برچسب‌زنی مقیاس‌پذیر تصاویر با خلاصه‌سازی نمونه‌ها به نماینده‌های برچسب‌دار

محیا محمدی کاشانی و سید حمید امیری*

دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران

چکیده

با افزایش روزافزون تصاویر، اندیس‌گذاری و جستجوی سریع آن‌ها در پایگاه داده‌های بزرگ، یک امر ضروری است. یکی از راه‌کارهای مؤثر، نسبت‌دادن یک یا چند برچسب به هر تصویر با هدف توصیف محتوای درون آن است. با وجود کارایی روش‌های خودکار برچسب‌زنی، یکی از چالش‌های اساسی آنها مقیاس‌پذیری با افزایش تصاویر پایگاه داده است. در این مقاله، با هدف حل این چالش، ابتدا براساس توصیف‌گر بصری تصاویر که از شبکه‌های یادگیری عمیق استخراج می‌شوند، نمایندگان مناسبی به‌دست می‌آیند. سپس، با استفاده از رویه انتشار برچسب بر روی گراف، برچسب‌های معنایی از تصاویر آموزشی به نمایندگان منتشر می‌شوند. با این راه‌کار، به یک مجموعه نمایندگان برچسب‌دار دست خواهیم یافت که می‌توان عمل برچسب‌زنی هر تصویر آزمون را بر اساس این نمایندگان انجام داد. برای برچسب‌زنی، یک رویکرد مبتنی بر آستانه‌گذاری وفقی پیشنهاد شده است. با روش پیشنهادی، می‌توان اندازه مجموعه‌داده آموزشی را به ۲۲/۶ درصد اندازه اولیه کاهش داد که منجر به تسریع حداقل ۴/۲ برابری زمان برچسب‌زنی خواهد شد. همچنین، کارایی برچسب‌زنی بر روی مجموعه‌داده‌های مختلف برحسب سه معیار دقت، یادآوری و F1 در حد مطلوبی حفظ شده است.

واژگان کلیدی: خلاصه‌سازی پایگاه داده، برچسب‌زنی تصویر، روش مبتنی برجستجو، مقیاس‌پذیری

Scalable Image Annotation by Summarizing Training Samples into Labeled Prototypes

Mahya Mohammadi Kashani & S. Hamid Amiri*

Computer Engineering Department, Shahid Rajaee Teacher Training University, Tehran, Iran

Abstract

By increasing the number of images, it is essential to provide fast search methods and intelligent filtering of images. To handle images in large datasets, some relevant tags are assigned to each image to for describing its content. Automatic Image Annotation (AIA) aims to automatically assign a group of keywords to an image based on visual content of the image. AIA frameworks have two main stages; Feature Extraction and Tag Assignment which are both important in order to reach a proper performance. In the first stage of our proposed method, we utilize deep models to obtain a visual representation of images. We apply different pre-trained architectures of Convolutional Neural Networks (CNN) to the input image including Vgg16, Dense169, and ResNet 101. After passing the image through the layers of CNN, we obtain a single feature vector from the layer before the last layer, resulting into a rich representation for the visual content of the image. One advantage of deep feature extractor is that it substitutes a single feature vector instead of multiple feature vectors and thus, there is no need for combining multiple features. In the second stage, some tags are assigned from training images to a test image which is called "*Tag Assignment*". Our approach for image annotation belongs to the search-based methods which have high performance in spite of simple structure. Although it is even more time-consuming due to its method of comparing the test image to every training in order to find similar images. Despite the efficiency of automatic Image annotation methods, it is challenging to provide a scalable method for large-scale datasets. In this paper, to solve this challenge, we propose a novel approach to summarize training database (images and their relevant tags) into a small number of

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال مقاله: ۱۳۹۸/۰۴/۲۳ ● تاریخ پذیرش: ۱۳۹۹/۰۵/۲۸ ● تاریخ انتشار: ۱۴۰۰/۱۲/۲۹ ● نوع مطالعه: کاربردی

[DOI: 10.52547/jsdp.18.4.49]

prototypes. To this end, we apply a clustering algorithm on the visual descriptors of training images to extract the visual part of prototypes. Since the number of clusters is much smaller than the number of images, a good level of summarization will be achieved using our approach. In the next step, we extract the labels of prototypes based on the labels of input images in the dataset. because of this, semantic labels are propagated from training images to the prototypes using a label propagation process on a graph. In this graph, there is one node for each input image and one node for each prototypes. This means that we have a graph with union of input images and prototypes. Then, to extract the edges of graph, the visual feature of each node on graph is coded using other nodes to obtain its *K*-nearest neighbors. This goal is achieved by using Locality-constraints Linear Coding algorithm. After construction the above graph, a label propagation algorithm is applied on the graph to extract the labels of prototypes. Based on this approach, we achieve a set of labeled prototypes which can be used for annotating every test image. To assign tags for an input image, we propose an adaptive thresholding method that finds the labels of a new image using a linear interpolation from the labels of learned prototypes. The proposed method can reduce the size of a training dataset to 22.6% of its original size. This issue will considerably reduce the annotation time such that, compared to the state-of-the-art search-based methods such as 2PKNN, the proposed method is at least 4.2 times faster than 2PKNN, while the performance of annotation process in terms of Precision, Recall and F1 will be maintained on different datasets.

Keywords: Database Summarization, Image Annotation, Search-Based method, Scalability

۱- مقدمه

جستجوی سریع و برخط، اندیس‌گذاری و بازیابی تصاویر به‌دلیل افزایش روزافزون تعداد تصاویر و ویدئو در مجموعه‌های شخصی و پایگاه‌های اینترنتی بسیار حائز اهمیت است. گسترش شبکه‌های اجتماعی این نیازمندی را جدی‌تر کرده است. برای مثال در روز ۳۵۰ میلیون تصویر در فیسبوک قرار می‌گیرد [1]. برای جستجو، بازیابی یا اندیس‌گذاری سریع و کارا، از رویکرد نسبت‌دادن برچسب‌ها به تصویر استفاده می‌شود. به فرآیند نسبت‌دادن خودکار مجموعه‌ای از واژگان کلیدی به یک تصویر جهت توصیف معنایی محتوای یک تصویر، برچسب‌زنی خودکار گویند. با توجه به حجم بالای مجموعه‌تصاویر و جستجو و بازیابی برخط، مقیاس‌پذیربودن سامانه برچسب‌زنی اهمیت بسیار بالایی دارد.

به‌طورعمومی در هر سامانهٔ برچسب‌زنی دو مرحله کلی مورد بررسی قرار می‌گیرد، استخراج ویژگی از تصاویر که به دنبال یک توصیف‌گر برای محتوای تصاویر است و در مرحله دوم، هدف ساخت یک مدل از تصاویر واقع در یک پایگاه داده آموزشی است که بتواند ارتباط بین محتوای تصاویر و کلمات نسبت داده شده به آن‌ها را استخراج کند. به‌طورمعمول پایگاه داده‌های برچسب‌زنی برای توصیف محتوای تصاویر از صدها کلمهٔ متمایز استفاده می‌کنند [6-2].

روش‌های متعددی جهت حل مسأله برچسب‌زنی وجود دارد. در سال‌های اخیر، روش‌های مبتنی بر جستجو تا حد زیادی در مسأله برچسب‌زنی تصاویر مطرح شده‌اند و با وجود معماری ساده آن‌ها، کارایی خوبی از خود نشان داده‌اند. این روش‌ها بر پایه این اصل بنا نهاده شده‌اند؛ تصاویری که از نظر محتوای بصری به یکدیگر شبیه هستند از کلمات مشترکی در برچسب‌گذاری استفاده می‌کنند [7].

به‌طورکلی روش‌های مبتنی برجستجو با وجود سادگی، کارایی بالایی از خود نشان داده‌اند، اما برای یافتن شباهت میان تصویر ورودی آزمون و تصاویر موجود در مجموعه آموزشی، ناگزیر به جستجوی تمام تصاویر آموزشی هستیم. به عبارتی دیگر، این گونه روش‌ها مقیاس‌پذیر نیستند؛ که سبب افزایش قابل توجه زمان اجرا برچسب‌زنی برای یک تصویر جدید می‌شوند. در این مقاله روش خلاصه‌سازی پایگاه داده به نماینده به همراه برچسب‌های فازی موثر، جهت حل مشکل پیش‌آمده ارائه شده است. به عبارت دقیق‌تر، در فاز آموزشی روش پیشنهادی، تصاویر با محتوای بصری نزدیک به یکدیگر، با یک نماینده جایگزین، سپس برچسب‌های این تصاویر به نماینده منتشر می‌شوند. درنهایت، یک مجموعه برچسب با مقادیر فازی برای نماینده به‌دست می‌آید؛ بنابراین، پس از انتشار برچسب، علاوه بر حضور تمامی برچسب‌های تصاویر در مجموعه برچسب‌های نماینده، یک درجه تعلق فازی برای هر برچسب نیز به‌دست می‌آید که بیان‌گر میزان اهمیت برچسب است؛ سپس، در فرآیند برچسب‌زنی یک تصویر جدید، از برچسب‌های فازی نمایندگان استفاده خواهد شد که یکی از دلایل موفقیت روش پیشنهادی است. به عبارت دیگر، فقط حضور یا عدم حضور برچسب مورد بررسی قرار نمی‌گیرد؛ بلکه درجه تعلق برچسب نیز مد نظر قرار خواهد گرفت.

[DOI: 10.52547/jsdp.18.4.49]

از طرف دیگر، در این مقاله از شبکه‌های یادگیری عمیق به همراه قطعه‌بندی تصویر طبق (شکل-۳)، جهت استخراج ویژگی استفاده شده است که قادر است محتوای بصری تصویر را با دقت بالایی در یک بردار ویژگی نمایش دهد.

روش پیشنهادی دارای دو فاز اصلی است. در فاز نخست، خلاصه‌سازی تصاویر پایگاه داده به نمایندگان صورت می‌گیرد به‌طوری‌که تعداد نمایندگان به‌طور قابل توجهی کمتر از تعداد تصاویر آموزشی است. پس از استخراج نمایندگان، یک گراف شامل تصاویر آموزشی و نمایندگان شکل می‌گیرد که در این گراف، یال‌های وزن‌دار بر اساس رویکرد ضرایب تُنُک محلی [8] تعیین می‌شوند؛ سپس، در طی یک فرآیند آموزش بر روی گراف، برچسب‌های معنایی تصاویر آموزشی به نمایندگان انتشار می‌یابند که در شکل (۴) نشان شده است. نتیجه این فرآیند، نمایندگانی است که دارای برچسب هستند. در نتیجه، در فاز دوم الگوریتم پیشنهادی، برچسب‌های یک تصویر ورودی جدید با استفاده از نمایندگان برچسب‌دار تعیین می‌شوند. در این فاز، برچسب‌ها با یک روش نوین مبتنی بر آستانه‌گذاری نسبت داده می‌شوند که در شکل (۵) نشان داده شده است.

ادامه مقاله بدین صورت تدوین شده است. در بخش دوم مروری بر کارهای مرتبط صورت می‌گیرد. در بخش سوم، روش پیشنهادی جهت خلاصه‌سازی پایگاه داده به نماینده و سپس روش پیشنهادی سامانه برچسب‌زنی ارائه شده است. در بخش چهارم، نتایج حاصل از آزمایش‌ها آمده است و درآخر به بحث و نتیجه‌گیری بر روی روش ارائه شده خواهیم پرداخت.

۲- کارهای پیشین

همان‌طور که در بخش پیشین گفته شد، برچسب‌زنی تصویر به‌طورعمومی دو مرحله کلی دارد؛ مرحله نخست استخراج ویژگی و مرحله دوم فرایند برچسب‌زنی از پایگاه داده به تصویر ورودی است. در رویکرد برچسب‌زنی، استخراج ویژگی غنی در جهت افزایش کارایی سامانه نقش بالایی دارد. طبق [9,10]، استفاده از یک بردار ویژگی استخراج‌شده از مدل‌های عمیق، کارایی بالاتری از چندین نوع بردار ویژگی استاندارد در حوزه برچسب‌زنی تصاویر دارد. اغلب روش‌های پیشین یک مدل برچسب‌زنی جهت افزایش کارایی ارائه داده‌اند.

در سال‌های اخیر، روش‌های برچسب‌زنی متعددی با دیدگاه مختلفی پیشنهاد شده‌اند. برای مثال، رویکردهای ورارسانی[1] در مقابل روش‌های استقرایی[2] و رویکردهای مبتنی بر آموزش[3] در مقابل رویکردهای مبتنی بر بازیابی[4][11] . طبق [12]، اغلب روش‌های برچسب‌زنی به پنج گروه تقسیم می‌شوند. مدل‌های مولد [2,6,13,14]، که تصاویر را در موضوعات مختلف گروه‌بندی کرده و یک توزیع احتمالاتی برای هر موضوع در نظر می‌گیرد. در این روش‌ها، تضمینی وجود ندارد که توزیع احتمالاتی مربوط به فضای ویژگی به‌درستی مدل شود. گروه دوم، مدل‌های تمایزی [15-19] هستند که مسأله برچسب‌زنی را مانند مسئله چندرده درنظر می‌گیرند و برای هر رده یک طبقه‌بند دودویی آموزش داده می‌شود. مشکل عمده این گروه، درنظرنگرفتن وابستگی میان برچسب‌ها و هم‌رخدادیشان است. گروه سوم روش‌های برچسب‌زنی مدل‌های مبتنی بر برچسب [20-25] است که تعدادی برچسب از تصویر آموزشی را نادرست فرض می‌کند؛ سپس، این مدل‌ها تعداد برچسب جدید به هر تصویر آموزشی نسبت داده و برچسب‌های نویزی را اصلاح می‌کنند.

