آشکارسازی و بازشناسی یکپارچه متن از تصاویر طبیعی با به‌کارگیری فرهنگ لغت

فاطمه نعیمی[۱]، دانشجوی دکتری، وحید قدس[۲]، استادیار، حسن خالصی[۳]، استادیار

۱- دانشجوی دکتری الکترونیک - واحد سمنان - دانشگاه آزاد اسلامی - سمنان - ایران - daneshjo_naimi@yahoo.com
۲- گروه مهندسی برق - واحد سمنان - دانشگاه آزاد اسلامی - سمنان - ایران - v.ghods@semnaniau.ac.ir
۳- گروه مهندسی برق - واحد گرمسار - دانشگاه آزاد اسلامی - گرمسار - ایران - h.khalesi@iau-garmsar.ac.ir

چکیده: در سال‌های اخیراً شکار سازی و باز شنا سی متن در ت صاویر طبیعی به‌طور گ سترده مورد مطالعه قرار گرفته ا ست. در این پژوهش، یک سیستم مکان‌یابی متن در صحنه چندجهته مقاوم برای به دست آوردن بازدهی بالا در آشکارسازی متن بر اساس شبکه عصبی پیچشی(CNN) ارائه شده ا ست. روش پی شنهادی شامل سه لایه ا ستخراج ویژگی، ادغام ویژگی و خروجی می‌با شد. در لایه ا ستخراج ویژگی، یک لایه ReLU بهبود یافته (i.ReLU) معرفی شـده اسـت. همچنین به‌منظورآشـکار سازی متون با ابعاد متنوع، یک لایه inception بهبود یافته (i.inception) ارائه شده است. سپس، برای بهبود استخراج ویژگی از یک لایه اضافی استفاده شده است که ساختار پیشنهادی را قادر می‌سازد متون چندجهته حتی منحنی و عمودی را آشکارسازی نماید. همچنین، یک چارچوب خط لوله برای بازشناسی کاراکتر پیشنهاد نموده‌ایم. چارچوب خط لوله پیشنهادی شامل دو خط لوله موازی ا ست که به‌طور هم‌زمان پردازش می شوند. خط لوله اول، متشکل از کلمات برش یافته و خط لوله دوم شامل زوایای متن می‌باشد. سپس، یک فرهنگ لغت جهت اصلاح خطای احتمالی کلمات بازشناسی شده استفاده نمودیم. آزمایش‌ها بر روی مجموعه داده‌های ICDAR 2013 ، ICDAR 2015 وICDAR 2019، نشان از برتری بارز سیستم پیشنهادی نسبت به کارهای پیشین دارد.

واژه‌های کلیدی: مکان‌یابی متن در صحنه، آشکارسازی تصویر متن، چندجهته، شبکه عصبی پیچشی، بازشناسی متن، بازشناسی یکپارچه متن، فرهنگ لغت.

End to End Text Detection and Recognition of Natural Images using Dictionary

Fatemeh Naiemi, Ph.D. [1], Vahid Ghods, Assistant professor[2], Hassan Khalesi[3], Assistant professor

1-Electronic Ph.D. student, Semnan Branch, Islamic Azad University, Semnan, Iran. Email: daneshjo_naimi@yahoo.com
2-Department of Electronic Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran. Email: v.ghods@semnaniau.ac.ir
3-Department of Electronic Engineering, Garmsar Branch, Islamic Azad University, Garmsar, Iran. Email: h.khalesi@iau-garmsar.ac.ir

Abstract: In recent years, text detection and recognition in natural images have been extensively studied.In this study, a robust multi-oriented scene text localization system was proposed to obtain high efficiency in text detection based on a convolutional neural network (CNN). The proposed method includes three layers of feature extraction, feature-merging, and output. An improved ReLU layer (i.ReLU) is introduced in the feature extraction layer. An improved inception layer (i.inception) is also provided to detect texts with valuable information.An extra layer has been used to improve the feature extraction, which enables the proposed structureto detect multi-oriented even curved and vertical texts. We have proposed a pipeline framework for character recognition.The proposed pipeline framework consists of two parallel pipelines that are processed at the same time, and can recognize 62 characters. The first pipeline consists of cropped words and the second pipeline consists of text angles. Then, we formed a dictionary and used it to correct the possible error of the recognized words. Experiments on the ICDAR 2013, ICDAR 2015 and ICDAR 2019 datasets demonstrated the architectural superiority of the proposed structure over the previous works.

Keywords: Scene text localization, Text image detection, Multi Oriented, Convolutional neural network, Text recognition, End to end recognition, Dictionary.

۱- مقدمه

آشکارسازی و بازشناسی متن (بازشناسی یکپارچه متن) معمولا در محیط‌های طبیعی از قبیل آرم‌های تجاری، پلاک‌های ماشین و علائم مورد استفاده قرار می‌گیرد و یک موضوع تحقیقاتی محبوب در انواع برنامه‌های کاربردی سیستم هوشمند مانند وسایل نقلیه خودمختار در زمان واقعی، سیستم‌های بازشناسی پلاک خودرو و سیستم‌های دستیار بینایی است [۱ و ۲].

به‌طور کلی آشکارسازی و بازشناسی متن به سه دسته آشکارسازی متن، بازشناسی متن و آشکارسازی و بازشناسی یکپارچه متن تقسیم می‌شوند [۳]. روش آشکارسازی مکان متن، شامل راه‌هایی برای یافتن نواحی که احتمال حضور متن در تصویر وجود دارد، می‌باشد. در روش بازشناسی متن، فرایند تبدیل نواحی متن آشکارسازی‌شده به نمادهای قابل ویرایش و قابل‌خواندن برای کامپیوتر انجام می‌گیرد. روش آشکارسازی و بازشناسی یکپارچه متن شامل سیستمی است که تواما هم به آشکارسازی مکان و هم بازشناسی متن می‌پردازد. از آنجا که تصاویر در صحنه واقعی به طور طبیعی تعداد زیادی از اشیاء بی‌ربط (پیچیدگی پس‌زمینه) را به همراه محتوای متن پوشش می‌دهند، به یک روش آشکار سازی و بازشناسی متن قوی نیاز است. با این حال، محلی سازی درست متن‌ها برای بازشناسی کاراکترها یک روش چالش‌برانگیز است [۱، ۸-۴]. اشکالاتی جدی برای الگوریتم‌های محلی سازی و بازشناسی متون در جهت افقی (یا تقریبا افقی) وجود دارد. هم‌چنین از آنجا که در دنیای واقعی متون در اندازه‌های مختلف[۱] دارای طیف گسترده‌ای از جهت‌ها هستند، چنین محدودیت‌هایی باعث می شود روند استخراج ویژگی در متون غیر افقی با مشکلات بیشتری روبرو شود [۱۱-۹]. در سال‌های اخیر برایآشکارسازی و محلی سازی متن از دو روش آنالیز اجزای متصل‌به‌هم و پنجره‌های لغزان استفاده شده است [۱۲ و ۱۳].

در روش اجزای متصل از ویژگی‌های رنگ، لبه، قلم (محاسبه چگالی و شیب پیکسل) و بافت برای تعیین و بررسی نواحی کاندیدای متن استفاده می‌نمایند. ابتدا از طریق خوشه‌بندی رنگ، لبه یا استخراج نواحی مفرط، اجزای کاندیدا تعیین می‌گردند (استخراج کاندیدای متن)، سپس اجزای غیر متن از طریق قوانین طراحی دستی یا طبقه‌بندی‌های آموزش دیده خودکارفیلتر می‌شوند (اصلاح نمودن کاندیدا از طریق حذف نمودن کاندیدای غیر متن). این روش بسیار کاربردی و مؤثر است زیرا تعداد اجزای قابل‌پردازش بسیار اندک هستند. مزیت اصلی این روش را می‌توان غیر حساس بودن به چرخش و جهت، تغییر مقیاس و تنوع فونت برشمرد. اما به درهم‌ریختگی و کلفت شدن متون که باعث تغییر اجزای متصل می‌شوند بسیار حساس هستند و از معایب این روش به شمار می‌رود[۱۳]. همچنین با اصلاح کاندیدا و حذف نمودن کاندیدای غیر متن امکان از بین رفتن تعدادی از کاراکترها و متن‌ها نیز وجود دارد که باعث کاهش نرخ فراخوانی می شود[۱۳]. همچنین در روش مبتنی بر پنجره‌های لغزان، آشکار سازی متن، به‌وسیله شیفت دادن یک پنجره به‌تمامی مکان‌ها در چندین مقیاس مختلف، تعریف می‌گردد. این روش، یک جستجوی جامع است، بنابراین نرخ فراخوانی آن بالا می‌باشد، اما به‌دلیل اینکه با استفاده از این روش تعداد زیادی کاندیدای متن به‌دست می‌آید و تعداد مثبت‌های کاذب بالا می‌رود، همچنین به‌دلیل اسکن نمودن کل پنجره‌ها حجم محاسباتی بالاست [۱۳]. در سال‌های اخیر پژوهش‌های بسیاری با استفاده از یادگیری عمیق جهت آشکار سازی و بازشناسی متن انجام گرفته است [۱۷-۱۵]. برای طبقه‌بندی متن از غیر متن در تصاویر و همچنین تعیین ویژگی‌های تصاویر، ویژگی‌های عمیق در چندین لایه از طریق شبکه‌های عصبی پیچشی CNN[۲]، محاسبه می‌گردند. روش‌های ذکر شده با معماری‌های مختلف از محدودیت‌های مختلف و نتایج نامطلوب در محلی سازی و بازشناسی جهت‌های مختلف متن با تغییرات کنتراست و تغییرات اندازه فونت در صحنه‌های زندگی واقعی رنج می‌برند. این اشکالات باعث عدم اطمینان در برخی از حوزه‌های کاربرد خاص، مانند ترافیک هوشمند و هوشمندسازی جدید در زیر سیستم‌های وسایل نقلیه می شود. با توجه به چالش‌های ذکر شده، هدف پژوهش حاضر، پیاده‌سازی یک سیستم کارا برای آشکارسازی و بازشناسی متن صحنه است. این سیستم پیشنهادی سازگار با آشکارسازی و بازشناسی متن در صحنه‌های واقعی است، حتی اگر متن صحنه دارای شکل خمیده باشد یا دارای چرخش ۹۰ درجه باشد. به‌علاوه، چند مرحله میانی مانند کاندیدای پیشنهادی، تقسیم‌بندی کلمات و آرایش نواحی متن از بین رفته است که باعث می‌گردد زمان اجرای الگوریتم پیشنهادی در مقایسه با الگوریتم‌های مشابه بهبود یابد. به‌منظور غلبه بر مشکلات بیان شده، مدل آشکار سازی و بازشناسی روش پیشنهادی با استفاده از ReLU[۳] و بلوک‌های inception و رویکرد خط لوله اجرا می شود. روش پیشنهادی نسبت به تغییرات رنگ مقاوم است و از پیچیدگی‌ها و مقیاس‌های مختلف پشتیبانی می‌کند. به طور خلاصه، نوآوری‌های پیشنهاد شده، در این پژوهش عبارت است از: (۱) یک چارچوب یکپارچه ارائه شده است که آشکارسازی و بازشناسی متن صحنه (سیستم بازشناسی یکپارچه) در دو مرحله متوالی و جداگانه ارائه می‌گردد. (۲) ابتدا، یک سیستم مکان‌یابی متن در صحنه مقاوم چندجهته برای به دست آوردن بازدهی بالا در آشکارسازی متن بر اساس شبکه عصبی پیچشی[۴] (CNN) ارائه شده است. در روش پیشنهادی، یک لایهReLU بهبود یافته[۵] (i.ReLU) و یک لایه inception بهبود یافته[۶] (i.inception) معرفی شده است. ساختار پیشنهادی برای استخراج ویژگی‌های دیداری سطح پایین استفاده می شود. سپس از یک لایه اضافی برای بهبود استخراج ویژگی استفاده شده است. لایه‌هایi.ReLU باعث می‌شوند ویژگی‌های سطح پایین بیشتری استخراج شوند. لایه‌های i.inception می‌توانند متنی با ابعاد

