دانشکده مهندسی کامپیوتر

بررسی یادگیری بدون داده مشابه و با مجموعه داده محدود بر روی مدل مبتنی بر ترنسفورمر برای دو تسک پرسش و پاسخ تصویری و توصیف تصویر

پروژه کارشناسی مهندسی کامپیوتر

شقایق مبشر و غزل زمانی‌نژاد

استاد راهنما

دکتر سید صالح اعتمادی

مهر ۱۴۰۱

بسم الله الرحمن الرحیم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: شقایق مبشر و غزل زمانی‌نژاد

عنوان پروژه : بررسی یادگیری بدون داده مشابه و با مجموعه داده محدود بر روی مدل مبتنی بر ترنسفورمر برای دو تسک پرسش و پاسخ تصویری و توصیف تصویر

تاریخ دفاع: مهر ۱۴۰۱

رشته: مهندسی کامپیوتر

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضـــــــــا
۱	استاد مشاور	دکتر سید صالح اعتمادی	استادیار	دانشگاه علم و صنعت ایران	

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب شقایق مبشر و غزل زمانی‌نژاد به شماره دانشجویی **۹۷۵۲۲۰۰۴** و **۹۷۵۲۲۱۶۶** 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. درصورت 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: شقایق مبشر و غزل زمانی‌نژاد

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.

استاد راهنما: دکتر سید صالح اعتمادی

تاریخ:

امضا:

تقدیم به:

مادر و پدر عزیزم

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست. در آغاز وظیفه خود می‌دانیم از زحمات بی‌دریغ استاد راهنمای خود، آقای دکتر سید صالح اعتمادی، صمیمانه تشکر و قدردانی کنیم.

هم‌چنین از خانم مهندس مریم سادات هاشمی به دلیل کمک‌های بسیاری که به ما کردند کمال امتنان را داریم.

از دوستان عزیزمان، خانم ملیکا نوبختیان و آقای محمدعلی فراهت که در انجام پروژه به ما یاری رساندند، سپاس فراوان داریم.

در پایان، بوسه زده بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیز و تشکر می‌کنیم از خانواده عزیز به پاس عاطفه سرشار و گرمای امیدبخش وجودشان، که بهترین پشتیبانمان بودند.

شقایق مبشر و غزل زمانی‌نژاد

مهر ۱۴۰۱

چکیده

امروزه یکی از داغ‌ترین مسائل در زمینهٔ هوش مصنوعی، توانایی ادراک تصویر به کمک مدل‌های کارآمد است. ظهور راه‌حل‌های یادگیری ماشین مانند توصیف تصاویر، برای افراد نابینا و کم‌بینا که قادر به درک تصاویر بصری نیستند، یک موهبت است. با استفاده از یک مدل توصیف تصویر، می‌توان توضیحات تصویر را برای افراد ارائه کرد. به کمک این تکنولوژی آن‌ها می‌توانند درک بهتری از محیط اطراف خود داشته باشند. علاوه بر آن، به کمک مدل‌های پرسش و پاسخ تصویری، می‌توان به پرسش‌های کاربر در مورد یک تصویر پاسخ مناسب داد. همچنین می توان سیستم‌های پرسش و پاسخ خودکار که توانایی ادراک تصاویر را دارند تولید کرد. این سیستم‌ها می‌توانند با انسان ارتباط برقرار کنند و در زندگی روزمره کاربردهای فراوانی داشته باشند.

در سال‌های اخیر در دو مسئله توصیف تصویر و پرسش و پاسخ تصویری تلاش‌های فراوانی صورت گرفته است اما بیشتر آن‌ها نمی‌توانند بر روی تصاویر مربوط به یک فرهنگ خاص عملکرد خوبی نشان دهند. در این پژوهش قصد داریم ضمن جمع‌آوری یک دادگان که نمایانگر فرهنگ ایران است، عملکرد یکی از مدل‌های موجود (مدل CLIP) را که در یادگیری بدون نمونه برای ارتباط بین تصویر و متن نسبت به مدل‌های دیگر عملکرد بهتری دارد، در این دو وظیفه با تصاویر مرتبط با فرهنگ ایران به دو زبان فارسی و انگلیسی آزمایش کنیم. نوع یادگیری‌های آزمایش شده، یادگیری بدون نمونه و یادگیری با نمونه‌های معدود است. در این آزمایش‌ها تلاش می‌کنیم تا عملکرد CLIP را روی دسته‌های خاصی از تصاویر به طور جداگانه آزمایش کنیم و با امتحان محرک‌های مختلف عملکرد آن را بهبود ببخشیم.

واژگان کلیدی: هوش مصنوعی، پردازش زبان‌های طبیعی، بینایی کامپیوتر، پرسش و پاسخ تصویری، توصیف تصویر، یادگیری بدون نمونه، یادگیری با نمونه‌های معدود، مدل‌های بینایی_ و_ زبان، مجموعه دادگان فارسی، بررسی عملکرد مدل‌ها با تصاویر مربوط به فرهنگ ایران

فهرست مطالب

فهرست تصاویر د

فهرست جداول ز

فهرست الگوریتم‌ها ژ

فهرست علائم اختصاری ژ

فصل ۱: مقدمه ۱

۱ـ۱ پیشگفتار ۲

۱ـ۲ بیان مسئله ۲

۱ـ۲ـ۱ تعریف مسائل پرسش و پاسخ تصویری و توصیف تصویر ۲

۱ـ۲ـ۲ به کارگیری مسائل پرسش و پاسخ تصویری و توصیف تصویر برای یک کشور و فرهنگ خاص ۳

۱ـ۲ـ۳ مفهوم یادگیری بدون نمونه و یادگیری با نمونه‌های معدود ۳

۱ـ۳ ترتیب نگارش گزارش ۵

فصل ۲: مروری بر مفاهیم اولیه ۶

۲ـ۱ مقدمه ۷

۲ـ۲ یادگیری بدون نمونه[۱] ۷

۲ـ۳ یادگیری با نمونه‌های معدود[۲] ۸

[۱] Zero-shot Learning
[۲] Few-shot Learning

۲ـ۴ یادگیری بر پایهٔ محرک[۳] ۸
۲ـ۴ـ۱ اضافه کردن پرامپت ۹
۲ـ۴ـ۲ جستجوی پاسخ ۹
۲ـ۴ـ۳ نگاشت پاسخ ۱۰
۲ـ۵ معیارهای به کار رفته برای ارزیابی مدل توصیف تصویر ۱۱
۲ـ۵ـ۱ معیار BLEU [۴] ۱۱
۲ـ۵ـ۲ معیار ROUGE[۵] ۱۳

فصل ۳: مروری بر منابع ۱۴
۳ـ۱ مقدمه ۱۵
۳ـ۲ مدل **GPT-2** ۱۵
۳ـ۳ مدل‌های **V&L** ۱۶
۳ـ۴ مدل CLIP ۱۶
۳ـ۴ـ۱ نتایج مدل CLIP در یادگیری بدون نمونه برای دسته‌بندی تصاویر ۱۷
۳ـ۴ـ۲ نتایج مدل CLIP در یادگیری با تعداد نمونه معدود برای طبقه بندی تصاویر . . ۱۹
۳ـ۴ـ۳ به‌کارگیری CLIP در مسئله پرسش و پاسخ تصویری ۱۹
۳ـ۴ـ۴ به‌کارگیری مدل CLIP در مسئله توصیف تصویر ۲۱

فصل ۴: شرح روش پیشنهادی ۲۵
۴ـ۱ مقدمه ۲۶
۴ـ۲ مدل CLIP برای وظیفه پرسش و پاسخ تصویری ۲۶
۴ـ۳ مدل CLIPCap برای وظیفه توصیف تصویر ۲۷

فصل ۵: پیاده‌سازی ۳۱
۵ـ۱ مقدمه ۳۲

[۳] Prompt-based Learning
[۴] Bilingual Evaluation Understudy
[۵] Recall-Oriented Understudy for Gisting Evaluation

۵ـ۲ جمع‌آوری داده ۳۲
۵ـ۲ـ۱ دستورالعمل جمع آوری دادهٔ توصیف تصویر ۳۶
۵ـ۲ـ۲ دستورالعمل جمع آوری دادهٔ پرسش و پاسخ تصویری ۳۷
۵ـ۲ـ۳ ترجمهٔ مجموعهٔ داده به زبان انگلیسی ۳۸
۵ـ۳ شرح نحوهٔ ایجاد و آموزش مدل‌های توصیف تصویر ۳۹
۵ـ۴ شرح نحوهٔ ایجاد و آموزش مدل‌های پرسش و پاسخ تصویری ۴۰

فصل ۶: بازبینی عملکرد ۴۱
۶ـ۱ مقدمه ۴۲
۶ـ۲ نتایج و ارزیابی ۴۲
۶ـ۲ـ۱ نمونه‌هایی از نتایج آزمایش مدل‌های توصیف تصویر ۴۲
۶ـ۲ـ۲ بررسی عملکرد مدل‌های توصیف تصویر ۷۱
۶ـ۲ـ۳ نمونه‌هایی از نتایج آزمایش مدل‌های پرسش و پاسخ تصویری ۷۶
۶ـ۲ـ۴ بررسی عملکرد مدل‌های پرسش و پاسخ تصویری ۹۷

فصل ۷: جمع‌بندی و پیشنهادها ۱۰۰
۷ـ۱ مقدمه ۱۰۱
۷ـ۲ نتیجه‌گیری ۱۰۱
۷ـ۳ نوآوری ۱۰۲
۷ـ۴ پیشنهادها ۱۰۳

مراجع ۱۰۴

واژه‌نامه فارسی به انگلیسی ۱۰۸

واژه‌نامه انگلیسی به فارسی ۱۱۰

فهرست تصاویر

۳ـ۱ مدل CLIP در یادگیری بدون نمونه، با مدل‌های کاملاً نظارت شده قابل رقابت است. . . ۱۸
۳ـ۲ مدل CLIP با یادگیری بدون نمونه بهتر از probe های خطی یادگیری با تعداد نمونه معدود است. ۱۹
۳ـ۳ نتایج مسئله توصیف تصویر ۲۱
۳ـ۴ مروری بر معماری توصیف تصویر مبتنی بر CLIP ۲۲
۳ـ۵ مروری بر روش توصیف تصویر مدل CLIP-S ۲۳
۳ـ۶ عملکرد مدل CLIP-S در مسئله توصیف تصویر با در نظر گرفتن پاداش‌های مختلف . . ۲۴

۴ـ۱ معماری مدل CLIPCap. مرجع تصویر: مقالهٔ CLIPCap[۲۶] ۲۸

۵ـ۱ تصویری از بخش جمع‌آوری دادهٔ پرسش و پاسخ تصویری در وب‌سایت ۳۵
۵ـ۲ تصویری از بخش جمع‌آوری دادهٔ توصیف تصویر در وب‌سایت ۳۵
۵ـ۳ تصویری از بخش بازبینی دادهٔ پرسش و پاسخ تصویری در وب‌سایت ۳۶
۵ـ۴ تصویری از بخش بازبینی دادهٔ توصیف تصویر در وب‌سایت ۳۶

۶ـ۱ نمونه‌ای از تصاویر مجموعه دادگان ۴۲
۶ـ۲ نمونه‌ای از تصاویر مجموعه دادگان ۴۴
۶ـ۳ نمونه‌ای از تصاویر مجموعه دادگان ۴۵
۶ـ۴ نمونه‌ای از تصاویر مجموعه دادگان ۴۶
۶ـ۵ نمونه‌ای از تصاویر مجموعه دادگان ۴۸
۶ـ۶ نمونه‌ای از تصاویر مجموعه دادگان ۴۹

۶ـ۷ نمونه‌ای از تصاویر مجموعه دادگان ۵۱
۶ـ۸ نمونه‌ای از تصاویر مجموعه دادگان ۵۲
۶ـ۹ نمونه‌ای از تصاویر مجموعه دادگان ۵۳
۶ـ۱۰نمونه‌ای از تصاویر مجموعه دادگان ۵۵
۶ـ۱۱نمونه‌ای از تصاویر مجموعه دادگان ۵۶
۶ـ۱۲نمونه‌ای از تصاویر مجموعه دادگان ۵۸
۶ـ۱۳نمونه‌ای از تصاویر مجموعه دادگان ۵۹
۶ـ۱۴نمونه‌ای از تصاویر مجموعه دادگان ۶۱
۶ـ۱۵نمونه‌ای از تصاویر مجموعه دادگان ۶۲
۶ـ۱۶نمونه‌ای از تصاویر مجموعه دادگان ۶۴
۶ـ۱۷نمونه‌ای از تصاویر مجموعه دادگان ۶۶
۶ـ۱۸نمونه‌ای از تصاویر مجموعه دادگان ۶۷
۶ـ۱۹نمونه‌ای از تصاویر مجموعه دادگان ۶۹
۶ـ۲۰نمونه‌ای از تصاویر مجموعه دادگان ۷۰
۶ـ۲۱بررسی یادگیری بدون نمونه و با تعداد نمونه محدود مدل CLIPCap هر دو زبان برپایه
معیارهای BLEU، ROUGE-L و ROUGE-2 . . . . . ۷۵
۶ـ۲۲نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۷۶
۶ـ۲۳نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۷۷
۶ـ۲۴نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصوری . . ۷۸
۶ـ۲۵نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۷۹
۶ـ۲۶نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۰
۶ـ۲۷نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۱
۶ـ۲۸نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۲
۶ـ۲۹نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۳
۶ـ۳۰نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۴
۶ـ۳۱نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۵

۶ـ۳۲نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۶
۶ـ۳۳نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۷
۶ـ۳۴نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۸
۶ـ۳۵نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۸۹
۶ـ۳۶نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۰
۶ـ۳۷نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۱
۶ـ۳۸نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۲
۶ـ۳۹نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۳
۶ـ۴۰نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۴
۶ـ۴۱نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۵
۶ـ۴۲نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری . . ۹۶
۶ـ۴۳بررسی یادگیری بدون نمونه و با تعداد نمونه محدود مدل CLIP برای مسئلهٔ پرسش و پاسخ تصویری ۹۹

فهرست جداول

۲ـ۱ اصطلاحات و نشانه‌گذاری روش‌های محرک. z^* پاسخ‌هایی را نشان می‌دهد که با خروجی واقعی y^* مطابقت دارند. ۱۱

۳ـ۱ مقایسه CLIP با نتایج قبلی دسته‌بندی تصویر یادگیری بدون نمونه. CLIP عملکرد هر سه مجموعه داده را مقدار زیادی بهبود می‌بخشد. ۱۸

۵ـ۱ تعداد تصاویر و توصیف‌های جمع‌آوری شده برای مسئلهٔ توصیف تصویر ۳۳

۵ـ۲ تعداد تصاویر، پرسش و پاسخ جمع‌آوری شده به تفکیک دسته‌بندی پرسش برای مسئله پرسش و پاسخ تصویری ۳۴

فهرست علائم اختصاری

توصیف تصویر IC

پرسش و پاسخ تصویری ..... VQA

پردازش زبان‌های طبیعی ..... NLP

بخش گفتار ..... PoS

فصل ۱

مقدمه

۱ ـ ۱ پیشگفتار

روش‌های مبتنی بر یادگیری برای بازشناسی اشیا و اتفاقات درون تصاویر، پیشرفت‌های بسیاری در سال‌های اخیر داشته‌اند. مدل‌های زیادی در زمینه‌های مختلف بینایی کامپیوتر و پردازش زبان‌های طبیعی ارائه شده‌اند که دقت‌های بالایی در دسته‌بندی، کشف و بازشناسی تصاویر و همچنین توصیف تصویر کسب کرده‌اند. با این وجود، این مدل‌ها نیاز به حجم زیادی دادهٔ برچسب زده شده دارند تا بتوانند به دقت بالا برسند و جمع‌آوری مجموعه دادگان این چنینی در بعضی حوزه‌ها و زبان‌ها می‌تواند بسیار هزینه‌بر و گران تمام شود.
برای زبان فارسی و فرهنگ ایران تاکنون مدل خوبی برای بسیاری از زمینه‌های بینایی کامپیوتر و پردازش زبان‌های طبیعی ارائه نشده است. مجموعه‌های داده فارسی نیز در این موارد بسیار محدود هستند، بدین ترتیب نمی‌توان از روش‌هایی که نیاز به داده‌های زیاد دارند برای بسیاری از حوزه‌ها در زبان فارسی استفاده کرد. با توجه به هزینه‌بر بودن گردآوری یک مجموعه دادگان، امروزه تحقیقات به سمت به‌کارگیری داده‌های کمتر برای آموزش پیش می‌رود و به نظر می رسد که آینده هوش مصنوعی در جهت کاهش تعداد دادهٔ مورد نیاز برای آموزش شبکه‌های عصبی قدم بردارد.[۱]
یادگیری بدون نمونه و یادگیری با نمونه‌های معدود دو نوع یادگیری است که نیاز به دادگان بسیار برای یادگیری دسته‌هایی که مدل تاکنون ندیده‌است ندارند. تعداد راهکارهای این نوع یادگیری‌ها هر ساله در حال افزایش است. بعضی از مدل‌هایی که در سال‌های اخیر معرفی شده‌اند توانسته‌اند نتایج قابل توجهی با استفاده از این دو روش به دست آورند. پژوهش برای بهبود عملکرد مدل‌ها به کمک این دو یادگیری هنوز ادامه داشته و جای پیشرفت زیادی دارد.

۱ ـ ۲ بیان مسئله

۱ ـ ۲ ـ ۱ تعریف مسائل پرسش و پاسخ تصویری و توصیف تصویر

در مسئله پرسش و پاسخ تصویری مدل سعی می‌کند به سوالاتی که در رابطه با تصویر ورودی پرسیده می‌شود؛ پاسخ مناسب دهد. ایدهٔ گسترده‌تر این مسئله، طراحی سیستم‌هایی است که می‌توانند محتویات یک تصویر را (مشابه آنچه انسان‌ها انجام می‌دهند) درک کنند و به طور موثر در مورد آن تصویر به زبان طبیعی ارتباط برقرار کنند. این فرآیند چالش‌های زیادی دارد زیرا به مدل‌های مبتنی بر تصویر و مدل‌های زبان طبیعی، برای تعامل

و تکمیل یک‌دیگر نیاز دارد.

توصیف تصویر، فرآیند تولید توضیح متنی برای یک تصویر است. از پردازش زبان طبیعی و بینایی کامپیوتر برای تولید توضیح تصویر استفاده می‌شود، بدین ترتیب یک مدل توصیف تصویر باید قابلیت درک تصویر ورودی و تولید یک متن خروجی به عنوان توصیف تصویر را داشته باشد.

۱ ـ ۲ ـ ۲ به کارگیری مسائل پرسش و پاسخ تصویری و توصیف تصویر برای یک کشور و فرهنگ خاص

هر دو مسئلهٔ پرسش و پاسخ تصویری و توصیف تصویر پیشرفت بسیاری در زبان انگلیسی و چند زبان دیگر داشته‌اند. با وجود پژوهش‌های بسیار در این دو حوزه، بیشتر مجموعه دادگان موجود به زبان انگلیسی است. از بین چند دادگان که به زبان‌های دیگر موجود است، بیشترشان از تصاویر متعلق به مجموعه‌های دادهٔ انگلیسی بهره گرفته‌اند و تنها پرسش و پاسخ و یا توصیف به زبان مورد نظر است [۱۷، ۲۰، ۱۸، ۴۰، ۲۲] و تنها تعداد دادگان محدودی مربوط به وظیفهٔ توصیف تصویر به جمع‌آوری تصاویر از فرهنگ کشور آن زبان دقت ورزیده‌اند [۴، ۳۱]. مشخص است که مدلی که به کمک تصاویر مجموعه دادگان انگلیسی آموزش داده شده باشد، احتمالا به سمت فرهنگ غربی سوگیری دارد و با فرهنگ کشورهای دیگر آشنا نیست، بدین ترتیب دقت کمتری روی تصاویر مربوط به یک کشور خاص خواهد داشت [۳۱].

