



دانشکده مهندسی کامپیوتر

بررسی هرس شبکه عصبی در پرسش و پاسخ تصویری

پروژه کارشناسی مهندسی کامپیوتر گرایش هوش مصنوعی

غزاله محمودی

استاد راهنما

دکتر سید صالح اعتمادی

شهریور ۱۴۰۰

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از گزارش پروژه پایانی

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: غزاله محمودی

عنوان گزارش پروژه پایانی: بررسی هرس شبکه عصبی در پرسش و پاسخ تصویری

تاریخ دفاع: شهریور ۱۴۰۰

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر صالح اعتمادی	استادیار	دانشگاه علم و صنعت ایران	

چکیده

مدل‌های از قبل آموزش دیده در مقیاس بزرگ مانند LXMERT در حال محبوب شدن برای یادگیری بازنمایی‌های متن و تصویر هستند. این مدل‌ها در مسایل مشترک بین بینایی و زبان کاربرد دارند. بر اساس فرضیه بلیط قرعه کشی^۱، شبکه‌های عصبی حاوی زیرشبکه‌های^۲ کوچکتری هستند که قادرند با آموزش در انزوا^۳ عملکردی مشابه شبکه کامل آموزش دیده داشته باشند. در این پروژه، وجود چنین زیرشبکه‌ای در شبکه LXMERT که بر روی مسئله پرسش و پاسخ تصویری آموزش دیده، بررسی می‌شود. همچنین مقادیر مختلف هرس شبکه و تاثیر آن بر کارایی شبکه مورد ارزیابی قرار می‌گیرد.

واژگان کلیدی: پرسش و پاسخ تصویری، شبکه LXMERT، فرضیه بلیط قرعه کشی.

¹Lottory Thicket Hypothesis

²Subnetwork

³Isolation

فهرست مطالب

ج	فهرست شکل‌ها
چ	فهرست جدول‌ها
۱	فصل ۱: مقدمه
۲	فصل ۲: پیش‌زمینه
۲	۱-۲ ساز و کار توجه
۴	فصل ۳: مروری بر کارهای مرتبط
۵	۱-۳ شبکه MiniVLM
۶	۲-۳ فرضیه بلیت قرعه‌کشی
۷	۱-۲-۳ هرس بر پایه وزن اتصالات
۷	۳-۳ فشرده‌سازی شبکه UNITER
۸	فصل ۴: پرسش و پاسخ تصویری
۸	۱-۴ تصاویر حقیقی مجموعه داده VQA
۹	۲-۴ تصاویر انتزاعی مجموعه داده VQA
۱۰	۳-۴ نوع سوالات و نحوه جمع‌آوری مجموعه داده VQA
۱۰	۴-۴ مجموعه داده VQA v2.0
۱۲	فصل ۵: شبکه LXMERT

۱۲	۱-۵ معماری شبکه
۱۳	۱-۱-۵ ورودی شبکه
۱۳	۲-۱-۵ رمزگذارهای شبکه
۱۳	۱-۲-۱-۵ Attention Layers
۱۴	۲-۲-۱-۵ رمزگذار Single-Modality
۱۵	۳-۲-۱-۵ رمزگذار Cross-Modality
۱۵	۳-۱-۵ خروجی شبکه
۱۵	۲-۵ استراتژی آموزش اولیه
۱۵	۱-۲-۵ روش‌های آموزش اولیه
۱۶	۱-۱-۲-۵ Language Task: Masked Cross-Modality LM
۱۶	۲-۱-۲-۵ Vision Task: Masked Object Prediction
۱۶	۳-۱-۲-۵ Cross-Modality Task
۱۷	۲-۲-۵ مجموعه داده استفاده شده در آموزش اولیه
۱۷	۳-۲-۵ نتایج

۱۸	فصل ۶: فشرده‌سازی شبکه LXMERT
۱۹	۱-۶ هرس اتصالات کم وزن
۲۰	۱-۱-۶ تحلیل نتایج
۲۱	۲-۶ هرس اتصالات به صورت تصادفی
۲۱	۱-۲-۶ تحلیل نتایج
۲۲	۳-۶ هرس اتصالات با وزن زیاد
۲۲	۱-۳-۶ تحلیل نتایج
۲۳	۴-۶ نحوه پیاده‌سازی و اجرا آزمایش‌ها

۲۵	فصل ۷: نتیجه‌گیری و پیشنهادات
۲۵	۱-۷ نتیجه‌گیری
۲۷	۲-۷ پیشنهادات و کارهای آینده

ث

فهرست مطالب

۲۸

مراجع

۳۰

واژه‌نامه فارسی به انگلیسی

فهرست شکل‌ها

۱-۳	مقایسه شبکه MiniVLM و شبکه OSCAR [۱۱]	۶
۱-۴	چند نمونه از تصاویر حقیقی مجموعه داده VQA v1 به همراه سوالات و زیر مجموعه‌ای از پاسخ‌ها. پاسخ‌های سبز رنگ با نگاه به تصویر داده شده است. پاسخ آبی بدون نگاه به تصویر داده شده است. [۱]	۹
۲-۴	چند نمونه از تصاویر انتزاعی مجموعه داده VQA v1 [۱]	۹
۳-۴	نمونه‌ای سوالات چند گزینه‌ای برای یک تصویر در VQA v1 [۱]	۱۰
۴-۴	چند نمونه از مجموعه داده VQA v2.0 [۶]	۱۱
۱-۵	معماری شبکه LXMERT [۱۰]	۱۳
۲-۵	رمزگذار Single-Modality در شبکه LXMERT [۱۰]	۱۴
۳-۵	رمزگذار Cross-Modality در شبکه LXMERT [۱۰]	۱۵
۴-۵	مثالی از آموزش اولیه شبکه LXMERT [۱۰]	۱۶
۵-۵	نتایج شبکه LXMERT [۱۰]	۱۷
۱-۶	نتایج حاصل از هرس اتصالات کم‌وزن	۲۰
۲-۶	نتایج حاصل از هرس اتصالات به صورت رندوم	۲۱
۳-۶	نتایج حاصل از هرس اتصالات با وزن زیاد	۲۲
۴-۶	میزان مصرف GPU	۲۳
۱-۷	نتایج انواع هرس به تفکیک درصد حذف اتصالات	۲۶

فهرست جدول‌ها

فصل ۱

مقدمه

در سال‌های اخیر و با پیشرفت‌های چشمگیر در حوزه هوش مصنوعی، پردازش زبان طبیعی و پردازش تصویر مسئله‌هایی با کاربرد عملی در زندگی روزمره انسان‌ها طراحی شده است. یکی از مواردی که اخیراً مورد توجه قرار گرفته است، بحث پرسش و پاسخ تصویری می‌باشد. این مسئله کاربردهای زیادی در کمک به نابینایان، دستیار هوشمند و موارد مشابه می‌تواند داشته باشد.

با توجه به اهمیت بحث پرسش و پاسخ تصویری در کمک به افراد کم‌بینا یا نابینا در زندگی روزمره و کمک به بهبود و تسهیل امور جاری روزانه، استفاده از مدل‌های آموزش دیده بر روی تلفن همراه یا وبسایت‌ها در قالب نرم‌افزارهای کاربردی از اهمیت بالایی برخوردار است. از سوی دیگر اغلب تلفن‌های همراه قدرت پردازش و حافظه محدودی دارند. استفاده بهینه از منابع موجود بسیار حائز اهمیت است. بنابراین علاوه بر آموزش مدل مناسب که برای این مسئله به دقت قابل قبولی برسد، لازم است مدل ساخته شده از حجم مناسبی برخوردار بوده و قابل استفاده بر روی تلفن همراه با استفاده از کمترین منابع باشد. به طوری که کارکرد تلفن همراه را دچار اختلال نکند.

با گسترش استفاده از شبکه‌های ترنسفورمر دقت‌های به دست آمده در مسئله پرسش و پاسخ تصویری به مقدار قابل قبولی رسیده است. اما شبکه‌های ترنسفورمری اغلب تعداد پارامترهای بالایی دارند. از این رو کوچک کردن مدل و فشردن آن از جمله مسائل داغ مورد بررسی است. در این پژوهش سعی شده است که هرس شبکه عصبی بر روی مسئله پرسش و پاسخ تصویری مورد بررسی قرار بگیرد و نتایج مدل فشرده شده با مدل اصلی مقایسه گردد. همچنین تاثیر فشردن سازی بر دقت مدل کاهش یافته و عملکرد آن بررسی شود.