متنوع و متفاوت را به‌طور مؤثرتر از زنجیره‌ای خطی لایه پیچشی (بدون لایه‌های inception) آشکارسازی نمایند. خروجی لایه‌های i.ReLU و لایه‌های i.inception به لایه اضافی تغذیه می‌شوند که ساختار پیشنهادی را قادر می‌سازد متون چندجهته حتی منحنی و عمودی را آشکارسازی نماید. (۳) بخش بازشناسی متن پیشنهادی از i.inception و i.ReLU بخش قبلی استفاده می‌نماید. الگوریتم پیشنهادی می‌تواند ۵۲ حرف انگلیسی (۲۶ حرف کوچک و ۲۶ حرف بزرگ)، اعداد انگلیسی ۰ تا ۹ (۱۰ رقم) را بازشناسی نماید که در کل ۶۲ کاراکتر را شامل می شود. در سیستم بازشناسی کاراکتر پیشنهادی، یک چارچوب خط لوله ایجاد می‌شود که لایه‌های موازی به طور هم‌زمان پردازش می‌شوند. (۴) سپس یک فرهنگ لغت متشکل از پایگاه‌های داده مذکور تشکیل داده و از آن جهت اصلاح خطای احتمالی کلمات بازشناسی شده از مرحله بازشناسی کلمات استفاده نمودیم.

ساختار مقاله حاضر بدین شرح است: در ادامه مقاله، در بخش ۲ مروری اجمالی بر پژوهش‌های پیشین صورت می‌گیرد. بخش ۳، به شرح کامل روش پیشنهادی اختصاص دارد. در بخش ۴، نتایج حاصل از روش پیشنهادی، بحث و ارزیابی آن‌ها ارائه شده است. در نهایت بخش ۵ به نتیجه‌گیری می‌پردازد.

۲- مروری بر پژوهش‌های پیشین

در سال‌های اخیر، مقالات متعدد‌وپیرامون شناخت متن صحنه منتشر شده و مطالعات جامع در [۱۸، ۱۹] ارائه شده است. در میان روش‌های سنتی، بسیاری از پژوهش‌ها رویکردهای از پایین به بالا را اتخاذ می کنند که در ابتدا هر کاراکتر با استفاده از روش‌های مبتنی بر پنجره لغزان [۲۰، ۲۱] و اجزای متصل‌به‌هم[۲۲]، آشکارسازی می‌شود. پس از آن، کاراکترهای آشکارسازی‌شده با استفاده از برنامه‌نویسی پویا، جستجوی واژگان [۲۰] و غیره به کلمات ادغام می‌شوند. برخی دیگر از پژوهش‌ها، رویکردهای بالا به پایین را اتخاذ می‌کنند. در این روش به جای آنکه در ابتدا آشکار سازی و بازشناسی هر کاراکتر به صورت جدا انجام گردد، متن مستقیما از کل تصاویر ورودی تشخیص داده می‌شود.

در سال‌های اخیر در روش‌های سنتی برای مکان‌یابی متن به دلیل توجه زیاد محققان به روش‌های مبتنی بر اجزای متصل به‌هم اکثر کارهای موجود در این حوزه است [۲۶-۲۳]. درروش ارائه شده توسط جین و همکاران [۲۴] تصاویر توسط خوشه‌بندی رنگی، مؤلفه‌های گروه‌بندی‌شده در خط متن، تجزیه تحلیل مؤلفه‌ها و حذف مؤلفه‌های غیرمتنی براساس قوانین هندسی به چندین بخش بدون تداخل تقسیم می‌شدند. این روش به دلیل تنظیم دستی قوانین و پارامترها برای تصاویر پیچیده‌ی طبیعی کند بود و عملکرد خوبی نداشت. اپستین و همکارانش [۱۴] روش استخراج ویژگی جدیدی به نام تبدیل پهنای قلم[۷] (SWT) را مطرح نمودند. این روش با تعیین نواحی لبه به بررسی چگالی و ضخامت کاراکترها پرداختکه در استخراج اجزای متن با سایزها و جهات مختلف در تصاویر پیچیده بسیار مؤثر و کاربردی بود. اما با توجه به محدودیت‌ها و قوانینی که برای آن در نظر گرفته شده بود، تنها برای آشکارسازی متون افقی کاربرد داشت. نویمن و همکارانش [۱] یک الگوریتم جامع آشکارسازی و بازشناسی نواحی حدی بیشینه پایدار[۸] (MSER) را معرفی نمودند که بر اساس نواحی حدی با پایداری بسیار بالا بود، ابتدا کاندیدای کاراکتر از تصویر استخراج شده و کاندیداهای نامعتبر با استفاده از یک طبقه‌بندی ماشین بردار پشتیبان حذف شدند. سپس در مرحله بعد، کاندیداهای باقیمانده از طریق مجموعه‌ای از قوانین اتصال، به خطوط متنی[۹] (کلمه) تبدیل شدند. در مرحله بعد برای بازشناسی به سیستم OCR[۱۰] که با استفاده از مقدار زیادی از کاراکترهای مصنوعی آموزش دیده بود اعمال شدند. این الگوریتم برای بازشناسی نیازمند فهرستی از کلمات نبود. با این‌حال، قوانین اتصال تنها می‌توانست متناسب بامتون افقی یا نزدیک به افقی باشد، بنابراین این الگوریتم مناسب پردازش متون با زاویه شیب بالا نبود. سپس نویمن و همکارانش [۱۹] یک روش جدید استخراج ویژگی آشکارسازی و بازشناسی متن و استراتژی ترکیبی در دو سطح آشکارسازی متن بر اساس مدل جامع پنجره‌های لغزان معرفی نمودند. برای گروه‌بندی کاراکترهای آشکارسازی شده به کلمات از خطوط متن استفاده نمودندکه با در نظر گرفتن ویژگی‌های رنگ، اندازه شیب و قلم (محاسبه چگالی و شیب پیکسل) و استفاده از الگوریتم تغییر پهنای قلم، در دقت و کارایی سیستم تأثیر بسزایی داشت. آن‌ها جهت اصلاح خطاهای مرحله بازشناسی سه فرهنگ لغت در اندازه‌های مختلف معرفی نمودند: فرهنگ لغت قوی شامل ۱۰۰ کلمه خاص برای هر تصویر، فرهنگ لغت ضعیف شامل کلمات موجود در مجموعه داده و فرهنگ لغت عمومی شامل کلمات انگلیسیK۹۰ بود.

در چند سال گذشته، شبکه‌های عصبی پیچشی عملکرد بسیار خوبی در محلی سازی متن (Deep CNN[۲۷] و R-CNN[۲۸]) و بازشناسی متن (شبکه MORAN [۲۹] و CNN [۳۰]) داشته‌اند. ژو و همکاران [۳۱] یک خط لوله آشکارساز متن صحنه کارآمد و دقیق[۱۱] (EAST) ارائه نمودند که با استفاده از یک شبکه عصبی واحد به‌طور مستقیم به پیش‌بینی سطح کلمه یا خط می‌پرداخت. این الگوریتم مراحل واسطه غیرضروری نظیر جمع نمودن کاندیدا و تقسیم‌بندی کلمات را حذف نمود. با ترکیب توابع ضرر مناسب، آشکارساز توانست مستطیل‌های چرخان یا چهارگوش‌های مربوط به مناطق متن را بسته به کاربردهای خاص پیش‌بینی کند. این روش متن‌های منحنی و عمودی را تشخیص نمی‌داد. لئو و همکاران [۳۲] یک مدل آشکارسازی متن را که از ترکیب CNN و RNN بود، معرفی نمودند و آن را آشکار سازی متن صحنه با استفاده از شبکه پیشنهاد متن مبتنی بر هرم[۱۲] (FTPN) نامیدند. این مدل با استفاده از ویژگی‌های هرمی موجود در FPN برای استخراج ویژگی در اندازه‌های مختلف و استفاده از توالی متن برای تولید یک سری از پیشنهادات متنی با استفاده از

Bi-LSTM، محقق شد. این روش با آشکارسازی متن در اندازه‌های متفاوت سازگار است و در پایگاه‌های داده عمومی نتایج خوبی را به‌دست آورده است. این روش برای متن منحنی مناسب نیست. تیان و همکاران [۳۳] یک شبکه پیشنهادی متن متصل‌کننده (CTPN)[۱۳] ارائه دادند، یک آشکارساز متناسب و کارآمد که قابلیت آموزش بازشناسی یکپارچه متن را داشت. CTPN یک خط متن را در یک دنباله از پیشنهادات متن به‌طور مستقیم در نقشه‌های پیچشی آشکارسازی می‌نمود. این روش مکانیسم لنگر عمودی را ایجادکرد که به‌طور مشترک مکان دقیق و امتیاز متن/غیرمتن را برای هر پیشنهادکه کلید محلی سازی دقیق متن بود، پیش‌بینی می‌نمود. آن‌ها همچنین یک لایه RNN درون شبکه‌ای ارائه دادند که پیشنهادات متن متوالی را با ظرافت به هم متصل و این امکان را فراهم نمود تا متون معنی‌دار را آشکارسازی نماید. هوانگ و همکاران [۳۴] یک سیستم مقاوم برای آشکارسازی و مکان‌یابی متن در صحنه در تصاویر طبیعی ارائه دادند. این روش تنها برای متون افقی مناسب بود. این روش بر پایه ترکیب مزایای روش نواحی حدی بیشینه پایدار مبتنی بر مدل یادگیری عمیق به جهت تفکیک مؤلفه‌های متنی از غیر متنی بود. همچنین یک مدل پنجره لغزان با طبقه‌بندی CNN با هم ترکیب‌شدند تا آشکارسازی متن را در تصاویر چالش‌برانگیز بهبود بخشد.

