مقاله مروری

مجله ماشین‌بینایی و پردازش‌تصویر
سال دهم، شماره سوم، پاییز ۱۴۰۲

بررسی روش‌های خودکار تفسیر تصاویر پزشکی بر پایه یادگیری عمیق

حسنیه ذوالفقاری[۱]، مریم رستگارپور[۲]، محمد تشنه‌لب[۳]، عباس کوچاری[۴]، علیرضا احسانبخش[۵]

چکیده

تفسیر خودکار تصاویر، زمینه جدیدی از هوش مصنوعی است که دو شاخه پردازش زبان‌طبیعی و یادگیری ماشین را به خدمت می‌گیرد. تحقیقاتی که در سال‌های اخیر بر روی این مقوله انجام شده و نتایج قابل قبولی که در این زمینه حاصل شده‌است از یک طرف و نیاز جامعه پزشکی به تفسیر خودکار تصاویر پزشکی از طرف دیگر، محققان را برآن داشته تا این رویکرد را در این زمینه نیز به‌کارگیرند. تفسیر خودکار تصاویر پزشکی نسبت به مسأله توصیف خودکار تصاویر طبیعی، چالش برانگیزتر می‌باشد. کمیت و کیفیت مجموعه‌داده‌های موجود در این مقوله نسبت به مجموعه‌داده‌های تفسیر تصاویر طبیعی کمتر است، تفسیرها غیرساختاریافته‌اند و تفسیر تصاویر طبیعی، شامل توصیف اشیاء و روابط بین آنها با یک یا چند جمله است درحالی‌که شرح تصاویر پزشکی شامل درک یافته‌های بالینی و ارائه یک گزارش دقیق از پاراگراف‌های مختلف است؛ تا فقط آنچه از نظر بالینی مهم است به جای آنچه در تصویر از نظر اشیاء وجود دارد برجسته گردد. در راستای رسیدن به نتایج مطلوب روش‌های متعددی پیشنهاد شده‌است که در این بین روش‌های مبتنی‌بر یادگیری‌عمیق، به نتایج بهتری دست یافته‌است. این مقاله به معرفی مجموعه‌داده‌ها، معیارهای ارزیابی و روش‌های توسعه‌یافته برپایه یادگیری‌عمیق در زمینه تفسیر خودکار تصاویر پزشکی می‌پردازد تا کمکی در راستای درک ادبیات موجود و برجسته نمودن مسیرهای آینده در این زمینه باشد.

کلیدواژه‌ها

تفسیر خودکار تصاویر پزشکی ، تفسیر خودکار تصاویر، شبکه عصبی کانولوشن، شبکه عصبی بازگشتی، مکانیزم توجه.

این مقاله آذر ۱۴۰۱ دریافت شد در اسفندماه بازنگری و سپس پذیرفته گردید.

[۱] دانشجوی دکترای هوش مصنوعی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران
رایانامه: zolfaghari@iaubir.ac.ir
[۲] دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد ساوه، ساوه، ایران
رایانامه: m.rastgarpour@iau-saveh.ac.ir
[۳] دانشکده فنی و مهندسی، گروه مهندسی کنترل، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران
رایانامه: teshnehlab@eetd.kntu.ac.ir
[۴] دانشکده مکانیک، برق‌وکامپیوتر، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران، ایران
رایانامه: koochari@srbiau.ac.ir
[۵] گروه تکنولوژی رادیولوژی، دانشگاه علوم پزشکی بیرجند، بیرجند، ایران
رایانامه: a.r.ehsanbakhsh@bums.ac.ir

نویسنده مسئول: مریم رستگارپور

dorl.net/dor/20.1001.1.23831197.1402.10.3.1.3

۱ مقدمه

توانایی درک تصویر و استخراج اطلاعات آن، از چالش‌های مشترک حوزه‌های بینایی ماشین و پردازش زبان طبیعی است. زیرا علاوه بر تشخیص، بخش‌بندی و طبقه‌بندی اشیاء، نیاز به درک رابطه بین اشیاء و اقدامات انجام شده توسط آنها و تبدیل آن به جملاتی صحیح در یک زبان طبیعی دارد. در سالهای اخیر پیشرفت‌های زیادی در این زمینه صورت گرفته است که شامل تولید مجموعه‌داده‌های غنی و مناسب [1]تا[4] و ارائه الگوریتم‌های کارا می‌باشد [5]تا[10].مسأله تفسیر تصاویر پزشکی و تولید خودکار گزارش نیز در این دسته از مسائل قرار می‌گیرد.

در علم پزشکی، استفاده از تصاویر پزشکی در زمینه‌های متعددی کاربرد دارد؛ به عنوان مثال متخصصان و رادیولوژیست‌هااز آن برای تشخیص و درمان بیماری‌ها استفاده

می‌کنند. داروسازان ممکن است آنها را برای کشف دارو به کار گیرند و جراحان نیز از این تصاویر، در قبل، بعد و همچنین در حین عمل برای نظارت بر روند درمان استفاده می‌کنند[11]. از آنجا که نوشتن گزارش پزشکی مستلزم آشنایی با آناتومی و فیزیولوژی طبیعی بدن،تصویربرداری پزشکی، درک عمیق بیماری و تجزیه و تحلیل کامل تصاویر مورد بررسی می‌باشد؛ این کار برای پزشکان کم تجربه، سخت و همراه خطاست و برای پزشکان مجرب کاری پرزحمت و تکراری است[12]؛ از این‌رو برای تسهیل فرایند تهیه گزارش، سیستم‌هایِ تولیدِ گزارش برای شرح تصاویر پزشکی برپایه استفاده از رایانه پیشنهاد شده است. این سیستم‌ها به طور خودکار یافته‌ها را از تصاویر استخراج کرده و مانند یک پزشکِ متخصص، تفسیری برای تصاویر پزشکی ارائه می‌کنند. این امر باعث صرفه‌جویی در وقت و کاهش حجم کار پزشکان می‌گردد. علاوه بر این، با توجه به کمبود نیروهای متخصص در اکثر مناطق تا حدودی نیاز به متخصصان برای نوشتن گزارش مرتفع می‌شود. همچنین بسیاری از کاربران بالقوه می‌توانند از گزارش‌های ایجاد شده استفاده کنند. به عنوان مثال، رادیولوژیست‌ها، پزشکان و همچنین تکنسین‌ها می‌توانند گزارش‌ها را برای یافتن اطلاعات فوری و دست اول به کار گیرند. بنابراین، تولید گزارش‌های خودکار تصاویر پزشکی، برای کمک به درمان و در جهت ارتقای سلامت جامعه، می‌تواند موثر باشد.

در تحقیقات اخیر تولید تفسیر خودکار تصاویر پزشکی، از انواع روش‌های مبتنی بر یادگیری عمیق استفاده شده است [11, 13]تا[17] اگر چه این روش‌ها بصورت خودکار تصاویر را به عنوان ورودی دریافت کرده و در نهایت گزارش را تولید می‌کند، اما از چالشهای متعددی رنج می برند. مثلاً، متخصصان رادیولوژی به مراحل مربوط به گردش کار، در تولید خودکار تفسیر تصاویر پزشکی واقف نیستند و اغلب در زمینه‌های مربوط به سلامت و درمان، جامعه پزشکی خواهان هوش مصنوعی قابل توضیح می‌باشند[18] و یا نمی‌توانند قالب یا ساختاری برای تولید گزارش توصیه کنند و نگرانی بالینی بیشتر از این بابت است که هیچ تضمینی وجود ندارد که مدل‌های آموزش دیده‌ی خودکار مولد تفسیر، یافته‌های کلینیکی کلیدی در تصویر را یاد بگیرند. همچنین برای یک گزارش پزشکی روان، به جای یک جمله، معمولاً یک پاراگراف ایجاد می‌شود که در مقایسه با مسأله توصیف خودکار تصاویر،چالشی قابل توجه است. مسأله دیگر پیش روی این سیستم‌ها نیاز به یک مجموعه‌داده به همراه گزارش‌های غنی می‌باشد، علی‌رغم اینکه مجموعه‌داده‌های متعددی در زمینه انواع تصاویر پزشکی برای سایر وظایف، مثلا تشخیص بیماری، بخش‌بندی تصاویر و ...وجود دارد اما برای توصیف تصاویر پزشکی،مجموعه‌داده‌ها محدودند زیرا علاوه بر تصاویر، گزارش پزشکی مرتبط با هر تصویر نیز مورد نیاز است و این درحالیست که اکثر مجموعه‌داده‌های موجود فاقد گزارش‌های نوشته شده توسط متخصصین این حوزهمی‌باشند. عدم توازن قابل توجه در کلاس‌های مجموعه‌داده‌ها، یک مشکل دیگر است؛ در بعضی موارد تصاویر هیچ یافته و بیماری را گزارش نمی‌کنندیا گاهی طیف وسیعی از بیماری‌های احتمالی می‌تواند مسبب یک ناهنجاری باشد و این درحالیست که برخی از آنها به ندرت در داده‌هاظاهر می‌شوند. چالش دیگر وجود مواردی با چندین تصویر اما تنها یک گزارش تشخیصی است [19]. از طرف دیگر مجموعه‌داده‌های مناسب این حوزه، نسبت به مجموعه‌داده‌های تفسیر تصاویر عمومی، در بسیاری از موارد تعداد داده‌ی کمتری دارد. از این رو یادگیری درست و کامل اطلاعات موجود در تصاویر مشکل است. معیارهای ارزیابی گزارش‌های تولید شده نیز چالش برانگیز است. بیشتر مدل‌های تولید گزارش به معیارهای ارزیابی مانند CIDEr[1][20]، ROUGE[2][21] و BLEU[3][22]، متکی‌اند که شباهت زبانی را بین دو توالی متن، اندازه می‌گیرند این معیارهای آماری برپایه انگرم[4] است، و در ارزیابی صحت و کیفیت کلی گزارش تولید شده کارایی لازم را ندارد[23].

با توجه به موارد ذکر شده، حل چالش‌های مساله مذکور، توجه بیشتر محققان را طلب می‌کند. مقالات مروری می تواند تسریع کننده حرکت محققین در این راستا باشد.

در سال ۲۰۱۹، پاولوپولوس و همکاران[24]، مروری کوتاه بر روی تفسیر تصاویر پزشکی ارائه کردند. نویسندگان تعدادی از مجموعه‌داده‌های موجود، معیارهای ارزیابی و برخی روش‌های پیشرفته را مورد بحث قرار دادند، بررسی مروری آنها فقط سه مجموعه‌داده را پوشش می‌داد و معیارهای ارزیابی را به صورت خلاصه موردبررسیقرار داده بود علاوه بر آن ساختار خاصی نیز در ارائه و دسته‌بندی روش‌ها نداشت.

در سال ۲۰۲۰، مسینا و همکاران [25] مقاله مروری‌شان را با تاکید بر روی توضیح پذیری روش‌های تفسیر تصاویر پزشکی ارائه کردند و مواردی که مورد بررسی قرار دادند، عبارت بود از: مجموعه‌داده‌ها، طراحی‌معماری، توضیح‌پذیری و معیارهای ارزیابی. مجموعه‌داده‌هایی که آنها بررسی کردند برای دو هدف مختلف تفسیر تصویر و کلاسه‌بندی تصویر بود. در بخش طراحی مدل، آنها بررسی‌هایشان را در چند زیر بخش ارائه کردند که عبارت بود از: ورودی و خروجی، مولفه بصری و مولفه زبان. آنها هر بخش را به صورت مختصر بررسی نمودند. نقطه قوت کار آنها بررسی مساله توضیح‌پذیری بود. آنها معیارهای ارزیابی را نیز بطور مختصر مانند پاولوپولوس و همکاران [24]، بررسی کردند.

در سال ۲۰۲۱، عایشه و همکاران[11]مقاله مروری جامعی در این زمینه منتشر کردند. آنها تجزیه و تحلیل و مقایسه‌ای از مطالعات موجود در مورد تفسیر تصاویر پزشکی با تمرکز بر رویکردهای مبتنی بر یادگیری عمیق ارائه دادند همچنین مجموعه‌داده‌های در

[1] Consensus based Image Description
[2] Recall-Oriented Understudy for Gisting Evaluation using LCS for longest matching
[3] BiLingual Evaluation Understudy
[4] N- gram

دسترس عموم، معیارهای ارزیابی و انواع روش‌های مبتنی بر یادگیری عمیق، را بررسی کردند. در بین مقالات مروری بررسی شده مقاله آنها جامع‌ترین مقاله مروری بود. اما در چند مورد مشکلاتی در آن وجود دارد. مجموعه‌داده‌های در دسترس عمومی که مورد بررسی قرار گرفت؛ پوشش کامل ندارد. در مبحث انواع مدل قابل طراحی در بخش کدگذار و کدگشا بر روی تنوع روش‌های پیاده‌سازی کدگذار و کدگشا بحث انجام نشده است و مباحث جدید شامل روش‌های جدید پیش‌پردازش و معماری‌های مبتنی بر ترنسفورمر بررسی نشده است.

مقاله حاضر بررسی های قبلی را با ارائه یک تحلیل عمیق‌تر از شرح تصاویر پزشکی گسترش می‌دهد. مزیت مقاله جاری به سایر مقالات مروری را می‌توان در موارد زیر خلاصه کرد:

- بررسی مجموعه‌داده‌ها در مقاله جاری، نسبت به سایر مقالات مروری جامع‌ترمی‌باشد.
- تفاوت و مزیت عمده این مقاله مروری نسبت به سایر مقالات مشابه در نحوه بررسی روش‌های کدگذار-کدگشاست. انواع معماری کدگذار و انواع معماری کدگشا در دسته‌بندی منظم بررسی شده، دلیل استفاده از آن و مقالاتی که بر پایه آن بنا شده، ذکر می‌گردد. این دسته‌بندی در ارائه دید مناسب به خواننده در جهت نوآوری و ایده‌پردازی راه گشاست.
- مسابقات ImageCLEF معرفی شده است که می‌تواند انگیزه‌بخش محققان، برای کار در این حیطه باشد.
- در مبحث پیش پردازش، نسبت به سایر مقالات بررسی بیشتری انجام شده است.
- اهمیت فراداده ها نیز با ارائه و بررسی مقالات مرتبط بررسی شده است؛ که البته این مورد در مقاله عایشه و همکاران[11] نیز لحاظ شده‌بود. در ضمن مدل‌های مبتنی بر الگو، مبتنی بر بازیابی و روش های ترکیبی که تکنیک‌های مبتنی بر یادگیری عمیق مولد و تکنیک‌های بازیابی را با هم ترکیب می‌کنند بررسی شده است؛ بدین جهت که همه زوایای مساله بررسی شده باشد.
- معیار های ارزیابی نیز به تفصیل بررسی شده است که نسبت به مقالات مسینا[25] و پاولوپولوس[24] کامل‌تر می باشد.

در ادامه بخش‌های مختلف مقاله به اینصورت سازماندهی شده است: بخش دوم به معرفی مجموعه‌داده‌های موجود می‌پردازد، در بخش سوم معیارهای ارزیابی معرفی می‌شود و در بخش چهارمروش‌های مختلف تفسیر تصاویر پزشکی با تاکید بر روش‌های مبتنی بر یادگیری عمیق مورد بررسی قرار می‌گیرد؛ بخش پنجم بحث و بخش پایانی نتیجه‌گیری مطالب خواهد بود.

۲ مجموعه‌داده‌ها

از آنجایی که مجموعه‌داده‌هایی که برای تفسیر تصاویر پزشکی استفاده می‌شود علاوه بر تصاویر پزشکی باید شامل گزارش‌های پزشکی مرتبط نیز باشد؛ لذا تعداد انگشت شماری مجموعه‌داده در دسترس عموم است و کاستی‌هایی در آنها وجود دارد که می‌توان ناسازگاری در مجموعه‌داده، تنوع بسیار زیاد در تصاویر [12], [28]–[26]، جملات توصیفی تکراری[28] ,[27] و گزارش‌هایی با طول‌های مختلف را برشمرد.بطور مثال در [28]طول تفسیر ارائه شده از ۱ تا ۸۱۶ کلمه متفاوت است[29]. مجموعه‌داده‌ها غالباً ناقص هستند به این معنا که در مواردی هیچ گزارشی برای تصاویر وجود ندارد و یا وجود مواردی که برای چند تصویر فقط یک گزارش پزشکی موجود است[30].تعداد نمونه‌های یک مجموعه‌داده برای روش‌های یادگیری عمیق بسیار حائز اهمیت است، اکثر این مجموعه‌داده‌ها شامل تعداد نمونه کمی است و تعداد نمونه‌ها برای بعضی از بیماریها کمتر از سایر بیماری‌ها و یا حالت نرمال است. این امر باعث ایجاد مشکلات در تعمیم مدل می‌شود. تعدادی از مجموعه‌داده‌ها که امکان دسترسی به آنها برای عموم وجود دارد در مقالات مختلف اغلب محققان از آنها استفاده کرده‌اند،در جدول ۱ ارائه شده است.و در ادامه معرفی می‌گردد. علاوه بر آن، تعدادی از محققان نیز مجموعه‌داده‌هایی گردآوری کرده‌اند؛ این مجموعه‌داده‌ها تکنیک‌های متفاوت تصویربرداری و اندام‌های متفاوت را پوشش می‌دهد، اما در دسترس عموم نمی باشد (جدول ۲).

اگر چه مجموعه‌داده می‌تواند در رسیدن به نتایج بهتر بسیار تاثیرگذار باشد اما در مورد مجموعه‌داده‌ای مانند IU Chest X-Ray که‌محبوب‌ترین مجموعه‌داده در زمینه تفسیر تصاویر پزشکی می‌باشد[31]، نتایج به عنوان مثال BLEU-1 در مقالات مورد بررسی، با توجه به مدل پیشنهادی بین ۰,۳۷[32] و ۰,۵۲۹ [33] متغیر بوده است. ولی آنچه غیرقابل انکار می‌باشد این است که، هر چه گزارش‌ها ساختار یافته‌تر و تشخیص‌ها محدودتر باشند نتایج بهتری حاصل می‌گردد. به عنوان مثال در مجموعه‌داده XRAYS FRONTAL PELVIC که حاوی تصاویر اشعه ایکس از شکستگی‌های لگن است اگرچه جملات توصیفی برای هر شکستگی و گزارش‌های اصلی در دسترس بوده است، اما این جملات ساختار و محتوای ناسازگاری داشتند، که آموزش سیستم برای تولید جملات مشابه را دشوار می‌نموده است. برای آموزش بهتر مدل، یک رادیولوژیست، مجموعه جدیدی از اصطلاحات توصیفی را به عنوان برچسب، به صورت دستی برای هر تصویر ایجاد کرده و برای هر کدام یک گزارش رادیولوژی بر اساس یک الگوی استاندارد ساده شده، آماده کرده است. نتایج حاصله برای گزارش‌های اصلی و گزارش‌های ساده شده تفاوت قابل توجهی دارد[34] (جدول۳).

از طرفی در صورت مقایسه نتایج حاصل برای روش‌های تفسیر خودکار تصاویر پزشکی با تفسیر تصاویر عمومی، نتایج برای تفسیر عمومی بصورت معناداری بیشتر می‌باشد به عنوان مثال الگوریتم پیشنهادی زنگ و همکاران[35] بر روی مجموعه‌داده COCO، معیار BLEU-1 را برای بهترین مدل پیشنهادی برابر

۰,۹۶۹ گزارش کرده است و این در حالی است که در بین مقالات مورد بررسی بهترین نتیجه BLEU-1 بر روی مجموعه داده XRAYS FRONTAL PELVIC برابر ۰,۹۱۹ بدست آمدهاست که همانطور که ذکر شد به دلیل بازنویسی گزارش‌ها می‌باشد. اما بر روی مجموعه‌داده IU Chest X-Ray، در بین مقالات مورد بررسی بهترین نتیجهBLEU-1 را یوان و همکاران [39] برابر 0.529 گزارش داده‌اند

جدول (۱): مجموعه‌داده‌هایی که در دسترس عموم می‌باشند.

Dataset	#Image	Image Type	#Reports	#Patients	Place & Time	Label Extraction
IU Chest x-ray	7470	Chest X-Ray	3955	3955	Indiana Network for Patient Care, 2016	Manual & MTI
ChestX-ray14	112120	Chest X-Ray	0 Label only	30805	National Institutes of Health (NIH) 2017	Automated from the impression and findings section
CheXpert	224316	Chest X-Ray	0 Label only	65240	Stanford Hospital 2002-2017	Automated Rule-based labeler
MIMIC_CXR	377110	Chest X-Ray	227827	227943	Beth Israel Deaconess Medical Center Emergency Department 2011–2016	Automated Rule-based labeler
PadChest	160868	Chest x-ray	109931	67625	San Juan Hospital (Spain) from 2009 to 2017	27% Manually 73% Automatic labeling
PEIR-Gross	7442	Medical Teaching Image	7442	-	Pathology Education Informational Resource 2018	Top TF-IDF caption words
ICLEF2017 ICLEF2018	184614 232305	Medical Images	184614 232305	- -	PubMed 2017 PubMed 2018	20,463 UMLS CUIs 111155 UMLS CUIs
DeepEyeNet	15709	Retinal Images (FA & CFP)	15709	-	2021	Manually labeled by Ophthalmologists
BCIDR	1000	The Bladder Cancer	5000	32	University of Florida 2017	Manually labeled by Pathologist & four doctors

جدول (۲): مجموعه‌داده‌هایی که به صورت خصوصی جمع‌آوری شده و در دسترس عموم نمی‌باشد.

