Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

Image Captioning Using Convolutional Neural Networks with Attention Mechanism

Fereshteh Ahmadi [a], **Fatemeh Amiri** [a, *]

[a] *Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran*

PAPER INFO

Paper history:

Received: 2022-03-13
Revised: 2022-05-12
Accepted: 2022-10-19

Keywords:
image captioning,
convolutional neural network,
Resnet,
attention mechanism.

ABSTRACT

Image captioning involves the process of assigning descriptive text to images or photographs. To create an accurate description, several steps are necessary: 1) Object Identification: Initially, the objects within the image must be correctly identified. This includes recognizing their specific features and understanding the relationships between them. 2) Sentence Generation: Once the objects are identified, grammatically and semantically correct sentences are generated to describe the image. In this research, an encoder-decoder architecture is employed for producing textual descriptions. The proposed model consists of three following components: 1) Encoder (ResNet): The ResNet network serves as the encoder, extracting visual features from the input image. 2) Decoder (Convolutional Network): In the decoding section, a four-layer convolutional neural network (CNN) generates descriptions within the language model. 3)Attention Mechanism: To enhance the representation of image features and understand object relationships, an attention mechanism is utilized. This mechanism allows the model to focus on both the input image and the language model. The performance of the proposed model is evaluated using the MSCOCO and Flickr datasets. Experimental results demonstrate that the proposed architecture outperforms state-of-the-art researches in terms of Bleu1 and Meteor measures, while also achieving reduced training time compared to them.

DOI: https://doi.org/10.22075/jme.2024.31282.2495

*** Corresponding author.**
E-mail address: f.amiri@hut.ac.ir

How to cite this article:

مقاله پژوهشی

توصیف خودکار تصویر مبتنی بر شبکه های عصبی کانولوشنی با بهره گیری از مکانیزم توجه

فرشته احمدی[۱] و فاطمه امیری[۱،*]

اطلاعات مقاله

دریافت مقاله:
بازنگری مقاله:
پذیرش مقاله:

واژگان کلیدی:
توصیف تصویر،
رمزگذار-رمزگشا،
سازوکار توجه،
شبکه های کانولوشن.

مقاله،
شیوه نامه تدوین،
نویسنده،
چاپ،
شکل،
جدول،

چکیده

به فرایند اختصاص دادن توضیحات یا شرح متنی به تصاویر یا عکس‌ها توصیف تصویر اطلاق می‌شود. برای توصیف تصویر ابتدا لازم است که اشیا درون تصویر، ویژگی این اشیا و ارتباط میان آنان به درستی تشخیص داده شود و سپس جملاتی که از نظر گرامری و معنایی درست هستند، تولید شوند. در این تحقیق از معماری رمزگذار-رمزگشا جهت تولید توصیفات متنی استفاده شده است. مدل پیشنهادی شامل یک شبکه ResNet به عنوان رمزگذار جهت استخراج ویژگی‌های بصری تصویر است. در بخش رمزگشا شبکه کانولوشنی با چهار لایه جهت تولید توصیفات در مدل زبانی ارایه شده است. برای نشان دادن موثرتر ویژگی های حاصل از تصویر و درک روابط بین اشیا از یک ساز و کارتوجه استفاده شده است که قابلیت توجه به تصویر ورودی و مدل زبانی را دارد. کارایی مدل پیشنهادی بر روی مجموعه داده های MSCOCO و Flickr مورد ارزیابی قرار گرفته است. نتایج آزمایشگاهی نشان می‌دهد کارایی معماری پیشنهادی بر اساس معیار Bleu1 و Meteor نسبت به پژوهش‌های جدید ارجحیت دارد، درحالیکه زمان آموزش مدل پیشنهادی در مقایسه با پژوهشهای جدید کاهش یافته است.

DOI: https://doi.org/10.22075/jme.2024.31282.2495

۱-مقدمه[۱]

روزانه ما با تعداد بی شماری از تصاویر از منابع گوناگون مانند اینترنت، مقالات، اخبار، متن و سندهای علمی و تبلیغات سروکار داریم. بیشتر این تصاویر توصیف و توضیح متنی همراه خود ندارند اما انسان‌ها به سادگی می‌توانند معنای تصویر را درک کنند . با این وجود کامپیوترها باید بتوانند که برای مواردی که انسان نیاز دارد به کمک انسان آمده و به صورت خودکار تصاویر را تفسیر و یک توصیف متنی از تصاویر را ارائه دهند[1].

به فرایند اختصاص دادن خودکار توضیحات یا شرح متنی به تصاویر، توصیف تصویر[۲] اطلاق می‌شود. تحقیقات در حوزه توصیف تصویر از ضرورت زیادی برخوردارند. این حوزه پژوهشی، فناوری بینایی ماشین[۳] و پردازش زبان طبیعی[۴] را ترکیب کرده است. توصیف تصویر کاربردهای گسترده‌ای از جمله در حوزه پزشکی، شبکه‌های اجتماعی، و کمک به افراد با محدودیت‌های بینایی در زندگی روزمره دارد. کاربردهای دیگر توصیف تصویر، شامل جستجو و بازیابی جملات تصویری و همچنین ارتقاء هوش بصری ربات‌ها می‌باشد. در مقایسه با وظایف دیگر بینایی ماشین،

* پست الکترونیک نویسنده مسئول: f.amiri@hut.ac.ir
۱. گروه مهندسی کامپیوتر، دانشگاه صنعتی همدان، همدان، ایران

[2] image captioning
[3] computer vision
[4] natural language processing

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

مانند طبقه‌بندی تصویر و تشخیص اشیا، تولید خودکار توصیف متنی برای تصاویر با چالش‌های بیشتری روبرو است. زیرا حوزه توصیف تصویر علاوه بر تشخیص اشیا داخل تصویر، شناسایی ویژگی و صفات مهم تصویر و شناسایی ارتباط میان اشیا درون تصویر، شامل تولید جملاتی است که از نظر معنایی و قواعد دستوری مشکلی ندارند. [2]. در این زمینه، موفقیت‌های چشم‌گیری در سال‌های اخیر حاصل شده است اما هنوز چالش‌های بسیاری وجود دارد. یکی از مسائل مهم، تمرکز اکثر تحقیقات بر روی شبکه‌های بازیابی و تولید جمله است که به نظر می‌آید مزایای حاصل از ویژگی‌های مشتق شده از تصویر را نادیده می‌گیرند[8].

