



دانشکده مهندسی کامپیوتر

تولید گزارش برای تصاویر رادیولوژی مربوط به قفسه سینه

پروژه کارشناسی مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک

سینا علی نژاد

اساتید راهنما

دکتر صالح اعتمادی و دکتر محمدرضا جاهد مطلق

اردیبهشت ۱۴۰۴



تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: سینا علی‌نژاد

عنوان پروژه: تولید گزارش برای تصاویر رادیولوژی مربوط به قفسه سینه

تاریخ دفاع: اردیبهشت ۱۴۰۴

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضاء
۱	استاد راهنما	دکتر صالح اعتمادی	دانشیار	دانشگاه علم و صنعت ایران	
۲	استاد داور داخلی	دکتر ...	استادیار	دانشگاه علم و صنعت ایران	

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب سینا علی‌نژاد به شماره دانشجویی ۹۹۵۲۱۴۶۹ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پروژه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: سینا علی‌نژاد

تاریخ و امضا:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

- ☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.
- ☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.
- ☐ بهره‌برداری از این پایان‌نامه تا تاریخ ممنوع است.

اساتید راهنما: دکتر صالح اعتمادی

دکتر محمدرضا جاهد

مطلق

تاریخ:

امضا:

تقديم به:

پدر و مادرم

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست. در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر ...، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید. از جناب آقای دکتر ... که زحمت مطالعه و مشاوره این رساله را تقبل فرمودند و در آماده سازی این رساله، به نحو احسن اینجانب را مورد راهنمایی قرار دادند، کمال امتنان را دارم. همچنین لازم می‌دانم از پدید آورندگان بسته زی‌پرشین، مخصوصاً جناب آقای وفا خلیقی، که این پایان‌نامه با استفاده از این بسته، آماده شده است و همه دوستانمان در گروه پارسی‌لاتک کمال قدردانی را داشته باشم. در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می‌کنم وجود مقدس‌شان را و تشکر می‌کنم از خانواده عزیزم به پاس عاطفه سرشار و گرمای امیدبخش وجودشان، که بهترین پشتیبان من بودند.

سینا علی‌نژاد

اردیبهشت ۱۴۰۴

چکیده

در این پژوهش، به بررسی و پیشنهاد مدلی برای تولید گزارش رادیولوژی برای تصاویر اشعه ایکس مربوط به قفسه سینه، پرداخته خواهد شد. تولید خودکار این گزارش‌ها به دلیل تعداد درخواست بالای بیماران برای این نوع تصاویر و کمبود رادیولوژیست‌ها بسیار حائز اهمیت است و می‌تواند به صورت دستیار در کنار شخص رادیولوژیست به تشخیص بهتر ناهنجاری‌ها و در نتیجه درمان سریع‌تر کمک کند. در این مقاله، ابتدا یک مدل پایه RAG ارائه خواهد شد که از روش تولید بر محور بازیابی (RAG) استفاده می‌کند. سپس مدل FRAG معرفی می‌شود که در آن علاوه بر بازیابی گزارش‌ها، برخی ویژگی‌ها یا ناهنجاری‌های مربوط به قفسه سینه نیز بازیابی می‌شوند و هردوی گزارش‌ها و ویژگی‌ها در فرایند تولید گزارش نهایی استفاده می‌شود. مدل FRAG به دو روش FRAG-A و FRAG-B پیاده‌سازی می‌شود که FRAG-A از ماژول استنتاج بدون نمونه استفاده می‌کند در حالیکه FRAG-B از دسته‌بند خطی بهره می‌گیرد. در نهایت تمام این روش‌ها با استفاده از معیارهای پزشکی و زبان طبیعی ارزیابی می‌شوند. نتایج ارزیابی نشان می‌دهد که مدل‌های FRAG با اختلاف زیادی در معیارهای پزشکی بهتر از مدل RAG هستند و در بین آن دو، مدل FRAG-A عملکرد بهتری دارد. همچنین در آزمایش‌هایی جدا، تاثیر روش بازیابی، مدل زبانی بزرگ و پرامپت استفاده شده در بخش تولید نیز مورد بررسی قرار خواهند گرفت. برای بازیابی از مدل بینایی-زبان ELIXR-B استفاده خواهد شد و پس از استخراج جانمایی‌ها، به غیر از انتخاب بهترین‌ها، از دو روش دیگر یعنی خوشه‌بندی K-Means و انتخاب بیشینه متنوع استفاده خواهد شد. نتیجه نهایی نشان خواهد داد دو مورد روش بازیابی و مدل زبانی بزرگ در عملکرد مدل تاثیرگذار هستند، در حالیکه پرامپت ورودی تاثیر زیادی ندارد.

واژگان کلیدی: RAG، FRAG، تولید بر محور بازیابی، استنتاج بدون نمونه، دسته‌بند خطی، مدل زبانی بزرگ، خوشه‌بندی K-Means، ELIXR-B، مدل بینایی-زبان، پرامپت، جانمایی

فهرست مطالب

خ	فهرست تصاویر
د	فهرست جداول
۱	فصل ۱: مقدمه
۲	۱-۱ لزوم خودکارسازی تولید گزارش برای عکس‌های رادیولوژی
۳	۱-۲ نقش هوش مصنوعی در خودکارسازی تولید گزارش رادیولوژی
۴	فصل ۲: مفاهیم و کارهای انجام شده
۵	۲-۱ مفاهیم اولیه
۱۳	۲-۲ کارهای انجام شده با روش‌های پیشین
۱۵	فصل ۳: شرح مسئله
۱۹	فصل ۴: روش پیشنهادی
۲۰	۴-۱ پردازش و جمع‌آوری داده
۲۲	۴-۲ مدل RAG
۲۵	۴-۳ مدل FRAG-A
۲۶	۴-۴ مدل FRAG-B
۲۹	فصل ۵: ارزیابی و معیارهای سنجش عملکرد
۳۰	۵-۱ معیارهای ارزیابی

۵-۲ نتایج ارزیابی ۳۵

فصل ۶: نتیجه‌گیری و کارهای آینده ۳۸

۶-۱ نتیجه‌گیری ۳۹

۶-۲ کارهای آینده ۳۹

کتاب‌نامه ۴۱

واژه‌نامه فارسی به انگلیسی ۵۱

فهرست تصاویر

- ۱-۲ فرایند آموزش و inference در مدل ELIXR-C ۱۲
- ۲-۲ فرایند آموزش و inference برای مدل ELIXR-B ۱۳
- ۱-۳ نمایی از وظیفه مدل تولید خودکار گزارش رادیولوژی ۱۷
- ۱-۴ ساختار مدل RAG ۲۳
- ۲-۴ ساختار مدل FRAG-A: اضافه کردن ماژول استنتاج بدون نمونه ۲۵
- ۳-۴ نمایی از عملکرد ماژول استنتاج بدون نمونه ۲۶
- ۴-۴ ساختار مدل دسته‌بند برای تشخیص ناهنجاری‌ها ۲۷
- ۵-۴ نمودار ROC مدل خطی ارائه شده برای تشخیص ناهنجاری‌ها ۲۷
- ۶-۴ ساختار مدل FRAG-B: جایگزینی ماژول استنتاج بدون نمونه با دسته‌بند خطی ۲۸

فهرست جداول

۱-۳	ناهنجاری‌های قفسه سینه	۱۸
۱-۴	آمار کلی مجموعه داده BioNLP: این مجموعه داده از ترکیب مجموعه داده‌های معروف در حوزه رادیولوژی قفسه سینه تشکیل شده است.	۲۰
۲-۴	نمایی از مجموعه داده کارگاه BioNLP	۲۱
۳-۴	تعداد نمونه‌های ناهنجاری‌های مختلف برای آموزش و اعتبارسنجی مدل دسته‌بند خطی	۲۸
۱-۵	نتایج ارزیابی مدل‌های مختلف ارائه شده	۳۵
۲-۵	مقایسه روش‌های مختلف در مازول بازیابی	۳۶
۳-۵	مقایسه دو مدل زبانی بزرگ	۳۶
۴-۵	مقایسه دو پرامپت مختلف	۳۶

فصل ۱

مقدمه

رشته رادیولوژی^۱ نقش بسیار مهمی در پزشکی مدرن ایفا می‌کند و از طریق تصویربرداری پزشکی، تشخیص، درمان و مدیریت بیمار را ممکن می‌سازد. با این حال، کمبود روزافزون جهانی رادیولوژیست^۲ و تقاضای روزافزون برای خدمات تصویربرداری پزشکی فشار زیادی را بر بخش‌های رادیولوژی وارد می‌کند. این روند می‌تواند ظرفیت رادیولوژیست برای ارائه گزارش‌های به موقع و دقیق را به خطر بیندازد و منجر به تاخیر و خطاهای احتمالی در تصمیم‌گیری‌های بالینی شود. تولید خودکار گزارش‌های رادیولوژی^۳ به عنوان یک راه‌حل حیاتی برای کاهش بار کاری رادیولوژیست، بهبود مراقبت از بیمار و افزایش کارایی کلی سیستم‌های مراقبت‌های بهداشتی پدیدار شده است.

۱-۱ لزوم خودکارسازی تولید گزارش برای عکس‌های رادیولوژی

رادیولوژی برای مراقبت‌های بهداشتی مدرن حیاتی است و ابزار تشخیصی اصلی برای نظارت بر درمان و پیش‌بینی نتایج است که توسط پزشکان استفاده می‌شود. رادیولوژی شامل روش‌های مختلف تصویربرداری مانند اشعه ایکس^۴ است و پس از انجام هر معاینه، گزارش رادیولوژی برای راهنمایی پزشک تهیه می‌شود. دقت تشخیص چنین گزارش‌هایی برای اطمینان از مراقبت بهینه بسیار مهم است، با این حال گزارش شده است که نرخ خطای ۳ تا ۵ درصد وجود دارد و همچنین بیان عدم قطعیت در ۳۵ درصد گزارش‌ها یافت شده است. این می‌تواند با کمبود رادیولوژیست در برخی مناطق تشدید شود، به عنوان مثال ۹۷ درصد از بخش‌های تصویربرداری در بریتانیا گزارش می‌دهند که نمی‌توانند با تقاضا پاسخگو باشند. تعداد کم پرسنل می‌تواند باعث تأخیر در گزارش‌دهی شود که منجر به اتخاذ تصمیم‌های حیاتی بدون بهره‌مندی از نظر رادیولوژیست، توسط پزشکان شود و به طور بالقوه به نتیجه‌گیری متفاوتی در مقایسه با یک رادیولوژیست حرفه‌ای برسد.

[۱۴]

¹Radiology²Radiologist³Automated Radiology Report Generation (ARRG)⁴X-ray

۱-۲ نقش هوش مصنوعی در خودکارسازی تولید گزارش رادیولوژی

مراقبت‌های بهداشتی یک کاربرد مهم از یادگیری عمیق^۵ است و با رادیولوژی که ۹۰ درصد داده‌های مراقبت‌های بهداشتی را تولید می‌کند، موضوعی جذاب برای محققان یادگیری عمیق است. روش‌هایی از بینایی کامپیوتر^۶ و پردازش زبان طبیعی^۷ در حال حاضر در حوزه مراقبت‌های بهداشتی برای بهبود دسترسی آسان و ارتقای استانداردهای مراقبت بهداشتی در حال به‌کارگیری هستند. یکی از کاربردهای یادگیری عمیق در مراقبت‌های بهداشتی که به سرعت در حال توسعه است، تولید خودکار گزارش رادیولوژی است، وظیفه‌ای که در هر دو حوزه‌ی زبان و بینایی قرار می‌گیرد که شباهت‌هایی به موضوع گسترده‌تری با عنوان تولید زیرنویس برای تصویر^۸ دارد. تولید خودکار گزارش رادیولوژی با توانایی خود در افزایش قابلیت‌های رادیولوژیست‌ها، ارزش بالینی قابل توجهی دارد و می‌تواند با تولید گزارش برای موارد نسبتاً ساده محدودیت‌های زمانی را کاهش دهد. همچنین می‌تواند برای یک رادیولوژیست بی‌تجربه با مشخص کردن خودکار هر گونه ناهنجاری‌های احتمالی، به عنوان خواننده دوم عمل کند.

⁵Deep Learning

⁶Computer Vision (CV)

⁷Natural Language Processing (NLP)

⁸Image Captioning

فصل ۲

مفاهیم و کارهای انجام شده

۲-۱ مفاهیم اولیه

۲-۱-۱ Neural Networks

شبکه‌های عصبی مدل‌های محاسباتی الهام گرفته از مغز انسان هستند که از لایه‌هایی از گره‌های متصل تشکیل شده‌اند و داده‌ها را از طریق اتصالات وزنی پردازش می‌کنند. این شبکه‌ها در تشخیص الگوها، پیش‌بینی‌ها و حل مسائل پیچیده مانند طبقه‌بندی تصاویر و پردازش زبان طبیعی عملکرد فوق‌العاده‌ای دارند. آموزش یک شبکه عصبی شامل تنظیم وزن‌ها با استفاده از تکنیک‌های بهینه‌سازی مانند پس‌انتشار^۱ و گرادیان نزولی^۲ است. با پیشرفت یادگیری عمیق، شبکه‌های عصبی تحولات بزرگی در حوزه‌هایی مانند مراقبت‌های بهداشتی، مالی و سیستم‌های خودران ایجاد کرده‌اند.

