مروری نقادانه بر روش‌های بازیابی محتوامحور و معناگرای تصاویر

محمدمهدی حاجی اسمعیلی[۱]، غلامعلی منتظر*[۲]
دانشجوی دکترای فناوری اطلاعات دانشگاه تربیت مدرس، تهران، ایران[۱]
استاد گروه مهندسی فناوری اطلاعات دانشگاه تربیت مدرس، تهران، ایران[۲*]

چکیده

تعداد، تنوع و پیچیدگی محتوای تصویری در دنیای رقمی به‌سرعت در حال افزایش است و این موضوع نیاز به طراحی و پیاده‌سازی سامانه‌های جویش و بازیابی محتوای تصویری را بسیار محسوس کرده‌است؛ در حال حاضر با مقیاس عظیمی از داده‌های تصویری در فضای وب روبه‌روهستیم که راه‌کارهای معمولِ مبتنی بر فراداده‌های دستی و انسانی پاسخ‌گوی تنوع و تعداد بسیار زیاد آن‌ها نیست. حجم عظیم داده‌های تولیدی در محیط وب، بدون راه‌کاری با دقت و سرعت بالا در درک و بازیابی آن‌ها، به آرشیوهای ابدی رقمی خواهند پیوست و هرگز دوباره پیدا نخواهند شد. در سال‌های اخیر تلاش‌های بسیاری برای بازیابی این تصاویر، به‌ویژه در مبحث «بازیابی محتوامحور» (CBIR) و «بازیابی معناگرای» (SIR) تصویر شده‌است. سامانه‌های بازیابی محتوامحور و معناگرای تصویر، توانایی جست‌وجو و بازیابی تصاویر بر اساس محتوای درونی و معانی سطح بالای انسانی را دارد، نه فراداده‌هایی که ممکن است، همراه با آن ثبت شده باشند. این مقاله، مروری جامع بر آخرین پیشرفت‌ها در زمینهٔ بازیابی محتوامحور تصاویر در سال‌های اخیر ارائه کرده و تلاش دارد با رویکردی نقادانه، نقاط مثبت و منفی هر حوزهٔ پژوهشی مطرح در مبحث بازیابی محتوامحور را بیان کند و نمایی کلی از چهارچوب این فرایند و پیشرفت‌های این حوزه ارائه دهد که شامل زمینه‌هایی همچون پیش‌پردازش تصویر، استخراج و تعبیهٔ ویژگی‌ها (Feature Embedding)، یادگیری ماشینی، مجموعه‌داده‌های مطرح در این حوزه، تطبیق شباهت و ارزیابی عملکرد است؛ درنهایت، رویکردهای پژوهشیِ اصیل، چالش‌ها و پیشنهادهایی برای پیشرفت بهتر پژوهش‌ها در این حوزه ارائه شده‌است.

واژگان کلیدی: بازیابی محتوامحور تصاویر، پردازش تصویر، بینایی ماشین، یادگیری ماشینی، یادگیری ژرف، شکاف معنایی.

a Critical Survey on Content-Based & Semantic Image Retrieval – Abstract

Mohammad Mahdi Haji-Esmaeili[1], Gholamali Montazer*[2]
PHD Student Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran[1]
Professor Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran*[2]

Abstract

The rapid increase in the volume, diversity, and complexity of visual content in the digital world has made the need for designing and implementing visual content search and retrieval systems highly evident. Currently, we are facing a massive scale of visual data on the web, for which the conventional approaches based on manual and human-generated metadata are not sufficient to handle the diversity and sheer volume. The enormous volume of data generated on the web, without a high-accuracy and high-speed solution for understanding and retrieving it, will join the digital archives forever and never be found again. Recently, there have been significant efforts for retrieving these images, particularly in the fields of Content-Based Image Retrieval (CBIR) and Semantic Image Retrieval (SIR). Content-based and semantic image retrieval systems have the capability to search and retrieve images based on their internal content and high-level human-understandable semantics, rather than just the metadata that may be associated with them.

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۳/۴/۲۳ • تاریخ پذیرش: ۱۴۰۳/۱۲/۲۵ • تاریخ انتشار: ۱۴۰۴/۳/۲۸ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.22.1.113]

This paper provides a comprehensive review of the latest advancements in the field of content-based image retrieval in recent years. It aims to critically discuss the strengths and weaknesses of each research area in content-based retrieval, and provide an overall framework of this process and the progress made in areas such as image preprocessing, feature extraction and embedding, machine learning, benchmark datasets, similarity matching, and performance evaluation. Finally, the paper presents novel research approaches, challenges, and suggestions for better advancing research in this field.

The sections of the paper are organized as follows: After the introduction, Section 2 describes the components of a CBIR system framework, and with a cursory look at classical and traditional methods, it will delve into the workings of modern approaches and their associated challenges. In Section 3, we will provide an overview of the concept of "relevance feedback" and explain the need for this method to enhance the retrieval performance in CBIR systems, followed by an introduction to the prominent solutions in this domain. Finally, in Section 4, we will present a review of the image datasets commonly used in the field of content-based image retrieval, along with a discussion of their characteristics.

IGiven the recent advancements in the field of computer vision and image processing, especially in the area of "image-text relationship" and how to integrate the two to improve retrieval performance, the focus of a large part of this study has been on the solutions in this area and the performance of the prominent methods.

The current main research in this field is monopolized by large companies and organizations with access to vast financial resources, which has slowed down the progress of research and academic work in this field. These companies, with access to unimaginable data and financial resources, have trained well-known and sometimes unknown models on a very large scale (billions of images and texts), and after the training is complete, they have placed the final model in various web services without publishing the details of the research conducted. The important point is that the scale law applies in this field, and any entity that has more access to computational and storage resources will be able to train better and more accurate models, which has made it less possible for small research units and universities to enter this field and wait for the publication of research by the aforementioned organizations and companies. There is a dire need to introduce effective solutions in this field that require limited resources and are capable of achieving high accuracy and competitiveness with the massive models, with a fraction of the budget required to train them. This has happened in the field of large language models, and after two years, multiple research groups have been able to achieve the accuracy of the Chat-GPT4 language model from OpenAI and with the ability to run on home devices, and it is necessary for research in this field to shift from focusing on achieving accuracy with greater scale to focusing on achieving accuracy with lower cost, otherwise this field will remain in the monopoly of companies focused on greater profits.

Keywords: Content-Based Image Retrieval, Image Processing, Computer Vision, Machine Learning, Deep Learning, Semantic Gap.

۱-مقدمه

برای انسان، نگاهی گذرا و کوتاه به یک تصویر یا منظره کافی است تا او را قادر به تشریح جزئیات پیچیده‌ای از آن تصویر کرده و به او امکان جست‌وجو و بازیابی تصاویری شبیه به آن را در مجموعه‌ای بزرگ بدهد. این فرایند ساده برای انسان‌ها، مسئله‌ای حل‌نشده در طول سال‌های گذشته برای ماشین‌ها محسوب می‌شود[۱]. برای جست‌وجو و بازیابی یک تصویرِ مناسب، نیاز به رویکرد و ابزاری فراتر از روش‌هایی همچون «دسته‌بندی و شناسایی اشیای درون تصویر»[1] یا «شناسایی مکان اشیا در تصویر»[2] است؛ بدون چنین ابزاری، تنها چیزی که از تصاویر به‌دست می‌آوریم، مجموعه‌ای از اطلاعات پردازش‌شده و خام است که به‌خودی‌خود قادر به پُرکردن خلأ بین درک معنایی و سطح بالای انسان از یک تصویر و شناسایی سطحِ پایینِ ماشین از آن نیست.

می‌توان سه درجهٔ مختلف از سختی برای جست‌وجو و بازیابی تصاویری شبیه به تصویر مدنظر کاربر درنظر گرفت[۲]:

۱. سطح نخست: جست‌وجو و بازیابی بر اساس ویژگی‌های اصلی تصویری همچون رنگ، شکل، بافت یا مکان فضایی عناصر درون تصویر. پرس‌وجوی عمومی این سطح، بیشتر به شکل «عکس‌هایی شبیه این عکس را پیدا کن!» است.

۲. سطح دوم: جست‌وجو و بازیابی اشیایی از یک نوع خاص (که توسط کاربر مشخص می‌شود) که ممکن است با استنتاج منطقی همراه باشد؛ برای مثال «تصویر یک زرافه را جست‌وجو کن!».

۳. سطح سوم: جست‌وجو و بازیابی بر مبنای ویژگی‌های انتزاعی سطح بالا که همراه با استنتاج گاهی قوی در مورد ساختار تصویر و روابط موجودیت‌های درونی آن است؛ برای مثال «تصویر کودکانی در حال بازی با یک گربه در پارک» از جملهٔ این پرس‌وجوهای سطح بالا است.

[1] Object Classification
[2] Object Detection

[DOI: 10.61186/jsdp.22.1.113]

(الف)

(ب)

(ج)

(شکل-۱): تصاویری با مفاهیم معنایی سطح بالا
(Figure-1): Images with high-level semantic meanings

سطوح دوم و سوم در کنار هم با نام «بازیابی معناگرای تصویر»[1] شناخته می‌شوند و شکافی که بین آن دو با سطح نخست به وجود می‌آید، با نام «شکاف معنایی»[2] معرفی می‌شود. به عبارت دیگر، اختلاف عمیق بین قدرت تشریحی ضعیف ویژگی‌های سطح پایین عکس و درخواست‌های گاهی غنی و پیچیدهٔ سطح بالای کاربر، به‌عنوان «شکاف معنایی» شناخته می‌شود. از سال ۲۰۱۲ میلادی با رشد پژوهش‌ها در حوزهٔ پردازش تصویر و یادگیری ماشینی ژرف[3]، از پیچیدگی مسائل سطح دوم (که بیشتر تمرکز را بر بازیابی تصاویر حاوی یک شیء خاص قرار داده بودند) کاسته شده‌است[۳]؛ اما مسائل سطح سوم هنوز مسئله‌ای حل‌نشده در این حوزه محسوب می‌شوند. این موضوع به عواملی همچون پیچیدگی استنتاج دربارهٔ روابط بین اشیا و نیاز به طراحی رابطی انسان‌پسند برای ارتباط بین درخواست‌های انسانی و خروجی ماشینی وابسته است. در (شکل-۱) نمونه‌ای از تصاویری نشان داده شده‌است که هم دارای مفاهیم سطح پایین (همچون اشیا) و هم سطح بالا (همچون رُخداد و عمل) هستند:

[1] Semantic Image Retieval
[2] Semantic Gap
[3] Deep Machine Learning

در جدول-۱) نمونه‌ای از تفاوت مفهومی بین «بازیابی محتوامحور»[4] و «بازیابی معناگرا» برای تصاویر (شکل-۱) نمایش داده شده‌است:

(جدول-۱): تفاوت مفهومی بین بازیابی محتوامحور و بازیابی معناگرا
(Table-1): Conceptual difference between Content-Based Image Retrieval and Semantic Image Retrieval

شکل	بازیابی محتوامحور	بازیابی معناگرا
الف	مرد/کودک/زره/سربازان	تعزیه‌خوان در حال آغوش‌کشیدن یک کودک
ب	پیرمرد/جوراب/کیسه	پیرمردِ جوراب‌فروش در کنار خیابان
ج	فرد/فانوس/مزار/قبرستان	مردی در حال زیارت در قبرستان

بحث بازیابی معناگرای تصاویر به چندین حوزهٔ پردازشی و گاهی متفاوت گسترش یافته‌است. تکیه بر روش‌های مجزای بینایی ماشینی برای بازیابی تصاویری که همراه با مفهوم باشند، کافی نیست و از طرفی وابستگی به قالب‌های ازپیش‌تعیین‌شده و ایستا نیز راه‌کاری کاربردی برای شناسایی معنا و مفاهیم نامحدودِ موجود در فضای تصویر محسوب نمی‌شود[۴]. روش‌های ارائه‌شده‌ای که در مرز دانش قرار دارند، هنوز فاصله‌ای قابل توجه با عملکرد انسانی داشته و حتی در بهترین حالت نیز در توصیف تصاویر غیرعادی به‌ندرت دقیق عمل می‌کنند[۵]. همان‌طورکه درک لطیفه‌ها و ضرب‌المثل‌ها و تلاش برای ایجاد آن‌ها به وسیلهٔ ماشین‌ها هنوز مسئله‌ای به‌طور کامل حل‌نشده است، درک «مفاهیم ضمنی»[5] در تصاویر نیز مسئله‌ای باز محسوب می‌شود؛ برای مثال در حال حاضر شاید شناسایی و دسته‌بندی تصاویر، دیگر چندان کار سختی به نظر نرسد (درحالی‌که هنوز این مسئله حل‌نشده است)، اما در قدم بعدی شناسایی و درک ضمنی تصاویری همچون کاریکاتورها و تصاویر طنزمحور، مسئله‌ای است که حتی در پژوهش‌های امروزی نیز به‌صورت جامع به آن‌ها پرداخته نشده‌است؛ به‌دلیل سختی بیش‌ازحد این مسئله، شاهد پژوهش‌های پراکنده‌ای در این حوزه‌ها در طول پنجاه سال اخیر بوده‌ایم[۶,۷,۸]. درک ضمنی محتوا و مفاهیم چه در حوزهٔ متن و چه در حوزهٔ تصویر نیازمند راه‌کاری است که در ابتدا قادر به درک کلی این حوزه‌ها باشد؛ در این دیدگاه، «معنا»، مفاهیم سطح بالای انتزاعی محسوب می‌شوند که در تصاویر می‌توان یافت؛ برای مثال با دیدن تصویری از «مراسم عزاداری» نه‌تنها باید بتوان تشخیص داد که این جمعیت برای چه هدفی دور هم جمع شده‌اند، بلکه باید مفاهیمی همچون «غم و غصه» یا «ازدست‌دادن یک شخص» را نیز برداشت کرد؛ شایان ذکر این‌که این

[4] Content-Based Image Retrieval (CBIR)
[5] Implicit Concepts

[DOI: 10.61186/jsdp.22.1.113]

استنتاج در مغز انسان در کسری از ثانیه انجام می‌شود[۱]؛ درحالی‌که انجام آن به کمک ماشین با فراز و نشیب‌های بسیاری همراه است؛ در این میان باید به نقش مهم «زبان طبیعی»[1] در بازیابی معناگرای تصاویر اشاره کرد؛ چراکه بازیابی چه بر اساس یک پرس‌وجو (به زبان طبیعی) و چه بر اساس یک تصویر اولیه، باید قادر به توصیف «تصویر» و «روابطِ درونِ آن» باشد و آن چیزی که ما را قادر به انجام این کار می‌کند، «زبان طبیعی» است. این موضوع خود به شکل‌گیری رویکردهایی برای توصیف تصاویر به‌کمک زبان طبیعی منجر شده‌است[۵]. پژوهش‌گران این حوزه، بازیابی معناگرا را مرتبط با استخراج معانی از تصاویر به‌کمک زبان طبیعی می‌دانند و تلاش دارند روش‌هایی برای تلفیق دو حوزهٔ تصویر و زبان طبیعی پیدا کنند که قادر به درک تصویر و حتی‌الامکان توصیف آن به بهترین روش به‌کمک زبان طبیعی باشد[۳]؛ برای مثال این مبحث می‌توان به تصاویر درون (شکل-۱) و توصیفات مربوط به هر شکل در جدول-۱) اشاره کرد. هر شکل شامل اطلاعات مورد نیاز برای بازیابی آن از طریق روشی محتوامحور (ستون نخست) و روشی مرتبط با بازیابی معناگرا است. طبیعی است که هر جملهٔ توصیفی در ستون دوم، حجم بیشتری از اطلاعات را نسبت به موجودیت‌های ستون نخست، برملا می‌کند.

(شکل-۲): بازیابی تصاویر مبتنی بر محتوا (و نه معنا)

(Figure-2): Content-Based Image Retrieval (Not Semantic Image Retrieval)

دانستن اینکه چه اشیا و موجودیت‌هایی در هر تصویر وجود دارند، در بازیابی محتوامحور کارایی دارد، اما در بازیابی معناگرا آن‌چنان کاربردی ندارد. تلاش برای بازیابی تصاویری که در آن موجودیت‌های «مرد/ کودک/ زره/ سربازان» وجود دارند، برابر با بازیابی تصاویری شبیه به (شکل-۲) خواهد شد.

برخلاف این نکته، این تصاویر که با درخواست محتوای «مرد/کودک/زره/سربازان» بازیابی شده‌اند، بسیار متفاوت از آن چیزی هستند که در متن توصیفی (شکل-۱) انتظار داشتیم؛ «تعزیه‌خوان در حال به آغوش‌کشیدن یک کودک»؛ این مثال تفاوت و شکاف بین بازیابی محتوامحور و بازیابی معناگرا را به‌خوبی نمایش می‌دهد.

از سوی دیگر، در حال حاضر با مقیاس عظیمی از داده‌های تصویری در فضای وب روبه‌روهستیم که راه‌کارهای معمول مبتنی بر فراداده‌های دستی و انسانی پاسخ‌گوی تنوع و تعداد بسیار زیاد آن‌ها نیست. آمار نشان می‌دهد جویش‌گر گوگل حدود ۳۵ هزار میلیارد صفحهٔ وب را نمایه کرده و پاسخ‌گوی بیش از ۵/۸ میلیارد جست‌وجوی روزانه است. فلیکر، به‌عنوان یک تارنمای اشتراک‌گذاری تصاویر، در حال حاضر شامل بیش از ده میلیارد تصویر است که روزانه ۲۵ میلیون تصویر به آن افزوده می‌شود. شبکهٔ اجتماعی اینستاگرام در هر روز شاهد بارگذاری حدود ۹۵ میلیون تصویر است و تا به امروز حدود ۵۵ میلیارد تصویر در آن به اشتراک گذاشته شده‌است[۹]. تارنمای یوتیوب شاهد بارگذاری روزانه ۷۲۰ هزار ساعت ویدئو است و در یک روز حدود پنجاه میلیارد بازیابی اطلاعات تصویری از آن صورت می‌گیرد[۱۰]. در سطح ملی نیز سازمان‌ها و رسانه‌های تصویری همچون سازمان صداوسیما دارای آرشیوهای عظیمی از اطلاعات تصویری و ویدئویی هستند که جست‌وجو در میان آن‌ها و بررسی منابع اطلاعاتی آن، فرایندی بسیار زمان‌گیر محسوب می‌شود؛ به‌خصوص با درنظرگرفتن این موضوع که در عموم این سازمان‌ها راه‌کارهای بازیابی محتوامحور تنها در سطح اطلاعات متنی و به‌صورتی گسسته و پراکنده استفاده می‌شوند[۱۱]. در این میان، جویش‌گرهای درون‌سازمانی، نیازمند راه‌کاری برای مرتب‌سازی و دسته‌بندی این اطلاعات بر مبنای محتوای درونی آن‌ها هستند. شناسایی تصاویری با دسته‌بندی‌های مشخص همچون تصاویر پویانمایی، سیاسی، طنز، عاطفی، خبری یا خانوادگی یکی از نیازهای این سازمان‌هاست که به دنبال آن، کاربران را قادر به دسترسی سریع و معنادار به اطلاعات بالا می‌کند. از طرفی گستردگی و تنوع منابع تصویری موجب شده‌است که بسیاری از راه‌کارهای «نظارت خانواده»[2] و «پالایش محتوا»[3] در امر کنترل محتوای قابل دسترسی توسط کودکان یا کارمندان سازمان شکست بخورند؛ زیرا عموم سامانه‌های پالایش محتوا، به‌دلیل تنوع بسیار زیاد تصاویر درون وب، عدم

[1] Natural Language

[2] Parental Control

[3] Content Filtering

[DOI: 10.61186/jsdp.22.1.113]

قابلیت شناسایی محتوای سطح بالا در تصاویر و همچنین نبود قابلیت بازیابی معناگرا تصاویری مرتبط با تصویر فعلی، خود را محدود به روش‌های پیش‌پاافتاده‌ای همچون «پالایش نشانی اینترنتی»[1] و یا «پالایش نام‌های دامنه»[2] می‌کنند که در بسیاری از موارد موجب عدم دسترسی مفید به بسیاری از منابع اینترنتی می‌شود.

