شناسایی و تشخیص چهره در زمان واقعی با استفاده از شبکه عصبی YOLO و یادگیری عمیق

،مریم فردی۲*کوروش داداش تبار احمدی۱

۱- هیات علمی و مدیر گروه هوش مصنوعی، مجتمع دانشگاهی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر

۲- کارشناسی ارشد هوش مصنوعی دانشگاه صنعتی مالک اشتر

چکیده

شناسایی و موقعیت‌یابی اشیا ازجمله زمینه‌های تحقیقاتی قدیمی و مهم در حوزه کامپیوتر است. شناسایی و موقعیت‌یابی اشیا در تصویر Object Detection نامیده می‌شود. یکی از مهمترین وظایف تشخیص اشیا، تشخیص چهره میباشد. به طور کلی اولین گام در تشخیص چهره، مرحله تشخیص الگو و احراز هویت افراد بوده است که در سال های اخیر، بسیار حائز اهمیت شده است، که بر پایه الگوریتم های مبتنی بر یادگیری عمیق در تشخیص اشیا استفاده میگردد.

الگوریتم ها را می توان به طور کلی به دو دسته تقسیم کرد، تشخیص دو مرحله ای مانند Faster R-CNN و تشخیص یک مرحله ای مانند YOLO، اگرچه YOLO از نظر دقت به اندازه آشکارسازهای دو مرحله ای نیستند، اما در تشخیص، از عملکرد بهتری برخوردار میباشد.YOLO هنگام برخورد با اشیا با اندازه طبیعی عملکرد خوبی دارد، اما در تشخیص اشیای کوچک از توانایی کمتری برخوردار میباشد. این دقت در هنگام برخورد با اشیایی همانند صورت، که در مقیاس کوچک مورد بررسی قرار میگیرد، کاهش می یابد. برای رفع این مشکل(تشخیص مقیاس های مختلف صورت)، یک آشکارساز چهره به نام YOLO-face را بر اساس YOLOv3 پیشنهاد نموده تا بتوان عملکرد تشخیص چهره را بهبود بخشید. در این روش، برای تشخیص چهره، استفاده ازanchor box مناسب تر برای صورت و همچنین regression loss function مناسبتر را انتخاب نموده ایم. این آشکارساز به طور قابل توجهی سرعت تشخیص و دقت را افزایش داده است.

این آزمایشات روی مجموعه داده های WIDER FACE و FDDB نشان می دهد که الگوریتم بهبود یافته ما بر روی YOLOو انواع آن عملکرد بهتری دارد.

کلمات کلیدی: Deep learning, Face Detection ,Anchor box, Loss function , Bounding Box

* **Corresponding author:** Kurosh Dadashtabar Ahmadi
Department of Computer Science, Malek-Ashtar University of Technology, Tehran, Iran
Email: dadashtabar@mut.ac.ir

۱. مقدمه

تشخیص چهره یکی از پر کاربردترین برنامه در بینایی کامپیوتر میباشد. از احراز هویت و شناسایی چهره در بسیاری از سناریوها می توان استفاده نمود. به طور کلی، اولین مرحله در تشخیص چهره، شناسایی و مکان یابی تصاویر یا فیلمها در چهره میباشد. بطورکلی، یک الگوریتم تشخیص چهره دقیق، می تواند به طور چشمگیری عملکرد سیستم تشخیص را بالا ببرد. به دلیل تکثیر دستگاه های تلفن همراه و دوربین های هوشمند، جمع آوری تصاویر و فیلم ها روز به روز بیشتر و راحت تر انجام می شود. با این حال، توانایی محاسباتی چنین دستگاه هایی نسبتاً محدود است. بهترین راه حل برای این مسئله، یافتن الگوریتم های سریعتر و کارآمدتر میباشد.

R-CNN اولین الگوریتم تشخیص شی، که مبتنی بر یادگیری عمیق میباشد معرفی شده است. در مقایسه با الگوریتم های سنتی مانند Adaboost ، DPM و غیره عملکرد بسیار بهتری پیدا نموده است.

الگوریتم های بعدی مانند SPP-Net، Fast R-CNN، Faster R-CNN و R-FCN بر اساس R-CNN بر روی سرعت یا دقت متمرکز بوده است. با این حال، سرعت پایین این آشکارسازها همیشه مطرح بوده که، جهت سرعت بخشیدن به روش تشخیص، جهت حل مشکل سرعت در زمان واقعی YOLOو SSD آشکارسازهای یک مرحله ای، برای کارهای تشخیص مورد استفاده قرار گرفته اند.[1]

به طور قابل توجهی، مشکل عمده در تشخیص چهره، از نظر دقت تشخیص برای مقیاس های مختلف صورت در یک تصویر، برای یک ردیاب[1] یکسان است. اخیراً، برخی از رویکردهای تشخیص چهره سعی در مقابله با مقیاس های مختلف با استفاده از چندین ساختار شبکه برای حل این مشکل ارائه نموده است. روش دیگر استفاده از سطوح مختلف ویژگی ها است که از چندین لایه آخر شبکه گرفته شده است. در عمل، میبایست بتوانیم انواع مختلف مقیاس صورت را در سناریوهای طبیعی تشخیص دهیم. [1]

بنابراین، ویژگی های مورد استفاده در تشخیص چهره برای تصویر با اندازه ۲۰۰×۲۰۰ پیکسل، با تشخیص چهره در مقیاس ۱۰×۱۰ پیکسل تفاوت زیادی دارد. YOLOv3 از ساختار شبکه ای مشابه FPN برای ادغام ویژگی های سطوح مختلف استفاده می نماید.

دلیل اصلی استفاده، بدین خاطر میباشدکه، الگوریتم می تواند اشیا چند مقیاسه را تشخیص دهد. YOLOv3 از مجموعه داده COCO نتایج پیشرفته ای را به دست آورده است. اما، برای شناسایی چهره، عملکرد آن مطابق انتظار نبوده است. از طرفی، اندازه anchor box در YOLOv3 در مجموعه COCO لزوماً برای تشخیص چهره مناسب نیستند. از طرف دیگر، تشخیص چهره فقط به شناسایی و مکان یابی چهره ها نیاز دارد و نیازی به طبقه بندی هشتاد نوع شی مانند مجموعه داده COCO نمیباشد.