در این پروژه ما می‌خواهیم دقت دو مدل فارسی و دو مدل انگلیسی را روی تصاویری که از ایران گرفته شده است آزمایش کنیم. هر کدام از مدل‌ها مربوط به یکی از وظایف پرسش و پاسخ تصویری و توصیف تصویر هستند و برای آموزش این مدل ها از تصاویر مجموعه دادگان انگلیسی استفاده شده است. این آزمایش‌ها به صورت یادگیری بدون نمونه و با نمونه معدود اجرا می‌شوند که خود مبحث جدیدی در هوش مصنوعی است.

۱ ـ ۲ ـ ۳ مفهوم یادگیری بدون نمونه و یادگیری با نمونه‌های معدود

شناسایی اشیا [۱] یک اصطلاح عمومی برای مجموعه‌ای از وظایف بینایی کامپیوتر است که نیاز به شناسایی و درک اشیا درون تصاویر دیجیتالی را دارند. شناسایی اشیا یکی از مهم‌ترین گام‌ها برای حل دو مسئلهٔ پرسش و پاسخ تصویری و توصیف تصویر است. در طی سالیان متمادی از یادگیری با ناظر [۲] برای شناسایی اشیا

[۱] Object Recognition
[۲] Supervised Learning

استفاده شده است. اما این نوع یادگیری نیازمند تعداد زیادی دادهٔ برچسب‌دار است.

به خصوص زمانی که تعداد دسته‌های اشیا زیاد شود، ممکن است نتوان تعداد نمونهٔ مناسب برای هر دسته را برای آموزش جمع‌آوری کرد. هم‌چنین در طول زمان، دسته‌های جدیدی از اشیا ممکن است اضافه شود که نمونه‌های مناسبی برای نشان دادن آن دسته وجود نداشته باشد. علاوه بر این، یادگیری با ناظر مفهوم خرد را نادیده می‌گیرد. انسان‌ها می‌توانند براساس توصیفی از یک شی جدید آن را شناسایی کنند که این امر معمولا با در نظر گرفتن شباهت آن شی به مفاهیم از پیش یاد گرفته‌شده است. در صورتی که مدل‌های آموزش‌دیده با روش یادگیری با ناظر متداول، این قابلیت را ندارند. بدین ترتیب این روش‌ها همیشه نمی‌توانند در دنیای واقعی پاسخگوی نیازهای ما باشند.

برای برطرف کردن کاستی‌های یاد شده در روش یادگیری با ناظر، روش دیگری به نام یادگیری بدون نمونه پیشنهاد شده‌است که هدف آن شناسایی دسته‌های از اشیا است که در مرحله آموزش دیده نشده‌اند. این نوع یادگیری تلاش می‌کند که بین دسته‌های دیده‌شده و دیده‌نشده یک پل ارتباطی به کمک اطلاعات معنایی ایجاد کند، که معمولا از توصیف‌های متنی مانند جانمایی[۳] و ویژگی‌های کلمه مشتق می‌شوند. نوع مشابه این یادگیری، یادگیری با تعداد نمونه معدود است که تعداد اندکی از داده‌های برچسب‌دار مربوط به هر دسته از دسته‌های دیده نشده هنگام آموزش به مدل نشان داده می‌شود [۳۲].

با وجود این که بیشتر مدل‌های state of the art به کمک این دو نوع یادگیری نتیجهٔ قابل توجهی نداشتند، چندین مدل که در سال‌های اخیر معرفی شده‌اند به دلیل معماری خاص و متفاوتشان توانسته‌اند دقت‌های خوبی به دست آورند. این نوع مدل‌ها به دلیل داشتن دو رمزگذار[۴] جدا برای تصویر و متن، ویژگی‌های مخصوص به آن‌ها را جداگانه استخراج کرده و با هم مقایسه می‌کنند. یکی از این مدل‌ها که CLIP نام دارد نتایج بسیار جالبی در این دو یادگیری داشته است. ما قصد داریم این مدل از پیش آموزش دیده[۵] را در معماری مدل‌های دو وظیفه مورد نظرمان به کار ببریم تا بتوانیم نتایج بهتری از این دو یادگیری کسب کنیم.

[۳] embedding
[۴] encoder
[۵] Pre-trained model

۱ـ۳ ترتیب نگارش گزارش

گزارش پیش رو در ۷ فصل تنظیم شده‌است. فصل اول (فصل حاضر) به معرفی و بیان مسئلهٔ مورد تحقیق می‌پردازد. در فصل دوم مفاهیم و تعاریفی که در پژوهش به طور مستقیم یا غیرمستقیم از آن‌ها استفاده می‌شود اشاره خواهیم کرد. فصل سوم به بررسی مدل‌های به کار برده شده در پژوهش و کارهای مرتبط می‌پردازد. در فصل چهارم معماری مدل‌های پیشنهادی به طور کامل توضیح داده می‌شوند. فصل پنجم دربارهٔ نحوهٔ پیاده‌سازی، گردآوری داده‌های مورد نیاز برای آزمایش و شرایط آزمایش می‌باشد. در فصل ششم نمونه‌هایی از نتایج آزمایش نشان داده شده و این نتایج همراه با چند نمودار به طور دقیق بررسی و ارزیابی می‌شوند. در فصل پایانی نیز به جمع‌بندی آزمایش‌های انجام شده و بررسی مشکلات آن و راه‌های بهبود عملکرد روش پیشنهادی پرداخته خواهد شد.

فصل ۲

مروری بر مفاهیم اولیه

۲ ـ ۱ مقدمه

در این فصل برخی از مفاهیم را که در طول گزارش به آن‌ها اشاره خواهد شد، معرفی می‌کنیم. ابتدا به شرح دو نوع یادگیری بدون نمونه و با نمونه‌های معدود می‌پردازیم. در ادامه مفهوم یادگیری برپایهٔ محرک‌ها و شیوهٔ اجرای این یادگیری را توضیح می‌دهیم. در پایان با دو معیار ROUGE و BLEU که در این پژوهش برای ارزیابی مدل‌های توصیف تصویر استفاده شده‌اند آشنا خواهیم شد.

۲ ـ ۲ یادگیری بدون نمونه[۱]

یادگیری بدون نمونه نوعی مسئله در یادگیری ماشین است که مدل در زمان آزمون، نمونه‌هایی از کلاس‌هایی را که در طول آموزش مشاهده نشده‌اند می‌بیند و باید کلاس مربوط به آن‌ها را پیش‌بینی کند. در یادگیری بدون نمونه رویکردهای متفاوتی وجود دارد، اما وجه مشترک رویکردها این است که به جای استفاده از برچسب‌های صریح، اطلاعات کمکی مانند توضیحات متنی در طول فرایند آموزش مورد استفاده قرار می‌گیرند. یکی از مدل‌هایی که اخیرا توانسته عملکرد خوبی در یادگیری بدون نمونه داشته باشد، CLIP است که توسط OpenAI ارائه شده است.

به جهت آشنایی بیشتر با این نوع یادگیری، به بررسی نحوه عملکرد یادگیری بدون نمونه در مدل CLIPبرای دسته‌بندی داده می‌پردازیم. هنگامی که مدل به اندازهٔ کافی بر روی جفت تصویر ـ متن آموزش دید، می‌توان از آن برای پیش بینی داده‌های دیده نشده استفاده کرد. در زمان آزمون، ابتدا با به دست آوردن فهرستی از همهٔ برچسب‌های ممکن، دسته بندی معمولی را آغاز می‌کنیم. به طور مثال اگر مسئلهٔ پیش‌بینی گونه‌های جانوری باشد، به فهرستی از همهٔ گونه‌های جانوری نیاز داریم. سپس کدگذاری هر برچسب را توسط کدگذار متنی از قبل آموزش دیده بدست می‌آوریم. بعد باید بردار ویژگی‌های تصویر مورد نظر را با کمک کدگذار تصویر استخراج کنیم. در آخر با استفاده از یک معیار به نام شباهت کسینوسی، شباهت بین بردار هر برچسب را با بردار تصویر محاسبه می‌کنیم. برچسب تصویر ورودی برابر با برچسب تصویری خواهد بود که بیشترین شباهت را با آن داشته باشد. این کار قابل انجام است زیرا مدل یاد گرفته است بیشترین شباهت را میان کدگذاری تصویر و برچسب مربوطه ایجاد کند.

[۱] Zero-shot Learning

۲ـ۳ یادگیری با نمونه‌های معدود[۲]

این مسئله شباهت زیادی به یادگیری بدون نمونه دارد. با این تفاوت که در این نوع یادگیری مدل در طول آموزش با چند نمونهٔ محدود از داده تحت آموزش قرار می‌گیرد. هدف از یادگیری با تعداد نمونه معدود این نیست که مدل تحت تصاویر مجموعه آموزشی آموزش ببیند و سپس به مجموعه آزمون تعمیم دهد؛ بلکه هدف "Learn to learn" است. این نوع از یادگیری در شاخه فرایادگیری[۳] قرار می‌گیرد. ایده اصلی یادگیری با تعداد نمونه معدود، آموزش تابعی است که شباهت را پیش‌بینی می‌کند. پس از آموزش، می‌توان از تابع آموخته‌شده برای پیش‌بینی نمونه‌های دیده نشده استفاده کرد [۲].

۲ـ۴ یادگیری بر پایهٔ محرک[۴]

یک روش مرسوم یادگیری در وظایف پردازش زبان‌های طبیعی این است که یک ورودی x (معمولا از جنس متن) را گرفته و به وسیله یک مدل $P(y|x;\theta)$ خروجی y را پیش بینی می‌کنیم. y می‌تواند از جنس برچسب، متن یا هر جنس دیگری باشد. برای یادگیری پارامتر θ ، از یک مجموعه دادگان متشکل از زوج‌های ورودی و خروجی استفاده می‌کنیم، و یک مدل را به کمک آن‌ها آموزش می‌دهیم تا بتواند احتمال شرطی مورد نظر ما را پیش‌بینی کند.

مشکل اصلی این نوع یادگیری، نیاز به داشتن تعداد زیادی داده برچسب گذاری شده است، در حالی که در بسیاری از وظایف هنوز چنین دادگانی گردآوری نشده است. روش‌های یادگیری براساس محرک تلاش می‌کنند این مشکل را با آموزش یک مدل زبانی که احتمال $P(x;\theta)$ را حدس می‌زند و سپس از این احتمال برای پیش‌بینی y استفاده می‌کند، حل کند. بدین ترتیب نیاز به مجموعه دادگان بزرگ که تحت نظارت گردآوری شده را کاهش می‌دهد. مدل پایه‌ای پرامپت بهترین y را در سه مرحله پیش‌بینی می‌کند.

[۲] Few-shot Learning

[۳] فرایادگیری یا Meta-learning با یادگیری سنتی تحت نظارت متفاوت است. یادگیری نظارت شده سنتی از مدل می‌خواهد که داده‌های آموزشی را تشخیص دهد و سپس به داده‌های آزمایشی دیده نشده تعمیم دهد. اما الگوریتم‌های فرایادگیری از خروجی سایر الگوریتم‌های یادگیری ماشینی که خود از داده‌ها یاد می‌گیرند، تحت آموزش قرار می‌گیرند. این بدان معناست که فرایادگیری مستلزم حضور سایر الگوریتم‌های یادگیری است که قبلاً روی داده‌ها آموزش دیده‌اند [۳].

[۴] Prompt-based Learning

۲ـ۴ـ۱ اضافه کردن پرامپت

در این مرحله یک تابع پرامپت $f_{prompt}(.)$ روی ورودی x اعمال می‌شود تا پرامپت $x' = f_{prompt}(x)$ به دست آید. در بیشتر کارهای انجام شده در این زمینه [۲۱، ۲۴، ۲۹] این تابع دو مرحله دارد:

۱. اعمال کردن یک قالب، که یک رشته متنی با دو جای خالی است، یک جای خالی [x] برای ورودی x و یک جای خالی [Z] برای جواب Z که در نهایت به خروجی y نگاشت می‌شود.

۲. پر کردن جای خالی ورودی با متن ورودی x

برای مثال اگر وظیفه پرسش و پاسخ تصویری را در نظر بگیریم، ورودی متنی x یک پرسش است مانند "هوا در تصویر چگونه است؟". قالب را به دلخواه رشته‌ای مانند "پرسش: [x] پاسخ: [Z]" در نظر می‌گیریم. بدین ترتیب حاصل تابع پرامپت برای پرسش مطرح شده رشته "پرسش: هوا در تصویر چگونه است؟ پاسخ: [Z]" خواهد بود. ممکن است قالب حتما از توکن‌های زبان طبیعی تشکیل نشده باشد. قالب‌ها می‌توانند دارای کلمات مجازی (شماره به جای کلمه) یا حتی بردارهایی که به وسیله روشی محاسبه شده‌اند باشند [۳۷]. لازم به ذکر است که قالب می‌تواند با توجه به مسئله بیشتر از یک جای خالی برای ورودی یا جواب داشته باشد.

۲ـ۴ـ۲ جستجوی پاسخ

در مرحله بعدی ما به دنبال جواب z ای می‌گردیم که امتیاز مدل زبانی را بیشینه کند. مجموعه Z مجموعه مقادیر ممکن برای پاسخ‌هاست. این مجموعه می‌تواند تمام زبان تعریف شده روی یک الفبا یا واژگان در وظایف تولید کننده دنباله، یا مجموعه محدودی از کلمات در مسائل دسته‌بندی باشد. برای مثال در وظیفه تحلیل احساسات [۶] که خروجی y از مجموعه {بسیار مثبت، مثبت، خنثی، منفی، بسیار منفی} انتخاب می‌شود مجموعه Z می‌تواند به شکل Z = {"عالی"، "خوب"،"متوسط"،"ناخوشایند"،"بسیار ناخوشایند"} تعریف شود.

در ادامه یک تابع $f_{fill(x',z)}$ که مکان [Z] را در x' با جواب z پر می‌کند تعریف می‌کنیم. به پرامپت‌های حاصل، پرامپت پر شده [۷] می‌گوییم. هم چنین، پرامپتی که با پاسخ درست پر شده است را پرامپت پاسخ داده

[۶] Sentiment Analysis
[۷] filled prompt

شده [۸] می‌خوانیم. در نهایت بین تمامی پاسخ‌های ممکن z جستجویی با محاسبه احتمال پرامپت پر شده آن‌ها به کمک یک مدل زبانی آموزش دیده شده (۲ـ۱) انجام می‌دهیم. تابع جستجو می‌تواند یک argmax باشد که پاسخ با بیشترین احتمال را انتخاب می‌کند، و یا تابع نمونه‌برداری باشد که به طور تصادفی پاسخ را از روی توزیع احتمال مدل زبانی تولید می‌کند.

$$\hat{z} = \underset{z \in Z}{search} P(f_{fill}(x', z); \theta). \qquad (۲ـ۱)$$

در مثال مربوط به مسئله پرسش و پاسخ تصویری، با توجه به این که به طور معمول این وظیفه یک مسئله دسته‌بندی در نظر گرفته می‌شود، تمام دسته‌های ممکن به صورت متنی، مجموعه Z را تشکیل می‌دهند. اگر دسته‌های این مسئله به طور مثال طبق جدول ۲ـ۱ به شکل متنی برابر با {"آفتابی"، "ابری"، "بارانی"، "بله"،"خیر"} باشند، این مجموعه تمام z های ممکن را نشان می‌دهد. بدین ترتیب، ابتدا پرامپت‌های پر شده مربوط به هر پاسخ را تشکیل داده و سپس با توجه به رابطه ۲ـ۱ بهترین پاسخ را انتخاب می‌کنیم.

۲ـ۴ـ۳ نگاشت پاسخ

در انتها، قصد داریم از جواب انتخاب شده z_{hat} به خروجی y_{hat} برسیم. در بعضی از مسائل، z_{hat} می‌تواند همان y_{hat} باشد (مانند مسائل تولید زبانی اعم از مسئله ترجمه و ...)، اما در بعضی مسائل این گونه نیست و نیاز است از جواب z_{hat} به خروجی نگاشت داده شود. در مثالی که بررسی کردیم، اگر خروجی را شماره دسته مورد نظر در نظر بگیریم، می‌توان از جواب z_{hat} پیدا شده به شمارهٔ آن رسید. اما اگر متن پاسخ مقصود ما باشد، z_{hat} خروجی ماست.

[۸] answered prompt

نام	نشانه‌گذاری	مثال	توضیح
ورودی	x	هوا در تصویر چگونه است؟	یک یا چند متن
خروجی	y	آفتابی	برچسب خروجی یا متن
تابع پرامپت	$f_{prompt}(x)$	پرسش: [x] پاسخ: [Z]	تابعی که با درج ورودی x و اضافه کردن یک [Z]، ورودی را به فرمی خاص تبدیل می‌کند. [Z] ممکن است بعدا با پاسخ z پر شود.
پرامپت	x'	پرسش: هوا در تصویر چگونه است؟ پاسخ: [Z]	متنی که در آن [X] با ورودی x جایگزین شده اما اسلات [Z] که پاسخ است، جایگزین نشده است.
پرامپت پر شده	$f_{fill(x',z)}$	پرسش: هوا در تصویر چگونه است؟ پاسخ: [ابری]	پرامپتی که در آن اسلات [Z] با هر پاسخی پر شده است.
پرامپت پاسخ داده شده	$f_{fill(x',z^*)}$	پرسش: هوا در تصویر چگونه است؟ پاسخ: [آفتابی]	پرامپتی که در آن اسلات [Z] با پاسخ صحیح پر شده است.
پاسخ	z	"آفتابی"، "ابری"، "بارانی"، "بله"، "خیر"	توکن، عبارت یا جمله‌ای که [Z] را پر می‌کند.

جدول ۲ ـ ۱: اصطلاحات و نشانه‌گذاری روش‌های محرک. z^* پاسخ‌هایی را نشان می‌دهد که با خروجی واقعی y^* مطابقت دارند.

۲ ـ ۵ معیارهای به کار رفته برای ارزیابی مدل توصیف تصویر

۲ ـ ۵ ـ ۱ معیار BLEU [۱۰]

معیاری برای ارزیابی یک جمله تولید شده نسبت به یک جمله مرجع است. در صورت تطابق کامل امتیاز ۱ و عدم تطابق کامل منجر به امتیاز ۰ می شود.

BLEU برای ارزیابی پیش‌بینی‌های انجام شده توسط سیستم‌های ترجمه ماشینی خودکار ایجاد شد. مزیت‌های آن عبارتند از: [۱۲]

- محاسبه سریع و درک آن آسان است.
- با روشی که یک انسان متن را ارزیابی می‌کند مطابقت دارد.
- مهمتر از همه، مستقل از زبان است و به همین دلیل استفاده از آن در مدل‌های NLP آسان است.
- حتی زمانی که بیش از یک جمله مرجع داشته باشیم می توان از آن استفاده کرد.
- بسیار گسترده استفاده می شود. در نتیجه مقایسه نتایج با کارهای دیگر را آسان تر می‌کند.

این امتیاز با وجود مشهور بودن، معایبی نیز دارد که عبارتند از:

- معنای کلمات را در نظر نمی‌گیرد. برای یک انسان کاملاً قابل قبول است که از یک کلمه متفاوت با همان معنی استفاده کند. به طور مثال می‌توان به جای «ماشین» از «خودرو» استفاده کرد اما امتیاز

[۱۰] Bilingual Evaluation Understudy

Bleu آن را کلمه‌ای نادرست می‌داند.

- فقط به دنبال تطابق دقیق کلمات است. گاهی اوقات می توان از یک نوع از همان کلمه استفاده کرد، به عنوان مثال "rain" و ،"raining" اما این امتیاز آن را خطا در نظر می‌گیرد.
- اهمیت کلمات را نادیده می‌گیرد. مثلا یک کلمه نادرست مانند "به" یا "با" که کمتر مرتبط با جمله است، به اندازه کلمه‌ای که به طور قابل توجهی به معنای جمله کمک می‌کند، در محاسبه امتیاز تاثیرگذار است.
- ترتیب کلمات را در نظر نمی‌گیرد. جمله "نگهبان به دلیل باران دیر رسید" و "باران به دلیل نگهبان دیر رسید" امتیاز (unigram Bleu) یکسانی دریافت می‌کنند در صورتی که جمله دوم از نظر معنا صحیح نیست.

