دسته‌بندی تصاویر در مدارک علمی بر اساس یک روش یادگیری عمیق

آزاده فخرزاده*[1]

استادیار، گروه سیستم‌های اطلاعاتی، پژوهشکده فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران

امیرحسین صدیقی

استادیار، گروه سیستم‌های اطلاعاتی، پژوهشکده فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران

مدیریت اطلاعات

دوره ۹، شماره ۱

بهار و تابستان ۱۴۰۲

چکیده: بازیابی اطلاعات از تصاویر، به دو روشِ بافت‌محور و محتوا محور امکان‌پذیر است. در روش محتوا محور، محتوای بصری تصویر، برای بازیابی اطلاعات در نظر گرفته می‌شود. برای استخراج اطلاعات از محتوای تصویرها و استفاده از روش‌های محتوا محور، ابتدا باید آن‌ها را دسته‌بندی کرد. در این پژوهش یک روش دسته‌بندی برای تصاویر علمی معرفی می‌شود. داده‌های آزمایشی این پژوهش، از رساله‌ها و پایان‌نامه‌های موجود در گنج، یکی از منابع غنی اسناد علمی فارسی، انتخاب شده است. داده‌های آموزشی شامل ۵۸۹۲ تصویر است که به‌صورت تصادفی از رساله‌ها و پایان‌نامه‌های گنج، در هفت حوزه مختلف انتخاب شده است و خبرگان آن‌ها را برچسب زده‌اند. تصاویر به شش دسته شامل عکس‌های طبیعی، نقشه‌ها، نمودارهای (x-y)، جدول‌ها، نمودارهای ساختارمند یا فلوچارت‌ها و نمودارهای آماری دسته‌بندی شدند. از آنجایی که داده آموزشی به‌شدت نامتقارن بودند، با استفاده از روش‌های افزونه، اعضای کلاس‌های کم‌جمعیت افزایش داده شد. به‌دلیل شباهت بصری بین تصاویر بعضی از دسته‌ها، در تصاویر علمی، استخراج ویژگی‌های متمایزکننده چالشی بود؛ بنابراین از روش‌های یادگیری عمیق که ویژگی‌ها را از خود تصاویر می‌آموزد، استفاده شد. با توجه به حجم کم داده آموزشی، شبکه عصبی با لایه‌ها و پارامترهای کمتر استفاده شد. بررسی‌ها نشان داد که شبکه‌هایی‌که روی یک پایگاه داده تصاویر بزرگ، از پیش آموزش داده شده‌اند، دقت بهتری دارند. بر اساس نتایج این پژوهش، شبکه از پیش آموزش داده شده VGG16، با ۱۶ لایه، با دقت ۹۷درصد روی داده آزمون، در دسته‌بندی تصاویر علمی عملکرد خوبی دارد.

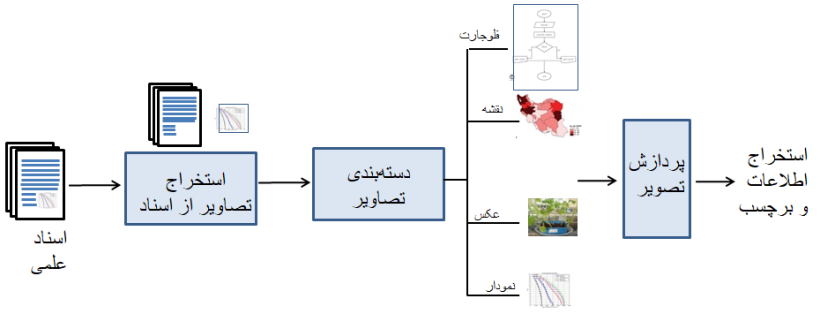
کلیدواژه‌ها: بازیابی تصاویر، دسته‌بندی تصاویر علمی، یادگیری عمیق، مدیریت اطلاعات.

---

۱. نویسنده مسئول: fakhrzadeh@irandoc.ac.ir

مقدمه

نویسندگان مقالات علمی، معمولاً از نمودارها یا تصویرها برای انتقال نتایج کمی آزمایش‌های خود یا فراهم‌کردن کمک بصری برای توضیح بهتر یا خلاصه‌کردن روش پیشنهادی خود بهره می‌برند (کلارک و دیوالا[1]، ۲۰۱۶). پردازش خودکار تصاویر موجود در اسناد علمی، امکان دسترسی به اطلاعات نهفته در تصاویر علمی را میسر می‌سازد. اغلب موتورهای جست‌وجوگر درکتابخانه‌های دیجیتال با متن کار می‌کنند؛ به این ترتیب که با وارد کردن یک عبارت، موتور جست‌وجوگر اسنادی را که شامل عبارت مدنظر یا عبارات نزدیک به آن را بازیابی می‌کند. چنانچه بخواهیم با همین روش تصاویر را بازیابی کنیم و تصاویر مرتبط با عبارت جست‌وجو شده را ببینیم، به ایجاد یک پایگاه داده از تصاویر همراه با برچسب‌های توصیف‌کنندهٔ آن‌ها نیازمندیم. استخراج اطلاعات از تصاویر علمی به برچسب زدن دقیق‌تر آن‌ها کمک می‌کند و دسترسی موتور جست‌وجو به آن‌ها را ممکن می‌سازد. از طرف دیگر استخراج اطلاعات از نمودارها و تصاویر، امکان تجزیه‌وتحلیل داده‌های موجود در آن‌ها را فراهم می‌کند. همان طور که در شکل ۱ نشان داده شده است، مراحل استخراج اطلاعات از تصاویر علمی عبارت‌اند از: ۱. استخراج تصاویر از اسناد؛ ۲. دسته‌بندی تصاویر؛ ۳. استفاده از ابزارهای مناسب پردازش تصویر برای هر دسته و استخراج اطلاعات و برچسب برای هر تصویر.

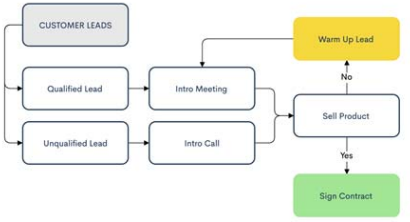


شکل ۱. مراحل استخراج اطلاعات از تصاویر موجود در مقالات علمی

برای به‌کاربردن ابزار پردازش مناسب برای هر تصویر، ابتدا تصاویر دسته‌بندی می‌شوند. تحقیقات زیادی در زمینه دسته‌بندی تصاویر طبیعی انجام شده است و دادگان آموزشی عظیمی از تصاویر و برچسب‌های آن‌ها ایجاد شده است. این دادگان عظیم، امکان استفاده از روش‌های یادگیری عمیق را در دسته‌بندی تصاویر ممکن ساخته است. پژوهش‌های بی‌شماری در زمینه پردازش تصاویر طبیعی انجام شده و مقاله‌های زیادی در این زمینه چاپ شده است؛ اما این پژوهش‌ها در زمینه پردازش و استخراج

1. Clark & Divvala

اطلاعات از تصاویر علمی، به‌دلیل چالش‌های موجود بسیار محدود است. تصاویر علمی از نظر بصری تفاوت زیادی با تصاویر طبیعی دارند. دسته‌بندی تصاویر علمی، به‌دلیل تنوع زیاد تصاویر در یک گروه و شباهت یک گروه به گروه دیگر، کار پیچیده‌ای است. برای مثال، در شکل ۲ یک فلوچارت و یک جدول نشان داده شده است، هر دو شکل از نظر ساختاری، شامل سلول‌هایی با متن هستند و ویژگی‌های بصری یکسان دارند؛ با این حال باید در دو گروه متفاوت دسته‌بندی شوند. در این پژوهش دسته‌بندی تصاویر موجود در اسناد علمی فارسی بررسی می‌شود و یک روش مبتنی بر الگوریتم عمیق، برای این امر معرفی می‌شود. داده آزمایشی این پژوهش تصویرهای موجود در اسناد گنج است. گنج یکی از غنی‌ترین پایگاه اطلاعات علمی اسناد علمی به زبان فارسی است که حاوی صدها هزاررساله و پایان نامه دانشگاه است. یک نمونه تصادفی از اسناد گنج انتخاب می‌شود و پس از استخراج تصاویر آن‌ها، یک الگوریتم برای دسته‌بندی تصاویر معرفی می‌شود.