گروه چهارم، روش‌های مبتنی بر جستجو [8,26-30] هستند، که بر این اصل بنا نهاده شده‌اند که تصاویر مشابه به‌احتمال دارای برچسب‌های مشترکی هستند. این روش‌ها برای یک تصویر ورودی، مشابه‌ترین تصاویر در مجموعه‌داده را بازیابی کرده و سپس برخی برچسب‌های تصاویر بازیابی‌شده را به تصویر ورودی نسبت می‌دهد. دو روش مرز دانش در این گروه 2PKNN [29] و انتشار برچسب (Tag-Prop) [30] هستند که چالش عدم توازن رده (به عبارتی، واریانس بالا در فراوانی برچسب‌های مختلف) و برچسب‌زنی ناقص (به عبارتی نبود برچسب‌های مرتبط در فهرست برچسب یک تصویر) در دادهٔ واقعی را حل کرده‌اند.

پنجمین گروه که در دو جنبهٔ مسأله برچسب‌زنی، شامل استخراج ویژگی و نسبت‌دهی برچسب وجود دارند، به عبارتی مسأله برچسب‌زنی و شبکه‌های عصبی کانولوشنی عمیق ترکیب شده است. در این گروه مبتنی بر ساختار شبکه‌ها، ویژگی‌های محلی [31-33] یا سراسری

[1] Transduction-based
[2] Induction-based
[3] Learning-based
[4] Retrieval-based

[DOI: 10.52547/jsdp.18.4.49]

[34] را استخراج کرده‌اند. این شبکه‌ها علاوه‌بر استخراج ویژگی غنی، با بهینه کردن تابع هزینه، برچسب‌های مناسبی به یک تصویر ورودی نسبت می‌دهد. برای مثال دودویی [35-37]، رتبه‌بندی [38-40] و غیره. طبق دسته بندی‌های ذکر شده، روش‌های مبتنی بر جستجو در کاربری‌های دنیای واقعی به دلیل صحت بالا و سادگیشان استفاده می‌شوند. گرچه، برچسب‌زنی یک تصویر جدید به وسیله این روش‌ها نیازمند مقایسه تصویر ورودی با تمام تصاویر در مجموعه آموزشی و انتخاب مشابه‌ترین تصاویر هستند؛ ازین رو برای مجموعه دادگان بزرگ به شدت زمان‌بر هستند. روش‌های مختلفی برای مقایس‌پذیری ارائه شده‌اند که چالش عدم توازن کلاس (به عبارتی، واریانس بالا در فراوانی برچسب‌های مختلف) و برچسب‌زنی ناقص (به عبارتی عدم وجود برچسب‌های مرتبط در فهرست برچسب یک تصویر) در داده‌ی واقعی را حل کرده‌اند.

روش‌های مختلفی برای مقیاس‌پذیری ارائه شده‌اند می‌توان آنها را در سه دسته کلی گروه‌بندی کرد: روش‌های مبتنی بر نماینده، روش‌های مبتنی بر کاهش بعد[1]، روش‌های مبتنی بر تبدیل[2]. روش‌های مبتنی بر نماینده[3]، نمونه‌ها را خوشه‌بندی می‌کنند و سپس یک یا چند نمونه از هر خوشه را انتخاب می‌نمایند. [41]، روش‌های مبتنی بر نماینده را در سه گروه دسته‌بندی می‌کنند: ۱) روش‌های چگالش[4] که تعداد نمایندگان را بدون تعمیم‌پذیری بهینه می‌کنند، ۲) روش‌هایی که نمایندگان نویزی را جهت حفظ تعمیم‌پذیری از بین می‌برند، ۳) روش‌های ترکیبی که تعداد نمایندگان را در عین تعمیم‌پذیری بهینه می‌کنند. یک رویکرد بسیار محبوب برای استخراج نمایندگان استفاده از روش خوشه‌بندی K-means [42] است. یکی از رویکردهای مهم جهت مقیاس‌پذیر کردن سامانه‌های برچسب‌زنی، روش‌های مبتنی بر کاهش بُعد است. در این رویکرد، یک ماشین کدگذاری به هر تصویر یک کد کوتاه نسبت می‌دهد. سپس، شباهت میان دو تصویر طبق شباهت کدهای درهم‌سازیشان اندازه‌گیری می‌شود. درنهایت، فرایند جستجو بر روی شباهت‌های میان کدهای تصاویر انجام می‌شود.

گروه سوم از روش‌های مقیاس‌پذیر، روش‌های مبتنی بر تبدیل هستند [27-29,42-44]، که تمام داده را به فضایی دیگر با قابلیت تفکیک‌پذیری بالاتر انتقال می‌دهند. یکی از رویکردهای محبوب، بازنمایش تُنُک است. موضوع فضای تُنُک به طور کلی دارای دو مسأله است: بازنمایش تُنُک [45-47] و کدگذاری تُنُک. بازنمایش تُنُک حجم وسیعی از نمونه‌های ورودی را به فضای تُنُک انتقال می‌دهد. این انتقال شامل آموزش ماتریس کتابچه (که یک بازنمایش جدید از فضای ویژگی است) و آموزش ماتریس ضرایب (که ترکیب خطی دودویی از تعداد کمی بردارهای کتابچه را نشان می‌دهد) خواهد بود. کدگذاری تُنُک تنها شامل بخش دوم بازنمایش تُنُک است؛ به عبارت دیگر، کتابچه ثابت است و تنها استخراج ماتریس ضرایب انجام می‌شود. در روش [48]، با استفاده از کدگذاری تُنُک، اتصال میان نمونه‌ها و نمایندگان بر روی یک گراف یادگیری، تعیین می‌شود. سپس، با استفاده از کدگذاری تُنُک، تِم‌ها استخراج می‌شود که در آن، تم به زیر مجموعه‌ای از نمونه‌های آموزشی اطلاق می‌شود محتوای بصری و معناییشان دارای اشتراک هستند. در روش [49]، یک کتابچه و ضرایب تُنُکش آموزش می‌بینند. هم‌چنین، ارتباط میان ضرایب تُنُک و برچسب‌ها را بررسی می‌کند. نقطه ضعف این روش حجم زیاد داده و تعداد پارامترهایی است که باید در هر تکرار محاسبه شودسبب افزایش پیچیدگی زمان اجرا می‌شود.

به طور کلی روش‌های مبتنی برجستجو با وجود سادگی، کارایی بالایی از خود نشان داده‌اند اما برای یافتن شباهت میان تصویر ورودی آزمون و تصاویر موجود در مجموعه آموزشی، روش‌های مبتنی برجستجو ناگزیر به جستجوی تمام تصاویر آموزشی است. به عبارتی دیگر، این گونه روش‌ها، مقیاس‌پذیر نیستند. روش خلاصه‌سازی پایگاه داده به نماینده جهت مقیاس‌پذیر کردن روش‌های مبتنی بر رویکرد جستجو ارائه شده است؛ که در عین حفظ یا بهبود قابل قبول کارایی، زمان اجرای برچسب‌زنی در زمان آزمون را کاهش می‌دهد؛ در ادامه این موضوع به تشریح آورده شده است.

۳- روش پیشنهادی خلاصه‌سازی پایگاه داده

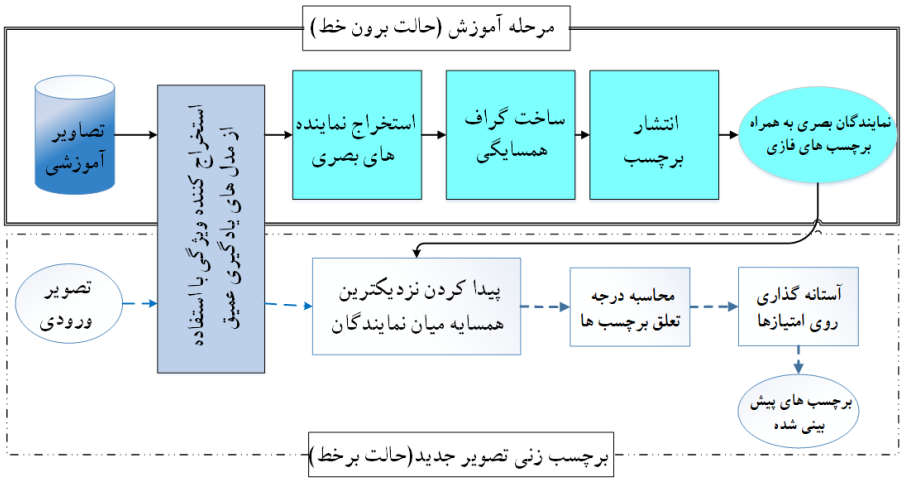
روند کلی الگوریتم پیشنهادی مقاله در (شکل-۱) قابل مشاهده است. الگوریتم پیشنهادی از دو بخش اصلی تشکیل شده است: ۱) مرحله آموزش به‌عنوان مرحله پیش پردازش، ۲) مرحله برچسب‌زنی تصویر ورودی. در ابتدا ویژگی‌های تصاویر آموزشی با استفاده از مدل‌های

[1] Dimension-reduction base
[2] Transform-based
[3] Prototype-based
[4] Condensation methods

[DOI: 10.52547/jsdp.18.4.49]

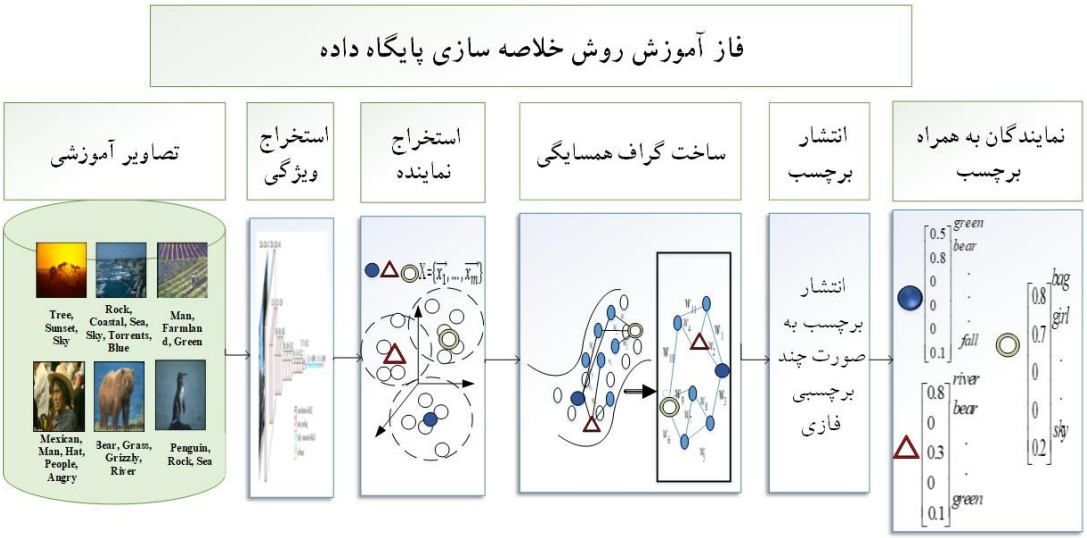
یادگیری عمیق استخراج شده، سپس نمایندگان بصری مناسب را از پایگاه داده آموزشی به‌دست می‌آوریم. سپس، یک گراف برای تصاویر آموزشی و نمایندگان ساخته می‌شود. بر روی این گراف، رویه آموزش جهت انتشار برچسب از تصاویر آموزشی به نمایندگان انجام می‌شود. درنهایت، در فاز آموزش، نمایندگان بصری به همراه بردارهای برچسب آنها ذخیره می‌شوند تا در مرحله برچسب‌زنی یک تصویر جدید، مورد استفاده قرار گیرند. برای این منظور، نمایندگان مرتبط با تصویر ورودی تعیین می‌شوند و بر اساس برچسب‌های فازی نمایندگان، برچسب‌های تصویر ورودی استخراج می‌شوند.

شمای کلی روش پیشنهادی خلاصه‌سازی پایگاه داده در شکل (۲) آمده است که در بخش‌های بعدی به بررسی کامل روش خواهیم پرداخت. ابتدا، در بخش (۳-۱) به نحوهٔ استخراج ویژگی از تصاویر پرداخته شده است. سپس روش پیشنهادی جهت خلاصه‌سازی ارائه شده است. در این قسمت به دو موضوع مهم پرداخته می‌شود: نحوهٔ استخراج نماینده و برچسب‌زنی نمایندگان. مجموعه‌دادهٔ نمایندگان به همراه برچسب‌های جدید جایگزین پایگاه داده اصلی می‌شوند. حال با استفاده از نماینده‌ها، می‌توان برای تصاویر آزمون ورودی، برچسب‌زنی را انجام داد.



(شکل-۱): روندنمای روش پیشنهادی

(Figure-1): Flowdiagram of the proposed method



(شکل-۲): فاز خلاصه‌سازی تصاویر آموزشی با روش پیشنهادی

(Figure-2): Summarization phase of training images using the proposed method

فرض کنید که مجموعه‌داده‌ی آموزشی به فرم $X = \{x_1, x_2, \dots, x_N\}$ را داریم؛ که N تعد تصاویر آموزشی در فضای $\mathbb{R}^d$ است و d بیانگر ابعاد بردارهای ویژگی داده‌های آموزشی است. برچسب‌های هر تصویر، زیر مجموعه‌ای از مجموعه واژگان $L = \{l_1, l_2, \dots, l_L\}$ هستند.