از مباحث فوق، قابل ذکر است که تمایز بسیاری در متون مختلف زبان وجود دارد. بنابراین، اکثر الگوریتم‌های تشخیص متن در زندگی واقعی برای تمامی زبان‌ها کاربرد ندارند و برای چندین زبان محدود، قابل‌اجرا هستند [۳۵]. وانگ وهمکاران [۳۵] سیستم آشکارسازی متون موجود در تصاویر صحنه طبیعی را پیشنهاد دادند. در این روش، به‌منظور حذف تداخل‌های پس‌زمینه، یک طرح تأیید کاندیدای کاراکتر، که براساس برخی از قوانین داوری و نقشه اطمینان استوار بود، به آشکارسازی متن می‌پرداخت. این روش برخلاف روش‌های مرسوم، بر ایجاد مدل نقشه اطمینان از طریق ادغام احتمال متن کاندیدای بذر و روابط با کاندیدای مجاور خود برای برجسته کردن متون از پیش‌زمینه‌ها متمرکز بود، کاندیداهای با ارزش اطمینان کم، که زیر یک آستانه مشخص حذف می‌شدند. به‌منظور بهبود نرخ فراخوانی، اطلاعات متن برای بازیابی مناطق متن از دست رفته استفاده می‌شدند. سرانجام، خطوط متنی، شکل گرفته و کلمات، با محاسبه آستانه‌ای برای جدا کردن حروف درون کلمه‌ای از حروف بین کلمه‌ای به دست می‌آمدند. ما و همکاران [۸] یک چارچوب آشکارسازی مبتنی بر چرخش[۱۴] (RRPN) برای آشکارسازی متن با جهات دلخواه معرفی نمودند. این روش قادر به آشکارسازی متن منحنی و عمودی نبود. در این روش، پیشنهادهای مستطیل شیب‌دار با اطلاعات زاویه جهت‌یابی منطقه متن از لایه‌های پیچشی بالاتر شبکه، تولید می‌شد. در نتیجه آشکارسازی متن با جهت‌گیری‌های متعدد بود. برای بهبود کارایی این روش، یک لایه پولینگ RRoI جدید نیز طراحی شده و با RoI چرخان سازگار شد. یانگ و همکاران [۳۶] آشکارسازی متن صحنه چندجهته را معرفی کردند. آن‌ها یک ماژول جدید متن Inceptionو پولینگ PSROI تغییر شکل یافته را برای آشکارسازی متن با مقیاس اندازه و جهت‌گیری‌های مختلف طراحی نمودند. در این آشکارساز از شبکه‌های ResNet۱۰۱ ، شبکه‌های ResNet ۵۰ و VGG استفاده شده است که این امر باعث افزایش زمان محاسباتی می‌شد. سیستم پیشنهادی آنان قادر به تقسیم دو کلمه با فاصله کم نبود. یکدیگر از محدودیت‌های سیستم، عدم آشکارسازی برخی از کاراکترها با پس‌زمینه شلوغ بود. این روش برای تصاویر با زوایای مختلف متن و متون منحنی و عمودی مناسب نیست. نعیمی و همکاران [۱۱] روش جدیدی جهت استخراج ویژگی، تحت عنوان HOG اصلاح‌شده ارائه نمودند.این روش در برابر تغییرات انتقالی و مقیاس کاراکترها مقاوم بوده و از لحاظ محاسباتی و زمانی مقرون به‌صرفه بود. همچنین از دو روش ضخیم‌سازی کاراکترهای نازک و عدم تبعیض در تشخیص حروف بزرگ و کوچک با شکل یکسان، برای بهینه‌سازی دقت روش پیشنهادی استفاده نمودند. در زمینه آشکارسازی متون فارسی نیز پژوهش‌هایی انجام گرفته است که می‌توان به قانعی و فائز [۳۷] و قویدل و همکاران [۳۸] اشاره نمود. قانعی و فائز [۳۷] به محلی‌سازی متون فارسی/عربی و انگلیسی در تصاویر دنیای واقعی پرداختند. در این روش متون در تصاویر طبیعی با اندازه‌ها، فونت‌ها و جهت‌های مختلف و روشنایی پس‌زمینه متفاوت آشکارسازی می‌شوند. در مرحله اول، از فیلتر میانه به عنوان یک فیلتر صاف کننده غیرخطی با حفظ لبه و سپس از رنگ‌زدایی با حفظ کنتراست رنگ استفاده می‌شود تا سیستم محلی سازی متن برای کنتراست‌های کم‌نور و متن‌های بی‌کیفیت آشکارسازی بهتری را انجام دهد. به‌منظور استخراج متن‌های فارسی/عربی و انگلیسی، یک چارچوب واحد پیشنهاد شده است که شامل نواحی حدی بیشینه پایدارو یک آشکارساز جدید منطقه پیشنهادی به نام مناطق با پهنای قلم پایداراست. سرانجام، برای استخراج خطوط متنی، از خوشه‌بندی mean-Shift و تبدیل رادون استفاده شد. قویدل و همکاران [۳۸] الگوریتمی برای شناسایی و محلی سازی متون فارسی و انگلیسی مبتنی بر رنگ لبه معرفی نمودند. در این روش، ابتدا هرمی با استفاده از مقیاس‌های مختلف تصویر ورودی ایجاد شدند. سپس برای هر سطح از هرم یک نقشه لبه استخراج شد. پس از آن، چندین ویژگی هندسی برای فیلتر کردن لبه‌های غیر متنی از لبه‌های استخراج شده استفاده شدند، و یک لبه را با استفاده از رنگ پیکسل‌های همسایه آن توصیف نمودند. برایبه دست آوردن حالت‌های رنگی اطراف هر پیکسل لبه، از الگوریتم mean-Shift استفاده شد. پس از آن، از الگوریتم خوشه‌بندی Single-Linkage برای ساخت گروه‌های معنی‌دار خوشه‌بندی استفاده شد. در آخر، هر یک از خوشه‌ها با استفاده از طبقه‌بندی کسکد مبتنی بر MLP به عنوان متن یا غیرمتن برچسب‌گذاری شدند.

همچنین پژوهش‌های کاربردی بسیاری در زمینه بازشناسی متن انجام گرفته است که به مهم‌ترین آن‌ها پرداخته‌ایم. اسلام و همکاران [۳۹]

یک روش متن صحنه مبتنی بر نواحی حدی بیشینه پایدار پیشرفته ارائه دادند. این سیستم قادر بود قسمت متن را از تصویر صحنه طبیعی متمایز نماید و سپس متن را از ناحیه متن انتخاب شده بازشناسی نماید. برای غلبه بر تاری و اندازه‌های کوچک تصویر و حروف، نواحی حدی بیشینه پایدار از لبه‌یاب کنی استفاده نمودند.

این الگوریتم برای تعیین دقیق پهنای قلم اجزای متصل به باینری، طراحی شده بود. سرانجام از OCR با توصیف کاراکتر متقاطع برای بازشناسی متن استفاده نمودند. ژانگ و همکاران [۴۰] یک مدل مقاوم جدید[۱۵] SSDAN برای بازشناسی تصویر متن ارائه نمودند که این اتصالات بازشناسی متن درون تصویر و سازگاری دامنه را می‌سازد. این روش قادر بود از مزایای داده‌های دنباله بدون نظارت برای یادگیری بازنمایی‌های قوی‌تر استفاده نماید. مدل پیشنهادی همچنین می‌توانست به صحنه‌های مختلف از جمله متن صحنه، متن دست‌نوشته و تشخیص بیان ریاضی تعمیم یابد. شی و همکاران [۵] RARE (بازشناسی متن مقاوم با تصحیح خودکار[۱۶]) را پیشنهاد نمودند. RARE یک مدل بازشناسی مقاوم برای متون منحنی و جهت‌دار بود. RARE از یک شبکه ترانسفورمر فضایی (STN) و یک شبکه تشخیص توالی (SRN) تشکیل شده بود. علاوه بر این، شبکه ترانسفورمر فضایی به یک بازشناسی دنباله‌ای متصل شده و به سیستم اجازه می‌داد کل مدل را به‌صورت یکپارچه آموزش دهد. وانگ و همکاران [۴۱] نشان دادند که بازشناسی متن صحنه در واقع یک مشکل پیش‌بینی مکانی–زمانی است. آن‌ها پیشنهاد دادند تا از دیدگاه مکانی-زمانی به حل این مشکل بپردازند (هم‌زمان از اطلاعات حوزه مکان و زمان تصویر استفاده کنند). به همین منظور، آن‌ها بازشناسی متن صحنه‌ای مؤثر با نام FACLSTM را ارائه دادند، جایی که ConvLSTM با ادغام مکانیسم توجه در ماژول رونویسی پی‌درپی اعمال شده و بهبود می‌یافت. همچنین آن‌ها یک ماژول توجه متمرکز در مرحله استخراج ویژگی رمزگذار-رمزگشا طراحی نمودند. FACLSTM پیشنهادی قادر به بازشناسی متون دارای قاعده و متون بدون قاعده (وضوح پایین، نویز دار و منحنی) بود. سیستم پیشنهادی توسط یائو و همکاران [۹] قادر به آشکارسازی و بازشناسی صحنه‌های متنوع در دنیای واقعی، متونی از مقیاس‌ها، رنگ‌ها، فونت‌ها و جهت‌های مختلف بود. محدودیت‌های الگوریتم پیشنهادی عمدتا به دلیل شرایط روشنایی غیریکنواخت، تاری، وضوح پایین و تضاد کم بین متن و پس‌زمینه بود. نقص‌های جزئی در بازشناسی، بخش‌بندی‌های نامناسب کلمه، فونت‌های نامنظم، کاراکترهای متصل، همگی باعث بروز خطاهای بازشناسی می‌شدند. لازم به ذکر است، معماری LSTM برای بازشناسی متن برون‌خط[۱۷] و برخط[۱۸] مورد استفاده قرار می‌گیرد.

اطلاعات استفاده شده در این پژوهش از نوع برون‌خط است که با توجه به مراجع [۵، ۹، ۳۹، ۴۰، ۴۱]، برای بازشناسی متن از CNN استفاده نمودیم.

۳- روش پیشنهادی

در این بخش یک چارچوب یکپارچه برای آشکارسازی و بازشناسی متن صحنه (سیستم بازشناسی یکپارچه) در دو مرحله متوالی و جداگانه بیان می‌شود.