به دلیل قابلیت استخراج ویژگی‌های پیچیده از تصویر و درک بهتر از محتوا و نیز انعطاف‌پذیری بالا و قابلیت تطبیق با تصاویر متنوع، در سالهای اخیر روشهای مبتنی بر شبکه‌های عصبی عمیق برای تولید خودکار توصیف تصاویر به شدت مورد توجه محققین قرار گرفته است. یکی از رایج ترین معماری های عمیق مورد استفاده در تولید توصیف تصویر به صورت خودکار، معماری رمزگذار - رمزگشا[5] است. این معماری از دو بخش تشکیل شده است. در بخش رمزگذار، ابتدا یک شبکه عصبی کانولوشنی مانند VGG، ResNet و AlexNet تمام ویژگی‌های بصری تصویر را استخراج می‌کند. این ویژگی‌های بصری شامل اشیا درون تصویر است. سپس در بخش رمزگشا، این اطلاعات استخراج شده به عنوان ورودی به شبکه عصبی دوم (مدل زبانی) وارد می‌شود تا اطلاعات استخراج شده را به کلمات تبدیل کند. این شبکه معمولاً یک شبکه بازگشتی[6] (RNN) می‌باشد. شبکه‌های عصبی بازگشتی با مشکل گرادیان محوشونده و یا انفجار گرادیان روبه رو بودند، از این‌رو شبکه‌های LSTM که هم مشکل گرادیان محوشونده را حل می‌کردند و هم قادر بودند که وابستگی داده‌ها را حفظ کنند، مورد توجه بیشتری قرار گرفتند. لایه های LSTM دارای پیچیدگی محاسباتی بودند و به صورت ذاتی ماهیت ترتیبی در طول زمان دارند [19,20,21,22]. Aneja و همکارانش [11] مدلی را برای توصیف خودکار متن ارائه دادند که برای بخش رمزگشا از لایه عصبی کانولوشنی استفاده می‌کردند. روش آنان سریع و کارامد بود.

با توجه به اختلاف بزرگ بزرگ میان بینایی و روش‌های پردازش زبان، اغلب روش‌های ارایه شده در توصیف خودکار تصویر با مشکل عدم تطابق معنایی کامل بین تصاویر و توصیف‌های تولید شده مواجه هستند. به منظور رفع این چالش، سازوکار توجه در نقشه ویژگی گرفته است که می‌تواند یک نقشه فضایی ایجاد کند تا برجستگی خاصی به مناطق مهم تصویری مرتبط با هر کلمه بدهد [24] بنابراین، مدل‌ها می‌توانند روابط بین ویژگی‌های تصویر خاص و کلمات مربوطه در توصیف‌ها را بیاموزند.

در این تحقیق از معماری توصیف تصویر مبتنی بر شبکه های عصبی عمیق با معماری رمزگزار-رمزگشا استفاده شده است که در بخش رمزگزار از شبکه ResNet و در بخش رمزگشا از لایه های عصبی کانولوشنی به همراه لایه های batch normalization استفاده شده است. از یک مکانیزم توجه ساده برای افزایش کارایی مدل پیشنهادی استفاده است. مدل پیشنهادی بر روی مجموعه دادگان MSCOCO و Flickr30k آموزش داده شده است. نتایج آزمایشگاهی نشان می‌دهد کارایی معماری پیشنهادی بر روی مجموعه MSCOCO بر اساس معیار BLEU1 72.7٪ و براساس معیار BLEU4 29.4٪ است.

در ادامه، بخش دو به بررسی مطالعات پیشین اختصاص یافته است. سپس در بخش سوم، مساله مورد بحث و معماری پیشنهادی برای حل آن به تفصیل تشریح شده است. در بخش چهارم، نتایج آزمایشگاهی بررسی می‌شود. در پایان، در بخش پنجم، نتیجه‌گیری حاصل از این پژوهش و پیشنهاداتی برای کارهای آینده بیان می‌شود.

۲- کارهای پیشین

پژوهشها در حوزه خودکار سازی توصیف تصویر را می‌توان به سه دسته اصلی تقسیم کرد. توصیف تصویر مبتنی بر جست و جو[7] [3]، توصیف تصویر مبتنی بر قالب های زبانی[8] [4] و توصیف تصویر مبتنی بر شبکه‌های عصبی عمیق[9] [23]. در روش مبتنی بر قالب های زبانی، از یک

[5] Encoder-Decoder
[6] recurrent neural network

[7] Search-based approache
[8] Language template-based approaches
[9] Deep learning based approche

ساختار زبانی ثابت برای تولید توصیفات استفاده می‌شود. روش مبتنی بر جست و جو بر اساس تکنیک‌های جست‌وجو، توصیف مناسبی را بر اساس محتوای تصویر جستجو می‌کند. در حالیکه در روش سوم، از شبکه‌های عصبی عمیق[10]، به ویژه شبکه‌های انتقالی[11]، برای استخراج ویژگی‌های پیچیده از تصویر و تولید توصیفات دقیق استفاده می‌شود.

در روشهای مبتنی بر قالبهای زبانی، قابلیت انعطاف پذیری مدل در تولید توصیفات برای تصاویر متنوع کمتر است. در روشهای مبتنی بر جستجو، نیاز به یک مکانیزم جست‌وجو پیچیدگی افزوده شده به مدل را افزایش می‌دهد و ممکن است مدل در مواجهه با تصاویر متفاوت، به تولید توصیفات غیر دقیق یا کلیشه‌ای منجر گردد.

برای تولید توصیف مبتنی بر جست وجو، ابتدا یک مجموعه داده "تصویر-توصیف" تهیه می‌شود. در انجام وظیفه توصیف متن، ابتدا تصویر مورد نظر با تمام تصاویر موجود در مجموعه داده مقایسه می‌شود. تصاویری با بیشترین شباهت به تصویر مورد نظر انتخاب شده و سپس توصیف‌های متناسب با این تصاویر انتخابی از مجموعه داده "تصویر-توصیف" به عنوان مجموعه توصیف‌های کاندید در نظر گرفته می‌شوند. سپس این توصیف‌های کاندید رتبه‌بندی می‌شوند و توصیفی با بالاترین رتبه به عنوان توصیف تصویر مورد نظر انتخاب می‌شود.

