کنفرانس پردازش سیگنال و سیستم های هوشمند
۷ الی ۸ دی ماه ۱۴۰۱
دانشگاه علم و فناوری مازندران

# بازیابی ریزدانه‌ای تصویر به کمک شبکه خودتوجهی مکانی و مکانیسم برجسته‌سازی

سید نیما سید آقا یزدی[۱]، نام و نام خانوادگی نویسنده دوم[۲]، نام و نام خانوادگی نویسنده سوم[۳]

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی، دانشکده فنی مهندسی، دانشگاه آزاد تهران‌جنوب ، تهران،
nima.yazdi۹۳@gmail.com

[۲] رتبه علمی نویسنده در صورت تمایل، گروه آموزشی یا واحد سازمانی مربوطه، نام سازمان ، شهر
آدرس پست الکترونیکی

[۳] رتبه علمی نویسنده در صورت تمایل، گروه آموزشی یا واحد سازمانی مربوطه، نام سازمان ، شهر
آدرس پست الکترونیکی

## 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 مقاله، 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 تصویر مبتنی بر محتوا - شبکه خودتوجهی مکانی – مکانیسم برجسته‌سازی

## ۱- مقدمه

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 شد که بر اساس ویژگی‌های استخراج شده از تصاویر، کار می‌کرد. این رویکرد به‌سرعت جایگزین رویکرد پیشین شد و در حوزه‌های پزشکی، احراز هویت، پیشگیری از وقوع جرم، امنیت محیط و... مورداستفاده 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 آمده‌است. به‌گونه‌ای که اکنون با نیاز به بررسی دقیق‌تر دسته‌بندهای تصاویر، بازیابی تصاویر ریزدانه‌ای[۳] معرفی شده‌است که در پیداکردن ویژگی‌های مشابه، تا حد ادراک انسان رفتار می‌کند.

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. این دسته بندی‌ها هر کدام دارای زیرروش های مختلفی هستند که می‌توان آن‌ها را با توجه به انواع

استخراج ویژگی، پردازش ویژگی و طبقه‌بندی تصاویر، دسته‌بندی کرد. در ادامه به بررسی نمونه‌هایی از این دسته‌ها پرداخته‌شده‌است.

در دسته روش‌های مبتنی بر طرح، در [۱] یک مدل بازیابی تصویر مبتنی بر طرح با نام صفرشات[۴] را بررسی می‌کند که در آن دسته‌های آزمایشی در مرحله آموزش ظاهر نمی‌شوند. سعی می‌شود از طریق گسستگی نامتقارن[۵] به بازیابی آگاهانه از ساختار رسید. 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 [۲] بازیابی تصویر مبتنی بر طرح به‌عنوان یک فرایند درشت به ریز فرموله شده است و یک مدل رتبه‌بندی متقابل با نام آبشاری عمیق[۶] پیشنهاد می‌شود که می‌تواند از تمام اطلاعات چندوجهی مفید در طرح‌ها و تصاویر حاشیه‌نویسی بهره‌برداری کند و کارایی بازیابی را بهبود بخشد. هدف ساختن بازنمایی‌های عمیق برای طرح‌ها، تصاویر، توضیحات و یادگیری همبستگی‌های عمیق بهینه شده در چنین حوزه‌های مختلف، متمرکز است؛ بنابراین برای یک طرح ورودی داده شده، تصاویر مربوط به آن با شباهت‌های ریز در سطح نمونه، در یک دسته خاص می‌توانند برگردانده شوند.

در دسته روش‌های مبتنی بر محتوا، در [۳] 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. در [٤] هدف ایجاد یک روش بازیابی تصویر برای مشخص‌کردن دسته‌بندی یک محصول است و یک مدل شبکه کانولوشنی سیامی[۷] پیشنهاد می‌شود که شامل برچسب‌های دسته و مورد در آموزش برای تولید ویژگی آگاه از دسته است. این مدل با اصلاح رویه آموزشی همراه می‌باشد که به طور هم‌زمان دسته و برچسب مورد را یاد می‌گیرد.

در [٥] یک روش بازیابی متقابل رسانه‌ای مبتنی بر ترکیب چند ویژگی[۸] پیشنهاد می‌شود. این روش قادر به ادغام چندین ویژگی برای ارتقای درک معنایی و اتخاذ یادگیری متخاصم[۹] برای بهبود بیشتر دقت بازنمایی زیرفضای عمومی است. سپس از شباهت در همان فضا برای مرتب‌سازی نتایج بازیابی استفاده می‌شود. در [٦] یک روش 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 ویژگی‌های رنگ و بافت را به تصویر بکشد. سپس هیستوگرام الگوهای باینری محلی یکنواخت[۱۰] تصویر را استخراج می‌کند تا اطلاعات تفاوت رنگ را به تصویر بکشد.

