مجله ماشین‌بینایی و پردازش‌تصویر
سال پنجم، شماره اول، تابستان ۱۳۹۷

پالایش شرح‌گذاری مجموعه تصاویر با مقیاس بزرگ با یادگیری انتقالی در شبکه عصبی کانولوشنال عمیق

شیما جوانمردی[۱] و محمدعلی زارع چاهوکی[۲]

چکیده

فرآیند پالایش شرح‌گذاری تصاویر، رویکردی موثر در بهبود بازیابی تصاویر مبتنی بر برچسب می‌باشد. در شبکه‌های اجتماعی و موتورهای جستجو بسیاری از تصاویر دارای تگ‌های مبهم، ناقص و بی‌ارتباط با محتوا هستند. وجود این تگ‌های غیرقابل اعتماد، موجب کاهش دقت بازیابی تصاویر می‌شود. از این‌رو در دهه اخیر، الگوریتم‌هایی با عنوان پالایش تگ (TR) مطرح شده‌اند که به رفع نویز و غنی‌سازی برچسب‌های تصاویر می‌پردازند. به‌منظور دستیابی به نتایج بهینه در TR، استخراج ویژگی‌هایی از تصویر که توصیف مناسبی از محتوای دیداری تصویر داشته باشند، تاثیر مستقیمی بر دقت فرآیند TR دارد. از جمله چالش‌های عمده در فرآیند پالایش شرح‌گذاری تصاویر، رسیدن به توصیفی مناسب و مرتبط با محتوای تصاویر می‌باشد. بدین منظور با توجه به کارآمدی فرآیند یادگیری عمیق در بسیاری از حوزه‌های پژوهشی، در این مقاله نیز به منظور استخراج ویژگی‌های کارآمد در تشابه دیداری تصاویر و ارتباط معنایی تصاویر با هم، از شبکه‌های عصبی کانولوشنال عمیق (DCNN) استفاده شده‌است. بهره‌گیری از فرآیند یادگیری انتقالی استفاده شده در DCNN مبتنی بر تصاویر ImageNet در توصیف و ایجاد ارتباط معنایی در مجموعه تصاویر با مقیاس بزرگ NUS-WIDE، بیانگر موثر بودن این رویکرد در کاربرد پالایش تگ تصاویر می‌باشد.

کلید واژه‌ها

پالایش شرح‌گذاری تصاویر، شبکه عصبی کانولوشنال عمیق، پالایش تگ، بازیابی تصاویر، یادگیری انتقالی

۱ مقدمه

با پیشرفت تکنولوژی تصویر برداری دیجیتال شاهد رشد بسیار سریع حجم دادگان تصویری می‌باشیم. این روند ضرورت وجود تکنولوژی‌های توسعه یافته برای حجم بالایی از تصاویر در کاربردهای مرتبط با بازیابی تصویر را ایجاد می‌کند. بازیابی تصاویر به دو صورت بازیابی مبتنی بر محتوا(CBIR) و بازیابی مبتنی بر برچسب[۱] (TBIR) صورت می‌گیرد[۱]. در CBIR با استفاده از ویژگی‌های دیداری تصویر، همچون رنگ، بافت و لبه‌های اشیاء درون تصویر، بازیابی تصاویر مرتبط با پرس‌وجوی کاربر انجام می‌شود. از جمله چالش‌های مطرح در CBIR می‌توان به وجود شکاف معنایی میان ویژگی‌های پایین رتبه و مفاهیم سطح بالای تصاویر و وجود ابعاد بالای ویژگی‌های دیداری تصاویر اشاره کرد. بنابراین وجود ویژگی‌هایی که به طور موثرتری قادر به توصیف محتوای تصویر باشند در این حوزه بسیار اثربخش خواهد بود. در مقابل، TBIR تنها با توجه به اطلاعات متنی تصاویر به جستجوی آنها می‌پردازد. اطلاعات متنی در مقایسه با اطلاعات دیداری با سادگی بیشتری گویای مفاهیم تصویر هستند. به همین دلیل TBIR می‌تواند به عنوان راه‌حلی ساده برای مقابله با

این مقاله در شهریورماه سال ۱۳۹۵ دریافت، دراسفندماه بازنگری و در اردیبهشت‌ماه سال ۱۳۹۷ پذیرفته شد.

[۱] دانشجوی دکتری مهندسی کامپیوتر گرایش هوش مصنوعی، دانشکده برق و کامپیوتر، دانشگاه یزد.
رایانامه: sh.javanmardi@stu.yazd.ac.ir
[۲] دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد.
رایانامه: iranpour@pnu.ac.ir

نویسنده مسئول: محمدعلی زارع چاهوکی

چالش‌های CBIR ارائه شود. اطلاعات متنی استفاده شده در فرآیند TBIR از عناوین تصاویر [۲]،متن، فرامتن‌های اطراف آنها [۳] و شرح‌گذاری‌های صورت گرفته توسط کاربران حاصل می‌شود [۴]. کارایی TBIR تا حد زیادی به کیفیت تگ‌های اعمال شده توسط کاربران بر تصاویر بستگی دارد. وجود نویز در اطلاعات متنی، شرح‌گذاری‌های غیر مرتبط با محتوای تصویر و عدم وجود اطلاعات مرتبط با تصاویر از جمله چالش‌های مطرح شده در این حوزه می‌باشد. شکل۱ نمونه‌ای از بازیابی تصاویر مبتنی بر جستجوی تگ "شکوفه" در موتور جستجوی پارسی‌جو[1] را نشان می‌دهد. وجود تصاویر غیر مرتبط با پرس‌وجوی مورد جستجوکه با کادر قرمز مشخص شده‌اند، در رتبه‌های بالای تصاویر بازیابی شده، از جمله مشکلات وجود تگ‌های نویز در تصاویر وب می‌باشد.

شکل۱ تصاویر حاصل از فرآیند بازیابی تصاویر مبتنی بر برچسب در موتور جستجوی پارسی‌جو

در دهه اخیر به‌منظور حل چالش مطرح شده، با استفاده رویکردهایی تحت عنوان پالایش تگ تصویر[2](ITR) به صورت خودکار به رفع نویز از شرح‌گذاری‌های تصاویر و تکامل تگ‌های مرتبط با محتوای تصویر می‌پردازند. پالایش شرح‌گذاری تصاویر فرآیندی است که در طی آن تگ‌های غیر مرتبط با مفاهیم تصویر از میان تگ‌های اولیه کاربران برروی تصاویر حذف می‌شوند و فرآیند غنی‌سازی تگ‌ها با دیگر برچسب‌های مرتبط صورت می‌گیرد. از آنجا که این فرآیند تا حد زیادی می‌تواند باعث کاهش شکاف معنایی میان ویژگی‌های سطح پایین و مفاهیم معنایی سطح بالای تصاویر شود، بنابراین رویکردی موثر در فرآیند بازیابی تصاویر مبتنی بر تگ به‌شمار می‌آید. با توجه به تقسیم‌بندی ارائه شده در[۵]، روش‌های پالایش تگ‌های تصاویر را می‌توان به دو دسته روش‌های مبتنی بر استقرا و روش‌های مبتنی بر استنتاج تقسیم کرد که در بخش بعد مورد بررسی قرار می‌گیرند. این دسته‌بندی بر اساس ایجاد و یا عدم ایجاد تمایز میان داده‌های آموزشی و آزمون و نحوه استخراج قوانین بر روی داده‌ها صورت می‌گیرد. روش‌هایی تحت عنوان روش‌های مبتنی بر نمونه،در دسته الگوریتم‌های استقرایی معرفی می‌شود که رویکرد پیشنهادی این پژوهش، در این گروه از الگوریتم‌ها قرار می‌گیرد. یکی از چالش‌های مطرح در حوزه پالایش تگ تصاویر، استخراج ویژگی‌های مهم از تصویر می‌باشد. از آنجا که ویژگی‌های مختلف، تاثیر متفاوتی در تشخیص تشابه دیداری میان تصاویر دارند، از این رو چگونگی استخراج ویژگی‌های تصویر به گونه‌ای که تا حد امکان کامل بوده و بیانگر محتوای ذاتی تصویر باشند، امری مهم تلقی می‌شود. تمام ویژگی‌های کنونی دارای محدودیت در توصیف تصاویر هستند و تحلیل و پردازش ویژگی‌های تصاویر با ابعاد بالا بسیار مسئله پیچیده‌ای است. همچنین عدم وجود پایگاه‌های داده تصویری قابل قبول و رایج برای فرآیند یادگیری و ارزیابی در ITR چالشی مطرح در این حوزه می‌باشد.

اخیرا روش‌هایی تحت عنوان یادگیری انتقالی در حوزه یادگیری ماشین مطرح شده‌اند که در یادگیری از داده‌های متفاوت از آزمون استفاده می‌کنند و به ارائه طبقه‌بندهایی برای داده‌های دیگر، در حوزه‌ای متفاوت می‌پردازند. مزیت کلی یادگیری انتقالی ایجاد طبقه‌بندهایی متناسب با داده‌های جدید می‌باشد که این موجب کاهش قابل توجه زمان و حجم عملیات مورد نیاز برروی دادهای آزمون می‌باشد. زمانی که حجم داده‌های نهایی نسبت به نمونه‌های آموزشی اولیه کوچکتر باشد، یادگیری انتقالی ابزاری بسیار قدرتمند به منظور یادگیری شبکه بزرگی از داده‌ها، با حداقل‌سازی بیش‌برازش ایجاد می‌نماید. از این رو در بسیاری از پژوهش‌ها از این مزیت استفاده می‌شود [۶].

با توجه به کارآمدی فرآیند یادگیری عمیق[3] در بسیاری از حوزه‌های پژوهشی و بکارگیری فرآیند یادگیری انتقالی در این حوزه، در این مقاله به‌منظور استخراج ویژگی‌های کارآمد در محاسبه تشابه دیداری تصاویر و مجموعه مفاهیم سطح بالای پایگاه‌های تصویری برچسب‌دار، از شبکه عصبی کانولوشنال عمیق[4](DCNN) استفاده شده‌است. در یادگیری انتقالی استفاده شده در DCNN، شبکه برروی مجموعه‌ای از داده‌های اولیه آموزش داده می‌شود و کلیه وزن‌های موجود در شبکه بر اساس آن نمونه‌های آموزشی بدست می‌آید. طبقه‌بند یادگیری شده از داده‌های آموزشی، به‌منظور پیش‌بینی و یا توصیف ویژگی‌های مجموعه داده‌های آزمون با دامنه متفاوت می‌باشد.