	Focus groups	Interviews	Surveys
Cons	Generative ideas	Personal approaches	Easy for large scale
pros	Expensive	Need trained interviewer	Difficulty in determining samples

شکل ۲. نمونه‌ای از دو تصویر علمی با شباهت‌های بصری مشابه

پیشینه پژوهش

با پیشرفت در زمینه ذخیره‌سازی تصاویر، اطلاعات تصویری سهم بزرگی از اطلاعات موجود در اینترنت را به خود اختصاص داده‌اند. تصویرها در مقاله‌های علمی، اطلاعات مهمی را شامل می‌شوند و در بسیاری از موارد، برای خلاصهٔ بصری از روش معرفی‌شده و نتایج مقاله استفاده می‌شوند. برای دسترسی به تصاویر نیاز است که اطلاعات آن‌ها استخراج و فهرست‌سازی شود. استخراج اطلاعات از تصاویر، یکی از موضوعات روبه‌رشد در ادبیات است (سینگ، سریواستاوا، پاتاک، تیواری و کائور[1]، ۲۰۲۰؛ سمیح، رادی و تارک غریب[2]، ۲۰۱۹؛ لی، یانگ و ما[3]، ۲۰۲۱). برای استخراج اطلاعات از تصاویر علمی، نیاز است که تصاویر دسته‌بندی شوند. روش‌های دسته‌بندی تصاویر، به دو دسته سنتی و مبتنی بر یادگیری عمیق طبقه‌بندی می‌شود. در روش‌های سنتی دسته‌بندی تصاویر، ابتدا با استفاده از فیلترهای از پیش تعیین شده،

1. Singh, Srivastava, Pathak, Tiwari & Kaur
2. Samih, Rady & Tarek Gharib
3. Li, Yang & Ma

ویژگی‌های تصاویر استخراج می‌شود؛ سپس الگوریتم دسته‌بندی با توجه به ویژگی‌های تصاویر، یک برچسب را به هر تصویر اختصاص می‌دهد.

گائو، ژائو و بارنر[1] (۲۰۱۲) VIEW را پیشنهاد دادند که به‌صورت خودکار اطلاعات را از نمودارهای علمی استخراج می‌کند. در روش معرفی شده، اجزای متنی و گرافیکی جدا می‌شوند و با استخراج ویژگی‌های اجزای بصری و با استفاده از ماشین بردار پشتیبانی[2]، نمودارها را دسته‌بندی می‌کنند.

چنگ، استنلی، آنتانی و توما[3] (۲۰۱۳) یک رویکرد چند حالته را معرفی کرده‌اند که هم‌زمان از ویژگی‌های متن و تصویر استفاده می‌کند. ویژگی‌های به‌دست‌آمده از تصویر و زیرنویس آن، به‌عنوان ورودی به یک شبکه عصبی پرسپترون چندلایه داده شده و دسته هر تصویر پیش‌بینی می‌شود. یک رویکرد رایج دیگر برای دسته‌بندی تصاویر، استفاده از بازنمایی کیسه کلمات[4] بصری است که در آن کلمات بصری ویژگی‌های سطح پایین تصویر، مانند گرادیان یا بافت ناحیه‌ای تصویر است (یانگ، جیانگ، هاوپتمن و انگو[5]، ۲۰۰۷). هر تصویر ورودی با استفاده از این کلمات به‌عنوان یک بردار ویژگی کدگذاری می‌شود؛ سپس با استفاده از روش‌های یادگیری ماشینی، بردار ویژگی‌ها و تصویر متناظر آن‌ها دسته‌بندی می‌شوند.

شائو و فوترل[6] (۲۰۰۶) اشکال سطح بالا (برای مثال، پاره خط) را از تصاویر برداری نمودارها استخراج کردند؛ سپس از این اشکال به‌عنوان ویژگی‌ها برای دسته‌بندی نمودارها استفاده کردند. ساوا و همکاران[7] (۲۰۱۱) روشی را با عنوان Revision معرفی کردند که نوع نمودار را با استفاده از ویژگی‌های سطح پایین تصویر و ویژگی‌های متنی تعیین کرده‌اند، پس از دسته‌بندی نمودارها، روشی برای استخراج اطلاعات و طراحی مجدد نمودارها ارائه داده‌اند.

ناگا پراساد، سدیکی، گلبک و دیویس[8] (۲۰۰۷) روشی را معرفی کرده‌اند که در آن با استفاده از هیستوگرام گرادیان‌های جهت‌دار[9] و تبدیل ویژگی مستقل از مقیاس، ویژگی‌های هر تصویر استخراج می‌شود و شباهت هر تصویر به تصاویر داده آزمایشی با استفاده از الگوریتم شباهت هرمی[10] محاسبه می‌شود؛ سپس با در نظر گرفتن شباهت‌ها با استفاده از ماشین بردار پشتیبانی، دستهٔ هر تصویر تعیین می‌شود.

با پیشرفت‌هایی که در پردازنده‌های اطلاعات شکل گرفت، روش‌های مبتنی بر یادگیری عمیق نقش مهمی در پردازش تصاویر و دسته‌بندی آن‌ها پیدا کرد.

---

1. Gao, Zhou & Barner
2. Support Vector Machine (SVM)
3. Cheng, Stanley, Antani & Thoma
4. Bag of words
5. Yang, Jiang, Hauptmann & Ngo
6. Shao & Futrelle
7. Savva et al.
8. Naga Prasad, Siddiquie, Golbeck & Davis
9. Histogram of Oriented Gradients
10. Pyramid Match algorithm

لکان، بوتو، بنژیو وهافنر[1] (۱۹۹۸) اولین مدل شبکه عصبی کانولوشنی[2] به اسم LeNet-5 را معرفی کردند؛ اما به‌دلیل نبود داده آزمایشی در مقیاس بزرگ و همچنین نداشتن ابزارهای مناسب محاسباتی، نتایج LeNet-5 برای تصاویر پیچیده رضایت‌بخش نبود.

هینتون، اوسندرو و ته[3] (۲۰۰۶) یک الگوریتم مؤثر برای یادگیری در شبکه‌های عصبی چند لایه پیشنهاد دادند و بعد از آن، فصل جدیدی در پردازش تصویر بر اساس یادگیری عمیق ایجاد شد. هم‌زمان محققان، عملگر کانولوشن روی GPU را تشخیص دادند که تأثیر چشمگیری در بهبود محاسبات شبکه‌ها داشت.

کریژفسکی، سوتسکور و هینتون[4] (۲۰۱۲) بر اساس LeNet-5 مدل AlexNet را طراحی کردند که در چالش ILSVRC 2012[5] با تفاوت زیاد از روش‌های دیگر پیشی گرفت. بعد از به‌دست آوردن نتایج موفق AlexNet در دسته‌بندی تصویرها، شبکه‌های عصبی پیچشی به یکی از روش‌های اصلی دسته‌بندی تصاویر مورد استفاده قرار گرفت.

