Text Recognition in Printed Persian Documents Based on Recurrent Neural Networks

Azadeh Fakhrzadeh*
PhD in Digital Image Processing; Uppsala University; Assistant Professor; Information Technology Research Department; Iranian Research Institute for Information Science & Technology (IranDoc); Tehran, Iran Email: fakhrzadeh@irandoc.ac.ir

Amir Hossein Seddighi
PhD in Industrial engineering; Amirkabir University of Technology; Assistant professor; Information Technology Research Department; Iranian Research Institute for Information Science & Technology (IranDoc); Tehran, Iran Email: seddighi@irandoc.ac.ir

Mohammad Eshratabadi
B.Sc. in Computer Engineering; Amirkabir University of Technology; Research assistant; Information Technology Research Department; Iranian Research Institute for Information Science & Technology (IranDoc); Tehran, Iran;
Email: m.e.abadi@aut.ac.ir

Alborz Esfandyari
M.Sc. in Software Engineering; Isfahan University of Technology (IUT); Research assistant; Information Technology Research Department; Iranian Research Institute for Information Science & Technology (IranDoc); Tehran, Iran;
Email: alborz.esfandyari@alumni.iut.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 40 | No. 4 | pp. 1283-1306
Summer 2025
https://doi.org/10.22034/jipm.2025.2052358.1926

Received: 02, Feb. 2025 **Accepted: 19, May. 2025**

Abstract: Automatic Persian text recognition has always been challenging due to the unique characteristics of the Persian script, including its connected structure, the high visual similarity between letters, and the significant variation in the shape of letters depending on their position within a word. The aim of this research is to develop an optical character recognition (OCR) model capable of converting Persian printed and scientific documents, including theses, articles, and books, into editable texts. Such a model is essential for tasks like labeling, indexing, and information retrieval in databases. This paper proposes a hybrid approach based on deep learning architectures for Persian text recognition. In this method, convolutional neural networks (CNNs) are used for feature extraction and recurrent neural networks (RNNs) for word recognition. The main advantage of this model is its ability to directly recognize Persian

* Corresponding Author

printed text without relying on complex preprocessing steps, such as letter segmentation. The proposed model is trained on a large and dedicated dataset, comprising over two million samples generated in five common Persian fonts. The model achieves an accuracy of 81 per cent in recognizing Persian letters and 60 per cent in recognizing words. The most common errors occur in words related to semi-spaces and signs.

Keywords: Optical Character Recognition, Long Short-Term Memory, Recurrent Neural Network, Convolutional Neural Network

تشخیص متن در اسناد فارسی چاپی بر اساس شبکه‌های عصبی بازگشتی

آزاده فخرزاده
دکتری تخصصی پردازش تصویر کامپیوتری؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط fakhrzadeh@irandoc.ac.ir

امیرحسین صدیقی
دکتری تخصصی مهندسی صنایع؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛ تهران، ایران؛
seddighi@irandoc.ac.ir

محمد عشرت‌آبادی
کارشناسی کامپیوتر؛ دستیار پژوهشی؛ پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛ تهران، ایران؛
m.e.abadi@aut.ac.ir

البرز اسفندیاری
کارشناسی ارشد نرم‌افزار؛ دستیار پژوهشی؛ پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛ تهران، ایران؛
alborz.esfandyari@alumni.iut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۳/۱۱/۱۴ **پذیرش:** ۱۴۰۴/۰۲/۲۹ مقاله برای اصلاح به مدت ۱۸ روز نزد پدیدآوران بوده است.

چکیده: تشخیص خودکار متن فارسی به‌دلیل ویژگی‌های یکتای خط فارسی از جمله ساختار پیوسته، اشتراک بالای ویژگی‌های بصری بین حروف، و تنوع بالای نوشتاری حروف با توجه به موقعیت آنان در کلمه همواره چالش‌برانگیز بوده است. هدف این پژوهش ارائه یک مدل نویسه‌خوانی نوری است که بتواند اسناد چاپی و علمی فارسی را که شامل پایان‌نامه‌ها، مقالات و کتب فارسی است، به متن قابل ویرایش تبدیل کند. این امر برای برچسب‌گذاری، فهرست‌بندی و بازیابی اطلاعات در پایگاه داده‌ها یک ضرورت محسوب می‌شود. این مقاله رویکردی ترکیبی مبتنی بر معماری‌های یادگیری عمیق برای تشخیص متن فارسی ارائه می‌دهد. در این روش از شبکه‌های عصبی پیچشی برای استخراج ویژگی‌ها و از شبکه‌های عصبی بازگشتی برای تشخیص کلمات استفاده می‌شود. مزیت اصلی این مدل، توانایی آن در تشخیص مستقیم متن چاپی فارسی بدون نیاز به پیش‌پردازش‌های پیچیده مانند ناحیه‌بندی حروف است. مدل پیشنهادی با استفاده از یک مجموعه داده اختصاصی و بزرگ، شامل بیش از دو میلیون نمونه که با پنج فونت متداول فارسی تولید شده، آموزش داده

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۴۰ | شمارهٔ ۴ | ص‌ص ۱۲۸۳-۱۳۰۶
تابستان ۱۴۰۴
https://doi.org/10.22034/jipm.2025.2052358.1926

شده است. مدل معرفی‌شده دقت ۸۱ درصد در تشخیص حروف فارسی و ۶۰ درصد در تشخیص کلمات دارد. عمده‌ترین خطاها در کلمات مرتبط با نیم‌فاصله و علائم بود

کلیدواژه‌ها: تشخیص کاراکتر نوری، حافظه طولانی کوتاه‌مدت، شبکه عصبی بازگشتی، شبکه عصبی پیچشی

۱. مقدمه

در دنیای امروز بیشتر اطلاعات به‌صورت دیجیتال ایجاد می‌شود و مورد بهره‌برداری قرار می‌گیرد. بیشتر مقالات، کتاب‌ها، و پایان‌نامه‌ها به‌صورت تصویر یا به‌صورت فرمت متن قابل حمل[1] در کتابخانه‌های دیجیتال و در شبکه اینترنت ذخیره شده‌اند. تصویر و یا فایل PDF، توسط کامپیوتر قابل ویرایش نیست. نویسه‌خوانی نوری[2]، فرایند استخراج کلمات و کاراکترها از عکس است که سند را تبدیل به متن قابل ویرایش می‌کند. نویسه‌خوانی نوری یکی از روش‌های تحقیقاتی مهم و پرکاربرد در حوزه بینایی کامپیوتر است (Moudgil, Singh & Gautam 2022; Avyodri, Lukas & Tjahyadi 2022; Raj & Kos 2022; Radwan, Khalil and Abbas 2018)

مراحل نویسه‌خوانی نوری در شکل ۱، نمایش داده شده و در ادامه، به‌شرح آن می‌پردازیم. در مرحله اول که پیش‌پردازش نام دارد، به‌منظور بهبود کیفیت تصاویر، مجموعه‌ای از روش‌ها بر روی تصاویر اعمال می‌شود. روش‌هایی همچون حذف اختلالات و تصحیح کجی در این مرحله انجام می‌شود

در ناحیه‌بندی، در مورد تصاویر ترکیبی، ابتدا ناحیه مربوط به داده متنی تشخیص داده می‌شود. سپس، در ناحیه متنی خطوط و کلمات ناحیه‌بندی و شناسایی می‌شوند تا حروف و کلمات به‌دست آیند. آنگاه، ویژگی‌هایی استخراج می‌شوند که به شناسایی حروف و کلمات کمک می‌کنند. در ادامه، در مرحله دسته‌بندی، پیش‌بینی نگاشت بین ویژگی‌ها و کاراکتر متناظر صورت می‌گیرد و سرانجام، در مرحله پس‌پردازش، با استفاده از فرهنگ لغت و یا مدل‌های آماری کلمات صحیح تشخیص داده می‌شوند

1. portable document format (PDF)
2. optical character recognition (OCR)