۳-۱-معماری روش پیشنهادی محلی سازی و آشکارسازی متن

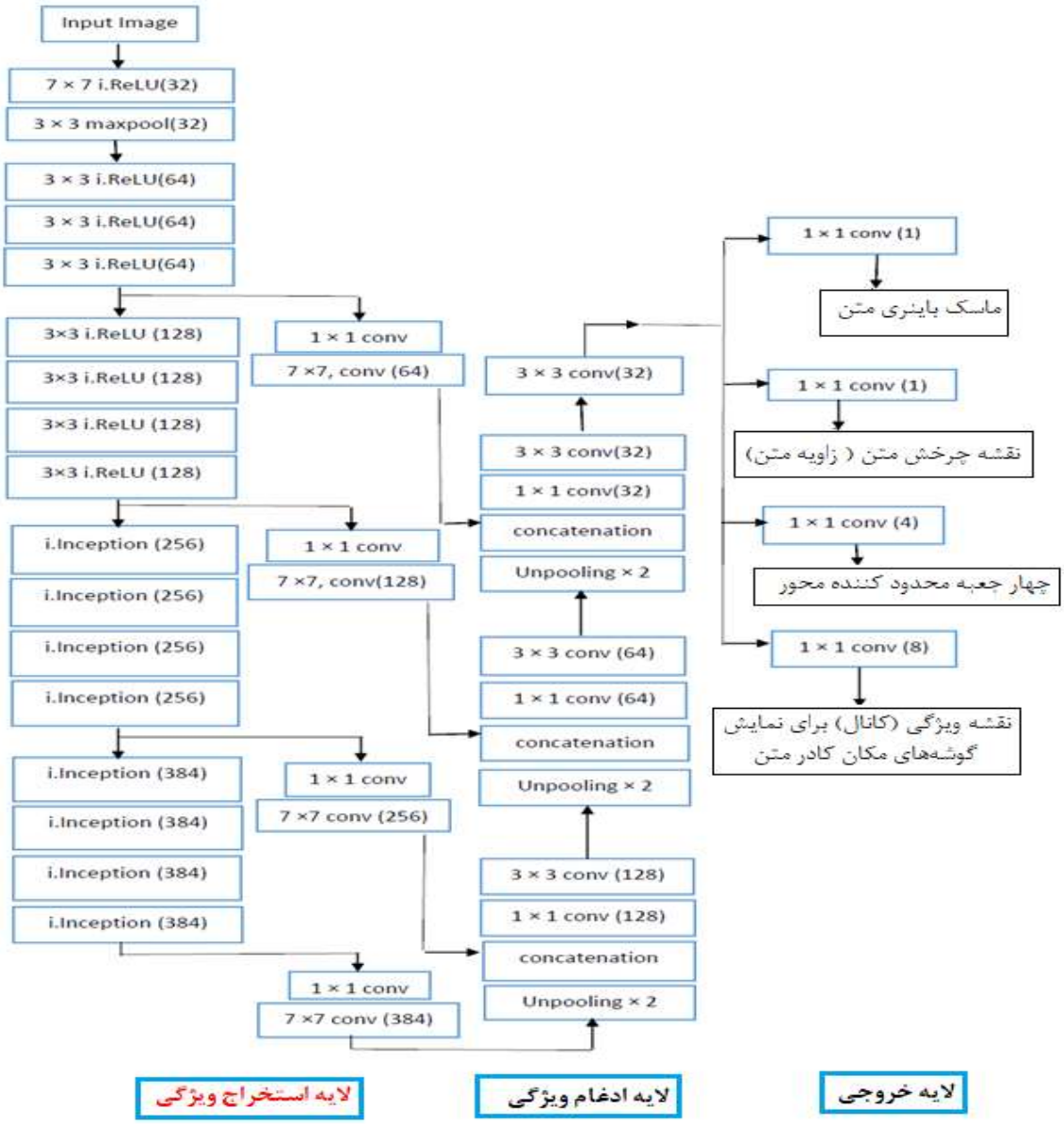
ساختار پیشنهادی مکان‌یابی متن صحنه در شکل ۱ نشان داده شده است. ساختار پیشنهادی از سه لایه تشکیل شده است که به نام‌های لایه استخراج ویژگی، لایه ادغام ویژگی[۱۹] ولایه خروجی نام‌گذاری شده‌اند. ساختار پیشنهادی قادر به آشکار سازی مکان و جهت‌گیری متن است. ساختمان داخلی ساختار پیشنهادی به صورت زیر می‌باشد:

در لایه‌های استخراج ویژگی، یک تصویر ورودی به لایه ۷ × ۷ i.ReLU تغذیه شده و سپس لایه ۳ × ۳ maxpool که حاوی ۳۲ کانال است را دنبال می‌کند. سپس، از هفت (سه به علاوه چهار) لایه ۳ × ۳ i.ReLU برای استخراج ویژگی‌های بصری سطح پایین استفاده می‌شود [۴۲، ۴۳]. خروجی لایه‌های ۳ ×۳i.ReLU (هر یک از سه لایه i.ReLU اول شامل ۶۴ کانال و هر چهار لایه دیگر شامل ۱۲۸ کانال) به هشت(چهار به علاوه چهار) لایه i.inception (هر یک از چهار لایه i.inception اول شامل ۲۵۶ کانال و هر چهار لایه دیگر شامل ۳۸۴ کانال است) داده می‌شود.

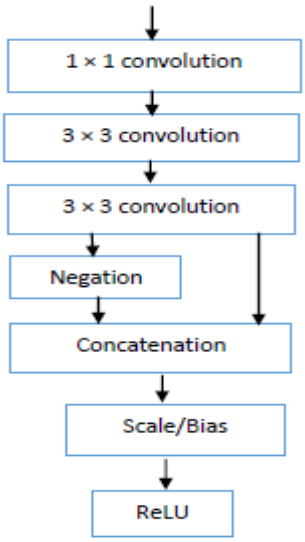
همان‌طور که در شکل ۱ نشان داده شده است، یک لایه استخراج ویژگی اضافی نیز برای بهبود روش استخراج ویژگی پیشنهادشده است. خروجی لایه های i.ReLU و i.inception در چهار نقطه به چهار لایه الحاقی[۲۰]×۱ ۱ تغذیه می شوند. این چهار لایه الحاقی، لایه‌های پیچشی ۷ × ۷ را دنبال می‌کنند که هدف آن آشکارسازی زاویه‌های مختلف متن حتی متن عمودی است. بلوک‌های حاوی لایه استخراج ویژگی که حاوی لایه‌های i.ReLU و لایه‌های i.inception و بلوک‌های لایه اضافی ساختار روش پیشنهادی را تشکیل می‌دهند و باعث می‌گردند ساختار پیشنهادی متون چندجهته حتی منحنی و عمودی را آشکارسازی نماید.

لایه‌های پیچشی ۷ × ۷ از آخرین مرحله لایه استخراج ویژگی ابتدا به یک لایه unpooling تغذیه می‌شوند تا اندازه آن در هر لایه ادغام ویژگی دو برابر شود و سپس با لایه‌های پیچشی ۷ × ۷ قبلی جمع می‌شود. سپس، یک پیچشی ۱ × ۱ تعداد کانال‌ها و محاسبات را کاهش می‌دهد. این لایه یک پیچشی ۳ × ۳ را دنبال می‌کند که اطلاعات را ادغام و در نهایت بازده این مرحله را تولید کند. در آخرین مرحله ادغام، یک لایه پیچشی ۳ × ۳ نقشه ویژگی نهایی را تولید می‌کند و آن را به لایه خروجی می‌دهد [۳۱، ۴۴]. همان‌طور که در شکل ۱ دیده می‌شود، در لایه‌های اولیه، i.ReLU پیشنهادی به کار می‌رود. بلوک ۳×۳ i.ReLU پیشنهادی در شکل ۲ نشان داده شده

است که از الگوهای فعال‌سازی واسطه‌ای[۲۱] در شبکه‌های عصبی پیچشی (CNNs) ایجاد شده است. در شکل ۲، گره‌های خروجی با طرف مقابلشان جفت می‌شوند. به همین دلیل، از آنجا که این ویژگی‌ها قبل از اعمال روی لایه ReLU با هم جمع می‌شوند، تعداد کانال‌های خروجی را می‌توان دو برابر کرد [۴۲، ۴۵]. علاوه بر این، یک لایه بایاس اضافه شده که باعث می شود فیلترهای همبسته[۲۲] دارای مقادیر بایاس مختلف باشند. در بلوک i.ReLU، خروجی پیچشی با ضرب ۱- در جعبه به دست می‌آید. همچنین، وزن و بایاس قابل آموزش توسط هر جعبه "مقیاس/بایاس" (Scale/Shift) اعمال می‌شود [۴۵]. لایه i.ReLU اجازه می‌دهد تا برخی از ویژگی‌های سطح پایین را به طور مناسب استخراج کنیم، به‌طوری‌که استفاده از دو لایه ۳ × ۳ ویژگی‌های سطح پایین بیشتری را استخراج می‌کند و با منفی نمودن کانال و جمع شدن با کانال مثبت، تعداد کانال‌ها دو برابر می‌شود.

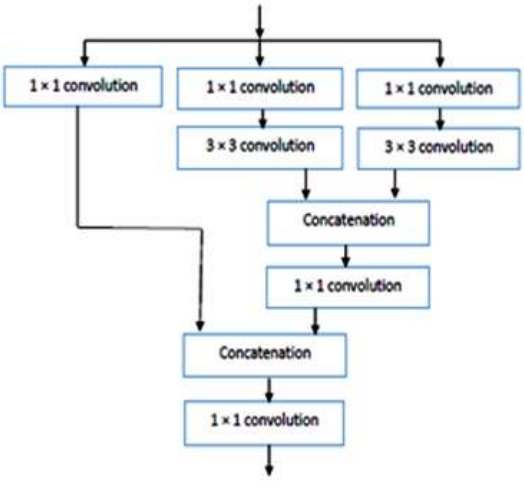


شکل ۱: ساختار پیشنهادی



شکل ۲: ساختار i.ReLU ۳×۳ پیشنهادی

همچنین از ساختار inception برای به دست آوردن اندازه‌های مختلف در آشکارسازی متن استفاده می‌شود [۴۵، ۴۶]. ساختار پیشنهادی i.inception در شکل ۳ نشان داده شده است. لایه i.inception پیشنهادی از برخی ایده‌های ارائه شده در [۴۶، ۴۷] الهام گرفته شده است. ایده اصلی که تو سط چارچوب inception پیشنهاد شده است، حذف لایه پیچشی و استفاده از تعدادی ساختارهای موازی است. این روش باعث کاهش تعداد پارامترها و تعدد نیروها[۲۳] بر روی ویژگی‌های خروجی هر لایه می شود. لایه‌های inception (پیچشی ۳× ۳ یا کرنل‌های بزرگ‌تر) می‌توانند متن با ابعاد مختلف را به‌طور مؤثرتر از لایه پیچشی زنجیره خطی[۲۴] آشکارسازی نمایند. به دلیل استفاده از این ماژول، فعالیت‌های خروجی را می‌توان با اندازه‌های تغییر یافته از گیرنده‌ها تولید کرد. این بدان معنی است که تنوع ابعاد گیرنده در لایه قبلی افزایش یافته است [۴۲، ۴۵].



شکل ۳:ساختار i.inception پیشنهادی

لایه خروجی از چهار نقشه ویژگی جداگانه تشکیل‌شده است. ابتدا، نقشه باینری تولید می‌شود که فقط پیکسل‌های داخل کادر متن مقدار ۱ (true) را به دست آورده و بقیه پیکسل‌ها ارزشان ۰ (false) می‌باشد. جعبه متن، شامل یک سری کاراکتر است که توسط یک خط مستقیم یا منحنی به هم وصل می‌شوند. برای تولید ماسک باینری متن (کلمه برش زده شده)، $LOC = \left\{ p_i \middle| i \in \{1,2,3,4\} \right\}$ را در نظر می‌گیریم، که چهار رأس نشان داده می شوند. همچنین، طول مرجع[۲۵] برای هر رأس به صورت زیر محاسبه می‌شود.

$$ref_i = \min(dist\left(p_i, p_{(i \bmod 4)+1}\right), dist(p_i, p_{((i+2) \bmod 4)+1})) \quad (1)$$

که $dist\left(p_i, p_j\right)$ فاصله اقلیدسی بین p_i و p_j را نشان می‌دهد. نقشه نمره[۲۶] (ماسک باینری) با کوچک کردن هر لبه جعبه به وسیله $0.4 \times ref_i$ و $0.4 \times ref_{(i \bmod 4)+1}$ به دست می‌آید.