تعداد معدودی از محققان به ایجاد داده‌های آموزشی از تصاویر علمی و استفاده از یادگیری عمیق در دسته‌بندی آن‌ها پرداخته‌اند. لیو و همکاران[6] (۲۰۱۵) از ترکیب شبکه عصبی پیچشی و شبکه باور عمیق[7] برای دسته‌بندی تصاویر علمی استفاده کرده‌اند. شبکه عصبی پیچشی برای استخراج ویژگی‌ها و شبکه باور عمیق برای پیش‌بینی کلاس تصاویر به کار رفته است. آن‌ها روش خود را کمابیش روی ۵۰۰۰ تصویر علمی آزمایش کردند. سیگل، هورویتز، لوین، دیوالا و فرهادی[8] (۲۰۱۶) یک روش را معرفی کردند که به‌طور خودکار تصویرها را از فایل مقاله استخراج می‌کرد و برای دسته‌بندی تصویرها از شبکه‌های Alexnet و (هی، ژنگ، رن و سان[9]، ۲۰۱۶) Reznet50 که با استفاده از ImageNet پیش‌آموخته شده بودند، بهره بردند. آن‌ها این شبکه‌ها را با استفاده از یک مجموعه‌داده محدود شامل ۶۰۰۰ تصویر علمی دقیق‌تر تنظیم کردند.

جوبین، موندال و جواهر[10] (۲۰۱۹) یک مجموعه داده (DocFigure) شامل ۳۳ هزار تصویری علمی در ۲۸ دسته معرفی کردند که می‌تواند برای آموزش شبکه‌های عصبی با هدف دسته‌بندی تصویرهای علمی استفاده شود. موریس، مولر بوداک و ایورث[11] (۲۰۲۰) مجموعه‌داده SlideImages را معرفی کردند و که شامل ۳۰۰۰ تصویر علمی در ۹ دسته است. آن‌ها برای سنجش اعتبار داده معرفی شده از شبکه

1. Lecun, Bottou, Bengio & Haffner
2. Convolutional Neural Networks (CNN)
3. Hinton, Osindero & Teh
4. Krizhevsky, Sutskever & Hinton
5. ImageNet Large Scale Visual Recognition Challenge
6. Liu et al.
7. Deep Belief Networks (DBN)
8. Siegel, Horvitz, Levin, Divvala & Farhadi
9. He, Zhang, Ren & Sun
10. Jobin, Mondal & Jawahar
11. Morris, Müller-Budack & Ewerth

MobileNetV2 (سندلر، هوارد، ژو، ژموگینوف و چن[1]، ۲۰۱۸) استفاده کرده‌اند. نتایج نشان داد با اینکه حجم SlideImages از داده DocFigure کمتر است؛ نتایج قابل مقایسه‌ای داشت.

کاواسیدس و همکاران[2] (۲۰۱۹) برای تشخیص جدول و نمودار از یک روش ترکیبی مبتنی بر شبکه عصبی پیچشی، مدل‌های بصری و مفهوم برجستگی استفاده کردند. در واقع آن‌ها یک شبکه عصبی پیچشی مبتنی بر برجستگی که استدلال چند مقیاسی را روی نشانه‌های بصری انجام می‌دهد، برای دسته‌بندی نمودارها و جدول‌ها پیشنهاد کردند که به‌دنبال آن یک میدان تصادفی شرطی کاملاً متصل[3] برای تشخیص ناحیه جدول‌ها و نمودارها اعمال می‌شود.

جانگ و همکاران[4] (۲۰۱۷) یک روش دسته‌بندی را بر اساس چارچوب Caffe معرفی کردند و از روش GoogleNet برای دسته بندی تصاویر استفاده کرده‌اند.

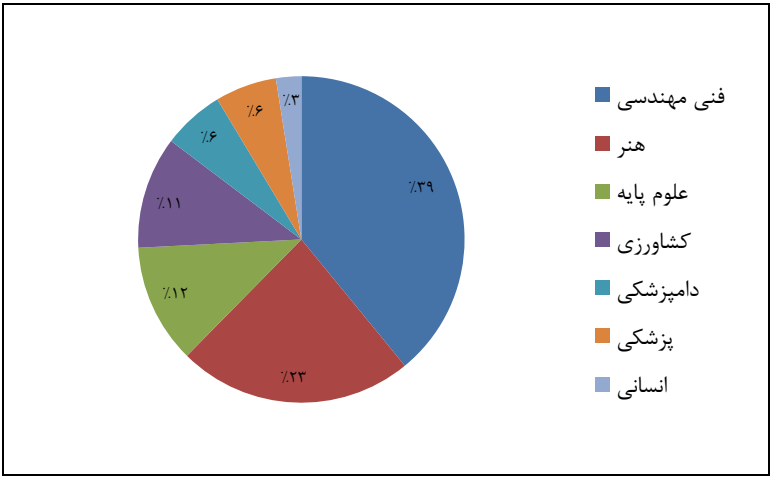
چاگاس و همکاران[5] (۲۰۱۸) مقایسه تحلیلی از تکنیک‌های مرسوم و سنتی در مقابل روش‌های CNN را فراهم کرده‌اند. آن‌ها رویکردهای سنتی یادگیری ماشین مانند طبقه‌بندی‌کننده‌های Naive Bayes ویژگی‌های HOG همراه با KNN، ماشین بردار پشتیبانی و جنگل تصادفی را با مدل‌های CNN مقایسه کرده‌اند. در تحلیل آن‌ها، مدل‌های CNN از پیش آموزش دیده با تنظیم دقیق آخرین لایه‌های شبکه استفاده شده است. نویسندگان به این نتیجه رسیدند که مدل‌های CNN با دقت ۷۷/۷۶ از روش‌های سنتی با دقت ۴۵/۰۳ پیشی گرفته‌اند. بر اساس دانش ما، غالب تحقیقاتی که تا به حال انجام گرفته، به چند دسته محدود از تصاویر علمی اختصاص یافته است. در تحقیقاتی هم که دسته‌های بیشتری در نظر گرفته شده، تصاویر از مجلات در حوزه خاص و با قالب‌های مشخص استخراج شده‌اند. در این تحقیق یک روش دسته‌بندی پیچشی برای تصاویر علمی معرفی می‌شود و روی داده آموزشی از تصاویر موجود در پایگاه اطلاعاتی گنج که یکی از مهم‌ترین کتابخانه‌های دیجیتال اسناد علمی فارسی است، آزمایش می‌شود. پایگاه اطلاعاتی گنج حاوی پایان‌نامه‌ها و رساله‌ها از دانشگاه‌های مختلف با حوزه‌ها و قالب‌های متنوع است. بنابراین دسته‌ها و دسته‌بندی معرفی شده نسبت به روش‌های دیگر عمومیت بیشتری دارد.

مجموعه داده

برای ایجاد مجموعه داده تصاویر از پایان نامه‌ها و رساله‌های موجود در پایگاه اطلاعاتی گنج استفاده شده است. در پایگاه اطلاعاتی گنج، تصاویر به‌صورت جداگانه ذخیره نمی‌شود و ابتدا باید تصاویر از اسناد استخراج شوند. فایل‌های گنج به دو صورت پی‌دی‌اف و ورد ذخیره می‌شوند. فخرزاده و صدیقی (۱۳۹۹) یک نرم‌افزار برای استخراج تصاویر موجود در ورد معرفی کرده‌اند. این نرم‌افزار تصویرهای الحاق شده را از فایل‌هایی که فرمت استاندارد ورد دارند، استخراج می‌کند. ۴۰۰ فایل به‌صورت تصادفی از پایگاه گنج

1. Sandler, Howard, Zhu, Zhmoginov & Chen
2. Kavasidis et al.
3. Fully-connected Conditional Random Field (CRF)
4. Jung et al.
5. Chagas et al.