در [۷] یک چارچوب چند وظیفه‌ای مبتنی بر جداسازی و بازسازی ویژگی [۱۱] برای بازیابی متقابل وجهی بر اساس روش‌های رایج یادگیری مکانی پیشنهاد می‌شود که ماژول جداسازی ویژگی را برای مقابله با عدم تقارن اطلاعات بین روش‌های مختلف معرفی می‌کند. در [۸] بازیابی تصویر مبتنی بر محتوا با یک شبکه عصبی سیامی کانولوشنی پیشنهاد می‌شود. ابتدا، تکه‌های ضایعه برای ایجاد دو مجموعه‌های داده، برش داده می‌شوند و جفت‌های دوتکه دلخواه یک مجموعه‌داده پچ-جفت را تشکیل می‌دهند. دوم، این مجموعه‌داده پچ-جفت برای آموزش یک شبکه استفاده می‌شود. سوم، یک پچ آزمایشی به‌عنوان یک پرس‌وجو در نظر گرفته می‌شود. فاصله بین این پرس‌وجو و بیست مورد در هر دو مجموعه‌داده با استفاده از شبکه عصبی کانولوشنی سیامی آموزش‌دیده محاسبه می‌شود. موارد نزدیک به پرس‌وجو برای ارائه پیش‌بینی نهایی با بالاترین امتیاز استفاده می‌شود. در [۹] روشی پیشنهاد می‌شود که از قدرت شبکه‌های عصبی کانولوشن برای پیش‌بینی دسته‌بندی تصویر ورودی برای همه کلاس‌های خروجی و بازیابی تصاویر با استفاده از تابع فاصله تغییریافته در فضای ویژگی موجک، استفاده می‌کند. در [۱۰] بافت کانال و اطلاعات توالی مکانی برای بازیابی مبتنی بر محتوا مورد تمرکز قرار می‌گیرند. ابتدا یک مدل عمیق پیشنهاد می‌شود که هدف آن استنباط نقشه‌های توجه، در امتداد ابعاد کانال و مکان است. با بهبود ماژول‌های توجه کانال و توجه مکانی و کاوش ترانسفورماتور[۱۲]، توانایی ساخت و درک مدل افزایش می‌یابد. در [۱۱] 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 می‌شود. این تابع در استخراج ویژگی برای تمایز در تقسیم مشابه‌ترین دسته‌ها کمک می‌کند. در [۱۲] یک شبکه ترکیبی مبتنی بر خودتوجهی[۱۳] برای یادگیری بازنمایی‌های رایج داده‌های رسانه‌های مختلف[۱٤] 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 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؛ در [۱۳] یک چارچوب سبک‌تر برای نمونه‌برداری تدریجی از قطعات متمایز، جهت یادگیری جزئیات ارائه می‌شود. در این روش ابتدا شیء از تصویر اصلی تقویت‌شده و سپس یک نمونه‌برداری خودتطبیقی برای شناسایی بیشتر منطقه تقویت‌شده انجام می‌گردد. پس این چارچوب می‌تواند از کل به شیء و از شیء به جزئیات برسد.

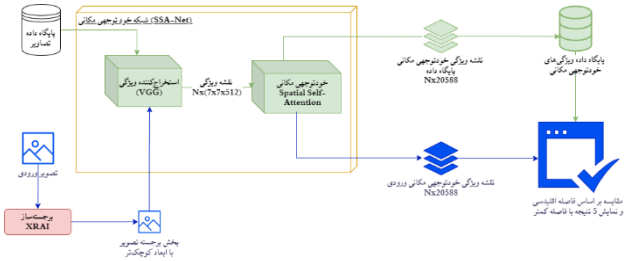
در دسته روش‌های مبتنی بر ریزدانه، در [۱٤] انتخاب توصیفگرهای عمیق مفید به خوبی به تشخیص تصویر با دانه‌ریز کمک می‌کند. به طور خاص، یک مدل جدید شبکه عصبی کانولوشنی ماسک دار[۱۵]، بدون لایه‌های کاملاً متصل پیشنهاد شده است. بر اساس حاشیه‌نویسی‌های بخش، مدل پیشنهادی شامل یک شبکه کاملاً کانولوشنی برای مکان‌یابی قسمت‌های متمایز و مهم‌تر از آن تولید ماسک‌های جسم/قطعه وزن‌دار برای انتخاب توصیفگرهای کانولوشنی مفید و معنادار است. پس از آن، یک مدل سه‌جریانی برای تجمیع توصیفگرهای انتخاب شده در سطح جسم و قطعه به طور هم‌زمان ساخته می‌شود. با کنار گذاشتن پارامتر لایه‌های کاملاً متصل اضافی، این شبکه دارای ابعاد کوچک و سرعت استنتاج کارآمد در مقایسه با سایر روش‌های ریزدانه است. در [۱۵] یک روش تخمین ریزدانه، برای تخمین نمره زیبایی‌شناسی[۱۶] پیشنهاد می‌شود و مکانیسم‌های توجه موقعیت و کانال را برای افزایش ترکیب ویژگی‌های زیبایی‌شناسی