در طی سال‌های اخیر، روش‌های بازیابی محتوامحور، به لطف پیشرفت‌های حوزهٔ محاسبات گرافی مقیاس بزرگ همچون شبکه‌های عصبی ژرف و شبکه‌های مبدل، رشد چشم‌گیری داشته‌اند. این مقاله مطالعه‌ای جامع و نقادانه دربارهٔ روش‌های مختلف در این حوزه است و تلاش می‌کند با تبیین ساختار کلّی یک سامانهٔ بازیابی محتوامحور تصاویر، روش‌های پیش‌پردازش تصاویر و استخراج ویژگی‌های بصری و متنی از آن‌ها، روش‌های کاهش ابعاد و تعبیه‌سازی بردارهای استخراجی از تصاویر و متون، معیارهای محاسبهٔ شباهت بین تصاویر و چگونگی ارزیابی عملکرد این سامانه‌ها را مورد بررسی قرار دهد.

بخش‌های بعدی مقاله بدین شرح سازماندهی شده‌است: بخش دو اجزای چهارچوب یک سامانهٔ بازیابی محتوامحور را تشریح کرده و با نگاهی گذرا بر روش‌های کلاسیک و قدیمی، به بررسی چگونگی عملکرد روش‌های نوین و مرتبط با معانی سطح بالا و چالش‌های آن‌ها خواهد پرداخت. در بخش سه مروری بر مفهوم «بازخورد ارتباط»[3] داشته و علت نیاز به این روش برای افزایش عملکرد بازیابی در سامانه‌های بازیابی محتوامحور تصویر تشریح و در ادامه به معرفی راه‌کارهای مطرح در این حوزه پرداخته می‌شود. در ادامه و در بخش چهار مروری بر مجموعه‌داده‌های تصویری مورد استفاده در حوزهٔ بازیابی محتوامحور تصاویر ارائه شده و به بیان ویژگی‌های آن‌ها خواهیم پرداخت.

۲- بازیابی محتوامحور تصاویر

در این بخش ساختار کلی یک سامانهٔ بازیابی محتوامحور تصاویر تشریح شده و روش‌های بازیابی اطلاعات تصویری بررسی خواهد شد. بازیابی اطلاعات تصویری، مقولهٔ پژوهشی فعالی از دههٔ ۱۹۷۰ میلادی بوده‌است. این حوزه در دو زمینهٔ زیر بررسی می‌شود[۱۲]:

۱. بازیابی مبتنی بر متون و فراداده[4]

۲. بینایی ماشینی[5]

در حوزهٔ نخست، مسئله به شکل زیر بیان می‌شود: «کاربر توضیحاتی را دربارهٔ هر تصویرِ درون پایگاه داده نگاشته و تصویر به همراه اطلاعات فرادادهٔ آن، در یک پایگاه داده ذخیره می‌شود. جست‌وجو برای تصویر، محدود به جست‌وجو برای فراداده‌ای مشابه با درخواست کاربر خواهد شد»؛ این روش از معروف‌ترین (و در عین حال قدیمی‌ترین) راه‌کارهای بازیابی اطلاعات تصویری است که می‌توان روش ارائه‌شده را از پژوهش‌های مطرح آن دوران دانست. بسیاری از پیشرفت‌های بازیابی اطلاعات از جمله مدل‌سازی داده‌ها[6]، نمایه‌سازی چندبُعدی[7] یا ارزیابی پرس‌وجو[8] در این حوزهٔ پژوهشی شکل‌گرفته‌اند[۱۲].

با وجود استفادهٔ گسترده از این راه‌کار، دو مشکل اصلی در استفادهٔ کارا از آن وجود دارد[۳]:

۱. مشکل افزودن فراداده به تعداد زیادی تصویر

۲. تفاوت درک انسان‌ها از یک تصویر

افزودن فراداده به ده، صد یا چند صد تصویر، شاید سخت به نظر نرسد، اما در میان حجم عظیم اطلاعات که روزانه در حال تبادل، پردازش و ذخیره‌سازی است، توانایی انسان در ثبت فراداده، نزدیک به صفر است؛ درواقع حجم عظیم اطلاعات، استفاده از عامل انسانی را برای معنادهی به آن و تبیین ساختار برای هر تصویر بی‌معنا می‌کند. مشکل دیگر، تفاوت سطح درک معنایی[9] انسان‌ها با یکدیگر است. طبیعی است که اتکای صرف به عامل انسانی با دیدگاه‌های متفاوت، موجب ثبت فراداده‌های مختلف برای تصاویری به طور اتفاقی با شباهتِ زیاد شده و بازیابی آن‌ها را با مشکل مواجه می‌کند.

در اوایل دههٔ ۱۹۹۰ میلادی با رشد حجم ذخیره‌سازی اطلاعات و به‌وجودآمدن مجموعه‌های تصویری بزرگ، بهره‌گیری از روش یادشده بیش‌ازپیش دچار مشکل شد. برای غلبه بر این مشکلات، مفهوم «بازیابی محتوامحور تصویر» ارائه شد که ایدهٔ آن استخراج ویژگی‌ها و مفاهیمی از خود تصویر (همچون رنگ یا بافت) و جست‌وجو بر اساس این ویژگی‌ها به‌جای استفاده از فراداده‌های متنی ثبت‌شده برای تصویر بود. از آن زمان تا به امروز روش‌های بسیاری در این حوزه مطرح و پیاده‌سازی شده و سامانه‌های بازیابی مختلفی (هم پژوهشی و هم تجاری) بر مبنای آن ارائه‌شده‌اند. تمرکز این مقاله بر راه‌کارهای بازیابی تصاویر به کمک محتوای آن است و سعی می‌کند از واردشدن به راه‌کار دیگر، یعنی بازیابی تصاویر بر مبنای فراداده، اجتناب کند.

[1] URL Filtering
[2] DNS Filtering
[3] Relevance Feedback
[4] Meta-data & Text-based Retrieval
[5] Computer Vision
[6] Data Modeling
[7] Multi-Dimensional Indexing
[8] Query Evaluation
[9] Semantic Understanding

[DOI: 10.61186/jsdp.22.1.113]

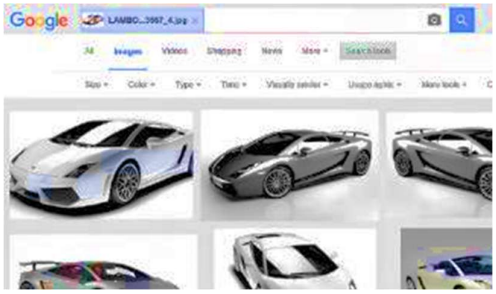
۲-۱-روش‌های مختلف بازیابی اطلاعات تصویری

مهم‌ترین تفاوت بین «بازیابی محتوامحور» و «متن‌محور» تصاویر این است که حضور عامل انسانی بخش جدایی‌ناپذیری از بازیابی متن‌محور محسوب می‌شود. اانسان‌ها از مفاهیم و ویژگی‌های سطح بالا برای توصیف یک تصویر بهره می‌برند؛ درحالی‌که ویژگی‌های استخراج‌شده به وسیلهٔ الگوریتم‌های بینایی ماشینی، بیشتر در سطح پایین هستند؛ برای مثال انسانی با تماشای یک تصویر از افرادی که در حال تحصن هستند، به راحتی مفهومی سطح بالا همچون «جمعیت ناراضی و معترض» در ذهنش تداعی می‌شود؛ درحالی‌که برای ماشین دست‌یابی به چنین مفهومی بسیار سخت است. در حالت کلی، هیچ رابطهٔ مستقیمی بین مفاهیم سطح بالا و ویژگی‌های سطح پایین وجود ندارد[۱۳].

مفاهیم سطح پایین، آن دسته از مشخصه‌هایی هستند که به تشریح اجزای مستقل پرداخته و رویکرد آن‌ها در توصیف شیء، تشریح جزئیات آن و نه کلیات شیء است؛ برای مثال ویژگی‌هایی همچون رنگ یا بافت اجزای درون یک تصویر، بخشی از ویژگی‌های سطح پایین آن محسوب می‌شوند؛ زیرا قادر به تشریح کلی آن تصویر نیستند. در طرف دیگر مفاهیم سطح بالا، آن مشخصه‌هایی هستند که انتزاعی‌تر بوده و تمرکزشان بر کلیات تصویر و یا اجزای جامع‌تر آن است[۱۴]؛

بااینکه الگوریتم‌های متنوع و گاه قدرتمندی در حوزهٔ شناسایی و تشخیص رنگ[۱]، شکل[۲]، بافت[۳] یا مکان فضایی[۴] طراحی و ارائه شده‌اند، بیشتر آن‌ها قادر به مدل‌کردن مفاهیم سطح بالای تصاویر نیستند و این موضوع در هنگام کار با پایگاه‌های دادهٔ تصویری عمومی (که شامل تصاویری از دسته‌بندی‌های مختلف و گاه غیرمرتبط‌اند) ضعف خود را بیش‌ازپیش نشان می‌دهد[۱۵]. آزمایش و پژوهش‌های بسیار بر روی سامانه‌های بازیابی فعلی نشان می‌دهند که ویژگی‌های سطح پایین بیشتر در تبیین و تشریح مفاهیم سطح بالایی که در ذهن انسان وجود دارد، شکست می‌خورند[۱۶].

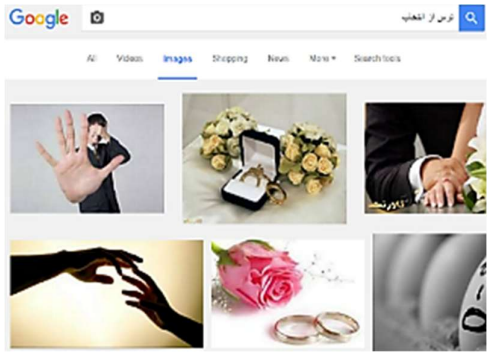
با توجه به سه سطح پرس‌وجویی که در بخشِ مقدمه معرفی شدند[۱۶]، کاربرانی که جست‌وجوهایی در سطح ۱ انجام می‌دهند، بیشتر نیاز به ارائهٔ یک تصویرِ نمونه به سامانهٔ بازیابی محتوامحور دارند تا سامانه به کمک آن تصویر، تصاویر مرتبط با آن را پیدا کند. در این سطح، سامانه تلاش می‌کند تصاویری را بیابد که ویژگی‌های سطح پایین آن‌ها شباهت بیشتری با ویژگی‌های سطح پایینِ تصویرِ نمونه داشته باشد؛ برای مثال در (شکل-۳)، عملکرد سامانهٔ بازیابی تصویر شرکت گوگل (Google Images) بر اساس تصویری نمونه نشان داده شده‌است. گزینه‌هایی که این سامانه در اختیار کاربر قرار می‌دهد همگی جزو ویژگی‌های سطح یک هستند؛ برای مثال رنگ یا اندازه، ویژگی‌های سطح پایینی هستند که برای شناسایی و بازیابی تصاویر غیرمفهومی مناسب است.



(شکل-۳): نمونه‌ای از جست‌وجوی سطح ۱ بر مبنای تصویر نمونه و با خروجی به طور کامل مرتبط با پرس‌وجو

(Figure-3): An instance of Level-1 CBIR query

در (شکل-۳)، تصویر نمونه‌ای از یک ماشین، به سامانهٔ بازیابی محتوامحور گوگل[۵] تحویل داده شده‌است. این سامانه نه تنها قادر به پیداکردن منبع اصلی تصویر شده‌است (تصویر بالا سمت چپ، کامل منطبق با تصویر نمونه است)، بلکه تصاویری مشابه آن را نیز برای کاربر نمایش می‌دهد؛ البته اگر کاربر تصویر نمونه‌ای در اختیار نداشته باشد، آنگاه سامانه نمی‌تواند جست‌وجو را به نحو مطلوبی انجام دهد. در اینجا، بازیابی معناگرای تصویر نیاز کاربر را پوشش می‌دهد؛ چراکه قابلیت‌هایی همچون پرس‌وجو بر مبنای «واژه‌های کلیدی»[۶] و یا «بافت» را در اختیار وی قرار می‌دهد.



(شکل-۴): نمونه‌ای از جست‌وجوی معناگرا با خروجی کم‌وبیش مرتبط با پرس‌وجو

(Figure-4): An instance of a Semantic query

در مثالی دیگر، سامانهٔ بازیابی تصاویر جویش‌گر گوگل در حوزهٔ بازیابی معناگرا تصاویر (سطوح ۲و۳) بسیار ضعیف‌تر از سطح ۱ عمل می‌کند. نمونه‌ای از عملکرد آن در (شکل-۴) نشان داده شده‌است. در این شکل، پرس‌وجویی با مفهوم سطح بالا «ترس از انتخاب» بر روی جویش‌گر گوگل انجام شده‌است. از آنجا که جویش‌گر درک درست و کاملی از مفهوم فوق ندارد،

[1] Color
[2] Shape
[3] Texture
[4] Spatial Location

[5] Google Reverse Image Search
[6] Keywords

[DOI: 10.61186/jsdp.22.1.113]

با تصاویری مواجهه می‌شویم که ارتباط کم‌وبیشی با پرس‌وجوی مورد نظر دارند و در بعضی موارد به‌طور حتی طنزگونه‌اند.

با توجه به این موضوع، نمی‌توان «معنا»ی درون یک تصویر را وابسته به یک و یا حتی چند توضیح به زبان طبیعی دانست؛ برای مثال، «گرفتنِ یک شیء از دست یک کودکِ گریان» را می‌توان همان‌قدر به «حفاظت از کودکان در برابر اشیای خطرناک» و توضیحات و معانی مربوط با آن مرتبط کرد که به «خشونت علیه کودکان». عامل اصلی و تأثیرگذار در برداشت‌های متفاوت انسان‌ها نسبت به یک تصویر، به «عقل سلیم»[1] و «دانش عمومی»[2] و موارد دیگر مانند احساسات لحظه‌ای و سایر موارد مرتبط با بیننده برمی‌گردد[۱۷].

بدون راه‌کار مشخصی برای حل مسئلهٔ «دانش عمومی»، رویکردهای علمی متفاوتی برای درک معنایی تصاویر، توصیف و درنهایت بازیابی آن‌ها ارائه شده‌است.گاه برخی پژوهش‌ها تلاش را بر محدودکردن دامنهٔ تصاویر و حل مسائل کوچک‌تر معطوف کرده‌اند[۱۸]، بعضی بر بحث «برجستگی»[3] و «توجه» به اطلاعات درون تصاویر تمرکز داشته‌اند[۱۹]، پژوهش‌هایی معدود، بر به‌کارگیری مبحث کمابیش جدید «یادگیری تقویتی»[4] و قدم‌های قابل‌توجهی که در سال‌های اخیر برداشته‌است، تأکید دارند[۲۰]، برخی تلاش بر نگاهی دوباره به مفاهیم معنایی و چگونگی بازسازی درک انسان در ماشین داشته‌اند[۲۱]، پاره‌ای از پژوهش‌ها بر درک روابط پیچیدهٔ بین اشیای درون تصاویر شکل گرفته‌اند[۲۲]، مجموعه‌داده‌های بزرگی برای توصیف این روابط پیچیده ایجادشده‌اند[۲۳]، برخی راه‌کارها از تلفیق اطلاعات کمکی و اضافی برای درک بهتر تصاویر بهره برده‌اند[۲۴] و روش‌های متعددی نیز به تلفیق و هم‌جوشی[5] توصیفات یک تصویر و محتوای آن روی آورده‌اند[۲۵]. شاید بتوان تأثیرگذارترین و موفق‌ترین روش برای درک دانش عمومی و عقل سلیم را، معرفی «مدل‌های زبانی بزرگ»[6] دانست. به لطف قابلیت «مقیاس‌پذیری»[7] این مدل‌ها و آموزش آن‌ها بر روی مجموعه‌داده‌های بسیار بزرگ (در مقیاس کل اینترنت)، شاهد یادگیری بسیاری از مفاهیم مرتبط با دانش عمومی و درک اجتماعی به‌وسیلهٔ آن‌ها بوده‌ایم. شاید بتوان مهم‌ترین نماد این عرصه را اوج‌گیری گپ‌بات[8]هایی همچون ChatGPT از شرکت OpenAI یا Gemini از شرکت Google دانست[۲۶].

با وجود تلاش‌های گوناگون و پژوهش‌های متعدد، هنوز نیز توصیف یک‌پارچه و مناسبی از «معنا»ی درون تصاویر ارائه نشده‌است؛ به‌طوری‌که به‌کارگیری این واژه در مباحثی از پردازش تصویر کاربردی گاهی متفاوت با منظور پژوهش‌گران در طیفی دیگر از مباحث این حوزه دارد؛ بااین‌حال، مبحث مشترک در این پژوهش‌ها، تلاش برای نزدیک‌کردن برداشت ماشین از یک تصویر به تجربه‌ای است که انسان از آن دارد.

۲-۲-معماری سامانه‌های بازیابی محتوامحور

در این بخش معماری سامانه‌های بازیابی محتوامحور تصویر را معرفی می‌کنیم. مطابق (شکل-۵)، هر سامانهٔ بازیابی محتوامحور تصویر از چند بخش زیر تشکیل شده‌است[۲۷]:

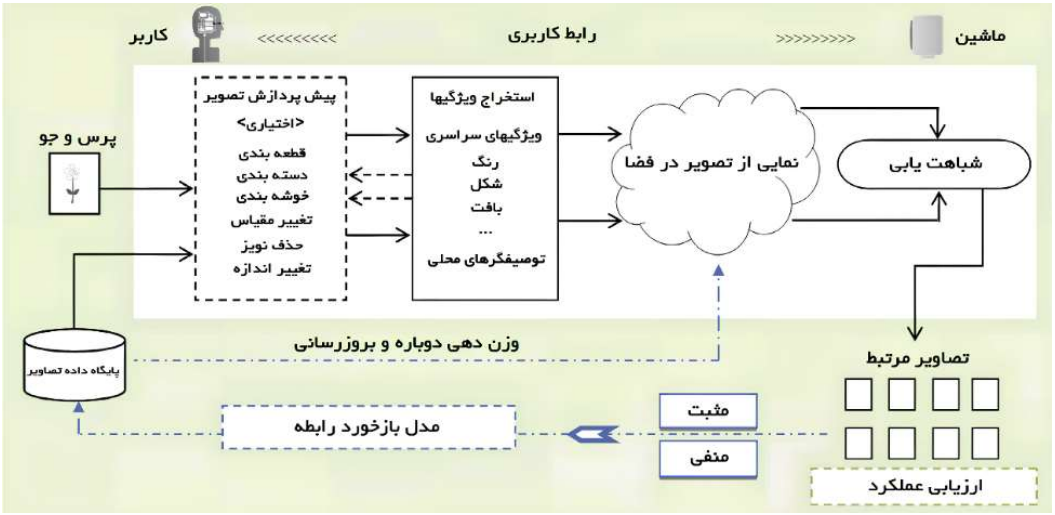
در مرحلهٔ نخست، پرس‌وجویی در سامانه مطرح می‌شود. این پرس‌وجو ممکن است یک نمونه تصویر باشد (درخواست سطح ۱) یا یک پرس‌وجوی متنی (درخواست سطوح ۲ و۳). در مرحلهٔ دوم، با توجه به وضعیت تصویر، پیش‌پردازش‌هایی[9] بر آن صورت می‌گیرد. فعالیت‌های پیش‌پردازشی، مبتنی بر هدف و عملکرد سامانهٔ بازیابی است؛ برای مثال اگر سامانه برای بازیابی تصویر چهره طراحی شده باشد، یکی از پیش‌پردازش‌ها، بریدن تصویر نمونه و حذف اجزای دیگر بدن است تا شناسایی ویژگی‌های چهره با خطای‌کمتری همراه باشد. در مرحلهٔ سوم ویژگی‌های بصری/متنی از درون تصویر استخراج می‌شود و سپس این ویژگی‌ها طبق قاعده‌ای مشخص در فضای داده دسته‌بندی و نگهداری می‌شود. برخی از ویژگی‌های معروف عبارت‌اند از: رنگ، بافت، شکل و توصیف‌گرهای محلی تصویر. برخی روش‌ها برای انجام استخراج ویژگی‌ها، پیش‌پردازش‌های ویژه همچون دسته‌بندی یا پردازش فضایی نیاز دارند و به همین دلیل پیش‌پردازش تصویر ممکن است، قبل یا بعد از استخراج ویژگی‌ها نیز صورت گیرد. در مرحلهٔ پایانی، سامانه برای شباهت‌یابی، فاصلهٔ بین ویژگی‌های استخراج‌شده را با ویژگی‌های ذخیره‌شده در پایگاه داده محاسبه می‌کند تا بتواند نزدیک‌ترین تصاویر به تصویر نمونه را شناسایی و به کاربر ارائه کند. برخی سامانه‌ها به کاربر قابلیت رتبه‌دهی به تصاویر بازگشتی را می‌دهند تا به کمک آن بتوانند بازخورد رابطهٔ تصویر نمونه و تصاویر بازگشتی را محک بزنند[۳].