برای x_i که مناظر با تصویر i-ام آموزشی است، برچسب‌های نسبت داده شده را به‌صورت τ_i (τ_iL) تعری می‌کنیم.

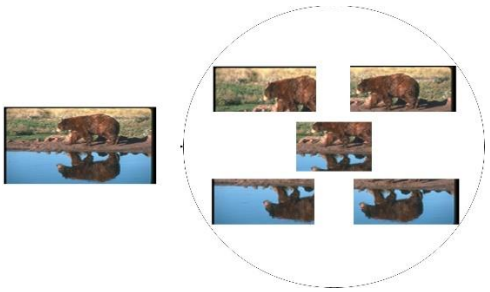
از طرفی $X_P = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_M\}$ بیانگر نمایندگان مستخرج از مجموعهٔ پایگاه داده آموزشی در فضای $\mathbb{R}^d$

[DOI: 10.52547/jsdp.18.4.49]

است که M تعداد نمایندگان مستخرج از تصاویر آموزشی است به طوری‌که $M \ll N$. هم‌چنین، K پارامتر نزدیک‌ترین همسایگی جهت ساخت گراف وزن‌دار مبتنی بر بازسازی تُنُک محلی جهت خلاصه‌سازی پایگاه داده آموزشی است.

۳-۱- استخراج ویژگی

برای کاربردی‌کردن سامانه پیشنهادی در دنیای واقعی ویژگی‌های تصویر را از مدل‌های یادگیری عمیق استخراج کرده‌ایم. از آنجا که سه معماری مطرح یادگیری عمیق VGGNet [51]، ResNet [52] و Dense-Net [53] در کارهای مربوط به بازشناسی اشیا بالاترین کارایی را برای مجموعه دادگان بزرگی چون Image Net[1] داشته‌اند، ما نیز از چنین معماری‌هایی استفاده نموده‌ایم. تمامی ویژگی‌های به‌دست‌آمده از لایهٔ یکی مانده به آخر مدل‌های مختلف استخراج شده، که این ویژگی‌ها دارای اطلاعات غنی از اشیای آموزش دیده شده هستند، در عین حال که ابعاد آن بیش از اندازه بزرگ نیست؛ می‌توان جهت سهولت کار، ابعاد هر ویژگی به‌دست‌آمده را با رویکرد آنالیز مؤلفه‌های اصلی[2] کاهش داد.نکته حائز اهمیت در این بخش، نحوهٔ قطعه‌بندی تصاویر و ارسال آن به شبکه جهت استخراج ویژگی است؛ از آن‌جایی‌که مسألهٔ برچسب‌زنی تصاویر اساساً رویکردی چند برچسبی است، استخراج ویژگی از نواحی مختلف تصویر به طور جداگانه دارای اهمیت است. درنهایت، میانگین‌گیری بر روی ویژگی‌های به‌دست‌آمده از تمامی نواحی صورت می‌گیرد تا بردار ویژگی تصویر به‌دست آید. در (شکل-۳)، بهترین شمای کلی، از میان شماهای دیگر، جهت قطعه‌بندی تصویر برای دادن به مدل‌های عمیق که بهترین کارایی را در سامانه‌های برچسب‌زنی دارد؛ ترسیم شده است.



(شکل-۳): بهترین شمای کلی قطعه‌بندی یک تصویر ورودی جهت استخراج ویژگی با استفاده از مدل‌های عمیق

(Figure-3): Best scheme for segmenting the image in the feature extraction step based on deep models

[1] http://www.image-net.org

[2] Principle Component Analysis (PCA)

۳-۲- خلاصه‌سازی پایگاه داده به نماینده

در این بخش به نحوهٔ استخراج نمایندگان می‌پردازیم. برای جایگزینی نمایندگان به جای کل پایگاه داده، انتخاب تعداد کم اما مناسب نمایندگان، نیازمند بررسی روش‌های متعددی هستیم. یکی از روش‌های رایج، روش اول، برای به‌دست‌آوردن نماینده از مجموعه‌داده‌ها، خوشه‌بندی است به طوری‌که مجموعه‌داده‌های مشابه، از نظر معیار فاصله، در یک خوشه قرار گیرند. هر خوشه دارای مراکزی است که می‌تواند بیانگر نماینده‌ای از آن خوشه باشد. به عبارت دیگر، مرکز خوشه، یک داده‌ی خلاصه شده و مفید از داده‌های درون خوشه است. در ادامه، روش دوم جهت به‌دست‌آوردن نماینده مطرح شده است. مراکز خوشه‌های نهایی، در واقع همان نمایندگان پایگاه داده آموزشی هستند.

جهت خوشه‌بندی در مجموعه پایگاه داده‌های بزرگ از خوشه‌بندی سلسله مراتبی شرطی استفاده می‌کنیم. در ابتدا همانند روش اول، کل مجموعه‌داده با استفاده از خوشه‌بندی K-means به Q بخش تقسیم می‌گردد؛ به طوری‌که Q<M و M تعداد نمایندگان است. سپس، هر خوشه از نظر گستردگی[3] بررسی می‌شود. طبق ضابطهٔ (۱) اگر شرایط برقرار باشد، یک مرحلهٔ دیگر عملیات خوشه‌بندی برای آن انجام می‌شود. در غیر این صورت، عملیات تقسیم‌بندی خوشه‌ی مدنظر متوقف خواهد شد. برای اندازه‌گیری گستردگی از معیار میانگین فاصلهٔ درون کلاستری، طبق رابطه (۲) استفاده می‌کنیم.

$$Sp_{\bar{x}_k} = \bar{D}_k > T_d \;\&\&\; |X_k| > T_n \qquad (۱)$$

$$\bar{D}_k = \frac{\sum_{i=1}^{M} dist(x_i, \bar{x})}{u} \qquad (۲)$$

به طوری‌که، $Sp_{\bar{x}_k}$ ضابطهٔ بررسی برای شکستن هر خوشه با نماینده $\hat{x}_k$ است. T_n بیانگر حد آستانه بر روی تعداد تصاویر موجود در هر خوشه در هر مرحله است و T_d آستانه‌گذاری بر روی $\bar{D}_k$ معیار میانگین فاصله درون کلاسی است. $\bar{x}$ بیان‌گر میانگین درون هر کلاس نمایندگان است. X_k تعداد تصاویر موجود در هر خوشهٔ مورد بررسی برای شکستن است و u تعداد کل داده‌های درون خوشه نماینده است. درنهایت، مراکز هر خوشه با

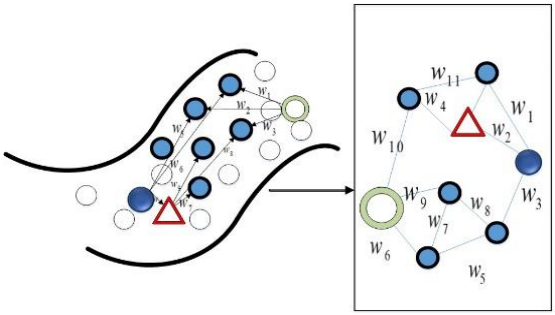
[3] Diversity

[DOI: 10.52547/jsdp.18.4.49]

میانگین‌گیری هر خوشه به‌عنوان نماینده در نظر گرفته می‌شوند.

۳-۳- چگونگی برچسب‌گیری نمایندگان با انتشار برچسب از داده‌های آموزشی

در ابتدا با استفاده از داده‌های آموزشی X و نمایندگان مستخرج‌شده از داده‌های آموزشی که آن‌ها را با X_p نشان می‌دهیم، گراف وزن‌دار مبتنی بر ضرایب تُنُک محلی $G =< \nu, \mathcal{E} >$ ساخته می‌شود که در شکل (۴) نشان داده شده است. رئوس گراف $\nu = X + X_p$ و $\mathcal{E}$ مجموعه یال‌های گراف است که e_{ij} بیان‌گر ارتباط بین دو راس x_i و x_j است. به عبارت دیگر در گراف ساخته شده، رئوس گراف با اجتماع تصاویر آموزشی و نمایندگان تعریف می‌شوند. P= (N+M) بیانگر تعداد کل رئوس گراف است.



(شکل-۴): گراف ساخته‌شده از نمایندگان و مجموعه آموزشی
(Figure-4): Constructed graph from prototypes and training images

در تکنیک‌های مبتنی بر بازنمایش تُنُک [49,50]، هر الگو تنها با استفاده از ترکیب خطی تعداد کم داده نمایش داده می‌شود. با به‌کارگیری این تکنیک‌ها، وزن‌های گراف G را با روش کدگذاری خطی محدودیت محلی[1] [9] به‌دست می‌آوریم. به عبارتی دیگر، در این روش با بهینه کردن تابع هدف (۳) وزن‌های گراف استخراج می‌شوند. طبق این الگوریتم ابتدا، K نزدیک‌ترین داده به کل نمایندگان و داده‌های آموزشی را پیدا کرده و سپس نمایندگان تنها با استفاده از نزدیک‌ترین تصاویر آموزشی مبتنی بر تُنُک بودن بازسازی می‌شوند.

$$c^* = \underset{s.t.\ \sum_j^K e_{ij}=1}{argmin} ||x_i - e^T B|| \qquad (۳)$$

در رابطه فوق، $B = \{b_j\}$، $\{j = 1,2,\dots,N\}$ کتابچه کدی[2] است که وزن یال‌ها براساس آن بازسازی می‌شود و برای ورودی x_i، کد وزن بازسازی‌شده e_i را داریم. e_i یک بردار به اندازه $P \times 1$ با K عنصر غیر صفر است به طوری‌که هر درایه e_i وزن میان یک نماینده و یک تصویر آموزشی است. با شرط $\sum_j^K e_j = 1$ تضمین می‌کنیم که الگوریتم LLC به‌ازای انتخاب هر K یال مختلف گراف، نسبت به جابه‌جایی تغییرناپذیر است. از آنجا که در این روش محلی[3] بودن اهمیت بالاتری نسبت به تُنُکی[4] دارد. یک عبارت دیگر جهت تطبیق‌دهندهٔ محلی برای هر پایه در کتابچه در نظر گرفته می‌شود و طبق رابطه (۴) به وزن‌های بازسازی تنک اضافه می‌کنیم، که $d_i = \exp(\frac{dist(e_i,B)}{\sigma})$ یک کرنل گوسی با درجهٔ تطبیق δ جهت تنظیم سرعت کاهش وزن برای تطبیق‌دهنده است. به‌عبارتی $dist(e_i,B) = [dist(e_i,b_1),\dots,dist(e_i,b_M)]^T$ این‌گونه تعریف می‌شود.

$$\min \sum_{i=1}^{N} \|e_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2 \qquad (۴)$$
$$s.t.\ 1^T c_i = 1, \forall i$$

فرض کنید F یک ماتریس به اندازهٔ P×L است، برای N سطر اول که مجموعه تصاویر آموزشی هستند، به ازای برچسب‌های مدنظر هر تصویر آموزشی مقدار غیر صفر و برای M سطر باقی مانده که نمایندگان مستخرج از مجموعه آموزشی هستند، مقدار صفر درنظر گرفته‌ایم. برای گراف ساخته شده $F = [F_1^T, F_2^T, \dots, F_P^T]^T$ متناظر با ν، رأس گراف تعریف می‌کنیم؛ به طوری‌که بردار F_i متناظر با نقطه x_i است و بیان‌گر یک نماینده یا تصویر آموزشی است. اگر $Y = (y_1, y_2, \dots, y_P)^T$ فرض شود به‌طوری‌که $y_i = \tau_i\ (i \le N, \tau_i \subset \mathcal{L})$ به‌ازای هر تصویر x_i در مجموعه کل که شامل برچسب‌های τ_i است، برابر یک قرار داده می‌شود و $y_i = 0\ (i > N)$، برای نمایندگان مقدار صفر تعیین می‌شود. از طرفی دیگر، مقدار اولیه ماتریس F برابر Y تعیین می‌شو ؛ به عبارتی دیگر، $F_0 = Y$ قرار می‌دهیم.

تابع هزینه مربوط به وزن‌های گراف طبق رابطه (۵) تعریف می‌شود.

$$Q(F) = \sum_{i=1}^{P} \sum_{j:x_j \in \mathcal{N}(x_j)} W_{ij} \|F_i - F_j\|^2 + \gamma \sum_{i=1}^{P} \|F_i - y_i\|^2 \qquad (۵)$$

عبارت اول سمت راست تابع هزینه فوق، یک محدودیت هموارساز[5] است؛ به این معنی است که یک

[1] Locality-constraints Linear Coding (LLC)
[2] Dictionary
[3] Locality
[4] Sparsity
[5] Smoothness Constraints

[DOI: 10.52547/jsdp.18.4.49]

تابع هدف خوب نباید میان نقاط نزدیک به هم تغییرات زیادی داشته باشد. عبارت دوم، یک محدودیت برازشی[1] است؛ به این معنی که تابع هدف نباید از مرحله نسبت‌دهی برچسب‌های اولیه تغییرات زیادی داشته باشد و توازن میان رقابت دو محدودیت، با استفاده از پارامتر مثبت $\gamma > 0$ کنترل می‌شود.