در ادامه ساختار بیان مطالب ارائه شده در این مقاله عنوان می‌گردد. در بخش دو به معرفی روش‌های پالایش تگ تصاویر

[1] http://parsijoo.ir/
[2]Image Tag Refinement (ITR)
[3]Deep Learning (DL)
[4]Deep Convolutional Neural Network(DCNN)

می‌پردازیم. در بخش سه فرآیند یادگیری عمیق و ساختار شبکه‌های عصبی کانولوشنال عمیق مورد بررسی قرار می‌گیرد. در بخش چهار به معرفی روش پیشنهادی در این پژوهش می‌پردازیم. در بخش پنج مشخصات داده‌های مورد استفاده ، معرفی معیارهای ارزیابی و نتایج آزمایش‌های صورت گرفته ارائه می‌شود. در نهایت در بخش شش نتیجه‌گیری کلی از پژوهش و بیان پیشنهادهایی برای پژوهش‌های آینده ارائه می‌گردد.

۲ روش‌های پالایش تگ تصاویر

بسته به اینکه الگوریتم‌های یادگیری تمایزی میان نمونه‌های آموزشی و آزمون قائل می‌شود یا خیر، روش‌های یادگیری به دو دسته مبتنی بر استقراء[1] و مبتنی بر استنتاج[2] تقسیم می‌شوند[۵]. در روش‌های مبتنی بر استقراء، یادگیرنده سعی می‌کند تا با استفاده از استقراء، تابع تصمیمی را نتیجه بگیرد که داراي نرخ خطای پایینی در تمامی توزیع‌های داده‌های آموزشی و آزمون براي یك یادگیری خاص باشد. از این رو در این روش‌ها بدون توجه به داده‌های آزمون، یکسری قواعد کلی بر اساس یادگیری نمونه‌های آموزشی نتیجه می‌شود و یا به تخمین مدلی عام بر اساس نمونه‌های یادگیری شده می‌پردازد. بر خلاف روش‌های استقرایی، در روش‌های استنتاجی قواعد بدست آمده از داده‌های آموزشی، تنها بر مجموعه‌ای خاص از داده‌های آزمون قابل اعمال می‌باشد. در این روش‌ها می‌توان بدون ایجاد تمایز میان داده‌های آموزشی و آزمون به استنتاج قوانین بر کل مجموعه داده‌های ارائه شده پرداخت.

الگوریتم‌های مبتنی براستقرا با توجه به اینکه از نوع جداساز-های پارامتری هستند یا از نوع مولدهای غیرپارامتری، خود به دودسته(۱) مبتنی برنمونه[3] و (۲) مبتنی بر مدل[4] تقسیم می‌شوند. الگوریتم‌های مبتنی بر استنتاج نیز با توجه به مدل یادگیری شامل دو زیرمجموعه(۱) مبتنی بر تجزیه ماتریس و (۲) مبتنی برگراف می‌باشند. درادامه هرکدام از این روش‌ها به طور خلاصه معرفی می‌شوند.

۲-۱ الگوریتم‌های مبتنی بر نمونه

در الگوریتم‌های مبتنی بر نمونه، هر نمونه‌ی آزمون با تمام نمونه‌های آموزشی مقایسه می‌شود. این روش‌ها زیرمجموعه‌ای از الگوریتم‌های غیر پارامتری هستند که در آنها کلیه‌ی فرضیه‌ها بر مبنای نمونه‌های آموزشی ساخته می‌شود. از جمله ویژگی‌های این دسته از الگوریتم‌ها این است که با افزایش تعداد نمونه‌های آموزشی، پیچیدگی آنها افزایش می‌یابد. انواع الگوریتم‌های رای-دهی همسایه از جمله این الگوریتم‌ها هستند [۷].

در [۸] لی[5] و همکارانش الگوریتمی ارائه می‌دهند که در آن میزان ارتباط تگ t به تصویر x بر اساس تعداد وقوع شرح‌گذاری با تگ t در همسایه‌های مشابه دیداری تصویر x می‌باشد. همسایه-های دیداری تصویر xبر اساس ویژگی‌های سراسری و با استفاده از روش‌های محاسبه معیار فاصله مشخص می‌شوند. یوریچیو[6] و همکارانش در [۹] چارچوبی برای پالایش تگ مبتنی بر تکنیکk همسایه نزدیکتر ارائه داده‌اند. ایده اصلی این تکنیک انتخاب مجموعه‌ای از تصاویر مشابه دیداری و سپس در نظر گرفتن مجموعه‌ای از تگ‌های مرتبط بر اساس یک روند انتقال تگ می-باشد. در پژوهش آنها، از معیاری برای ارتباط تگ استفاده می‌شود که میزان توزیع هر تگ در مجموعه تصاویر همسایه تصویر آزمون و در کل مجموعه تصاویر را محاسبه می‌کند. بابهیت[7]و همکارانش همکارانش در [۱۰] به منظور کاهش شکاف معنایی میان ویژگی-های سطح پایین تصاویر و مفاهیم معنایی سطح بالا از گراف مستقیم ترکیب شده از گراف ترکیبی تصویر-تگ و گراف تشابه تصویر استفاده می‌کنند. در پژوهش آنها به منظور بهبود عملیات بازیابی تصاویر از اعمال رویکرد قدم‌زنی تصادفی[8] برروی گراف ترکیبی ایجاد شده استفاده می‌شود.

۲-۲ الگوریتم‌های مبتنی بر مدل

مبنای این خانواده از الگوریتم‌های یادگیری، مدل‌های پارامتری هستند که از یادگیری نمونه‌های آموزشی حاصل شده‌اند. انواع طبقه‌بند SVM در این مجموعه از الگوریتم‌ها قرار می‌گیرند. در [۱۱] چن[9]و همکارانش چارچوبی مشتمل از فاز یادگیری و تخمین ضرایب مرتبط پرس‌وجو ارائه می‌دهند. در پژوهش آنها از روشی تحت عنوان AFSVM[10] به منظور تخمین میزان ارتباط تگ‌ها به تصاویر استفاده می‌شود. سپس با بهره‌گیری از فاز پالایشی تحت عنوان LapRLS[11] عملیات پالایش برچسب‌های پیش‌بینی شده برای بهبود بازیابی تصاویر فلیکر[12]مبتنی بر تگ انجام می‌گیرد. در [۱۲]مدلی ارائه می‌شود که طی سه فازفیلتر، پالایش و غنی‌سازی تگ‌هابه پالایش برچسب‌های تصاویر می-پردازد. در پژوهش آنها برای کاهش زمان محاسباتی ابتدا الگوریتم K-Means بر روی تصاویر پیاده‌سازی شده تا تصاویر مشابه دسته‌بندی شوند. سپس تشابه دیداری بر مبنای ویژگی‌های کم‌رتبه

[1] Inductive Based
[2] Transductive Based
[3] Instance Based
[4] Model Based

[5] Li
[6] Uricchio
[7] Bobhate
[8] Random Walk
[9] Chen
[10] Augmented Feature Support Vector Machine (AFSVM)
[11] Laplacian Regularize Least Square (LapRLS)
[12] Flickr

و تشابه معنایی میان تگ‌ها و تصاویر محاسبه می‌شود و با توجه به فرض مقاله به دنبال حداقل شدن فاصله ایندو معیار می‌باشند.

۲-۳ الگوریتم‌های مبتنی بر تجزیه ماتریس

در این روش‌ها ورودی، ماتریسی حاوی ضرایب ارتباط تگ‌ها و تصاویر می‌باشد و خروجی الگوریتم، ماتریس بازسازی شده می‌باشد. در [۱۳] چارچوبی تحت عنوان تکامل تگ‌ها به‌وسیله بازسازی پراکنده خطی نمای دوجانبه[1] (DLSR) ارائه می‌شود. در این روش بابکارگیری ماتریس تگ‌های اولیه و تجزیه آن بصورت سطری و ستونی، به بازسازی تصاویر در هر ردیف و تگ‌ها در هر ستون می‌پردازد. لین[2] و همکارانش با بکارگیری ویژگی‌های کم‌رتبه رتبه تصاویر و بردار تگ استفاده شده در دو دیدگاه تصویر و تگ، به محاسبه ماتریس‌های شباهت تصاویر و همبستگی تگ‌ها می‌پردازند، سپس با ترکیب خطی این دو ماتریس بردارتگ‌گذاری نهایی برای تصاویرایجاد می‌کنند. در [۱۴] مسئله پالایش تگ بصورت تجزیه ماتریس تگ‌های ایجاد شده توسط کاربر به یک ماتریس تگ تصفیه شده و یک ماتریس پراکندگی خطا، انجام شده‌است. هدف پژوهش آنها بهینه سازی اندازه‌گیری چهار جنبه (۱) همبستگی معنایی میان تگ‌های کم‌رتبه فراهم شده توسط کاربر، (۲) سازگاری محتوای تصویر با تگ‌ها، (۳) همبستگی میان تگ‌ها و (۴) ماتریس پراکندگی خطا می‌باشد. در [۱۵] نیز به منظور بهبود دقت فرآیند بازیابی تصاویر، رویکرد تکامل تگی ارائه می‌شود که با توجه به تگ‌های اولیه ایجاد شده توسط کاربران، به ایجاد ماتریس تگ-تصویر می‌پردازد. سپس با استفاده از ویژگی‌های دیداری و هم‌پوشانی میان تگ‌ها ماتریس ضرایب تگ و تصویر را ایجاد کرده و بصورت فرآیندی تکرارشونده در هر مرحله کلیه ضرایب متعلق به تگ‌های تمام تصاویر به‌روز می‌شوند. ماتریس ضرایب نهایی برای فرآیند جستجوی تصاویر مبتنی بر تگ مورد استفاده قرار می‌گیرد.