نحوه محاسبه: این رویکرد با شمارش تعداد تطبیق n-gram ها در ترجمه کاندید و n-gram ها در متن مرجع کار می کند، که در آن 1-gram برای هر توکن و bigram یک جفت کلمه خواهد بود. مقایسه بدون توجه به ترتیب کلمات انجام می شود. فرمول محاسبه آن به صورت زیر است: [۱۶]

$$\begin{aligned}\text{Geometric Average Precision (N)} &= \exp(\sum_{n=1}^{N} w_n \log p_n) \\ &= \prod_{n=1}^{N} p_n^{w_n} \\ &= (p_1)^{\frac{1}{4}}.(p_2)^{\frac{1}{4}}.(p_3)^{\frac{1}{4}}.(p_4)^{\frac{1}{4}}\end{aligned}$$

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if c > r} \\ e^{(1-\frac{r}{c})}, & \text{if c <= r} \end{cases} \qquad (۲ ـ ۲)$$

$$\text{Bleu (N)} = \text{Brevity Penalty.Geometric Average Precision Scores (N)} \quad (۳ـ۲)$$

۲ـ۵ـ۲ معیار ROUGE[۱۲]

مجموعه‌ای از معیارهاست که به طور خاص برای ارزیابی خلاصه‌سازی خودکار طراحی شده است، اما می‌تواند برای سایر وظایف نیز استفاده شود.

ROUGE-N تعداد تطبیق n-gram ها بین متن تولید شده توسط مدل و مرجع تولید شده توسط انسان را اندازه‌گیری می‌کند. برای هر N می‌توانیم معیارهای precision و recall و F1-score را محاسبه کنیم.

ROUGE-L بر اساس طولانی‌ترین دنباله مشترک بین خروجی مدل و مرجع است، یعنی طولانی‌ترین دنباله کلمات (نه لزوماً متوالی، اما همچنان به ترتیب) که بین هر دو مشترک است. یک دنباله به اشتراک گذاشته شده طولانی تر باید نشان دهنده شباهت بیشتر بین دو دنباله باشد. [۱۴]

[۱۲] Recall-Oriented Understudy for Gisting Evaluation

فصل ۳

مروری بر منابع

۳ـ۱ مقدمه

در این فصل ابتدا به مدل‌های GPT-2 و CLIP که در پژوهش به کار برده شده می‌پردازیم. در ادامه به عملکرد عالی مدل CLIP در یادگیری بدون نمونه و با نمونه‌های معدود می‌پردازیم که دلیل به کار بردن این مدل را در پژوهش مورد نظر مشخص می‌کند. در پایان به چند کار مرتبط برای مسائل پرسش و پاسخ تصویری و توصیف تصویر اشاره خواهیم کرد و روش کار آن‌ها را توضیح خواهیم داد.

۳ـ۲ مدل GPT-2

GPT-2 یک مدل زبانی [۱] است که برای پیش‌بینی کلمه بعدی بر روی ۴۰ گیگابایت متن اینترنتی آموزش داده شده است. این مدل بزرگ مبتنی بر ترنسفورمر دارای 1.5 میلیارد پارامتر است که بر روی یک دادگان از ۸ میلیون صفحه وب با یک هدف ساده آموزش داده می شود: پیش‌بینی کلمه بعدی با توجه به تمام کلمات قبلی در متن. تنوع دادگان آموزشی باعث می‌شود عملکرد خوبی در مسائل مختلف پردازش زبان طبیعی داشته باشد.

این مدل قابلیت‌های فراوانی دارد. به طور مثال قادر است با دریافت چند کلمه به عنوان ورودی، یک نمونه متن مصنوعی با کیفیت بی‌سابقه به عنوان ادامهٔ آن چند کلمهٔ ورودی تولید کند. علاوه بر این، GPT-2 از سایر مدل‌های زبانی که در حوزه‌های خاص آموزش داده شده‌اند (مانند ویکی‌پدیا، اخبار یا کتاب‌ها) بهتر عمل می‌کند. در کارهای زبانی مانند پرسش و پاسخ، درک مطلب، خلاصه‌نویسی و ترجمه، GPT-2 شروع به یادگیری از متن خام و بدون استفاده از داده‌های آموزشی مختص به هر وظیفه می‌کند. با وجود اینکه امتیازات مربوط به این وظایف پایین‌دستی [۲] با مدل‌های دارای آخرین فن‌آوری بسیار فاصله دارد، آنها نشان می‌دهند که این وظایف می‌توانند با استفاده از داده‌های کافی (بدون برچسب) از تکنیک‌های یادگیری بدون ناظر بهره ببرند. [۲۷]

[۱] مدل زبانی یک مدل یادگیری ماشینی است که برای نشان دادن حیطه زبان طراحی شده است. می توان از آن برای تعدادی از وظایف مختلف مبتنی بر زبان از جمله پرسش و پاسخ، خلاصه‌سازی و ترجمه ماشینی استفاده کرد.

[۲] downstream tasks

۳ـ۳ مدل‌های V&L

وظایف بینایی‌ـ‌و‌ـ‌زبان [۳] که به اختصار وظایف V&L نامیده می‌شوند، توانایی یک سیستم را در درک و استنتاج معنایی دنیای بصری به کمک یک زبان طبیعی آزمایش می‌کنند. هر دو وظیفه پرسش و پاسخ تصویری و توصیف تصویر در این دسته وظایف قرار می‌گیرند. مدل‌هایی که برای این وظایف طراحی می‌شوند اغلب یک رمزگذار تصویری [۴] ، یک رمزگذار متنی [۵] و یک ماژول cross-modal برای ارتباط این دو با هم دارند. [۳۵]

۴ـ۳ مدل CLIP

مدل CLIP [۶] یک شبکه عصبی است که مفاهیم بصری را از نظارت زبان طبیعی یاد می گیرد. شامل یک کدگذار تصویر و یک کدگذار متن است که با استفاده از تابع هزینه تقابلی [۷] آموزش داده می شود و در خروجی پیش‌بینی می کند که کدام تصاویر با کدام متن در دادگان منطبق شده‌اند. کدگذار تصویر و کدگذار متن هر کدام به طور جداگانه، ویژگی‌های تصویر و متن ورودی را استخراج می‌کنند.سپس ضرب داخلی [۸] خروجی‌های این دو کدگذار به عنوان امتیاز هم‌ترازی [۹] محاسبه می‌شود و از آن برای اندازه‌گیری شباهت متن و تصویر ورودی استفاده می‌شود. برای آموزش CLIP از ۴۰۰ میلیون زوج تصویرـ متن یافته شده از اینترنت استفاده شده است.

از مدل CLIP ، می‌توانیم برای حل مسئله طبقه‌بندی تصاویر با استفاده از یادگیری بدون نمونه استفاده کنیم. برای این کار کافیست تمام دسته‌های تصاویر را به توضیحاتی مانند «a photo of a dog» تبدیل کنیم. سپس میزان شباهت تصویر مورد نظر را با هر یک از توضیحات دسته‌ها بدست آوریم. بیشترین میزان شباهت تصویر با هر یک از توضیحات، نشان‌دهنده ی دسته آن تصویر خواهد بود.

مدل CLIP توانسته است چند مشکل اساسی در حوزه بینایی ماشین را حل کند:

[۳] Vision-and-Language tasks
[۴] Vision encoder
[۵] Text Encoder
[۶] Contrastive Language–Image Pre-training
[۷] Contrastive loss
[۸] dot product
[۹] Alignment score

- دادگان پرهزینه: رویکرد یادگیری عمیق به داده‌های زیادی نیاز دارد و مدل‌های بینایی معمولا بر روی دادگان‌هایی که به صورت دستی برچسب‌گذاری شده‌اند آموزش داده می‌شوند. ایراد این دادگان‌ها آن است که ساخت آن‌ها پرهزینه است و تنها تعداد محدودی از مفاهیم بصری از پیش تعیین‌شده را پشتیبانی می‌کند. در مقابل، CLIP از جفت‌های متن‌ـ‌تصویر که به صورت عمومی در اینترنت در دسترس هستند، برای یادگیری استفاده می‌کند.

- محدودیت در حل مسائل گوناگون: یک مدل ImageNet در صورتی که به خوبی آموزش دیده باشد می‌تواند عملکرد مناسبی در پیش‌بینی ۱۰۰۰ دسته‌بندی موجود در این مجموعه داده داشته باشد. اما اگر همین مدل بخواهد مسئله دیگری را حل کند، در ابتدا یک متخصص باید دادگان جدید بسازد، سپس یک لایه خروجی جدید اضافه کند و مدل را تنظیم دقیق [۱۰] کند. در مقابل، CLIP را می‌توان برای انجام طیف گسترده‌ای از مسائل دسته بندی تصویر بدون نیاز به نمونه‌های آموزشی اضافی تطبیق داد. برای استفاده از CLIP در یک مسئله جدید، تنها کاری که باید انجام دهیم این است که نام مفاهیم بصری را به کدگذار متنی CLIP بگوییم، و CLIP یک دسته‌بندی خطی از بازنمایی‌های بصری را تولید می‌کند. دقت این دسته‌بند اغلب با مدل‌های کاملاً نظارت شده قابل رقابت است.

- نتیجه ضعیف در دنیای واقعی: اغلب گزارش می‌شود که سیستم‌های یادگیری عمیق به عملکردی مشابه انسان دست پیدا کرده‌اند اما زمانی که در دنیای واقعی مورد آزمایش قرار می‌گیرند، عملکرد آن‌ها می‌تواند بسیار کمتر از انتظار تعیین شده توسط معیارها باشد. به عبارت دیگر، بین «عملکرد معیار» و «عملکرد واقعی» فاصله وجود دارد. مدل CLIP قابلیت ارزیابی بر روی معیارها بدون آموزش دیدن روی داده‌های آن‌ها را دارد. [۲۸]

۳ـ۴ـ۱ نتایج مدل CLIP در یادگیری بدون نمونه برای دسته‌بندی تصاویر

در این حوزه CLIP به نتایج فوق‌العاده‌ای دست می‌یابد، و دقت آزمون یادگیری بدون نمونه را در دادگان ImageNet از ٪۱۱٫۵ به ٪۷۶٫۲ بهبود می‌بخشد.

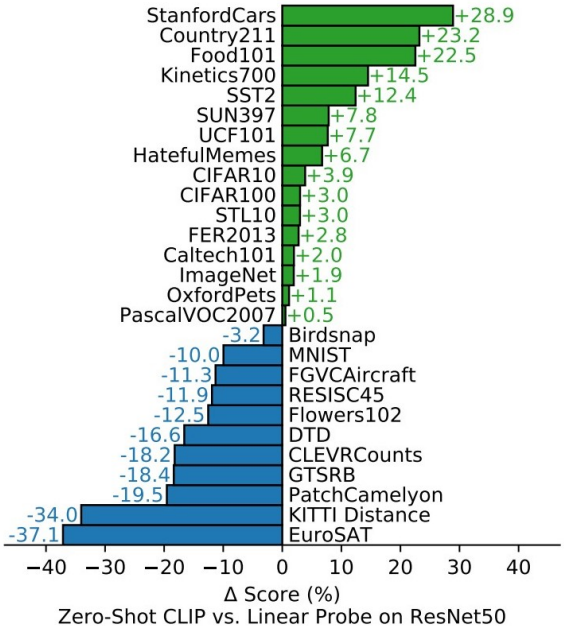
وقتی عملکرد یادگیری بدون نمونه CLIP با یک دسته بند خطی نظارت شده که با مدل ResNet50 از پیش آموزش‌دیده مقایسه می‌شود، CLIP به نتایج قابل‌توجهی در مجموعه داده‌های مختلف می‌رسد. این

[۱۰] fine-tune

SUN	ImageNet	aYahoo	
۲۳/۰	۱۱/۵	۷۲/۴	Visual N-Grams
۵۸/۵	۷۶/۲	۹۸/۴	CLIP

جدول ۳ـ۱: مقایسه CLIP با نتایج قبلی دسته‌بندی تصویر یادگیری بدون نمونه. CLIP عملکرد هر سه مجموعه داده را مقدار زیادی بهبود می‌بخشد.

مدل در ۱۶ مورد از مجموع ۲۷ دادگانی که مورد مطالعه قرار گرفت عملکرد بهتری دارد، علی‌رغم اینکه هرگز حتی یک نمونه آموزشی از این مجموعه‌داده‌ها را مشاهده نکرده است.

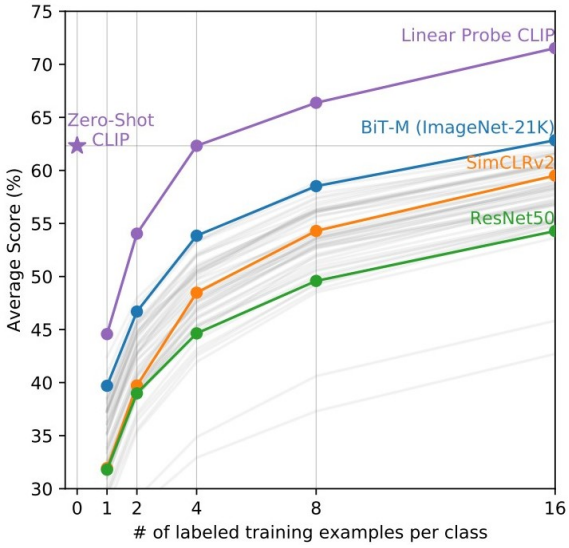


شکل ۳ـ۱: مدل CLIP در یادگیری بدون نمونه، با مدل‌های کاملاً نظارت شده قابل رقابت است.

مطابق تصویر ۳ـ۱ هنگامی که عملکرد روی هر یک از این دادگان‌ها تجزیه و تحلیل می‌شود، مشخص می شود که CLIP در دادگان‌های طبقه‌بندی عمومی اشیا (به عنوان مثال ImageNet یا CIFAR10/100) و حتی در دادگان‌های تشخیص فعالیت بهتر عمل می‌کند. در مقابل، در دادگان‌های پیچیده و تخصصی مانند طبقه‌بندی تصاویر ماهواره‌ای و تشخیص تومور ضعیف‌ترین عملکرد را دارد.

۳ـ۴ـ۲ نتایج مدل CLIP در یادگیری با تعداد نمونه معدود برای طبقه بندی تصاویر

عملکرد یادگیری بدون نمونه در مدل CLIP با میانگین عملکرد دسته‌بندهای خطی با تعداد نمونه معدود ـ پس از مشاهده چهار مثال آموزشی در هر کلاس ـ مطابقت دارد. علاوه بر این، CLIP از همه دسته بندهای خطی با تعداد نمونه معدود بهتر عمل می‌کند. این نتایج در شکل ۳ـ۲ خلاصه شده است. [۳۹]



شکل ۳ـ۲: مدل CLIP با یادگیری بدون نمونه بهتر از probe های خطی یادگیری با تعداد نمونه معدود است.

۳ـ۴ـ۳ به‌کارگیری CLIP در مسئله پرسش و پاسخ تصویری

در مقاله [۳۵] مدلی به نام CLIP-VIL ارائه شده است که کدگذار تصویر دو مدل Pythia و MCAN را با کدگذارتصویر CLIP جایگزین کرده است. مدل Pythia [۱۹] به کمک یک attention-based GRU پرسش را رمزگذاری کرده و خروجی به دست آمده را با یک multi-modal factorized bilinear pooling network ترکیب می‌کند. مدل MCAN [۴۲] یک LSTM به عنوان کدگذار پرسش و یک encoder-decoder based modular co-attention network را با ورودی رمزگذاری‌های تصویر و پرسش دارد. هر دو مدل یک دسته‌بند خروجی [۱۱] در انتهای شبکه‌شان دارند. هر کدام از این مدل‌ها با انواع مختلف کدگذار تصویر CLIP آزمایش شده‌اند.

[۱۱] Output classifier

طبق نتایج این مقاله، بیشتر مدل‌هایی که از کدگذار تصویر CLIP استفاده کرده اند، عملکرد بهتری نسبت به کدگذارهای تصویری این مدل‌ها (که همه نسخه‌ای از ImageNet-ResNet هستند) داشتند. بهترین مدل CLIP-VIL مدلی‌ست که از کدگذار تصویر CLIP-Res50x4 استفاده می‌کند. همچنین در مقاله [۳۶] برای استفاده از CLIP در وظیفه پرسش و پاسخ تصویری، یک پرامپت ایجاد می‌کنند. تولید پرامپت در دو مرحله صورت می‌گیرد:

۱. ابتدا پرسش را به یک قالب [۱۲] تبدیل می‌کنند که یک عبارت با یک توکن پنهان (masked) است. برای تولید این قالب دو روش به کار برده‌اند. اولین روش استفاده از مدل از پیش آموزش دیده T5 [۳۰] است که با نمونه‌هایی از تبدیل پرسش به قالب آموزش داده شده‌است. روش دوم، تجزیه وابستگی [۱۳] ، پرسش را به کمک برچسب‌زنی بخش گفتار [۱۴] و نتایج تجزیه ساختار جمله پرسش، با قوانین دستور زبان به قالب مورد نظر تبدیل می‌کند. این روش در معدود نمونه‌هایی که مدل T5 نتیجه قابل قبولی برای آن‌ها نداشت به کار برده شد. بدین ترتیب که برای پیش‌بینی این مدل یک حد اطمینان تعیین شد و اگر میزان اطمینان به یک قالب تولید شده توسط مدل کمتر از این حد بود از روش دوم برای ساختن قالب استفاده کردند.

۲. از بین مجموعه لغات [۱۵] یا دسته‌های مشخص شده برای پاسخ‌ها، قرار دادن بعضی از کلمات به جای پاسخ در قالب تولید شده از نظر معنایی نادرست خواهد بود. به کمک مدل‌های زبانی از پیش آموزش دیده می‌توان این کلمات را از مجموعه پاسخ‌های قابل قبول خارج کرد. سپس تمامی پاسخ‌های دیگر را به جای توکن پنهان شده قرار داده تا تمامی پرامپت‌ها برای یک پرسش به دست آیند.

پس از تولید پرامپت‌ها برای هر پرسش، مدل CLIP را به صورت یادگیری بدون نمونه روی آن‌ها آزمایش کردند. به دلیل تولید یک عبارت به عنوان پرامپت از روی پرسش و پاسخ، که بیشتر مشابه متن‌هاییست که CLIP در هنگام آموزش مشاهده کرده است، نتایج قابل قبولی از یادگیری بدون نمونه حاصل شد.