انتخاب شد. نرم‌افزار استخراج تصاویر از ۲۷۹ فایل در هفت رشته دام‌پزشکی، انسانی، فنی، هنر، کشاورزی، علوم پایه و پزشکی، تصاویر را استخراج کرد. فراوانی حوزه اسناد به‌کار رفته در شکل ۳ نمایش داده شده است. همان طور که مشاهده می‌شود، دو حوزه هنر و فنی مهندسی بیشترین فراوانی را دارند. تصاویر استخراج شده از اسناد توسط دو خبره برچسب‌گذاری شد.



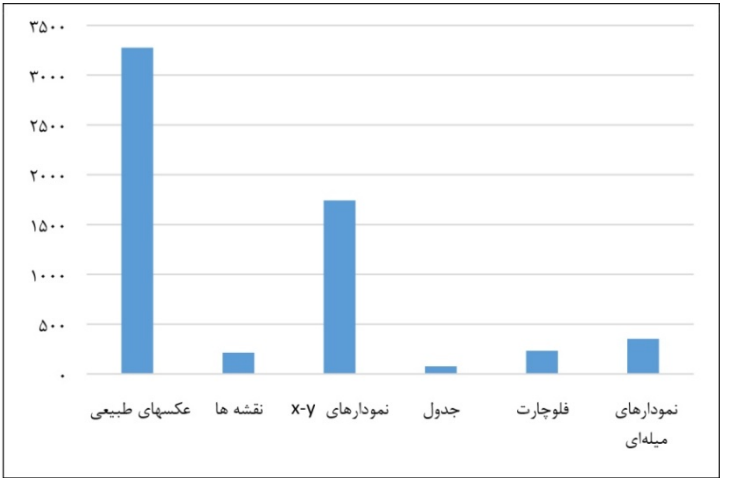
شکل ۳. فراوانی اسناد در حوزه‌های مختلف

با بررسی مقالات موجود در حوزه دسته‌بندی تصاویر علمی، شش برچسب برای دادگان در نظر گرفته شد. مجموع کل تصاویر برابر با ۵۸۹۲ بوده و شامل شش کلاس با برچسب‌های صفر تا پنج به شرح زیر می‌شود. در شکل ۴ فراوانی کلاس‌ها نشان داده شده است.

- برچسب صفر شامل عکس‌های طبیعی[1] است. این دسته از تصاویر بر اثر تابش نور بر یک سطح حساس به نور به‌دست می‌آید. تصاویری که با دوربین عکاسی یا وسایل تصویربرداری پزشکی ثبت می‌شود، در این دسته قرار می‌گیرند.
- برچسب ۱ برای نقشه‌هاست. نقشه‌ها نمایش نمادین از ویژگی‌های یک مکان یا توزیع داده در آن مکان هستند.
- برچسب ۲ به نمودارهای (x-y) اختصاص داده شده است. نمودارهای (x-y) روابط بین دو متغیر را نشان می‌دهد.
- برچسب ۳ جدول‌ها هستند. جدول ساختاری با سلول‌های حاوی متن و داده‌های عددی است. جدول‌ها در ادبیات برای کارهایی مانند مقایسه روش‌های موجود، خلاصه کردن مجموعه داده‌ها، برجسته کردن مشاهدات و غیره استفاده می‌شوند.

1. Natural images

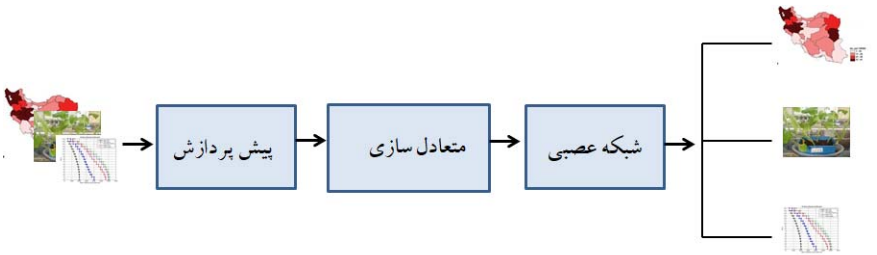
- برچسب ۴ به نمودارهای ساختارمند یا فلوچارت اختصاص یافته است. این نمودارها برای نشان دادن مراحل مورد نیاز برای حل یک مشکل استفاده می‌شوند.
- برچسب ۵ به نمودارهای آماری اختصاص یافته است که شامل نمودارهای میله‌ای، نمودارهای دایره‌ای و نمودار جعبه‌ای می‌شود.



شکل ۴. فراوانی تصاویر در دسته‌های مختلف

روش پیشنهادی

تصاویر استخراج شده از اسناد نیاز است برای اعمال شبکه عصبی مناسب‌سازی شود. همان طور که شکل ۴ مشاهده می‌شود، داده آموزشی به‌شدت نامتعادل است. نامتعادل بودن داده آموزشی برای دسته‌بندی مشکل‌ساز است (هی و گارسیا[1]، ۲۰۰۹). الگوریتم‌های یادگیری ماشین که روی مجموعه داده نامتعادل آموزش داده می‌شوند، تمایل دارند به‌سمت کلاس اکثریت سوگیری کنند.



شکل ۵. مراحل دسته‌بندی تصاویر علمی

1. De & Garcia

معمولاً الگوریتم‌های دسته‌بندی با به حداقل رساندن خطاها در کلاس اکثریت، تابع ضرر را بهینه می‌کنند و این کار به عملکرد ضعیف در کلاس اقلیت منجر می‌شود. برای بهبود عملکر دسته‌بندی، نیاز است داده آموزشی در دسته‌ها متعادل‌تر شود. در شکل ۵ مراحل دسته‌بندی نشان داده شده است که در آن، ابتدا تصاویر پیش پردازش می‌شوند؛ سپس داده متعادل‌سازی شده و در نهایت شبکه عصبی اعمال می‌شود و تصاویر دسته‌بندی می‌شوند. در ادامه هر قسمت توضیح داده می‌شود.

پیش‌پردازش تصاویر

در مرحله پیش‌پردازش ابتدا رنگ هر تصویر به دامنه خاکستری تبدیل شد؛ سپس اندازه هر تصویر به ۲۲۴ در ۲۲۴ پیکسل تغییر یافت. در ادامه میانگین و انحراف استاندارد مقادیر تمامی پیکسل‌ها برای داده‌های آموزش محاسبه شد. سپس با استفاده از این میانگین و انحراف استاندارد، تمامی تصاویر موجود در داده‌های آموزش و آزمون نرمال‌سازی شدند.

متعادل‌سازی

مجموعه داده‌های آموزش در این پژوهش بسیار نامتعادل هستند. به‌منظور از بین بردن این مشکل، برای کلاس‌های اقلیت که شامل یک، سه، چهار و پنج است، با استفاده از روش‌های زیر به ایجاد تصاویر افزوده از تصاویر موجود اقدام کردیم. روش‌های افزونه به‌کار رفته عبارت‌اند از: چرخش ۹۰ درجه، چرخش ۱۸۰ درجه، چرخش ۲۷۰ درجه، جابه‌جایی از چپ به راست، جابه‌جایی از بالا به پایین، جابه‌جایی از چپ به راست و سپس چرخش ۹۰ درجه، جابه‌جایی از چپ به راست و سپس چرخش ۲۷۰ درجه، فیلتر مات گاوسی[1]، فیلتر Unsharp Mask، فیلتر Contour، چرخش ۴۵ درجه، چرخش ۱۳۵ درجه، چرخش ۲۲۵ درجه، چرخش ۳۱۵ درجه، فیلتر Blur و فیلتر Sharpen.