۲-۴ الگوریتم‌های مبتنی بر گراف

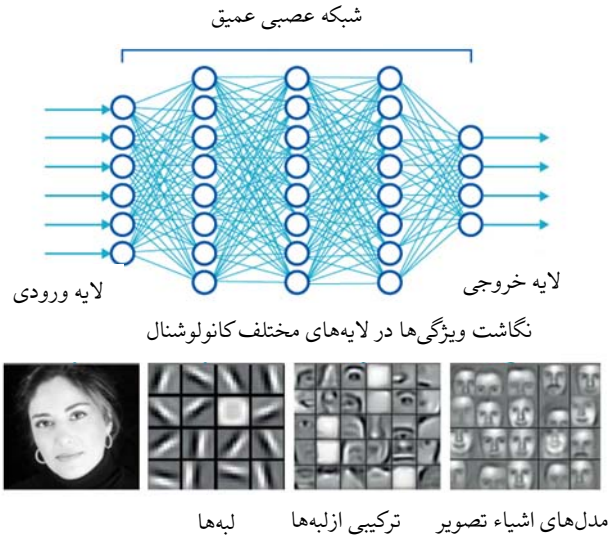
این الگوریتم‌ها دسته دیگری از الگوریتم‌های مبتنی بر استنتاج هستند که به محاسبه ضرایب ارتباط میان تگ و تصویر می‌پردازند. در[۱۶] چارچوبی ارائه می‌شود که در آن به‌منظور بازیابی تگ‌های مرتبط از دست رفته و حذف تگ‌های مبهم، ابتدا تصاویر مرتبط معنایی با توجه به تگ‌های اولیه کاربران از مجموعه تصاویر آموزشی انتخاب می‌شوند، سپس بر اساس شباهت دیداری میان تصاویر، گراف ستاره‌ای تصاویر مرتبط معنایی ساخته می‌شود. در مرحله بعد، از میان تصاویر موجود تصاویری که هم از لحاظ دیداری و هم از لحاظ محتوایی مشابه هستند انتخاب شده و مجموعه واژگانی از تگ‌های کاندید برای تصاویر مشخص می‌شود و با استفاده از یک تابع هدف احتمالی، هم‌ترازی میان تصویر و تگ‌های کاندید بر اساس مفاهیم وابسته به دیدگاه ارائه شده، اندازه‌گیری می‌شود. لیو[3] و همکارانش در [۱۷] به منظور رفع تگ‌های ناقص و مبهم شمای تگ‌گذاری مجددتصاویر با هدف پالایش تگ‌ها ارائه می‌دهند. در این فرآیند برچسب‌گذاری مجدد به عنوان مسئله یادگیری چند برچسبی مبتنی بر گراف چندگانه فرموله‌سازی می‌شود که بطور همزمان به بررسی محتوای دیداری تصاویر، همبستگی معنایی برچسب‌ها و همچنین اطلاعات پیشین منتقل شده توسط کاربران می‌پردازد. الگوریتم ارائه شده به انتشار اطلاعات هر تگ در امتداد گراف شباهت مختص تگ‌ها می‌پردازد.

۳ یادگیری عمیق

هدف از این پژوهش برقرای ارتباط معنایی مناسب بین ویژگی‌های سطح پایین تصویر و مفاهیم سطح بالای آن به‌منظور کاهش شکاف معنایی در بازیابی تصاویر می‌باشد. در سال‌های اخیر روشی با عنوان یادگیری عمیق به منظور کاهش این شکاف معنایی ارائه شده است [۱۸]. یادگیری عمیق تکنیکی مبتنی بر شبکه‌های عصبی است و به عنوان زیر مجموعه‌ای از یادگیری ماشین به‌شمار می‌رود. معماری‌های یادگیری عمیق متنوعی وجود دارد که از جمله‌ی آنها می‌توان به شبکه عصبی کانولوشن[4] [۱۹](CNN)، شبکه باور عمیق[5] [۲۰]و شبکه عصبی بازگشت‌کننده[6] [۲۱] اشاره کرد. موثر بودن معماری‌های معرفی‌شده در زمینه‌های متنوعی از قبیل تشخیص اشیاء[7] تصویر [۲۲]، تشخیص خودکار گفتار[8] [۲۳]، شناسایی چهره[9] [۱۹]، پردازش زبان طبیعی[10] [۲۴]به اثبات رسیده است. CNN از معروف‌ترین و موفق‌ترین معماری‌های یادگیری عمیق در حوزه آنالیز تصاویر بوده و لایه کانولوشن (Conv)[11] آن، هسته اصلی تشکیل دهنده این شبکه می‌باشد. وظیفه اصلی این لایه‌ها استخراج سلسه مراتبی از ویژگی‌های غیرخطی تصاویر می‌باشد. به‌منظور استخراج نگاشت ویژگی[12] متفاوت از تصاویر در هر لایه کانولوشن، این فیلترها در امتداد پهنا و ارتفاع بر سطح تصویر غلتانده می‌شود. ویژگی‌های استخراج شده خاصیت سلسله مراتبی دارند به این معنی که در لایه‌های ابتدایی،بطور مثال گوشه‌ها و خط‌ها و لبه‌ها، یادگرفته می‌شوند و در لایه‌های عمیق‌تر به ترتیب ویژگی‌هایی با سطوح بالاتر،

[1] Dual-view linear sparse reconstructions (DLSR)
[2] Lin
[3] Liu
[4] Convolution Neural Network (CNN)
[5] Deep Belief Network
[6] Recurrent Neural Network
[7] Object Detection
[8] Automatic Speech Recognition
[9] Face Recognition
[10] Natural Language Processing
[11] Convolution (Conv)
[12] Features Map

همانند اشیاء درون تصویر، استخراج می‌شوند. شکل۲ نمایی از سلسله مراتب ویژگی‌های استخراج شده از چهره انسان طی سه لایه کانولوشن متوالی را نشان می‌دهد.



شکل ۲استخراج ویژگی های سلسله مراتبی از تصویر صورت انسان. همانطور که مشاهده می شود در هر لایه کانولوشن تعدادی فیلتر بر سطح تصویر اعمال می شود و یکسری نگاشت ویژگی حاصل می شود که در لایه‌های پایین این ویژگی‌ها، سطح پایینی دارند و با افزایش تعداد لایه ها سطوح ویژگی افزایش می یابند[1].

جز لایه کانولوشن، لایه‌ی ادغام[2]و لایه تماما متصل (FC)[3]نیز دیگر لایه‌های تشکیل‌دهنده این شبکه می‌باشند. کارکرد لایه ادغام، کاهش اندازه عرض و ارتفاع تصویر ورودی به‌جهت کاهش تعداد پارامترها و محاسبات در داخل شبکه و بنابراین کنترل بیش‌برازش[4] می‌باشد. در لایه FC، نورون‌هایی که در یک لایه قرار دارند، دقیقا همانند شبکه های عصبی معمولی، با تمام نورون‌های موجود در لایه قبلی ارتباط دارند. تنها تفاوت بین لایه تماما متصل و کانولوشنی این است که نورون‌ها در هر لایه کانولوشن تنها به ناحیه‌ای محلی از نورون‌های لایه قبل متصل هستند و به اشتراک پارامترها با یکدیگر می‌پردازند. آخرین لایه شبکه لایه سافت‌مکس[5] است که دسته‌بندی تصاویر در ۱۰۰۰ طبقه معرفی شده را نشان می‌دهد.

یکی از رویکردهایی که تابه‌حال در حل چالش دقت محاسبات پالایش تگ‌های تصاویر مبتنی بر ویژگی‌های دیداری و معنایی، به آن پرداخته نشده است استفاده از CNN از روش‌های یادگیری عمیق می‌باشد. معماری‌های متفاوتی برای CNN وجود دارد که تفاوت آنها در تعداد و ساختار لایه‌های میانی می‌باشد. از جمله معماری‌های مفید در حوزه پردازش و استخراج ویژگی[6] از تصاویر، معماری VGGNet[7]می‌باشد. VGGNet شبکه عصبی کانولوشن بسیار عمیقی است که توسطکارن سیمونیان[8] واندرو زیسرمن[9] توسعه داده شده است. این معماری به عنوان دومین شبکه پیشنهادی برنده مسابقات ILSVRC2014 در عملیات طبقه‌بندی تصاویر[10]اما با بهترین عملکرد در عملیات تشخیص اشیاء نسبت به دیگر معماری‌های ارائه شده در آن رقابت، انتخاب شده است [۲۵]. نسخه از پیش‌آموزش دیده شده از این شبکه، توسط ۱,۲ میلیون تصویر موجود در دادگان ImageNet[۲۶] با رزولوشن بالا و اندازه ۲۲۷×۲۲۷ در ۱۰۰۰ طبقه یادگیری شده است. این معماری دارای چندین پیکربندی می‌باشد که تنها تفاوت آنها در عمق شبکه می‌باشد. حداقل عمق این شبکه دارای ۱۱ لایه است که شامل هشت لایه کانولوشن و سه لایه تمام متصل می‌باشد و حداکثر عمق آن ۱۹ لایه است که دارای ۱۶ لایه کانولوشن و سه لایه تمام اتصال می‌باشد. ویژگی شاخص ارائه شده توسط این معماری بالا بودن عمق شبکه به عنوانمولفه‌ای حیاتی است که موجب عملکرد مناسب این شبکه شده‌است. نسخه نهایی بهترین شبکه آن شامل ۱۶ لایه Conv/FC بوده که در این پژوهش مورد استفاده قرار می‌گیرد.

عملکرد مناسب معماری VGGNet در پردازش تصاویر با مقیاس بزرگ و برتری این معماری نسبت به دیگر معماری‌های موجود از جمله[۲۷]MSRA، کلاریفی[11][۲۸]، اورفیت[12][۲۹]،و الکس‌نت[13][۳۰]، در مقایسه معیار پنج خطای برتر[14] محاسبه شده توسط این معماری‌ها برروی مجموعه تصاویر ImageNet در [۲۵] نشان داده شده‌است. برخلاف قدرت کمتر دسته بندیVGGNet نسبت بهGoogLeNet، این شبکه در وظایف یادگیری انتقالی چندگانه[15]از GoogLeNet بهتر عمل می‌کند. بنابراین شبکه VGGNet در حال حاضر محبوب‌ترین انتخاب برایاستخراج ویژگی از تصاویر می‌باشد[۳۱].از این‌رو در این پژوهش، با استفاده از ویژگی‌ها و برچسب‌های استخراج شده از شبکه‌های عصبی کانولوشن بسیار عمیق VGGNet-16، به پالایش تگ‌های تصاویر می‌پردازیم.