برای حل این مشکل، یک روش مبتنی بر YOLOv3 را برای تشخیص چهره پیشنهاد نموده، که ایده اصلی آن، بر استراتژی انتخاب یک سری anchor box مناسب تر و استفاده از یک تابع خطا[2] جدید تمرکز دارد.

با آموزش روی مجموعه داده های آموزشی WIDER FACE، در کنار چهره یاب پیشنهادی YOLO-based براساس YOLOv3 در مقایسه با YOLOv3عملکرد بسیار بهتری را ارائه نموده است.

بر اساس سه مرحله موجود در مجموعه WIDER FACE ، داده های اعتبارسنجی[3]، YOLO-face به ترتیب ۲۱٪ ، ۱۸٪ و دقت[4] ۱۸٪ نسبت به YOLOv3 دارد، در حالی که سرعت تشخیص سریع به اندازه YOLOv3 را حفظ مینماید.[2]

1. detector
2. loss function
3. validation dataset
4. accuracy

مشارکت های اصلی به شرح زیر است:

۱. یک ساختار شبکه اصلی ستون فقرات به نام darknet deeper (عمیق تر) ارائه نموده،که عملکرد بهتری در darknet-53 دارد، به ویژه در تشخیص چهره های کوچک.[1]

۲. یک تابع کاهش رگرسیون جدید که MSE و GIoU را با هم ترکیب مینماید، پیشنهاد شده است.[1]

۳. anchor box های که برای تشخیص چهره مناسب ترند، توسط خوشه بندی k-means آموخته می شوند.[1]

تمامی روش های تشخیص اشیا، مبتنی بر یادگیری ماشین و یادگیری عمیق میباشند. مانند Scale-invariant feature transform (SIFT) و Histogram of oriented gradients (HOG) ، روش‌های نواحی دربرگیرنده اشیاء در تصویر[5]، نظیر R-CNN،Fast R-CNN و Faster R-CNN وهمچنین روش Single Shot MultiBox Detector یا SSD و YOLO را میتوان نام برد.

۲. شرح مساله

به طور کلی عملکرد یک الگوریتم تشخیص، بدین صورت میباشد که یک منطقه مورد نظر (ROI) ،در یک تصویر مشخص به عنوان یک منطقه نامزد انتخاب گردیده، ویژگی های منطقه کاندید شده، مانند یک ویژگی باینری محلی (LBP) یا یک هیستوگرام شیب جهت دار (HOG) را استخراج نموده، و از طریق آموزش (training classifier)، آن منطقه را طبقه بندی می نماید.[۲] الگوریتم پیشنهادی جهت تشخیص چهره، استفاده از الگوریتم YOLO-face میباشد .مدل یکپارچه YOLO مزایای زیادی نسبت به روش‌های سنتی تشخیص اشیا دارد که به شرح آن میپردازیم.

دو نوع تشخیص شی بر اساس یادگیری عمیق وجود دارد: سری R-CNN مبتنی بر تشخیص منطقه ای، Fast R-CNN و Faster R-CNN و دیگری رگرسیون SSD و YOLO . تفاوت بین شبکه YOLO و RCNN را میتوان بدین صورت مطرح نمود که، آموزش و شناسایی و استخراج ویژگی و طبقه بندی رگرسیون YOLO ، همه در یک شبکه واحد انجام می شود. YOLO یک شبکه end-to-end جداگانه است. YOLO تشخیص شی را به عنوان یک مسئله رگرسیون در نظر می گیرد هنگامی که یک تصویر به شبکه وارد شود، موقعیت همه اشیا موجود در تصویر، دسته های آنها و نمره اطمینان از خطا[6] مربوطه را میتواند بدست آورد. در RCNN تشخیص در دو مرحله صورت میگیرد، دسته شی[7] ، مکان شی و جعبه[8] میباشد.

YOLO برای پیش‌بینی تشخیص، به صورت کلی[9] به تصویر نگاه می‌کند. برخلاف تکنیک‌های پنجره‌های لغزان (اسلاید) و پروپوزال، YOLO به کل تصویر نگاه می‌کند.YOLO تعمیم‌پذیری بالایی دارد. زمانی‌که تصاویر به شبکه آموزش داده می‌شوند و سپس شبکه آموزش‌دیده روی کارهای هنری تست می‌شود (در واقع منظورمان همان تغییر حوزه داده‌های ورودی است) شبکه YOLO با فاصله زیادی بهتر از شبکه‌هایی مانند DPM و R-CNN کار می‌کند.

۲-۱ تشخیص چهره

تشخیص چهره بخشی از تشخیص اشیا است و طیف وسیعی از الگوریتم های تشخیص چهره از الگوریتم های تشخیص اشیا بهبود می یابند.

5. Region Proposals
6.Confidence probabilities
7. classification
8.bounding box
9.Global

قبل از فراگیری مقدمه یادگیری عمیق در این زمینه، بیشتر الگوریتم های تشخیص شی از ویژگی های دستی[10] برای شناسایی استفاده میکنند.

محققان به دلیل عدم توانایی کافی در نمایش ویژگی ها مجبور به طراحی الگوریتم های متنوع تشخیص شی هستند. علاوه بر این، در بیشتر موارد نمودارهای پیچیده برای تسریع الگوریتم ها مورد نیاز است.

قدرت عملکرد آشکارسازها تا حد زیادی به کارآیی محاسباتی و توانایی بیان ویژگی ها بستگی دارد. آشکارسازهای مشهور چهره بر اساس این ویژگی های دستی، مانند Viola–Jones، هیستوگرام شیب های جهت دار(HOG) و مدل تغییر شکل پذیر(DPM)، از الگوریتم های معمول میباشند.[1]

اولین آشکارساز معمول چهره در یادگیری عمیق، Cascade CNN است. از سه شبکه عصبی کانولوشن آبشاری برای تشخیص چهره استفاده میکند. جعبه های همپوشانی با حذف غیر حداکثرها (NMS) [11]حذف میگردند. MTCNN از یک ساختار Cascade مشابه استفاده می کند.