شکل ۱. مراحل نویسه‌خوان نوری

روش‌های نویسه‌خوانی نوری در زبان لاتین با پیشرفت‌های به‌سزایی روبه‌رو شده‌اند. این روش‌ها، مبتنی بر پردازش تصویر هستند. بنابراین، شکل حروف و کلمات در به‌کارگیری آن‌ها تأثیرگذار است. در برخی از زبان‌ها مانند زبان فارسی و یا عربی که دارای خط شکسته و نویسه‌های متصل هستند، روش‌های تشخیص حروف در آن‌ها نسبت به روش‌های تشخیص حروف لاتین با چالش بیشتری همراه است. بعضی از چالش‌های زبان فارسی (Zand, Naghsh Nilchi and Monadjemi 2008; Khosravi, and Kabir 2009; Khosrobeigi et al. 2020) عبارت‌اند از

- بعضی از حروف در فارسی از نظر ساختاری کاملاً شبیه به هم هستند و تفاوت جزیی دارند؛ مثال: حرف «گ» و «ک» که فقط در یک سرکش با هم فرق دارند و گاهی با هم اشتباه گرفته می‌شوند؛
- در زبان فارسی حروفی وجود دارند که از نظر ساختاری مشابه هستند و بر اساس تعداد نقطه تفاوت پیدا می‌کنند. مثال: حروف «ت»، «ث» و «ب» و یا «ز» و «ژ» که ممکن است با هم اشتباه گرفته شوند؛
- تشخیص دندانه‌ها در کلمه و جداسازی آن‌ها در زبان فارسی سخت است؛
- یک حرف با توجه به موقعیتش ممکن است به‌صورت‌های متفاوت ظاهر شود. به‌طور مثال: حرف «ه» به چهار صورت (در واژه‌های هدایت، مهتاب، شکوفه، و ساده) ظاهر می‌شود؛
- بعضی کلمات در فارسی از چند زیرکلمه[1] تشکیل شده‌اند. زیرکلمه در واقع، بخش‌های به‌هم‌پیوسته یک کلمه است؛ به‌عنوان نمونه، کلمه «فارسی» شامل سه زیرکلمه «فا»، «ر»، و «سی» است و کلمه «معلم» از هیچ زیرکلمه‌ای ساخته نشده است. این امر باعث دشواری در ناحیه‌بندی کلمات می‌شود؛

1. subword

◇ برخی از اسناد شامل ترکیبی از زبان فارسی و انگلیسی است و این امر فرایند ناحیه‌بندی را با چالش مواجه می‌سازد.

در روش‌های نویسه‌خوانی نوری ابتدایی، متن در تصویر به حروف ناحیه‌بندی می‌شد، و تصویر هر حرف به‌همراه برچسب متناظر آن در یک مجموعه داده ذخیره می‌گشت و این مجموعه داده برای آموزش مدل و یا طراحی فیلترِ[1] مناسب به‌کار می‌رفت (Mori, Suen & Yamamoto 1992; Mithe Indalkar & Divekar 2013). محققان الگوریتم‌های گوناگونی را برای بهبود عملکرد بخش‌های تقسیم‌بندی و تشخیص سیستم OCR آزمایش کرده‌اند (Javed et. al. 2010; Hussain, & Ali, Akram Qu. Nastalique 2015). در این روش‌ها میزان همپوشانی بین الگو و ناحیه مورد نظر در تصویر با محاسبه تفاوت بین مقادیر پیکسل‌ها[2] اندازه‌گیری می‌شود. با مرور زمان، تطبیق الگو با راهکارهای دیگری مانند استخراج ویژگی، تحلیل ساختاری، تحلیل نحوی و تحلیل آماری ترکیب شده‌اند تا عملکرد و استحکام OCR بهبود یابد. امروزه با پیشرفت‌هایی که در زمینه یادگیری عمیق صورت گرفته، حوزه‌های مختلفی از جمله بینایی کامپیوتر و OCR نیز دستخوش تحول شده، و روش‌هایی با دقت بالاتر به‌وجود آمده است. تمایز روش‌های یادگیری عمیق با روش‌های اولیه در این است که فیلتر و یا الگوی مناسب برای تشخیص حروف از قبل تعیین نمی‌شود و در حین فرایند آموزش برای داده مورد نظر، یاد گرفته می‌شود. یکی از چالش‌های روش‌های یادگیریِ عمیق، داشتن یک نمونه داده با اندازه مناسب برای آموزش است. از جمله روش‌های یادگیری عمیق که در OCR به‌کار رفته، می‌توان به روش‌هایی همچون شبکه‌های عصبی پیچشی[3]، شبکه‌های عصبی بازگشتی[4] و ترنسفورمرها[5] اشاره کرد. شبکه‌های CNN ویژگی‌های محلی و عمومی را از تصاویر با استفاده از فیلترهای پیچشی و عملیات ادغام استخراج می‌کنند. شبکه‌های RNN می‌توانند داده‌های متوالی را با استفاده از لایه‌های پنهان و حلقه‌های بازخورد مدل کنند. پس از استخراج ویژگی‌های حروف توسط CNN، کلمات توسط RNN تشخیص داده می‌شوند

به‌دلیل چالش‌های مطرح‌شده در زبان‌های پیوسته، روش‌های نویسه‌خوانی نوری

1. filter
2. pixel
3. convolution neural network (CNN)
4. recurrent neural network (RNN)
5. transformers

در فارسی محدود است. نویسه‌خوانی نوری در پایگاه‌های اطلاعاتی اسناد علمی فارسی کاربرد بسیار مهمی دارد. در این پایگاه‌ها تعداد زیادی از اسناد ممکن است به‌صورت PDF یا تصویر باشند. محتوای این اسناد به‌طور خودکار قابل جست‌وجو یا تحلیل نیست. با استفاده از OCR مناسب برای زبان فارسی می‌توان متن‌های این اسناد را استخراج و فهرست‌بندی کرد و در اختیار سیستم‌های هوشمند قرار داد. این امر دسترسی کاربران به اطلاعات را تسهیل می‌کند و زمینه را برای استفاده از فناوری‌های پردازش زبان طبیعی، آموزش مدل‌های زبانی و تولید دانش مبتنی بر داده فراهم می‌آورد. بنابراین، توسعه و بهبود سیستم‌های OCR برای زبان فارسی در این حوزه از اهمیت بالایی برخوردار است

در این پژوهش یک روش نویسه‌خوانی نوری معرفی می‌شود که بر روی اسناد علمی چاپی فارسی کاربرد داشته باشد. بر این اساس، ابتدا روش‌های نویسه‌خوانی نوری را که در زبان‌های پیوسته کاربرد دارند، بررسی می‌کنیم. سپس بر اساس روش‌های موجود در پیشینه، روشی معرفی می‌کنیم که متن را از فایل‌های PDF فارسی چاپی استخراج نماید. پایگاه اطلاعاتی «گنج»، یکی از بزرگ‌ترین پایگاه‌های اطلاعاتی اسناد علمی به زبان فارسی است که شامل پایان‌نامه‌ها و رساله‌ها (پارساها)ی فارسی دانشگاه‌های کشور می‌شود. دقت روش‌های معرفی‌شده بر روی یک نمونه داده از مجموعه داده پایگاه اطلاعاتی «گنج» بررسی شده است. همچنین برای آموزش سیستم نویسه‌خوانی نوری یک داده آموزشی فارسی ساخته شد. این مجموعه داده شامل دو میلیون تصویر و متن فارسی متناظر با آن است که می‌تواند در تحقیقات این حوزه مورد استفاده قرار گیرد

۲. پیشینه پژوهش

روش‌های نویسه‌خوانی نوری مبتنی بر پردازش تصویر است؛ به همین دلیل شکل کلمات و حروف در طراحی روش مؤثر خواهد بود. «پنگ» و همکاران مجموعه‌ای از روش‌ها و تکنیک‌های اولیه را که در OCR چندزبانه استفاده می‌شود، از جمله پیش‌پردازش، تبدیل به تصویر دوتایی[1]، ناحیه‌بندی، استخراج ویژگی و شناسایی بررسی کرده‌اند (Peng et al. 2013). آن‌ها در ناحیه‌بندی از سه رویکرد مختلف از جزء به کل، از کل به جزء، و ترکیبی بهره گرفته‌اند که شامل ناحیه‌بندی صفحه، خط و کاراکتر می‌شود. این تحقیق

1. binary

همچنین چالش‌ها و محدودیت‌های سیستم‌های OCR چندزبانه را بررسی می‌کند. «حمیدا» و همکاران یک توصیفگر ویژگی کارآمد برای بهبود تشخیص کلمات دست‌نویس عربی ارائه کرده‌اند. این رویکرد از سه ویژگی تصویر، گرادیان جهت‌دار هیستوگرام[1]، فیلتر گابور[2] و الگوی باینری محلی[3] برای استخراج ویژگی استفاده می‌کند و یک الگوریتم بازشناخت الگو مبتنی بر نزدیک‌ترین همسایگی[4] را برای ساخت مدل‌ها آموزش می‌دهد (Hamida et al. 2022).