[1]http://www.amax.com/blog/wp-content/uploads/2015/12/blog_deeplearning3.jpg
[2]Pooling
[3]Fully Connected (FC)
[4]Overfittng
[5]Soft-Max

[6]Feature Extraction
[7] Visual Geometric Group Network (VGGNet)
[8]Karen Simonyan
[9]Andrew Zisserman
[10] Image Classification
[11]Clarifi
[12]Overfeat
[13]AlexNet
[14]Top-5-error
[15]Multiple transfer learning tasks

۴ روش پیشنهادی

در این بخش رویکرد پالایش تگ تصویر ارائه شده در این پژوهش معرفی می‌شود. در ادامه از این رویکرد با نام DCNN-TR[1] یاد می‌کنیم. رویکرد ارائه شده علاوه بر بهره‌گیری از مزایای فرآیند یادگیری انتقالی از جمله کاهش زمان و حجم عملیات مورد نیاز در آنالیز تصاویر، در فرآیند پالایش مجموعه تصاویر با مقیاس بزرگ نیز مفید خواهد بود. لذا در این پژوهش با بهره‌گیری از فرآیند یادگیری انتقالی از شبکه از پیش یادگیری‌شده با تصاویر آموزشی ImageNet استفاده می‌کنیم. بدیهی است که با توجه به تنوع و همه منظوره بودن مجموعه داده آموزشی ImageNet[۲۶]، بارگذاری وزن‌های یادگیری شده برای اهداف مختلف در مجموعه داده‌های دیگر امکان‌پذیر بوده و موجب صرفه‌جویی در زمان یادگیری شبکه و انجام محاسبات خواهد شد. مراحل چارچوب رویکرد پیشنهادی بصورت گام‌های الگوریتم ۱ ارائه می‌گردد.

الگوریتم ۱ مراحل چارچوب رویکرد پیشنهادی این مقاله

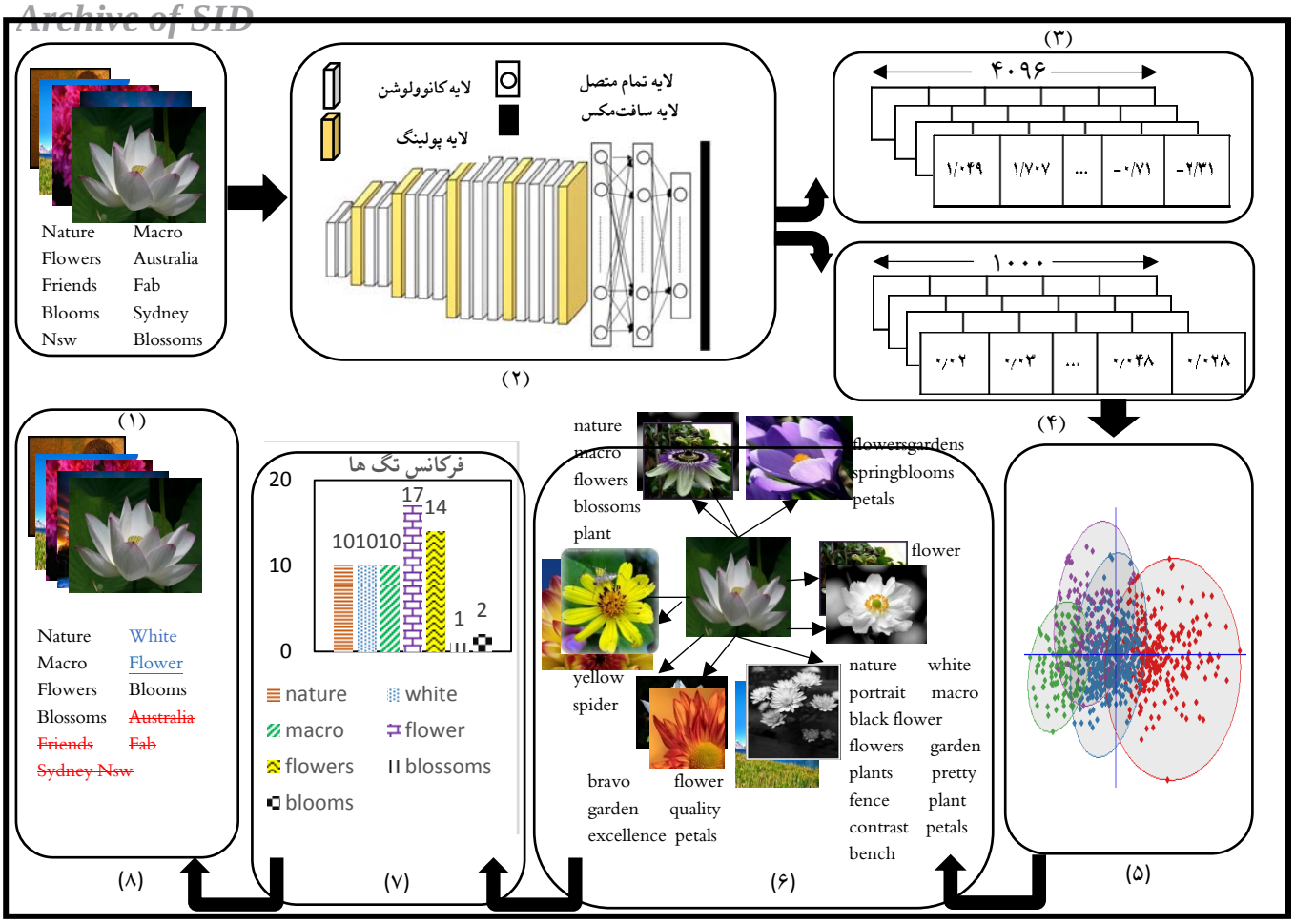
ورودی : تصاویر برچسب‌دار حاوی تگ‌های پالایش نشده
گام ۱: پیش‌پردازش تصاویر، انتخاب تگ‌های با بیشترین فرکانس
گام ۲ : یادگیری انتقالی تصاویر با استفاده از شبکه عصبی کانولوشنال عمیق(DCNN)
گام ۳ : استخراج بردار ویژگی و بردار برچسب پیشنهادی به ازای هر تصویربا DCNN
گام۴: خوشه بندی تصاویر با استفاده از بردار ویژگی‌های استخراج شده ازتصاویر درگام قبل
گام۵: محاسبه شباهت معنایی میان تصاویر هر خوشه با ارزیابی همپوشانی میان تگ‌ها
گام۶: انتخابnنزدیکترین تصویر مشابه دیداری و معنایی به هر تصویر
گام۷: محاسبه فرکانس برچسب‌های تصاویر همسایه
گام۸: انتقال برچسب های با فرکانس بالا از تصاویر همسایه به تصویر آزمون
گام۹: حذف برچسبهای با فرکانس پایین در تصاویر همسایه از تصویر آزمون
خروجی : تصاویر حاوی تگ‌های پالایش شده

همانطور که در گام‌های چارچوب ارائه شده بیان شده است، نخست کلیه تصاویر مورد پیش‌پردازش قرار می‌گیرند و تصاویر قابل دسترس با بالاترین فرکانس تگ‌ها انتخاب می‌شوند. تصاویر حاصل از فاز پیش‌پردازش با هدف محاسبه ویژگی‌ها و درجه احتمال تعلق هر نمونه به هر کلاس را به عنوان ورودی، به یک شبکه عصبی کانولوشنال عمیق داده می‌شود. در فاز سوم، یک بردار ویژگی به ابعاد ۴۰۹۶×۱ و یک بردار برچسب به ابعاد ۱۰۰۰×۱ با ضرایب احتمال عضویت هر نمونه به هر کلاس، به ازای هر تصویر از شبکه استخراج می‌شود. در مرحله چهارم به منظور مقیاس‌پذیری فرآیند پالایش مجموعه داده‌های با مقیاس بزرگ، با استفاده از بردار ویژگی‌های پایین رتبه، تصاویر خوشه‌بندی شده تا تصاویر مشابه دیداری دسته‌بندی شوند. در مرحله پنجم در هر خوشه با بررسی همبستگی میان تگ‌ها، تشابه معنایی میان تصاویر با استفاده از ضرایب تعلق برچسب‌ها ارزیابی شده و تصاویر همسایه متعلق به هر تصویر محاسبه می‌شود. در نهایت در گام‌های بعد، با بکارگیری تکنیک رای‌گیری نزدیک‌ترین همسایه[2] عملیات انتقال تگ‌های مرتبط و با فرکانس بالا از تصاویر تصاویر همسایه به تصویر آزمون و حذف تگ‌های غیر مرتبط از آن، عملیات پالایش تگ صورت می‌گیرد. نوآوری این مقاله، بهره‌گیری از شبکه‌های عصبی کانولوشنال عمیق به‌منظور استخراج ویژگی‌های دیداری و معنایی از تصاویر برای نخستین بار در فرآیند پالایش تگ‌های تصاویر می‌باشد که نتایج آزمایش‌های تجربی موثر بودن این رویکرد را در این حوزه نشان می‌دهد.طرح کلی چارچوب روش پیشنهادی در شکل۳ نشان داده شده است. در ادامه به توضیح مراحل آن می‌پردازیم.

فرض شود$X = \{x1, x2, \dots, xN\}$مجموعه تصاویر اولیه باشد و$T = \{w1, w2, \dots, wM\}$به مجموعه تگ‌های پیش‌بینی شده برای تصاویر اشاره می‌کند که N بیانگر تعداد تصاویر و M بیانگر تعداد تگ‌های تصاویر می‌باشد. به منظور بیان تعلق و عدم تعلق تگ w_Mبه تصویر x_N از $y_{N,t} \epsilon \{0,1\}$ استفاده شده است. برای ارائه نتایج پالایش تصاویر، ماتریس دیگری تحت عنوان Y معرفی می‌کنیم که هر عضو آن بصورت$Y_{i,j} \geq 0$است و به درجه اطمینان تعلق هر تگ wi به xj اشاره می‌کند. به ازای هر تصویر دو بردار$yi = (yi1, yi2, \dots, yim)$و$vi = (vi1, vi2, \dots, viz)$معرفی می‌شود که به ترتیب بیانگر بردار درجه اطمینان تخصیص تگ‌ها به iمین تصویر و بردار ویژگی‌های دیداری تصاویر می‌باشد. همانطور که در شکل نشان داده شده است بعد از ورود تصاویر، در اولین مرحله از فرآیند پالایش تگ‌های تصاویر، به استخراج ویژگی‌های دیداری و معنایی از تصاویر، با استفاده از شبکه کانوولوشن VGGNet می‌پردازیم. در این راستا ابتدا هرکدام از نمونه‌های x_i به شبکه کانوولوشن VGGNet عنوان ورودی داده می‌شود و به ازای هر نمونه بردارy_i با m=۱۰۰۰ ضریب برچسب از آخرین لایه FC و بردار v_i باz=۴۰۹۶ ویژگی از دومین لایه FC شبکه استخراج می‌شود. ازy_i برای انتخاب تصاویر مشابه معنایی و ازv_i برای انتخاب تصاویر مشابه دیداری استفاده می‌شود.

