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

بازیابی ریزدانه‌ای تصویر مبتنی بر محتوا

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

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

[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

در این مقاله، شیوه نگارش یك مقاله برای کنفرانس پردازش سیگنال و سیستم‌های هوشمند تشریح می‌شود. روش قالب‌بندی مقاله، بخش‌های مختلف آن، انواع قلم‌ها و اندازه آن‌ها، به طور کامل مشخص شده است. کلیه سبك (Style) های مورد نیاز برای بخش‌های مختلف مقاله، از جمله عنوان‌ها، نویسندگان، چکیده، متن، و ... از پیش تعریف شده‌اند و تنها کافی است سبك مورد نظر را برای بخشی از مقاله انتخاب کنید. نویسندگان محترم مقاله‌ها باید توجه داشته باشند، کنفرانس از پذیرش مقاله‌هایی که خارج از این چارچوب تهیه شده باشند، معذور است. چکیده مقاله باید در یك بند (پاراگراف) تهیه شود و حداکثر شامل ۲۰۰ کلمه باشد. چکیده باید بطور صریح و شفاف موضوع پژوهش و نتایج آن را مطرح کند؛ یعنی بیان کند چه کاری، چگونه، و برای چه هدفی انجام و چه نتایجی حاصل شده است. در چکیده از ذکر جزییات کار، شکل‌ها، جدول‌ها، فرمول‌ها، و مراجع پرهیز کنید.

کلمات کلیدی

بازیابی تصویر – بازیابی ریزدانه‌ای تصویر – بازیابی تصویر مبتنی بر محتوا

۱- مقدمه

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. ازاین‌رو، 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 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 آمده است. به‌گونه‌ای که اکنون با نیاز به بررسی دقیق‌تر دسته‌بندهای

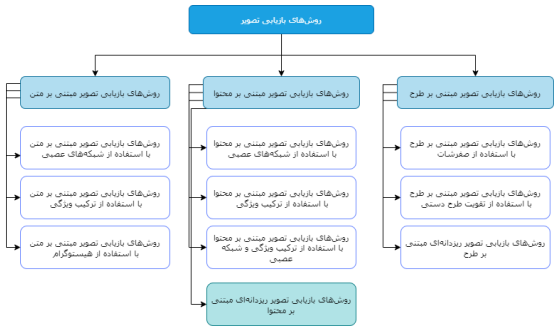
[۱] Text Based Image Retrieval

[۲] Content Based Image Retrieval

تصاویر، بازیابی تصاویر ریزدانه‌ای[۳] معرفی شده است که در پیداکردن ویژگی‌های مشابه، تا حد ادراک انسان رفتار می‌کند.

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 پیداکردن شباهت میان نمونه‌های تصویری مورد بررسی، می‌تواند وظایف مربوط به جست‌وجو را سریع‌تر و کم‌هزینه‌تر انجام دهد.

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 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.



شکل ۱. دسته‌بندی شاخه‌های مختلف روش‌های بازیابی تصویر

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] که هدف آن ایجاد یک روش بازیابی تصویر برای مشخص‌کردن دسته‌بندی یک محصول است، یک مدل شبکه کانولوشنی سیامی[۴] پیشنهاد می‌شود که شامل برچسب‌های دسته و آیتم در آموزش برای تولید 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 مجموعه‌داده به‌عنوان ستون فقرات و شبکه تک‌لایه برای یادگیرنده با ویژگی متوسط پیاده‌سازی می‌شود.

در [۳] 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[۵]، 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 وزن‌دار برای انتخاب توصیف‌گرهای کانولوشنی مفید و معنادار است. پس از آن، یک مدل سه‌جریانی برای تجمیع 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.

در [۴] 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[۶] به‌عنوان معیار ارزیابی اصلی عادت کرده‌اند، که در اندازه‌گیری خطای هر بازه ناکافی است. به‌منظور درنظرگرفتن کامل تصاویر، بخش‌های مختلف امتیاز زیبایی‌شناختی، به‌جای تمرکز بر بخش‌های نمره زیبایی‌شناختی متوسط به دلیل عدم تعادل مجموعه‌داده‌های زیبایی‌شناختی، یک معیار ارزیابی جدید به نام خطاهای میانگین مربع تقسیم شده[۷] برای اثبات مزایا پیشنهاد می‌شود.

در [۵] 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 خصمانه برای بهبود بیشتر دقت بازنمایی زیر فضای عمومی است. سپس از شباهت در همان فضا برای مرتب‌سازی نتایج بازیابی استفاده می‌شود.

در [۶] 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 ویژگی‌های رنگ و بافت را به تصویر بکشد. سپس هیستوگرام الگوهای باینری محلی یکنواخت[۹] 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 درنظرگرفتن شباهت‌های تصاویر قابل‌مقایسه در مجموعه‌داده به‌روزرسانی می‌شود. براین‌اساس، این روش شباهت‌های نهایی تصاویر را به دست می‌آورد.