دسته‌بندی تصاویر

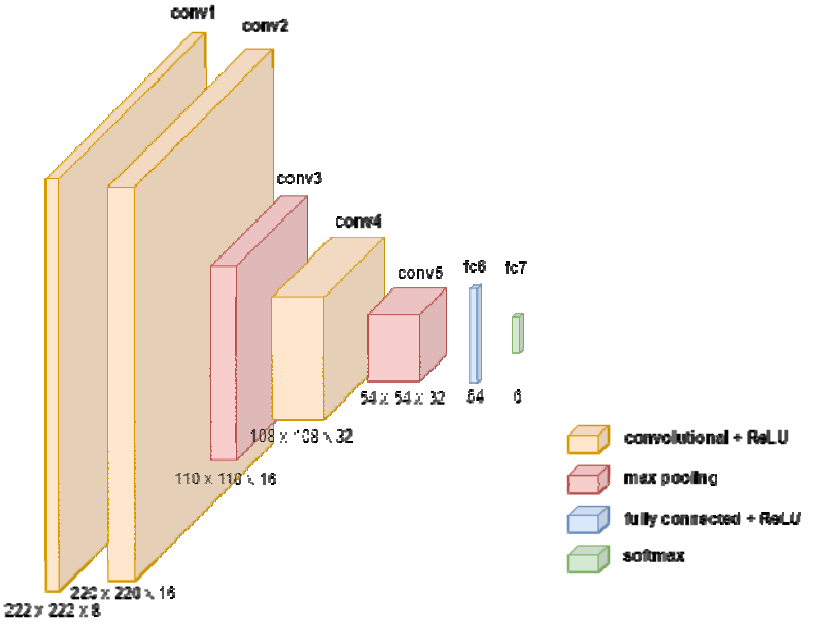
شبکه عصبی پیچشی (CNN) (شین و همکاران[2]، ۲۰۱۶) یکی از روش‌های پُرکاربرد در دسته‌بندی تصاویر است. بخش‌های اصلی CNN شامل لایه کانولوشنی[3]، لایه ادغام[4]، لایه فعال‌سازی غیرخطی و لایه کاملاً متصل (FC)[5] است. تصویر پیش‌پردازش شده از طریق لایه ورودی وارد شبکه CNN می‌شود، سپس توسط چندین لایه کانولوشنی و ادغام و لایه‌های کاملاً متصل پردازش و دسته‌بندی می‌شود. در CNN لایه‌های کانولوشنی نقش استخراج ویژگی‌ها را برعهده دارد. لایه کانولوشنی، به تصویر ورودی یا نقشه ویژگی‌ها[6] لایهٔ قبل اعمال می‌شود. لایه ادغام بلافاصله بعد از لایه کانولوشن می‌آید و دلیل اصلی استفاده

1. Gaussian Blur
2. Shin et al.
3. Convolutional
4. Pooling
5. Fully connected
6. Feature map

از این لایه، انجام نمونه‌برداری پایین[1] و کاهش ابعاد به‌منظور کاهش ارتباطات لایه‌های کانولوشنی است که نتیجهٔ آن کاهش بار محاسباتی است. تابع فعال‌ساز رابطهٔ بین ورودی و خروجی نرون را تعیین می‌کند. توابع غیرخطی باعث ایجاد رابطهٔ غیرخطی بین ورودی و خروجی می‌شوند. لایهٔ کاملاً متصل (FC) بعد از لایه‌های کانولوشنی و ادغام می‌آید و نورون‌های آن به تمام نورون‌های لایه قبلی متصل است. این لایه وظیفه دارد که کار دسته‌بندی را بر اساس ویژگی‌های استخراج شده از لایه‌های قبلی انجام دهد. هر چه تعداد لایه‌های شبکه عصبی بیشتر باشد، تعداد پارامترها بیشتر خواهد بود و داده بیشتری برای آموزش شبکه و تعیین پارامترها نیاز است. با توجه به کم بودن حجم داده و عدم اطمینان از تعداد مناسب لایه‌ها، ابتدا یک شبکه عصبی کم‌عمق روی داده اعمال می‌شود و عملکرد آن با شبکه عمیق تر مقایسه می‌شود.

شبکهٔ عصبی کم‌عمق

با توجه به حجم داده آموزشی شبکه عصبی پیچشی کم‌عمق با هفت لایه طراحی شد. مطابق شکل ۶ در این شبکهٔ عصبی از پنج لایه پیچشی و دو لایه تماماً متصل استفاده شد. در تمامی لایه‌های پیچشی از کرنل ۳ در ۳ استفاده شد.



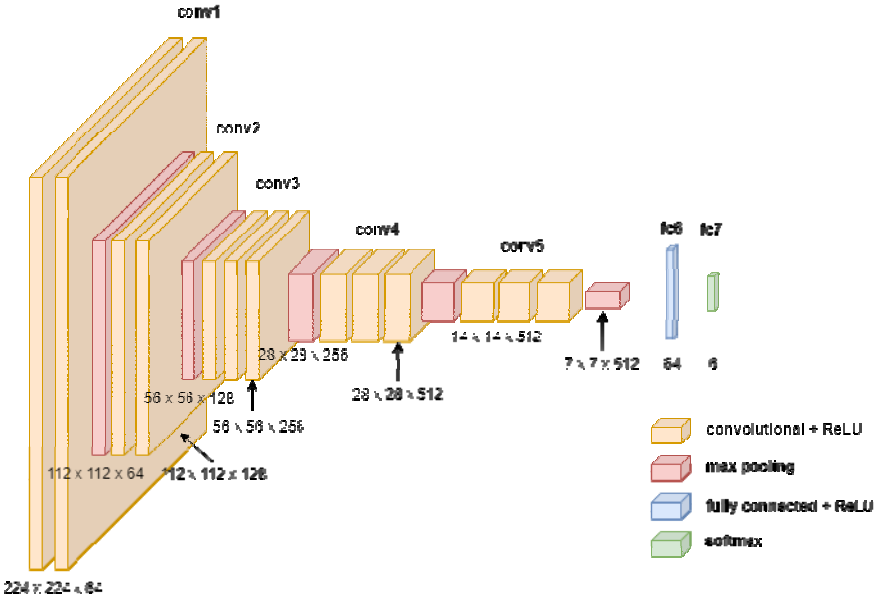
شکل ۶. لایه‌های شبکه کم‌عمق معرفی شده

1. Down-sampling

به‌علاوه تابع فعالیت در تمامی لایه‌ها به غیر از لایه آخر Relu در نظر گرفته شد. در لایه‌های اول، دوم، چهارم و پنجم از نرمال‌سازی روی بسته دادگان ورودی[1] بهره گرفته شد. این نرمال‌سازی ورودی‌ها را به گونه‌ای تغییر می‌دهد که میانگین آن‌ها نزدیک به صفر و انحراف استاندارد آن‌ها نزدیک به یک باشد. برای جلوگیری از بیش برازش[2] در لایه‌های سوم و پنجم و ششم از نرمال‌سازی مبتنی بر حذف[3] استفاده شد. در این روش واحدهای ورودی به طور تصادفی ولی با نرخی ثابت در هر مرحله از آموزش صفر در نظر گرفته می‌شوند. مابقی ورودی‌ها نیز به نحوی تغییر داده می‌شوند که مجموع کل ورودی‌ها ثابت باقی بماند. این تغییر تصادفی در ورودی‌ها، فقط در مرحله آموزش انجام شده است و روی ساختار نهایی شبکه یا داده‌های آزمون تأثیری ندارد.