Dataset	Year	Image Type	#Images
XRAYS FRONTAL PELVIC[34]	۲۰۲۱	اشعه ایکس از شکستگی‌های لگن	۵۰۳۶۳
DDSM[36]	۲۰۰۰	ماموگرافی	۲۶۲۰
PACS of NIH clinical centre[37]	۲۰۱۵	اشعه ایکس چندگانه (گردن، استخوان، کبد، مغز، قلب و ...)	۲۱۶۰۰۰
CX-CHR[38]	۲۰۱۸	اشعه ایکس قفسه سینه	۳۵۵۰۰
ROCO[39]	۲۰۱۸	سونوگرافی، اشعه ایکس، PET ، فلوروسکوپی، MRI ، PET-CT، ماموگرافی	۸۱۸۲۵
INBreast[40]	۲۰۱۲	ماموگرافی	۵۱۰
RDIF	۲۰۱۹	بیوپسی کلیه	۱۱۵۲
[41]	۲۰۱۸	ستون فقراتMRI	۲۵۳
MICCAI 2017 LiTS Chalenge [42]	۲۰۱۸	تومور کبدCT	۲۰۱

جدول (۳): مقایسه معیار ارزیابیBLEU در مدل پیشنهاد شده برای مجموعه‌داده شکستگی‌های لگن، بر روی هر دو پیکره متنی:جملات گزارش‌های اصلی بدون ویرایش با جملات ساده و ویرایش شده [34].

	جملات اصلی	جملات ساده شده
1-gram	۶۵,۰	**۹۱,۹**
2-gram	۳۷,۹	**۸۳,۸**
3-gram	۲۴,۲	**۷۶,۱**
4-gram	۱۵,۹	**۶۷,۷**
میانگین	۲۵,۶۷	**۷۷,۹۷**

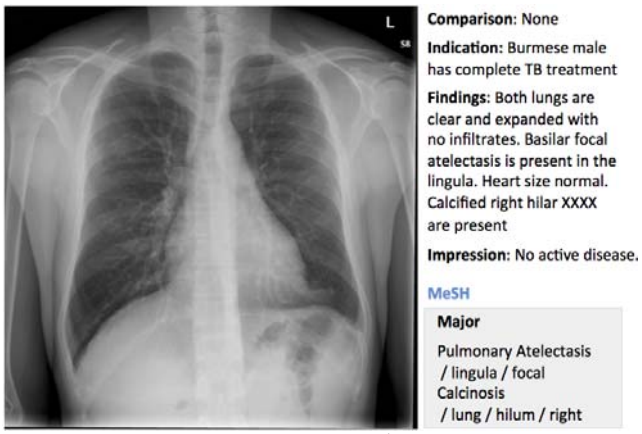
از طرفی در مجموعه‌دادهMIMIC-CXR که از نظر تعداد تصاویر نزدیک به مجموعه‌داده COCO می‌باشد بهترین نتیجه BLEU-1 در مقالات مورد بررسی برابر ۰,۳۹۴ [13] می‌باشد؛ که تاکیدی بر این موضوع است که تفسیر تصاویر پزشکی نسبت به تفسیر تصاویر عمومی چالشی‌تر می‌باشد و کمیت و کیفیت تصاویر و گزارش‌ها توامان بر نتایج تاثیرگذار است.

۲-۱ مجموعه‌داده‌های IU Chest X-Ray و Chest X-Ray14

دمنر-فوشمن و همکاران[30] مجموعه‌ای از تصاویر اشعه ایکس قفسه سینه[1] و تفسیرهای مربوطه را جمع‌آوری و در دسترس محققان قرار دادند. مجموعه‌داده شامل ۷۴۷۰ تصویر اشعه ایکس قفسه سینه شامل نمای جلویی و جانبی برای هر بیمار و 3955 گزارش رادیولوژی است که از این تعداد، ۱۵۲۶ مورد(۳۸٪) نرمال است. برای هر تصویر، یک گزارش رادیولوژی متنی وجود دارد که شامل بخشهای متفاوتی است (شکل۱). اولین بخش، مقایسه[2] اطلاعات مربوط به درمان‌های پزشکی قبلی بیمار را ارائه می‌دهد. در قسمت علائم[3] نشانه های بیماریِ ارائه شده توسط بیمار و فرادادههای بیمار نشان داده می‌شود در بخش یافته‌ها[4]، رادیولوژیست مشاهدات خود را می‌نویسد. در بخش

[1] https://openi.nlm.nih.gov/faq.php
[2] Comparison
[3] Indication
[4] Findings

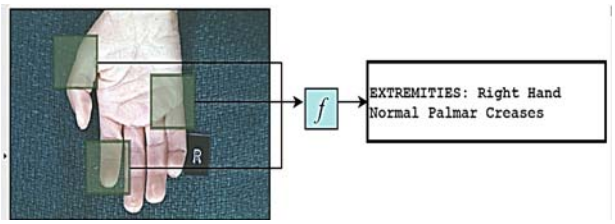
تشخیص[1]نتیجه گیری نهایی عنوان می‌گردد. بخش برچسب ها[2] کلمات کلیدی را نشان می‌دهد که اطلاعات مهم بخش یافته‌هاست. این کلمات کلیدی به دو روش دستی و خودکار تهیه شده و از نمایه کننده متن پزشکی (MTI)[3] برای کلمات کلیدی استفاده شده است. اکثر محققان، برای تفسیر تصویر فقط تولید بخشهای تشخیص و یافته‌هارا در دستور کار خود قرار داده‌اند[12]. موسسه ملی سلامت[4]آمریکا، مجموعه‌داده‌ای در مقیاس بزرگتر از IU Chest X-Ray با عنوان[43]Chest X-Ray14[5] را منتشر کرده است که شامل ۱۱۲۱۲۰ تصویر اشعه ایکس قفسه سینه از ۳۰۸۰۵ بیمار مختلف است. به هر تصویر یک یا چند برچسب از ۱۴ بیماری قفسه سینه (آتلکتازی، تثبیت، نفوذ، پنوموتوراکس، ادم، آمفیزم، فیبروز، افیوژن، ذات‌الریه، ضخیم شدن پلور، کاردیومگالی، ندول، توده و فتق[6]) اختصاص داده شده است. برچسب‌ها در اینجا نیز از بخش تشخیص و یافته‌هایگزارش‌های رادیولوژی با استفاده از ابزارهای استخراج برچسب بدست آمده است.



شکل(۱):نمونه تصویر و گزارش IU Chest X-Ray[44] . در اینجا MeSH همان بخش برچسب‌ها می‌باشد، که دربرگیرنده کلمات کلیدی است.

۲-۲ مجموعه‌دادهPEIR Gross

این مجموعه‌داده[7] جزء کتابخانه دیجیتالی منابع اطلاعاتی آموزش پاتولوژی PEIR می‌باشد که یک کتابخانه تصویری در زمینه آموزش پزشکی عمومی است. PEIR Gross زیرمجموعه‌ای از این کتابخانه و شامل ۷۴۴۳ تصویر و تفسیرهای مرتبط، از ضایعات درشت (قابل مشاهده با چشم غیر مسلح)، از ۲۱ زیرگروه مختلف آلبوم‌های PEIR است. هر تصویر دارای وضوح ۵۲۸ در ۷۹۲ است و فقط یک جمله برای تفسیر هر تصویر در نظر گرفته شده است (شکل۲). تعداد کلمات منحصربه‌فرد پس از پیش‌پردازش پیکره متنی گزارش‌ها برابر ۴۴۵۲ می‌باشد؛ بطور متوسط هر تصویر دارای ۱۲ کلمه است. برای هر تفسیر ۵ کلمه با بالاترین امتیاز[8]TF-IDF به عنوان برچسب انتخاب شده است [12].



شکل (۲): نمونه تصویر و گزارش PEIR Gross[17]

۲-۳ مجموعه‌داده‌های CheXpert و MIMIC-CXR V2.0.0

مجموعه‌داده‌های MIMIC-CXR V2.0.0[26][9][10] و CheXpert[45][11] حاوی تصاویر اشعه ایکس قفسه سینه هستند. در هر دو مجموعه‌داده، گزارش‌های رادیولوژی برای تصاویر، شناسایی و از پرونده الکترونیکی سلامت بیماران استخراج شده است. گزارش ها با استفاده از یک رویکرد مبتنی بر قاعده[12] همراه با یک رویکرد شبکه عصبی تازه توسعه یافته، شناسایی شدند[46]تا[48]. مجموعه‌داده CheXpert شامل ۲۲۴۳۱۶ تصویر اشعه ایکس قفسه سینه از ۶۵۲۴۰ بیمار است. به هر تصویر یک یا چند برچسب از ۱۴ بیماری قفسه سینه یا "بدون یافته" اختصاص داده شده است[45]. در این مجموعه‌داده فقط برچسب‌ها در اختیار عموم می‌باشد.احتمال رخداد برچسب‌های مختلف در این مجموعه‌داده در شکل ۳ نشان داده شده است.MIMIC-CXR V2.0.0 اولین مجموعه‌داده‌ی عمومی‌اشعه ایکس قفسه سینه است که تعداد نمونه‌های قابل توجهی دارد و شامل ۳۷۷۱۱۰ رادیوگرافی قفسه سینه ۲۲۷۸۳۵ گزارش تشخیصی است؛ ساختار گزارش‌ها مشابه مجموعه‌داده IU Chest X-Ray می‌باشد.

۲-۴ مجموعه‌داده PadChest

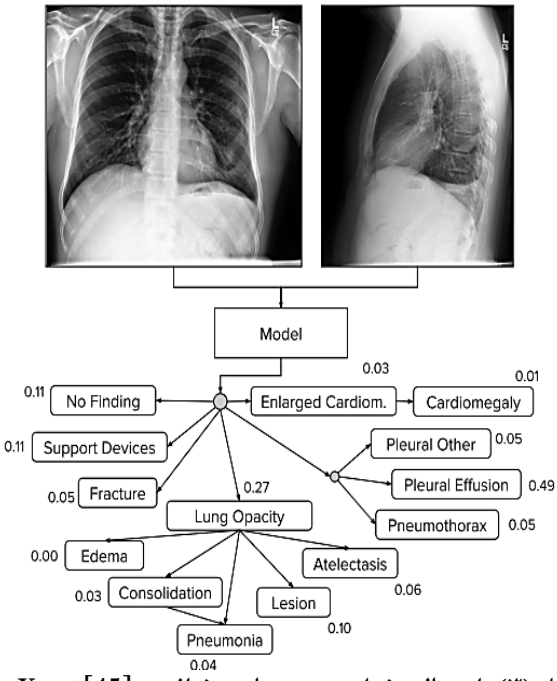
مجموعه‌داده PadChest[49][13] شامل تصاویر اشعه ایکس قفسه سینه با مقیاس بزرگ و وضوح بالا و گزارش‌های مرتبط با آنهاست و شامل ۱۶۰۸۶۸ اشعه ایکس قفسه سینه از ۶ نمای مختلف همراه با ۱۰۹۹۳۱ گزارش به زبان اسپانیایی از ۶۹۸۸۲

[1] Impression
[2] Tages
[3]Medical Text Indexer (MTI is the main product of the Indexing Initiative project and has been providing indexing recommendations based on the Medical Subject Headings (MeSH®) vocabulary since 2002)[118].
[4] National Institutes of Health
[5]https://nihcc.app.box.com/v/ChestXray-NIHCC
[6] (atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia)
[7]http://peir.path.uab.edu/library/index.php?/category/106
[8]Term Frequency–Inverse Document Frequency
[9]https://archive.physionet.org/physiobank/database/mimiccxr./
[10] Medical Information Mart for Intensive Care-Chest X-ray
[11]https://stanfordmlgroup.github.io/competitions/chexpert./
[12] Rule-based approach
[13]http://bimcv.cipf.es/bimcv-projects/padchest./

بیمار است. فراوانی انواع ۶ نمای مختلف تصاویر این مجموعه‌داده در (جدول ۴)نشان داده شده است. تفسیرهای تصاویر توسط رادیولوژیست ها در بیمارستان سن‌خوان اسپانیا از سال ۲۰۰۹ تا۲۰۱۷ نوشته شده است. برچسب بیش از ۲۰۰۰۰ تصویر (۲۷%) با کمک پزشکان متخصص از گزارش‌ها به صورت دستی استخراج شده است. برچسب بقیه تصاویر بصورت خودکار با کمک یک شبکه عصبی بازگشتی تولید شده است. به طور کلی، ۲۹۷ برچسب استخراج شده است که به دسته‌های مختلف از جمله ۱۷۴ برچسب به عنوان یافته‌های مختلف رادیوگرافی، ۱۹ برچسب تشخیصی و ۱۰۴ محل آناتومیک تقسیم بندی شدند. این برچسب‌ها براساس اصطلاحات استاندارد یکپارچه سیستم زبان پزشکی[1](UMLS) انتخاب شده‌اند.

جدول (۴): فراوانی انواع ۶ نمای مختلف تصاویر درمجموعه‌داده‌ی PadChest[49]

Projection and Positioning	Images
PA	96010
L	51124
AP-Horizontal	12355
AP-Vertical	5158
Costal	631
Pediatric	274



شکل (۳): احتمال رخداد برچسب‌های مختلف در[45]CheXpert

۲-۵ مجموعه‌دادهICLEFCaption

مارک ساندرسون و پل کلاف از گروه مطالعات اطلاعات دانشگاه شفیلد، کمپین ارزیابی[2]ImageCLEF را پیشنهاد دادند که از سال ۲۰۰۳، هر سال به عنوان بخشی از آزمایشگاه‌های CLEF[3] برگزار می‌شود؛ مجموعه‌داده‌های معرفی شده در این کمپین هر سال تغییر می کند و وظایف مختلف شامل: بازیابی چند رسانه‌ای، تفسیر، و نمایه سازی را دربرمی‌گیرد. شرکت کنندگان زیادی از سراسر جهان برای انتشار پیشنهادهای نوآورانه، بر اساس مجموعه‌داده‌های ارائه شده در آن شرکت می کنند.تفسیر تصاویر پزشکی از سال ۲۰۱۷ در زمره وظایف این کمپین قرار گرفته است.

مجموعه‌داده[27]ICLEF-Caption[4]که در سال ۲۰۱۷ ارائه شد؛ برای دو هدف جمع‌آوری شده بود: پیش‌بینی مفاهیم و تولید تفسیر. به هر تصویرشناسه‌های منحصربه‌فرد مفهومی CUI[5] اختصاص یافته است که براساس اصطلاحات استاندارد یکپارچه سیستم زبان پزشکی[50]می‌باشد.یک تصویر معمولاً با چندین CUI مرتبط است(شکل۴).گردآوری کنندگانِمجموعه‌داده، از یک روش خودکارِ سلسه مراتبی براساس نوع تصاویر پزشکی[51]، برای طبقه بندی ۵,۸ میلیون تصویر استخراج شده از مقالات پزشکی PubMed[6] استفاده کردند. هدف هدف آنها استخراج تصاویر بالینی و یا غیر بالینی و همچنین حذف تصاویر ترکیبی (مانند تصاویر متشکل از چندین تصویرِ اشعه ایکس) بود. از آنجا که استخراج تصاویر خودکار انجام می‌شد، آنها نویز کلی در مجموعه‌دادهرا ۱۰٪ الی ۲۰٪ پیش‌بینی کردند. در نهایت۱۸۴۶۱۴ تصویر استخراج شده و تفسیرهای هر تصویر معمولاً از چند جمله تشکیل شده است.کاستی‌های متعددی در این مجموعه‌داده وجود دارد که شامل تناقض در مجموعه‌داده، تنوع تصاویر، تفسیرهای تکراری، تفسیرهایی با طول متفاوت و همچنین تفسیرهایی که هیچ برچسب مرتبطی با آنها وجود ندارد، می‌باشد [27].

در سال ۲۰۱۸، سازمان‌دهندگان از یک شبکه عصبی کانولوشنال (CNN) استفاده کردند تا همان ۵,۸ میلیون تصویر را بر اساس نوع آن‌ها طبقه‌بندی کنند که منجر به ۲۳۲۳۰۵ تصویر به همراه زیرنویس‌های مربوط به آنها شد. اگرچه آنها گزارش دادند که تصاویر ترکیبی کاهش یافته است، اما آنها اشاره کردند که نویز و تصاویر غیربالینی (مثلاً، تصاویر نقشه‌ها)هنوز وجود دارد(شکل۵)، و همچنین مجموعه‌داده با مشکلات قابل توجهی هنوز درگیر است که می توان موارد زیر را ذکر کرد:

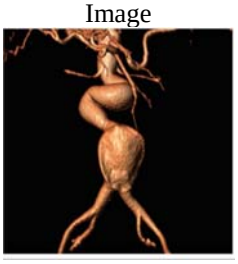
- ناسازگاری در مجموعه‌داده وجود دارد.
- تصاویر بسیار متنوع است.
- تفسیر تکراری دارد (۱,۴%).
- طول جملات در تفسیرها یکسان نیست از ۱ تا ۸۱۶ کلمه متغیر است.
- طول تفسیر بطور متوسط ۲۱ کلمه و تعداد واژگان برابر ۱۵۷۲۵۶ می‌باشد.

[1]Unified Medical Language System
[2]Image Cross Language Evaluation Forum
[3]Cross Language Evaluation Forum (CLEF)

[4]http://www.imageclef.org/
[5] Concept Unique Identifiers(CUI)
[6]https://www.ncbi.nlm.nih.gov/pmc/

- برای تعدادی از تفسیرها هیچ برچسب خاصی تعریف نشده است.
- ۱۱۱۱۵۵ مفهوم منحصربه‌فرد از ۲۲۲۳۰۵ تفسیر استخراج شده‌است. هر تصویر بطور متوسط با ۳۰ مفهوم از اصطلاحات یکپارچه در زبان پزشکی مرتبط است، درحالی که کمتر از ۶۰۰۰ تصویر دارای یک یا دو مفهوم هستند، تصاویری وجود دارد که با هزاران مفهوم مرتبط می‌باشند. تهیه کنندگان به وجود نویز در تعیین مفاهیم معترفند و آن را ناشی از فرایند خودکار تولید مفاهیم می‌دانند که منجر به استخراج مفاهیم بی‌ربط شده است.

Image



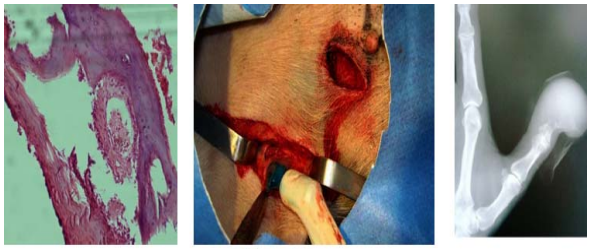
Concept detection:

C0002940: Aneurysm
C0002978: angiogram
C0027530: Neck
C0087111: Therapeutic procedure
C0524425: inside the blood vessel
C0524865: Reconstructive Surgical Procedures
C3887704: treatment – ActInformationManagementReason

Caption prediction:

Preoperative computed tomographic angiography reconstruction showing hostile neck anatomy amenable to treatment with endovascular aneurysm sealing (EVAS).

شکل (۴): نمونه‌ای از یک تصویر و اطلاعات ارائه شده در ICLEFCaption2017[27]



(a) Relevant images.

(b) Irrelevant images.