---

[۳] Fine-Grained Content Based Image Retrieval

[۴] Siamese Convolutional Network

[۵] Mask-Convolutional Neural Network

[۶] Mean Square Errors

[۷] Segmented Mean Square Errors

[۸] Multi-feature Fusion based Cross-Media Retrieval

[۹] Uniform local binary patterns

در [۷] یک چارچوب چند وظیفه‌ای جدید مبتنی بر جداسازی و بازسازی ویژگی [۱۰] 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 و تصویر را معرفی می‌کند و ماژول بازسازی متن برای بهبود کیفیت ماژول جداسازی ویژگی.

این مطالعه [۸] 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 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 ۲۰ وصله در هر دو مجموعه‌داده با استفاده از شبکه عصبی کانولوشنی سیامی آموزش‌دیده محاسبه می‌شود. وصله‌های نزدیک به پرس‌وجو برای ارائه پیش‌بینی نهایی با رأی اکثریت استفاده می‌شود.

در [۹] روشی را پیشنهاد می‌شود که از قدرت شبکه‌های عصبی کانولوشن برای پیش‌بینی عضویت کلاس تصویر پرس‌وجو برای همه کلاس‌های خروجی و بازیابی تصاویر با استفاده از تابع فاصله تغییریافته در فضای ویژگی موجک استفاده می‌کند.

در [۱۰] 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[۱۱] به بازیابی آگاهانه از ساختار رسید. برای این منظور، روش جداسازی نامتقارن آگاه از ساختار [۱۲] 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 [۱۱] بازیابی تصویر ریزدانه‌ای مبتنی بر طرح به‌عنوان یک فرایند درشت به ریز فرموله شده است و یک مدل رتبه‌بندی متقابل آبشاری عمیق[۱۳] 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.

در [۱۲] که روی طبقه‌بندی تصاویر گلبول‌های سفید 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 کرده و طبقه‌بندی می‌کند.

در [۱۳] 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 [۱۴] 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[۱۵] یک ژنراتور لنگر استخراج ویژگی محلی[۱۴] 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[۱۵] 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.

در [۱۶] 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 بهر می‌گیرد و پس از آموزش به یک چارچوب پیشنهادی منتقل می‌کند. سپس ویژگی‌های عمومی را از دو شبکه گرفته و مورد بررسی قرار می‌دهد

در [۱۷] روش یادگیری هش با دو مشکل بررسی می‌شود: ۱ـ ویژگی‌های با ابعاد کم فرایند بازیابی را تسریع می‌بخشند اما به دلیل ازدست‌رفتن اطلاعات، دقت را کاهش می‌دهند. ۲ـ تصاویر ریزدانه منجر به ایجاد کدهای هش جستجوی یکسان در خوشه‌های مختلف در فضای پنهان پایگاه‌داده می‌شوند. پس این پژوهش به یک شبکه پاک‌کننده توجه مبتنی بر ثبات ویژگی[۱۶] می‌پردازد. برای مشکل نخست، از یک ماژول پاک‌کردن ناحیه انتخاب‌شده[۱۷] استفاده می‌کند که با پوشش تطبیقی برخی از مناطق تصاویر خام، شبکه را در برابر تفاوت‌های ظریف ریزدانه‌ای مقاوم می‌کند. پس کدهای هش متمایزتری

---

[۱۰] multi-task framework based on feature separation and reconstruction

[۱۱] Asymmetric Disentanglement

[۱۲] STRucture-aware Asymmetric Disentanglement (STRAD)

[۱۳] Deep Cascaded Cross-modal Ranking Model

[۱۴] Local Feature Extraction Anchor Generator

[۱۵] Symmetrized Local Feature Extraction Module

[۱۶] Feature Consistency Driven Attention Erasing Network: FCAENet

[۱۷] Selective Region Erasing Module: SREM

در پایگاه‌داده هش ذخیره می‌شوند. سپس برای پایدارترکردن رابطه بین کد هش جستجو و کد هش پایگاه‌داده از ماژول افزایش خطای رابطه مکانی[۱۸] استفاده می‌کند.