۲-۱-۲ RNN

شبکه‌های عصبی بازگشتی^۳ نوعی از شبکه‌های عصبی هستند که برای پردازش داده‌های ترتیبی طراحی شده‌اند، به‌طوری که اطلاعات گذشته بر پیش‌بینی‌های آینده تأثیر می‌گذارد. برخلاف شبکه‌های عصبی سنتی، این شبکه‌ها دارای حلقه‌هایی هستند که به آن‌ها امکان حفظ حافظه ورودی‌های قبلی را می‌دهد و آن‌ها را برای وظایفی مانند تشخیص گفتار و پیش‌بینی سری‌های زمانی مؤثر می‌سازد. با این حال، این شبکه‌ها در حفظ وابستگی‌های طولانی‌مدت به دلیل مشکلاتی مانند گرادیان ناپدیدشونده^۴ دچار چالش می‌شوند. مدل‌های پیشرفته‌تر مانند LSTM و GRU این محدودیت‌ها را برطرف کرده و عملکرد بهتری در پردازش دنباله‌های طولانی ارائه می‌دهند. [۱۳]

۲-۱-۳ LSTM

به مرور زمان و با آموزش مدل‌های شبکه‌ی بازگشتی، محققان شاهد مشکل گرادیان ناپدیدشونده و انفجار گرادیان^۵ در این نوع شبکه‌های عصبی بوده‌اند. یعنی به مرور زمان دچار فراموشی داده‌های قبلی و در نتیجه

^۱Backpropagation^۲Gradient Descent^۳Recurrent Neural Networks (RNN)^۴Vanishing Gradient^۵Exploding Gradient

ساختار کلی متن را فراموش می‌کردند. سپس با ارائه مدل حافظه طولانی کوتاه مدت^۶ توانستند بر این مشکل غلبه کنند به طوری که با تعریف دروازه های ورودی، فراموشی و خروجی داده های مورد نیاز را نگه می‌داشتند و داده های غیرقابل استفاده را از درون حافظه پاک می‌کردند. [۱۰]

۲-۱-۴ Attention

مدل شبکه عصبی حافظه طولانی کوتاه مدت ارائه شده همچنان دچار فراموشی هایی به مرور زمان می‌شد و نمی‌توانست یک دنباله با طول بسیار زیاد را به خاطر بسپارد و همچنان مشکل ناپدید شدن گرادیان مشاهده می‌شد. برای برطرف کردن مشکل ذکر شده پژوهشگران به ایده‌ی استفاده کردن از سازوکار توجه رسیدند که به قدر خوبی تمام مشکلاتی که تا کنون مطرح شد را حل می‌کرد. [۱۷]

۲-۱-۵ Transformer

در جدیدترین پژوهش، مدل های مبدل^۷ مطرح شده اند که علم پردازش زبان طبیعی را متحول کردند. آن‌ها در مقاله‌ی خود ساختاری جدید را معرفی کردند که دیگر ساختار این مدل ها بر پایه شبکه های عصبی بازگشتی نمی‌باشند. راه حل نوینی برای بردارهای وابسته به متن ارائه کرده‌اند که نویسندگان این مقاله با ارائه سازوکار توجه به خود واقع به این معناست اگر دو کلمه‌ی هم شکل با معنای متفاوت درون متن قرار بگیرند، این سازوکار متوجه تفاوت این دو کلمه خواهد شد. آنها همچنین فرآیند وابسته بودن هر بخش در شبکه‌های عصبی بازگشتی که منجر به کند بودن آن می‌شد را با استفاده از مدل جدید خود کاملاً به طور موازی درآوردند که بسیار به کار سرعت می‌بخشید. قدم بزرگ دیگر این ساختار، آموزش دیدن مدل های کارآمدی می‌باشند که در واقع با استفاده از این مدل ها که بر روی حجم بسیار عظیمی از متن‌ها آموزش دیده‌اند، می‌توانیم از وزن^۸ های آموزش دیده‌ی آن‌ها در مسائل مختلف استفاده کنیم و مدل ها را برای انجام وظایف جدید بدون نیاز به آموزش دوباره از ابتدا به کار بگیریم. این ویژگی به ویژه در مواقعی که داده‌های آموزشی محدود هستند، بسیار مفید است. به عنوان مثال، مدل های مبدل که بر روی مقادیر عظیمی از داده‌های عمومی آموزش دیده‌اند، می‌توانند با تنظیمات اندک و استفاده از وزن های یادگیری شده، به سرعت به مسائل خاصی مانند ترجمه، خلاصه سازی

^۶Long Short-Term Memory (LSTM)

^۷Transformer

^۸Weight

متن، یا پاسخ به سوالات پاسخ دهند.

این مدل‌ها همچنین به دلیل ساختار مقیاس‌پذیری که دارند، امکان پردازش موازی را فراهم می‌کنند که این امر منجر به کاهش زمان آموزش و پیش‌بینی می‌شود. علاوه بر این، مدل‌های مبدل به دلیل حذف محدودیت‌های مربوط به وابستگی‌های طولانی مدت در متن، قادر به درک بهتر و عمیق‌تری از توالی‌های طولانی هستند.

مدل‌های مبدل، از جمله معروف‌ترین آنها یعنی Bert و GPT به عنوان پایه‌ای برای بسیاری از کاربردهای عملی در پردازش زبان طبیعی مورد استفاده قرار گرفته‌اند. این مدل‌ها توانسته‌اند در بسیاری از معیارهای استاندارد، عملکردی بهتر از مدل‌های پیشین ارائه دهند و در واقع، انقلابی در این حوزه به وجود آورده‌اند. در نهایت، مبدل‌ها با ارائه رویکردی جدید به پردازش زبان طبیعی، امکان توسعه سیستم‌های هوشمندتر و کارآمدتر را فراهم کرده‌اند که می‌توانند در طیف وسیعی از کاربردها از جمله ترجمه ماشینی، تولید متن، تحلیل احساسات و بسیاری دیگر به کار گرفته شوند. این مدل‌ها به دلیل انعطاف‌پذیری و قدرت پیش‌بینی بالای خود، همچنان در حال پیشرفت و بهبود هستند و تحقیقات بیشتری در این زمینه در حال انجام است تا از توانایی کامل آن‌ها بهره‌برداری شود. [۱۷]

۲-۱-۶ CNN

شبکه‌های عصبی کانولوشنی^۹ نوعی از شبکه‌های عصبی هستند که برای پردازش داده‌های شبکه‌ای مانند تصاویر طراحی شده‌اند. این شبکه‌ها با استفاده از لایه‌های کانولوشنی، سلسله‌مراتب مکانی ویژگی‌ها را به صورت خودکار یاد می‌گیرند و در وظایفی مانند طبقه‌بندی تصاویر، تشخیص اشیا و تحلیل تصاویر پزشکی بسیار مؤثر هستند. شبکه‌های عصبی کانولوشنی با بهره‌گیری از نواحی پذیرش محلی، اشتراک وزن و لایه‌های تجمعی، نیاز به استخراج ویژگی‌های دستی را کاهش می‌دهند. [۷]

۲-۱-۷ Vision Transformers

مدل‌های مبدل بینایی^{۱۰} معماری یادگیری عمیقی هستند که با استفاده از مکانیزم خودتوجهی در پردازش تصاویر، رویکردی متفاوت از روش‌های سنتی کانولوشنی ارائه می‌دهند. این مدل‌ها با تقسیم تصویر به

^۹Convolutional Neural Networks (CNN)

^{۱۰}Vision Transformers (ViT)

وصله‌هایی^{۱۱} و پردازش آن‌ها به‌عنوان دنباله‌ای از توکن^{۱۲}‌ها، وابستگی‌های بلندمدت را مؤثرتر از مدل‌های کانولوشنی یاد می‌گیرند. این معماری در وظایفی مانند طبقه‌بندی و بخش‌بندی تصاویر عملکرد پیشرفته‌ای نشان داده‌اند، به‌ویژه زمانی که روی مجموعه داده‌های بزرگ آموزش ببینند. [۲]

۲-۱-۸ Vision-Language Models

مدل‌های بینایی-زبانی^{۱۳} برای پردازش و ادغام اطلاعات از داده‌های تصویری و متنی طراحی شده‌اند و درک جامع‌تری از ورودی‌های چندوجهی ارائه می‌دهند. این مدل‌ها با الهام از کاربردهای دنیای واقعی مانند رانندگی خودران و تشخیص سرطان، پل ارتباطی بین پردازش زبان طبیعی و بینایی کامپیوتری ایجاد می‌کنند. با مدل‌سازی همزمان این دو حوزه، این مدل‌ها وظایفی مانند توصیف تصویر، پاسخ‌دهی به سؤالات تصویری و تصمیم‌گیری‌های پزشکی را بهبود می‌بخشند. این مدل‌ها نقش کلیدی در پیشرفت سیستم‌های هوش مصنوعی ایفا می‌کنند که نیاز به استدلال در زمینه‌های چندوجهی دارند.

Dual Stream vs. Single Stream VLM

مدل‌های زبان-بینایی بر اساس نحوه ترکیب داده‌های متنی و تصویری به دو دسته تک‌جریانی و دوجریانی^{۱۴} تقسیم می‌شوند. مدل‌های تک‌جریانی هر دو نوع داده را در یک ماژول یکپارچه پردازش کرده و با ترکیب زودهنگام ویژگی‌ها، کارایی محاسباتی و صرفه‌جویی در پارامترها را بهبود می‌بخشند. در مقابل، مدل‌های دوجریانی مسیرهای جداگانه‌ای برای پردازش متن و تصویر دارند و سپس با استفاده از مکانیزم‌های توجه، این دو نوع داده را ادغام می‌کنند، که منجر به تعاملات پیچیده‌تر میان آن‌ها می‌شود. مدل‌های تک‌جریانی سریع‌تر و کم‌هزینه‌تر هستند، در حالی که مدل‌های دوجریانی انعطاف‌پذیری بیشتری در پردازش اطلاعات چندوجهی دارند. انتخاب بین این دو معماری بستگی به نیاز بهینه‌سازی بین کارایی و قدرت بازنمایی در یک وظیفه خاص دارد.

¹¹ Patch

¹² Token

¹³ Vision-Language Models (VLM)

¹⁴ Single Stream and Dual Stream

¹⁵ Module

Encoder-only vs. Encoder-Decoder VLM

مدل‌های زبان-بینایی بر اساس نحوه پردازش بازنمایی‌های چندوجهی^{۱۶} به دو دسته رمزگذار-محور^{۱۷} و رمزگذار-رمزگشا^{۱۸} تقسیم می‌شوند. مدل‌های رمزگذار-محور بر یادگیری بازنمایی‌های کارآمد تمرکز دارند و به دلیل سادگی پردازش و کاهش پیچیدگی محاسباتی، برای استخراج ویژگی‌های فشرده مناسب هستند، اما در تولید خروجی‌های پیچیده و دقیق محدودیت دارند. در مقابل، مدل‌های رمزگذار-رمزگشا با افزودن یک مرحله رمزگشایی، قادر به تولید خروجی‌های متنوع و غنی هستند که برای وظایفی مانند توصیف تصویر و ترجمه مفید است. این انعطاف‌پذیری با افزایش هزینه‌های محاسباتی همراه است. انتخاب بین این دو معماری بستگی به نیاز بهینه‌سازی بین کارایی و قدرت تولیدی مدل دارد.

۲-۱-۹ VLM Training

Transfer Learning

یادگیری انتقالی^{۱۹} یک رویکرد رایج در یادگیری ماشین است که از مدل‌های از پیش آموزش دیده استفاده کرده و آن‌ها را برای وظایف خاص تطبیق می‌دهد. این فرآیند معمولاً با تنظیم دقیق^{۲۰} پارامترهای مدل بر روی مجموعه داده‌های کوچک‌تر و مخصوص هر وظیفه انجام می‌شود تا چالش‌های خاص آن را برطرف کند. در برخی موارد، تغییراتی در معماری مدل، مانند اصلاح لایه‌های نهایی، برای هماهنگی با نیازهای وظیفه جدید ضروری است. با حفظ دانش کسب‌شده از آموزش اولیه، یادگیری انتقالی به یادگیری کارآمدتر و بهبود عملکرد در مقایسه با آموزش مدل از ابتدا کمک می‌کند.

Curriculum Learning

یادگیری برنامه‌ریزی‌شده^{۲۱} یک رویکرد ساختاریافته است که داده‌های آموزشی یا وظایف را بر اساس میزان پیچیدگی به ترتیب پیش‌رونده سازمان‌دهی می‌کند. این روش با شروع از نمونه‌های ساده‌تر و معرفی تدریجی

¹⁶Cross-modal representation¹⁷Encoder-only models¹⁸Encoder-Decoder models¹⁹Transfer Learning²⁰Fine-Tuning²¹Curriculum Learning

نمونه‌های پیچیده‌تر، به مدل‌ها کمک می‌کند تا به شکل مؤثرتری یاد بگیرند. به عنوان مثال، مدل زبانی پزشکی LLaVa-Med از یادگیری برنامه‌ریزی شده در فرآیند آموزش خود استفاده می‌کند. این راهبرد گام‌به‌گام توانایی مدل را در پردازش وظایف پیچیده با دقت و کارایی بیشتر بهبود می‌بخشد.

Self-Supervised Learning

یادگیری خودنظارتی^{۲۲} یک روش کلیدی در آموزش مدل‌های بینایی-زبانی است که به آن‌ها امکان می‌دهد بدون وابستگی به داده‌های برچسب‌گذاری شده یاد بگیرند. در این روش، مدل‌ها با بهره‌گیری از ساختارهای درونی داده، برچسب‌های خود را تولید می‌کنند، که این امر به‌ویژه در شرایطی که داده‌های برچسب‌دار به‌سختی در دسترس یا پرهزینه هستند، بسیار مفید است. این رویکرد به مدل‌ها کمک می‌کند تا نمایش‌های معنایی مناسبی را در میان حالات مختلف داده بدون نظارت مستقیم بیاموزند. تکنیک‌های رایج در یادگیری خودنظارتی شامل یادگیری تضادمحور^{۲۳}، مدل‌سازی زبان ماسک‌شده^{۲۴}، و مدل‌سازی تصویر ماسک‌شده^{۲۵} هستند که توانایی مدل را در یادگیری مؤثر از داده‌های خام تقویت می‌کنند.

یادگیری تضادمحور، مدل را وادار می‌کند که نمونه‌های مشابه را به هم نزدیک و نمونه‌های نامرتبط را از هم دور کند. مدل‌سازی زبان ماسک‌شده شامل حذف برخی از کلمات در متن و پیش‌بینی آن‌ها توسط مدل است. مدل‌سازی تصویر ماسک‌شده نیز مشابه این روش، بخش‌هایی از تصویر را حذف کرده و مدل را برای بازسازی آن‌ها آموزش می‌دهد.

In-Context Learning

یادگیری درون‌متنی^{۲۶} روشی برای انطباق مدل‌های بینایی-زبانی بدون تغییر پارامترهای آن‌ها است که تنها از اطلاعات ورودی استفاده می‌کند. یکی از تکنیک‌های مهم در این روش مهندسی پرسش^{۲۷} است که شامل طراحی دستورالعمل‌های خاص برای هدایت مدل به سمت تولید خروجی‌های دقیق‌تر می‌شود. این روش می‌تواند شامل ارائه چندین نمونه مرتبط یا ساختاردهی تدریجی پرسش‌ها برای بهبود پاسخ مدل باشد.

²² Self-Supervised Learning (SSL)

²³ Contrastive Learning (CL)

²⁴ Masked Language Modeling (MLM)

²⁵ Masked Image Modeling (MIM)

²⁶ In-Context Learning

²⁷ Prompt Engineering

تولید افزوده‌شده با بازیابی^{۲۸} نیز نوعی مهندسی پرسش است که ترکیبی از بازیابی اطلاعات و تولید متن را ارائه می‌دهد. در این روش، ابتدا یک مدل بازیابی اطلاعات مرتبط را از مجموعه داده‌های گسترده استخراج می‌کند و سپس یک مدل زبانی بزرگ^{۲۹} بر اساس این اطلاعات، خروجی تولید می‌کند. این معماری با کاهش وابستگی به داده‌های برچسب‌دار، دقت و تطبیق‌پذیری مدل را بهبود می‌بخشد.