شکل (۵): نمونه‌هایی از تصاویر مرتبط(الف)، نمونه‌هایی از تصاویر غیر مرتبط (ب) در مجموعه‌داده ICLEFCaption2018[28]

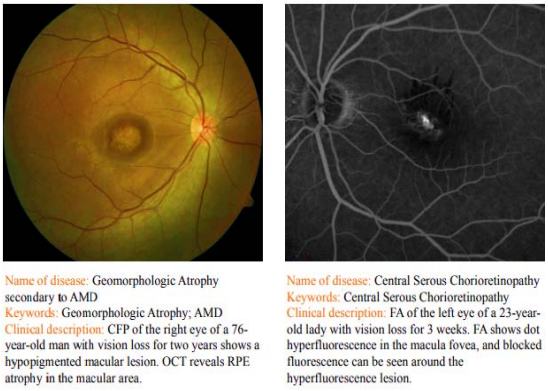
تعداد مفاهیم منحصربه‌فرد در نسخه‌هایی که درسال‌های بعد ارائه شد،روند کاهشی داشت. در مجموعه داده‌ی سال ۲۰۱۹ تعداد مفاهیم برابر ۵۵۲۸ و فقط دربرگیرنده تصاویر رادیولوژی بود؛در سال ۲۰۲۰ تعداد مفاهیم برابر۳۰۴۷ و انواع تصاویر پزشکی در مجموعه‌داده گنجانده شده‌بود. مفاهیم منحصربه‌فرد در سال ۲۰۲۱ به ۱۵۸۵ مفهوم کاهش یافت؛تصاویر موجود، تصاویر رادیولوژی واقعی است که تفسیر آن توسط پزشکان انجام شده است. همچنین شامل داده‌هایی از مجموعه‌دادهROCO[39]نیز می‌باشد [52].

مجموعه‌داده‌ی ارائه شده در نسخه ۲۰۲۱، حاوی تصاویر رادیولوژی واقعی بود که توسط پزشکان تفسیر شده بود. که این تغییر منجر به مفاهیم با کیفیت بالا شد. از آنجا که دستیابی به داده‌هایی با کیفیت مشابه به سختی ممکن است، بنابراین در سال ۲۰۲۲ تصمیم گرفته شد از نسخه توسعه‌یافته مجموعه‌داده ۲۰۲۰ استفاده شود. آنها برای کاهش دامنه و اندازه مفاهیم، چندین ابزار استخراج مفهوم مورد تجزیه و تحلیل قرار دادند و مفاهیم با تکرار کمتر را حذف کردند[1]. تعداد کل تصاویر برابر ۹۷,۹۶۵ است[53].

۲-۶ مجموعه‌داده DeepEyeNet

این مجموعه‌داده[2] شامل تصاویر شبکیه چشم به همراه تفسیر و کلمات کلیدی برای هر تصویر می‌باشد. چشم‌پزشکان باتجربه در جمع‌آوری این مجموعه‌داده همکاری داشته و با استفاده از نظرات تخصصی آنها، مجموعه‌داده، براساس ۲۶۵ مورد منحصربه‌فرد از علائم شبکیه با تفسیر بالینی گردآوری شده‌است. در این مجموعه‌داده، دو نوع تصویر شبکیه وجود دارد: خاکستری FA[3] و رنگی CFP[4]. تعداد کل تصاویر ۱۵۷۰۹، شامل ۱۸۱۱ تصویر FA و ۱۳۸۹۸ تصویر CFP است[54]. هر تصویر شبکیه دارای سه برچسب متناظر است که شامل نام بیماری ، کلمات کلیدی و توضیحات بالینی است. مجموعه‌داده شامل ۲۶۵ بیماری مختلف شبکیه از جمله موارد شایع و غیر شایع می‌باشد و حاوی ۱۵۷۰۹ تفسیر و ۱۵۷۰۹ برچسب کلمات کلیدی است. کلمات کلیدی نشان دهنده اطلاعات مهم در روند تشخیص است و توصیف بالینی مربوط به یک تصویر شبکیه را نشان می‌دهد. تمام برچسب ها توسط متخصصان شبکیه یا چشم پزشکان تعریف شده‌اند (شکل ۶). همچنین، بیشترین فراوانی در تعداد کلمه برای کلمات کلیدی بیشتر از ۱۵ و برای توصیفات بالینی بیشتر از ۵۰ کلمه است.

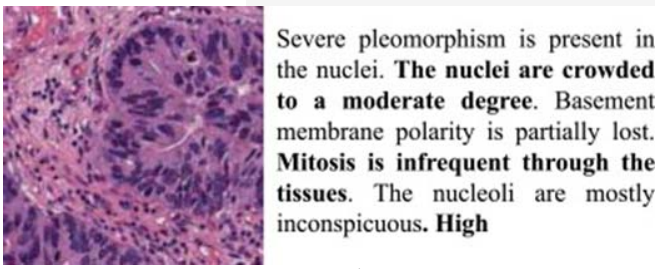
[1]https://www.imageclef.org/2022/medical/caption
[2]Dataset request email: deepeyenet.den@gmail.com
[3]FluoresceinAngiography
[4]Color Fundus photography



شکل (۶): نمونه تصویر و گزارش مجموعه‌داده[54]DeepEyeNet

BCIDR ۲-۷

این مجموعه‌داده[1] شامل تصویر و گزارش تشخیصی سرطان مثانه و حاوی ۱۰۰۰ تصویر رنگی ۵۰۰*۵۰۰ به همراه گزارش مربوطه است [55] که به کمک پاتولوژیست‌ها جمع آوری شده‌است. گزارش پاتولوژی توصیف مشاهدات با بررسی پنج نوع از ویژگی‌های ظاهری سلول شامل وضعیت پلومورفیسم هسته‌ای، ازدحام سلول، قطب سلول، میتوز و برجستگی هسته‌ها می‌باشد و در انتها تشخیص ارائه شده‌است (شکل۷). نتیجه گیری در چهار کلاس انجام شده‌است. همچنین، چهار پزشک (غیر متخصص در سرطان مثانه) هر کدام یک گزارش برای هر تصویر نوشته‌اند. اما برای تضمین صحت، اغلب به گزارش توصیفی پاتولوژیست مراجعه می‌شود. بنابراین در کل پنج گزارش وجود دارد. طول هر گزارش بین ۳۰ تا ۵۹ کلمه متفاوت است.



شکل (۷): نمونه تصویر و گزارشمجموعه‌داده[11]BCIDR

۳ معیارهای ارزیابی

در ارزیابی خروجی روش‌های تفسیر خودکار تصاویر، پیچیدگی‌های زیادی وجود دارد. برای توصیف یک تصویر، می‌توان جملات متنوعی عنوان کرد که همه به نوعی درست باشد یا حتی یک مفهوم را با لغات مترادف بیان کرد که همه آنها به درستی تصویر را توصیف کند؛ از این رو ارزیابی خروجی روش‌های توصیف خودکار تصاویر دشوار است. انسان معتبرترین و تواناترین ارزیاب برای تعیین کیفیت جمله تولید شده توسط الگوریتم‌های مولد تفسیر خودکار می‌باشد. ارزیابی گزارش‌هایی که با روش‌های مختلف ایجاد می‌شوند را می‌توان بوسیله قضاوت‌های انسانی و به صورت شهودی انجام داد. اما این امر مستلزم تلاش و زمان قابل توجهی است که باعث می‌شود فرایند ارزیابی گران و دشوار شود. علاوه بر آن ارزیابی‌های انجام شده، قابل تعمیم نیست. بنابراین لزوم بررسی کیفیت تفسیرهای تولید شده با استفاده از معیارهای ارزیابی خودکار احساس می‌شود[11]. وظیفه این معیارها، مقایسه جملات تولید شده در سیستم‌های تولید تفسیر خودکار با جملات مرجع[2] می‌باشد؛ در ادامه به بررسی این معیارها پرداخته می‌شود.

BLEU ۳-۱

این معیار ارزیابی، اولین بار توسط کیشور پاپینی و همکاران[22] برای ترجمه ماشینی و به منظور تحلیل همبستگی بین ترجمه تولید شده توسط ماشین و ترجمه مرجع ارائه شد. ایده اصلی به این صورت است که هرچه ترجمه ماشینی به ترجمه حرفه‌ای انسانی از لحاظ تعداد کلمات و ترتیب رخداد آنها شبیه‌تر باشد، ترجمه انجام شده بهتر خواهد بود. در این روش عبارات تولید شده با طول‌های مختلف با جملات اصلی نوشته شده توسط انسان مطابقت داده می‌شود. مقایسه بین جمله تولید شده و جمله ترجمه مرجع برحسب انگرم (از ۱ تا ۴ گرم) انجام می شود. برای جملات کوتاه، ترم مجازات اختصار، به هر نمره اضافه می شود. میانگین هندسی همه نمرات در امتیاز مجازات ضرب می‌شود.BLEU با استفاده از رابطه ۱ محاسبه می شود[22]:

$$\text{Modified Unigram Precision}(p_n) = \frac{\text{Sum of clipped unigrams of candidate caption}}{\text{Sum of all unigrams of all groundtruth captions}} \quad (1)$$

$$\text{BrevityPenality}(BP) = \begin{cases} 1 & if\ c > g \\ e^{(1-\frac{g}{c})} & if\ c \le g \end{cases}$$

$$BLEUscore = BP * e^{\sum_{n=1}^{N} w_n \log p_n}$$

که c طول جمله پیشنهادی کاندید است و g طول جمله اصلی مرجع می‌باشد. N طول کل انگرم مورد استفاده برای محاسبه p_n است و w_n وزن و برابر عددی مثبت بین ۰ تا ۱ است. محدوده نمره BLEU از ۰ تا ۱ است که هرچه به ۱ نزدیکتر باشد تطابق جمله پیشنهادی به جمله مرجع بیشتر است.اکثر محققان از BLEU در ارزیابی نتایج استفاده می کنند,[56] [33]تا[59].

ROUGE-L ۳-۲

معیار ارزیابی ROUGE-L[3]، مقادیر Precision, Recall و معیار F1 را بر اساس طول طولانی‌ترین زیرترتیب مشترک بین جملات تولید شده و جملات مرجع محاسبه می‌کند. ROUGE انواع مختلفی دارد که بر اساس حذف ایست‌واژه‌ها[4]، ریشه‌یابی[1] و انگرم

[1]https://figshare.com/projects/nmi-wsi-diagnosis/61973

[2]Ground Truth
[3]Recall Oriented Understudy for Gisting Evaluation- Longest common subsequence
[4]Stop words

متفاوت است، اما بیشتر برای ارزیابی توصیف تصاویر پزشکی ازROUGE-L که در رابطه ۲ آمده است استفاده می‌شود[21]:

$$Recall(R_{lcs}) = \frac{LCS(c,g)}{m} \quad (2)$$

$$Precision(P_{lcs}) = \frac{LCS(c,g)}{n}$$

$$ROUGH-L\ score = F_{lcs} = \frac{R_{lcs}P_{lcs}(1+\beta^2)}{R_{lcs}+P_{lcs}\beta^2}$$

که در اینجا $LCS(c,g)$طول طولانی‌ترین زیرترتیب مشترک در تفسیر پیشنهادی c و تفسیر مرجع g است و m طول جمله مرجع و n طول جمله پیشنهادی می‌باشد. β بصورت رابطه۳ محاسبه می‌شود:

$$\beta = \frac{P_{lcs}}{R_{lcs}} \quad when \quad \frac{\partial F_{lcs}}{\partial R_{lcs}} = \frac{\partial F_{lcs}}{\partial P_{lcs}} \quad (3)$$

METEOR۳-۳

این معیار ابتدا تفسیر را با محاسبهBLEU-1بین جمله مرجع و جمله تولیدی ارزیابی می‌کند. سپس میانگین هارمونیک بر اساس Precision و Recall نتایج تطابق یافته محاسبه می‌شود. علاوه بر تکگرم[۲]؛ ریشه‌ها، مترادف کلمات و جدول ترجمه نیز برای مطابقت دو جمله در نظر گرفته می‌شود. برای تفسیرهای طولانی‌تر، در صورت وجود شکاف، کلمات مرتب شده مختلف و مطابقت‌های غیر معمول تکگرام تا ۵۰ درصد مجازات در نظر گرفته می شود. نمره METEOR[۳]با استفاده از رابطه۴ محاسبه می شود[60]:

$$Unigram\ Precision(P) = \frac{u_{cg}}{U_c} \quad (4)$$

$$Unigram\ Recall(R) = \frac{u_{cg}}{U_g}$$

$$Harmonic\ mean\ (F\ mean) = \frac{10PR}{R+9P}$$

$$Penalty(p) = \left(\frac{C}{U_m}\right)^3 * 0.5$$

$$METEOR\ score = F\ mean * (1-F)$$

در اینجا U_{cg} مجموع تعداد تکگرم های تطبیق یافته بین تفسیر پیشنهادی و تفسیر مرجع است. U_c و U_g مجموع تعداد تکگرم‌های موجود در متن کاندید تولید شده و متن اصلی مرجع است. C مجموع chunkهاست (مجموعه‌ای از تکگرم‌هایی که در شرح‌نویسی مرجع وجود دارد و آنها در شرح‌نویسی پیشنهادی نیز هم‌جوارند). U_mمجموع تعداد تکگرام‌های تطبیق یافته‌اند.

CIDEr۳-۴

معیار[۴]CIDEr به طور خاص برای ارزیابی تفسیر تصویر توسعه یافته‌است همچنین برای ارزیابی زیرنویس فیلم نیز استفاده می‌شود. در این روش ابتدا برای همه لغات موجود در جملات مرجع و جملات پیشنهادی ریشه آنها در نظر گرفته می‌شود سپس انگرم برای هر جمله محاسبه می‌شود. آنگاه شباهت کسینوسی بین تفسیر ایجاد شده و تفسیر مرجع بر اساس فرکانس (TF-IDF) هر انگرم محاسبه می‌شود. میانگین نمرات انگرم به عنوان نمره نهایی بازگردانده می‌شود. روش محاسبه CIDEr در رابطه ۵ آورده شده است:

$$CIDEr_n = \frac{1}{m}\sum_j \frac{g^n(c_i).g^n(s_{ij})}{\|g^n(c_i)\|\|g^n(s_{ij})\|} \quad (5)$$

$$CIDEr\ score = \sum_{n=1}^{N} w_n CIDEr_n$$

که در آن $g^n(c_i)$و$g^n(s_{ij})$ بردارهای تشکیل شده از وزن TF-IDFبرای هر انگرم به ترتیب برای جملات پیشنهادی و جملات مرجع است. w_n وزن نرمال یکنواخت و برابر1/Nمی‌باشد[20].

SPICE۳-۵

معیار جدید اندازه گیری SPICE[۵] بر اساس مفاهیم معنایی است. این معیار اشیاء، ویژگیها و روابط بین آنها را از زیرنویس‌های مرجع اصلی استخراج می‌کند، و در برابر هر زیرنویس یک چندتایی تشکیل می‌دهد که حاوی تمام اطلاعات فوق است. Precision, Recall و معیار F1برای یافتن نتایج نهایی محاسبه می‌شود. SPICE با استفاده از رابطه ۶ محاسبه می‌شود[61]:

$$Percision = \frac{T(G(c))\otimes T(G(g))}{T(G(c))} \quad (6)$$

$$Recall\ (R) = \frac{T(G(c))\otimes T(G(g))}{T(G(g))}$$

$$SPICE\ score = F_1 = \frac{2PR}{P+R}$$

که در آن T یک تابع است که چندتایی‌هایی را از زیرنویس‌های پیشنهادی ((G (c) و زیرنویس مرجع ((G (g) برمی گرداند. ⊗ تابعی است که تاپل های منطبق با هم را از هر دو نمودار برمی‌گرداند.

۳-۶ مقایسه معیارهای ارزیابی

از میان معیارهای فوق BLEU و ROUGE بیشتر از سایر معیارها استفاده می‌شوند. هیچ کدام از معیارهای فوق کاملاً با تصمیم‌های انسانی مطابقت ندارد و همه آنها نقاط قوت و محدودیت‌هایی دارند. اگرچه معیار BLEU با قضاوت‌های انسانی ارتباط دارد، اما در تطبیق صریح انگرم، عملکرد پایینی دارد. در BLEU، ROUGE-L و CIDEr، ترتیب کلمات مهم است زیرا در این معیارها از تطبیق دقیق انگرم استفاده می‌شود[62]، اما ترتیب بر SPICE تأثیر نمی‌گذارد زیرا تطبیق مترادف را در سطح جمله انجام می‌دهد. METEOR ابتدا ریشه‌یابی را انجام می‌دهد و تطبیق ترجمه و مترادف را انجام می‌دهد، بنابراین در سطح جمله به

[1]Stemming
[2]Unigram
[3] Metric for Evaluation of Translation with Explicit Ordering
[4]Consensus-based Image Description Evaluation
[5]Semantic Propositional Image Caption Evaluation

خوبی عمل می‌کند، اما گاهی شباهت معنایی آسیب‌زننده است. اگر کلمات با مترادف آنها جایگزین شوند، نمره ارزیابی همه معیارها کاهش می‌یابد[62]. الیوت و همکاران[63] بیان کردند که نمی‌توان همیشه بین روش‌های خودکار ارزیابی تفسیر تصویر و قضاوت‌های انسانی همبستگی کامل یافت.

۴ تفسیر خودکار تصاویر پزشکی

در راستای تسهیل فرایند تفسیر خودکار تصاویر پزشکی، با توجه به امکانات سخت‌افزاری و نرم‌افزاری،روش‌های متعددی در پیش گرفته شده است. در مطالعات ابتدایی به علت عدم رونق روش‌های مبتنی بر یادگیری عمیق از سایر روش‌های یادگیری ماشین استفاده می‌شد و البته نتایج چندان مطلوب نبود و همچنین سیستم تمام اتوماتیکی که تصویر را دریافت و در خروجی تفسیر را تولید نماید؛ پیاده‌سازی نشد.

اولین مطالعاتی که در زمینه تولید تفسیر خودکار تصاویر پزشکی انجام شد،در سال ۲۰۰۳ توسط دی هوسکه- کراوس و همکاران[65] ,[64] صورت گرفت که تاکیدی بر این نکته داشت که جامعه انفورماتیک پزشکی باید خود را متعهد به استفاده از ایده NLG[1]در پزشکی کنند؛آنها سیستمی‌را طراحی نمودند که پاره‌ای از لغات استاندارد یکپارچه در زبان پزشکی و تعدادی لغات فوق تخصصی را به زبان آلمانی به عنوان ورودی دریافت می‌کرد و گزارش‌های قابل قبول از نظر پزشکان تولید می‌کرد. این برنامه در بیمارستان به کار گرفته‌شد و عملاً تلاشی در جهت تولید گزارش‌های خودکار بود در این پژوهش استخراج خصوصیات از تصاویر پزشکی توسط پزشک متخصص انجام می‌شد و سیستم پیشنهادی تنها گزارش را تولید می‌کرد.

سباستین وارژس و همکاران [66] در مقاله تحقیقاتی خود از روشی مانند دی هوسکه- کراوس پیروی کردند؛ با این تفاوت که بجای استفاده از لغات استاندارد یکپارچه در زبان پزشکی و لغات فوق تخصصی، از یک پیکره متنی با ۷۰ گزارش نوشته شده توسط پزشکان، با ۷۲۴ عبارت منحصربه‌فرد استفاده کردند.گزارش‌هاییکه آنها تولید کردند از نظر محققان از لحاظ کمی و کیفی مورد تایید قرارگرفت. در روش آنها نیز واژگان کلیدی بصورت دستی وارد می‌شد و بصورت خودکار از تصاویر استخراج نمی‌شد.

پی کیسیلف و همکاران [67]، از408تصویر سونوگرافی و ۲۰۳ تصویر ماموگرافی استفاده کردند و سیستمی برای تولید گزارش‌های پزشکی بصورت نیمه‌اتوماتیک ارائه کردند به این صورت که در ابتدا رادیولوژیست باید محل تقریبی ضایعه را مشخص کند سپس با توجه به نواحی تشخیص داده شده به جای تولید گزارش ساختاریافته، لغات مهم و تاثیرگذار در تصمیم تولید شود. پی کیسیلف و همکاران [68] در تلاشی دیگر سعی کردند برای تولید برچسب‌های هر تصویر از SVM استفاده کنند و نتایج بهتری نسبت به مقاله قبلی شان[67]،ارائه دادند.

آنچه در بالا ذکر شد جزء اولین تلاشها در جهت تفسیر تصاویر پزشکی بود؛ سیستم‌های پیشنهادی خودکار نبودند و در هر کدام، بخشی با دخالت انسان انجام می‌شد. رویکردهایی که در سالهای بعد در پیش گرفته شد به طور کلی در سه دسته طبقه بندی می‌شوند:

- روش‌های مبتنی برالگو[2]
- روش‌های مبتنی بر بازیابی[3]
- روش‌های مبتنی بر یادگیری عمیق (کدگذار- کدگشا)[4]

پیش از آنکه به بررسی رویکردهای عنوان شده، پرداخته شود ذکر این نکته ضروری به نظر می‌رسد که یک گزارش پزشکی استاندارد معمولاً شامل بخش مهمی است که در آن اطلاعات فراداده‌های بیمار، شامل علائم بیماری و اطلاعات مربوط به درمانهای قبلی بیمار ثبت می‌شود. استفاده از این اطلاعات همانطورکه در تصمیم گیری پزشک در تفسیر تصویر موثر است می‌تواند در تولید خودکار تفسیر نیز تاثیرگذار باشد. تعدادی از محققان این موضوع را در مدل پیشنهادی‌شان اعمال کردند [57].

