مجله ماشین‌بینایی و پردازش‌تصویر
سال هفتم، شماره اول، تابستان ۱۳۹۹

ویژگی‌های آگاه به محتوا برای قطعه‌بندی معنایی تصویر

مجید نصیری[۱]، حمیدرضا رشیدی کنعان[۲] و سید حمید امیری[۳]

چکیده

قطعه‌بندی معنایی تصویر مبتنی بر شبکه‌های عصبی عمیق، از رویکردهای مهم محققان بینایی ماشین می باشد. در روش های مبتنی بر شبکه‌های عصبی عمیق، بطور کلی از یک شبکه-پایه که برای کاربرد شناسایی تصویر، آموزش دیده است، بمنظور استخراج ویژگی از تصویر استفاده می‌شود. از آنجا که ابعاد ویژگی‌های خروجی از این شبکه‌های-پایه کوچکتر از تصویر ورودی می‌باشد، لذا با اضافه کردن چندین لایه کانولوشنی به انتهای این شبکه‌های-پایه، ابعاد ویژگی‌های خروجی از این شبکه‌ها را به اندازه ابعاد تصویر ورودی می‌رسانند. استفاده از ویژگی‌های محلی خروجی از شبکه‌های-پایه، بدون در نظر گرفتن ارتباط کلی بین این ویژگی‌های محلی، منجر به قطعه‌بندی ضعیف و ناهموار می‌شود. بر این اساس، در این تحقیق واحدی با نام "واحد ویژگی‌های آگاه به محتوا" پیشنهاد می‌شود. این واحد با کمک ویژگی‌های محلی خروجی از شبکه‌های-پایه، ویژگی‌های سطح-تصویر ایجاد می‌کند. واحد پیشنهادی را می‌توان در معماری‌های مختلف قطعه‌بندی معنایی تصویر قرار داد. در این تحقیق، با اضافه کردن واحد پیشنهادی CAF به معماری‌های پایه FCN و DeepLab-v3-plus، به ترتیب معماری‌های FCN-CAF و DeepLab-v3-plus-CAF پیشنهاد شده است. بمنظور آموزش معماری‌های پیشنهادی از دادگان PASCAL VOC2012 استفاده شده است. نتایج آزمایشات نشان می دهد که معماری‌های پیشنهادی نسبت به معماری‌های پایه مربوطه، به ترتیب ۲/۷ و ۱/۸۱ درصد بهبود دقت (mIoU) دارد.

کلیدواژه‌ها

قطعه‌بندی معنایی تصویر، شبکه‌های عصبی عمیق، شبکه‌های عصبی کانولوشنی، واحد ویژگی‌های آگاه به محتوا

۱ مقدمه

با پیشرفت‌های انجام شده در زمینه تصویر برداری و تولید تصاویر دیجیتال با تفکیک‌پذیری بالا، نیاز به قطعه‌بندی دقیق تصویر بیش از پیش احساس می‌گردد، به نحوی که یکی از اساسی‌ترین مراحل پردازش تصویر، قطعه‌بندی تصویر می‌باشد. از طرف دیگر از آنجا که قطعه‌بندی تصویر از مراحل اولیه الگوریتم‌های بینایی‌ماشین می‌باشد، لذا انجام دقیق قطعه‌بندی، تاثیر زیادی در نتایج مراحل بعدی خواهد داشت.

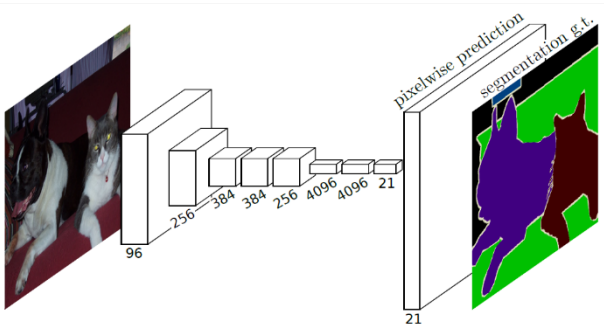
در سال‌های اخیر روش‌های مبتنی بر شبکه‌های عصبی عمیق، پیشرفت بسیاری نسبت به روش‌های پیشین که بر پایه ویژگی‌های مهندسی شده است، داشته‌اند. این روش‌ها در کاربردهای متفاوتی مانند طبقه‌بندی تصویر [۱]، قطعه‌بندی تصویر [۲]، بازشناسایی صوت [۳]، بازشناسایی رفتار [۴] پیشرفت‌های زیادی نسبت به روش‌های پیشین داشته‌اند.

به فرآیندی که پیکسل های مشابه را در تصویر در یک کلاس قرار می‌دهد، قطعه‌بندی تصویر می‌گویند [۵]، در حالی که در قطعه‌بندی معنایی تصویر، بخش‌هایی از تصویر که مربوط به یک

این مقاله در خرداد ماه سال ۱۳۹۸ دریافت، در آبان ماه بازنگری و در آذر ماه همان سال پذیرفته شد.
این پژوهش با پشتیبانی مالی دانشگاه تربیت دبیر شهید رجایی بر اساس قرارداد شماره ۱۹۹۷۴ مورخ ۹۷/۸/۳۰ انجام شده است.
[۱] دانش آموخته کارشناسی ارشد هوش مصنوعی دانشگاه شهید رجایی
رایانامه: majid.nasiri@sru.ac.ir
[۲] دانشکده مهندسی کامپیوتر، دانشگاه شهید رجایی
رایانامه: h.rashidykanan@sru.ac.ir
[۳] دانشکده مهندسی کامپیوتر، دانشگاه شهید رجایی
رایانامه: s.hamidamiri@sru.ac.ir
نویسنده مسئول: حمیدرضا رشیدی کنعان

شیء از یک کلاس خاص می‌باشد در یک دسته‌بندی قرار می‌گیرد (شکل (۱)). به عبارت دیگر، قطعه‌بندی اجزای تصویر به کلاس‌های از پیش تعیین شده طوریکه هر کلاس معنای یک بخش را مشخص کند، قطعه‌بندی معنایی تصویر نامیده می‌شود. قطعه‌بندی معنایی تصویر از عملیات پایه در بسیاری از کاربردهای بینایی ماشین مانند، خودروهای بدون سرنشین[1] [۶]، تجزیه اجزای بدن[2] [۷] و درک صحنه[3] [۸] می‌باشد.

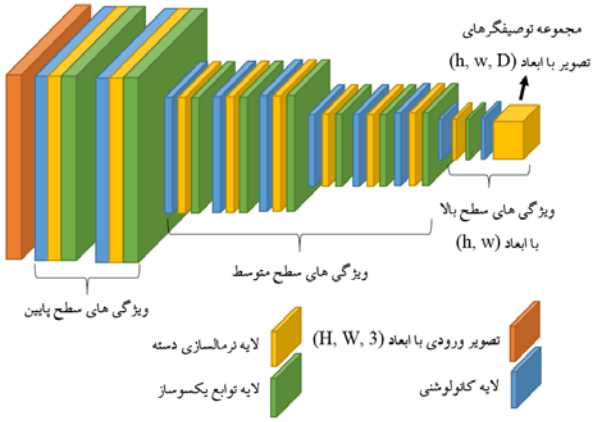
بیشتر رویکردهای موجود در مرزهای دانش، مبتنی بر معماری FCN [۲] می‌باشد. این معماری از معماری‌های مبتنی بر شبکه‌های عصبی عمیق می‌باشد. در روش های مبتنی بر شبکه‌های عصبی عمیق، بطور کلی از یک شبکه-پایه[4]، که برای کاربرد شناسایی تصویر با دادگان بسیار بزرگ مانند ImageNet [۹] آموزش دیده است، برای استخراج ویژگی از تصویر استفاده می‌شود. از جمله شبکه‌های-پایه استفاده شده در این مقالات، شبکه‌های AlexNet [۱]، VGG [۱۰]، ResNet [۱۱]، GoogleNet [۱۲] و Xception [۱۳] می‌باشند. تمامی شبکه‌های-پایه مذکور، از نوع شبکه‌های عصبی کانولوشنی می‌باشند، که معمولا از پشت سرهم قرار دادن لایه‌های کانولوشنی و max-pooling بوجود آمده‌اند. با استفاده از لایه‌های کانولوشنی، ویژگی‌های تصویر استخراج می‌شوند و با استفاده از لایه‌های max-pooling، مقدار بیشینه این ویژگی‌ها در پنجره لایه max-pooling انتخاب شده و به لایه بعدی انتقال داده می‌شوند. در لایه max-pooling با انتخاب مقدار بیشینه و حذف بقیه مقادیر، اندازه ابعاد مکانی تصویر ورودی (طول و عرض تصویر) کاهش می‌یابد. لازم به ذکر است که با حذف مقادیر زیادی از ویژگی‌ها در لایه max-pooling، اطلاعات زیادی از بین می‌رود.