فصل ۲

پیش زمینه

۱-۲ ساز و کار توجه

هدف استفاده از ساز و کار توجه^۱ بازیابی اطلاعات از بردارهای زمینه^۲ y_j در رابطه با بردار پرس و جو^۳ x می باشد. ساز و کار توجه ابتدا امتیاز α_j را بین بردار پرس و جو x و بردار زمینه y_j محاسبه می کند.

$$a_j = \text{Score}(x, y_j) \quad (۱-۲)$$

$$\alpha_j = \frac{\exp(a_j)}{\sum_k \exp(a_k)} \quad (۲-۲)$$

خروجی لایه توجه میانگین وزن دار امتیاز α_j به ازای بردارهای زمینه می باشد. محاسبات انجام شده مشابه لایه softmax است.

$$\text{Att}_{x \rightarrow \{y_j\}} = \sum_j \alpha_j y_j \quad (۳-۲)$$

^۱ Attention

^۲ Context Vector

^۳ Query Vector

اگر بردار پرس‌وجو x مجموعه‌ای از بردار زمینه $\{y_z\}$ باشد، امتیاز به دست آمده از معادله ۲-۳ توجه به خود^۴ نامیده می‌شود.

^۴self-attention

فصل ۳

مروری بر کارهای مرتبط

در حال حاضر تمرکز اصلی پژوهش‌های انجام شده در حوزه متن و تصویر دستیابی به دقت بالا با استفاده از شبکه‌های بزرگ ترنسفورمر می‌باشد. مدل‌های یادشده علاوه بر دقت بالا در مسائل، پیچیدگی محاسباتی و تعداد پارامترهای زیادی دارند. همین عامل موجب می‌شود استفاده از این مدل‌ها در کاربردهای زندگی حقیقی با مشکلاتی روبرو شود؛ زیرا به دستگاه‌هایی با پردازنده قوی و حافظه بالا نیاز دارند. بنابراین استفاده از مدل‌ها بر روی دستگاه‌های تلفن همراه که اغلب محدودیت پردازش و حافظه دارند، با چالش‌هایی روبرو شده است.

از این رو امروزه پژوهش‌هایی در زمینه کاهش پارامترهای مدل و کاهش پیچیدگی محاسباتی انجام شده است. هدف این پژوهش‌ها کاهش پیچیدگی محاسباتی و کاهش تعداد پارامترها می‌باشد به طوری که دقت مدل دست خوش تغییر نشود. پژوهش‌ها را در دو دسته کلی زیر می‌توان طبقه‌بندی کرد.

۱. معرفی معماری و ساختار جدید برای شبکه.

۲. فشردن سازی شبکه‌های از قبل معرفی شده با روش‌های مختلف.

در ادامه به بررسی پژوهش‌های صورت گرفته در این حوزه پرداخته می‌شود.

۳-۱ شبکه MiniVLM

شبکه MiniVLM^۱ یک مدل کوچک‌تر و سریع‌تر برای مسئله‌های تصویری-زبانی می‌باشد. با توجه به ویژگی یاد شده، این شبکه مناسب استفاده در دستگاه‌هایی با محدودیت حافظه و قدرت پردازش است. ساختار MiniVLM با هدف کاهش محاسبات ناشی از ساختار ترنسفورمرها طراحی شده است. در پژوهش‌های انجام شده برای طراحی مدل MiniVLM ابتدا معماری مدل برای دستیابی به سرعت و دقت مناسب، بهبود یافته و سپس بخش آموزش اولیه^۲ ارتقا داده شده است [۱۱].

برای بخش تصویر از Two-stage Efficient feature Extractor (TEE) استفاده شده است. استفاده از این ماژول هزینه استخراج ویژگی‌های تصویر را به میزان ۹۵ درصد در مقایسه با شبکه OSCAR [۷]^۳ کاهش داده است. در Two-stage Efficient feature Extractor (TEE) به جای استفاده از لایه‌های پیچشی^۴ معمولی از لایه پیچشی نقطه‌ای^۵ و لایه پیچشی عمقی^۶ استفاده شده است.

برای ارتقا بخش آموزش اولیه و جبران کاهش تعداد پارامترها در شبکه MiniVLM، از مجموعه‌های بسیار بزرگ استفاده شده است.

در نتیجه همه تلاش‌ها زمان ابتدا به انتها این مدل در مقایسه با شبکه OSCAR میزان ۶٪ کاهش پیدا کرده است. تعداد پارامترهای شبکه MiniVLM ۲۷٪ پارامترهای OSCAR می‌باشد. همچنین شبکه MiniVLM، به ۹۴٪-۹۷٪ دقت شبکه OSCAR دست یافته است. موارد مطرح شده نشان می‌دهد با وجود کاهش قابل توجه در تعداد پارامترها و همچنین افزایش سرعت مدل، دقت تغییر چندانی نکرده است. نتایج مقایسه دو شبکه در شکل ۳-۱ قابل مشاهده است.

^۱Mini Vision-Language Model

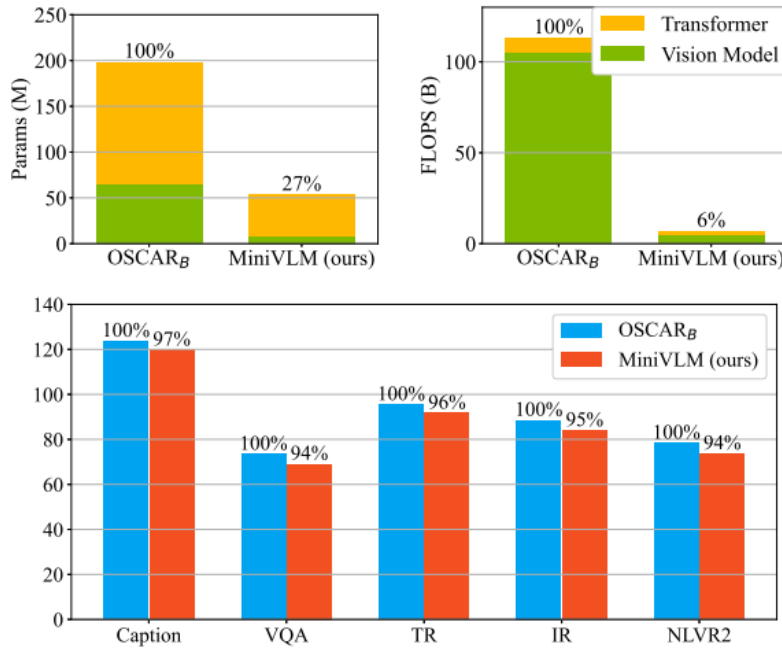
^۲pre-training

^۳State Of The Art

^۴Convolution

^۵Pointwise

^۶Depthwise



شکل ۳-۱: مقایسه شبکه OSCAR و MiniVLM [۱۱]

۳-۲ فرضیه بلیت قرعه کشی

در یادگیری ماشین تکنیک هرس^۷ شبکه عصبی، پارامترهای غیر ضروری شبکه را حذف می‌کند. با این روش می‌توان بدون کاهش چشم‌گیر در دقت شبکه، پارامترهای شبکه را به میزان قابل قبولی کاهش داد. همچنین سرعت شبکه نسبت به حالت قبل هرس، افزایش پیدا می‌کند. هدف اصلی در پژوهش یافتن بازنمایی کم‌حجم‌تر برای شبکه‌های عصبی کاملاً متصل^۸ می‌باشد تا موجب کاهش پیچیدگی محاسباتی شود. در مقاله فرضیه زیر اثبات شده است: یک شبکه کاملاً متصل که به صورت تصادفی مقدار دهی اولیه شده است شامل یک زیر شبکه^۹ می‌باشد به طوری که اگر آن زیر شبکه را به تعداد تکرار^{۱۰} مشابه شبکه اصلی آموزش دهیم، دقت روی داده تست در هر دو حالت یکسان خواهد شد [۴].