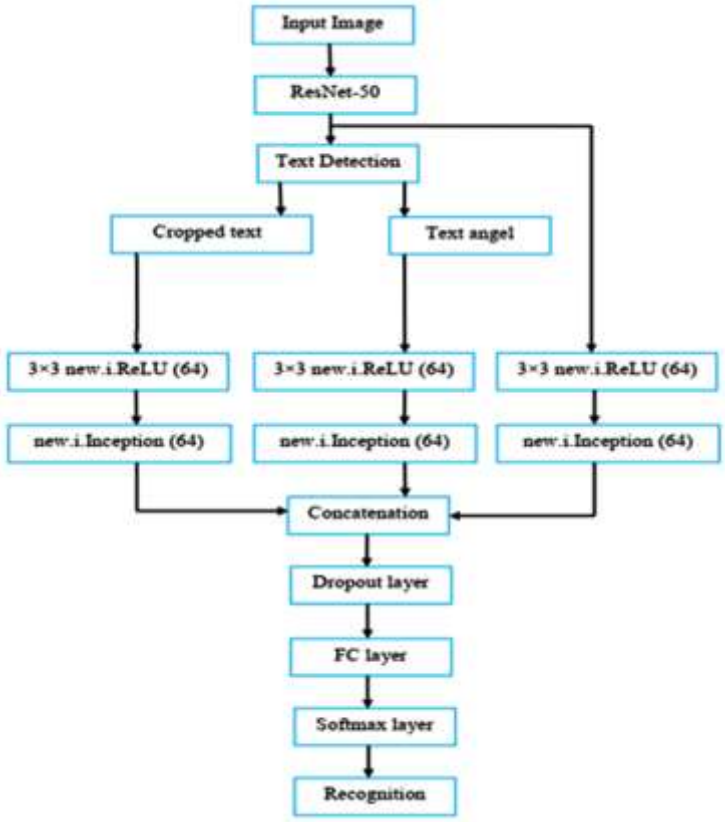
در مرحله بعد، نقشه چرخش متن (زاویه متن) محاسبه می‌شود.این نقشه نشان‌دهنده چرخش هر کاراکتر با دقت قابل قبول می‌باشد که در داخل کادر تو سط یک نقشه ویژگی (کانال)، تعریف شده است[۳۱]. سوم، چهار جعبه محدود کننده محور[۲۷] (AABB) مطابق با [۲۰، ۴۸] محاسبه می شوند.فاصله پیکسل‌ها تا مرزهای بالا، راست، پایین، سمت چپ چهارگوشه کادر محدودکننده متن توسط چهار نقشه ویژگی (کانال) محاسبه می شود. سرانجام، هشت نقشه ویژگی (کانال) برای نمایش گوشه‌های مکان کادر متن در دو جهت x و y تولید می‌شود.

به‌طورکلی، تغییرات در i.ReLU استخراج ویژگی‌های سطح پایین را بهبود می‌بخشد و تغییر در i.inception باعث بهبود آشکارسازی تصاویر در اندازه‌های مختلف می شود. همچنین، اضافه کردن یک لایه اضافی به لایه‌های استخراج ویژگی (ستون دوم در شکل ۱)، که لایه‌های i.ReLU و لایه‌های i.inception در چهار نقطه به پیچشی ۱×۱ تغذیه می‌شوند، ساختار پیشنهادی را قادر ساخته است تا باعث استخراج ویژگی های بیشتر و مکان متن را در جهات مختلف آشکارسازی نماید.

۳-۱-شبکه پیچشی پیشنهادی جهت بازشناسی کاراکتر

از آنجا که ما اطلاعات مهم و اساسی را از فرآیند آشکارسازی متن استخراج نمودیم، در این مرحله اطلاعات کمی برای بازشناسی هر کلمه برش یافته نیاز است. این بدان معنا است که در لایه خروجی مرحله آشکارسازی با توجه به چهار نقشه ویژگی (ماسک باینری متن (کلمه برش زده شده)، نقشه چرخش متن (زاویه متن)، چهار جعبه محدود کننده محور و هشت نقشه ویژگی (کانال) برای نمایش گوشه‌های مکان کادر متن در دو جهت x و y)، الگوریتم پیشنهادی برای هر کلمه آشکارسازی شده، یک پنجره مستطیل شکل را انتخاب می‌کند که شامل یک کلمه است و آن را به سیستم بازشناسی کاراکتر تحویل می‌دهد. عملگر بازشناسی متن پیشنهادی همچنین از i.inception و i.ReLU

بخش قبلی استفاده می‌نماید الگوریتم پیشنهادی می‌تواند ۵۲ حرف انگلیسی (۲۶ حرف کوچک و ۲۶ حرف بزرگ)، اعداد انگلیسی ۰ تا ۹ (۱۰ رقم) را بازشناسی نماید که در کل ۶۲ کاراکتر را شامل شود. در سیستم بازشناسی کاراکتر پیشنهادی، یک چارچوب خط لوله ایجاد می‌شود که لایه‌های موازی به‌طور هم‌زمان پردازش می‌شوند. این چارچوب باعث بهبود دقت سیستم می‌شود. چارچوب خط لوله پیشنهادی برای بازشناسی کاراکتر در شکل ۴ نشان داده شده است. در واقع الگوریتم پیشنهادی بازشناسی متناز بلوک‌های پیشنهادی i.inception و i.ReLU که در مرحله آشکارسازی نیز استفاده گردید و همچنین خروجی مرحله آشکارسازی (کلمات برش یافته و زوایای متن) جهت بازشناسی متن استفاده می‌نماید. این چارچوب شامل ۲ خط لوله موازی است. چارچوب خط لوله برای بازشناسی کاراکتر عبارتند از: خط لوله اول، متشکل از کلمات برش یافته (نقشه باینری) و خط لوله دوم از زوایای متن تشکیل شده است که هر دو خروجی مرحله آشکارسازی متن است. خروجی‌های کلمات برش یافته و زاویه متن برای استخراج ویژگی‌های مناسب به‌صورت جداگانه به بلوک i.ReLU تغذیه می‌شوند. پس از بررسی، دریافتیم که لایهi.inception پس از لایه i.ReLU منجر به شناخت بهتر و دقیق‌تر کاراکترهای منحنی می‌شود. بنابراین، لایه i.inception پس از هر لایه i.ReLU قرار داده شده است. این دو خط لوله موازی هم‌زمان روند و کار خود را انجام می‌دهند. سرانجام، خروجی دو خط لوله موازی، با یکدیگر جمع می‌شود. همچنین برای کاهش تأثیر بیش برازش[۲۸] با تغییر روند برازش[۲۹]، یک لایه حذف تصادفی[۳۰] با احتمال حذف تصادفی ۴۰٪ در معماری پیشنهادی CNN گنجانیده شده است [۴۹].

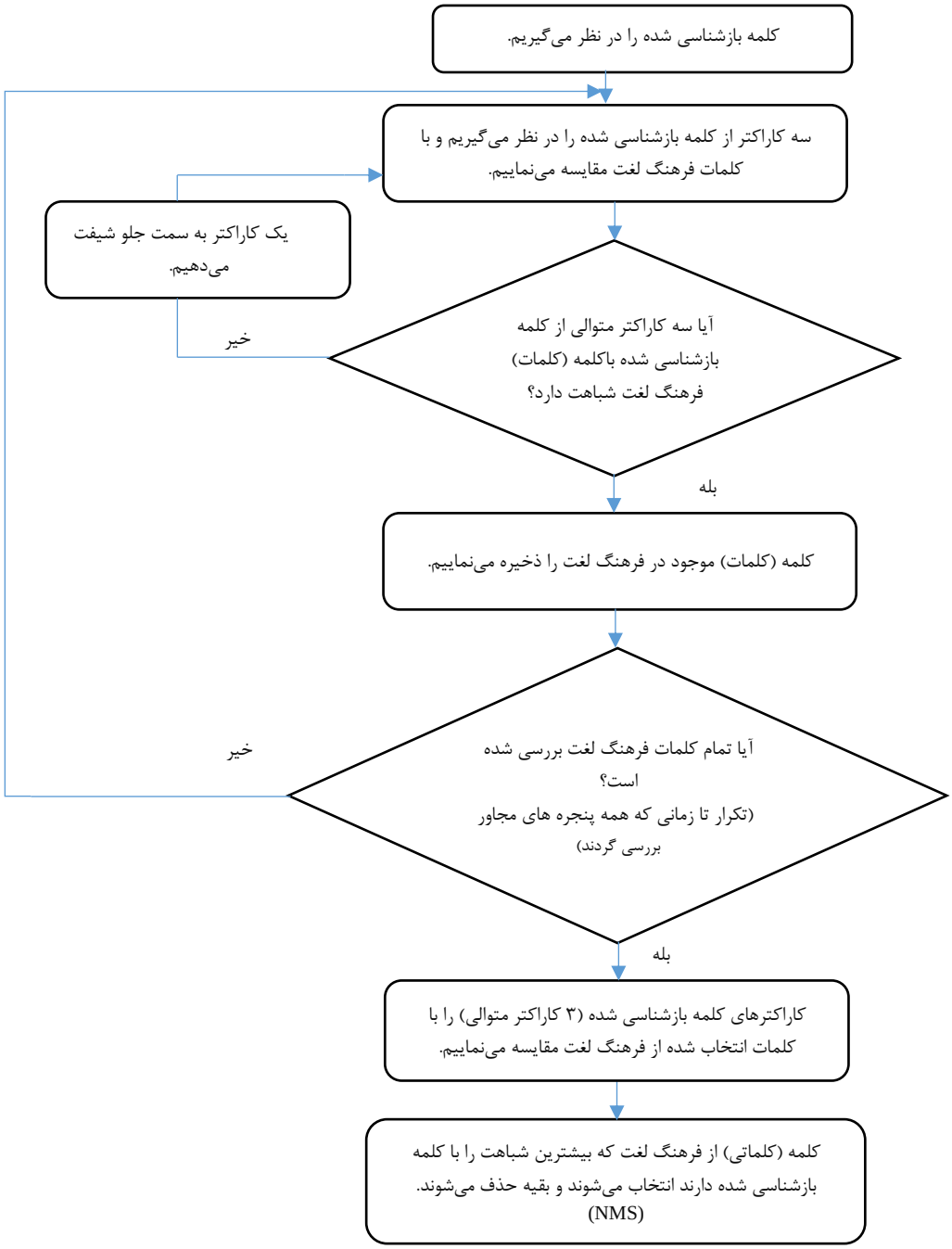


شکل ۴: خط لوله پیشنهادی برای بازشناسی کاراکتر

۳-۳ به‌کارگیری فرهنگ لغت جهت اصلاح خطای احتمالی کلمات بازشناسی شده

همان‌طور که در [۹، ۱۹ و ۵۰] اشاره شده است پس از مرحله آ شکار سازی و باز شنا سی متن، در مرحله بازیابی از یک فرهنگ لغت پویا جهت تصحیح خطاهای مرحله بازشناسی استفاده نموده‌اند. بازیابی شامل استخراج بهترین نامزدهای استخراج شده از فرهنگ لغت می با شد. در این پژوهش، برای باز شنا سی بهتر کلمات، یک فرهنگ لغت متشکل از پایگاه‌های ICDAR 2013 [۵۱]، ICDAR 2015 [۵۲] و ICDAR 2019 [۵۳] تشکیل می‌دهیم. سپس هر کلمه بازشناسی شده را در نظر می‌گیریم و سه کاراکتر اول از کلمه بازشناسی شده را با کلمات فرهنگ لغت مقایسه می‌نماییم. اگر کلماتی موجود بود که سه کاراکتر متوالی آن‌ها با کلمه باز شنا سی شده شباهت دا شت آن‌ها را انتخاب می‌نماییم، در غیر این صورت یک کاراکتر از کلمه بازشناسی شده را به سمت جلو شیفت می‌دهیم، مجددا سه کاراکتر متوالی از کلمه باز شنا سی شده را با کلمات فرهنگ لغت مقایسه می‌نماییم و این

روال را تکرار می‌نماییم (۱ کاراکتر ۱ کاراکتر به سمت جلو شیفت می‌دهیم تا به آخرین کاراکتر کلمه بازشناسی شده برسیم). زمانیکه سه کاراکتر متوالی از کلمه باز شنا سی شده با کلمه (کلمات) فرهنگ لغت شباهت داشت،کلمه (کلمات) موجود در فرهنگ لغت را ذخیره می‌نماییم. لازم به ذکر است تمامی کلمات فرهنگ لغت باید بررسی گردند. حال از طریق الگوریتم سرکوب غیر حداکثری[۳۱] کاراکترهای کلمه بازشناسی‌شده (۳ کاراکتر متوالی) را با کلمات انتخاب شده از فرهنگ لغت مقایسه می‌نماییم و کلمه‌ای که بیشترین شباهت را با کلمه بازشناسی شده دارد در نظر می‌گیریم و بقیه کلمات فرهنگ لغت را حذف می‌نماییم. در شکل ۵، فلوچارت مقایسه کلمات بازشناسی شده با کلمات موجود در فرهنگ لغت نشان داده شده است. این مرحله باعث می‌گردد تا اگر در مرحله بازشناسی کاراکتری حذف شده است و یا اشتباه تشخیص داده شده است اصلاح گردد.



