., Zhao, D., Yu, T., Yue, Y., Yao, Y., & Qian, W. (2022). Content-based image retrieval with a Convolutional Siamese Neural Network: Distinguishing lung cancer and tuberculosis in CT images. In *Computers in Biology and Medicine* (Vol. 140, p. 105096). Elsevier BV. <https://doi.org/10.1016/j.compbimed.2021.105096>
- [23]Sun, M., Zou, W., Hu, N., Wang, J., & Chi, Z. (2022). Iterative brain tumor retrieval for MR images based on user’s intention model. In *Pattern Recognition* (Vol. 127, p. 108650). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.108650>
- [24]Shamna, P., Govindan, V. K., & Abdul Nazeer, K. A. (2022). Content-based medical image retrieval by spatial matching of visual words. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 34, Issue 2, pp. 58–71). Elsevier BV. <https://doi.org/10.1016/j.jksuci.2018.10.002>
- [25]Yelchuri, R., Dash, J. K., Singh, P., Mahapatro, A., & Panigrahi, S. (2022). Exploiting deep and hand-crafted features for texture image retrieval using class membership. In *Pattern Recognition Letters* (Vol. 160, pp. 163–171). Elsevier BV. <https://doi.org/10.1016/j.patrec.2022.06.017>

-
- [26]Devulapalli, S., Potti, A., Krishnan, R., & Khan, Md. S. (2021). Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques. In *Materials Today: Proceedings*. Elsevier BV. <https://doi.org/10.1016/j.matpr.2021.04.326>
- [27]Li, J., Ling, Z., Niu, L., & Zhang, L. (2022). Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement. In *Computer Vision and Image Understanding* (Vol. 218, p. 103412). Elsevier BV. <https://doi.org/10.1016/j.cviu.2022.103412>
- [28]Tursun, O., Denman, S., Sridharan, S., Goan, E., & Fookes, C. (2022). An efficient framework for zero-shot sketch-based image retrieval. In *Pattern Recognition* (Vol. 126, p. 108528). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.108528>
- [29]Zheng, Y., Yao, H., Sun, X., Zhang, S., Zhao, S., & Porikli, F. (2021). Sketch-specific data augmentation for freehand sketch recognition. In *Neurocomputing* (Vol. 456, pp. 528–539). Elsevier BV. <https://doi.org/10.1016/j.neucom.2020.05.124>
- [30]Zhang, X., Shen, M., Li, X., & Feng, F. (2022). A deformable CNN-based triplet model for fine-grained sketch-based image retrieval. In *Pattern Recognition* (Vol. 125, p. 108508). Elsevier BV. <https://doi.org/10.1016/j.patcog.2021.108508>
- [31]Dai, D., Tang, X., Liu, Y., Xia, S., & Wang, G. (2022). Multi-granularity association learning for on-the-fly fine-grained sketch-based image retrieval. In *Knowledge-Based Systems* (p. 109447). Elsevier BV. <https://doi.org/10.1016/j.knosys.2022.109447>
- [32]Wang, Y., Huang, F., Zhang, Y., Feng, R., Zhang, T., & Fan, W. (2020). Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. In *Pattern Recognition* (Vol. 100, p. 107148). Elsevier BV. <https://doi.org/10.1016/j.patcog.2019.107148>
- [33]Xu, P., Yin, Q., Huang, Y., Song, Y.-Z., Ma, Z., Wang, L., Xiang, T., Kleijn, W. B., & Guo, J. (2018). Cross-modal subspace learning for fine-grained sketch-based image retrieval. In *Neurocomputing* (Vol. 278, pp. 75–86). Elsevier BV. <https://doi.org/10.1016/j.neucom.2017.05.099>
- [34]Ha, Y., Du, Z., & Tian, J. (2022). Fine-grained interactive attention learning for semi-supervised white blood cell classification. *Biomedical Signal Processing and Control*, 75, 103611. <https://doi.org/10.1016/j.bspc.2022.103611>
- [35]Liu, X., Wang, L., & Han, X. (2022). Transformer with peak suppression and knowledge guidance for fine-grained image recognition. *Neurocomputing*, 492, 137–149. <https://doi.org/10.1016/j.neucom.2022.04.03>
- [36]Chen, Y., Zhang, Z., Wang, Y., Zhang, Y., Feng, R., Zhang, T., & Fan, W. (2022). AE-Net: Fine-grained sketch-based image retrieval via attention-enhanced network. *Pattern Recognition*, 122, 108291. <https://doi.org/10.1016/j.patcog.2021.108291>
- [37]Chen, H., Sun, C., Liao, P., Lai, Y., Fan, F., Lin, Y., Deng, Z., & Zhang, Y. (2022). A fine-grained network for human identification using panoramic dental images. *Patterns*, 100485. <https://doi.org/10.1016/j.patter.2022.100485>
- [38]Zeng, X., Liu, S., Wang, X., Zhang, Y., Chen, K., & Li, D. (2021). Hard Decorrelated Centralized Loss for fine-grained image retrieval. *Neurocomputing*, 453, 26–37. <https://doi.org/10.1016/j.neucom.2021.04.030>