۲-۲-۱- تعبیه‌سازی[10]

در این مقاله ارجاعات متعددی به مفهوم «تعبیه‌سازی» شده‌است که این مفهوم در این قسمت بیشتر تشریح خواهد شد:

[1] Common Sense
[2] General Knowledge
[3] Saliency
[4] Reinforcement Learning
[5] Fusion
[6] Large Language Models
[7] Scalability
[8] Chatbot
[9] Preprocessing
[10] Embedding

[DOI: 10.61186/jsdp.22.1.113]



(شکل-۵): معماری عمومی سامانه‌های بازیابی محتوامحور تصویر

(Figure-4): The overall architecture of a CBIR system

تعبیه‌سازی به معنای نمایش داده‌ها در قالب بردارهایی در یک فضای برداری کم‌بعد و پیوسته است که اطلاعات کلیدی آن‌ها را به شکلی فشرده و قابل استفاده برای الگوریتم‌های یادگیری ماشین حفظ می‌کند. هدف از این کار، ایجاد بازنمایی‌هایی است که بتوانند شباهت‌ها و تفاوت‌های داده‌ها را بازتاب دهند. تعبیه‌سازی به‌ویژه برای کار با داده‌هایی که در حالت خام آن‌ها برای ماشین قابل درک نیستند (مانند متن یا تصاویر) بسیار مهم است.

در پردازش زبان طبیعی، تعبیه‌سازی واژه‌ها، جملات یا اسناد به بردارهایی در یک فضای عددی، یکی از ابزارهای کلیدی برای درک و پردازش زبان انسانی است. روش‌های مختلف تعبیه‌سازی می‌توانند روابط معنایی و نحوی زبان را استخراج کنند. تعبیه‌سازی در این حوزه می‌تواند شامل تعبیه‌سازی واژه‌ها (Word Embedding) یا جملات (Sentence Embedding) و اسناد (Document Embedding) به فضای برداری جدید باشد.

در بینایی ماشین، تعبیه‌سازی برای نمایش داده‌های تصویری در قالب بردارهای فشرده استفاده می‌شود تا ویژگی‌های کلیدی تصویر (مانند شکل، رنگ، و الگوها) را حفظ کند. تعبیه‌سازی در این حوزه، می‌تواند شامل تعبیه‌سازی تصاویر (Image Embedding) یا یک ویدئو (Video Embedding) به فضای برداری جدید باشد. در ادامه، به بررسی راه‌کارهای استخراج ویژگی استفاده شده در سامانه‌های بازیابی محتوامحور پرداخته و نقاط قوت و ضعف این راه‌کارها را تشریح خواهیم کرد.

۲-۲-۲-بازیابی محتوامحور با ویژگی‌های سطح پایین

الگوی عمومی برای پردازش تصاویر شامل مراحل زیر است[۲۸]:

۱. پردازش‌های رده‌پایین[1]: این دسته فرایندهایی با هدف حذف نوفه[2]، افزایش تباین[3]، آستانه‌گیری[4] و قطعه‌بندی[5] سادهٔ تصویر است.

۲. پردازش‌های میان‌رده[6]: مواردی همچون قطعه‌بندی تصویر به نواحی کوچک‌تر، تشریح نواحی قطعه‌بندی‌شده و اشیای آن و دسته‌بندی اشیای درون تصویر.

۳. پردازش‌های رده‌بالا[7]: انجام پردازش‌های شناختی، مفهومی و معناگرا بر تصویر و تلاش برای شناسایی موارد بالا از آن.

در پردازش‌های رده‌پایین، ورودی و خروجی، هر دو از جنس تصویر است. این موضوع در سطح میانی متفاوت است؛ چراکه ورودی این سطح تصویر و خروجی آن ویژگی‌هایی است که از تصویر استخراج شده‌اند؛ همچون لبه‌ها، مرزهای درون تصویر یا شناسه‌های اشیای درون آن. در مرحلهٔ سوم، نوعی سامانهٔ بازیابی، تصویر و ویژگی‌های آن را که در مراحل نخست و دوم به‌دست‌آمده‌اند، برای انجام پردازش‌های شناختی و معنایی مورد استفاده قرار می‌دهد.

پردازش‌های سطح نخست و دوم، شاهد حجم عظیمی از پژوهش‌ها در سه دههٔ گذشته بوده‌است. نکتهٔ بارز این پژوهش‌ها، که بیشتر با نام «روش‌های کلاسیک» از آن‌ها یاد می‌شود، نزدیکی عملکرد آن‌ها به یکدیگر است؛ بدین معنی که از لحاظ دقت و عملکرد در مسائل دنیای واقعی، عملکرد کمابیش

[1] Low-level Processes
[2] Noise Removal
[3] Contrast
[4] Thresholding
[5] Segmentation
[6] Medium-level Processes
[7] High-level Processes

[DOI: 10.61186/jsdp.22.1.113]

یکسانی دارند. در (جدول-۲) و **Error! Reference source not found.** خلاصه‌ای از انواع این روش‌ها، ویژگی‌ها و محدودیت‌های آن‌ها آورده شده‌است. با توجه به گسترهٔ عظیم روش‌های ارائه‌شده برای مبحث بازیابی تصاویر در طول سال‌ها، به‌طور طبیعی نمی‌توان تمامی این روش‌ها را در جداولی محدود مقایسه کرد؛ بااین‌حال تلاش شده‌است نقطهٔ مشترک بسیاری از روش‌ها در این جدول‌ها خلاصه‌سازی شود. به‌علت اینکه تمرکز این مقاله بر روی روش‌های سطح سوم است، به تشریح کلیات عملکرد روش‌های سطوح نخست و دوم در این جداول بسنده کرده‌ایم.

۲-۲-۳- بازیابی محتوامحور با ویژگی‌های سطح بالا

چنان‌که اشاره شد، حوزهٔ بازیابی محتوامحور تصاویر و پردازش معناگرا دارای نقاط مشترک بسیاری با حوزهٔ «عنوان‌بندی و توصیف تصاویر» است. با توجه به این موضوع نیاز به بررسی گرایش‌های مطرح در این عرصه داریم که روش غالب انجام آن، تلاش برای تلفیق ویژگی‌های تصویر با ویژگی‌های به‌دست‌آمده از متن است؛ به‌طوری که بتوان با یادگیری چگونگی این تلفیق به توصیف مناسبی برای تصاویر دست یافت. با داشتن توصیف مناسبی از هر تصویر می‌توان بازیابی آن‌ها را با دقت بالاتری انجام داد[۲۹].

روش کلی این رویکرد در طول سال‌های اخیر به زیرساختی موفق در توصیف تصاویر به‌کمک زبان طبیعی تبدیل شده‌است. در این روش هدف اصلی، ایجاد ارتباط بین اجزای تصویر و اجزای متن توصیفی آن است و برای نیل به این هدف از تلفیق یک یا چند شبکهٔ عصبی ژرف (در سمت تصاویر) و یک یا چند مؤلفهٔ پردازش متن (در سمت متون) استفاده شده‌است[۳۰]. برای محاسبهٔ ارتباط میان متن و تصویر، لازم است متن به فضایی جدید منتقل شود تا بتوان محاسبات بین ویژگی‌های استخراج‌شدهٔ تصویری و قطعات درون متن را انجام داد. اشاره به این نکته ضروری است که برخلاف مؤلفهٔ تصویری که در آن بیشتر از روش‌های معدود و کمابیش یکسانی برای پردازش تصویر و استخراج ویژگی استفاده می‌شد، برای پردازش متن روش‌های متعدد و مختلفی به‌کار گرفته شده‌است.

در رایج‌ترین روش، ابتدا متنِ توصیفی تصویر به قطعات کوچک‌تری تقسیم می‌شود؛ بدین منظور، از یک درخت تجزیهٔ وابستگی استفاده می‌شود و از هر ارتباط درون این درخت (یالی که بین گره‌ها برقرار است) یک قطعه متن دو واژه‌ای استخراج شده‌است[۳۱]. در مرحلهٔ بعد، باید واژه‌ها به فضایی برداری و قابل محاسبه منتقل شود؛ برای این کار در ابتدا هر واژهٔ درون متن را به‌کمک روش کدبندی وان هات[1] به یک بردار تبدیل کرده و سپس آن‌ها را با یک ضرب ماتریسی (تبدیل خطی) به فضایی جدید می‌برند. پژوهش‌های گسترده‌ای برای پیداکردن یک ماتریس انتقال فضای مناسب برای واژه‌ها صورت گرفته‌است که در بین آن‌ها می‌توان به [۳۲] و یا روش Word2Vec از شرکت گوگل [۳۳] اشاره کرد. ایدهٔ اصلی چنین روش‌هایی، تعبیهٔ واژه‌ها درون یک فضای برداری جدید[2] است؛ به‌طوری‌که واژه‌هایی که از لحاظ معنایی و بافتاری در نزدیکی هم قرار می‌گیرند، در این فضای برداری پیوسته نیز در نزدیکی یکدیگر قرار داشته باشند.

با داشتن قطعات تصویر و متن در فضایی قابل محاسبه، ضربی داخلی از تمامی قطعات تصویر و متن محاسبه و خروجی آن به‌عنوان یک شاخص امتیازدهی بین این دو زوج (قطعهٔ تصویر و قطعهٔ متن) درنظر گرفته می‌شود.

در (شکل-۶) روند کار تا بدین مرحله نمایش داده شده‌است. در این شکل از روش ترازبندی متن و تصویر [۳۱]، سه قطعهٔ تصویر (کل تصویر، سگ، کودک) و پنج قطعهٔ متنی از زوج‌های به هم وابسته داریم. ضرب داخلی تمامی این قطعات برابر پانزده حالت مختلف است. نقاط پررنگ‌تر نشان‌دهندهٔ امتیاز مثبت (پشتیبانی از شباهت بین تصویر و قطعهٔ متن) و نقاط کم‌رنگ‌تر نشان‌دهندهٔ امتیاز منفی (عدم هم‌ترازی بین تصویر و قطعهٔ متن) است؛ برای مثال در ردیف سوم (ضرب داخلی تصویر سگ و پنج قطعهٔ متن) شاهد دو امتیاز مثبت برای متن‌های [black,dog] و [chasing, dog] و سه امتیاز منفی برای دیگر حالات هستیم. این امتیاز، مقداری نسبی است که به‌خودی‌خود معنایی قابل درک ندارد، اما با ایجاد یک تابع هدف می‌توان به امتیاز به‌دست‌آمده و تغییرات آن نسبت به متن یا تصویر معنا داد؛ پس از یادگیری و بهینه‌سازی، مدل نهایی قادر است با گرفتن یک تصویر، بهترین جملهٔ توصیفی به زبان طبیعی را از میان جمله‌هایی که پیش‌تر ندیده، برای این تصویر انتخاب و از طرفی با گرفتن یک متن، بهترین تصویر با قطعاتی هم‌تراز قطعات متن را بازیابی کند. در این مثال AMOD[3] به معنای ویراستار صفت است که برای تغییر معنی یک عبارت اسمی عمل می‌کند، DEP[4] به‌معنای وابسته، یک برچسب کلی است که زمانی استفاده می‌شود که سامانه نتواند رابطهٔ دقیق‌تر وابستگی بین دو واژه را تعیین کند، DOBJ[5] به معنای مفعول مستقیم بوده و عبارت اسمی است که فاعل آن را انجام می‌دهد و DET[6] به معنای معین‌کننده

[1] One-Hot Encoding
[2] Word Embeddings
[3] Adjective Modifier
[4] Dependent
[5] Direct Object
[6] Determiner

[DOI: 10.61186/jsdp.22.1.113]

است و اسم‌ها را با بیان کمیت، تعیین یا ویژگی‌های دیگر اصلاح می‌کنند.

(جدول-۲): مقایسهٔ روش‌های پیش‌پردازش تصاویر سطح یک (پردازش‌های رده پایین)

(Table-2): Comparison between Level-1 Preprocessing methods for CBIR

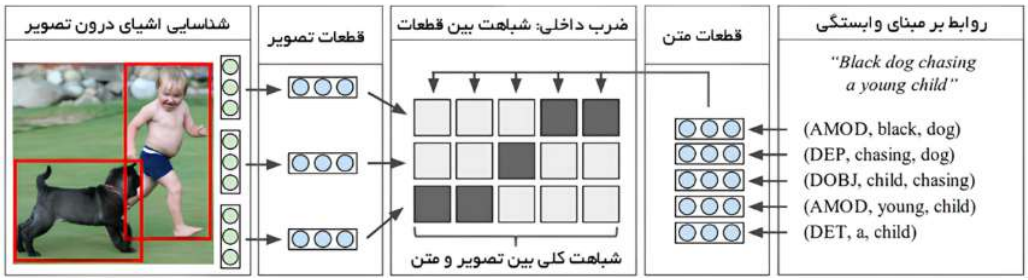
روش	ویژگی‌ها	محدودیت‌ها	نقش در کاهش شکاف معنایی
مبتنی بر گراف [۴۴، ۴۶، ۴۷، ۴۸، ۴۹، ۵۰]	هم خودکار و هم به کمک کاربر روابط بین قطعات را به خوبی مشخص می‌کند	مسئله از نوع NP-HARD نبود کمیتی برای ارزیابی کیفیت عملکرد	افراز یک گراف به چندین زیر گراف که اشیایی معنادار دارند
مبتنی بر لبه [۲۸،۴۴،۴۵]	بر مبنای تغییرات ناگهانی در شدت یا رنگ تصویر	حساس به نوفه کارایی پایین در تصاویر ساده و کم لبه	شبیه به ساختار بصری چشم انسان در قطعه‌بندی تصاویر
مبتنی بر ناحیه [۴۱،۴۲،۴۳]	یک حلقه آن‌قدر تکرار می‌شود تا یک شرط یکنواختی معین به‌دست آید	نیاز به شناسایی اولیهٔ نواحی حساس به معیارهای مختلف شباهت‌یابی	پیداکردن کاندیداهای کیفیت بالا برای اشیا
فازی [۳۸، ۳۹،۴۰]	مقاومت در برابر نوفه جزئیات تصویر را نگه می‌دارد	از لحاظ محاسباتی پرهزینه است	ایجاد ارتباط بین چندین مفهوم با درجه‌ای از عضویت
محاسبات تکاملی [۳۷]	بیشتر از روش‌های ژنتیک برای حل مسائل بهینه‌سازی یا مشکلات در سطح پیکسل استفاده می‌کنند	محدود به دامنه ارزیابی از لحاظ محاسباتی پرهزینه است	ایجاد ویژگی‌های سطح بالا برای اصلاح عملکرد فرایند قطعه‌بندی
هم قطعه‌بندی [۳]	دستیابی به قطعه‌بندی متناسب با قطعات چندین تصویر	تصاویر باید شامل اشیای یکسان باشند	ایجاد معیار شباهت بین چند تصویر
آستانه‌گیری [۳۶]	جداسازی اشیا از پس‌زمینه سرعت بالا و نیاز به حافظه کم	حساس به نوفه، نور محیطی و شدت نور	بر کارایی مراحل بعدی همچون دسته‌بندی و بازیابی تأثیر دارد
مبتنی بر شکل[۳۴، ۳۵]	مدل‌های آماری مدل‌های قابل‌تغییر	مشکل وقتی اشیا بر روی‌هم قرار دارند حساس به انسداد اشیا	شناسایی اشیای موردنظر کاربر

(جدول-۳): مقایسهٔ روش‌های پیش‌پردازش تصاویر سطح دو (پردازش‌های میان‌رده)

(Table-3): Comparison between Level-2 Preprocessing methods for CBIR

روش	ویژگی‌ها	محدودیت‌ها	نقش در کاهش شکاف معنایی
SIFT [۵۹ ،۳۵]	شباهت‌سنجی قدرتمند حتی با وجود نوفه و تغییرات آفین و تغییر در روشنایی و شدت نور نامتغیر نسبت به تغییر در مقیاس یا دوران	شباهت‌سنجی در ابعاد بالا نیاز به رمزگذاری در آرایه‌های محدود	استخراج ویژگی‌های پایا و مقاوم به تغییرات در تصویر اولیه
SURF [۵۸]	مبتنی بر ماتریکس هسین (Hessian) وابسته به تصاویر انتگرالی برای کاهش محاسبات مورد نیاز	ضعف در تخمین میزان دوران در نقاط کلیدی ضعف در شناسایی میزان دوران	همچون SIFT
Local Patterns [۵۷ ،۵۶ ،۴۴،۵۵]	مقاوم به تغییرات پیوستهٔ شدت روشنایی پیچیدگی محاسباتی کم	حساسیت به نوفه در نواحی پیوستهٔ درون تصویر	استخراج ویژگی‌های پیوسته و ناپیوسته از بافت‌های درون تصویر
HOG [۵۴]	عدم نیاز به لبه‌های دقیق در تصویر مقاومت به شدت نور در نواحی محلی درون تصویر	استخراج نامزدهای متعدد شناسایی اشیا در یک ناحیه	تشخیص شکل و بافت محلی اشیا
GIST [۵۳]	توصیف‌کنندهٔ ابعاد پایین بدون نیاز به پیش‌پردازش	حساس به تغییر در مقیاس یا دوران	ویژگی‌های ادراکی و معنادار را برای انسان فراهم می‌کند (مانند طبیعی‌بودن، بازبودن و انبساط)
ORB [۵۲]	ویژگی‌های دودویی و حجم کم مقاوم به بلورشدن تصویر و تغییر در پرسپکتیو	حساس به تغییر در مقیاس	همچون SIFT و SURF
Sliding Windows [۵۱]	سادگی پیاده‌سازی ویژگی‌های محلی	از لحاظ محاسباتی سنگین حساس به تغییرات ریز در تصویر اشیا	کارایی در درک بهتر رخدادهای درون تصویر

[DOI: 10.61186/jsdp.22.1.113]



(شکل-۶): محاسبهٔ شباهت بین اشیای درون تصویر و قطعات متن [31]

(Figure-6): Finding similarities between Image patches and parts of Text [31]

البته اشاره به این نکته ضروری است که این روش قدیمی و عملکرد آن نسبت به دیگر روش‌های مطرح در این حوزه ضعیف‌تر است؛ بااین‌حال روشی پرکاربرد در سال‌های گذشته بود که در آن نیاز به تلفیق ویژگی‌های متن و تصویر است؛ برای مثال برای روش‌های کمابیش مشابه می‌توان به پژوهش‌هایی از دانشگاه استنفورد [۳۰، ۶۰]، شرکت گوگل [۴]، دانشگاه مونترال [۶۱]، دانشگاه کالیفرنیا [۶۲] و یا پژوهش‌هایی در چالش مایکروسافت کوکو[1] اشاره کرد[۱۹]. نقطه ضعف اصلی روش‌های مبتنی بر ترازبندی همچون روش [۳۱]، نیازمند یک تجزیه‌گر وابستگی متن[2] همچون Stanford NLP یا NLTK است تا بتوان روابط بین واژه‌ها را استخراج کرده و سپس آن‌ها را تحویل شبکهٔ پردازش متنی داد. طبیعی است عملکرد این تجزیه‌گر، تأثیر مستقیمی بر خروجی نهایی مدل خواهد داشت و درصورت عملکرد نامناسب یا ضعیف، خود تجزیه‌گر به نقطهٔ شکستِ عملکرد تبدیل خواهد شد.