⁷Pruning

⁸Dense Neural Network

⁹Subnetwork

¹⁰Iteration

نویسندگان مقاله فرضیه را بلیت قرعه‌کشی^{۱۱} نام‌گذاری کردند و برای اثبات فرضیه علاوه بر شبکه کاملاً متصل^{۱۲}، صحت فرضیه مطرح شده را بر روی شبکه پیچشی^{۱۳} بررسی کردند. در این مقاله الگوریتمی ارائه شده است که می‌تواند بلیط برنده^{۱۴} را پیدا کند. برای یافتن بلیت برنده در یک فرایند تکرار شونده^{۱۵}، اتصالات کم‌وزن شبکه را حذف می‌کنیم. سپس زیر شبکه جدید را دوباره آموزش می‌دهیم. این روش شبکه را کوچک‌تر می‌کند در حالی که دقت با حالت شبکه کامل تفاوتی ندارد. فرایند انجام آزمایش‌ها در این مقاله با هرس تکرار شونده^{۱۶} صورت گرفته است.

۳-۲-۱ هرس بر پایه وزن اتصالات

در هرس بر پایه وزن اتصالات^{۱۷} هدف این است که اتصالات شبکه با وزن کم نسبت به سایر اتصالات از شبکه حذف شود. ایده روش مطرح شده این است که اتصالات کم وزن، تاثیر کمتری در نتایج به دست آمده دارد. بنابراین حذف این اتصالات تغییر ناچیزی در عملکرد مدل دارد. پس می‌توان این اتصالات را حذف کرد.

۳-۳ فشرده‌سازی شبکه UNITER

شبکه UNITER^{۱۸} [۲] یکی از شبکه‌های معرفی شده در مسئله پرسش و پاسخ تصویری می‌باشد که به نتایج بسیار قابل قبولی دست یافته است. در راستا فشرده‌سازی شبکه‌های حجیم ترنسفورمری، پژوهشی به بررسی فشرده‌سازی و کاهش تعداد پارامترهای شبکه UNITER پرداخته است. نتایج به دست آمده از این پژوهش نشان می‌دهد برای مسئله پرسش و پاسخ تصویری اگر ۷۰ درصد اتصالات کم‌وزن شبکه کامل UNITER را هرس کنیم، دقت زیر شبکه جدید ۹۹ درصد دقت شبکه کامل می‌باشد [۵].

¹¹ Lottery Thicket Hypothesis

¹² Dense Neural Network

¹³ Convolutional Neural Network

¹⁴ Winning Ticket

¹⁵ Iterative

¹⁶ Iterative Pruning

¹⁷ Magnitude Pruning

¹⁸ UNiversal Image-TEXT Representation Learning

فصل ۴

پرسش و پاسخ تصویری

مجموعه داده **Visual Question Answering (VQA v1.0)** یکی از غنی‌ترین و معروف‌ترین مجموعه داده در مسئله پرسش و پاسخ تصویری می‌باشد. یک تصویر و یک سوال مرتبط با تصویر به عنوان ورودی به سیستم داده می‌شود. هدف این است که با توجه به تصویر، دقیق‌ترین پاسخ به سوال داده شود. این مجموعه داده شامل ۲۵۰ هزار تصویر و ۷۶۰ هزار سوال و ۱۰ میلیون پاسخ می‌باشد. تصاویر این مجموعه داده از دو بخش تصاویر حقیقی^۱ و تصاویر انتزاعی^۲ تشکیل شده است [۱].

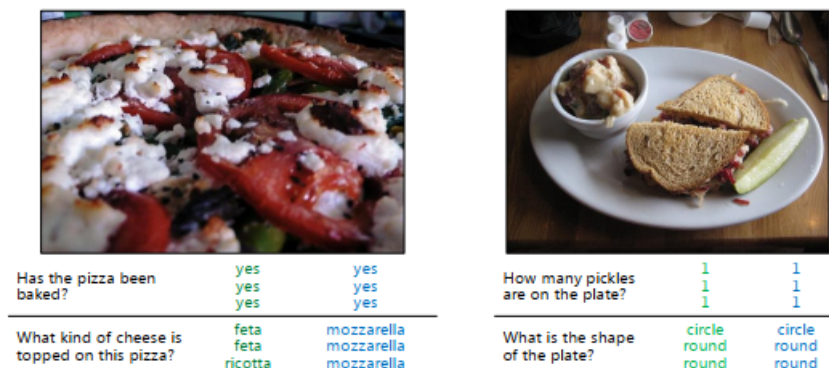
۴-۱ تصاویر حقیقی مجموعه داده VQA

برای تصاویر حقیقی (که شامل ۱۲۳۲۸۹ داده آموزشی و ۸۱۴۳۴ داده ارزیابی می‌باشد) از تصاویر موجود در مجموعه داده MS COCO^۳ [۸] استفاده شده است. هر تصویر مجموعه داده MS COCO شامل چندین شی است. درجه سختی تصاویر، این مجموعه داده را مناسب برای تسک VQA کرده است. در شکل ۴-۱ نمونه‌ای از تصاویر حقیقی به همراه پرسش و پاسخ مربوطه آورده شده است.

¹Real Image

²Abstract Image

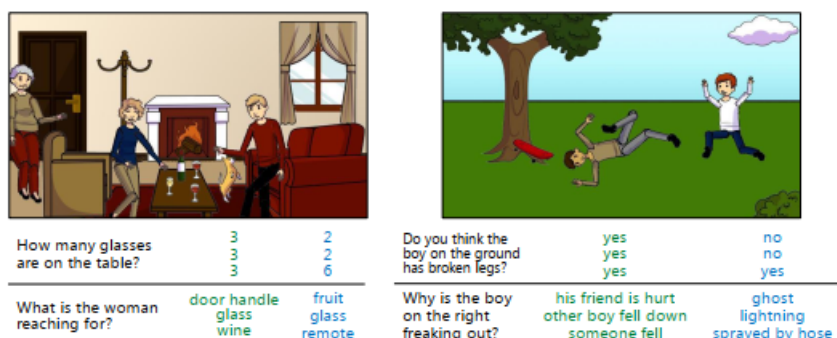
³Microsoft Common Objects In Context (MS COCO)



شکل ۴-۱: چند نمونه از تصاویر حقیقی مجموعه داده VQA v1 به همراه سوالات و زیر مجموعه‌ای از پاسخ‌ها. پاسخ‌های سبز رنگ با نگاه به تصویر داده شده است. پاسخ آبی بدون نگاه به تصویر داده شده است. [۱]

۲-۴ تصاویر انتزاعی مجموعه داده VQA


تصاویر انتزاعی شامل تصاویر کارتونی می‌باشد. علت قرار دادن تصاویر انتزاعی در کنار تصاویر حقیقی این است که با از بین بردن نیاز به تجزیه و تحلیل تصویر واقعی، تمرکز مدل بر روی استدلال‌های سطح بالاتر افزایش یابد. تصاویر انتزاعی شامل هر دو محیط داخل خانه و خارج خانه می‌باشند. این مجموعه شامل ۱۰۰ شی و ۳۱ حیوان در موقعیت‌های مختلف می‌باشد. ۵۰ هزار داده از تصاویر انتزاعی در مجموعه VQA موجود است. در شکل ۲-۴ نمونه‌ای از تصاویر انتزاعی به همراه پرسش و پاسخ مربوطه آورده شده است.



شکل ۲-۴: چند نمونه از تصاویر انتزاعی مجموعه داده VQA v1 [۱]

۳-۴ نوع سوالات و نحوه جمع‌آوری مجموعه داده VQA

به ازای هر تصویر در مجموعه داده VQA حداقل ۳ سوال (به طور میانگین ۴ یا ۵ سوال) وجود دارد که ۱۰ کاربر مختلف به هر سوال پاسخ داده‌اند. پرسش‌های بله/خیر، تعداد اشیا و دیگر پرسش‌ها دسته‌بندی انواع سوالات در این مجموعه داده می‌باشد. برای پاسخ‌گویی به سوالات از کاربران حقیقی استفاده شده است و کاربر می‌بایست از بین گزینه‌های موجود پاسخ مناسب سوال را انتخاب کند (شکل ۳-۴).



Q: Why are they standing?

(a) yes	(b) no	(e) 3	(f) 4
(c) 1	(d) 2	(i) blue	(j) yellow
(g) white	(h) red	(m) waiting	(n) no where to sit
(k) playing game	(l) sheepskin	(q) forks	(r) waiting for train
(o) firestone	(p) rugby		

Q: Is the TV on?

