مجله ماشین‌بینایی و پردازش‌تصویر
سال ششم، شماره دوم، زمستان ۱۳۹۸

بهبود شبکه عمیق R-FCN در آشکارسازی و برچسب‌زنی اشیاء

علی قنبری سرخی[۱]، حمید حسن‌پور[۲] و منصور فاتح[۳]

چکیده

امروزه آشکارسازی و برچسب‌زنی اشیاء در تصاویر یکی از چالش‌های اساسی در برخی از کاربردهای بینایی‌ماشین می‌باشد. در سال‌های اخیر استفاده از یادگیری عمیق مورد توجه محققان قرار گرفته است. در همین راستا، در این مقاله ابتدا جدیدترین شبکه‌های عمیق موجود معرفی، سپس نقاط قوت و ضعف آنها تحلیل می‌شود. در ادامه شبکه‌ای بهبود یافته از شبکه R-FCN ارائه می‌شود. روش پیشنهادی بر پایه معماری ResNet و شبکه تمام کانولوشن است. در این روش، معماری جدیدی مبتنی بر شبکه عمیق برای پیشنهاد ناحیه کاندید و روشی ترکیبی مبتنی بر SVM فازی دوکلاسه و SVR برای آشکارسازی و برچسب‌زنی اشیاء ارائه شده است. در این روش از تابع زیان جدید با عنوان اختلاف کوشی-شوارتز استفاده شده است. این تابع زیان از لحاظ سرعت و دقت، عملکرد بهتری از خود نشان داده است. روش پیشنهادی با معماری ResNet-۱۰۱ بر روی مجموعه داده SUN برای آشکارسازی و برچسب‌زنی ۳۶ شی مورد آزمایش قرار گرفت و نتایج بدست آمده نشان دهنده بهبود عملکرد این روش نسبت به روش پایه شبکه R-FCN است. روش پیشنهادی از لحاظ معیار mAP، عملکرد ۴۸/۳۸% و مدت زمان متوسط برای هر تصویر ۰/۱۳ را دارد، و نسبت به بهترین روش در این حوزه تقریبا ۲% در عملکرد و ۰/۰۴ ثانیه در زمان بهتر عمل کرده است.

کلیدواژه‌ها

آشکارسازی و شناسایی اشیاء، یادگیری عمیق، R-FCN، ماشین بردار پشتیبان دودویی فازی، اختلاف کوشی-شوارتز

۱ مقدمه

بدست آوردن خودکار اطلاعات معنایی در تصویر یکی از موضوعات مهم در بینایی ماشین می‌باشد. روش‌های زیادی بدین منظور در سال‌های اخیر معرفی شده است [۱-۳]. در برخی از روش‌هایی که پیشتر در این حوزه معرفی شده، فرض می‌شود که هر تصویر فقط شامل یک شی می‌باشد و آن روش‌ها با ویژگی‌های سطح پایین عمل دسته‌بندی را انجام می‌دهند. این روش‌ها به طور معمول، فقط برای دسته‌بندی صحنه‌های محدود قابل استفاده هستند.

در سال‌های اخیر از نمایش سطح بالای تصویر با انتخاب چندین ناحیه کاندید از تصویر جهت دسته‌بندی استفاده شده است. استفاده از نمایش سطح بالا عملکرد خوبی را در دسته‌بندی صحنه تصویر از خود نشان داده است. در روش‌های نوین برای دسته‌بندی صحنه، استفاده از مدل‌های موضوعی ساخته شده توسط متغیرهای پنهان مطرح شده است. در این روش‌ها دسته‌بندی تصویر برپایه معناشناسی تصویر انجام می‌شود. این روش‌ها برای مواردی که تعداد کلاس‌های صحنه بالا است مورد استفاده قرار می‌گیرند [۴-۶].

هدف از دسته‌بندی تصویر، توصیف سراسری یک تصویر با برچسب توصیف‌کننده است. ساحل، فضای باز، درون شهر و غیره نمونه‌هایی از برچسب‌های یک تصویر هستند. برچسب‌زنی تصویر، به برچسب‌گذاری محتوایی محلی یک تصویر تاکید دارد. برای نمونه، یک تصویر می‌تواند شامل "آسمان"، "ماشین"، "درخت" و غیره باشد. این دو مسئله به هم مرتبط هستند. پس تلاش برای حل مشترک این مسائل امری ضروری است. برای نمونه در یک تصویر با برچسب خیابان، احتمال توصیف تصویر با حضور "ماشین"، "انسان" یا "ساختمان" بیشتر از "ساحل" یا "آب دریا" است. با توجه به توضیحات ارائه شده، بدست آوردن

این مقاله در بهمن‌ماه ۱۳۹۶ دریافت، در مهرماه ۱۳۹۷ بازنگری و در آبان‌ماه پذیرفته شد.

[۱] دانشجوی دکتری مهندسی کامپیوتر، دانشگاه صنعتی شاهرود
رایانامه: ali.ghanbari289@gmail.com
[۲] دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود
رایانامه: h.hassanpour@shahroodut.ac.ir
[۳] دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود
رایانامه: mansoor_fateh@Shahroodut.ac.ir

نویسنده مسئول: علی قنبری سرخی

محتوای تصویر در بالا بردن دقت دسته‌بندی بسیار تاثیرگذار است. در همین راستا در این مقاله، راهکاری به منظور شناسایی و برچسب‌گذاری محتوای تصویر معرفی می‌شود [۷, ۸]. لازم به ذکر است که منظور از محتوای تصویر، اشیاء موجود در تصویر می‌باشد.

طراحی یک سیستم برای آشکارسازی و شناسایی اشیاء موجود در تصویر، چالش‌های اساسی و پیچیده‌ای دارد. در تصاویر طبیعی معمولا ویژگی‌هایی نظیر پس‌زمینه، محل و زاویه قرار گرفتن اشیاء، نورپردازی صحنه، تعدد اشیاء و همپوشانی جزئی یا کلی آنها، کار آشکارسازی و برچسب‌زنی اشیاء را دچار مشکل می‌کند. روش‌های استفاده شده در این نوع از سیستم‌ها بر اساس این چالش‌ها سنجیده می‌شوند. هدف غالب این روش‌ها، فائق آمدن بر این چالش‌ها است [۹-۱۲].

سیستم‌های معرفی شده تا سال ۲۰۱۴ به منظور تشخیص اشیاء، به الگوریتم‌های انتخاب ناحیه کاندید برای تخمین مکان اشیاء بسیار وابسته بودند. از مهم‌ترین سیستم‌های تشخیص اشیاء می‌توان به SPPnet [۱۳] و R-CNN Fast [۱۴] اشاره نمود که مورد توجه بسیاری از محققان و کاربردهای صنعتی قرارگرفته است. در این روش‌ها [۱۳, ۱۴] به دلیل محدودیت‌های زمانی و حالت‌های مختلف اشیاء، نمی‌توان سراسر تصویر را جستجو کرد. در همین راستا، ابتدا باید ناحیه‌های کاندید وجود شی استخراج شوند. استفاده از شبکه‌های عمیق به همراه روش‌های استخراج ناحیه کاندید، سبب بهبود زمان اجرای شبکه‌های تشخیص شده است ولی تعیین نواحی کاندید همچنان مهم‌ترین تنگنای محاسباتی این‌گونه روش‌ها محسوب می‌شود. در همین راستا در مقاله [۱۴]، یک شبکه پیشنهاد ناحیه کاندید[1] (RPN) معرفی شده است. در این شبکه، ویژگی‌های استخراج شده با کانولوشن‌های مختلف از سراسر تصویر به اشتراک گذاشته می‌شود. درنتیجه، زمان مربوط به پیشنهادهای ناحیه کاهش می‌یابد. RPN به‌طور هم‌زمان مرزهای اشیاء و امتیاز شی بودن را در هر موقعیت از تصویر پیش‌بینی می‌کند. در ایده‌ای متفاوت، مرجع [۱۵] محاسبات بالا در شبکه‌های عصبی کانولوشن بر پایه ناحیه را به‌عنوان یک اصل بیان کرده است که این محاسبات را با اشتراک‌گذاری کاندیدهای بدست آمده از کانولوشن به‌شدت کاهش می‌دهد[۱۳, ۱۴]. در آزمایش‌های انجام شده در سال‌های اخیر، R-CNN سریع [۱۴]، نرخی نزدیک به زمان بی‌درنگ در استفاده از شبکه‌های عمیق بدست آورده است [۱۶].

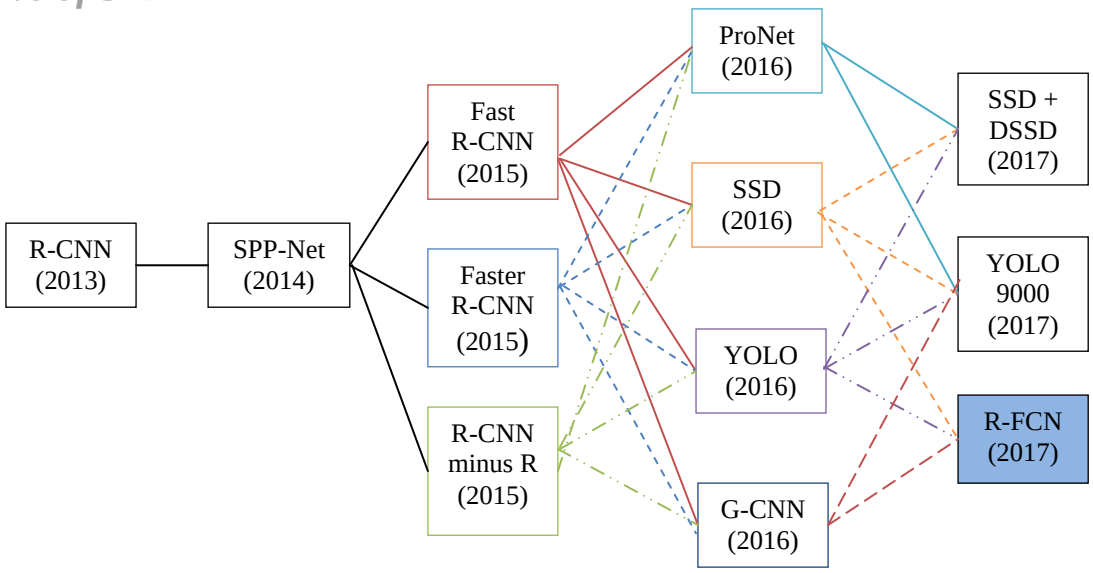
روش‌های پیشنهاد ناحیه کاندید به‌طورمعمول بر ویژگی‌های ساده و سریع تکیه دارند. جستجوی انتخابی [۱۷] یکی از معروف‌ترین روش‌های پیشنهاد ناحیه است که به‌صورت حریصانه ابرپیکسل‌های طراحی شده بر اساس ویژگی‌های سطح پایین (مانند میزان شدت روشنایی و رنگ در فضاهای رنگی متفاوت) را ادغام می‌کند. اما این روش در مقایسه با شبکه‌هایی مشابه R-CNN سریع، کندتر عمل می‌کند. اخیراً روش EdgeBoxs پیشنهاد شده در [۱۸]، یک تقابل بین کیفیت کاندید و زمان را فراهم کرده است. با این وجود مراحل تولید نواحی کاندید، بار محاسباتی زیادی (ازلحاظ زمان اجرا) در شبکه‌های تشخیص ایجاد می‌کنند. در واقع بدست آوردن نواحی کاندید بارمحاسباتی سیستم را بالا می‌برد. در نتیجه، باید سیستمی با قابلیت انجام همزمان عمل آشکارسازی و برچسب‌زنی طراحی نمود.