ترکیب می‌کند. با آموزش شبکه رگرسیون، جدا از شبکه طبقه‌بندی، طبقه‌بندی وظیفه رگرسیون را تکمیل می‌کند. محققان به استفاده از میانگین مربع خطا[۱۷] به‌عنوان معیار ارزیابی اصلی عادت کرده‌اند، که در اندازه‌گیری خطای هر بازه ناکافی است. به‌منظور درنظرگرفتن کامل تصاویر، بخش‌های مختلف امتیاز زیبایی‌شناسی، به‌جای تمرکز بر بخش‌های این نمره‌ها به دلیل عدم تعادل مجموعه‌داده‌ها ، یک معیار ارزیابی جدید به نام خطاهای میانگین مربع تقسیم شده[۱۸] برای اثبات مزایا پیشنهاد می‌شود. در [۱۶] که روی طبقه‌بندی تصاویر تمرکز دارد، یک سیستم یادگیری نیمه نظارت تهیه‌شده 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 و طبقه‌بندی می‌کند. در [۱۷] روش یادگیری هش[۱۹] با دو مشکل بررسی می‌شود: ۱ ـ ویژگی‌های با ابعاد کم، فرایند بازیابی را تسریع می‌بخشند اما به دلیل ازدست‌رفتن اطلاعات، دقت را کاهش می‌دهند. ۲ ـ تصاویر ریزدانه منجر به ایجاد کدهای هش جستجوی یکسان در خوشه‌های مختلف در فضای پنهان پایگاه‌داده می‌شوند. از این‌رو به یک شبکه پاک‌کننده توجه مبتنی بر ثبات ویژگی[۲۰] پرداخته‌می‌شود. برای مشکل نخست، از یک ماژول پاک‌کردن ناحیه انتخاب‌شده[۲۱] استفاده می‌کند که با پوشش تطبیقی برخی از مناطق تصاویر خام، شبکه را در برابر تفاوت‌های ظریف ریزدانه‌ای مقاوم می‌کند. بدین ترتیب کدهای هش متمایزتری در پایگاه‌داده هش ذخیره می‌شوند. سپس برای پایدارترکردن رابطه بین کد هش تصویر ورودی و کد هش پایگاه‌داده، از ماژول افزایش خطای رابطه مکانی[۲۲] استفاده می‌کند. در [۱۸] 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 ـ زیرشبکه بازیابی اطلاعات متمایز ریز[۲۳] ۲ ـ زیرشبکه شناسایی با رابطه معنایی خطای تقطیر[۲۴]. 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 پیداکردن جزئیات دقیق و قابل‌اعتماد کمک کند. در [۱۹] روشی برای استفاده از یک مدل توجه چند سطحی[۲۵] 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 ـ که برای طبقه‌بندی ریزدانه‌ای مفید هستند ـ ارائه می‌دهد و شامل دو مرحله است: ۱ ـ انتخاب نورون‌هایی که حداکثر فعال‌سازی را در سه نقشه ویژگی انتخاب‌شده دارند. این نقشه‌ها خروجی مدل‌های شبکه عصبی کانولوشنی هستند که از قبل روی تصاویر اندازه کامل، آموزش‌داده‌شده‌اند. ۲ ـ آموزش شبکه‌های ظریف با این مناطق چند مقیاسی ایجادشده. هر منطقه متمایزشده را می‌توان به‌عنوان یکی از ویژگی‌ها در نظر گرفت. سپس این نتایج برای پیش‌بینی نهایی ادغام می‌شوند. در [۲۰] به یکی از مشکلات بازیابی تصویر ریزدانه‌ای پرداخته می‌شود: تنوع کم در بین کلاس‌های مختلف و درعین‌حال تنوع زیاد در هر کلاس. این پژوهش با بررسی خطای آنتروپی متقاطع[۲۶] 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[۲۷] معرفی می‌کند که پیاده‌سازی آن بسیار آسان‌تر است.

در ادامه در بخش ۲ به مرور مقاله پایه پرداخته می‌شود. در بخش ۳ راهکار پیشنهادی مبتنی بر روش برجستگی ارائه می‌گردد. در بخش ٤ به ارزیابی دقت و پایداری نتایج آزمایشگاهی و در نهایت در بخش ۵ به جمع بندی این مقاله پرداخته شده است.