(a) yes	(b) no	(e) 3	(f) 4
(c) 1	(d) 2	(i) blue	(j) yellow
(g) white	(h) red	(m) sports	(n) between big elephants
(k) shag	(l) jeopardy	(q) tv show	(r) white streak on face
(o) edinburgh	(p) strawberries		

شکل ۳-۴: نمونه‌ای سوالات چند گزینه‌ای برای یک تصویر در VQA v1 [۱]

۴-۴ مجموعه داده VQA v2.0

مجموعه داده VQA v2.0 در سال ۲۰۱۷ در تکمیل و بهبود مجموعه داده VQA v1.0 معرفی شد. مشکل اصلی مجموعه داده VQA v1.0 تعصبات زبانی^۴ موجود می‌باشد. به عنوان مثال اگر سوال با Is there a clock آغاز شود، با احتمال ۹۵ درصد پاسخ yes می‌باشد. غلبه بر این مشکل با جمع‌آوری تصاویر مکمل میسر شد. به این صورت که به ازای هر پرسش یکسان دو تصویر وجود دارد که پاسخ‌های متفاوتی دارند. در شکل ۴-۴ نمونه‌ای از مجموعه داده VQA v1.0 قابل مشاهده است [۶].

⁴language-bias



شکل ۴-۴: چند نمونه از مجموعه داده VQA v2.0 [۶]

با جمع‌آوری مجموعه داده VQA v2.0 به عنوان نسخه متعادل شده مجموعه داده معروف VQA v1.0، تعصبات زبانی به میزان قابل توجهی کاهش پیدا کرد. همچنین وجود تصاویر مکمل باعث شد دقت به دست‌آمده در مدل‌های زبانی-بصری قابل اطمینان‌تر باشد و فهم مدل از تصویر را انعکاس دهد.

فصل ۵

شبکه LXMERT

استدلال در ترکیب تصویر و زبان نیاز به فهم بصری، فهم زبانی و ارتباط مابین فهم بصری و فهم زبانی دارد. شبکه LXMERT^۱ برای حل مسئله‌های زبانی-بصری طراحی شده است. LXMERT یک شبکه عصبی از قبل آموزش دیده^۲ از نوع ترنسفورمر می‌باشد که بر خلاف ترنسفورمرهای معمول از ۳ رمزگذار^۳ تشکیل می‌شود. در ادامه درباره معماری شبکه، ساختار رمزگذار و نحوه آموزش اولیه^۴ شبکه توضیحاتی داده می‌شود [۱۰].

۱-۵ معماری شبکه

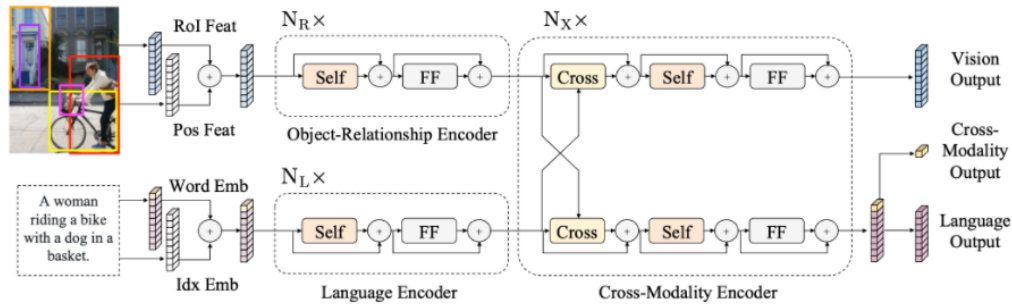
همان‌طور که در شکل ۱-۵ قابل مشاهده است، شبکه LXMERT از لایه‌های self-attention و cross-attention تشکیل شده است. یک تصویر و جمله مرتبط با آن به عنوان ورودی به شبکه داده می‌شود. هر تصویر شامل دنباله‌ای از object می‌باشد. شبکه LXMERT بازنمایی مناسبی از تصویر و زبان و cross-modality ایجاد می‌کند.

^۱Language Cross-Modality Encoder Representation from Transformers

^۲pre-train

^۳Encoder

^۴pre-train



شکل ۵-۱: معماری شبکه LXMERT [۱۰]

۵-۱-۱ ورودی شبکه

ورودی شبکه LXMERT از دو بخش Word-Level Sentence Embedding و Object-Level Image Embeddings تشکیل شده است.

در بخش Word-Level Sentence Embedding جملات توسط WordPiece tokenizer به توکن‌هایی جدا می‌شوند. در ادامه هر توکن توسط لایه embedding به بردار بازنمایی تبدیل می‌شود. مشابه شبکه BERT [۳] برای نمایش محل دقیق توکن در جمله از Index Embedding استفاده می‌شود. در بخش Object-Level Image Embeddings به جای استفاده از CNN's feature map، ویژگی‌های پیدا شده توسط Faster-RCNN [۹] مورد استفاده قرار می‌گیرد.

۵-۱-۲ رمزگذارهای شبکه

شبکه LXMERT شامل Language Encoder، Object-relationship Encoder و Cross-modality Encoder می‌باشد. هر رمزگذار از self-attention و cross-attention تشکیل شده است.

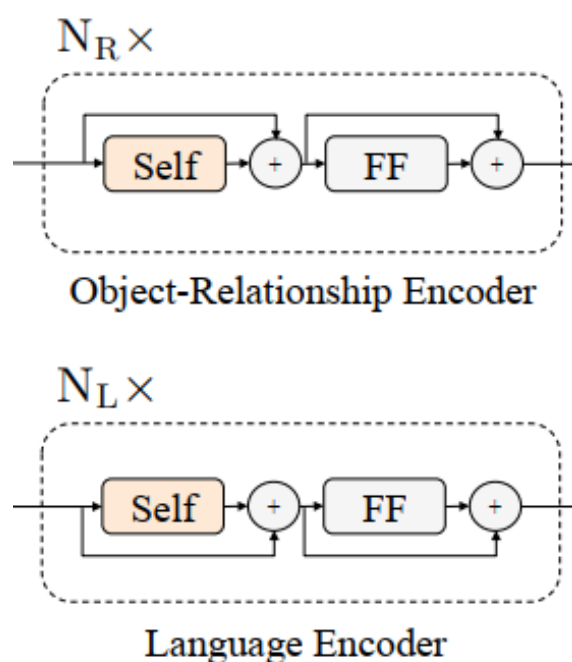
۵-۱-۲-۱ Attention Layers

مشابه مکانیزم توجه^۵ در ترنسفورمر می‌باشد. توضیحات دقیق‌تر در بخش ۲-۱ داده شده است.

^۵ Attention

۵-۲-۱-۵ رمزگذار Single-Modality

بعد از لایه embedding برای هر کدام از ورودی‌های زبان و تصویر دو رمزگذار^۶ مجزا وجود دارد. محاسبات در این دو رمزگذار از یکدیگر مستقل هستند. هر رمزگذار شامل مکانیزم توجه به خود^۷ و شبکه Feed Forward می‌باشد. همچنین بعد از هر زیر لایه اتصال رو به جلو^۸ و لایه نرمال‌سازی^۹ وجود دارد که با نماد "+" در شکل ۵-۲ نشان داده شده است.



شکل ۵-۲: رمزگذار Single-Modality در شبکه LXMERT [۱۰]

^۶Encoder

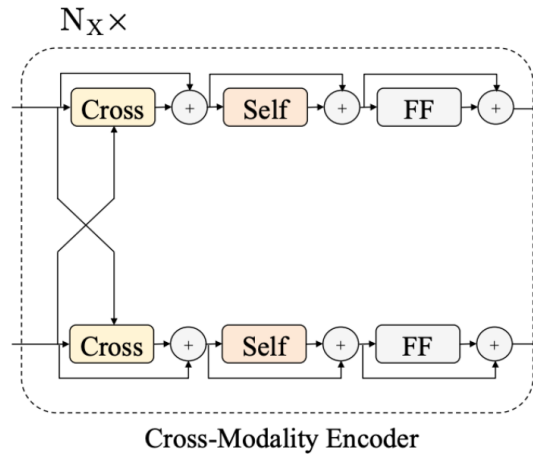
^۷self-attention

^۸Residual Connection

^۹Normalization Layer

۵-۲-۳ رمزگذار Cross-Modality

برای به دست آوردن بازنمایی مشترک مابین تصویر و زبان استفاده می‌شود. این رمزگذار مشابه رمزگذار Single-Modality می‌باشد با این تفاوت که دارای cross attention نیز می‌باشد (شکل ۵-۳).