شکل ۱: قطعه بندی معنایی تصویر [۲].

با توجه به مقاله [۱۴]، ویژگی‌های استخراج شده در لایه اول شامل خطوط، نقاط و خم‌های ساده می‌باشد. به این ویژگی‌های استخراج شده در لایه اول، ویژگی‌های سطح-پایین می‌گویند. این ویژگی‌های سطح-پایین هر چند در کاربرد شناسایی تصویر (کاربردی که برای آن طراحی شده‌اند) بسیار مفید می‌باشد، ولی برای کاربرد قطعه‌بندی معنایی تصویر مشکل‌ساز می‌باشند. مشکل ساز بودن لایه max-pooling بدین علت است که، این لایه، ویژگی‌های سطح-پایین را حذف می‌کند، در حالی که این ویژگی‌ها برای بازسازی حاشیه کلاس‌های موجود در تصویر قطعه‌بندی شده خروجی، ضروری می‌باشد.

در شبکه‌های عصبی کانولوشنی با قرار دادن لایه‌های مختلف کانولوشنی و max-pooling به صورت پشت سر هم، رفته رفته، ابعاد مکانی ویژگی‌های استخراج شده کوچکتر، عمق ویژگی‌های استخراج شده بیشتر و پیچیدگی ویژگی‌های استخراج شده بیشتر می‌شود. بستگی به میزان پیچیدگی ویژگی‌های استخراج شده می‌توان این ویژگی‌ها را به سه دسته ویژگی‌های سطح-پایین، ویژگی‌های سطح-متوسط و ویژگی‌های سطح-بالا دسته‌بندی کرد (شکل (۲)).



شکل ۲: معماری شبکه های عصبی کانولوشنی عمیق.

در رویکردهای مبتنی بر معماری FCN، از آنجا که ابعاد مکانی ویژگی‌های خروجی از شبکه‌های-پایه، کوچکتر از اندازه تصویر ورودی می‌باشد، با اضافه کردن چندین لایه کانولوشنی به انتهای این شبکه‌های-پایه، اندازه ویژگی‌های خروجی از این شبکه‌ها را با لایه های deconvolutional به اندازه تصویر ورودی می‌رسانند.

مجموعه ویژگی‌های خروجی از یک شبکه-پایه، تصویر ورودی را توصیف می‌کنند. به عبارت دیگر، خروجی این شبکه‌های-پایه مجموعه‌ای از بردارهای توصیفگر محلی است که کل این مجموعه، تصویر ورودی را توصیف می‌کند. در قطعه‌بندی معنایی تصویر، استفاده از بردارهای توصیفگر محلی خروجی از شبکه‌های-پایه، بدون در نظر گرفتن ارتباط کلی بین این ویژگی‌ها، منجر به قطعه‌بندی ضعیف و ناهموار می‌شود. لذا، در قطعه‌بندی معنایی تصویر، برای بدست آوردن تصویر قطعه‌بندی شده با دقت بالا و یکنواختی بیشتر بایستی ارتباط بین ویژگی‌های سطح-بالا را در نظر گرفت. بعبارت دیگر، در قطعه‌بندی معنایی تصویر به

[1] Self-drive
[2] Human Parsing
[3] Scene Understanding
[4] Backend

ویژگی‌های سطح-تصویر[1]، یعنی ویژگی‌هایی که کل تصویر را توصیف می‌کنند، نیازمند می‌باشیم.

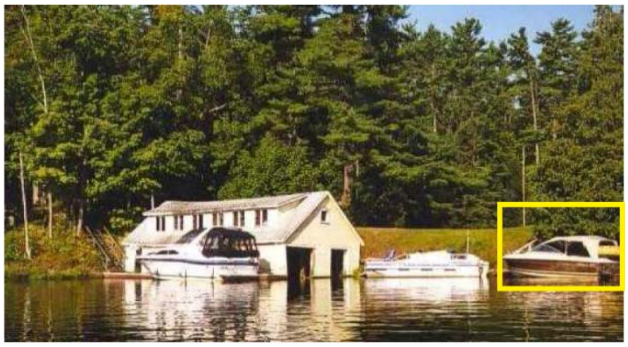
با توجه به نقاط ضعف دیده شده در قطعه‌بندی معنایی تصویر در قسمت ویژگی‌های سطح-تصویر، تحقیقات انجام شده در این مقاله در جهت بهبود این ضعف متمرکز شده است. برای واضح شدن اهمیت ویژگی‌های سطح-تصویر، این موضوع را با یک تصویر تشریح می‌کنیم. شکل (۳) را در نظر بگیرید، با استفاده از معماری FCN [۲]، در این تصویر قایق روی آب به اشتباه به عنوان خودرو، قطعه‌بندی شده است. این اشتباه به این علت رخ می‌دهد که بردارهای توصیفگر محلی استخراج شده از شبکه‌های عمیق (شبکه-پایه) برای دو شیء قایق و خودرو بسیار به هم شبیه می‌باشند. برای جلوگیری از این اشتباه، کافیست ویژگی‌های سطح-تصویر بطور موثری در کلاسبندی پیکسل‌ها دخالت داده شوند، بعبارت دیگر، در صورتی که بدانیم در تصویر، مجموعه کلاس‌های درخت، سبزه، برکه آب و کلبه جنگلی وجود دارد، می‌توان از قطعه-بندی اشتباه قایق بعنوان خودرو جلوگیری کرد.

بر اساس اشتباهات بوجود آمده ناشی از کمبود اطلاعات مرتبط با محتویات کلی در قطعه‌بندی معنایی تصویر، در این مقاله واحد "ویژگی‌های آگاه به محتوا" برای تقویت ویژگی‌های سطح-تصویر پیشنهاد می‌شود. این واحد، بردارهای توصیفگر محلی خروجی از شبکه‌های-پایه را در ورودی دریافت کرده و در خروجی، بردار ویژگی‌های آگاه به محتوای تصویر را تولید می‌کند. ویژگی‌های خروجی از واحد ویژگی‌های آگاه به محتوا، ویژگی‌هایی سطح-تصویر هستند. وجود ویژگی‌های سطح-تصویر مانند این است که بدانیم محتویات درون تصویر ورودی، چه فضایی را نمایش می‌دهند (بطور مثال فضای جنگلی، فضای داخل خانه، فضای بازار و ...).

در این تحقیق با اضافه کردن واحد ویژگی‌های آگاه به محتوا به معماری پایه FCN، معماری FCN-CAF پیشنهاد می‌گردد. با آموزش این معماری با استفاده از دادگان PASCAL VOC2012 [۳۳] مشاهد شد که این معماری ۲/۷ درصد دقت (mIoU) بهتری نسبت به معماری پایه FCN دارد.

بصورت خلاصه می توان گفت نوآوری ما در این تحقیق، طراحی یک واحد شبکه عصبی برای استخراج ویژگی‌های سطح-تصویر به نام واحد "ویژگی‌های آگاه به محتوا" (شکل (۴)) می-باشد. این واحد را می‌توان در معماری‌های دیگر قطعه‌بندی معنایی تصویر نیز بکار برد.