[۱۲] Template
[۱۳] Dependency Parsing
[۱۴] Part-of-Speech tagging
[۱۵] Vocabulary

۳ـ۴ـ۴ به‌کارگیری مدل CLIP در مسئله توصیف تصویر

آزمایش CLIP با روش SCST [۱۶] : روش SCST یک راهکار رایج برای آموزش مدل‌های تولیدکننده دنباله [۱۷] است. این روش برای بهینه‌سازی دنباله‌های تولید شده بر یک هدف مشتق ناپذیر، عموما معیارهای ارزیابی، مانند CIDEr برای همین وظیفه است. مقاله [۲۳] نوع بهتری از SCST را ارائه می‌دهد و مقاله [۳۵] از این روش آموزش برای آزمودن عملکرد CLIP در وظیفه توصیف تصویر استفاده می‌کند. معماری مدل یک مدل ترنسفورمر پایه با اقتباس از [۳۸] است. مطابق جدول ۳ـ۳ مدل‌های CLIP با کدگذارهای تصویری ResNet50 و ResNet101 از تمامی مدل‌های ImageNet از پیش آموخته شده بهتر عمل کرده‌اند. هر چه اندازه مدل‌های CLIP بزرگتر شده‌اند، نتایجشان نیز پیشرفت داشته و مدل CLIP-Res50x4 بهترین عملکرد را از خود نشان داده است. با این حال، نتایج CLIP با نتایج مدل‌هایی که مرحله pre-training تعاملی تصویری‌ـ متنی داشته‌اند (مانند $Oscar_{base}$ و $VinVL_{base}$) فاصله دارد. [۳۵]

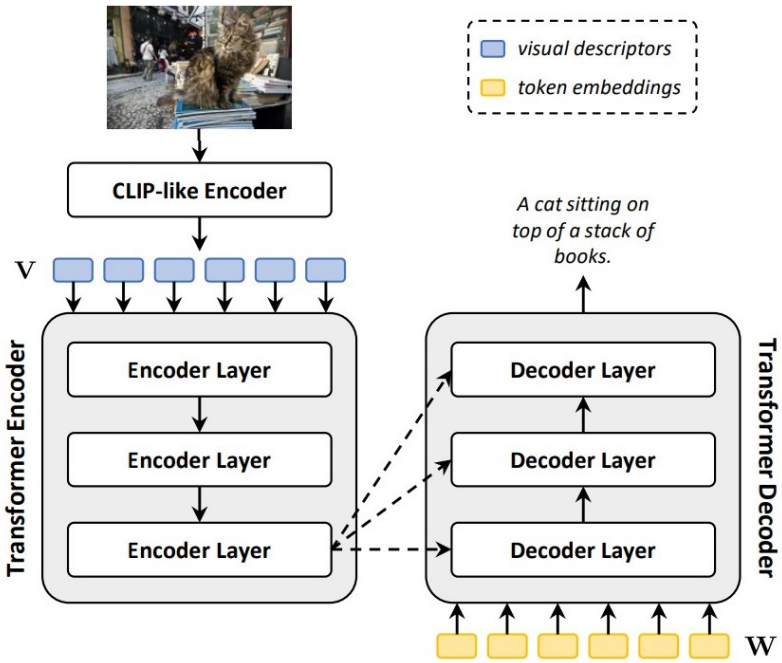
Model	B@4	M	C	S
BUTD (Anderson et al., 2018a)	36.3	27.7	120.1	21.4
VLP (Zhou et al., 2020)	39.5	29.3	129.8	22.4
AoANet (Huang et al., 2019b)	38.9	29.2	129.8	22.4
$\text{Oscar}_{\text{base}}$ (Li et al., 2020)	40.5	29.7	137.6	22.8
$\text{VinVL}_{\text{base}}$ (Zhang et al., 2021)	**40.9**	**30.9**	**140.4**	**25.1**
$\text{BUTD}_{\text{Transformer}}$* (Luo, 2020)	-	-	**127.7**	**22.5**
$\text{ImageNet-Res50}_{\text{Transformer}}$	36.2	27.6	118.8	21.2
$\text{ImageNet-Res101}_{\text{Transformer}}$	36.8	27.8	121.1	21.5
$\text{CLIP-Res50}_{\text{Transformer}}$	38.6	28.8	127.9	22.7
$\text{CLIP-Res101}_{\text{Transformer}}$	39.2	29.1	130.3	23.0
$\text{CLIP-Res50x4}_{\text{Transformer}}$	**40.2**	**29.7**	**134.2**	**23.8**
$\text{CLIP-ViT-B}_{\text{Transformer}}$	21.1	19.4	58.0	12.2

شکل ۳ـ۳: نتایج مسئله توصیف تصویر

CLIP-Captioner : این مدل یک مدل کدگذارـ کدگشا مطابق شکل ۳ـ۴ است که از یک کدگذار CLIP برای به دست آوردن ویژگی‌های تصویر استفاده کرده و آن را به یک کدگذار و سپس یک کدگشا برای تولید توصیف تصویر می‌دهد. کدگشا در هر گام زمانی یک توکن تولید می‌کند و دنباله توکن‌ها، توصیف

[۱۶] Self-Critical Sequence Training
[۱۷] Sequence Generation Model

آن تصویر خواهد بود. مدل ارائه شده با معیار CIDEr و همچنین آموزش Cross-Entropy ارزیابی شده است. مقایسه انواع مدل‌های CLIP با مدل‌های تصویری دیگر مانند Faster R-CNN به کار رفته در مدل CLIP-Captioner نشان می‌دهد ویژگی‌های استخراج شده از کدگذارهای تصویری CLIP که هنگام pre-training از مقایسه بین تصویر و متن یاد گرفته شده‌اند بسیار موثرتر از ویژگی‌هاییست که از مدل‌های آموزش دیده روی تصاویر به دست می‌آیند. بدین ترتیب کدگذارهای تصویر CLIP از بهترین مدل‌ها در استخراج ویژگی‌های تصویر است. [۱۰]



شکل ۳ـ۴: مروری بر معماری توصیف تصویر مبتنی بر CLIP

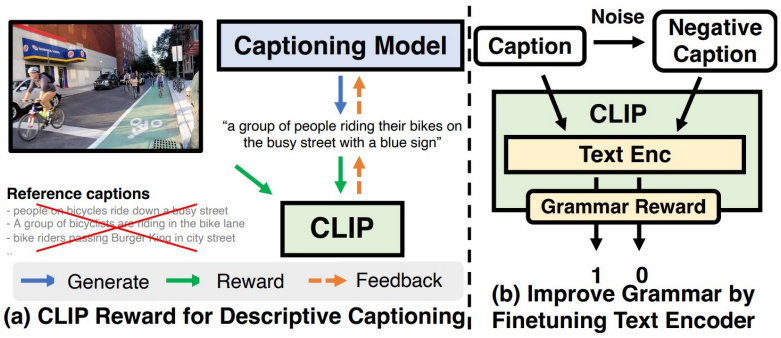
CLIP-S و CLIP-S+Grammar : در مقاله [۱۵] از امتیاز شباهت تصویر و متن CLIP برای آموزش یک مدل توصیف تصویر استفاده کرده‌اند، بدین ترتیب که CLIP-S را به عنوان پاداش طبق رابطه ۳ـ۱ تعریف می‌کنند. در این رابطه، I و c تصویر و توصیف آن، و f^I و f^T کدگذارهای تصویر و متن CLIP هستند و $w = ۲/۵$ می‌باشد. با یادگیری بیشینه‌سازی شباهت متن و تصویر یک مدل مقابله‌ای [۱۸] ، یک مدل توصیف تصویر به سمت تولید توصیف‌هایی که اطلاعات بیشتری از تصویر را دارند تشویق می‌شود. روش یادگیری در قسمت الف شکل ۳ـ۵ نشان داده شده است.

[۱۸] Contrastive model

به این دلیل که مدل CLIP با هدف مدل‌سازی زبان آموزش ندیده است، آموزش مدل توصیف تصویر به تنهایی با امتیاز CLIP-S اغلب به ایجاد توصیف‌هایی با دستور زبان اشتباه منجر می‌شود. برای حل این مشکل، یک شبکه عصبی کوچک دو لایه‌ای تشکیل داده که خروجی کدگذار متنی CLIP به نام c را به عنوان ورودی می‌گیرد و احتمال درست بودن آن (g(c)) را از جنبه دستور زبانی بررسی می‌کند. همچنین از توصیف‌های مرجع چندین توصیف منفی را با تکرار عملیات حذف، اضافه، جابه‌جایی، تکرار و به هم ریختن توکن‌های توصیف مرجع به صورت تصادفی تولید می‌کنند. به کمک توصیف‌های تولید شده کدگذار متنی CLIP را همراه با شبکه دو لایه مربوط به دستور زبان تنظیم [۱۹] می‌کنند. با داشتن این شبکه عصبی، پاداش مربوط به آموزش مدل توصیف تصویر به رابطه ۳ـ۲ تبدیل می‌شود.

نتایج به‌کارگیری این پاداش‌ها را در آموزش مدل توصیف تصویر در شکل ۳ـ۶ مشاهده می‌کنید.

$$\text{CLIP-S(I, c)} = w * max(\frac{f^I(I)^\intercal f^T(C)}{|f^I(I)|.|f^T(C)|}, ۰) \qquad (۱ـ۳)$$



شکل ۳ـ۵: مروری بر روش توصیف تصویر مدل CLIP-S

$$R(I, c) = \text{CLIP-S}(I, c) + \lambda g(c), \text{where } \lambda = ۲/۰ \qquad (۲ـ۳)$$

[۱۹] Fine-tune

Reward	N-gram based				Embed based			Text-to-Image Retrieval			FineCapEval			
	Text based					Image-Text based					Overall	Bg.	Obj.	Rel.
	BLEU-4	CIDEr	METEOR	ROUGE-L	BERT-S	CLIP-S	RefCLIP-S	R@1	R@5	R@10	CIDEr	R_{word}	R_{word}	R_{word}
Reference captions								29.5*	54.2*	65.0*				
MLE	32.5	110.3	27.2	55.2	0.937	0.758	1.12	21.8	45.6	58.0	13.5	11.6	13.0	19.8
CIDEr	**38.2**	**124.9**	28.7	**58.5**	**0.942**	0.759	1.13	20.9	45.6	58.2	12.8	13.1	23.1	22.4
CLIP-S	6.2	11.2	18.7	31.6	0.882	**0.860**	**1.17**	**42.5**	**71.6**	**82.2**	13.9	20.8	**26.4**	24.9
CIDEr+CLIP-S	37.7	124.6	**28.8**	58.3	0.941	0.772	1.14	24.4	50.2	63.1	13.0	13.0	23.4	21.7
CLIP-S+Grammar	16.9	71.0	24.9	47.3	0.924	0.793	1.15	35.8	64.0	75.8	**19.3**	**21.8**	25.5	**27.5**

شکل ۳ـ۶: عملکرد مدل CLIP-S در مسئله توصیف تصویر با در نظر گرفتن پاداش‌های مختلف

فصل ۴

شرح روش پیشنهادی

۴ـ۱ مقدمه

در این فصل به شرح کامل معماری مدل‌های به کار برده شده برای دو مسئلهٔ پرسش و پاسخ تصویری و توصیف تصویر می‌پردازیم و طرز به کارگیری محرک‌ها در مسئلهٔ پرسش و پاسخ تصویری و شیوهٔ ایجاد یک مدل توصیف تصویر به کمک دو مدل از پیش آموزش دیدهٔ CLIP و GPT-2 را توضیح می‌دهیم.

۴ـ۲ مدل CLIP برای وظیفه پرسش و پاسخ تصویری

با توجه به این که مدل CLIP تنها یک کدگذار متنی و یک کدگذار تصویر دارد، برای به کارگیری این مدل در وظیفهٔ پرسش و پاسخ تصویری نیاز به یادگیری بر پایهٔ پرامپت داریم. بدین ترتیب، لازم نیست دسته‌های پاسخ را از خروجی مدل به دست آوریم و این پاسخ‌ها درون قالب پرامپت جاسازی می‌شوند. برای زبان فارسی از سه قالب "پرسش: [Q] پاسخ: [A]" ، قالب "[Q] [A]" و قالب "[A] پاسخ [Q] است." و برای زبان انگلیسی از سه قالب "Retrieve [A] from [Q]"، "[Q] [A]" و "Question: [Q] Answer: [A]"را به عنوان قالب پرامپت در نظر گرفته‌ایم. دلیل انتخاب چند قالب مقایسه نتایج آن‌ها با یکدیگر است.

هنگام آموزش، هر پرسش را با پاسخ درست آن در قالب پرامپت قرار می‌دهیم. سپس پرامپت پر شده را به کدگذار متنی و تصویر را به کدگذار تصویر مدل CLIP می‌دهیم. مدل به سمت شبیه کردن ویژگی‌های به دست آمده از این دو کدگذار آموزش داده می‌شود.CLIP از شباهت کسینوسی برای محاسبه شباهت این دو استفاده می‌کند. توجه داشته باشید که مدل CLIP برای وظیفهٔ پرسش و پاسخ تصویری از پیش آموخته نشده و مرحلهٔ آموزش را تنها برای یادگیری با تعداد نمونهٔ محدود پیاده‌سازی کرده‌ایم.

در هنگام آزمودن مدل، تمامی پاسخ‌های موجود در مجموعهٔ آزمون را به عنوان دسته‌های ممکن برای پاسخ (مجموعه Z) در نظر می‌گیریم. سپس برای هر دادهٔ آزمون، $|Z|$ عدد پرامپت می‌سازیم که پرسش همان داده و تمامی پاسخ‌های ممکن (Z) را در آن‌ها تعبیه کرده‌ایم. پرامپت‌های پر شده را به کدگذار متنی داده و سپس بردار ویژگی‌های آن را با ویژگی‌های گرفته شده از کدگذار تصویر به کمک شباهت کسینوسی مقایسه می‌کنیم. برچسب متناظر با پرامپتی که بیشترین امتیاز شباهت را دارد به عنوان پاسخ درست انتخاب می‌شود.

۴ـ۳ مدل CLIPCap برای وظیفه توصیف تصویر

مدل CLIP تنها از دو کدگذار متنی و تصویری ساخته شده‌است که می‌توانند متن و تصویر را به ویژگی‌های آن‌ها تبدیل کنند و این ویژگی‌های cross-modal قابلیت قیاس شدن با یک‌دیگر را دارند. بدین ترتیب اگر بخواهیم از مدل CLIP به عنوان یک مدل توصیف تصویر استفاده کنیم، باید تمامی توصیف‌های ممکن را خود تولید کرده و در اختیار کدگذار متنی CLIP قرار دهیم تا ویژگی‌های استخراج شده آن‌ها را با ویژگی‌های استخراج شده از کدگذار تصویر مقایسه کرده و متنی که ویژگی‌هایش بیشترین شباهت را با تصویر داشت به عنوان توصیف انتخاب شود. واضح است که اجرای این روش اصلا میسر نیست.

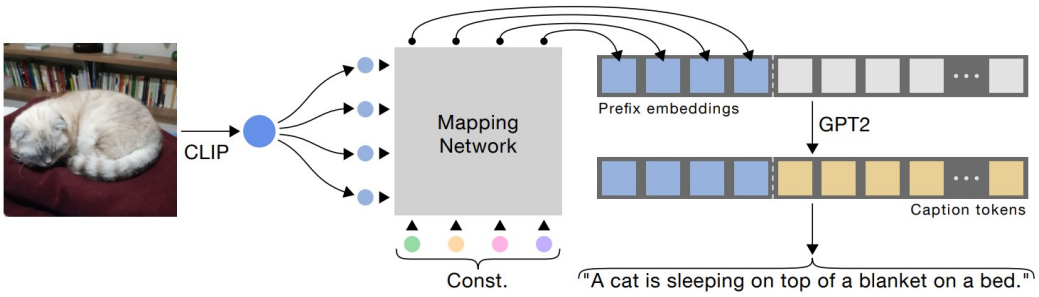
برای وظیفهٔ توصیف تصویر نیاز به مدلی داریم که قابلیت تولید دنباله‌ای از توکن‌ها را با توجه به تصویر داشته باشد. اگر چه زمان زیادی از انتشار CLIP نمی‌گذرد، چندین مقاله برای به‌کارگیری این مدل در وظیفه توصیف تصویر راهکارهایی ارائه داده‌اند که در بخش ۳ـ۴ـ۴ تعدادی از آن‌ها بررسی شد. یکی دیگر از این مقالات[۲۶] مدلی با نام CLIPCap (و یا CLIP Prefix Caption) را پیشنهاد می‌دهد. در پروژه این مدل را برای وظیفه توصیف تصویر به کار برده‌ایم.

معماری کل مدل به شرح زیر است:

این مدل از سه بخش تشکیل شده است: کدگذار تصویر مدل CLIP برای درک تصویر و بازگرداندن ویژگی‌های آن، مدل GPT-2 برای تولید توصیف و یک شبکهٔ نگاشت بین این دو برای نگاشت ویژگی‌های تصویر گرفته شده از کدگذار کلیپ به embeddingهای ورودی مدل GPT-2 (تا GPT-2 از embeddingهای حاصل شده از تصویر برای تولید توصیف استفاده کند).

ابتدا ویژگی‌های تصویر را با استفاده از کدگذار تصویر مدل CLIP استخراج می‌کنیم. برای استخراج یک prefix با طول ثابت، یک شبکه نگاشت آموزش می‌دهیم که ورودی‌های آن embeddingهای CLIP و مقادیر ثابتی هستند. بخش اصلی آموزش مربوط به همین قسمت است به همین جهت در بیشتر فرایند آموزش، مدل CLIP و GPT-2 تحت آموزش قرار نمی‌گیرند و freeze هستند. خروجی تولید شده توسط شبکه نگاشت را prefix embedding می‌نامیم. prefix embedding را با بردار embedding توصیف همان تصویر در کنار هم قرار می‌دهیم و از آن به عنوان ورودی GPT-2 استفاده می‌کنیم تا مدل را برای تخمین و تولید توصیف آموزش دهیم. برای پیش بینی کردن یک توصیف هم تنها prefix embedding را به عنوان ورودی به GPT-2 می‌دهیم تا ادامه توصیف را تولید کند. ساختار مدل در شکل ۴ـ۱ رسم شده است. لازم به توضیح است که برای

زبان انگلیسی از مدل‌های CLIP و GPT-2 و برای زبان فارسی از مدل‌های CLIPfa[۵] و GPT2-Persian[۱۱] استفاده کرده‌ایم. حال به شرح نحوهٔ آموزش مدل می‌پردازیم. ابتدا لازم است دادگان را تجزیه کنیم. برای این



شکل ۴ـ۱: معماری مدل CLIPCap. مرجع تصویر: مقالهٔ CLIPCap[۲۶]

کار، هر تصویر را به کمک پیش‌پردازش‌گر مدل CLIP پیش‌پردازش کرده و سپس با استفاده از کدگذار تصویر CLIP، ویژگی‌های آن را استخراج و ذخیره می‌کنیم Embeddingهای تصویر به همراه توصیف‌های آن را در یک فایل pickle ذخیره می‌کنیم تا در آموزش از آن‌ها استفاده کنیم. برای دسترسی ساده‌تر و منظم‌تر به داده‌های ذخیره شده درون فایل pickle، از DataLoader استفاده می‌کنیم. با مشخص کردن آرگومان‌ها، فرایند آموزش Transformer به تنهایی یا همراه با GPT-2 را آغاز می‌کنیم. آرگومان‌ها عبارتند از: تعداد epochهای آموزش، اندازه batch، انتخاب بین اینکه تنها شبکه نگاشت آموزش داده شود یا علاوه بر آن GPT-2 نیز آموزش داده شود، نوع شبکه نگاشت، تعداد لایه‌های شبکه نگاشت، طول prefix، اندازه prefix، طول CLIP prefix و زبان آموزش (فارسی یا انگلیسی). در ادامه به شرح آرگومان‌های مربوط به ساختار مدل می‌پردازیم:

- نوع شبکه نگاشت: برای نگاشت بین ویژگی‌های کدگذار CLIP و مقادیر ثابت و مدل GPT-2، دو نوع شبکه برای مدل CLIPCap پیاده سازی شده است. یکی از مدل‌ها یک شبکه عصبی چند لایه است و دیگری یک شبکه ترنسفورمر[۳۸] می‌باشد. ما در این پروژه از مدل ترنسفورمر استفاده کرده‌ایم که به دلیل داشتن attention با ساختاری کوچکتر از شبکه ساده چند لایه کارآیی بالاتری دارد.

- تعداد لایه‌های شبکه نگاشت: تعداد لایه‌های مدل ترنسفورمر به عنوان شبکهٔ نگاشت است.

- اندازهٔ prefix و طول CLIP prefix: اندازهٔ prefix همان اندازهٔ بردار خروجی کدگذار تصویر CLIP است. این بردار از یک شبکه عصبی یک لایه می‌گذرد و حاصل آن یک بردار دو بعدی با اندازهٔ طول CLIP prefix و اندازهٔ embeddingهای ورودی GPT-2 خواهد بود.