-
- [39]Tang, H., Yuan, C., Li, Z., & Tang, J. (2022). Learning attention-guided pyramidal features for few-shot fine-grained recognition. In *Pattern Recognition* (Vol. 130, p. 108792). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.108792>
- [40]Yang, M., Xu, Y., Wu, Z., & Wei, Z. (2022). Symmetrical irregular local features for fine-grained visual classification. In *Neurocomputing* (Vol. 505, pp. 304–314). Elsevier BV. <https://doi.org/10.1016/j.neucom.2022.07.056>
- [41]Araújo, V. M. , Britto Jr. , A. S. , Oliveira, L. S. , & Koerich, A. L. (2022). Two-view fine-grained classification of plant species. *Neurocomputing*, 467, 427–441. <https://doi.org/10.1016/j.neucom.2021.10.015>
- [42]Yuan, P. , Qian, S. , Zhai, Z. , FernánMartínez, J. , & Xu, H. (2022). Study of chrysanthemum image phenotype on-line classification based on transfer learning and bilinear convolutional neural network. *Computers and Electronics in Agriculture*, 194, 106679. <https://doi.org/10.1016/j.compag.2021.106679>
- [43]Bhupendra, Moses, K. , Miglani, A. , & Kumar Kankar, P. (2022). Deep CNN-based damage classification of milled rice grains using a high-magnification image dataset. *Computers and Electronics in Agriculture*, 195, 106811. <https://doi.org/10.1016/j.compag.2022.106811>
- [44]Yan, T. , Shi, J. , Li, H. , Luo, Z. , & Wang, Z. (2022). Discriminative information restoration and extraction for weakly supervised low-resolution fine-grained image recognition. *Pattern Recognition*, 127, 108629. <https://doi.org/10.1016/j.patcog.2022.108629>
- [45]Zhao, Q. , Wang, X. , Lyu, S. , Liu, B. , & Yang, Y. (2022). A feature consistency driven attention erasing network for fine-grained image retrieval. *Pattern Recognition*, 128, 108618. <https://doi.org/10.1016/j.patcog.2022.108618>
- [46]Chen, Y. , Song, J. , & Song, M. (2022). Hierarchical gate network for fine-grained visual recognition. *Neurocomputing*, 470, 170–181. <https://doi.org/10.1016/j.neucom.2021.10.096>
- [47]Zhou, Y. , Li, X. , Zhou, Y. , Wang, Y. , Hu, Q. , & Wang, W. (2022). Deep collaborative multi-task network: A human decision process inspired model for hierarchical image classification. *Pattern Recognition*, 124, 108449. <https://doi.org/10.1016/j.patcog.2021.108449>
- [48]Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (2021). Spatial self-attention network with self-attention distillation for fine-grained image recognition. *Journal of Visual Communication and Image Representation*, 81, 103368. <https://doi.org/10.1016/j.jvcir.2021.103368>
- [49]Santra, B. , Shaw, A. K. , & Mukherjee, D. P. (2022). Part-based annotation-free fine-grained classification of images of retail products. *Pattern Recognition*, 121, 108257. <https://doi.org/10.1016/j.patcog.2021.108257>
- [50]Zeng, X. , Zhang, Y. , Wang, X. , Chen, K. , Li, D. , & Yang, W. (2019). Fine Grained Image Retrieval via Piecewise Cross Entropy loss. *Image and Vision Computing*. <https://doi.org/10.1016/j.imavis.2019.10.006>
- [51]Shan, W., Huang, D., Wang, J., Zou, F., & Li, S. (2022). Self-Attention based fine-grained cross-media hybrid network. In *Pattern Recognition* (Vol. 130, p. 108748). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.108748>

-
- [52] Yu, X., Zhao, Y., & Gao, Y. (2022). SPARE: Self-supervised part erasing for ultra-fine-grained visual categorization. In *Pattern Recognition* (Vol. 128, p. 108691). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.108691>
- [53] Peng, J., Wang, Y., & Zhou, Z. (2022). Progressive Erasing Network with consistency loss for fine-grained visual classification. In *Journal of Visual Communication and Image Representation* (Vol. 87, p. 103570). Elsevier BV. <https://doi.org/10.1016/j.jvcir.2022.103570>
- [54] Yu, Y., Zhang, D., Wang, S., Ji, Z., & Zhang, Z. (2022). Local spatial alignment network for few-shot learning. In *Neurocomputing* (Vol. 497, pp. 182–190). Elsevier BV. <https://doi.org/10.1016/j.neucom.2022.05.020>
- [55] Xu, S., Muselet, D., & Trémeau, A. (2022). Sparse coding and normalization for deep Fisher score representation. In *Computer Vision and Image Understanding* (Vol. 220, p. 103436). Elsevier BV. <https://doi.org/10.1016/j.cviu.2022.103436>