با پیشرفت در زمینه یادگیری عمیق و نتایج امیدبخش آن در بینایی کامپیوتر، به‌کارگیری روش‌های مبتنی بر یادگیری عمیق در حوزه OCR نیز قوت گرفته است. نخستین‌بار، «شی، بای و یائو» روشی بر اساس یادگیری عمیق معرفی نمودند که در آن به‌طور مستقیم کلمات موجود در متن برای آموزش مورد استفاده قرار می‌گرفت و لزومی بر ناحیه‌بندی حروف و کلمات نبود (Shi, Bai and Yao 2017). مدل پیشنهادی آن‌ها با عنوان شبکه عصبی بازگشتی پیچشی[5]، ترکیبی از شبکه عصبی پیچشی و شبکه عصبی بازگشتی ارائه داد که در آن نیازی به برچسب نظیربه‌نظیر حروف وجود نداشت.

در مطالعه‌ای که توسط «ناز» و همکاران انجام شد، ادغام شبکه پیچشی و شبکه عصبی تکرارپذیر برای تشخیص حروف نستعلیق در زبان اردو به‌کار برده شد و نتایج قابل قبولی به‌دست آمد (Naz et al. 2017). در مطالعه‌ای دیگر، «جیان» و همکاران، با رویکردی مشابه، با در نظر گرفتن فونت‌های متفاوت و نسخه‌های چاپی اسکن‌شده در زبان اردو، متن را از تصویر استخراج کردند (Jain, Mathew & Jawahar 2017). «رحمتی» و همکاران یک سیستم OCR مبتنی بر CNN و شبکه حافظه طولانی کوتاه‌مدت[6] برای پردازش زبان فارسی معرفی کردند و خروجی آن را برای تشخیص «لا» در صورت‌های نگارشی مختلف، و در نظر گرفتن شکل‌های مختلف اتصال حروف در فونت‌های مختلف، با استفاده از روش‌های پردازش تصویر بهبود دادند (Rahmati et al. 2020). «اسدی زیدآبادی» و همکاران یک مجموعه داده فارسی برای تشخیص نویسه‌خوانی نوری معرفی کردند و یک

1. histogram of oriented gradients (HOG)
2. Gabor filter (GF)
3. local binary pattern (LBP)
4. k-nearest neighbors (KNN)
5. convolutional recurrent neural networks (CRNN)
6. long short-term memory (LSTM)

روش مبتنی بر CRNN را بر روی این مجموعه داده به‌کار برده و دقت ۷۸ درصد را گزارش کردند (Asadi-zeydabadi et al. 2023). «خسروبیگی» و همکاران از نرم‌افزار فارسی خوانا (Mirzaee 2012) استفاده کرده، سپس خروجی آن را بر اساس یک رویکرد پس‌پردازش مبتنی بر قانون[1] بهبود دادند و دقت را از ۵۸ درصد به ۹۰ درصد بر روی مجموعه داده مورد بررسی افزایش دادند (Khosrobeigi et al. 2020). «الخوالده» از شبکه ترکیبی CNN-LSTM برای تشخیص اعداد دست‌نویس عربی استفاده کرده است. در بخش اول این پژوهش از VGG-16 و LeNet برای استخراج ویژگی‌های بصری استفاده شده و سپس در بخش دوم از LSTM بهره‌گیری شده است تا به کمک آن وابستگی‌های طولانی‌مدت میان ویژگی‌های استخراج‌شده در مرحله قبل حفظ شود (Alkhawaldeh 2020). «فاشا» و همکاران یک مدل CNN-BiLSTM را برای تشخیص متن عربی چاپ‌شده بدون ناحیه‌بندی کلمات پیشنهاد کرده‌اند. در این روش، پنج لایه CNN به‌عنوان استخراج‌کننده ویژگی به‌کار برده شده؛ در حالی که دو لایه BiLSTM برای وابستگی‌های طولانی‌مدت استفاده شده است. مجموعه داده آن‌ها شامل دو میلیون کلمه با هجده فونت متفاوت در زبان عربی است. این مطالعه توانسته است انواع مختلفی از مجموعه‌های آزمایشی را برای مدل پیشنهادی با تغییر فونت‌ها، سایز کلمات و اضافه کردن نویز مورد آزمایش قرار دهد (Fasha et al. 2020).

در سال ۲۰۱۷، مدل ترنسفورمر[2] برای تحلیل داده‌های متوالی توسط گوگل معرفی شد که از چند جهت بر شبکه‌های عصبی بازگشتی ارجحیت دارد (Vaswani et al. 2017). شبکه‌های عصبی بازگشتی، داده را به‌صورت ترتیبی پردازش می‌کنند و یک واحد حافظه به‌صورت لایه‌های پنهان در خود جای داده‌اند که اطلاعات مربوط به واحد (توکن)[3] قبلی را در خود ذخیره می‌کند. اما ترنسفورمرها به‌صورت موازی همه عناصر دنباله را پردازش می‌کنند و توکن‌های مربوط به هم در جمله، با بردار عددی نزدیک به هم متناظر می‌شوند. ترنسفورمرها از مکانیزم توجه[4] برای درک چگونگی ارتباط بخش‌های مختلف دنباله با یکدیگر استفاده می‌کنند، بدون اینکه لزوماً ترتیب توکن‌ها در جمله اهمیت داشته باشد. از ترنسفورمرها در OCR نیز استفاده شده است (Li et al. 2021; Mostafa

1. rule-based
2. transformer
3. token
4. attention

.(et al. 2021; Momeni et al. 2022

«لی» و همکاران مدل TrOCR را معرفی کردند که از معماری ترنسفورمر برای درک تصویر و تولید متن در سطح کلمه استفاده می‌کند (Li et al. 2021). آن‌ها از ساختار رمزگذار-رمزگشای[1] ترنسفورمر ساده (Vaswani et al. 2017) در TrOCR استفاده کردند. رمزگذار[2] برای به‌دست آوردن نمایش تکه‌های تصویر طراحی شده است و رمزگشا[3] برای تولید توالی قطعه کلمه با هدایت ویژگی‌های بصری و پیش‌بینی‌های قبلی است

«مصطفی» و همکاران یک روش بر اساس CNN-Transformer معرفی کرده‌اند که بر روی یک مجموعه داده از دست خط عربی آموزش داده شده است (Mostafa et al. 2021). معماری شبکه معرفی‌شده از ۱۰۱ لایه Resnet101 تشکیل شده و چهار رمزگذار و رمزگشای ترنسفورمر دارد که برای توالی چندکلمه‌ای آموزش دیده شده‌اند. خلاصه روش‌های معرفی‌شده در این بخش در جدول ۱، آمده است

اگر چه روش‌های مبتنی بر ترنسفورمرها نتایج خوبی در پردازش داده‌های تصویری و متنی در زبان‌های لاتین داشته است، اما این روش‌ها با میلیون‌ها پارامتر نیازمند مجموعه داده آموزشی بسیار عظیم و زیرساخت‌های پردازشی مناسب است. از این‌رو، در این تحقیق روشی مبتنی بر شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی پیشنهاد شده که نتایج امیدبخشی در زبان‌های پیوسته داشته است

جدول ۱. خلاصه روش‌های معرفی‌شده

مقاله	زبان	روش‌شناسی	کاربرد
Peng et al. (2013)	چندزبانه	پیش‌پردازش، ناحیه‌بندی (صفحه، خط، کاراکتر)، استخراج ویژگی و شناسایی	بررسی چالش‌های OCR چندزبانه و استفاده از رویکردهای جزءبه‌کل، کل‌به‌جزء و ترکیبی در OCR
Hamida et al. (2022)	عربی	GB + LBP + HOG+ KNN	بهبود تشخیص کلمات دست‌نویس عربی با استخراج ویژگی‌های کارآمد
Shi, Bai and Yao (2017)	چندزبانه	ترکیب CNN + RNN (بدون نیاز به ناحیه‌بندی حروف)	تشخیص مستقیم کلمات بدون نیاز به برچسب‌گذاری نظیربه‌نظیر