در ادامه مقاله، در بخش ۲، مروری بر کارهای گذشته بر پایه رویکردهای مبتنی بر یادگیری عمیق را خواهیم داشت، سپس در بخش ۳، روش‌های مرتبط با روش پیشنهادی را بطور مختصر شرح خواهیم داد، در بخش ۴، روش پیشنهادی را تشریح و ارائه خواهیم کرد، در نهایت به ارائه نتایج و نتیجه‌گیری خواهیم پرداخت.



شکل ۳: قطعه‌بندی اشتباه، بدون در نظر گرفتن ویژگیهای سطح-تصویر [۱۸].

۲ مرور کارهای گذشته

همانطور که در مقالات [۲،۱۵،۱۶] نشان داده شده است، استفاده از ویژگی‌های سطح-بالا برای کلاسبندی دقیق پیکسل‌های تصویر، بسیار مهم می‌باشد. همچنین استفاده از ویژگی‌های سطح-تصویر برای کلاسبندی دقیق و یکنواختی قطعه‌بندی بسیار مفید است [۱۷،۱۸]. روش‌های زیادی برای بهبود ویژگی‌های سطح-تصویر ارائه شده است.

استفاده از ویژگی‌های سطح-تصویر برای اولین بار در مقاله [۳۱] ارائه شد. فرض می‌کنیم مجموعه بردارهای توصیفگر محلی خروجی از شبکه-پایه، یک ماتریس [h, w, D] است، که در این ماتریس h ارتفاع، w عرض و D عمق ماتریس خروجی است (شکل (۲)). در این مقاله، ابتدا از این ماتریس ویژگی‌ها در راستای w و h میانگین گرفته و به بردار میانگین با ابعاد [۱،۱،D] می‌رسد، سپس این بردار را به اندازه قبلی خود تغییر اندازه داده و به ماتریسی به ابعاد [h, w, D] می‌رسد. از آنجا که این ماتریس بدست آمده، میانگین بردارهای توصیفگر محلی تصویر در هر بعد (در راستای عمق) می‌باشد، حاوی اطلاعاتی است که مربوط به کل تصویر است. مقاله [۴۶] با استفاده از هسته‌های مشبک کانولوشنی و همچنین استفاده از پس-پردازش CRF[2] توانسته ویژگی‌های عمومی‌تری را جهت کلاسبندی بکار گیرد. در مقاله [۴۹]، برای بهبود کلاسبندی با استفاده از ماژول GCN[3] بصورت سلسله مراتبی ویژگی‌های سطح-پایین، سطح-متوسط و سطح-تصویر را ترکیب کرده است. استفاده از لایه‌های max-pooling و کانولوشنی بصورت زنجیره‌ای برای ترکیب ویژگی‌های استخراج شده از تصویر ورودی با ابعاد متفاوت در مقاله [۴۷]، منجر به بهبود تشخیص کلاس‌های تصویر شده است. مقاله [۴۸]، با استفاده از معماری DenseNet [۴۴] در بخش رمزکننده[4] و الگوبرداری از این معماری در بخش رمزگشا[5] و همچنین استفاده از

[1] Image-Level Features

[2] Conditional Random Field
[3] Global Convolutional Network
[4] Encoder
[5] Decoder

ارتباطات میانبر[1]، توانسته اطلاعات تصویر را به لایه کلاسبندی انتقال دهد تا بتواند دقت بالاتری در قطعه‌بندی بدست آورد. مقاله [۵۰]، با اضافه کردن یک تابع هزینه بیشتر بر روی ماژول رمزکننده محتوای[2] ارائه شده، توانسته ارتباط ویژگی‌های سطح-تصویر را بهتر استخراج کند. مقاله [۱۸]، برای افزایش میزان ویژگی‌های سطح-تصویر، بر روی ویژگی‌های خروجی از شبکه-پایه هسته-های۳ کانولوشنی با ابعاد مکانی متفاوتی را اعمال کرده است. اعمال هسته‌های کانولوشنی با ابعاد مختلف باعث جمع کردن اطلاعات در ابعاد متفاوت می‌شود. این هسته‌ها در بزرگترین اندازه خود تقریباً تمامی ویژگیهای محلی را در بر می‌گیرند. این بدان معنیست که با این ایده می‌توان ویژگی‌های در سطح-ناحیه و هم ویژگی‌های سطح-تصویر را استخراج کرد. با توجه به افزایش چشمگیر میزان پارامترها با استفاده از این روش، مقاله [۳۲] هسته‌های کانولوشنی را با هسته‌های مشبک کانولوشنی جایگزین کرده و میزان پارامترهای کمتری را برای استخراج ویژگی‌های سطح-تصویر بکار گرفته است. در حقیقت مقاله [۳۲] با استفاده از اعمال هسته‌های مشبک کانولوشنی با اندازه های متفاوت، ویژگی‌های در سطح-ناحیه و هم ویژگی‌های سطح-تصویر را با تعداد کمتری پارامتر استخراج کرده است. در [۴۵] استراتژی استخراج ویژگی‌های سطح-تصویر عینا مانند روش مورد استفاده در [۳۲] می باشد با این تفاوت که به علت استفاده از دو مرحله افزایش نرخ نمونه برداری به جای یک مرحله، عملکرد بهتری بدست آمده است.

متفاوت با تمامی روش‌های ارائه شده در رویکردهای مبتنی بر یادگیری عمیق، واحد پیشنهادی CAF مجموعه بردارهای توصیفگر محلی خروجی از شبکه-پایه را گرفته و در خروجی، ویژگی‌هایی سطح-تصویر را تولید می‌کند. واحد پیشنهادی CAF در مرحله اول، مجموع فاصله‌ی تک تک بردارهای توصیفگر محلی را با تک تک مراکز خوشه‌های خود حساب کرده و یک ماتریس تولید می‌کند، سپس مقادیر این ماتریس را با استفاده از یک شبکه عصبی چند لایه به یک بردار تبدیل می‌کند. بردار بدست آمده در حقیقت یک بردار توصیفگر سطح-تصویر است.

۳ روش‌های مرتبط با روش پیشنهادی

استفاده از ویژگی‌های سطح-تصویر که نماینده کل تصویر باشد، در کاربردهای مختلف بینایی ماشین دیده شده است [۱۹،۲۰]. از جمله الگوریتم هایی که تصویر را با ویژگی‌های سطح بالا ارائه می-دهند، می‌توان BOW [۲۱]، HOG [۲۲]، SIFT [۲۳]، Fisher Vector [۲۴] و VLAD [۲۵] را نام برد. این ویژگی های سطح-بالا داری حجم کمتر و اطلاعات سطح بالاتری می-باشند. در ادامه، این الگوریتم‌ها را بطور مختصر شرح خواهیم داد.

۳-۱ الگوریتم BOW

یکی از قدیمی‌ترین الگوریتم های استخراج کننده ویژگی‌های سطح-تصویر، الگوریتم BOW [۲۱] می‌باشد. این الگوریتم بردارهای توصیفگر محلی را در گروه‌هایی قرار می‌دهد، بطوری که بایستی یک codebook با تعداد k مرکز تعریف کرد (عبارات بصری[4]) که معمولا این مراکز به کمک الگوریتم k-means [۲۶] بدست می‌آید. سپس هر یک از بردارهای توصیفگر محلی به یکی از نزدیکترین مراکز، اختصاص داده می‌شوند. در نهایت، خروجی الگوریتم عبارتند از هیستوگرام تعداد بردارهای محلی اختصاص داده شده به هر عبارت بصری (مرکز خوشه).

۳-۲ الگوریتم HOG

الگوریتم HOG[۲۲]بطور محلی هیستوگرام گرادیان جهت‌دار را در محل مورد نظر در تصویر استخراج کرده و از این اطلاعات بردارهایی را استخراج می‌کند که این بردارها توصیفگر تصویر هستند.

