Semnan University

Journal of Modeling in Engineering

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

ISSN: 2783-2538

Research Article

Medical Report Generation for Chest X-rays Using Convolutional Recurrent and Attention-Based Architectures

Fardin Ghaderi[a], Mohammad Bagher Khodabakhshi[a,*], Shahriar Jamasb[a]

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

PAPER INFO

Paper history:

Received: 2024-11-05
Revised: 2025-05-30
Accepted: 2025-06-21

Keywords:

Medical image processing;
Recurrent deep neural networks;
Automatic image captioning;
Encoder;
Decoder;
Attention mechanism.

ABSTRACT

Medical images are extensively used in medical science for diagnosis and treatment protocol design. Writing medical reports in text form can be error-prone for inexperienced physicians due to the deep understanding of the disease and its analysis. It is also time-consuming and laborious for experts due to the large number of patients they see in a day. Also, the existence of template reports for physicians can significantly increase their accuracy in diagnosing diseases and reduce errors caused by inattention to details. This research presents a deep learning-based model for the automatic generation of radiology reports. This model is based on a combination of a convolutional recurrent structure and an attention-based architecture called Res-LSTM-Attn. In this model, features are first extracted from medical images using a convolutional residual network, and based on a multi-label word model, a report is predicted. Then, using the LSTM recurrent neural network and multi-head attention layers, the final report is generated. The performance of the proposed models was evaluated based on the BLEU 1-4, ROUGE-L, and CIDEr-D criteria. The results showed that the proposed model outperformed previous studies in generating long reports in terms of CIDEr-D and ROUGE-L metrics, with improvements of 7.2% and 3.2%, respectively.

DOI: https://doi.org/10.22075/jme.2025.35811.2749

*** Corresponding author.**
E-mail address: mb.khodabakhshi@hut.ac.ir

How to cite this article:
F. Ghaderi , M. B. Khodabakhshi and S. Jamasb, "Medical Report Generation for Chest X-rays Using Convolutional Recurrent and Attention-Based Architectures," Journal of Modeling in Engineering, 24 85 (2026): 1-15, doi: 10.22075/jme.2025.35811.2749

مقاله پژوهشی

تولیدخودکار گزارش برای تصاویر قفسه‌سینه با استفاده از ترکیب مدل کانولوشنی بازگشتی و معماری توجه محور

فردین قادری[۱]، محمدباقر خدابخشی [۱*]، شهریار جاماسب[۱]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۳/۰۸/۱۵
بازنگری مقاله: ۱۴۰۴/۰۳/۰۹
پذیرش مقاله: ۱۴۰۴/۰۳/۳۱

واژگان کلیدی:
تصاویر پزشکی،
شبکه عصبی عمیق بازگشتی،
تولیدخودکار گزارش،
رمزگشا،
رمزگذار،
مکانیسم توجه.

چکیده

در مطالعات علوم پزشکی، از تصاویر پزشکی برای تشخیص و طراحی پروتکل درمان بیماری‌ها بصورت گسترده استفاده می‌شود. برای پزشکان کم‌تجربه، نوشتن گزارش پزشکی به شکل متنی ممکن است مستعد خطا باشد، زیرا این کار نیازمند درک عمیق نسبت به بیماری و تجزیه و تحلیل آن است. همچنین برای متخصصان، این کار به دلیل تعدد بیمارانی که در یک روز مراجعه می‌کنند زمان‌بر و پر زحمت است. از دیدگاه دیگر، وجود گزارش‌های الگو برای پزشکان می‌تواند به میزان قابل توجهی دقت آن‌ها را در تشخیص بیماری و کاهش خطای ناشی از عدم توجه به جزئیات کاهش دهد. این پژوهش یک مدل مبتنی بر یادگیری عمیق را برای تولیدخودکار گزارش‌های تصاویر رادیولوژی ارائه نموده است. این مدل بر پایه ترکیب یک ساختار کانولوشنی بازگشتی و معماری توجه محور با نام Res-LSTM-Attn معرفی گردیده است. در این مدل ابتدا از تصاویر پزشکی با استفاده از شبکه عصبی کانولوشنی رِزنت ویژگی‌ها استخراج خواهند شد و بر اساس یک مدل چند برچسبی کلمات یک گزارش پیش‌بینی خواهند شد. در ادامه با استفاده از شبکه عصبی بازگشتی LSTM و لایه‌های توجه چندسر گزارش نهایی تولید می‌شود. عملکرد مدل‌های پیشنهادی بر اساس معیارهای BLEU 1-4 و ROUGE-L و CIDEr-D مورد ارزیابی قرار گرفت. نتایج نشان داد مدل پیشنهادی از نظر معیار CIDEr-D و ROUGE-L در تولید گزارشات طولانی بر مطالعات پیشین غلبه کرده است و این مقادیر به ترتیب به میزان ۲/۷ و ۲/۳ درصد بهبود یافته‌اند.

DOI: https://doi.org/10.22075/jme.2025.35811.2749

۱- مقدمه

تولید خودکار گزارش برای تصویر، از طریق روش‌های مختلف استخراج ویژگی و توصیف آن‌ها با استفاده از پردازش زبان طبیعی صورت می‌گیرد. گزارش یا توصیف تصویر ترکیبی از دو زمینه هوش مصنوعی است که شامل بینایی‌کامپیوتری برای استخراج بازنمایی‌های بصری و پردازش زبان طبیعی برای توضیح آن بازنمایی‌ها به جملات زبان انگلیسی است [۱]. تولیدخودکار زیرنویس تصویر یک کار چالش برانگیز است که فراتر از تشخیص، تقسیم بندی و طبقه بندی اشیاء است، زیرا این فرایند نیازمند درک روابط میان اشیاء مختلف در تصویر، بازنمایی بصری آن‌ها و تبدیل این بازنمایی‌ها به جملات معنادار است. با در دسترس بودن مجموعه داده‌های بزرگ، روش‌های توصیف تصاویر بر اساس الگوریتم‌های مبتنی بر یادگیری ماشینی

* پست الکترونیک نویسنده مسئول: mb.khodabakhshi@hut.ac.ir
۱. گروه مهندسی پزشکی، دانشکده مهندسی پزشکی و مکانیک، دانشگاه صنعتی همدان، همدان، ایران

استناد به این مقاله:
فردین قادری , محمدباقر خدابخشی و شهریار جاماسب، "تولید خودکار گزارش برای تصاویر قفسه سینه با استفاده از ترکیب مدل کانولوشنی بازگشتی و معماری توجه محور،" مدل سازی در مهندسی، ۲۴ ۸۵ (۱۴۰۵): ۱-۱۵، doi: 10.22075/jme.2025.35811.2749