در تمام روش‌های معرفی شده در سال‌های اخیر، یک گام مهم تشخیص نواحی کاندید می‌باشد. در واقع بدست آوردن نواحی که می‌توانند به عنوان شی مطرح شوند بخش مهمی در دسته‌بندی تصویر می‌باشد. در همین راستا همانطور که در کارهای پیشین اشاره شد، با تغییر معماری و ساختار شبکه‌های عمیق عمل آشکارسازی و شناسایی به صورت همزمان صورت گرفته است. ولی چالش‌هایی به مانند انتخاب نواحی کاندید نادرست، همپوشانی نواحی کاندید، نواحی که به اشتباه به عنوان پس‌زمینه انتخاب شده و تعداد بسیار زیاد نواحی کاندید در این کاربرد وجود دارد. در همین راستا در این مقاله از روش ترکیبی از شبکه عمیق تمام کانولوشن و ماشین بردار پشتیبان فازی دو کلاس برای آشکارسازی نواحی مرتبط به اشیاء استفاده شده است. وظیفه اصلی این روش تشخیص نواحی کاندید می‌باشد و بر ناحیه‌هایی تاکید داردکه می‌توانند شی باشند و با پس‌زمینه متفاوت هستند.

در ادامه‌ی این مقاله، در بخش دوم به معرفی اجمالی شبکه‌های عصبی عمیق برای آشکارسازی و دسته‌بندی اشیاء پرداخته می‌شود. در گام بعدی معماری شبکه عصبی پیشنهادی ارائه می‌شود. در این بخش روش R-FCN[2] بهبود یافته با یک تابع زیان جدید با تغییر در ساختار لایه آخر ارائه شده است. بخش چهارم شامل تحلیل نتایج و پیاده‌سازی‌های انجام شده است و در نهایت نتیجه‌گیری و کارهای آینده در بخش پنجم معرفی می‌گردد.

[1] Region Proposal Network

[2] Region-based Fully Convolutional Networks(R-FCN)

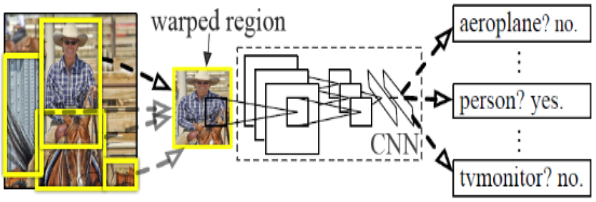


شکل ۱ شمای کلی از کارهای انجام شده برپایه شبکه های عمیق در آشکارسازی و برچسب‌زنی اشیاء

۲ استفاده از شبکه‌های عمیق در آشکارسازی اشیاء

در سال‌های اخیر استفاده از شبکه‌های عمیق به منظور آشکارسازی و برچسب‌گذاری اشیاء موجود در تصویر مورد توجه محققین و پژوهشگران پردازش تصویر و بینایی ماشین قرار گرفته است. شکل ۱ سیر تکامل مقاله‌های منتشر شده در مجلات معتبر دنیا برای شناسایی شی توسط شبکه با مفهوم عمیق را نشان می‌دهد. معماری R-CNN مطرح شده در مرجع [۱۵] یکی از اولین شبکه‌های عصبی عمیق کانولوشن برپایه ناحیه است. در این معماری برای تشخیص اشیاء از شبکه کانولوشن عمیق با معماری AlexNet استفاده شده است (شکل ۲). شبکه R-CNN از نواحی کاندید بدست آمده به کمک روش جستجوی انتخابی [۱۷] استفاده می‌کند. در این روش از یک شبکه کانولوشن برای استخراج ویژگی استفاده می‌شود. همچنین برای دسته‌بندی برچسب اشیاء از یک ماشین بردار پشتیبان خطی باینری استفاده شده است.

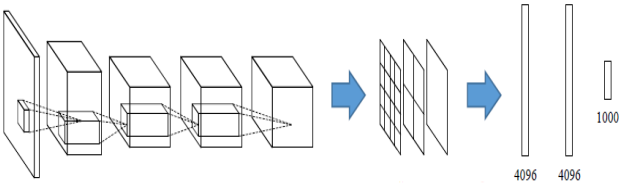
شبکه SPP-Net در سال ۲۰۱۴ معرفی شده است. این ساختار در تشخیص اشیاء نسبت به شبکه‌های پیشین خود عملکرد بهتری دارد [۱۳]. این روش، ۲۰ تا ۶۰ برابر سریعتر از شبکه R-CNN عمل می‌کند. در این روش، اندازه تصویر ورودی



شکل ۲ شمای کلی از روشR-CNN ارائه شده در مرجع [۱۵]

می‌تواند متغیر باشد. در کارهای قبلی، ثابت بودن تصویر ورودی برای لایه تمام متصل[1] امری ضروری بوده است.

در این روش، با طراحی یک لایه نظرسنجی هرمی-فضایی[2] قبل از لایه تمام متصل، مشکل ثابت بودن تصویر ورودی رفع شده است. در شمای کلی روش مرجع [۱۳] که در شکل ۳ نشان داده شده است، بعد از لایه کانولوشن یک لایه هرمی-فضایی تعبیه شده است. در شبکه R-CNN به ازای هر ناحیه کاندید یک شبکه عمیق آموزش داده می‌شود. در مرجع [۱۳] ادعا شده که تنها یک شبکه به ازای کل تصاویر، آموزش داده می‌شود. سپس از خروجی لایه کانولوشن، نواحی مربوطه استخراج و برچسب‌گذاری می‌شوند. در واقع در این روش، سرعت بسیار بهبود یافته است.

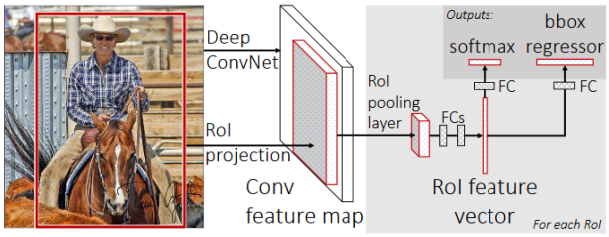


شکل ۳ شمای کلی از روش SPP-Net ارائه شده در مرجع [۱۳]

با ایده مطرح شده در SPP-Net چندین تحقیق در سال ۲۰۱۵ انجام شده است. از معروف‌ترین این شبکه‌ها می‌توان به Fast-RCNN [۱۴]، Faster R-CNN [۱۹] و R-CNN minus R [۲۰] اشاره کرد. معماری مطرح شده در Fast-RCNN در شکل ۴ نشان داده شده است. در این معماری، تصویر ورودی و چندین ناحیه مطلوب[3] (RoI) به عنوان ورودی شبکه تمام کانولوشن هستند. هر RoI توسط لایه نظرسنجی به یک نقشه ویژگی با طول ثابت تبدیل می‌شود. این نقشه‌ی ویژگی، در گام

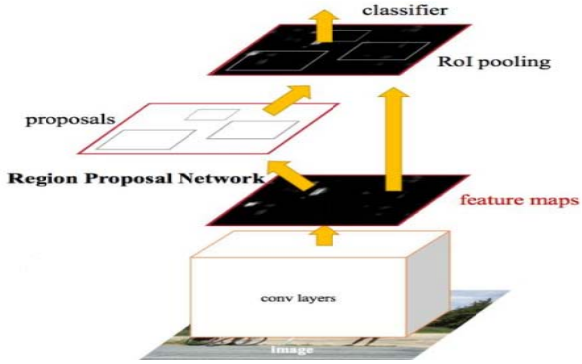
[1] Fully connected
[2] Spatial Pyramid Pooling Layer
[3] Region-of-Interest (RoI)

بعدی توسط لایه‌های تمام متصل به بردار ویژگی نگاشت می‌شود. شبکه برای هر RoI دو بردار خروجی دارد که یک خروجی احتمال هر کلاس و خروجی دیگر نواحی مرتبط به محدوده‌ی شی را مشخص می‌کند. شبکه Faster-RCNN یک شبکه تنها است. در این شبکه بعد از استخراج نقشه‌های ویژگی، شبکه‌ای با عنوان شبکه پیشنهاد ناحیه برای استخراج نواحی کاندید تعبیه شده است. بعد از شبکه پیشنهادی ناحیه (RPN) لایه نظرسنجی RoI قرار می‌گیرد. در ادامه مشابه Fast-RCNN موقعیت مرتبط با نواحی محدوده شی تخمین زده می‌شود. در این روش از مفهوم لنگر[1] استفاده شده است. لنگر، هر ناحیه پیشنهاد شده را توسط پنجره لغزان به محدوده‌هایی با اندازه متفاوت از لحاظ نسبت ابعاد تقسیم می‌کند. شکل ۵ شمای کلی روش پیشنهاد شده در شبکه Faster-RCNN را نشان می‌دهد.



شکل ۴شمای کلی از روش Fast-RCNN ارائه شده در مرجع [۱۴]

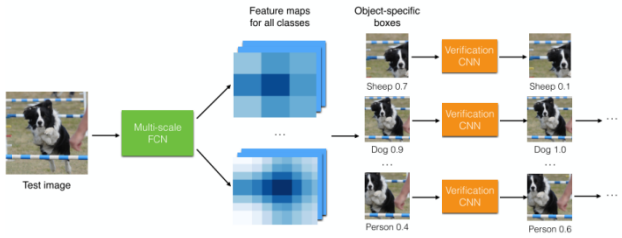
شبکه R-CNN minus R شکل ساده شده‌ی روش RCNN است. در مقاله [۲۰]، نقش تولید نواحی کاندید در آشکارسازهای مبتنی بر CNN مورد بررسی قرار گرفته است. در این مقاله، اشاره شده است که اطلاعات مهم هندسی در CNN دیده نشده است. در همین راستا، یک آشکارساز جدید برای تولید ناحیه کاندید معرفی شده است. ترکیب این آشکارساز با SPP-Net عملکرد بهتر و سریعتری ایجاد می‌کند. در ساده‌سازی آشکارسازهای مبتنی بر CNN چندین مرحله یادگیری در یک الگوریتم تلفیق شده است. یکی دیگر از شبکه‌های عمیق معرفی شده در سال ۲۰۱۶، ProNet [۲۱]



شکل ۵ شمای کلی از روش Faster-RCNN ارائه شده در مرجع [۱۹]

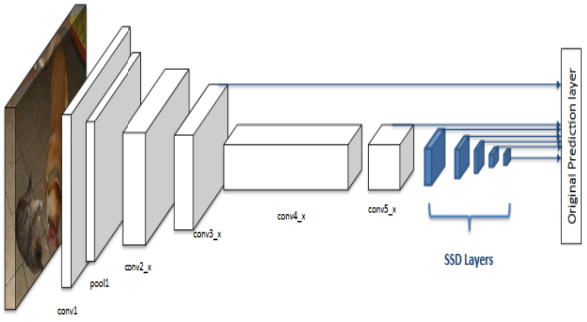
[1] anchors

است. در این روش از شبکه‌های عصبی کارامدتر برای پیشنهاد نواحی شامل شی و هم چنین شبکه تمام کانولشن با چندین مقیاس استفاده شده است. این روش، از شبکه‌هایی با قدرت بیشتر ولی کندتر برای پیشنهاد ناحیه استفاده می‌کند. این شبکه، امتیازی به محدوده مربوط به اشیاء با در نظر گرفتن مکان‌ها و مقیاس‌های متفاوت اختصاص می‌دهد. برای انتخاب برچسب شی از روش‌های آبشاری یا درختی استفاده می‌شود. در واقع در این روش، نواحی مختلف از یک شی با مکان‌ها و مقیاس‌های متفاوت انتخاب می‌شوند. در گام بعدی با استفاده از CNN احتمال شی بودن و برچسب شی به هر ناحیه اختصاص داده می‌شود. در نهایت برچسب شی و ناحیه مربوط به آن با توجه به ساختار درخت آبشاری انتخاب می‌شود.