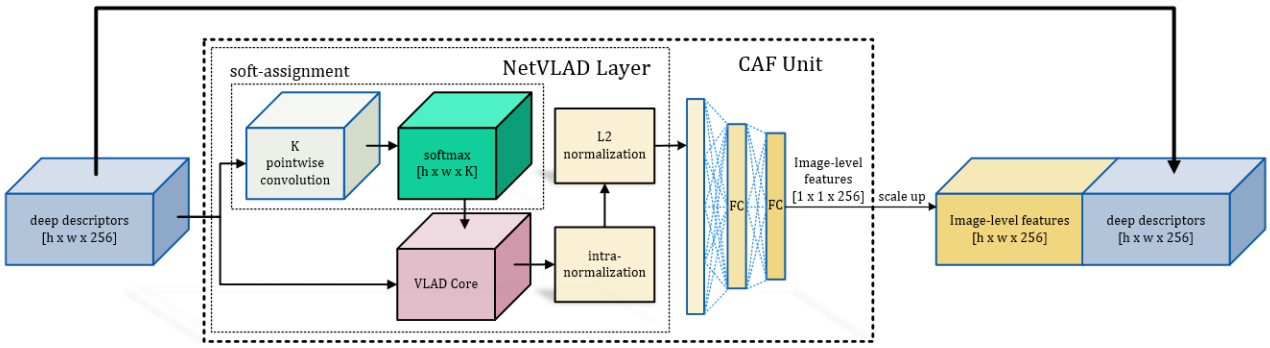
۳-۳ الگوریتم SIFT

الگوریتم SIFT [۲۳] که از الگوریتم های بسیار قوی در این حوزه می‌باشد، برای استخراج ویژگی‌هایی با حجم کم، برای ارائه تصاویر در دادگان با حجم (تعداد) بالا ارائه شده است. این الگوریتم برای یک تصویر، بردارهای توصیفگری را ارائه می‌دهد که این بردارهای توصیفگر نسبت به انتقال، تغییر اندازه و چرخش اشیاء در تصویر، مقاوم هستند. بردارهای ویژگی استخراج شده در این الگوریتم، محلی و به شکل ظاهری اشیاء وابسته هستند. این الگوریتم در کاربردهای شناسایی اشیاء [۲۷]، شناسایی چهره [۲۸]، شناسایی رفتار [۲۹] و دیگر کاربردها مورد استفاده قرار گرفته است.

۳-۴ الگوریتم FV

الگوریتم Fisher Vector (FV) [۳۰]، یک الگوریتم برای ارائه تصویر در فضایی دیگر است. بردارهای توصیفگر حاصل شده از FV، با جمع‌آوری بردارهای توصیفگر محلی از تصویر بدست می-آیند. این الگوریتم، در کاربردهایی که بر روی تصویر اعمال می-شود، ویژگی‌های سطح-تصویر را استخراج می‌کند. این الگوریتم با استفاده از برازش مدل مخلوط گوسین (GMM)، توزیع بردارهای توصیفگر تصویر را بدست می‌آورد. برخلاف الگوریتم BOW که اطلاعات آماری مرتبه-صفر را استخراج می‌کند، این الگوریتم اطلاعات آماری مرتبه-یک (میانگین) و مرتبه-دو (پراکندگی) را نیز استخراج می‌کند.

[1]Skip Connections
[2]Context Encoding Module
[3]Kernel
[4]Visual Words



شکل ۴: نمای کلی واحد ویژگیهای آگاه به محتوا (CAF).

۳-۵ الگوریتم VLAD

الگوریتم VLAD [۲۵]، از دیگر الگوریتم های ارائه بردارهای ویژگی سطح بالا از تصویر می‌باشد. این الگوریتم که بخشی از واحد پیشنهادی CAF را تشکیل می‌دهد، با دقت بیشتری تشریح خواهد شد.

در این الگوریتم، ابتدا با استفاده از الگوریتم k-means [۲۶] مجموعه مراکز خوشه‌ها، $C = \{c_k, i = 1 \dots K\}$ از مجموعه بردارهای توصیفگر ورودی $X = \{x_i, i = 1 \dots n\}$ یاد گرفته می‌شود.این مراکز خوشه‌ها در واقع همان عبارات بصری می‌باشند (همانند الگوریتم BOW). برای هر بردار ویژگی ورودی x_i ، فاصله این بردار تا تمامی مراکز خوشه‌ها با هم جمع می‌شوند (رابطه (۱)).

$$v_i = \sum_{i=1}^{k} x_t - \mu_i \qquad (۱)$$

خروجی حاصل از این الگوریتم، یک ماتریس K×D بعدی بصورت $V = [v_1, v_2, \dots, v_n]$ می‌باشد که از الطاق K بردار D بعدی بدست می‌آید که K تعداد خوشه‌ها و D ابعاد بردار ویژگی ورودی می‌باشد.

VLAD مانند الگوریتم FV است، با این تفاوت که در VALD از الگوریتم k-means استفاده شده است ولی در FV از تابع GMM استفاده شده است. خروجی حاصل از الگوریتم VLAD همانند الگوریتم FV نرمال‌سازی می‌شود. نرمال‌سازی ماتریس خروجی VLAD در دو مرحله انجام می‌شود. در مرحله اول بردارهای v_i با استفاده از روش نرمالسازی L2، نرمال می‌شوند (رابطه (۲)).

$$v_i = \frac{v_i}{\|v_i\|_2} \qquad (۲)$$

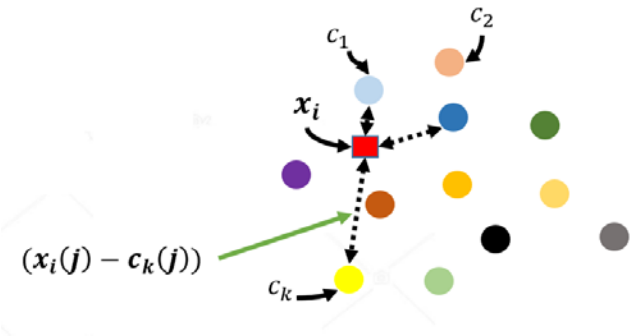
در مرحله دوم، کل ماتریس V خروجی از الگوریتم، با استفاده از روش نرمالسازی L2، نرمال می‌شوند (رابطه (۳)).

$$V = \frac{V}{\|V\|_2} \qquad (۳)$$

۴ روش قطعه‌بندی معنایی تصویر پیشنهادی

همانطور که قبلاً گفته شد، الگوریتم VLAD [۲۵] یک روش برای جمع آوری اطلاعات سطح-تصویر از بردارهای توصیفگر محلی می‌باشد. با اعمال این الگوریتم بر روی بردارهای ویژگی خروجی از شبکه‌های عصبی عمیق (شبکه-پایه) می‌توان این بردارها را بگونه‌ای دیگر ارائه کرد. بعبارت دیگر می‌توان گفت، این الگوریتم، بردارهای توصیفگر تصویر را خوشه‌بندی کرده و در خروجی، مجموع فاصله این بردارها را از مراکز خوشه‌ها ارائه می‌کند.

خروجی شبکه‌های عصبی عمیق، مجموعه‌ای از بردارهای توصیفگر تصویر هستند که به عنوان ورودی به الگوریتم VLAD داده می‌شود. فرض کنیم بردارهای توصیفگر خروجی از شبکه-پایه برای تصویر j ام مجموعه $X_j = \{x_{ij}, i = 1 \dots n_j\}$ باشد که برای سادگی، این مجموعه را برای یک تصویر به صورت $X = \{x_i, i = 1 \dots n\}$ نمایش می‌دهیم (شکل (۵)). این مجموعه شامل $n = h \times w$ بردارتوصیفگر محلی با عمق D می‌باشند (شکل (۲)). الگوریتم VLAD دارای K مرکز خوشه است که برای این الگوریتم، ابتدا با استفاده از الگوریتم k-means [۲۶]، مجموعه مراکز خوشه‌ها $C = \{c_k, k = 1 \dots K\}$ را بدست می‌آوریم. برای بدست آوردن مقدار بهینه برای K، مقادیر ۳۲، ۴۸ و ۶۴ را انتخاب کرده و در شبکه تست کردیم، برای مقادیر متفاوتی از k، دقت متفاوتی را برای شبکه خواهیم داشت که این نتایج در جدول(۱) آورده شده است. نتایج این جدول نمایانگر این است که تعداد ۳۲ مرکز، دقت بالاتری را ارائه می دهد.