Ordonez و همکاران [3] با ارائه مدل IM2TEXT ، یک روش برای تولید توصیف متنی بر اساس تصاویر ارائه کردند. ابتدا، با جمع‌آوری یک مجموعه بزرگ تصاویر از اینترنت، برای هر تصویر، یک عنوان و یک توصیف دستی ایجاد کردند. در مرحله بعد، با پیمایش تصاویر موجود در مجموعه داده، تصویری با بیشترین شباهت به تصویر مورد نظر را شناسایی کردند. سپس با استفاده از ویژگی‌های سطح پایین تصویر، یک فضای تصویری چندبعدی را ایجاد کردند که مناظرهای مشابه را نزدیک به یکدیگر نگاشت. در نهایت، توصیفهای مرتبط با تصاویری که بیشترین شباهت را با تصویر مورد نظر داشتند را رتبه‌بندی کرده و جمله‌ای که بالاترین رتبه را داشت به عنوان خروجی نهایی را انتخاب می‌کردند.

در روش پیشنهادی توسط Yang و همکاران[4]، از قالب زبانی چهارگانه (اسامی، افعال، حروف اضافه، مکان وقوع) به عنوان قالب زبانی استفاده می‌شود. این روش با استفاده از الگوریتم‌های تشخیص اشیا و مکان تصویر، ابتدا اشیا و مکان‌های موجود در تصویر تشخیص داده می‌شود. در نهایت، یک توصیف متنی مناسب با تصویر تولید می‌شود.

از سال‌های قبل معماری های بسیار برای توصیف خودکار تصویر توسط محققین معرفی شده اند که عمده آن ها از روش های یادگیری با نظارت و معماری رمزگزار و رمزگشا استفاده می کنند [2]. در روش‌های یادگیری با نظارت داده‌های آموزش همراه با خروجی مطلوب به سیستم اعمال می‌شوند که به آن خروجی‌های مطلوب برچسب گفته می‌شود. الگوریتم‌های یادگیری با نظارت در زمینه‌های مختلف، مانند دسته‌بندی تصاویر، تشخیص اشیا و استخراج ویژگی‌ها، در سال‌های اخیر به موفقیت‌های قابل توجهی دست یافته‌اند. این پیشرفت‌ها محققان را تشویق کرده‌اند تا از الگوریتم‌های یادگیری با نظارت برای تولید توصیف‌های متنی مرتبط با تصاویر استفاده کنند. در این روش‌ها، مجموعه‌ی تصاویر همراه با توصیف‌های متنی مرتبط به شبکه‌های عصبی اعمال می‌شوند[5].

Mao و همکارانش[6] شبکه‌های عصبی بازگشتی چندگانه را برای تولید توصیف متنی تصاویر ارایه کردند. برای این منظور، یک شبکه عصبی بازگشتی برای جملات و یک شبکه عصبی کانولوشنی برای تصاویر استفاده شد که شبکه عصبی مورد استفاده برای مدل زبانی یک لایه تعبیه ویژگی های متراکم را برای هر لایه می گیرد.

Vinyalsو همکارانش[8] از معماری رمزگزار-رمزگشا به نام NIC برای تولید خودکار توصیف تصاویر استفاده کردند. روش پیشنهادی آنان شامل یک شبکه عصبی کانولوشنی برای استخراج ویژگی‌های بصری تصویر و یک شبکه عصبی LSTM برای تولید جملات است. اصلی‌ترین ایده در روش NIC این است که از تکنیک‌های ماشین ترجمه بهره گرفته شود، جایی که یک متن به زبان مبدا دریافت‌شده و متن ترجمه‌شده به زبان مقصد تولید می‌شود.

[10] Deep Neural Networks
[11] Transformer

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

Xu و همکارانش [7] مدل توصیف تصویر مبتنی بر ساز و کار توجه[۱۲] را معرفی کردند. مدل آن‌ها نقاط برجسته تصویر را به صورت پویا توصیف می‌کرد . مدل پیشنهادی از یک شبکه عصبی کانولوشنی برای استخراج مجموعه بردارهای ویژگی‌های بصری تصویر استفاده می‌کند. این بردارها هر کدام دارای d بعد هستند که به صورت مجزا برای هر بخش از تصویر تولید می‌شوند. مدل پیشنهادی از شبکه عصبی LSTM به عنوان بخش رمزگشا استفاده می‌کرد.

روش پیشنهادی Chen و همکارانش[9] قادر بود که توصیفات جدیدی برای تصاویر تولید کند و همچنین ویژگی های بصری را از توصیفات دریافت‌شده بازیابی کند. روش آن‌ها به صورت پیوسته نمایش بصری تصاویر را از کلمات تولید شده را بازیابی می‌کرد.

در سالهای اخیر Ruifan و همکاران [10] رمزگشای CNN-Dual با حافظه طولانی مدت و محاسبه موازی پیشنهاد دادند. برخلاف بعضی از روش‌ها، روش مبتنی بر معماری چند حالته بر الگوی خاصی تکیه نمی‌کند. اما در عوض به ویژگی‌های سطح بالای تصویر و نمایش کلمات یاد گرفته شده از شبکه‌های عصبی عمیق و زبان عصبی چندگانه[۱۳] وابسته است.