۲-۱-۱۰ Report Generation

تولید گزارش^{۳۰} یکی از وظایف مهم مدل‌های بینایی-زبانی در حوزه پزشکی است که بر ایجاد خلاصه‌ای جامع از داده‌های تصویری تمرکز دارد. این فناوری نقش کلیدی در خلاصه‌سازی خودکار نتایج تصویربرداری پزشکی و کاهش بار کاری نگارش گزارش‌ها ایفا می‌کند. به‌عنوان مثال، در رادیولوژی، سیستم تولید گزارش می‌تواند تصاویر پزشکی مانند اشعه ایکس، سی‌تی‌اسکن یا ام‌آرآی را تحلیل کرده و گزارشی دقیق از ناهنجاری‌های مشاهده‌شده، محل آن‌ها و پیامدهای احتمالی برای تشخیص یا درمان تولید کند. گزارش‌های رادیولوژی معمولاً شامل بخش‌هایی مانند نوع معاینه^{۳۱}، دلایل انجام آن^{۳۲}، مقایسه با تصاویر قبلی^{۳۳}، روش اسکن^{۳۴}، یافته‌های دقیق^{۳۵} و جمع‌بندی نتایج اصلی^{۳۶} هستند. در این فرآیند، مدل بینایی-زبان عمدتاً برای تولید بخش‌های یافته‌ها (Findings) و جمع‌بندی (Impression) طراحی شده‌اند.

۲-۱-۱۱ معرفی مدل ELIXR-B

این مدل برای تولید گزارش رادیولوژی مربوط به قفسه سینه می‌باشد و توسط گوگل ارائه شده است و در کنار آن مدل دیگری به اسم ELIXR-C نیز معرفی شد. در ELIXR-C از یک رمزگذار تصویر و یک رمزگذار متن برای دستیابی به جانمایی‌های عکس و متن استفاده شده و سپس از طریق تابع ضرر در معماری CLIP سعی بر نزدیک کردن عکسها و تصاویر مشابه در فضای جانمایی و دور کردن موارد غیریکسان شده است. فرایند

²⁸ Retrieval-Augmented Generation (RAG)

²⁹ Large Language Model (LLM)

³⁰ Report Generation (RG)

³¹ Examination

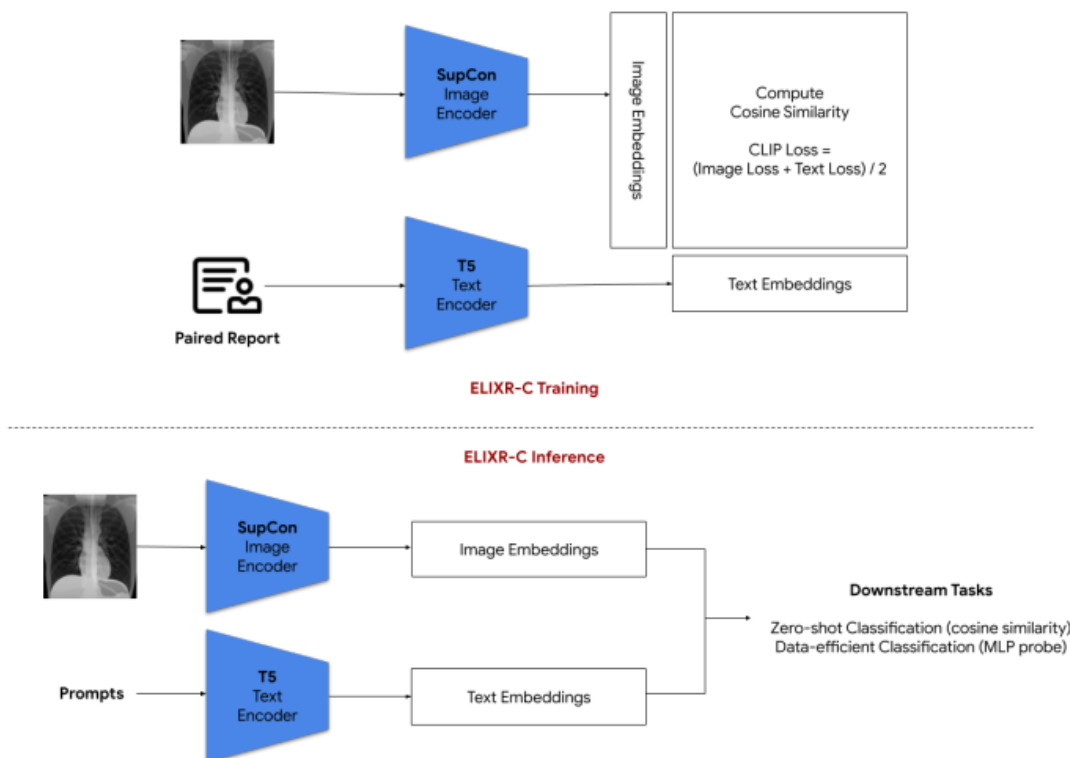
³² Indication

³³ Comparison

³⁴ Technique

³⁵ Findings

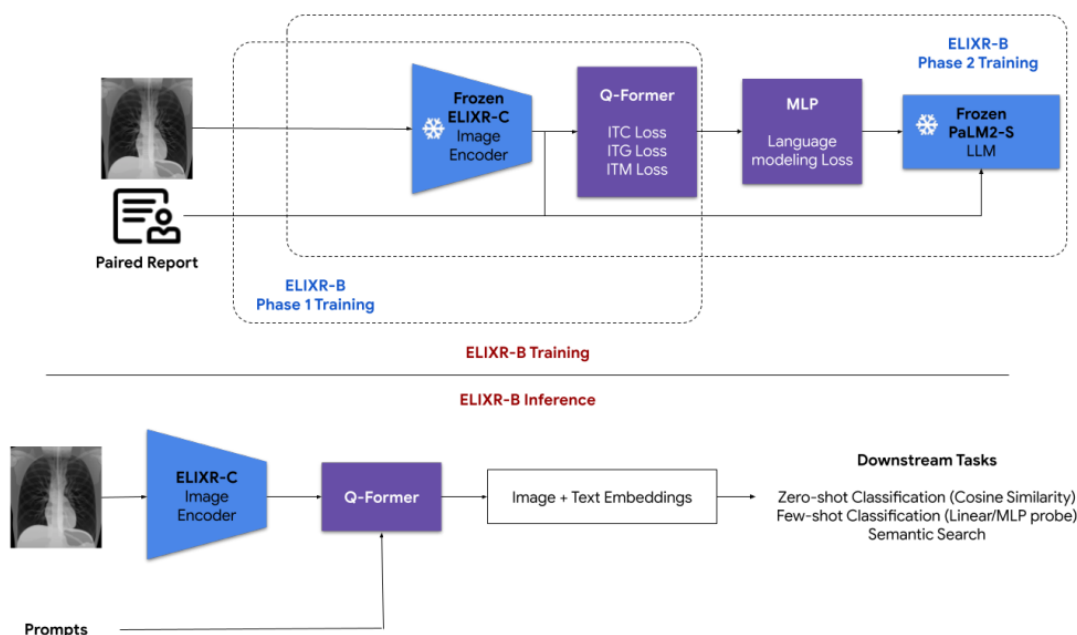
³⁶ Impression



شکل ۲-۱: فرایند آموزش و inference در مدل ELIXR-C

آموزش و inference این مدل را در شکل ۲-۱ ببینید.

نویسندگان همچنین شبکه‌ای به نام ELIXR-B را بر پایه معماری BLIP-2 آموزش دادند که به عنوان یک آداپتور بین انکودر تصویر (ELIXR-C) و مدل زبانی (PaLM 2-S) عمل می‌کند. این شبکه ویژگی‌های مکانی از تصاویر را استخراج کرده و آن‌ها را به فضای توکن‌های زبانی مدل منتقل می‌کند. آموزش ELIXR-B در دو مرحله انجام می‌شود: در مرحله اول، Q-Former با استفاده از سه وظیفه یادگیری تضاد تصویر-متن، تولید متن مبتنی بر تصویر، و تطابق تصویر-متن، بازنمایی مشترکی از تصویر و گزارش یاد می‌گیرد. در مرحله دوم، Q-Former و یک پرسپترون چندلایه آموزش داده می‌شوند تا بخش "impressions" گزارش‌های رادیولوژی را بر اساس تصویر تولید کنند. در نهایت، Q-Former قادر است اطلاعات مرتبط تصویری را به توکن‌هایی سازگار با مدل زبانی تبدیل کرده و اطلاعات غیرضروری را حذف کند. شکل ۲-۲ فرایند آموزش و inference را برای مدل ELIXR-B نشان می‌دهد. [۱۸]



شکل ۲-۲: فرایند آموزش و inference برای مدل ELIXR-B

۲-۲ کارهای انجام شده با روش‌های پیشین

حوزه‌ی تولید گزارش‌های رادیولوژی برای تصاویر اشعه‌ی ایکس از قفسه‌ی سینه شاهد پیشرفت‌های قابل توجهی بوده است، به‌ویژه از طریق یکپارچه‌سازی هوش مصنوعی و تکنیک‌های یادگیری عمیق. پژوهشگران مدل‌ها و روش‌های نوآورانه‌ای را توسعه داده‌اند که هدف آن‌ها خودکارسازی فرایند تولید گزارش‌های رادیولوژی منسجم و مرتبط با متن از تصاویر اشعه‌ی ایکس قفسه‌ی سینه است و در نتیجه بهبود جریان کاری بالینی و مراقبت از بیماران را به دنبال دارد.

کارهای اخیر در حوزه‌ی تولید گزارش‌های رادیولوژی این مسئله را به‌عنوان یک وظیفه‌ی تولیدی^{۳۷} در نظر گرفته‌اند، مانند مقاله [۱] که از معماری مبدل رمزگشا^{۳۸} در مدل R2Gen استفاده کرد و همچنین مقاله [۱۱] که بر تولید گزارش‌های کامل، منسجم و از نظر بالینی دقیق با استفاده از یک رویکرد یادگیری تقویتی مبتنی بر پاداش^{۳۹} تحت عنوان M2 Trans تمرکز داشت.

مقاله [۲] در کار خود مدل CXR-RePaiR را معرفی کردند و مسئله‌ی تولید گزارش‌های رادیولوژی را

³⁷Generative

³⁸Transformer decoder

³⁹Reward-based Reinforcement Learning

با استفاده از یک رویکرد بازیابی^{۴۰} مورد بررسی قرار دادند و یک معیار جدید و قابل اطمینان‌تر از نظر بالینی برای ارزیابی این سیستم‌ها ارائه کردند که به‌عنوان SOTA^{۴۱} جدیدی شناخته شد. فرآیند بازیابی بر اساس یک مدل بینایی-زبانی که با روش یادگیری تضادمحور پیش‌آموزش‌یافته بود و از مجموعه داده‌ی MIMIC-CXR معرفی شده در مقاله [۶] استفاده می‌کرد، انجام شد. در این پژوهش، یک معیار جدید برای سنجش شباهت معنایی با نام S_{emb} معرفی شد که شباهت معنایی بین گزارش مرجع و گزارش پیش‌بینی‌شده را با استفاده از نمایش‌های پنهان آخرین لایه^{۴۲} مدل CheXbert، معرفی شده در مقاله [۱۵] محاسبه می‌کرد. این مقاله همچنین از معیار BERTScore در مقاله [۲۰] به‌عنوان یک معیار دیگر برای سنجش شباهت معنایی استفاده کرد.

نویسندگان در مقاله [۱۲] به یکی از چالش‌های کلیدی در هم‌پوشانی خودکار تولید گزارش‌های رادیولوژی پرداختند که شامل ارجاعات به گزارش‌های پیشین^{۴۳} در گزارش‌های رادیولوژی بود، عاملی که بر کیفیت تولید گزارش‌ها تأثیرگذار است. آن‌ها مجموعه داده‌ی جدیدی با نام CXR-PRO ایجاد کردند که این مشکل را در مجموعه داده‌ی MIMIC-CXR برطرف می‌کرد. آن‌ها همچنین CXR-RePai را با استفاده از مجموعه داده‌ی CXR-PRO و یک معماری به‌روزرسانی‌شده به نام ALBEF معرفی شده در مقاله [۹]، بازآموزی کردند و جدیدترین SOTA را برای وظیفه‌ی تولید گزارش‌های رادیولوژی ثبت کردند. آن‌ها همچنین از معیار RadGraph F1 معرفی شده در مقاله [۱۹]، به‌عنوان یک معیار اضافی برای سنجش کامل بودن و دقت موجودیت‌های بالینی^{۴۴} موجود در گزارش تولیدشده با استفاده از مدل RadGraph معرفی شده در مقاله [۵]، بهره بردند.

با ظهور مدل‌های زبانی بزرگ، روش تولید افزوده‌شده با بازیابی (RAG) در پژوهش [۸] معرفی شد که برخی مزایای کلیدی را از طریق بهره‌گیری از منابع دانش خارجی برای تقویت دانش مدل‌های زبانی بزرگ (LLMs) ارائه کرد. با این روش، تولیدات مدل‌های زبانی بزرگ به‌شدت بر دانش واقعی و مستند تکیه دارند که این ویژگی باعث می‌شود کمتر دچار توهم^{۴۵} شوند و خروجی‌های واقعی‌تر و مبتنی بر حقایق تولید کنند. این معماری می‌تواند در سناریوهای مختلف مانند حوزه پزشکی تأثیر مهمی داشته باشد.

⁴⁰Retrieval-based approach

⁴¹State-of-the-Art (SOTA)

⁴²Last Hidden Representations

⁴³Prior report references

⁴⁴Clinical entities

⁴⁵Hallucination

فصل ۳

شرح مسئله

در پروژه حاضر، هدف اصلی توسعه یک سیستم هوشمند است که بتواند عکس پزشکی را به عنوان ورودی بگیرد و پس از طی چند مرحله، یک خروجی به صورت متن بدهد. عکس ورودی تصویر رادیولوژی از قفسه سینه افراد است و متن خروجی باید گزارش مرتبط با آن عکس باشد. شکل ۳-۱ نمایی از وظیفه مدل تولید خودکار گزارش رادیولوژی را نشان می‌دهد. پس ما با یک مسئله تشریح تصویر^۱ روبرو هستیم. گزارشی که برای تصویر ورودی تولید می‌شود باید از دو جنبه کیفیت بالایی داشته باشد. اولین جنبه زبان طبیعی است یعنی گزارش خروجی باید مانند هر متن دیگری از لحاظ گرامری و چینش کلمات یک روند طبیعی را دنبال کند. جنبه دوم پزشکی است یعنی گزارش تولیدی باید ناهنجاری‌های موجود در تصویر را به درستی شناسایی کند و در متن خروجی نشان دهد. جدول (۳-۱) این ناهنجاری‌ها و تعریف هر کدام را نشان می‌دهد.

تولید خودکار گزارش رادیولوژی از تصاویر قفسه سینه یک چالش میان‌رشته‌ای است که در آن نیاز به درک عمیق از تصویر پزشکی و همچنین تولید زبان طبیعی دقیق و تخصصی وجود دارد. تصاویر رادیولوژی به دلیل پیچیدگی‌های بصری، نیازمند تحلیل دقیق توسط رادیولوژیست‌های باتجربه هستند. هدف ما این است که این تحلیل را با کمک مدل‌های یادگیری عمیق و تکنیک‌های پیشرفته در حوزه پردازش زبان طبیعی (NLP) خودکار کنیم تا در شرایطی که نیروی متخصص محدود است یا بار کاری زیاد می‌باشد، از آن به عنوان یک ابزار کمکی استفاده شود.