1. encoder-decoder
2. encoder
3. decoder

مقاله	زبان	روش‌شناسی	کاربرد
Naz et al. (2017)	اردو	ترکیب CNN + RNN	تشخیص حروف نستعلیق اردو با نتایج قابل قبول
Jain, Mathew, Jawahar (2017)	اردو	روش مشابه CRNN با در نظرگیری فونت‌های مختلف	استخراج متن از تصاویر اسکن‌شده چاپی با فونت‌های متنوع
Rahmati et al. (2020)	فارسی	CNN + LSTM + پردازش تصویر (پیش‌پردازش)	بهبود تشخیص حروف پیوسته مانند «لا» و «ه» آخر چسبان
Alkhawaldeh (2020)	هندی، عربی	ترکیب VGG-16/LeNet+ LSTM	تشخیص اعداد دست‌نویس عربی
Fasha et al. (2020)	عربی	ترکیب CNN + BiLSTM (بدون ناحیه‌بندی)	تشخیص متن چاپی عربی آموزش دیده‌شده با دو میلیون کلمه و ۱۸ فونت مختلف
Li et al. (2021)	چندزبانه	Transformer	بهره بردن از ارتباط کلمات در جمله بهترین خروجی روی زبان انگلیسی چاپی
Mostafa et al. (2021)	عربی	ResNet101 + Transformer	پردازش توالی‌های چندکلمه‌ای در دست‌نویس عربی

۳. روش پژوهش

روش پژوهش حاضر برای بررسی انواع روش‌ها و الگوریتم‌های موجود، روش مطالعات کتابخانه‌ای است. با مطالعه تحقیقات پیشین و بررسی چالش‌های مسئله پیش رو، روش جدیدی پیشنهاد می‌شود. روش‌های اخیر پیشنهادشده برای نویسه‌خوانی نوری بررسی می‌شود و بر این اساس، الگوریتم مناسب پیشنهاد می‌شود. سپس روش پیشنهادی بر روی مجموعه‌ای از داده‌ها آزمایش می‌شود تا صحت و دقت، و خطای احتمالی آن بررسی گردد و در صورت نیاز، بهبودهایی روی آن صورت گیرد. بنابراین، آماده‌سازی این مجموعه داده برای انجام آزمایش و اندازه‌گیری دقت روش نیز قسمتی از فعالیت‌های این پژوهش محسوب می‌شود

۳-۱. داده آموزشی

مجموعه داده آموزشی در این پژوهش شامل عکس و متن فارسی متناظر با آن است. این مجموعه داده از دو منبع مختلف کلمات فارسی ساخته شد که در زیر جزییات آن شرح داده شده است (جدول ۲)

◇ لیست کلمات داده آموزشیِ «پکا»[1]

لیست کلمات داده آموزشیِ «پکا» از پیکره کتاب‌های دیجیتال ایرانداک (پکا) ساخته شده است (Alayiaboozar & Hojjatpanah 2022). «پکا» از متون کتاب‌های دیجیتال «پژوهشگاه علوم و فناوری اطلاعات» ایجاد شده است. برای ساخت این پیکره، پس از طی فرایند نمونه‌گیری، ۶۵ کتاب فارسی انتخاب شد که حوزه‌های میان‌رشته‌ای کتابداری، مدیریت، زبان‌شناسی، مدیریت اطلاعات و غیره را پوشش می‌دهند. این پیکره شامل ۱۶۸۹۹۹ کلمه یکتاست که افزون بر حروف فارسی، حروف انگلیسی و چند زبان دیگر را نیز شامل می‌شود. برای ایجاد داده آموزشی «پکا»ی تمیزشده، فقط کاراکترهای فارسی «پکا» در نظر گرفته شده است و کلماتی که شامل کاراکترهای غیرفارسی بودند، حذف شده است. همچنین اعداد ۱ تا ۱۰۰۰۰ فارسی و ۱ تا ۱۰۰۰۰ انگلیسی به «پکا»ی تمیزشده اضافه شده است

◇ لیست کلمات داده آموزشی ترکیبی

داده ترکیبی شامل کلمات موجود در «پکا» و همچنین کلمات موجود در منابع «تسراکت»[2] گوگل است. «تسراکت» یک نرم‌افزار منبع باز است که در تشخیص کاراکتر نوری به‌کار می‌رود و در زبان‌های مختلف کاربرد دارد (Kay 2017). در مخزن این نرم‌افزار یک لیست از کلمات فارسی شامل ۱۳۸۹۲ کلمه ارائه شده است. این کلمات به کلمات «پکا» اضافه شده و داده آموزشی ترکیبی ساخته شد. این مجموعه شامل ۴۵۵۱۹ کلمه است

جدول ۲. مشخصات آماری مجموعه داده‌های مختلف

نام مجموعه داده	تعداد کلمات	تعداد کاراکترهای یکتا
پکا	۱۶۸۹۹۹	۲۹۷
پکا تمیزشده	۴۰۴۶۱	۶۴
ترکیبی	۴۵۵۱۹	۹۶

داده آموزشی با استفاده از لیست‌های معرفی‌شده ساخته شد. در لیست‌های ایجادشده

۱. پیکره کتاب‌های دیجیتال ایرانداک (پکا)

2. Tesseract

هر ردیف شامل یک کلمه، عدد، علامت و یا نشانه است. با انتخاب تصادفی هر ردیف و کنار هم قرار دادن آن‌ها، ۴۰۰ هزار خط به‌عنوان داده آموزشی ساخته شد. شرایط ساخته شدن این خطوط به‌شرح زیر است

- ◇ هر ردیف باید کمینه ۲۰ بار در جملات مختلف به‌کار رفته باشد؛
- ◇ خطوط، کمینه ۵۰ و بیشینه ۶۰ کاراکتر داشته باشند.

سپس، پنج فونت متداول در اسناد فارسی شامل Times New Roman، IRNazanin، IRLotus، IRMitra و IRZar در نظر گرفته شدند. به‌ازای هر یک از فونت‌ها ۴۰۰ هزار خط داده آموزشی (متن تصادفی ساخته‌شده) و تصویر متناظر با آن ایجاد شد. بدین‌ترتیب، مجموعه داده آموزشی شامل دو میلیون زوج متن تک‌خط و تصاویر آن‌ها در قالب فونت‌های مختلف فراهم گردید

۳-۲. داده آزمایشی

برای ایجاد داده آزمایشی از پایان‌نامه‌ها و رساله‌های موجود در پایگاه اطلاعاتی «گنج» بهره‌گیری شد. به‌صورت تصادفی شش فایل «ورد»[1] از حوزه‌های هنر، انسانی، فنی، علوم پایه، کشاورزی و پزشکی انتخاب شد. سپس، ۵۱ پاراگراف از این فایل‌ها در نظر گرفته شدند. در ادامه، پاراگراف‌هایی که در آن‌ها بیش از یک فونت استفاده شده است، به زیرپاراگراف‌هایی شکسته شدند تا سرانجام، هر پاراگراف به‌دست‌آمده، از نظر فونت استفاده‌شده منسجم باشد. سپس متن این پاراگراف‌ها از حروف انگلیسی و علائم یونانی مربوط به فرمول‌های ریاضی پاک‌سازی شد. همچنین لازم به ذکر است که هیچ‌گونه جدول و یا تصویری به‌عنوان پاراگراف در نظر گرفته نشده است. در انتها، پس از این مراحل، ۵۵ پاراگراف شامل ۸۰۲۹۷ کاراکتر به‌دست آمد. از آنجا که داده‌های آموزشی به شکل تک-خط در نظر گرفته شده‌اند، یک مرحله شکستن پاراگراف‌ها بر اساس علائم نگارشی نیز انجام شد؛ به شکلی که هیچ متن تک-خطی بیش از ۱۰۰ کاراکتر نباشد. پس از این مرحله از روی تک-خط‌های به‌دست‌آمده، زوج متن مرجع تک-خط و تصویر متناظر آن ایجاد شد. تعداد تک-خط‌های به‌دست‌آمده برابر با ۸۷۶ است