شکل ۵: فلوچارت مقایسه کلمات بازشناسی شده با کلمات موجود در فرهنگ لغت

۴- ارزیابی روش پیشنهادی

در مقاله حاضر هدف استخراج و آشکارسازی متن در تصاویر طبیعی است. برای نشان دادن کارآیی ساختار پیشنهادی، از پایگاه‌های داده ICDAR 2013 [۵۱]، ICDAR 2015 [۵۲] و ICDAR 2019 [۵۳] استفاده شده است. این پایگاه‌ها در بسیاری از مطالعات تحقیق متنی صحنه اخیر مورد استناد قرار گرفته‌اند. در ICDAR 2013، تمام نمونه‌های متن به‌صورت افقی هستند و ۲۲۹ تصویر طبیعی برای آموزش و ۲۳۳ تصویر طبیعی برای آزمون وجود دارد. این مجموعه داده از آدرس https://rrc.cvc.uab.es/?ch=2&com=downloads قابل بارگذاری است. مجموعه داده ICDAR 2015 شامل ۱۰۰۰ تصویر آموزش و ۵۰۰ تصویر آزمون برای محلی سازی مقاوم متن است و در https://rrc.cvc.uab.es/?ch=4&com=downloads در دسترس است. مجموعه داده سوم ICDAR 2019 است که از ۱۰۰۰ تصویر آموزش و ۱۰۰۰ تصویر آزمون تشکیل شده است و در https://rrc.cvc.uab.es/?ch=15 در دسترس است. استحکام و توانایی بالای سیستم آشکارسازی و بازشناسی متن پیشنهادی (بازشناسی یکپارچه متن) در شکل ۶ نشان داده شده است. در این شکل حتی تصاویری که در آن‌ها متن‌های بسیار مبهم وجود دارد، مکان و جهت یابی آن‌ها توسط الگوریتم پیشنهادی با موفقیت انجام شده است. آزمایش‌ها و نتایج در این مطالعه با استفاده از Python (Keras) بر روی رایانه Intel Core I7 ۳،۲ گیگاهرتز تحت سیستم ۶۴ گیگابایتی و NVIDIA(R)، تسلا (TM) و 40KGPU انجام شد. به منظور ارزیابی عملکرد و استحکام روش پیشنهادی، آن را با هفت روش آشکارسازی متن اخیر و پنج روش بازشناسی متن اخیر مقایسه کردیم. ما از معیار امتیاز F1 به‌عنوان اندازه‌گیری دقت سیستم محلی سازی (آشکارسازی) و بازشناسی متن در مقایسه با روش‌های اخیر استفاده نمودیم. زیرا این معیار یک معامله بین نتیجه نرخ فراخوان و دقت است [۲۱]. برای مقایسه روش پیشنهادی با مقالات اخیر در زمینه آشکارسازی و بازشناسی متن، پارامترهای دقت[۳۲] و فراخوانی[۳۳] و معیار امتیاز F1 هر یک را در نظر گرفتیم. برتری مقدار پارامترهای دقت و فراخوانی و معیار امتیاز F1 در روش پیشنهادی توانایی بالای سیستم عملکرد آشکارسازی متن را نشان می‌دهد. این نتایج رضایت‌بخش از آشکارسازی و بازشناسی متن به ترتیب در جداول ۱ و ۲ برای پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019 ارائه شده است. برای هر شاخص در جداول، بالاترین مقادیر یادآوری، دقت و نمره F1 پر رنگ شده است. این نتایج اثبات می‌کند که چارچوب پیشنهادی حتی در تصاویر با متون تار، متون منحنی یا جهت‌های مختلف متن در صحنه‌های واقعی، عملکرد آشکارسازی و بازشناسی متن را به میزان قابل‌توجهی بهبود می‌بخشد. از نتایج نشان داده شده در جدول ۱ پایگاه داده ICDAR 2013، می‌توان دریافت که چارچوب پیشنهادی در دقت و فراخوان بهتر عمل می‌کند. الگوریتم پیاده‌سازی شده در [۳۴] قادر به آشکارسازی متون عمودی و منحنی نیست و باعث کاهش فراخوان می‌شود. علاوه بر این، بین استراتژی‌های گزارش شده در [۳۲، ۳۳] اختلاف کمی در مقدار دقت وجود دارد، اما تفاوت زیادی در مقدار فراخوان وجود دارد. همچنین، جدول ۱ (برای مجموعه داده ICDAR 2015) نشان می‌دهد که بدترین نتیجه توسط شبکه متن اتصال دهنده (اجزای متصل‌به‌هم) به دست آمده است [۳۴] و خط لوله پیشنهادی ما پیشرفت چشمگیری را نسبت به سایر چارچوب‌های دیگر نشان می‌دهد. استفاده از روش پیشنهادی بهترین نتیجه را در میان سایر روش‌های جدول ۱ در مجموعه داده ICDAR 2019 به دست آورد. همچنین در جدول ۲ نشان داده شده است که ساختار پیشنهادی بهترین نتایج را برای بازشناسی متن در مقایسه با پنج روش اخیر دیگر ارائه می‌دهد.

میانگین زمان اجرای مدل پیشنهادی آشکارسازی متن در پایگاه داده ICDAR 2015 در حدود ۰،۲۰ ثانیه برای آزمون هر تصویر می‌باشد. مقایسه زمان اجرای مدل پیشنهادی با سایر روش‌ها در جدول ۳ نشان داده شده است (محاسبه زمان اجرا برای تمام روش‌ها با شرایط کاملا یکسان، بر روی رایانه Intel CoreI7 ۳،۲ گیگاهرتز تحت سیستم ۶۴ گیگابایتی و NVIDIA(R)، تسلا (TM) و 40KGPU انجام شد). لازم به ذکر است زمان اجرای روش ژو [۳۱] در حدود ۰،۱۸ ثانیه می‌باشد اما هیچ یک از روش‌های آشکارسازی متن اخیر، قادر به آشکارسازی و محلی‌سازی متون منحنی و عمودی نمی‌باشند.

همچنین میانگین زمان اجرا در مرحله بازشناسی در پایگاه داده ICDAR 2015 در حدود ۰،۱۱ ثانیه برای آزمون هر تصویر می‌باشد. مقایسه زمان اجرای مدل پیشنهادی با سایر روش‌ها برای مرحله بازشناسی در جدول ۴ نشان داده شده است. (محاسبه زمان اجرا برای تمامی روش‌ها با شرایط کاملایکسان، بر روی رایانه Intel Core I7 ۳،۲ گیگاهرتز تحت سیستم ۶۴ گیگابایتی و NVIDIA(R)، تسلا (TM) و 40KGPU انجام شد). روش پیشنهادی بازشناسی متن در مقایسه زمان اجرای روش بازشناسی وانگ [۴۱] در حدود ۰،۰۵ ثانیه کاهش یافته است. بنابراین نتیجه می‌گیریم میانگین زمان اجرای سیستم آشکارسازی و بازشناسی یکپارچه متن صحنه پیشنهادی در پایگاه‌های داده ICDAR 2015 در حدود ۰،۳۲ ثانیه برای آزمون هر تصویر می‌باشد.

در جدول ۵ تأثیر لایه‌های i.ReLU و i.inception و لایه اضافی روش پیشنهادی بر روی نرخ فراخوانی، دقتوامتیاز F1 در پایگاه داده ICDAR 2019 نشان داده شده است.

جدول۱: مقایسه نتایج آشکارسازی متن در پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019 (دقت (P)،فراخوانی (R)و معیار امتیاز F1 ((F)

ICDAR 2019			ICDAR 2015			ICDAR 2013			
F	P	R	F	P	R	F	P	R	روش
0.4735	0.4912	0.4571	0.5001	0.5063	0.4940	0.5213	0.5248	0.5178	هوانگ [۳۴]
0.4776	0.4801	0.4752	0.5943	0.6419	0.5532	0.6760	0.7721	0.6011	وانگ [۳۵]
0.5732	0.5849	0.5619	0.7744	0.8217	0.7323	0.8002	0.9022	0.7189	ما [۸]
0.5773	0.5904	0.5647	0.8072	0.8327	0.7833	0.8215	0.9128	0.7468	ژو [۳۱]
0.6106	0.6311	0.5914	0.6085	0.7422	0.5156	0.8797	0.9311	0.8336	تیان [۳۳]
0.6495	0.6573	0.6418	0.7277	0.6820	0.7800	0.9257	0.9325	0.9190	لئو [۳۲]
0.6546	0.6591	0.6501	0.8533	0.9055	0.8068	0.9239	0.9388	0.9094	یانگ[۳۶]
0.6906	**0.7086**	**0.6735**	**0.8619**	**0.9109**	**0.8179**	**0.9304**	**0.9411**	**0.9201**	روش پیشنهادی

جدول۲: مقایسه نتایج بازشناسی متن در پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019 (دقت (P)،فراخوانی (R)و معیار امتیاز F1 ((F)

ICDAR 2019			ICDAR 2015			ICDAR 2013			
F	P	R	F	P	R	F	P	R	روش
0.6700	0.7535	0.6031	0.6642	0.7465	0.5983	0.6997	0.7758	0.6372	یائو[۹]
0.7168	0.8170	0.6385	0.7055	0.7831	0.6419	0.7281	0.8139	0.6586	اسلام [۳۹]
0.8038	0.8756	0.7429	0.8049	0.8628	0.7543	0.8198	0.8875	0.7617	شی[۵]
0.8480	0.9236	0.7839	0.8345	0.8925	0.7836	0.8360	0.9011	0.7796	ژانگ [۴۰]
0.8508	0.9215	0.7901	0.8472	0.9013	0.7993	0.8498	0.9125	0.7951	وانگ [۴۱]
0.8626	**0.9331**	**0.8021**	**0.8742**	**0.9278**	**0.8265**	**0.8618**	**0.9240**	**0.8074**	روش پیشنهادی