شکل ۵-۳: رمزگذار Cross-Modality در شبکه LXMERT [۱۰]

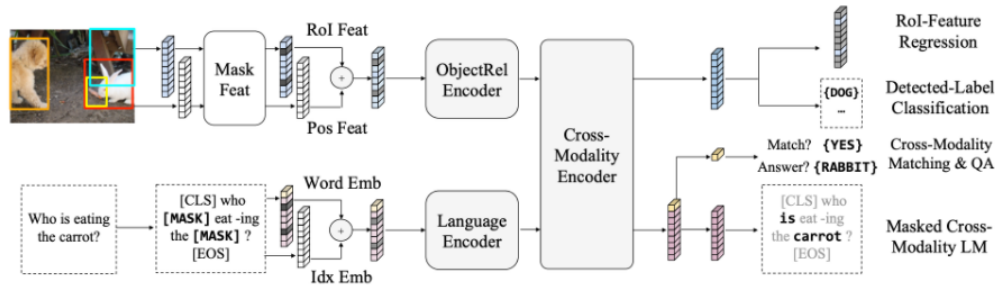
۵-۱-۳ خروجی شبکه

خروجی شبکه LXMERT از سه بخش تصویر، زبان و Cross-Modality تشکیل می‌شود. بخش زبان و تصویر توسط رمزگذار Cross-Modality و با توجه به دنباله ورودی هر مورد تولید شده است. خروجی Cross-Modality از توکن CLS تشکیل شده و کاربردی همانند آنچه در BERT ذکر شده دارد.

۵-۲ استراتژی آموزش اولیه

۵-۲-۱ روش‌های آموزش اولیه

شبکه LXMERT به طور کلی توسط سه نوع روش pre-train می‌شود. در ادامه توضیحی مختصری از هر روش آورده شده است.



شکل ۵-۴: مثالی از آموزش اولیه شبکه LXMERT [۱۰]

۵-۲-۱-۱ Language Task: Masked Cross-Modality LM

در این روش همچون روشی که در شبکه BERT استفاده شده، پانزده درصد توکن‌های ورودی با توکن Mask جایگزین می‌شوند. فرق اجرا این روش در LXMERT با BERT در این است که در BERT تشخیص توکن Mask تنها با استفاده از توکن‌های جمله ورودی انجام می‌شود. این در حالی است که در LXMERT علاوه بر توکن‌های جمله ورودی، از ویژگی‌های تصویر هم در تشخیص توکن Mask استفاده می‌شود. همان‌طور که در شکل ۵-۴ قابل مشاهده است، در صورتی که کلمه Carrot برای Mask شدن انتخاب شود، وجود تصویر در تشخیص درست شبکه برای کلمه Mask شده از اهمیت بالایی برخوردار است.

۵-۲-۱-۲ Vision Task: Masked Object Prediction

مشابه قسمت زبانی، در این قسمت نیز بخشی از ویژگی ورودی تصویر Mask می‌شود و شبکه به کمک جمله مرتبط با تصویر و سایر ویژگی‌های تصویر که Mask نشده‌اند، قسمت مورد نظر را تشخیص می‌دهد. با توجه به نوع ورودی‌ها در این روش همچون روش قبلی مدل cross-modality alignment را نیز فرا می‌گیرد.

۵-۲-۱-۳ Cross-Modality Task

برای آموزش بهتر بخش Cross-Modality از دو روش دیگر نیز استفاده شده است.

۱. روش Cross-Modality Matching

هر جمله به احتمال ۵۰ درصد با یک جمله نامرتبط با تصویر جایگزین می‌شود. سپس یک رده‌بند^{۱۰}

¹⁰classifier

آموزش داده می‌شود تا مطابقت تصویر و جمله را بررسی کند. این مسئله شبیه به پیشبینی جمله بعدی^{۱۱} در آموزش اولیه شبکه BERT می‌باشد.

۲. روش Image Question Answering

در این روش یک تصویر و سوال مرتبط با تصویر داده می‌شود و وظیفه مدل پیش‌بینی پاسخ می‌باشد.

۵-۲-۲ مجموعه داده استفاده شده در آموزش اولیه

برای آموزش اولیه از ۵ مجموعه داده استفاده شده است که شامل COCO-Cap^{۱۲}، VG-Cap^{۱۳}، VQA^{۱۴}، GQA^{۱۵} و VG-QA^{۱۶} می‌باشد. فقط از مجموعه آموزشی^{۱۶} و ارزیابی^{۱۷} مجموعه داده‌های فوق استفاده شده است. برای هر تصویر چندین پرسش و پاسخ موجود است.

۵-۲-۳ نتایج

در نهایت دقت این شبکه بر روی ۳ مجموعه داده NLVR، VQA و GQA مورد بررسی قرار گرفته است. در هر سه مورد نتایج نسبت به State Of The Art بهبود قابل توجهی داشته است (شکل ۵-۵).

Method	VQA				GQA			NLVR ²	
	Binary	Number	Other	Accu	Binary	Open	Accu	Cons	Accu
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	88.2	54.2	63.1	72.5	77.8	45.0	60.3	42.1	76.2

شکل ۵-۵: نتایج شبکه LXMERT [۱۰]

¹¹Next Sentence Prediction

¹²Visual Genom Caption

¹³Visual Question Answering

¹⁴A New Dataset for Real-World Visual Reasoning and Compositional Question Answering

¹⁵Visual Genom - Question Answering

¹⁶Train

¹⁷Dev

فصل ۶

فشرده‌سازی شبکه LXMERT

در این پژوهش به سه روش فشرده‌سازی (هرس) شبکه LXMERT انجام شده و تلاش برای یافتن بهترین روش فشرده‌سازی صورت گرفته است. هر سه روش هرس بر پایه وزن اتصالات می‌باشند. بررسی نتایج و تاثیر هرس بر دقت شبکه بر روی مجموعه داده **VQA v2.0** بررسی شده است. در اجرا از مقادیر از پیش تعیین شده ابرپارامترها^۱ در شبکه LXMERT استفاده شده است.

به جز اتصالات embedding ورودی و اتصالات لایه خروجی، برای سایر اتصالات احتمال حذف شدن وجود دارد. این روش‌ها از میزان هرس ۱۰ درصد شبکه تا هرس ۹۰ درصد شبکه در سه seed تکرار شد؛ تا علاوه بر بررسی تاثیر نوع و میزان حذف اتصالات بر دقت نهایی، با تکرار در سه seed میزان قابل اطمینان بودن نتایج به دست‌آمده مورد بررسی قرار گیرد.

^۱Hyperparameters

۶-۱ هرس اتصالات کم وزن

این روش همانند فرضیه بلیت قرعه‌کشی^۲ می‌باشد. صحت این فرضیه تا به حال در شبکه‌های کاملاً متصل و شبکه پیچشی مورد بررسی قرار گرفته است. حال قرار است صحت آن بر یک ترنسفورمر زبانی-تصویری دو جریانه^۳ مورد بررسی قرار دهیم. در این روش به صورت هرس تکرار شونده^۴ عمل شد. مراحل الگوریتم به صورت زیر می‌باشد.

۱. وزن‌های از قبل آموزش دیده مدل LXMERT به همراه رده‌بند VQA به شبکه داده می‌شود. این مقادیر برای مراحل بعد ذخیره می‌شود.

۲. مدل بر روی ۳۱۲۹ جواب پرتکرار مجموعه داده VQA v2.0 آموزش می‌بیند. آموزش ۴ بار تکرار^۵ می‌شود. در نهایت مدل finetune شده بر روی مسئله VQA به دست می‌آید. مقادیر دقت به دست آمده از این مرحله با برچسب Unpruned Baseline در نمودار مشاهده می‌شود.

۳. در این مرحله هر بار ۱۰ درصد از اتصالات کم‌وزن شبکه به صورت تکرار شونده حذف می‌شود. این مرحله تا زمانی که درصد مشخصی از کل اتصالات شبکه حذف شود ادامه پیدا می‌کند. برای همه اتصالات به جز اتصالات لایه embedding و اتصالات لایه خروجی، احتمال حذف در این مرحله وجود دارد. پس از رسیدن به درصد هرس مشخص، دقت شبکه هرس شده بر روی مجموعه داده VQA با برچسب pruned در نمودار مشخص می‌شود.

۴. پس حذف مقدار مشخصی از اتصالات (اتمام هرس) وزن‌های ذخیره شده در مرحله اول به شبکه بازنشانی^۶ می‌شود. مقادیر دقت به دست آمده از این مرحله با برچسب reset initial weight در نمودار مشاهده می‌شود.