## ۲- بازیابی تصویر ریزدانه‌ای با استفاده از شبکه خودتوجهی مکانی و مکانیسم برجسته‌سازی

**شکل ۱. شمای کلی شبکه خودتوجهی مکانی با مکانیسم برجسته‌سازی**




در [۲۱] شبکه خودتوجهی مکانی[۲۸] پیشنهاد شده‌است که شامل دو جزء اصلی می‌باشد. ابتدا یک شبکه عصبی کانولوشنی به‌عنوان استخراج‌کننده ویژگی[۲۹] 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.

### ۲ـ۱ـ 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 می‌توانند ویژگی‌های سطح بالا را از تصاویر استخراج کنند. برای مقایسه منصفانه با سایر روش‌های پیشرفته، از VGG-۱۶ از پیش آموزش‌دیده بر روی مجموعه‌داده ImageNet استفاده می‌شود.

برای استخراج اولیه، سه‌لایه آخر که کاملاً متصل هستند حذف می‌شوند. نقشه‌های ویژگی هر تصویر از مجموعه‌داده به عنوان X از خروجی لایه کانولوشنی نهایی گرفته‌می‌شود. این فرآیند به صورت زیر نشان داده شده است

$$F = VGG(X) \qquad (۱)$$

به طور خاص، استخراج‌کننده ویژگی، یک تصویر ورودی $X$ را به یک نقشه ویژگی با ابعاد $F \in R^{H\times W\times K}$ نگاشت می‌کند، که در آن $H$،

W و $K$ نشان دهنده ارتفاع مکانی، عرض مکانی، تعداد کانال ها/کرنل حاوی کانال هستند.

## ۲ـ۲ـ ماژول خودتوجهی مکانی

ماژول خودتوجهی مکانی از مکانیسم خودتوجهی پیشنهاد شده استفاده می‌کند که توجه محلی را از طریق یک تابع فعال‌ساز softmax جمع می‌کند. این ایده گسترش می‌یابد تا به موقعیت‌های پیکسل مکانی ویژگی‌های اصلی توجه و از تجمیع ویژگی‌ها برای به‌دست‌آوردن نقشه‌های ویژگی خودتوجهی مکانی استفاده شود. باتوجه‌به نقشه‌های ویژگی اولیه $F \in R^{\mathrm{H\times W\times K}}$ به دست آمده از استخراج‌کننده ویژگی، ابتدا سه نقشه ویژگی جدید $A$، $B$ و $C$ با استفاده از کانولوشن ۱×۱ تولید می‌شود. $\{\mathrm{A,B,C}\} \in R^{\mathrm{H\times W\times K}}$ همان ابعاد فضای F را داراست. سپس A و B و C به $R^{\mathrm{N\times K}}$ تغییر شکل می‌یابد، که در آن $N = H\mathrm{x}W$ تعداد پیکسل‌ها است. سپس، ضرب عناصر بین $A$ و ترانهاده $B$ محاسبه می‌شود. تابع فعال‌ساز softmax از نظر مکانی برای محاسبه نقشه خودتوجهی مکانی اعمال می‌شود $S \in R^{\mathrm{N\times K}}$ که:

$$S_{ij} = \frac{exp(A_i \otimes B_j)}{\sum_{i=1}^{N} exp(A_i \otimes B_j)} \qquad (۲)$$

در فرمول (۲) $\otimes$ نشان‌دهنده ضرب عنصر به عنصر است. $S_{ij}$ نشان می‌دهد که چگونه شبکه تاثیر $i$مین موقعیت مکانی را بر موقعیت مکانی $j$مین اندازه گیری 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 مکانی، ضرب عناصر بین $S \in R^{\mathrm{N\times K}}$ و $C \in R^{\mathrm{N\times K}}$ انجام می‌شود و نتایج به $R^{\mathrm{H\times W\times K}}$ تغییر شکل داده‌می‌شود.

در نهایت، یک مکانیسم تجمیع ویژگی برای بررسی تأثیر مناطق خودتوجهی مکانی در همه موقعیت‌ها، در نقشه ویژگی اصلی از طریق فرمول (۳) پیاده‌سازی می‌شود:

$$H_j = \sum_{i=1}^{N} (s_{ij} C_i) \oplus F_j \qquad (۳)$$

می‌توان از فرمول (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.

## ۲ـ۳ـ 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 ۲ـ۲ و با استفاده از معادله زیر مقایسه می‌شوند:

$$D(X,Y) = \sqrt{\sum_{i=۱}^{N=۲۰۵۸۸} (X_i - Y_i)^۲} \qquad (٤)$$

در فرمول (٤) X نقشه ویژگی خودتوجهی مکانی تصویر ورودی و Y نقشه ویژگی خودتوجهی مکانی هر تصویر از پایگاه داده است. سپس فاصله‌های به‌دست‌آمده، که هرکدام نگاشتی به تصویری از پایگاه داده دارند، به صورت نزولی مرتب شده و پنج نتیجه برتر بازیابی می‌شود. خروجی سیستم بر اساس کلاسی که بیشترین احتمال را در بین این پنج نتیجه دارد، تعیین می‌گردد.