از آن‌جا که تابع هزینه مسدله محدب[2] است، پس دارای حداقل یک بهینه سراسری[3] است. جهت به‌دست‌آوردن مینمم بهینه، گرادیان $Q(F)$ نسبت به $F = [F_1^T, F_2^T, \dots, F_P^T]^T$ طبق رابطه (۶) محاسبه می‌شود.

$$\frac{\partial Q(F)}{\partial F} = [(I - W) + (I - W)^T]\mathrm{F} + 2\gamma(F - Y) \qquad (۶)$$

طبق [50]، با تقریب‌زدن حل مسأله مینمم‌سازی $Q(F)$ و صفر قرار دادن رابطه (۶)، جواب مسأله یک رابطه بازگشتی مطابق رابطه (۷)، خواهد بود.

$$F = (I - \alpha)(I - \alpha W)^{-1}Y \qquad (۷)$$

به طوری‌که $\alpha = \frac{1}{1+\gamma}$ است.

درنهایت خروجی خلاصه‌سازی پایگاه داده به‌صورت مجموعه نمایندگان X_p و برچسب‌های آن‌ها حاصل می‌شود. حال، این خلاصه‌شدهٔ داده‌های آموزشی را به یک الگوریتم برچسب‌زنی پیشنهادی می‌دهیم.

با توضیحات فوق الگوریتم خلاصه‌سازی پایگاه داده به‌صورت زیر عمل می‌کند:

۱. در ابتدا W، وزن گراف برای رئوس نمایندگان و داده‌های آموزشی از طریق بازسازی تقریبی کدگذاری خطی محدودیت محلی طبق رابطه (۳) محاسبه می‌شود.

۲. از آن‌جایی که ارتباط منفی معنایی در روابط بین رئوس گراف وجود ندارد، آستانه‌گذاری بر روی وزن W صورت می‌گیرد و یال‌های با وزن منفی حذف می‌شوند.

۳. با رابطه $W = \frac{W^T+W}{2}$، متقارن‌سازی بر روی ماتریس وزن انجام می‌شود. بدین معنی که، ارتباط بین هر درایه x_i به x_j با x_j به x_i تفاوتی وجود ندارد.

۴. رابطه $F(t+1) = \alpha F(t) + (1-\alpha)Y$ تا هم‌گرایی تکرار می‌شود که $0 < \alpha < 1$ است. در هر تکرار، هر درایه از ماتریس F با استفاده از اطلاعات همسایگانش و هم‌چنین اطلاعات اولیه مقداردهی شده تعیین می‌گردد. به عبارتی دیگر، برچسب‌ها از مجموعه‌داده آموزشی به نمایندگان انتشار داده می‌شوند.

۵. $F^{(t+1)}(u)_{u\in X} = F^{(t)}(u)_{u\in X}$، برای تثبیت اطلاعات مجموعه‌داده آموزشی، مقادیر F مربوط به آن‌ها را به حالت اولیه باز می‌گرداند.

طی عملیات تکرارشونده فوق، برچسب‌ها به نمایندگان بدون برچسب، انتشار پیدا می‌کنند تا به هم‌گرایی برسیم. خروجی خلاصه‌سازی پایگاه داده به‌صورت مجموعه نمایندگان X_p و برچسب‌های آن‌ها به‌دست می‌آید. به دو صورت برچسب‌های انتشار یافته برای نمایندگان در نظر گرفته می‌شود؛ تنها خود برچسب‌ها یا برچسب به همراه مقدار تأثیر آن برچسب. در حالت اول، اگر از داده‌ی x_i به y_i $(y_i \neq x_i)$ برچسبی انتشار یابد، $Y_{ij} = 1$ است درغیر این صورت صفر خواهد بود. برای حالت دوم، برچسب‌های نهایی نمایندگان به‌صورت مجموع ترکیب خطی وزن‌های گراف در مقادیر نهایی F به‌ازای تمام برچسب‌های تعریف شده است. از آنجا که وزن‌های میان مجموعه آموزشی در تعیین برچسب‌های نمایندگان ندارد، این مقادیر از ترکیب خطی حذف شده و طبق رابطه (۸) خواهد بود.

$$\sum_{i=M}^{P} F_{ij} \times W_i, \qquad j \in \{1,2,..,L\} \qquad (۸)$$

برای مثال، نمایندهٔ i-ام، $\hat{x}_i$، دارای مجموعه برچسب‌های $\hat{y}_i \in Y$ با مقادیر حقیقی است. به طوری‌که برچسب‌های با مقادیر بالاتر، احتمال تعلق بالاتری به هر نماینده دارند. پس از خلاصه‌سازی می‌توان از نمایندگان به همراه برچسب‌های آن‌ها برای تعیین برچسب‌های هر تصویر ورودی استفاده کرد که در بخش بعدی یک روش پیشنهاد شده است.

۳-۴- روش پیشنهادی نسبت‌دهی برچسب

سامانهٔ پیشنهادی، پایگاه داده واقعی آموزشی را با یک پایگاه داده خلاصه شده شامل نماینده‌ها و برچسب‌های آن‌ها جایگزین می‌کند و تنها با استفاده از نمایندگان،

[1] Fitness Constraints
[2] convex

[DOI: 10.52547/jsdp.18.4.49]

عملیات برچسب‌زنی یک تصویر ورودی را انجام می‌دهد. روند کلی فرآیند در شکل (۵) نمایش داده شده است. فرض کنید داده‌های آموزشی، به عبارتی مجموعه‌داده خلاصه‌شدهٔ آموزشی در بخش قبل به همراه برچسب‌هایشان را داریم و مانند هر روش برچسب‌زنی دیگر، برچسب‌های پیش‌بینی شده را به تصویر ورودی x_t نسبت می‌دهیم.

پس در ابتدا گراف وزن‌داری بین تصویر ورودی و نمایندگان تشکیل می‌دهیم. وزن رئوس گراف ساخته شده، $S(t,P)$ است که از رابطه (۹) به‌دست می‌آید، به عبارتی دیگر میزان شباهت تصویر ورودی به نمایندگان توسط این رابطه حاصل می‌شود. δ، مقادیر وزن بین مجموعه خلاصه‌شده و تصویر ورودی را تنظیم می‌کند.

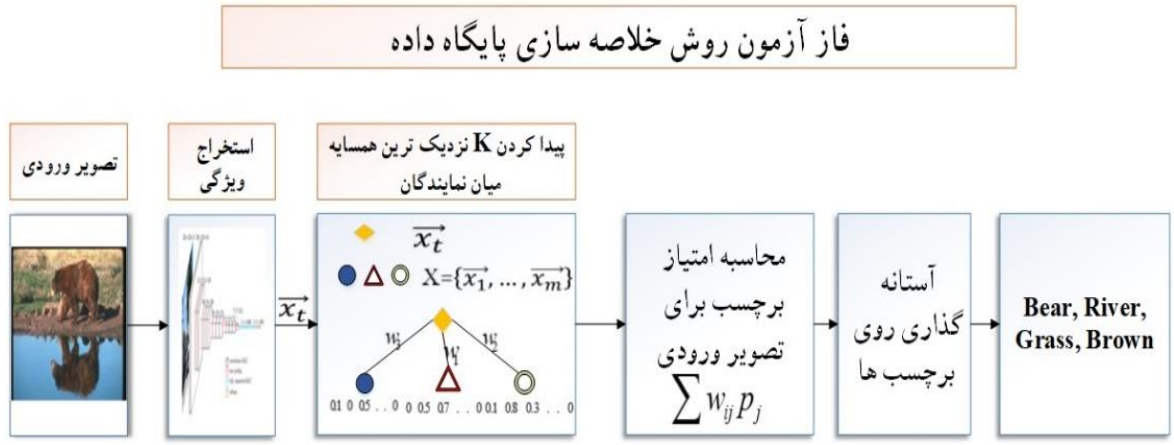
$$S(t,P) = e^{-\frac{\|x_t - x_p\|^2}{2\delta^2}} \qquad (۹)$$

سپس با ترکیب خطی مقادیر وزن میان نمایندگان و تصویر آزمون ورودی و مقدار برچسب‌های هر نماینده طبق رابطه (۱۰)، احتمال تعلق هر برچسب به تصویر ورودی، محاسبه خواهد داشت.

$$Score_t(j) = \sum_{p=1}^{M} S(t,P) \times F_{pj} \qquad (۱۰)$$

درنهایت، با نسبت‌دهی برچسب تطبیقی به وسیلهٔ آستانه‌گذاری بر روی احتمال برچسب‌ها، برچسب‌های نهایی یک تصویر ورودی تعیین می‌شود. امتیاز برچسب‌ها برای هر تصویر ورودی یک ماتریس $Score_t$ به اندازهٔ $1 \times L$ برای یک تصویر ورودی t است که طبق رابطه (۱۱) تعیین می‌شود.

$$Score_t = [Score_t(1), \dots, Score_t(L)] \qquad (۱۱)$$



(شکل-۵): فاز برچسب‌زنی روش پیشنهادی
(Figure-5): Tag Assignment phase of the proposed method

۴- آزمایش‌ها و نتایج

جهت ارزیابی سامانه‌های برچسب‌زنی طراحی شده، مجموعه‌داده‌های استاندارد برچسب‌زنی به نام ESP-Game و IAPRTC12 استفاده شده‌اند. این مجموعه‌داده‌ها شامل صحنه‌های طبیعی مختلف، اشیا، طبیعت و ... هستند؛ هم‌چنین این مجموعه‌داده‌ها دارای صدها کلمهٔ متمایز هستند. جهت ارزیابی و مقایسه عادلانه با روش‌های پیشین برچسب‌زنی، از تقسیم‌بندی یکسان برای مجموعه آموزشی و آزمون استفاده می‌کنیم. اطلاعات آماری این مجموعه‌داده‌ها در جدول (۱) قابل مشاهده است.

(جدول-۱): اطلاعات آماری مجموعه‌داده مورد استفاده در آزمایش‌های تجربی
(Table-1): Statistical information about Datasets in our experiments

پایگاه داده	تعداد تصاویر	تعداد برچسب‌ها	تعداد تصاویر آموزشی	تعداد تصاویر تست	تعداد برچسب به‌ازای هر تصویر(حداکثر، میانه، میانگین)
ESP-Game	20770	268	18689	2081	4.7, 5, 15
IAPR-TC12	19672	291	17665	1962	5.7, 5, 23

[DOI: 10.52547/jsdp.18.4.49]

۴-۱- معیار ارزیابی کارایی سامانه‌های برچسب‌زنی

برای تجزیه و تحلیل داده‌ها از سه معیار میانگین دقت[1]، میانگین یادآوری[2] و F1 بر روی کل برچسب‌ها استفاده می‌کنیم. معیار دقت برای یک برچسب، تعداد تصاویری است که برچسب به درستی نسبت داده شده تقسیم بر کل تعداد تصاویر پیش‌بینی شده برای آن برچسب طبق رابطه (۱۲) تعریف می‌شود.

یادآوری، تعداد تصاویر با برچسب به درستی نسبت داده شده تقسیم بر تعداد تصاویری که دارای این برچسب هستند و طبق رابطه (۱۳) است. حال معیار مناسب‌تر جهت ارزیابی بهتر می‌توان با معیاری که هر دو معیار دقت و یادآوری را در نظر می‌گیرد، سامانه را بسنجیم. برای این منظور، میانگین هارمونی دقت و یادآوری که به آن امتیاز F متعادل یا F1 گویند، طبق رابطه (۱۴) تعریف می‌شود.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (۱۲)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (۱۳)$$

$$\text{F1} = \frac{2}{(\frac{1}{\text{recall}} + \frac{1}{\text{precision}})} \qquad (۱۴)$$

به طوری‌که به‌ازای هر برچسب مقادیر دقت و یادآوری محاسبه شده و سپس میانگینی بر روی کل برچسب گرفته می‌شود. TP، مثبت صحیح[3]، تعداد تصاویری است که سامانه برچسب را به درستی نسبت داده و FN، منفی غلط[4]، تعداد تصاویری است که برچسب توسط سامانه به نادرستی پیش‌بینی شده، ولی برچسب واقعی تصویر است و درآخر FP، مثبت غلط[5]، تعداد تصاویری است که سامانه برچسب را پیش‌بینی کرده اما برچسب واقعی تصویر نیست.

طی بررسی آزمایش‌ها مقادیر بهینه پارامترها برای هر مجموعه‌داده بدین صورت درنظر گرفته شده است. مقدار K که پارامتر نزدیک‌ترین همسایه جهت بازسازی وزن گراف میان نمایندگان و تصاویر پایگاه داده آموزشی برای مجموعه‌دادهٔ IAPRTC12، مساوی مقدار ۵ و مجموعه‌دادهٔ ESP-Game، مساوی مقدار ۹ قرار داده شده است که این مقادیر سبب بیشینه‌شدن کارایی سامانهٔ برچسب‌زنی پیشنهادی و کاهش زمان اجرا خواهد شد با بررسی گراف ساخته شده مطمئن می‌شویم که به تعداد کافی تصاویر مؤثر آموزشی به هر نماینده متصل شده است. جهت بررسی مناسب در تنظیم پارامترهای قابل تنظیم، داده‌های آموزش به‌صورت تصادفی به دو بخش آموزش و اعتبارسنجی به اندازه‌ی ۷۵ و ۲۵ درصد تقسیم می‌شوند. این کار برای ۱۰ مرتبه انجام شده است و سپس میانگین نتایج حاصل از هر ۱۰ بخش برای به دست آوردن پارامترها استفاده شده است. برای مقایسه بهتر کارایی روش پیشنهادی جهت خلاصه‌سازی مجموعه آموزشی، روشی پایه یا همان 2PKNN در نظر گرفته شده است. این روش در میان روش‌های پیشین برچسب‌زنی بالاترین کارایی را دارد.