این مسئله همچنین با چالش‌های خاصی در حوزه یادگیری ماشین روبرو است. تصاویر ورودی از کیفیت‌ها و ویژگی‌های متنوعی برخوردارند و توصیف دقیق آن‌ها در قالب متن نیازمند تطابق محتوای بصری با اطلاعات متنی معتبر است. علاوه بر این، وجود برجستگی‌های محدود و عدم توازن در توزیع بیماری‌ها در مجموعه داده‌های پزشکی، فرایند آموزش مدل را دشوارتر می‌کند. از این رو، طراحی سیستم‌هایی که بتوانند از دانش موجود (مثلاً گزارش‌های گذشته یا اطلاعات زمینه‌ای) برای بهبود عملکرد خود استفاده کنند، یکی از اهداف کلیدی در این پروژه است.

¹Image Captioning

عکس اشعه ایکس قفسه سینه



ورودی

مدل تولید خودکار
گزارش رادیولوژی

خروجی

Findings : The lungs appear hyperexpanded. There is mild increased pulmonary vascular congestion from _____. A small right pleural effusion is likely present with mild right basilar atelectasis. Right base consolidation is not entirely excluded. No significant left pleural effusion or pneumothorax is detected. Suture chain material and scarring in the left upper-to-mid lung zone is not significantly changed. Multiple mediastinal surgical clips are compatible with history of CABG surgery. The cardiac silhouette is top normal in size but unchanged. The mediastinal and hilar contours are within normal limits with moderate tortuosity of the descending thoracic aorta. Lobulation at the apex of the left hemi thorax along the mediastinal border is stable, residual of slowly resolving hematoma.

Impression : 1. Increased mild pulmonary vascular congestion from _____ with small right pleural effusion and right basilar atelectasis. Right basilar opacity may be combination of above but underlying consolidation due to infection is not excluded.

2. Staple suture material and scar in the left upper-to-mid lung.

شکل ۳-۱: نمایی از وظیفه مدل تولید خودکار گزارش رادیولوژی

جدول ۳-۱: ناهنجاری‌های قفسه سینه

نام ناهنجاری	توضیح
Atelectasis	فروریزش جزئی یا کامل بخشی از ریه که باعث کاهش یا عدم وجود هوا در آن ناحیه می‌شود. در عکس قفسه سینه به صورت ناحیه‌ای با تراکم بالا یا کاهش حجم دیده می‌شود.
Cardiomegaly	افزایش اندازه قلب که ممکن است نشانه‌ای از نارسایی قلبی یا بیماری‌های قلبی باشد. در عکس قفسه سینه به صورت گسترش سایه قلب نمایان می‌شود.
Consolidation	پر شدن بافت ریوی با مایع (مانند چرک، خون یا آب) که باعث سفت شدن آن ناحیه می‌شود. معمولاً در اثر عفونت‌هایی مانند ذات‌الریه دیده می‌شود.
Edema	تجمع مایع در ریه‌ها که اغلب به دلیل نارسایی قلبی ایجاد می‌شود. در عکس قفسه سینه به صورت تاری یا نمای بال‌های خفاش در اطراف ریه مرکزی دیده می‌شود.
Pleural Effusion	تجمع مایع اضافی بین لایه‌های پلور (پوشش ریه و دیواره قفسه سینه). به صورت تاری در زاویه دنده‌ای-دیافراگمی یا سطح مایع مشاهده می‌شود.
Pneumonia	عفونت بافت ریوی که باعث التهاب و تراکم می‌شود. در عکس قفسه سینه به صورت کدورت یا سایه‌های منطقه‌ای یا لوبی دیده می‌شود.
Pneumothorax	ورود هوا به فضای پلور که منجر به فروریزش ریه می‌شود. به صورت ناحیه‌ای بدون علامت‌های ریوی و با خط پلور مشخص نمایان می‌شود.
Enlarged Cardiom.	بزرگ شدن سایه قلب یا مدیاستین در عکس قفسه سینه که می‌تواند نشان‌دهنده بزرگی قلب یا بیماری‌های ناحیه مدیاستین باشد.
Lung Lesion	ناحیه‌ای غیرطبیعی در ریه که ممکن است خوش‌خیم یا بدخیم باشد. به صورت ندول یا توده در عکس نمایان می‌شود.
Lung Opacity	هر ناحیه‌ای در ریه که نسبت به حالت طبیعی سفیدتر یا کدرتر دیده شود، به علت‌هایی مانند مایع، عفونت، یا توده.
Pleural Other	هر ناهنجاری پلورال غیر از افیوژن یا پنوموتوراکس، مانند ضخیم‌شدگی، پلاک‌ها یا کلسیفیکاسیون‌ها.
Fracture	شکستگی در استخوان، که معمولاً شکستگی دنده‌ها بوده و به صورت گسستگی در ساختار استخوانی دیده می‌شود.
Support Devices	وسایل پزشکی مانند ضربان‌ساز، لوله تنفسی، کاتتر و لوله‌های قفسه سینه که در عکس قفسه سینه قابل مشاهده‌اند.

فصل ۴

روش پیشنهادی

۴-۱ پردازش و جمع‌آوری داده

برای اجرا و ارزیابی مدل به مجموعه داده برچسب‌دار نیاز داشتیم. این مجموعه داده از لینک پانویس شده^۱ قابل مشاهده است.

جدول ۴-۱: آمار کلی مجموعه داده BioNLP: این مجموعه داده از ترکیب مجموعه داده‌های معروف در حوزه رادیولوژی قفسه سینه تشکیل شده است.

Dataset	Findings Count	Impressions Count
PadChest	101,752	-
BIMCV-COVID19	45,525	-
CheXpert	45,491	181,619
OpenI	3,252	3,628
MIMIC-CXR	148,374	181,166
Total	344,394	366,413





این مجموعه داده مربوط به مسابقه BioNLP-2024 است و ترکیبی از مجموعه داده‌های شناخته شده می‌باشد. شکل اطلاعات کلی این دیتاست را نشان می‌دهد و در جدول ۴-۲ می‌توانید چند نمونه از ردیف‌های این مجموعه داده را مشاهده کنید. همانطور که مشاهده می‌کنید، برخی از ردیف‌ها فقط ستون Impression و برخی دیگر تنها ستون Findings و برخی نیز هر دو در آن‌ها پر شده است.

برای این پروژه، از زیرمجموعه‌ای از این مجموعه داده استفاده شده است. این زیرمجموعه به صورت تصادفی و به تعداد ۱۰۰۰۰ از بین ردیف‌هایی که متعلق به مجموعه داده CheXpert هستند، انتخاب شده است. از این تعداد ردیف، ۹۵۰۰ مورد برای پیکره‌گزارش‌ها^۲ و مابقی برای تست مدل‌های معرفی شده استفاده شده است. دلیل استفاده محدود از این مجموعه داده این است که بتوان آزمایش‌های بیشتری را در زمان کمتر انجام داد و این تعداد برای بررسی مؤثر بودن ایده‌های معرفی شده در این پروژه کافی است. در این پروژه تنها از محتویات ستون Impression برای پیکره‌گزارش‌ها و ارزیابی مدل‌ها استفاده شده است.

^۱<https://huggingface.co/datasets/StanfordAIMI/rrg24-shared-task-bionlp>

^۲Report Corpus

جدول ۴-۲: نمایی از مجموعه داده کارگاه BioNLP

Source	Image	Impression	Findings
CheXpert		DECREASED BIBASILAR VOLUMES	-
PadChest		-	NORMAL CARDIOME- DIASTINAL SILHOUETTE
BIMCV- COVID19		-	NO PLEURAL EFFUSION OR PNEUMOTHORAX
CheXpert		CARDIOMEGALY WITH STABLE RETROCARDIAC OPACITY	-

۴-۱-۱ پیش‌پردازش متن

در بخش پردازش متنی، گزارش‌های رادیولوژی ابتدا به حروف کوچک تبدیل شدند و سپس با استفاده از ماژول پیش‌پردازش BERT از کتابخانه TensorFlow، به رشته‌هایی از شناسه‌های واژگانی (کدهای عددی هر واژه) تبدیل شدند. در این مرحله، همچنین ماسک‌هایی برای مشخص کردن جایگاه واژه‌های واقعی در برابر بخش‌های خالی (پر شده با صفر، به اصطلاح لبه‌گذاری^۳) تولید شد. برخی نشانه‌های خاص، مانند علامت پایان جمله، برای حفظ ساختار یکنواخت، به صورت خنثی در نظر گرفته شدند. در نهایت، تمامی داده‌ها به صورت آرایه‌ای با اندازه (128,1,1) بازآرایی^۴ شدند تا با ساختار ورودی مدل سازگار باشند.

^۳Padding^۴Reshape

۴-۱-۲ پیش پردازش تصویر

در پردازش تصویر، داده‌های خام تصویری (آرایه‌های عددی تصویر اشعه ایکس) ابتدا به نوع عددی شناور (float32) تبدیل و نرمال‌سازی شدند، به گونه‌ای که کوچک‌ترین مقدار به صفر منتقل شد. اگر نوع داده اولیه تصویر از نوع ۸ بیتی بود، بدون تغییر باقی ماند؛ در غیر این صورت، تصویر برای استفاده از کل دامنه ۱۶ بیت مقیاس‌بندی شد. سپس تمامی تصاویر در قالب تصویر خاکستری دوبعدی به فرمت فشرده‌ی PNG با عمق ۸ یا ۱۶ بیت رمزگذاری شدند و به همراه فراداده‌ها (مانند نوع فرمت تصویر) در قالب ساختار استاندارد tf.train.Example ذخیره گردیدند تا برای ورود به مدل آماده باشند. کد مربوط به پیش‌پردازش تصویر و متن در قسمت پیوست قابل مشاهده است.

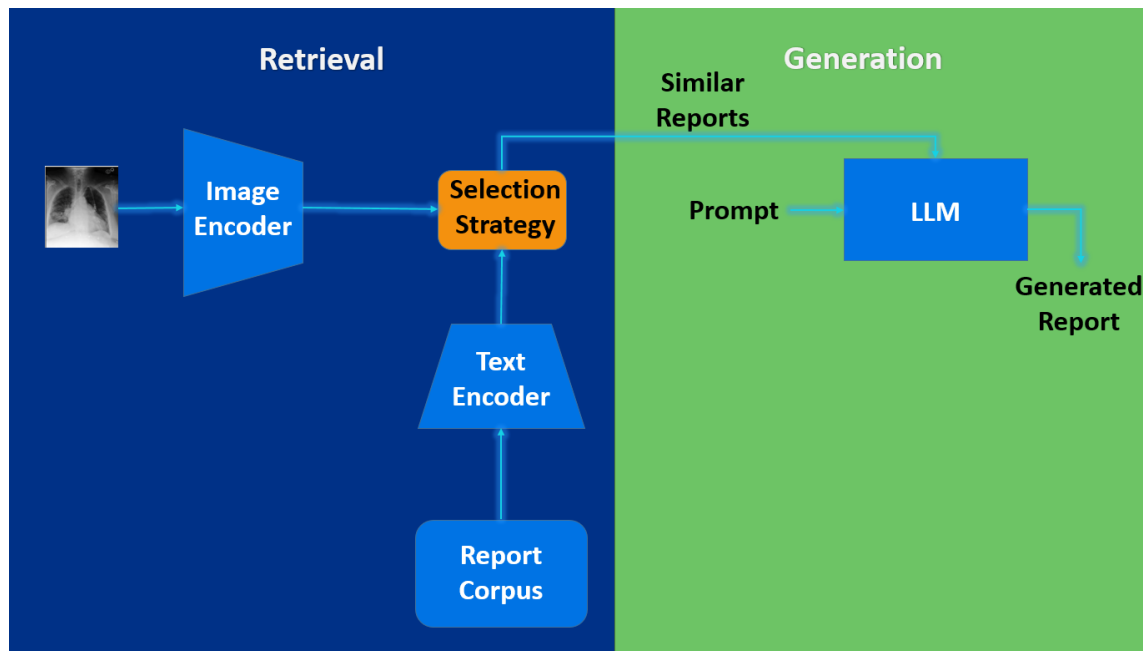
۴-۲ مدل RAG

در مدل RAG^۵ برای دستیابی به یک گزارش نهایی برای هر عکس ورودی، از ساختاری مانند RAG^۶ استفاده شده است. از این مدل به عنوان مدل پایه^۷ استفاده شده است. در این ساختار دو بخش بازیابی و تولید حائز اهمیت است. در بخش بازیابی از یک مدل بینایی-زبان در حوزه تصاویر اشعه ایکس قفسه سینه استفاده می‌شود و در بخش تولید از یک مدل زبانی بزرگ بهره‌برده می‌شود. مدل بینایی-زبان کمک می‌کند گزارش‌های نزدیک به تصویر ورودی شناسایی شود و در پرامپت ورودی به مدل زبانی بزرگ گنجانده شود تا این مدل بتواند خروجی نهایی را براساس این گزارش‌ها خلاصه کند و تحویل دهد. این کار نیازمند یک دیتاست از گزارش‌ها است که از آن به عنوان پیکره گزارش‌ها^۸ یاد می‌شود. در شکل ۴-۱ ساختار کلی سیستم ارائه شده را مشاهده می‌کنید.

۴-۲-۱ ماژول بازیابی

برای بازیابی گزارش‌های مرتبط با یک تصویر از ماژول Q-Former در مدل ELIXR-B [۱۸] استفاده شده است که در بخش مفاهیم اولیه به معرفی این مدل پرداخته شده است. در این روش با استفاده از این مدل،

^۵Report Augmented Generation^۶Retrieval-Augmented Generation^۷Baseline^۸Report Corpus



شکل ۴-۱: ساختار مدل RAG

جانمایی‌های عکس ورودی و تمام گزارشهای موجود در پیکره گزارش بدست می‌آید و با استفاده از تابع شباهت کسینوسی تشابه عکس ورودی با تمام گزارش‌ها محاسبه شده و ۱۰ عدد از مشابه‌ترین گزارش‌ها استخراج می‌شود. از این ۱۰ گزارش نهایتاً به سه گزارش نهایی خواهیم رسید. دستیابی به این ۳ گزارش از طریق ماژول Selection Strategy در شکل انجام می‌شود.

۴-۲-۲ ماژول انتخاب یا Selection Strategy

این سه گزارش به روشهایی مختلفی می‌توانند بدست بیایند. در روش اول صرفاً همان بهترین و مشابه‌ترین گزارش‌ها انتخاب می‌شوند. یعنی از بین ۱۰ گزارش بازیابی شده، ۳ گزارشی که بیشترین تشابه کسینوسی را داشتند انتخاب می‌شوند. اما انتخاب مشابه‌ترین گزارش‌ها از تنوع در آنها اطمینان حاصل نمی‌کند. برای ایجاد تنوع می‌توان از روش‌های دیگری برای انتخاب این سه گزارش استفاده کرد. در روش دوم از الگوریتم خوشه‌بندی K-Means استفاده می‌شود و سپس مراکز خوشه‌ها انتخاب می‌شوند. در روش سوم که به آن انتخاب بیشینه متنوع^۹ گفته می‌شود، ابتدا یک گزارش برای مثال مشابه‌ترین گزارش از بین ۱۰ گزارش انتخاب می‌شود

^۹Maximally Diverse Selection

و سپس از بین گزارش‌های باقیمانده، گزارشی که حداقل فاصله آن تا گزارش‌های انتخاب شده، بیشتر از تمام گزارش‌های باقیمانده باشد، به گزارش‌های انتخاب شده اضافه می‌شود. این دو روش اخیر سعی می‌کنند گزارش‌های متنوع‌تری را خروجی دهند.