روز به روز محبوبیت بیشتری پیدا می‌کنند.
استفاده از تصاویر اخذ شده از بدن در عمده تخصص‌های پزشکی دارای اهمیت است، به عنوان مثال، متخصصان پزشکی و رادیولوژیست‌ها از تصاویر پزشکی برای تشخیص و درمان بیماری‌ها استفاده می‌کنند. برای پزشکان کم‌تجربه، نوشتن گزارش پزشکی به شکل متنی ممکن است با خطا همراه باشد. علاوه بر آن، این امر زمان‌بر و دشوار است و پزشکان ناگزیر هستند تا روزانه تعداد زیادی از این تصاویر را بررسی کنند که به نوبه خود منجر افزایش بروز خطای انسانی ناشی از خستگی خواهد شد. این مسئله در مناطقی که از کمبود امکانات درمانی رنج می‌برند نیز اهمیت بیشتری دارد. برای تسهیل فرآیند گزارش‌دهی تصویر پزشکی، بسیاری از سیستم‌های تولید گزارش به کمک رایانه مبتنی بر توصیف تصویر پیشنهاد شده‌اند که به‌طور خودکار یافته‌ها را از تصاویر پزشکی استخراج می‌کنند.
یک گزارش تصویر پزشکی معمولاً شامل حداقل یک پاراگراف متشکل از چندین جمله است که برای ضایعات غیرطبیعی می‌تواند بسیار طولانی تر باشد. اولین کار برای تولید گزارش‌های رادیولوژی واقعی با جملات طولانی، مدل یادگیری چندکاره با مکانیسم توجه مشترک است که توسط Jing و همکاران [۱] ارائه گردیده است. آنها یک مدل خافظه دار سلسله مراتبی متشکل از یک LSTM جمله و یک LSTM کلمه برای تولید گزارش‌های طولانی اشعه‌ایکس قفسه‌سینه، با الهام از شبکه‌های بازگشتی سلسله مراتبی پیشنهاد کردند. در این ساختار، LSTM کلمه به تحلیل و یادگیری ویژگی‌های کلمات و نحوه ترکیب آن‌ها در جملات کمک می‌کند، در حالی که LSTM جمله می‌تواند ساختار کلی جملات و ارتباط بین آن‌ها را در یک متن طولانی‌تر درک کند [۲].
Harzig و همکاران [۳] برای جلوگیری از سوگیری مدل نسبت به داده‌ها، از LSTM های دو کلمه‌ای استفاده کردند. در این روش، یکی از کلمات، غیرطبیعی و دیگری مربوط به کلمات معمولی بود همچنین، آنها یک پیش‌بینی‌کننده جملات غیرعادی تنظیم کردند تا تعیین شود آیا از جملات تولید شده توسط LSTM دو کلمه‌ای استفاده شود یا خیر. Yang و همکاران [۴] مدلی را پیشنهاد دادند که ابتدا ویژگی تصاویر اولتراسوند را با استفاده از یک شبکه کانولوشنی استخراج و سپس از این ویژگی‌ها برای پیش بینی خواص محلی تصاویر بوسیله یک مدل دسته بندی چند برچسبی[۲] استفاده می‌کردند. در ادامه، ویژگی‌های خروجی شبکه کانولوشنی و خواص محلی تلفیق می‌گردد و به یک مدل LSTM مبتنی بر توجه برای ساخت گزارش منتقل می‌شوند.
برای برچسب زدن جفت تصاویر و گزارش‌های پزشکی، Han و همکاران چارچوبی با نظارت ضعیف پیشنهاد کردند که این چارچوب بدون نیاز به گزارش رادیولوژیست، گزارش‌های یکپارچه تولید می‌کند [۵]. به طور مشابه، Xue و همکاران مدلی تکرارشونده برای توصیف تصاویر طراحی کردند که یافته‌های گزارش پزشکی را جمله به جمله تولید می‌کند، که در آن هر جمله متوالی براساس ورودی‌های چندوجهی، از جمله اصلی و جمله قبلی است [۶]. Li و همکاران [۷] اولین مدل بازیابی را با یک شبکه عصبی مولد معرفی کردند که این مدل ویژگی‌های بصری اشعه‌ایکس قفسه‌سینه را از آخرین لایه کانولوشن استخراج می‌کند و با تقویت RNN با مکانیزم توجه، تولید متن را بهبود می‌بخشد.
Shi و همکاران [۸] با پیشنهاد مدل AIMNet که حاوی یک مکانیسم ادغام مبتنی بر اهمیت و هدایت برچسب‌های بیماری است، با هدف کاهش مشکل سوگیری داده‌ها به منظور تولید گزارش‌های پزشکی دقیق در مطالعه خود از مدلی استفاده کردند که ابتدا ویژگی‌های تصاویر اشعه‌ایکس قفسه‌سینه به وسیله یک شبکه ResNet50 استخراج کردند و تگ‌های بیماری را با یک شبکه چندبرچسبی پیشبینی و سپس اهمیت ویژگی‌های تصویر و ویژگی تگ‌ها به وسیله یک دروازه ادغام تطبیقی[۳] در تولید هر جمله تنظیم کردند.
همچنین پیشرفت‌های اخیر در پردازش زبان طبیعی زیست‌پزشکی، با معرفی مدل‌های زبانی پیش‌آموزش‌دیده مانند BioBERT و BioGPT، امکانات جدیدی برای تحلیل و تولید متون پزشکی فراهم کرده است. BioBERT، توسط Lee و همکاران [۹] معرفی‌شده ، که یک مدل مبتنی بر BERT است که با پیش‌آموزش روی ۴/۵ میلیارد کلمه از چکیده‌های PubMed ، ۱۳/۵ میلیارد کلمه از مقالات کامل PubMed Central و داده‌های

[2] multi-label classification

[3] Adaptive Merging Gate

عمومی Wikipedia و BooksCorpus برای وظایف درک زبان زیست‌پزشکی بهینه شده است. این مدل با بهبود ۰/۶۲ درصد در امتیاز F1 برای شناسایی موجودیت‌های نام‌دار[۴]، ۸۰/۲ درصد برای استخراج روابط[۵]، و ۱۲/۲۴ درصد در میانگین رتبه معکوس[۶] برای پرسش و پاسخ[۷] نسبت به مدل‌های پیشرفته قبلی، توانایی بالایی در پردازش متون پیچیده پزشکی نشان داده است . با این حال، BioBERT به دلیل تمرکز بر وظایف تفکیکی[۸]، برای تولید متن‌های مولد مانند گزارش‌های رادیولوژی محدودیت دارد. در مقابل، BioGPT، توسط Luo و همکاران [۱۰] ارائه‌شده، یک مدل مولد مبتنی بر ترنسفورمر[۹] است که با آموزش از ابتدا روی ۱۵ میلیون چکیده PubMed و واژگان تخصصی طراحی شده است. BioGPT با معماری GPT-2 متوسط با ۳۴۷ میلیون پارامتر و مدل‌سازی زبانی خود-بازگشتی، در وظایف مولد مانند استخراج رابطه سرتاسری[۱۰] با ۹۸/۴۴ درصد امتیاز F1 در مجموعه داده زیست‌پزشکی برای روابط شیمیایی-بیماری[۱۱] و پرسش و پاسخ دقت ۷۸/۳ درصد در پاسخ به سؤالات [۱۲] PubMed عملکرد برتری نسبت به سایر مدل‌های پایه ارائه می‌دهد. این مدل با تولید جملات طبیعی و حذف نیاز به حاشیه‌نویسی میانی، پتانسیل بالایی برای تولید گزارش‌های رادیولوژی دقیق و روان دارد، هرچند برای پردازش ویژگی‌های بصری نیازمند یکپارچگی با مدل‌های بینایی است.

مروری به مطالعات فوق نشان می‌دهد یکی از چالش‌های اصلی در این زمینه، تولید گزارشات دقیق و جامع از تصاویر پزشکی تنها با استفاده از تصویر و یک گزارش در دسترس بدون استفاده از کلاس تصاویر استخراج شده توسط متخصص است. بنابراین در این مطالعه با در نظر گرفتن چالش‌های اشاره شده اقدام به معرفی مدل جدید مبتنی بر ساختار کانولوشنی بازگشتی با در نظر گرفتن مکانیزم توجه در بخش رمزگشا نموده‌ایم. این مدل که Res-LSTM-Attn نام گذاری می‌گردد؛ علاوه بر این که قادر است ویژگی معنایی تصاویر را بدون نیاز به کلاس تصاویر و با استفاده از یک مدل چند برچسبی تطبیق تصاویر-کلمات استخراج کند، دارای توانایی تولید پاراگراف طولانی به طور مکرر و کلمه به کلمه نیز خواهد بود. به منظور ارزیابی عملکرد مدل پیشنهادی، آن را بر روی مجموعه داده IU X-Ray اعمال نمودیم. این دادگان حاوی تصاویر اشعه‌ایکس قفسه‌سینه است که در کنار هر تصویر گزارشات پزشکی به عنوان معیار استاندارد برای ارزیابی عملکرد مدل در اختیار قرار دارد.

ساختار مقاله حاضر به شرح زیر است: ابتدا در بخش۲، مراحل مختلف الگوریتم روش پیشنهادی شرح داده شده است. در بخش۳، مراحل اجرای الگوریتم و نتایج آزمایش ارائه شده به همراه بحث و بررسی آنها ارائه خواهد شد. در نهایت بخش ۴ نیز به جمع بندی مطالعه اختصاص یافته است.