شکل ۵: مراکز خوشه‌ها (دایره های رنگی) و بردارهای توصیفگر خروجی از شبکه پایه.

سپس برای هر بردار توصیفگر محلی x_i، همانند [۲۵]، فاصله این بردار تا تمامی مراکز خوشه‌ها در ضریب $a_k(x_i)$ ضرب شده و سپس با هم جمع می‌شوند (رابطه (۴)).

$$V(j,k) = \sum_{i=1}^{n} a_k(x_i)(x_i(j) - c_k(j)) \quad (۴)$$

در این رابطه، j اندیس عمق بردار توصیفگر و k اندیس خوشه می‌باشد. بعبارت ساده‌تر، $x_i(j)$ مقدار بردار توصیفگر iام در بعد jام و $c_k(j)$ مقدار مرکز خوشه kام در بعد jام است. مقدار $a_k(x_i)$ نمایانگر میزان عضویت بردار توصیفگر x_i به خوشه kام می‌باشد. در صورت استفاده از الگوریتم پایه VLAD [۲۵]، مقدار عضویت صفر یا یک می‌باشد که در این صورت می‌گوییم الگوریتم به صورت انتصاب-سخت است. ولی با روش ارائه شده در [۳۴] مقدار عضویت می‌تواند عددی بین صفر و یک باشد، که در این صورت می‌گوییم الگوریتم به صورت انتصاب-نرم است. در نهایت ماتریس خروجی $V(j,k)$ را بصورت ستونی نرمال‌سازی کرده تا مقادیر در بعدهای مختلف جداگانه نرمال شوند (همانند [۳۴])، که این عمل همان نرمالسازی برون-خوشه‌ای است. بعد از نرمالسازی برون-خوشه‌ای، کل آرایه‌های ماتریس را دوباره نرمال-سازی می‌کنیم. همانند مقاله [۳۴] برای انتصاب-نرم، از رابطه (۵) برای بدست آوردن مقادیر $a_k(x_i)$ استفاده شده است.

$$\bar{a}_k(x_i) = \frac{e^{-\alpha\|x_i - c_k\|^2}}{\sum_{k'} e^{-\alpha\|x_i - c_{k'}\|^2}} \quad (۵)$$

رابطه (۵) با توجه به نزدیکی بردار توصیفگر x_i به خوشه c_k ، وزنی را به آن اختصاص می‌دهد. α یک مقدار ثابت مثبت است که میزان وزن اختصاص داده شده را با نسبت اندازه فاصله بردار ویژگی تا مرکز خوشه، کنترل می‌کند. یعنی هر چه میزان α بزرگتر باشد، میزان عضویت اختصاص داده شده به بردارهایی که فاصله بیشتری تا مرکز خوشه دارند، کمتر می‌شود. در صورتی که مقدار α به سمت مثبت بینهایت میل کند، عملکرد الگوریتم به سمت انتصاب-سخت میل خواهد کرد. با بسط دادن رابطه و حذف ترم $e^{-\alpha\|x_i\|^2}$ از صورت و مخرج، رابطه (۶) حاصل می‌شود[۳۴].

$$\bar{a}_k(x_i) = \frac{e^{W_k^T x_i + b_k}}{\sum_{k'} e^{W_{k'}^T x_i + b_{k'}}} \quad (۶)$$

که بردار $W_k = 2\alpha c_k$ و $b_k = -\alpha\|c_k\|^2$ می‌باشد. در نهایت رابطه (۴) به صورت رابطه (۷) بازنویسی خواهد شد.

$$V(j,k) = \sum_{i=1}^{n} \frac{e^{W_k^T x_i + b_k}}{\sum_{k'} e^{W_{k'}^T x_i + b_{k'}}} (x_i(j) - c_k(j)) \quad (۷)$$

در این رابطه، مجموعه‌های $\{W_k\}$، $\{b_k\}$ و $\{c_k\}$، پارامترهای هر خوشه k هستند که در حین فرآیند آموزش، یاد گرفته می‌شوند. این پارامترها همگی به صورت سرتاسر[1] در شبکه یاد گرفته می‌شوند. همچنین می‌توان در یک مرحله، مراکز خوشه‌ها را با استفاده از الگوریتم k-means بدست آورد و در لایه VLAD قرار داد و حین آموزش سرتاسری شبکه، دوباره مقادیر این مراکز را دقیق‌تر کرد.

در ادامه فرآیند، خروجی V از الگورتم VLAD که یک ماتریس $K \times D$ است را به صورت یک بردار $N = [1, \square]$ در می آوریم ($N = K \times D$). این بردار را به ورودی یک شبکه عصبی متشکل از دولایه تماماً-متصل[2] می‌دهیم، که لایه‌های آن به ترتیب دارای ۱۰۲۴ و ۲۵۶ نورون می‌باشند. روابط این دو لایه در معادلات (۸) و (۹) آورده شده‌اند.

$$V_{fc1} = H_1(W_1 v^T + b_1) \quad (۸)$$

$$V_{fc2} = H_2(W_2 V_{fc1}^T + b_2) \quad (۹)$$

در روابط (۸) و (۹)، W_1 یک ماتریس [۱۰۲۴، N]، W_2 یک ماتریس [۲۵۶، ۱۰۲۴]، b_1 یک بردار ۱۰۲۴ المانی و b_2 یک بردار ۲۵۶ المانی می‌باشند. همچنین توابع H_i، توابع فعالساز[3] می-باشند. V_{fc1} خروجی اولین لایه تماماً متصل و V_{fc2} خروجی دومین لایه می‌باشد که این خروجی‌ها به ترتیب بردارهای ۱۰۲۴ و ۲۵۶ المانی هستند. بردار خروجی V_{fc2} برداری توصیفگر سطح-تصویر می‌باشد که بایستی بطور موثری در معماری قرار داده شود. بردار خروجی V_{fc2} را بصورت یک بردار [۱،۱،۲۵۶] تغییر شکل داده و سپس آن را تا ابعاد F=[h, w,۲۵۶] بزرگ می‌کنیم که در آن، h و w ارتفاع و عرض مجموعه بردارهای توصیفگر خروجی از شبکه-پایه برای تصویر ورودی به ابعاد [H, W,۳] است. ماتریس خروجی نهایی (F)، به ماتریس بردارهای توصیفگر محلی خروجی از شبکه-پایه الطاق[4] شده، تا عمل قطعه‌بندی در مراحل بعدی شبکه هم به کمک بردارهای توصیفگر محلی و هم به کمک ویژگی‌های ماتریس F صورت گیرد.

از آنجایی که بردار خروجی نهایی V_{fc2} حاوی اطلاعات محتوایی تصویر می‌باشد، ما مجموعه لایه VLAD [۲۵] به همراه لایه‌های اضافه شده را به عنوان واحد ویژگی‌های آگاه به محتوا معرفی می-

[1] End-to-end
[2] Fully-connected
[3] Activation function
[4] Concatenate

کنیم. این واحد در ورودی، خود ویژگی های سطح بالای تصویر را دریافت کرده و با استفاده از روش خوشه‌بندی، این ویژگی‌های سطح بالا را خوشه‌بندی می‌کند. در این واحد، اطلاعات ویژگی-های خوشه‌بندی بندی شده با استفاده از دو لایه شبکه عصبی تماماً متصل به ویژگی‌های سطح-تصویر تبدیل می‌شود. تست های انجام شده نشان می‌دهد که این واحد قدرت زیادی در استخراج ویژگی‌های سطح-تصویر را دارد.