1. Word

۳-۳. روش پیشنهادی

روش پیشنهادی شامل سه بخش استخراج ویژگی‌های بصری، برچسب‌گذاری توالی و تشخیص حروف و رونویسی تشخیص کلمات است. ابتدا از هر تصویر، ویژگی‌های بصری استخراج می‌شود. برچسب متناظر با هر تصویر یک عبارت متنی است که ویژگی‌های داده توالی را دارد. در بخش برچسب‌گذاری از سیستم‌های بازگشتی استفاده می‌شود که متناسب با داده‌های توالی‌دار است و حروف تشخیص داده می‌شود. سپس، در بخش رونویسی به مدل اجازه داده می‌شود تا توالی‌های خروجی با طول‌های متغیر را تولید کند و کلمات را با ترکیب حروف تشخیص دهد. در ادامه، بخش‌های مختلف روش پیشنهادی معرفی شده و سرانجام، معماری آن ارائه می‌شود

۳-۳-۱. استخراج ویژگی‌های بصری

اولین قدم در روش‌های نویسه‌خوانی نوری استخراج ویژگی‌ها از تصویر است. برای استخراج ویژگی‌ها، از شبکه‌های عصبی پیچشی استفاده شده است. لایه پیچشی در این شبکه‌ها فیلترهایی را به‌صورت مکرر به ناحیه‌های ادراکی -در موقعیت‌های متفاوت ورودی- از طریق یک پنجره لغزنده اعمال می‌کند. پارامترهای فیلترهای ذکرشده متشکل از ماتریسی از وزن‌هاست که طی فرایند آموزش شبکه آموخته می‌شوند. لایه پیچشی دارای مشخصات ناحیه‌های ادراکی محلی است که می‌تواند شکل ورودی را حفظ کند؛ به‌طوری که همبستگی بین ویژگی‌های پیکسل‌های تصویر در دو جهت طول و عرض به‌طور مؤثر شناسایی شود. فیلتر یا کرنل[1] بر روی تصویر ورودی می‌لغزد و عملیات پیچشی را در پیکسل‌های متناظر اعمال می‌کند. پیچش تصویر دوبعدی $I[i,j]$ با فیلتر (کرنل) $K[i,j]$ با اندازه m در n به‌صورت زیر محاسبه می‌شود

$$X[i,j] = \sum_m \sum_n I[i-m, j-n]K[m,n] \quad (۱$$

لایه ادغام در این شبکه‌ها بلادرنگ بعد از لایه پیچشی می‌آید و دلیل اصلی استفاده از این لایه، کاهش نمونه‌برداری[2] و ابعاد به‌منظور کاهش ارتباطات لایه‌های پیچشی است، که نتیجه آن کاهش بار محاسباتی است. تابع فعال‌سازی، رابطه بین ورودی و

1. Kernel
2. down-sampling

خروجی نورون[1] را تعیین می‌کند. توابع غیرخطی[2] باعث ایجاد رابطه غیرخطی بین ورودی و خروجی می‌شوند. انتخاب مناسب یک تابع فعال‌سازی[3] غیرخطی مناسب می‌تواند عملکرد شبکه را به‌شدت بهبود بخشد. در لایه‌های پیچشی به‌طور معمول، از توابع ReLU[4] استفاده می‌شود. با اعمال شبکه عصبی پیچشی نقشه ویژگی‌ها از تصویر ورودی استخراج می‌شود. در مراحل بعدی این نقشه ویژگی‌ها به حرف و کلمه نگاشت می‌شود.

۳-۳-۲. برچسب‌گذاری توالی و تشخیص حروف

پس از استخراج ویژگی‌ها توسط شبکه عصبی پیچشی، نیاز است این ویژگی‌ها به حروف متناظر نگاشت شود. از آنجا که حروف در کلمات به‌صورت توالی هستند، شبکه‌های برگشت‌پذیر که برای تحلیل داده‌های زمانی و دنباله‌ای به‌کار می‌روند، پیشنهاد می‌شود. در این شبکه‌ها اطلاعات از یک مرحله به مرحله بعد انتقال داده می‌شود. شبکه حافظه کوتاه‌مدت طولانی (LSTM) بهبودیافته شبکه‌های بازگشتی است که مشکل محوشدگی گرادیان[5] در آن‌ها حل شده است. این شبکه مشکل محوشدگی گرادیان را با معرفی یک سلول حافظه برطرف کرده است که محفظه‌ای است که می‌تواند اطلاعات را برای مدت طولانی نگهداری کند. شبکه‌های LSTM قادر به یادگیری وابستگی‌های طولانی‌مدت در داده‌های متوالی هستند

یک واحد LSTM افزون بر حالت مخفی، به‌طور معمول، از یک سلول حافظه و سه دروازه تشکیل می‌شود: دروازه ورودی[6]، دروازه فراموشی[7]، و دروازه خروجی[8]. همان‌طور که گفته شد سلول حافظه حالت بلندمدت را نگهداری می‌کند و این دروازه‌ها هستند که جریان اطلاعات (شامل ورودی جدید، حالت بلندمدت قبلی و حالت مخفی قبلی) را کنترل می‌کنند. دروازه فراموشی با محاسبه مقداری بین ۰ و ۱ بر حسب حالت مخفی قبلی و ورودی جاری، تعیین می‌کند که کدام بخش‌های حالت بلندمدت باید حذف شوند.

1. neuron
2. nonlinear
3. activation function
4. rectified linear unit
5. vanishing gradient problem
6. input gate
7. forget gate
8. output gate

خروجی دروازه فراموشی برای ورودی $x_t \in \mathbb{R}^d$ به‌صورت زیر محاسبه می‌شود

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (۲$$

در معادله بالا، $f_t \in (0,1)^h$ بردار فعال‌سازی دروازه فراموشی، $W \in \mathbb{R}^{h\times d}, U \in \mathbb{R}^{h\times d}, b \in \mathbb{R}^{h\times d}$ وزن‌ها و یا پارامترهایی هستند که توسط شبکه یادگیری می‌شوند. متغیرهای d و h به‌ترتیب، تعداد ویژگی‌های ورودی و تعداد واحدهای مخفی است. همچنین h_{t-1} بردار حالت مخفی قبل و σ_g تابع فعال‌سازی سیگموئید[1] هستند. در ادامه، دروازه ورودی با استفاده از مکانیزمی مشابه با دروازه فراموشی، تعیین می‌کند که کدام بخش از حالت مخفی قبلی و ورودی جاری باید در حالت بلندمدت سلول اضافه شود تا حالت بلندمدت جدید برای گام بعد ایجاد شود که به‌صورت زیر محاسبه می‌شود

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (۳$$

که در آن، $i_t \in (0,1)^h$ بردار فعال‌سازی دروازه ورودی است. در ادامه، دروازه خروجی نیز تعیین می‌کند که کدام‌یک از بخش‌های حالت بلندمدت، بایستی به‌عنوان خروجی سلول، در این گام و حالت مخفی بعدی سلول، برای گام بعدی قرار داده شود

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (۴$$

که در آن، $o_t \in (0,1)^h$ بردار فعال‌سازی دروازه خروجی است. سرانجام، خروجی شبکه به‌صورت زیر محاسبه می‌شود

$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (۵$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (۶$$

$$h_t = o_t \odot \sigma_h(c_t) \quad (۷$$

در معادله بالا، $\tilde{c}_t \in (-1,1)^h$، بردار فعال‌سازی ورودی سلول، σ_c، تابع فعال‌سازی تانژانت هایپربولیک و σ_h تابع فعال‌سازی هایپربولیک، $c_t \in \mathbb{R}^h$ بردار حالت سلول است. مقادیر اولیه $c_0 = 0$ و $h_0 = 0$ و عملگر $\odot$ نشان‌دهنده ضرب عنصر به عنصر است. اندیس t نیز نشان‌دهنده گام زمانی است. حروف تشخیص داده‌شده در این مرحله با کمک لایه رونویسی به کلمات معنادار تبدیل می‌شوند.