## ۲ـ٤ـ مکانیسم برجسته‌سازی

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 مقاله، روش XRAI برجسته‌سازی[۳۰] پیشنهاد شده‌است. 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، ۱۵۰، ۲۵۰، ۵۰۰، ۱۲۰۰] استفاده شده و بخش‌های کوچکتر از بیست پیکسل نادیده گرفته‌می‌شود. یک پارامتر واحد، اجتماع بخش‌ها در کل تصویر را محاسبه می‌کند. بنابراین، اجتماع همه بخش ها مساحتی برابر با شش برابر مساحت تصویر ایجاد می‌کند 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 تصاویر در دو طرف یک لبه نازک اغلب به یکدیگر مرتبط هستند. برای این منظور، ماسک های بخش‌ها پنج پیکسل گسترش می‌یابند تا مجموعه نهایی قطعات به دست آید.

**XRAI Algorithm:**

*Given Image I, model f and attribution method g*
*Over-segment I to segments $s \in S$*
*Get attribution map $A = g(f, I)$*
*Let saliency mask $M = 0$, trajectory $T = []$*
*while $S \neq \emptyset$ and area(M) < area(I) do*
  *for $s \in S$*
    *Compute gain: $g_s = \sum_{i \in S \setminus M} \frac{A_i}{area(S \setminus M)}$*
  *end for*
  *$\hat{s} = argmax_s g_s$*
  *$S = S \setminus \hat{s}$*
  *$M = M \cup \hat{s}$*
  *Add M to list T*
*end while*
*return T*

## ۳- ارزیابی دقت و نتایج آزمایشگاهی

در تکمیل بهبود استفاده از شبکه خودتوجهی مکانی، از روش برجسته‌سازی XRAI استفاده شده‌است. مجموعه داده Stanford Dogs [۲۲] شامل ۲۰۵۸۰ تصویر از ۱۲۰ نژاد سگ از سراسر جهان است. این مجموعه‌داده با استفاده از تصاویر و حاشیه نویسی از ImageNet، برای طبقه بندی تصاویر ریز دانه ساخته شده است. این مجموعه‌داده یک مشکل چالش برانگیز داشت، زیرا برخی از نژادهای سگ ویژگی های تقریباً یکسانی دارند یا از نظر رنگ و سن متفاوت

هستند. در آزمایش پیش رو، برای شروع، تصویر 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 ۲ و با استفاده از معادله (٤) مقایسه می‌شوند.سپس فاصله‌های به‌دست‌آمده، که هرکدام نگاشتی به تصویری از پایگاه داده دارند، به صورت نزولی مرتب شده و ۵ نتیجه برتر بازیابی می‌شود. امتیاز سیستم بر اساس کلاسی که بیشترین احتمال را در بین این ۵ نتیجه دارد، تعیین می‌گردد.

$$Score = \frac{True\ Prediction}{All\ Prediction} \qquad (٦)$$

در این آزمایش از یک کامپیوتر شخصی مدل Macbook Air با پردازنده M۱ و حافظه اصلی ۸ گیگابایت استفاده شده است. محیط توسعه و زبان مورد استفاده python می‌باشد.

**شکل۲۱. نمونه‌ای از تصاویر مجموعه‌داده [۲۳]**

**جدول۱. نتایج آزمایش روی داده‌های آزمایشی**

| Input | Outputs | | | | | Score |
|---|---|---|---|---|---|---|
| Lhasa | Lhasa | Lhasa | Lhasa | Lhasa | silky_terrier | ۸۰٪ |
| Irish_setter | Irish_setter | Irish_setter | Irish_setter | Irish_setter | flat_coated_retriever | ۸۰٪ |
| Pomernian | Pomernian | Pomernian | Pomernian | Pomernian | Pomernian | ۱۰۰ % |
| Afghan_hound | bloodhound | Afghan_hound | bloodhound | Afghan_hound | Afghan_hound | ٦۰٪ |
| otterhound | otterhound | otterhound | otterhound | otterhound | otterhound | ۱۰۰ % |
| Afghan_hound | Afghan_hound | Afghan_hound | Afghan_hound | Afghan_hound | Aghan_hound | ۱۰۰ % |
| Pekinese | Pekinese | Shih-Tzu | Shih-Tzu | Pekinese | Pekinese | ٦۰٪ |
| Bernese_mountain | Bernese_mountain | Bernese_mountain | Bernese_mountain | Bernese_mountain | Bernese_mountain | ۱۰۰ % |
| chow | chow | chow | chow | chow | chow | ۱۰۰ % |
| clumber | clumber | clumber | clumber | clumber | clumber | ۱۰۰ % |