در (شکل-۸) چندین نمونه از روش‌های مطرح و موفق در بنچمارک‌های بین‌المللی همچون Microsoft COCO و Pascal VOC که در سال‌های گذشته به موفقیت‌های خوبی در چالش‌های توصیف تصاویر دست‌یافته‌اند، نمایش داده شده‌است.

یکی از نقاط قوّت راه‌کارهای یادشده، ادعای آن‌ها در زایایی و تولید متن است، لیکن نمی‌توان به قدرت زایایی مدل‌هایی که از شبکه‌های بازگشتی و سازوکار توجهی استفاده می‌کنند، اطمینان کرد[۲۹]. تجربه نشان داده‌است؛ باوجود حضور روش‌های رقیبی مبتنی بر «نزدیک‌ترین همسایه»[3] و دست‌یابی برخی از آن‌ها به امتیازاتی بالا در چالش Microsoft COCO، روش‌های فوق قادر به زایایی متن نیستند که این موضوع از کارایی آن‌ها در پردازش تصاویری با اجزای کمتر دیده شده و غیرمعمول می‌کاهد[۶۴، ۶۵].

روش‌های مبتنی بر نزدیک‌ترین همسایه بیشتر به امتیازاتی مناسب برای وظایفی می‌رسند که شباهت به عملکرد انسانی را می‌طلبند؛ این موضوع در ابتدا عجیب و گمراه‌کننده به نظر می‌رسد، اما با دقت در جزئیات عملکرد آن‌ها متوجه علت این برتری نه‌چندان ارزشمند می‌شویم؛ برای مثال در بین روش‌های ارائه‌شده در Microsoft COCO 2015 یک روش با وجود سادگی به دقت بیشتری نسبت به دقت انسانی رسیده‌است[۶۶]. علت برتری این روش نسبت به ادراک انسانی نه به‌دلیل دقت بالای آن در زایایی و تولید جملات و درک تصاویر بلکه به‌دلیل استفادهٔ مستقیم از جملات توصیفی موجود در داده‌های آموزشی است[۶۷].

با وجود این‌که در حال حاضر روش‌های مبتنی بر هم‌ترازی متن و تصویر، کاراترین و نزدیک‌ترین عملکرد را به ادراک انسانی دارند، اما این روش‌ها با چند ضعف اصلی مواجه‌اند: چنان که اشاره شد، یکی از این نقاط ضعف، حفظ‌کردن متون به جای تلاش برای تولید آن‌هاست. اشکال دیگر، وابستگی شدید روش‌ها به مجموعه‌داده‌ها و توصیفات درون آن‌هاست. بهطور معمول، مجموعه‌داده‌های مطرح تصاویر و توصیفات، شامل تصاویری عادی و گاه تکراری از روزمرگی و اتفاقات اطراف ماست. این تکرار مکررات باعث تنبلی مدل‌ها و عدم انعطاف آن‌ها نسبت به تصاویری می‌شود که کمتر دیده و یا هرگز دیده نشده‌اند؛ برای مثال **Error! Reference source not found.** نمونه‌ای از تصاویر «خارج از بافتار»[4] را نشان می‌دهد که در مجموعه‌داده‌های معمولی همچون Flickr یا Visual Genome [۲۳] موجود نیست[۶۸].

این مشکل ناشی از آن است که روش‌های کاربردی بیش از یادگیری و توجه به جزئیات، تلاش در حفظ‌کردن و بازیابی متون و سپس چسباندن آن‌ها به هم دارند و به همین دلیل، تصاویری غیر قابل قبول را در خروجی تولید می‌کند.

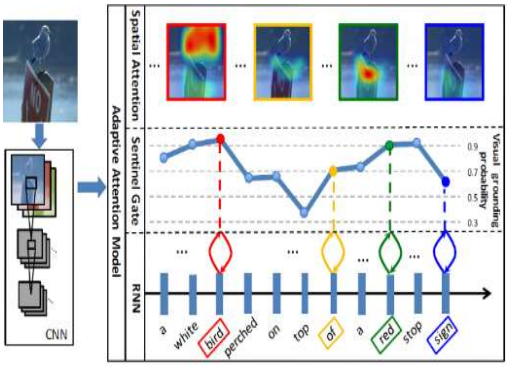
[1] Microsoft COCO
[2] Dependency Parser
[3] Nearest Neighbour
[4] Out of context

[DOI: 10.61186/jsdp.22.1.113]

(شکل-۸): رویکردهای کمابیش یکسان در توصیف تصاویر [۴، ۱۹، ۳۰، ۶۱، ۶۳]

(Figure-7): Common patterns in connecting images and captions [4] [19] [30] [61] [63]

روش‌های اخیر در حوزهٔ حل مسائل خارج از بافتار، تلاش در نزدیک‌کردن عملکرد معناگرای ماشین در پردازش تصاویر به آنچه در مغز روی می‌دهد، دارند[۱]. به‌کارگیری دقیق‌تر «سازوکار توجه» و توجه به برجستگی‌های تصویری یکی از این راه‌کارهاست[۶۹]. در پژوهشی تلاش شده‌است با تکیه بر این موضوع در هر لحظه به بخشی مهم از تصویر نگاه شده و توصیفی مناسب از آن بخش نسبت به کل تصویر ارائه شود[۱۹]. نقطه ضعف روش‌های مبتنی بر سازوکار توجه این است که این روش‌ها بیشتر توضیحی برای انتخاب قطعات مورد توجه‌شان ارائه نمی‌دهند و در صورتی که این قطعات بد یا اشتباه انتخاب شوند، راه‌کاری برای تفسیر عملکرد آن‌ها نیست؛ بااین‌حال، عملکرد کلان این روش‌ها نسبت به مدل‌هایی که از آن‌ها استفاده نمی‌کنند، در بیشتر موارد برتری چشم‌گیری دارد[۶۹]. مجموعه‌داده‌های نمونه‌ای از عملکرد این شیوه در (شکل-۹) نمایش داده شده‌است:



(شکل-۹): توجه به برجستگی‌ها و جزئیات تصویر به طور مستقل و درعین‌حال در ارتباط با کل تصویر [۱۹]

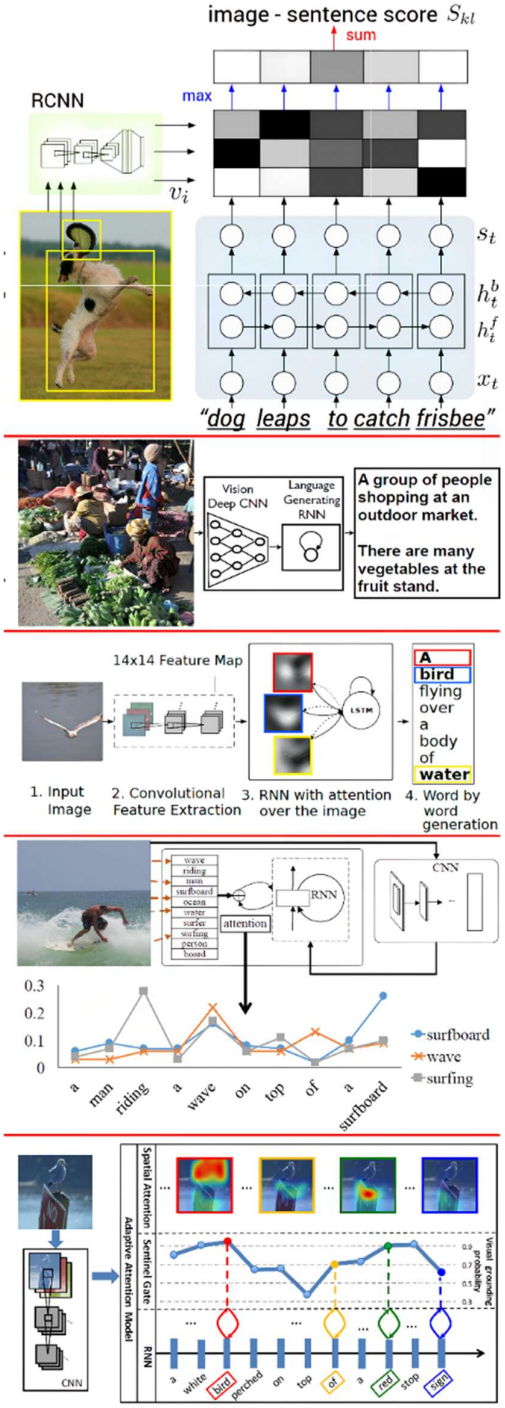
(Figure-9): Attending to different parts of an image based on text

در روشی دیگر برای غلبه بر این مشکل تلاش شده‌است که تا جای ممکن اطلاعات تصویر و موجودیت‌های درون آن و روابط آن‌ها، به‌صورت گراف از تصویر استخراج شود[۶۰]. با داشتن بیشینهٔ اطلاعات قابل استخراج از کل تصویر و تک‌تک جزئیات درونِ آن، می‌توان از بافتار کلیِ حاصل، بهره برد و احتمال رخداد خطا را در توصیفات پایین آورد. متأسفانه استخراج گراف تصویر[1] وظیفه‌ای به غایت پیچیده است و برخلاف دیگر روش‌های مطرح در بینایی ماشینی که در سال‌های اخیر به دقت[2] و بازیابی[3]های بالای ۹۰٪ رسیده‌اند، روش‌های

[1] Scene Graph Generation
[2] Precision
[3] Recall

(شکل-۷): تصاویر خارج از بافتار [۶۸]

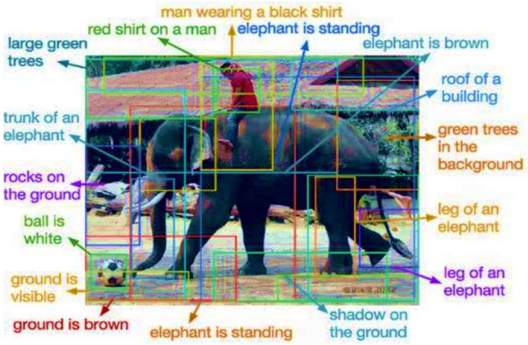
(Figure-8): Out-of-Context images



[DOI: 10.61186/jsdp.22.1.113]

این حوزه هنوز در حال رقابت برای رسیدن به بازیابی‌های بالای ۳۰٪ بر روی مجموعه‌داده‌های مطرح بازیابی همچون Microsoft COCO یا Stanford Online Products هستند [۷۰, ۷۱] که حکایت از پیچیدگی زیاد وظایف مطرح در این دامنه از مسائل است.

در (شکل-۱۰) نمونه‌ای از خروجی به‌دست‌آمده از چنین روش‌هایی نمایش داده شده‌است:



(شکل-۱۰): توجه به برجستگی‌ها و جزئیات تصویر به‌طور مستقل و در عین حال در ارتباط با کل تصویر [۱۹]

(Figure-10): Connecting different parts of an image in regards to the whole image

۲-۳- هم‌ترازی[1] بهتر قطعات متن و تصویر

با علم به این نکته که برای بازیابی موفق محتوامحور نمی‌توان از نیاز به ایجاد ارتباط بین دو دامنهٔ متن و تصویر چشم‌پوشی کرد. با داشتن راه‌کاری که قادر به ایجاد ارتباط بین دو دامنهٔ تصویر و متن است، می‌توان بازیابی مؤثرتر و با دقت بالاتری انجام داد؛ به همین دلیل در این بخش به مرور راه‌کارها و پژوهش‌هایی خواهیم پرداخت که تلاش بر ایجاد ارتباطی هر چه بهتر بین این دو دامنه داشته‌اند.

۲-۳-۱- دامنهٔ تصاویر

در سال‌های اخیر، روش‌های یادگیری ماشینی در حوزهٔ بینایی ماشینی، تمرکز خود را بر استخراج ویژگی‌ها و یادگیری بازنمایی‌های ارزشمند از تصاویر قرار داده‌اند. یکی از مهم‌ترین روش‌های مورد استفاده در این زمینه روش «شبکه‌های عصبی هم‌آمیز»[2] است[۷۲]. تمرکز این روش بر استفاده از لایه‌های هم‌آمیز و کم‌هزینه (از لحاظ محاسباتی در مقایسه با لایه‌های مطرح در شبکه‌های «پرسپترون چندلایه») است. با توزیع این لایه‌ها بر روی حافظهٔ کارت‌های گرافیکی (به‌جای حافظهٔ اصلی رایانه)، پژوهش‌گران قادر به افزایش سرعت و دقت در یادگیری بازنمایی‌های تصویری و نیز در حوزهٔ دسته‌بندی تصاویر هستند[۷۳].

در شکل (۱۱) شاهد دو لایهٔ پایانی پرسپترونی با ابعاد ۲۵۶ و ۱۲۸ هستیم. کاهش ابعادی که در معماری این شبکه در حال رخداد است، موجب اعمال فشار بر شبکه برای فشرده‌سازی بازنمایی‌های یادگیری‌شده در لایه‌های قبلی می‌شود[۷۲، ۷۴]. با تلاش برای فشرده‌سازی ویژگی‌های استخراج‌شده از لایه‌های میانی، شاهد غنی‌ترشدن لایه‌های پایانی شبکه خواهیم بود به‌طوری‌که فقط چندین پرسپترون در این لایه، وظیفهٔ هزاران پالایه و پارامتر در لایه‌های میانی را بر عهده می‌گیرند؛ برای مثال اگر تنها یکی از پالایه‌های لایه‌های میانی که مخصوص شناسایی و بازنمایی تصاویر آموزش‌دیده فعال شود، به‌احتمال بسیار زیاد یکی از پرسپترون‌های لایه‌های پایانی در ارتباط با آن فعال خواهد شد[۷۳].

نقش این لایه‌های پایانی به حدی زیاد است که با نام «لایهٔ تعبیه»[3] شناخته می‌شوند. این لایه‌ها نگاشت قدرتمندی از فضای تصویر به فضای پالایه‌ها (فضای برداری) می‌دهند. «بردار تعبیه»[4] حجم بسیار کمی دارد ضمن اینکه وظیفهٔ تعبیهٔ تمامی مفاهیم یادگیری‌شده به‌وسیلهٔ لایه‌های میانی شبکه را نیز بر عهده دارد و اگر هدف، آموزش برای دسته‌بندی تصاویر باشد، این بردارها به‌صورت ضمنی قادر به جداسازی فضای دستهٔ خود از دستهٔ تصاویر دیگر هستند. نکتهٔ اصلی در این حوزه، قابلیت به‌کارگیری مفاهیم «شباهت[5]» و «فاصله[6]» است.

در حالت کلی در فضای n بُعدی، باید بتوان با تعریف معیاری برای فاصله، شباهت بین دو بردار اندازه‌گیری‌شده را مشخص کرد[۷۵]. در

(جدول-۴) مهم‌ترین توابع و معیارهای فاصله‌یابیِ مطرح در مباحث یادگیری ماشینی و نظریهٔ اطلاعات نمایش داده شده‌است.

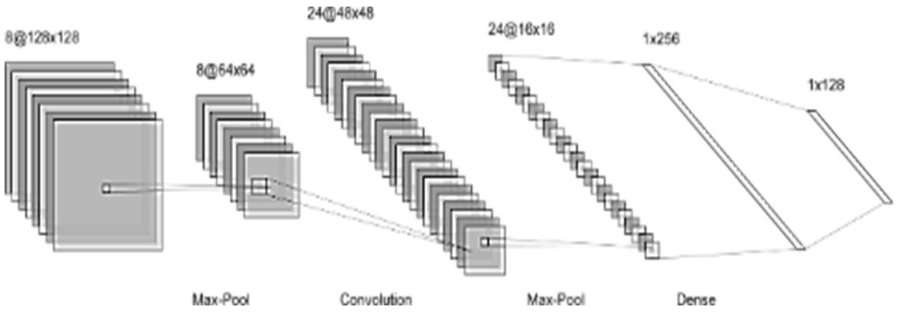
فاصلهٔ اقلیدسی، اندازه‌گیری فاصلهٔ خطِ مستقیمِ بینِ دو نقطه در فضایی n بُعدی است. این فاصله، در هنگام مقایسهٔ بردارهای با مختصات دکارتی مفید است. این فاصله، جهت و اندازهٔ بردارها را در کنار هم درنظر می‌گیرد. این موضوع بدین معنی است که این فاصله، برای بردارهایی با بزرگی متفاوت، می‌تواند تحت تأثیر عوامل پرت[7] قرار گیرد؛ بدین معنا که بردارهایی که از نظر بزرگی با هم متفاوت‌اند، بر حسب معمول فاصلهٔ اقلیدسی بزرگ‌تری خواهند داشت[۷۵]. «شباهت کسینوسی» معیاری برای تشابه بین دو بردار غیر صفر از فضای حاصل‌ضرب داخلی است که کسینوس زاویهٔ بین آن‌ها را اندازه‌گیری کرده و اهمیتی به بزرگی[8] بردارها نمی‌دهد؛

[1] Alignment
[2] Convolutional Neural Networks

[3] Embedding Layer
[4] Embedding Vector
[5] Similarity
[6] Distance
[7] Outlier
[8] Magnitude

[DOI: 10.61186/jsdp.22.1.113]

این ویژگیِ شباهت کسینوسی آن را زمانی مفید می‌کند که بزرگی بردارها می‌تواند به طور گسترده‌ای متفاوت باشد، اما جهت آن‌ها هنوز نسبت به هم معنی‌دار است. بی‌اعتنایی شباهت کسینوسی به بزرگی بردارها موجب می‌شود که مقادیر این شباهت بین ۱- الی ۱+ قرار گرفته و بتوان با هنجارسازیِ آن بین صفر الی ۱+، این معیار را به صورت درصد (٪) نمایش داد[۷۵].



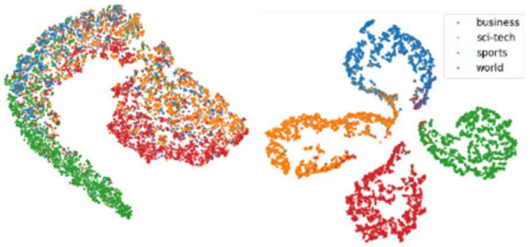
(شکل-۱۱): معماری شبکه‌های هم‌آمیز
(Figure-11): CNN Architecture

«فاصلهٔ چِبی‌شِف»، تعمیم فاصلهٔ اقلیدسی محسوب می‌شود و می‌تواند جنبه‌های مختلفی از رابطهٔ بین دو نقطه را به تصویر بکشد، اما در اصل مزیتی نسبت به فاصلهٔ اقلیدسی ندارد[۷۹]. فواصل «لونشتاین» [۸۱] و «همینگ» [۸۰]، در درجهٔ نخست برای مقایسهٔ دنباله‌ها یا رشته‌ها استفاده می‌شوند و بیشتر کاربرد مناسبی برای داده‌هایی خارج از این دامنه ندارند. «فاصلهٔ ماهالانوبیس» برای مقایسهٔ توزیع‌های چندمتغیره و برای مقایسهٔ فواصل مجموعه‌ها از «فاصلهٔ جاکارد» استفاده می‌شود[۸۴]؛ از طرفی معیارهایی همچون «واگرایی‌های جنسن-شانون» یا «کولبک-لیبلر»[1] برای محاسبهٔ میزان شباهت بین دو توزیعِ احتمال به‌کار می‌روند و نه مقایسهٔ میزان شباهتِ دو بردارِ متفاوت و به همین دلیل در فرایند بازیابی نزدیک‌ترین همسایه‌های مربوط به یک بردار کاربرد چندانی ندارند[۹۰].