۴-۲- بررسی و مقایسه ویژگی‌های مختلف

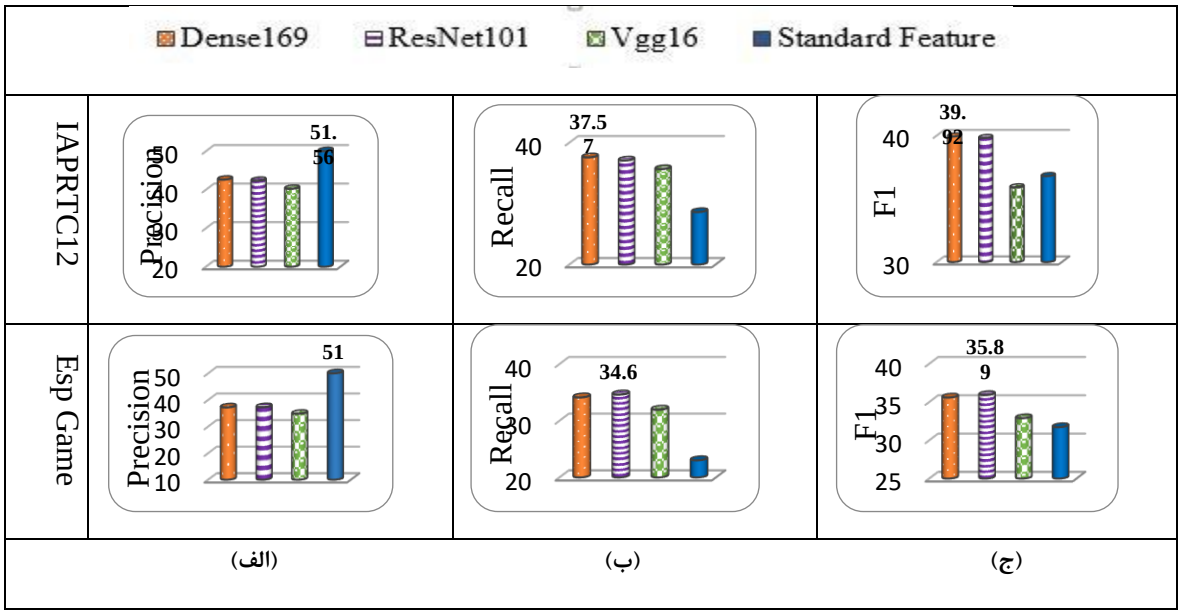
کارایی سامانه‌ی برچسب‌زنی طراحی شده (مبتنی بر دقت، یادآوری و معیار F1) برای ویژگی‌های مختلف در شکل (۶) بررسی شده است.

طبق (شکل (۶)، ارزیابی کارایی سامانهٔ برچسب‌زنی پیشنهادی بر حسب سه معیار ذکر شده، بر روی ویژگی‌های مختلف انجام شده است. هر سطر بیانگر یک مجموعه‌داده است. در هر شکل مقایسه‌ای بین ویژگی‌های مختلف انجام شده است. طبق این مقایسه، ویژگی‌های یادگیری عمیق همواره نتایج بالاتری (برحسب معیار یادآوری و F1) نسبت به ویژگی‌های استاندارد دارند. گرچه معیار دقت ویژگی‌های استاندارد کمی بالاتر است، اما معیار ارزیابی مناسب‌تر F1 حاکی از برتری ویژگی‌های یادگیری عمیق دارد. از طرفی در میان ویژگی‌های ذکر شده، ویژگی Dense169 بهترین عملکرد F1 را دارد. هم‌چنین با توجه به سرعت بالای استخراج ویژگی در این معماری، یک ویژگی بسیار مناسب محسوب می‌شود. تعداد نمایندگان در این آزمایش ۳ هزار است.

در ادامه مقایسه‌ای میان سرعت پیش‌پردازش استخراج ویژگی با استفاده از پردازنده هسته مرکزی با مشخصات Intel (R) Core (TM) i7-6700 HQ 3.1GHz و کارت پردازنده گرافیک Geforce1060-GTX در جدول (۲) آورده شده است. طبق جدول (۲) زمان اجرای استخراج ویژگی بر روی کارت گرافیک حدود ۲ میلی‌ثانیه است که مقدار بسیار کمی است. از طرفی، در روش‌های قبلی برچسب‌زنی [30-29]، از ۱۴ نوع بردار ویژگی استفاده کرده‌اند که استخراج آنها به مراتب بیشتر از روش پیشنهادی است.

[1] Precision
[2] Recall
[3] True Positive
[4] False Negative
[5] False Positive

[DOI: 10.52547/jsdp.18.4.49]



(شکل-۶): عملکرد سامانهٔ برچسب‌زنی برحسب میانگین دقت (الف) ، میانگین یادآوری (ب) و F1 (ج) ، در هر شکل مقایسه بین ویژگی‌ها به‌ترتیب Dense169, ResNet101, Vgg16, Standard Features است. هر سطر مربوط به یک مجموعه‌داده است.

به‌ترتیب IAPRTC12، ESP- Game

(Figure-6): Annotation performance in terms of mean precision (a), mean recall (b), and F1 score (c) (vertical axis) for different CNN features. In each figure, from left to right, the results are related to Dense169, ResNet101, VGG16, and Standard features, respectively. Each row of plots correspondes to one database. The first one is IAPRTC12 and second is ESP-Game.

(جدول-۲): میانگین زمان استخراج ویژگی برای یک تصویر بر حسب میلی‌ثانیه بر روی CPU و GPU

(Table-2): Average feature extration time (in millisecond) for an image on two types of hardwares

سخت‌افزار	CPU	GPU
زمان (میلی‌ثانیه)	۱۰/۷۸	۱/۵۷

۴-۳-آنالیز حساسیت بر روی تعداد نمایندگان

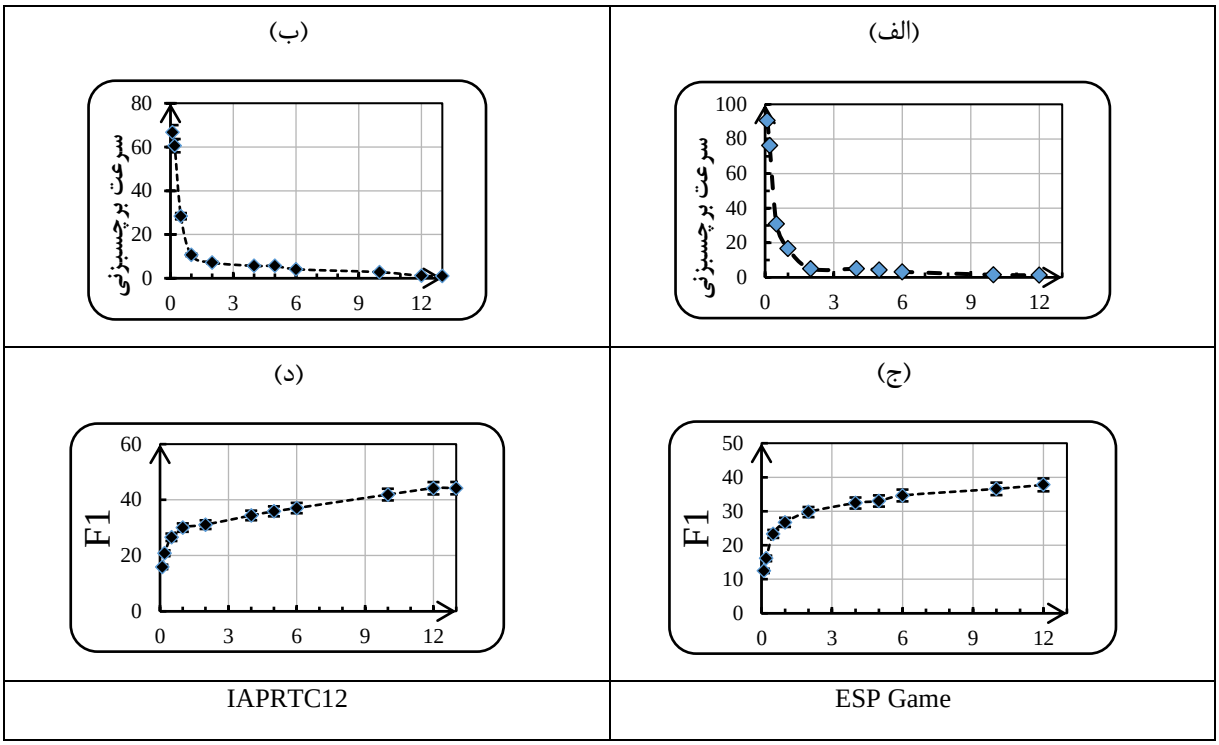
یکی از مسائل مهم در مسائل با مقیاس بزرگ آنالیز حساسیت بر روی تعداد نمایندگان است. برای بررسی این موضوع، زمان متوسط برچسب‌زنی برای هر تصویر را بر روی یک کامپیوتر شخصی با مشخصات حافظهٔ ۱۶ گیگابایت و واحد پردازنده Intel (R) Core (TM) i7-6700 HQ 3.1GHz و پردازنده‌ی گرافیکی Geforce1060-GTX در محیط برنامه‌نویسی MATLAB، محاسبه کرده‌ایم و آن را با زمان یک روش برچسب‌زنی مبتنی بر جستجو مثل 2PKNN با ویژگی یکسان، مقایسه کرده‌ایم.

در شکل (۷) سطر اول، بهبود سرعت برچسب‌زنی پیشنهادی نسبت به روش پایه به‌ازای تعداد نمایندگان مختلف برای دو مجموعه‌داده IAPRTC12 و Esp-Game نشان داده شده است. برای مجموعه‌داده IAPRTC12 با تعداد ۶ هزار نماینده، کارایی سامانه را نسبت به روش پایه طبق معیار F1، ۱۰٪ و نسبت به معیار یادآوری ۳۰/۸٪ افزایش داده است. از طرفی زمان اجرای مصرفی ۴/۲ برابر کم‌تر شده است. در صورتی که اگر با تعداد ۳ هزار نماینده، خلاصه‌سازی انجام شود؛ کارایی سامانه نسبت به معیار F1، ۹۲٪ حفظ خواهد شد و نسبت به معیار یادآوری ۱۵/۶٪ افزایش خواهیم داشت. در عین حال که زمان اجرای برچسب‌زنی ۷/۳ برابر سریع‌تر خواهد بود.

در مجموعه دادهٔ ESP-Game با تعداد ۶ هزار نماینده، کارایی سامانه نسبت به روش پایه طبق معیار F1، ۹۷/۳٪ حفظ می‌شود و زمان اجرای مصرفی جهت برچسب‌زنی ۳/۴ برابر سریع‌تر از روش پایه است. اگر با در نظر گرفتن تعداد ۳ هزار نماینده، خلاصه‌سازی انجام شود؛ کارایی سامانه طبق معیار F1، ۷۸/۹٪ کارایی حفظ خواهد شد. در عین حال زمان اجرای برچسب‌زنی ۶/۲ برابر سریع‌تر خواهد بود. از طرفی، تنها با جایگزین‌کردن هزار نماینده جایگزین کل پایگاه داده برای مجموعه‌دادهٔ ESP-Game با سرعت ۱۶/۶ برابر روش پایه، برچسب‌زنی را با کارایی متوسط نسبت به معیار F1، ۶۹/۹٪ نسبت به روش پایه خواهد بود.