[1] Deep Convolutional Neural Network Tag Refinement (DCNN-TR)

[2] Nearest Neighbor Voting



شکل۳چارچوب پالایش تگ تصویر DCNN-TR ارائه شده در این پژوهش.در مرحله (۱)تصاویر برچسب‌دار به سیستم وارد می‌شوند. در مرحله (۲) یادگیری انتقالی تصاویر با شبکه عصبی کانوولوشنال بسیار عمیقVGGNet صورت گرفته و در مرحله (۳) و (۴) استخراج ویژگی‌های تصاویر از لایه دوم تمام متصل و استخراج برچسب‌های تصاویر از لایه سوم تمام متصل انجام می‌شود. درمرحله (۵) تصاویر با استفاده از بردار ویژگی‌های استخراج شده از شبکه در مرحله (۳) خوشه بندی می‌شود. در مرحله (۶) با استفاده از ویژگی‌های استخراج شده در مرحله (۴)، همسایه‌های متعلق به هر تصویر با استفاده از تشابه معنایی میان تصاویر در هر خوشه انتخاب شده ودر مرحله (۷)پالایش برچسب‌های تصاویر آزمون با استفاده از رویکرد رای‌گیری از نزدیک‌ترین تصاویر همسایه و انتخاب برچسب‌های با فرکانس بالا صورت می‌گیرد. در نهایت در مرحله (۸) تصاویر با تگ‌های پالایش شده استخراج می‌شوند.

در مرحله بعد به منظور ایجاد مقیاس‌پذیری سیستم پالایش تگ‌های تصاویر و ساده‌سازی محاسبات به خوشه‌بندی تصاویر می‌پردازیم. از این‌رو با ورود بردارهای v_i به خوشه‌بند K-Means، تصاویر مشابه دیداری دسته‌بندی می‌شوند. بعد از خوشه‌بندی نمونه‌ها عملیات پالایش در هر خوشه به‌طور مجزا صورت می‌گیرد. از آنجاکه ویژگی‌های دیداری تصویر به تنهایی قادر به شناسایی دقیق محتوای تصویر نیستند، از این رو با استفاده از بردار y_i متعلق به هر تصویر، به محاسبه تشابه معنایی میان تصاویر مشابه دیداری در هر خوشه می‌پردازیم. استفاده ترکیبی از تشابه دیداری و معنایی به طور موثرتری می‌تواند از وقوع پدیده پالایش مشابه تصاویر متفاوت با شرح‌گذاری‌های مشابه جلوگیری کند. بنابراین ابتدا با در نظر گرفتن یک حد آستانه، بردار y_i که حاوی ضرایب تعلق تصاویر به ۱۰۰۰ طبقه از تصاویر ImageNet می‌باشد را به $y_{NM} \in \{0,1\}$ تبدیل می‌کنیم. به این صورت که ضرایب بیشتر از β را به ۱ و سایر ضرایب به صفر که نشان از عدم تعلق تگ w_mبه تصویر x_N می‌باشد، تبدیل می‌شود. سپس با محاسبه تشابه میان تگ‌ها و همپوشانی میان آنها به ارزیابی تشابه معنایی میان تصاویر می‌پردازیم. برای محاسبه تشابه معنایی میان تگ‌های w_i و w_j از رابطه زیر استفاده می‌شود:

$$S(wi,wj) = exp\,(-d(wi,wj)) \qquad (۱)$$

در این رابطه $d(wi,wj)$ فاصله میان تگ‌های wiو wjمی‌باشد

که برای محاسبه آن مشابه [۱۴] از روش فاصله‌سنجی گوگل[1] استفاده می‌شود که به صورت رابطه۲ به محاسبه هم‌پوشانی میان تگ‌ها می‌پردازد.

$$d(wi,wj) = \frac{max\,[log\,q(wi), log\,q(wj)] - log\;\; q(wi,wj)}{log\,N - min[\,log\,q(wi), log\,q(wj)]} \qquad (۲)$$

در این رابطه $q(w_i)$ و $q(w_j)$ تعداد تصاویر حاوی تگ w_i و w_j می‌باشد. $q(w_i,w_j)$ تعداد تصاویری می‌باشد که حاوی هر دو تگ w_i و w_j است. N تعداد کل تصاویر موجود در مجموعه داده می‌باشد. به‌منظور محاسبه تشابه معنایی میان تصاویر مشابه روش ارائه شده در [۱۲] استفاده می‌کنیم. به این صورت که از ضرب نقطه‌ای ماتریس تشابه میان تگ‌ها (S) و بردار درجه اطمینان تعلق هر تگ به تصویر (Y) استفاده می‌کنیم. رابطه ارائه شده بصورت زیر می‌باشد:

$$y_i S y_i^T = \sum_{k,l=1}^{m} Y_{ik}\, S_{kl} Y_{j,l} \qquad (۳)$$

در نهایت با استفاده از تشابه دیداری و معنایی بکار گرفته‌شده، تصاویر همسایه متعلق به هر تصویر محاسبه می‌شود. از آنجا که هدف ما در راستای پالایش تگ‌های تصاویر انتخابی، پیش‌بینی لیستی از تگ‌ها بر مبنای تگ‌های تصایر همسایه مشابه می‌باشد، از این رو در چارچوب پالایش تگ ارائه شده مشابه [۸]،[۹]،[۳۲] از روش رای‌گیریهمسایه[2] استفاده شده است.به‌منظور انجام عملیات پالایش، تگ‌های هر تصویر با تگ‌های تصاویر همسایه مقایسه می‌شود. کلیه عملیات مقایسه تصاویر و پالایش تگ با توجه به برچسب‌های متعلق به دادگان آزمون صورت می‌گیرد. بدیهی است که با توجه به بهره‌مندی از فرآیند یادگیری انتقالی، دامنه برچسب‌های دادگان آزمون با دامنه برچسب‌های بکارگرفته شده در شبکه کانولوشن VGGNet متفاوت می‌باشد و فرض بر این است که هیچ‌گونه ارتباطی میان برچسب‌های استخراج شده از شبکه کانولوشن و برچسب‌های دادگان آزمون وجود ندارد. هرچند که وجود اشتراک میان این دودسته برچسب ۱۰۰۰ تایی عملا امکان‌پذیر می‌باشد. عملیات پالایش برچسب‌های تصاویر طی دو مرحله زیر انجام می‌گیرد:

۱. با در نظر گرفتن حد آستانه α، تمامی تگ‌های با فرکانس وقوع بالاتر از این حد آستانه در تصاویر همسایه، به تصویر آزمون منتقل می‌شود.
۲. تگ‌هایی از تصویر آزمون که در تصاویر همسایه آن رخ نداده‌اند از آن حذف می‌شود و در صورت وقوع در تصاویر همسایه با فرض محتمل شدن صحت آن تگ، حفظ می‌شود. ماتریس برچسب حاصل از فرآیند پالایش به منظور محاسبه میزان کارایی رویکرد پیشنهادی با استفاده از معیارهای ارزیابی مورد بررسی قرار می‌گیرد.

[1] Google Distance
[2] Neighbor Voting

۵ نتایج تجربی

در این بخش نتایج ارزیابی روش DCNN-TR ارائه شده بر مجموعه‌ای از داده‌ها با مقیاس بزرگ ارائه شده است. از آنجایی که پالایش تگ‌های تصاویر مسئله‌ای چالش برانگیز به‌شمار می‌رود، لذا به‌منظور نشان دادن صلاحیت روش ارائه شده، از مجموعه‌ای داده با مقیاس بزرگ استفاده شده است. بدین ترتیب در بخش ۱-۵ به معرفی داده‌های مورد استفاده می‌پردازیم. در بخش ۲-۵ معیارهای ارزیابی بکار گرفته شده مورد معرفی قرار می‌گیرد. در بخش ۳-۵ نحوه تنظیم پارامترهای بهینه آزمایش بیان می‌شود و در نهایت در بخش ۴-۵ نتایج حاصل از پیاده‌سازی روش DCNN-TR ارائه خواهد شد.

۵-۱ مجموعه دادهای مورد استفاده

به‌منظور ارزیابی کیفیت رویکرد DCNN-TRاز مجموعه تصاویر با مقیاس بزرگ NUS-WIDE-270K استفاده شده است[۳۳]. مجموعه داده NUS-WIDE یکی از مجموعه داده‌های تصویری در مقیاس بزرگ است که توسط آزمایشگاه اطلاعات چندرسانه‌ای دانشگاه ملی سنگاپور فراهم شده است. این مجموعه شامل ۲۶۹۶۴۸ تصویر جمع‌آوری شده از سایت فلیکر[3] به همراه ۵۰۱۸ تگ ایجاد شده توسط کاربران می‌باشد که به‌صورت دستی شرح‌گذاری مجدد شده‌اند. از آنجا که تگ‌های ابتدایی ایجاد شده توسط کاربران به شدت ذهنی و یا شخصی هستند بنابراین به نظر می‌رسد حذف برچسب‌های شخصی راهی ساده برای بهبود کیفیت فرآیند پالایش برچسب‌های تصویر است. به همین دلیل به‌عنوان اولین گام در چارچوب محاسبه ضرایب ارتباط تگ و تصویر مورد توجه قرار می‌گیرد. با فرض موثرتر بودن برچسب‌های با فرکانس بالا مشابه [۱۵]، [۱۳] در فرآیند پالایش شرح‌گذاری تصاویر کلیه آزمایش‌ها بر مجموعه تصاویر حاوی ۱۰۰۰ تگ با بیشترین فرکانس صورت گرفته است. همچنین طی یک مرحله عملیات پیش‌پردازشی، با حذف آدرس‌های نامعتبر و تصاویر غیرقابل دسترس از کل مجموعه تصاویر NUS-WIDE، زیر مجموعه‌ای با تعداد ۲۲۱۸۱۷ تصویر حاصل شد که کلیه آزمایش‌ها بر این زیر مجموعه از تصاویر انجام می‌گیرد. در ادامه از این مجموعه به عنوان NUS-WIDE-220K یاد می‌کنیم. در این مجموعه به ازای هر تصویر ۵ دسته ویژگی سراسری محاسبه شده است. این ویژگی‌ها عبارتند از: ۱- هیستوگرام رنگ[4]LAB با ۶۴ بعد، ۲- ممان رنگ[5] مبتنی بر بلاک با ۲۲۵ بعد، ۳- کورلوگرام رنگ[6]HSV با