- طول prefix: پارامتری دو بعدی در هنگام آموزش مدل با اندازهٔ طول prefix و اندازهٔ embeddingهای ورودی GPT-2 تنظیم می‌شود که به عنوان مقادیر ثابت به خروجی شبکهٔ عصبی یک لایه می‌پیوندد و حاصل (یک بردار دو بعدی با اندازهٔ طول prefix + طول CLIP prefix و اندازهٔ embeddingهای GPT-2) به شبکهٔ ترنسفورمر به عنوان ورودی داده می‌شود. خروجی ترنسفورمر همان prefix embedding خواهد بود که به GPT-2 به عنوان قسمتی از کپشن داده می‌شود.

به مقادیر آرگومان‌هایی که در این پژوهش مورد استفاده قرار گرفتند در بخش ۵ـ۳ اشاره خواهد شد.
پس از پایان آموزش وزن‌های محاسبه شده را ذخیره می‌کنیم تا برای پیش‌بینی نمونه‌های جدید از آن استفاده کنیم.
در فرایند پیش‌بینی (تولید نوصیف برای تصویر جدید) مشابه آموزش باید ابتدا embeddingهای تصویر را استخراج کنیم. سپس آن را از شبکه mapping که از قبل آموزش داده‌ایم عبور می‌دهیم. بعد prefix embedding را به GPT-2 می‌دهیم. از آنجایی که GPT-2 به عنوان خروجی لیستی از واژگان به همراه احتمال رخ دادن هر یک ارائه می‌دهد، باید میان این واژگان یکی را انتخاب کنیم. برای این کار می‌توانیم یکی از دو الگوریتم حریصانه[۱] و جستجوی پرتو[۲] را استفاده کنیم که در ادامه به توضیح آن‌ها خواهیم پرداخت.
الگوریتم حریصانه: یک راه برای انتخاب یک واژه از میان واژگان این است که کلمه‌ای را که بیشترین احتمال رخ دادن را در هر موقعیت دارد انتخاب کنیم. محاسبه آن سریع و آسان است و اغلب نتیجه درستی را به دنبال دارد.
جستجوی پرتو: این الگوریتم نسبت به الگوریتم حریصانه دو مزیت دارد:

۱. با جستجوی حریصانه، فقط بهترین کلمه در هر موقعیت انتخاب می‌شود. در مقابل، جستجوی پرتو از میان "N" عدد از کلمات با احتمال بیشتر انتخاب می‌کند.

۲. در جستجوی حریصانه هر موقعیت به صورت مجزا در نظر گرفته می‌شود. هنگامی که بهترین کلمه را برای آن موقعیت تشخیص دادیم، آنچه را که قبل یا بعد از آن آمده بود، بررسی نکردیم. در مقابل، جستجوی پرتو "N" عدد از بهترین دنباله‌های تا کنون را انتخاب می‌کند و احتمال ترکیب همه کلمات قبلی را به همراه کلمهٔ فعلی در نظر می‌گیرد.

[۱] Greedy
[۲] Beam Search

به طور شهودی منطقی است که این الگوریتم نتایج بهتری نسبت به جستجوی حریصانه داشته باشد. زیرا اگر فقط بهترین کلمه را در هر موقعیت انتخاب کنیم ممکن است کامل‌ترین جمله را از دست بدهیم.

فرایند تولید توکن به عنوان بخشی از توصیف تا زمانی ادامه می‌یابد که توکن پایان جمله که در توکنایزر GPT-2، توکن [SEP] در نظر گرفته شده تولید شود. این توکن، توکن توقف[۳] برای ایجاد توصیف است.

در یادگیری با تعداد نمونه معدود یکی از آرگومان‌ها تعداد شات است. پس از مشخص کردن شات، به همان تعداد به صورت تصادفی با یک seed از داده‌های آموزشی نمونه‌برداری می‌کنیم. سپس فرایند آموزش را با همان داده‌ها پیش می‌بریم. و بعد مدل را روی باقی داده‌ها امتحان می‌کنیم.

در یادگیری بدون نمونه همانطور که از پیش توضیح داده شد، مدل را روی هیچ یک از داده‌های جدید آموزش نمی‌دهیم و از تمامی داده‌ها برای پیش‌بینی استفاده می‌کنیم.

برای سنجش عملکرد مدل، تصمیم بر آن شد که از دو معیار BLEU و ROUGE استفاده شود که در فصل معیارهای به کار رفته برای ارزیابی مدل توصیف تصویر به توضیح کامل آن‌ها پرداخته‌ایم.

[۳] Stop token

فصل ۵

پیاده‌سازی

۵ـ۱ مقدمه

در این فصل ابتدا به شیوهٔ گردآوری داده‌های فارسی مورد نیاز برای انجام آزمایش‌ها، پیش پردازش آن‌ها و ترجمه آن‌ها به زبان انگلیسی می‌پردازیم و در ادامه شرایط آزمایش شامل هایپرپارامترهای تعیین شده، نحوهٔ آموزش مدل توصیف تصویر و تنظیم دقیق[۱] آن را برای هر کدام از مسائل پرسش و پاسخ تصویری و توصیف تصویر شرح خواهیم داد.

۵ـ۲ جمع‌آوری داده

از آن جایی که یکی از اهداف اصلی این پروژه آموزش یک مدل با عملکرد قابل قبول بر روی تصاویر نشان‌دهنده فرهنگ ایرانی است، در ابتدا لازم دیدیم یک دادگان مناسب جمع‌آوری شود. برای پیدا کردن تصاویر مناسب در صفحات مجازی جستجو کردیم و در نهایت تصاویر در ۷ دسته‌بندی با عناوین افراد، فضای داخلی، غذا، ورزش، مراسم، ماشین و فضای خارجی به صورت دستی جمع شدند. به گردآوری تصاویری که دارای ویژگی‌های خاص ایران مانند حجاب، معماری ایرانی ساختمان‌ها، لباس‌های محلی، غذاها و مراسم محلی هستند و به چشم یک خارجی ناآشنا آیند اهتمام ورزیده شد. بیشتر عکس‌ها از سایت pinterest به دست آمده‌اند.

در ادامه به شرح تصاویر هر دسته می‌پردازیم. توجه اصلی تصاویر دسته افراد، روی فرد یا افرادی‌ست که در تصویر وجود دارند. تصاویر فضای خارجی همان طور که از نامش مشخص است فضای بیرون از ساختمان‌ها را نشان می‌دهند. تصاویر دسته ماشین‌ها نیز تصویری از ماشین دارند. در این تصاویر محیط اطراف ماشین (که معمولا فضای بیرون هستند و به همین دلیل به تصاویر دسته فضای خارجی شباهت دارند) نیز می‌تواند در توصیف تصویر وجود داشته باشد و تنها تمرکز روی ماشین‌های درون تصویر نیست. تصویر دسته مراسمات محیطی از جشن‌ها، اعیاد و مراسم‌های دیگر فرهنگ ایران را داراست. تصاویر دسته‌های غذا، مکان داخلی و ورزش نیز به ترتیب نشان‌دهنده غذاهای ایرانی، داخل ساختمان‌های ایرانی و ورزش‌های مخصوص ایران (مانند کشتی، وزنه برداری و ...) و یا تصویری از ورزشکاران ایرانی سایر رشته‌هاست.

پس از ذخیره کردن حدود ۱۰۰ تصویر برای هر دسته، اقدام به پیاده‌سازی یک وبسایت[۲] به جهت جمع‌آوری

[۱] Fine-tuning
[۲] https://nlplab.iust.ac.ir/pvqa

دادگان کردیم. بخش "جمع‌آوری دیتاست" در سایت شامل دو سناریو است؛ وارد کردن پرسش و پاسخ برای تصویر و وارد کردن متن برای تصویر. برای فراهم کردن داده تصمیم بر آن شد که از ۴ نفر که زبان مادری‌شان فارسی است کمک بگیریم. بدین جهت راهنمای جمع‌آوری داده برای هر دو سناریو در اختیارشان قرار گرفت که در بخش‌های بعد به دستورالعمل جمع‌آوری داده اشاره خواهد شد.

برای وظیفه توصیف تصویر، از ۶۱۰ عدد از تصاویر جمع شده استفاده شد که تصاویر تمام دسته‌ها به جز فضای خارجی است. هر فرد یک توصیف برای هر تصویر نوشته و بدین ترتیب برای هر تصویر ۴ عدد توصیف گردآوری شد. برای وظیفه پرسش و پاسخ تصویری، هر تصویر تنها یک پرسش و یک پاسخ کوتاه و بلند همراه با دسته‌بندی پرسش (وجود شیء، شناسایی ورزش، استدلال مکانی، درک عواطف، ویژگی رنگی، شمارش شیء، شناسایی فعالیت، یافتن شیء، جنس شیء، شناسایی جنسیت و سایر) را دارد که به طور تصادفی توسط یکی از کارمندان نوشته شده است (هر فرد تمام داده‌های یک تصویر را نوشته‌است). تمام داده‌های وارد شده توسط ما بررسی و در صورت نیاز اصلاح شد. تعداد تصاویر و توصیف‌ها و همچنین جفت پرسش و پاسخ‌های تصاویر به طور دقیق‌تر در جدول ۵ ـ ۱ و ۵ ـ ۲ آمده‌است.

دستهٔ تصاویر	تعداد تصویر	تعداد توصیف نوشته‌شده
افراد	۱۰۱	۴۰۴
فضای داخلی	۱۰۰	۴۰۰
غذا	۱۰۱	۴۰۴
ورزش	۱۰۲	۴۰۸
مراسم	۱۰۰	۴۰۰
ماشین	۱۰۶	۴۲۴
فضای خارجی	۱۰۳	–
مجموع	۷۱۳	۲۴۴۰

جدول ۵ ـ ۱: تعداد تصاویر و توصیف‌های جمع‌آوری شده برای مسئلهٔ توصیف تصویر

برای پیش‌پردازش پرسش، پاسخ بلند و توصیف تصویر از نرمالایزر هضم[۳۳] و برای پیش‌پردازش پاسخ‌های کوتاه از نرمالایزر پارسی‌ور[۲۵] استفاده کرده‌ایم. این دو ابزار تا حدی توانستند مشکلات نیم‌فاصله نداشتن بعضی کلمات را اصلاح کنند. سپس به طور دستی بعضی پاسخ‌های کوتاه را که در یک دسته قرار می‌گرفتند اما متن متفاوتی داشتند یکسان کردیم (به طور مثال "سمت راست" و "راست"، هر دو به کلمه

دستهٔ پرسش	تعداد تصویر، پرسش و پاسخ این دسته
وجود شیء	۷۷
شناسایی ورزش	۱۳
استدلال مکانی	۶۵
درک عواطف	۵
ویژگی رنگی	۹۴
شمارش شیء	۱۲۵
شناسایی فعالیت	۳۱
یافتن شیء	۶۳
جنس شیء	۲۷
شناسایی جنسیت	۷۲
سایر	۱۵۱
مجموع	۷۲۳

جدول ۵ـ۲: تعداد تصاویر، پرسش و پاسخ جمع‌آوری شده به تفکیک دسته‌بندی پرسش برای مسئله پرسش و پاسخ تصویری

"راست" تبدیل شدند)، نقطه در پایان توصیف و پاسخ بلند و علامت سوال در پایان پرسش‌ها گذاشتیم. نمونه‌هایی از دادگان گردآوری و پیش‌پردازش شده را می‌توانید در بخش‌های ۶ـ۲ـ۱ و ۶ـ۲ـ۳ مشاهده کنید.

طراحی وب‌سایت در دو بخش سمت کاربر (با کتاب‌خانه ReactJS) و سمت سرور (با چارچوب Django و زبان پایتون) انجام شد. علاوه بر بخش جمع‌آوری دیتاست (شکل ۵ـ۱ و ۵ـ۲) که در بخش قبل ذکر شد، سه بخش بازبینی، آمار و راهنمای دیتاست نیز پیاده‌سازی شد. بخش بازبینی به منظور اصلاح و تایید داده‌های وارد شده توسط کارمندان ایجاد شد (شکل ۵ـ۳ و ۵ـ۴). آمار دیتاست تعداد داده‌هایی که هر حساب شخصی ثبت کرده را نشان می‌دهد. بخش راهنمای دیتاست با هدف نوشتن نکات و قوانین مربوط به جمع‌آوری دیتاست برای کارمندان ایجاد شد. این قوانین در دو بخش بعد شرح داده می‌شوند.

شکل ۵ـ۱: تصویری از بخش جمع‌آوری دادهٔ پرسش و پاسخ تصویری در وب‌سایت

شکل ۵ـ۲: تصویری از بخش جمع‌آوری دادهٔ توصیف تصویر در وب‌سایت

شکل ۵ـ۳: تصویری از بخش بازبینی دادهٔ پرسش و پاسخ تصویری در وب‌سایت

شکل ۵ـ۴: تصویری از بخش بازبینی دادهٔ توصیف تصویر در وب‌سایت

۵ـ۲ـ۱ دستورالعمل جمع آوری دادهٔ توصیف تصویر

ابتدا یک تصویر به صورت تصادفی به کاربر نشان داده شد. سپس از او خواسته شد تا یک جمله مرتبط با تصویر بسازد. در ساختن جملات از کاربران خواسته شد تا نکات زیر را رعایت کنند:

- متن باید حداقل ۸ کلمه داشته باشد.
- متن باید توصیف‌کننده منظره، رویداد، فعالیت و موجودات درون تصویر (اشخاص، جانداران دیگر و اشیا) باشد.

- از نوشتن جزئیات ریز تصویر خودداری شود.
- متن حتما به صورت نوشتاری باشد و از به کار بردن کلمات محاوره‌ای خودداری شود.
- متن تک جمله‌ای باشد.
- از نوشتن کلمات یا صفات سلیقه‌ای خودداری کنید. مثلا "گربه زیبا" اشتباه است و "گربه" درست است.
- از به کار بردن عباراتی که به تصویر یا شکل اشاره می‌کنند خودداری شود. به طور مثال "در این تصویر ..." یا "در این عکس ..." یا ".... در تصویر است." یا "متن تصویر عبارت است از ...".
- از نوشتن اسم خاص افراد مانند اسامی افراد مشهور خودداری شود.

۵ ـ ۲ ـ ۲ دستورالعمل جمع آوری دادهٔ پرسش و پاسخ تصویری

مانند بخش قبل در ابتدا یک تصویر به صورت تصادفی به کاربر نشان داده شد. سپس از او خواسته شد تا یک پرسش در رابطه با تصویر مطرح کند و به آن به دو صورت پاسخ (پاسخ کوتاه و پاسخ بلند) دهد و در نهایت مشخص کند که پرسش در چه دسته‌ای قرار می‌گیرد. در این فرایند از آنان خواسته شد تا نکات زیر را رعایت کنند:

- سختی سوال محدودیتی ندارد اما باید رسمی و نوشتاری باشد.
- هدف از جمع‌آوری پرسش و پاسخ اشاره به فرهنگ ایرانی‌ست. بدین ترتیب تا حد امکان به نکات مربوط به فرهنگ ایرانی و موضوعات یا چیزهایی که در خارج از ایران رایج نیست اشاره شود، به طور مثال حجاب، غذا و آداب و رسوم ایرانی.
  این پرسش و پاسخ باید تا جای ممکن به شکلی باشد که شخصی که با ایران آشنایی ندارد به سختی بتواند به آن جواب دهد، اما جواب دادن به آن غیر ممکن نباشد. (مثلا پرسش درباره نام خاص یک فرد یا مکان نباشد.)
- برای پاسخ دادن به پرسش تولید شده نیازی به داشتن دانش خارجی نباشد. اطلاعاتی به غیر از مفاهیمی که انسان با نگاه کردن به تصویر و خواندن پرسش متوجه می‌شود، لازم نباشد. برای مثال سوال در

مورد یک تصویر اینگونه پرسیده نشود که «این تصویر مربوط به دهه‌ی ۶۰ است یا دهه‌ی ۵۰؟» زیرا برای پاسخ به این سوال فرد باید دانش خارجی در مورد نوع مد لباس، وسایل نقلیه و ... در هر دهه داشته باشد تا بتواند تشخیص دهد که تصویر مربوط به کدام دهه است.

- پرسش باید بدون ابهام مطرح شود و یک پاسخ مشخص داشته باشد به طوری که اگر آن سوال را از یک فرد دیگر بپرسیم دقیقا همان پاسخ مد نظر را بدهد.
- پرسش نباید به گونه‌ای باشد که پاسخ آن تنها به اطلاعات تصویر و یا تنها به اطلاعات پرسش نیاز داشته باشد. بلکه باید از مجموع اطلاعاتی که از تصویر و پرسش دریافت می‌شود؛ پاسخ سوال داده شود.

۵ـ۲ـ۳ ترجمهٔ مجموعهٔ داده به زبان انگلیسی

از کتابخانه deep-translator[۳] برای ترجمه داده‌های جمع‌آوری شده استفاده کردیم. به کمک GoogleTranslator API داده‌های فارسی را به انگلیسی ترجمه کردیم تا بتوانیم از تصاویر مرتبط با فرهنگ فارسی برای آموزش مدل به زبان انگلیسی استفاده کنیم. چند نمونهٔ فارسی را همراه با نمونهٔ انگلیسی آن در ادامه مشاهده می‌کنید:

- مجموعه دادگان توصیف تصویر:
  - توصیف فارسی: خانمی با چادری رنگی همراه کودک در بغلش در بازار محلی است .
    توصیف انگلیسی:A woman with a colorful chador with a child in her arms is in the local market.
  - توصیف فارسی: دو کشتی گیر با لباس‌های قرمز و آبی روی تشک در حال کشتی گرفتن هستند .
    توصیف انگلیسی:.Two wrestlers in red and blue clothes are wrestling on the mat
  - توصیف فارسی: یک بشقاب پلو و کباب با مخلفات در سینی است .
    توصیف انگلیسی:.A plate of pilaf and kebab with side dishes is in the tray
- مجموعه دادگان پرسش و پاسخ تصویری:

[۳] https://deep-translator.readthedocs.io/en/latest/

- پرسش و پاسخ فارسی: آیا در هر دو سمت بزرگراه ترافیک است؟ خیر
  پرسش و پاسخ انگلیسی: Is there traffic on both sides of the highway? no
- پرسش و پاسخ فارسی: نمازگزاران زن هستند یا مرد؟ زن
  پرسش و پاسخ انگلیسی: Are the worshipers women or men? Female
- پرسش و پاسخ فارسی: لباس مرد حاضر در عزاداری چه رنگی است؟ مشکی
  پرسش و پاسخ انگلیسی: What color is the dress of the man in mourning? Black

۵ـ۳ شرح نحوهٔ ایجاد و آموزش مدل‌های توصیف تصویر

مدل CLIPCap استفاده شده برای زبان انگلیسی از سه بخش کدگذار تصویر CLIP ResNet50x4، یک ترنسفورمر به عنوان تابع نگاشت و GPT-2 تشکیل شده است. مدل‌های CLIP و GPT-2 مدل‌های از پیش آموزش دیده مدل‌های اصلیشان هستند و برای وظیفه توصیف تصویر آموزش داده نشده‌اند.ترنسفورمر بین این دو مدل که نگاشت بردارهای ویژگی استخراج شده از کدگذار CLIP به embeddingهای کپشن GPT-2 را انجام می‌دهد بر روی دیتاست COCO[۱۳] برای انجام این وظیفه از پیش آموزش دیده شده است. به طور دقیق‌تر، شبکه نگاشت بین کدگذار CLIP و GPT-2 از ۸ لایه با طول prefix و طول CLIP prefix برابر با ۴۰ تشکیل شده و اندازهٔ بردارهای تصویر خروجی از کدگذار CLIP ۶۴۰ است.