۴-۲-۳. ماژول تولید

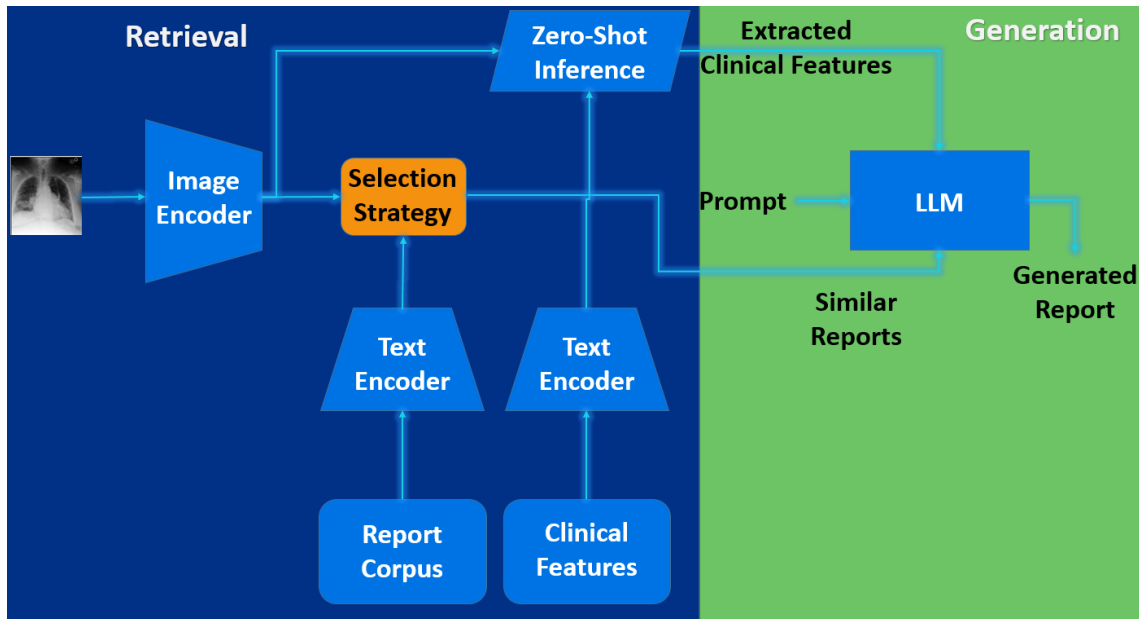
در بخش تولید از مدل زبانی Gemini-2.0-Flash استفاده شده است. به دو دلیل اصلی این ماژول اضافه شده است. اول اینکه گزارش‌های استخراج شده همان بخش impression در مجموعه داده ما هستند و محتویات این بخش معمولاً به طور خلاصه و فاقد ساختاری طبیعی است و کلمات کلیدی پشت هم ظاهر شده‌اند. دلیل دوم این است که در گزارش‌های استخراجی ممکن است برخی موارد به صورت تکراری آمده باشند، برای مثال در هر سه گزارش به یکی از ناهنجاری‌های مربوط به قفسه سینه اشاره شده باشد. به کمک یک مدل زبانی و یک پرامپت مناسب می‌توان از ظاهر شدن چندباره این ناهنجاری‌ها جلوگیری کرد. البته برای بررسی تاثیر مدل‌های زبانی مختلف، در آزمایشی جدا مقایسه این مدل زبانی با مدل Qwen2.5-70B-Instruct نیز صورت گرفت. هرچند باقی آزمایش‌ها همگی با همان مدل Gemini بودند.

۴-۲-۴. چالش‌های مدل RAG

در این مدل با برخی چالش‌ها روبرو هستیم. برای مثال در هریک از گزارش‌های استخراج شده ممکن است مواردی باشد که لزوماً تمام این موارد در عکس ورودی وجود نداشته باشد یا اینکه به برخی ناهنجاری‌ها که در تصویر رادیولوژی حضور دارند، اشاره‌ای صورت نگیرد. این مشکل به این خاطر ایجاد می‌شود که هر گزارش معمولاً به موارد متعددی اشاره می‌کند و لزوماً نمی‌توان دقیقاً گزارشی را که تمام موارد آن تصویر را پوشش دهد پیدا کرد. این مشکل را می‌توان با بزرگ کردن فضای پاسخ تا حدودی حل کرد. بزرگ کردن فضای پاسخ می‌تواند از چند طریق انجام شود. یکی از آن‌ها این است که از تعداد گزارش بیشتری استفاده شود. در حالتی دیگر می‌توان هر گزارش را به جملات تشکیل دهنده آن تجزیه کرد و به جای پیکره گزارش یک پیکره جملات^{۱۰} تشکیل داد. اما در این پروژه از روشی دیگر استفاده شده است که حاصل آن مدل FRAG^{۱۱} است. چالش دیگر وجود برخی ارجاعات به گزارشات قبلی است که مطلوب ما نیست. برای مثال در عبارت

¹⁰Sentence Corpus

¹¹Feature and Report Augmented Generation(FRAG)

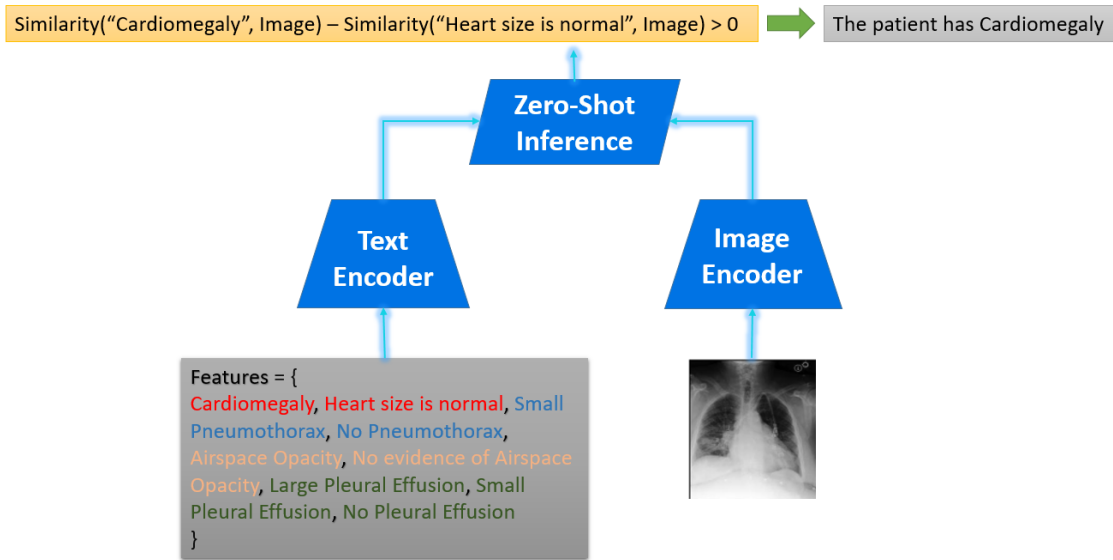


شکل ۴-۲: ساختار مدل FRAG-A: اضافه کردن ماژول استنتاج بدون نمونه

و مقایسه با آن مشهود است. در این پروژه از پرامپت مناسب برای حل این مشکل استفاده شده است و این وظیفه را بر دوش مدل زبانی در بخش انتهایی خواهیم گذاشت.

۴-۳ مدل FRAG-A

اولین مدل از مدل‌های FRAG (Feature and Report Augmented Generation) مدل FRAG-A می‌باشد. در این مدل ما با استفاده از روش استنتاج بدون نمونه، برخی از ناهنجاری‌ها را به صورت جداگانه برای تصویر ورودی پیدا می‌کنیم. در اینجا همچنان از جانمایی‌های مدل ELIXR-B استفاده شده است. به عنوان مثال در این روش برای اینکه مشخص شود ناهنجاری Cardiomegaly در تصویر ورودی وجود دارد یا خیر، embedding دو جمله Cardiomegaly و Heart size is normal را که ویژگی‌های متضاد یکدیگر هستند، بدست آورده و با تصویر ورودی و به روش شباهت کسینوسی شباهت سنجی می‌شود. اختلاف امتیاز شباهت این دو مشخص کننده وجود یا عدم وجود این ناهنجاری خواهد بود. شکل ۴-۲ مدل تغییر یافته را نشان می‌دهد. همچنین شکل ۴-۳ طریقه عملکرد ماژول استنتاج بدون نمونه را نشان می‌دهد.



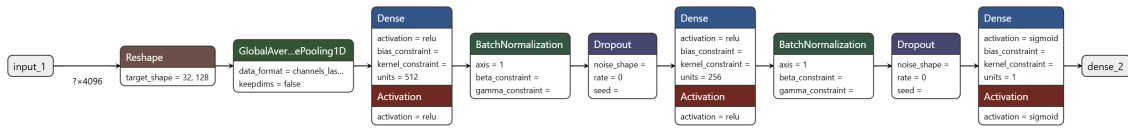
شکل ۴-۳: نمایی از عملکرد ماژول استنتاج بدون نمونه

۴-۳-۱ چالش‌های مدل FRAG-A

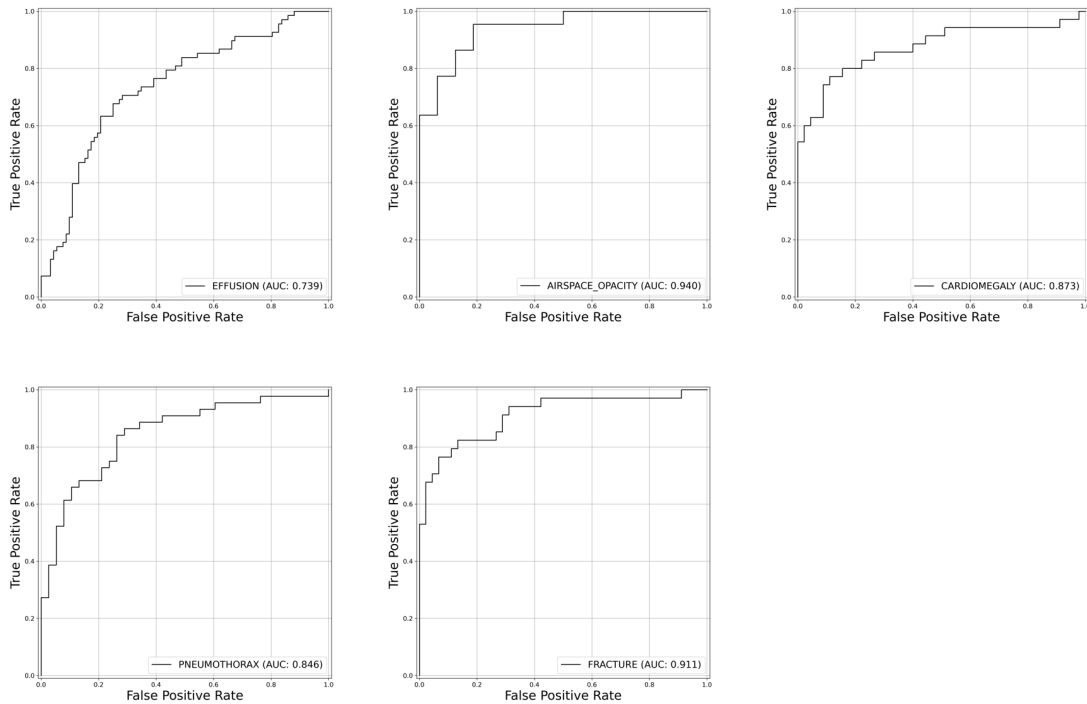
چالش این مدل در واقع در ضعف مدل استخراج کننده جانمایی یا همان ELIXR-B است که در واقع این چالش برای مدل RAG نیز وجود دارد. این باعث می‌شود گزارش‌های غیرمرتبط استخراج شوند یا در بخش استنتاج بدون نمونه ناهنجاری‌ها به درستی شناسایی نشوند. برای حل این مشکل نیاز به آموزش داریم. در این پروژه بخش استنتاج بدون نمونه با یک ماژول دیگر که آموزش دیده است، جایگزین می‌شود.

۴-۴ مدل FRAG-B

برای تقویت قدرت تشخیص ناهنجاری‌ها، به جای روش استنتاج بدون نمونه از چندین دسته‌بند (هر ناهنجاری یک دسته‌بند) استفاده شده است. خروجی این دسته‌بندها به صورت صفر و یکی است و ورودی آن جانمایی‌های از پیش محاسبه شده است. این جانمایی‌ها از همان ماژول Q-Former در مدل ELIXR-B می‌آیند. استفاده از این جانمایی‌های از پیش محاسبه شده کمک می‌کند که بتوانیم با مجموعه داده کوچکتری یک دسته‌بند با دقت مناسب بسازیم. برای آموزش این دسته‌بند از تعدادی از رکوردهای مجموعه داده NIH ChestX-ray14 [۱۶] استفاده شده است. تعداد داده‌های آموزشی و اعتبارسنجی متعلق به هر کدام از ناهنجاری‌ها در جدول ۴-۳ قابل مشاهده است. در این پروژه برای ۵ مورد از ناهنجاری‌ها دسته‌بند در نظر گرفته شده است.