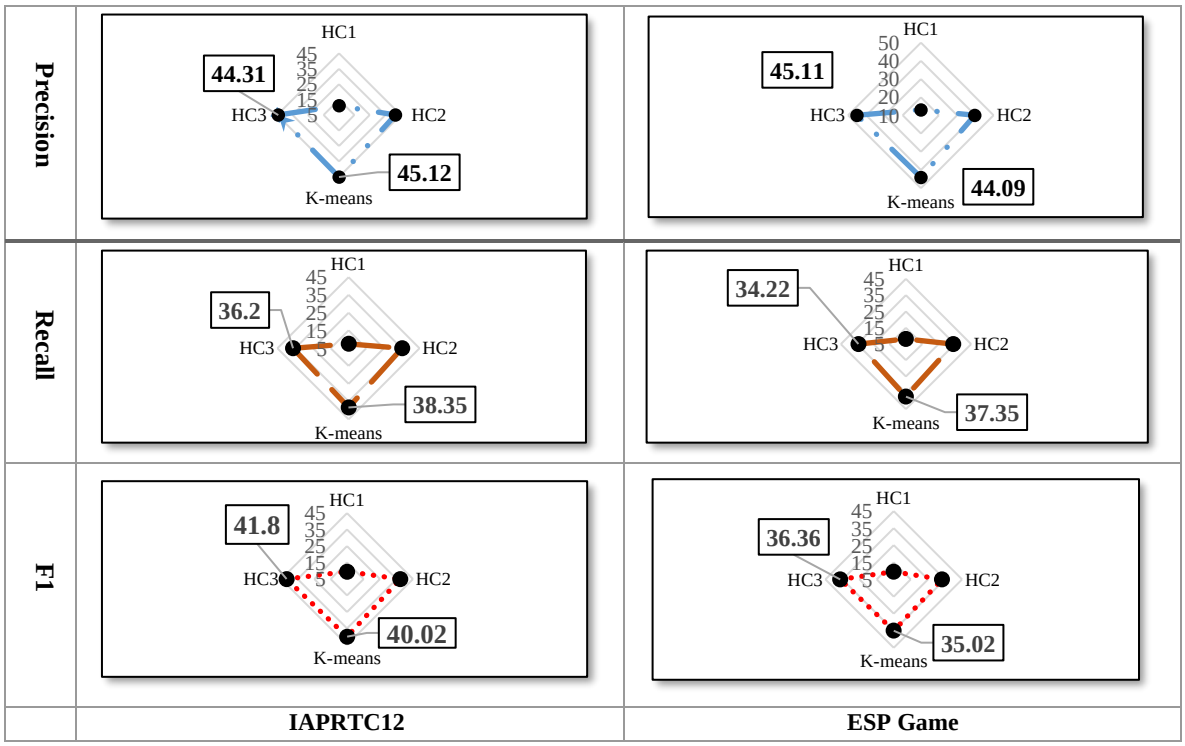
با افزایش تعداد نمایندگان کارایی نسبت به دو معیار F1 و یادآوری افزایش می‌یابد. درنهایت با جایگزین کردن تنها تعداد نمایندگان به اندازهٔ ۲۲/۶۷٪ پایگاه داده آموزشی برای مجموعه‌دادهٔ IAPRTC12 و تعداد نمایندگان ۳۱/۷۸٪ به جای کل مجموعه‌داده آموزشی ESP-Game به طور قابل توجهی سبب تسریع زمان اجرای برچسب‌زنی نسبت به روش پایه شده است و به طور خلاصه، کارایی سامانه نسبت به دو معیار F1 و یادآوری بر روی مجموعه دادگان برچسب‌زنی افزایش داده شده یا حفظ شده است، که با این جایگزینی نمایندگان به پایگاه داده آموزشی، توانسته‌ایم سامانه برچسب‌زنی را مقیاس‌پذیر کنیم.

[DOI: 10.52547/jsdp.18.4.49]



(شکل–۷): سطر اول، نسبت سرعت زمان برچسب‌زنی روش پیشنهادی نسبت به روش پایه به‌ازای تعداد نمایندگان (مضرب ۱۰۰۰) قرار داده شده است. سطر دوم، میزان کارایی سامانه به‌ازای نمایندگان مختلف نشان داده است. نمودارهای سمت راست متناظر با مجموعه‌داده ESP Game و سمت چپ،متناظر با مجموعه دادهٔ IAPRTC12 هستند.

(Figure-7): First row, the proportion of annotation time for the proposed approach in comparison to the baseline method for different number of prototypes. Second row, performance of the proposed method for different number of prototypes. Each column of plots corresponds to a certain database. The right one is IAPRTC12 and the left one is ESP Game.

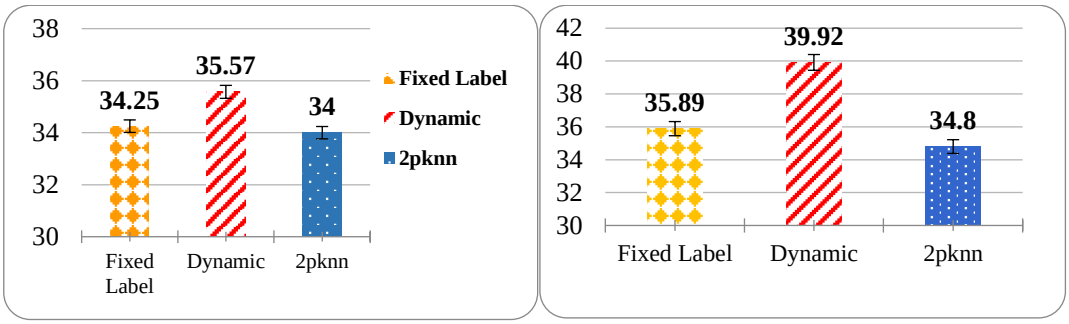


(شکل–۸): مقایسه دو روش استخراج نماینده و تأثیر آن بر روی نتیجه نهایی برچسب‌زنی. هرچه تعداد لایه خوشه‌بندی سلسله مراتبی بیش‌تر شود، کارایی سامانه نزدیک یا حتی بالاتر از روش خوشه‌بندی رایج خواهد شد.

هر ستون مربوط به یک مجموعه‌داده است.

(Figure-8): A comparison between the two different types of prototype extraction. Increasing the layers in hierarchical clustering improves the performance of our proposed method and even higher than K-means prototype extraction. Each column of plots is dedicated certain database.

[DOI: 10.52547/jsdp.18.4.49]



(شکل-۹): مقایسه کارایی (نسبت به معیار F1) سامانه پیشنهادی بین روش‌های نسبت دهی برچسب پویا و تعداد ثابت و روش پایه برای دو مجموعه‌داده به‌ترتیب از راست، ESP Game ,IAPRTC12

(Figure-9): Comparison performance (F1) of the proposed method based on different tag assignments (Dynamic Label, Fixed Label, and 2PKNN); each plot corresponds to a certain database. The left one is IAPRTC12 and right one is ESP Game.

۴-۴- مقایسه روش‌های استخراج نماینده

در این بخش به مقایسه دو روش استخراج نماینده مبتنی بر خوشه‌بندی K-means و خوشه‌بندی سلسله مراتبی با ضابطه، نسبت به سه معیار دقت، یادآوری و F1 پرداختیم که در شکل (۸) نشان داده شده است. طبق شکل، مقادیر ممکن جهت ارزیابی کارایی به شکل لوزی‌های محیطی به اندازه ۱۰ از هم قرار گرفته‌اند؛ به طوری‌که هرچه کارایی سامانه برچسب‌زنی با استفاده از روش‌های استخراج نماینده‌ای کم‌تر باشد، به مرکز شکل نزدیک‌تر و هر چه کارایی بهتر باشد، در گوشه‌های شکل قرار می‌گیرند. روش استخراج نماینده با K-means دارای کارایی مناسبی نسبت به روش خوشه‌بندی سلسله مراتبی یک لایه[1] و دو لایه[2] است به طوری‌که مقایسه کارایی سامانه برچسب‌زنی بر روی مجموعه‌دادهٔ IAPRTC12، میان دو روش استخراج نماینده دو لایه نسبت به تک لایه، طبق سه معیار دقت و یادآوری و F1، به‌ترتیب ۲۰۶٪، ۲۷۱٪ و ۲۹۲٪ و کارایی سامانه میان دو روش سلسله مراتبی دو لایه و K-means روش خوشه‌بندی K-means، نسبت به معیار دقت، یادآوری F1، به‌ترتیب ۱۰/۸٪، ۹/۲۵٪ و ۸٪ افزایش داشته‌ایم. در نتیجه، طبق (شکل-۸) هر چه تعداد لایه بیش‌تر شود، کارایی سامانه نسبت به هر سه معیار بر روی دو مجموعه دادهٔ IAPRTC12 و ESP-Game افزایش می‌یابد. دو روش خوشه‌بندی یک لایه و دو لایه سلسله مراتبی کارایی کم‌تری دارند؛ اما زمانی که تعداد لایه‌های استخراج نماینده بیش‌تر می‌شود، مثلا سه لایه[3]، نتیجه بهتر از خوشه‌بندی سلسله مراتبی با تعداد لایه‌های کم و حتی روش استخراج نماینده با استفاده از K-means است. به طوری‌که کارایی سامانه برچسب‌زنی نسبت به معیار F1، روش خوشه‌بندی ۳ لایه روی مجموعه‌داده IAPRTC12، ۴/۴ درصد و مجموعه دادهٔ ESP-Game، ۳/۸ درصد افزایش نسبت به روش استخراج نماینده با استفاده از K-means داشته‌ایم.

۴-۵- مقایسه روش برچسب‌زنی پیشنهادی با2PKNN

در این بخش به مقایسه سه روش برچسب‌زنی پرداختیم. روش اول، خلاصه‌سازی پایگاه داده و نسبت‌دهی برچسب با 2PKNN[4] است و روش دوم، نحوهٔ نسبت‌دهی برچسب به‌صورت پویا[5] است که بر روی امتیاز به‌دست‌آمده از رابطه (۷) آستانه‌گذاری می‌شود. از آن‌جا که مقدار این آستانه برای هر مجموعه‌داده با تغییر اندکی کارایی تغییر می‌کند، روشی تطبیقی مبتنی بر داده‌های آموزشی آستانه را آموزش می‌دهیم، با استفاده از روش‌های K-Cross Validation مجموعه‌دادهٔ آموزشی را به K (۱۰) بخش تقسیم کرده و سپس به‌ازای هر بخش بهترین آستانه مبتنی بر کارایی سامانه را به‌دست می‌آوریم. درنهایت میانگینی از این مقادیر را به‌عنوان آستانه مجموعه‌داده در نظر می‌گیریم. روش سوم نسبت‌دهی برچسب ثابت است[6]، به ازای یک تعداد ثابت برچسب‌های با بالاترین امتیاز به‌عنوان برچسب پیشنهادی انتخاب می‌شود. مقایسه میان چنین روش‌هایی در شکل (۹) آورده شده است. با توجه به شکل (۹)، روش ارائه شده با داشتن پایگاه دادهٔ یکسان، نتایجی بهتر از روش پایه، یا همان 2PKNN، دارد. به هم‌چنین با روش نسبت‌دهی پویا، کارایی بر مبنای F1 ۱۱/۷۸٪ برای مجموعه‌دادهٔ ESP Game به‌طور معناداری افزایش داشته است و برای مجموعه‌دادهٔ IAPRTC12 کارایی سامانه با استفاده از نسبت‌دهی پویا، ۱۳٪ بهبود داشته است.

[1] Hierarchical Clustering by 1 layer (HC1)
[2] Hierarchical Clustering by 2 layer (HC2)
[3] Hierarchical Clustering by 3 layer (HC3)

[4] Training Database Summarization by neighborhood Label propagation (TDSNLP+2PKNN)
[5] Training Database Summarization by neighborhood Label propagation with Dynamic Tags (TDSNLPDT)
[6] Training Database Summarization by neighborhood Label propagation with Fixed Tags (TDSNLPFT)

[DOI: 10.52547/jsdp.18.4.49]

۴-۶- مقایسه روش خلاصه‌سازی با روش‌های پیشین برچسب‌زنی

طبق جدول (۳) کارایی روش‌های پیشنهادی برچسب‌زنی نسبت به روش‌های مولد و تمایزی مانند KSVM-VT، SVM-DMBRM و MBRM بر حسب معیار F1 بر روی دو مجموعه‌دادهٔ IAPRTC12 و مجموعه دادهٔ ESP Game با وجود استفاده از تعداد محدود نمایندگان، افزایش قابل توجهی داشته‌اند. از طرفی در مجموعه‌دادهٔ ESP Game، روش NMF-KNN مقدار F1 برابر ۲۹٪ است در حالی‌که روش پیشنهادی با نسبت‌دهی پویای برچسب با تعداد ۶ هزار نماینده، بر روی معیار F1 ۱۶٪۲۲ و بر روی معیار یادآوری ٪۳۶ نسبت به این روش افزایش داشته‌ایم. نسبت به روش پایه یا 2PKNN با ویژگی‌های مستخرج از یادگیری عمیق نیز، کارایی سامانه برحسب معیار F1، بر روی مجموعه‌دادهٔ IAPRTC12 حفظ شده است و مقایسه میان کارایی روش پایه با روش پیشنهادی TDSNLPDT افزایش ٪۱۰ داشته‌ایم. روش پیشنهادی ارائه شده در مقایسه با روش‌های مقیاس‌پذیر ارائه شده مانند MVSAE بر روی مجموعه‌دادهٔ IAPRTC12 معیار دقت، یادآوری و F1، به‌ترتیب ٪۹/۳، ٪۱۰/۵ و ٪۱۰ بر روی مجموعه‌دادهٔ ESP Game برحسب معیار یادآوری و F1 ترتیب، ٪۳۲ و ٪۱۰/۵۲ افزایش داشته‌ایم.

روش پیشنهادی نسبت به روش Mvg-NMF، بر حسب معیار دقت، یادآوری و F1، به‌ترتیب ٪۲/۱، ٪۵/۱ و ٪۲/۳۲ بر روی مجموعه‌دادهٔ IAPRTC12 و بر روی مجموعه‌دادهٔ ESP Game کارایی سامانه بر حسب معیار یادآوری و F1 به‌ترتیب ٪۱۲/۱۲ و ٪۲/۲۷ افزایش قابل مشاهده است، البته قابل توجه است که روش اخیر مقایسه شده با روش پیشنهادی، علاوه بر کارایی سامانه، سامانهٔ برچسب‌زنی با زمان اجرای ده برابر بهتر از روش مبتنی بر جستجویی مثل 2PKNN ارائه کرده است.

روش پیشنهادی TDSNLPDT بر حسب معیار F1 بر روی مجموعه‌دادهٔ IAPRTC12 ٪۳/۲ و بر حسب معیار یادآوری ٪۱۰ افزایش داشته، هم‌چنین بر روی مجموعه‌دادهٔ ESP Game کارایی سامانه مبتنی بر F1 با وجود استفاده از نمایندگان به اندازه ٪۲۸ از تعداد کل مجموعه‌داده آموزشی در روش CCA-KNN کارایی سامانه برحسب معیار F1، ٪۹۳ حفظ شده است و بر روی معیار یادآوری ٪۲/۷ افزایش داشته‌ایم.