**جدول ۲. نتایج آزمایشگاهی**

| Method | Score |
|---|---|
| FCAN [۲٤] | ۸٤/۵٪ |
| PDFR [۲۵] | ۷۱/۹٪ |
| PC-DenseNet-۱٦۱[۲٦] | ۸۳/٦٪ |
| HDWE [۲۷] | ۷۹/٦٪ |
| EfficientNet-B۰[۲۸] | ٦۱/۲٪ |
| PC [۲۹] | ٦۱/۹٪ |
| SSA [۳۰] | ۸٦٪ |
| روش پیشنهادی | ۸۸٪ |

جدول ۲ بیان‌کننده مقایسه نتایج عملکرد روش‌های مختلف در بازیابی تصویر ریزدانه‌ای بر روی مجموعه‌داده Stanford Dogs می‌باشد. در روش EfficeintNet-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 استفاده‌شده‌است که نسبت به روش پیشین ۰/۷ درصد بهبود داشته‌است. روش PDFR، از فیلترهای متمایز و آشکارسازی استفاده می‌کند که نسبت به روش قبلی ۱۰ درصد بهبود ایجاد کرده‌است. در روش‌های FCAN، PC-DenseNet-۱۱٦ و HDWE نیز از شبکه‌های کانولوشنی متعددی استفاده‌شده که نتیجه آن‌ها نسبت به موارد قبلی بین ۸ الی ۱۳ درصد افزایش امتیاز صورت گرفته‌است. 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.

## ٤ ـ نتیجه‌گیری

در این مقاله یک روش بازیابی تصویر ریزدانه‌ای با کمک شبکه خودتوجهی مکانی و مکانیسم توجه پیشنهاد شده‌است. شبکه خودتوجهی مکانی با استخراج ویژگی و استفاده از مکانیسم توجه، ویژگی‌های جدید تصاویر را ذخیره می‌کند و مکانیستم برجسته‌سازی با حذف حاشیه‌های کم‌اهمیت تصویر ورودی، عملکرد سیستم را ارتقا می‌بخشد. نتایج نشان می‌دهند که روش پیشنهادی در بازیابی تصویر ریزدانه‌ای عملکرد مطلوبی دارد.

## مراجع

[۱] Li, J., Ling, Z., Niu, L., & Zhang, L. (۲۰۲۲). Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement. In Computer Vision and Image Understanding (Vol. ۲۱۸, p. ۱۰۳٤۱۲). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.cviu.۲۰۲۲,۱۰۳٤۱۲

[۲] Wang, Y., Huang, F., Zhang, Y., Feng, R., Zhang, T., & Fan, W. (۲۰۲۰). Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. In Pattern Recognition (Vol. ۱۰۰, p. ۱۰۷۱٤۸). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.patcog.۲۰۱۹,۱۰۷۱٤۸

[۳] Bressan, R. S., Bugatti, P. H., & Saito, P. T. M. (۲۰۲۲). Optimum-path forest and active learning approaches for content-based medical image retrieval. In Optimum-Path Forest (pp. ۹۵–۱۰۷). Elsevier. https://doi.org/۱۰,۱۰۱٦/b۹۷۸-۰-۱۲-۸۲۲٦۸۸-۹,۰۰۰۱۲-۸

[٤] Rahman, A., Winarko, E., & Mustofa, K. (۲۰۲۲). Product image retrieval using category-aware siamese convolutional neural network feature. In Journal of King Saud University - Computer and Information Sciences (Vol. ۳٤, Issue ٦, pp. ۲٦۸۰–۲٦۸۷). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.jksuci.۲۰۲۲,۰۳,۰۰۵

[۵] Jiang, Y., Du, J., Xue, Z., & Li, A. (۲۰۲۲). Cross-Media Retrieval of Scientific and Technological Information Based on Multi-Feature Fusion. In Neurocomputing. Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.neucom.۲۰۲۲,۰٦,۰٦۱

[٦] Niu, D., Zhao, X., Lin, X., & Zhang, C. (۲۰۲۰). A novel image retrieval method based on multi-features fusion. In Signal Processing: Image Communication (Vol. ۸۷, p. ۱۱۵۹۱۱). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.image.۲۰۲۰,۱۱۵۹۱۱

[۱۹] Ji, J. , Guo, Y. , Yang, Z. , Zhang, T. , & Lu, X. (۲۰۲۱). Multi-level dictionary learning for fine-grained images categorization with attention model. Neurocomputing, ٤٥٣, ٤٠٣–٤١٢. https://doi.org/١٠,١٠١٦/j.neucom. ٢٠٢٠. ٠٧. ١٤٧