۲- مواد و روش

۲-۱- معرفی مجموعه داده

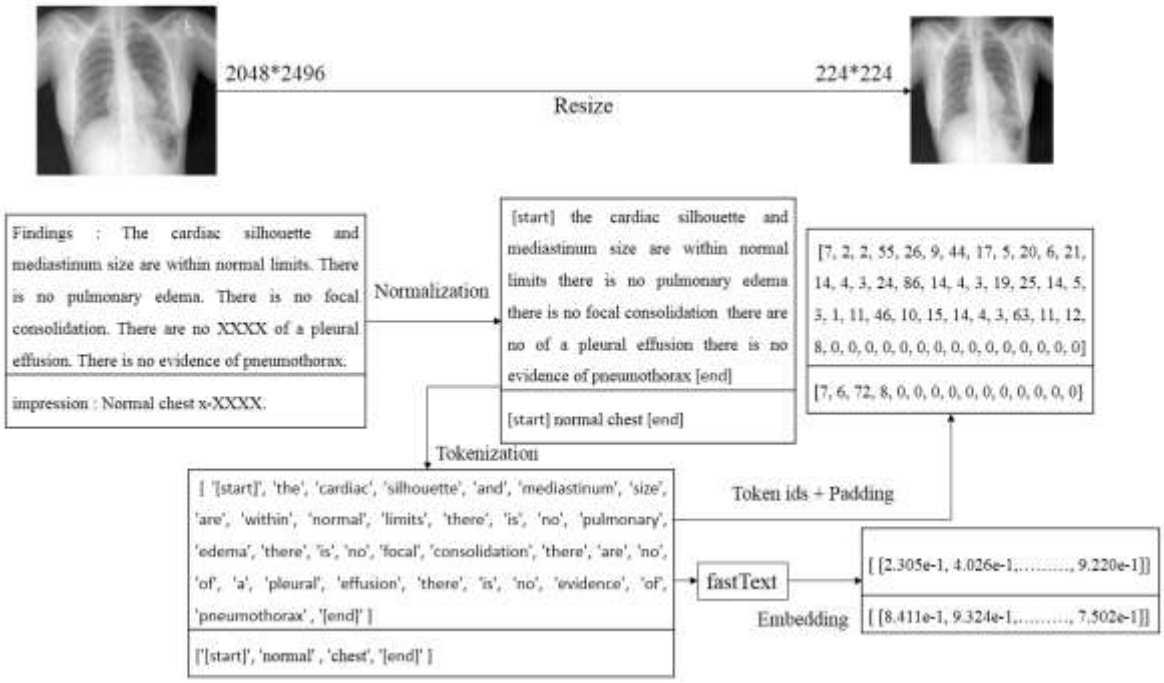
در این پژوهش از مجموعه داده IU X-Ray استفاده شده این مجموعه داده یکی از رایج ترین مجموعه داده‌های مورد استفاده در زمینه تولید گزارش برای تصاویر پزشکی است که شامل ۷۴۷۰ تصویر رادیوگرافی قفسه‌سینه است که برای هر بیمار از نمای جلویی و جانبی تشکیل شده است. به طور معمول، گزارش‌های تصاویر رادیولوژی برای مسئله تولید گزارش شامل چندین بخش ثابت به شرح زیر است:

- بخش نشانه[۱۳]، دلیلی برای درخواست اشعه‌ایکس است.
- بخش یافته‌ها[۱۴]، شامل توضیحات مشاهدات رادیولوژیست از تصویر است.
- بخش برداشت[۱۵]، خلاصه‌ای از یافته‌های مرتبط برای تشخیص احتمالی یا تعیین مراحل بعدی ارائه می‌دهد.

هر گزارش معمولاً با یک یا چند تصویر اشعه‌ایکس قفسه‌سینه که از نماهای مختلف، مانند جلو یا جانبی گرفته شده است، مرتبط است. هدف روش‌های تولید گزارش رادیولوژی معمولاً تولید بخش یافته‌ها، بخش برداشت یا هر دو است. مجموعه داده اشاره شده شامل ۳۹۵۵ گزارش رادیولوژی است که هر کدام با تصاویر اشعه‌ایکس از جلو و/یا جانبی قفسه‌سینه و در مجموع ۷۴۷۰ تصویر با ابعاد ۲۰۴۸ در ۲۴۹۶ مرتبط است. در این پژوهش، ما تصاویر بدون دو نمای تصویر کامل یا بدون بخش‌های یافته‌ها و برداشت حذف گردیده است، که منجر به ایجاد مجموعه داده‌ای

[4] Named Entity Recognition
[5] Relation Extraction
[6] Mean Reciprocal Rank
[7] Question Answering
[8] discriminative
[9] Transformer
[10] end-to-end relation extraction
[11] Biomedical Corpus for Chemical-Disease Relations
[12] PubMed Question Answering
[13] indication
[14] findings
[15] impression

کوچک‌تر با ۲۸۶۷ گزارش مرتبط با تصاویر از نمای جلو شد. این حذف برای اطمینان از وجود داده‌های کامل برای آموزش مدل Res-LSTM-Attn، که به‌منظور تولید همزمان یافته‌ها و برداشت طراحی شده، ضروری بود. با این حال، حذف گزارش‌های ناقص ممکن است به سوگیری در داده‌ها منجر شده باشد، زیرا این گزارش‌ها می‌توانستند شامل موارد بالینی خاص، مانند بیماری‌های نادر یا ناهنجاری‌های غیرمعمول، باشند. ابعاد تصاویر پس از انجام پیش‌پردازش به ۲۲۴ در ۲۲۴ پیکسل تغییر می‌کند. در بخش پیش پردازش تمامی کلمات معنا دار در" یافته‌ها" و " برداشت" نشانه گذاری می‌شوند؛ اما کلمات که اطلاعات شخصی افراد بودند، در کنار اعداد، علائم نگارشی، حروف یونانی و اختصار نیز حذف کرده گردیده‌اند. این کار با الهام از [۱۱] صورت گرفته است که در نهایت منجر به ایجاد ۱۶۴۵ کلمه منحصر به فرد گردید. علاوه بر این، دو نشانه ویژه [start] و [end] مطابق بخش نرمالسازی شکل (۱) اضافه می‌شوند که شبکه بتواند شروع و پایان جمله را تشخیص دهد. سپس کلمات را جدا و به رمزهای عددی مطابق بخش توکن سازی[16] و توکن آیدی[17] شکل (۱) تبدیل می‌کنیم و اندازه تمام جملات با استفاده از لایه گذاری[18] صفر به اندازه طول بزرگترین جمله تبدیل می‌شوند. همچنین تمام مراحل بالا بطور جداگانه برای هر بخش یافته‌ها و برداشت صورت می‌پذیرد که به ترتیب برای هر بخش ۱۴۰۰ و ۱۰۷۰ کلمه منحصر به فرد به‌دست می‌آید. حداکثر تعداد کلمات یک گزارش برای گزارش‌های که از دوبخش یافته‌ها و برداشت تشکیل شده‌اند، ۱۶۳ و برای گزارش‌های که فقط از یافته‌ها و فقط از برداشت تشکیل شده‌اند، به ترتیب برابر ۱۴۵ و ۱۰۷ کلمه است. همچنین به منظور ارزیابی مدل پیشنهادی، به طور تصادفی ۲۸۰ گزارش (۱۰٪) را برای تشکیل مجموعه آزمایشی انتخاب کردیم و تمام ارزیابی‌ها را روی مجموعه آزمایشی انجام می‌دهیم.



شکل ۱- مراحل پیش پردازش گزارش‌ها و تصاویر

۲-۲- مدل پیشنهادی Res-LSTM-Attn

ساختار مدل پیشنهادی این مطالعه بر اساس معماری رمز گذار-رمزگشا طراحی شده است. در این مدل برای استخراج ویژگی و بخش رمزگذار شبکه کانولوشنی Resnet50 استفاده شده است. در ادامه، از شبکه LSTM بعنوان معماری بازگشتی در بخش رمزگشا استفاده گردیده است. بخش رمزگشا وظیفه اصلی تولید کلمات و درک ویژگی‌های معنایی آنها را به عهده دارد. در نتیجه این بخش به لایه‌های مبتنی بر مکانیسم توجه مجهز گردیده تا بتواند درک سطح بالاتری نسبت به ارتباطات معنایی بین کلمات ایجاد نماید.

[16] Tokenization
[17] Token ids
[18] padding

از آنجایی که ساختار کلی مدل رمزگذار-رمزگشا حاصل ترکیب دو شبکه کانولوشنی و حافظه محور در کنار معماری مبتنی بر مکانیسم توجه است، نام Res-LSTM-Attn به آن اطلاق می گردد. نمای کلی این مدل برای تولید گزارش تصویر پزشکی در شکل (۲) ارائه شده است. چارچوب مدل شامل چهار بخش است:

۱) رمزگذار تصویر برای استخراج ویژگی‌های بصری از تصویر اشعه‌ایکس قفسه‌سینه.