اما همچنین پنج نقطه عطف (چشم ، گوش و بینی و گوشه های دهان) را برای بازگشت به عقب پیش بینی می نماید. DenseBox شبکه کامل تکاملی (FCN) را در تشخیص چهره معرفی می کند. Faceness-Net یک رویکرد دو مرحله ای را برای تشخیص چهره پیشنهاد می کند.[3]

در مرحله اول شبکه های آگاه از ویژگی (attribute-aware) برای تولید نقشه پاسخ[12]، از قسمت های مختلف صورت استفاده می نماید، در مرحله دوم پنجره کاندیدای تولید شده را توسط یک شبکه عصبی چند منظوره کانولوشن (CNN) اصلاح مینماید. اگرچه الگوریتم های تشخیص چهره مبتنی بر یادگیری عمیق عملکرد به مراتب بهتری نسبت به روش های مرسوم به دست آورده اند، اما با مقیاس های کوچک و چهره هایی پوشیده، این دقت به طرز چشمگیری کاهش می یابد.

برای حل این مشکلات، روشهای مختلفی پیشنهاد شده است.

Face R-CNN بر اساس Faster R-CNN ساخته شده است.R-CNN از روش(OHEM) و آموزش multi-scale training، برای بهینه سازی مدل استفاده مینماید. یک روش چند شاخه، که از لایه های مختلف در شبکه VGG برای شناسایی چهره های multi-scale استفاده می کند، را پیشنهاد کرده است. FDNet، برای بهبود عملکرد تشخیص چهره (Light-Head Faster R-CNN) پیشنهاد می کند و به طور همزمان از آموزش و آزمایش چند مقیاس و شبکه عصبی کانولوشن قابل تغییر استفاده می کند.[۱] Face R-FCN براساس R-FCN ساخته شده است. این روش از هسته کوچکتر[13] حساس به موقعیت RoI و لنگرهای اضافی استفاده می کند.

PyramidBox از ویژگی های سطح پایین شبکه هرمی، PyramidAnchors و ماژول context-sensitive را برای حل مشکل تشخیص چهره پوشیده پیشنهاد نمود. علاوه بر این، یک روش نمونه گیری لنگر داده [15]برای افزایش نمونه های آموزش در مقیاس های مختلف ارائه نموده است.

10.handcraft
11.non-maximum suppression
12.response maps
13. pooling kernel
14. smaller anchors
15. dataanchor-samplin

YOLO ۲-۲

الگوریتم های تشخیص اشیا مبتنی بر یادگیری عمیق ابتدا توسط R-CNN معرفی شدند. این الگوریتم در مقایسه با الگوریتم DPM، که بهترین ردیاب قبل از آن است، بیش از ۵۰٪ قدرت عملکرد تشخیص را افزایش داده است. در راستای حل مشکل اندازه تصویر ورودی، و سرعت بخشیدن به عملکرد شناسایی و تشخیص ، SPP-Net، هرم فضایی (SPP) را ارائه نمود.

با توجه به این ساختار، سرعت تشخیص به طور قابل توجهی بیشتر از R-CNN شده است. همچنین Fast R-CNN تجمع ROI را ارائه می دهد که در آموزش و تشخیص سریعتر از R-CNN است. Fast R-CNN با ارائه ROI pooling نسبت به R-CNN، در بخش آموزش و تشخیص، عملکرد سریعتری داشته است. علاوه بر این، استفاده از روش softmax به جای SVM به عنوان طبقه بندی کننده ارائه گردیده است.Faster R-CNN شبکه region proposal (RPN)را ارائه داد.

FCN مفاهیم نقشه حساس به موقعیت[16] را معرفی نموده است، که از لایه های شبکه مشترک عمیق تری استفاده می کند و سرعت شناسایی را به طرز چشمگیری تسریع میبخشد. YOLO اولین آشکارساز یک مرحله ای است که براساس CNN ساخته شده است.YOLO از یک شبکه عصبی واحد به طور مستقیم از تصاویر ورودی، برای پیش بینی box bounding و احتمالات کلاس در یک ارزیابی استفاده می کند. YOLO تصویر ورودی را به سلولهای شبکه تقسیم میکند و سپس مستقیماً مختصات و تقسیم بندی Grid ها را برای هر سلول پیش بینی مینماید. اگرچه سرعت آن، چندین برابر بیشتر از آشکارسازهای دو مرحله ای است، اما دقت تشخیص نسبتاً کمتر از نمونه های مشابه است.

YOLOv2پیشرفت بسیاری، از جمله استفاده از معماری شبکه های عمیق تر، anchor box خودکار آموخته شده، بهبودloss function، آموزش چند مقیاس[17] , افزایش داده [18]و غیره را انجام داده است. نسخه های بهبود یافته YOLO در PASCAL VOC عملکرد خوبی با سرعت بالا نشان داده است، که باعث می شود این روش در عمل، نیاز به تشخیص در زمان واقعی را برآورده کند.[۱]YOLOv3 در ستون فقرات[19] شبکه جدیدی به نام darknet-53 را اعمال نموده و نتایج جالبی را در مجموعه داده COCO بدست آورد. در این روش، از معماری YOLOv3 به عنوان ساختار شبکه اصلی استفاده نموده و از چندین جنبه پیشرفت نموده است. این روش را بر روی مجموعه داده WIDER FACE و مجموعه داده FDDB ارزیابی نموده و عملکرد نسبتاً بهتری را بدست آورده است. از روش پیشنهادی می توان در کارهای زمان واقعی تشخیص چهره در مقیاس های مختلف استفاده نمود.

۳. روش(Method)

در این راستا با هدف حل مشکل تشخیص مقیاس های مختلف صورت، روشی را به نام YOLO-face پیشنهاد مینماییم. معماری YOLO-face بر اساس YOLOv3 ساخته شده است. مدل پیشنهادی را با بهبود ستون فقرات (backbone)، جعبه های لنگر(anchor box) و عملکرد (loss function) بهبود بخشیدیم تا آن را برای تشخیص چهره در مقیاس چند منظوره مناسب تر نماییم.