شکل ۶ شمای کلی از روش ProNet ارائه شده در مرجع [۲۱]

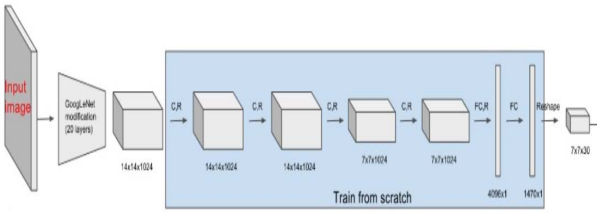
شبکه SSD[2] [۲۲] یکی دیگر از شبکه‌های مطرح شده در حوزه تشخیص شی است. این روش تنها از یک شبکه عصبی عمیق برای تشخیص شی استفاده می‌کند. شمای کلی این روش در شکل ۷ نشان داده شده است. در این روش، یک پیش‌بینی کننده کانولوشن برای تشخیص شی معرفی شده است. در این پیش‌بینی کننده، بالای هر نقشه ویژگی، کانولوشنی به همراه مجموعه‌ای از فیلترها تعبیه شده است. وظیفه این قسمت، پیش‌بینی دسته‌ی کلاس‌ها و نسبت ابعاد است. در این روش مشابه مفهوم لنگرها در Faster R-CNN استفاده شده است. با این تفاوت که SSD این مفهوم را در چندین نقشه ویژگی در تفکیک‌پذیری[3] متفاوت اعمال می‌کند. این روش، عملکرد بهتری از روش Faster R-CNN از خود نشان داده است.



شکل ۷ شمای کلی از روشSSD ارائه شده در مرجع [۲۲]

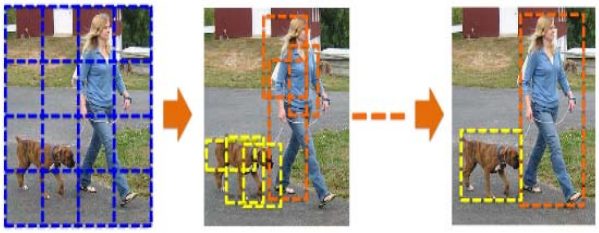
[2] Single Shot MultiBox Detector (SSD)
[3] Resolution

روش YOLO[1] [۲۳] یکی دیگر از روش‌های مطرح شده در سال ۲۰۱۶ است. روش‌های قبلی با دو شبکه پیشنهادی RPN و تخمین کلاس، به دلیل خط لوله[2] کند، سختی در بهینه‌سازی خط لوله‌های مجزا و نیاز به آموزش خط لوله‌ها برای هر بخش، پیچیده بودند. در همین راستا، در این مقاله آشکارسازی به عنوان یک مسئله تخمین در نظر گرفته شده است. همچنین شبکه کانولوشن مجزا برای طراحی معرفی شده است که از لحاظ پیاده‌سازی ساده و سریع است. در این روش، تصویر به تعداد ثابتی شبکه مشبک[3] تبدیل می‌شود. در صورتی که مرکز یک شی در محدوده یک شبکه مشبک قرار بگیرد این شبکه می‌تواند برای تشخیص شی قابل قبول باشد. در ادامه برای هر شبکه مشبک قابل قبول، تعداد ثابتی محدوده در نظر گرفته می‌شود که از معماری شکل ۸ برای تشخیص شی بودن استفاده می‌شود. معماری پیشنهادی شامل ۲۴ لایه کانولوشن و دو لایه تمام متصل می‌باشد، و ۲۰ لایه اول مشابه معماری GoogleNet می‌باشد. محدودیت این روش در تشخیص اشیاء با اندازه کوچک و اشیاء با ابعاد متفاوت است. همچنین تابع خطا استفاده شده تخمینی بوده و میزان خطای تخمینی برای اشیاء با نسبت ابعاد متفاوت می‌تواند یکسان باشد [۲۳].

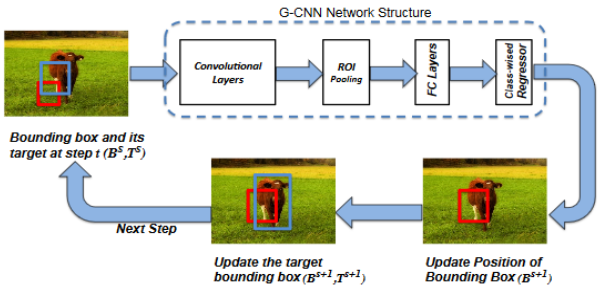


شکل ۸ شمای کلی از روشYOLO ارائه شده در مرجع [۲۳]

روش G-CNN [۲۴] یک تکنیک تشخیص برپایه CNN و بدون استفاده از الگوریتم پیشنهاد کاندید است. همانطور که در شکل ۹ نشان داده می‌شود G-CNN با یک شبکه مشبک چند مقیاسه با محدوده ثابت شروع می‌شود. یک رگرسیون تکراری برای جابجایی محدوده و مقیاس عناصر شبکه به منظور استخراج اشیاء آموزش داده می‌شود. در واقع G-CNN مسئله تشخیص اشیاء را به عنوان پیدا کردن مسیر از شبکه مشبک ثابت به محدوده اطراف اشیاء مدل می‌کند. G-CNN با محدوده حدود ۱۸۰ ناحیه در شبکه‌های مشبک با مقیاس متفاوت، عملکردی مشابه با روش Fast R-CNN با k^2 محدوده تولید شده توسط تکنیک‌های پیشنهاد کاندید دارد. در واقع این روش، بدلیل حذف مرحله پیشنهاد ناحیه و کاهش تعداد نواحی پیشنهادی سریعتر از روش Fast R-CNN است. شکل ۱۰ شمای کلی این روش برای تشخیص شی را نشان می‌دهد. در این روش یک تابع هدف تعریف می‌شود و با استفاده از شیب نزولی تصادفی بهینه می‌شود. در همین راستا، G-CNN در مرحله آموزش با تعدادی تکرار، محدوده تخمین را به محدوده واقعی نزدیک می‌کند. معماری شبکه CNN استفاده شده در این روش AlexNet و VGG است.

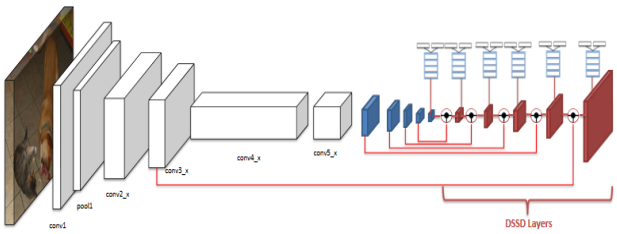


شکل ۹مراحل انتخاب نواحی در G-CNN ارائه شده در مرجع [۲۴]



شکل ۱۰ شمای کلی از روش G-CNN ارائه شده در مرجع [۲۴]

از مهمترین کارهای انجام شده در سال ۲۰۱۷ می‌توان به روش SSD+DSSD [۲۵]، YOLO9000 [۲۳] و RFCN [۲۶] اشاره کرد. در روش DSSD[4] ابتدا دسته‌بند پیشرفته ۱۰۱-Residual [۲۷] با سیستم تشخیص سریع SSD [۲۲] ترکیب شده است. در ادامه شبکه ۱۰۱-SSD+Residual با معرفی لایه‌های کانولوشن معکوس[5] تقویت شده تا بتواند عملکرد خوبی را در تشخیص اشیاء کوچک و بهبود عملکرد تشخیص شی داشته باشد. روش پیشنهادی در شکل ۱۱ نشان داده شده است.



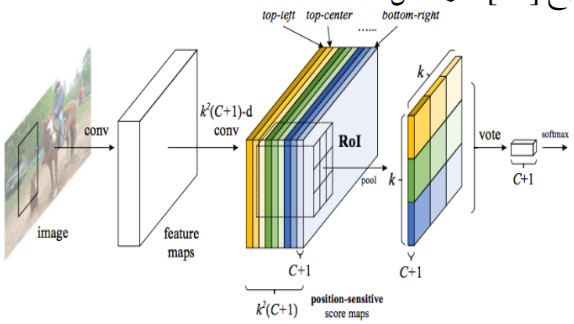
شکل ۱۱ شمای کلی روش SSD+DSSD ارائه شده در مرجع [۲۵]

روش YOLO9000 نسبت به YOLO سریعتر، قویتر و بهتر بوده است. در این روش از نرمال‌سازی دسته‌ای [۲۸]، دسته‌بند با تفکیک‌پذیری بالاتر، کانولوشن با محدوده‌های لنگر، آموزش با چندین مقیاس، استفاده از دسته‌بند سلسله مراتبی و اتصال دسته‌بندی با آشکارسازی استفاده شده است.

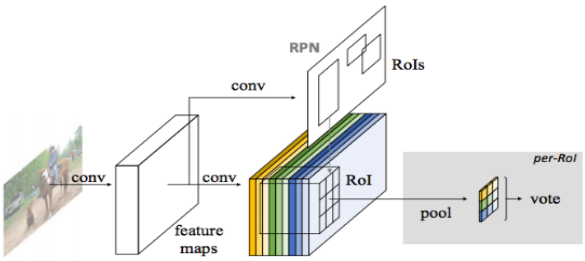
[1] You Only Look Once (YOLO)
[2] Pipeline
[3] Grid
[4] Deconvolutional Single Shot Detector (DSSD)
[5] deconvolution

با توجه به مطالب ارائه شده در این بخش، استفاده از شبکه‌های عصبی عمیق برای تشخیص شی را می‌توان به وسیله لایه نظرسنجی ناحیه مطلوب به دو زیرشبکه تقسیم نمود [۱۴]. دسته اول، زیرشبکه تمام کانولوشن مستقل از RoIs با اشتراک‌گذاری محاسبات و دسته دوم زیرشبکه بر پایه RoI بدون اشتراک‌گذاری محاسبات است. این تحلیل [۸] به لحاظ تاریخی از معماری‌های پیشگام طبقه‌بندی مانند AlexNet [۲۹] و VGG Nets [۳۰] بدست آمده است. این شبکه‌ها از دو زیرشبکه تشکیل شده است. یک زیرشبکه کانولوشن که در انتهای آن یک لایه نظرسنجی فضایی تعبیه شده و زیرشبکه دیگر شامل چندین لایه تمام متصل است. بنابراین آخرین لایه نظرسنجی فضایی در شبکه‌های دسته‌بندی تصویر به لایه نظرسنجی RoI در شبکه‌های تشخیص اشیاء تبدیل می‌شوند [۱۳, ۱۴, ۱۹]. اما شبکه‌های دسته‌بندی تصویر ResNets [۲۷] و GoogleNets [۳۱, ۳۲] در طراحی خود از لایه تمام کانولوشن استفاده می‌کند. در این ساختار، فقط آخرین لایه به صورت تمام متصل است. این لایه برای تشخیص اشیاء به شبکه اضافه می‌شود. در این نوع از شبکه‌ها، از لایه‌های تمام کانولوشن برای ایجاد اشتراک‌گذاری، زیرشبکه کانولوشن در معماری تشخیص شیء و زیرشبکه مبتنی بر RoI بدون لایه مخفی استفاده می‌شود. با این حال، در مرجع [۲۶] اشاره شده است که اشیایی که با دقت طبقه‌بندی برتر شبکه مطابقت ندارند دارای دقت تشخیص پایین‌ترند. برای اصلاح این مسئله، در مقاله ResNet [۲۷]، لایه نظرسنجی RoI آشکارساز Faster R-CNN [۱۹] به صورت غیرطبیعی بین دو مجموعه لایه‌های کانولوشن قرار گرفته شده است. این کار، زیرشبکه مبتنی بر RoI عمیق‌تری را ایجاد می‌کند که سبب بهبود دقت می‌شود. محاسبات هر RoI اشتراک گذاشته نمی‌شود و سرعت سیستم را کاهش می‌دهد.