مدل فارسی از کدگذار تصویر فارسی CLIPfa[۵] که یک کدگذار CLIP VIT-B/32 آموزش دیده شده روی ۴۰۰ هزار دادهٔ ترجمه شده از مجموعه دادگان COCO، Flickr30k[۴۱] و Conceptual Captions 3m[۳۴] است [۸، ۹، ۷، ۶] استفاده می‌کند. به جای GPT-2، از GPT-2 persian[۱۱] که یک مدل معادل آن در زبان فارسی است بهره می‌برد.

به دلیل نداشتن یک شبکهٔ نگاشت بین این دو مدل برای زبان فارسی، خود این شبکه را ساخته و آموزش داده‌ایم. این شبکه، یک ترنسفورمر ۸ لایه با طول prefix برابر با ۱۰ و طول CLIP prefix برابر با ۲۰ است. اندازه بردارهای خروجی کدگذار CLIPfa نیز برابر با ۷۶۸ است. ترنسفورمر ساخته شده به تعداد ۵ epoch به تنهایی با ضریب یادگیری ۲×۱۰^{-۵} آموزش دیده و در دو epoch با همان ضریب یادگیری به همراه مدل GPT-2 persian تنظیم شد. برای این کار، از مجموعه‌دادگان آموزش مدل CLIPfa استفاده کرده‌ایم.

۵ـ۴ شرح نحوهٔ ایجاد و آموزش مدل‌های پرسش و پاسخ تصویری

برای این وظیفه، از مدل CLIP با کدگذار تصویری VIT-B/32 با وزن‌های منتشر شده توسط OpenAI[۴] برای زبان انگلیسی و مدل CLIPfa از پیش آموخته شده با همین نوع کدگذار[۵] استفاده کرده‌ایم. توجه کنید که مدل‌ها را برای وظیفه پرسش و پاسخ تصویری آموزش نداده‌ایم.

[۴] https://huggingface.co/openai/clip-vit-base-patch32

[۵] clip-fa-vision و clip-fa-text

فصل ۶

بازبینی عملکرد

۶ـ۱ مقدمه

در ابتدا برای هر کدام از مسائل توصیف تصویر و پرسش و پاسخ تصویری نمونه‌هایی را از عملکرد مدل‌ها نشان خواهیم داد. سپس به ارزیابی آزمایش‌های انجام شده می‌پردازیم. برای ارزیابی عملکرد مدل توصیف تصویر از دو معیار BLEU و ROUGE استفاده شد که در فصل ۲ـ۵ به توضیح این دو معیار پرداختیم. جهت ارزیابی مدل پرسش و پاسخ از معیار دقت استفاده کردیم.

۶ـ۲ نتایج و ارزیابی

۶ـ۲ـ۱ نمونه‌هایی از نتایج آزمایش مدل‌های توصیف تصویر

شکل ۶ـ۱: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"گروهی از مردان جوان در حال اجرای رقص سنتی در یک روستا."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"گروهی از مردان جوان در حال اجرای موسیقی سنتی."

توصیف‌های مرجع:

"چهار مرد با لباس‌های یک شکل محلی در حال دست زدن هستند .".

"چهار مرد با لباس سنتی بلوچی در محدوده‌ای قدیمی دست می‌زنند .".

"مردان با لباس محلی رقص و پایکوبی می‌کنند .".

"چهار مرد با لباس محلی سفید در حال دست زدن هستند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A man holding a kite on top of a hill."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A man holding a kite on top of a hill."

توصیف‌های مرجع:

"Four men in the same local clothes are shaking hands."

"Four men in traditional Balochi clothes are shaking hands in an old area."

"Men dance and stomp in local clothes."

"Four men in white local clothes are shaking hands."

شکل ۶ـ۲: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"گروهی از زنان در یک کتابخانه کتاب می خوانند."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"گروهی از زنان در یک کتابخانه کتاب می خوانند."

توصیف‌های مرجع:

"خانمی با شال آبی و زرد و پیراهن قرمز درحال خواندن کتاب است .".

"سه زن در کتابخانه‌ای با قفسه‌های پر از کتاب هستند .".

"سه زن در یک کتاب‌خانه در حال انتخاب کتاب هستند .".

"سه خانم در قفسه‌های کتابخانه می‌گردند و به کتب نگاه می‌کنند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A group of people in a library with books."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A group of people in a library.”

توصیف‌های مرجع:

”A lady with a blue and yellow shawl and a red shirt is reading a book.”

”Three women are in a library with shelves full of books.”

”Three women are choosing books in a library.”

”Three women go around the library shelves and look at the books.”

شکل ۶ـ۳: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”دو زن با چتر در خیابان راه می روند.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”دو زن با چتر در خیابان راه می روند.”

توصیف‌های مرجع:

”دو زن چادری چتر به دست در پیاده‌روی خیس قدم می‌زنند .“،

”دو زن با چادرهای مشکی و چتر در حال عبور از خیابان هستند .“،

”دو خانم در روز بارانی چتر روی سرشان گرفته‌اند .“،

”دو زن با چادر مشکی و چتر ، در حال پیاده‌روی زیر باران هستند .“

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”Two women are holding umbrellas in the rain.“

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”Two women are holding umbrellas in the rain.“

توصیف‌های مرجع:

”Two women wearing umbrellas are walking on the wet sidewalk.“

”Two women with black tents and umbrellas are crossing the street.“,

”Two women are holding an umbrella on their heads on a rainy day.“,

”Two women with black tents and umbrellas are walking in the rain.“

شکل ۶ـ۴: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"زن و شوهری که با هم غذا می خورند."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"زن و شوهری در حال خوردن غذا در بازار."

توصیف‌های مرجع:

"چهار خانم با لباسهای محلی دور میزی نشسته‌اند .".

"چند زن به دور میزی با رومیزی راه راه جمع شده‌اند .".

"چند پیرزن با لباس‌های محلی دور کرسی نشسته‌اند .".

"چهار خانم با لباس‌های سنتی پشت کرسی پر از خوراکی نشسته‌اند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A group of people sitting around a table."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A group of people sitting around a table."

توصیف‌های مرجع:

"Four women in local clothes are sitting around a table.",

"Several women are gathered around a table with a striped tablecloth.",

"Some old women in local clothes are sitting around the chair.",

"Four women in traditional clothes are sitting behind a chair full of food."

شکل ۶ـ۵: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"مردی با پیراهن قرمز و شلوار جین در حال کندن گل های تازه است."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"دو مرد در حال کندن گل های تازه هستند."

توصیف‌های مرجع:

"دو مرد با گونی‌های پر از گل مشغول گلاب‌گیری هستند .".

"دو مرد گل‌های محمدی چیده شده را از کیسه خالی می‌کنند .".

"دو مرد در حال خالی کردن گل‌های محمدی از کیسه هستند .".

"دو مرد کیسه بزرگ گل محمدی را خالی می‌کنند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A man that is holding a bunch of flowers."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A man that is holding a bunch of flowers.”

توصیف‌های مرجع:

”Two men are picking rose water with sacks full of flowers.”

”Two men empty the picked flowers from the bag.”

”Two men are emptying flowers from a bag.”

”Two men empty a big bag of roses.”

شکل ۶ـ۶: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”گروهی از مردان سوار بر موتور سیکلت.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”گروهی از مردان سوار بر موتور سیکلت در کنار جاده ای ایستاده اند.”

توصیف‌های مرجع:

"دو مرد روی موتور نشسته و یک مرد کنار موتور خود در کنار دیوار ایستاده است .".

"سه مرد با موتور کنار هم جلوی دیوار ایستاده‌اند .".

"سه مرد با دو موتورسیکلت به دیوار تکیه داده‌اند .".

"سه مرد با کاپشن‌های مشکی کنار موتور ایستاده‌اند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"Two men are sitting on a motorcycle."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"Two men are sitting on a motorcycle."

توصیف‌های مرجع:

"Two men are sitting on the motorcycle and one man is standing next to his motorcycle by the wall."

"Three men with motorcycles are standing together in front of the wall."

"Three men with two motorcycles are leaning against the wall."

"Three men in black jackets are standing next to the engine."

شکل ۶ـ۷: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"یک ماشین پارک شده در یک پارکینگ."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"یک ماشین پارک شده در یک پارکینگ."

توصیف‌های مرجع:

"خودروی زرد در گوشه ی حیاط پارک‌شده‌است ."

"یک ماشین قدیمی زرد رنگ که در پارکینک پارک‌شده‌است ."

"یک ماشین زرد در پارکینگ خانه‌ای پارک‌شده‌است ."

"یک ماشین زرد در پارکینگ خانه پارک‌شده‌است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A yellow car parked next to a parked car."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A yellow car parked in a parking lot.”

توصیف‌های مرجع:

”The yellow car is parked in the corner of the yard.”

”An old yellow car parked in the parking lot.”

”A yellow car is parked in the parking lot of a house.”

”A yellow car is parked in the parking lot of the house.”

شکل ۶ـ۸: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”ماشین های پارک شده در پارکینگ”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”ماشین های پارک شده در پارکینگ”

توصیف‌های مرجع:

”یک ماشین سفید بر اثر تصادف دچار آسیب‌شده‌است .“

”یک ماشین سفید با در آسیب‌دیده کنار خیابان است .“

”یک خودروی سفید تصادف کرده کنار خیابان پارک کرده است .“

”یک ماشین سفید رنگ که تصادف کرده است در حال حرکت از یک خیابان است .“

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”A car that has been hit by a car.“

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A car is parked next to a tree.“

توصیف‌های مرجع:

”A white car was damaged in an accident.“

”A white car with a damaged door is on the side of the street.“

”A white car crashed and parked on the side of the street.“

”A white car that has an accident is moving down a street.“

شکل ۶ـ۹: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"گروهی از مردم در حال عبادت در یک کلیسا."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"گروهی از مردم در حال عبادت در یک کلیسا."

توصیف‌های مرجع:

"مردان با لباس مشکی در حرم گل به دست دارند ."

"جمعیت زیادی با لباس‌های مشکی و گل در دست در حرم حضور دارند ."

"مردم در یک مکان مذهبی جمع شده‌اند و گل در دست دارند ."

"یک مرد گل به دست با کلاه سبز در جمعیتی سیاه پوش است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A group of people standing next to a building."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A group of people standing next to a building."

توصیف‌های مرجع:

"Men in black clothes are holding flowers in the shrine."

"Many people are present in the shrine with black clothes and flowers in their hands."

"People have gathered in a religious place and are holding flowers."

"A man with a flower in his hand and a green hat is in a crowd dressed in black."

شکل ۶ـ۱۰: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"زنی که در خیابان های شهر قدم می زند"

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"زنی که در خیابان با تلفن همراه صحبت می کند، در حالی که در خیابان راه می رود، در حالی که در خیابان راه می رود، در حالی که در خیابان راه می رود، در حالی که در خیابان راه می رود، در حالی که در خیابان راه می رود، در حالی که در خیابان راه می رود."

توصیف‌های مرجع:

"یک مرد با لباس قرمز درخیابان ایستاده است ."

"مردی سیاه پوست با طبلی در دست در پیاده‌رو ایستاده است ."

"مردی با صورت سیاه و لباس قرمز و دایره زنگی به دست در پیاده‌رو ایستاده است ."

"مردی با لباس و کلاهی قرمز و صورت سیاه در خیابان ایستاده است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”A man in a red jacket is standing in the street.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A man in a red jacket and a red cap.”

توصیف‌های مرجع:

”A man in a red dress is standing on the street.”

”A black man is standing on the sidewalk with a drum in his hand.”

”A man with a black face and red clothes and a bell ring is standing on the sidewalk.”

”A man in a red dress and hat with a black face is standing on the street.”

شکل ۶ـ۱۱: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”زن و مردی که چتر در دست دارند در حالی که یک چتر در دست دارند در خیابان قدم می زنند.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"زن و شوهری که چتر در دست دارند در حالی که چتر در دست دارند با هم صحبت می کنند."

توصیف‌های مرجع:

"عروس و داماد کنار هم نشسته‌اند و عروس چیزی می‌نویسد ."

"یک مرد در حالی که دسته گلی در دست دارد ، در کنار یک زن با چادر سفید نشسته است ."

"مرد با کت و شلوار خاکستری و پیراهن قهوه‌ای دسته گل در دست دارد ."

"زن با چادر سفید کنار مردی که دسته گل قرمز در دست دارد نشسته است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A person is holding a piece of paper."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A person is holding a piece of paper."

توصیف‌های مرجع:

"The bride and groom are sitting together and the bride is writing something."

"A man is sitting next to a woman with a white veil while holding a bouquet of flowers."

"A man in a gray suit and a brown shirt is holding a bouquet of flowers."

"A woman in a white veil is sitting next to a man holding a bouquet of red flowers."

شکل ۶ـ۱۲: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"گروهی از مردم در حالی که پرچم های خود را بالا می برند، از میان جمعیت عبور می کنند."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"گروهی از مردم در حالی که پرچم های خود را بالا می برند، از میان جمعیت عبور می کنند."

توصیف‌های مرجع:

"جمعیت بسیاری برای برگزاری مراسم عزاداری در محوطه‌ی بزرگی جمع شدند ."

"تعداد زیادی از مردم با لباس مشکی برای برگزاری مراسم عزاداری جمع شده‌اند ."

"رنگین‌کمان در آسمان است و عزاداری مذهبی برپاست ."

"جمعیت زیادی از مردم با لباس‌های سیاه و پرچم‌های سبز رنگ در یک محوطه بزرگ جمع شده‌اند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A group of people standing in the middle of a field."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A large group of people standing in the middle of a field.”

توصیف‌های مرجع:

”A large crowd gathered in a large area for the funeral ceremony.”

”A large number of people dressed in black have gathered for the mourning ceremony.”

”There is a rainbow in the sky and there is a religious mourning.”

”A large crowd of people in black clothes and green flags have gathered in a large area.”

شکل ۶ـ۱۳: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”یک کاسه پر از سبزیجات و گوشت.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”یک کاسه پر از سبزیجات و گوشت.”

توصیف‌های مرجع:

"کنار ظرف سفید رنگ خورش کرفس یک قاشق است ."

"خورش کرفس در ظرف سفید رنگ سروشده‌است ."

"یک کاسه سفید با قاشق در آن پر از خورشت کرفس است ."

"یک ظرف خورش کرفس که قاشقی در آن قرار دارد ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A white plate topped with a bunch of food."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A plate of food with meat and vegetables."

توصیف‌های مرجع:

"Next to the white dish of celery stew is a spoon."

"Celery stew is served in a white dish."

"A white bowl with a spoon is full of celery stew."

"A dish of celery stew with a spoon in it."

شکل ۶ـ۱۴: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"یک تکه کیک در بشقاب."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"یک تکه کیک با یک چنگال روی آن."

توصیف‌های مرجع:

"یک بشقاب ته چین طلایی که با زرشک تزئین‌شده‌است ."

"روی ظرف چینی پلو زرشک ریخته‌شده‌است ."

"یک قسمت از ته چین قالبی بریده‌شده‌است ."

"در یک ظرف سفید ته‌چین مرغ با زرشک‌پلو است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A plate of food with a cake on it."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A plate filled with food on a table.”

توصیف‌های مرجع:

”A golden plate decorated with barberry.”

”Barberry rice is poured on the porcelain dish.”

”A part is cut from the bottom of the mold.”

”In a white dish, the bottom of the chicken is filled with barberry.”

شکل ۶ـ۱۵: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”قطاری که در ایستگاه راه آهن در حال حرکت است.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”قطاری که در ایستگاه مترو در حال حرکت است.”

توصیف‌های مرجع:

”ایستگاه مترو خلوت است و کسی در آن نیست .”

”ردیفی از صندلی‌های سبز و زرد در سالن قرار دارد .”

”یک سالن که دارای صندلی‌های زرد و سبز است .“

”یک سالن با دیوارهای سنگی که خالی از افراد است .“

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”A train is traveling down the tracks at a station.“

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A train is traveling down the tracks at a station.“

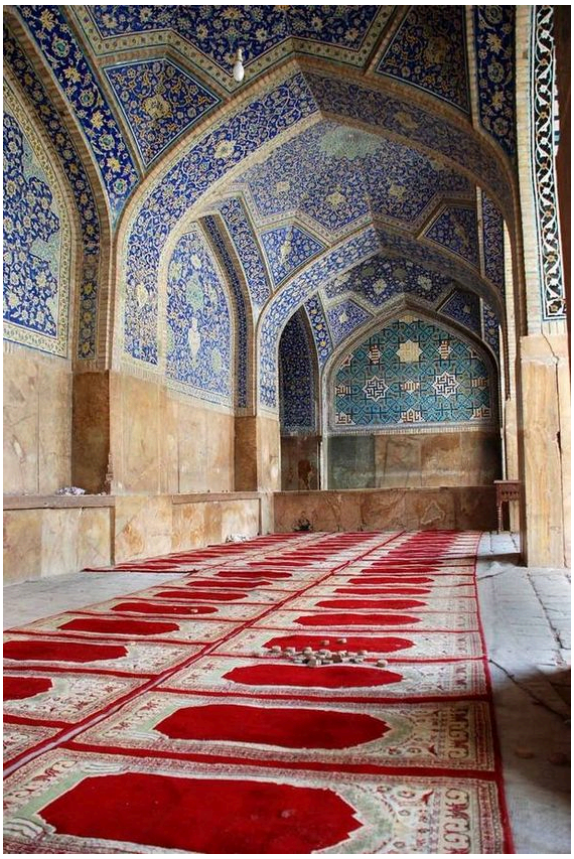
توصیف‌های مرجع:

”The metro station is quiet and there is no one there.“

”There is a row of green and yellow chairs in the hall.“

”A hall with yellow and green chairs.“

”A hall with stone walls that is empty of people.“



شکل ۶ـ۱۶: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"عبادتگاه بودایی در یک شهر"

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"عبادتگاه بودایی در یک شهر"

توصیف‌های مرجع:

"فرش‌های قرمز در مسجد برای خواندن نماز پهن شده‌اند ."

"نمازخانه‌ای که با قالی‌های قرمز پوشیده‌شده‌است ."

"مکانی مذهبی که با فرش قرمز پوشیده‌شده‌است ."

"در راهرو با سقف گنبدی لاجوردی ، قالیچه‌هایی پهن‌شده‌است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”A large room with a large clock on it.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A large room with a large rug.”

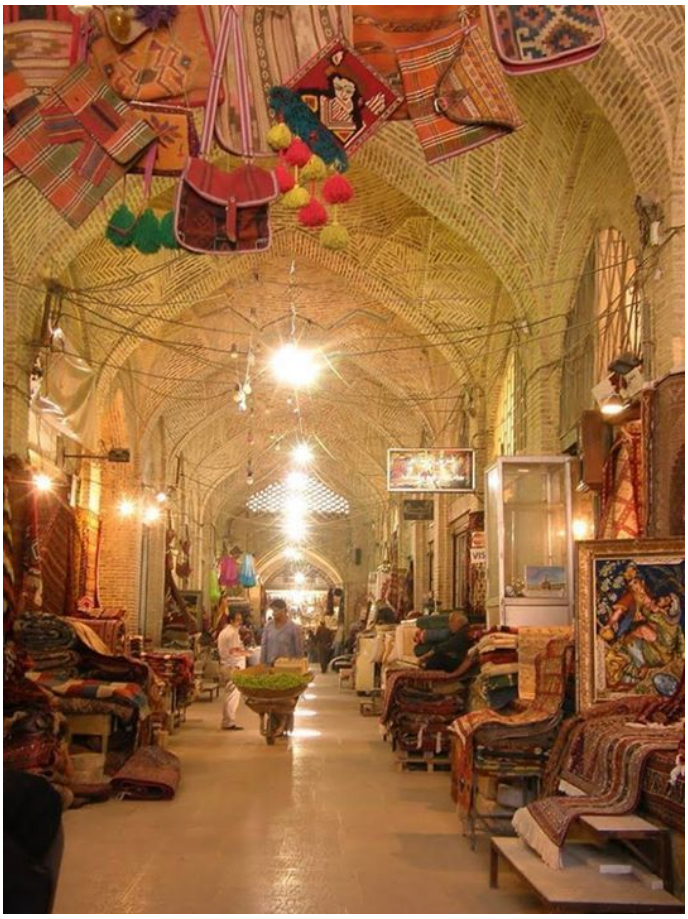
توصیف‌های مرجع:

”Red carpets are spread in the mosque for praying.”