- [56] Wang, X., Zeng, X., Zhang, Y., Chen, K., & Li, D. (2022). Improved fine-grained object retrieval with Hard Global Softmin Loss objective. In *Signal Processing: Image Communication* (Vol. 100, p. 116515). Elsevier BV. <https://doi.org/10.1016/j.image.2021.116515>
- [57] Hu, Y., Zhang, H., Jiang, H., Bi, Y., & Yin, B. (2022). CGNN: Caption-Assisted Graph Neural Network for Image-Text Retrieval. In *Pattern Recognition Letters*. Elsevier BV. <https://doi.org/10.1016/j.patrec.2022.08.002>
- [58] Ji, J. , Guo, Y. , Yang, Z. , Zhang, T. , & Lu, X. (2021). Multi-level dictionary learning for fine-grained images categorization with attention model. *Neurocomputing*, 453, 403–412. <https://doi.org/10.1016/j.neucom.2020.07.147>
- [59] Guo, C. , Lin, Y. , Chen, S. , Zeng, Z. , Shao, M. , & Li, S. (2022). From the whole to detail: Progressively sampling discriminative parts for fine-grained recognition. *Knowledge-Based Systems*, 235, 107651. <https://doi.org/10.1016/j.knosys.2021.1076>
- [60] Ahmad, J., Muhammad, K., Bakshi, S., & Baik, S. W. (2018). Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets. In *Future Generation Computer Systems* (Vol. 81, pp. 314–330). Elsevier BV. <https://doi.org/10.1016/j.future.2017.11.002>
- [61] Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (2021). Spatial self-attention network with self-attention distillation for fine-grained image recognition. *Journal of Visual Communication and Image Representation*, 81, 103368. <https://doi.org/10.1016/j.jvcir.2021.103368>
- [62] Kapishnikov, A., Bolukbasi, T., Viégas, F., & Terry, M. (2019). XRAI: Better Attributions Through Regions (Version 2). *arXiv*. <https://doi.org/10.48550/ARXIV.1906.02825>
- [63] Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (2012). Novel Dataset for Fine-Grained Image Categorization : Stanford Dogs.
- [64] Liu, X., Xia, T., Wang, J., & Lin, Y. (2016). Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 1(2), 4.

-
- [65]X. Zhang, H. Xiong, W. Zhou, W. Lin and Q. Tian, "Picking Deep Filter Responses for Fine-Grained Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1134–1142, doi: 10.1109/CVPR.2016.128
- [66]Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N. (2018). Pairwise Confusion for Fine-Grained Visual Classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11216. Springer, Cham. https://doi.org/10.1007/978-3-030-01258-8_5
- [67]67 Yu, J., Huang, Y., Gbur, G., Wang, F., & Cai, Y. (2019). Enhanced backscatter of vortex beams in double-pass optical links with atmospheric turbulence. In Journal of Quantitative Spectroscopy and Radiative Transfer (Vol. 228, pp. 1–10). Elsevier BV. <https://doi.org/10.1016/j.jqsrt.2019.02.021>
- [68]D. Haase and M. Amthor, "Rethinking Depthwise Separable Convolutions: How Intra-Kernel Correlations Lead to Improved MobileNets," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14888–14897, doi: 10.1109/CVPR.42600.2020.1461
- [69]Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., & Naik, N. (2017). Training with confusion for fine-grained visual classification. CoRR
- [70]Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (2021). Spatial self-attention network with self-attention distillation for fine-grained image recognition. Journal of Visual Communication and Image Representation, 81, 103368. <https://doi.org/10.1016/j.jvcir.2021.103368>
- [71]X. Zhang, H. Xiong, W. Zhou, W. Lin and Q. Tian, "Picking Deep Filter Responses for Fine-Grained Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1134–1142, doi: 10.1109/CVPR.2016.128
- [72]72 Yu, J., Huang, Y., Gbur, G., Wang, F., & Cai, Y. (2019). Enhanced backscatter of vortex beams in double-pass optical links with atmospheric turbulence. In Journal of Quantitative Spectroscopy and Radiative Transfer (Vol. 228, pp. 1–10). Elsevier BV. <https://doi.org/10.1016/j.jqsrt.2019.02.021>
- [73]Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (2021). Spatial self-attention network with self-attention distillation for fine-grained image recognition. Journal of Visual Communication and Image Representation, 81, 103368. <https://doi.org/10.1016/j.jvcir.2021.103368>