طبق بررسی‌های انجام‌شده، روش پیشنهادی با وجود جایگزینی نمایندگان بجای مجموعه‌داده‌های آموزشی،کارایی روش برچسب‌زنی مبتنی بر جستجوی پیشنهادی بهتر یا قابل مقایسه با روش‌های پیشین با رویکردهای مختلف برچسب‌زنی است. هم‌چنین در مقایسه با روش‌های برچسب‌زنی مقیاس‌پذیر روش برچسب‌زنی پیشنهادی بهتر از روش‌های پیشین عمل می‌کند.

(جدول-۳): مقایسه روش‌های پیشنهادی (TDSNLP، TDSNLPFT، TDSNLPDT) با روش‌های قبلی برای مجموعه‌داده IAPRTC12 وESP Game

(Table-3): Comparing the proposed methods (TDSNLP ،TDSNLPFT، TDSNLPDT) against the state-of-the-art methods for two datasets: IAPRTC12 and ESP Game

Method	Feature	ESP Game			IAPRTC-12		
		P	R	F1	P	R	F1
JEC [7]	Standard	0.23	0.19	0.21	0.25	0.16	0.19
MBRM [2]	Standard	0.21	0.17	0.19	0.21	0.14	0.17
Tag Prop (ML) [30]	Standard	0.49	0.2	0.28	0.48	0.25	0.33
Tag Prop (δML) [30]	Standard	0.39	0.27	0.32	0.46	0.35	0.4
2PKNN [29]	Standard	0.51	0.23	0.32	0.49	0.32	0.39
Fast Tag [4]	Standard	0.46	0.22	0.3	0.47	0.26	0.34
NMF-KNN [43]	Standard	0.33	0.26	0.29	-	-	-
CCA-KNN [10]	CNN(VGG16)	0.46	0.36	**0.41**	0.45	0.38	0.41
KSVM-VT [36]	Standard	0.33	0.32	0.33	0.47	0.29	0.36
SVM-DMBRM [53]	Standard	0.55	0.25	0.34	0.56	0.29	0.34
SKL-CRM [54]	Standard	0.41	0.26	0.32	0.47	0.32	0.38
MVSAE [40]	SAE	0.47	0.28	0.34	0.43	0.38	0.4
MSRRVK [55]	Standard	0.27	0.28	0.27	0.32	0.34	0.33
Mvg-NMF [42]	Standard	0.41	0.33	0.37	0.46	0.4	0.43
2PKNN	CNN(Dense169)	0.49	0.32	0.39	0.531	0.321	0.4
TDSNLP+2PKNN (our method)	CNN(Dense169)	0.44	0.28	0.342	0..45	0.34	0.39
TDSNLPFT (our method)	CNN(Dense169)	0.38	0.35	0.36	0.45	0.39	0.42
TDSNLPDT (our method)	CNN(Dense169)	0.40	**0.37**	0.38	0.47	**0.42**	**0.44**

[DOI: 10.52547/jsdp.18.4.49]

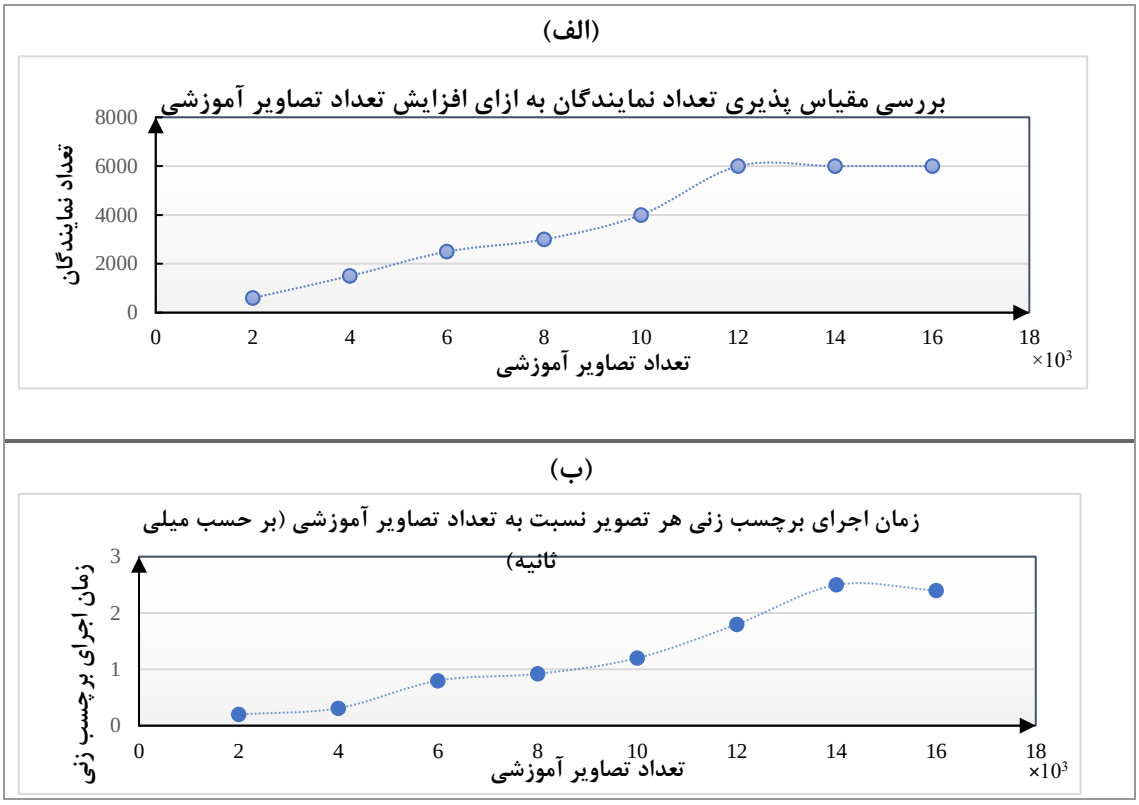
۴-۷- بررسی مقیاس‌پذیری روش پیشنهادی

به جهت بررسی مقیاس‌پذیری سامانه برچسب‌زنی پیشنهادی دو آزمایش و تحلیل انجام شده است.

در تحلیل نخست، تعداد نمایندگان مؤثر به‌ازای تعداد تصاویر آموزشی مختلف مورد بررسی قرار گرفته است. هدف این تحلیل، بررسی رشد تعداد نماینده‌های مناسب به‌ازای تعداد داده‌های آموزشی مختلف است. برای این منظور، به‌ازای تعداد مختلف تصاویر آموزشی (از محدودهٔ ۲۰۰۰ تا ۱۶۰۰۰ تصویر) که به‌صورت تصادفی از کل مجموعه آموزشی نمونه‌برداری شده‌اند، تعداد نماینده‌های مناسب و مؤثر برای سامانه برچسب‌زنی روش پیشنهادی به‌دست آمده‌اند. شایان ذکر است که تعداد موثر نماینده برای یک مجموعه تصاویر، بدین گونه محاسبه می‌شود: با افزایش تعداد نماینده‌ها در جایی که رشد معیار F1 سامانه کند شود (معیار F1 به حد اشباع برسد)، به‌عنوان تعداد نماینده مؤثر درنظر گرفته می‌شود. با این رویکرد، حداقل تعداد نماینده با F1 بالا به‌دست می‌آید.

طبق شکل (۱۰-الف) هر چه تعداد تصاویر آموزشی افزایش می‌یابد، تعداد نمایندگان مؤثر باید افزایش یابد اما این افزایش تا حدود ۶۰۰۰ یا ۷۰۰۰ نماینده ادامه دارد و بعد از آن روند ثابتی را پیش می‌گیرد. بر اساس این نمودار، رشد تعداد نماینده‌ها در مقایسه با رشد تعداد تصاویر آموزشی بسیار کم است که مشخص می‌کند تعداد نماینده‌ها لزوماً به صورت خطی بر اساس تعداد تصاویر افزایش نمی‌یابد.

در تحلیل دوم، زمان اجرای برچسب‌زنی به‌ازای تعداد تصاویر آموزشی مختلف مورد بررسی قرار گرفته است. در این تحلیل، مشابه تحلیل قبلی تعداد تصاویر آموزشی تغییر داده شده‌اند و همان تعداد نمایندهٔ مؤثر در شکل (۱۰-الف) مورد استفاده قرار گرفته‌اند. سپس، برچسب‌زنی تصاویر آزمون بر اساس این نماینده‌ها انجام می‌شود و میانگین زمان برچسب‌زنی محاسبه می‌شود. مطابق نتایج در شکل (۱۰-ب)، با افزایش تعداد تصاویر آموزشی، زمان اجرای برچسب‌زنی تا حدی افزایش و بعد از آن به حالت اشباع می‌رسد. البته ذکر این نکته نیز مهم است که زمان برچسب‌زنی بین ۰/۵ تا ۲/۵ میلی ثانیه متغیر است که مقدار بسیار ناچیزی است.



(شکل-۱۰): نمودار (الف) میزان مقیاس‌پذیری تعداد نمایندگان مؤثر به‌ازای تعداد تصاویر آموزشی است. نمودار (ب) زمان اجرای برچسب‌زنی به ازای یک تصویر ورودی بر حسب تعداد تصاویر آموزشی را نشان می‌دهد. هر دو نمودار بر روی مجموعه‌دادهٔ IAPRTC12 هستند.

(Figure-10): First row represents scalability of number of Training data for different number of prototypes. Second row represents scalability of Annotating input image for different number of prototypes. All plots corresponds to a IAPRTC12 database.

[DOI: 10.52547/jsdp.18.4.49]

از آنجا که تنها افزایش سرعت برچسب‌زنی سبب مقیاس‌پذیرشدن سامانه نخواهد شد، در مرحله بعدی با استفاده از سخت افزار یادشده در ابتدای بخش چهارم و در محیط کدنویسی MATLAB، مرحله آزمون الگوریتم پیشنهادی را به‌صورت پردازش توزیع‌شده شبیه‌سازی کردیم. در این شبیه‌سازی به دنبال الگوریتمی توزیع شده هستیم که متوسط زمان اجرای برچسب‌زنی را کاهش داده و از طرفی دقت سامانه برچسب‌زنی هم‌چنان حفظ شود.

با انجام موازی‌سازی و در نظر گرفتن تعداد ۵ خوشه جهت برچسب‌زنی با ۵ هزار نماینده، به‌طور متوسط در هر خوشه محاسباتی هزار نماینده قرار می‌گیرد. در هر خوشه، وزن میان نمایندگان داخل آن خوشه با تصویر ورودی به همراه ترکیب خطی وزن‌های نمایندگان و برچسب‌هایشان برای به دست‌آوردن درجه تعلق برچسب‌ها به تصویر ورودی محاسبه می‌شود؛ سپس نتایج درجه امتیاز برچسب‌های مربوط به هر خوشه بهم الحاق و تجمیع می‌شوند. پس از محاسبه زمان اجرای برچسب‌زنی در این حالت، به طور متوسط به ازای هر تصویر ورودی حدود ۱/۵۹ میلی‌ثانیه است. در حالت قبلی (توزیع نشده) متوسط زمان برچسب‌زنی حدود ۲/۳ میلی‌ثانیه است. بنابراین، در حالت توزیع شده ٪۶۹ زمان اولیه مورد نیاز است و از طرفی دقت سامانه حفظ خواهد شد.

در این بخش به بررسی دو مورد مهم، مقیاس‌پذیری روش پیشنهادی به‌ازای تعداد تصاویر آموزشی مختلف و مقیاس‌پذیر کردن روش پیشنهادی با استفاده از امکانات محاسباتی جهت کار بر روی مجموعه آموزشی با مقیاس بزرگ پرداخته شده است.

۴-۸- نمونه تصاویر ورودی برچسب‌زده شده

در این بخش برچسب‌های پیش‌بینی شده توسط سامانهٔ پیشنهادی به یک تصویر ورودی دلخواه بررسی شده است؛ که در شکل (۱۱) آمده است. هم‌چنان برچسب‌های واقعی نیز آورده شده است. رنگ سبز به معنی پیش‌بینی درست سامانه است و رنگ قرمز به معنای پیش‌بینی اشتباه است. در این بخش دو نکته حائز اهمیت است. یک تصاویر به‌صورت تصادفی بررسی شده اما سامانه پیشنهادی به خوبی برچسب‌زنی کرده است. نکته دوم، با بررسی تصاویر می‌توان دریافت که برخی برچسب‌های پیش‌بینی شده توسط سامانه پیشنهادی غلط نسبت داده شده‌اند، اما نامرتبط با محتوای بصری تصویر ورودی نیستند. تصویر (ب)، برچسب "sky" و "Shadow" را نادرست درنظر گرفته است اما تصویر دارای چنین کلمات معنایی هست.