”A prayer hall covered with red carpets.”

”A religious place covered with a red carpet.”

”Carpets are spread in the corridor with azure dome ceiling.”



شکل ۶ـ۱۷: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"یک کلیسا در یک روز آفتابی در یک شهر شلوغ."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"مردم در حال قدم زدن در یک بازار سنتی هستند."

توصیف‌های مرجع:

"کیف‌های دوشی از سردر مغازه آویزان است ."

"بازار فرش فروشان که یک مرد با چرخ دستی در وسط بازار در حال عبور است ."

"بازار قدیمی که کیف‌های سنتی از بالا آویزان است ."

”تعدادی مغازه فرش‌فروشی در یک بازار سرپوشیده قدیمی در کنار یکدیگر قرار دارند .“

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”A group of people sitting in a room with umbrellas.“

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”A group of people sitting in a room.“

توصیف‌های مرجع:

”Doshi bags are hanging from the front of the shop.“

”The market of carpet sellers, where a man is passing by with a wheelbarrow in the middle of the market.“

”The old market where traditional bags hang from above.“

”A number of carpet shops are located next to each other in an old covered market.“

شکل ۶ـ۱۸: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"دو تیم فوتبال در حال تلاش برای گرفتن توپ هستند."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"دو تیم فوتبال در حال تلاش برای گرفتن توپ هستند."

توصیف‌های مرجع:

"بازیکن‌ها با لباس سفید و قرمز درحال بازی فوتبال هستند ."

"در مسابقه فوتبال بازیکنان برای گرفتن توپ بالا پریده‌اند ."

"تعدادی از جوانان با لباس‌های قرمز و سفید مشغول بازی فوتبال هستند ."

"دو فوتبالیست با لباس سفید برای زدن توپ با سر به هوا پریدند ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A group of men playing soccer on a field."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A group of men playing soccer on a field."

توصیف‌های مرجع:

"The players are playing football with white and red clothes."

"In the football match, the players have jumped up to catch the ball."

"A number of young people in red and white clothes are playing football."

"Two soccer players in white clothes jumped in the air to hit the ball."

شکل ۶ـ۱۹: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"مردی با ژاکت آبی در حال اسکی روی آب."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"مردی با ژاکت آبی در حال پارو زدن در آب است."

توصیف‌های مرجع:

"قایق‌ران خانم ایرانی درحال پارو زدن در آب است ."

"دختری با پرچم ایران در قایق در حال پارو زدن است ."

"یک زن ورزش‌کار در حال قایق‌رانی در دریاچه می‌باشد ."

"یک زن قایق‌ران وسط دریاچه پارو می‌زند و به هر دو پارو پرچم وصل‌شده‌است ."

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A man on a skis in the water."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A man on a skis in the water."

توصیف‌های مرجع:

”An Iranian woman rower is rowing in the water.”

”A girl with an Iranian flag is rowing in a boat.”

”A sports woman is rowing in the lake.”

”A rowing woman is rowing in the middle of the lake and a flag is attached to both oars.”

شکل ۶ـ۲۰: نمونه‌ای از تصاویر مجموعه دادگان

فارسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

”ورزشکار المپیکی در حالی که ورزشکار المپیکی در حال مبارزه با حریف خود است، به نظر می رسد که به او ضربه می زند.”

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

”ورزشکار المپیکی در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت های کشتی آزاد و فرنگی مردان در طول رقابت”

توصیف‌های مرجع:

"بازیکن‌های کشتی با لباس کشتی آبی و قرمز درحال مسابقه هستند ."

"دو کشتی گیر با لباس‌های قرمز و آبی روی تشک در حال کشتی گرفتن هستند ."

"دو مرد کشتی‌گیر روی تشک کشتی با هم مسابقه می‌دهند ."

"دو کشتی‌گیر مرد روی تشک مسابقه برای ضربه زدن به یکدیگر تلاش می‌کنند .".

انگلیسی:

توصیف CLIPCapبا یادگیری بدون نمونه:

"A man standing on top of a basketball court."

توصیف CLIPCap با یادگیری با ۱۶ نمونه:

"A man standing on top of a basketball court."

توصیف‌های مرجع:

"Wrestling players in blue and red wrestling uniforms are competing."

"Two wrestlers in red and blue clothes are wrestling on the mat."

"Two male wrestlers compete on the wrestling mat."

"Two male wrestlers try to hit each other on the mat."

۶ـ۲ـ۲ بررسی عملکرد مدل‌های توصیف تصویر

نمودار ۶ـ۲۱ امتیاز مدل‌های دو زبان را در دو یادگیری بدون نمونه و با تعداد نمونه محدود بر پایه معیارهای BLEU و ROUGE نشان می‌دهد. توجه داشته باشید که در مرحلهٔ آموزش یادگیری با تعداد نمونه محدود، تعداد epoch برابر با ۵ و ضریب یادگیری ۲×۱۰^{-۶} در نظر گرفته شده است.

روند تغییر امتیاز BLEU و ROUGE تقریبا یکسان است و در بیشتر تست‌ها دسته‌ها به همان ترتیب افزایش امتیاز BLEU، امتیاز ROUGEشان نیز افزایش پیدا کرده‌است. با این حال، تفاوت اولین دسته‌ها در امتیاز ROUGE-2 نسبت به دسته‌های بعدیشان به طور مشخصی بیشتر از دو امتیاز دیگر است. نکته قابل توجه دیگر این است که دسته ماشین‌ها تنها در زبان فارسی و امتیاز ROUGE-L از دسته افراد بالاتر است.

دسته‌ای که در زبان انگلیسی بیشترین امتیاز را کسب کرده دستهٔ "افراد" است و دستهٔ "ماشین‌ها" نیز به آن نزدیک است. در این زبان ورزش کمترین امتیاز را دارد، اما در زبان فارسی بیشترین امتیاز را به خود تخصیص داده است. تقریبا ترتیب دقت مدل‌های هر دو زبان در همه دسته‌ها یکسان است و این موضوع تنها برای دستهٔ

ورزش صدق نمی‌کند. توصیف‌های تولید شده توسط مدل انگلیسی در دستهٔ ورزش در مقایسه با توصیف‌های مدل فارسی بسیار متفاوت از توصیف‌های مرجع است (به طور مثال، توصیف‌های مدل فارسی فعالیت‌های ورزشی افراد درون تصویر را مانند پرتاب توپ، تیراندازی و ... در بسیاری از موارد دربردارد که مشابه با توصیف مرجع است (مانند شکل ۶ـ۱۸). همچنین مدل فارسی از کلمهٔ ورزشکار در بسیاری از توصیف‌ها بهره برده در حالی که در توصیف‌های مدل انگلیسی این موارد وجود ندارند یا بسیار کم هستند.). همچنین مدل فارسی بهتر از مدل انگلیسی ورزش درون تصویر را حدس می‌زند و تشخیص مدل انگلیسی در بسیاری از نمونه‌ها اشتباه است. این‌ها دلایلی برای تفاوت زیاد دقت در دستهٔ ورزش هستند. در دیگر دسته‌ها تفاوت توصیف‌های تولید شده مدل انگلیسی و فارسی کمتر است.

مدل فارسی CLIPCap به طور جالبی عملکرد بسیار بهتری از مدل انگلیسی آن را دارد. (برای مثال مدل فارسی برای دسته افراد امتیاز BLEU حدودا ۴ برابر، امتیاز ROUGE-2 حدودا دو برابر و ROUGE-L سه برابر بیشتر کسب کرده است.) همان طور که در بخش شرح نحوهٔ ایجاد و آموزش مدل‌های توصیف تصویر توضیح دادیم، آموزش مدل فارسی در کل با حجم کمتری از داده اجرا شده است (بخش CLIP در مدل فارسی تنها با ۴۰۰ هزار داده آموزش داده شده، در حالی که CLIP انگلیسی از میلیون‌ها داده برای آموزش استفاده کرده‌است. البته این مدل ازپیش آموزش داده شده برای شروع آموزش CLIP فارسی استفاده شده است.) و شبکه نگاشت در مدل فارسی بسیار کوچک‌تر از شبکهٔ نگاشت مدل انگلیسی است؛ به همین دلیل انتظار می‌رفت که عملکرد ضعیف‌تری نسبت به مدل انگلیسی داشته باشد. از دلایلی که ممکن است باعث بهتر شدن عملکرد مدل فارسی شده باشد می‌توان به آموزش اضافهٔ شبکه نگاشت و GPT-2 با هم برای ۲ epoch اشاره کرد که این مرحله برای مدل انگلیسی انجام نشده و تنها آموزش شبکه نگاشت به تنهایی برای مسئله توصیف تصویر انگلیسی اجرا شده است. دلیل دیگر می‌تواند در نظر گرفتن ویژگی‌های تصویر بیشتری نسبت به مدل انگلیسی باشد (مدل فارسی ۶۷۸ عدد ویژگی تصویر استخراج شده از کدگذار را برای نگاشت استفاده می‌کند در حالی که مدل انگلیسی ۶۴۰ ویژگی را به شبکهٔ نگاشت می‌دهد). مورد دیگری که می‌توان به آن اشاره کرد موضوع ترجمهٔ اتوماتیک مجموعه دادگان برای زبان انگلیسی است، که ممکن است باعث ایجاد برخی توصیف‌های مرجع نادرست دربارهٔ تصویر شود و مدل در این موارد امتیاز کمتری با مقایسه با توصیف‌های مرجع نادرست به دست آورد. به طور کلی با توجه به امتیازهای کسب شده، نتایج یادگیری بدون نمونه و با نمونه‌های معدود روی این مدل به نظر زیاد مطلوب نیست. با این حال، با بررسی نمونه توصیف‌های تولید شده توسط این مدل‌ها می‌توان پی برد که این مدل‌ها با وجود تنها آموزش دادن یک شبکهٔ نگاشت بین دو شبکهٔ مخصوص درک تصویر و تولید

متن، توصیف‌های مناسبی برای تصاویر تولید می‌کنند. این مورد به خصوص در زبان فارسی اهمیت دارد که مدل‌های زیادی با عملکرد خوب در مسئله توصیف تصویر برای این زبان وجود ندارد. از دلایلی که باعث پایین آمدن امتیاز این مدل‌ها با وجود تولید توصیف‌های قابل قبول می‌شود نداشتن توصیف‌های مرجع بیشتر برای پوشش کامل تمام توصیف‌های ممکن برای یک تصویر است. ما تلاش کردیم با انتخاب چهار توصیف نوشته شده توسط چهار فرد مختلف به عنوان توصیف مرجع این مشکل را کاهش دهیم، اما با توجه به نمونه‌های قرار داده شده مشخص است که بسیاری از توصیف‌های تولید شده توسط مدل با وجود درست بودنشان، با توصیف‌های جمع‌آوری شده تفاوت زیادی دارند.

حال به بررسی جزئی‌تر عملکرد مدل‌ها در موضوع‌های مختلف مربوط به فرهنگ ایران می‌پردازیم. یکی از مسائل مهم در فرهنگ ایران، حجاب است. ما تلاش کردیم با گردآوری تصاویری از پوشش‌های مختلف زنان ایرانی آزمایش کنیم که آیا مدل می‌تواند جنسیت فرد درون تصویر را تشخیص دهد یا خیر. مدل‌ها در این مورد عملکرد قابل قبولی داشتند (مراجعه شود به شکل‌های ۶ـ۲ و ۶ـ۳) و می‌توانستند به خوبی جنسیت خانم‌ها با پوشش اسلامی را تشخیص دهند. البته در حالاتی تشخیص آن‌ها اشتباه می‌شد. یکی از حالات پوشش‌های ورزشی اسلامی خانم‌ها بود که به دلیل نوع خاص آن‌ها و نبود ویژگی‌هایی که این نوع لباس‌ها زنانه به نظر برسند (رنگ، وجود دامن یا طرح‌های مخصوص) باعث می‌شد مدل جنسیت فرد را مذکر حدس بزند یا به جنسیت او اشاره نکند (شکل ۶ـ۱۹). این مورد یکی از اشتباه‌های پرتکرار مدل (به خصوص مدل فارسی) برای دستهٔ ورزش بود. حالت دیگر وجود تعداد زیادی (معمولا بیشتر از ۳ نفر) خانم در تصویر بود که باعث اشتباه مدل می‌شد (شکل ۶ـ۴ و توصیف انگلیسی شکل ۶ـ۲). این مورد ممکن است به دلیل ناتوانی مدل در توجه به تعداد زیادی از افراد و دسته‌بندی کردن آن‌ها براساس ویژگی مشابهشان هنگام تولید توصیف باشد. با توجه به نتایج، مدل‌ها درباره فعالیت‌ها و مکان‌های اسلامی دچار اشتباه می‌شوند، به طوری که یا آن‌ها را از مذهب و دین دیگری در نظر می‌گیرند (اشاره به کلیسا در شکل ۶ـ۹ و عبادتگاه بودایی در شکل ۶ـ۱۶) یا در تشخیص خاص بودن آن‌ها ناتوانند و ویژگی‌های کلی آن تصاویر را بیان می‌کنند (شکل ۶ـ۱۲). احتمال می‌دهیم در آموزش کدگذار تصویری CLIP داده‌های مربوط به مذاهب دیگر بیشتر بوده و به همین دلیل تمایل بیشتری به اشاره به آن‌ها نسبت به اسلام دارد.

یکی دیگر از موارد بررسی شده، دقت مدل‌ها در تشخیص تصاویر مربوط به قوم‌های مختلف ایران و لباس‌ها و مراسم محلی آنان است. مدل فارسی در این مورد عالی عمل کرده و توانسته محلی بودن لباس و فعالیت‌ها را به خوبی تشخیص دهد، اما مدل انگلیسی عملکردش متوسط است. (شکل ۶ـ۱ و ۶ـ۱۷) مورد دیگر،

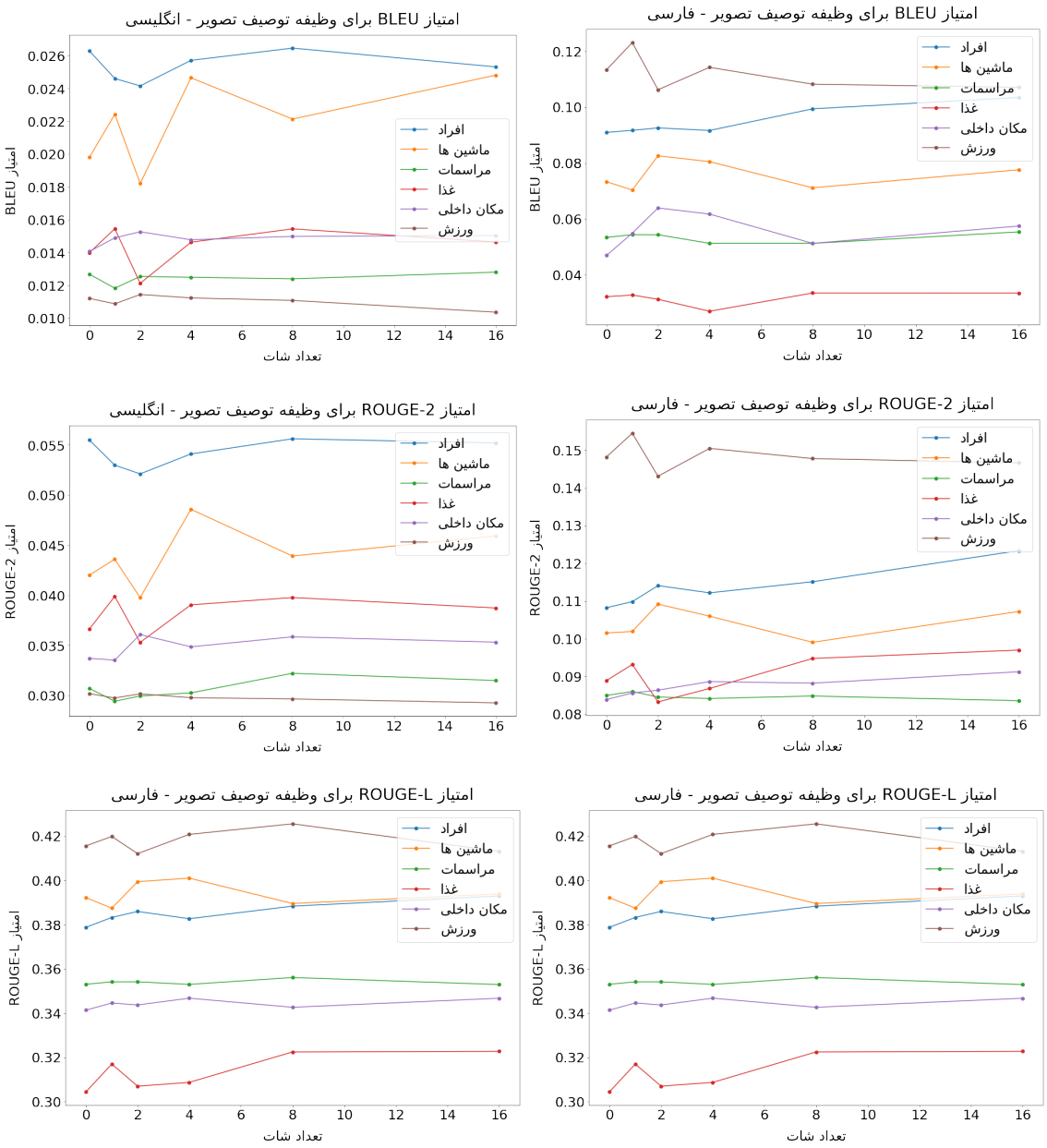
ورزش‌هایی‌ست که در ایران شهرت بیشتری دارند، مانند ورزش‌های وزنه‌برداری، کشتی و .... مدل‌ها در تشخیص این نوع ورزش‌ها به مشکل خورده و یا ورزش را به کلی اشتباه تشخیص می‌دهند یا درباره ورزش درون تصویر صحبت نمی‌کنند (مدل فارسی توصیفی از ورزش شکل ۶ـ۲۰ تولید نکرده و مدل انگلیسی به اشتباه آن را تشخیص داده‌است). البته تنها همین ورزش‌ها نیستند که به اشتباه تشخیص داده شده‌اند و مدل‌ها به خصوص مدل انگلیسی در تشخیص ورزش ضعیف عمل کرده‌اند (شکل ۶ـ۱۹).

توصیف مدل‌ها از غذاهای ایرانی بسیار جالب است. در یادگیری بدون نمونه و با نمونه‌های معدود، مدل تلاش می‌کند براساس دسته‌هایی که تاکنون دیده است دسته‌های جدید را توصیف کند و این اتفاق به طور واضحی در دستهٔ غذاهای ایرانی دیده می‌شود. به طور مثال در شکل ۶ـ۱۳ مدل‌ها خورش کرفس را به صورت کاسه ای پر از سبزیجات و گوشت توصیف کرده‌اند و در شکل ۶ـ۱۴ ته چین وجود دارد که به دلیل مشابهت کیک توصیف شده‌است. ما امیدوار بودیم که مدل‌ها در این دسته عملکرد بهتری داشته باشند و توصیف دقیق‌تری از غذاها بدهند اما نتیجه این چنین نبود.

با توجه به این که تمرکز پروژه روی آزمایش شناخت مدل از فرهنگ ایران است، مدل‌ها در بعضی موارد نمی‌توانند توصیف درستی دربارهٔ اشیا، افراد و ویژگی‌های محیط که ربطی به فرهنگ ایران ندارد ارائه بدهند. این مشکل نیز باعث کاهش امتیازها شده‌است. به طور مثال مدل فارسی در بعضی نمونه‌ها توصیف‌هایی با عبارات تکراری ایجاد می‌کند (شکل ۶ـ۱۰ و ۶ـ۲۰). مدل انگلیسی نیز در بسیاری از نمونه‌ها توصیف بسیار کلی و یا کوتاه ارائه می‌دهد که نمی‌توان به وسیلهٔ آن‌ها میزان درک مدل را از ویژگی‌های خاص تصویر را بررسی کرد (شکل ۶ـ۱۱ و شکل ۶ـ۹). علاوه بر این، در بعضی موارد مدل‌ها تنها به شمایل کلی تصویر توجه کرده و توصیفی می‌سازند که دربارهٔ تصویر صدق نمی‌کند (شکل ۶ـ۱۵).