فامیل ستاری و همکاران [۲۸] یک روش جدید را پیشنهاد کردند که چارچوب رمزگذار-رمزگشا را با ساز و کار توجه و سازوکار توجه بر توجه ادغام کردند. بخش رمزگذار مدل شامل چند بخش ResNet، Attention-LSTM، Multi Head Attention و سازوکار توجه بر توجه بود. از ResNet برای استخراج ویژگی‌های کلی تصویر استفاده شده بود. لایه‌های Language-LSTM مسئولیت رمزگشایی را بر عهده داشتند. سازوکار توجه از شواهد محلی برای افزایش نمایش ویژگی‌ها و استدلال در تولید توصیفات تصویری بهره برد و سازوکار توجه بر توجه می‌توانست روابط اشیای داخل تصاویر را به خوبی درک کند

Ding و همکاران [۲۹] یک روش یادگیری باقیمانده برای تولید توصیف خودکار برای هر تصویر داده‌شده ارائه دادند. در این روش ,از یک شبکه عصبی کانولوشنی برای استخراج ویژگی‌های طیفی و مکانی تصویر استفاده می کردند که برای حل مساله تولید عنوان ضروری است. در روش معرفی شده توسط آنان دو نوع مکانیزم توجه پیشنهاد شد :محرک محور و مفهوم محور مکانیزم توجه پیشنهادی آنان به ترکیبی از شبکه عصبی کانولوشنی بر روی تصاویر و شبکه حافظه بلند مدت بر روی جملات تکیه دارد علاوه بر این، آنان کیفیت متفاوت زبان را با استفاده از یک مولد حاشیه نویسی تصویر در سیستمی که توصیه شده است در نظر می گرفتند.

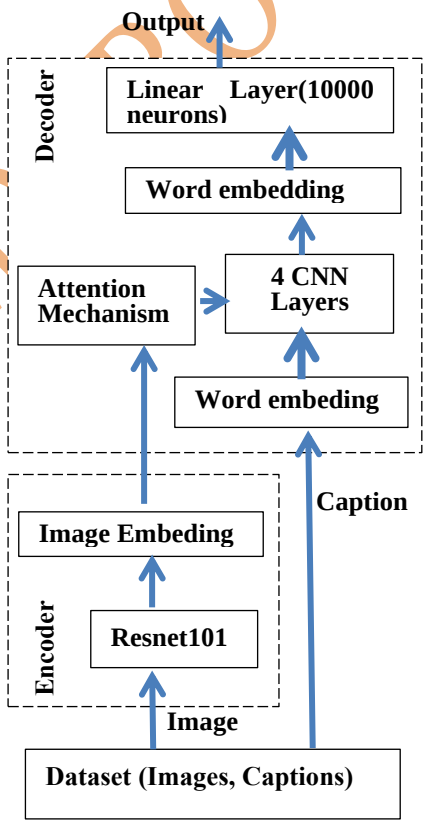
Aneja و همکارانش[11] برای اولین بار از یک شبکه عصبی کانولوشنی برای تولید جمله در بخش رمزگشا استفاده کردند. در این تحقیق از شبکه VGG16 به عنوان رمزگزار استفاده شده است و شبکه های CNN به عنوان رمزگشا استفاده شده است. با بررسی کارهای گذشته و پیچیدگی محاسباتی کمتر برای شبکه های کانولوشن در مقایسه با شبکه های بازگشتی، در مدل پیشنهادی این پژوهش نیز از این نوع شبکه های کانولوشن استفاده شده است.

۳- روش پیشنهادی

در روش پیشنهادی از یک معماری رمزگزار-رمزگشا استفاده می‌شود. برای بخش رمزگزار با توجه به دقت بالایی شبکه ResNet در تشخیص اشیا درون تصویر، از این نوع شبکه استفاده شده است. همچنین با هدف یادگیری مجدد وزن‌های دولایه آخر شبکه متناسب با مجموعه دادگان مورد استفاده، از شبکه ResNet101 با قابلیت دوباره یادگیری وزن‌ها استفاده شده است. برای افزایش کارای مدل پیشنهادی از ساز و کار توجه در بخش رمزگشا استفاده می‌شود که قابلیت توجه به تصویر ورودی و مدل زبانی را دارد. در شکل ۱ یک معماری پیشنهادی از مراحل یادگیری مدل توصیف خودکار تصویر مشاهده می‌شود. همانطور که مشاهده می‌شود، تصویر ورودی به شبکه ResNet وارد می‌شود تا اشیا درون تصویر استخراج شوند. خروجی شبکه ResNet به لایه تعبیه تصویر[۱۴] وارد می‌شود و به یک بردار ۵۱۲ بعدی تقلیل می‌یابد. در مولفه مربوط به ساز و کار توجه، برای هر کلمه یک بردار

[12] Attention mechanism
[13] Neural Language MultiModal
[14] Image embedding

توجه به ابعاد ۵۱۲ محاسبه می‌شود. از سوی دیگر توصیف متنی موجود در مجموعه داده به صورت کلمه به کلمه به لایه تعبیه کلمات[15] وارد می‌شوند و ضمن رقمی‌سازی هر کلمه به ابعاد ۵۱۲ بعد تقلیل می یابد. در لایه الحاق[16]، خروجی لایه تعبیه کلمات هر کلمه با خروجی لایه تعبیه تصویر متناسب با آن کلمه با هم متصل می‌شوند. لازم است قبل از اینکه ورودی دو لایه به لایه الحاق اعمال شود، از لایه‌های نرمال‌سازی دسته‌ای[17] به منظور نرمال‌سازی خروجی لایه ها استفاده شود تا از بزرگ شدن بیش از حد خروجی لایه های قبلی جلوگیری کند.



شکل ۱: معماری پیشنهادی برای توصیف خودکار تصاویر که با عنوان ResNet+CNN به آن اشاره می‌شود.

خروجی لایه الحاق با بردار توجه مربوط به همان کلمه با هم ترکیب می‌شوند و به لایه‌های کانولوشن وارد می‌شوند و ضمن عبور از چهار لایه کانولوشن هر کلمه به ابعاد ۵۱۲ تقلیل می‌یابد. لازم به ذکر است بردار توجه به تمام لایه‌های کانولوشنی(CNN) وارد می‌شود. خروجی لایه کانولوشن به یک لایه تعبیه کلمات وارد می‌شود و هر کلمه به ابعاد ۲۵۶ تقلیل می‌یابد.

در ادامه، خروجی۲۵۶ بعدی لایه تعبیه کلمات به یک لایه خطی به ابعاد ده هزار نورون وارد می‌شود. در این مرحله فرض می‌شود حداکثر تعداد کلمات متفاوت در مجموعه داده ده هزار کلمه است و هر نورون معادل یک کلمه در نظر گرفته می‌شود. این لایه از تابع فعال‌ساز softmax استفاده می‌کند و احتمال تولید هر کلمه را به شرط مشاهده تصویر I و کلمه ایی که در مرحله پیش تولید شده، محاسبه می‌کند. در ادامه توضیحات بیشتری در مورد عناصر اصلی معماری بیان می‌شود.