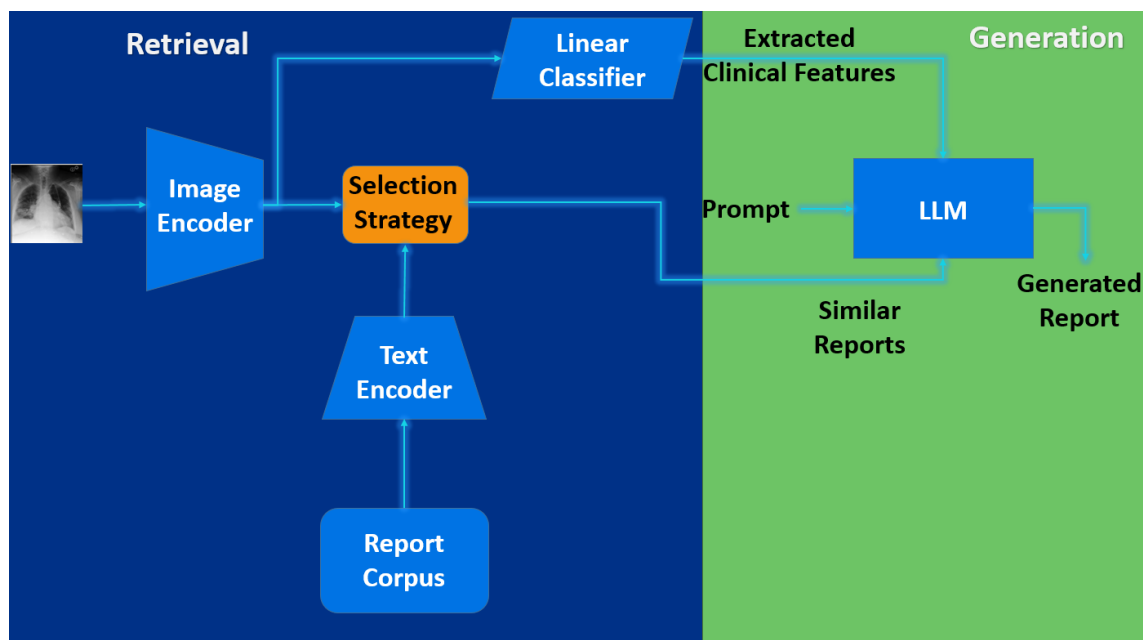
شکل ۴-۴: ساختار مدل دسته‌بند برای تشخیص ناهنجاری‌ها



شکل ۴-۵: نمودار ROC مدل خطی ارائه شده برای تشخیص ناهنجاری‌ها

این کار می‌تواند برای تمام ناهنجاری‌ها انجام شود اما همین تعداد نیز برای بررسی موثر بودن آن کافی خواهد بود. شکل ۴-۴ ساختار این دسته‌بند را نشان می‌دهد. شکل ۴-۵ نیز نتایج این دسته‌بند را بر روی داده تست نشان می‌دهد. با توجه به این نمودار، حد آستانه مناسب برای هر دسته‌بند و با روش Youden's J statistic ۴-۱ تعیین می‌شود. شکل ۴-۶ ساختار بروز شده پس از اضافه کردن دسته‌بند را نشان می‌دهد.

$$\tau^* = \arg \max_{\tau} (TPR(\tau) - FPR(\tau)) \quad (۴-۱)$$



شکل ۴-۶: ساختار مدل FRAG-B: جایگزینی ماژول استنتاج بدون نمونه با دسته‌بند خطی

جدول ۴-۳: تعداد نمونه‌های ناهنجاری‌های مختلف برای آموزش و اعتبارسنجی مدل دسته‌بند خطی

Feature	Training	Validation
EFFUSION	1,440	160
AIRSPACE OPACITY	339	38
CARDIOMEGALY	720	80
PNEUMOTHORAX	738	82
FRACTURE	705	79

فصل ۵

ارزیابی و معیارهای سنجش عملکرد

برای وظایف تولید متن و مقایسه دو متن با یکدیگر معیارهای مختلفی معرفی شده است. در این پروژه به دلیل ارتباط آن با حوزه پزشکی، باید از معیارهای پزشکی نیز در کنار معیارهای زبان طبیعی استفاده شود. برخی از این موارد را در ادامه توضیح خواهیم داد.

۵-۱ معیارهای ارزیابی

۵-۱-۱ امتیاز BLEU

امتیاز BLEU^۱ معیاری است که در ابتدا برای ارزیابی ترجمه ماشینی طراحی شد و بعدها برای تولید گزارش و پاسخ به پرسش‌های تصویری نیز به کار رفت. این معیار سنجش شباهت متن تولیدشده توسط مدل با متن مرجع انسانی است.

برای فهم فرمول این امتیاز ابتدا باید با فرمول دقت n-gram^۲ ۵-۱ آشنا شویم. فرمول ۵-۲ طریقه محاسبه امتیاز BLEU را نشان می‌دهد. عبارت BP^۳ یا همان جریمه کوتاه بودن نیز در فرمول ۵-۳ نشان داده شده است که در آن c طول متن تولیدی و r طول متن مرجع است. معمولاً مقدار n=۴ در نظر گرفته می‌شود. این نمره بین ۰ تا ۱ است و مقدار بالاتر نشان‌دهنده تطابق بهتر با متن مرجع می‌باشد.

$$\text{Precision}(n) = \frac{\text{تعداد } n\text{-gram های مشترک}}{\text{تعداد کل } n\text{-gram ها در متن تولید شده}} \quad (۵-۱)$$

$$\text{BLEU-n} = BP \times \exp \left(\frac{1}{n} \sum_{k=1}^n \log(\text{Precision}(k)) \right) \quad (۵-۲)$$

^۱Bilingual Evaluation Understudy (BLEU)

^۲Precision(n)

^۳Brevity Penalty (BP)

$$BP = \begin{cases} 1 & c \geq r \\ e^{(1-\frac{r}{c})} & c < r \end{cases} \quad (۳-۵)$$

۵-۱-۲ امتیاز ROUGE

همانند امتیاز BLEU، معیار ROUGE^۴ نیز برای سنجش شباهت دو متن مرجع و تولید شده توسط مدل استفاده می‌شود. این امتیاز حالات مختلف دارد که یکی از آنها در فرمول ۴-۵ تحت عنوان ROUGE-n آمده است. حالت دیگر، ROUGE-L است که از طول بزرگترین زیردنباله مشترک بین متن مرجع و تولید شده استفاده می‌کند. فرمول ۵-۵ و ۶-۵ برای بدست آوردن این امتیاز مورد استفاده قرار می‌گیرند. در این روابط، $LCS(X, Y)$ طول بزرگترین زیردنباله مشترک بین متن مرجع X و متن تولید شده Y است. m و n نیز به ترتیب طول متن مرجع X و طول متن تولید شده Y هستند. ضریب β نیز برای تنظیم اهمیت دقت^۵ نسبت به بازخوانی^۶ استفاده می‌شود.

$$ROUGE-n = \frac{\text{تعداد } n\text{-gram های مشترک}}{\text{تعداد کل } n\text{-gram ها در متن مرجع}} \quad (۴-۵)$$

$$ROUGE-L = \frac{(1 + \beta^2) \times R \times P}{R + \beta^2 \times P} \quad (۵-۵)$$

$$R = \frac{LCS(X, Y)}{m}, \quad P = \frac{LCS(X, Y)}{n} \quad (۶-۵)$$

^۴Recall-Oriented Understudy for Gisting Evaluation

^۵Precision

^۶Recall

۵-۱-۳ معیار METEOR

معیار METEOR^۷ یک معیار ارزیابی برای سنجش کیفیت متن تولیدشده توسط مدل زبانی است که در ابتدا برای ترجمه ماشینی طراحی شد. برخلاف معیارهایی مانند BLEU که عمدتاً بر اساس دقت هستند، METEOR سعی می‌کند معنا و روانی جمله را بهتر در نظر بگیرد.

امتیاز METEOR بر پایه ترکیبی از دقت و بازخوانی تعریف می‌شود، که در آن بازخوانی نشان می‌دهد چند درصد از 1-gram های موجود در متن مرجع در متن تولیدی نیز وجود دارند (فرمول ۵-۷). دقت درصدی از 1-gram های تولیدی است که در متن مرجع نیز وجود دارند (فرمول ۵-۸). در این معیار از مفهومی با عنوان جریمه^۸ نیز استفاده می‌شود. این پارامتر برای جلوگیری از نمره‌دهی بالا به جملاتی که فقط شامل کلمات درست ولی با ترتیب نادرست هستند، مورد استفاده قرار می‌گیرد. این جریمه بر اساس تعداد قطعه‌ها^۹ محاسبه می‌شود (فرمول ۵-۹). یک قطعه گروهی از 1-gram های مجاور است که در هر دو متن (مرجع و تولیدی) به صورت پشت سرهم آمده‌اند. اگر متن تولیدی ترتیب واژگان را به درستی رعایت نکرده باشد، تعداد قطعات زیاد خواهد شد و در نتیجه جریمه افزایش می‌یابد. در نهایت فرمول معیار METEOR از ترکیب این پارامترها بدست می‌آید (فرمول ۵-۱۰). نمره نهایی METEOR در بازه‌ی صفر تا یک قرار دارد. مقدار بالاتر به معنای تطابق بهتر با متن مرجع از نظر واژگان، ترتیب و معنا است.

$$R = \frac{\text{تعداد 1-gram های مشترک}}{\text{تعداد کل 1-gram ها در متن مرجع}} \quad (۷-۵)$$

$$P = \frac{\text{تعداد 1-gram های مشترک}}{\text{تعداد کل 1-gram ها در متن تولید شده}} \quad (۸-۵)$$

^۷Metric for Evaluation of Translation with Explicit ORDERing (METEOR)

^۸Penalty

^۹Chunks

$$\text{Penalty} = \frac{1}{4} \times \left(\frac{\text{تعداد قطعات}}{\text{تعداد 1-gram های مشترک}} \right)^3 \quad (۹-۵)$$

$$\text{METEOR} = \frac{10 \times P \times R}{R + 9 \times P} \times (1 - \text{Penalty}) \quad (۱۰-۵)$$

۵-۱-۴ امتیاز RadGraph F1

این معیار یک معیار نوین برای ارزیابی میزان همپوشانی بین موجودیت‌های بالینی و روابط استخراج شده از گزارش‌های رادیولوژی است [۱۹]. این معیار به‌طور خاص برای درک ساختاری و معنایی دقیق‌تر از محتوای بالینی طراحی شده است و به جای مقایسه سطحی متون، به مقایسه گراف‌های معنایی حاصل از آن‌ها می‌پردازد. محاسبه نمره RadGraph F1 به صورت زیر انجام می‌شود: ابتدا، مدل RadGraph هر دو گزارش (گزارش تولیدشده توسط مدل و گزارش مرجع انسانی) را به نمایش گرافی تبدیل می‌کند؛ در این گراف‌ها، موجودیت‌های بالینی به عنوان گره‌ها (nodes) و روابط بین آن‌ها به عنوان یال‌ها (edges) مدل‌سازی می‌شوند.

در مرحله‌ی دوم، تعداد گره‌هایی که بر اساس متن موجودیت بالینی و برچسب آن (نوع موجودیت) در هر دو گراف هم‌خوانی دارند، محاسبه می‌شود.

در مرحله‌ی سوم، تعداد یال‌هایی که با در نظر گرفتن موجودیت‌های آغاز و پایان رابطه و نوع رابطه (برچسب) در دو گراف مشابه هستند، تعیین می‌گردد.

در نهایت، نمره F1 به صورت جداگانه برای موجودیت‌ها و روابط محاسبه می‌شود، و نمره RadGraph F1 برای یک جفت گزارش برابر با میانگین این دو نمره است.

۵-۱-۵ استفاده از مدل (S_{emb}, Micro F1, Macro F1) CheXbert

برای ارزیابی کیفیت مفهومی یا همان پزشکی گزارش‌های رادیولوژی تولیدشده، از مدل CheXbert استفاده کردیم که نسخه‌ای بهبودیافته از CheXpert labeler [۴] است و با استفاده از معماری BERT برای استخراج برچسب‌های بالینی از متون گزارش آموزش دیده است [۱۵].

این مدل برای هر گزارش رادیولوژی یک بازنمایی یا embedding تولید می‌کند. سپس با استفاده از این بازنمایی و اضافه کردن سر^{۱۰} به یک بردار دودویی ۱۴ بعدی شامل وضعیت وجود یا عدم وجود بیماری‌های رایج قفسه سینه می‌رسد. (برای مثال: Consolidation, Edema, Cardiomegaly, ...). سپس با استفاده از این بردارهای مفهومی، شباهت بین گزارش‌ها از سه منظر محاسبه می‌شود. اولین معیار که S_{emb} نام دارد، از طریق محاسبه شباهت کسینوسی بین بازنمایی اولیه گزارش مرجع و گزارش تولیدی بدست می‌آید. دو معیار دیگر بر روی بردار دودویی ۱۴ بعدی تعریف می‌شوند. امتیاز Micro F1 دقت در سطح برچسب که تمام پیش‌بینی‌ها را بدون توجه به نوع بیماری با هم مقایسه می‌کند. امتیاز Macro F1 میانگین F1 برای هر بیماری به‌صورت جداگانه که تأکید یکسانی بر تمام بیماری‌ها دارد، حتی بیماری‌هایی با شیوع کمتر. فرمول‌های ۵-۱۱، ۵-۱۲، ۵-۱۳، ۵-۱۴ و ۵-۱۵ طریقه محاسبه این دو معیار را نشان می‌دهند. در این روابط C به تعداد ناهنجاری‌ها یا همان کلاس‌ها اشاره دارد که در اینجا تعدادشان ۱۴ است. عبارت TP, FP, FN نیز به ترتیب به معنای تعداد موارد درست مثبت برای برچسب، تعداد مثبت‌های اشتباه پیش‌بینی شده و تعداد مواردی که باید مثبت پیش‌بینی می‌شدند ولی مدل آن‌ها را از دست داده است، می‌باشند. زیروند i در این عبارات نشان‌دهنده این است که آن عبارت برای کدام یک از ۱۴ بیماری می‌باشد.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۵-۱۱)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (۵-۱۲)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۵-۱۳)$$

¹⁰Head

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^C \text{F1}_i \quad (۱۴-۵)$$

$$\text{Micro-F1} = \frac{2 \times \sum_{i=1}^C TP_i}{2 \times \sum_{i=1}^C TP_i + \sum_{i=1}^C FP_i + \sum_{i=1}^C FN_i} \quad (۱۵-۵)$$

۵-۲ نتایج ارزیابی

نتایج ارزیابی مدل‌های مختلف معرفی شده در بخش‌های قبل در جدول ۵-۱ قابل مشاهده است. در این آزمایش‌ها از همان روش بهترین گزارشها در ماژول انتخاب یا Selection Strategy و از مدل Gemini برای ماژول تولید استفاده شده است.

جدول ۵-۱: نتایج ارزیابی مدل‌های مختلف ارائه شده

Model	BLEU-1	BLEU-3	ROUGE-L	METEOR	S_{emb}	Macro F1	Micro F1
RAG	0.06	0.02	0.06	0.09	0.18	0.18	0.41
FRAG-A	0.06	0.01	0.06	0.14	0.28	0.26	0.50
FRAG-B	0.06	0.01	0.05	0.14	0.24	0.22	0.43

در آزمایشی که در ماژول بازیابی انجام شد، سه روش مورد مقایسه قرار گرفتند. در روش اول صرفاً سه مشابه‌ترین گزارشها انتخاب شدند. در روش دوم پس از اجرای الگوریتم خوشه‌بندی K-Means، مراکز خوشه‌ها انتخاب شدند. در روش سوم که به آن انتخاب بیشینه متنوع^{۱۱} گفته می‌شود، ابتدا یک گزارش برای مثال مشابه‌ترین گزارش از بین ۱۰ گزارش انتخاب می‌شود و سپس از بین گزارش‌های باقیمانده، گزارشی که حداقل فاصله آن تا گزارش‌های انتخاب شده، بیشتر از تمام گزارش‌های باقیمانده باشد، انتخاب می‌شود. در تمام این روش‌ها نهایتاً سه گزارش به عنوان خروجی انتخاب می‌شوند. نتایج این آزمایش‌ها در جدول ۵-۲ قابل مشاهده است.