یکی دیگر از روش‌های مشهور ارائه شد در کاربرد شناسایی اشیاء روش R-FCN است. این روش در سال ۲۰۱۷ معرفی شده است. این روش بهترین عملکرد را نسبت به روش‌های مطرح شده در سال‌های اخیر داشته است. در روش R-FCN ارائه شده در مرجع [۲۶]، از شبکه‌های کانولوشن کاملا متصل مبتنی بر ناحیه برای تشخیص دقیق و کارآمد شیء استفاده شده است. این روش در مقایسه با آشکارسازهای قبلی مبتنی بر ناحیه مانند Fast / Faster R-CNN، از کانولوشن‌های کاملا متصل استفاده می‌کند. این کانولوشن‌های کاملا متصل، تقریبا تمام محاسبات را در تمام تصویر به اشتراک می‌گذارند. برای رسیدن به این هدف، از مفهومی به عنوان نگاشت‌های امتیازدهی حساس به موقعیت[1] استفاده شده است. در شکل ۱۲ شمای کلی از این روش نشان داده شده است. در این روش، کانال‌ها نشان دهنده مکان هر RoI در تصویر ورودی می‌باشند. هر کانال برای مکان خاصی طراحی شده است. معماری کلی ارائه شده در مرجع [۲۶]، در شکل ۱۳ نشان داده شده است.



شکل ۱۲ شمای کلی از روش نگاشت‌های امتیازدهی حساس به موقعیت ارائه شده در مرجع [۲۶]

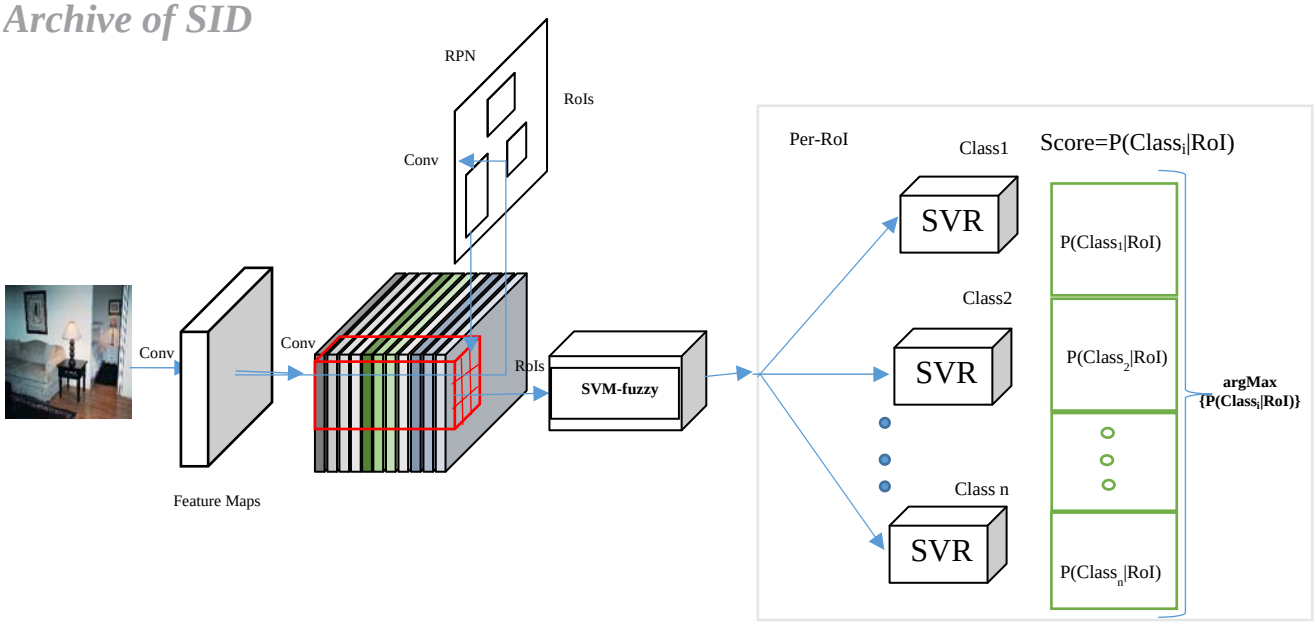


شکل ۱۳ شمای کلی از روش R-FCN ارائه شده در مرجع [۲۶]

همانطور که در شکل مشخص شده است، یک شبکه RPN[۱۹] برای پیشنهاد RoIهای کاندید معرفی شده است. RoIهای کاندید به نقشه امتیاز[2] اعمال می‌شوند. در این شبکه وزن لایه‌های قابل یادگیری، کانولوشن‌ها هستند که برای تصویر ورودی محاسبه می‌شوند. هزینه محاسبه هر RoI ناچیز است. RPN یک شبکه تمام کانولوشن است. در این شبکه، ویژگی‌های استخراج شده بین RPN و R-FCN به اشتراک گذاشته می‌شود. معماری R-FCN به منظور دسته‌بندی RoIها به دسته‌های اشیاء و پس‌زمینه، طراحی شده است. آخرین لایه کانولوشن، مجموعه‌ای از k^2 نگاشت امتیازدهی حساس به موقعیت می‌باشد که برای هر دسته (C) ارائه خواهد داد. در واقع تعداد کانال لایه خروجی، $(C+1)k^2$ است. مجموعه‌ای از k نقشه‌های امتیاز به k*k شبکه مشبک فضایی توصیف‌کننده مکان نسبی وابسته است. برای نمونه به ازای k=3، نه نگاشت امتیاز به صورت { top-left, top-center,top-right, ..., bottom-right} برای هر کلاس اختصاص داده می‌شود. آخرین

[1] Position-Sensitive Score Maps

[2] Score Maps



شکل ۱۴: نمای کلی روش پیشنهادی برای تشخیص و برچسب‌زنی اشیاء موجود در تصویر

لایه R-FCN، یک لایه نظرسنجی RoI حساس به موقعیت است. این لایه خروجی آخرین لایه کانولوشن و تولید کننده امتیاز برای هر RoI را تجمیع می‌کند. شبکه عمیق استفاده شده در این مقاله دارای معماری ResNet-۱۰۱ [۹] است. ۱۰۱-ResNet شامل ۱۰۰ لایه کانولوشن، به دنبال آن لایه نظرسنجی متوسط سراسری و لایه تمام متصل ۱۰۰۰ کلاسه است. در این مقاله دو لایه آخر حذف شده و فقط لایه کانولوشن برای محاسبه نقشه‌های ویژگی استفاده شده است. جدول ۱ نمایی کلی از ساختار شبکه‌های ResNet را نشان می‌دهد. وزن‌دهی اولیه استفاده شده در این مقاله از وزن‌های اولیه کانولوشن حاصل از آموزش بر روی مجموعه ImageNet است [۲۶].

جدول ۱ معماری ResNet برای آموزش ImageNet

Layer name	Output size	18-layer	34-layer	50-layer	101-layer
Conv1	112×112	7×7,64 Stride 2			
Conv2.x	56×56	3×3 Max pool Stride 2			
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3.x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4.x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$
Conv5.x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pool, 1000-d fc, softmax			
FLOPs		$1.8*10^9$	$3.6*10^9$	$3.8*10^9$	$7.6*10^9$

۳ استفاده از R-FCN با معماری جدید

همانطورکه در بخش قبل توضیح داده شد، شبکه R-FCN یکی از جدیدترین شبکه‌های عمیق ارائه شده برای تشخیص اشیاء است. این شبکه از ترکیب شبکه RPN و شبکه تمام کانولوشن ساخته شده است. خروجی این شبکه‌ها با توجه به موقعیت قرار گرفتن اشیاء، برای هر RoI امتیازدهی شده‌اند. برای طبقه‌بندی از تابع زیان آنتروپی متقابل استفاده شده که در رابطه (۱) نشان داده شده است. تابع زیان[1] استفاده شده در مرجع [۲۶] برای هر RoI، جمع شده است و زیان تخمین محدوده[2] به صورت رابطه (۲) است.

$$CEL = -\sum_j y^{*(j)} \log \sigma(o)^{(j)} \qquad (۱)$$

$$L(S,t_{x,y,w,h}) = L_{cls}(S_{c^*}) + \lambda[c^* > 0]L_{reg}(t,t^*) \qquad (۲)$$

در رابطه (۱) y^* خروجی مطلوب، O خروجی شبکه و (j) نشان دهنده زامین بعد بردار است. در رابطه (۲)، c^* برچسب مطلوب RoI است. $c^*=0$ به معنی پس‌زمینه است. در واقع $[c^*>0]$ شرط مشخص کننده پس‌زمینه است. در بخش‌های غیر از پس‌زمینه، c^* برابر با یک و در بقیه‌ی نواحی صفر است. هر RoI با همپوشانی بیش از ۵۰% با نواحی مطلوب، به عنوان نمونه مثبت در نظر گرفته می‌شود. t و t^* به ترتیب مقادیر مرتبط با ناحیه تخمینی و ناحیه مطلوب هستند. $L_{cls}(S_{c^*}) = -\log(S_{c^*})$ زیان آنتروپی متقابل برای طبقه‌بندی است. $L_{reg}(t,t^*)$ تابع زیان تخمین محدوده بر اساس مرجع [۱۴] است. تابع زیان تخمین محدوده در رابطه (۳) و (۴) بیان شده‌اند.

[1] Loss Function
[2] Box regression Loss

$$L_{reg}(t,t^*) = \sum_{i\in\{x,y,w,h\}} smooth_{L_1(t_i - t_i^*)} \quad (۳)$$

$$smooth_{L_1(x)} = \begin{cases} 0.5x^2 & if\ x<1 \\ |x|-0.5 & otherwise \end{cases} \quad (۴)$$

در این رابطه‌ها از تابع زیان $L1(.)$ استفاده شده که یک تابع زیان نرم یک است. دلیل استفاده از نرم یک حساسیت کمتر آن به داده‌های پرت در مقایسه با نرم دو استفاده شده در مرجع [۱۵] است.