جدول ۱: نتایج بدست آمده معماری FCN-CAF با مقادیر متفاوتی از مراکز خوشه ها.

معماری	K	Pixel Accuracy (PA)	mIoU
FCN-CAF-8s	۳۲	۹۲/۱	۶۴/۹
FCN-CAF-8s	۴۸	۹۱/۸	۶۴/۰
FCN-CAF-8s	۶۴	۹۱/۶	۶۳/۶

۵ نتایج آزمایشات و شبیه‌سازیها

در این بخش ابتدا به بیان معیارهای استفاده شده برای ارزیابی قطعه‌بندی معنایی تصویر می‌پردازیم، سپس شرایط پیاده سازی را تشریح خواهیم کرد. در نهایت، نتایج حاصل از قطعه‌بندی معنایی تصویر با معماریهای پیشنهادی FCN-CAF و DeepLab-v3-plus-CAF که به ترتیب با اضافه شدن واحد CAF به معماری‌های FCN و DeepLab-v3-plus [۴۵] بدست آمده-اند را ارائه می‌کنیم.

۵-۱ معیارهای ارزیابی

روش‌های مختلفی برای ارزیابی خروجی الگوریتم‌های قطعه‌بندی معنایی تصویر وجود دارد که در ادامه، دو نمونه از پرکاربردترین-های آنها که در تمامی مقالات مورد استفاده قرار می‌گیرد را در روابط (۱۰) و (۱۱) توضیح خواهیم داد. در این روابط $k \in \mathbb{N}$ تعداد کلاس ها را مشخص می‌کند. همچنین $C_{ij} \in \mathbb{N}_0$ با $i,j \in 1,\dots,k$ که C_{ij} تعداد پیکسل های متعلق به کلاس i هستند که به کلاس j نسبت داده شده‌اند، C_{ij} ماتریس پراکندگی نام دارد و G_i با رابطه $G_i = \sum_{j=1}^{k} C_{ij}$، نشان دهنده تمامی پیکسل های کلاس i می‌باشد، و $P_j = \sum_i C_{ij}$ تمامی پیکسل‌هاییست که به کلاس j نسبت داده شده اند.

دقت پیکسلی (PA[1]): این معیار نسبت تمامی پیکسل‌های درست کلاس‌بندی شده به کل پیکسل‌ها می‌باشد. دقت در این روش از رابطه‌ی (۱۰) محاسبه می‌شود.

$$PA = \frac{\sum_{i=1}^{k} C_{ij}}{\sum_{i=1}^{k} G_i} \qquad (۱۰)$$

دقت میانگین اشتراک به اجتماع پیکسل‌ها(mIoU[2]): در این معیار ابتدا نسبت اشتراک پیکسل‌های درست کلاس‌بندی شده، به اجتماع کل پیکسل‌های کلاس مربوطه و پیکسل‌هایی که به آن کلاس نسبت داده شده‌اند، محاسبه شده و سپس میانگین این مقادیر برای همه کلاس‌ها در نظر گرفته می‌شود. دقت در این روش از رابطه‌ی (۱۱) محاسبه می‌شود.

$$mIoU = \frac{1}{k}\sum_{i=1}^{k}\frac{C_{ij}}{G_i + P_i - C_{ii}} \qquad (۱۱)$$

۵-۲ دادگان مورد استفاده

در تست‌های انجام شده برای آموزش شبکه های پیشنهادی، از مجموعه دادگان PASCAL VOC2012 [۳۳] استفاده شده است. این مجموعه دادگان دارای ۱۴۶۴ تصویر استاندارد و ۱۰۵۸۲ تصویر از مجموعه تصاویر تهیه شده توسط مقاله [۳۵] برای بخش آموزش و تعداد ۱۴۴۹ تصویر برای ارزیابی می باشد. لازم به ذکر است، بسیاری از مقالات برای ارزیابی عملکرد الگوریتم ارائه شده خود و مقایسه آن با دیگر الگوریتم ها، از این مجموعه دادگان استفاده می‌کنند.

۵-۳ شرایط پیاده سازی

برای آموزش معماری‌های پیشنهادی FCN-CAF و DeepLab-v3-plus-CAF، از واحد پردازنده گرافیکی GTX1070 TI با حافظه ۸ گیگابایت استفاده شده است. همچنین، تمامی کدهای مربوط به پیاده‌سازی این معماریها در چهارچوب تنسورفلو [۳۶] و به زبان پایتون نوشته شده و به ترتیب در گیت هابهای https://github.com/m-nasiri/FCN_CAF و https://github.com/m-nasiri/Deeplab-v3-plus-CAF در دسترس می باشد.

برای بهینه سازی شبکه طراحی شده از تابع هزینه آنتروپی-متقابل[3] (رابطه (۱۲))، استفاده شده است.

$$loss = -\sum_{i=1}^{P}\sum_{c=1}^{C} Y_{ic}.\log(\bar{Y}_{ic}) \qquad (۱۲)$$

که در این رابطه، P تعداد پیکسل‌ها، C تعداد کلاس‌ها (در این مجموعه دادگان ۲۱)، Y_{ic} مقدار مورد انتظار در پیکسل iام برای کلاس cام است و $\hat{Y}_{ic}$ مقدار پیش‌بینی شده توسط شبکه در پیکسل iام برای کلاس cام است. به منظور کمینه کردن هزینه، از الگوریتم پس انتشار خطا [۳۷] و بهینه‌ساز Adam Optimizer [۳۸] با نرخ آموزش بسیار کوچک (جهت آموزش بهتر) و برابر با 10^{-4} استفاده شده است.

۵-۴ نتایج آزمایشات

معماری FCN-CAF با قرار دادن واحد CAF در معماری FCN [۲] بدست آمده است. این واحد به تشخیص صحیح پیکسل‌ها و یکنواختی قطعه‌بندی کمک می‌کند. شکل (۶) نمای

[1]Pixel Accuracy

[2]Mean Intersection Over Union (mIoU)

[3]Cross-Entropy

کلی شبکه FCN-CAF را نشان می‌دهد. این شبکه بعد از آموزش دیدن، عملکرد آن بر روی تصاویر بخش تست از مجموعه دادگان VOC2012[۳۳] با معیارهای PA و mIoU مورد ارزیابی قرار گرفته و نتایج حاصل در جدول (۲) آورده شده است. همچنین در این جدول معماری FCN-CAF با معماری‌های [۴۰،۴۱،۴۲] مقایسه شده است.

در الگوریتم [۴۰]، ابتدا پنجره‌هایی در تصویر به عنوان نواحی کاندید برای حضور شیء انتخاب می شود، سپس با استفاده از شبکه عصبی کانولوشنی، بطور جداگانه ویژگی‌های کل نواحی کاندید و ویژگی شیء موجود در پنجره استخراج می‌شود، در ادامه با استفاده از الگوریتم SVM [43] کلاس پنجره مورد نظر بدست آمده و در نهایت با استفاده از ویژگی‌های خاص شیء منتصب شده به پنجره، نواحی شیء موجود در پنجره قطعه بندی می‌شود. در [۴۱]، برای کلاسبندی هر پیکسل، از تمامی ویژگی‌های استخراج شده در لایه های شبکه عصبی کانولوشنی در راستای پیکسل مورد نظر استفاده شده است. بعبارت دیگر، بمنظور کلاسبندی هر پیکسل، هم از خود پیکسل و هم از ویژگی‌های استخراج شده در سطح بالا، سطح متوسط و سطح پایین (شکل (۲)) استفاده شده است. روش استفاده شده در [۴۲] شباهت زیادی به روش [۴۰] دارد، با این تفاوت که ابتدا ویژگی های تصویر با استفاده از شبکه عصبی کانولوشنی استخراج شده، سپس برای کلاسبندی پنجره‌های کاندید شده، از ویژگی‌هایی که با پنجره کاندید ماسک می‌شوند، استفاده شده است. این تغییر باعث افزایش سرعت مدل تا ۵۰ برابر شده است.