۳-۳-۳. لایه رونویسی و تشخیص کلمه

از ترکیب حروف تشخیص داده‌شده در مرحله قبل، لزوماً کلمات معنادار ایجاد نمی‌شود.

1. sigmoid

برای حل این مشکل دو دیدگاه وجود دارد، دیدگاه اول ایجاد یک مجموعه داده آموزشی است که در آن حروف به‌طور دقیق ناحیه‌بندی شده باشد. در این صورت، خروجی شبکه بازگشتی باید به‌طور دقیق، با حروف خروجی منطبق شود. اما ایجاد چنین مجموعه داده‌ای، به‌ویژه برای متون پیوسته با چالش روبه‌رو بوده و دور از دسترس است. دیدگاه دوم ، استفاده از لایه رونویسی با تابع زیان دسته‌بندی زمانی ارتباط‌گرایانه[1] است (Graves et al. 2006). در این روش تمامی توالی‌هایی که توسط شبکه قبل از حروف تشخیص داده شده است، در نظر گرفته می‌شود و سپس توالی با احتمال بیشتر انتخاب خواهد شد. در این صورت امکان خطا برای مرحله قبل وجود خواهد داشت، اما سرانجام، از ترکیب حروف، کلماتی که احتمال معنادار بودن آن‌ها بیشتر و در داده آموزشی پرتکرار است، انتخاب خواهد شد. چنانچه ورودی لایه رونویسی توالی $y = y_1, \dots, y_T$ در نظر گرفته شود در حالی که T طول دنباله و $y_t \in \mathbb{R}^{|\mathcal{L}'^T|}$ در نظر گرفته شود، $\mathcal{L}' = \mathcal{L} \cup$ و $\mathcal{L}$ شامل همه کاراکترها و همچنین کاراکتر خالی[2] است. B توالی $\pi \in \mathcal{L}'^T$ را به توالی l با حذف کاراکترهای خالی و تکراری نگاشت می‌کند. در نتیجه، l^* به‌عنوان خروجی به‌صورت زیر محاسبه می‌شود

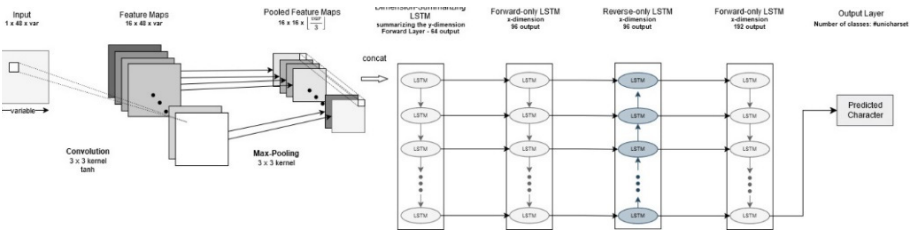
$$l^* = \mathrm{B}\left(argmax_\pi p(\pi|y)\right) \quad (۸$$

۳-۳-٤. معماری شبکه

معماری پایه شبکه پیشنهادشده تا قبل از لایه رونویسی، در شکل ۲، نمایش داده شده است. ورودی شبکه یک تصویر سیاه‌وسفید با عرض ۴۸ و طول متغیر است. عکس ورودی باید تنها یک خط را دربربگیرید. تشخیص در مرحله اول به‌صورت کاراکتری است و سپس در لایه آخر و با اعمال CTC به کلمه معنادار تبدیل می‌شود. در روش پایه، ابتدا یک لایه پیچشی شامل ۱۶ فیلتر (۳ در ۳) با تابع فعال‌سازی هایپربولیک[3] اعمال شده و ویژگی‌های تصویر استخراج می‌شود. بعد از آن لایه Max-pooling با سایز (۳ در ۳) اعمال می‌شود که اندازه نقشه ویژگی‌ها را نصف می‌کند. بعد از استخراج ویژگی، یک لایه LSTM جلوسو در بعد y با ۶۴ خروجی یک LSTM جلوسو در بعد x با ۹۶ خروجی و یک LSTM معکوس (عقب‌سو) در بعد x با ۹۶ خروجی و یک LSTM در بعد x با ۱۹۲ خروجی اعمال می‌شود. کاراکترهای تشخیص داده‌شده وارد لایه خروجی می‌شوند و با تابع زیان CTC و

1. connectionist temporal lassification (CTC)
2. blank
3. hyperbolique

softmax کلمات تشخیص داده می‌شود. در معماری عمیق‌تر لایه LSTM در بعد x با ۱۹۲ خروجی با لایه LSTM در بعد x با ۲۵۶ خروجی جایگزین می‌شود. وزن پنج لایه اول از بهترین مدل انتقال داده می‌شود و وزن دو لایه آخر از ابتدا آموزش داده می‌شود



شکل ۲. معماری شبکه پیشنهادی

٤. یافته‌ها

برای پیاده‌سازی روش معرفی‌شده از کتابخانه‌های «تسراکت» گوگل استفاده شد. زیرساخت مورد نیاز برای پیاده‌سازی این روش شامل یک پردازنده Intel Core i7-13700K با ۱۶ هسته و ۳۲ گیگابایت حافظه است. داده آموزشی و آزمایشی مطابق آنچه پیشتر توضیح داده شد، ساخته شده است. داده آموزشی که از «پیکره کتاب‌های دیجیتال ایرانداک» ساخته‌شده و به‌اختصار «پکا» نامیده می‌شود، و پیکره ترکیبی شامل داده‌هایی است که از ترکیب کلمات موجود در منابع «تسراکت» و «پکا»ی تمیزشده ساخته شده است. شبکه پایه در شکل ۲، نشان داده شده است. شبکه‌ها به سه روش آموزش از ابتدا، تنظیم دقیق، و آموزش انتقالی آموزش داده شده‌اند. در آموزش از ابتدا، شبکه از ابتدا آموزش داده می‌شود. در تنظیم دقیق، وزن‌های شبکه پایه از منابع «تسراکت» به‌عنوان وزن‌های اولیه در نظر گرفته می‌شود و سپس با داده‌های جدید آموزش شروع می‌شود. در آموزش انتقالی، وزن لایه‌های اول از منابع «تسراکت» انتقال داده می‌شود. در مدل عمیق‌تر، وزن پنج لایه اول مدل انتقالی است و دو لایه آخر از ابتدا آموزش می‌بینند. از دو نرخ خطای مرسوم در روش‌های نویسه‌خوانی نوری، نرخ خطای کاراکتری[1] و نرخ خطای کلمه‌ای[2] برای ارزیابی مدل معرفی شده، استفاده شد

نرخ CER به‌عنوان معیاری برای اندازه‌گیری تفاوت میان خروجی مدل و متن مرجع

1. character error rate (CER)
2. word error rate (WER)

در سطح کاراکتر به‌کار می‌رود. این نرخ نشان می‌دهد که چه درصدی از کاراکترها نیاز به جایگزینی[1]، حذف[2] یا افزودن[3] در خروجی مدل دارند تا با متن مرجع همخوانی یابند. این معیار به تغییرات جزئی مانند اشتباهات املایی بسیار حساس است. برای محاسبه CER از فاصله ویرایشی[4] بین دو رشته استفاده می‌شود. فاصله ویرایشی، حداقل تعداد عملیات مورد نیاز برای تبدیل یک رشته به رشته دیگر را نشان می‌دهد که شامل عملیات‌های زیر است

- ◇ جایگزینی: تعویض یک کاراکتر با کاراکتر دیگر
- ◇ حذف: حذف یک کاراکتر
- ◇ افزودن: اضافه کردن یک کاراکتر جدید

فرمول کلی برای محاسبه CER به‌صورت زیر است:

$$WER = \frac{S_w + D_w + I_w}{N_w} \quad (۹$$

در معادله بالا، S_c تعداد کاراکتر جایگزین‌شده، D_c تعداد کاراکتر حذف‌شده، و I_c تعداد کاراکتر اضافه‌شده و N_c تعداد کل کاراکترهای متن مرجع است