جدول ۳:مقایسه میانگین زمان اجرا مدل پیشنهادی آشکارسازی متن با سایر روش‌ها در پایگاه داده ICDAR 2015

ICDAR 2015	
زمان اجرا برای آزمون هر تصویر (ثانیه)	روش
0.48	هوانگ [۳۴]
0.41	وانگ [۳۵]
0.30	ما [۸]
0.18	ژو [۳۱]
0.24	تیان [۳۳]
0.23	لئو [۳۲]
0.25	یانگ[۳۶]
0.20	روش پیشنهادی

جدول ۴: مقایسه میانگین زمان اجرا مدل پیشنهادی بازشناسی متن با سایر روش‌ها در پایگاه داده ICDAR 2015

ICDAR 2015	
زمان اجرا برای آزمون هر تصویر (ثانیه)	روش
0.78	یائو[۹]
0.33	اسلام [۳۹]
0.25	شی[۵]
0.21	ژانگ [۴۰]
0.16	وانگ [۴۱]
0.11	روش پیشنهادی

شکل ۶: نمونه‌ای از محلی سازی متن (آشکارسازی) و بازشناسی توسط سیستم پیشنهادی در پایگاه داده ICDAR 2013 (ردیف اول و دوم)، ICDAR 2015 (ردیف سوم و چهارم) و ICDAR 2019 (ردیف پنجم و ششم)

جدول ۵: تأثیر لایه‌های i.ReLU و i.inception و لایه اضافی روش پیشنهادی بر روی نرخ فراخوانی، دقت و امتیاز F1 در پایگاه داده ICDAR 2019

لایه i.ReLU	لایه i.inception	لایه اضافی	فراخوانی	دقت	امتیاز F1
x	x	x	0.5794	0.5991	0.5891
√	x	x	0.6013	0.6239	0. 6124
√	√	x	0.6486	0.6626	0.6555
√	√	√	0.6735	0.7086	0.6906

یکی از مهم‌ترین پارامترهای یادگیری شبکه عصبی پیچشی CNN، میزان یادگیری (قابلیت تنظیم ابر پارامترها[۴۴]) است که باید به‌دقت مورد بررسی قرار گیرد. میزان یادگیری یا اندازه گام به ارزش وزنی که طی فرایند آموزش به روز می‌شوند، وابسته است. به عبارت دیگر، سرعت یادگیری مدل به میزان یادگیری بستگی دارد [۵۴، ۵۵]. نرخ یادگیری کوچک (اندازه گام) باعث می شود که شبکه به آهستگی و با دقت تنظیم شود (مجموعه‌ای از وزن‌های مطلوب در سطح عمومی)، در حالی که میزان یادگیری بزرگ به‌سرعت تنظیم می‌شود؛ اما ممکن است فرا جهش داشته باشد و مقدار بهینه را رد نماییم [۳۳، ۴۰ و ۴۱]. در این مطالعه، حرکت و کاهش وزن را به ترتیب ۰،۹ و $3 * 10^{-4}$ تعیین نمودیم. ابتدا مقدار اولیه میزان یادگیری را برابر با ۰،۰۰۱ قرار می‌دهیم و پس از تکرار آموزش ۴۰K، مقدار بهینه ۰،۰۰۰۲ به دست آمد.

سیستم پیشنهادی، در موارد نادر دارای خطاهای آشکارسازی است. زمینه‌های شلوغی که باعث ایجاد تصاویر شبیه به متن می‌شوند باعث می‌شود سیستم به اشتباه آن را متن آشکارسازی نماید. در شکل ۷ (ردیف اول) این خطا نشان داده شده است. همچنین اگر زاویه متن بسیار بالاتر از ۹۰ درجه باشد، اصولا سیستم قادر نیست بخشی از متن را آشکارسازی نماید. در شکل ۷ (ردیف دوم سمت چپ) بعضی از کاراکترها به دلیل زاویه چرخش غیر معمول آن‌ها، شناسایی نشده‌اند. همچنین امکان دارد متن‌هایی با فونت‌های غیرمتعارف آشکارسازی نشوند. در شکل ۷ (ردیف دوم سمت راست) قلم شکسته خاص، آشکارسازی نمی شود. به دلیل استفاده کم از این فونت، سیستم این فونت را به‌عنوان متن آموزش نمی‌بیند. با این‌حال، روش پیشنهادی دارای مزیت بسیار بالاتری نسبت به سایر روش‌های محلی سازی متن مشابه است.

شکل ۷: نمونه‌ای از خطاهای محلی سازی (آشکارسازی) و بازشناسی متن توسط سیستم پیشنهادی در پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019

۴-۱-ارزیابی به‌کارگیری فرهنگ لغت جهت اصلاح خطای احتمالی کلمات بازشناسی شده

همان‌طور که ذکر شد، برای بازشناسی بهتر کلمات، یک فرهنگ لغت متشکل از پایگاه های داده ICDAR 2013، ICDAR 2015و ICDAR 2019 تشکیل می‌دهیم.

با توجه به جدول ۶، در روش پیشنهادی تعداد کاراکترهای مشابه متوالی کلمه بازشناسی شده با کلمات موجود در فرهنگ لغت را سه در نظر گرفتیم. در صورتی که تعداد کاراکترهای متوالی را دو در نظر بگیریم، به دلیل مشابهت تعداد بسیاری از کاراکترهای کلمات موجود در فرهنگ لغت با کاراکترهای کلمه بازشناسی شده، تعداد مثبت‌های کاذب (FP) افزایش می‌یابد. همچنین در صورتی که تعداد کاراکترها را

بیشتر از سه در نظر بگیریم مقادیر مثبت‌های درست (TP) کاهش می یابد.

جدول ۶: دقت بازشناسی با توجه به مقایسه تعداد کاراکترهای کلمه بازشناسی شده با کلمات موجود در فرهنگ لغت

تعداد کاراکترهای متوالی	دقت
2	0.883
3	0.952
4	0.913

در شکل ۸، نمونه‌ای از اصلاحات کاراکترهایی که به اشتباه بازشناسی شده‌اند و یا بازشناسی نشده‌اند با استفاده از کلمات موجود در فرهنگ لغت نشان داده شده است.

همچنین پارامترهای دقت (P) و فراخوانی (R) و معیار امتیاز F1 برای بازشناسی متن روش پیشنهادی با استفاده از فرهنگ لغت، در جدول ۷ برای پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019 ارائه شده است.

متن بازشناسی شده	ACOFFECT	ARBUC
کلمه مشابه آشکارسازی شده در فرهنگ لغت	COFFEE	STARBUCKS

متن بازشناسی شده	Road	POClifton	1520	NURSING	OF	SCHOOL	OODRUP	HODGSON	NELL	MORV
کلمه مشابه آشکارسازی شده در فرهنگ لغت	Road	Clifton	1520	NURSING	OF	SCHOOL	WOODRUFF	HODGSON	NELL	EMORY

شکل ۸:نمونه‌ای از اصلاحات کلمات بازشناسی شده با استفاده از کلمات موجود در فرهنگ لغت (کاراکترهایی که زیرشان خط کشیده شده است به اشتباه بازشناسی شده‌اند که توسط فرهنگ لغت تصحیح شده است).

جدول ۷: محاسبه پارامترهای دقت (Precision) و فراخوانی (Recall) و معیار امتیاز F1 برای بازشناسی متن روش پیشنهادی با استفاده از فرهنگ لغت، در پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019

	ICDAR 2013			ICDAR 2015			ICDAR 2019		
روش	R	P	F	R	P	F	R	P	F
روش پیشنهادی با فرهنگ لغت	0.8813	0.9421	0.9138	0.9123	0.9611	0.9390	0.8813	0.9712	0.9300

۵- نتیجه‌گیری

در سال‌های اخیر از روش‌های شبکه‌های عصبی پیچشی (CNN) و یادگیری عمیق جهت آشکارسازی و بازشناسی متون صحنه استفاده شده است. روش‌های اخیر در زمینه‌های آشکارسازی و بازشناسی متن با معماری‌های مختلف از محدودیت‌های مختلف و نتایج نامطلوب در محلی سازی و بازشناسی جهت‌های مختلف متن با تغییرات کنتراست و تغییرات اندازه فونت در صحنه‌های زندگی واقعی رنج می‌برند. این اشکالات باعث عدم اطمینان در برخی از حوزه‌های کاربرد خاص، مانند ترافیک هوشمند و هوشمند سازی جدید در زیر سیستم‌های و سایل نقلیه می‌شود. همان‌طور که می‌دانیم، بازشناسی متون منحنی و عمودی با رنگ های نزدیک به هم تغییر اندازه فونت و پیش‌زمینه پیچیده، بسیار چالش‌برانگیز است. در این پژوهش ابتدا، یک سیستم مکان‌یابی متن در صحنه مقاوم چندجهته برای به دست آوردن بازدهی بالا در آشکارسازی متن بر اساس شبکه عصبی پیچشی (CNN) ارائه شده است.

در روش پیشنهادی، یک لایه ReLU بهبود یافته (i.ReLU) و یک لایه inception بهبود یافته (i.inception) معرفی شدند. در مرحله اول، ساختار پیشنهادی برای استخراج ویژگی‌های دیداری سطح پایین استفاده می‌شود. سپس از یک لایه اضافی برای بهبود استخراج ویژگی استفاده شده است. لایه‌هایi.ReLU باعث می شوند ویژگی‌های سطح پایین بیشتری استخراج شوند. لایه‌های i.inception (با پیچشی ۳×۳) می‌توانند متنی با ابعاد متنوع و متفاوت را به‌طور مؤثرتر از زنجیره‌ای خطی لایه پیچشی (بدون لایه‌هایinception) به دست آورند. خروجی لایه‌های i.ReLU و لایه‌های i.inception به لایه اضافی تغذیه می‌شوند که ساختار پیشنهادی را قادر می سازد متون چندجهته حتی منحنی و عمودی را آشکارسازی نماید. به‌طورکلی، تغییرات در i.ReLU استخراج ویژگی‌های سطح پایین را بهبود می‌بخشد و ویژگی‌های سطح پایین بیشتری از متن استخراج شوند. تغییر در i.inception باعث بهبود آشکارسازی تصاویر در اندازه‌های مختلف می‌شود. لایه‌های i.inception می‌توانند متن‌های با اندازه‌های متفاوت را به‌طور مؤثر آشکارسازی نمایند. همچنین، اضافه کردن یک لایه اضافی به لایه‌های استخراج ویژگی که لایه‌هایi.ReLU و لایه‌های i.inception در چهار نقطه به آن تغذیه می‌شوند، ساختار پیشنهادی را قادر ساخته است تا باعث استخراج ویژگی‌های بیشتر و مکان متن را در جهات مختلف حتی منحنی و عمودی را آشکارسازی نماید. ساختار پیشنهادی، زوایای مختلف متن را محاسبه و ماسک باینری موقعیت مکانی متن را به‌وسیله چرخش کادر دور متن، تولید می‌کند.