شبکه عصبی VGG16

یکی از شبکه‌های عمیق و پُر کاربرد **که** در دسته‌بندی تصاویر استفاده می‌شود، شبکه VGG (سیمونیان و زیسرمن[4]، ۲۰۱۵) است. این شبکه به دو صورت ۱۶ لایه ۱۹ لایه ارائه شده است. شبکه VGG16 شامل ۱۶ لایه است و در شکل ۷ نشان داده شده است.



شکل ۷. شبکه VGG16

1. Batch Normalization
2. Overfitting
3. Dropout Regularization
4. Simonyan & Zisserman

در این شبکه در ابتدا دولایه کانولوشنی با ۶۴ فیلتر ۳×۳ پشت‌سرهم قرار گرفته‌اند. سپس، یک‌لایه بیشترین ادغام۲×۲ با پرش به اندازه ۲ قرار گرفته است. **در ادامه** دو لایه کانولوشنی دیگر با ۱۲۸ فیلتر ۳×۳ و یک لایه بیشترین ادغام ۲×۲ و پرش ۲ قرار گرفته‌اند. به‌طور مشابه، سه لایه کانولوشنی با ۲۵۶ فیلتر ۳×۳ و یک لایه بیشترین ادغام ۲×۲ با پرش ۲ قرار گرفته‌اند. سه لایه کانولوشنی با ۵۱۲ فیلتر ۳×۳ و یک لایه بیشترین ادغام ادامه این شبکه هست که البته دو بار تکرار می‌شود. درنهایت، ویژگی‌ها تبدیل به یک بردار ویژگی می‌شوند تا در اختیار لایه‌های FC قرار گیرند.

به‌منظور مقایسه شبکه VGG16 با شبکه کم‌عمق معرفی شده، لایه‌های پیچشی از شبکه VGG16 استفاده شده و لایه‌های انتهایی، دقیقاً مشابه با شبکه کم‌عمق در نظر گرفته شده است. به‌علاوه برای مقایسه بهتر بین این دو شبکه، یک بار از وزن‌های مربوط به شبکه آموزش داده شده روی مجموعه داده ImageNet استفاده شد و یک بار با استفاده از وزن‌های تصادفی، به آموزش شبکه VGG16 روی دادگان آموزش مدنظر این پژوهش اقدام شد. این شبکه‌ها را در ادامه به ترتیب با VGG16-l و VGG16-ll نشان می‌دهیم. هنگام استفاده از وزن‌های ImageNet، این وزن‌ها را در شبکه VGG16-l ثابت در نظر گرفتیم و تغییری در وزن لایه‌های پیچشی در طول فرایند آموزش داده نشد.

بررسی نتایج

برای آموزش و آزمون شبکه معرفی شده، مجموعه داده به صورت تصادفی با نسبت ۸۰ به ۲۰ به دو مجموعه داده آموزش و آزمون تقسیم شد. این تفکیک داده‌ها، به‌نحوی انجام شد که نسبت هر کلاس در داده‌های آموزش به آزمون، کمابیش برابر با نسبت ۸۰ به ۲۰ باشد. جدول ۱ وضعیت آماری داده‌های آموزش و آزمون را نشان می‌دهد.

جدول ۱. توزیع داده‌های آزمون و آزمایش در دسته‌های مختلف

برچسب	داده آموزش	داده آزمون	مجموع
۰	۲۵۸۸	۶۸۸	۳۲۷۶
۱	۱۷۴	۳۹	۲۱۳
۲	۱۴۱۶	۳۲۶	۱۷۴۲
۳	۶۷	۹	۷۶
۴	۱۸۷	۴۶	۲۳۳
۵	۲۸۲	۷۰	۳۵۲
مجموع	۴۷۱۴	۱۱۷۸	۵۸۹۲

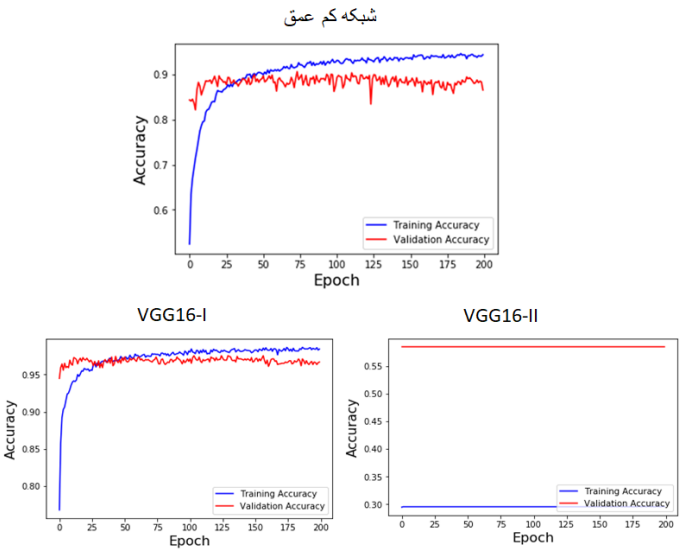
مجموعه داده‌های آموزش در این پژوهش بسیار نامتعادل هستند. برای کم‌کردن اثر این عدم تعادل، تعداد اعضای کلاس‌های یک، سه، چهار و پنج با استفاده از روش‌های افزونه معرفی افزایش داده شد. شایان ذکر است که برای پیاده‌سازی روش‌های افزوده از کتابخانه Pillow در زبان پایتون استفاده شد. با

توجه به تعداد اعضای کلاس‌های نامتعادل، برای هر کلاس از تعداد مشخصی از روش‌های افزونه برای ایجاد تصاویر افزوده استفاده کردیم. در نهایت تعداد کل داده‌های آموزش به ۸۷۵۹ داده رسید که جزئیات آن در جدول ۲ آمده است.

جدول ۲. توزیع داده‌های آموزشی بعد از اعمال افزونه

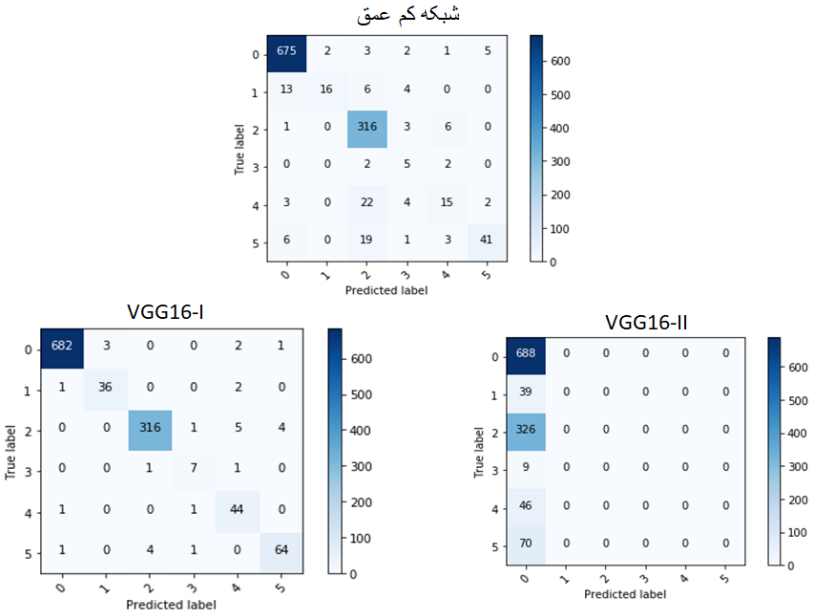
برچسب	داده آموزش
۰	۲۵۸۸
۱	۱۲۱۸
۲	۱۴۱۶
۳	۱۰۰۵
۴	۱۱۲۲
۵	۱۴۱۰
مجموع	۸۷۵۹

از سه شبکه عصبی معرفی شده، شبکه عصبی کم عمق، VGG16-l و VGG16-ll برای دسته‌بندی داده استفاده شد. برای آموزش شبکه‌های عصبی از یک کامپیوتر با پردازنده مرکزی i7-6700K با ۳۲ گیگابایت حافظه و یک پردازنده گرافیکی GTX1050 با ۲ گیگابایت حافظه بهره گرفته شد. شکل ۸ نمودار دقت سه شبکه عصبی روی دادگان آموزش و آزمون را در طول مرحله یادگیری نشان می‌دهد.