استفاده از تابع زیان مناسب، در دقت شبکه بسیار تاثیرگذار است. به همین دلیل در این مقاله راهکاری برای بهبود شبکه R-FCN با استفاده از تابع زیان جدید معرفی شده است. مرجع [۳۳]، دوازده تابع زیان معروف را معرفی و تحلیل کرده است که این توابع قابلیت استفاده در شبکه‌های عمیق را دارند. نتایج حاصل نشان می‌دهد که استفاده از تابع زیان وابستگی زیادی به کاربرد مورد نظر دارد. در این مقاله به تحلیل و مقایسه رابطه‌ی بین تابع زیان اختلاف کوشی-شوارتز[1] و زیان آنتروپی متقابل پرداخته شده است. شایان ذکر است، این تابع زیان در شبکه‌های عمیق برای شناسایی اشیاء و روش R-FCN استفاده نشده است. آزمایش‌ها نشان می‌دهند که تابع زیان اختلاف کوشی-شوارتز نسبت به زیان آنتروپی متقابل از لحاظ سرعت و عملکرد بهینه‌تر است [۳۳]. در این مقاله به جای تابع زیان شبکه R-FCN اصلی (تابع زیان آنتروپی) از تابع اختلاف کوشی-شوارتز استفاده شده است. این تابع زیان به صورت رابطه (۵) تعریف شده است [۳۴].

$$\mathrm{DCS} = -\log \frac{\sum_j \sigma(o)^j y^{*(j)}}{\|\sigma(o)\|_2 \|y\|_2} \quad (۵)$$

در رابطه (۵) y^* خروجی مطلوب، O خروجی شبکه، σ تخمین احتمالی[2] و (j) نشان دهنده زامین بعد بردار است.

در R-FCN پس‌زمینه به عنوان یک کلاس مجزا در نظر گرفته شده است. در روش پیشنهادی نمی‌خواهیم تعداد کلاس‌ها را افزایش دهیم در همین راستا یک فاز مجزا برای جدا نمودن پس‌زمینه با اشیاء در نظر گرفته شده است. در همین راستا ابتدا یک سیستم فازی دوکلاسه مبتنی بر SVM[3] طراحی کردیم که مشخص می‌کند آیا ناحیه مربوط به شی می‌باشد یا نه، در صورتی که متعلق به شی نباشد برچسب‌زنی برای این ناحیه انجام نمی‌شود. معمولا برای هر تصویر تعداد زیادی ناحیه کاندید معرفی می‌شود. با این عمل فقط نواحی که می‌توانند به عنوان شی باشند مورد ارزیابی قرار می‌گیرند. این عمل زمان برچسب‌زنی برای کل تصویر را بسیار کاهش می‌دهد. برای مشخص نمودن نواحی مرتبط با اشیاء از یک رویکرد جدید استفاده شد که در بخش ۱-۳ به تشریح آن می‌پردازیم.

روش R-FCN از رای دهی برای انتخاب برچسب هر RoI استفاده کرده است. ولی در روش پیشنهادی از SVR[4] استفاده شده است. در واقع به ازای هر کلاس از اشیاء، یک ماشین بردار در نظر گرفته شده است. در روش پیشنهادی هر SVR میزان تعلق هر RoI به هر کلاس را مشخص می‌کند. در نهایت در RoIهای متعلق به شی، عمل ماکزیمم‌گیری انجام و برچسب هر RoI مشخص می‌شود. در روش پیشنهادی همه‌ی SVRها به صورت خط لوله اجرا شدند و زمان پردازش کاهش یافته است. برای هر شی یک SVR اجرا شده است که برای این منظور هر SVR در یک خط لوله مجزا اجرا شده است. در واقع به ازای هر شی یک SVR به صورت همزمان اجرا شده است. شمای کلی روش پیشنهادی در شکل ۱۴ نشان داده شده است. در واقع در هر SVR میزان تعلق هر RoI به یک کلاس مشخص محاسبه می‌شود و در نهایت با استفاده از رابطه ۶ برچسب کلاس تعیین می‌شود. در این رابطه n تعداد کلاس‌ها است.

$$ClassNumber = \arg\max_i^n \{P(Class_i \mid RoI)\} \quad (۶)$$

۳-۱ ماشین بردار پشتیبان فازی

ماشین بردار پشتیبان به عنوان یک ابزار قدرتمند برای طبقه‌بندی و رگرسیون، در بسیاری از کاربردهای عملی مورد استفاده قرار گرفته شده‌است [۳۵-۳۷]. نسخه‌های بسیاری از ماشین بردار پشتیبان در سال‌های اخیر معرفی شده است. یکی از معروف‌ترین آنها می‌توان نسخه مبتنی بر فازی اشاره نمود. اغلب SVM فازی برای حل مسائلی استفاده می‌شود که الگوهای متعلق به یک کلاس اغلب نقش‌های مهمتری در طبقه‌بندی دارند. در روش پیشنهادی بعد از مشخص نمودن RoIهای موجود در هر تصویر توسط ماشین SVM معمولی میزان تعلق هر ناحیه به کلاس مشخص می‌باشد. در این گام می‌خواهیم از SVM فازی دوکلاسه برای برچسب‌زنی تصویر استفاده نمائیم. در این راستا از روش معرفی شده در مرجع [۳۸] برای طبقه‌بندی استفاده می‌کنیم. خروجی این طبقه‌بند دوکلاسه فازی مشخص کننده شی یا پس‌زمینه بودن است. در صورتی که پس‌زمینه تشخیص داده شود این ناحیه کنار گذاشته می‌شود ولی در صورت تشخیص وجود شی، در مرحله بعدی برچسب کلاس مشخص می‌شود. این عمل سبب کاهش عمل برچسب‌زنی تصویر می‌شود.

فرض شود که مجموعه آموزش برای مسئله طبقه‌بندی دوکلاسه به صورت زیر تعریف شود:

[1] Cauchy-Schwarz Divergence Loss (D_{CS})
[2] Probability estimate
[3] Support Vector Machine (SVM)
[4] Support Vector Regression(SVR)

$$T^* = \{(x_1, m_1), (x_2, m_2), ..., (x_l, m_l)\} \quad (۷)$$

در رابطه (۷)، x_i نمونه‌ها، $m_i \in [0,1]$ عضویت فازی می‌باشد که میزان تعلق iامین مشاهده x_i به کلاس مثبت را ارزیابی می‌کند و l تعداد نمونه‌های مجموعه آموزش می‌باشد. در این روش P نمونه از مشاهدات به عنوان $\{(\tilde{x}_1, \tilde{m}_1), (\tilde{x}_2, \tilde{m}_2), ..., (\tilde{x}_P, \tilde{m}_P)\}$ مشاهدات مثبت و q نمونه از $\{(\hat{x}_1, \hat{m}_1), (\hat{x}_2, \hat{m}_2), ..., (\hat{x}_q, \hat{m}_q)\}$ مشاهدات به عنوان نمونه‌های منفی در نظر گرفته می‌شود. برچسب نمونه‌ها از رابطه $Y_i = 2m_i - 1$ محاسبه می‌شود که برای نمونه‌های مثبت و منفی ماتریس‌های قطری $Y_1 = diag(\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_p)$ و $Y_2 = diag(\hat{y}_1, \hat{y}_2, ..., \hat{y}_q)$ تعریف می‌شود. در اینجا $\tilde{y}_i = 2\tilde{m}_i - 1, (i = 1,2,..., p)$ و $\hat{y}_j = 2\hat{m}_j - 1, (j = 1,2,..., q)$ می‌باشد. $m_i = 1$ زمانی می‌باشد که x_i نمونه مثبت بوده و برچسب مرتبط به آن $Y_i = 2m_i - 1 = 1$ می‌باشد. $m_i = 0$ در صورتی x_i نمونه منفی بوده و برچسب مرتبط به آن $Y_i = 2m_i - 1 = -1$ می‌باشد. در این مقاله کلاس مثبت نشان‌دهنده شی و کلاس منفی نشان‌دهنده پس‌زمینه می‌باشد. مسئله بهینه سازی به صورت زیر تعریف می‌شود:

$$\min_{w_1, b_1, \zeta_2} \frac{1}{2}\|Aw_1 + e_1 b_1\|_2^2 + \frac{1}{2}c_1\left(w_1^2 + b_1^2\right) + c_2 e_2^T \zeta_2 \quad (۸)$$
$$s.t. \quad Y_2(Bw_1 + e_2 b_1) \geq Y_2^2 e_2 - Y_2^2 \zeta_2, \zeta_2 \geq 0$$

$$\min_{w_2, b_2, \zeta_1} \frac{1}{2}\|Bw_2 + e_2 b_2\|_2^2 + \frac{1}{2}c_3\left(w_2^2 + b_2^2\right) + c_4 e_1^T \zeta_1 \quad (۹)$$
$$s.t. \quad Y_1(Aw_2 + e_1 b_2) \geq Y_1^2 e_1 - Y_1^2 \zeta_1, \zeta_1 \geq 0$$

در رابطه فوق c_1 ، c_2 ، c_3 و c_4 پارامترهای جریمه مثبت، ζ_1 و ζ_2 متغیرهای نرم، ماتریس $A \in \Re^{p \times n}$ نمونه‌های مرتبط با کلاس مثبت و $B \in \Re^{q \times n}$ نمونه‌های مرتبط با کلاس منفی می‌باشند. بعد از حل این رابطه توسط راگرانژ چندگانه که در مرجع [۳۸] به تفصیل بیان شده به رابطه‌های زیر می‌رسیم.

$$v_1 = (H^T H + c_1 I)^{-1} G^T Y_2^2 \alpha \quad \text{where} \quad v_1 = \begin{bmatrix} w_1^T b_1 \end{bmatrix}^T, H = [B\, e_2] \quad (۱۰)$$

$$v_2 = (G^T G + c_3 I)^{-1} H^T Y_1^2 \gamma \quad \text{where} \quad v_2 = \begin{bmatrix} w_2^T b_2 \end{bmatrix}^T, G = [H\, A\, e_1] \quad (۱۱)$$

بعد از بدست آوردن v_1 و v_2 ، برای هر نمونه جدید $x \in \Re^n$ برچسب نمونه توسط رابطه (۱۲) بدست می‌آید.

$$x \in W_k, k = \arg\min_{i=1,2} \left\{ \frac{\left|w_1^T x + b_1\right|}{\|w_1\|}, \frac{\left|w_2^T x + b_2\right|}{\|w_2\|} \right\} \quad (۱۲)$$

برای محاسبه مقدار عضویت از تابع عضویت معرفی شده در مرجع [۳۹] استفاده شده است. برای نمونه x_i با برچسب مثبت (+۱) عضویت فازی به صورت رابطه (۱۳) بیان می‌شود

$$m_1(x_i) = 0.5 + \frac{\exp(C_0(d_{-1}(x_i) - d_1(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - \exp(-C_0))} \quad (۱۳)$$
$$m_{-1}(x_i) = 1 - m_1(x_i)$$

برای نمونه x_i با برچسب منفی (-۱) عضویت فازی به صورت رابطه (۱۴) بیان می‌شود.

$$m_{-1}(x_i) = 0.5 + \frac{\exp(C_0(d_1(x_i) - d_{-1}(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - \exp(-C_0))} \quad (۱۴)$$
$$m_1(x_i) = 1 - m_{-1}(x_i)$$

$d_1(x_i)$ فاصله بین x_i و میانگین کلاس مثبت، $d_{-1}(x_i)$ فاصله بین x_i و میانگین کلاس منفی، d فاصله بین میانگین کلاس‌های مثبت و منفی، C_0 یک ثابت به عنوان کنترل‌کننده تابع عضویت می‌باشد.