سیستم بازشناسی متن پیشنهادی نیز از i.inception و i.ReLU بخش قبلی استفاده می‌نماید. الگوریتم پیشنهادی می‌تواند ۵۲ حرف انگلیسی (۲۶ حرف کوچک و ۲۶ حرف بزرگ)، اعداد انگلیسی ۰ تا ۹ (۱۰ رقم) را بازشناسی نماید. در سیستم بازشناسی کاراکتر پیشنهادی، یک چارچوب خط لوله ایجاد می‌شود که لایه‌های موازی به‌طور هم‌زمان پردازش می شوند. این چارچوب باعث بهبود دقت سیستم می شود و شامل دو خط لوله موازی است. این دو خط لوله موازی، هم‌زمان کار خود را انجام می‌دهند. سرانجام، خروجی دو خط لوله موازی، با یکدیگر جمع می‌شود. سیستم آشکارسازی و بازشناسی متن صحنه پیشنهادی در پایگاه های داده ICDAR 2013، ICDAR2015 و ICDAR 2019 آزمایش شد و نتایج در مقایسه با مقالات مرتبط نشان از برتری روش پیشنهادی دارد. روش پیشنهادی، رویکرد بالا به پایین دارد. در این روش به جای آن‌که در ابتدا آشکارسازی و بازشناسی هر کاراکتر به‌صورت جدا انجام گردد و مبتنی بر کلمه و کاراکتر باشد، متن مستقیما از کل تصاویر ورودی تشخیص داده می‌شود. اگر کاراکتری به صورت اشتباه بازشناسی گردد در قسمت فرهنگ لغت تصحیح می‌گردد. همچنین پایگاه‌های داده استفاده شده در این پژوهش شامل کلمه و جمله هستند و معماری استفاده شده به صورت آشکارسازی و بازشناسی یکپارچه متن می باشد. بنابراین هم زمان بازشناسی انجام می گردد. همچنین میانگین زمان اجرا برای روش پیشنهادی آشکار سازی متن در پایگاه داده ICDAR 2015 در حدود ۰،۲۰ ثانیه و برای روش پیشنهادی بازشناسی متن در حدود ۰،۱۱ ثانیه برای آزمون هر تصویر می‌باشد. سپس، یک فرهنگ لغت متشکل از پایگاه‌های داده ICDAR 2013، ICDAR 2015 و ICDAR 2019 تشکیل داده و از آن جهت اصلاح خطای احتمالی کلمات بازشناسی شده استفاده نمودیم که نتایج، بهبود مطلوبی یافت.

مراجع

[1] Neumann, L. and Matas, J., 2010, November. A method for text localization and recognition in real-world images. In *Asian conference on computer vision* (pp. 770-783). Springer, Berlin, Heidelberg.

[2] Chen, J., Zhao, H., Yang, J., Zhang, J., Li, T. and Wang, K., 2017. An intelligent character recognition method to filter spam images on cloud. *Soft Computing*, *21*(3), pp.753-763.

[3] Zhu, Y., Yao, C. and Bai, X., 2016. Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science*, *10*(1), pp.19-36.

[4] Zhu, W., Lou, J., Chen, L., Xia, Q. and Ren, M., 2017. Scene text detection via extremal region based double threshold convolutional network classification. *PloS one*, *12*(8), p.e0182227.

[5] Shi, B., Wang, X., Lyu, P., Yao, C. and Bai, X., 2016. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4168-4176).

[6] Ren, X., Zhou, Y., Huang, Z., Sun, J., Yang, X. and Chen, K., 2017. A novel text structure feature extractor for Chinese scene text detection and recognition. *IEEE Access*, *5*, pp.3193-3204.

[7] Hanif, S.M. and Prevost, L., 2009, July. Text detection and localization in complex scene images using constrained adaboost algorithm. In *2009 10th international conference on document analysis and recognition* (pp. 1-5). IEEE.

[8] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y. and Xue, X., 2018. Arbitrary-oriented scene text detection via rotation

[30] Zheng, Y., Iwana, B.K. and Uchida, S., 2019. Mining the displacement of max-pooling for text recognition. *Pattern Recognition*, *93*, pp.558-569.

[31] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. and Liang, J., 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551-5560).

[32] Liu, F., Chen, C., Gu, D. and Zheng, J., 2019. FTPN: scene text detection with feature pyramid based text proposal network. *IEEE Access*, *7*, pp.44219-44228.

[33] Tian, Z., Huang, W., He, T., He, P. and Qiao, Y., 2016, October. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision* (pp. 56-72). Springer, Cham.

[34] Huang, W., Qiao, Y. and Tang, X., 2014, September. Robust scene text detection with convolution neural network induced mser trees. In *European conference on computer vision* (pp. 497-511). Springer, Cham.

[35] Wang, R., Sang, N. and Gao, C., 2015. Text detection approach based on confidence map and context information. *Neurocomputing*, *157*, pp.153-165.

[36] Yang, Q., Cheng, M., Zhou, W., Chen, Y., Qiu, M., Lin, W. and Chu, W., 2018. Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection. *arXiv preprint arXiv:1805.01167*.

[37] Ghanei, S. and Faez, K., 2015. Robust localization of texts in real-world images. International Journal of Pattern Recognition and Artificial Intelligence, 29(07), p.1555012.

[38] Ghavidel, J., Ahmadyfard, A. and Zahedi, M., 2019. Natural scene text localization using edge color signature. International Journal of Nonlinear Analysis and Applications, 10(1), pp.229-237.

[39] Islam, M.R., Mondal, C., Azam, M.K. and Islam, A.S.M.J., 2016, May. Text detection and recognition using enhanced MSER detection and a novel OCR technique. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 15-20). IEEE.

[40] Zhang, Y., Nie, S., Liu, W., Xu, X., Zhang, D. and Shen, H.T., 2019. Sequence-to-sequence domain adaptation network for robust text image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2740-2749).

[41] Wang, Q., Huang, Y., Jia, W., He, X., Blumenstein, M., Lyu, S. and Lu, Y., 2020. FACLSTM: ConvLSTM with focused attention for scene text recognition. *Science China Information Sciences*, *63*(2), pp.1-14.

[42] Hong, S., Roh, B., Kim, K.H., Cheon, Y. and Park, M., 2016. PVANet: Lightweight deep neural networks for real-time object detection. *arXiv preprint arXiv:1611.08588*.

[43] Zhan, F., Zhu, H. and Lu, S., 2019. Scene text synthesis for efficient and effective deep network training. *arXiv preprint arXiv:1901.09193*.

[44] Huang, L., Yang, Y., Deng, Y. and Yu, Y., 2015. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*.

[45] Kim, K.H., Hong, S., Roh, B., Cheon, Y. and Park, M., 2016. Pvanet: Deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*.

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

[47] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[48] Bissacco, A., Cummins, M., Netzer, Y. and Neven, H., 2013. Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the ieee international conference on computer vision* (pp. 785-792).

proposals. *IEEE Transactions on Multimedia*, *20*(11), pp.3111-3122.

[9] Yao, C., Bai, X. and Liu, W., 2014. A unified framework for multioriented text detection and recognition. *IEEE Transactions on Image Processing*, *23*(11), pp.4737-4749.

[10] Liao, M., Shi, B. and Bai, X., 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, *27*(8), pp.3676-3690.

[11] Naiemi, F., Ghods, V. and Khalesi, H., 2019. An efficient character recognition method using enhanced HOG for spam image detection. *Soft Computing*, *23*(22), pp.11759-11774.

[12] Ye, Q. and Doermann, D., 2014. Text detection and recognition in imagery: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *37*(7), pp.1480-1500.

[13] Cho, H., Sung, M. and Jun, B., 2016. Canny text detector: Fast and robust scene text localization algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3566-3573).

[14] Epshtein, B., Ofek, E. and Wexler, Y., 2010, June. Detecting text in natural scenes with stroke width transform. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2963-2970). IEEE.

[15] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A., 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

[16] Wang, T., Wu, D.J., Coates, A. and Ng, A.Y., 2012, November. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)* (pp. 3304-3308). IEEE.

[17] Jaderberg, M., Vedaldi, A. and Zisserman, A., 2014, September. Deep features for text spotting. In *European conference on computer vision* (pp. 512-528). Springer, Cham.

[18] Vasilopoulos, N. and Kavallieratou, E., 2017. Unified layout analysis and text localization framework. *Journal of Electronic Imaging*, *26*(1), p.013009.

[19] Neumann, L. and Matas, J., 2015. Real-time lexicon-free scene text localization and recognition. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), pp.1872-1885.

[20] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A., 2014. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*.

[21] Jeong, M. and Jo, K.H., 2015, January. Multi language text detection using fast stroke width transform. In *2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)* (pp. 1-4). IEEE.

[22] Ye, Q., Huang, Q., Gao, W. and Zhao, D., 2005. Fast and robust text detection in images and video frames. *Image and vision computing*, *23*(6), pp.565-576.

[23] Pan, Y.F., Hou, X. and Liu, C.L., 2010. A hybrid approach to detect and localize texts in natural scene images. *IEEE transactions on image processing*, *20*(3), pp.800-813.

[24] Jain, A.K. and Yu, B., 1998. Automatic text location in images and video frames. *Pattern recognition*, *31*(12), pp.2055-2076.

[25] Koo, H.I. and Kim, D.H., 2013. Scene text detection via connected component clustering and nontext filtering. *IEEE transactions on image processing*, *22*(6), pp.2296-2305.

[26] Yao, C., Bai, X., Liu, W., Ma, Y. and Tu, Z., 2012, June. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1083-1090). IEEE.

[27] Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W., 2017, February. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

[28] Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P. and Luo, Z., 2017. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*.

[29] Luo, C., Jin, L. and Sun, Z., 2019. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, *90*, pp.109-118.

[49] Amin, K.M., Shahin, A.I. and Guo, Y., 2016. A novel breast tumor classification algorithm using neutrosophic score features. *Measurement, 81*, pp.210-220.

[50] Jemni, S.K., Kessentini, Y. and Kanoun, S., 2019. Out of vocabulary word detection and recovery in Arabic handwritten text recognition. *Pattern Recognition, 93*, pp.507-520.

[51] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A. and De Las Heras, L.P., 2013, August. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1484-1493). IEEE.

[52] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S. and Shafait, F., 2015, August. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1156-1160). IEEE.

[53] Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C.V., Valveny, E. and Karatzas, D., 2019, September. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1563-1570). IEEE.

[54] Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (pp. 437-478). Springer, Berlin, Heidelberg.

[55] Breuel, T.M., 2015. The effects of hyperparameters on SGD training of neural networks. *arXiv preprint arXiv:1508.02788*.

زیرنویس‌ها

[1] multi-scale
[2] convolutional neural network
[3] Rectified linear unit
[4] convolutional neural network
[5] improved ReLU layer
[6] improved inception layer
[7] Stroke width transform
[8] maximally stable extremal region
[9] text lines
[10] optical character recognition
[11] efficient and accurate scene text detector
[12] feature pyramid-based text proposal network
[13] connectionist text proposal network
[14] rotation region proposal networks
[15] Sequence-to-sequence domain adaptation network
[16] robust text recognizer with automatic rectification
[17] Offline
[18] Online
[19] feature-merging layer
[20] concatenation layer
[21] intermediate activation patterns
[22] correlated filters
[23] forces multiplicity
[24] linear chain of convolution layer
[25] reference lengthhref_i
[26] score map
[27] four axis-aligned bounding box
[28] overfitting
[29] fitting
[30] dropout layer
[31] non-maximum suppression (NMS)
[32] Precision
[33] Recall
[34] hyperparameter