شکل ۸. نمودار دقت سه شبکه

دقتی که در این شکل نشان داده شده است، میانگین روی دقت تشخیص در همه دسته‌هاست. همان‌طور که در این شکل دیده می‌شود، در شبکه عصبی کم‌عمق تقریباً با ۵۰ بار تکرار الگوریتم، دقت شبکه عصبی در دادگان آموزش به ۹۰ درصد می‌رسد و پس از آن، شیب یادگیری شبکه کمتر می‌شود. برای دادگان آزمون روند، تقریباً یکنواختی مشاهده می‌شود و شبکه در ۵۰ تکرار، به دقت نزدیک به ۹۰درصد می‌رسد. در نهایت دقت شبکه روی دادگان آموزش به ۹۶ درصد و روی دادگان آزمون نیز به دقت ۹۱ درصد می‌رسد. شبکه VGG16-I با استفاده از وزن‌های ImageNet و شبکه VGG16-II با شروع از وزن‌های تصادفی کار می‌کند. همان‌طور که در این شکل دیده می‌شود در VGG16-I تقریباً با ۲۵ بار تکرار الگوریتم، دقت شبکه عصبی در دادگان آموزش به ۹۵ درصد می‌رسد و پس از آن، شیب یادگیری شبکه کمتر می‌شود. برای دادگان آزمون روند تقریباً ثابتی مشاهده می‌شود و شبکه در ۲۵ تکرار به دقت ۹۵ درصد می‌رسد. در نهایت دقت شبکه روی دادگان آموزش به ۱۰۰ درصد و روی دادگان آزمون نیز به دقت ۹۷ درصد می‌رسد. در شبکه VGG16-II، دقت شبکه در دادگان آموزش هیچ‌گاه بیش از ۳۰ درصد نمی‌شود و روند کاملاً ثابتی دارد. برای دادگان آزمون نیز، یک روند ثابت با دقت ۵۸ درصدی مشاهده می‌شود. نامتقارن بودن نسبی دادگان آموزش روی شبکه اثر گذاشته است و به نفع دسته اکثریت دسته‌بندی را انجام داده است. در این مثال، داده آزمون هم بیشتر از دسته اکثریت بوده است و به همین دلیل، دقت روی دادگان آزمون بیشتر شده است. با توجه به این نتایج از نظر دقت شبکه VGG16-I بهتر از بقیه کار می‌کند.



شکل ۹. ماتریس درهم‌ریختگی سه شبکه

شکل ۹ ماتریس درهم‌ریختگی برای دادگان آزمون را نشان می‌دهد. هر چه اعداد روی قطر این ماتریس بزرگ‌تر باشد، عملکرد دسته‌بندی بهتر است. بهترین ماتریس درهم‌ریختگی مربوط به VGG16-l و بدترین عملکر مربوط به VGG16-ll است. در شبکه کم‌عمق بدتربن عملکرد در مورد کلاس چهار یا دیاگرام‌هاست که از تعداد ۴۶ نمونه، ۱۵نمونه برچسب صحیح و ۲۲ نمونه آن‌ها برچسب دو (نمودارهای x-y) را دریافت کرده‌اند. همچنین در خصوص نمودارهای آماری (برچسب ۵)، الگوریتم اکثریت را درست تشخیص داده است؛ اما ۱۹ نمونه آن‌ها به اشتباه برچسب ۲ را دریافت کرده‌اند. بهترین عملکرد الگوریتم در تشخیص برچسب صفر (عکس‌های طبیعی) و برچسب ۲ (نمودارهای x-y) است. VGG16-ll همه را در کلاس صفر قرار داده است و VGG16-l با خطای جزئی کلاس‌ها را به‌درستی تشخیص داده است.

مقایسه پارامترها

جدول ۳ تعداد پارامترها در شبکه پیشنهادی را در مقایسه با شبکه‌های VGG16 نشان می‌دهد. مطابق این جدول دیده می‌شود که مجموع پارامترهای شبکه کم عمق تقریبا ۳۷ درصد تعداد پارامترهای شبکه‌های VGG16 است. با این حال تعداد پارامترهای قابل آموزش شبکه VGG16-l در مقایسه با شبکه پیشنهادی ۲۷ درصد است. در نهایت برای انجام هر Epoch در حین آموزش شبکه کم عمق به صورت متوسط به ۵۰ ثانیه زمان نیاز است در حالی که این زمان برای شبکه VGG16-l در حدود ۱۵۰ ثانیه و برای VGG16-ll در حدود ۸۸۰ ثانیه است. این امر نشان می‌دهد که تعداد کل پارامترهای شبکه نسبت به تعداد پارامترهای قابل آموزش تأثیر بیشتری بر روی زمان یادگیری دارد.

جدول ۳. مقایسه پارامترهای سه روش

نوع شبکه	پارامترهایی قابل آموزش	پارامترهای ثابت	مجموع
شبکه کم عمق	۵,۹۷۸,۵۵۰	۲۴۰	۵,۹۷۸,۷۹۰
VGG16-l	۱,۶۰۶,۲۱۴	۱۴,۷۱۴,۸۱۶	۱۶,۳۲۱,۰۳۰
VGG16-ll	۱۶,۳۱۹,۷۵۰	۱۲۸	۱۶,۳۱۹,۸۷۸

بحث

شبکه VGG16-ll و شبکه کم‌عمق هر دو با وزن‌های تصادفی اولیه آموزش داده شده است و نتایج نشان می‌دهد که شبکه کم‌عمق عملکرد بسیار بهتری دارد. شبکه VGG16-ll به همه تصاویر که همان تصاویر طبیعی است، برچسب صفر داده است. شبکه VGG16-I با استفاده از یادگیری انتقالی آموزش داده شده است؛ به این معنا که وزن‌های لایه‌های پیچشی از شبکه VGG16 که با استفاده از داده ImageNet آموزش داده شده، منتقل شده است و وزن‌های لایه‌های آخر با استفاده داده آموزشی تعیین شده است. شبکه VGG16-I نسبت به دو شبکه دیگر، دقت بهتری دارد. شبکه کم‌عمق پارامترهای کمتری دارد و با سرعت بیشتر و حافظه کمتر نسبت به شبکه VGG16 داده را با دقت قابل قبول دسته‌بندی می‌کند. نتایج این پژوهش نشان می‌دهد که اگرچه داده مورد تحقیق از نظر ویژگی‌ها با داده ImageNet متفاوت است،

انتقال وزن‌ها از لایه‌های کانولوشنی شبکه آموزش دیده شده با ImageNet، به دسته‌بندی بهتر کمک می‌کند. این نشان می‌دهد لایه‌های اولیه کانولوشونی در شبکه‌هایی که روی ImageNet آموزش داده شده‌اند، ویژگی‌هایی استخراج می‌کند که برای دسته‌بندی انواع تصویر می‌تواند مفید باشد.