۳-۱ شبکه ResNet

شبکه Resnet[18][17] نوع خاصی از شبکه عصبی عمیق است و نوآوری اصلی این شبکه آن است که برای حل مشکل زیاد شدن تعداد لایه‌ها، بلوک جدیدی به نام بلوک باقیمانده معرفی شد. در شبکه‌ی ResNet ، اتصالی به نام اتصالات میانبر وجود دارد که از یک یا چند لایه عبور می‌کند و مسئله‌ی گرادیان محوشونده را در شبکه‌های عصبی عمیق حل می‌کند و امکانی را فراهم می‌آورد که گرادیان از مسیرهای میانبر عبور کند.

۳-۲ شبکه های عصبی کانولوشنی

نوعی از شبکه‌های عصبی عمیق است که برای پردازش داده‌های تصویری طراحی شده است. این شبکه‌ها، به ویژه در زمینه بینایی ماشین، مورد استفاده قرار می‌گیرند. شبکه عصبی کانولوشن با الهام از ساختار قشر بینایی[19] در مغز انسان ارائه شده است. این لایه ها شامل یک پنجره فیلتر است که هر عضو این پنجره دارای وزن خاصی هستند و این پنجره بر روی داده ورودی می‌لغزد [27].

۳-۳ مکانیزم توجه

مهم‌ترین بخش در سیستم بینایی انسان ، قابلیت توجه به بخش‌های مختلف تصویر است، این موضوع ایده اصلی مکانیزم توجه می‌باشد به صورتی که به شبکه این قابلیت را میدهد که به بخش های مختلف تصویر در صورت لزوم

[15] Word embedding
[16] Concatenate layer
[17] Batch normalization

[18] Residual Network
[19] Visual Cortex

دقت و توجه بیشتری نشان دهد و ناحیه های تصویر جهت شناسایی اشیا درون آن ناحیه، به صورت پویا مورد توجه شبکه قرار بگیرند.[7] برای هر کلمه بردار توجه و پارامتر توجه محاسبه می‌شود. برای محاسبه بردار توجه متناسب با هر کلمه، پارامتر توجه مخصوص به آن کلمه به ابعاد ۷*۷ مفروض می‌باشد و از معادله (۱) محاسبه می‌شود.

$$a_{i,j} = \frac{\exp\left(W(d_j)^T c_i\right)}{\sum_i \exp\left(W(d_j)^T c_i\right)} \quad (1)$$

که d_j بردار محاسبه شده در خروجی لایه تعبیه کلمه برای کلمه j ام می‌باشد. c_i همان بردار خروجی شبکه ResNet برای ناحیه i ام می‌باشد. W یک لایه خطی است که بر d_j اعمال می‌شود. کلمه انتخاب شده در هر مرحله به شبکه وارد می‌شود تا از طریق ساز و کار توجه در تولید کلمه بعدی به کار گرفته شود. بیشترین طول توصیف تولید شده توسط مدل، بیست کلمه است. این روند تولید کلمات تا رسیدن به توکن نهایی یا حداکثر بیست کلمه ادامه می‌یابد.

۴- نتایج پیاده‌سازی

در این فصل کارایی مدل پیشنهادی بررسی می‌شود. برای پیاده سازی مدل پیشنهادی، از زبان برنامه نویسی پایتون نسخه ۳.۱۰.۱۲ در چارچوب کتابخانه Pytorch استفاده شد. در این تحقیق داده‌ها به نسبت ۲۰ به ۸۰ برای بخش‌های آزمایش و آموزش تقسیم شدند. همچنین یک کامپیوتر با یک پردازنده Corei7 شرکت اینتل با ۲۴ گیگابایت حافظه رم و یک پردازنده گرافیکی NVIDIA GTX 3070 مورد استفاده قرارگرفت. فرایند آموزش در مدل پیشنهادی در ۵۰ دور انجام می‌شود. لازم به ذکر است نتایج معیارهای ارزیابی در این بخش به صورت درصد ارایه شده است و مدل پیشنهادی به صورت مخفف با عبارت Resnet+CNN در جدولها نمایش داده می‌شود و بر اساس شکل ۱، منظور مدلی است که در بخش رمزگزار از شبکه ResNet و در بخش رمزگشا از شبکه های کانولوشنی بهره برده است.

۴-۱ مجموعه داده مورد استفاده

در این بخش مجموعه داده‌های متداول برای ارزیابی کارایی مدلهای توصیف خودکار شرح داده می شود.

MSCOCO [12] یک مجموعه داده معروف در زمینه بینایی ماشین و پردازش زبان طبیعی است. این مجموعه داده شامل تصاویر پیچیده و متنوع است و تعداد زیادی توصیف مختلف برای هر تصویر فراهم می‌کند، که باعث ارتقاء دقت و تنوع در مدل‌های یادگیری ماشین می‌شود. این مجموعه داده شامل ۱۲۳،۲۸۷ تصویر دارد. هر تصویر دارای حداقل پنج عنوان توصیف می‌باشد که می‌تواند تنوع در زبان و توصیف ایجاد کند. این ساختار چندمتغیره باعث می‌شود تا مدل‌ها بتوانند اطلاعات گوناگون را درک کرده و توصیف دقیقی از تصویر ارائه دهند.

Flickr30K[25] مجموعه داده‌ای برای توصیف خودکار تصویر و درک زبان پایه است. شامل ۳۰ هزار تصویر جمع‌آوری شده از فلیکر با ۱۵۸ هزار توصیف است که توسط حاشیه نویسان انسانی ارائه شده است. هیچ گونه تقسیم ثابتی از تصاویر را برای آموزش، آزمایش و اعتبار سنجی ارائه نمی‌دهد. مجموعه داده همچنین حاوی آشکارسازهایی برای اشیاء معمولی، طبقه‌بندی‌کننده رنگ و سوگیری نسبت به انتخاب اشیاء بزرگ‌تر است.