ژانگ و همکاران[69] اولین اثری را ارائه کردند که در آن از اطلاعات سابقه بیمار برای تولید بخش تشخیص در گزارش‌های پزشکی استفاده می‌شد. آنها معتقد بودند که بخش پس‌زمینه در گزارش رادیولوژی مهم است، زیرا اطلاعات مهمی مانند هدف مطالعه، عضو درگیر و وضعیت بیمار اغلب فقط در پس‌زمینه ذکر می‌شوند. یک راه ساده برای ترکیب اطلاعات پس‌زمینه این است که تمام متن پس‌زمینه را به یافته‌ها اضافه کنیم و با توجه به آن شبکه آموزش ببیند اما این روش ساده، در واقع نه تنها به بهبود نتایج کمک نمی‌کند بلکه به کیفیت نتیجه نیز آسیب می‌زند، احتمالاً به این دلیل که مدل نمی‌تواند به اندازه کافی بین یافته‌ها و اطلاعات پیش‌زمینه تمایز قائل شود، که در نتیجه منجر به مدل‌سازی ناکافی یافته‌ها و پیش‌زمینه می‌شود. برای حل این مشکل، پیشنهاد شد بخش سابقه بیمار با یک کدگذار توجه جداگانه کدگذاری شود و از نمایش حاصل برای هدایت فرآیند کدگشایی در مدلِ تولید گزارش استفاده شود. مدل ارائه شده بر اساس دو چارچوب بود؛ یک مدل شبکه عصبی دنباله به دنباله[5] و یک شبکه مولد اشاره گر.آنها هر دو بخش یافته‌هاو اطلاعات پیش‌زمینه را به طور جداگانه با استفاده از LSTM دوطرفه کدگذاری کردند و سپس مکانیزم توجه برای هر دو اطلاعات کدگذاری شده محاسبه شد. در نتایج حاصل بهبود قابل توجهی دیده شد.

[1] Natural Language Generation

[2] Template Based
[3] Retrival Based
[4] Deep Learning Based (Encoder-Decoder)
[5] Seq to Seq

در ادامه بطور مختصر روش‌های مبتنی بر الگو و روش‌های مبتنی بر بازیابی بررسی می‌شود. تاکید مقاله در توصیف روش‌های مبتنی بر یادگیری عمیق (کدگذار-کدگشا) خواهد بود.

۴-۱ تفسیر تصاویر پزشکی مبتنی بر الگو

در روش مبتنی بر الگو، ابتدا اشیاء و ویژگی‌های موجود در تصویر بوسیله یک الگوریتم یادگیری ماشین شناسایی می‌شوند و سپس تفسیرها با تبعیت از قوانین خاص (معمولاً اگر-آنگاه) و یا از طریق الگوهایی که برای جملات در نظر گرفته می‌شود، ایجاد می‌شوند. تفسیرهای ایجاد شده ساده و از نظر دستوری صحیح هستند[11]. اما، نقطه ضعف این رویکرد این است که این تفسیرها دارای قالب خاصی هستند و تنوع و انعطاف پذیری در آنها وجود ندارد.در راستای حل این مشکل در پاره‌ای از مقالات روش‌های ترکیبی پیشنهاد شده‌است. در این روش‌ها علاوه بر این که از الگوها استفاده می‌شود در کنار آن امکان ایجاد جملات،کلمه به کلمه وجود دارد. که در نهایت با انتخاب بین یک الگو یا جمله ایجاد شده از ابتدا[38]، یا با ویرایش و یا نقل به مضمون کردن یک الگوی انتخاب شده قبلی[71][70]گزارش نهایی تولید می‌شود. در جدول ۵ نمونه‌ای از الگوهای استفاده شده وانگ و همکاران[72] وجملاتی که بر اساس آن الگوها، مدل پیشنهادی‌شان تولید کرده، نشان داده شده است.آنها از یک شبکه از پیش آموزش دیده، مفاهیم پزشکی را استخراج کردند. مفاهیم پزشکی به چهار دسته معنایی: نوع تصویربرداری، ساختار آناتومیک، یافته‌ها و موارد دیگر تقسیم شدند. و، تصاویر بر این اساس تفسیر شدند.

جدول (۵): نمونه‌ای از الگوهای پیشنهادی برای تولید تفسیر تصاویر پزشکی و جمله تولید شده بر اساس آن[72]

الگو	جمله تولید شده
<image> of <body> demonstrate / show /suggest <findings>	**synpic24243: Sagittal T1-weighted image** of **the cervical spine** demonstrates **cord expansion.**
<image> demonstrate / show / suggest <findings> in/of/within <body>	**synpic19193: Lateralradiograph of the skull** shows **lytic lesions** in **the temporoparietal region**.

در میان مقالات بررسی شده، CLARA[71] یک گردش کار صریح با تعامل انسانی را هدف قرار می‌دهد، که در آن گزارشی با همکاری پزشک تولید می‌شود بدین صورت که متن اولیه توسط پزشک تولید می‌شود و سیستم گزارش به صورت خودکار آن را تکمیل می‌کند.

۴-۲ تفسیر تصاویر پزشکی مبتنی بر بازیابی

رویکرد دوم مبتنی بر بازیابی است؛ این روش بر این فرض تکیه می‌کند که تصاویر مشابه دارای توصیف یکسانی هستند در اینجا،شرح تصویر ورودی با بازیابی یک تفسیر یا مجموعه‌ای از تفسیرها از پیکره‌متنیموجود تولید می‌شود. تفسیر تولید شده جدید می‌تواند برابر توصیف شبیه‌ترین تصویر بازیابی شده یا ترکیب/تجمیع همه تفسیرهای بازیابی شده باشد[73]. اگرچه شرح ایجاد شده از نظر گرامری روان و صحیح است اما نقطه ضعف این روش این است که تفسیرها در برابر ویژگی‌های تصویر موجود نمی‌توانند خود را با اشیاء و صحنه‌های جدید تطبیق دهند و این ممکن است منجر به تولید یک تفسیر بی‌ربط شود.

چارالامپاکوس و همکاران[52] یک رویکرد بازیابی تصاویر بر اساس K-NNپیشنهاد دادند. آنها ابتدا، با استفاده از کدگذارِ از قبل آموزش دیده، یک بردار تعبیه[1] برای هر تصویر آموزشی ایجاد کردند. در طول استنتاج، همان کدگذار، یک بردار تعبیه برای تصویر آزمایشی ایجاد می‌کند، و k تصویر آموزشی با مشابه‌ترین بردار تعبیه بازیابی می‌شوند (از شباهت کسینوسی استفاده شده است). تفسیرهای تصاویر بازیابی‌شده سپس با هم ترکیب می‌شوند تا یک تفسیر برای تصویر آزمایشی ایجاد شود؛ به این ترتیب که r جمله پرتکرار با یکدیگر الحاق شدند تا تفسیر تصویر تولید شود. همچنین آنها از 1-NN نیز استفاده کردند که در صورت استفاده از یک همسایگی عملکرد بهتری گزارش شد.

در مورد بازیابی تصاویر نیز روش‌های ترکیبی پیشنهاد شده است. لیانگ و همکاران[74] از رویکرد بازیابی فقط به عنوان مکملی برای تفسیر تولید شده توسط مدل CNN-LSTM استفاده کردند. از آنجا که تعداد کلمات در جملات مجموعه‌داده بسیار متغیر (بین ۱ تا ۶۰۶) بود و این تنوع، تولید تفسیر را سخت می‌کرد؛ از این رو آنها ابتدا مجموعه‌داده‌ی آموزشی را به سه زیر مجموعه، با طول تفسیر مختلف (۰-۱۳، ۱۳-۳۰، بالای ۳۰) تقسیم کردند. معیار این تقسیم‌بندی به این صورت بود که هر زیرمجموعه ۱/۳ داده ها را پوشش دهد. مدل پیشنهادی آنها شامل سه قسمت بود. ابتدا برای استخراج ویژگی‌های بصری تصویر، از VGGNet از پیش آموزش دیده استفاده کردند و برای ایجاد تفسیر LSTM به کار گرفته شد. برای هر یک از زیرمجموعه‌ها یک مدل مجزا آموزش داده شد، به طوری که در نهایت سه مدل آموزش‌دیده‌ی متفاوتِ کدگذار- کدگشا وجود داشت. بخش دوم، آموزش طبقه‌بندی کننده SVM بود. استفاده از SVMبدین منظور بود که با دریافت ویژگی‌های بصری تصویر پیش بینی کند طول گزارش تولید شده در کدام یک از سه زیرمجموعه قرار خواهد گرفت و در نتیجه مشخص کند از کدام یک از سه مدل در هنگام تولید تفسیر برای یک تصویر آزمایشی استفاده شود.

قسمت سوم مدل پیشنهادی، بازیابی تفسیر بود. ابتدا فاصله اقلیدسی بین ویژگی‌های بصری تصویر آزمایشی با ویژگی‌های بصری سایر تصاویر محاسبه و شبیه ترین تصویر انتخاب می شود. اگر این فاصله از یک مقدار آستانه بزرگتر بود تفسیرِ تصویرِ بازیابی شده به عنوان مکمل به تفسیر تولید شده در مرحله قبل اضافه و تفسیر نهایی تولید می شود. نویسندگان اذعان داشتند که،

[1]Embedding Vector

مدل پیشنهادی برای ایجاد تفسیرهای کاملاً توصیفی و پیچیده مناسب نیست.

اگرچه می‌توان با ترکیب تفسیرهای مبتنی بر بازیابی با شبکه عصبی عمیق، به عملکرد بهتری دست یافت، اما هنوز تحقیقات کمی در تولید گزارش‌های پزشکی یا شرح تصاویر پزشکی بر اساس بازیابی تصاویر انجام شده است.

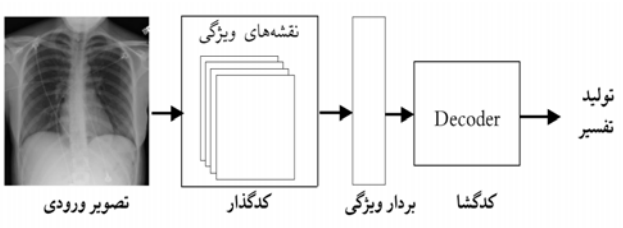
۴-۳ تفسیر تصاویر پزشکی بر پایه یادگیری عمیق (کدگذار- کدگشا)

اخیراً پردازش زبان طبیعی با استفاده از شبکه‌های یادگیری عمیق [76] ,[75] در نوشتن تفسیر تصاویر به موفقیت‌های قابل توجهی دست یافته‌است. این امر محققان را بر آن داشته است تا از روش‌های یادگیری عمیق برای شرح تصاویر پزشکی نیز استفاده کنند.رویکرد تحقیقات اخیر بیشتر استفاده از معماری کدگذار- کدگشا می‌باشد. معماری کدگذار- کدگشا که توسط کیروس و همکاران [77]پیشنهاد شده است،نوعی از شبکه‌های عصبی است که برای ترجمه ماشینی طراحی شده‌است که در آن ابتدا هر جمله به کدگذار وارد می‌شود تا به بردار حاوی اطلاعات معنایی جمله ورودی کدگذاری شود. این بردار سپس به کدگشا وارد شده تا به جمله‌ای به زبان مقصد ترجمه شود. همین ایده برای تولید تفسیر تصاویر اعمال می‌شود و موثر است. به این صورت که کدگذار تصویر را دریافت و ویژگی‌های تصویر را استخراج کرده و آن را در بردار ویژگی با طول ثابت (بردار تعبیه) کدمی‌کند سپس بردار ویژگی به کدگشا داده می‌شود تا کلمه به کلمه تفسیر را تولید کند. (شکل۸).

در اکثر موارد کدگذار یک شبکه عصبی کانولوشن است که ویژگی‌های بصری تصاویر را به صورت سلسله مراتبی استخراج می‌کند.برای آموزش شبکه می‌توان از ابتدا آموزش را بر روی مجموعه‌داده انجام داد و یا از مدل‌های از پیش آموزش دیده مانند VGGNet[78]، ResNet[79] و Inception-V3[80] استفاده کرد و با استفاده از مجموعه‌داده شبکه را تنظیم دقیق کرد.کدگشا یک ماژول تولید کننده زبان است که معمولاً در پیاده‌سازی آن از شبکه‌های عصبی معروف پرکاربرد در حوزه‌پردازش زبان طبیعی (مانند LSTM،GRU و Transformer) استفاده می شود و وظیفه تولید تفسیر را عهده‌دار است. بطور معمول در روند تولید گزارش مراحل زیر دنبال می‌شود:

- پیش پردازش
- استخراج و کدگذاری ویژگی‌های تصویر
- تولید گزارش

که در ادامه مورد بررسی قرار می‌گیرد.



شکل (۸): معماری کدگذار- کدگشا

۴-۳-۱ پیش‌پردازش

به منظور بهبود کیفیت تصویر، حذف نویز و گاهی افزایش اندازه مجموعه‌دادهنیاز به استفاده از تکنیک‌های مختلف پیش پردازش احساس می‌شود واین امر در حصول نتایج بهتر در تولید تفسیر تصاویر موثر است[81]. این تکنیک‌ها ممکن است شامل حذف نویز از تصاویر با استفاده از فیلترهای مختلف، نرمال‌سازی، آینه سازی، تغییر اندازه، برش و چرخاندن باشد. همچنین لازم است بر روی پیکره متنی تفسیرهای مرجع هم پیش‌پردازش انجام شود، که می‌توان جداسازی توکن‌ها، تبدیل همه توکن‌ها به حروف کوچک، حذف نشانه‌های خاص، حذف ایست‌واژه‌ها، ریشه‌یابی، حذف جداکننده‌ها و حذف اعداد را نام برد.

با توجه به رونق گرفتن Contrastive Learning در سالهای اخیر روش‌های پیش پردازش Contrastive نیز در دستور کار تعدادی از پژوهشگران قرار گرفته است. تاکید Contrastive Learning بر اهمیت پیش پردازش و تنظیم دقیق می‌باشد و تلاش آن در این جهت است که پیش از آنکه مدل بر روی مساله خاص آموزش ببیند، ابتدا بصورت یادگیری خودنظارتی یک بازنمایی مناسب از داده‌ها را بیاموزد و سپس این مدلِ از پیش آموزش داده شده، بر روی داده‌های برچسب‌دار تنظیم دقیق شود. در این راستا با استفاده از توابع تبدیل هر داده، اندکی تغییر داده می‌شود بطوری که تفاوت داده جدید با داده اصلی زیاد نباشد. وظیفه تابع هزینه Contrastive این است که فاصله بازنمایی داده اصلی و داده تغییر داده شده آن را، کمینه کند[82].در این راستا پیش پردازش‌هایی که به عنوان مثال در حوزه تصویر برای خلق تصاویر جدید، می‌تواند استفاده شود، عبارتند از: آینه سازی، چرخش، برش و ... .

چارالامپاکوس و همکاران[52]، مراحل پیش‌پردازشی که بر روی پیکره متنی انجام دادند عبارت بود از: تبدیل کل حروف موجود در متن به حروف کوچک، ریشه‌یابی، حذف علائم گزارشی و ایست‌واژه ها. از آنجا که آنها برای آموزش مدل، از Contrastive Learning استفاده کردند؛ در همین راستا پیش‌پردازشContrastive را نیز با هدف اینکه در فضای ویژگی‌ها، ویژگی‌های مربوط به هر کلاس را تا جای ممکن به هم نزدیک و ویژگی‌های مربوط به کلاس های مختلف را از هم دور کرد، به‌کار گرفتند. بدین منظور هر تصویر را بصورت افقی به چهار تکه تقسیم کردند سپس تصاویر را به نویز گوسی آلوده نمودند. بعلاوه، بر روی تصاویرِ پیش‌آموزش چرخش و آینه‌سازی را نیز اعمال کردند.

لی و همکاران[83] نیز برای کاهش شکاف دامنه بین کدگذار تصویر و مجموعه‌داده از پیش‌پردازش Contrastive استفاده کردند. بدین منظور آنها از روش MOCO[84] برای پیش‌پردازش Contrastive استفاده کردند و پیش پردازش‌هایی که انجام دادند به اینصورت بود که: یک برش ۲۲۴*۲۲۴ پیکسل از هر تصویر به طور تصادفی انتخاب شد و بر روی تصویر حاصل، تغییر رنگ تصادفی، چرخش افقی تصادفی و خاکستری کردن تصویر با احتمال p انجام شد.

۲-۳-۴استخراج و کدگذاری ویژگی‌های تصویر

استخراج ویژگی فرآیندی است که منتهی به درک ماشین از تصویر می‌شود و گام مهمی است که در تولید نتایج بهتر تفسیر،تاثیر مستقیم دارد. می توان آن را یک روش کاهش ویژگی، خلاصه کردن، یا مکانیسم کدگذاری برای ایجاد ویژگی‌های مربوطه دانست.استخراج ویژگی به دو صورت ویژگی‌های بصری و ویژگی‌های معنایی می تواند لحاظ شود. شبکه‌های عصبی کانولوشنال و اخیراً شبکه‌های ترنسفورمر بینایی کانولوشنال[1] برای این منظور به کار می‌روند. از خروجی‌های لایه قبل از لایه‌های کاملاً متصل به عنوان ویژگی‌های بصری استفاده می شود و در صورت استفاده از ویژگی‌های معنایی تعدادی از کلمات کلیدیبه عنوان ویژگی معنایی از شبکه استخراج می‌شود. با توجه به نوع ویژگی‌های استخراج شده و در ادامه تولید متن، دو رویکرد وجود دارد که عبارتند از: رویکردهای بالا به پایین و پایین به بالا. در رویکرد بالا به پایین، یک تصویر با ترجمه بازنمایی‌های بصری به متن توصیف می‌شود و در رویکرد پایین به بالا، ویژگی‌های معنایی تولید و سپس با استفاده از مدل های زبانی در جملات ترکیب می‌شوند[85].

روش‌های عمده‌ای که کدگذار بر پایه آن بنا می شود عبارتند از:

۱- کدگذار مبتنی بر شبکه‌های عصبی کانولوشنال[2]

۲- کدگذار مبتنی بر مکانیزم توجه[3]

۳- کدگذار مبتنی بر گراف[4]

۴- کدگذار مبتنی بر مکانیزم توجه به خود[5]

مقالات بررسی شده، براساس نوعی کدگذاری که در روش پیشنهادی بکار رفته است؛ درجدول ۶درج شده است. در ادامه این روش‌ها بررسی می‌شود.

- **کدگذار مبتنی بر شبکه‌های عصبی کانولوشنال**

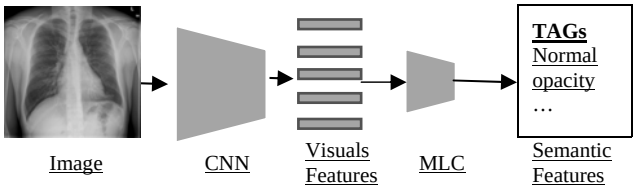
با ظهور شبکه‌هایCNN بهبود قابل توجهی در کارایی تمام مدل‌هایی‌که ورودی تصویر دریافت می‌کردند، ایجاد شد؛ تفسیر تصویر نیز از این قاعده مستثنی نیست. بدین منظور برای استخراج ویژگی‌های بصری، از یکی از آخرین لایه‌های CNNقبل از لایه‌های تمام متصل، استفاده می‌شود (شکل ۹).

اولین بار ایده‌ی تولید بردار ویژگی توسطCNN به عنوان کدگذار، برای تولید تفسیر تصویر در سال ۲۰۱۵ توسط وینالز و همکاران [75] با به کارگیری GoogleNetدر مقاله"Show and tell" پیشنهاد شد. در سال ۲۰۱۷ وو و همکاران[86] همان روش را برای تفسیر تصاویر رتینوپاتی دیابتی به کار بردند و به نتایج قابل توجهی دست یافتند.

کاسترو و همکاران در سال ۲۰۲۱ [87]، برای شرکت در کمپین ICLEF-Captionسه مدل پیشنهاد دادند که یکی از آنها مبتنی بر تولید تفسیر بر پایه ویژگیهای معنایی بود. آنها چندین معماری از جمله ResNetو DensNet که از ابتدا بر روی ImageNet آموزش دیده شده بود؛ را با و بدون تنظیم دقیق به کار گرفتند. بعلاوه یک مدل DenseNet121 که از قبل بر روی مجموعه‌داده-ChestX ray14آموزش داده شده بود، را نیز تنظیم دقیق کردند. برای استخراج ویژگی‌های معنایی آخرین لایه شبکه با یک لایه کاملاً متصل جایگزین شد که خروجی آن با ابعاد واژگان آموزشی مطابقت داشت.N کلمه با بالاترین امتیاز به عنوان خروجی این مرحله برای تولید تفسیر تصویر استفاده می‌شود. سپس، یک رویکرد آماری برای مرتب کردن کلمات در یک جمله منطقی به کار گرفته شد. نتایج حاصله از ResNet منجر به کسب مقام اول در کمپین ICLEF-Caption 2021 گشت.