[۲۰] Zeng, X. , Zhang, Y. , Wang, X. , Chen, K. , Li, D. , & Yang, W. (۲۰۱۹). Fine Grained Image Retrieval via Piecewise Cross Entropy loss. Image and Vision Computing. https://doi.org/١٠,١٠١٦/j. imavis.٢٠١٩,١٠. ٠٠٦

[۲۱] Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (۲۰۲۱). Spatial self-attention network with self-attention distillation for fine-grained image recognition. Journal of Visual Communication and Image Representation, ٨١, ١٠٣٣٦٨. https: //doi. org/١٠. ١٠١٦/j. jvcir. ٢٠٢١. ١٠٣٣٦٨

[۲۲] Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (۲۰۱۲). Novel Dataset for Fine-Grained Image Categorization : Stanford Dogs.

[۲۳] Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (۲۰۱۲). Novel Dataset for Fine-Grained Image Categorization : Stanford Dogs.

[۲٤] Liu, X., Xia, T., Wang, J., & Lin, Y. (۲۰۱٦). Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. arXiv preprint arXiv:١٦٠٣,٠٦٧٦٥, ١(٢), ٤.

[۲٥] X. Zhang, H. Xiong, W. Zhou, W. Lin and Q. Tian, "Picking Deep Filter Responses for Fine-Grained Image Recognition," ٢٠١٦ IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ٢٠١٦, pp. ١١٣٤-١١٤٢, doi: ١٠,١١٠٩/CVPR.٢٠١٦,١٢٨.

[۲٦] Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N. (۲۰۱۸). Pairwise Confusion for Fine-Grained Visual Classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV ۲۰۱۸. ECCV ۲۰۱۸. Lecture Notes in Computer Science(), vol ١١٢١٦. Springer, Cham. https://doi.org/١٠,١٠٠٧/٩٧٨-٣-٠٣٠-٠١٢٥٨-٨_٥

[۲۷] Yu, J., Huang, Y., Gbur, G., Wang, F., & Cai, Y. (۲۰۱۹). Enhanced backscatter of vortex beams in double-pass optical links with atmospheric turbulence. In Journal of Quantitative Spectroscopy and Radiative Transfer (Vol. ٢٢٨, pp. ١–١٠). Elsevier BV. https://doi.org/١٠,١٠١٦/j.jqsrt.٢٠١٩,٠٢,٠٢١

[۲۸] D. Haase and M. Amthor, "Rethinking Depthwise Separable Convolutions: How Intra-Kernel Correlations Lead to Improved MobileNets," ٢٠٢٠ IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), ٢٠٢٠, pp. ١٤٥٨٨-١٤٥٩٧, doi: ١٠,١١٠٩/CVPR٤٢٦٠٠,٢٠٢٠,٠١٤٦١.

[۲۹] Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., & Naik, N. (۲۰۱۷). Training with confusion for fine-grained visual classification. CoRR

[۳۰] Baffour, A. A. , Qin, Z. , Wang, Y. , Qin, Z. , & Choo, K. -K. R. (۲۰۲۱). Spatial self-attention network with self-attention distillation for fine-grained image recognition. Journal of Visual Communication and Image Representation, ٨١, ١٠٣٣٦٨. https: //doi. org/١٠. ١٠١٦/j. jvcir. ٢٠٢١. ١٠٣٣٦٨زیرنویس‌ها

[۷] Zhang, L., & Wu, X. (۲۰۲۲). Multi-task framework based on feature separation and reconstruction for cross-modal retrieval. In Pattern Recognition (Vol. ١٢٢, p. ١٠٨٢١٧). Elsevier BV. https://doi.org/١٠,١٠١٦/j.patcog.٢٠٢١,١٠٨٢١٧

[۸] Zhang, K., Qi, S., Cai, J., Zhao, D., Yu, T., Yue, Y., Yao, Y., & Qian, W. (۲۰۲۲). 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. ١٤٠, p. ١٠٥٠٩٦). Elsevier BV. https://doi.org/١٠,١٠١٦/j.compbiomed.٢٠٢١,١٠٥٠٩٦

[۹] Yelchuri, R., Dash, J. K., Singh, P., Mahapatro, A., & Panigrahi, S. (۲۰۲۲). Exploiting deep and hand-crafted features for texture image retrieval using class membership. In Pattern Recognition Letters (Vol. ١٦٠, pp. ١٦٣–١٧١). Elsevier BV. https://doi.org/١٠,١٠١٦/j.patrec.٢٠٢٢,٠٦,٠١٧