۴ آزمایش‌ها

در این بخش از مقاله مجموعه داده، معیارهای ارزیابی و نتایج بدست آمده معرفی می‌شوند. پیاده‌سازی‌های انجام شده در زبان برنامه‌نویسی پایتون بر روی سخت‌افزاری با مشخصات:کارت گرافیک GTX 1080، حافظه ۳۲G و پردازنده Core i7 4790k 4GHz انجام شده است. از چارچوب معروف و سریع Caffe برای یادگیری عمیق استفاده شده است. لازم به ذکر است، تمامی نتایج بدست آمده از روش‌های دیگران که برای مقایسه در مقاله معرفی شده بر روی یک کامپیوتر با مشخصات یکسان بوده است.

جدول ۲ اشیاء موجود پایگاه داده SUN

شی	Countertop	chair occluded	Chair	Ceiling	Cabinet	Buildings	Building	bed crop	Bed
تعداد تصاویر	۱۰۹	۱۰۵۷	۳۴۳	۱۱۱۵	۱۳۲۵	۴۵۸	۷۹۵	۲۳۳	۵۰۷
شی	night table	Mirror	microwave	Flowers	Floor	Faucet	desk lamp	Cushion	curtain
تعداد تصاویر	۲۰۶	۳۲۹	۱۴۹	۱۰۹	۱۵۶۶	۵۵۳	۵۵۲	۴۹۳	۸۷۰
شی	Stove	Skyscraper	Sky	Sink	Plant	pillow occluded	Pillow	Painting	night table occluded
تعداد تصاویر	۲۳۱	۶۲۲	۲۴۵	۲۵۵	۴۵۱	۳۱۴	۳۸۷	۵۵۷	۲۸۴
شی	Worktop	Window	washbasin	Wall	trees	Tree	Towel	Toilet	Table
تعداد تصاویر	۵۱۹	۱۳۹۳	۱۶۲	۴۴۶۲	۱۴۳	۱۱۲	۳۰۲	۱۱۹	۱۹۱

در آزمایشات انجام شده توسط روش پیشنهادی، از هسته[1] گوسین $K(x,y)=e^{-\|x-y\|^2/2\sigma^2}$ برای SVM فازی دوکلاسه و SVR استفاده شده است.

۴-۱ پایگاه داده

در این مقاله برای آشکارسازی و برچسب‌زنی تصاویر از پایگاه داده SUN[2] [۴۰] استفاده شده است. این پایگاه داده یک مجموعه بزرگ از تصاویر برای دسته‌بندی صحنه‌ی تصویر است. این پایگاه داده، شامل ۱۳۱۰۶۷ تصویر در ۹۰۸ دسته مختلف و شامل ۳۸۱۹ شی متفاوت در تصاویر است. تصاویر بدست آمده در حالت‌های مختلف به مانند محیط بسته، باز، طبیعی و غیره تهیه شده است. از آنجایی که هدف در این مقاله، آشکارسازی و برچسب‌زنی اشیاء می‌باشد، اشیایی انتخاب شده‌اند که از نظر تکرار قابل قبول بوده‌اند. در واقع اشیا با تعداد تکرار اندک نادیده گرفته شده‌اند. در همین راستا، در این مقاله ۲۱۵۱۸ تصویر در پنج دسته مختلف در شرایط مختلف انتخاب شده است. تعداد اشیاء موجود در این مجموعه داده ۳۶ شی است که در جدول ۲ تعداد هر شی و تصاویری شامل این اشیاء، نشان داده شده است. در هر تصویر امکان تکرار یک شی وجود دارد. در این مجموعه داده، ۶۰% برای آموزش و ۴۰% برای آزمون در نظر گرفته شده‌اند.

۴-۲ معیار ارزیابی

معیارهای دقت و یادآوری مهم‌ترین و رایج‌ترین پارامترها برای ارزیابی و مقایسه سامانه برچسب‌گذاری تصاویر هستند. اما معیار دیگری به نام میانگین دقت متوسط[3] (mAP)، کیفیت رتبه‌بندی تصاویر توسط روش را ارزیابی می‌کند. معیار ارزیابی استفاده شده در این مقاله میانگین دقت متوسط است [۴۱].

رتبه‌بندی تصاویر یک ویژگی مهم و ذاتی برای روش‌های بازیابی تصاویر است. میانگین دقت متوسط عملکرد بازیابی را ارزیابی می‌کند. اگر تصاویر مربوط در رتبه‌های اول بازیابی شوند، این معیار به عدد یک نزدیک خواهد شد. هر چه تصاویر مربوط در رتبه‌های آخر قرار گیرند، این معیار کمتر خواهد بود.

برای محاسبه میانگین دقت متوسط، ابتدا دقت متوسط (AP) برای پرس و جوی q محاسبه می‌شود. دقت متوسط همانند مرجع [۴۲] با رابطه‌ی (۱۵) محاسبه می‌شود.

$$AP(q)=\frac{\sum_{i=1}^{n}(P(i)\times rel(i))}{|releventimages|} \qquad (۱۵)$$

در این رابطه i رتبه در دنباله بازیابی شده، n تعداد تصاویر بازیابی شده، P(i) دقت بازیابی در i تصویر اول (تعداد تصاویر مرتبط تا رتبه iام تقسیم بر i) و rel(i) یک تابع است. مقدار این تابع، با قرارگیری تصویر مرتبط با پرس‌وجو در رتبه iام، برابر یک و در غیر این صورت برابر صفر است. $|releventimages|$ تعداد تصاویر استفاده شده در پایگاه داده را نشان می‌دهد.

میانگین دقت متوسط، با رابطه‌ی (۱۶) محاسبه می‌شود. در این رابطه، Nq تعداد پرس‌وجوها را نشان می‌دهد.

$$mAP=\frac{\sum_{i=1}^{n}AP(q)}{N_q} \qquad (۱۶)$$

در مرحله محاسبه‌ی نواحی مربوط به یک شی، از میزان هم‌پوشانی نواحی استفاده شده است. میزان مقبولیت نواحی با استفاده از رابطه (۱۷) محاسبه می‌شود [۱۹].

$$R=\frac{B_p\cap B_{gt}}{B_p\cup B_{gt}} \qquad (۱۷)$$

در رابطه‌ی (۱۷)، Bp نواحی پیش‌بینی شده و Bgt نواحی مرتبط با نواحی مطلوب است. با افزایش میزان R، دقت سیستم بالا می‌رود. در این مقاله، میزان همپوشانی بیش از ۵۰% مطلوب در نظر گرفته شده است.

۴-۳ نتایج آزمایش‌ها

در آزمایش‌های انجام شده برای هر شبکه عمیق از وزن‌های اولیه از پیش آموزش داده شده بر روی مجموعه داده ImageNet [۴۳] استفاده شده است. در روش Faster R-CNNT از شبکه‌های عمیق با معماری مختلف استفاده شده است. از مهم‌ترین معمارهای عمیق استفاده شده می‌توان به VGG[۱۶]، ZF [۴۴] و ResNet [۲۷] اشاره نمود. برای مقایسه بهتر از چهار روش معروف Fast RCNN [14]، Faster R-CNN [19]، SPPNet [۱۳] و R-FCN [۲۶] استفاده شده است. در مراحل آموزش و آزمون در همه‌ی روش‌های معرفی شده از مجموعه داده SUN استفاده شده است. نتایج حاصل از آزمایش‌ها بر روی ۳۶ شی مجموعه داده SUN در جدول ۳ نشان داده شده است. همانطور که در این جدول نشان داده شده، روش R-FCN نسبت به سایر روش‌ها نتایج بهتری از خود نشان می‌دهد. همچنین روش پیشنهادی بهترین نتایج را از خود نشان می‌دهد.

[1] Kernel
[2] Scene Understanding
[3] Mean Average Precision (mAP)با

جدول ۳ مقایسه روش‌های مختلف برای هر شی مجموعه داده SUN

روش / شی	Fast RCNN[۱۴]	Faster RCNN[۱۹]	SPPNet[۱۳]	R-FCN[۲۶]	Proposed Method
Bed	۷۸/۸۸	**۷۹/۳۱**	۷۸/۵۸	۷۶/۵۴	۷۷/۴۴
bed crop	۴۲/۵۸	۴۶/۱۶	۴۰/۴۳	**۵۴/۶۵**	۵۲/۸۳
Building	۴۵/۷۰	۴۸/۲۶	۴۵/۵۶	۴۸/۳۹	**۴۹/۰۰**
Buildings	۱۷/۴۲	۲۰/۹۱	۱۹/۱۸	۲۲/۹۴	**۲۵/۱۴**
Cabinet	۲۹/۴۶	۳۵/۱۹	۳۲/۶۷	**۳۳/۸۵**	۳۳/۱۵
Ceiling	۷۶/۴۷	۷۶/۲۵	۷۴/۹۱	۷۷/۸۵	**۷۷/۹۶**
Chair	۴۱/۱۸	۴۴/۵۰	۳۸/۵۲	۴۷/۱۸	**۴۹/۰۱**
chair occluded	۱۸/۸۴	۲۰/۹۵	۱۷/۸۶	۳۰/۲۳	**۳۲/۰۱**
Countertop	۴۲/۸۷	۴۶/۷۵	۴۰/۰۴	۴۶/۵۷	**۵۲/۱۸**
Curtain	۴۹/۴۷	۵۲/۲۰	۵۰/۴۵	**۵۵/۶۴**	۵۵/۲۸
Cushion	۳۷/۲۷	۳۵/۲۲	۳۳/۹۹	۵۱/۸۲	**۵۲/۵۳**
desk lamp	۵۶/۸۶	۵۹/۰۴	۵۵/۱۶	۷۰/۶۲	**۷۰/۸۰**
Faucet	۶/۰۷	۹/۱۲	۵/۱۰	۲۴/۶۶	**۲۸/۷۹**
Floor	۷۲/۰۶	۷۱/۸۸	۷۳/۵۷	۷۵/۵۸	**۷۷/۲۱**
Flowers	۲۵/۱۷	۲۷/۴۱	۲۴/۸۳	۳۴/۲۳	**۴۲/۳۷**
Microwave	۲۵/۸۴	۳۳/۶۸	۲۰/۹۰	۴۵/۸۶	**۴۸/۰۷**
Mirror	۳۲/۶۷	**۳۵/۱۵**	۲۹/۹۷	۳۰/۶۹	۳۵/۱۰
night table	۶۸/۵۵	۶۸/۰۱	۶۷/۱۲	۶۷/۸۸	**۷۴/۰۴**
night table occluded	۲۸/۵۷	۳۸/۱۹	۳۲/۶۴	۵۲/۳۸	**۵۳/۸۰**
Painting	۴۴/۷۵	۴۴/۹۲	۴۲/۹۵	۴۸/۳۴	**۴۸/۷۸**
Pillow	۳۰/۲۷	۳۲/۸۱	۳۲/۶۳	۴۴/۱۳	**۴۷/۰۲**
pillow occluded	۱۳/۹۰	۶/۸۲	۱۲/۰۱	۱۷/۳۰	**۲۱/۳۷**
Plant	۲۴/۰۳	۲۴/۵۶	۲۴/۱۱	۳۰/۲۲	**۳۴/۸۹**
Sink	۲۵/۰۴	۲۱/۸۲	۲۳/۲۵	۲۵/۶۴	**۳۱/۸۸**
Sky	۸۹/۸۶	۸۹/۵۹	۸۹/۶۲	**۹۰/۱۴**	۹۰/۱۳
Skyscraper	۶۰/۹۸	۶۱/۵۱	۵۷/۴۱	۶۵/۷۲	**۶۷/۱۲**
Stove	۳۰/۱۷	۳۰/۱۳	**۳۳/۷۹**	۲۹/۱۳	۲۷/۴۱
Table	۲۷/۶۲	**۲۹/۱۱**	۲۸/۳۴	۲۵/۶۹	۲۶/۰۷
Toilet	۵۶/۱۴	**۶۹/۲۲**	۵۳/۵۱	۶۶/۱۷	۶۶/۵۹
Towel	۸/۲۴	۱۵/۸۶	۱۰/۵۳	**۲۱/۲۹**	۲۰/۲۵
Tree	۲۹/۴۵	۳۱/۵۲	۳۰/۲۲	۳۴/۵۲	**۳۵/۰۴**
Trees	۴۳/۲۱	۴۴/۸۶	۴۰/۷۹	۴۶/۶۳	**۴۹/۳۸**
Wall	۵۹/۴۷	۶۱/۶۳	۵۹/۸۰	۶۲/۶۱	**۶۳/۶۹**
Washbasin	۲۳/۱۷	۳۶/۶۲	۲۱/۲۹	**۴۰/۰۷**	۳۳/۶۴
Window	۳۴/۰۱	۳۶/۶۱	۳۲/۲۴	۴۵/۳۶	**۴۸/۰۹**
Worktop	۲۸/۲۲	۳۳/۸۲	۲۳/۷۷	۳۹/۱۹	**۴۳/۶۸**
متوسط	۳۹/۵۷	۴۲/۲۲	۳۸/۸۳	۴۶/۶۶	**۴۸/۳۸**