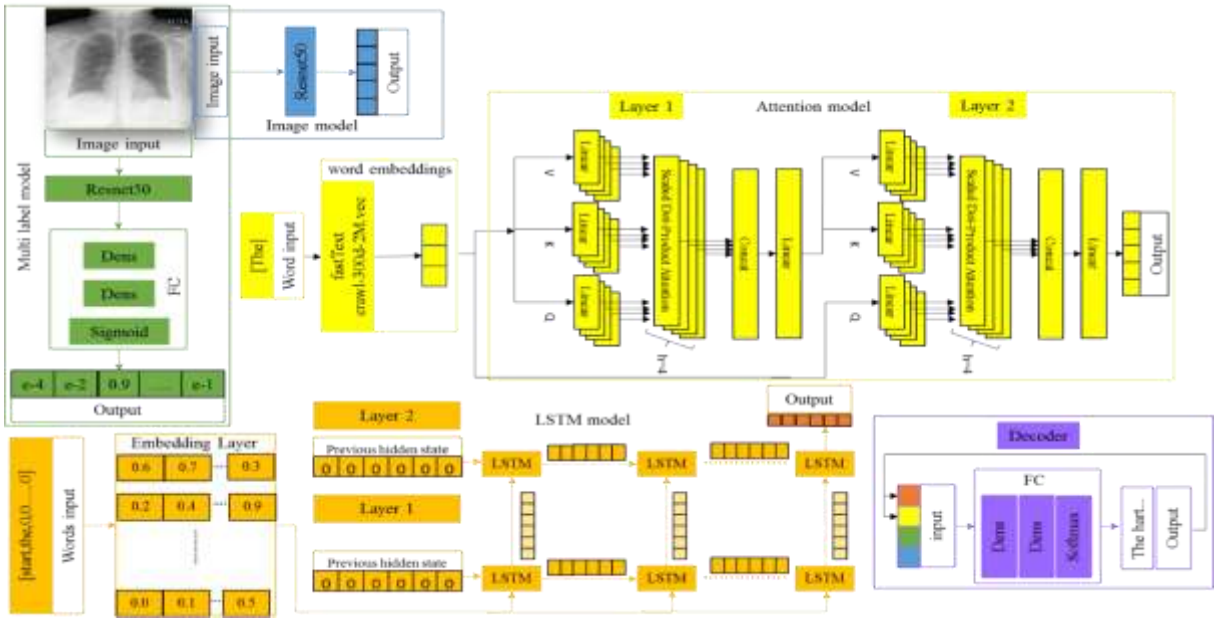
۲) یک مدل چند برچسبی برای پیش بینی لغات یک گزارش از ویژگی‌های بصری استخراج شده.

۳) مدل زبانی LSTM که کلمات پیشبینی شده در زمان قبلی مدل را به عنوان ورودی می‌گیرد و خروجی این بخش به همراه ویژگی سایر بخش‌ها به لایه‌های تمام متصل داده می‌شود.

۴) مدل زبانی با لایه‌های MultiHead self-Attention که ویژگی‌های خروجی مدل fastText برای اخرین کلمه پیشبینی شده مدل را بعنوان ورودی می‌گیرد. fastText یک مدل پیش‌آموزش‌شده برای بردارسازی کلمات است که بردارهای کلمات تولید شده توسط fastText می‌توانند در طیف گسترده‌ای از کاربردهای پردازش زبان طبیعی، از طبقه‌بندی متن گرفته تا تجزیه و تحلیل احساس، به کار گرفته شوند. خروجی این بخش بعنوان ویژگی معنایی کلمات همانند بخش ۳ به بخش تمام متصل داده می‌شود تا گزارش پزشکی را کلمه به کلمه تولید می‌کند.

معماری پیشنهادی از چند بلوک مختلف تشکیل شده است که در شکل (۲) نشان داده شده‌اند.

بخش Multi label model از یک شبکه عصبی کانولوشن برای استخراج ویژگی‌های بصری از تصویر اشعه‌ایکس قفسه‌سینه استفاده می‌کند. این ویژگی‌ها به یک مدل طبقه‌بندی چند برچسبی داده می‌شوند تا لغات گزارش مربوط به تصویر را پیش‌بینی کند. خروجی این بخش به عنوان ورودی مدل تولید گزارش استفاده می‌شود. بخش Image model از یک شبکه ResNet50 برای استخراج ویژگی‌های بصری از تصویر اشعه‌ایکس استفاده می‌کند. این ویژگی‌ها در مدل تولید گزارش استفاده می‌شوند. بخش Attention model از یک مدل توجه معنایی استفاده می‌کند که از ویژگی‌های معنایی خروجی مدل FastText بهره می‌برد تا اطلاعات معنایی مربوط به تصویر را استخراج کند. بخش LSTM model یک مدل زبانی LSTM است که بر اساس کلمات تولید شده تا کنون کلمات بعدی را پیش‌بینی می‌کند. بخش FC لایه‌های تمام متصل که با ترکیب اطلاعات بصری از مدل‌های Multi label model و Image model، اطلاعات معنایی از Attention model و اطلاعات زبانی از LSTM model، گزارش تصویر پزشکی را کلمه به کلمه تولید می‌کنند. این ترکیب اطلاعات بصری، معنایی و زبانی در بخش FC به مدل کمک می‌کند تا گزارش‌های دقیق‌تری از تصاویر اشعه‌ایکس قفسه‌سینه تولید کند.



شکل ۲- ساختار کلی مدل Res-LSTM-Attn

۲-۲-۱- مدل رمزگذار تصویر

ما در این مدل از شبکه ResNet50 مطابق شکل (۲) بخش Image model استفاده کرده‌ایم و ثابت شده است که یکی از کارآمدترین استخراج ویژگی و شبکه‌های عصبی کانولوشن رمزگذار است و بر بسیاری از اشکالاتی کهCNN های سنتی با آن مواجه هستند، غلبه می‌کند [۱۲]. افزایش تعداد لایه‌ها در شبکه‌های عصبی موجب افزایش دقت در مدل می‌شود، اما مشکل از بین رفتن گرادیان‌ها را ایجاد می‌کند درResNet50 ، همراه با اتصال به هر لایه بعدی، یک اتصال جهشی[19] از ۲ یا ۳ لایه قبل وجود دارد. این اتصال جهشی متفاوت از شبکه‌های عصبی سنتی است و به بهبود عملکرد مدل کمک می‌کند. به طور کلی، این اتصالات جهشی یک معماری متفاوت از شبکه‌های عصبی سنتی را فراهم می‌کنند و به بهبود عملکرد مدل در مسائل پیچیده کمک می‌کنند. از خروجی شبکه resnet50 در مدل چند برچسبی و در مدل تولید گزارش بعنوان ویژگی‌های بصری استفاده شده است.

۲-۲-۲- مدل چندبرچسبی

در مسئله طبقه‌بندی چند برچسبی، هر نمونه داده ممکن است با چندین برچسب مرتبط باشد. مقدار پیش‌بینی‌شده هر گره خروجی نشان‌دهنده احتمال تعلق ورودی به آن کلاس است. پیکربندی شبکه طبقه‌بندی‌کننده چند کلاسه در شکل ۲ بخش Multi label model نشان داده شده است. ورودی طبقه‌بندی‌کننده تصویر ورودی و خروجی‌ها کلمات گزارش مرتبط هستند که توسط پزشکان نوشته شده‌اند. همانطور که در شکل (۲) نشان داده شده است، ابتدا ویژگی‌های تصویر با استفاده از شبکه ResNet50 استخراج می‌شوند. سپس این ویژگی‌ها به دو لایه تمام متصل داده می‌شوند. خروجی‌های نهایی، احتمال تعلق تصویر ورودی به هر یک از کلمات دیکشنری است. این بخش به‌صورت جداگانه با تابع هزینه آنتروپی متقاطع دودویی[20] مطابق رابطه زیر آموزش می‌بیند و خروجی آن بعنوان ورودی در هنگام تولید گزارش به مدل تمام متصل انتقال داده می‌شود.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i . log(\hat{y}_i) + (1-y_i).log\,(1-\hat{y}_i)] \quad (۱)$$

که در آن y_i برچسب واقعی برای هر کلمه در دیکشنری، $\hat{y}_i$ احتمال پیش‌بینی‌شده، و N تعداد کلمات در دیکشنری است. تابع هزینه آنتروپی متقاطع دودویی برای طبقه‌بندی چندبرچسبی مناسب است، زیرا هر تصویر ممکن است با چندین کلمه مرتبط باشد، و به پیش‌بینی دقیق کلمات کلیدی کمک کرد.