16. sensitive score maps
17. multi-scale training
18. data augmentation
19. backbone

۳-۱ ستون فقرات

در این بخش از darknet-53 به عنوان ستون فقرات شبکه استفاده می کنیم. این معماری از یک شبکه استخراج ویژگی و سه شبکه شناسایی تشکیل شده است. شبکه استخراج ویژگی بر اساس darknet-53 میباشد. Darknet-53 ترکیبی از darknet-19 و ResNet است. دارای لایه های کانولوشن ۳×۳ و ۱×۱ پی در پی و برخی اتصالات میانبر. Darknet-53 شامل ۵۳، لایه کانولوشن و به طور قابل توجهی بزرگتر از darknet 19 است. این شبکه بسیار قدرتمندتر از darknet-19 و کارآمدتر از ResNet-101 یا ResNet-152 است. عملکردی مشابه ResNet-152 دارد اما دو برابر سریعتر است.[1]
برای دستیابی به یکپارچه سازی در مقیاس چندگانه، ویژگی های سطح پایین با ویژگی های سطح بالا مانند شبکه های هرم ویژگی (FPN) ادغام می شوند. این طراحی میتواند استفاده بهتری از مقیاس های اطلاعات تصویر داشته، و بدین ترتیب عملکرد بهتری را برای تشخیص چند مقیاس[20] از یک تصویر بدست آورد. در تشخیص اشیا در مقیاس کوچک، عملکرد YOLOv3 چالش برانگیز میباشد. بعبارتی، ساختار شبکه نقشی اساسی در استخراج ویژگی و مکان یابی اشیا دارد.
از آنجا که ویژگی ها در مقیاس کوچک[21] پس از چندین بار کاهش ابعادی حتی در نقاط مختلف روی feature map بسیار کوچک می شوند، لایه های بعدی نمی توانند اطلاعات کافی را بدست آورند، که این امر بر کارایی استخراج ویژگی و دقت تشخیص تأثیر می گذارد.
بنابراین، استخراج ویژگی های کافی، زودتر از ویژگی های مقیاس کوچک بر روی feature map به دست می آید که این امر تشخیص دقیقتر small object را تسهیل می بخشد. با در نظر گرفتن این فاکتورها، با افزایش تعداد لایه های شبکه در، ساختار شبکه darknet-53 اصلی، آن را بهبود بخشیدیم تا ویژگی های چهره، در مقیاس کوچک را بدست آوریم. تجربه نشان می دهد که ستون فقرات بهبود یافته عملکرد قابل توجهی را برای تشخیص چهره داشته است.

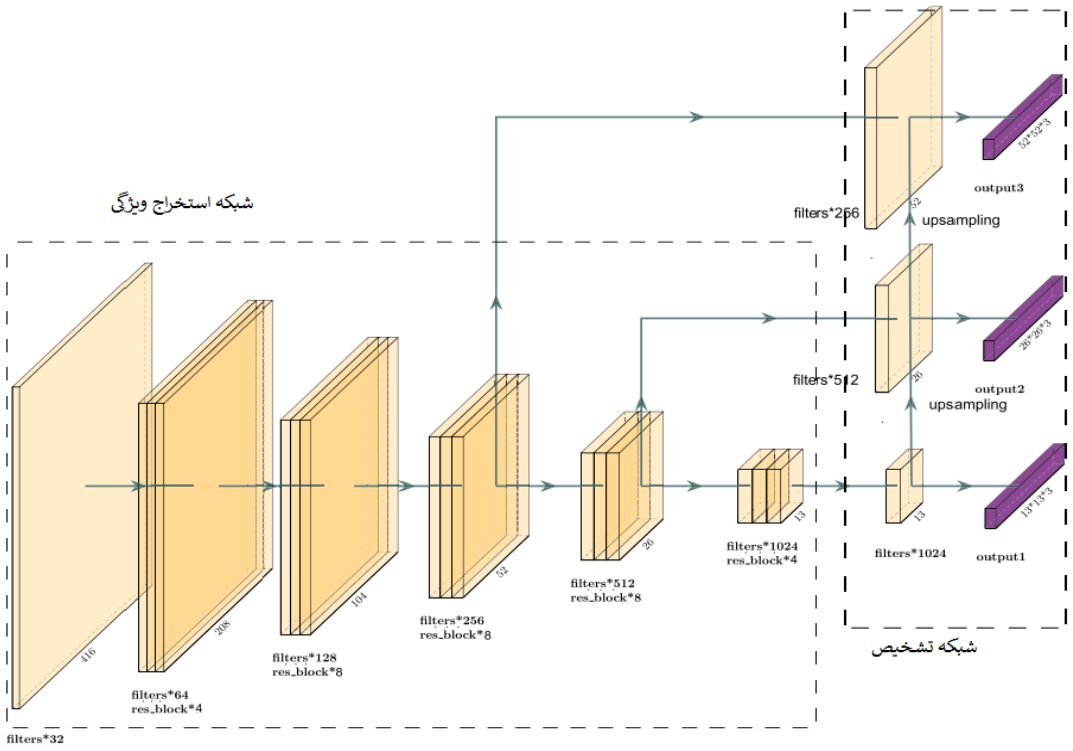
۳-۲ anchor boxes مناسب برای تشخیص چهره

به طور کلی مقیاس و نسبت anchor box از پارامترهای بسیار مهم در تشخیص شی میباشند. بدیهی است که شکل anchor box باید ارتباط زیادی با اهداف شناسایی شده داشته باشد. برای تشخیص عمومی اشیا، شکل anchor box ها، تا حد امکان میبایست انواع مختلفی را در بر داشته باشد. به طور شهودی، برای بیشتر چهره ها در یک تصویر، ارتفاع صورت همیشه بیشتر از عرض صورت میباشد، بنابراین شکل anchor box برای تشخیص چهره نباید با شکل کلی تشخیص اشیا یکسان در نظر گرفته شود.[4]
برای انتخاب جعبه های لنگر مناسب برای تشخیص صورت، دو نوع جعبه لنگر را در نظر میگیریم و با هم جمع می کنیم. یک نوع anchor box از YOLOv3 اصلی گرفته شده است، ولی از جعبه های مسطح[22] به جعبه های باریک تبدیل میشود(ارتفاع جعبه ها از عرض کمتر است) و بعدی anchor box باریک و بلند میباشند. به دنبال YOLOv2 و YOLOv3 ، نوع دیگری از anchor ها از اجرای k-means clustering در مجموعه داده های آموزش WIDER FACE برای بدست آوردن ابعاد جعبه های مرزی استفاده می شود.