نرخ WER به‌عنوان معیاری برای ارزیابی تفاوت میان خروجی مدل و متن مرجع در سطح کلمه مورد استفاده قرار می‌گیرد. این نرخ نشان‌دهنده درصد کلماتی است که نیاز به جایگزینی، حذف یا افزودن در خروجی مدل دارند تا با متن مرجع هماهنگ شوند. این معیار نسبت به حذف یا افزودن کلمات حساسیت بیشتری دارد و می‌تواند میزان موفقیت مدل در انتقال معنای کلی را نمایش دهد. WER نیز مشابه CER بر اساس فاصله ویرایشی بین دو دنباله کلمه‌ای محاسبه می‌شود. فرمول محاسبه WER، مشابه فرمول CER است؛ با این تفاوت که به‌جای کاراکترها، بر روی کلمات اعمال می‌شود. فرمول کلی برای محاسبه WER در زیر آمده است

$$WER = \frac{S_w + D_w + I_w}{N_w} \quad (۱۰$$

در معادله بالا، S_w تعداد کلمات جایگزین‌شده، D_w تعداد کلمات حذف‌شده، و I_w

1. substitution
2. deletion
3. insertion
4. edit distance

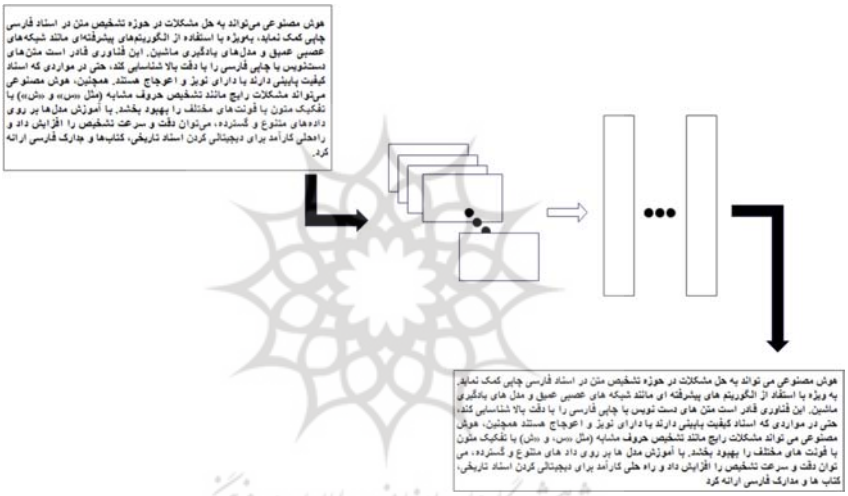
تعداد کلمات اضافه‌شده و N_w تعداد کل کلمات متن مرجع است

نتایج پیاده‌سازی روش‌های معرفی‌شده، در جدول ۳، نمایش داده شده است. در این جدول اعمال روش بر روی مجموعه داده‌های مختلف و روش‌های آموزش متفاوت بررسی شده است. این جدول درصد نرخ خطای کاراکتر (CER) و نرخ خطای کلمات (WER) را نشان می‌دهد. ردیف اول نتایج اعمال وزن‌های ارائه‌شده در منابع «تسراکت» است. ستون آخر ده کاراکتری را نمایش می‌دهد که بیشترین خطا را داشته‌اند. در کلیه روش‌ها نیم‌فاصله بیشترین تعداد خطا را دارد. خطای حرف «ی» نیز در همه موارد دیده می‌شود. بیشترین اشتباهات مربوط به علائم و نشانه‌ها است و حروف درست تشخیص داده شده‌اند. بهترین خروجی با نرخ خطای کلمه ۴۰/۲۹ درصد و نرخ خطای کاراکتر ۱۹/۵۴ درصد مربوط به تنظیم دقیق روش پایه با داده آزمایش «پکا»ی اصلی است. نرخ خطای کلمات ۱۹/۵۴ درصد به معنای دقت ۸۱ درصد است که دقت قابل قبولی در زبان فارسی است. در مجموعه داده ترکیبی، از «پکا»ی تمیزشده و کلمات موجود در منابع «تسراکت» استفاده شده است. برای این مجموعه داده روش پایه و روش عمیق‌تر برای آموزش به‌کار رفت. همان‌طور که در جدول ۳، دیده می‌شود، عمیق‌تر کردن شبکه نتایج را بهبود نداد. بنا بر نتایج این جدول، داده آموزشی اصلی بدون پاک‌سازی، که به داده آزمایشی از نظر ساختاری نزدیک‌تر است، خروجی بهتری دارد. بر اساس نتایج تحقیقات انجام‌شده، با اضافه کردن نمونه جدید داده آموزشی و یا عمیق‌تر کردن شبکه، آنچه که شبکه از قبل یادگرفته است، دستخوش تغییرات می‌شود و ممکن است به بهبود عملکرد کلی شبکه منجر نشود. خروجی شبکه بر روی یک متن ورودی در شکل ۳، نمایش داده شده است

جدول ۳. مقایسه عملکرد روش‌های پیشنهادشده بر روی مجموعه داده‌های متفاوت

مجموعه داده	مدل	نوع آموزش	WER	CER	نوع خطا و تکرار آن
تسراکت	پایه	از ابتدا	۴۰/۸	۲۰/۹۸	نیم‌فاصله: ۳۶۱۴ «–»: ۸۴۸ «،»: ۶۱۵ «۲»: ۴۲۲ «ی»: ۴۰۵ «۱»: ۳۶۱۴ «و»: ۸۴۸ «.»: ۶۱۵ «}»: ۴۲۲ «۰»: ۴۰۵
پکا	پایه	تنظیم دقیق	۴۰/۲۹	۱۹/۵۴	نیم‌فاصله: ۳۶۴۹ «–»: ۸۴۸ «ا»: ۵۶۹ «)»: ۴۹۳ «ی»: ۴۹۲ «۲»: ۴۲۲ «و»: ۳۹۸ «۱»: ۳۷۵ «ه»: ۳۰۲ «}»: ۲۸۵
پکای تمیزشده	پایه	آموزش از ابتدا	۵۹/۰۹	۳۰/۲۷	نیم‌فاصله: ۳۰۶۱ «)»: ۱۳۱۶ «–»: ۸۴۸ «(»: ۷۶۹ «ی»: ۷۰۹ «.»: ۶۰۲ «،»: ۵۸۸ «ا»: ۵۱۴ «۲»: ۴۲۲ «۱»: ۳۷۵
ترکیبی	پایه	تنظیم دقیق	۴۹/۱۷	۲۴/۲۹	نیم‌فاصله: ۲۸۱۹ «–»: ۸۴۸ «)»: ۵۶۳ «و»: ۵۱۳ «(»: ۴۶۴ «»: ۴۳۱ «.»: ۴۲۶ «۲»: ۴۲۲ «ی»: ۳۸۲ «۱»: ۳۷۵

مجموعه داده	مدل	نوع آموزش	WER	CER	نوع خطا و تکرار آن
ترکیبی	مدل عمیق‌تر	آموزش انتقالی	۴۹/۴۲	۲۵/۰۳	نیم‌فاصله: ۳۵۹۴ «–»: ۸۴۸ «.»: ۷۹۸ «()»: ۵۹۰ «)»: ۵۶۴ «ه»: ۴۹۶ «۲»: ۴۲۲ «ی»: ۴۰۲ «و»: ۳۹۴ «۱»: ۳۷۵



شکل ۳. خروجی شبکه پیشنهادی برای یک ورودی نمونه

۵. نتیجه‌گیری و کارهای آینده

در این پژوهش یک روش ترکیبی مبتنی بر شبکه‌های پیچشی و شبکه‌های بازگشتی برای نویسه‌خوانی نوری اسناد چاپی فارسی معرفی شد. در این روش ابتدا حروف تشخیص داده می‌شود و با استفاده از لایه رونویسی کلمات تشخیص داده می‌شود. برای آموزش مدل یک مجموعه داده بر اساس متون کتاب‌های فارسی ایجاد شد. این مجموعه داده شامل دو میلیون خط با پنج فونت متفاوت است. سپس، خروجی شبکه بر روی یک نمونه داده از پایان‌نامه‌های موجود در «گنج» بررسی شد. بر اساس نتایج روش معرفی‌شده، عملکرد قابل

قبولی در تشخیص درست کلمات و حروف دارد. افزون بر این، نشانه‌ها و علائم بیشترین خطا را شامل می‌شود. از مشکلات این روش این است که از ارتباط معنایی بین کلمات در جمله برای تشخیص استفاده نمی‌کند. در تحقیق‌های آتی استفاده از روش‌های مبتنی بر ترنسفورمر بررسی خواهد شد که تعبیه‌سازی بهتری از کلمات در جمله ارائه می‌دهند و از ارتباط کلمات در جمله برای تشخیص کلمات بهتر بهره می‌برند

References

Alayiaboozar, E. and A. Hojjatpanah. 2022. Steps for creating two Persian specialized corpora. International *Journal of Information Science and Management (IJISM)* 20 (4): 231-243.