۲-۲-۳- مدل LSTM

در این مدل، از یک ساختار LSTM دولایه مطابق شکل (۲) بخش LSTM model استفاده می‌شود. در این ساختار، دو لایه LSTM به صورت سری قرار داده می‌شوند. حالت‌های پنهان لایه اول به عنوان ورودی لایه دوم استفاده می‌گردد. این ساختار چندلایه به مدل امکان می‌دهد تا ویژگی‌های سطوح مختلف را استخراج و در نهایت به یک بردار بهینه‌تری از ویژگی‌ها دست یابد.با حالت پنهان h_t^1 و سلول حافظه m_t^1 ، فرآیند اولین LSTM مطابق Layer 1 بخش LSTM model شکل(۲) را می‌توان به صورت رابطه زیر نشان داد

$$(h_t^1, m_t^1) = LSTM_1([W_e, \bar{x} + C_{t-1}], (h_{t-1}^1, m_{t-1}^1)) \quad (۲)$$

که در آن W_e ماتریس جاسازی کلمات جمله است

$\bar{x} = \frac{1}{k}\sum_i x_i$ نشان‌دهنده ادغام میانگین ویژگی x است و برای اطلاعات کلی ارائه شده به بردار زمینه C_{t-1} اضافه می‌شودLSTM های دولایه همیشه ماژول توجه را روی ویژگی متن X قرار می‌دهد و بردار متن C_t را تولید می‌کند، به این معنا که:

$$\hat{a}_t = f_{att}(h_t^1, X) \quad (۳)$$

که در آن $f_{att}(.)$ نشان دهنده ماژول توجه است .سپس، $\hat{a}_t$ با h_t^1 به دومین LSTM تغذیه می‌شود، یعنی،

$$(h_t^2, m_t^2) = LSTM_2([\hat{a}_t, h_t^1], (h_{t-1}^2, m_{t-1}^2)) \quad (۴)$$

در این زمان، بردار زمینه خروجی پنهان m_t^2 است. در نهایت، خروجی LSTM دوم برای پیش‌بینی توزیع احتمال به softmax داده می‌شود، یعنی،

$$p(Z_t|Z_{1:t-1}) = softmax(C_t W_p + b_p) \quad (۵)$$

[19] skip connection

[20] Binary Cross entropy

در مقایسه با LSTM منفرد، LSTM دو لایه با مکانیسم‌های توجه، نمایش قدرتمندتری دارد و نشان‌دهنده جدیدترین مدل‌های زبانی پیشرفته قبل از تولید توصیف‌های مبتنی بر ترانسفورمر است [۱۳].

۲-۲-۴- مدل توجه

در بخش Attention model شکل (۲) از مکانیسم خودتوجهی چند سر استفاده کرده‌ایم که به نتایج عالی در ترجمه ماشینی دست می‌یابد .مکانیسم توجه به خود می‌تواند داده‌های ورودی را در فواصل طولانی یاد بگیرد تا توجه را در حین محاسبات موازی بهبود بخشد . مکانیسم توجه به خود نقش مهمی در ترجمه ماشینی، سیستم مکالمه و سایر زمینه‌ها ایفا می‌کند و می‌تواند مشکلات استفاده ناکافی از اطلاعات از راه دور، ناپدید شدن گرادیان و انفجار مدل RNN را به دلیل ویژگی‌های موازی بالا و وابستگی به اطلاعات در فواصل طولانی حل کند. همانطور که در شکل(۲) نشان داده شده است، مدل خودتوجهی چند سر از چندین ماژول توجه نقطه-محصول مقیاس شده تشکیل شده است که در کنار هم قرار گرفته‌اند . ماتریس ورودی حاوی $Q \in R^n$، $K \in R^{n*d_k}$ و $V \in R^{n*d_v}$ است. توجه را می‌توان به صورت زیر محاسبه کرد:

$$attention(q_i, k, v) = \sum_{j=1}^{m} softmax(\frac{q_i . k_j^T}{\sqrt{d_k}})\, v_j$$
$$attention(Q, K, V) = softmax(\frac{Q.K^T}{\sqrt{d_k}})V \quad (۶)$$

که در آن d بعد داده‌های ورودی است. ابعاد ماتریس‌های ورودی Q, K و V با توجه به اینکه توجه چند سر مکانیسم توجه به خود را اتخاذ می‌کند، یکسان است. همانطور که در معادله نشان داده شده است، بردارهای Q, K و V د ر مدل توجه چند سر ابتدا به صورت خطی تبدیل می‌شوند. محاسبه خود توجه بر روی ورودی q_i انجام می‌شود: ضرب q_i با هر k_j^T برای محاسبه شباهت و به دست آوردن وزن، و سپس با تقسیم بر $\sqrt{d_k}$ برای جلوگیری از صفر شدن گرادیان مقیاس[21] می‌شود. تابع $softmax$ برای نرمال سازی وزن اعمال می‌شود .در نهایت، از مجموع وزنی وزن‌ها و مقادیر کلیدی متناظر برای به دست آوردن ارزش توجه یک بلوک استفاده می‌شود و مقدار توجه سر پس از محاسبه nتوجه به دست می‌آید .توجه نقطه-محصول مقیاس بندی در معادله (۶-۳) بیان می‌شود N سر به دست می‌آید و سر هر زمان پس از N بار محاسبه توجه به هم متصل می‌شود . نتیجه نهایی با همان ابعاد ورودی از طریق تبدیل خطی به صورت زیر به‌دست می‌آید:

$$head_i = attention(QW_i^Q, KW_i^K, VW_i^V)$$
$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_N)W^o \quad (۷)$$

که در آن $W_i^Q \in R^d\ model^{d_k}$، $W_i^K \in R^d\ model^{d_k}$، $W_i^V \in R^d\ model^{d_k}$ ماتریس پارامتر نگاشت خطی است.

$W^o \in R^d\ model^{d_k}$ وزن تبدیل خطی است، Concat عملیات اتصال برداری است، و MultiHead (Q, K, V) نتایج نهایی محاسبه از طریق سر چندگانه است.

محاسبات مکانیسم توجه چند سر می‌تواند اطلاعات معنایی اضافی را از فضاهای مختلف بیاموزد. مدل ما از دو لایه Multi-head self-attention بصورت شکل (۲) بخش Attention model استفاده می‌کند که از خروجی مدل FastText برای کلمه پیشبینی شده حال حاضر تغذیه می‌کند.

۲-۲-۵- مدل تمام متصل

خروجی بخش‌های قبل با یکدیگر ادغام شده بعنوان ورودی ترکیبی از طریق یک لایه کاملاً متصل متشکل از۲۰۴۸ نورون و به یک لایه softmax مطابق بخش FC شکل (۲) داده می‌شود که اندازه آن با تعداد کلمات موجود در واژگان مطابقت دارد. در طول آزمایش، کلمه جاسازی گزارش اصلی در دسترس نیست .برای راه‌اندازی تولید گزارش، نشانه [start]در کنار جاسازی تصویر و خروجی مدل چند برچسبی وارد می‌شود. مدل یک کلمه را در خروجی بصورت حریصانه تولید میکند. سپس به همین ترتیب یک توالی کامل از کلمات پیش بینی شده ایجاد می‌شود:

$$logp(X_{0:T}|M, I; \theta) = \sum_{t=0}^{T} P(x_t|M, I, x_{t-1}, x_{0:t-1}; \theta) \quad (۸)$$

که I رمزگذاری تصویر، M خروجی مدل چندبرچسبی و θ پارامترهای مدل را نشان می‌دهد. متغیر $X_{0:T}$ بیانگر

[21] scale

کلماتی است که در موقعیت‌های ۰ تا T ایجاد شده‌اند و x_{t-1} بیانگر اخرین کلمه تولید شده است. کلمه با بیشترین احتمال بعنوان کلمه منتخب گزارش انتخاب می‌شود. مدل تولید گزارش با آنتروپی متقاطع دسته‌ای[۲۲] مطابق رابطه زیر آموزش می‌بینند:

$$CE = -\sum_{i=1}^{C} y_i . log(\hat{y}_i) \quad (۹)$$

که در آن y_i توزیع واقعی (one-hot) برای کلمه صحیح، $\hat{y}_i$ توزیع پیش‌بینی‌شده توسط لایه softmax، و C تعداد کلمات در واژگان است. آنتروپی متقاطع برای مسائل تولید توالی (مانند پیش‌بینی کلمه بعدی) استاندارد است، زیرا مدل را برای حداکثر کردن احتمال کلمه صحیح در هر مرحله آموزش می‌دهد.

۳- یافته‌ها و بحث

برای پیاده سازی مدل‌های معرفی شده از زبان برنامه نویسی پایتون و پکیج تنسورفلو اخرین نسخه (۲.۱۶.۲) تحت محیط برنامه نویسی Google Colab pro با GPU GB ۱۶ T4 و RAM ۲۵ GB استفاده گردید.