۴-۳- معیارهای ارزیابی

برای ارزیابی مسائل خودکارسازی توصیف تصویر، چندین معیار ارزیابی استفاده می‌شود. مقادیر محاسبه‌شده برای معیارها مقداری بین ۰ و ۱ است که بالاتر بودن این مقدار به معنای عملکرد بالاتر جملات تولید شده می‌باشد. به صورت معمول، نتایج حاصل از معیارها به صورت درصد نمایش داده می شود. در محاسبه تمام معیارهای مذکور، هر کلمه "یک گرم[20]" نامیده می‌شود.

معیار BLEU[21] [13] :این معیار بر اساس انطباق میان توصیفات تولید شده توسط مدل با توصیفات مرجع (توصیفات اصلی) است. BLEU تعیین میکند چند تعداد از کلمات و عبارات مدل با توصیفات مرجع همخوانی دارند. این معیار بطور مخفف B نامیده می‌شود. در پژوهشهای قبلی معیارBLEU به تطبیق یک گرم‌ها تا چهار گرم‌ها می‌پردازد و به صورت B1,…,B4 نمایش

[20] Uni-gram
[21] Bilingual Evaluation Understudy

داده می‌شود. این معیار به موقعیت کلمات در متن ترجمه و متن منبع توجهی نمی‌نماید. همواره مقدار B1 از معیارهای دیگر بالاتر محاسبه می‌شود زیرا فقط به یک-گرم‌ها دقت می‌نماید

معیار METEOR[22] [14]: یک معیار ارزیابی اتوماتیک است که از انطباق لغات، نحوه‌ی جمله‌سازی، و ساختار جمله برای ارزیابی توصیف تصویر استفاده می‌کند. این معیار بطور مخفف M نامیده می‌شود. در این معیار می‌توان بر اساس کلمات مترادف، کلمات هم‌ریشه و کلمات هم‌معانی نیز یک گرم‌ها را تشخیص داد.

معیار ROUGE-L[23] [16]: انطباق بین توصیف تولیدی و توصیف مرجع (واقعی) را ارزیابی میکند. این معیار از طولترین زیردنباله مشترک بین دو توصیف برای اندازه‌گیری انطباق استفاده می‌کند. این معیار بطور مخفف R نامیده می‌شود. در این معیار می‌توان بر اساس کلمات مترادف، کلمات هم‌ریشه و کلمات هم‌معانی نیز یک گرم‌ها را تشخیص داد.

معیار CIDEr[24] [15]: به عنوان یک معیار ارزیابی اتوماتیک برای توصیف تصویر شناخته می‌شود. این معیار از ارزیابی تنوع و جلب‌توجه در توصیف تولیدی استفاده می‌کند. این معیار بطور مخفف C نامیده می‌شود..

۴-۴- نتایج آزمایشگاهی

در گام اول آزمایشات، نتایج کارایی مدل پیشنهادی به ازای تعداد لایه های کانولوشنی متفاوت بررسی می‌شود. نتایج ارزیابی بر روی مجموعه داده MSCOCO در جدول ۱ بیان شده است. در این جدول سطرها معیارهای ارزیابی و ستونها معماری پیشنهادی با تعداد لایه‌های کانولوشن متفاوت است. نتایج سطرها به صورت درصد نمایش داده شده است. نتایج آزمایشگاهی نشان می‌دهد بالاترین کارایی در معیارهای B2، B4، M و C با چهارلایه کانولوشن بر روی مجموعه داده MSCOCO ایجاد شده است.

در ادامه نتایج کارایی مدل پیشنهادی با ۴ لایه کانولوشنی با نزدیکترین پژوهشها بر روی مجموعه داده MSCOCO و Flickr30K مقایسه می‌شود. نتایج ارزیابی در دو جدول ۲ و۳ قابل مشاهده است. سطرها معیار ارزیابی و ستونها مدلهای مختلف ارایه شده را نمایش می‌دهد. نتایج به صورت درصد در جدولها نمایش داده شده است.

جدول ۱: نتایج کارایی Resnet+CNN با تعداد متفاوت لایه کانولوشنی بر روی مجموعه داده MSCOCO

معیار ارزیابی	۳ لایه کانولوشنی	۴لایه کانولوشنی	۵لایه کانولوشنی
B1	۶۹.۹	**۷۲.۷**	۷۱.۷
B2	۵۰.۵	**۵۴.۰**	۵۲.۷
B3	۳۹.۱	۳۸.۵	۳۵.۰
B4	۲۷.۶	**۲۹.۴**	۲۷.۱
M	۲۱.۹	**۲۴.۷**	۲۲.۳
R	۵۲.۲	۵۱.۰	۵۱.۰
C	۷۷.۰	**۸۸.۴**	۸۳.۶

جدول ۲: مقایسه کارایی مدل پیشنهادی Resnet+CNN با پژوهش‌های دیگر برروی مجموعه داده MSCOCO

	Chen et.al. [9]	Xu et al [7]	Famil sattari et. al.[28]	Aneja et. al [11]	**Our model**
B1	۷۲.۵	۷۱.۸	۷۰.۲	۷۱.۰	**۷۲.۷**
B2	**۵۵.۶**	۵۰.۴	۵۱.۰۵	۵۳.۷	۵۴.۰
B3	۴۱.۴	۳۵.۷	۳۵.۸	۳۹.۱	۳۸.۵
B4	۳۰.۶	۲۵.۰	۲۵.۳	۲۸.۱	۲۹.۴
M	۲۴.۶	۲۳.۴	۲۷.۳	۲۴.۱	**۲۴.۷**
R	۵۲.۸	-	۵۷.۳	۵۱.۹	۵۱.۰
C	۹۱.۱	-	۴۹.۳	۸۹.۰	۸۸.۴

همان‌طور که در جدول ۲ مشخص است، مدل پیشنهادی نسبت به مدل Aneja و همکاران [11] و فامیل ستاری و همکاران [28] و Chen و همکاران [9] در معیار B1 عملکرد بهتری داشته است. مدل پیشنهادی بالاترین کارایی براساس معیار B2 را نسبت به مدل Aneja و همکاران، Ding و همکاران و فامیل ستاری و همکاران نمایش می‌دهد. همچنین در معیار B3 از مدل پیشنهادی Ding و همکاران عملکرد بهتری داشت. مدل پیشنهادی در معیار B4 عملکرد خوبی از خود نشان داد و نسبت به مدل پیشنهادی فامیل ستاری و همکاران و Ding و

[22] Metric for Evaluation of Translation with Explicit ORdering

[23] Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence

[24] Consensus-based Image Description Evaluation

همکاران و Aneja و همکاران بهبود یافته است.