20.multi-scale
21. small-scale
22. flat

فرآیند به شرح زیر است:

مرحله اول تعیین تعداد دانه ها، k برای anchor box خوشه ای و سپس انتخاب k anchor box به طور تصادفی به عنوان مراکز خوشه بندی اولیه، سپس محاسبه IoU جعبه های لنگر و سایر anchor box انجام میشود. با استفاده از IoU به عنوان فاصله جعبه های لنگر، تمام برچسب های چهره به کلاس های k تقسیم میشوند. سپس، mean values اندازه جعبه لنگر کلاس k را به عنوان مراکز خوشه های جدید در نظر میگیرد. این روند را تا زمان همگرایی تکرار مینماید. در این آزمایش، مراکز خوشه ای اولیه k را روی ۹ تنظیم نموده. جعبه های لنگر افقی به صورت عمودی جابجا شدند تا متناسب با تشخیص چهره باشد. ۹ شکل نهایی anchor عبارتند از: (۳ ، ۳) ؛ (۴ ، ۵) ؛ (۶ ، ۸) ؛ (۳۰ ، ۶۱) ؛ (۴۵،۶۲) ؛ (۵۹ ، ۱۱۹) ؛ (۹۰ ، ۱۱۶) ؛ (۱۵۶ ، ۱۹۸) و (۳۲۶ ، ۳۷۳).



(شکل۱: معماری پیشنهادی شبکه استخراج ویژگی دارای ۷۱ لایه کانولوشن است و مقیاس feature mapsرا توسط لایه کانولوشن با گام stride 2 می دهد. شبکه تشخیص ساختار مشابه FPN برای استخراج ویژگی ها از مقیاس های مختلفfeature maps)

Loss function ۳-۳

عملکرد مدل، YOLO در مرحله آموزش از چهار قسمت تشکیل شده است: regression loss, confidence loss, classification loss , loss responsible، نسبت های تعیین شده به صورت ۱: ۱: ۱: ۱ است. این توزیع وزن برای تشخیص شی multi-class طراحی شده است.

با این حال، تشخیص چهره مشکلی در binary classification است. [1]

برای اینکه عملکرد کلی تشخیص چهره مناسب تر باشد آن را به صورت وزنهای ۵,۰:۵,۰:۱:۲ اصلاح مینمایم.

(1)

$$L= 2 .\ \Sigma L_{reg} + \Sigma L_{objconf} + 0.5\ \Sigma L_{noobjconf} + 0.5 .\ \Sigma L_{cls}$$

L_{reg} : Regression loss است.

$L_{objconf}$: محاسبه Confidence نمره اطمینان از خطا هنگامی که یک شی در anchor box شناسایی می شود.

$L_{noobjconf}$: محاسبه Confidence نمره اطمینان از خطا هنگامی که یک شی در anchor box شناسایی نشده است.

L_{cls} : Classification ، احتمال تخمین شی و کلاس واقعی درهر grid cell، میباشد.

در YOLO از تابع اتلاف MSE یا Mean Squared Error استفاده شده است، چون بهینه‌سازی این تابع اتلاف آسان است و با مساله رگرسیون که در YOLO مطرح شده سازگار است.در این بخش جهت، محاسبه MSE در راستای، پیش بینی anchor box هدف، در YOLO v2 استفاده شده است.

از حد آستانه IoU، برای فیلتر کردن کادرهایی به کار می‌رود که یک شیء واحد و یکسان در آن‌ها شناسایی شده است. با این حال بین بهینه سازی MSE و به حداکثر رساندن مقدار IoU اختلاف وجود دارد. به طور خاص، بهینه سازی در مورد جعبه های اتصال بدون همپوشانی غیرممکن است، برای رفع این ضعف، یک پیشنهاد کلی را برای IoU به عنوان یک معیار جدید، یعنی GIoU ارائه داده شد.که در این معیار، ارتباط زیادی بین بهینه سازی عملکرد MSE و خود وجود دارد.

با تلفیق خطای اصلی l_n-norm با کاهش وزن GIoU ، وکاهش رگرسیون آنرا بهبود داده ایم.

$$GIoU = IoU- \frac{A_c - U}{A_c} \qquad (2)$$

$$L_{GIoU} = 1- GIoU \qquad (3)$$

$$\begin{aligned} L_{reg} &= \Sigma_{c=x,y,w,h} \ \Sigma\ (|\Delta c_{pred} - \Delta c_{truth}| + \alpha\ .\ L_{GIoU})^2 \\ &= \Sigma\ (|\Delta x_{pred} - \Delta x_{truth}| + \alpha\ .\ L_{GIoU})^2 \\ &+ \Sigma\ (|\Delta y_{pred} - \Delta y_{truth}| + \alpha\ .\ L_{GIoU})^2 \\ &+ \Sigma\ (|\Delta w_{pred} - \Delta w_{truth}| + \alpha\ .\ L_{GIoU})^2 \\ &+ \Sigma\ (|\Delta h_{pred} - \Delta h_{truth}| + \alpha\ .\ L_{GIoU})^2 \end{aligned} \qquad (4)$$

در اینجا، AC مکان پیش بینی شده کوچکترین محصور[23] میباشد، α یک فاکتور real-valued است، و x، y، w و h به ترتیب مکان و اندازه anchor box هستند. در مدل پیشنهاد شده، ضریب α را روی ۰,۱ قرار می دهیم.