نمونه‌هایی از مقالاتی که از شبکه های عصبی کانولوشنال برای استخراج ویژگیهای بصری و یا ویژگی‌های معنایی در تفسیرتصاویر پزشکی استفاده کردند،عبارتند از: [27], [15], [13], [88][84], [72], [57], [39], [29]تا[91].

سادگی و فشرده بودن، مزیت اصلی ویژگی‌های بصری استخراج شده توسط CNN می‌باشد. این شبکه‌ها ظرفیت استخراج و فشرده کردن اطلاعات از کل تصویر ورودی، با در نظر گرفتن زمینه کلی تصویر را دارا می‌باشند[92]. اما از سوی دیگر این مقوله می‌تواند، منجر به فشرده سازی بیش از حد اطلاعات و از دست رفتن جزئیات شود، که باعث می‌شود مدل تفسیر خودکار نتواند توضیحات خاص و دقیق را ارائه دهد.



شکل (۹): کدگذار مبتنی بر شبکه‌های عصبی کانولوشنالCNN. شبکهCNN برای استخراج ویژگی تصویر بکار گرفته می‌شود. برای استخراج ویژگی‌های بصری معمولا از لایه قبل از تمام متصل استفاده می شود و برای استخراج ویژگی‌های معنایی از یک طبقه‌بندِ چند کلاسه استفاده می‌شود که تعدادی از ویژگی‌های برجسته در تصویر را تشخیص و در قالب تعدادی برچسب ارائه می‌دهد.(بخش کدگذار از مقاله[12])

[1]Convolutional Vision Transformer(CVT)
[2]Convolutional Neural Network (CNN)
[3]Additive Attention
[4]Graph-based Encoding
[5]Self-Attention

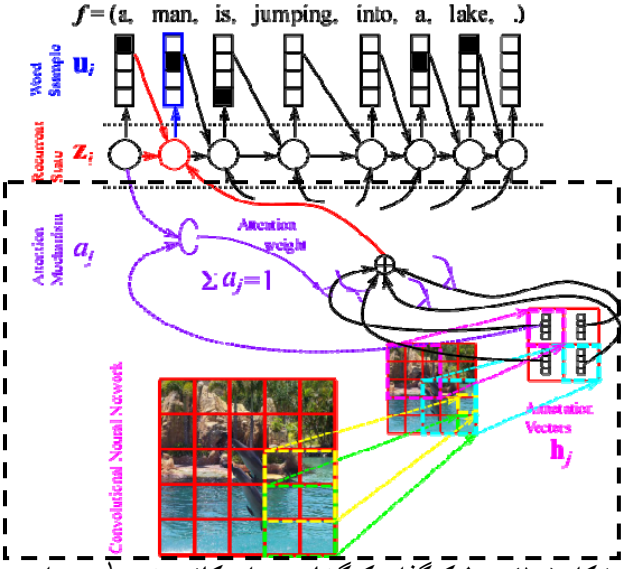
- **کدگذار مبتنی بر مکانیزم توجه**

روند کار در روش‌های کدگذار – کدگشا در سمت کدگشا به این صورت است که (با فرض این که کدگشا، شبکه بازگشتی باشد) در هر مرحله، ویژگی‌های بصری تولید شده توسط کدگذار به همراه کلمه تولید شده در مرحله قبلِ کدگشا، به عنوان ورودی به شبکه بازگشتی وارد شده و حالت پنهان را به‌روز می‌کند، سپس شبکه احتمالِ شرطی کلمه‌ی بعد را تخمین می‌زند. در واقع حالات پنهان به عنوان محورهایی عمل می‌کنند که بین حوزه بصری و زبانی ارتباط برقرار می‌کنند. ویژگی‌های بصری استخراج شده از تصویر معمولاً با استفاده از شبکه کانولوشنال بدست می‌آید. ویژگی متمایز شبکه‌های کانولوشنال این است که هر عنصر خروجی با یک منطقه محلی در ورودی مطابقت دارد. این ویژگی باعث می شود که ویژگی‌های فضایی تصویر، توسط نقشه‌های ویژگی در سراسر لایه‌ها نگهداری شوند[93]. که انگیزه‌ای است که بتوان از مکانیزم توجه[94] استفاده کرد. مکانیزم توجه به این صورت عمل می‌کند که در قسمت کدگشا، برای محاسبه حالت‌های نهان هر گام، یک بردار حاصل از جمع وزن‌دار بردارهای تعبیه قسمت کدگذار محاسبه می‌شود که این وزن‌ها تابعی از میزان مشابهت ویژگی‌های بصری استخراج شده از کدگشا با حالت های نهان کدگشا هستند. نمایی از این ایده را می‌توان در شکل ۱۰ دید.

از آنجا که برخلاف روش استاندارد، مکانیزم توجه قادر به حفظ موقعیت فضایی است، بنابراین ممکن است نقش ساختارهای بصری را در فرآیند تولید تفسیر تقویت کند[93].

هدف از استفاده از مکانیزم توجه این است که تفسیر در برابر اطلاعات و ویژگی‌های برجسته تصویر ایجاد شود و آنها را در اولویت قرار دهد. در مواردی که از مکانیزم توجه استفاده می‌شود نسبت به موارد مشابه که از این تکنیک استفاده نکرده‌اند، معمولاً نتایج بهتری بدست آمده است,[54], [27], [14], [12],[32]تا [56] در ادامه چند نمونه از این مقالات بررسی می‌شود.

مکانیسم توجه ابتدا توسط ژانگ و همکاران [96] در گزارش پزشکی MDNet استفاده شد. مجموعه‌داده انتخابی آنها تصاویر پاتولوژی سرطان مثانه بود. نویسندگان اظهار کردند که در تصاویر پاتولوژی مثانه، تغییر در اندازه و تراکم هسته‌های سلول‌های ادراری یا ضخیم شدن نئوپلاسم یوروتلیال بافت مثانه، نشان دهنده سرطان است. توصیف دقیق این ویژگی‌ها، تشخیص دقیق را تسهیل می‌کند و برای شناسایی سرطان مثانه در مراحل اولیه حیاتی است. اما تشخیص دقیق این تغییرات ظاهری ظریف حتی برای متخصصان با تجربه نیز چالش برانگیز است. آنها از یک ماژول توجه برای کارآمدتر کردن مدل‌شان و توضیح پذیری آن استفاده کردند.ResNet به عنوان کدگذار در جهت استخراج ویژگی‌های تصویر استفاده شد؛ همچنین با حذف بعضی از اتصالات مشکل ناپدید شدن گرادیان را حل کردند. ویژگی‌های بصری استخراج شده فقط به حالت اولیه LSTM وارد شد. از آنجا که اندازه مجموعه‌داده کوچک بود، آنها Regularization را اعمال کردند تا از بیش برازش جلوگیری شود. داده‌ها به روی سه شبکه عصبی کانولوشن آزمایش شد: شبکه عصبی کانولوشن از پیش آموزش دیده، شبکه کانولوشن از پیش آموزش دیده با تنظیم دقیق، و شبکه ساخته شده‌ای که از ابتدا آموزش داده بودند. مدل آموزش دیده آنها به عملکرد بهتری نسبت به شبکه‌های از پیش آموزش دیده و تنظیم دقیق شده، دست یافت. در اینجا استفاده از مکانیزم توجه مؤلفه اصلی است که از تفسیرپذیری بصری شبکه پشتیبانی می کند. آنها مکانیزم توجه را همزمان بر روی، خروجی LSTM، ویژگی‌های بصری و معنایی در جهت مشخص نمودن بخش‌های تاثیرگذار تصویر در تولید گزارش به کار بردند.



شکل (۱۰): مدل کدگذار- کدگشا به‌همراه مکانیزم توجه[1]. پس از استخراج ویژگی‌های بصری توسط شبکه کانولوشنال، مکانیزم توجه بر روی ویژگی‌های بصری استخراج شده اعمال می‌گردد و جمع وزن‌دار ویژگی‌های بصری به کدگشا ارسال می‌گردد. بخش داخل کادر خط‌چین سیاه کدگذار مبتنی بر مکانیزم توجه می‌باشد.

جینگ و همکاران[12] مدل سلسله مراتبی با مکانیزم توجه متقابل[2] ارائه کردند. در روش پیشنهادی آنها، تصویر به مناطقی با اندازه‌های یکسان تقسیم شده و برای استخراج ویژگی به VGG-19 داده شد[78]. ویژگی‌های بصری از آخرین لایه کانولوشن استخراج شد؛ ویژگی‌های بصری استخراج شده به یک شبکه طبقه بندی چند کلاسه داده شد تا احتمال برچسب‌ها را بر روی واژگانی که به عنوان برچسب از پیش مشخص شده، تعیین کند. علاوه بر اینبرچسب‌ها بصورت بردارهای تعبیه کلمه حاوی اطلاعات سطح بالا از ویژگی‌های معنایی نشان داده شدند. هر دو ویژگی معنایی و بصری به مدل توجه متقابل داده شد که امتیازاتی را به بردارهای ویژگی بصری و معنایی اختصاص داد و مجموع وزن آنها جداگانه محاسبه شد که به ترتیب منجر به بردارهای ویژگی بصری و بردارهای ویژگی معنایی شد. هر دو بردار به هم متصل شدند و

[1] https://developer.nvidia.com/blog/introduction-neural-machine-translation-gpus-part-3/

[2] Co-attention

یک بردار مشترک حاوی اطلاعات سطح بالا، بصری و معنایی تصویر تولید شد و به عنوان ورودی کدگشا در نظر گرفته شد.

جدول (۶): روش‌های عمده‌ای که کدگذار بر پایه آن بنا می شود.

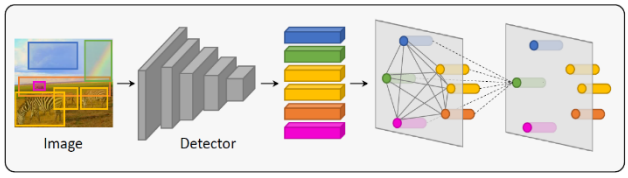
	GoogleNet	VGG	Inception –V3	DenseNet	ResNet
Non Attentive Global CNN Features or Semantic Features	[97][86]	[74][32]	[73][91][88][86][89]	[34][15][90][13][32]	[59][29][33][13][87][96]
Additive Attention	[12]	[29]			[57][56][58][14]
Graph-based	[70][90]			[68]	
Self Attention	[13]	[98]			[98][99]

- **کدگذار مبتنی بر گراف**

در بعضی از مطالعات در جهت کدگذاری بهتر و یافتن روابط بین مناطق مختلف تصویر و گنجاندن روابط معنایی و مکانی تصویر، استفاده از شبکه‌های گراف (شکل۱۱) پیشنهاد شده‌است.اولین تلاش در این راستا در زمینه تصاویر طبیعی (مجموعه‌داده COCO) توسط یائو و همکاران انجام شده است[100]و پس از آن گوو و همکاران[101]، استفاده از یک شبکه کانولوشن گراف [102](GCN)را برای ادغام روابط معنایی و فضایی بین اشیا پیشنهاد کردند(مجموعه‌داده COCO). در زمینه تفسیر تصاویر پزشکی نیز این رویکرد به کار گرفته شده است.

کریستی و همکاران[70] مراحل کار را به یک فرآیند کدگذاری دانش محور، بازیابی و تفسیر تقسیم کردند. به این صورت که ابتدا یک ماژول کدگذار ویژگی‌های بصری را به گراف ناهنجاری تبدیل می‌کند؛ در این گراف هر گره نشان‌دهنده یک ناهنجاری بالینی احتمالی است؛ گراف توسط پزشک طراحی شده و ویژگی‌های معنایی ناهنجاری‌ها را نشان می‌دهد. همبستگی گره‌های ناهنجاری به عنوان وزن لبه ناهنجاری کدگذاری می‌شود تا روابط بین یافته‌های غیرطبیعی مختلف هنگام تصمیم‌گیری تشخیص بالینی در نظر گرفته شود. سپس دنباله‌ای از الگوها را با توجه به ناهنجاری‌های شناسایی شده از طریق یک ماژول، بازیابی می‌کند. کلمات الگوهای شناسایی شده بیشتر بسط می‌یابند و توسط یک ماژول Paraphrase به یک گزارش تبدیل می‌شوند.

ایشاو و همکاران[90]نیز از گراف احتمالاتی که بر اساس دانش پزشک آماده شده بود، استفاده کردند. ماژول کدگذار ویژگی‌های بصری را به گراف ناهنجاری تبدیل می‌کند و خروجی این مرحله به عنوان ورودی به کدگشای LSTMسلسله مراتبی وارد می‌شود که وظیفه تولید جملات گزارش را عهده‌دار است.



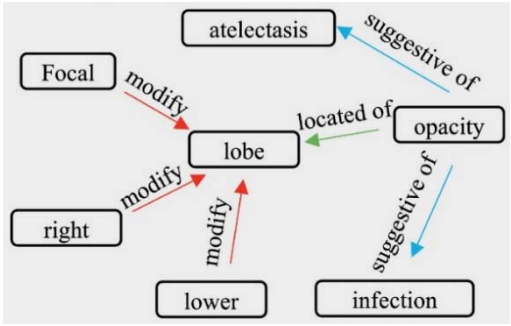
شکل (۱۱): کدگذار مبتنی بر گراف [103]

یانگ و همکاران[99] معتقد بودندکه در حوزه پزشکی، مدل‌های کدگذار–کدگشا با دو مشکل روبرو هستند:

- سوگیری بصری و متنی: انتظار می رود که مدل در هنگام تهیه گزارش رادیولوژی به ناهنجاری ها توجه بیشتری داشته باشد. اما، در بیشتر موارد، مناطق غیر طبیعی تنها بخش کوچکی از تصویر رادیولوژی را اشغال می کنند. و توصیف‌های مربوط به این مناطق تنها بخش کوتاهی را در گزارش نهایی به خود اختصاص می‌دهند. در نتیجه، رویکردهای کدگذار–کدگشا مبتنی بر داده خالص می‌توانند به سمت توصیف‌های معمولی متمایل شوند و در کشف ناهنجاری‌ها شکست بخورند.
- فقدان دانش تخصصی: رویکردهای مبتنی بر کدگذار–کدگشای خالص نمی‌توانند دانش تخصصی را دقیقاً برای تولید گزارش‌ها ترکیب کنند، که باعث می‌شود که به عنوان یک روش بالینی مورد استفاده قرار نگیرد.

آنها برای ادغام بهتر دانش موجود در تولید گزارش رادیولوژی قفسه سینه،دانش پزشکی را بر اساس [104]RadGraph[1]به دانش عمومی و دانش خاص دسته‌بندی کردند. دانش عمومی به عنوان دانش مستقل از تصویر،با استفاده از یک پایگاه دانش استاندارد تعریف می‌شود و دانش خاص به عنوان دانش وابسته به تصویر درنظر گرفته‌می‌شود که به تصویر ورودی فعلی مربوط می‌شود. برای هر تصویر رادیولوژی، تصاویر مشابه بازیابی شده و مجموعه‌ای از دانش سفارشی از گزارش‌های آنها جمع‌آوری می‌شود. آنها برای ادغام دانش عمومی، دانش خاص و ویژگی‌های بصری تصویر رادیولوژی، یک مکانیسم توجه چندسر پیشنهاد دادند. آنها نمودار دانش را به صورت یک شبکه معنایی نشان دادند که رابطه بین موجودیت‌ها را مشخص می‌کند. در گراف حاصل گره‌ها معمولاً موجودیت‌های مختلف و یال‌ها معمولاً روابط متفاوت را در نمودار دانش نشان می‌دهد(شکل ۱۲). آنها با ادغام ویژگی‌های بصری تصویر با دانش عمومی و دانش خاص، با کمک توجه چندسر کیفیت گزارش‌های تولید شده را بهبود بخشیدند.

[1] RadGraph یک نمودار دانش در مقیاس بزرگ شامل ۶ میلیون موجودیت است، که از مجموعه‌داده های MIMIC-CXR استخراج شده است.



شکل (۱۲): بخشی از گرافشبکه معنایی دانش عمومی[104]

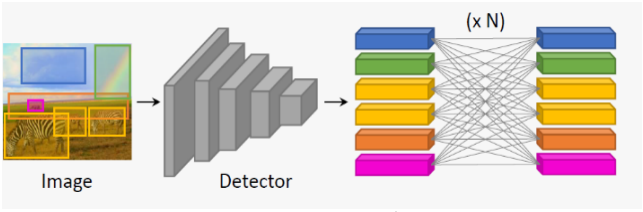
- **کدگذار مبتنی بر مکانیزم توجه به خود**

توجه به خود یک مکانیسم توجه است که در آن هر عنصر از یک مجموعه با همه عناصر دیگر همان مجموعه مرتبط است، این ارتباط برای محاسبه یک نمایش دقیقو تاثیرگذاری مجموعهعناصر بر یکدیگر از طریق اتصالات باقیمانده اتخاذ می‌شود(شکل ۱۳) و اولین بار توسط واسوانی و همکاران[105] معرفی شد. توجه به خود در ترجمه ماشینی، درک زبان، ایجاد معماری ترنسفورمر و انواع آن، تاثیر بسیار داشته است.در میان اولین مدل‌های تفسیر تصویر که از این رویکرد استفاده می‌کنند، یانگ و همکاران[106] از یک ماژول خود توجه برای کدگذاری روابط بین ویژگی‌هایی که از یک آشکارساز شی بدست می‌آمد، استفاده کردند.

در زمینه تفسیر تصاویر پزشکی زی‌هونگ‌چنگ و همکاران [98]، با استفاده از شبکه های عصبی کانولوشنال از پیش آموزش دیده (VGG, ResNet) ویژگی‌های بصری تصاویر رادیولوژی قفسه سینه را استخراج کردند؛ نتایج کدگذاری شده به عنوان دنباله اصلی برای همه ماژول‌های بعدی استفاده شد. آنها در مدل خود، از کدگذار استاندارد ترنسفورمر استفاده کردند(شکل۱۵).

به غیر از اعمال عملگر توجه به خود بر روی ویژگی‌های استخراج شده توسط CNN، عملگر توجه بطور مستقیم بر روی تکه‌های تصویر هم اعمال می‌شود که در این راستا دو رویکرد وجود دارد.

رویکرد اول با نام ترنسفورمر بینایی[1] شناخته می شود(به اختصار ViT) و توسط دوسوویتسکی و همکاران پیشنهاد شد[107]. از آنجا که ترنسفورمرها در حوزه پردازش زبانهای طبیعی خوب عمل کرده‌اند، نویسندگان مقاله ایده به کارگیری مکانیزمی شبیه به آن را در حوزه پردازش تصویر مطرح کردند. در روش پیشنهادی آنها، ابتدا، تصاویر به تکه‌های مجزایی که همپوشانی ندارند تقسیم می‌شوند (مثلاً ۱۶×۱۶). سپس، این تکه‌ها به بردار تبدیل شده[2] و با تعبیه مکانی[3] الحاق می‌شود. پارامترهای آموزشی آنها به ترنسفورمر منتقل می‌شودو وظیفه تشخیص کلاس توسط یک طبقه‌بندی کننده چندکلاسه انجام می‌گردد (شکل۱۴).



شکل (۱۳): کدگذار مبتنی بر توجه به خود[103]

رویکرد دوم با عنوان شبکه‌های ترنسفورمر بینایی کانولوشنال (به اختصارCVT) شناخته می‌شود، این روش توسط وو و همکاران [108] ارائه شد. آنها معتقد بودند که ViT فاقد ویژگی‌های مطلوب خاصی است که ذاتاً در معماری CNN تعبیه شده‌است و باعث عملکرد منحصربه‌فردCNNها برای حل وظایف بینایی می‌شوند.از آنجا که تصاویر دارای ساختار محلی دو بعدی قوی هستند و پیکسل های همسایه معمولاً همبستگی بالایی دارند، معماری CNN، این موضوع را پوشش می‌دهد و به تبع آن باعث می‌شود مدلدر برابر تغییر مقیاس، چرخش و ... پایدارتر باشد. علاوه بر این، ساختار سلسله مراتبی هسته‌های کانولوشنال، الگوهای بصری را به‌خوبی یادمی‌گیرد بطوری که بافت فضایی محلی را در سطوح مختلف پیچیدگی، از لبه‌ها و بافت‌های ساده سطح پایین گرفته تا الگوهای معنایی مرتبه بالاتر،در نظر می‌گیرد. از اینرو آنها مدلی پیشنهاد کردند که هم از مزایای ترنسفورمرها و هم شبکه‌های کانولوشن بهره‌برداری نمایند. مدلCVT، ترنسفورمرهای بینایی معمولی (ViT) هستند، با این تفاوت که دیگر تصاویر به بردار تبدیل نمی‌شود پس طبیعتاً دیگر نیازی به تعبیه مکانی نیست؛ علاوه بر آن آموزش با استفاده از کانولوشن انجام داده می‌شود (شکل۱۶). جدول (۷) دو روش CVTوViT را با هم مقایسه می‌نماید.این دو روش در اکثر وظایف بینایی ازجمله در زمینه تفسیر تصویر نیز به کار گرفته می‌شود.