[3] www.flickr.com
[4] Color Histogram
[5] Color Moment
[6] Color Correlogram

۱۴۴ بعد، ۴- هیستوگرام جهت لبه[1] با ۷۳ بعد و ۵- موجک بافت[2] با ۱۲۸ بعد. به‌منظور بررسی میزان کارایی رویکرد DCNN-TR پیشنهاد شده بر فرآیند پالایش شرح‌گذاری تصاویر، به مقایسه نتایج حاصل از آزمایش‌های انجام شده با بکارگیری شبکه کانوولوشن VGGNet و بدون استفاده از آن می‌پردازیم. از این رو مشابه [۳۴]،[۳۵] و [۳۶]کلیه نتایج حاصل از DCNN-TR، با استفاده از سه دسته ویژگی سراسری موجود برای تصاویر این مجموعه که دارای اهمیت بیشتری نسبت سایر ویژگی‌ها می‌باشد،ارائه می‌گردد. این دسته ویژگی عبارت‌اند از: (۱) کورلوگرام رنگ HSV با ۱۴۴ بعد، (۲) هیستوگرام جهت لبه با ۷۳ بعد و (۳) موجک بافت با ۱۲۸ بعد. بردار ویژگی سطح پایین هر تصویر از کنار هم قرار دادن اینسه بردار ویژگی به دست می‌آید، بنابراین هر تصویر دارای برداری به طول ۳۴۵ ویژگی می‌باشد. در ادامه از آزمایش‌های صورت گرفته با استفاده از این دسته از ویژگی‌های دادگان NUS-WIDE، با عنوان NUS-TR یاد می‌کنیم.

۵-۲ معیارهای ارزیابی

به‌منظور ارزیابی نتایج حاصل از رویکرد DCNN-TRارائه شده، مشابه روش ارائه شده در [۱۴]، [۹]،[۳۲]نسخه ۸۱ تگ NUS-WIDE-220k شرح‌گذاری شده توسط افراد خبره را به عنوان ماتریس پایه درنظر می‌گیریم. برای ارزیابی نتایج پالایش تگ‌های تصاویر از سه معیار ارزیابی شناخته‌شده میانگین دقت[3]، میانگین فراخوان[4]و $F1$استفاده شده است. اگر TP[5]تعداد نمونه‌های مثبتی باشد که به درستی مثبت پیش‌بینی شده‌اند، TN[6]بیانگر تعداد نمونه‌های منفی باشد که به درستی منفی پیش‌بینی شده‌اند، FP[7]تعداد نمونه‌های منفی باشد که به اشتباه مثبت پیش‌بینی شده‌اند و FN[8] بیانگر تعداد نمونه‌های مثبتی باشد که به اشتباه منفی در در نظر گرفته شده‌اند، در این صورت معیارهای ارزیابی دقت، فراخوان و $F1$ به ترتیب توسط روابط ۴ تا ۶ محاسبه می‌شود:

$$\text{دقت} = \frac{TP}{TP + FP} \quad (۴)$$

$$\text{فراخوان} = \frac{TP}{TP + FN} \quad (۵)$$

$$F1 = \frac{\text{دقت} \times \text{فراخوان} \times ۲}{\text{دقت} + \text{فراخوان}} \quad (۶)$$

در محاسبه مقادیر این سه معیار در مسائل چند برچسبه از استراتژی‌های آماری میکرو و ماکرو استفاده می‌شود. از اینرو مشابه[۹] مقادیر میکرو و ماکرو به ازای هر معیار محاسبه شده است. در میانگین‌گیری ماکرو ابتدا بر اساس مقادیر TP، TN، FP و FN معیارهای ارزیابی به‌ازای هر برچسب بطور جداگانه محاسبه شده و سپس از آن مقادیر، میانگین گرفته می‌شود. در حالی‌که در میانگین‌گیری میکرو، ابتدا از تمامی مقادیر TP، TN، FP و FN میانگین گرفته می‌شود و سپس معیارهای ارزیابی محاسبه خواهند شد. روابط ۷ و ۸ نحوه محاسبه میانگین‌گیری ماکرو و میکرو را نشان می‌دهد:

$$\text{میانگین‌گیری‌ماکرو} = \sum_{i=1}^{M} (\text{محاسبه بهارزیابیمعیارازای } TPi، TNi، FPi و FNi) \quad (۷)$$

$$\text{میکرومیانگین‌گیری} = \left(\sum_{i=1}^{M} TPi, \sum_{i=1}^{M} TNi, \sum_{i=1}^{M} FPi, \sum_{i=1}^{M} FNi\right) \text{به‌ازایارزیابیمعیارمحاسبه} \quad (۸)$$

در این روابط M بیانگر تعداد تگ‌های تصاویر می‌باشد. بدین ترتیب در این پژوهش معیارهای دقت ماکرو، فراخوان ماکرو، $F1$ ماکرو، دقت میکرو، فراخوانی میکرو و $F1$ میکرو، به‌عنوان معیار ارزیابی نتایج پالایش تگ‌های تصاویر استفاده شده است.

۵-۳ تنظیم پارامترهای بهینه آزمایش

با توجه به حجم بالای داده‌های مورد استفاده، به‌منظور تنظیم پارامترهای مورد نیاز در رویکرد DCNN-TR، نسخه کاهش یافته از دادگان اصلی مورد استفاده قرار گرفته است. به این صورت که به صورت تصادفی ۲۵k تصویر از داده‌های NUS-WIDE-220K استخراج و کلیه آزمایش‌ها ابتدا بر روی این زیرمجموعه از تصاویر به ازای مقادیر مختلف از پارامترها انجام شد. در نهایت،کلیه آزمایش‌ها با فرض مناسب بودن پارامترهای بدست آمده از زیر مجموعه‌ای از داده‌ها برای مجموعه کل داده‌ها، انجام گرفت. دستیابی به نتایج مناسب حاکی از عملکرد درست فرضیه درنظر گرفته‌شده می‌باشد. از این‌رو در این پژوهش مقادیر بهینه محاسبه شده برای پارامتر β به عنوان حد آستانه انتخابی برای تبدیل احتمال عضویت برچسب‌های استخراج شده از شبکه کانوولوشن VGGNetبه ۱۰۰۰ کلمه کلیدی، مقادیر ۱۰، ۵ در نظر گرفته شده است. همچنین مقدار پارامتر K به عنوان تعداد خوشه‌ها و α به عنوان حد آستانه برای انتخاب تگ‌های تصاویر با فرکانس بالا به ترتیب ۱۰۰ و ۱۰ می‌باشد. همچنین با توجه به بهره‌مندی از فرآیند خوشه‌بندی تصاویر به‌منظور مقیاس‌پذیری سیستم،کوچک‌ترین خوشه دارای ۶۲ تصویر و بزرگ‌ترین آن با ۹۱۴۱ تصویر حاصل شد. از این‌رو حداکثر تعداد همسایگی‌های در نظر گرفته شده ازای هر نمونه آزمون، N=۶۰ می‌باشد و به ازای تعداد همسایگان N=۱۰,۲۰,۳۰,۴۰,۵۰ نیز نتایج آزمایش‌ها ارائه می‌گردد.

[1] Edge Direction Histogram
[2] Wavelet Texture
[3] Average Perecision (AP)
[4] Average Recall(AR)
[5] True Posetive (TP)
[6] True Negetive (TN)
[7] False Posetive (FP)
[6] False Negetive (FN)

۴-۵ ارائه و تحلیل نتایج تجربی

با توجه با اینکه در اکثریت پژوهش‌ها معیار ماکرو *F1* مورد محاسبه قرار می‌گیرد، بنابراین تمرکز ما در این پژوهش بر معیار ماکرو *F1* می‌باشد. نتایج حاصل از آزمایش‌های صورت گرفته با استفاده از رویکرد DCNN-TR برروی دادگان NUS-WIDE-۲۲۰k با ۱۰۰۰ تگ با بیشترین فرکانس در جدول۱ ارائه می‌گردد. همانطور که مشخص‌است، نتایج حاصل از پیاده سازی رویکرد DCNN-TR با بهره‌گیری از فرآیند یادگیری انتقالی در CNN در بهترین حالت منجر به دستیابی به مقادیر *F1* ماکرو ۶۵/۴۷٪ و *F1* میکرو ۳۱/۷۶٪می‌باشد. این در حالی است که با توجه به نتایج ارائه شده در جدول۲، مقادیر بدست آمده برای معیارهای فوق در صورت استفاده از ویژگی‌های پایین رتبه ارائه شده توسط NUS-WIDE در رویکرد NUS-TR در بهترین حالت ماکرو *F1* برابر با ۴۷/۷۳٪و در بهترین حالت میکرو*F1* برابر با ۴۵/۴۰٪شده است. اختلاف مشاهده شده در مقادیر بدست آمده حاکی از موثر بودن فرآیند یادگیری انتقالی پیاده‌سازی شده در شبکه کانوولوشن VGGNet در فرآیند پالایش شرح‌گذاری تصاویر می‌باشد. همچنین این بهبود نشان از اهمیت بکارگیری این روش در پالایش شرح‌گذاری مجموعه تصاویر در مقیاس بزرگ نیز می‌باشد.