در گام نهایی تست‌ها، برای اثبات توانمندی واحد پیشنهادی CAF، این واحد را به معماری Deeplab-v3-plus که از قدرتمندترین معماریهای اخیر قطعه بندی معنایی تصویر می باشد، اضافه کردیم، از آنجا که این معماری در بخش کدکننده، از شبکه-پایه ۱۰۱ لایه resnet استفاده کرده است، آموزش این شبکه به واحد پردازش گرافیکی با حجم حافظه بیشتری نیاز دارد، با توجه به محدودیتهای سخت افزاری که در اختیار داشتیم، متاسفانه امکان آموزش این معماری و همچنین معماری Deeplab-v3-plus-CAF میسر نبود، لذا این دو معماری را برای مقایسه منصفانه با منابع و شرایط موجود خودمان یعنی با تصاویری با اندازه 385×385 و اندازه دسته (batch size) برابر ۶ آموزش دادیم. در این شرایط همانطورکه انتظار داشتیم، نتیجه بدست آمده برای معماری deeplab-v3-plus از مقداری که در مقاله مربوطه [۴۵] گزارش شده، ضعیف‌تر است. نتایج حاصل از آموزش ۱۰ مرتبه و میانگین‌گیری برای هر یک از معماری‌ها، نشان‌دهنده بهبود ۱/۸۱ درصدی (mIoU) در معماری پایه Deeplab-v3-plus می‌باشد. در جدول (۳)، نتایج قطعه‌بندی برای این دو معماری آورده شده است.

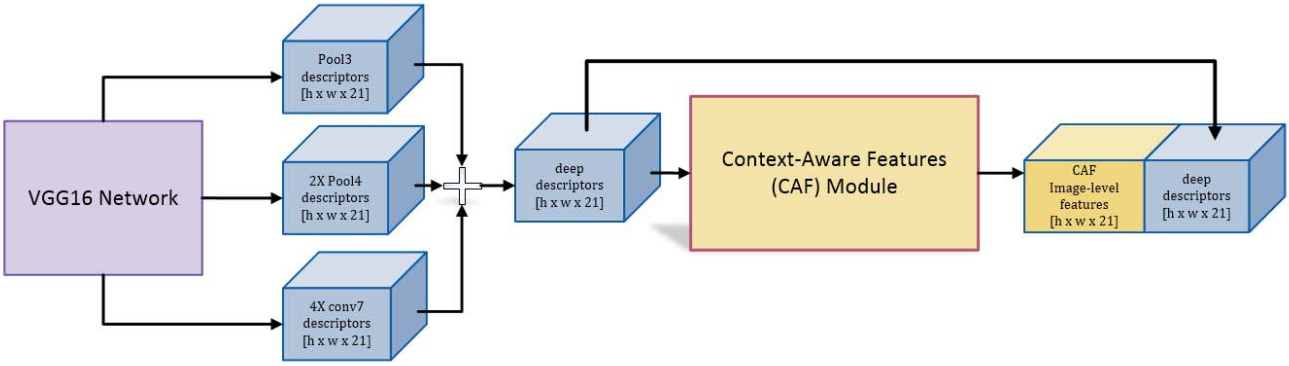
جدول ۲: مقایسه معماری FCN-CAF با معماری پایه FCN

معماری	Pixel Accuracy (PA)	mIoU
SDS [۴۰]	۴۹/۷	۵۱/۶
Hypercolum [۴۱]	۶۰	۵۴/۶
CFM [۴۲]	۶۰/۷	۶۱/۸
FCN-8s [۲]	۹۰/۳	۶۲/۲
FCN-CAF-8s	**۹۲/۱**	**۶۴/۹**

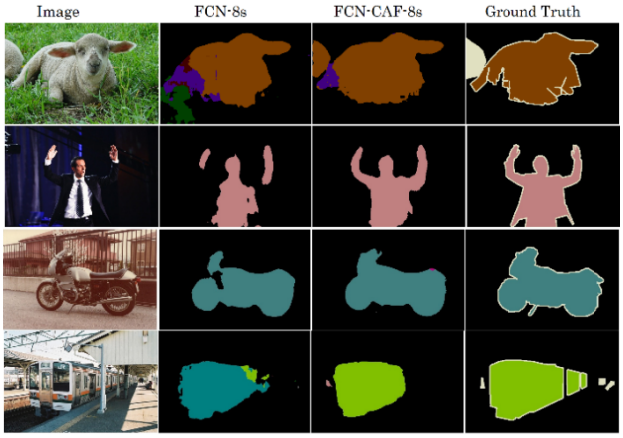
جدول ۳: مقایسه معماری DeepLab-v3-plus-CAF با معماری پایه DeepLab-v3-plus آموزش دیده شده با تصاویر با ابعاد ۳۸۵×۳۸۵ و اندازه دسته برابر با ۶.

معماری	Pixel Accuracy (PA)	mIoU
Deeplab-v3-plus [۴۵]	۹۲/۰۵	۶۸/۳۳
Deeplab-v3-plus-CAF	**۹۲/۹۵**	**۷۰/۱۴**

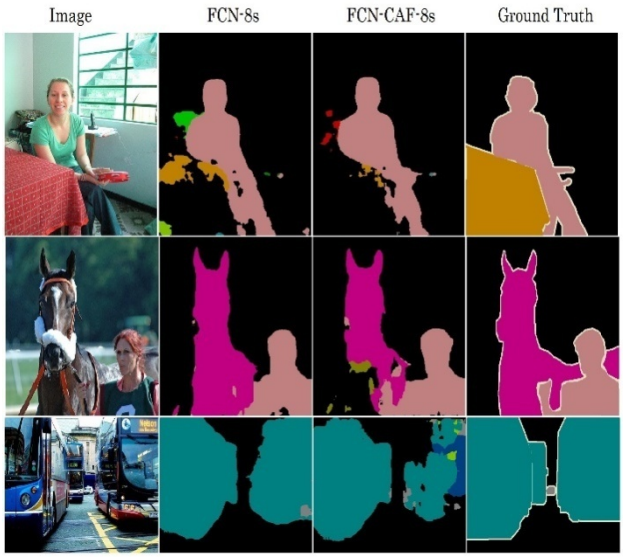
شکل (۷)، نمونه‌هایی از تصاویر قطعه‌بندی شده توسط معماری پیشنهادی FCN-CAF را نمایش می‌دهد. همچنین، شکل (۸) نمونه‌هایی از تصاویری که معماری پیشنهادی FCN-CAF عملکرد مطلوبی را در آنها بدست نیاورده است، نشان می دهد. همانطورکه مشاهده می شود، در این نمونه‌ها خروجی بدست آمده شباهت کمتری به خروجی مطلوب دارد. لازم به ذکر است در این نمونه‌ها، معماری FCN-CAF از لحاظ معیارهای mIoU و Pixel Accuracy (PA) از معماری پایه FCN ضعیفتر می باشد.



شکل ۶: نمای کلی معماری FCN-CAF که با جاسازی واحد CAF در معماری FCN بدست می آید.



شکل ۷: نمونه هایی از تصاویر قطعه بندی شده توسط معماری FCN-CAF و معماری پایه FCN.