با توجه به نکات یادشده، دو معیار «فاصلهٔ اقلیدسی» و «شباهت کسینوسی»، معیارهای مناسبی برای محاسبهٔ فاصلهٔ بین دو بردار به نظر می‌رسند. طبق پژوهش [۹۱]، در فضاهای با ابعاد بالا، بین این دو فاصله تفاوت خاصی وجود ندارد و در وظایف بازیابی اطلاعات، شبیه به یکدیگر عمل می‌کنند؛ بدین ترتیب می‌توان یک سامانهٔ بازیابی محتوامحور تصویر طراحی کرد که هدف آن دریافت یک تصویر، استخراج بردارهای تعبیه[2] و جست‌وجو و بازیابی تصاویر شبیه به آن باشد. این مراحل سال‌هاست که برای بازیابی محتوامحور تصاویر به‌کار می‌رود و از مقیاس‌های کوچک چند صد تصویره (نرم‌افزارهای خانگی) تا مقیاس‌های بزرگ چندمیلیاردی (قابلیت جست‌وجوی تصاویر در گوگل و بینگ[3]) همگی از این مراحل پیروی می‌کنند. با توجه به این موضوع، آنچه وجه تمایز و قدرت رقابتی راه‌کارهای مختلف بازیابی تصاویر است «استفاده از بردارهای تعبیهٔ بهتر» است و برای رسیدن به چنین بردارهایی ناگزیر به تلفیق دامنه‌های تصویر و متن هستیم.

۲-۳-۲- دامنهٔ متن

تفاوت بنیادی متن با تصویر، «توالی و دنباله‌داربودن»[4] متون است؛ برای مثال، وقتی در حال پردازش یک تصویر به‌کمک شبکه‌های عصبی هم‌آمیز هستیم، فیلترهای این شبکه به‌صورت یک‌پارچه بر کل تصویر اعمال می‌شوند و این موضوع تا انتهای شبکه و استنتاج خروجی آن صدق می‌کند؛ درحالی‌که یک متن می‌تواند شامل تعداد متناهی یا نامتناهی از واژه‌ها باشد، به‌طوری‌که در بسیاری از شرایط، حافظهٔ کافی برای پردازش کل آن متن موجود نیست. برای حل این مشکل و پردازش متون کوتاه و بلند، معماری‌های متعددی همچون شبکه‌های عصبی «حافظهٔ طولانی کوتاه‌مدت»[5] [۹۲]، «واحد بازگشتی دروازه»[6] [۹۳] و «شبکه‌های مبدّل»[7] [۹۴] معرفی‌شده‌اند. وجه اشتراک تمام این روش‌ها، یکسان‌بودن مفهوم «بردارهای تعبیه» در بین آن‌هاست.



(شکل-۱۲): مصورسازی متون مجموعه‌داده‌ای شامل اخبار بر مبنای لایهٔ تعبیهٔ هر متن در فضای دو بعدی. از چپ به راست: بردارهای تعبیه در ابتدای آموزش مدل، بردارهای تعبیه در پایان آموزش مدل
(Figure-12): a News dataset embedding in 2D using t-SNE. Left: Embeddings in the start of training. Right: Embeddings at the end of the training.

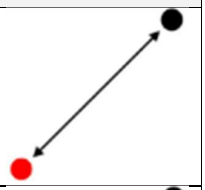
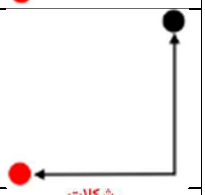
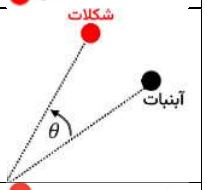
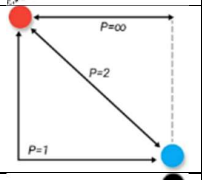
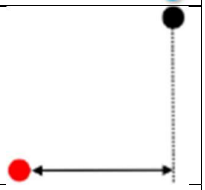
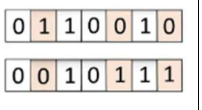
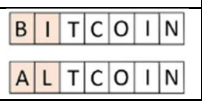
[1] Kullback-Leibler Divergence
[2] Embedding Vector
[3] Bing
[4] Sequential
[5] Long Short-Term Memory
[6] Gated Recurrent Unit
[7] Transformer

[DOI: 10.61186/jsdp.22.1.113]

در روش‌های مبتنی بر متون نیز همچون روش‌های مبتنی بر تصاویر، مدل محاسباتی در آخرین لایه‌ها دارای بردارهایی است که اطلاعات ارزشمندی را در خود نهفته دارند و از قواعدی که در مورد بردارهای تعبیهٔ تصویر یاد کردیم (همچون وجود مفهوم فاصله در بین بردارها) تبعیت می‌کنند. در **Error! Reference source not found.** نمونه‌ای از کاهش ابعاد بر روی بردارهای تعبیهٔ متنی نمایش داده شده‌است. بردارهای ابعاد در این شکل به کمک یک شبکهٔ BERT [۹۵] از مجموعه‌داده‌ای ۱۲۰ هزار متنی استخراج شده‌اند. تصویر سمت چپ وضعیت این بردارها را در شروع آموزش شبکه و تصویر سمت راست وضعیت آن‌ها را در پایان آموزش و پس از یادگیری مفاهیم به وسیلهٔ شبکه، نشان می‌دهد. می‌توان دید که در این حوزه نیز مفاهیم مرتبط با بردارهای تعبیه صدق می‌کنند.

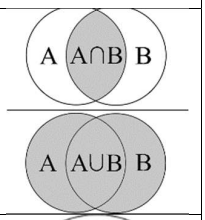
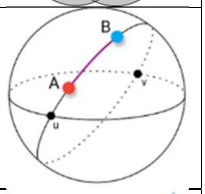
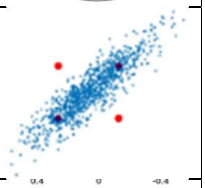
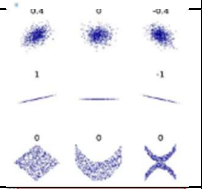
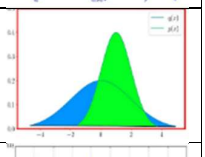
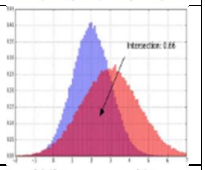
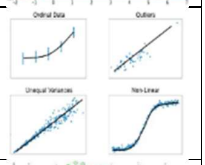
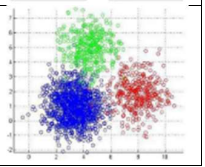
(جدول-۴): اصلی‌ترین توابع و معیارهای فاصله‌یابی

(Table-4): Main similarity and distance functions

تصویر معیار	نام معیار	ویژگی	کاربردها	نقاط ضعف
	فاصلهٔ اقلیدسی [۷۵]	محاسبهٔ اندازهٔ خط مستقیم بین دو نقطه در فضای n بعدی	فاصله‌یابی کلی، خوشه‌بندی، دسته‌بندی، رگرسیون	حساس به مقادیر پرت قابل تأثیر از بردارهایی با بزرگی متفاوت
	فاصلهٔ منهتن[1] [۷۶]	محاسبهٔ اندازهٔ فاصلهٔ بین دو نقطه بر روی یک شبکهٔ گریدی با حرکت محدود (بالا، پایین، چپ، راست)	فاصله‌یابی در شبکه‌های گریدی، الگوریتم‌های مسیریابی، پردازش تصویر	بی‌اعتنا به حرکت مورب در گرید غیر قابل استفاده در ابعاد بالا کم استفاده در داده‌های پیوسته
	شباهت کسینوسی [۷۷]	محاسبهٔ زاویهٔ کسینوس بین دو بردار در فضای n بعدی	بازیابی اطلاعات، پردازش متن، خوشه‌بندی، دسته‌بندی، سیستم‌های توصیه‌گر	بی‌اعتنا به بزرگی بردارها
	فاصلهٔ مینکوسکی[2] [۷۸]	محاسبهٔ فاصلهٔ بین نقاط در فضای n بعدی	معیاری کلی برای محاسبهٔ فواصل	حساس به مقادیر پرت
	فاصلهٔ چِبی‌شِف[3] [۷۹]	محاسبهٔ فاصلهٔ بیشینه بین اجزای همسان دو بردار	محاسبهٔ فاصلهٔ بیشینه، خوشه‌بندی، تشخیص ناهنجاری[4]	مناسب برای داده‌های پیوسته، حساس به مقادیر پرت، کم استفاده در داده‌هایی با هم‌بستگی بالا
	فاصلهٔ همینگ[5] [۸۰]	محاسبهٔ تعداد مکان‌هایی که مقادیر بردار با یکدیگر متفاوت هستند	محاسبهٔ شباهت رشته‌ها، کدهای تصحیح خطا، توالی‌یابی DNA[6]	مناسب برای توالی‌هایی با اندازهٔ یکسان، کم استفاده در داده‌های پیوسته
	فاصلهٔ لِونشتاین[7] [۸۱]	محاسبهٔ تعداد کمینهٔ مورد نیاز برای تبدیل یک کاراکتر از بردار	محاسبهٔ شباهت رشته‌ها	هزینهٔ محاسباتی بالا برای رشته‌های طولانی و بزرگ

[1] Manhattan Distance
[2] Minkowski Distance
[3] Chebyshev Distance
[4] Anomaly Detection
[5] Hamming Distance
[6] DNA Sequencing
[7] Levenshtein Distance

[DOI: 10.61186/jsdp.22.1.113]

تصویر معیار	نام معیار	ویژگی	کاربردها	نقاط ضعف
		نخست به کاراکتر بردار دوم		
	شباهت جاکارد[1] [۸۲]	محاسبهٔ میزان شباهت بین دو مجموعه با مقایسهٔ اشتراک[2] و اجتماع[3] بین آن‌ها	مقایسهٔ مجموعه‌ها، تحلیل متن، سیستم‌های توصیه‌گر	بی‌اعتنا به بزرگی مجموعه‌ها، کم استفاده در داده‌های پیوسته
	فاصلهٔ هاورسین[4] [۸۳]	محاسبهٔ فاصلهٔ ژئودزیک بین دو نقطه بر روی یک کُره	محاسبهٔ فاصله روی کُره، محاسبات جغرافیایی	نامناسب برای فواصل کوتاه بر روی کره‌هایی با مقیاس عظیم
	فاصلهٔ ماهالانوبیس[5] [۸۴]	محاسبهٔ فاصلهٔ بین نقاط در فضای n بعدی با درنظرگرفتن هم‌بستگی بین متغیرها	تجزیه‌وتحلیل آماری چندمتغیره، شناسایی داده‌های پرت، خوشه‌بندی	نیاز به ماتریس کوواریانس، نامناسب برای مجموعه‌داده‌هایی با تعداد دادهٔ زیاد
	هم‌بستگی پیرسون[6] [۸۵]	محاسبهٔ هم‌بستگی خطی بین دو متغیر	محاسبهٔ همبستگی خطی	نیازمند وجود هم‌بستگی خطی
	واگرایی جنسن-شانون[7] [۸۶]	معیاری برای شباهت و تفاوت بین توزیع‌های احتمال	محاسبهٔ شباهت بین توزیع‌های احتمال، خوشه‌بندی، سامانه‌های توصیه‌گر	فقط قابل اعمال بر روی بردارهای مثبت
	فاصلهٔ مربع کای[8] [۸۷]	محاسبهٔ شباهت بین دو هیستوگرام با مقایسهٔ واگرایی مربع کای	محاسبهٔ شباهت بین هیستوگرام‌ها	فقط قابل اعمال بر روی بردارهای مثبت
	هم‌بستگی اسپیرمن[9] [۸۸]	محاسبهٔ وابستگی بین دو متغیر در یک مجموعه‌داده بر مبنای ترتیب رتبه	محاسبهٔ هم‌بستگی رتبه	فقط قابل اعمال بر روی داده‌های ترتیبی
	فاصلهٔ کانبرا[10] [۸۹]	محاسبهٔ فاصلهٔ بین دو بردار با درنظرگرفتن نسبت بزرگی آن‌ها	محاسبهٔ فاصله برای داده‌های پراکنده	حساس به تفاوت‌های کوچک

[1] Jaccard Distance (Index)
[2] Intersection
[3] Union
[4] Haversine Distance
[5] Mahalanobis Distance
[6] Pearson Correlation
[7] Jensen-Shanon Divergence
[8] Chi-Squared Distance
[9] Spearman Correlation
[10] Canberra Distance

[DOI: 10.61186/jsdp.22.1.113]

اشاره به این نکته ضروری است که در سال‌های اخیر حوزهٔ پردازش متون (برخلاف حوزهٔ پردازش تصاویر) در انحصار کامل شرکت‌های فناوری دنیا و دانشگاه‌هایی با دسترسی به زیرساخت‌های محاسباتی قدرتمند بوده‌است[۹۶] .علت این امر نیاز به آموزش مدل‌های زبانی و متنی در مقیاس‌های بسیار بزرگ است؛ زیرا عملکرد این مدل‌ها در صورت آموزش بر روی متون در مقیاس‌های کوچک قابل‌قبول نیست؛ برای مثال در بین ده مدل برتر استخراج بردارهای تعبیه در سال ۲۰۲۳، هیچ‌کدام از آن‌ها کمتر از یک میلیارد پارامتر برای یادگیری ندارد؛ به همین دلیل برای آموزش این مدل‌ها نیاز به ده‌ها کارت محاسباتی و صدها گیگابایت حافظه است[۹۷].

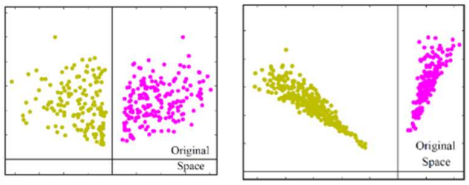
مثال دیگر، شرکت OpenAI است که برای آموزش یکی از مدل‌های GPT خود در سال ۲۰۲۰ بیش از چهار میلیون دلار هزینهٔ انرژی برق برای کارسازهایش[1] پرداخت کرده‌است[۹۸]. این تفاوت‌های فاحش در مدل‌های زبانی و متنی موجب شده‌است که پژوهش‌گران این حوزه، تلاش‌های خود را بر «تنظیم»[2] این مدل‌ها متمرکز سازند و نه معرفی و آموزش یک مدل جدید؛ برای مثال، روش ارائه‌شده در مقالهٔ شبکه‌های مبدّل، با نام « Attention is All You Need» در سال ۲۰۱۷، هنوز پس از هفت سال از معرفی به وسیلهٔ شرکت گوگل، رتبهٔ نخست خود را در حوزه‌هایی همچون «ترجمهٔ ماشینی»[3] یا «شناسایی احساسات در سخنرانی»[4] حفظ کرده‌است که به معنی عدم دسترسی پژوهش‌گران به منابع محاسباتی است که قادر به رقابت با این روش باشد؛ زیرا توسعهٔ روشی رقیب و برتر از این روش نیازمند هزینه‌های گزاف برای آموزش‌های مکرر و آزمون‌وخطاهای متعدد است[۹۴].

۲-۳-۳-هم‌جوشی ویژگی‌های تصویری و متنی با هدف بازیابی

برای تلفیق دو حوزهٔ تصویری و متنی با یکدیگر، روش‌های هم‌جوشی متنوعی ارائه شده‌اند. برخی از این روش‌ها با تلفیق گراف‌های درون تصویر با متون توصیفی آن تلاش در هم‌جوشی این دو حوزه دارند[۹۹].

برخی روش‌ها با به‌کارگیری راه‌کارهای مبتنی بر «توجه» تلاش در ایجاد ارتباط بین این دو حوزه دارند[۱۹] و برخی دیگر از روش‌ها رویکردهای مبتنی بر «یادگیری فاصلهٔ معیار»[5] یا «یادگیری تضادی»[6] را برای نزدیک‌کردن این دو فضا پیشنهاد داده‌اند. در بین تمامی روش‌های ارائه‌شده در این حوزه، روش‌های مبتنی بر «یادگیری فاصله» بیشترین دقت و بهترین عملکرد را از خود نشان داده‌اند[۹۶، ۱۰۰، ۱۰۱].

چنان که اشاره شد، در روش‌های یادگیری فاصله، تمرکز بر تعبیه‌کردن داده‌های ورودی در فضا به‌گونه‌ای است که تمامی داده‌های مربوط به یک «دسته»[7] در نزدیکی هم و تمامی داده‌های نامربوط خارج از آن قرار گیرند. در **Error! Reference source not found.** بردارهای فضای تعبیهٔ هر دسته تلاش داشتند تا با یک ابرصفحه خود را از فضای تعبیهٔ دسته‌های دیگر جدا نگه دارند.



(شکل-۱۳): مصورسازی داده‌های آموزش‌دیده بدون راه‌کارهای مبتنی بر یادگیری فاصله (تصویر چپ) و مبتنی بر یادگیری فاصله (تصویر راست) [۱۰۲]

(Figure-13): Visualizing the learned embeddings of a dataset based on non-metric-learning approaches (Left) and metric-learning approaches (Right)

بااین‌حال، در صورتی که آن مجموعه‌داده‌ها به وسیلهٔ یک روش «یادگیری فاصله» در فضا تعبیه شوند، شاهد بردارهای تعبیه‌ای شبیه به **Error! Reference source not found.** خواهیم بود. در این شکل، تصاویر چهرهٔ انسان، یک‌بار از طریق شبکه‌ای که بر مبنای روش‌های دسته‌بندی استاندارد و مطرح آموزش‌دیده است تبدیل به بردارهای تعبیه شده‌اند و بار دیگر از طریق شبکه‌ای که بر مبنای روش‌های مبتنی بر یادگیری فاصله آموزش‌دیده است تبدیل به بردارهای تعبیه شده‌اند. بردارهای تصاویری (چهره‌هایی) که به لطف به‌کارگیری راه‌کارهای مبتنی بر یادگیری فاصله، در فضای دوبعدی تعبیه شده‌اند، تمایز بیشتری را نسبت به دسته‌های دیگر (چهره‌های دیگر) از خود نمایش می‌دهند[۱۰۲]. نکتهٔ دیگر این‌که با وجود عملکرد قدرتمند روش‌های مبتنی بر یادگیری فاصله در تعبیهٔ بردارهای ورودی در فضایی جدید، نمی‌توان آن‌ها را بر روی همه نوع مجموعه‌داده اعمال کرد. ویژگی متمایز این روش، متوازن‌بودن تعداد هر دسته از تصاویر یا متون، حساسیت روی مقیاس داده‌ها و «اندازهٔ دسته»[8] به هنگام آموزش است؛ برای مثال در حوزهٔ «تشخیص چهره»[9] نیاز

[1] Server
[2] Fine-tune
[3] Machine Translation
[4] Speech Emotion Recognition
[5] Metric Learning
[6] Contrastive Learning
[7] Class
[8] Batch Size
[9] Face Recognition

[DOI: 10.61186/jsdp.22.1.113]

است، نخست میلیون‌ها تصویر از چهره‌های افراد مختلف موجود باشد؛ سپس از چهرهٔ هر فرد حداقل ده تصویر وجود داشته باشد و مهم‌تر از همه، به هنگام آموزش هر چقدر «اندازهٔ دسته» بزرگ‌تر باشد دقت نهایی مدل آموزش‌دیده بر روی این تصاویر و کیفیت بردارهای تعبیه‌شده به وسیلهٔ آن بالاتر خواهد بود[۱۰۰]. به‌طبع این موضوع به دسترسی به سخت‌افزارهای قدرتمند با حافظهٔ زیاد نیاز دارد که در عمل باعث کاهش تعداد پژوهش‌گرانی شده‌است که در این حوزه فعالیت می‌کنند.