۵. حال شبکه هرس شده به تعداد تکرار مشابه مرحله ۲ آموزش می‌بیند. در نهایت دقت شبکه در این حالت بررسی می‌شود. مقادیر دقت به دست آمده از این مرحله با برچسب retrain در نمودار مشاهده می‌شود.

²Lottery Ticket Hypothesis

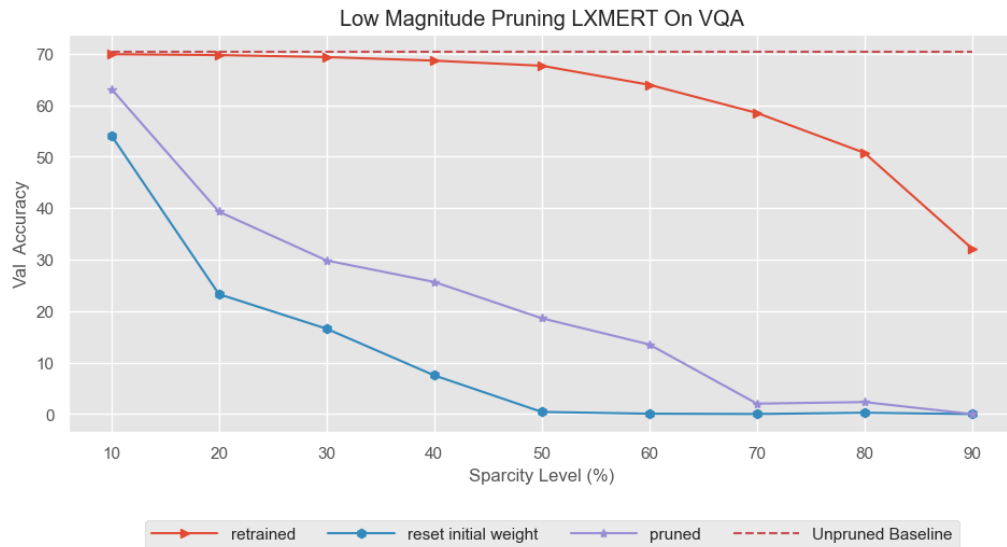
³cross-modal

⁴Iterative Pruning

⁵Iteration

⁶Reset

نتایج هرس اتصالات کم‌وزن شبکه LXMERT در نمودار شکل ۶-۱ قابل مشاهده است. لازم به ذکر است همه دقت‌های گزارش شده، از ارزیابی بر روی مجموعه داده Test_dev (Validaion) به دست آمده است.



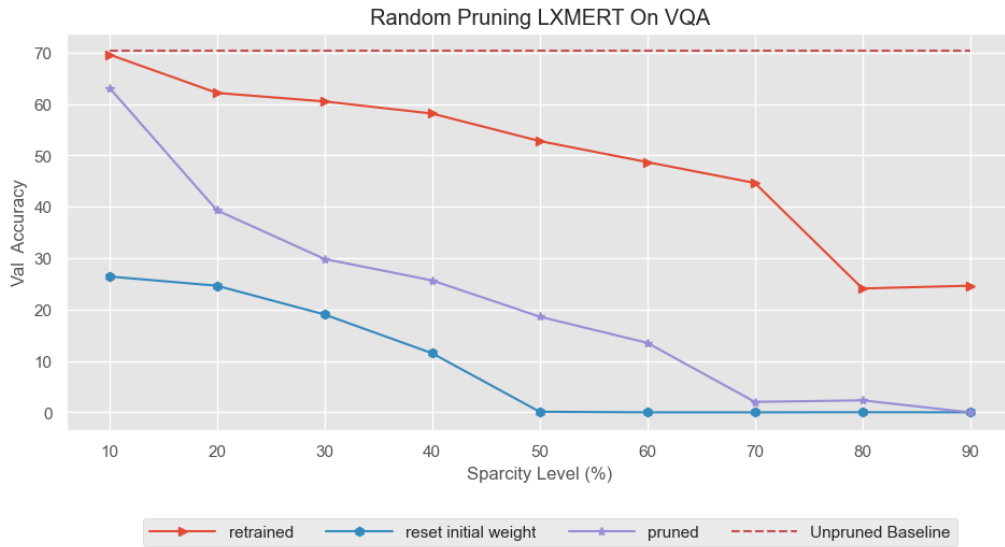
شکل ۶-۱: نتایج حاصل از هرس اتصالات کم‌وزن

۶-۱-۱ تحلیل نتایج

از آزمایش‌های انجام شده این نتیجه حاصل می‌شود که می‌توان ۵۰ تا ۶۰ درصد اتصالات کم‌وزن شبکه LXMERT را بدون کاهش شدیدی در دقت نهایی حذف کرد. این نتیجه نمایان‌گر آن است که ۵۰ درصد اتصالات عملاً تاثیری چندانی ندارند و قابل حذف هستند. همچنین فرضیه بلیت قرعه‌کشی در شبکه دو جریانه LXMERT برقرار است. همان‌طور که از نمودارهای رسم شده مشخص است، همه نمودارها روند نزولی دارند. به این صورت که هر چه تعداد بیشتری از اتصالات را حذف کنیم، دقت شبکه کاهش پیدا می‌کند. اگر میزان حذف اتصالات بیش از ۵۰ درصد باشد، شاهد کاهش دقت شدیدتری هستیم. همچنین نتایج نشان می‌دهد اتصالات کم‌وزن در شبکه عصبی تاثیر کمتری در کارایی نهایی شبکه دارند. بنابراین اگر نصف اتصالات کم‌وزن حذف شود تاثیر چندانی در دقت نهایی ندارد. پس آموزش اتصالات و مقدار وزن آن‌ها به بهترین شکل صورت گرفته و وزن اتصالات نشان‌دهنده اهمیت آن‌ها می‌باشد.

۲-۶ هرس اتصالات به صورت تصادفی

در این روش به صورت کاملاً تصادفی تعدادی از وزن‌های شبکه را حذف کرده و سایر اتصالات باقی‌مانده را نگه می‌داریم. نتایج هرس اتصالات شبکه LXMERT به صورت تصادفی در نمودار شکل ۲-۶ قابل مشاهده است.



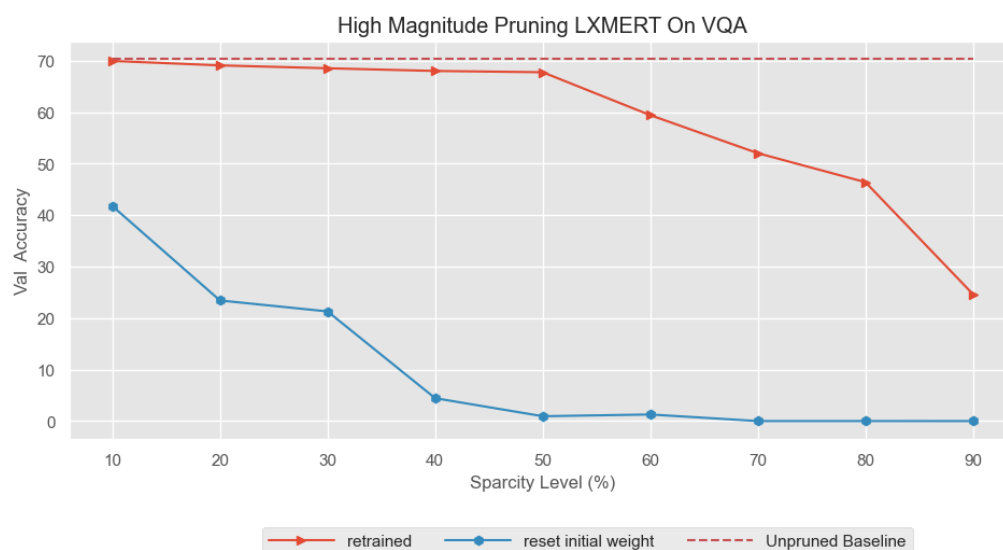
شکل ۲-۶: نتایج حاصل از هرس اتصالات به صورت رندوم

۱-۲-۶ تحلیل نتایج

همچون شکل ۱-۶ که نتایج قسمت ۱-۶ را نشان می‌دهد، در این قسمت نیز نمودار نزولی داریم (شکل ۲-۶). به این صورت که با افزایش میزان هرس شبکه، دقت نیز به همان نسبت کاهش پیدا می‌کند. با توجه به اینکه در این حالت اتصالات رندوم انتخاب می‌شوند از میزان هرس ۲۰ درصد و بیشتر کاهش دقت بیشتری رخ می‌دهد. پس می‌توان به این صورت مطرح کرد که اتصالاتی که وزن کمتری دارند، تاثیر کمتری در دقت نهایی شبکه داشتند. به همین علت بود که در هرس اتصالات کم وزن (قسمت ۱-۶) با حذف ۵۰ درصد اتصالات کم وزن دقت تغییر چشم‌گیری نداشت.