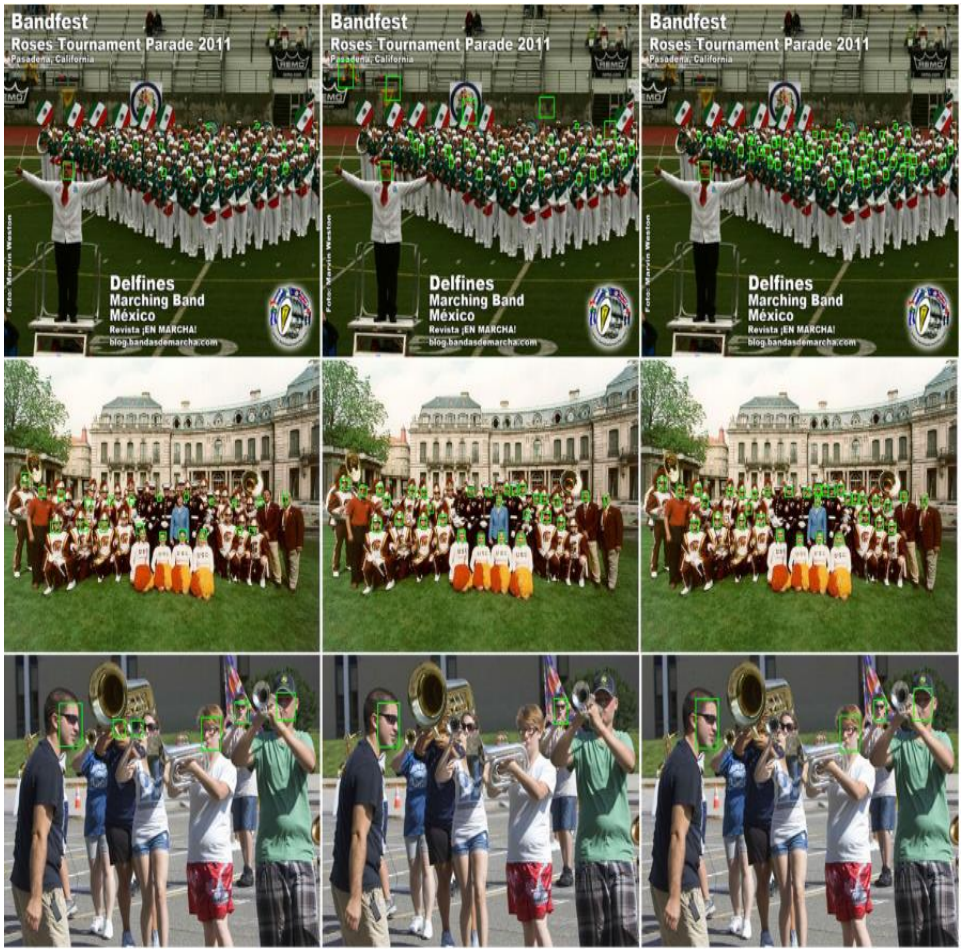
۳-۴ مجموعه داده

از مجموعه داده های WIDER FACE به عنوان آموزش استفاده نموده و داده ها را ارزیابی می کنیم. WIDER FACE یک مجموعه داده بسیار بزرگ برای تشخیص چهره است. داده ها پس ازجمع آوری، به صورت دستی تمیز شدند. این مجموعه داده شامل ۳۹۳۷۰۳ عدد جعبه های لنگر تصویر صورت[24] در ۳۲۲۰۳ تصویر است. تشخیص چهره در این مجموعه داده به دلیل تغییرات غنی در ژست، مقیاس، حالت چهره و شرایط روشنایی بسیار چالش برانگیز است.

23.smallest enclosing
24. face bounding box annotations

WIDER FACE با توجه به مشکلات تشخیص، برای ارزیابی بیشتر عملکرد ردیاب، داده ها را به سه گروه "آسان" ، "متوسط" و "سخت" تقسیم بندی مینماید. همچنین داده ها به سه زیر مجموعه تقسیم می شود، training (۴۰٪) ، validation (۱۰٪) و testing (۵۰٪). محبوب ترین و پر کاربردترین مجموعه داده WIDER FACE برای تشخیص صورت معرفی شده است.

یکی دیگر از معیارهای محبوب ارزیابی الگوریتم های تشخیص چهره FDDB [25] میباشد. این دیتا ست شامل ۲۸۴۵ تصویر و که در آن تعداد ۵۱۷۱ چهره وجود دارد. با آزمایش بر روی مجموعه داده های FDDB پیشرفت های چشمگیری حاصل گردید.



(شکل۲: نمونه های نتیجه تشخیص چهره سمت چپ: نتایج توسط YOLOv2 شناسایی شده است. وسط: نتایج توسط YOLOv3 شناسایی شده است. راست: نتایج توسط YOLO-face شناسایی شده است)

۳-۵ آموزش

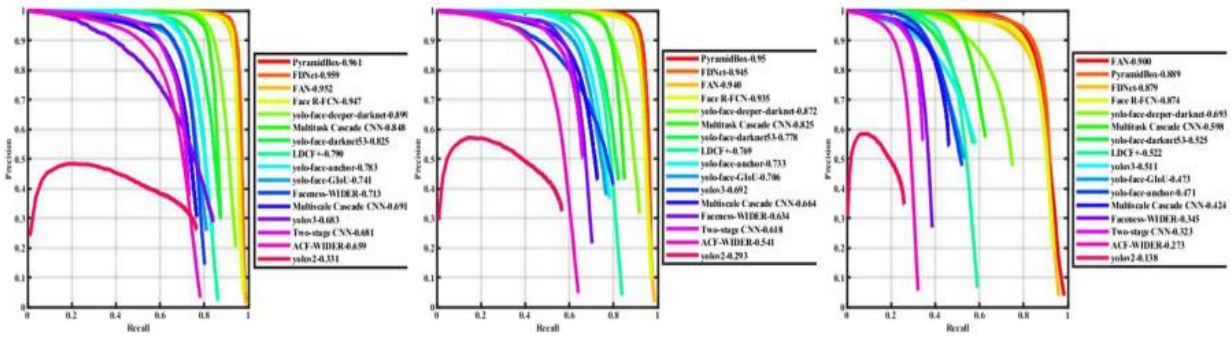
مدل پیشنهادی YOLO-face را با استفاده از darknet در GPU NVIDIA GeForce GTX 1080Ti آموزش داده ایم. اندازه تصویر ورودی روی ۴۱۶×۴۱۶ و اندازه batch size =۶۴ تنظیم گردید، و از بهینه سازSGD استفاده شده است.