¹¹Maximally Diverse Selection

جدول ۵-۲: مقایسه روش‌های مختلف در مازول بازیابی

Model	BLEU-1	BLEU-3	ROUGE-L	METEOR	S_{emb}	Macro F1	Micro F1
Top 3	0.12	0.02	0.19	0.09	0.18	0.18	0.41
K-means Clustering	0.12	0.02	0.19	0.09	0.18	0.18	0.40
Maximally Diverse	0.11	0.02	0.18	0.09	0.21	0.20	0.43

جدول ۵-۳ نشان‌دهنده نتایج آزمایش با استفاده از دو مدل زبانی مختلف است. این دو آزمایش هر دو با استفاده از مدل FRAG-A که شامل مازول استنتاج بدون نمونه^{۱۲} و مدل زبانی می‌شود، اجرا شدند و تنها بخش مدل زبانی آن متفاوت است.

جدول ۵-۳: مقایسه دو مدل زبانی بزرگ

Model	BLEU-1	BLEU-3	ROUGE-L	METEOR	S_{emb}	Macro F1	Micro F1
Gemini	0.06	0.01	0.06	0.14	0.28	0.26	0.50
Qwen2.5	0.06	0.01	0.07	0.13	0.25	0.25	0.45

در نهایت در آزمایشی که نتایج آن در جدول ۵-۴ آمده است، دو پرامپت مختلف مورد آزمایش قرار گرفتند. در هر دو مورد از مدل Gemini استفاده شده است. در پرامپت اول صرفاً به مدل زبانی بهترین گزارشها و ناهنجاری‌های استخراج شده داده می‌شوند و از آن خواسته می‌شود بر اساس این موارد گزارش نهایی را تولید کند. در پرامپت دوم برخی از جزئیات نیز به آن گوشزد می‌شود. برای مثال نباید ناهنجاری‌های متناقض را در گزارش نهایی بیاورد و همچنین نباید به گزارش‌های گذشته اشاره کند. متن دقیق این دو پرامپت را می‌توانید در بخش پیوست مشاهده کنید.

جدول ۵-۴: مقایسه دو پرامپت مختلف

Model	BLEU-1	BLEU-3	ROUGE-L	METEOR	S_{emb}	Macro F1	Micro F1
Prompt 1	0.06	0.01	0.06	0.14	0.26	0.27	0.49
Prompt 2	0.06	0.01	0.06	0.14	0.28	0.26	0.50

¹²Zero-Shot Inference(ZSI)

۵-۲-۱ تحلیل نتایج ارزیابی

در آزمایش اول که مدل‌های مختلف مورد بررسی قرار گرفتند، نشان می‌دهد بهترین عملکرد در معیارهای پزشکی برای مدلی است که از ماژول استنتاج بدون نمونه استفاده می‌کند. دلیل اینکه در حالات استفاده از مدل زبانی بزرگ معیارهای زبان طبیعی ضعیفتر هستند، در این است که گزارشهای مرجع همان محتوای ستون Impression هستند و این فیلد دارای گزارشهایی است که فاقد جریان طبیعی در زبان هستند و تنها شامل کلمات کلیدی می‌شوند که پشت سر هم آمده‌اند. این آزمایش نشان داده است که اضافه کردن ماژول دسته‌بند خطی کمکی در بهبود معیارهای پزشکی نکرده است و حتی باعث تضعیف آن شده است. این احتمالاً به این دلیل رخ داده است که مجموعه داده‌ای که این دسته‌بند بر روی آن آموزش دیده است (NIH)، منبع متفاوتی از مجموعه داده‌ای دارند که در ماژول ارزیابی از آن استفاده می‌شود (CheXpert). دلیل دیگر می‌تواند تعداد کم داده‌های تست باشد که نتوانسته‌اند نماینده خوبی از لحاظ تنوع باشند و مدل ما بر روی آن‌ها بیش‌برازش^{۱۳} شده باشد.

نتایج آزمایش دوم نشان می‌دهد که روش انتخاب بیشینه متنوع عملکرد بهتری نسبت به دو روش دیگر دارد و این نشان می‌دهد ایجاد تنوع در گزارش‌های تولیدی تاثیرگذار است اما باید با روش مناسب انجام شود. آزمایش سوم نشان می‌دهد مدل Gemini بهتر عمل کرده است و ایجاد تغییراتی در این بخش می‌تواند عملکرد کلی مدل را بهبود بخشد. در نهایت، آزمایش چهارم نشان می‌دهد که مدل‌های زبانی بزرگ خودشان تشخیص می‌دهند که در گزارش نهایی تولید شده باید برخی موارد رعایت شود و آنها خودشان این کار را می‌کنند و طولانی کردن پرامپت در این آزمایش به طور خاص کمکی به بهبود عملکرد نکرده است.

¹³Overfitting

فصل ۶

نتیجه‌گیری و کارهای آینده

۶-۱ نتیجه‌گیری

در این پروژه، برخی از ایده‌ها برای بهبود گزارش‌های تولیدی به کار برده شد. نشان داده شد که تشخیص جداگانه برخی ناهنجاری‌ها و اضافه کردن آن به گزارش‌های بازیابی شده می‌تواند باعث بهبود عملکرد مدل شود. این کار به کمک دو روش استنتاج بدون نمونه و دسته‌بند خطی انجام شد. علاوه بر آن، تاثیر ایجاد تنوع در مازول بازیابی و گزارش‌های بازیابی شده مورد بررسی قرار گرفت و روش انتخاب بیشینه متنوع از روشهای دیگر عملکرد بهتری داشت. در آزمایشی جدا نشان داده شد که مدل زبانی که در مرحله آخر استفاده می‌شود نیز تاثیرگذار است. در نهایت دو پرامپت مختلف مورد آزمایش قرار گرفت اما تاثیر چندانی روی نتیجه نهایی نداشت. در این پروژه در راستای بهبود عملکرد مدل پایه کاری انجام نشد و فرض بر این بود که مدل پایه عملکرد خوبی دارد، هرچند در این پروژه، این مدل پایه جای بهبود و تقویت داشت.

۶-۲ کارهای آینده

۶-۲-۱ استفاده بهتر از متخصصان بالینی

با توجه به اینکه هدف نهایی ARRГ تولید گزارش‌هایی در سطح دقت گزارش‌های متخصصان رادیولوژی است، بهره‌گیری مستقیم از دانش و بازخورد این متخصصان در فرآیند ارزیابی مدل‌ها اهمیت بالایی دارد. اگرچه برخی پژوهش‌ها از ارزیابی کیفی توسط پزشکان استفاده کرده‌اند، فقدان یک چارچوب استاندارد برای ارزیابی‌های کیفی، مقایسه منصفانه بین مدل‌ها را دشوار کرده است. تدوین یک پروتکل استاندارد برای این ارزیابی‌ها می‌تواند به شفاف‌سازی و اعتبارسنجی بهتر دستاوردها منجر شود.

۶-۲-۲ یادگیری تقویتی با بازخورد انسانی

در حالی که یادگیری تقویتی در حوزه ARRГ در حال گسترش است، استفاده از بازخورد انسانی در تعریف تابع پاداش هنوز مورد استفاده قرار نگرفته است. یادگیری پاداش از بازخورد متخصصان بالینی، می‌تواند به بهبود ملاحظات مهمی مانند ایمنی، اخلاق و کیفیت بالینی کمک کند و در عین حال هزینه آموزش مدل‌ها را کاهش دهد.

۶-۲-۳ بهبود معیارهای ارزیابی کمی

ارزیابی کیفی هرچند ارزشمند است، اما انجام آن برای تمامی موارد بسیار پرهزینه و زمان‌بر است. به همین دلیل، نیاز به معیارهای کمی استاندارد برای سنجش دقت و درستی محتوای گزارش‌ها وجود دارد. بیشتر پژوهش‌ها از معیارهای سنتی NLP مانند BLEU و ROUGE استفاده کرده‌اند، اما این معیارها در حوزه پزشکی ناکافی هستند. استفاده از معیارهایی نظیر BERTScore که بر پایه شباهت معنایی و نه صرفاً تطابق واژگانی عمل می‌کنند، می‌تواند رویکرد ارزیابی را بهبود دهد.

۶-۲-۴ بهره‌گیری از مدل‌های زبانی بزرگ

با وجود پیشرفت‌های چشمگیر در مدل‌های زبانی، استفاده از مدل‌های نسل جدید مانند LLaMA یا Falcon در ARRg بسیار محدود بوده است. بیشتر مطالعات به GPT-2 یا رابط وب ChatGPT بسنده کرده‌اند. در حالی که مدل‌های متن‌باز جدید با میلیاردها پارامتر اکنون در دسترس جامعه علمی هستند، استفاده از آن‌ها می‌تواند موجب جهش چشمگیری در کیفیت گزارش‌های تولیدی شود.

۶-۲-۵ استفاده از مجموعه داده‌های جدید

بیشتر مطالعات در این حوزه بر دو مجموعه داده IU-Xray و MIMIC-CXR متمرکز بوده‌اند که هر دو از تصویر قفسه سینه استفاده می‌کنند. با این حال، مجموعه داده‌هایی مانند PadChest و مجموعه داده سه‌بعدی جدید بر پایه سی‌تی‌اسکن می‌توانند به افزایش تنوع و قدرت تعمیم‌پذیری مدل‌ها کمک کنند و مسیرهای پژوهشی جدیدی را باز کنند.

کتاب نامه

- [1] Chen, Z., Song, Y., Chang, T.-H., and Wan, X. Generating radiology reports via memory-driven transformer, 2022.
- [2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [3] Endo, M., Krishnan, R., Krishna, V., Ng, A. Y., and Rajpurkar, P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. in *Proceedings of Machine Learning for Health* (04 Dec 2021), volume 158 of *Proceedings of Machine Learning Research*, PMLR, pp. 209–219.
- [4] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Illcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.
- [5] Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., Chambon, P., Zhang, Y., Lungren, M. P., Ng, A. Y., Langlotz, C. P., and Rajpurkar, P. Radgraph: Extracting clinical entities and relations from radiology reports, 2021.
- [6] Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., ying Deng, C., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.
- [7] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

- [8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
- [9] Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation, 2021.
- [10] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., and Weyrich, M. A survey on long short-term memory networks for time series prediction. *Procedia CIRP* 99 (2021), 650–655. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.
- [11] Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P., and Jurafsky, D. Improving factual completeness and consistency of image-to-text radiology report generation, 2021.
- [12] Ramesh, V., Chi, N. A., and Rajpurkar, P. Improving radiology report generation systems by removing hallucinated references to non-existent priors, 2022.
- [13] Sherstinsky, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.
- [14] Sloan, P., Clatworthy, P., Simpson, E., and Mirmehdi, M. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering* (2024).
- [15] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., and Lungren, M. P. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.
- [16] Summers, R. Nih chest x-ray dataset of 14 common thorax disease categories. *NIH Clinical Center: Bethesda, MD, USA* (2019).
- [17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.
- [18] Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.-H., Kiraly, A., Kazemzadeh, S., Melamed, Z., Park, J., Strachan, P., Liu, Y., Lau, C., Singh, P., Chen, C., Etemadi, M., Kalidindi, S. R., Matias, Y., Chou, K., Corrado, G. S., Shetty, S., Tse, D., Prabhakara, S., Golden, D., Pilgrim, R., Eswaran, K., and Sellergren, A. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders, 2023.

- [19] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Lee, H. M. H., Abad, Z. S. H., Ng, A. Y., Langlotz, C. P., Venugopal, V. K., and Rajpurkar, P. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns* 4, 9 (2023), 100802.
- [20] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert, 2020.

پیوست ۱ - آماده سازی داده ها

تمامی کد های پروژه و فایل های Latex در لینک زیر موجود است:

To be done!

آماده سازی گزارش ها برای ماژول بازیابی

```
dataset = load_dataset("StanfordAIMI/rrg24-shared-task-bionlp", split='train', streaming=True,
                        cache_dir='./radiology')
reports = []
for example in dataset:
    if example['source'] == 'CheXpert' and len(example['impression'].strip()) > 0:
        reports.append(example['impression'])
        if len(reports) == 10000:
            break

# convert to Json
json_str = json.dumps(reports)

# Store the JSON data in a file
with open("all_reports.json", "w") as file:
    json.dump(json_str, file)
```

پیش پردازش ورودی متن و عکس برای مدل Elixr-b

```
# Helper function for tokenizing text input
def bert_tokenize(text):
    """Tokenizes input text and returns token IDs and padding masks."""
    preprocessor = tf_hub.KerasLayer(
        "https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3")
    out = preprocessor(tf.constant([text.lower()]))
    ids = out['input_word_ids'].numpy().astype(np.int32)
    masks = out['input_mask'].numpy().astype(np.float32)
    paddings = 1.0 - masks
    end_token_idx = ids == 102
    ids[end_token_idx] = 0
    paddings[end_token_idx] = 1.0
    ids = np.expand_dims(ids, axis=1)
    paddings = np.expand_dims(paddings, axis=1)
    assert ids.shape == (1, 1, 128)
    assert paddings.shape == (1, 1, 128)
    return ids, paddings

# Helper function for processing image data
def png_to_tfexample(image_array: np.ndarray) -> tf.train.Example:
```



```

"""Creates a tf.train.Example from a NumPy array."""
# Convert the image to float32 and shift the minimum value to zero
image = image_array.astype(np.float32)
image -= image.min()

if image_array.dtype == np.uint8:
    # For uint8 images, no rescaling is needed
    pixel_array = image.astype(np.uint8)
    bitdepth = 8
else:
    # For other data types, scale image to use the full 16-bit range
    max_val = image.max()
    if max_val > 0:
        image *= 65535 / max_val # Scale to 16-bit range
    pixel_array = image.astype(np.uint16)
    bitdepth = 16

# Ensure the array is 2-D (grayscale image)
if pixel_array.ndim != 2:
    raise ValueError(f'Array must be 2-D. Actual dimensions: {pixel_array.ndim}')

# Encode the array as a PNG image
output = io.BytesIO()
png.Writer(
    width=pixel_array.shape[1],
    height=pixel_array.shape[0],
    greyscale=True,
    bitdepth=bitdepth
).write(output, pixel_array.tolist())
png_bytes = output.getvalue()

# Create a tf.train.Example and assign the features
example = tf.train.Example()
features = example.features.feature
features['image/encoded'].bytes_list.value.append(png_bytes)
features['image/format'].bytes_list.value.append(b'png')

return example

```

محاسبه بازنمایی گزارش‌ها به کمک مدل Elixr-b

```

snapshot_download(repo_id="google/cxr-foundation", local_dir='/content/hf',
                  allow_patterns=['elixr-c-v2-pooled/*', 'pax-elixr-b-text/*'])
for i in tqdm(range(10000)):
    # Run QFormer with text only.
    # Initialize image input with zeros
    tokens, paddings = bert_tokenize(reports[i])
    qformer_input = {
        'image_feature': np.zeros([1, 8, 8, 1376], dtype=np.float32).tolist(),
        'ids': tokens.tolist(),
        'paddings': paddings.tolist(),
    }

    if 'qformer_model' not in locals():
        qformer_model = tf.saved_model.load(
            "/content/drive/MyDrive/FinalProject/CXR_Foundation/hf/pax-elixr-b-text"
        )

    qformer_output = qformer_model.signatures['serving_default'](**qformer_input)
    text_embeddings = qformer_output['contrastive_txt_emb']
    embeddings.append(text_embeddings)

text_embeddings = np.array(embeddings)