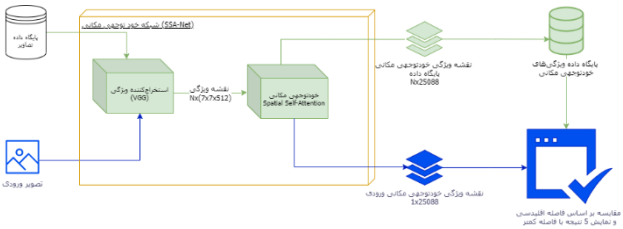
در [۱۸] به یک طرح پیشنهادی برای طبقه‌بندی ریزدانه‌ای انواع محصولات خرده‌فروشی در قفسه سوپرمارکت‌ها پرداخته می‌شود. این طرح، به طور هم‌زمان، نشانه‌های سطحی شیء[۱۹] و نشانه‌های سطحی بخشی از تصاویر محصول[۲۰] را ضبط می‌کند. نشانه‌های سطح شیء تصاویر محصول توسط یک شبکه جدید طبقه‌بندی بازسازی[۲۱] تولید می‌شود. برای مدل‌سازی بدون حاشیه‌نویسی نشانه‌های سطح جزئی، قسمت‌های تبعیض‌آمیز، تصاویر محصول در اطراف نقاط کلیدی شناسایی می‌شوند. این بخش‌ها به‌صورت توالی‌های مرتب‌شده توسط یک حافظه کوتاه‌مدت ـ بلندمدت کانولوشنی کدگذاری می‌شوند و محصولات را به طور منحصربه‌فرد توصیف می‌کنند.

در [۱۹] یک شبکه ترکیبی مبتنی بر خودتوجهی[۲۲] برای یادگیری بازنمایی‌های رایج داده‌های رسانه‌های مختلف[۲۳] 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؛ بنابراین، رویکرد پیشنهادی می‌تواند به طور مؤثر شکاف بین توزیع‌های ویژگی‌های مختلف در وظایف بازیابی بین رسانه‌ای را کاهش دهد.

در [۲۰] 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 این میان ویژگی‌های سلسله‌مراتبی نیز سنجیده می‌شوند که هزینه‌های محاسباتی را کاهش می‌دهد.

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

شبکه خودتوجهی مکانی[۲۴] پیشنهادی از دو جزء اصلی تشکیل شده است (شکل ۲ را ببینید). ابتدا یک شبکه عصبی کانولوشنی به‌عنوان استخراج‌کننده ویژگی[۲۵] پیاده‌سازی می‌شود که ویژگی‌های اولیه را از تصاویر ورودی از طریق چندین لایه کانولوشن و ادغام استخراج می‌کند.



شکل ۲. شبکه خودتوجهی مکانی پیشنهادی

۲ـ۱- 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 مقایسه منصفانه با سایر روش‌های پیشرفته، از VGG-۱۶ از پیش آموزش‌دیده بر روی مجموعه‌داده ImageNet استفاده می‌شود.

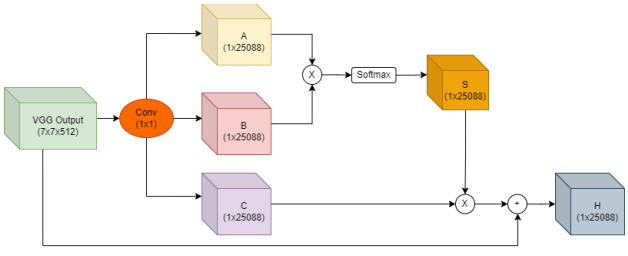
برای استخراج اولیه، سه‌لایه آخر که کاملاً متصل هستند حذف می‌شوند. ویژگی نقشه‌ها از تصاویر ورودی برای یک تصویر ورودی X یک مجموعه داده، خروجی نقشه ویژگی از لایه کانولوشنی نهایی گرفته‌می‌شود. این فرآیند به صورت نشان داده شده است

$$F = VGG(X) \quad (۱)$$

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.

۲ـ۲- ماژول خودتوجهی 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

---

[۱۸] Enhancing Space Relation Loss: ESRL
[۱۹] Object-level
[۲۰] Part-level
[۲۱] Reconstruction-Classification Network: RC-Net
[۲۲] Self-Attention Network
[۲۳] Cross-Media
[۲۴] Spatial Self-Attention Network (SSA.Net)
[۲۵] Feature Extractor: FE

همان‌طور که در شکل ۳ نشان‌داده‌شده است، باتوجه‌به 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 ۱ × ۱ 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 ترانهاده B محاسبه می‌شود. "softmax" از نظر مکانی برای محاسبه نقشه خودتوجهی مکانی اعمال می‌شود $S \in R^{N\times N}$ که:

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

که در آن $\otimes$ نشان‌دهنده ضرب عنصر است. S_{ij} نشان می‌دهد که چگونه شبکه تاثیر iمین موقعیت مکانی را بر موقعیت مکانی jمین 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 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 داده‌می‌شود.

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

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

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.

۳ـ۲- 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:

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

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 هرکدام 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.