در مدل Res-LSTM-Attn ابتدا بخش کلاس بندی چند برچسبی را با استفاده از تابع خطا آنتروپی متقاطع دودویی و تابع بهینه سازی Adam برای ۳۵ دوره با نرخ یادگیری اولیه ۰/۰۰۱ آموزش می‌بینند و مدل در جایی که کمترین خطا برای داده‌های Validation وجود دارد ذخیره و از ان استفاده می‌کنیم. همچنین مدل تولید گزارش را با تابع خطا آنتروپی متقاطع و تابع بهینه سازی Adam برای ۱۰ دوره با نرخ یادگیری اولیه ۰/۰۰۱ آموزش می‌دهیم.

ساختار رمزگشا از یک LSTM دولایه با ۲۵۶ نرون در هر لایه و مکانیسم توجه چندسر با ۴ سر تشکیل شده است. تعداد لایه‌ها از طریق آزمایش با ساختارهای تک‌لایه، دولایه، و سه‌لایه تعیین شد. ساختار تک‌لایه برای مدل‌سازی توالی‌های طولانی گزارش‌های پزشکی دقت کافی نداشت و CIDEr پایین‌تری کسب کرد. ساختار سه‌لایه به بیش‌برازش در مجموعه داده IU X-Ray منجر شد و زمان آموزش را افزایش داد. دولایه، با انتقال حالت‌های پنهان، بهترین عملکرد را در CIDEr و ROUGE-L داشت، مشابه [۱] و [۶] تعداد ۲۵۶ نرون از آزمایش با ۱۲۸ و ۵۱۲ نرون انتخاب شد؛ ۱۲۸ نرون ظرفیت محدودی داشت و ۵۱۲ نرون پیچیدگی را بدون بهبود قابل‌توجه افزایش داد. تعداد ۴ سر توجه از آزمایش با ۲ و ۸ سر انتخاب شد، که ۴ سر با ابعاد جاسازی کلمه[۲۳] ۳۰۰ بهینه بود و ۸ سر بدون اثر قابل‌توجه در معیار ها، زمان محاسبات را افزایش داد. به دلیل زمان‌بر بودن آموزش، آزمایش‌ها یک‌بار انجام شدند و نتایج عددی ذخیره نشدند، اما مقادیر بهینه با معیارهای جدول ۱ تأیید شدند.

جهت ارزیابی عملکرد مدل از امتیازات ROUGE-L[۲۴] [۱۴]، CIDEr[۲۵] [۱۵] و BLEU[۲۶] [۱۶] استفاده گردیده است که معیارهای مرسوم و پر استناد برای مطالعات این حوزه هستند. مدلها را در حالتی که گزارش شامل برداشت و یافته‌ها است با مدل مرجع [۱۷] مقایسه می‌کنیم که در آن از ویژگی‌های بصری مختلف استخراج شده از یک ResNet۱۵۲ بعنوان رمزگذار و از رمزگشا معماری ترانسفورمر برای تولید گزارش استفاده شده است. علاوه بر این، از مدل‌های مرسوم تولید گزارش تصاویر نیز برای مقایسه استفاده گردیده است.

در این پژوهش، مدل Res-LSTM-Attn را در شرایطی که گزارش شامل بخش "یافته‌ها" است با مدل‌های پژوهش‌های قبلی CLARA [۲۳]، CMAS [۲۴] و RTMIC [۲۵] مقایسه کرده‌ایم. سپس، برای ارزیابی بیشتر عملکرد مدل، از آن برای تولید گزارشات بخش "برداشت" استفاده کرده‌ایم. این بخش در پژوهش‌های پیشین کمتر مورد بررسی قرار گرفته و تنها با مدل CMAS [۲۴] مقایسه شده است.

[22] Categorical Cross-Entropy
[23] Word Embedding
[24] Recall-Oriented Understudy for Gisting Evaluation –Longest common subsequence
[25] Consensus-based Image Description Evaluation rater
[26] Bilingual Evaluation Understudy

جدول ۱- مقایسه نتایج مدل پیشنهادی با سایر پژوهش‌ها

Paper	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
یافته‌ها + برداشت						
Huang et al.[20]	۰/۴۷۶	۰/۳۴۰	۰/۲۳۸	۰/۱۶۹	۰/۳۴۷	۰/۲۹۷
Singh et al.[19]	۰/۳۷۴	۰/۲۲۴	۰/۱۵۳	۰/۱۱۰	۰/۳۰۸	۰/۳۶۰
mDiNAP-transformer-ewp[17]	۰/۳۷۳۱	۰/۲۲۶۰	۰/۱۴۷۳	۰/۱۰۱۰	۰/۲۹۳۰	۰/۳۱۹۱
CDGPT2[18]	۰/۳۸۷	۰/۲۴۵	۰/۱۶۶	۰/۱۱۱	۰/۲۸۹	۰/۲۵۷
Xue et al.[6]	۰/۴۶۴	۰/۳۵۸	۰/۲۷۰	۰/۱۹۵	۰/۳۳۶	-
A3FN[21]	۰/۴۴۳	۰/۳۳۷	۰/۲۳۶	۰/۱۸۱	۰/۳۴۷	۰/۳۷۴
Zhang et al.[11]	۰/۴۴۱	۰/۲۹۱	۰/۲۰۳	۰/۱۴۷	۰/۳۶۷	۰/۳۰۴
Harzig et al.[3]	۰/۳۷۳	۰/۲۴۶	۰/۱۷۵	۰/۱۲۶	۰/۳۱۵	۰/۳۵۹
Yin et al.[22]	۰/۴۴۵	۰/۲۹۲	۰/۲۰۱	۰/۱۵۴	۰/۳۱۵	۰/۳۴۲
Res-LSTM-Attn	۰/۴۵۵۸	۰/۳۰۸۰	۰/۲۱۹۱	۰/۱۵۵۸	**۰/۳۹۰۰**	**۰/۴۰۱۰**
یافته‌ها						
KERP[26]	۰/۴۸۲	۰/۳۲۵	۰/۲۲۶	۰/۱۶۲	۰/۳۳۹	۰/۲۸۰
RTMIC[25]	۰/۳۵۰	۰/۲۲۴	۰/۱۴۳	۰/۰۹۶	-	۰/۳۲۳
CLARA[23]	**۰/۴۷۱**	**۰/۳۲۴**	**۰/۲۱۴**	**۰/۱۹۹**	-	۰/۳۵۹
HRGR[7]	۰/۴۳۸	۰/۲۹۸	۰/۲۰۸	۰/۱۵۱	۰/۳۲۴	۰/۳۴۳
CMAS[24]	۰/۴۶۴	۰/۳۰۱	۰/۲۱۰	۰/۱۵۴	**۰/۳۶۲**	۰/۲۵۷
Res-LSTM-Attn	۰/۴۲۷۵	۰/۲۸۷۰	۰/۲۰۳۸	۰/۱۴۷۸	۰/۳۵۳۴	**۰/۴۱۷۰**
برداشت						
CMAS[24]	۰/۴۰۱	۰/۲۹۰	۰/۲۲۰	۰/۱۶۶	۰/۵۲۱	۱/۴۵۷
Res-LSTM-Attn	**۰/۵۶۲۴**	**۰/۴۹۶۴**	**۰/۴۲۱۱**	**۰/۳۲۷۱**	**۰/۶۰۴۴**	**۲/۳۴۱۹**

جدول ۱ نتایج پیاده سازی مدل این مطالعه در مقایسه با دیگر مطالعات را به تفصیل بیان می کند. همانگونه که پیداست، مدل ارائه شده برای تولید گزارش یافته‌ها + برداشت از همه مدل‌های پایه بهتر عمل می‌کند و بهترین نمرات CIDEr، ROUGE را کسب کرده است. برای تولید گزارش یافته‌ها، مدل پیشنهاد شده نسبت به مدل CMAS از نظر معیار ROUGE عملکرد ضعیف تری داشته اما همچنان بهترین نمره CIDEr را کسب کرده است. همچنین می توان گفت، مدل ارائه شده در مقایسه با مدل CMAS برای تولید گزارش بخش برداشت بسیار عملکرد