در سال‌های اخیر، روش‌های متنوع و مختلفی برای «یادگیری فاصله» معرفی‌شده‌اند که هر کدام بر روی وظیفه‌ای خاص متمرکز بوده‌اند، اما می‌توان گفت که مهم‌ترین عامل تأثیرگذار بر عملکرد و دقت این روش‌ها، حجم مجموعه‌دادهٔ استفاده‌شده و «اندازهٔ دسته»ها به هنگام آموزش مدل است[۹۶]. شرکت OpenAI با ارائهٔ روش CLIP[1] در سال ۲۰۲۱ برای نخستین‌بار، تلاش کرد روش «یادگیری فاصله» (یا «یادگیری تضادی») را در مقیاس چهارصد میلیون تصویر و متن پیاده‌سازی کند. هدف این روش تعبیه‌کردن هر متن و توضیح مربوط به آن متن در فضایی متریک است که در صورت موفقیت پژوهش‌گر را قادر به رمزگذاری تصاویر و متون در یک فضای یکسان، مرتبط و نزدیک به هم می‌کند؛ به‌طوری‌که هر تصویر و متن مرتبط با آن در این فضا و هر تصویر و متنی دیگر، در مکانی دیگر و نزدیک به هم قرار بگیرند و بین این گروه‌های «تصویر/ متن» فاصله باشد.

تمامی تلاش‌هایی که قبل از این برای یادگیری فضای تعبیه‌ای یکسان برای تصاویر و متون پیشنهاد شده بود، صحت بسیار پایینی در حوزهٔ «دسته‌بندی تصاویر بدون آموزش بر روی مجموعه‌دادهٔ تصاویر» یا «صحت تلاش نخست»[2] داشتند[۹۶]. برای مشخص‌کردن موفقیت عملکرد روش‌های تعبیه‌سازی تصاویر و متون، از شاخصِ «صحت تلاش نخست» استفاده می‌شود. این شاخص، معیاری برای ارزیابیِ مدل‌هایی است که برای دسته‌بندی داده‌هایی استفاده می‌شوند که در داده‌های آموزشی آن‌ها وجود نداشته‌اند. صحتِ تلاش نخست، نسبتِ تعداد نمونه‌هایی که به‌درستی طبقه‌بندی شده‌اند به تعداد کل نمونه‌ها در مجموعه‌داده است[۹۶].

در **Error! Reference source not found.** تفاوت دقت روش یادشده (مبتنی بر آموزش از تصویر به متن بر روی یک مدل زبانی) با روش CLIP (مبتنی بر تعبیه‌سازی تصاویر و متون در یک فضای یکسان و نزدیک به هم) نمایش داده شده‌است. در این شکل، نتایج عملکرد این مدل‌ها پس از به‌کارگیری شانزده عدد GPU برای آموزش روی مجموعه‌دادهٔ چهارصد میلیون تصویر/متنی LAION نمایش داده شده‌است. می‌توان دید که روش مبتنی بر مدل زبانی پس از آموزش بر روی چهارصد میلیون تصویر، فقط به دقت ۱۶٪ تلاش نخست روی مجموعه‌دادهٔ ImageNet دست می‌یابد؛ درحالی‌که روش مبتنی بر CLIP بسیار بهتر است و در زمانی ده برابر سریع‌تر از روش قبلی به همین دقت (۱۶٪) دست می‌یابد[۹۶].

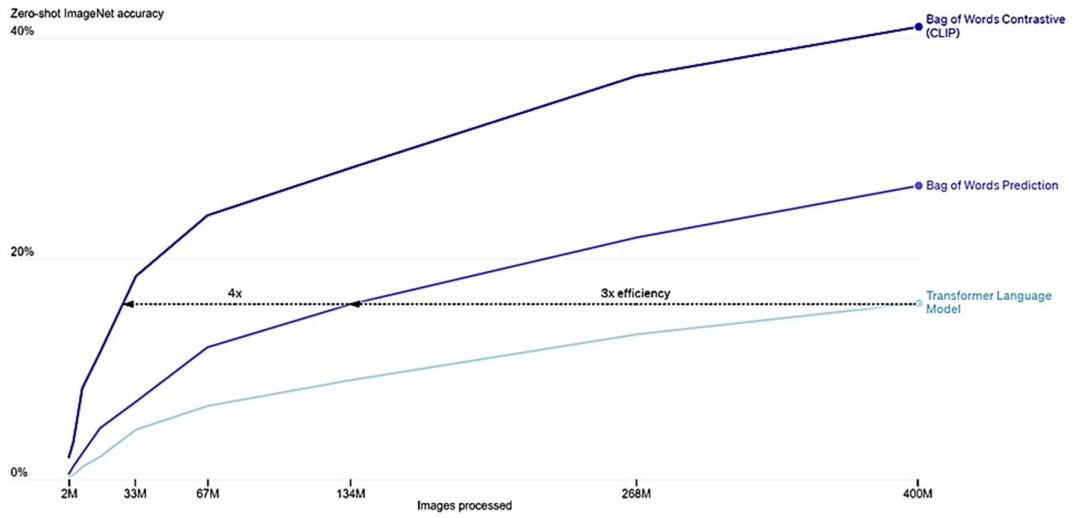
تفاوت روش CLIP با دیگر روش‌ها، افزایش دقت و یادگیری روابط واژه‌ها درون متن بدون نیاز به کتابخانه‌های بیرونی است. در (شکل-۶) به روش‌هایی اشاره کردیم که از طریق یک کتابخانهٔ مجزا (همچون NLTK یا Stanford NLP) در ابتدا روابط بین واژه‌های درون متن را استخراج کرده و سپس آن‌ها را تحویل به مدل اصلی برای انجام محاسبات مربوط به متن می‌دادند. علت انجام این کار نبود راه‌کار مناسبی برای تبدیل متن از واژه‌های اولیهٔ گسسته به بردارهای تعبیهٔ پیوسته بود که موجب شکل‌گیری روش‌های دیگری همچون Word2Vec نیز شده بود[۳۳]. روش CLIP به لطف به‌کارگیری شبکه‌های مبدل برای پردازش متون، نیازی به کتابخانه‌های مجزای پردازش متن نداشته و مستقیم قادر به پردازش متون و واژه‌های داخل آن‌هاست و این موضوع موجب افزایش دقت و عملکرد این روش در دامنهٔ متنی و همچنین بهبود عملکرد آن در ارتباط‌دادن دو دامنهٔ متنی و بصری شده‌است.

موفقیت روش CLIP را می‌توان ناشی از موفقیت روش‌های مبتنی بر «یادگیری فاصله» یا «یادگیری تضادی» دانست: آموزش بر مجموعه‌داده‌های مقیاس بزرگ در کنار دسترسی به قدرت محاسباتی عظیم برای آموزش بر روی «اندازه دسته»های بسیار بزرگ (برای مثال اندازهٔ دسته‌های هزارتایی که در آن قادر به پردازش هزار زوج «تصویر / متن» باشد).

روش آموزش این مدل، کمابیش ساده و بدین صورت است که در آن از دو شبکهٔ عصبی برای تعبیهٔ تصاویر و متون استفاده می‌شود. در این روش یک شبکه برای تصاویر و شبکهٔ دیگر برای متون درنظر گرفته شده و هرکدام وظیفهٔ گرفتن تصویر یا متن به‌عنوان ورودی و تعبیهٔ آن‌ها در فضایی جدید را بر عهده می‌گیرند[۹۶]. طبیعی است که در ابتدای کار فضای بردارهای تعبیه‌شدهٔ تصاویر و متون یکسان نیستند و به همین دلیل به‌کمک یک تابع خطای «آنتروپی متقابل»[3] تلاش می‌شود تا هر بردار تعبیه با بردار متنِ متناظرش در یک دسته قرار بگیرند.

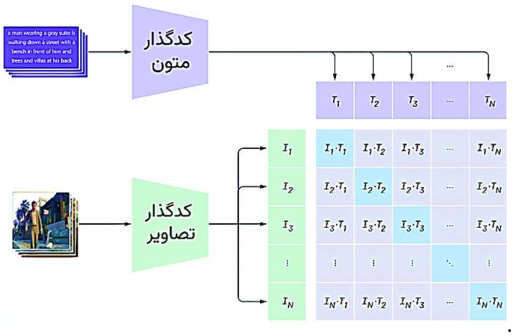
[1] Contrastive Language-Image Pre-training
[2] Zero-Shot Classification Accuracy
[3] Cross Entropy Loss

[DOI: 10.61186/jsdp.22.1.113]



(شکل-۱۴): افزایش دقت در دسته‌بندی تلاش نخست بر روی مجموعه‌دادهٔ ImageNet در صورت به‌کارگیری روش CLIP نسبت به دیگر روش‌ها [۹۶]

(Figure-14): Increase in Zero-shot classification on ImageNet when using the CLIP approaches v ersus other approaches

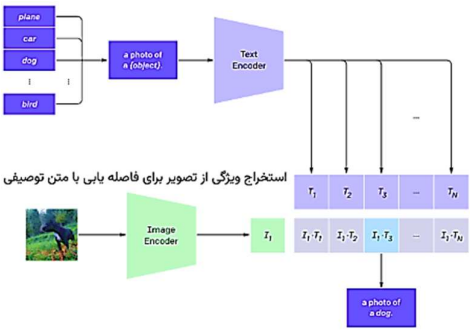


(شکل-۱۵): فرایند آموزش CLIP در کنار «یادگیری فاصله» برای ایجاد ارتباط بین تصاویر و متون متناظر

(Figure-15): The CLIP approach to learn common embeddings for the same image and caption

در Error! Reference source not found. جزئیات آموزش یک مدل CLIP برای تعبیهٔ تصاویر و متون نمایش داده شده‌است. در این شکل شاهد تغذیهٔ شبکه‌های عصبی کدگذار برای تصویر و متن هستیم که از طریق آن‌ها تصاویر و متون تبدیل به بردارهای تعبیه‌شده‌ای در فضاهای غیریکسان می‌شوند (این فضاها با T برای متون و I برای تصاویر نمایش داده شده‌اند).

سپس به کمک تابع خطای «آنتروپی متقابل برای دسته‌بندی‌های چندتایی»[1] تلاش می‌شود هر دو شبکهٔ تصویری و متنی طوری بهینه شوند تا قادر به تعبیهٔ تصاویر و متون در فضاهای نزدیک به هم و به‌طورتقریبی یکسان باشند. با بهینه‌سازی این خطا، شبکه را به سمتی هدایت می‌کنیم تا فاصلهٔ دامنه‌های تصویر و متنِ متناظر را نسبت به هم کم کند. پس از پایان فرایند آموزش مدل CLIP، می‌توان از این شبکه برای پیش‌بینی بردارهای تعبیه و به‌کارگیری این بردارها برای انواع و اقسام اهداف محاسباتی استفاده کرد. از آنجا که کدگذارهای متن و تصویرِ شبکه، پس از آموزش، قادر به عملکرد مستقیم‌اند، می‌توان از هر کدام از آن‌ها مستقل استفاده کرد. با داشتن هر عکس یا متن، می‌توان آن را تحویلِ کدگذارِ مرتبط کرده و آن را تبدیل به یک بردار تعبیه ارزشمند کرد و با داشتن این بردارهای تعبیه (چه تصویر و چه متن) می‌توان فاصلهٔ بین آن‌ها را به کمک روش‌های متفاوت فاصله‌یابی محاسبه کرد و از این فواصل برای کاربردهای مختلف همچون بازیابی اطلاعات استفاده کرد.



(شکل-۱۶): مراحل به‌کارگیری معماری CLIP برای دسته‌بندی تلاش نخست

(Figure-16): Diagram to use the CLIP architecture for Zero-shot classification

در **Error! Reference source not found.** شاهد مراحل به‌کارگیری از معماری CLIP برای دسته‌بندی تلاش نخست یک مجموعه‌دادهٔ تصاویر و برچسب آن‌ها پس از آموزش هستیم. با داشتن مجموعه‌داده‌ای از تصاویر

[1] Multi-Class Cross Entropy Loss

[DOI: 10.61186/jsdp.22.1.113]

و برچسب آن‌ها، در ابتدا هر برچسب را تبدیل به متنی ساده و سپس به‌کمک رمزگذار متنی CLIP، متن یادشده را به بردار ویژگی متنی و به همین ترتیب، تصویر مربوط به برچسب را نیز از طریق رمزگذار تصویری CLIP به بردار ویژگی تصویر تبدیل می‌کنیم. با داشتن ویژگی‌های یادشده، می‌توان فاصلهٔ بردار تصویر را با تمامی متون حاصل از برچسب‌های مجموعه‌داده محاسبه و برچسبی را که بردار متنی‌اش کمترین فاصله را با بردار تصویر دارد، به‌عنوان برچسب مناسب آن انتخاب کرد. بدین ترتیب بدون نیاز به آموزش بر روی مجموعه‌داده‌ای با برچسب‌های یادشده، می‌توان مدل CLIP را برای پیش‌بینی برچسب درست هر مجموعه‌دادهٔ تصویری و بازیابی تصاویر استفاده کرد.

تمرکز بر جمع‌آوری مجموعه‌داده‌های به‌روز و در مقیاس عظیم، موجب افزایش دقت و عملکرد مدل‌های مبتنی بر معماری CLIP شده‌است. این افزایش عملکرد در طی سه سال از معرفی راه‌کار CLIP، از دقت ۵۹.۶٪ در سال ۲۰۲۱ به دقت ۸۸.۵٪ در سال ۲۰۲۴ برای دسته‌بندی تلاشِ نخست تصاویرِ درونِ مجموعه‌دادهٔ ImageNet (با یک میلیون تصویر) رسیده‌است[۱۰۳]. نگاهی به معماری داخلی راه‌کارهای ارائه‌شده در طی این مدت، بیشتر نشان از تمرکز بر مهندسی معماری ارتباطی مدل‌های شناخته‌شده بر روی سخت‌افزار و چگونگی آموزش آن‌ها در مقیاس‌های عظیم می‌دهد؛ به تعبیر دیگر، این بخش از پژوهش‌ها بیش از اینکه شاهد معرفی راه‌کارهای نوین و کم هزینه‌تر از راه‌کاری همچون CLIP باشد، خود را درگیر مدل‌های بزرگ‌تر، ژرف‌تر و آموزش‌دیده بر روی مجموعه‌داده‌های بسیار بزرگ و کمابیش دست‌نیافتنی برای بخش‌های پژوهش‌هایی کوچک‌تر (همچون دانشگاه‌ها و شرکت‌های فنی غیر بین‌المللی) کرده‌است[۹۷].

در (جدول-۵) ویژگی‌های روش‌های اصلی برای دسته‌بندی تصاویر مجموعه‌دادهٔ ImageNet خلاصه شده‌است. گفتنی این‌که عملکرد این روش‌ها، بر مبنای «دقت تلاش نخست» آن‌ها در دسته‌بندی تصاویر گزارش شده‌است.

همچنان که ملاحظه می‌شود، در حال حاضر حتی روش‌های آموزش نظارت‌شده‌ای همچون Swin Transformer V2 (با دقت ۹۰.۱۷٪ و سه میلیارد پارامتر) نیز قابلیت رقابت با روش‌های آموزش خودنظارتی[1] و به‌ویژه روش‌های خودنظارتی مبتنی بر ارتباط بین تصویر و متن را ندارند و این روش‌ها در صدر جدول عملکرد بازیابی تصاویر بر روی مجموعه‌دادهٔ ImageNet هستند که نشان از گذر پژوهش‌های این حوزه از آموزش‌های نظارتی به آموزش‌های خودنظارتی است.

به نظر می‌رسد برای دست‌یابی به بازیابی‌های موفقی که مورد تأیید کاربر باشد، نیاز به درک هر چه بهتر و بیشتر دنیای اطراف داریم و گام اساسی در این مبحث، دست‌یابی به راه‌کاری مؤثر برای ایجاد ارتباط بین انسان و ماشین محسوب می‌شود که این پل ارتباطی، در وهلهٔ نخست زبان طبیعی و در مراحل بعدی، ارتباط مغزی مستقیم خواهد بود تا بتوان نیت اصلی کاربر را برای بازیابی یا خلق آن چیزی که مدنظر دارد، تشخیص داد. مقالهٔ CLIP [۹۶] نشان داد که حتی با به‌کارگیری شبکه‌های عصبی که نزدیک به یک دهه از معرفی آن‌ها می‌گذرد، می‌توان مدل‌هایی موفق و با دقت تا اندازه‌ای زیاد روی مجموعه‌داده‌های میلیاردی آموزش داد که قادر به تعبیهٔ تصاویر و متون در فضای یکسان باشند و حل این مسئله، بیش از اینکه درگیر طراحی مدلی بهتر باشد (که بی‌شک مؤثر است)، وابسته به درک بهتر از دنیای اطراف ماست که نیازمند مجموعه‌داده‌هایی عظیم‌تر، با پوشش هر چه بیشتر از دنیای واقعی، با افزونگی زیاد و در کنار هزینهٔ سنگین آموزش چنین مدل‌هایی است.

۳- مدل بازخورد ارتباط

فرایند «بازخورد ارتباط»[2] فرایندی است که در بسیاری از سامانه‌های بازیابی اطلاعات به‌کار گرفته می‌شود و دقت جست‌وجو و بازیابی و رضایت کاربر را بهبود می‌بخشد. علت شکل‌گیری این حوزهٔ پژوهشی، به‌دلیل ضعف مدل‌های بازیابی اطلاعات بوده‌است. در بسیاری از سناریوهای دنیای واقعی، تصاویر بازیابی‌شده پاسخ‌گوی نیاز واقعیِ کاربر نیستند و نیاز به راه‌کاری برای هدایت سامانه بازیابی به سمت بازیابی‌های با دقت بالاتر و هم‌سو با نیاز واقعی کاربر است. سامانه‌های «بازخورد ارتباط»، با گرفتن بازخوردی از کاربر، تلاش در پالایش جست‌وجو و بازیابی تصاویر مناسب‌تری با نیاز کاربر در چندین مرحله دارند. علت نیاز به چنین سامانه‌هایی در این حقیقت نهفته است که رابط‌های کاربری سامانه‌های بازیابی تصاویر، بیشتر انعطاف کافی برای بازیابی‌های دقیق‌تر و مرتبط با نیاز واقعی ندارند.

با معرفی مدل‌های نوینی که از تلفیق دامنه‌های تصویری و زبانی بهره می‌برند، انعطاف سامانه‌های بازیابی در حوزهٔ یافتن نیاز واقعی کاربر افزایش یافته است. بدین

[1] Self-Supervised Learning

[2] Relevance Feedback

[DOI: 10.61186/jsdp.22.1.113]

ترتیب در صورتی‌که بازیابی مورد پسند کاربر نباشد، تنها لازم است کاربر توصیف زبانی خود را از آنچه بازیابی شده‌است تغییر دهد و جزئیات بیشتری به این توصیف اضافه کند تا سامانه با دقت بهتری بازیابی را انجام دهد؛ بااین‌حال این راه‌کار همیشه پاسخ‌گو نیست؛ برای مثال ممکن است با سامانه‌ای طرف باشیم که پشتیبانی از قابلیت زبانی نداشته باشد. در این سناریو، به راه‌کاری نیاز است تا بتوان فارغ از فعل‌وانفعالات داخلی یک مدل، خروجی‌های به‌دست‌آمده از آن را به کمک راه‌کاری مستقل پالایش و آن‌ها را تا جای ممکن هم‌سو با نظر کاربر بازیابی کرد. با توجه به این موضوع، هنوز نمی‌توان ادعا کرد که راه‌کارهای سنتی بازخورد ارتباط کارایی خود را از دست داده‌اند و به همین دلیل در این قسمت، به بررسی راه‌کار کلی این روش‌ها خواهیم پرداخت و نمونه‌هایی از آن را بیان خواهیم کرد.