٤ـ۲- نوآوری

به عنوان نوآوری در این پژوهش، از متد XRAI شفاف‌سازی[۲۶] 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، ۱۵۰،۲۵۰، ۵۰۰، ۱۲۰۰] استفاده شده و بخش‌های کوچکتر از ۲۰ پیکسل نادیده گرفته‌می‌شود (پارامتر مقیاس عمدتاً بر اندازه بخش‌ها تأثیر می‌گذارد). برای یک پارامتر واحد، اتحاد بخش‌ها کل تصویر را محاسبه می‌کند. بنابراین، اتحاد همه بخش ها 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، ماسک های بخش را ۵ پیکسل گشاد می‌شود تا مجموعه نهایی قطعات به دست آید.

مراجع

[۱] 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/۱۰,۱۰۱۶/b۹۷۸-۰-۱۲-۸۲۲۶۸۸-۹,۰۰۰۱۲-۸

[۲] 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

[۳] 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. ۷٦, pp. ۷۰٤–۷۱٤). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.patcog.۲۰۱۷,۱۰,۰۰۲

[۲۶] Saliency

[۱۲] Ha, Y. , Du, Z. , & Tian, J. (۲۰۲۲). Fine-grained interactive attention learning for semi-supervised white blood cell classification. Biomedical Signal Processing and Control, ۷۵, ۱۰۳۶۱۱. https://doi.org/۱۰,۱۰۱۶/j.bspc.۲۰۲۲,۱۰۳۶۱۱

[۱۳] 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, ۱۲۲, ۱۰۸۲۹۱. https://doi.org/۱۰,۱۰۱۶/j. patcog. ۲۰۲۱,۱۰۸۲۹۱

[۱۴] Zeng, X. , Liu, S. , Wang, X. , Zhang, Y. , Chen, K. , & Li, D. (۲۰۲۱). Hard Decorrelated Centralized Loss for fine-grained image retrieval. Neurocomputing, ۴۵۳, ۲۶–۳۷. https: //doi. org/۱۰. ۱۰۱۶/j. neucom. ۲۰۲۱. ۰۴. ۰۳۰

[۱۵] Yang, M., Xu, Y., Wu, Z., & Wei, Z. (۲۰۲۲). Symmetrical irregular local features for fine-grained visual classification. In Neurocomputing (Vol. ۵۰۵, pp. ۳۰۴–۳۱۴). Elsevier BV. https://doi.org/۱۰,۱۰۱۶/j.neucom.۲۰۲۲,۰۷,۰۵۶

[۱۶] Yuan, P. , Qian, S. , Zhai, Z. , FernánMartínez, J. , & Xu, H. (۲۰۲۲). Study of chrysanthemum image phenotype on-line classification based on transfer learning and bilinear convolutional neural network. Computers and Electronics in Agriculture, ۱۹۴, ۱۰۶۶۷۹. https: //doi. org/۱۰. ۱۰۱۶/j. compag. ۲۰۲۱. ۱۰۶۶۷۹

[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, ۱۲۸, ۱۰۸۶۱۸. https: //doi. org/۱۰. ۱۰۱۶/j. patcog. ۲۰۲۲. ۱۰۸۶۱۸

[۱۸] Santra, B. , Shaw, A. K. , & Mukherjee, D. P. (۲۰۲۲). Part-based annotation-free fine-grained classification of images of retail products. Pattern Recognition, ۱۲۱, ۱۰۸۲۵۷. https://doi.org/۱۰,۱۰۱۶/j.patcog.۲۰۲۱,۱۰۸۲۵۷

[۱۹] Shan, W., Huang, D., Wang, J., Zou, F., & Li, S. (۲۰۲۲). Self-Attention based fine-grained cross-media hybrid network. In Pattern Recognition (Vol. ۱۳۰, p. ۱۰۸۷۴۸). Elsevier BV. https://doi.org/۱۰,۱۰۱۶/j.patcog.۲۰۲۲,۱۰۸۷۴۸

[۲۰] 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, ۲۳۵, ۱۰۷۶۵۱. https://doi.org/۱۰,۱۰۱۶/j. knosys.۲۰۲۱,۱۰۷۶

[٤] 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/۱۰,۱۰۱٦/j.cogr.۲۰۲۲,۰۷,۰۰۳

[٥] 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

[۷] Zhang, L., & Wu, X. (۲۰۲۲). Multi-task framework based on feature separation and reconstruction for cross-modal retrieval. In Pattern Recognition (Vol. ۱۲۲, p. ۱۰۸۲۱۷). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.patcog.۲۰۲۱,۱۰۸۲۱۷

[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. ۱٤۰, p. ۱۰۵۰۹٦). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.compbiomed.۲۰۲۱,۱۰۵۰۹٦

[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. ۱٦۰, pp. ۱٦۳–۱۷۱). Elsevier BV. https://doi.org/۱۰,۱۰۱٦/j.patrec.۲۰۲۲,۰٦,۰۱۷

[۱۰] 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

[۱۱] 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