۳-۶ هرس اتصالات با وزن زیاد

در این روش اتصالاتی که در هرس ۱-۶ نجات یافته‌اند، حذف می‌شوند. به عبارت دیگر در این روش اتصالات با وزن زیاد حذف می‌شوند. نتایج حاصل از اجرا در شکل ۳-۶ قابل مشاهده است.



شکل ۳-۶: نتایج حاصل از هرس اتصالات با وزن زیاد

۱-۳-۶ تحلیل نتایج

همان‌طور که در شکل ۳-۶ مشاهده می‌شود، با افزایش میزان حذف اتصالات دقت نهایی شبکه کاهش می‌یابد. شیب نمودار از نمودار ۱-۶ تندتر است بدین معنی که تغییرات با شدت بیشتری رخ داده است. انتظار می‌رفت تغییرات نمودار و دقت نهایی از دو بخش قبل بدتر باشد ولی نتایج به دست آمده با پیش‌بینی‌ها مغایرت دارد.

۴-۶. نحوه پیاده‌سازی و اجرا آزمایش‌ها

پیاده‌سازی پروژه در اینجا قابل مشاهده می‌باشد که با زبان پایتون و فریم‌ورک پایتورچ انجام شده است. برای شبکه LXMERT از پیاده‌سازی اصلی مقاله و ابر پارامترهای تعیین شده که در گیت‌هاب موجود می‌باشد، استفاده شد.

برای اجرا از سخت‌افزار GPU.1080Ti.xlarge با رم 31.3GB استفاده شد. دستور nvidia-smi میزان استفاده از GPU را نمایش می‌دهد. خروجی این دستور در هنگام اجرا آزمایشات به صورت شکل ۴-۶ شد.

```
ubuntu@nlp992:~$ nvidia-smi
Tue Aug 10 13:47:17 2021

+-----+
| NVIDIA-SMI 470.57.02    Driver Version: 470.57.02    CUDA Version: 11.4     |
+-----+-----+
| GPU   Name                               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|=====+=====+
| 0     NVIDIA GeForce ...      Off          | 00000000:00:05.0 Off |           N/A       |
| 58%   73C   P2     282W / 250W | 10267MiB / 11178MiB |    97%    Default   |
+-----+-----+

+-----+
| Processes:                                                       GPU Memory |
|  GPU   GI    CI          PID    Type   Process name                  Usage       |
|=====+=====+
|  0     N/A   N/A         52444    C     python3                       10263MiB   |
+-----+-----+

ubuntu@nlp992:~$
```

شکل ۴-۶: میزان مصرف GPU

هر دوره^۷ اجرا نزدیک به ۲ ساعت زمان می‌گیرد. فرآیند هرس شبکه هم حدوداً نیم ساعت زمان می‌خواهد. تعداد اجرا پیش فرض ۴ می‌باشد. بنابراین هر آزمایش شامل فرآیند هرس به میزان مشخص، بازنمایی^۸ وزن‌های اولیه و آموزش شبکه هرس شده نزدیک به ۹ ساعت زمان می‌گیرد. با توجه به این موضوع برای اجرا آزمایش‌ها کد اتوماتیک نوشته شد که برای ۳ seed به ازای ۳ مدل هرس (اتصالات کم‌وزن، تصادفی، اتصالات با وزن زیاد) و برای میزان هرس ۱۰ درصد تا ۹۰ درصد (۹ مقدار) به صورت پشت سر هم اجرا شود. به این صورت از منابع GPU به بهترین صورت ممکن استفاده شد. با توجه به سنگین بودن مجموعه داده VQA v2.0 رم GPU می‌بایست از 6900 MiB بیشتر باشد.

^۷Epoch

^۸Reset

برای اجرا آزمایش‌ها لازم است با دستور `pip3 install -r requirements.txt` پکیج‌های مورد نیاز برای اجرا کد را نصب کنید. با توجه به طولانی بودن زمان اجرا کل آزمایش‌ها بهتر است با دستور `screen` یک اسکرین جدید برای اجرا بسازید و به شکل معمول دستورات را اجرا کنید. در این صورت حتی با بستن ترمینال اجرا ادامه پیدا می‌کند. با `screen -r` در صورت باز کردن مجدد ترمینال می‌توانید فرآیند و میزان پیش‌رفت اجرا را ببینید. راه حل دیگر برای تداوم اجرا در صورت بستن ترمینال استفاده از دستور `nohup` است. کافیت `nohup bash run/vqa_run.bash` را اجرا کنید. خروجی‌های اجرا در `nohup.out` قابل مشاهده می‌باشد. در صورتی که یک `screen` جدید برای خود ساختید، با دستور `bash run/vqa_run.bash` کلیه آزمایش‌ها به صورت متوالی اجرا می‌شود. در نهایت نتایج به دست‌آمده در فایل `json` در پوشه `result` ذخیره می‌شود. همچنین `mask` های هرس و وزن‌های شبکه در پوشه `models` قابل مشاهده می‌باشند.

فصل ۷

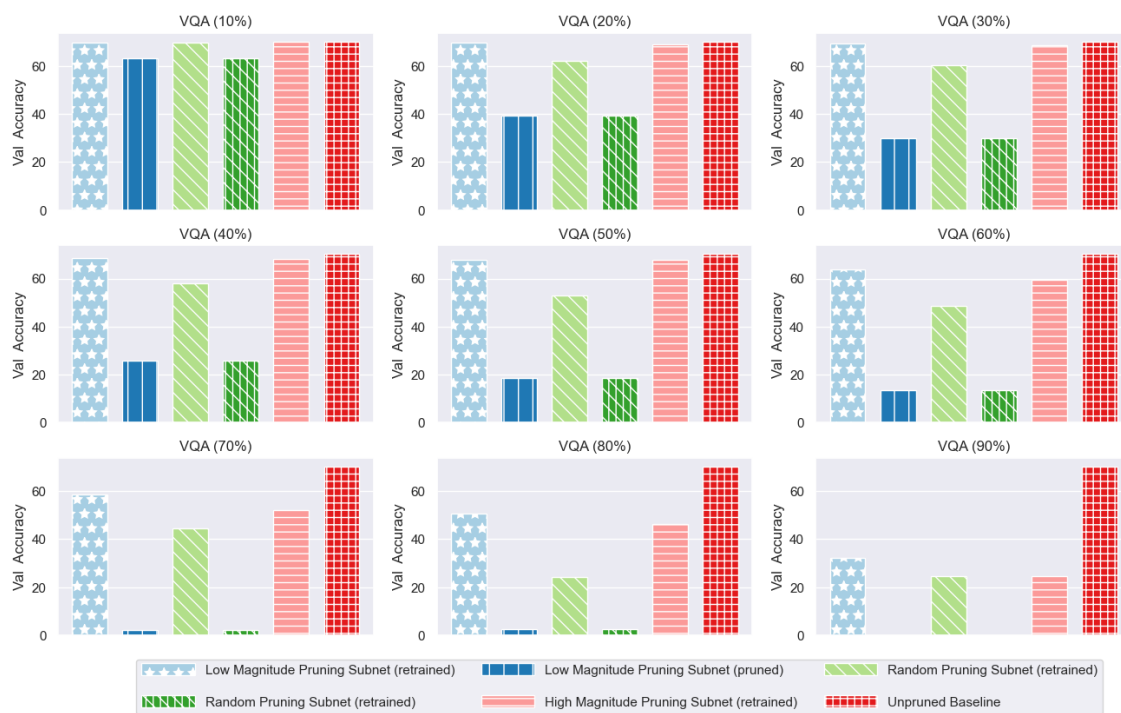
نتیجه گیری و پیشنهادات

۷-۱ نتیجه گیری

با توجه به نتایج به دست آمده از آزمایش های انجام شده مشاهده می شود ۵۰ درصد اتصالات کم وزن شبکه LXMERT تاثیر چندانی بر کارایی نهایی شبکه ندارند. بنابراین برای کاهش پارامترها و اندازه شبکه حذف اتصالات کم وزن یکی از بهترین روش های ممکن است. ولی اگر اتصالات به صورت تصادفی انتخاب شود نتایج متفاوت است. بدین معنی که علاوه بر اتصالات نوع اتصالاتی که در زیر شبکه هرس شبکه نگه می داریم عاملی مهم و تاثیر گذار است. در شکل ۷-۱ خلاصه نتایج سه نوع هرس معرفی شده به تفکیک میزان هرس و بر اساس نوع هرس قابل مشاهده است. همچنین در جدول ۷-۱ اعداد دقیق برای بررسی های احتمالی گزارش شده است.