تصویر ورودی			
برچسب پیش بینی شده	Girl, house	Door, stone, white	Forest,palm,river,tree
برچسب واقعی	Girl, house, pullover, terrain, waistcoat	Door, shadow, sky, stone, white	Mountain,river,tree
	(الف)	(ب)	(ج)

(شکل-۱۱): برچسب‌زنی تصاویر، مقایسهٔ برچسب‌های حدس زده شده با حقیقی. شکل (الف) و از مجموعه‌دادهٔ IAPRTC12 و دو شکل دیگر (ب) و (ج) از مجموعه‌دادهٔ مستخرج از ESP Game

(Figure-11): Predicted and ground-truth tags for some images of two datasets. Image (a) is selected from IAPR-TC12 dataset. Images (b-c) are selected from ESP-Game dataset.

[DOI: 10.52547/jsdp.18.4.49]

۵- بحث و نتیجه‌گیری

در روش‌های مرسوم برچسب‌زنی خودکار تصاویر جهت نسبت‌دهی برچسب به تصویر آزمون ورودی، از کل پایگاه داده آموزشی استفاده می‌شود. روش پیشنهادی خلاصه‌سازی پایگاه داده، نمایندگان به همراه برچسب‌های معنایی آموزش داده شده آن‌ها به جای استفاده از تمام تصاویر آموزشی ارائه شده است. پس، این مجموعه به جای پایگاه داده اصلی در فرایند برچسب‌زنی تصویر استفاده می‌شود.

جهت ارزیابی این خلاصه‌سازی پایگاه داده، روش پیشنهادی برچسب‌زنی مبتنی بر نزدیک‌ترین همسایگی ارائه شده است که مقایسه‌ای بین یکی از بهترین روش برچسب‌زنی مبتنی بر جستجو یا همان روش 2PKNN (پایه) و روش پیشنهادی انجام گیرد. با بررسی و آزمایش‌های مختلف، خلاصه‌سازی پایگاه داده به نماینده، مجموعه‌داده آموزشی را به ٪۲۲/۶ از کل تصاویر آموزشی در مجموعه‌داده IAPRTC12 و ٪۲۸ از کل تصاویر آموزشی در مجموعه‌داده ESP Game، کاهش داده است و کارایی روش پیشنهادی بهتر یا نزدیک به روش پایه برحسب سه معیار مطرح دقت، یادآوری و F1 است. از دلایل موفقیت این روش، نسبت‌دهی پویای برچسب به تصویر ورودی به جای نسبت‌دهی با تعدادی ثابت برچسب است. از طرفی با جایگزین کردن نمایندگان به جای مجموعه پایگاه داده آموزشی، زمان اجرای برچسب‌زنی کاهش قابل توجهی داشته است، در عین حال تا حد ممکن کارایی سیستم برچسب‌زنی قابل مقایسه با روش‌های پیشین است.

از مزیت‌های اساسی این روش به غیر از استفاده در سیستم‌های با مقیاس بزرگ، می‌توان به کنار گذاشتن کل پایگاه داده آموزشی و استفاده از پایگاه داده آموزشی جایگزین و کوچک‌تر اشاره کرد.

یکی از مسائل مهم برای مقیاس‌پذیرکردن سامانه‌های برچسب‌زنی، حافظه مصرفی مناسب در طراحی آن است. در کارهای آینده می‌توان رویکرد تقریبی یا محلی را برای آموزش برچسب‌های معنایی اتخاذ نمود؛ به طوری‌که به جای آموزش کل نمایندگان به‌صورت همزمان، کل مجموعه به مؤلفه‌های کوچک تقسیم و عملیات انتشار برچسب انجام گیرد. یکی دیگر از مسائل مهم در برچسب‌زنی، عدم توازن کلاس در مجموعه دادگان برچسب‌زنی است. حتماً رویکردی با بررسی انتشار برچسب‌های کم‌یاب نیز درنظر گرفته شود. به عبارتی دیگر با کنار گذاشتن کل مجموعه آموزشی ممکن است اطلاعاتی از تصاویر آموزشی حاوی برچسب‌های کمیاب از دست بروند. یکی از روش‌های مفید می‌تواند در نظر گرفتن توأم تصاویر آموزشی و نمایندگان باشد. به دلیل محدودیت‌های سخت‌افزای اجرای روش پیشنهادی بر روی مجموعه دادگان با مقیاس بزرگ ممکن نبود؛ در صورت اجرا بر روی چنین مجموعه‌دادگانی به نتایج قابل استنادتری دست خواهیم یافت.

سپاس و قدردانی

این پژوهش با حمایت مالی دانشگاه تربیت دبیر شهید رجایی طبق قرارداد شماره ۳۰۴۶۸ مورخ ۹۸/۱۲/۱۴ انجام گردیده است.

۶- مراجع 6- References

[1] V. N. Murthy, E. F. Can, and R. Manmatha, "A hybrid model for automatic image annotation," *in Proceedings of International Conference on Multimedia Retrieval*, pp. 355-369, 2014.

[2] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," *in Computer Vision and Pattern Recognition (CVPR)*, 2004.

[3] P. Ji, X. Gao, and X. Hu, "Automatic image annotation by combining generative and discriminant models," *Neurocomputing*, 2016.

[4] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo, "A cross-media model for automatic image annotation," *in Proceedings of International Conference on Multimedia Retrieval*, 2014, pp. 73.

[5] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003, pp. 119-126.

[6] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, pp. 1075-1088, 2003.

[7] A. Makadia and V. Pavlovic, "Baselines for image annotation." *International Journal of Computer Vision*, pp. 88-105, 2010.

[8] Wang, J., Yang, J., Lv, F., Huang, T., "Locality-constrained linear coding for image classification," 2010.

[DOI: 10.52547/jsdp.18.4.49]

[9] M. M. Kashani and S. H. Amiri, "Leveraging deep learning representation for search-based image annotation," *in Artificial Intelligence and Signal Processing Conference (AISP)*, 2017, pp. 156-161.

[10] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," *in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 603-606.

[11] X. Li, T. Uricchio, L. Ballan, M. Bertini, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval." *ACM Computing Surveys (CSUR)*, 2016, 49(1): 14.

[12] Q. Cheng, Q. Zhang, P. Fu, C. Tu, S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognition*, pp. 242-259, 2018.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J*ournal of machine Learning research*, vol. 3, pp. 993-1022, 2003.

[14] F. Monay and D. Gatica-Perez, "PLSA-based image auto-annotation: constraining the latent space," *in Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 348-351.

[15] A. Llorente, R. Manmatha, S. Ruger, Image retrieval using markov random Fields and global image features, *in Proceedings of the ACM International,Conference on Image and Video Retrieval, ACM*, 2010, pp. 243-250.

[16] Y. Xiang, X. Zhou, T.-S. Chua, C.-W. Ngo, A revisit of generative model for automatic image annotation using markov random _elds, i*n Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE Conference on, IEEE*, 2009, pp. 1153-1160.

[17] I. Dimitrovski, D. Kocev, S. Loskovska, S. D_zeroski, Hierarchical annotation of medical images, *Pattern Recognition* 44 (10-11), pp. 2436-2449, 2011.

[18] J. Wang and J. Hu, Multi-label image annotation via maximum consistency, *in Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE*, 2010, pp. 2337-2340.

[19] H.Wang, H. Huang, C. Ding, Image annotation using the bi-relational graph of images and semantic labels, *in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on IEEE*, 2011, pp. 793-800.

[20] Z. Lin, G. Ding, M. Hu, J. Wang, X. Ye, Image tag completion via image-specific and tag-specific linear sparse reconstructions, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1618-1625.

[21] L. Wu, R. Jin, A. K. Jain, Tag Completion for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), (2013), pp. 716-727.

[22] Z. Qin, C.-G. Li, H. Zhang, J. Guo, Improving tag matrix completion for image annotation and retrieval, *in Visual Communications and Image Processing (VCIP)*, IEEE, 2015, pp. 1-4.

[23] X.-Y. Jing, F. Wu, Z. Li, R. Hu, D. Zhang, Multi-label dictionary learning for image annotation, *IEEE Transactions on Image Processing* 25 (6) (2016),2712-2725.

[24] Y. Hou, Z. Lin, Image tag completion and refinement by subspace clustering and matrix completion, *in Visual Communications and Image Processing(VCIP)*, 2015, IEEE, 2015, pp. 1-4.

[25] Z. Lin, G. Ding, M. Hu, Y. Lin, S. S. Ge, Image tag completion via dual-view linear sparse reconstructions, *Computer Vision and Image Understanding*, 124 (2014) 42-60

[26] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, *Journal of Machine Learning Research*, 10, pp. 207-244, 2009.

[27] E. P. Xing, M. I. Jordan, S. J. Russell, A. Y. Ng, Distance metric learning with application to clustering with side-information, in *Advances in neural information processing systems*, pp. 521-528, 2003.

[28] S. C. Hoi, W. Liu, M. R. Lyu, W.-Y. Ma, Learning distance metrics with contextual constraints for image retrieval, *in Computer vision and pattern recognition, IEEE computer society conference*, Vol. 2, 2006, pp. 2072-2078.

[29] Y. Verma and & C. V. Jawahar, Image annotation by propagating labels from semantic neighbourhoods. *International Journal of Computer Vision*, 2017, *121*. 1., pp. 126-148.

[30] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in 2009 IEEE 12th international conference on computer vision, 2009, pp. 309-316.

[31] L. Wu, S. C. Hoi, R. Jin, J. Zhu, N. Yu, Distance metric learning from uncertain side information with application to automated photo tagging, *in Proceedings of the 17th ACM international conference on Multimedia,*, 2009, pp. 135-144.

[32] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning a Mahalanobis metric from equivalence constraints, *Journal of*

[DOI: 10.52547/jsdp.18.4.49]

Machine Learning Research, 6, pp. 937-965, Jun 2005.

[33] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, C. Sun, Semantic regularisation for recurrent image annotation, *in Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference, 2017, pp. 4160-4168.

[34] J. Johnson, L. Ballan, L. Fei-Fei, Love the neighbors: Image annotation by exploiting image metadata, *in Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4624-4632.

[35] H.-F. Yu, P. Jain, P. Kar, I. Dhillon, Large-scale multi-label learning with missing labels, i*n International conference on machine learning*, 2014, pp. 593-601.

[36] Y. Verma, C. Jawahar, Exploring svm for image annotation in presence of confusing labels, *in BMVC*, 2013, pp. 1-25.

[37] B. Hariharan, L. Zelnik-Manor, M. Varma, S. Vishwanathan, Large scale max-margin multi-label classification with priors, *in Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Citeseer, 2010, pp. 423-430.

[38] Y. Li, Y. Song, J. Luo, Improving pairwise ranking for multi-label image classification, *in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3617-3625.

[39] T. Lan, G. Mori, A max-margin riffled independence model for image tag ranking, *in IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013, pp. 3103-3110.

[40] Y. Yang, W. Zhang, and Y. Xie, "Image automatic annotation via multi-view deep representation," *Journal of Visual Communication and Image Representation*, vol. 33, 2015, pp. 368-377.

[41] H. K. Shooroki, M. A. Z. Chahooki, Selection of effective training instances for scalable automatic image annotation, *Multimedia Tools and Applications*, 2017, 76 (7) (2017), pp. 9643-9666.

[42] S. H. Amiri and M. Jamzad. "Leveraging multi-modal fusion for graph-based image annotation.", *Journal of Visual Communication and Image Representation*, 2018, 55, pp. 816-828.

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

[44] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 184-191.

[45] Sun, Y., Liu, Q., Tang, J., Tao, D., "Learning discriminative dictionary for group sparse representation." *IEEE transactions on image processing*, 2014, 23(9): 3816-3828.

[46] XC. Deng, X. Liu, Y. Mu, J. Li, Large-scale multi-task image labeling with adaptive relevance discovery and feature hashing, *Signal Processing* 112 , 2015, pp. 137-145.

[47] J. Wang, G. Li, A multi-modal hashing learning framework for automatic image annotation, *in IEEE Second International Conference on Data Science in Cyberspace (DSC)*, IEEE, 2017, pp. 14-21.

[48] Wang, Changhu, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. "Multi-label sparse coding for automatic image annotation." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 1643-1650.

[49] Q. Zhang and B. Li, 2015. Dictionary learning in visual computing. *Synthesis Lectures on Image, Video, & Multimedia Processing*, *8*(2), pp.1-151.

[50] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 55-67, 2008.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition,", *arXiv preprint arXiv:1409.1556*, 2014.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[53] G. Huang and Z. Liu, "Densely connected convolutional networks," *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3.

محیا محمّدی کاشانی تحصیلات خود را در مقطع کارشناسی رشتهٔ مهندسی کامپیوتر در دانشگاه تربیت دبیر شهید رجایی در سال ۱۳۹۴ به پایان رساند. وی هم اکنون دانشجوی مقطع کاشناسی ارشد (هوش مصنوعی) در دانشگاه تربیت دبیر شهید رجایی است. موضوعات مورد علاقه ایشان برچسب‌زنی خودکار تصاویر، شناسایی الگو و پردازش تصویر است. نشانی رایانامه ایشان عبارت است از:

mahya.mkashani@sru.ac.ir

[DOI: 10.52547/jsdp.18.4.49]

سیّد حمید امیری تحصیلات خود را در مقطع دکتری رشته‌ی مهندسی کامپیوتر (هوش مصنوعی) در سال ۱۳۹۳ به پایان رساند. وی هم اکنون استادیار دانشکده مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی است. موضوعات مورد علاقه ایشان بینایی ماشین، شناسائی الگو و برچسب‌زنی خودکار و بازیابی تصویر است.

نشانی رایانامه ایشان عبارت است از:

s.hamidamiri@sru.ac.ir

[DOI: 10.52547/jsdp.18.4.49]