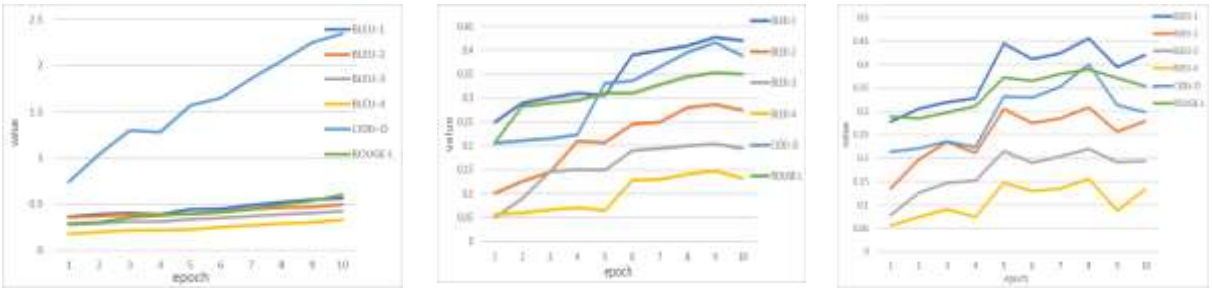
بهتری از منظر تمامی معیارهای ارزیابی داشته است. در ارزیابی کیفیت گزارش‌های خودکار در حوزه پزشکی، علاوه بر صحت[۲۷]، پوشش[۲۸] نیز اهمیت زیادی دارد. امتیازات BLEU که میزان سازگاری گزارش خودکار را با گزارش اصلی انسانی اندازه‌گیری می‌کنند، تنها بر روی صحت تمرکز دارند و پوشش را در نظر نمی‌گیرند. در واقع، گزارش‌های خودکار ممکن است اطلاعات کلیدی در مورد بیماری را از دست بدهند، اما همچنان به امتیازات BLEU بالایی دست یابند. این موضوع نشان می‌دهد که صحت به تنهایی برای ارزیابی کیفیت این گزارش‌ها در کاربردهای بالینی کافی نیست و پوشش نیز باید در نظر گرفته شود [۶]. بنابراین، در ارزیابی کیفیت گزارش‌های خودکار در کاربردهای بالینی، معیارهایی که هم پوشش و هم صحت را اندازه‌گیری می‌کنند، مناسب‌تر هستند. در این راستا ROUGE نسبت به BLEU مناسب‌تر است، زیرا علاوه بر صحت، پوشش را نیز در نظر می‌گیرد. همچنین، CIDEr به دلیل توجه به مفاهیم دستوری، برجسته بودن، اهمیت و دقت محتوا، برای ارزیابی کیفیت گزارش‌های خودکار پزشکی مناسب‌تر است. CIDEr از تکنیک TF-IDF برای تمرکز بر روی کلمات کلیدی مرتبط با بیماری و فیلتر کردن کلمات رایج غیرمهم استفاده می‌کند [۲۷]. بنابراین، نمرات بالاتر ROUGE و CIDEr نشان‌دهنده عملکرد برتر مدل پیشنهادی ما در تولید گزارش‌های تصاویر پزشکی است. همان‌گونه که در جدول ۱ مشاهده می‌شود، مدل Res-LSTM-Attn در تولید گزارش‌های یافته‌ها + برداشت در مقایسه با مدل‌های مرجع، از جمله مدل مبتنی بر ترنسفورمر [۱۷]، عملکرد برتری داشت. انتخاب LSTM دولایه به‌عنوان مدل زبانی به دلیل محدودیت‌های محاسباتی و مناسب بودن آن برای گزارش‌های پزشکی انجام شد. مدل‌های پیچیده مانند ترنسفورمرها به دلیل محاسبات موازی و تعداد پارامترهای بالا به توان محاسباتی بیشتری نیاز دارند. در مقابل، LSTM با ۲۵۶ نرون پیچیدگی کمتری دارد و ترکیب با مکانیسم توجه چندسر، با استفاده از خروجی‌های FastText، درک معنایی را تقویت کرد و برخی مزایای ترنسفورمرها را بدون افزایش پیچیدگی فراهم نمود. با وجود این نقاط قوت، مدل Res-LSTM-Attn معایبی دارد که بر کیفیت گزارش‌ها تأثیر گذاشت. نخست، تابع هزینه آنتروپی متقاطع بر پیش‌بینی کلمات تمرکز دارد و صحت نحوی جملات را در نظر نمی‌گیرد، که به ضعف در تولید جملات با دستور زبان صحیح منجر شد، مطالعات مرجع مانند [۷] از معیارهای تقویت‌شده مانند CIDEr یا ROUGE برای بهبود پوشش و انسجام استفاده کرده‌اند، که می‌توانست روانی گزارش‌ها را ارتقا دهد. دوم، محدودیت تنوع زبانی در مجموعه داده IU X-Ray عملکرد مدل را تحت تاثیر قرار داد و به نمره پایین BLEU-4 منجر شد. سوم، استفاده از رویکرد حریصانه به‌جای Beam Search تنوع جملات را کاهش داد از طرفی رویکرد حریصانه، که کلمه با بالاترین احتمال را در هر مرحله انتخاب می‌کند، به دلیل سادگی و سرعت استفاده شد. این روش تنوع جملات را کاهش داد و باعث BLEU-4 پایین شد اما Beam Search، که چندین مسیر محتمل را حفظ می‌کند، می‌توانست روانی را بهبود بخشد و احتمالا BLEU-4 را ارتقا دهد، با این حال Beam Search پیچیدگی محاسباتی و ناپایداری نتایج تست در مقایسه معیارها به دلیل تولید گزارش های متنوع در هر اجرا را موجب گردید.

همانطور که در نمودارهای الف، ب و ج شکل (۳) مشاهده می‌شود، روند افزایشی این معیارها در طول دوره‌های آموزش به ترتیب برای سه حالت مختلف تولید گزارش یافته‌ها + برداشت، یافته‌ها و برداشت به خوبی نمایان است.

[27] precision

[28] recall

الف ب ج



شکل ۳- روند افزایش معیارها برای مدل Res-LSTM-Attn

در شکل (۴) نمونه‌ای از توصیفات تولید شده توسط مدل پیشنهادی برای هر سه نوع گزارش اورده شده است. از نتایج مثال در شکل (۴)، بخش c می‌توانیم ببینیم که برداشت با گزارش اصلی در مدلهای Res-LSTM-Attn کاملا مطابقت دارد. با این حال، در b و a ، یافته‌ها و برداشت‌ها ایجاد شده، با گزارش اصلی دقیقا مطابقت ندارد اما تا حد زیادی مشابه هستند. دلیل اصلی ممکن است این باشد که ما مدل را در یک مجموعه آموزشی کوچک اموزش می‌دهیم، و برخی ناهماهنگی‌ها و همچنین نویز در گزارش‌های اصلی وجود دارد. علاوه بر این، مدل ما جملات جدیدی را که هرگز در مجموعه آموزشی ظاهر نشده‌اند به خوبی ایجاد نمی‌کند. این می‌تواند به دلیل مشکل در یادگیری دستور زبان صحیح از یک مجموعه کوچک باشد زیرا تابع هدف برای آموزش درستی نحوی را در نظر نمی‌گیرد.

برداشت + یافته‌ها-a	یافته‌ها-b	برداشت-c
		
211_IM-0740-1001.dcm.png	138_IM-0244-1001.dcm.png	1481_IM-0312-3001.dcm.png
GT sentence= normal heart size clear lungs trachea is midline no pneumothorax no pleural effusion no acute cardiopulmonary abnormality	GT sentence = cardiac and mediastinal contours are within normal limits the lungs are clear bony structures are intact	GT sentence = no acute cardiopulmonary abnormality
Predicted sentence = the heart is normal in size the lungs are clear no pneumothorax or pleural effusion no acute cardiopulmonary abnormality	Predicted sentence = the heart size and mediastinal contours are within normal limits the lungs are clear bony structures are intact	Predicted sentence = no acute cardiopulmonary abnormality