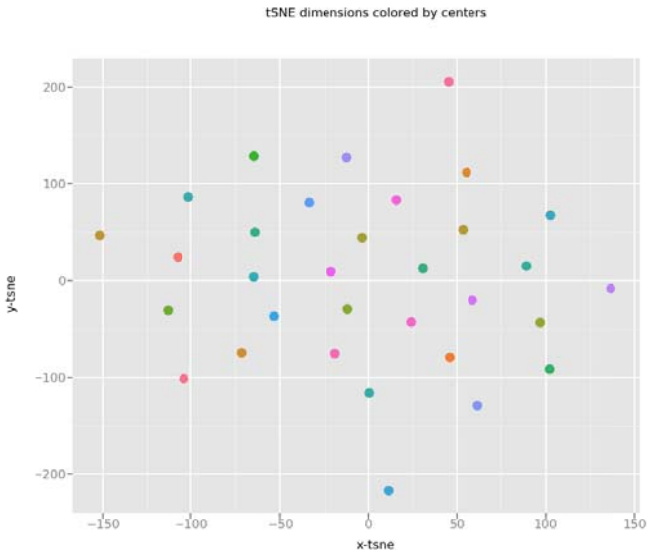
شکل ۸: نمونه‌های قطعه بندی شده توسط معماری FCN-CAF که عملکرد بهتری نسبت به FCN نداشته است.

۶ نتیجه

در این مقاله، واحدی با نام ویژگی‌های آگاه به محتوا (CAF) برای تقویت ویژگی‌های در سطح تصویر پیشنهاد شد. تست های انجام شده وتصاویر قطعه‌بندی شده با اضافه کردن این واحد به معماریهای FCN و Deeplab-v3-plus نشان دهنده عملکرد قدرتمند این واحد می‌باشد. نتایج بصری حاصل، نشان دهنده اهمیت هر چه بیشتر ویژگی‌های سطح-تصویرو قدرت واحد CAF برای استخراج این ویژگی‌ها در قطعه‌بندی معنایی صحیح تصاویر می‌باشد. بدیهی است، استفاده از این واحد در دیگر معماری‌های قطعه‌بندی معنایی تصویر، باعث افزایش دقت این معماری‌ها خواهد شد.

۷ پیوست‌ها

در مرحله طراحی بخش VLAD از واحد CAF برای اینکه مراکز خوشه‌ها را با استفاده از الگوریتم k-means بدست آوردیم، ابتدا شبکه-پایه را انتخاب می کنیم. سپس تمامی تصاویر بخش آموزش دادگان را یکی پس از دیگری به ورودی شبکه-پایه داده، و بردارهای توصیفگر خروجی را ذخیره می‌کنیم. آنگاه مجموعه تمامی بردارهای توصیفگر ذخیره شده را با الگوریتم k-means خوشه‌بندی کرده و مراکز خوشه‌ها را بدست می‌آوریم. این مراکز خوشه‌های بدست آمده را بعنوان مقادیر اولیه مراکز خوشه‌ها استفاده کرده و حین آموزش سرتاسری شبکه دوباره مقادیر این مراکز را دقیق‌تر می‌کنیم. شکل (۹) موقعیت مراکز خوشه‌های بدست آمده را نمایش می‌دهد. این تصویر با اعمال الگوریتم کاهش ابعاد t-SNE[۳۹]، در بر روی مراکز و کاهش ابعاد آنها از ۲۱ به ۲ بعد ترسیم شده است.



شکل ۹:ترسیم مراکز خوشه ها با الگوریتم t-SNE برای ۳۲ خوشه.

مراجع

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[3] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.

[4] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).

segmentation by global convolutional network. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 1743-1751). IEEE.

[18] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017, July). Pyramid scene parsing network. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (pp. 2881-2890).

[19] Yang, J., Yu, K., Gong, Y., & Huang, T. (2009, June). Linear spatial pyramid matching using sparse coding for image classification. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1794-1801). IEEE.

[20] Bicego, M., Lagorio, A., Grosso, E., & Tistarelli, M. (2006, June). On the use of SIFT features for face authentication. In Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on (pp. 35-35). IEEE.

[21] Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization with bags of keypoints. In Workshop on statistical learning in computer vision, ECCV (Vol. 1, No. 1-22, pp. 1-2).

[22] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[23] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[24] Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. International journal of computer vision, 105(3), 222-245.

[25] Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., & Schmid, C. (2012). Aggregating local image descriptors into compact codes. IEEE transactions on pattern analysis and machine intelligence, 34(9), 1704-1716.

[26] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[27] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on (Vol. 2, pp. 1150-1157). Ieee.

[28] Bicego, M., Lagorio, A., Grosso, E., & Tistarelli, M. (2006, June). On the use of SIFT features for face authentication. In Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on (pp. 35-35). IEEE.

[29] Scovanner, P., Ali, S., & Shah, M. (2007, September). A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th ACM

[5] Thoma, M. (2016). A survey of semantic segmentation. arXiv preprint arXiv:1602.06541.

[6] Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2722-2730).

[7] Liang, X., Xu, C., Shen, X., Yang, J., Liu, S., Tang, J., ...& Yan, S. (2015). Human parsing with contextualized convolutional neural network. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1386-1394).

[8] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ...& Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

[9] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). Ieee.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ...& Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[13] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.

[14] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.

[15] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[16] Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S., ...& Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 891-898).

[17] Peng, C., Zhang, X., Yu, G., Luo, G., & Sun, J. (2017, July). Large kernel matters—improve semantic

[44] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[45] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[46] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[47] Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1925-1934).

[48] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 11-19).

[49] Peng, C., Zhang, X., Yu, G., Luo, G., & Sun, J. (2017). Large Kernel Matters--Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4353-4361).

[50] Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., & Agrawal, A. (2018). Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7151-7160).

international conference on Multimedia (pp. 357-360). ACM.

[30] Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. International journal of computer vision, 105(3), 222-245.

[31] Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.

[32] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[33] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[34] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5297-5307).

[35] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., & Malik, J. (2011). Semantic contours from inverse detectors.

[36] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ...& Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning. In OSDI (Vol. 16, pp. 265-283).

[37] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

[38] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[39] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

[40] Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014, September). Simultaneous detection and segmentation. In European Conference on Computer Vision (pp. 297-312). Springer, Cham.

[41] Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 447-456).

[42] Dai, J., He, K., & Sun, J. (2015). Convolutional feature masking for joint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3992-4000).

[43] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters9.3 (1999): 293-300.

مجید نصیری مدرک کارشناسی خود را در رشته مهندسی برق (گرایش الکترونیک) در سال ۱۳۸۷ از دانشگاه صنعتی شیراز و مدرک کارشناسی ارشد خود را از دانشگاه تربیت دبیر شهید رجایی در رشته مهندسی کامپیوتر (گرایش هوش مصنوعی) در سال ۱۳۹۷ دریافت کرد. زمینه اصلی فعالیت‌های پژوهشی ایشان در حوزه بینایی ماشین، یادگیری ماشین و یادگیری عمیق می‌باشد.

حمیدرضا رشیدی کنعان مدرک کارشناسی خود را در سال ۱۳۷۹ از دانشگاه تبریز، مدرک کارشناسی ارشد خود را در سال ۱۳۸۱ از دانشگاه تربیت مدرس و مدرک دکترای خود را در سال ۱۳۸۷ از دانشگاه صنعتی امیرکبیر هر سه در رشته مهندسی برق (گرایش الکترونیک) دریافت کرد. وی در دوران تحصیل خود در مقطع دکتری، به مدت یک سال در دانشگاه Griffith استرالیا به عنوان دانشجوی بورسیه مشغول به تحقیق بوده اند. نامبرده قبل از پیوستن به گروه مهندسی کامپیوتر دانشگاه شهید رجایی، از سال ۱۳۸۸ تا ۱۳۹۴ عضو هیات علمی گروه مهندسی برق دانشگاه بوعلی سینا بوده‌اند. زمینه های اصلی تحقیقاتی مورد علاقه ایشان پردازش تصویر و ویدئو، بازشناسی الگو، بینایی ماشین و یادگیری عمیق است.

سید حمید امیری مدرک کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر (گرایش هوش مصنوعی) از دانشگاه صنعتی شریف به ترتیب در سالهای ۱۳۸۷ و ۱۳۹۳ دریافت کرد. وی در حال حاضر استادیار گروه مهندسی کامپیوتر دانشگاه شهید رجایی می باشد. زمینه های اصلی تحقیقاتی مورد علاقه ایشان پردازش تصویر و ویدئو، بازشناسی الگو، بینایی ماشین و یادگیری مبتنی بر گراف است.