روش پیشنهادی در واقع، بهبود یافته روش R-FCN می‌باشد و از SVMفازی با SVR بجای رای‌گیری استفاده شده است. همچنین از تابع زیان جدید استفاده شده است. در جدول ۴ مقایسه‌ای بین روش‌های معرفی شده با معماری‌های متفاوت انجام شده است. همانطور که در این جدول نشان داده شد. روش پیشنهادی با معماری ResNet در مقایسه با روش‌های دیگر بهترین نتایج را از خود نشان می‌دهد.

مقایسه دیگری که در این مقاله انجام شد، مقایسه بر اساس اندازه لایه‌های ResNet و تعداد کانولوشن است. معمولا شبکه با تعداد لایه‌های ۵۰، ۱۰۱ و ۱۵۲ در بحث شناسایی معرفی شده و وزن‌های اولیه برای این شبکه با توجه به مجموعه ImageNet موجود می‌باشد. در همین راستا در این مقاله این تعداد لایه مورد آزمایش قرار گرفته شده است. همانطور که در جدول ۵ نشان داده شد روش پیشنهادی با معماری ResNet با

۱۰۱ لایه مختلف بهترین نتایج را ارائه داده است. همانطور که اشاره شد، استفاده از تابع زیان مناسب در افزایش دقت عملکرد شبکه عمیق بسیار موثر است. در این راستا در جدول ۶ مقایسه‌ای بین روش R-FCN با تابع زیان‌های متفاوت انجام شده است. نتایج جدول ۶ نشان می‌دهد، روش R-FCN با تابع زیان اختلاف کوشی-شوارتز نسبت به آنتروپی متقابل، عملکرد بهتری از لحاظ دقت و زمان داشته است. در این جدول نشان داده شده که متوسط زمان اجرا برای هر تصویر در تابع زیان اختلاف کوشی-شوارتز بهتر از زمان R-FCN با تابع زیان آنتروپی متقابل است. معماری استفاده شده برای هر دو شبکه عمیق، ResNet با ۱۰۱ لایه است.

جدول ۴ مقایسه روش های مختلف با معماری های مختلف

معماری / روش	ResNet	ZF	VGG
Fast RCNN[۱۴]	۳۸/۲۱	۳۲/۱۷	۳۹/۵۷
Faster RCNN[۱۹]	۴۰/۲۳	۳۹/۰۹	۴۲/۲۲
SPPNet[۱۳]	۳۷/۹۸	۳۸/۸۳	۳۶/۳۳
R-FCN[۲۶]	۴۶/۶۶	۴۱/۰۷	۴۲/۹۹
Proposed Method	**۴۸/۳۸**	**۴۴/۶۷**	**۴۵/۰۸**

جدول ۵ مقایسه معماری ResNet با تعداد لایه متفاوت

تعداد لایه / روش	ResNet-50	ResNet-101	ResNet-152
R-FCN[۲۶]	۴۳/۲۱	۴۶/۶۶	۴۶/۰۵
Proposed Method	۴۵/۹۳	**۴۸/۳۸**	۴۷/۶۷

زمان مرحله آزمون، یک معیار مهم و اساسی در بررسی روش‌های آشکارسازی و برچسب‌زنی تصاویر است. در همین راستا، جدول ۷ مقایسه‌ای بین روش‌های مطرح شده و روش پیشنهادی بر اساس متوسط زمان برای هر تصویر را نشان می‌دهد. همانطور که در این جدول نشان داده شده، روش پیشنهادی از نظر زمان اجرا، عملکرد خوبی را داشته است. نتایج جداول ۳، ۶ و ۷، مبین این نکته می‌باشد که متوسط زمان برای هر تصویر در روش پیشنهادی نسبت به R-FCN با تابع زیان اختلاف کوشی-شوارتز افزایش یافته ولی همچنان از روش R-FCN با تابع زیان آنتزوپی متقابل بهتر بوده و عملکرد بهتری از لحاظ دقت نسبت به بقیه روش‌ها دارد.

شکل ۱۵ نمونه‌های از نتایج بدست آمده از R-FCN بهبود یافته شده را نشان می‌دهد. همانطور که در شکل نشان داده شده، نتایج حاصل نشان دهنده عملکرد مناسب روش پیشنهادی در آشکارسازی و برچسب‌زنی اشیاء است.

جدول ۶ مقایسه تابع زیان در شبکه R-FCN

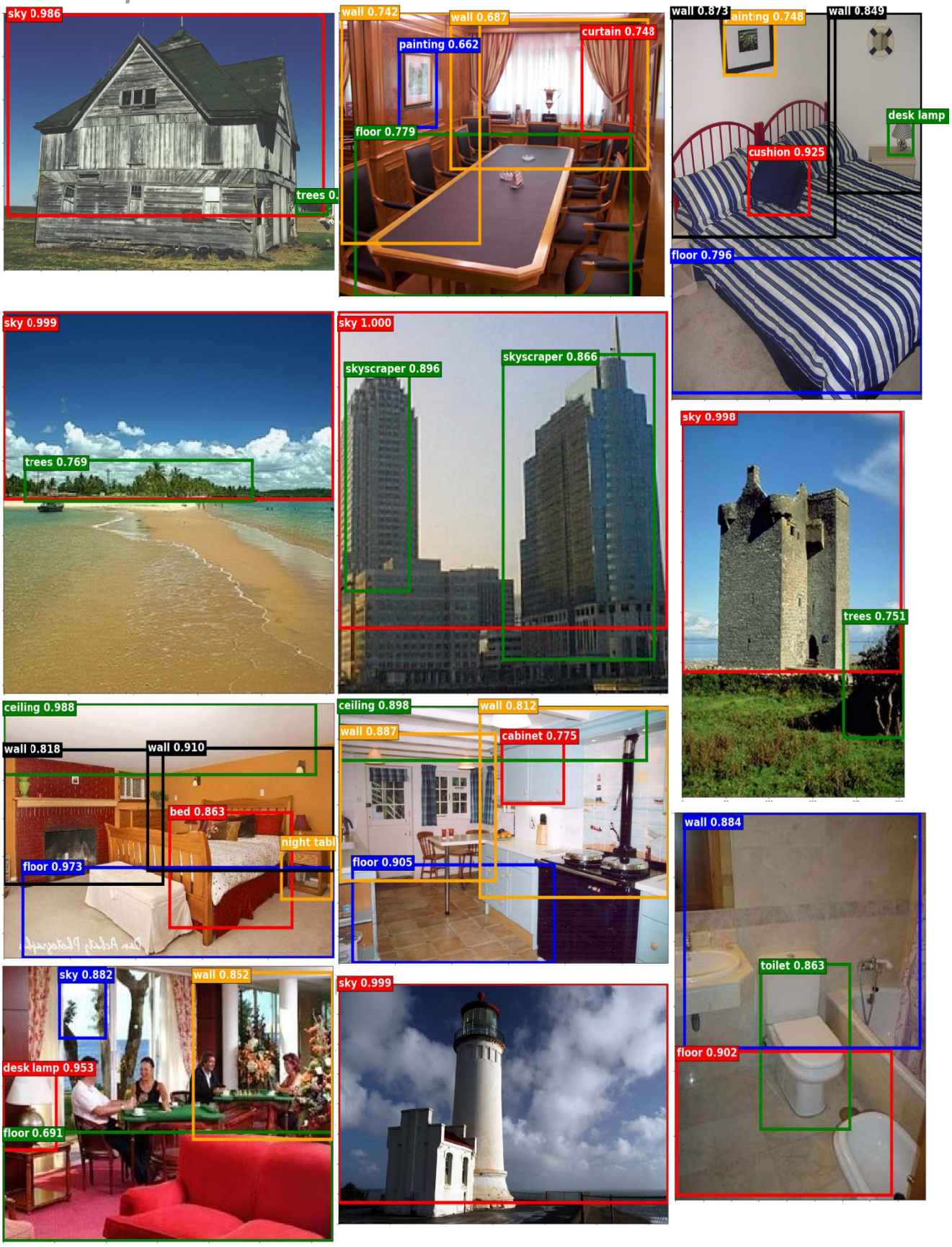
روش	R-FCN [۲۶]	R-FCN With Cauchy-Schwarz Divergence Loss
دقتmAP	۴۶/۶۶	**۴۷/۰۴**
زمان (ثانیه)	۰,۱۷	**۰,۱۱**