Asadi-zeydabadi, F., A. Afkari-Fahandari, A. Faraji, E. Shabaninia, & H. Nezamabadi-pour. 2023. IDPL-PFOD2: A New Large-Scale Dataset for Printed Farsi Optical Character Recognition. arXiv: 2312.01177

Alkhawaldeh, R.S. 2020. Arabic (Indian) digit handwritten recognition using recurrent transfer deep architecture. *Soft Comput.* 25: 3131–3141.

Avyodri, R., S. Lukas, & H. Tjahyadi. 2022. Optical Character Recognition (OCR) for Text Recognition and its Post-Processing Method: A Literature Review. In Proceedings of the 1st International Conference on Technology Innovation and Its Applications (ICTIIA), Tangerang, Indonesia, pp. 1–6.

Fasha, M., B. Hammo, N. Obeid, & J. Al Widian. 2020. A Hybrid Deep Learning Model for Arabic Text Recognition. *10.48550/arXiv.2009.01987.*

Feng, W., G. Naiyang, L. Yuan, Z. Xiangand, & L. Zhigang, 2017, Audio visual speech recognition with multimodal recurrent neural networks. *International Joint Conference on neural networks (IJCNN)* May 14: 681-688). IEEE.

Graves A., S. Fern´andez, F. J. Gomez, and J. Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 369-376)

Hamida, S., O. E. Gannour, B. Cherradi, O. Hassan, & A. Raihani. 2022. Efficient feature descriptor selection for improved Arabic handwritten words recognition. *International Journal of Electrical & Computer Engineering* 12 (5): 2088-8708.

Hussain, S., S. Ali, Akram QU. Nastalique2015 .. Segmentation-based approach for Urdu OCR. *International Journal on Document Analysis and Recognition (IJDAR),* 18, 357–374.

Jain, M., M. Mathew, & C.V. Jawahar. 2017. Unconstrained OCR for urdu using deep CNN-RNN hybrid networks. 2017 4th IAPR Asian Conf. on Pattern Recognition (ACPR), Nanjing, People's Republic of China, 2017, pp. 747–752.

Javed, S.T., S. Hussain, A. Maqbool, S. Asloob, S. Jamil, & H. Moin. 2010. Segmentation free nastalique urdu ocr. *World Academy of Science, Engineering and Technology* 46: 456-461.

Kay, Anthony. 2007. Tesseract: an Open-Source Optical Character Recognition Engine. *Linux Journal* 159: 2.

Khosravi, H., and E. Kabir. 2009. Blackboard approach towards integrated Farsi OCR system. *International Journal of Document Analysis and Recognition (IJDAR)* 12 (1): 2132.

Khosrobeigi, Z., H. Veisi, H. Ahmadi, H. Shabanian. 2020. based post-processing approach A rule- to improve Persian OCR performance. *Scientia Iranica* 27 (6): 3019-3033. doi: 10.24200/sci.2020.53435.3267

Li, M., T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. 2021. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv. /abs/2109.10282

Mithe R., S. Indalkar, & N. Divekar. 2013. Optical Character Recognition. *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878 2 (1): 72-75.

Mirzaee, M. 2012. Text detection in images for Persian optical character recognition. MSc Thesis, University Of Tehran. Iran.

Momeni, S., & B. Babaali. 2022. Arabic Offline Handwritten Text Recognition with Transformers. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-2300065/v1

Mori, S., C. Y. Suen, and K. Yamamoto. 1992. Historical review of OCR research and development. Proceedings of the IEEE, 80 (7), 1029–1058. doi:10.1109/5.156468 10.1109/5.156468

Mostafa, A., O. Mohamed, A. Ashraf, A. Elbehery, S. Jamal, G. Khoriba, & A.S. Ghoneim. 2021. OCFormer: A Transformer-Based Model for Arabic Handwritten Text Recognition. In Proceedings of the International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 27; pp. 182–186.

Moudgil, A., S. Singh, & V. Gautam. 2022. Recent Trends in OCR Systems: A Review. In Machine Learning for Edge Computing; CRC Press: Boca Raton, FL, USA.

Naz, S., A. I. Umar, R. Ahmad, I. Siddiqi, S. B. Ahmed, M. I. Razzak, & F. Shafait. 2017. Urdu Nastaliq recognition using convolutional–recursive deep learning. *Neurocomputing* 243: 80–87.

Peng, X., H. Cao, S. Setlur, V. Govindaraju, & P. Natarajan. 2013. Multilingual OCR research and applications: An overview. In Proceedings of the International Workshop on Multilingual OCR, Washington, DC, USA,; pp. 1–8.

Radwan, M. A., M. I. Khalil, and H. M. Abbas. 2018. Neural networks pipeline for off line machine printed Arabic OCR. *Neural Process. Lett*. 48 (2): 769–787.

Rahmati M., M. Fateh, M. Rezvani, A. Tajary, V. Abolghasemi. 2020. Printed Persian OCR system using deep learning. IET Image Processing, 14: 3920-3931. https://doi.org/10.1049/iet-ipr.2019.0728

Raj, R., & A. Kos. 2022. A Comprehensive Study of Optical Character Recognition. In Proceedings of the 29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES), Lodz, Poland, 25–27 June 2022; pp. 151–154.

Shi, B., X. Bai, and C. Yao. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (11): 2298-2304.

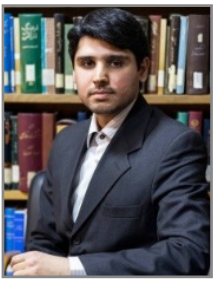
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. 2017. *Attention Is All You Need*. ArXiv. /abs/1706.03762

Zand, M., A. Naghsh Nilchi, & S.A Monadjemi. 2008. Recognition-based segmentation in Persian character recognition, Proceedings of World Academy of Science, Engineering and Technology. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 2 (1): 14-18.

آزاده فخرزاده

دارای مدرک تحصیلی دکتری در رشته پردازش تصویر از دانشگاه اوپسالای سوئد است. ایشان هم‌اکنون استادیار پژوهشکده فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

پردازش تصویر، یادگیری ماشین، کلان‌داده‌ها، و یادگیری عمیق از جمله علایق پژوهشی وی است.



امیرحسین صدیقی

دانش‌آموخته دکتری تخصصی در رشته مهندسی صنایع از دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) است. ایشان هم‌اکنون استادیار گروه پژوهشی سیستم‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران است.

هوش مصنوعی، یادگیری عمیق، مدل‌های زبانی بزرگ، کلان داده، و بهینه‌سازی از جمله علایق پژوهشی وی است.

محمد عشرت‌آبادی

دانشجوی کارشناسی مهندسی کامپیوتر در دانشگاه صنعتی امیرکبیر است. او در پروژه‌های تحقیقاتی مرتبط با شناسایی متون فارسی در تصاویر و ترکیب اطلاعات متنی و تصویری فعالیت داشته است.

یادگیری ماشین، مدل‌های تصویری-زبانی و پردازش زبان طبیعی از جمله علایق پژوهشی وی است..

البرز اسفندیاری

دارای مدرک تحصیلی کارشناسی ارشد در رشته مهندسی نرم‌افزار-کامپیوتر از دانشگاه صنعتی اصفهان است. ایشان هم‌اکنون دستیار پژوهشی در پژوهشکده هوش مصنوعی و پردازش داده پژوهشگاه علوم و فناوری اطلاعات ایران است.

هوش مصنوعی، یادگیری ماشین و کاربرد آن‌ها در حوزه‌های مختلف، به‌ویژه پردازش و بخش‌بندی تصاویر پزشکی از جمله علایق پژوهشی وی است.

پژوهش‌نامه
پردازش و مدیریت اطلاعات