شکل ۴- نمونه‌ای از توصیفات تولید شده توسط مدل پیشنهادی برای هر سه نوع گزارش

۴- نتیجه‌گیری

در این پژوهش، ما یک مدل پیشرفته بازگشتی را برای تولید گزارش‌های اشعه‌ایکس قفسه‌سینه پیشنهاد کرده‌ایم. در مدل Res-LSTM-Attn ما یک شبکه طبقه‌بندی چند برچسبی تطبیق تصویر- کلمات گزارش، برای یادگیری ویژگی‌های معنایی تصاویر اشعه‌ایکس قفسه‌سینه پیشنهاد داده‌ایم که به برچسب‌های تصاویر نیازی ندارد و یک پاراگراف طولانی و مفصل را می‌تواند به طور مکرر و کلمه به کلمه ایجاد کند. این مدل می‌تواند نقش مهمی در بهبود روند تشخیص و گزارش‌نویسی ایفا کند. مدل پیشنهادی ما به طور مؤثری می‌توانند مشکلاتی را که در حال حاضر در تولید گزارش‌های پزشکی وجود دارد، برطرف کنند. یکی از چالش‌های اصلی در این زمینه، تولید گزارشات دقیق و جامع از تصاویر پزشکی تنها با استفاده از تصویر و یک گزارش در دسترس است. مدل ما با بهره‌مندی از تکنیک‌های یادگیری عمیق، می‌تواند تصاویر نمای جلوی اشعه‌ایکس قفسه‌سینه را به طور کامل تحلیل کند و نتایج دقیق‌تری از نظر پزشکی ارائه دهد. لازم به ذکر است که این مدل تنها از تصاویر نمای جلوی اشعه‌ایکس قفسه‌سینه استفاده می‌کند و این ویژگی به ما این امکان را می‌دهد تا تمرکز بیشتری بر روی تحلیل و استخراج اطلاعات از این نوع تصاویر داشته باشیم. به علاوه، این مدل می‌تواند به تولید گزارش‌های طولانی و جامع کمک کند که به پزشکان و متخصصان این حوزه اجازه می‌دهد تا ارزیابی‌های دقیق‌تری انجام دهند و تصمیمات بهتری بگیرند. با وجود مزایای مدل پیشنهادی در تولید گزارش‌های پزشکی همانطور که در بخش قبل به ان اشاره شد، مجموعه داده‌های استفاده شده در این پژوهش، تنوع محدودی از بیماری‌های مرتبط با قفسه‌سینه را در بر می‌گیرد. لازم است مدل در آینده بر روی مجموعه داده‌های گسترده‌تر با پوشش بیماری‌های متنوع‌تر آزمایش شود همچنین زمان دوره آموزش و تست مدل بالا می‌باشد و نیازمند سخت افزار مناسبتر است. از دیگر اقدامات پیشنهادی برای ارتقای مدل می‌توان به جایگزینی شبکه‌های کانولوشنی با مدل‌های مبتنی بر مکانیسم توجه در بینایی ماشین، مانند ویژن ترانسفورمر، اشاره کرد که در استخراج ویژگی‌های بصری عملکرد بهتری دارند. علاوه بر این، بهبود مدل چندبرچسبی و رمزگذار تصویر، و جایگزینی مدل بازگشتی LSTM با رمزگشای ترانسفورمر، که در وظایفی مانند ترجمه ماشینی عملکرد مطلوبی نشان داده است، می‌تواند تولید گزارش‌های منسجم‌تر و دقیق‌تر را تسهیل کند. در بخش استخراج ویژگی‌های کلمات، پیشنهاد می‌شود به‌جای مدل fastText، از BioBERT، که یک مدل زبانی پیش‌آموزش‌دیده تخصصی برای متون زیست‌پزشکی است، استفاده شود. همچنین، یکپارچگی مدل‌های مولد پیش‌آموزش‌دیده مانند BioGPT، می‌تواند تولید توصیف‌های روان و دقیق را تقویت کند، با بهره‌گیری از BioBERT برای استخراج ویژگی‌های متنی و BioGPT برای تولید متن، می‌تواند دقت و پوشش گزارش‌های تولیدشده، به‌ویژه در بخش‌های «یافته‌ها» و «برداشت»، را به‌طور قابل‌توجهی بهبود داد. این رویکرد ترکیبی، با ادغام مدل‌های تخصصی زیست‌پزشکی، می‌تواند محدودیت‌های فعلی مانند ناهماهنگی‌های زبانی و کمبود تنوع داده را برطرف کرده و عملکرد مدل را در کاربردهای بالینی ارتقا دهد.

تعارض منافع

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

تاییدیه اخلاقی

نویسندگان متعهد میشوند که مطالب این مقاله را در هیچ مجله دیگری به چاپ نرسانده اند.

مشارکت های نویسندگان

اقای فردین قادری: ایده، پیاده سازی، نوشتن نسخه ابتدایی مقاله، ویرایش، ارائه نتایج

دکتر محمد باقر خدابخشی: ایده، نظارت، بررسی نتایج، نهایی سازی نسخه

دکتر شهریار جاماسب: ویرایش علمی، بررسی نتایج

منابع مالی

برای این پژوهش هیچگونه حمایت مالی دریافت نشده است.

مراجع

[1] Jing, Baoyu, Pengtao Xie, and Eric Xing. “On the Automatic Generation of Medical Imaging Reports.” ArXiv Preprint ArXiv:1711.08195, 2017.

[2] Krause, Jonathan, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. “A Hierarchical Approach for Generating Descriptive Image Paragraphs.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 317–25, 2017.

[3] Harzig, Philipp, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. “Addressing Data Bias Problems for Chest X-Ray Image Report Generation.” ArXiv Preprint ArXiv:1908.02123, 2019.

[4] Yang, Shaokang, Jianwei Niu, Jiyan Wu, Yong Wang, Xuefeng Liu, and Qingfeng Li. “Automatic Ultrasound Image Report Generation with Adaptive Multimodal Attention Mechanism.” Neurocomputing 427 (2021): 40–49.

[5] Han, Zhongyi, Benzheng Wei, Stephanie Leung, Jonathan Chung, and Shuo Li. “Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework.” In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, 185–93. Springer, 2018.

[6] Xue, Yuan, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. “Multimodal Recurrent Model with Attention for Automated Radiology Report Generation.” In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, 457–66. Springer, 2018.

[7] Li, Yuan, Xiaodan Liang, Zhiting Hu, and Eric P Xing. “Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation.” Advances in Neural Information Processing Systems 31 (2018).

[8] Shi, Jijun, Shanshe Wang, Ronggang Wang, and Siwei Ma. “AIMNet: Adaptive Image-Tag Merging Network For Automatic Medical Report Generation.” In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7737–41. IEEE, 2022.

[9] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” Bioinformatics 36, no. 4 (2020): 1234–40.

[10] Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. “BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining.” Briefings in Bioinformatics 23, no. 6 (2022): bbac409.

[11] Zhang, Yixiao, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. “When Radiology Report Generation Meets Knowledge Graph.” In Proceedings of the AAAI Conference on Artificial Intelligence, 34:12910–17, 2020.

[12] Goel, Navya, Aditi Arora, Priyanshu Kashyap, and Sagar Varshney. “An Analysis of Image Captioning Models Using Deep Learning.” In 2023 International Conference on Disruptive Technologies (ICDT), 131–36. IEEE, 2023.

[13] Xu, Liming, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. "Deep image captioning: A review of methods, trends and future challenges." Neurocomputing 546 (2023): 126287.

[14] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out, pp. 74-81. 2004.

[15] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575. 2015.

[16] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: A Method for Automatic Evaluation of Machine Translation.” In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–18. 2002.

[17] Park, Hyeryun, Kyungmo Kim, Seongkeun Park, and Jinwook Choi. “Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation.” IEEE Access 9 (2021): 150560–68.

[18] Alfarghaly, Omar, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. “Automated Radiology Report Generation Using Conditioned Transformers.” Informatics in Medicine Unlocked 24 (2021): 100557.

[19] Singh, Sonit, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. "From chest x-rays to radiology reports: a multimodal machine learning approach." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8. IEEE, 2019.

[20] Huang, Xin, Fengqi Yan, Wei Xu, and Maozhen Li. “Multi-Attention and Incorporating Background Information Model for Chest x-Ray Image Report Generation.” IEEE Access 7 (2019): 154808–17.

[21] Xie, Xiancheng, Yun Xiong, Philip S. Yu, Kangan Li, Suhua Zhang, and Yangyong Zhu. "Attention-based abnormal-aware fusion network for radiology report generation." In International Conference on Database Systems for Advanced Applications, pp. 448-452. Cham: Springer International Publishing, 2019.

[22] Yin, Changchang, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network." In 2019 IEEE International Conference on Data Mining (ICDM), pp. 728-737. IEEE, 2019.

[23] Biswal, Siddharth, Cao Xiao, Lucas M. Glass, Brandon Westover, and Jimeng Sun. "Clara: clinical report auto-completion." In Proceedings of the Web Conference 2020, pp. 541-550. 2020.

[24] Jing, Baoyu, Zeya Wang, and Eric Xing. “Show, Describe and Conclude: On Exploiting the Structure Information of Chest x-Ray Reports.” ArXiv Preprint ArXiv:2004.12274, 2020.

[25] Xiong, Yuxuan, Bo Du, and Pingkun Yan. "Reinforced transformer for medical image captioning." In International Workshop on Machine Learning in Medical Imaging, pp. 673-680. Cham: Springer International Publishing, 2019.

[26] Li, Christy Y., Xiaodan Liang, Zhiting Hu, and Eric P. Xing. "Knowledge-driven encode, retrieve, paraphrase for medical image report generation." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 6666-6673. 2019.

[27] Li, Xin, Rui Cao, and Dongxiao Zhu. "Vispi: Automatic visual perception and interpretation of chest x-rays." arXiv Preprint arXiv:1906.05190 (2019).