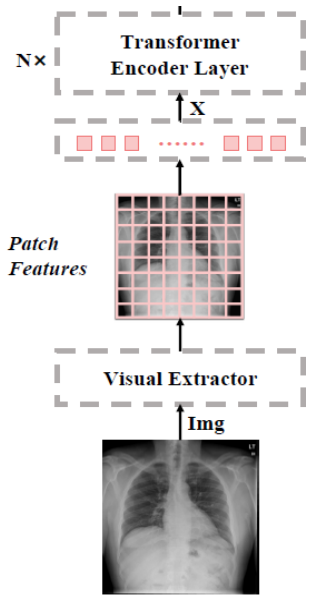
آرون نیکلسون و همکاران [13]در یک پژوهش جامع روش‌های مختلف ممکن برای کدگذار و کدگشا را بررسی کردند و نتایج آنرا در دسترس عموم قرار دادند[4]. تحقیقات آنها بیانگر این‌است که در صورت استفاده ازشبکه‌های ترنسفورمر بینایی کانولوشنال به عنوان کدگذار نتایج حاصله بهتر خواهد بود.

[1]Vision Transformer
[2]Flatten
[3]Position Embedding

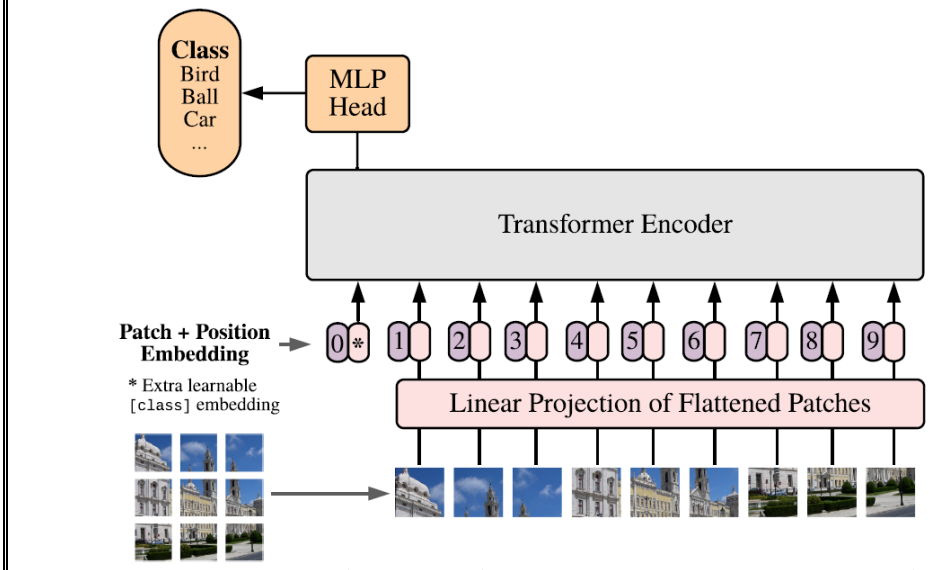
[4]https://github.com/aehrc/cvt2distilgpt2

جدول (۷):مقایسه ViT و CVT

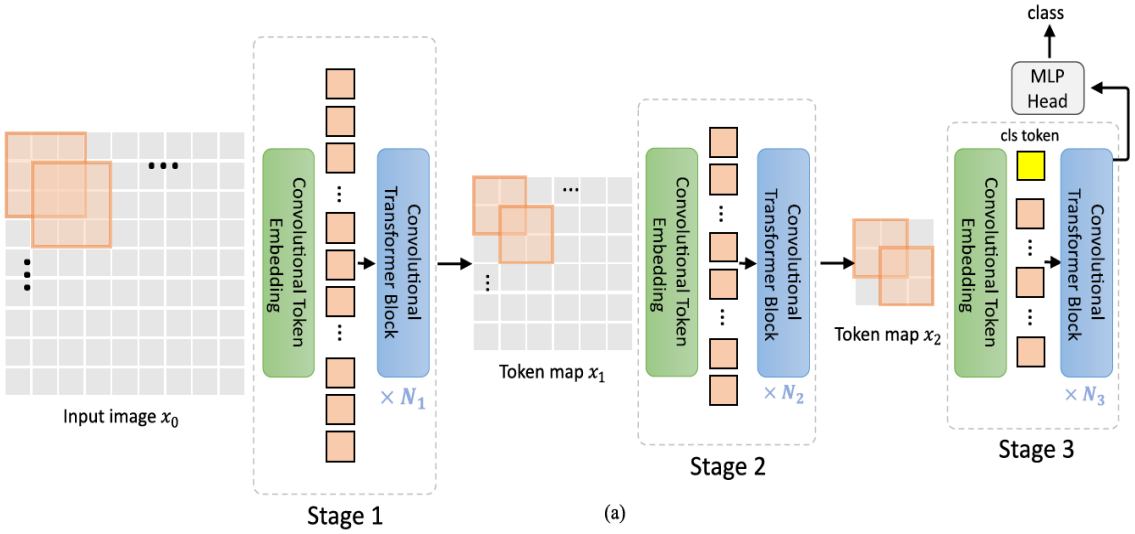
Method	Needs Position Encoding (PE)	Token Embedding	Projection for Attention	Hierarchical Transformers
ViT	yes	non-overlapping	linear	no
CVT	no	overlapping (convolution)	convolution	yes



شکل (۱۵): کدگذار بر پایه توجه به خود [98]



شکل (۱۴): ترنسفورمر بینایی، ابتدا، تصاویر به تکه‌های مجزای که همپوشانی ندارند تقسیم می‌شوند (مثلاً ۱۶×۱۶). سپس، این تکه‌ها به بردار تبدیل می‌شوند و در ادامه با تعبیه مکانی الحاق می‌شود. پارامترهای آموزشی آنها به ترنسفورمر منتقل می‌شود. وظیفه تشخیص کلاس توسط یک طبقه‌بندی کننده چندکلاسه انجام می‌گردد[107].



شکل (۱۶): شبکه ترنسفورمربینایی کانولوشنال[108]

جدول (۸): روش‌های عمده‌ای که کدگشا بر پایه آن بنا می شود.

Methods		Papers
RNN-Based	LSTM	[97][59][86][74][109][108][107][95][89][35][80][29][91][58]
	BiLSTM	[34]
	Hierarchical LSTM	[88][34][32]
	Hierarchical LSTM sentence & word	[57][33][90][14][12]
	Hierarchical BiLSTM	[56]
Transformer-Based	Multi-Layer Transformer	[14][98][13]
	Transformer-based Architectures	[15][89][13][52][99]
	BERT- like	

۳-۳-۴ تولید گزارش

در مرحله تولید گزارش، ویژگی‌های استخراج شده از کدگذار برای تولید تفسیر تصویر استفاده می‌شود. روش‌های عمده‌ای که برای پیاده‌سازی کدگشا به کار می‌رود(جدول۸) عبارتند از:

- کدگشا مبتنی بر شبکه های عصبی بازگشتی
- کدگشا مبتنی بر ترنسفورمر

در ادامه به بررسی آنها پرداخته می‌شود.

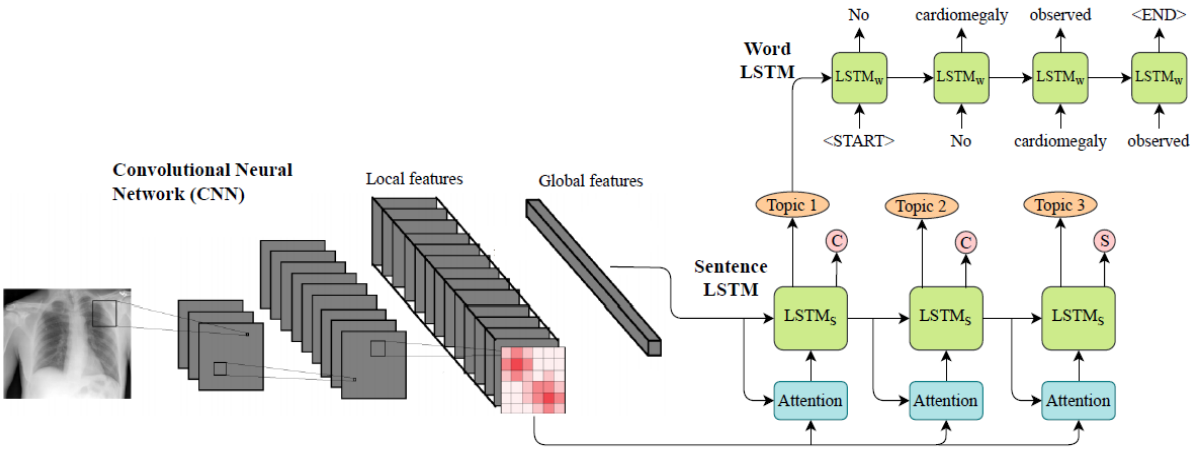
- **کدگشا مبتنی بر شبکه های عصبی بازگشتی**

رایج‌ترین روش پیاده‌سازی کدگشا استفاده از انواع شبکه‌های عصبی بازگشتی[1] می‌باشد. یک شبکه عصبی بازگشتی بسیار شبیه به یک شبکه عصبی معمولی به نظر می‌رسد، با این تفاوت که دارای اتصالات بازگشتی به نورون‌هاست. ساده‌ترین RNN، متشکل از یک نورون است که ورودی‌ها را دریافت و خروجی تولید می‌کند و آن خروجی را در مرحله زمانی بعد به همراه ورودی برای آموزش به خودش برمی‌گرداند، از آنجایی که خروجی یک نورون بازگشتی در مرحله زمانی t، تابعی از تمام ورودی‌های مراحل زمانی قبلی است، می‌توان گفت شبکه نوعی حافظه دارد.

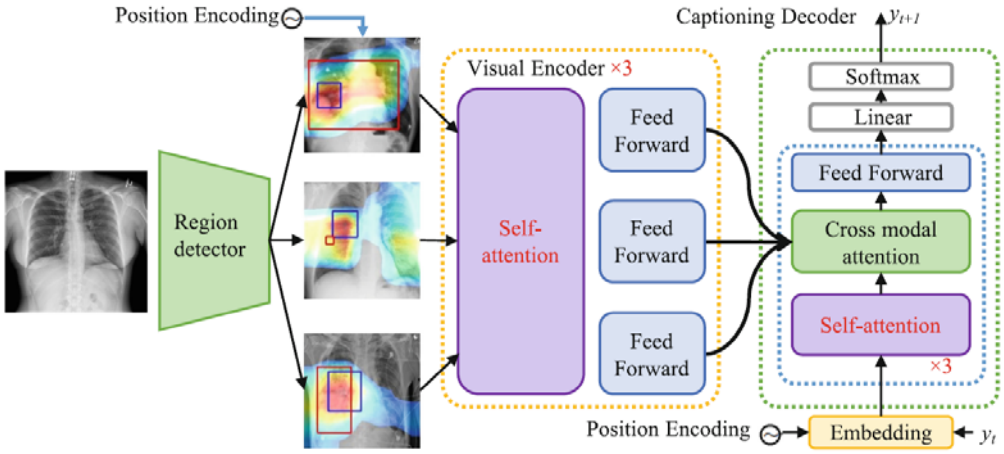
برای آموزش یک RNNبر روی دنباله‌های طولانی، باید RNNبه اندازه طول دنباله اجرا شود. و عملاً RNN باز شده به یک شبکه عمیق تبدیل می‌شود که عمق آن به طول دنباله وابسته است. در اینجا نیز درست مانند شبکه‌های عصبی عمیق، با افزایش عمق شبکه ممکن است مشکل گرادیان‌های ناپایدار رخ دهد و علاوه بر این، هنگامی‌که یک RNN یک دنباله طولانی را پردازش می‌کند، به دلیل دگرگونی‌هایی که داده‌ها هنگام عبور از آنطی می‌کنند، در هر مرحله زمانی مقداری از اطلاعات از بین می‌رود. پس از مدتی، وضعیت جاریRNN عملاً هیچ اثری از اولین ورودی ها ندارد. اصطلاحاً گفته می‌شود با طولانی شدن طول دنباله، شبکه ورودی‌های اولیه را فراموش می‌کند. برای مقابله با این مشکل، انواع مختلفی از سلول‌ها با حافظه بلند مدت معرفی شده‌اند. سلول LSTM از این نوع سلولها می‌باشد. سلول LSTM نسبت به سلول اصلیRNN عملکرد بسیار بهتری دارد، آموزش سریعتر همگرا می‌شود و وابستگی های طولانی مدت در داده‌ها را تشخیص می‌دهد؛ اما پارامترهایی که در این شبکه باید آموزش ببینند بیشتر از RNN است. ایده کلیدی در LSTMاین است که شبکه بتواند یاد بگیرد که چه چیزی را در بلند مدت ذخیره کند، چه چیزی را دور بریزد و چه چیزی را از آن بخواند. در حالیکه LSTM می‌تواند توالی‌هایطولانی‌تری نسبت به RNNهای ساده را مدیریت کند، اما حافظه کوتاه مدت نسبتاً محدودی دارد و برای یادگیری الگوهای طولانی‌مدت، به مشکل برمی‌خورد[110]. از آنجا تفسیرهای موجود برای تفسیر تصاویر پزشکی با طول‌های متغیر و در مواردی طولانی می‌باشند، محققان برای حل مشکل از روش‌های متعددی بهره برده‌اند که در ادامه معرفی می‌شود.

استفاده از LSTM جمله و LSTM کلمه می‌تواند راهکاری برای مقابله با جملات طولانی و البته بدون ساختار باشد. LSTMجمله دنباله‌ای از بردارهای موضوعی را تولید می‌کند، وظیفه LSTM کلمه تولید کلمه به کلمه جملات است. و مکانیزم توجه در جهت وصول نتایج بهتر می‌تواند در هر دو سطح جمله و کلمه و یا هر دو استفاده شود[25][12](شکل ۱۷). همچنین از آنجا که معمولاً وابستگی هر لغت نه تنها به لغت قبلی آن، بلکه به لغت بعدی آن نیز می‌باشد؛ استفاده از LSTM های دوطرفه در تولید بهتر گزارش موثر است[56] و البته استفاده از مکانیزم توجه در حصول نتیجه بهتر تاثیرگذار است. روش‌های مختلف ترکیب موارد فوق در جدول ۸ ذکر شده است.

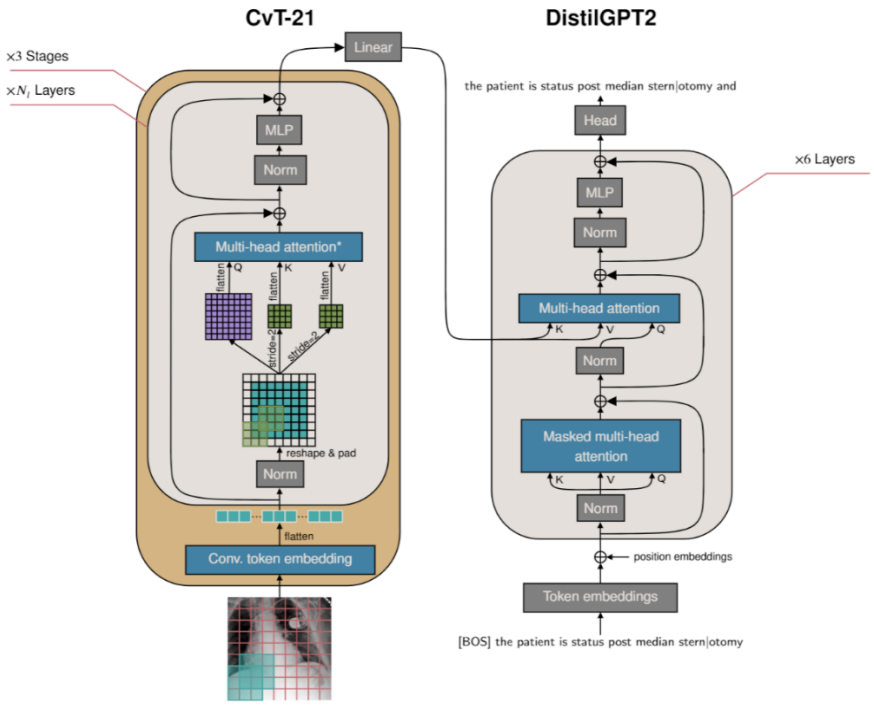
[1]Recurent Neural ntwork



شکل (۱۷): معماری کدگذار- کدگشا با LSTM سلسله مراتبی با رویکرد توجه در سطح جمله[25]



شکل (۱۸): کاربرد ترنسفورمر در کدگشا[111]



شکل (۱۹): مدلCvT-212DistilGPT2[13]

جدول (۹): مقایسه جدیدترین روش‌های شرح تصاویر پزشکی بر روی مجموعه‌داده‌ها با استفاده از معیارهای ارزیابی. خط تیره (-) در صورتی استفاده می شود که محققان از آن معیار ارزیابی استفاده نکرده باشند.(B=BLEU, R=ROUGE, M=METERO, C=CIDEr)

Method	Encoder	Decoder	Datasets	References	B_1	B2	B3	B4	R	M	C	Mean-B
Encoder-Decoder	CNN	RNN Attention	RDIF	[112]	0.49	0.35	0.28	0.23	0.39	0.31	-	-
	CNN	LSTM Attention	MICCAI 2017 LiTS Chalenge	[42]	0.883	0.837	0.798	0.7506	0.892	-	-	-
	CNN	Hierarchical LSTM + Attention	IU Chest Xray	[32]	0.373	0.246	0.175	0.118	0.315	0.163	0.359	-
	CNN	LSTM + Attention	BCIDR	[96]	91.2	82.9	75.0	67.7	70.1	39.6	2.04	-
	CNN	LSTM	ICLEFCaption 2017	[91]	-	-	-	-	-	-	-	0.0749
	CNN	LSTM	Chest XRay14	[59]	0.286	0.159	0.103	0.073	0.226	0.106	-	-
	CNN	LSTM	DIARETDB0 DIARETDB1	[84]	-	-	-	-	-	-	-	-
	CNN	Hierarchical LSTM	ICLEFCaption 2017	[88]	-	-	-	-	-	-	-	0.0982
	CNN	Transformer-based Architectures (GPT2)	IU X-RAY	[15]	0.387	0.245	0.166	0.111	0.289	0.164	0.257	-
	CNN	Transformer (BERT)	Chest X ray	[87]	0	-	-	-	-	0.763	0.773	-
		LSTM			-	-	-	-	-	0.716	0.707	-
		GRU			-	-	-	-	-	0.682	0.669	-
	CNN + Attention	Two layer LSTM	XRAYS FRONTAL PELVIC	[34]	91.9	83.8	76.1	67.7	-	-	-	77.97
	CNN +Co Attention	Scentenc-Word LSTM	IU Chest Xray PEIR Gross	[12]	0.517 0.300	0.386 0.218	0.306 0.165	0.247 0.113	0.447 0.279	0.217 0.149	0.327 0.329	- -
	CNN + Attention	LSTM Attention	CheXpert	[33]	0.529	0.372	0.315	0.255	0.453	0.343	-	-
	CNN + Attention	LSTM	ICLEFcaption 2018	[29]	-	-	-	-	-	-	-	0.1800
	CNN + Attention	Hierarchical BiLSTM	IU Chest XRay	[56]	0.466	0.358	0.270	0.195	0.366	0.274	-	-
	CNN + Attention	LSTM	ICLEFcaption 2019	[58]	-	-	-	-	-	-	-	0.2316
	CNN + Attention	Hierarchical LSTM sentence & word	IU X-RAY	[14]	0.298	0.187	0.126	0.088	0.273	0.165	0.186	-
	CNN + Attention	Multi-Layer Transformer			0.373	0.226	0.147	0.101	0.293	0.182	0.319	-
	Transformer Self Attention	Transformer-based Architectures	Retina ImBank	[83]	0,638	0.561	0.508	0.456	0.676	0.332	-	-
			Retina Chinese		0.371	0.181	0.181	0.143	0.336	0.168	-	-
			IU X-RAY		0.479	0.213	0.213	0.155	0.363	0.95	-	-
			MIMIC-CXR RI		0.362	0.227	0.155	0.113	0.283	0.142	-	-
	CVT	DistilGPT2	IU Chest XRay	[13]	0.477	0.308	0.227	0.177	0.377	0.203	0.681	-
			MIMIC-CXR		0.394	0.249	0.172	0.127	0.155	0.287	0.379	-
	CNN + Self Attention	Multi-Layer Transformer	IU X-RAY	[98]	0.470	0.304	0.219	0.165	0.371	0.187	-	-
			MIMIC-CXR		0.353	0.218	0.145	0.103	0.277	0.142	-	-
Others	CNN + Attention	GPT2 +Image Retrieval	ICLEFcaption 2021	[52]	-	-	-	0.553	-	-	-	-
	CNN	Stack RNN+ Retrival	U Chest Xray	[38]	0.438 0.673	0.298 0.597	0.208 0.530	0.151 0.486	0.322 0.612	-	0.343 2.895	-