جدول۱ نتایج حاصل از پیاده‌سازی رویکرد CNN-TR@n بر داده‌های NUS-WIDE-220k به ازای *n* همسایه مختلف

روش	دقت میکرو (%)	فراخوان‌میکرو (%)	میکرو F1 (%)	دقت ماکرو (%)	فراخوان ماکرو (%)	ماکرو F1 (%)
CNN-TR@10	۹۲/ ۴۲	۵۳/ ۷۵	۶۹/ ۷۸	۳۸/ ۸۱	۹۲/ ۴۹	۵۵/ ۶۶
CNN-TR@20	۹۲/ ۱۱	۱۳/۶۵	۷۶/ ۳۱	۴۹/ ۳۹	۸۷/ ۶۰	۶۳/ ۱۷
CNN-TR@30	**۷۷/ ۴۹**	**۷۱/ ۱۲**	**۷۴/ ۱۷**	**۵۵/ ۲۲**	**۸۰/ ۴۰**	**۶۵/ ۴۷**
CNN-TR@40	۹۹/ ۴۲	۵۳/ ۷۵	۶۹/ ۷۷	۵۹/ ۱۳	۷۳/ ۱۸	۶۵/ ۴۰
CNN-TR@50	۹۲/ ۱۱	۶۵/ ۱۳	۶۹/ ۷۸	۶۲/ ۲۱	۶۷/ ۰۸	۶۴/ ۵۵
CNN-TR@60	۴۸/ ۱۱	۷۹/ ۹۴	۶۰/ ۰۶	۶۴/ ۶۰	۶۱/ ۹۵	۶۳/ ۲۵

همانطورکه در جدول۲ نشان داده شده است بهترین معیار ماکرو F1 محاسبه شده با استفاده از سه دسته از ویژگی‌های پایین رتبه ارائه شده توسط NUS-WIDE، به ازای همسایگی n=۶۰ حاصل شده است که اختلاف آن با معیار ماکرو F1 نهایی محاسبه شده توسط رویکرد DCNN-TR نزدیک به ۷/۱۷٪ می‌باشد.

جدول۲ نتایج حاصل از پیاده‌سازی رویکرد NUS-TR@n بر داده‌های NUS-WIDE-220k به ازای *n* همسایه مختلف

روش	دقت میکرو (%)	فراخوان‌میکرو (%)	میکرو F1 (%)	دقت ماکرو (%)	فراخوان ماکرو (%)	ماکرو F1 (%)
NUS-TR@10	۷۹/ ۱۷	۳۱/ ۸۳	۴۵/ ۴۰	۱۲/ ۶۲	۹۳/ ۵۹	۲۲/ ۲۳
NUS-TR@20	۲۳/ ۷۲	۴۶/ ۶۰	۳۱/ ۴۴	۲۱/ ۳۸	۷۷/ ۳۷	۵۱/۳۳
NUS-TR@30	۱۷/ ۳۳	۵۵/ ۸۹	۲۶/ ۴۶	۲۸/ ۷۰	۷۰/ ۴۸	۴۰/ ۷۹
NUS-TR@40	۱۳/ ۹۹	۶۱/ ۷۳	۲۲/ ۸۱	۳۳/ ۹۵	۶۴/ ۷۴	۴۴/ ۵۴
NUS-TR@50	۱۱/ ۹۷	۶۶/ ۱۱	۲۰/ ۲۷	۳۸/ ۲۷	۵۹/ ۴۴	۴۶/ ۵۶
NUS-TR@60	**۱۰/ ۶۱**	**۶۹/ ۵۲**	**۱۸/ ۴۲**	**۴۵/ ۵۳**	**۵۰/ ۱۵**	**۴۷/ ۷۳**

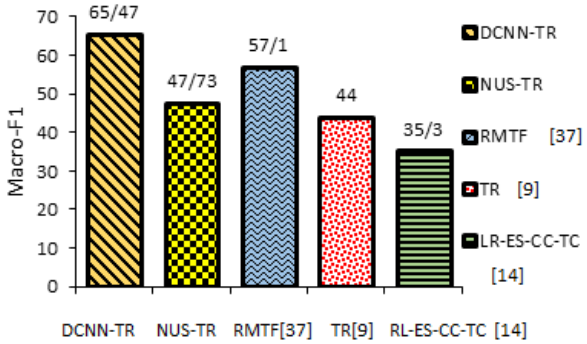
شکل۵ نمونه‌ای از تصاویر پالایش شده با استفاده از دو روش DCNN-TR و NUS-TR را نشان می‌دهد. در این تصویر تگ‌های جدید و مرتبط با محتوای افزوده شده به تصویر به رنگ آبی است و در زیر آن‌ها خط کشیده شده است. در صورت عدم پیش‌بینی تگ‌های مرتبط اولیه طی فرآیند پالایش، تگ‌ها به رنگ آبی و بر روی آنها خط کشیده شده است. تگ‌های قرمز رنگ، برچسب‌های غیر مرتبط با محتوای تصویر هستند که در صورت پالایش و حذف آنها از مجموعه تگ‌های تصویر، بر روی آنها خط کشیده شده است. همانطور که مشاهده می‌شود بسیاری از برچسب‌های غیر مرتبط با محتوای تصویر بعد از پالایش حذف شده‌اند و برچسب‌های مرتبط بعد از اعمال رویکرد پالایش تگ به تصاویر افزوده شده‌اند. عملکرد مناسب‌تر روش DCNN-TR نسبت به‌روش NUS-TR عدم وجود

			نوع برچسب
Canada, England, bird, country, Britain, hawk	explore, wildlife, animals, Africa, family, elephant, safari	sunset, England, colours, village, Britain, heritage, cottage, fire	برچسب‌های اولیه
hawk, bird, brown, birds, eagle, nature, clouds, sky, ,beautiful, England,Canada, ~~country~~, ~~Britain~~,	Wildlife, animals, family,nature, sky, blue, trees, water, clouds, explore, art, life, ~~safari~~, ~~elephant ,Africa~~	England,village, sunset ,tree, park, blue, grass mountains,woman, people,~~cottage~~, ~~Britain~~,~~colours ,heritage~~, ~~fire~~	برچسب‌های پالایش شده با روش NUS-TR
bird, hawk, birds, eagle , England, ~~Canada~~, ~~Britain~~,~~country~~	wildlife, animals, Africa, elephant, safari,nature, elephants, tusks, ~~family~~	sky, clouds, explore, sunset, England, village, cottage, Britain, silhouette, ~~colours~~, heritage	برچسب‌های پالایش شده با روش DCNN-TR
			نوع برچسب
sunset, beach, trees, reflection, winter, snow, lake, sand, silhouette, ice, path, Michigan	water, tree, birds, reflections, eyes, flight, earth, branch	reflection, river, house, Ireland, capital	برچسب‌های اولیه
sunset, trees, winter, nature, sky, water, landscape, tree, autumn, fall, blue, beautiful, photo, color, photograph, brown, colorful, ~~sand~~, ~~path~~, ~~silhouette~~, ~~Michigan~~	branch, water, reflection, tree, birds, earth, nature, sky, landscape, sea, building,beach, bravo, art, sun, snow, earth, winter, church, ~~eyes~~, ~~flight~~	reflection, river, house, sky, water, blue, clouds, sea, ocean, seascape, rock, sunlight, sunset, mountain, ~~Ireland~~, ~~capital,rocks~~, ~~island~~	برچسب‌های پالایش شده با روش NUS-TR
Sunset, beach, trees, winter, snow, ice,nature, sea, reflection, ~~sand~~, ~~path~~, ~~silhouette~~, ~~Michigan~~	water, tree, reflections, branch, reflection, lake, house, nature, flight, ~~eyes~~, ~~earth~~	city, reflection, river, house, sea, ocean, sky, seascape, night, ~~capital~~, ~~ireland~~	برچسب‌های پالایش شده با روش DCNN-TR

شکل ۴ نمونه ای از تصاویر پالایش شده توسط رویکرد پیشنهاد شده در این پژوهش

برچسب‌های غیر مرتبط با محتوای تصویر با بکارگیری این روش پالایش برچسب تصویر، محسوس‌تر می‌باشد. بررسی رویکردهای پالایش تگ ارائه شده در دیگر پژوهش‌ها نشان می‌دهد که بهبود دقت روش DCNN-TR ارائه شده در معیار ماکرو *F1* نسبت به NUS-TR و همچنین سایر روش‌های پالایش تگ مجموعه

تصاویر در مقیاس بزرگ NUS-WIDE بسیار چشم‌گیر است. مقایسه میان این رویکردها درشکل۵ نشان داده شده است.



شکل ۵نمودار مقایسه روش DCNN-TR ارائه شده با سایر روش‌های پالایش تگ در مجموعه تصاویر NUS-WIDE-۲۷۰k

در [۱۴] ژو و همکارانش نتایج فعالیت‌های پالایش تگ خود را بر روی مجموعه NUS-WIDE-270k با ۵۲۱ تگ پیش‌پردازش شده توسط ویکی‌پدیا ارائه داده‌اند. بهترین معیار ماکرو *F1* محاسبه شده در پژوهش آنها در ازای پیاده‌سازی رویکرد LR-TC-CC-ES[1] مقدار *F1*=۳/۳۵ درصد را به خود اختصاص می‌دهد. بکارگیری رویکرد رای‌دهی نزدیک‌ترین همسایه به‌عنوان روش پالایش تگ ارائه شده در این پژوهش، مشابه با پژوهش یوریچیو و همکارانش در[۹] می‌باشد، با این تفاوت که تعداد نمونه‌های مورد بررسی در این پژوهش ۲۳۸۲۵۱ تصویر با ۶۸۴ تگ تطبیق داده شده با پایگاه داده وردنت می‌باشد. همچنین در این پژوهش ضریب α=۵ در نظر گرفته شده است. این در حالی است که در پژوهش ما تعداد نمونه‌های مورد بررسی ۲۲۱۸۱۷ تصویر با ۱۰۰۰ تگ با بیشترین فرکانس بوده و ضریب α=۱۰ در نظر گرفته شده است. بهترین میزان معیارماکرو *F1* ارائه شده در پژوهش آنها حاصل از رویکرد TR[۸][2] برابر با ۴۴٪می‌باشد. در [۳۷] روشی تحت عنوان RMTF ارئه شده است که در آن بر اساس یک ساختار یکپارچه سه‌گانه از تصویر، تگ و کاربر به پالایش‌تگ‌های تصاویر می‌پردازد. سانگ[3] و همکارانش در این پژوهش با انتخاب ۱۲۴۰۹۹ تصویر از میان ۲۴۷۸۴۹ تصویر قابل دسترس و بارگذاری شده توسط ۵۰۱۲۰ کاربر از دادگان NUS-WIDE به رفع نویز از تگ‌های تصاویر می‌پردازند. معیار*F1*محاسبه شده توسط این رویکرد میزان ۱/۵۷٪ را به خود اختصاص می‌دهد و ضریب α نیز مشابه پژوهش ما ۱۰ در نظر گرفته شده است. در کلیه پژوهش‌های معرفی شده مقایسه نتایج پالایش تگ‌های تصاویر با نسخه ۸۱ تگی از دادگان NUS-WIDE-270k که توسط خبره‌ها شرح‌گذاری شده‌اند صورت گرفته است. همانطور که مشخص است معیار ماکرو *F1* محاسبه شده حاصل از رویکرد پالایش تگ DCNN-TR در این پژوهش با مقدار ۴۷/۶۵٪، با اختلاف بسیار زیادی بهتر از دیگر رویکردها می‌باشد. این در حالی است که درصورت عدم بهره‌مندی از ویژگی‌های استخراج شده از شبکه کانولوشن VGGNet معیار *F1*=۴۷/۳٪محاسبه شده است. علت برتری روش پیشنهادی بر دیگر روش‌ها، بهره‌مندی از فرآیند یادگیری انتقالی در شبکه‌های عصبی کانولوشنال عمیق به‌منظور استخراج ویژگی‌های کارآمد در محاسبه تشابه دیداری تصاویر و مجموعه مفاهیم سطح بالای پایگاه‌های تصویری برچسب‌دار، به ازای هر تصویر است.بهره‌گیری از این مجموعه از ویژگی‌های استخراج شده از شبکه DCNN با بکارگیری فرآیند یادگیری انتقالی و خوشه‌بندی تصاویر مشابه موجب افزایش دقت محاسبات و کاهش قابل توجه زمان و حجم عملیات مورد نیاز برروی مجموعه دادهای با مقیاس بزرگ‌شده است.