جدول ۷-۱: نتایج مدل هرس شده آموزش دیده برای انواع هرس به تفکیک درصد حذف اتصالات

درصد هرس	اتصالات با وزن کم	رندوم	اتصالات با وزن زیاد
۱۰	69.94 ± 0.03	69.58 ± 0.02	69.75 ± 0.13
۲۰	69.59 ± 0.11	62.21 ± 0.06	69.17 ± 0.09
۳۰	69.23 ± 0.07	60.23 ± 0.24	68.60 ± 0.08
۴۰	68.70 ± 0.05	57.49 ± 0.62	67.87 ± 0.18
۵۰	67.44 ± 0.14	52.63 ± 0.13	67.78 ± 0.04
۶۰	63.94 ± 0.01	48.68 ± 0.00	58.76 ± 0.63
۷۰	58.50 ± 0.04	44.96 ± 0.34	52.28 ± 0.26
۸۰	50.63 ± 0.21	24.36 ± 0.26	46.54 ± 0.15
۹۰	28.12 ± 4.02	24.36 ± 0.26	25.40 ± 0.77



شکل ۷-۱: نتایج انواع هرس به تفکیک درصد حذف اتصالات

۲-۷. پیشنهادات و کارهای آینده

با توجه به پیشرفت چشم‌گیر در هوش مصنوعی و حرکت سریع به سمت استفاده از ابزارهای هوشمند، برای ادامه تحقیقات موارد زیر پیشنهاد می‌شود.

۱. در این پژوهش مدل LXMERT بر روی مجموعه داده VQA بررسی شد. در ادامه می‌توان دو مجموعه داده دیگر از جمله GQA و NLVR2 را بررسی کرد.

۲. می‌توان کارایی شبکه هرس شده و آموزش دیده روی مسئله VQA را بر روی سایر مجموعه داده‌ها (GQA, NLVR2) بررسی نمود. بدین ترتیب تاثیر انتقال یادگیری^۱ در LXMERT مشخص می‌شود.

۳. روش بررسی شده در این پژوهش، هرس اتصالات شبکه می‌باشد. از این رو زمان اجرا ابتدا به انتها شبکه تفاوتی چندانی نمی‌کند. می‌توان انواع دیگر هرس از جمله هرس ساختاری^۲ را مورد بررسی قرار داد.

۴. نتایج به دست‌آمده در روش اتصالات با وزن زیاد (بخش ۶-۳) دور از انتظار بود. می‌توان در ادامه تحقیقات معماری شبکه LXMERT را به صورت دقیق‌تر بررسی کرد. این بررسی ممکن است به معرفی مدل دیگری با ساختار جدید و بهبود دقت در مسئله پرسش و پاسخ تصویری ختم شود.

۵. بیشتر پژوهش‌ها در موضوع فشرده‌سازی شبکه به صورت تئوری است. وقت آن است نتیجه این پژوهش‌ها در عمل و برنامه‌های کاربردی^۳ مورد استفاده در روزمره انسان‌ها مورد بررسی قرار گیرد.

۶. مجموعه داده VQA که در آزمایش‌ها مورد استفاده قرار گرفت به زبان انگلیسی می‌باشد. با توجه به نبود مجموعه داده مناسب به زبان فارسی، یکی دیگر از کارهای ارزشمند جمع‌آوری مجموعه داده فارسی پرسش و پاسخ تصویری می‌باشد. بدین ترتیب برنامه‌های کاربردی طراحی شده برای کمک به کم‌بینایان، نابینایان یا استفاده‌های دیگر می‌توانند به زبان فارسی باشند.

¹Transfer Learning

²Structural Pruning

³Application

- [1] ANTOL, S., AGRAWAL, A., LU, J., MITCHELL, M., BATRA, D., LAWRENCE ZITNICK, C., AND PARIKH, D. Vqa: Visual question answering. in *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2425–2433.
- [2] CHEN, Y.-C., LI, L., YU, L., KHOLY, A. E., AHMED, F., GAN, Z., CHENG, Y., AND LIU, J. Uniter: Universal image-text representation learning, 2020.
- [3] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
- [4] FRANKLE, J., AND CARBIN, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks, 2019.
- [5] GAN, Z., CHEN, Y.-C., LI, L., CHEN, T., CHENG, Y., WANG, S., AND LIU, J. Playing lottery tickets with vision and language, 2021.
- [6] GOYAL, Y., KHOT, T., SUMMERS-STAY, D., BATRA, D., AND PARIKH, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
- [7] LI, X., YIN, X., LI, C., HU, X., ZHANG, P., ZHANG, L., WANG, L., HU, H., DONG, L., WEI, F., CHOI, Y., AND GAO, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020* (2020).
- [8] LIN, T.-Y., MAIRE, M., BELONGIE, S., BOURDEV, L., GIRSHICK, R., HAYS, J., PERONA, P., RAMANAN, D., ZITNICK, C. L., AND DOLLÁR, P. Microsoft coco: Common objects in context, 2015.

- [9] SHAOQING REN, KAIMING HE, R. G. J. S. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).
- [10] TAN, H. H., AND BANSAL, M. Lxmert: Learning cross-modality encoder representations from transformers. in *EMNLP/IJCNLP* (2019).
- [11] WANG, J., HU, X., ZHANG, P., LI, X., WANG, L., ZHANG, L., GAO, J., AND LIU, Z. Minivlm: A smaller and faster vision-language model, 2020.

واژه‌نامه فارسی به انگلیسی

pre-train	آموزش اولیه
Train	آموزشی
Hyperparameter	ابریارمتر
Residual Connection	اتصال رو به جلو
Dev	ارزیابی
Transfer Learning	انتقال یادگیری
Isolation	انزوا
Reset	بازنشانی
Label	برچسب
Context Vector	بردار زمینه
Query Vector	بردار پرس و جو
Visual Question Answering	پرسش و پاسخ تصویری
Abstract Image	تصاویر انتزاعی
Real Image	تصاویر حقیقی
language-bias	تعصبات زبانی
Iteration	تکرار
Iterative	تکرار شونده
Attention	توجه
self-attention	توجه به خود
Token	توکن
Epoch	دوره
cross-modal	دو جریانه
Classifier	رده‌بند
Subnetwork	زیرشبکه
Encoder	رمزگذار
Seed	سید

Convolutional Neural Network	شبکه عصبی پیچشی
Dense Neural Network	شبکه کاملاً متصل
Depthwise	عمقی
Lottory Thicket Hypothesis	فرضیه بلیط قرعه کشی
Normalization Layer	لایه نرمال سازی
Pointwise	نقطه‌ای
Pruning	هرس
Magnitude Pruning	هرس اتصالات بر اساس وزن
Low Magnitude Pruning	هرس اتصالات کم وزن
High Magnitude Pruning	هرس اتصالات با وزن زیاد
Iterative Pruning	هرس تکرارشونده
Structural Pruning	هرس ساختاری
Random Pruning	هرس تصادفی
Structural Pruning	هرس ساختاری

Abstract:

Large-scale pretrained models such as LXMERT are becoming popular for learning cross-modal representations on text-image pairs for vision-language tasks. According to the lottery ticket hypothesis, NLP and computer vision models contain smaller subnetworks capable of being trained in isolation to full performance. In this project, we combine these observations to evaluate whether such trainable subnetworks exist in LXMERT when fine-tuned on the VQA task. In addition, we perform a model size cost-benefit analysis by investigating how much pruning can be done without significant loss in accuracy.

Keywords: Visual Question Answering, LXMERT, Lottery Ticket Hypothesis.



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

LXMERT Model Compression for Visual Question Answering

Bachelor of Science Thesis in Computer Engineering

By:

Ghazaleh Mahmoodi

Supervisor:

Dr. Sayyed Sauleh Eetemadi

August 2021