	Template	CX-CHR									
CNN + Attention+ Graph Convolution	Hierarchical LSTM sentence & word+ att	IU Chest XRay	[90]	0.441	0.291	0.203	0.147	0.367	-	0.304	-
CNN	LSTM + Image Retrieval	ICLEFcaption 2017	[74]	0.134	0.061	0.026	0.012	0.113	0.43	0.053	-
CNN	Graph-base + Templeate-base +interactive	IU Chest XRay MGH(EEG) TUH(EEG)	[71]	0.489 0.762 0.764	0.356 0.684 0.659	0.225 0.614 0.624	0.225 0.464 0.483	- - -	- - -	0.374 0.443 0.425	- - -
Statistical approach		ICLEFcaption 2021	[87]	-	-	-	-	-	-	-	0.378
CNN	Multi-label Classification + Statistical Rule			-	-	-	-	-	-	-	0.510
CNN	Similarity-based (Image Retrieval)			-	-	-	-	-	-	-	0.442
CNN + Multi Head Att and Knowledge Retrieval	Knowledge Graph + Multi-Head Attention	IU X-RAY	[99]	0.496	0.327	0.238	0.178	0.381	-	0.382	-
		MIMIC-CXR		0.363	0.228	0.156	0.115	0.284	-	0.203	-
CNN + Attention	Scentenc-Word LSTM MetaData	IU Chest XRay	[57]	0.476	0.340	0.238	-	0.347	-	0.297	-
CNN	Image Retrieval	ICLEFcaption 2018	[73]	-	-	-	-	-	-	-	0.2501
Graph Transformer	Hybrid retrieval and paraphrasing	IU X-Ray CX-CHR	[70]	0.482 0.673	0.325 0.588	0.226 0.532	0.162 0.473	0.322 0.618	- -	o.280 2.850	-

- **کدگشا مبتنی بر ترنسفورمر**

محققان گوگل در سال ۲۰۱۷[105]، موفق به ایجاد معماری به نام ترنسفورمر شدند که به طور قابل توجهی ترجمه ماشینی عصبی را بدون استفاده از هیچ گونه لایه بازگشتی یا کانولوشن، فقط با مکانیسم‌های توجه، بهبود بخشید. آموزش این معماری بسیار سریع‌تر و موازی‌سازی آن آسان‌تر بود. اندکی پس از آن، مدل ترنسفورمر به بلوک ساختمانی برای پیشرفت‌های دیگر در پردازش زبان طبیعی، مانند BERT و GPT و معماری‌های مشابه، برای بسیاری از وظایف درک زبان تبدیل شد.

تفسیر تصاویر نیز این معماری را به طرق گوناگون به خدمت گرفته است.استفاده از بلوک‌های متعدد ترنسفورمر جایگزین مناسبی برای شبکه های عصبی بازگشتی است که مشکلات آنها نظیر فراموشی و ناپدید شدن گرادیان را ندارد. این ایده توسط تعدادی از محققان به‌کار گرفته شده است[13][98][14].

BERT به عنوان یک مدل زبانی قوی[113]و مدل‌های زبانی مبتنی بر ترنسفورمر در حل اکثر مسائل زبان طبیعی موفق عمل کرده است. این مدل‌ها و گاهی مدل‌های تقطیر شده آنها به عنوان کدگشا در تفسیر تصاویر به‌کار گرفته شده است [13][89][15]. اما مدل BERT بر اساس توالی‌های متنی آموزش دیده‌است؛ از آنجا که در مسائل شرح تصویر استفاده همزمان از تعبیه بصری و معنایی منجر، به نتایج بهتر می‌گردد از اینرو استفاده از مدل VL-BERT[114][1] برای شرح تصاویر پیشنهاد شده است. این مدل‌ها چون همزمان از اطلاعات بصری تصویر نیز استفاده می کنند معمولا به نتایج قابل قبول‌تری دست می‌یابند. مکانیزم این مدل‌ها به این طریق است که فقط از متن ماسک شده استفاده نمی‌کند بلکه همزمان علاوه بر گزارش‌ها که بصورت ماسک شده به مدل تزریق می شود تصویر نیز بصورت ماسک شده به عنوان ورودی برای آموزش در نظر گرفته می شود.

اگرچه در زمینه تفسیر تصاویر طبیعی این روش به‌کار برده شده است اما پژوهشی که در زمینه شرح تصاویر پزشکی مدل شبیه به BERT را به‌کار گرفته باشد دیده نشد.

یکی از اولین تلاش ها برای ادغام ترنسفورمر در تولید گزارش بالینی توسط شیونگ و همکاران انجام شده است[111]. آنها ترنسفورمر تقویتی را برای تفسیر تصویر پزشکی (RTMIC) پیشنهاد کردند و از یک DenseNet از قبل آموزش دیده برای شناسایی منطقه مورد نظر در تصویر ورودی، و به دنبال آن یک کدگذار مبتنی بر ترنسفورمر برای استخراج ویژگی های بصری استفاده کردند. این ویژگی ها به عنوان ورودی به کدگشا برای تولید جملات داده می‌شود. مجموعه‌داده مورد استفاده IU-XRay بود (شکل ۱۸).

[1]Visual-Linguistic BERT

نیکلسون و همکاران[13]در یک تحقیق جامع انواع روش‌های کدگذار و کدگشا را بررسی نمودند (جدول۹). کدگذار و کدگشاهای منتخب آنها در جدول ۱۰ ذکر شده‌است. نتایجدر صورت استفاده از CVT به عنوان کدگذار و DistilGPT2به عنوان کدگشا (شکل ۱۹) بهتر از سایر مدل‌های ترکیبی بوده است.

جدول (۱۰):کدگذار و کدگشاهای بررسی شده در مقاله نیکسون و همکاران[13]

کدگشا	کدگذار
Transformer	ResNet
GPT2	DendeNet
BERTBase	CheXNet
Distil BERTBase	EfficientNet
DistilGPT2	ViTBase
BioBERTBase	CVT
Sci BERTBase	DietBsae
Sci BERTBase	XCit
Clinical BERTBase	BEiTBase
Blue BERTBase	
PubMed BERTBase	

۵ بحث

با توجه به نتایج استخراج شده از مقالات مورد بررسی (جدول۹) علی‌رغم پیشرفت‌های صورت گرفته در تولید تفسیرتصاویر، کاربرد آن در حوزه پزشکی، همچنان چالش برانگیز است؛ که این به دلیل ماهیت تصاویر و گزارش‌های پزشکی است که با تصاویر طبیعی و تفسیرهای عمومی متفاوت است؛ در واقع، شرح تصاویر عمومی شامل توصیف اشیا و روابط بین آنها با استفاده از یک یا چند جمله است.در حالی که، تفسیر تصویر پزشکی شامل درک یافته‌های بالینی و ارائه یک گزارش دقیق متشکل از پاراگراف‌های مختلف است تا فقط آنچه از نظر بالینی مهم است به جای آنچه در تصویر از نظر اشیاء وجود دارد برجسته گردد.

هدف از این پژوهش بررسی مساله تفسیر تصاویر پزشکی با تاکید بر روش‌های مبتنی بر یادگیری عمیق است. در راستای درک بهتر موضوع و ادبیات موجود در این حوزه، سه موضوع عمده‌ی: مجموعه‌داده‌ها، معیارهای‌ارزیابی و روش‌های تفسیر تصاویر پزشکی مورد بررسی قرار گرفت. به عنوان زیرشاخه‌های اصلیروش‌های تفسیر تصاویر پزشکی: تفسیر مبتنی بر الگو، تفسیر مبتنی بر بازیابی و تفسیر مبتنی بر یادگیری عمیق (کدگذار-کدگشا) مورد تحلیل قرار گرفت. تاکید اصلی برروی معماری کدگذار-کدگشا گذاشته شده است.

مجموعه داده‌ی مناسب، یکی از عوامل تاثیرگذار در کارایی مدل‌های مبتنی بر یادگیری عمیق است که این امر در مساله تفسیر تصاویر پزشکی نیز صادق است. در مورد ارائه یک تفسیر قابل اعتماد ممکن است به ده‌ها میلیون نمونه تصویر/متن نیاز باشد که هنوز به آسانی در دسترس نیست[115]. علاوه بر تعداد داده‌ها، باید نمونه‌ها بدون اطلاعات پراکنده و عاری از نویز باشد تا فرایند یادگیری در روش‌های یادگیری عمیق تسهیل شود؛ به عبارت دیگر کمیت و کیفیت مجموعه‌داده بر نتایج تاثیر مستقیم دارد. به عنوان مثال تعداد نمونه در مجموعه داده MIMIC-CXR برابر ۳۷۷۱۱۰ و در مجموعه داده IU Chest X-Rayبرابر ۷۴۷۰ می‌باشد. هر دو مجموعه‌داده شامل رادیوگرافی‌های قفسه سینه و با ساختار مشابه می باشند؛ اما محققانی که همزمان مدل پیشنهادی‌شان را بر روی هر دو مجموعه‌داده اعمال کردند [83],[13], [98], [99]علی‌رغم بزرگتر بودن مجموعه‌داده MIMIC-CXR نتایج حاصله بر روی IU Chest X-Rayرا به مراتب بهتر گزارش کرده‌اند (جدول۹) که این نشان دهنده کیفیت بالاتر این مجموعه‌داده می‌باشد.

چالش دیگر در مجموعه‌داده‌ها، سوگیری داده‌ها است. این بدین معنی است که، هنگام در نظر گرفتن کل جمعیت، موارد بیمار بسیار نادرتر از موارد سالم هستند، به عنوان مثال در مجموعه داده IU Chest x-ray موارد نرمال ۳۷% (۲۶۹۶ تصویر) از کل مجموعه داده (۷۲۸۴ تصویر) را تشکیل می دهند، که در مقایسه، "کدری" که شایع ترینناختلال غیرطبیعی در تصاویر رادیولوژی قفسه سینه می‌باشد ۱۲٪ (۸۴۰ تصویر) را در برمی‌گیرد و یا "کاردیومگالی" که دومین عارضه غیرطبیعی در این نوع تصاویر است، تنها ۹٪ (۶۵۵ تصویر) از کل مجموعه داده را تشکیل می‌دهد[116].

برتری یک مدل به مدل‌های دیگر تا حدودی به کیفیت زیرنویس‌های تولید شده بستگی دارد، و اگر معیارهای ارزیابیِ تعیین کننده این کیفیت، قابل اعتماد نباشد به نتیجه حاصل از آن نمی‌توان تکیه کرد. متاسفانه معیارهای ارزیابی مورد استفاده برای سنجشِ کیفیت تفسیر خودکار، چالش برانگیز است. اکثر این معیارها شباهت زبانی را بین دو توالی متنی، اندازه می‌گیرند که بر پایه ترتیب رخداد کلمات می‌باشد، و برای ارزیابی سایر وظایف پردازش زبان طبیعی، نظیر: ترجمه‌ماشینی، تشابه متون و ... پیشنهاد شده‌اند و نمی‌توانند دقت و کیفیت کلی گزارش تولید شده را نشان دهند[23]. از این رو نیاز به توسعه معیارهای ارزیابی مناسب برای ارزیابی تفسیر تصاویر پزشکی خودکار احساس می‌شود؛ در کنار آن استفاده از الگوهای استاندارد برای تولید گزارش می‌تواند مفید باشد. همچنین مشارکت دادن پزشک در روند تولید گزارش می‌تواند کارگشا باشد[71]، زیرا به پزشک اجازه می‌دهد گزارش تولید شده به‌طور خودکار را ببیند و تصحیح و یا تأیید کند. در کنار آن به کارگیری راه‌حل‌های مبتنی بر هوش مصنوعی قابل توضیح می تواند، تقریب‌ها و تجسم مدل‌های یادگیری عمیق را برجسته کند تا درک نتایج را آسان‌تر نماید و به این ترتیب اجماع بین پزشکان افزایش یابد.

در تولید تفسیر خودکار برای تصاویر پزشکی انتظار می‌رود که یک مدل دو ویژگی کلیدی را مدیریت کند: (۱) روان بودن زبان برای خوانایی انسان (۲) دقت بالینی برای شناسایی صحیح بیماری همراه با علائم مرتبط[117].روش‌های مبتنی بر الگو، مورد اول را به خوبی پوشش می‌دهند و نگرانی از قسمت مدیریتِ مولفه‌ی تولید زبان وجود ندارد. اما در این روش نگرانی بیشتر در این زمینه است که ممکن است یافته‌هایی در تصویر موجود باشد که در قالب

الگوهای پیش فرض نگنجد. از این رو در جهت بهبود عملکرد مدل استفاده از این روش به تنهایی توصیه نمی‌گردد و معمولاً روش‌ها ترکیبی استفاده و نتایج بهبود پیدا می کند[68] (جدول ۹). یک اصل ساده وجود دارد که اگر دو تصویر شبیه به هم است احتمالاً تفسیر دو تصویر نیز با هم شباهت خواهد داشت. مدل‌های مبتنی بر بازیابی با تکیه بر این اصل کار می کنند و نتایج خوبی ارائه می کنند. بطور مثال در مسابقات ICLEFcaption معمولاً مدل‌هایی که از این رویکرد استفاده کرده بودند به نتایج بهتری دست یافتند اما هیچ کدام از آنها مدل پیشنهادی خود را تنها با تکیه بر روش‌های مبتنی بر بازیابی بنا نکردند و روش‌های بازیابی را با سایر روش‌ها ترکیب کردند[87][52] (جدول۹).

علی‌رغم پیشرفت هایی که در تولید خودکار گزارش‌های بالینی از تصاویر پزشکی با استفاده از یادگیری عمیق صورت گرفته است با این حال، تولید تفسیر از داده‌های تصویربرداری پزشکی به دلیل تنوع در گزارش‌های رادیولوژیست‌های مختلف، طول توالی طولانی (برخلاف زیرنویس‌های تصویر طبیعی)، و سوگیری مجموعه داده‌ها (داده‌های نرمال‌تر در مقایسه با غیر طبیعی) چالش برانگیز است. علاوه بر این، یک مدل خوب همانطور که ذکر شد علاوه بر انتظار روان بودن از لحاظ ادبیات و گرامر، باید دقت بالینی برای شناسایی صحیح بیماری همراه با علائم مرتبط را نیز داشته باشد. از آنجا که هم پردازش تصویر و هم تولید متن مد نظر است؛ شبکه‌های کدگذار-کدگشا با استفاده از CNN برای کدگذار و RNN بعنوان کدگشا بسیار محبوب می باشند. CNN وظیفه استخراج ویژگی‌های بصری و گاهی (در جهت بهبود عملکرد مدل) ویژگی‌های معنایی را بر عهده دارد و شبکه بازگشتی وظیفه تولید تفسیر را عهده‌دار است. از آنجا که مکانیزم توجه بصری، می تواند در لحظه‌ی تولید هر لغت تعیین کند که کدام بخش تصویر مهم‌تر و تاثیرگذارتر است؛ هم به تفسیر پذیری و هم به تولید گزارش بهتر کمک می کند و به کرات مورد استفاده قرار گرفته و در بهبود نتایج دخیل است؛ تا آنجا که مدل پیشنهادی جینگ و همکاران [12] در مقالات مورد بررسی بهترین نتیجه را بر روی مجموعه داده IU ChestX-ray داشته است (جدول۹). نویسندگان مقاله مکانیزم توجه را همزمان بر روی ویژگی‌ها بصری و معنایی اعمال کردند و بهبود در نتایج را ثبت نمودند.

شبکه های گراف در جهت دخیل کردن دانش رادیولوژی در تولید گزارش پیشنهاد شده است. از آنجا که بین علائم بیماری و تشخیص بیماری یک رابطه استنتاجی وجود دارد این ایده علاوه بر شفاف کردن تصمیم گیری بر روی نتایج نیز تاثیر گذار است.

با معرفی ترنسفورمرها و کاربرد موفق آنها در زمینه پردازش زبان طبیعی در چند سال اخیر به کارگیری آنها در حوزه تفسیر تصاویر پزشکی نیز دیده می‌شود و دامنه‌ی این استفاده متنوع است؛ گاهی فقط در یک مدل پیشنهادی پس از استخراج ویژگی‌ها توسط CNNاز توجه به خود استفاده می‌شود و گاهی شبکه CNNبه‌طور کامل حذف و جایگزین آن شبکه های CVT می شود. در این دامنه متنوع، از ایده‌های موجود مقایسه مدل‌ها به دلیل مجموعه‌داده‌های متفاوت امکان پذیر نیست ولی در یک پژوهش[13] که اکثر روش‌ها در شرایط یکسان بر روی مجموعه‌داده‌های خاص انجام شده است با انتخاب شبکه‌ی CVT به عنوان کدگذار نتایج بهبود پیدا کرده است.

در سمت کدگشا نیز دو ایده اصلی شبکه های بازگشتی و ترنسفورمرها وجود دارد. برای مدیریت طول بسیار متغیر و طولانی پاراگراف و این که هر جمله در پاراگراف بر موضوع خاص تکیه می‌کند، معمولاً شبکه بازگشتی به‌صورت سلسله‌مراتبی، دوطرفه و گاهی به همراه مکانیزم توجه استفاده می‌شود که بر بهبود نتایج تاثیرگذار است.

در راستای اعمال یادگیری انتقال در حوزه پردازش زبانهای طبیعی مدل های مبتنی بر ترنسفورمر نظیر BERT و GPTمورد توجه قرار گرفته‌اند. در تحقیقات اخیر این مدل‌ها در سمت کدگشا بکار گرفته شده و نتایج بهبود یافته است. بطور مثال در مدل پیشنهای [13] در صورت استفاده از CVT بعنوان کدگذار در صورتی که از Distil GPT به عنوان کدگشا استفاده گردد نتایج بهتری نسبت به بقیه ترکیب‌های ممکن کدگذار- کدگشا گزارش شده‌است.

۶ نتیجه گیری

با توجه به همه پیشرفت‌هایی که در حوزه تفسیر خودکار تصاویر پزشکی وجود داشتهاست اما رویکردهای موجود هنوز از محدودیت‌های خاصی رنج می‌برند؛ به نظر می‌رسد یک روند جدید از رویکردهای ترکیبی، مثلا ترکیب مدل‌های مولد و مبتنی بر بازیابی، امیدوارکننده باشد.

علاوه بر این، نیاز به توسعه مجموعه‌داده های تصاویر پزشکی همراه با گزارش از قسمت‌های مختلف بدن مانند مغز و سینه و ... با روش‌های گوناگون تصویر پزشکی احساس می‌شود.

از آنجا که معیارهای ارزیابی موجود دقیق نیستند، ارائه معیارهای ارزیابی مناسب برای ارزیابی تفسیرهای تولید شده مورد نیاز می‌باشد.افزایش تعامل انسانی نیز می‌تواند در مرحله ارزیابی دقت، با ترکیب ارزیابی دستی توسط پزشکان واجد شرایط برای گزارش‌های بهتر مفید باشد.مشارکت دادن پزشک در تولید گزارش به او اجازه می‌دهد گزارش تولید شده به‌طور خودکار را تصحیح یا تأیید کند.

به عنوان یک نتیجه گیری کلی، می توان گفت که تکنیک‌های توسعه‌یافته در زمینه نوشتن تفسیر تصاویر پزشکی هنوز با مشکلات متعددی روبرو هستند و هنوز نیاز به رسیدگی برای چالش‌های موجود احساس می‌شود. که این مهم، همکاری جامعه پزشکی و متخصصان علم داده را بطور همزمان طلب می کند.

مراجع

[1] X. Chen *et al.*, "Microsoft COCO Captions: Data Collection and Evaluation Server." 2015, [Online]. Available: http://arxiv.org/abs/1504.00325.

[2] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," in *IJCAI International Joint Conference on Artificial Intelligence*, 2015, vol. 2015-Janua, pp. 4188–4192.

[3] 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," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter. pp. 2641–2649, 2015, doi: 10.1109/ICCV.2015.303.