۶ نتیجه‌گیری و پیشنهاد برای پژوهش‌های آتی

در این پژوهش به منظور حل چالش پایین بودن دقت فرآیند بازیابی تصاویر در موتورهای جستجو، به ارائه رویکردی در دسته الگوریتم‌های مبتنی بر نمونه در پالایش تگ‌های تصاویر پرداختیم. در این رویکرد به‌منظور استخراج ویژگی‌هایی با توصیف دیداری مناسب از تصویر، از فرآیند یادگیری انتقالی به‌کار گرفته شده در شبکه عصبی کانولوشنال عمیق VGGNet استفاده شده‌است. به‌همین منظور ابتدا کلیه تصاویر به شبکه کانولوشنال VGGNet، که از قبل با تصاویر ImageNetیادگیری شده‌اند، به عنوان ورودی به شبکه داده شد واز شبکه، بردارهایی با ۴۰۹۶ ویژگی و ۱۰۰۰ برچسب به‌ازای هر تصویر استخراج شد. به‌منظور مقیاس‌پذیری سیستم پالایش تگ‌های تصاویر، ابتدا با استفاده از ویژگی‌های دیداری استخراج شده توسط VGGNet خوشه‌بندی تصاویر صورت گرفت، سپس در هر خوشه با استفاده از تشابه معنایی میان تصاویر و محاسبه هم‌پوشانی میان تگ‌ها، تصاویر همسایه انتخاب شد. در نهایت با بهره‌گیری از تکنیک رای‌گیری از تصاویر همسایه، پالایش تگ‌های تصاویر صورت گرفت. نتایج حاصل، بیانگر موثر بودن فرآیند یادگیری انتقالی بکارگرفته شده در CNN در حوزه پالایش تگ‌های تصاویر می‌باشد. این روش دارای مزیت عمده‌ای چون سادگی و مقیاس‌پذیری آن است. بر همین اساس در پالایش تگ‌های تصاویر در پایگاه‌های داده‌های با مقیاس بزرگ استفاده از رویکرد ارائه شده، توصیه می‌گردد. در پژوهش‌های آتی سعی داریم که به بررسی تاثیر رویکرد پالایش تگ ارائه شده بر فرآیند شرح‌گذاری خودکار تصاویر بپردازیم. همچنین از آنجا که پایگاه‌های داده با مقیاس بزرگ همواره از مهمترین چالش‌های موجود در این حوزه می‌باشد، لذا سعی داریم تا با بکارگیری تکنیک

[1] Tag Correlation-Content Consistency-Error Sparcity-Low Rank(LR-ES-CC-TC)

[2]Learning Tag Relevance from Visual Neighbors (TR)

[3]Sung

نگاشت/کاهش[1] به کنترل حجم دادگان و بهبود این چالش از این منظر نیز بپردازیم. همچنین تصاویر با ابعاد بزرگ نیز چالش عمده دیگری است که در حوزه یادگیری عمیق به آن پرداخته می‌شود. ارائه روشی سازگار برای تصاویر با ابعاد بزرگ در پایگاه‌های تصویری با مقیاس بزرگ استراتژی آینده پژوهشی ما در این حوزه می‌باشد.

مراجع

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, p. 5, 2008.

[2] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, "Image annotation by large-scale content-based image retrieval," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 607–610.

[3] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu, "Bipartite graph reinforcement model for web image annotation," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 585–594.

[4] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[5] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. Del Bimbo, "Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval," *ACM Comput. Surv.*, vol. 49, no. 1, p. 14, 2016.

[6] J. Donahue *et al.*, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.," in *ICML*, 2014, pp. 647–655.

[7] S. Lee, W. De Neve, and Y. M. Ro, "Visually weighted neighbor voting for image tag relevance learning," *Multimed. Tools Appl.*, vol. 72, no. 2, pp. 1363–1386, 2014.

[8] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1310–1322, 2009.

[9] T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo, "An evaluation of nearest-neighbor methods for tag refinement," in *Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[10] G. Y. Bobhate and U. A. Jogalekar, "An Efficient Algorithm to Reduce the Semantic Gap between Image Contents and Tags," *Int. J. Comput. Appl.*, vol. 72, no. 11, pp. 38–44, 2013.

[11] L. Chen, D. Xu, I. W. Tsang, and J. Luo, "Tag-based image retrieval improved by augmented features and group-based refinement," *IEEE Trans. Multimed.*, vol. 14, no. 4, pp. 1057–1067, 2012.

[12] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proceedings of the international conference on Multimedia*, 2010, pp. 491–500.

[13] Z. Lin, G. Ding, M. Hu, Y. Lin, and S. S. Ge, "Image tag completion via dual-view linear sparse reconstructions," *Comput. Vis. Image Underst.*, vol. 124, pp. 42–60, 2014.

[14] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," *Proc. Int. Conf. Multimed. - MM '10*, p. 461, 2010.

[15] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, 2013.

[16] J. Fu, J. Wang, Y. Rui, X. Wang, T. Mei, and H.-H. Lu, "Image Tag Refinement with View-Dependent Concept Representations," 2014.

[17] J. Fu, J. Wang, Y. Rui, X.-J. Wang, T. Mei, and H. Lu, "Image Tag Refinement With View-Dependent Concept Representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1409–1422, 2015.

[18] Y. Bengio, "Learning deep architectures for AI," *Found. trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[19] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Networks,* vol. 8, no. 1, pp. 98–113, 1997.

[20] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[21] T. Mikolov, M. Karafiát, L. Burget, J. Cernock\`y, and S. Khudanpur, "Recurrent neural network based language model.," in *Interspeech*, 2010, vol. 2, p. 3.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR),* 2014, pp. 580–587.

[23] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Process. Mag. IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

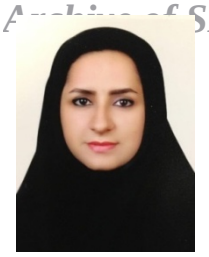
[24] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference onComputer Vision and Pattern Recognition,* 2009, pp. 248–255.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid

[1] Map/Reduce



شیما جوانمردی کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۹۰از دانشگاه دولتی جهرم دریافت کردند. سپس کارشناسی ارشد خود را در همان رشته و گرایش در سال ۱۳۹۵ از دانشگاه یزد اخذ نمودند. در حال حاضر ایشان دانشجوی دکتری در گرایش هوش مصنوعی در دانشگاه یزد هستند. حوزه‌های پژوهشی ایشان شامل یادگیری ماشین، بینایی ماشین (شرح‌گذاری خودکار تصاویر، پالایش شرح‌گذاری تصاویر و بازیابی تصاویر) و یادگیری عمیق است.

محمدعلی زارع چاهوکی کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۷۸ از دانشگاه شهید بهشتی دریافت کردند. سپس کارشناسی ارشد و دکتری خود را در همان رشته و گرایش در سال های ۱۳۸۳ و ۱۳۹۲ از دانشگاه تربیت مدرس اخذ نمودند. در حال حاضر ایشان استادیار گروه مهندسی کامپیوتر دانشگاه یزد هستند. حوزه‌های پژوهشی ایشان شامل یادگیری ماشین، عقیده‌کاوی، مهندسی نرم‌افزار (متدولوژی و کاربردهای مرتبط با یادگیری)، بینایی ماشین (شرح‌گذاری خودکار تصاویر و بازیابی تصاویر)، و نهان‌نگاری است.

pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, 2014, pp. 346–361.

[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[29] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv Prepr. arXiv1312.6229*, 2013.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] V. V. Gomez, A. S. Cortes, and F. M. Noguer, "Object Detection for Autonomous Driving Using Deep Learning," 2015.

[32] S. Zhu, S. Aloufi, and A. El Saddik, "Utilizing image social clues for automated image tagging," *IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.

[33] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, p. 48.

[34] H. K. Shooroki and M. A. Z. Chahooki, "Selection of effective training instances for scalable automatic image annotation," *Multimed. Tools Appl.*, pp. 1–24, 2016.

[35] H. Kargar-Shooroki, M. A. Z. Chahooki, and S. Javanmardi, "MLENN-KELM: a Prototype Selection Based Kernel Extreme Learning Machine Approach for Large-Scale Automatic Image Annotation," *Adv. Comput. Sci. an Int. J.*, vol. 4, no. 5, pp. 95–100, 2015.

[36] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimed.*, vol. 14, no. 4, pp. 1021–1030, 2012.

[37] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. Multimed.*, vol. 14, no. 3, pp. 883–895, 2012.