جدول ۳: مقایسه کارایی مدل پیشنهادی Resnet+CNN با پژوهش‌های دیگر برروی مجموعه داده Flickr30k

	Chen et.al [9]	Xu et al [7]	Famil Sattari et al.[28]	Aneja et. al [11]	**Our model**
B1	۶۶.۲	۶۶.۹	۶۹.۷	-	**۶۸.۱**
B2	۴۶.۸	۴۳.۹	۵۱.۷	-	۴۵.۸
B3	۳۲.۲	۲۹.۶	۳۸.۱	-	**۳۲.۴**
B4	۲۲.۳	۱۹.۹	۱۸.۷	-	۲۱.۷
M	۱۹.۵	۱۸.۴۶	۲۸.۶	-	**۱۹.۷**
R	-	-	۵۷.۹	--	۴۶.۵
C	-	-	۵۲.۱	-	۷۹.۹

در جدول ۳ نتایج نشان می‌دهد براساس معیار B1 کارایی مدل پیشنهادی از پژوهشهای Xu و همکاران و Chenو همکاران بالاتر است. براساس معیار B3 مدل پیشنهادی بر پژوهشهای [7] و [9] برتری دارد.
نزدیکترین پژوهش به مدل پیشنهادی معماری رمزگذار – رمزگشا ارایه شده در [11] است که در بخش رمزگذار از VGG و در بخش رمزگشا از سه لایه‌ی کانولوشنی و ساز و کار توجه استفاده کرده‌اند. کارایی مدل پیشنهادی ما براساس معیارهای B1,B2,B4 و M (جدول ۲) برروی مجموعه داده MSCOCO بالاتر از مدل ارائه شده توسط Aneja و همکاران [11] است و کارایی پژوهش [11] توسط نویسندگان بر روی این مجموعه داده (جدول ۳) بررسی نشده است.
در گام بعدی آزمایشات، تاثیر انواع شبکه های کانولوشنی به عنوان رمزگذار بررسی می‌شود. باتوجه به اینکه در پژوهش ارائه شده در Aneja و همکاران [۱۱] از شبکه VGG در بخش رمزگذار استفاده شده است، در ادامه، تاثیر استفاده از شبکه VGG به‌جای Resnet101 در بخش رمزگذار مدل پیشنهادی بررسی می‌شود. به عبارت دیگر در معماری پیشنهادی این پژوهش (شکل ۱) به جای Resnet101، از شبکه VGG استفاده می‌شود. به مدل حاصل با عبارت VGG+CNN اشاره می‌شود. نتایج ارزیابی کارایی مدل حاصل در جدول ۴ بر روی سه مجموعه داده نمایش داده شده است. در این جدول، در هر دو مدل از چهارلایه کانولوشن به عنوان رمزگشا استفاده شده است. در این جدول نتایج مدل پیشنهادی ResNet+CNN با مدل VGG+CNN مقایسه شده است. همان‌طور که در این جدول مشاهده می‌شود معماری پیشنهادی بر روی مجموعه داده MSCOCO با حضور Resnet کارایی بهتری در معیارهای B1,B3,B4,R,C نسبت به VGG داشته است. مدل پیشنهادی با حضور ResNet بر روی مجموعه داده Flickr30k در معیارهای B1، B2، B3، B4، R و C عملکرد بهتری داشته است.
همان‌طور که در جدول ۴ مشخص است، مدل پیشنهادی بر روی مجموعه داده MSCOCO به نتایج بهتری نسبت به Flickr30K دست یافته است که تعداد زیاد داده‌های این مجموعه داده را می‌توان علت عملکرد بهتر مدل بر روی این مجموعه داده در نظر گرفت.

در این مرحله، کارایی مدل پیشنهادی از نظر زمان اجرا بررسی می‌شود. نظر به اینکه ویژگی‌های سخت افزاری برای پیاده‌سازی مدل در تعداد اندکی از پژوهش‌های قبلی ارایه شده است. در نتیجه در این مرحله، زمان اجرای مدل پیشنهادی ResNet+CNN تنها با پژوهش [28] مقایسه می‌شود. در پژوهش [28] آموزش مدل بر روی مجموعه داده MSCOCO حدود ۱۸۰ ساعت زمان برای ۵۰ دور آموزش اعلام شده است. در مدل پیشنهادی در [28]، در بخش رمزگشا از شبکه عصبی بازگشتی استفاده شده است. این در حالی است که در مدل پیشنهادی این پژوهش (یعنی Resnet+CNN) در بخش رمزگزار از شبکه ResNet و در بخش رمزگشا از معماری شبکه های کانولوشنی بهره برده است، زمان آموزش مدل بر روی مجموعه داده MSCOCO حدود ۱۱۰ ساعت و بر روی مجموعه داده Flickr30k حدود ۳۷ ساعت است. این نتایج نشان می‌دهد لایه‌های عصبی کانولوشنی در بخش رمزگشای مدل پیشنهادی Resnet+CNN و نیز مکانیزم توجه ساده باعث کاهش زمان آموزش در مقایسه با پژوهش [28] شده است که از لایه های عصبی بازگشتی در بخش رمزگشا استفاده می‌کنند. بنابراین می‌توان نتیجه گرفت که معماری پیشنهادی پیچیدگی زمانی کمتر و کارایی بالاتر در مقایسه با کارهای پیشین را نشان می‌دهد.

۵- نتیجه گیری و کارهای آینده

در این پژوهش یک معماری رمزگذار-رمزگشا برای توصیف خودکار تصاویر ارایه شده است. تلاش شد تا با بهره گیری از شبکه ResNet در بخش رمزگزار و لایه های عصبی کانولوشنی در بخش رمزگشا و حضور مکانیزم توجه علاوه بر کاهش زمان آموزش شبکه به دقت بهتری در معیارها ارزیابی دست یابیم و جملاتی با کیفیت و منطبق با محتوای تصویر تولید شود.