راه‌کارهای بازخورد ارتباط، به دو دسته تقسیم می‌شوند: راه‌کارهای بازیابی سریع و کوتاه‌مدت و راه‌کارهای بازیابی بلندمدت و با توجه به ذائقهٔ کلی کاربر. اشکال اصلی راه‌کارهای کوتاه‌مدت، کمبود اطلاعات مورد نیاز برای شناسایی نظر اصلی کاربر است. در سامانه‌های بازیابی مبتنی بر متن این مشکل وجود ندارد؛ زیرا خود کاربر می‌تواند نظرش را با دقت و جزئیات قابل توجه بیان کند، اما در صورت نبود چنین قابلیتی یا عملکرد ضعیف آن، نیاز به جمع‌آوری داده‌های اولیه برای تشخیص علت اشتباه در بازیابی و اصلاح آن است و برای رسیدن به این داده‌ها، کاربر، با بازیابی‌های شکست‌خورده و نه چندان مناسب روبه‌رو خواهد شد.

مراحل بازخورد ارتباط، به شرح زیر است[۱۱۱]:

۱. کاربر پرس‌وجوی خویش را از سامانه می‌پرسد.
۲. سامانه درخواست را پردازش می‌کند و نزدیک‌ترین همسایه‌های یافت‌شده برای آن را به کاربر نمایش می‌دهد.
۳. کاربر در بین تصاویر بازیابی‌شده، موارد مثبت (مرتبط با درخواست خود) و منفی (بی‌ربط با درخواست) را مشخص می‌کند.
۴. بر اساس این بازخورد، سامانه، فرایند بازیابی و پالایش نزدیک‌ترین همسایه‌ها را انجام داده و تصاویر جدیدی را برای کاربر بازیابی می‌کند.

این مراحل تا زمان رسیدن سامانه به یک بازیابی متناسب با نیاز کاربر ادامه پیدا می‌کند.

در ادامه به معرفی راه‌کارهای اصلی برای تعیین تناسب میان درخواست و بازیابی می‌پردازیم:

بازخورد ارتباط به کمک دسته‌بند: این راه‌کار که از محبوب‌ترین روش‌های بازخورد محسوب می‌شود، تمرکز خود را بر پالایش بازیابی‌های به‌دست‌آمده از سامانه به‌کمک یک دسته‌بند (همچون دسته‌بند SVM) [۱۱۲]. قرار می‌دهند. در این حالت، بیشتر بازیابی دارای چند نمونهٔ مثبت و چند نمونهٔ منفی از نظر کاربر است و به‌صورت نظری می‌توان انتظار داشت که با تحویل‌دادن این نمونه‌ها بتوان مدلی را آموزش داد که قادر به درک ذائقهٔ کاربر در بازیابی بهتر تصاویر باشد. مشکل اصلی این روش، نامتوازن‌بودن[1] نمونه‌های مثبت و منفی و تعداد کم آن‌ها در مقایسه با مجموعه تصاویر کلی است که موجب آموزش نادرست دسته‌بندها می‌شود. روش‌های متفاوتی برای حل مسئلهٔ عدم توازن در این حوزه معرفی شده‌اند که هر کدام به صورتی نسبی در حل این مسئله موفق بوده‌اند[۱۱۳-۱۱۵].

بازخورد ارتباط به‌کمک تحلیل افتراقی[2]: این روش‌ها نیز مشابه روش‌های مبتنی بر دسته‌بند، تلاش در تحلیل بازخورد کاربر به‌کمک روش‌های تحلیل افتراقی دارند و پس از یادگیری، بازیابی‌های بعدی را چنان وزن‌دهی می‌کنند تا ویژگی‌های موافق با نظر کاربر بیش از دیگر ویژگی‌ها در بازیابی تأثیرگذار شوند[۱۱۶]. این روش‌ها نیز نقاط ضعف روش‌های مبتنی بر دسته‌بند دارند.

بازخورد ارتباط به کمک روش‌های فرا ابتکاری[3]: هر زمانی که بازیابی‌های متعددی داریم که باید از میان آن‌ها بهترین را برگزید، می‌توان از روش‌های فراابتکاری همچون الگوریتم ژنتیک[4]، هوش ازدحامی[5]، تبرید شبیه‌سازی‌شده[6] و تپه‌نوردی[7] استفاده کرد؛ برای مثال با به‌کارگیری الگوریتم ژنتیک، سامانهٔ بازیابی پس از دریافت بازخورد کاربر، جمعیتی از تصاویر را مبتنی بر پرس‌وجوی کاربر بازیابی کرده و از بین آن‌ها مستعدترین جمعیت را انتخاب می‌کند (میزان مستعدبودن جمعیت نیز از کاربر گرفته می‌شود)؛ سپس به‌کمک الگوریتم ژنتیک تصاویر جدیدی بازیابی می‌شود و این فرایند آنقدر ادامه پیدا می‌کند تا سامانه قادر به بازیابی بهترین نمونه‌ها برای کاربر باشد[۱۱۶، ۱۱۷]. مشکل اصلی روش‌های تکاملی، نیاز آن‌ها به انجام تعداد زیادی حلقهٔ محاسباتی برای رسیدن به پاسخ‌های مناسب است که در بسیاری از موارد بیش از زمانی است که یک کاربر برای یک بازیابی موفق انتظار دارد.

[1] Imbalance
[2] Discriminant Analysis
[3] Meta-Heuristics
[4] Genetic Algorithm
[5] Swarm Intelligence
[6] Simulated Annealing
[7] Hill Climbing

[DOI: 10.61186/jsdp.22.1.113]

(جدول-۵): مقایسهٔ روش‌های مطرح در سال‌های اخیر برای دسته‌بندی تصاویر مجموعه‌دادهٔ ImageNet (مرتب‌شده بر مبنای دقت)

(Table-5): Classification accuracy comparison of different approaches for the ImageNet dataset

روش	ویژگی	تعداد پارامتر	دقت↓	نوع آموزش			سال
				نظارتی	خودنظارتی فقط بر روی تصویر	خودنظارتی با تلفیق متن و تصویر	
OmniVec [۱۶]	▪ آموزش چندحالته بر روی تصویر، متن، صوت و داده‌های سه‌بعدی	ذکر نشده	۹۲.۴٪	×	×	✓	۲۰۲۴
CoCa [۱۰۴]	▪ آموزش دو حالته بر روی متن و تصویر ▪ به‌کارگیری تابع خطای یادگیری تضادی برای کم‌کردن فاصلهٔ تصاویر و متون ▪ به‌کارگیری تابع خطای توصیف تصاویر برای یادگیری توصیف‌کردن تصاویر در شبکه	۲.۱ میلیارد	۹۱.۰٪	×	×	✓	۲۰۲۲
Model Soups [۱۰۵]	▪ تلفیق و میانگین‌گیری از وزن‌های چندین مدل آموزش‌دیده با تنظیمات متفاوت بر روی یک مجموعه‌دادهٔ بزرگ	۲.۴ میلیارد	۹۰.۹۴٪	×	×	✓	۲۰۲۲
Swin Transformer V2 [۱۰۵]	▪ به‌کارگیری یکی از بزرگ‌ترین و سنگین‌ترین شبکه‌های مبدل با بیش از سه میلیارد پارامتر	۳ میلیارد	۹۰.۱۷٪	✓	×	×	۲۰۲۱
DINOv2 [۱۰۶]	▪ آموزش یک مدل سنگین بر روی مجموعه‌داده‌های با کیفیت و تقطیر آن در مدل‌های کوچک‌تر	۱ میلیارد	۸۶.۷٪	×	✓	×	۲۰۲۳
M2-Encoder [۱۰۳]	▪ آموزش بر روی مجموعه‌داده‌ای شش میلیاردی متشکل از تصویر و متن به زبان‌های انگلیسی و چینی	۱۰ میلیارد	۸۸.۵٪	×	×	✓	۲۰۲۴
BASIC [۱۰۷]	▪ آموزش بر روی مجموعه‌داده‌ای شانزده برابر بزرگتر، مدلی ۳.۷۵ برابر بزرگتر و اندازهٔ دسته‌ای دو برابر بزرگتر از مدل CLIP از OpenAI	۲.۴ میلیارد	۸۸.۳٪	×	×	✓	۲۰۲۱
LiT-22B [۱۰۸]	▪ افزایش مقیاس مدل‌های مبدل بصری از میانگین چهار میلیارد پارامتری به ۲۲ میلیارد پارامتر	۲۲ میلیارد	۸۵.۹٪	×	✓	×	۲۰۲۳
MIM Refiner [۱۰۹]	▪ قابل رقابت با مدل‌هایی همچون DINOv2 که بر روی دوهزار برابر دادهٔ بیشتر آموزش دیده‌اند	۶۳۲ میلیون	۸۴.۷٪	×	✓	×	۲۰۲۴
UNICOM [۱۱۰]	▪ به‌کارگیری تابع خطای ArcFace برای یادگیری متریک فاصلهٔ بین خوشه‌های مجموعه‌داده	۶۳۲ میلیون	۸۲.۷٪	×	×	✓	۲۰۲۳
CLIP [۹۶]	▪ راه‌کار اولیه از OpenAI برای ارتباط فضای دو دامنهٔ متن و تصویر	۲۰۰ میلیون	۵۹.۶٪	×	×	✓	۲۰۲۱

۴- مجموعه‌داده‌ها

به‌دلیل شباهت مفهومی، بسیاری از مجموعه‌داده‌های مطرح در حوزه‌های متفاوت پردازش تصویر و بینایی ماشینی، قابل به‌کارگیری در حوزهٔ بازیابی محتوامحور تصویر نیز هستند؛ به‌طورکلی می‌توان مجموعه‌داده‌های این حوزه را به دو دستهٔ «دارای توضیحات» و «بدون توضیحات» تقسیم کرد. این تقسیم‌بندی در گذشته که مدل‌های زبانی در فرایند بازیابی نقشی نداشتند، وجود نداشت؛ اما با رشد روش‌های آموزش چندحالته[1] و تلفیق دامنه‌های تصویری، ویدئویی و زبانی با یکدیگر، این نوع تقسیم‌بندی نیز موضوعیت یافته است. در جدول-۶)، مهم‌ترین ویژگی این مجموعه‌داده‌ها آورده شده‌است.

۵. تحلیل و جمع‌بندی

در این پژوهش، مهم‌ترین موضوعات مربوط به حوزهٔ «بازیابی محتوامحور تصاویر» و پژوهش‌های صورت‌گرفته

[1] Multi-Modal Learning

[DOI: 10.61186/jsdp.22.1.113]

در سال‌های اخیر و روش‌های بازیابی کلاسیک و نوین به‌صورت مشروح مورد بررسی قرار گرفت. با توجه به پیشرفت‌های اخیر در حوزهٔ بینایی ماشینی و پردازش تصویر به‌ویژه در حوزهٔ «ارتباط تصویر و متن و چگونگی تلفیق آن دو برای افزایش عملکرد بازیابی»، تمرکز بخش اعظمی از این مطالعه، بر راه‌کارهای این حوزه و عملکرد روش‌های مطرح معطوف شده‌است.

بخش نخست مقاله با تعریف مسئله و توضیح سه سطح اصلی برای بازیابی تصاویر، به تشریح مفهوم «بازیابی معناگرا» اختصاص یافت. در ادامه با داشتن درک درستی از مفهوم «معنا» در مبحث بازیابی، به نقش «زبان طبیعی» در افزایش کیفیت و عملکرد بازیابی تصویری پرداخته و آن را جزئی جدایی‌ناپذیر از فرایند «بازیابی معناگرای» تصویر دانستیم. در بخش دوم، به پیشینهٔ پژوهش و تاریخچهٔ پژوهش‌های دیگر در این حوزه پرداخته شد. در ابتدا با بیان سطوح مختلف عملکردی سامانه‌های بازیابی محتوامحور، قلمرو عملکرد آن‌ها را از یکدیگر جدا کرده و نقطه ضعف آن‌ها در مبحث «بازیابی معناگرا» مورد بحث قرار گرفت. در ادامه معماری سامانه‌های بازیابی محتوامحور و عملکرد عناصر مختلف آن تشریح شد و به‌کمک آن، راه‌کارهای کلاسیک استخراج ویژگی سطح پایین برای بازیابی تصاویر و نقاط قوت و ضعف آن‌ها تشریح شدند.

در ادامه بازیابی محتوامحور با تکیه بر ویژگی‌های سطح بالای درون تصویر مورد توجه قرار گرفت و روش‌های نوین مطرح در این حوزه بررسی شدند. با نگاهی بر این روش‌ها، مشخص شد تمرکز صرف بر کاربرد ویژگی‌های تصویری برای بازیابی موفق کافی نیست و نیاز به تلفیق قلمرو زبان طبیعی با قلمرو تصاویر است و به همین دلیل به تشریح راه‌کارهای مطرح در این حوزه و به خصوص روش CLIP برای ارتباط بین این دو دامنه پرداخته شد که راه‌کاری مشترک بین روش‌های نوین استخراج ویژگی برای بازیابی محتوامحور تصاویر است.

در بخش سوم مفهوم «بازخورد ارتباط» را معرفی و علت نیاز به این روش برای افزایش عملکرد بازیابی در سامانه‌های بازیابی محتوامحور را شرح دادیم و در ادامه به معرفی راه‌کارهای مطرح در این حوزه پرداختیم. در بخش چهارم و نهایی نیز مجموعه‌داده‌های مورد استفاده در حوزهٔ پژوهشی «بازیابی محتوامحور تصاویر» را معرفی کردیم.

در حال حاضر، پژوهش‌های اصلی در این حوزه در انحصار شرکت‌ها و سازمان‌های بزرگ با دسترسی به منابع مالی عظیم است و همین موضوع موجب کندشدن پیشرفت‌های پژوهشی و دانشگاهی در این عرصه شده‌است. این شرکت‌ها با دسترسی به منابع داده‌ای و مالی غیرقابل تصور، به آموزش مدل‌های شناخته‌شده و گاه ناشناخته و در مقیاس‌های بسیار بزرگ (میلیاردها تصویر و متن) پرداخته و پس از اتمام آموزش، بدون انتشار جزئیات پژوهش‌های انجام‌شده، مدل نهایی را در خدمات وب‌های متفاوتی قرار می‌دهند. نکتهٔ مهم اینکه قانون مقیاس[1] در این حوزه برقرار است و هر موجودیتی که دسترسی بیشتری به منابع محاسباتی و ذخیره‌سازی داشته باشد، قادر به آموزش مدل‌هایی بهتر و دقیق‌تر خواهد بود و این موضوع باعث شده‌است واحدهای پژوهشی کوچک و دانشگاه‌ها کمتر بتوانند در این حوزه وارد شوند و منتظر منتشرشدن پژوهش‌های سازمان‌ها و شرکت‌های یاد شده باشند. در حال حاضر نیاز شدیدی به معرفی راه‌کارهایی کارا در این حوزه است که نیازمند منابع محدود بوده تا قادر به دست‌یابی به دقت‌های بالا و قابل رقابت با مدل‌های عظیم و با کسری از بودجهٔ مورد نیاز برای آموزش آن‌ها باشد. این اتفاق در حوزهٔ مدل‌های زبانی بزرگ رخ داده و پس از دو سال، گروه‌های پژوهشی متعددی با معرفی روش‌های کارا، قادر به رسیدن به دقت مدل زبانی Chat-GPT4 از شرکت OpenAI و با قابلیت اجرا بر روی دستگاه‌های خانگی شده‌اند [۱۲۸] و نیاز است که پژوهش‌هایی در این حوزه از تمرکز بر رسیدن به دقت با مقیاس بیشتر، به تمرکز بر رسیدن به دقت با هزینهٔ کمتر متمایل شود وگرنه این حوزه همچنان در انحصار شرکت‌هایی متمرکز بر سود بیشتر باقی خواهد ماند.

با توجه به دلایل یادشده، پژوهش‌گران بر ایجاد ارتباط بین دامنه‌های مختلف متمرکز شده‌اند و به همین دلیل، پژوهش‌های این حوزه نیازمند توجه بیشتری به حوزه‌های زیر است:

۱.افزایش عملکرد و کارایی در مقیاس

الف) معرفی روش‌های کارا و بهینه در آموزش مدل‌های عظیم با هدف کاهش نیازهای محاسباتی عظیم این مدل‌ها

ب) معرفی راه‌کارهای جدید هرس[2] که قادر به حفظ اطلاعات ارزشمند مدل در عین کاهش تعداد پارامترهای آن باشد.

ج) معرفی روش‌های کوانتیزه‌کردن[3] مدل بدون توجه به معماری زیرین سخت‌افزار و با قابلیت اجرا روی سخت‌افزار ثالث برای کاهش بیت‌های مورد استفادهٔ هر وزن شبکهٔ عصبی از ۳۲ بیت به شانزده بیت و کمتر و با هدف کاهش میزان حافظهٔ مورد نیاز مدل و افزایش سرعت استنتاج آن

۲.افزایش عمومیت مدل‌ها

الف) معرفی راه‌کارهایی برای «گسترش دامنهٔ»[4] عملکرد مدل‌ها برای پشتیبانی از داده‌هایی که تابه‌حال بر آن‌ها آموزش ندیده‌است.

[1] Scalability Law
[2] Pruning
[3] Quantization
[4] Domain Expansion & Adaptation

[DOI: 10.61186/jsdp.22.1.113]

ب) ارائهٔ راه‌کارهایی برای آموزش چندحالته برای هم‌جوشی دامنه‌های جدید (همچون تصویر، متن، صوت، ویدئو و حتی بو) با هدف افزایش دقت و عملکرد

(جدول-۶): مقایسهٔ مجموعه‌داده‌های حوزهٔ بازیابی محتوامحور تصویر

(Table-6): Comparison of datasets used in the evaluation of CBIR methods

مجموعه‌داده	تعداد تصاویر	تعداد توصیف	تعداد دسته‌ها
LAION-5B [۱۱۸]	5×10^9	5×10^9	بدون دسته
LAION-400M[۱۱۹]	4×10^8	4×10^8	بدون دسته
WIT [۱۲۰]	37.8×10^6	17×10^6	بدون دسته
ImageNet [۱۲۱]	14×10^6	بدون توصیف	≈۲۱۰۰۰
Conceptual Captions [۱۲۲]	3.3×10^6	3.3×10^6	بدون دسته
VQA V2 [۱۲۳]	≈۲۰۴۰۰۰	$\approx 1 \times 10^6$	بدون دسته
MS-COCO [۱۲۴]	≈۱۲۳۰۰۰	≈۶۱۶۰۰۰	بدون دسته
Visual [۲۳]Genome	≈۱۰۸۰۰۰	$\approx 2 \times 10^6$	بدون دسته
[۱۲۵]CIFAR	6×10^5	بدون توصیف	۱۰
FLICKR-30K[۱۲۶]	۳۱۰۰۰	16×10^4	بدون دسته
PASCAL-VOC2007 [۱۲۷]	≈۱۰۰۰۰	بدون توصیف	۲۰

۳.کاربرد در روش‌های نوین

الف) به‌کارگیری مدل‌های هم‌جوشی برای خلق محتوای جدید در مدل‌های متن به تصویر، تصویر به تصویر، متن به ویدئو و تصویر به ویدئو

ب) توسعهٔ روش‌هایی برای تفسیر عملکرد داخلی مدل‌های بازیابی و توضیح عملکرد آن‌ها و پیش‌بینی‌پذیرکردن رفتارشان

6-References ۶-مراجع

[1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *J. Vis.*, vol. 7, no. 1, p. 10, Jan. 2007, doi: 10.1167/7.1.10.

[2] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval Using Deep Learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2687–2704, 2022, doi: 10.1109/TCSVT.2021.3080920.

[3] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 20–54, 2015, doi: 10.1016/j.jvcir.2015.07.012.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[5] C. Li *et al.*, "mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections," *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022*, pp. 7241–7259, 2022.