جدول ۷ مقایسه زمان اجرای تصویر آزمون

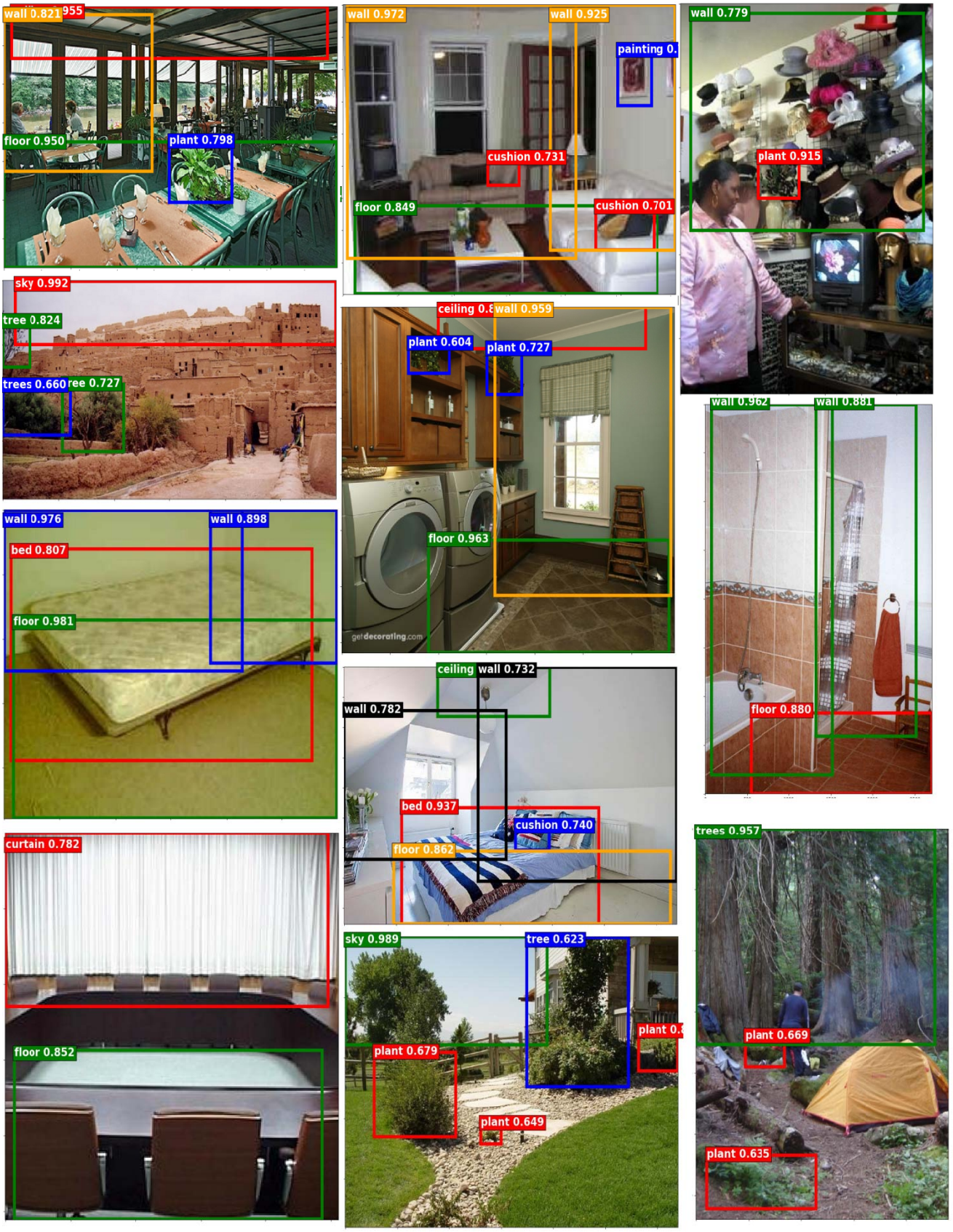
روش / زمان (ثانیه)	Fast RCNN [۱۴]	Faster RCNN [۱۹]	SPPNet [۱۳]	R-FCN [۲۶]	Proposed Method
Time	۰/۴۹	۰/۴۳	۰/۳۸	./۱۷	**۰/۱۳**

۵ نتیجه‌گیری

در این مقاله به بررسی شبکه‌های عصبی عمیق به منظور آشکارسازی و برچسب‌زنی اشیاء موجود در تصاویر پرداخته شده است. بهترین شبکه‌ی عصبی عمیق معرفی شده در سال‌های اخیر، شبکه R-FCN است. در همین راستا، در این مقاله روشی برای بهبود R-FCN ارائه شده است. یک تابع زیان جدید برای این شبکه معرفی شده و برای برچسب‌زنی نواحی استخراج شده، به ازای هر RoI یک شبکه ترکیبی SVMفازی دو کلاسه و SVR برای پیش‌بینی میزان تعلق به یک شی معرفی شده است. تابع زیان اختلاف کوشی-شوارتز از نظر سرعت و دقت عملکرد بهتری از تابع زیان آنتروپی متقابل استفاده شده در R-FCN داشته است. نتایج بدست آمده بر روی مجموعه داده SUN با شبکه بهبود یافته عملکرد بهتری از لحاظ سرعت و دقت از خود نشان داده است.



شکل ۱۵ نمای کلی از نتایج بدست آمده روش پیشنهادی در مجموعه داده SUN



دنباله تصاویر نتایج بدست آمده از روش پیشنهادی شکل ۱۵

مراجع

[1] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 524-531.

[2] E. B. Sudderth, A. Torralba, W. T .Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, pp. 1331-1338.

[3] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 966-973.

[4] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, pp. 2175-2184, 2015.

[5] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "A deep and autoregressive approach for topic modeling of multimodal data," IEEE transactions on pattern analysis and machine intelligence, vol. 38, pp. 1056-1069, 2016.

[6] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos, "Scene classification with semantic fisher vectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2974-2983.

[7] X. Li and Y. Guo, "Multi-level adaptive active learning for scene classification," in European Conference on Computer Vision, 2014, pp. 234-249.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," arXiv preprint arXiv:1412.6856, 2014.

[9] L.-J. Li, H. Su, Y. Lim, and F.-F. Li, "Objects as Attributes for Scene Classification," in ECCV Workshops (1), 2010, pp. 57-69.

[10] X. Qi, C.-G. Li, G. Zhao, X. Hong, and M. Pietikäinen, "Dynamic texture and scene classification by transferring deep image features," Neurocomputing, vol. 171, pp. 1230-1241, 2016.

[11] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorization," in Pattern recognition applications and methods, ed: Springer, 2015, pp. 209-224.

[12] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification ",in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1287-1295.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in European Conference on Computer Vision, 2014, pp. 346-361.

[14] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

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

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

[17] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International journal of computer vision, vol. 104, pp. 154-171, 2013.

[18] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in European Conference on Computer Vision, 2014, pp. 391-405.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91-99.

[20] K. Lenc and A. Vedaldi, "R-cnn minus r," arXiv preprint arXiv:1506.06981, 2015.

[21] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bourdev, "Pronet: Learning to propose object-specific boxes for cascaded neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3485-3493.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, et al., "Ssd: Single shot multibox detector," in European conference on computer vision, 2016, pp. 21-37.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. ۷۸۸-۷۹۷

[24] M. Najibi, M. Rastegari, and L. S. Davis, "G-cnn: an iterative grid based object detector," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2369-2377.

[25] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional Single Shot Detector," arXiv preprint arXiv:1701.06659, 2017.

[26] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional

IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 693-699, 1985.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo ",in Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, 2010, pp. 3485-3492.

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

[42] Z. Li, Z. Shi, X. Liu, and Z. Shi, "Modeling continuous visual features for semantic image annotation and retrieval," Pattern Recognition Letters, vol. 32, pp. 516-523, 2011.

[43] L. Fei-Fei, "ImageNet: crowdsourcing, benchmarking & other cool things," in CMU VASC Seminar, 2010, pp. 18-25.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014, pp. 818-833

علی قنبری سرخی مدرک کارشناسی خود را در رشته کامپیوتر گرایش نرم‌افزار از دانشگاه علم و صنعت ایران در سال ۱۳۸۹ دریافت کرد. سپس در سال ۱۳۹۱ موفق به اخذ مدرک کارشناسی ارشد در گرایش هوش مصنوعی از دانشگاه صنعتی شاهرود گردید. در حال حاضر نیز، دانشجو مقطع دکتری در دانشگاه صنعتی شاهرود در گرایش هوش مصنوعی است. موضوع پایان‌نامه دکتری ایشان، "دسته‌بندی محتوایی تصویر بر پایه مدل‌های موضوعی" می‌باشد. علایق پژوهشی او پردازش تصویر و یادگیری عمیق است.

حمید حسن‌پور استاد تمام دانشکده مهندسی کامپیوتر دانشگاه شاهرود می‌باشند. ایشان در سال ۱۳۷۲ مدرك کارشناسی مهندسی کامپیوتر خود را از دانشگاه علم و صنعت و در سال ۱۳۷۵ مدرك کارشناسی ارشد خود را در گرایش هوش ماشین از دانشگاه صنعتی امیرکبیر دریافت نمود. در سال ۱۳۸۳ موفق به اخذ مدرک دکتری خود از دانشگاه صنعتی کوئینزلند استرالیا در گرایش پردازش سیگنال شد. از سال ۱۳۸۴ الی ۱۳۸۶ نامبرده به عنوان عضو هیئت علمی در دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی بابل فعالیت داشت؛ سپس به دانشکده مهندسی کامپیوتر دانشگاه شاهرود انتقال یافت. زمینه‌های علمی مورد علاقه ایشان پردازش سیگنال، پردازش تصویر، داده‌کاوی، و پردازش متن می‌باشد.

networks," in Advances in neural information processing systems, 2016, pp. 379-387.

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

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning, 2015, pp. 448-456.

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

[30] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761-769.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition ,۲۰۱۵ ,pp. 1-9.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818-2826.

[33] K. Janocha and W. M. Czarnecki, "On Loss Functions for Deep Neural Networks in Classification," arXiv preprint arXiv:1702.05659, 2017.

[34] W. M. Czarnecki, R. Jozefowicz, and J. Tabor, "Maximum entropy linear manifold for learning discriminative low-dimensional representation," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2015, pp. 52-67.

[35] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, vol. 2, pp. 121-167, 1998.

[36] D. Isa, L. H. Lee, V. Kallimani, and R. Rajkumar, "Text document preprocessing with the Bayes formula for classification using the support vector machine," IEEE Transactions on Knowledge and Data engineering, vol. 20, pp. 1264-1272, 2008.

[37] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu, "A support vector machine-based context-ranking model for question answering," Information Sciences, vol. 224, pp. 77-87, 2013.

[38] S.-G. Chen and X.-J. Wu, "A new fuzzy twin support vector machine for pattern classification," International Journal of Machine Learning and Cybernetics, pp. 1-1 ,۲ .۲۰۱۷

[39] J. M. Keller and D. J. Hunt, "Incorporating fuzzy membership functions into the perceptron algorithm,"

منصور فاتح مدرک کارشناسی خود را در رشته مهندسی برق از دانشگاه صنعتی شاهرود در سال۱۳۸۶دریافت کرد. سپس کارشناسی ارشد و دکتری خود را در رشته‌های مهندسی پزشکی و الکترونیک دیجیتال درسالهای ۱۳۸۸ و ۱۳۹۳ از دانشگاه تربیت مدرس دریافت کرد. پروژه کارشناسی ارشد خود را با عنوان" بررسی نقش و اثر نور پلاریزه در درماتوسکپی از بدن با استفاده از شبیه سازی" و پروژه دکتری خود را با عنوان" خواندن خودکار نقشه‌های دستی فرش" به انجام رسانید. از سال ۱۳۹۴ ایشان عضو هیئت علمی دانشگاه صنعتی شاهرود بوده و زمینه تحقیقاتی ایشان پردازش تصویر و یادگیری تقویتی میباشد.