[۱۰] Chen, Y. , Zhang, Z. , Wang, Y. , Zhang, Y. , Feng, R. , Zhang, T. , & Fan, W. (۲۰۲۲). AE-Net: Fine-grained sketch-based image retrieval via attention-enhanced network. Pattern Recognition, ١٢٢, ١٠٨٢٩١. https://doi.org/١٠,١٠١٦/j. patcog. ٢٠٢١,١٠٨٢٩١

[۱۱] Zeng, X. , Liu, S. , Wang, X. , Zhang, Y. , Chen, K. , & Li, D. (۲۰۲۱). Hard Decorrelated Centralized Loss for fine-grained image retrieval. Neurocomputing, ٤٥٣, ٢٦–٣٧. https: //doi. org/١٠. ١٠١٦/j. neucom. ٢٠٢١. ٠٤. ٠٣٠

[۱۲] Shan, W., Huang, D., Wang, J., Zou, F., & Li, S. (۲۰۲۲). Self-Attention based fine-grained cross-media hybrid network. In Pattern Recognition (Vol. ١٣٠, p. ١٠٨٧٤٨). Elsevier BV. https://doi.org/١٠,١٠١٦/j.patcog.٢٠٢٢,١٠٨٧٤٨

[۱۳] Guo, C. , Lin, Y. , Chen, S. , Zeng, Z. , Shao, M. , & Li, S. (۲۰۲۲). From the whole to detail: Progressively sampling discriminative parts for fine-grained recognition. Knowledge-Based Systems, ٢٣٥, ١٠٧٦٥١. https://doi.org/١٠,١٠١٦/j. knosys.٢٠٢١,١٠٧٦

[۱٤] Wei, X.-S., Xie, C.-W., Wu, J., & Shen, C. (۲۰۱۸). Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. In Pattern Recognition (Vol. ٧٦, pp. ٧٠٤–٧١٤). Elsevier BV. https://doi.org/١٠,١٠١٦/j.patcog.٢٠١٧,١٠,٠٠٢

[۱٥] Jin, X., Deng, Q., Lou, H., Li, X., & Xiao, C. (۲۰۲۲). Fine-grained Regression for Image Aesthetic Scoring. In Cognitive Robotics. Elsevier BV. https://doi.org/١٠,١٠١٦/j.cogr.٢٠٢٢,٠٧,٠٠٣

[۱٦] Ha, Y. , Du, Z. , & Tian, J. (۲۰۲۲). Fine-grained interactive attention learning for semi-supervised white blood cell classification. Biomedical Signal Processing and Control, ٧٥, ١٠٣٦١١. https://doi.org/١٠,١٠١٦/j.bspc.٢٠٢٢,١٠٣٦١١

[۱۷] Zhao, Q. , Wang, X. , Lyu, S. , Liu, B. , & Yang, Y. (۲۰۲۲). A feature consistency driven attention erasing network for fine-grained image retrieval. Pattern Recognition, ١٢٨, ١٠٨٦١٨. https: //doi. org/١٠. ١٠١٦/j. patcog. ٢٠٢٢. ١٠٨٦١٨

[۱۸] Yan, T. , Shi, J. , Li, H. , Luo, Z. , & Wang, Z. (۲۰۲۲). Discriminative information restoration and extraction for weakly supervised low-resolution fine-grained image recognition. Pattern Recognition, ١٢٧, ١٠٨٦٢٩. https: //doi. org/١٠. ١٠١٦/j. patcog. ٢٠٢٢. ١٠٨٦٢٩

---

۱ Text Based Image Retrieval

۲ Content Based Image Retrieval

۳ Fine-Grained Content Based Image Retrieval

[۴] Zero-Shot Sketch based Image Retrieval (ZS-SBIR)

۵ Asymmetric Disentanglement

۶ Deep Cascaded Cross-modal Ranking Model

۷ Siamese Convolutional Network

۸ Multi-feature Fusion based Cross-Media Retrieval

۹ Adversarial Learning

۱۰ Uniform local binary patterns

۱۱ multi-task framework based on feature separation and reconstruction

۱۲ Transformator

۱۳ Self-Attention Network

۱۴ Cross-Media

۱۵ Mask-Convolutional Neural Network

۱۶ Image Aesthetic Scoring

۱۷ Mean Square Errors

۱۸ Segmented Mean Square Errors

[۱۹] Hash Learning Method

۲۰ Feature Consistency Driven Attention Erasing Network: FCAENet

۲۱ Selective Region Erasing Module: SREM

۲۲ Enhancing Space Relation Loss: ESRL

[۲۳] Fine-Grained discriminative Information Restoration: FDR

۲۴ Semantic Relation Distillation Loss: SRD-Loss

۲۵ Multi-level Attention Model

[۲۶] Cross Entropy Loss

۲۷ Piecewise Cross Entropy loss

۲۸ Spatial Self-Attention Network (SSA.Net)

۲۹ Feature Extractor: FE

۳۰ Saliency