[6] R. Mihalcea, "The multidisciplinary facets of research on humour," *Appl. Fuzzy Sets Theory*, pp. 412–421, 2007, doi: 10.1007/978-3-540-73400-0_52.

[7] A. Chandrasekaran *et al.*, "We Are Humor Beings: Understanding and Predicting Visual Humor," Dec. 2015.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," 2022, [Online]. Available: http://arxiv.org/abs/2204.06125

[9] "Internet LiveStats 2024." [Online]. Available: https://www.internetlivestats.com

[10] "DemandSage 2024." [Online]. Available: https://www.demandsage.com/social-media-users/

[11] "Gartner 2024." [Online]. Available: https://www.gartner.com

[12] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999, doi: 10.1006/jvci.1999.0413.

[13] D. Stan and I. K. Sethi, "Mapping low-level image features to semantic concepts," M. M. Yeung, C.-S. Li, and R. W. Lienhart, Eds., Jan. 2001, pp. 172–179. doi: 10.1117/12.410925.

[14] R. Datta, J. Li, and J. Z. Wang, "Content-based image retrieval: approaches and trends of the new age," in *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, New York, NY, USA: ACM, Nov. 2005, pp. 253–262. doi: 10.1145/1101826.1101866.

[15] A. Mojsilovic and B. Rogowitz, "Capturing image semantics with low-level descriptors," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, IEEE, pp. 18–21. doi: 10.1109/ICIP.2001.958942.

[16] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000, doi: 10.1109/34.895972.

[17] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Common sense computing: from the society of mind to digital intuition and beyond," in *European Workshop on Biometrics and Identity Management*, Springer, 2009, pp. 252–259.

[18] E. Perez, H. de Vries, F. Strub, V. Dumoulin, and A. Courville, "Learning Visual Reasoning

Without Strong Priors," 2017, [Online]. Available: http://arxiv.org/abs/1707.03017

[19] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning," 2016, doi: 10.1109/CVPR.2017.345.

[20] H. Ling and S. Fidler, "Teaching Machines to Describe Images via Natural Language Feedback," 2017, [Online]. Available: http://arxiv.org/abs/1706.00130

[21] A. Chandrasekaran *et al.*, "We Are Humor Beings: Understanding and Predicting Visual Humor," *Cvpr 2016*, p. 17, Dec. 2015, doi: 10.1109/CVPR.2016.498.

[22] D. Raposo, A. Santoro, R. Pascanu, T. Lillicrap, P. Battaglia, and U. Kingdom, "Discovering objects and their relations from entangled scene representations," *Iclr*, pp. 1–10, 2017.

[23] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.07332

[24] L. Ladický, J. Shi, and M. Pollefeys, "Pulling things out of perspective," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 89–96, 2014, doi: 10.1109/CVPR.2014.19.

[25] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting Image Captioning with Attributes," 2016, [Online]. Available: http://arxiv.org/abs/1611.01646

[26] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.02155

[27] N. Ghosh, S. Agrawal, and M. Motwani, "A Survey of Feature Extraction for Content-Based Image Retrieval System," *Lect. Notes Networks Syst.*, vol. 34, pp. 305–313, 2018, doi: 10.1007/978-981-10-8198-9_32.

[28] R. C. Gonzalez and R. E. Woods, *Digital Image Processing, Global Edition*. Pearson Education, 2018.

[29] J. Devlin *et al.*, "Language Models for Image Captioning : The Quirks and What Works," *Acl-2015*, no. Me Lm, pp. 100–105, 2015, doi: 10.1103/PhysRevE.92.022112.

[30] A. Karpathy and F. F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2015, pp. 3128–3137. doi: 10.1109/CVPR.2015.7298932.

[31] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," in *Proceedings of NIPS 2014*, 2014, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1406.5679

[32] E. H. Huang, R. Socher, C. D. Manning, and A. Ng, "Improving word representations via global context and multiple word prototypes," *Proc. 50th Annu. Meet. Assoc. Comput. Linguist. Long Pap. 1*, pp. 873–882, 2012, [Online]. Available: http://nlp.stanford.edu/pubs/HuangACL12.pdf

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Nips*, pp. 1–9, 2013, doi: 10.1162/jmlr.2003.3.4-5.951.

[34] J. Graham, T. Cootes, C. Taylor, and D. Cooper, "Active shape models-their training and application," *Comput. Vis. image Underst.*, vol. 61, 1995.

[35] S. Ali and A. Madabhushi, "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery," *IEEE Trans. Med. Imaging*, vol. 31, no. 7, pp. 1448–1460, 2012, doi: 10.1109/TMI.2012.2190089.

[36] Mehmet Sezgin Bu¨ lent Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[37] Y. Liang, M. Zhang, and W. N. Browne, "Image segmentation: A survey of methods based on evolutionary computation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8886, pp. 847–859, 2014, doi: 10.1007/978-3-319-13563-2_71.

[38] B. M. Carvalho, C. J. Gau, G. T. Herman, and T. Y. Kong, "Algorithms for Fuzzy Segmentation," *Int. Conf. Adv. Pattern Recognit.*, pp. 154–163, 1999, doi: 10.1007/978-1-4471-0833-7_16.

[39] B. M. Carvalho, G. T. Herman, and T. Y. Kong, "Simultaneous Fuzzy Segmentation of Multiple Objects," *Electron. Notes Discret. Math.*, vol. 12, pp. 3–22, 2003, doi: 10.1016/S1571-0653(04)00470-6.

[40] J. K. Udupa, P. K. Saha, and R. A. Lotufo, "Relative fuzzy connectedness and object definition: Theory, algorithms, and applications in image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1485–1500, 2002, doi: 10.1109/TPAMI.2002.1046162.

[41] Hanan Same, "The Quadtree and Related Hierarchical Data Structures," *ACM Comput. Surv.*, vol. 16, pp. 187–260, 1984.

[42] J. P. Marques de Sá, "Structural Pattern Recognition," *Pattern Recognit.*, pp. 243–289, 2001, doi: 10.1007/978-3-642-56651-6_6.

[43] I. Karoui, R. Fablet, J. M. Boucher, and J. M. Augustin, "Unsupervised region-based image segmentation using texture statistics and level-set methods," *2007 IEEE Int. Symp. Intell. Signal Process. WISP*, 2007, doi: 10.1109/WISP.2007.4447617.

[44] L. Grady, "Multilabel random walker image segmentation using prior models," *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, vol. I, pp. 763–771, 2005, doi: 10.1109/CVPR.2005.239.

[45] X. Yu and J. Yla-Jaaski, "A new algorithm for image segmentation based on region growing and edge detection," *Proc. - IEEE Int. Symp. Circuits Syst.*, vol. 1, pp. 516–519, 1991, doi: 10.1109/iscas.1991.176386.

[46] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004, doi: 10.1023/B:VISI.0000022288.19776.77.

[DOI: 10.61186/jsdp.22.1.113]

[58] X. Yang and K. T. Cheng, "Accelerating SURF detector on mobile devices," *MM 2012 - Proc. 20th ACM Int. Conf. Multimed.*, pp. 569–578, 2012, doi: 10.1145/2393347.2393427.
[59] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2003, doi: 10.1109/cvpr.2003.1211478.
[60] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," *arXiv Prepr.*, 2015, doi: 10.1109/CVPR.2016.494.
[61] K. Xu, J. L. B. R. Kiros, K. C. A. Courville, and R. S. R. S. Z. Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, Feb. 2015, doi: 10.1109/72.279181.
[62] J. Mao *et al.*, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," *To Appear ICLR-2015*, vol. 1090, no. 2014, pp. 1–14, 2015, [Online]. Available: http://cbmm.mit.edu/sites/default/files/publications/CBMM Memo 033.pdf
[63] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image Captioning with Semantic Attention," *Cvpr*, p. 10, 2016, doi: 10.1109/CVPR.2016.503.
[64] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1724–1734, 2014, doi: 10.3115/v1/D14-1179.
[65] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning," *arXiv Prepr.*, pp. 1–6, 2015, [Online]. Available: http://arxiv.org/abs/1505.04467
[66] M. Kolář, M. Hradiš, and P. Zemčík, "Technical Report: Image Captioning with Semantically Similar Images," p. 3, 2015, [Online]. Available: http://arxiv.org/abs/1506.03995
[67] J.-B. Michel *et al.*, "Quantitative analysis of culture using millions of digitized books.," *Science*, vol. 331, no. 6014, pp. 176–82, 2011, doi: 10.1126/science.1199644.
[68] M. J. Choi, A. Torralba, and A. S. Willsky, "Context models and out-of-context objects," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 853–862, 2012, doi: 10.1016/j.patrec.2011.12.004.
[69] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.
[70] A. Zhang *et al.*, "Fine-Grained Scene Graph Generation with Data Transfer," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13687 LNCS, pp. 409–424, 2022, doi: 10.1007/978-3-031-19812-0_24.
[71] J. Kim, J. Park, J. Park, J. Kim, S. Kim, and H. J. Kim, "Groupwise Query Specialization and
[47] L. Lucchese and S. K. Mitray, "Color image segmentation: A state-of-the-art survey," *Proc. Indian Natl. Sci. Acad. (INSA-A). Delhi, Indian Natl Sci Acad*, vol. 67, pp. 207–221, 2001, [Online]. Available: http://ultra.sdk.free.fr/docs/Image-Processing/filters/Color Image Segmentation-A State-of-the-Art Survey.pdf%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.4896
[48] R. Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1101–1113, 1993, doi: 10.1109/34.244673.
[49] L. Grady and G. Funka-Lea, "Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3117, pp. 230–245, 2004, doi: 10.1007/978-3-540-27816-0_20.
[50] L. Grady, "Random Walks for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, 2006.
[51] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008, doi: 10.1109/CVPR.2008.4587586.
[52] C. H. Lin, R. T. Chen, and Y. K. Chan, "A smart content-based image retrieval system based on color and texture feature," *Image Vis. Comput.*, vol. 27, no. 6, pp. 658–665, 2009, doi: 10.1016/j.imavis.2008.07.004.
[53] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008, doi: 10.1109/CVPR.2008.4587633.
[54] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients A low bit-rate feature descriptor," pp. 2504–2511, 2010, doi: 10.1109/cvpr.2009.5206733.
[55] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognit.*, vol. 43, no. 3, pp. 706–719, 2010.
[56] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, 2010.
[57] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4778 LNCS, pp. 168–182, 2007, doi: 10.1007/978-3-540-75690-3_13.

[DOI: 10.61186/jsdp.22.1.113]

[87] C. Shen, S. Panda, and J. T. Vogelstein, "The Chi-Square Test of Distance Correlation," *J. Comput. Graph. Stat.*, vol. 31, no. 1, pp. 254–262, 2022, doi: 10.1080/10618600.2021.1938585.
[88] C. Spearman, "The Proof and Measurement of Association between Two Things," *Am. J. Psychol.*, vol. 15, no. 1, p. 72, 1904, doi: 10.2307/1412159.
[89] G. N. Lance and W. T. Williams, *Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses")*, vol. 9, no. 1. 1966. doi: 10.1093/comjnl/9.1.60.
[90] J. Shlens, "Notes on Kullback-Leibler Divergence and Likelihood," 2014, [Online]. Available: http://arxiv.org/abs/1404.2000
[91] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between euclidean and cosine angle distance for nearest neighbor queries," *Proc. ACM Symp. Appl. Comput.*, vol. 2, pp. 1232–1237, 2004, doi: 10.1145/967900.968151.
[92] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[93] Y. Merri, Bart Van; Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," 1997.
[94] A. Vaswani *et al.*, "Attention Is All You Need," *Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762
[95] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.
[96] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00020
[97] Z. Jiang *et al.*, "MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs," 2024, [Online]. Available: http://arxiv.org/abs/2402.15627
[98] D. MEYER, "The cost of training AI could soon become too much to bear." [Online]. Available: https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/
[99] J. Johnson, A. Gupta, and L. Fei-Fei, "Image Generation from Scene Graphs," 2018, [Online]. Available: http://arxiv.org/abs/1804.01622
[100] X. An *et al.*, "Unicom: Universal and Compact Representation Learning for Image Retrieval," 2023, [Online]. Available: http://arxiv.org/abs/2304.05884
[101] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2018, [Online]. Available: http://arxiv.org/abs/1801.07698
[102] B. Dhingra, H. Liu, R. Salakhutdinov, and W. W. Cohen, "A Comparative Study of Word Embeddings for Reading Comprehension," Quality-Aware Multi-Assignment for Transformer-based Visual Relationship Detection," 2024, [Online]. Available: http://arxiv.org/abs/2403.17709
[72] A. Krizhevsky, Ii. Sulskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Nips*, 2012, pp. 1–9.
[73] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689 LNCS, no. PART 1, Springer, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.
[74] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.
[75] D. Lay, *Linear algebra and its applications*, vol. 41. 2016. doi: 10.1016/0024-3795(81)90106-3.
[76] Schrutka, "Geometrie der Zahlen," *Monatshefte für Math. und Phys.*, vol. 22, no. 1, pp. A30–A30, 1911, doi: 10.1007/bf01742861.
[77] Amit Singhal, "Modern information retrieval: a brief overview," *Bull. Ieee Comput. Soc. Tech. Comm. Data Eng.*, vol. 24, 2001.
[78] H. Minkowski, "The Fundamental Equations for Electromagnetic Processes in Moving Bodies," *Math. Klasse*, pp. 53–111, 1908.
[79] É.O. Rodrigues, "Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier," *Pattern Recognit. Lett.*, vol. 110, pp. 66–71, 2018.
[80] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, 1950, doi: 10.1002/j.1538-7305.1950.tb00463.x.
[81] I. Levenshtein, Vladimir, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Sov. Phys. Dokl.*, vol. 10, p. 707, 1966.
[82] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull. la Société Vaudoise des Sci. Nat.*, vol. 37, pp. 547–579, 1901.
[83] G. Van Brummelen, *Heavenly mathematics: The forgotten art of spherical trigonometry*. 2012. doi: 10.33137/aestimatio.v11i0.26065.
[84] P. C. Mahalanobis, "On the Generalised Distance in Statistics," *Proc. Natl. Inst. Sci. India*, vol. 2, pp. 49–55, 1936.
[85] K. Pearson, *Mathematical contributions to the theory of evolution*, vol. 60, no. 1834. 1896. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=aIU_AQAAIAAJ&oi=fnd&pg=PA1&dq=Mathematical+Contributions+to+the+Theory+of+Evolution&ots=6q0ynawAzT&sig=FdqqMWpdG0a5gRGfvPbW2BRUw8I
[86] Manning C.D. and Schutze H., "Foundations of statistical natural language processing.," *MIT Press*, 1999.

[DOI: 10.61186/jsdp.22.1.113]

"Relevance feedback based on genetic programming for image retrieval," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 27–37, 2011, doi: 10.1016/j.patrec.2010.05.015.
[118] C. Schuhmann *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," 2022, [Online]. Available: http://arxiv.org/abs/2210.08402
[119] C. Schuhmann *et al.*, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," 2021, [Online]. Available: http://arxiv.org/abs/2111.02114
[120] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning," *SIGIR 2021 - Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 2443–2449, 2021, doi: 10.1145/3404835.3463257.
[121] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
[122] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 2556–2565, 2018, doi: 10.18653/v1/p18-1238.
[123] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 398–414, 2019, doi: 10.1007/s11263-018-1116-0.
[124] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.
[125] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *... Sci. Dep. Univ. Toronto, Tech. ...*, pp. 1–60, 2009, doi: 10.1.1.222.9220.
[126] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. April, pp. 67–78, 2014, [Online]. Available: http://nlp.cs.illinois.edu/HockenmaierGroup/Papers/DenotationGraph.pdf
[127] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010, doi: 10.1007/s11263-009-0275-4.
[128] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971
2017, [Online]. Available: http://arxiv.org/abs/1703.00993
[103] Q. Guo *et al.*, "M2-Encoder: Advancing Bilingual Image-Text Understanding by Large-scale Efficient Pretraining," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.15896
[104] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive Captioners are Image-Text Foundation Models," 2022, [Online]. Available: http://arxiv.org/abs/2205.01917
[105] M. Wortsman *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," *Proc. Mach. Learn. Res.*, vol. 162, pp. 23965–23998, 2022.
[106] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," 2023, [Online]. Available: http://arxiv.org/abs/2304.07193
[107] H. Pham *et al.*, "Combined scaling for zero-shot transfer learning," *Neurocomputing*, vol. 555, 2023, doi: 10.1016/j.neucom.2023.126658.
[108] M. Dehghani *et al.*, "Scaling Vision Transformers to 22 Billion Parameters," *Proc. Mach. Learn. Res.*, vol. 202, pp. 7480–7512, 2023.
[109] B. Alkin, L. Miklautz, S. Hochreiter, and J. Brandstetter, "MIM-Refiner: A Contrastive Learning Boost from Intermediate Pre-Trained Representations," 2024, [Online]. Available: http://arxiv.org/abs/2402.10093
[110] X. An *et al.*, "Unicom: Universal and Compact Representation Learning for Image Retrieval," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.05884
[111] X. S. Zhou, T. S. Huang, and N. M. Ave, "Relevance Feedback in Image Retrieval: A Comprehensive Review," *ACM Multimed. Syst. J.*, vol. 544, no. 2003, pp. 536–544, 2001.
[112] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[113] L. Zhang, L. Wang, and W. Lin, "Semisupervised biased maximum margin analysis for interactive image retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2294–2308, 2012, doi: 10.1109/TIP.2011.2177846.
[114] W. Bian and D. Tao, "Biased discriminant euclidean embedding for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 545–554, 2010, doi: 10.1109/TIP.2009.2035223.
[115] G. T. Ngo, T. Q. Ngo, and D. D. Nguyen, "Image Retrieval with Relevance Feedback using SVM Active Learning," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 6, p. 3238, 2016, doi: 10.11591/ijece.v6i6.pp3238-3246.
[116] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using BiasMap," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, doi: 10.1109/cvpr.2001.990450.
[117] C. D. Ferreira, J. A. Santos, R. Da, M. A. Gonalves, R. C. Rezende, and W. Fan,

محمد مهدی حاجی اسمعیلی مدرک کارشناسی فناوری اطلاعات خود را در سال ۱۳۹۱ از دانشگاه صنعتی شاهرود و مدرک کارشناسی‌ارشد خود را نیز در رشتهٔ یادشده و در گرایش تجارت الکترونیک در سال ۱۳۹۳ از دانشگاه قم دریافت کرد. موضوع پایان‌نامهٔ کارشناسی‌ارشد ایشان طراحی و پیاده‌سازی سامانه‌ای برای ارتباط با پایگاه‌های دادهٔ تجاری به واسطهٔ زبان طبیعی بوده‌است. وی هم اکنون دانشجوی مقطع دکترا در همان رشته در دانشگاه تربیت مدرس است. زمینه‌های پژوهشی مورد علاقهٔ وی یادگیری ماشینی، یادگیری تقویتی، پردازش تصویر و پردازش زبان طبیعی است.

نشانی رایانامهٔ ایشان عبارت است از:

MohammadHaji@Modares.ac.ir

غلامعلی منتظر مدرک کارشناسی مهندسی برق خود را در سال ۱۳۷۰ از دانشگاه خواجه نصیرالدین طوسی دریافت کرد. وی مدارک کارشناسی‌ارشد و دکترای خود را نیز در همین رشته در سال‌های ۱۳۷۳ و ۱۳۷۷ از دانشگاه تربیت مدرس دریافت کرد. وی هم اکنون استاد گروه مهندسی فناوری اطلاعات دانشگاه تربیت مدرس است. زمینه‌های پژوهشی مورد علاقهٔ ایشان هوش مصنوعی، روش‌های مبتنی بر نرم رایانش از جمله نظریهٔ فازی و یادگیری ماشینی، یادگیری الکترونیکی و دولت الکترونیک است.

نشانی رایانامهٔ ایشان عبارت است از:

Montazer@Modares.ac.ir

[DOI: 10.61186/jsdp.22.1.113]