```

پیوست ۲ - ساخت مدل و پارامترها

پرامپت‌های استفاده شده برای مدل زبانی بزرگ

```
PROMPT1 = """Assume that you are a radiologist. The following are top 3 similar reports to a chest xray image: \n{top3_reports_str}. We have also some clinical facts related to the image as follows: \n{clinical_facts_str}. You should generate the final report. Try to mimic the style from similar reports and include the facts from both similar reports and clinical facts. Your generated report can contain maximum of 6 sentences. Your output should be in json format with a key of 'report' which contains the final report."""

PROMPT2 = """Assume that you are a radiologist. The following are top 3 similar reports to a chest xray image: \n{top3_reports_str}. We have also some clinical facts related to the image as follows: \n{clinical_facts_str}. You should generate the final report. Try to mimic the style from similar reports and include the facts from both similar reports and clinical facts. Try to avoid including both sides of opposite facts, for example 'cardiomegaly' and 'heart size is normal' do not appear at the same time. Other examples are 'no evidence of pleural effusion' and 'presence of pleural effusion', 'no pneumothorax' and 'pneumothorax is seen', etc. Remove any information not directly observable from the current imaging study. For instance, remove any patient demographic data, past medical history, or comparison to prior images or studies. The generated report should not reference any changes based on prior images, studies, or external knowledge about the patient. Rewrite such comparisons as a status observation based only on the current image or study. Your generated report can contain maximum of 6 sentences. Your output should be in json format with a key of 'report' which contains the final report."""
```

مدل استفاده شده برای دسته‌بند خطی

```
def create_model(heads,
                 token_num,
                 embeddings_size,
                 learning_rate=0.1,
                 end_lr_factor=1.0,
                 dropout=0.0,
                 decay_steps=1000,
                 loss_weights=None,
                 hidden_layer_sizes=[512, 256],
                 weight_decay=0.0,
                 seed=None) -> tf.keras.Model:
    """
    Creates linear probe or multilayer perceptron using LARS + cosine decay.
    """
    inputs = tf.keras.Input(shape=(token_num * embeddings_size,))
    inputs_reshape = tf.keras.layers.Reshape((token_num, embeddings_size))(inputs)
```

```

inputs_pooled = tf.keras.layers.GlobalAveragePooling1D(data_format='channels_last')(inputs_reshape)
hidden = inputs_pooled
# If no hidden_layer_sizes are provided, model will be a linear probe.
for size in hidden_layer_sizes:
    hidden = tf.keras.layers.Dense(
        size,
        activation='relu',
        kernel_initializer=tf.keras.initializers.HeUniform(seed=seed),
        kernel_regularizer=tf.keras.regularizers.l2(l2=weight_decay),
        bias_regularizer=tf.keras.regularizers.l2(l2=weight_decay))(
        hidden)
    hidden = tf.keras.layers.BatchNormalization()(hidden)
    hidden = tf.keras.layers.Dropout(dropout, seed=seed)(hidden)

output = tf.keras.layers.Dense(
    units=1, # Single head for binary classification
    activation='sigmoid',
    kernel_initializer=tf.keras.initializers.HeUniform(seed=seed)
)(hidden)

model = tf.keras.Model(inputs, output)
learning_rate_fn = tf.keras.experimental.CosineDecay(
    tf.cast(learning_rate, tf.float32),
    tf.cast(decay_steps, tf.float32),
    alpha=tf.cast(end_lr_factor, tf.float32))
model.compile(
    optimizer=tfm.optimization.lars.LARS(
        learning_rate=learning_rate_fn),
    loss='binary_crossentropy',
    weighted_metrics=[
        tf.keras.metrics.FalsePositives(),
        tf.keras.metrics.FalseNegatives(),
        tf.keras.metrics.TruePositives(),
        tf.keras.metrics.TrueNegatives(),
        tf.keras.metrics.AUC(),
        tf.keras.metrics.AUC(curve='PR', name='auc_pr')])
return model

```

آموزش دسته‌بند خطی

```

class AUCLoggerAndEarlyStopping(tf.keras.callbacks.Callback):
    def __init__(self, threshold):
        super().__init__()
        self.threshold = threshold

    def on_epoch_end(self, epoch, logs=None):
        train_auc = logs.get('auc')
        val_auc = logs.get('val_auc')

        if (train_auc - val_auc) > self.threshold:
            print(f'Stopping training: AUC difference
                  {abs(train_auc - val_auc):.4f} exceeded threshold {self.threshold}')
            self.model.stop_training = True

TOKEN_NUM = 32
EMBEDDINGS_SIZE = 128

# Prepare the training and validation datasets using embeddings and diagnosis labels
training_data = create_tf_dataset_from_embeddings(
    embeddings=df_train["embeddings"].values,
    labels=df_train[DIAGNOSIS].values,
    embeddings_size=TOKEN_NUM * EMBEDDINGS_SIZE)

```

```
validation_data = create_tf_dataset_from_embeddings(  
    embeddings=df_validate["embeddings"].values,  
    labels=df_validate[DIAGNOSIS].values,  
    embeddings_size=TOKEN_NUM * EMBEDDINGS_SIZE)  
  
model = create_model(  
    [DIAGNOSIS],  
    token_num=TOKEN_NUM,  
    embeddings_size = EMBEDDINGS_SIZE,  
)  
  
threshold = 0.095 # Set your desired threshold  
auc_logger_and_early_stopping = AUCLoggerAndEarlyStopping(threshold)  
  
model.fit(  
    x=training_data.batch(512).prefetch(tf.data.AUTOTUNE).cache(),  
    validation_data=validation_data.batch(1).cache(),  
    epochs=100,  
    callbacks=[auc_logger_and_early_stopping]  
)
```

پیوست ۳ - ارزیابی مدل‌ها

توابع استفاده شده برای محاسبه معیارهای ارزیابی مدل

```
# This tokenizer performs the following steps:
# split standard contractions, e.g. don't -> do n't and they'll -> they 'll
# treat most punctuation characters as separate tokens
# split off commas and single quotes, when followed by whitespace
# separate periods that appear at the end of line
def tokenize(s):
    return TreebankWordTokenizer().tokenize(s)

# Cumulative scores of bluee, n gram scores.
# Match ngrams from candidate to n-grams in reference text.
# Regardless of word order.
def get_bleu(query, groundtruth):
    # Initialize smoothing function
    smooth = SmoothingFunction().method4
    reference = [tokenize(groundtruth)]
    candidate = tokenize(query)
    bleu_1 = sentence_bleu(
        reference, candidate, weights=(1, 0, 0, 0), smoothing_function=smooth)
    bleu_2 = sentence_bleu(
        reference, candidate, weights=(0.5, 0.5, 0, 0), smoothing_function=smooth)
    bleu_3 = sentence_bleu(
        reference, candidate, weights=(0.33, 0.33, 0.33), smoothing_function=smooth)
    bleu_4 = sentence_bleu(
        reference, candidate, weights=(0.25, 0.25, 0.25, 0.25), smoothing_function=smooth)
    return {"bleu_1": bleu_1, "bleu_2": bleu_2, "bleu_3": bleu_3, "bleu_4": bleu_4}

# Rouge has different variants, the recommended one is rouge-1,
# which stands calculates precision, recall and F1-measure based
# on the length of the longest common subsequence.
# The desired metrics result is the F1-measure.
def get_rouge(query, groundtruth, variant="rouge-1", measure="f"):
    rouge = Rouge()
    rouge_scores = rouge.get_scores(query, groundtruth)
    return rouge_scores[0][variant][measure]

# Meteor evaluates the caption by first calculating bleu_1
# between generated and ground truth to find matching results.
# Computes harmonic mean.
def get_meteor(query, groundtruth):
    hypothesis = [tokenize(query)]
    reference = tokenize(groundtruth)
    return meteor_score(hypothesis, reference)

def get_scores(query, groundtruth):
    scores = dict()
```

```

for k, v in get_bleu(query, groundtruth).items():
    scores[k] = v
scores["rouge-1"] = get_rouge(query, groundtruth)
scores["meteor"] = get_meteor(query, groundtruth);
return scores

def calculate_f1(generated_labels_path, gt_labels_path):
    cxr_labels = ['Atelectasis', 'Cardiomegaly', 'Consolidation', 'Edema',
                  'Enlarged Cardiomeastinum', 'Fracture', 'Lung Lesion',
                  'Lung Opacity', 'No Finding', 'Pleural Effusion', 'Pleural Other',
                  'Pneumonia', 'Pneumothorax', 'Support Devices']
    useful_labels = cxr_labels
    true_labels = pd.read_csv(gt_labels_path).fillna(0)[useful_labels]

    pred_labels = pd.read_csv(generated_labels_path, index_col=False).fillna(0)[useful_labels]

    np_true_labels = true_labels.to_numpy()
    np_pred_labels = pred_labels.to_numpy()
    np_pred_labels[np_pred_labels == -1] = 0
    np_true_labels[np_true_labels == -1] = 0
    opts = np.array([0,1])
    assert np.all(np.isin(np_pred_labels, opts))

    f1_macro = f1_score(np_true_labels, np_pred_labels, average='macro')
    f1_micro = f1_score(np_true_labels, np_pred_labels, average='micro')
    return f1_macro, f1_micro

def calculate_s_emb(generated_reports_path, gt_embeddings_path):
    label_embeds = torch.load(gt_embeddings_path)
    np_label_embeds = label_embeds.numpy()

    pred_embeds = torch.load(generated_reports_path)
    np_pred_embeds = pred_embeds.numpy()
    assert np_label_embeds.shape == np_pred_embeds.shape
    # calc cosine sim
    sim_scores = (np_label_embeds * np_pred_embeds).sum(axis=1)/
        (np.linalg.norm(np_pred_embeds, axis=1)*np.linalg.norm(np_label_embeds, axis=1))
    assert len(sim_scores) == np_label_embeds.shape[0]
    return np.mean(sim_scores)

def present_scores(ground_truth, generated_reports):
    scores = []
    # headers = ['bleu_1', 'bleu_2', 'bleu_3', 'bleu_4', 'rouge-1', 'meteor']
    for i in range(len(ground_truth)):
        score = get_scores(generated_reports[i], ground_truth[i])
        scores.append(score)
    averaged_scores = {}
    bleu_1, bleu_2, bleu_3, bleu_4, rouge_1, meteor = 0,0,0,0,0,0
    for score in scores:
        bleu_1 += score['bleu_1']
        bleu_2 += score['bleu_2']
        bleu_3 += score['bleu_3']
        bleu_4 += score['bleu_4']
        rouge_1 += score['rouge-1']
        meteor += score['meteor']
    averaged_scores['bleu_1'] = bleu_1 / len(ground_truth)
    averaged_scores['bleu_2'] = bleu_2 / len(ground_truth)
    averaged_scores['bleu_3'] = bleu_3 / len(ground_truth)
    averaged_scores['bleu_4'] = bleu_4 / len(ground_truth)
    averaged_scores['rouge-1'] = rouge_1 / len(ground_truth)
    averaged_scores['meteor'] = meteor / len(ground_truth)
    return averaged_scores, scores

```

واژه‌نامه فارسی به انگلیسی

Automated Radiology Report Generation	تولید خودکار گزارش رادیولوژی
Class	کلاس
Method	تابع
Module	ماژول
Rule Based System	سیستم قانون‌محور
Technique	روش
Recursive Neural Network	شبکه عصبی بازگشتی
Neuron	نورون
Loss Function	تابع ضرر
Gradient	مشتق
Vanishing Gradient	محوشدگی مشتق
Exploding Gradient	انفجار مشتق
Long Short Term Memory	مدل حافظه طولانی کوتاه مدت
Attention	سازوکار توجه
Transformer	مبدل
Natural Language Processing	پردازش زبان‌های طبیعی
Weight	وزن
Self-Attention	مکانیسم خودتوجهی
Encoder	کدگذار
Decoder	کدگشا
Machine Learning	یادگیری ماشین
Threshold	حد آستانه
Metric	معیار
Deep Learning	یادگیری عمیق
Convolutional Network	شبکه‌های پیچشی
Singel-Label Classification	دسته‌بندی تک برچسبی

Multi-Label Classification	دسته بندی چند برچسبی
Tokenize	واژه‌بندی
Multi Layer Perceptron	شبکه عصبی چندلایه
Entropy	آنتروپی
Quantize	فشرده‌سازی
Accuracy	دقت
Example-Based Precision	دقت نمونه
Example-Based Recall	بازخوانی نمونه
Example-Based F1 Score	امتیاز F1 نمونه
Micro-Averaged Precision	دقت خرد
Micro-Averaged Recall	بازخوانی خرد
Micro-Averaged F1 Score	امتیاز F1 خرد
Activation Function	تابع فعال‌ساز
True Positive	مثبت صحیح
False Positive	منفی صحیح
True Negative	مثبت غلط
False Negative	منفی غلط
Open Source	متن باز
Hyper-Parameter	ابر متغیر

Abstract:

In this study, we propose a model for generating radiology reports for chest X-ray images. Automatic report generation is crucial due to the high volume of imaging requests and the shortage of radiologists. Such models can assist radiologists by facilitating faster and more accurate detection of abnormalities. We first introduce a baseline Report-Augmented Generation (RAG) model that generates reports by retrieving similar past reports. Then, we propose an enhanced model, Feature and Report Augmented Generation (FRAG), which retrieves not only past reports but also relevant clinical features or abnormalities associated with the input image. These retrieved elements are jointly used during the report generation process. The FRAG model is implemented in two variants: FRAG-A, which employs a zero-shot inference module, and FRAG-B, which utilizes a linear classifier for abnormality detection. All models are evaluated using both medical and natural language metrics. Evaluation results demonstrate that FRAG models significantly outperform the baseline RAG model in clinical metrics, with FRAG-A showing superior performance. Additionally, we conduct further experiments to examine the effect of the retrieval strategy, large language model, and prompting used in the generation stage. The ELIXR-B vision-language model is used for retrieval, and after embedding extraction, three selection strategies—top-k, K-Means clustering, and maximally diverse selection—are explored. The findings reveal that both the retrieval method and the choice of language model substantially impact performance, whereas the prompt template has a minimal effect.

Keywords: RAG, FRAG, Zero-Shot Inference, ELIXR-B, K-Means Clustering, Maximally Diverse Selection, LLM



**Iran University of Science and Technology
Computer Engineering Department**

Automated Chest X-ray Radiology Report Generation

Bachelor of Science Thesis in Computer Engineering

By:

Sina Alinejad

Supervisor:

Dr. Sauleh Eetemadi

May 2025