25.Dataset and Benchmark Detection Face

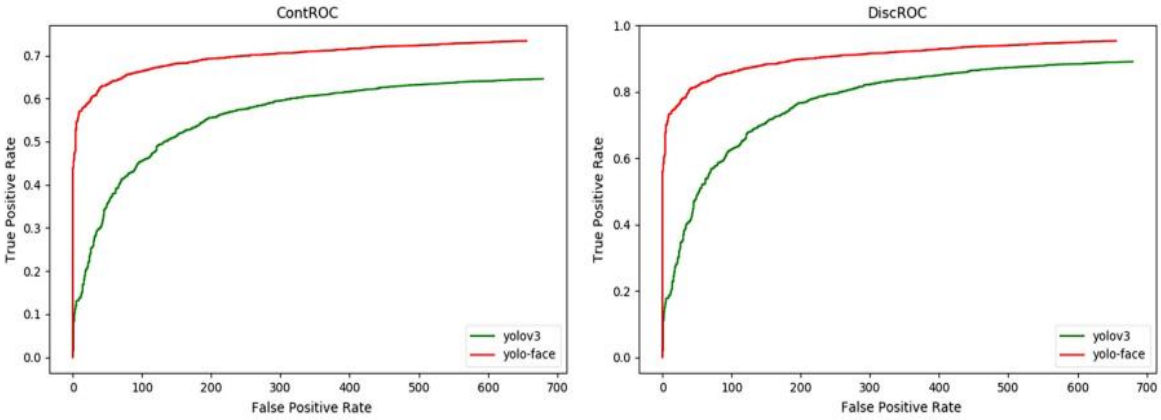
میزان یادگیری به ۰,۰۰۱ رسیده است. برای اینکه مدل قابل تعمیم باشد، از سه نوع افزایش داده[26] تغییر در روشنایی و رنگ و saturation ، استفاده نموده و مدل را آموزش داده ایم. پس از ۲۰۰۰۰ دورآموزش، بر روی مجموعه داده های اعتبارسنجی WIDER FACE و مجموعه داده های FDDB ارزیابی شد.[1]

۳-۶ نتیجه آزمایش

برای ارزیابی YOLO پیشنهادی، آزمایشات جامعی را روی مجموعه validation (اعتبارسنجی) WIDER FACE و مجموعه داده FDDB انجام گردید. شکل ۲ نمونه ای از نتایج شناسایی چهره را روی مجموعه داده نشان می دهد. به راحتی میتوانیم ببینیم که آشکارساز صورت YOLO بهبود یافته از YOLOv2 وYOLOv3 بهتر عمل نموده است. آشکارسازهای YOLOv2 وYOLOv3 کمتر قادر به تشخیص چهره های در مقیاس کوچک و چهره پوشیده با چادر هستند. YOLO-face در مواجهه با این مسائل از عملکرد بهتری برخوردار است. این بدان معناست که نسبت لنگر و مقیاس های یادگیری برای تشخیص چهره، منطقی میباشد، به ویژه برای تشخیص چهره های پوشیده شده و صورت های کوچک عملکرد بهتری داشته است. علاوه بر این، YOLO-face نسبت به ردیاب های اصلی، چهره های بسیار بیشتری را شناسایی کرده و جعبه های اشتباه[27] را نیز کمتر شناسایی مینماید(شکل ۲ را ببینید). دلیل آن انطباق و سازگاری GIoU و loss function بهبود یافته میباشد.



(شکل۳: نتایج تشخیص در مجموعه داده WIDER FACE)



(شکل۴: نتایج تشخیص در مجموعه داده FDDB. سمت چپ: ContROC. راست: DiscROC)

26.data augmentation
27.wrong boxes

(جدول۱: تشخیص YOLOv2، YOLOv3، YOLO چهره در مجموعه WIDER FACE)

	Easy	Medium	Hard
LDCF	0.790	0.769	0.522
Faceness-WIDER	0.713	0.634	0.456
Multi-scale cascade CNN	0.691	0.664	0.424
Multitask cascade CNN	0.848	0.825	0.598
Two-stage CNN	0.681	0.618	0.323
ACF-WIDER	0.659	0.541	0.273
PyramidBox	0.961	0.95	0.889
FDNet	0.959	0.945	0.879
FAN	0.952	0.940	0.900
Face R-FCN	0.947	0.935	0.874
YOLOv2	0.331	0.293	0.138
YOLOv3	0.683	0.692	0.511
YOLO-face-anchor	**0.783**	**0.733**	**0.471**
YOLO-face-GIoU	**0.741**	**0.706**	**0.473**
YOLO-face-darknet-53	**0.825**	**0.778**	**0.525**
YOLO-face-deeper darknet	**0.899**	**0.872**	**0.693**

(جدول۲: سرعت تشخیص YOLO-face، FAN، Face R-FCN و PyramidBox)

	Speed(fps)
FAN-1200	11(Titan xp GPU)
FAN-400	42(Titan xp GPU)
Face R-FCN	3(K80 GPU)
PyramidBox	3(Titan RTX GPU)
YOLO-face(darknet-53)	**45(1080Ti GPU)**
YOLO-face(deeper darknet)	**38(1080Ti GPU)**

(جدول۳: فراخوانی تشخیص FAN-400 و YOLO-face در مجموعه WIDERFACE)

	Recall
FAN-400	0.546
YOLO-face(darknet-53)	**0.693**

قابل توجه است، دراین مدل آشکارساز بهبود یافته، هزینه محاسباتی[28] به میزان کمی افزایش میابد، بنابراین سرعت شناسایی سریعتری را دارد. از طرفی، مدل معرفی شده به دلیل مشکلات تشخیص، در سه زیر مجموعه داده ارزیابی گردیده است. ("آسان" ، "متوسط" و "سخت"). ارزیابی جداگانه بر روی این زیر مجموعه ها می تواند سازگاری ردیاب را در سناریوهای متفاوت کشف کند.