با توجه به ارائه مدل های گوناگون از مکانیزم توجه ما در نظر داریم تا با بررسی مکانیزم های توجه گوناگون کارایی مدل پیشنهادی را بهبود دهیم و در ضمن معماری ViT بر پایه‌ی مکانیزم توجه و معماری ترنسفورمر طراحی شده است و توجه بیشتری به ارتباطات جهانی و کلیت تصویر دارد، که می‌تواند برای استخراج مفاهیم سطح بالا در این حوزه نیز مفید باشد بنابراین برانیم در کارهای آینده کارایی معماری ViT در بخش رمزگشا را بررسی نماییم.

جدول ۴- مقایسه کارایی مدل پیشنهادی با حضور VGG بجای Resnet در بخش رمزگذار

مجموعه داده	معماری مورد آزمایش	B1	B2	B3	B4	M	R	C
MSCOCO	ResNet+CNN	**۷۲/۷**	۵۴/۰	**۳۸/۵**	**۲۹/۴**	۲۴/۷	**۵۱.۲**	**۸۸.۴**
	VGG+CNN	۷۲/۱	**۵۵/۳**	۳۸/۱	۲۷/۰	**۲۷/۹**	۴۹/۹	۷۶/۱
Flickr30k	ResNet+CNN	**۶۸/۱**	**۴۵/۸**	**۳۲/۴**	**۲۱/۷**	۱۹/۷	**۴۶/۵**	**۷۹/۹**
	VGG+CNN	۶۴.۶	۴۴.۴	۳۰.۲	۱۸.۰	**۲۲.۲**	۴۵.۱	۷۶.۶

تعارض منافع

نویسندگان اعلام می‌کند که در مورد انتشار این مقاله تعارض منافع وجود ندارد.

مراجع

[1] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank, “ Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures”. Journal of Artificial Intelligence Research (JAIR), no. 55 (2016): 409–442.

[2] S. Bai and S. An, “A Survey on Automatic Image Caption Generation”. Neurocomputing. 2018

[3]V. Ordonez, G. Kulkarni, and T.L.Berg, “Im2text: Describing images using 1 million captioned photographs”. In Advances in Neural Information Processing Systems. No.24 (2011) :1143–1151.

[4]Y. Yang, C. Teo, H.Daumé III, and Y. Aloimonos, “Corpus-Guided Sentence Generation of Natural Images”. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (2011): 444–454.

[5]A.Kumar and S. Goel, “A survey of evolution of image captioning techniques”. International Journal of Hybrid Intelligent Systems Preprint 14, no. 3 (2017): 123-139.

[6] J. Mao, W.Xu, Y.Yang, J. Wang, Z.Huang, and A.Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn)”. In International Conference on Learning Representations (ICLR), 2015.

[7] K.Xu, J.Ba, R.Kiros, K.Cho, A.Courville, R.Salakhudinov, R.Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention”, International Conference on Machine Learning (2015): 2048–2057.

[8] O.Vinyals, A.Toshev, S.Bengio, and D. Erhan. “Show and tell: A neural image caption generator”, In Proceedings of the IEEE conference on computer vision and pattern recognition (2015): 3156–3164.

[9] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," The IEEE conference on computer vision and pattern recognition (2015): 2422-2431.

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

[10] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: a convolutional language decoder for paragraph image captioning," Neurocomputing, vol. 396 (2020): 92-101.

[11] J. Aneja, A. Deshpande, and A.G. Schwing. “Convolutional image captioning”. The IEEE conference on computer vision and pattern recognition (2018) : 5561-5570.

[12] T. Lin, M.Maire, S. Belongie, J.Hays, P.Perona, D.Ramanan, P.DollÃąr, and C.Zitnick, “ Microsoft coco: Common objects in context. In European conference on computer vision (2014), 740–755.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (2002) : 311-318

[14] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (2005): 65-72.

[15] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," In Proceedings of the IEEE conference on computer vision and pattern recognition (2015): 4566-4575.

[16] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out (2004): 74-81.

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

[18] L. Chen et al., "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition (2017): 5659-5667.

[19] Q. Wu, C. Shen, P. Wang, A. Dick, and A.V.D.Hengel, “Image captioning and visual question answering based on attributes and external knowledge”. IEEE transactions on pattern analysis and machine intelligence 40, no. 6 (2017): 1367-1381

[20] L.Zhang, F.Sung, F.Liu, T.Xiang, S. Gong, Y.Yang, and T.M.Hospedales.” Actor-critic sequence training for image captioning”, arXiv preprint arXiv:1706.09601 (2017).

[21] S.Venugopalan, L.A.Hendricks, M. Rohrbach, R.Mooney, T. Darrell, and K.Saenko, “Captioning images with diverse objects”. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017) : 1170–1178.

[22] S.J.Rennie, E.Marcheret, Y.Mroueh, J.Ross, and V.Goel, “Self-critical sequence training for image captioning”. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017) : 1179–1195.

[23] K.Cho, B.V. Merriënboer, C.Gulcehre, D.Bahdanau, F. Bougares, H. Schwenk, & Y. Bengio,” Learning phrase representations using RNN encoder-decoder for statistical machine translation”. arXiv preprint arXiv:1406.1078 (2014).

[24] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition (2017) : 375-383.

[25] B.A. Plummer, L.Wang, C.M. Cervantes, J.C. Caicedo, J.Hockenmaier, and S. Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." In Proceedings of the IEEE international conference on computer vision (2015) : 2641-2649.

[26] Z. F. Sattari, H. Khotanlou, and E. Alighardash, "Improving Image Captioning with Local Attention Mechanism," in 2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2022: IEEE, pp. 1-5 .

[27] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into imaging 9, no. 4 (2018) :611-629.

[28] Z. F. Sattari, H. Khotanlou, and E. Alighardash, "Improving Image Captioning with Local Attention Mechanism," In 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (IEEE CFIS)(2022): pp. 1-5

[29] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," Neurocomputing 398 (2020): 520-530.