[4] H. Wang, Y. Zhang, and X. Yu, "An Overview of Image Caption Generation Methods," *Comput. Intell. Neurosci.*, vol. 2020, 2020, doi: 10.1155/2020/3062706.

[5] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 21–29, doi: 10.1109/CVPR.2016.10.

[6] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *33rd International Conference on Machine Learning, ICML 2016*, 2016, vol. 5, pp. 3574–3583.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems*, 2016, pp. 289–297.

[8] J. Lu, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via," *Openaccess.Thecvf.Com*۲۰۱۷ ., [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Knowing_When_to_CVPR_2017_paper.pdf.

[9] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 4651–4659, doi: 10.1109/CVPR.2016.503.

[10] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua. pp. 6298–6306, 2017, doi: 10.1109/CVPR.2017.667.

[11] H. Ayesha *et al.*, "Automatic medical image interpretation: State of the art and future directions," *Pattern Recognit.*, vol. 114, 2021, doi: 10.1016/j.patcog.2021.107856.

[12] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1. pp. 2577–2586, 2018, doi: 10.18653/v1/p18-1240.

[13] A. Nicolson, J. Dowling, and B. Koopman, "Improving Chest X-Ray Report Generation by Leveraging Warm-Starting," 2022, [Online]. Available: http://arxiv.org/abs/2201.09405.

[14] H. Park, K. Kim, S. Park, and J. Choi, "Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation," *IEEE Access*, vol. 9, pp. 150560–150568, 2021, doi: 10.1109/ACCESS.2021.3124564.

[15] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics Med. Unlocked*, vol. 24, p. 100557, Jan. 2021, doi: 10.1016/J.IMU.2021.100557.

[16] J. H. Huang, T. W. Wu, C. H. H. Yang, and M. Worring, "Deep Context-Encoding Network for Retinal Image Captioning," *Proc. - Int. Conf. Image Process. ICIP*, vol. 2021-Septe, pp. 3762–3766, 2021, doi: 10.1109/ICIP42928.2021.9506803.

[17] E. Pahwa, D. Mehta, S. Kapadia, D. Jain, and A. Luthra, "MedSkip: Medical Report Generation Using Skip Connections and Integrated Attention," *CVPR*, pp. 3402–3408, Nov. 2021, doi: 10.1109/ICCVW54120.2021.00380.

[18] D. Beddiar, M. Oussalah, and S. Tapio, "Explainability for Medical Image Captioning," *2022 11th Int. Conf. Image Process. Theory, Tools Appl. IPTA 2022*, 2022, doi: 10.1109/IPTA54936.2022.9784146.

[19] V. Kougia, "Medical Image Labeling and Report Generation," *Nes.Aueb.Gr*. [Online]. Available: https://nes.aueb.gr/ipl/nlp/theses/kougia_msc_thesis.pdf.

[20] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

[21] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proceedings of the workshop on text summarization branches out (WAS 2004)*, no. 1. pp.

25–26, 2004, [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, pp. 311–318, 2001, doi: 10.3115/1073083.1073135.

[23] J. T. Wu *et al.*, "AI Accelerated Human-in-the-loop Structuring of Radiology Reports," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2020, pp. 1305–1314, 2020.

[24] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," 2019, pp. 26–36, doi: 10.18653/v1/w19-1803.

[25] P. Messina *et al.*, "A Survey on Deep Learning and Explainability for Automatic Image-based Medical Report Generation." 2020, [Online]. Available: http://arxiv.org/abs/2010.10563.

[26] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, 2019, doi: 10.1038/s41597-019-0322-0.

[27] C. Eickhoff, I. Schwall, A. G. S. De Herrera, and H. Müller, "Overview of imageclefcaption 2017 - Image caption prediction and concept detection for biomedical images," in *CEUR Workshop Proceedings*, 2017, vol. 1866.

[28] A. G. Seco De Herrera, C. Eickhoff, V. Andrearczyk, and H. Müller, "Overview of the ImageCLEF 2018 caption prediction tasks," in *CEUR Workshop Proceedings*, 2018, vol. 2125.

[29] Y. Su, F. Liu, and M. P. Rosen, "UMass at ImageCLEF caption prediction 2018 task," in *CEUR Workshop Proceedings*, 2018, vol. 2125.

[30] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 2, pp. 304–310, 2016, doi: 10.1093/jamia/ocv080.

[31] D. R. Beddiar, M. Oussalah, and T. Seppänen, *Automatic captioning for medical imaging (MIC): a rapid review of literature*, no. Mic. Springer Netherlands, 2022.

[32] P. Harzig, Y. Y. Chen, F. Chen, and R. Lienhart, "Addressing data bias problems for chest x-ray image report generation," *30th Br. Mach. Vis. Conf. 2019, BMVC 2019*, 2020.

[33] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11769 LNCS, pp. 721–729, 2019, doi: 10.1007/978-3-030-32226-7_80.

[34] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Producing radiologist-quality reports for interpretable artificial intelligence," *undefined*, 2018, Accessed: Sep. 13, 2021. [Online]. Available: http://arxiv.org/abs/1806.00340.

[35] P. Zhang *et al.*, "VinVL: Revisiting Visual Representations in Vision-Language Models Pengchuan," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 5575–5584, 2021, [Online]. Available: https://ieeexplore.ieee.org/document/9577951/.

[36] K. W. Heath M, Bowyer K, Kopans D, Moore R, "THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY," *Proc. 5th Int. Work. Digit. Mammogr.*, 2000, [Online]. Available: https://www.ptonline.com/articles/how-to-get-better-mfi-results.

[37] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artif. Intell. Med.*, no. January, 2020.

[38] C. Y. Li, Z. Hu, X. Liang, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Advances in Neural Information Processing Systems*, 2018, vol. 2018-Decem, pp. 1530–1540.

[39] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in COntext (ROCO): A multimodal image dataset," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11043 LNCS, pp. 180–189, doi: 10.1007/978-3-030-01364-6_20.

[40] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, Feb. 2012, doi: 10.1016/j.acra.2011.09.014.

[41] Z. Han, B. Wei, S. Leung, J. Chung, and S. Li, *Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework*, vol. 11073 LNCS. Springer International Publishing, 2018.

[42] J. Tian, C. Li, Z. Shi, and F. Xu, *A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism*, vol. 11071 LNCS. Springer International Publishing, 2018.

[43] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 3462–3471, doi: 10.1109/CVPR.2017.369.

[44] H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman,

J. Yao, and R. M. Summers, “Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2497–2506, Mar. 2016, doi: 10.1109/CVPR.2016.274.

[45] J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 590–597, 2019, doi: 10.1609/aaai.v33i01.3301590.

[46] M. Douglass, “Computer-Assisted De-Identification of Free-text Nursing Notes,” 2005.

[47] M. M. Douglass, G. D. Cliffford, A. Reisner, W. J. Long, G. B. Moody, and R. G. Mark, “De-identification algorithm for free-text nursing notes,” *Comput. Cardiol.*, vol. 32, pp. 331–334, 2005, doi: 10.1109/CIC.2005.1588104.

[48] I. Neamatullah *et al.*, “BMC Medical Informatics and Decision Making Automated de-identification of free-text medical records,” 2008, doi: 10.1186/1472-6947-8-32.

[49] A. Bustos, A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá, “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” *Med. Image Anal.*, vol. 66, p. 101797, Dec. 2020, doi: 10.1016/j.media.2020.101797.

[50] L. Soldaini and N. Goharian, “Quickumls: a fast, unsupervised approach for medical concept extraction,” *MedIR Work. sigir*, pp. 1–4, 2016.

[51] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani, “Creating a classification of image types in the medical literature for visual categorization,” *Med. Imaging 2012 Adv. PACS-based Imaging Informatics Ther. Appl.*, vol. 8319, no. May 2014, p. 83190P, 2012, doi: 10.1117/12.911186.

[52] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, and Androutsopoulos, “AUEB NLP Group at ImageCLEFmed Caption Tasks 2021,” *CEUR Workshop Proc.*, vol. 2380, 2021.

[53] B. Ionescu *et al.*, “Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2022, vol. 13390 LNCS, pp. 541–564, doi: 10.1007/978-3-031-13643-6_31.

[54] J. H. Huang *et al.*, “DeepOpht: Medical report generation for retinal images via deep models and visual explanation,” *Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021*, pp. 2441–2451, 2021, doi: 10.1109/WACV48630.2021.00249.

[55] L. Y. Zizhao Zhang, Pingjun Chen, Manish Sapkota, Z. Zhang, P. Chen, M. Sapkota, and L. Yang, *TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References*, vol. 10435 LNCS. 2017, pp. 320–328.

[56] Y. Xue *et al.*, “Multimodal recurrent model with attention for automated radiology report generation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11070 LNCS, pp. 457–466, 2018, doi: 10.1007/978-3-030-00928-1_52.

[57] X. Huang, F. Yan, W. Xu, and M. Li, “Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation,” *IEEE Access*, vol. 7, pp. 154808–154817, 2019, doi: 10.1109/ACCESS.2019.2947134.

[58] J. Xu *et al.*, “Concept detection based on multi-label classification and image captioning approach - DAMO at ImageCLEF 2019,” *CEUR Workshop Proc.*, vol. 2380, 2019.

[59] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, “TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 9049–9058, 2018, doi: 10.1109/CVPR.2018.00943.

[60] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72, [Online]. Available: https://www.aclweb.org/anthology/W05-0909.

[61] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9909 LNCS, pp. 382–398, 2016, doi: 10.1007/978-3-319-46454-1_24.

[62] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning,” *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, vol. 1. pp. 199–209, 2017, doi: 10.18653/v1/e17-1019.

[63] D. Elliott and F. Keller, “Comparing automatic evaluation measures for image description,” *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 2. pp. 452–457, 2014, doi: 10.3115/v1/p14-2074.

[64] D. Hüske-Kraus, “Text generation in clinical medicine - A review,” in *Methods of Information in Medicine*, 2003, vol. 42, no. 1, pp. 51–60, doi: 10.1055/s-0038-1634209.

[65] D. Hueske-kraus, "Suregen-2: a shell system for the generation of clinical documents," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, 2003, pp. 215–218.

[66] S. Varges, H. Bieler, M. Stede, L. C. Faulstich, K. Irsig, and M. Atalla, "SemScribe: Natural language generation for medical reports," *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*. pp. 2674–2681, 2012.

[67] P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul, "From medical image to automatic medical report generation," *IBM Journal of Research and Development*, vol. 59, no. 2–3. 2015, doi: 10.1147/JRD.2015.2393193.

[68] P. Kisilev, E. Walach, S. Hashoul, E. Barkan, B. Ophir, and S. Alpert, "Semantic description of medical image findings: structured learning approach," 2015, pp. 171.1-171.11, doi: 10.5244/c.29.171.

[69] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to Summarize Radiology Findings," 2018. doi: 10.18653/v1/w18-5623.

[70] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 6666–6673, 2019, doi: 10.1609/aaai.v33i01.33016666.

[71] S. Biswal, C. Xiao, L. M. Glass, B. Westover, and J. Sun, "CLARA: Clinical Report Auto-completion," *Web Conf. 2020 - Proc. World Wide Web Conf. WWW 2020*, pp. 541–550, 2020, doi: 10.1145/3366423.3380137.

[72] X. Wang, Z. Guo, C. Xu, L. Sun, and J. Li, "ImageSem group at ImageCLEFmed caption 2021 task: Exploring the clinical significance of the textual descriptions derived from medical images," *CEUR Workshop Proc.*, vol. 2936, pp. 1387–1393, 2021.

[73] Y. Zhang, X. Wang, Z. Guo, and J. Li, "ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning," in *CEUR Workshop Proceedings*, 2018, vol. 2125.

[74] S. Liang, X. Li, Y. Zhu, X. Li, and S. Jiang, "ISIA at the ImageCLEF 2017 image caption task," *CEUR Workshop Proceedings*, vol. 1866. 2017.

[75] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 3156–3164, Oct. 2015, doi: 10.1109/CVPR.2015.7298935.

[76] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *undefined*, 2015.

[77] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models." 2014, [Online]. Available: http://arxiv.org/abs/1411.2539.

[78] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2015, Accessed: Sep. 13, 2021. [Online]. Available: https://arxiv.org/abs/1409.1556v6.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[80] C. Szegedy *et al.*, "Going deeper with convolutions," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1–9, 2015.

[81] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Systems with Applications*, vol. 167. p. 114095, Apr. 2021, doi: 10.1016/j.eswa.2020.114095.

[82] A. Tamkin, M. Wu, and N. Goodman, "Viewmaker Networks: Learning Views for Unsupervised Representation Learning," 2021, [Online]. Available: http://arxiv.org/abs/2010.07432.

[83] Z. Lin *et al.*, "Contrastive pre-training and linear interaction attention-based transformer for universal medical reports generation," *J. Biomed. Inform.*, vol. 138, no. January, p. 104281, 2023, doi: 10.1016/j.jbi.2023.104281.

[84] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 9726–9735, 2020, doi: 10.1109/CVPR42600.2020.00975.

[85] M. Alsharid, H. Sharma, L. Drukker, P. Chatelain, A. T. Papageorghiou, and J. A. Noble, "Captioning ultrasound images automatically," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11767 LNCS, no. 1, pp. 338–346, 2019, doi: 10.1007/978-3-030-32251-9_37.

[86] L. Wu, C. Wan, Y. Wu, and J. Liu, "Generative caption for diabetic retinopathy images," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2017*, 2018, vol. 2018-Janua, pp. 515–519, doi: 10.1109/SPAC.2017.8304332.

[87] V. Castro, P. Pino, D. Parra, and H. Lobel, "PUC Chile team at Caption Prediction: ResNet visual encoding and caption classification with parametric ReLU," in *CEUR Workshop Proceedings*, 2021, vol. 2936, pp. 1174–1183.

[88] D. Lyndon, A. Kumar, and J. Kim, "Neural captioning for the ImageCLEF 2017 medical image challenges,"

CEUR Workshop Proceedings, vol. 1866. 2017.

[89] V. Chhatbar, M. Gondhalekar, S. Pimple, and R. Pawar, "Machine Interpretation of Medical Images Using Deep Learning," *2021 2nd Glob. Conf. Adv. Technol. GCAT 2021*, pp. 3–7, 2021, doi: 10.1109/GCAT52182.2021.9587518.

[90] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 12910–12917, 2020, doi: 10.1609/aaai.v34i07.6989.

[91] O. Pelka and C. M. Friedrich, "Keyword generation for biomedical image retrieval with recurrent neural networks," in *CEUR Workshop Proceedings*, 2017, vol. 1866.

[92] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[93] B. Dai, D. Ye, and D. Lin, "Rethinking the Form of Latent States in Image Captioning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11209 LNCS, pp. 294–310, 2018, doi: 10.1007/978-3-030-01228-1_18.

[94] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[95] X. Zeng, L. Wen, B. Liu, and X. Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392. pp. 132–141, 2020, doi: 10.1016/j.neucom.2018.11.114.

[96] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network," *CVPR*, Jul. 2017.

[97] H. C. Shin *et al.*, "Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation," *J. Mach. Learn. Res.*, vol. 17, pp. 1–31, 2016, [Online]. Available: http://www.jmlr.org/papers/volume17/15-176/15-176.pdf.

[98] Z. Chen, Y. Song, T. H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1439–1449, 2020, doi: 10.18653/v1/2020.emnlp-main.112.

[99] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, p. 102510, 2022, doi: 10.1016/j.media.2022.102510.

[100] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring Visual Relationship for Image Captioning Ting," *Eur. Conf. Comput. Vis.*, vol. 1, pp. 711–727, 2018.

[101] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 765–773, 2019, doi: 10.1145/3343031.3350943.

[102] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–14, 2017.

[103] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8828, no. c, 2022, doi: 10.1109/TPAMI.2022.3148210.

[104] S. Jain *et al.*, "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports," no. NeurIPS, 2021, [Online]. Available: http://arxiv.org/abs/2106.14463.

[105] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[106] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 4249–4259, 2019, doi: 10.1109/ICCV.2019.00435.

[107] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, [Online]. Available: http://arxiv.org/abs/2010.11929.

[108] H. Wu *et al.*, "CvT: Introducing Convolutions to Vision Transformers," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 22–31, 2021, doi: 10.1109/ICCV48922.2021.00009.

[109] S. A. Hasan *et al.*, "PRNA at ImageCLEF 2017 caption prediction and concept detection tasks," *CEUR Workshop Proceedings*, vol. 1866. 2017.

[110] A.Géron, *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2019.

[111] Y. Xiong, B. Du, and P. Yan, "Reinforced Transformer for Medical Image Captioning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11861 LNCS, pp. 673–680, 2019, doi: 10.1007/978-3-030-32692-0_77.

[112] S. Maksoud, A. Wiliem, K. Zhao, T. Zhang, L. Wu, and B. Lovell, "CORAL8: Concurrent object regression for area localization in medical image panels," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11764 LNCS, pp. 432–441, 2019, doi: 10.1007/978-3-030-32239-7_48.

[113] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[114] W. Su *et al.*, "VL-BERT: Pre-training of Generic Visual-Linguistic Representations," pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1908.08530.

[115] K. D. W. F. Casalino LP, "Deep Learning in Medicine—Promise, Progress, and Challenges," *JAMA Intern. Med.*, vol. 179, no. 3, p. 293, Mar. 2019, doi: 10.1001/jamainternmed.2018.7117.

[116] H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, no. June, pp. 2497–2506, 2016, doi: 10.1109/CVPR.2016.274.

[117] F. Shamshad *et al.*, "Transformers in Medical Imaging: A Survey," pp. 1–41, 2022, [Online]. Available: http://arxiv.org/abs/2201.09873.

[118] "National Library of Medicine." https://lhncbc.nlm.nih.gov/ii/tools/MTI.html (accessed Feb. 25, 2022).

محمد تشنه‌لب مدرک دکترای خود را در سال ۱۹۹۵ از دانشگاه ساگا، ژاپن در رشته‌ی مهندسی برق دریافت نمودند و در حال حاضر با مرتبه علمی استاد تمام مشغول به تدریس در دانشکده مهندسی برق دانشگاه خواجه نصیرالدین طوسی می‌باشند. ایشان عضو مرکز کنترل صنعتی و بنیانگذار آزمایشگاه سیستم‌های هوشمند هستند همچنین یکی از بنیانگذاران و اعضای انجمن علمی سیستم‌های هوشمند ایران (ISSSI) و عضو هیئت تحریریه مجله بین المللی تحقیقات فناوری اطلاعات و ارتباطات (IJICTR) می‌باشد.

مریم رستگارپور مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۸۲ از دانشگاه خوارزمی و مدرک کارشناسی ارشد و دکتری خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه علوم و تحقیقات به ترتیب در سال‌های ۱۳۸۶ و ۱۳۹۲ و در زمینه‌های تخصصی قطعه‌بندی کلمات دستنویس و قطعه‌بندی تصاویر پزشکی اخذ نموده است. ایشان در حال حاضر استادیار گروه مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد ساوه است. علایق تحقیقاتی او شامل یادگیری ماشین، بهینه‌سازی، پردازش‌تصویر و بینایی‌ماشین است.

حسنیه ذوالفقاری مدرک کاردانی خود را در رشته کاربرد کامپیوتر از دانشگاه بیرجند و مدرک کارشناسی را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۸۲ از دانشگاه آزاد تهران‌جنوب و مدرک کارشناسی ارشد را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه علوم و تحقیقات در سال ۱۳۸۵اخذ کرده‌است. ایشان اکنون دانشجوی دکترا در رشته هوش مصنوعی دانشگاه علوم و تحقیقات و همچنین هیات‌علمی گروه مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد بیرجند می‌باشد. علایق تحقیقاتی وی پردازش تصویر و پردازش زبان‌های طبیعی است.

عباس کوچاری مدرک کارشناسی ارشد و دکتری خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی به ترتیب در سال ۱۳۸۴ از دانشگاه صنعتی امیرکبیر و سال ۱۳۹۱ از دانشگاه علم و صنعت اخذ نموده است. ایشان در حال حاضر استادیار گروه مهندسی کامپیوتر دانشگاه علوم و تحقیقات تهران است. علایق تحقیقاتی وی شامل پردازش‌تصویر و پردازش زبان‌های طبیعی و پردازش‌صوت می‌باشد.

علیرضا احسانبخش مدرک پزشکی عمومی خود را در سال۱۳۷۱ از دانشگاه مشهد اخذ کرد و در سال ۱۳۷۸ تخصص خود را در رشته رادیولوژی از دانشگاه زاهدان دریافت کرد. ایشان در حال حاضر دانشیار گروه تکنولوژی رادیولوژی در دانشگاه علوم پزشکی بیرجند می‌باشد.