نتیجه‌گیری

در این مقاله دسته‌بندی تصاویر علمی در اسناد فارسی بررسی شد. مجموعه تصاویر آموزشی از اسناد گنج که یکی از مهم‌ترین منابع اسناد علمی فارسی است، استخراج شد. تصاویر استخراج شده از این اسناد توسط خبرگان برچسب دریافت کردند و در شش دسته قرار گرفتند؛ سپس با استفاده از سه شبکه عصبی متفاوت تصاویر دسته‌بندی شدند. نتایج نشان داد که شبکه یادگیری عمیق VGG16 که روی داده ImageNet پیش آموزش داده شده است، بهترین دقت را دارد. این نتیجه نشان می‌دهد که شبکه‌های پیش‌پردازش شده، دانشی را در خود دارند که در دسته‌بندی تصاویر با ویژگی‌های بصری متفاوت با داده اولیه می‌تواند مفید باشد. بعد از دسته‌بندی تصاویر با استفاده از ابزارهای پردازش تصویر مناسب هر دسته می‌توان برچسب‌هایی را استخراج کرد که محتوای تصویر را توصیف می‌کنند. از این برچسب‌ها می‌توان در فهرست‌بندی و بازیابی تصاویر از اسناد در کتابخانه دیجیتال استفاده کرد.

فهرست منابع

فخرزاده، آزاده و امیرحسین صدیقی (۱۳۹۹). ارائه روشی ساختارمحور برای ایجاد پایگاه داده از تصاویر مستخرج از اسناد علمی؛ مورد مطالعه: پایگاه اطلاعات علمی ایران (گنج). *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۵ (۱۰۱)، ۷۲۹-۷۵۴.

Chagas, P., Akiyama, R., Meiguins, A., Santos, C., Saraiva, F., Meiguins, B. & Morais, J. (2018). Evaluation of Convolutional Neural Network Architectures for Chart Image Classification, *International Joint Conference on Neural Networks (IJCNN)*, 1–8

Cheng, B., Stanley, R., Antani, S. & Thoma, G. (2013). Graphical Figure Classification, Using Data Fusion for Integrating Text and Image Features. *Proceedings of the International Conference on Document Analysis and Recognition*, ICDAR, Aug. 2013.

Clark, C. & Divvala, S. (2016). Pdf figures 2.0: Mining figures from research papers. In Digital Libraries (JCDL), *IEEE/ACM Joint Conference*, pp. 143–152.

Gao, J., Zhou, Y. & Barner, K. E. (2018). View: Visual Information Extraction Widget for improving chart images accessibility. *19th IEEE International Conference on Image Processing*, pp. 2865–2868, Sept.2012. ISSN: 2381-8549.

He, H. & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engine*, 21(9), 1263–1284.

He, K., Zhang X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30; pp. 770–778

Hinton, G.E, Osindero, S. & Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput*, 18(7), 1527-54. doi: 10.1162/neco.2006.18.7.1527. PMID: 16764513.

Jobin, K.V., Mondal A. & Jawahar, C. V. (2019). DocFigure: a dataset for scientific document figure classification. *13th IAPR International Workshop on Graphics Recognition.* GREC 2019, Sydney, Australia, 20–22.

Jung, D., Kim, W., Song, H., Hwang, J., Lee, B., Kim, B. H. & Seo J. (2017). ChartSense: Interactive Data Extraction from Chart Images, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* Pages 6706–6717.

Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D. & Messina P. (2019). A saliency-based convolutional neural network for table and chart detection in digitized documents. *Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds) Image Analysis and Processing – ICIAP 2019.* ICIAP 2019. Lecture Notes in Computer Science. Vol 11752. Springer, Cham.

Krizhevsky, A., Sutskever, I. & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *2012 Neural Inf. Process. Syst.*, 25, 1097–1105

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE*, 86, 2278–2324

Li, X., Yang, J. & Ma, J. (2021). Recent developments of content-based image retrieval (CBIR), *Neurocomputing*, 452 (675-689).

Liu, X., Tang, B., Wang, Z., Xu, X., Pu, S., Tao, D. & Song, M. (2015). Chart classification by combining deep convolutional networks and deep belief networks. *In 2015 13th International Conference on Document Analysis and Recognition (ICDAR*), pp. 801–805, Aug. 2015.

Morris, D., Müller-Budack, E. & Ewerth, R. (2020). SlideImages: A Dataset for Educational Image Classification. *Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science*, Vol 12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_36

Naga Prasad, V.S., Siddiquie, B., Golbeck, J. & Davis, L.S. (2007) Classifying computer generated charts. *International Workshop on Content-Based Multimedia Indexing*. CBMI'07. IEEE, 85-92.

Samih, H., Rady, S. & Tarek Gharib, F. (2020). Enhancing image retrieval for complex queries using external knowledge sources. *Multimedia Tools and Applications*, 79(27633–27657).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), PP. 4510-4520

Savva, M., Kong, N., Chhajta A., Fei-Fei L., Agrawala, M. & Heer, J. (2011). *ReVision: Automated Classification, Analysis and Redesign of Chart Images.* ACM User Interface Software & Technology (UIST)

Shao, M. & Futrelle, R. (2006) . Recognition and classification of figures in pdf documents. *W. Liu and J. Llados, editors, ´ Graphics Recognition. Ten Years Review and Future Perspectives*, Vol. 3926 (231-242) of Lecture Notes in Computer Science. Springer Berlin.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*. May; 35(5):1285-98. doi: 10.1109/TMI.2016.2528162. Epub 2016 Feb 11. PMID: 26886976; PMCID: PMC4890616

Siegel, N., Horvitz, Z., Levin, R., Divvala, S., Farhadi, A. (2016). FigureSeer: Parsing Result-Figures in Research Papers. *Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, Vol. 9911. Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_41

Simonyan, K. & Zisserman, A. (2015). *AVery Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556v6

Singh, V.P., Srivastava, R., Pathak, Y., Tiwari, S. & Kaur, K. (2019). Content-based image retrieval based on supervised learning and statistical-based moments. *World Scientific*, 33 (19), 1-23.

Yang, J., Jiang, Y.G., Hauptmann, A.G. & Ngo, C.W. (2007) . Evaluating bag-of-visual-words representations in scene classification. *Workshop on Multimedia Information Retrieval*, pages 197–206.

Classification of Figures in Scientific Documents Based on a Deep Learning Method

Azadeh Fakhrzadeh[*1]
Assistant Prof., Department of Information Technology Research, Iranian Research Institute for Information Science and Technology (IranDoc). Tehran. Iran

Amir Hossein Seddighi
Assistant Prof., Department of Information Technology Research, Iranian Research Institute for Information Science and Technology (IranDoc). Tehran. Iran

Abstract
There are two ways to retrieve information from figures: context-oriented and content-oriented. The content-oriented methods use the visual content of the figures for retrieval. However, scientific figures are complex, so they need to be classified first before using content-oriented methods to extract information from them. This paper presents a classification method for scientific figures. The training data for the classification task was chosen from Ganj, a rich source of Persian scientific documents. The training data consisted of 5892 figures randomly selected from dissertations and theses of Ganj in seven different fields. Experts labeled the figures into six classes: natural photos, maps, x-y diagrams, tables, structured diagrams or flowcharts, and statistical diagrams. The training data was unbalanced, so augmentation methods were used to increase the number of figures in underrepresented classes. Scientific images from different classes, in some cases, look very similar, so finding features that can distinguish them is difficult. We applied deep learning methods that learn the features directly from the images. Due to the scarcity of data, we used neural network with fewer layers and parameters. We found that networks that were pre-trained on a large image database performed better. Our research shows that the pre-trained VGG16 network with sixteen layers can classify scientific images with 97% accuracy.

Keywords: Image retrieval, Scientific image classification, Deep learning, Information management.

1. Corresponding Author: fakhrzadeh@irandoc.ac.ir