از طرف دیگر، برای از بین بردن تأثیر نسبت های بهبود یافته anchor box، معیار GIoU و ستون فقرات[29] بهبود یافته، روش ها را به طور جداگانه فقط با استفاده از یک بهبود و ترکیب آنها ارزیابی می نماید.

28.computational cost
29. backbone

نتایج تجربی در شکل ۳ نشان داده شده است. همچنین YOLO-face و YOLOv3 اصلی را در مجموعه داده FDDB مقایسه نموده و نتایج درشکل۴ نشان داده شده است.

در آزمایشات انجام شده، اگر IoU بین آن و جعبه های لنگر پیش بینی شده[30] از ۰,۵ بیشتر باشد، bounding box پیشنهادی مثبت تلقی می شود در غیر اینصورت، منفی برچسب گذاری میگردد.

در زیرمجموعه "سخت"، فقط با استفاده از یکی از پارامترهای پیشرفت (anchor box)، یا معیار GIoU ، نتایج ضعیفی در مقایسه با YOLOv3 نشان می دهد، اما ترکیب این دو پارامتر، عملکرد بسیار بالایی را نسبت به YOLOv3 نشان داده است.

دلیل این موضوع، بدین صورت میباشدکه جهت شناسایی anchor box کوچک، loss functions خطای بین bounding box هدف و bounding box های شناسایی شده را به درستی نشان نمیدهد. استفاده از پارامتر GIoU به تنهایی نیز، loss functions را افزایش میدهد بنابراین، جهت شناسایی چهره با مقیاس کوچک نیاز به آموزش نسبتاً کمتری میباشد. به عبارتی، استفاده از دو بهبود به دست آمده در کنار یکدیگر تأثیرات مثبتی را جهت شناسایی چهره های کوچک از خود نشان داده است. ستون فقرات پیشنهادی نیز به بهبود عملکرد شناسایی کمک بسیار زیادی نموده است.

همانطور که در جدول ۱ و ۲ نشان داده شده است. FAN چهار برابر کندتر از YOLO میباشد. همچنین آزمایشاتی با PyramidBox بر روی همان مجموعه داده انجام شده است و نتیجه نشان می دهد که PyramidBox حدود ده برابر کندتر عمل نموده است. همچنین نتایج مشابهی با Face R-FCN بدست آمده است. سرعت FDNet در مقاله گزارش نشده است، اما FDNet یک آشکارساز دو مرحله ای است و از یک آزمون چند مقیاسی و تعداد زیادی پیشنهاد (۶۰۰۰) در RPN استفاده میکند. عملکرد آن، بعید است سریعتر از YOLO صورت پذیرد.

لازم به ذکر است که هنگام استفاده از FAN-400 ، به عنوان مثال، اندازه ورودی روی ۴۰۰ تنظیم شده، FAN دارای سرعت تشخیصی مشابه با YOLO-face است، اما روش پیشنهادی، عملکرد بهتری در قسمت"سخت" مجموعه داده اعتبارسنجی WIDER FACE دارد.

۴. نتیجه‌گیری

در این تحقیق از YOLOv3 به عنوان ستون فقرات[31] برای شناسایی چهره پیشنهادی خود استفاده نموده، و از چندین جنبه آن را بهبود داده ایم، از جمله یادگیری مقیاس ها و نسبت های خاص anchor box ها، برای شناسایی چهره های انسان، همچنین معرفی GIoU به تابع loss function جدید و استفاده آن درساختار شبکه صورت پذیرفته است.

معرفی یک روش بهبود یافته، بر روی مجموعه داده WIDER FACE و آموزش آن بر روی مجموعه داده FDDB و مجموعه داده WIDER FACE صورت گرفته و نهایتا ارزیابی گردید.

آزمایشات جامعی برای مقایسه روش پیشنهادی با برخی از آشکارسازهای محبوب چهره انجام شده است، که نتایج نشان می دهد که روش بهبود یافته می تواند تعادل بین عملکرد[32] و سرعت را به دست آورد، همچنین روش پیشنهادی سازگار و انعطاف پذیر بوده، و ممکن است با انطباق متناسب با سناریوهای خاص به نتایج دقیقتری نیز برسد. برخی از پیشرفت ها ممکن است مانند اندازه تصویر ورودی بزرگتر، مقیاس های anchor box مناسب برای سناریوهای خاص و داده های آموزش بیشتر مورد استفاده قرار گیرند. تشخیص چهره یکی از موارد تشخیص اشیا با خواص ویژه در حالات خاص میباشد.

30.groundtruth bounding box
31. backbone
32.performance

۵. قدردانی

بدینوسیله از استاد محترم جناب آقای دکتر علی اکبرکیایی که مرا در انجام این تحقیق یاری فرمودند، صمیمانه تشکر می‌نمایم.

۶. مراجع

[1] Chen, W., et al., *YOLO-face: a real-time face detector*. The Visual Computer, 2020: p. 1-9.
[2] Yang, W. and Z. Jiachun. *Real-time face detection based on YOLO*. in *2018 1st IEEE international conference on knowledge innovation and invention (ICKII)*. 2018 . IEEE.
[3] Bochkovskiy, A., C.-Y. Wang, and H.-Y.M. Liao, *Yolov4: Optimal speed and accuracy of object detection.* arXiv preprint arXiv:2004.10934, 2020.
[4] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates, Inc., New York (2012)
[5] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
[6] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (1), vol. 1, pp. 511–518 (2001)
[7] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010). https://doi.org/10.1109/tpami.2009.167
[8] He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1904–1916 (2015)
[9] Girshick, R.: Fast r-CNN. arXiv preprint arXiv:1504.08083 (2015) 7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards realtime object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing, pp. 91–99 (2015)
[10] Dai, J., Li, Y., He, K., Sun, J.: R-fcn: object detection via regionbased fully convolutional networks. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 379–387. Curran Associates, Inc., New York (2016)
[11] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)