با توجه به نمودار ۶ـ۲۱ دربارهٔ یادگیری بدون نمونه و یا با نمونه‌های معدود نیز می‌توان به چند نکته اشاره کرد. می‌توان دریافت که تقریبا از تعداد ۸ به بالا نمونهٔ آموزشی، عملکرد مدل‌ها حدودا ثابت می‌شود و عملکرد یادگیری با ۱۶ نمونه نسبت به ۸ نمونه تغییر زیادی نمی‌کند. همچنین عملکرد مدل‌ها با افزایش تعداد نمونه‌های آموزشی همواره بهبود نمی‌یابد، در صورتی که این نمونه‌های آموزشی هر چند در تعداد کم، از بین نمونه‌های تست جدا شده‌اند و توزیعی مشابه آنان دارند پس انتظار می‌رود که با دیدن نمونه‌هایی مشابه نمونه‌های تست عملکرد مدل‌ها بهتر شود. در بیشتر نمونه‌های تست بررسی شده تعداد نمونهٔ آموزشی (شات) تاثیری در تغییر توصیف تولید شده برای یک تصویر ندارد (شکل ۶ـ۱۶ و ۶ـ۱۱ و ۶ـ۹ و ۶ـ۷ برای توصیف فارسی و شکل ۶ـ۱ و شکل‌های ۶ـ۵ و ۶ـ۲۰ و ۶ـ۱۹ و ۶ـ۱۵ برای توصیف انگلیسی) و در بعضی موارد

باعث بهبود (توصیف فارسی شکل‌های ۶ـ۱۷ و ۶ـ۲۰ و ۶ـ۱۹ و ۶ـ۵ و توصیف انگلیسی شکل‌های ۶ـ۱۶ و ۶ـ۱۳ و ۶ـ۷) و در تعدادی باعث بدتر شدن توصیف (توصیف فارسی شکل‌های ۶ـ۱، ۶ـ۱۰ و ۶ـ۱۴ و توصیف انگلیسی شکل ۶ـ۸) شده است.



شکل ۶ـ۲۱: بررسی یادگیری بدون نمونه و با تعداد نمونه محدود مدل CLIPCap هر دو زبان برپایه معیارهای BLEU، ROUGE-L و ROUGE-2

۶ـ۲ـ۳ نمونه‌هایی از نتایج آزمایش مدل‌های پرسش و پاسخ تصویری

در ادامه به بررسی چند نمونه از پیش‌بینی‌های مدل پرسش و پاسخ تصویری می‌پردازیم. پرامپت (محرک)های استفاده شده در زبان فارسی به ترتیب عبارتند از:

- "پرسش: [Q] پاسخ: [A]"
- "[Q] [A]"
- "[A] پاسخ [Q] است."

و در زبان انگلیسی عبارتند از:

- "Question: [Q] Answer: [A]"
- "[Q] [A]"
- "Retrieve [A] from [Q]"

شکل ۶ـ۲۲: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: در کاسه کوچکتر سالاد است؟

پاسخ اصلی: بله

یادگیری بدون نمونه:

۱. پرسش:در کاسه کوچکتر سالاد است؟ پاسخ:بازار
۲. در کاسه کوچکتر سالاد است؟ لوستر
۳. چهار پاسخ در کاسه کوچکتر سالاد است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:در کاسه کوچکتر سالاد است؟ پاسخ:بازار
۲. در کاسه کوچکتر سالاد است؟ لوستر
۳. چهار پاسخ در کاسه کوچکتر سالاد است؟ است.

شکل ۶ـ۲۳: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: مرد از روی چه چیزی می‌پرد؟

پاسخ اصلی: آتش

یادگیری بدون نمونه:

۱. پرسش:مرد از روی چه چیزی می‌پرد؟ پاسخ:آتش
۲. مرد از روی چه چیزی می‌پرد؟ آتش
۳. آتش پاسخ مرد از روی چه چیزی می‌پرد؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:مرد از روی چه چیزی می‌پرد؟ پاسخ:آتش
۲. مرد از روی چه چیزی می‌پرد؟ آتش
۳. آتش پاسخ مرد از روی چه چیزی می‌پرد؟ است.

شکل ۶ـ۲۴: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصوری

پرسش اصلی: تصویر چه مکانی را نشان می‌دهد؟

پاسخ اصلی: خیابان

یادگیری بدون نمونه:

۱. پرسش:تصویر چه مکانی را نشان می‌دهد؟ پاسخ:خیابان
۲. تصویر چه مکانی را نشان می‌دهد؟ خیابان
۳. خیابان پاسخ تصویر چه مکانی را نشان می‌دهد؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:تصویر چه مکانی را نشان می‌دهد؟ پاسخ:خیابان
۲. تصویر چه مکانی را نشان می‌دهد؟ خیابان
۳. خیابان پاسخ تصویر چه مکانی را نشان می‌دهد؟ است.

شکل ۶ـ۲۵: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: زنی که عینک دارد در کدام سمت تصویر است؟

پاسخ اصلی: راست

یادگیری بدون نمونه:

۱. پرسش:زنی که عینک دارد در کدام سمت تصویر است؟ پاسخ:جلوی عروس و داماد
۲. زنی که عینک دارد در کدام سمت تصویر است؟ جلوی عروس و داماد
۳. سه پاسخ زنی که عینک دارد در کدام سمت تصویر است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:زنی که عینک دارد در کدام سمت تصویر است؟ پاسخ:جلوی عروس و داماد
۲. زنی که عینک دارد در کدام سمت تصویر است؟ جلوی عروس و داماد
۳. سه پاسخ زنی که عینک دارد در کدام سمت تصویر است؟ است.

شکل ۶ـ۲۶: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: بازیکن تیم ایران لباس چه رنگی پوشیده؟

پاسخ اصلی: سفید

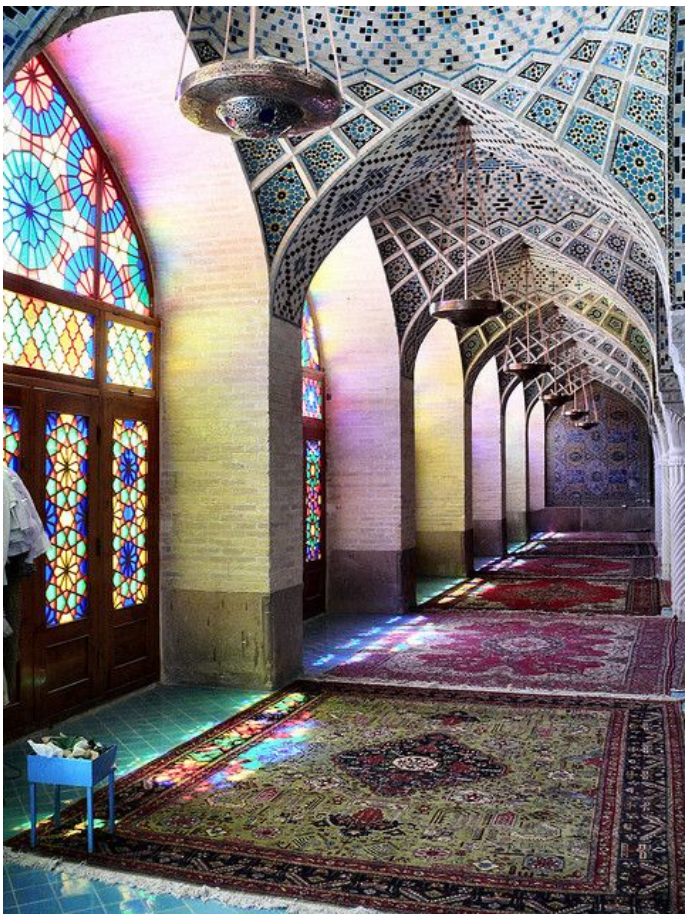
یادگیری بدون نمونه:

۱. پرسش:بازیکن تیم ایران لباس چه رنگی پوشیده؟ پاسخ:بسکتبال

۲. بازیکن تیم ایران لباس چه رنگی پوشیده؟ بسکتبال
۳. بسکتبال پاسخ بازیکن تیم ایران لباس چه رنگی پوشیده؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:بازیکن تیم ایران لباس چه رنگی پوشیده؟ پاسخ:بسکتبال
۲. بازیکن تیم ایران لباس چه رنگی پوشیده؟ بسکتبال
۳. بسکتبال پاسخ بازیکن تیم ایران لباس چه رنگی پوشیده؟ است.



شکل ۶ـ۲۷: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: سقف مکان چه شکلی است؟
پاسخ اصلی: گنبدی

یادگیری بدون نمونه:

۱. پرسش:سقف مکان چه شکلی است؟ پاسخ:نماز جماعت
۲. سقف مکان چه شکلی است؟ گنبدی
۳. شیشه پاسخ سقف مکان چه شکلی است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:سقف مکان چه شکلی است؟ پاسخ:نماز جماعت
۲. سقف مکان چه شکلی است؟ گنبدی
۳. شیشه پاسخ سقف مکان چه شکلی است؟ است.

شکل ۶ـ۲۸: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: علاوه بر زعفران چه چیزی روی برنج است؟

پاسخ اصلی: مرغ

یادگیری بدون نمونه:

۱. پرسش:علاوه بر زعفران چه چیزی روی برنج است؟ پاسخ:مرغ
۲. علاوه بر زعفران چه چیزی روی برنج است؟ دلمه ی بادمجان
۳. خورش کرفس پاسخ علاوه بر زعفران چه چیزی روی برنج است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:علاوه بر زعفران چه چیزی روی برنج است؟ پاسخ:مرغ
۲. علاوه بر زعفران چه چیزی روی برنج است؟ دلمه ی بادمجان
۳. خورش کرفس پاسخ علاوه بر زعفران چه چیزی روی برنج است؟ است.

شکل ۶ـ۲۹: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟

پاسخ اصلی: بله

یادگیری بدون نمونه:

۱. پرسش:آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ پاسخ:کافه‌های خیابانی

۲. آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ پارک
۳. خیابان پاسخ آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ پاسخ:کافه‌های خیابانی
۲. آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ پارک
۳. خیابان پاسخ آیا همه ی دوچرخه سواران در تصویر کلاه ایمنی بر سر دارند؟ است.

شکل ۶ـ۳۰: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: چه غذایی در بشقاب است؟

پاسخ اصلی: دلمه ی بادمجان

یادگیری بدون نمونه:

۱. پرسش:چه غذایی در بشقاب است؟ پاسخ:مرغ
۲. چه غذایی در بشقاب است؟ مشکی و آبی
۳. پرتاب نیزه پاسخ چه غذایی در بشقاب است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:چه غذایی در بشقاب است؟ پاسخ:مرغ
۲. چه غذایی در بشقاب است؟ مشکی و آبی
۳. پرتاب نیزه پاسخ چه غذایی در بشقاب است؟ است.

شکل ۶ـ۳۱: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: ماشین سمت راست چه رنگی است؟

پاسخ اصلی: زرد

یادگیری بدون نمونه:

۱. پرسش:ماشین سمت راست چه رنگی است؟ پاسخ:صورتی
۲. ماشین سمت راست چه رنگی است؟ مقنعه
۳. پرتاب تیر پاسخ ماشین سمت راست چه رنگی است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:ماشین سمت راست چه رنگی است؟ پاسخ:صورتی
۲. ماشین سمت راست چه رنگی است؟ مقنعه
۳. پرتاب تیر پاسخ ماشین سمت راست چه رنگی است؟ است.

شکل ۶ـ۳۲: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: زن در حال انجام چه ورزشی است؟

پاسخ اصلی: وزنه برداری

یادگیری بدون نمونه:

۱. پرسش:زن در حال انجام چه ورزشی است؟ پاسخ:پرتاب نیزه

۲. زن در حال انجام چه ورزشی است؟ اسکی
۳. اسکی پاسخ زن در حال انجام چه ورزشی است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:زن در حال انجام چه ورزشی است؟ پاسخ:پرتاب نیزه
۲. زن در حال انجام چه ورزشی است؟ اسکی
۳. اسکی پاسخ زن در حال انجام چه ورزشی است؟ است.

شکل ۶ـ۳۳: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: آیا در تصویر مردی دیده می‌شود؟

پاسخ اصلی: خیر

یادگیری بدون نمونه:

۱. پرسش:آیا در تصویر مردی دیده می‌شود؟ پاسخ:عروسی
۲. آیا در تصویر مردی دیده می‌شود؟ پایکوبی
۳. عروسی پاسخ آیا در تصویر مردی دیده می‌شود؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:آیا در تصویر مردی دیده می‌شود؟ پاسخ:عروسی
۲. آیا در تصویر مردی دیده می‌شود؟ پایکوبی
۳. عروسی پاسخ آیا در تصویر مردی دیده می‌شود؟ است.

شکل ۶ـ۳۴: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: آیا در تصاویر لامپ دیده می شود؟
پاسخ اصلی: بله
یادگیری بدون نمونه:

۱. پرسش:آیا در تصاویر لامپ دیده می شود؟ پاسخ:مرکز خرید
۲. آیا در تصاویر لامپ دیده می شود؟ مرکز خرید
۳. مرکز خرید پاسخ آیا در تصاویر لامپ دیده می شود؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:آیا در تصاویر لامپ دیده می شود؟ پاسخ:مرکز خرید
۲. آیا در تصاویر لامپ دیده می شود؟ مرکز خرید
۳. مرکز خرید پاسخ آیا در تصاویر لامپ دیده می شود؟ است.

شکل ۶ـ۳۵: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: صاحب مغازه پیرزن است؟

پاسخ اصلی: خیر

یادگیری بدون نمونه:

۱. پرسش:صاحب مغازه پیرزن است؟ پاسخ:مرد
۲. صاحب مغازه پیرزن است؟ مرد
۳. مرد پاسخ صاحب مغازه پیرزن است؟ است.

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. پرسش:صاحب مغازه پیرزن است؟ پاسخ:مرد
۲. صاحب مغازه پیرزن است؟ مرد
۳. مرد پاسخ صاحب مغازه پیرزن است؟ است.

شکل ۶ـ۳۶: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: What is the man holding?

پاسخ اصلی: Bread

یادگیری بدون نمونه:

۱. Question: What is the man holding? Answer: Bread
۲. What is the man holding? Bread

۳. Retrieve Bread from What is the man holding?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: What is the man holding? Answer: Bread

۲. What is the man holding? Bread

۳. Retrieve Bread from What is the man holding?

شکل ۶ـ۳۷: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: What sports are the people in the picture doing?

پاسخ اصلی: Ski

یادگیری بدون نمونه:

۱. Question: What sports are the people in the picture doing? Answer: Ski

۲. What sports are the people in the picture doing? Ski

۳. Retrieve Ski from What sports are the people in the picture doing?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: What sports are the people in the picture doing? Answer: Ski
۲. What sports are the people in the picture doing? Ski
۳. Retrieve Ski from What sports are the people in the picture doing?

شکل ۶ـ۳۸: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: What sport is the woman doing?

پاسخ اصلی: weightlifting

یادگیری بدون نمونه:

۱. Question: What sport is the woman doing? Answer: Iran
۲. What sport is the woman doing? Iran
۳. Retrieve weightlifting from What sport is the woman doing?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: What sport is the woman doing? Answer: Iran
۲. What sport is the woman doing? Iran

۳. Retrieve weightlifting from What sport is the woman doing?

شکل ۶ـ۳۹: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: What are the men doing?

پاسخ اصلی: worship

یادگیری بدون نمونه:

۱. Question: What are the men doing? Answer: shrine

۲. What are the men doing? Congregational Prayer

۳. Retrieve Congregational Prayer from What are the men doing?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: What are the men doing? Answer: shrine

۲. What are the men doing? Congregational Prayer

۳. Retrieve Congregational Prayer from What are the men doing?

شکل ۶ـ۴۰: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: Are all people male?

پاسخ اصلی: no

یادگیری بدون نمونه:

۱. Question: Are all people male? Answer: Car exhibition

۲. Are all people male? Car exhibition

۳. Retrieve Car exhibition from Are all people male?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: Are all people male? Answer: Car exhibition

۲. Are all people male? Car exhibition

۳. Retrieve Car exhibition from Are all people male?

شکل ۶ـ۴۱: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: What is spread on the floor?

پاسخ اصلی: Carpet

یادگیری بدون نمونه:

۱. Question: What is spread on the floor? Answer: rug

۲. What is spread on the floor? carpet weaving

۳. Retrieve rug from What is spread on the floor?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: What is spread on the floor? Answer: rug

۲. What is spread on the floor? carpet weaving

۳. Retrieve rug from What is spread on the floor?

شکل ۶ـ۴۲: نمونه‌ای از تصاویر مجموعه دادگان برای بررسی عملکرد مدل پرسش و پاسخ تصویری

پرسش اصلی: Is there a red chair in the picture?

پاسخ اصلی: Yes

یادگیری بدون نمونه:

۱. Question: Is there a red chair in the picture? Answer: basketball

۲. Is there a red chair in the picture? basketball

۳. Retrieve basketball from Is there a red chair in the picture?

یادگیری با تعداد نمونه معدود ـ ۱۶ شات:

۱. Question: Is there a red chair in the picture? Answer: basketball

۲. Is there a red chair in the picture? basketball

۳. Retrieve basketball from Is there a red chair in the picture?

۶ـ۲ـ۴ بررسی عملکرد مدل‌های پرسش و پاسخ تصویری

نمودار ۶ـ۴۳ میزان دقت مدل در دو یادگیری بدون نمونه و با تعداد نمونه معدود را نشان می‌دهد. نمودار هر زبان نشان دهندهٔ آزمایش مدل آن زبان با ۳ محرک یا پرامپت است. مطابق نمودارها، در زبان فارسی پرامپت "[پرسش] [پاسخ]" و در زبان انگلیسی پرامپت "Retrieve [A] from [Q]" بهترین درصد دقت را کسب کرده‌اند. به طور کلی تفاوت محسوسی میان دقت مدل انگلیسی و فارسی مشاهده نمی‌شود اما مدل‌های زبان انگلیسی برای هر ۳ پرامپت دقت بالاتری کسب کرده‌اند.

حال به بررسی بیشتر نمودارها می‌پردازیم. در مجموعه دادهٔ انگلیسی، ۱۸۸ عدد برچسب متفاوت وجود دارد؛ به عبارتی دیگر سطح تصادفی دقت[1] حدود ۰.۵۳ درصد است. بهترین مدل توانسته به دقت ۵.۶ درصد برسد و دقت آن بیشتر از ۱۰ برابر بهتر از پیش‌بینی تصادفی است. در مجموعه داده فارسی، ۱۹۲ عدد برچسب متفاوت وجود دارد؛ یعنی سطح تصادفی دقت حدود ۰.۵۲ درصد است. بهترین مدل توانسته به دقت ۴.۶ درصد برسد و دقت آن حدود ۸.۸ برابر پیش‌بینی تصادفی است.

با وجود این که انتظار می‌رفت با زیاد کردن تعداد شات‌ها در یادگیری با تعداد نمونه معدود، دقت مدل بیشتر شود اما در نمودارهای هر دو زبان با زیاد کردن تعداد شات‌ها همانند نمودارهای مدل‌های توصیف تصویر، نوسان دیده می‌شود.

تعداد نمونه‌هایی که همهٔ پرامپت‌ها در هر دو نوع یادگیری به آن پاسخ صحیح داده‌اند، کم است؛ شکل‌های ۶ـ۲۳ و ۶ـ۲۴ از مثال‌هایی هستند که پاسخ‌شان مناسب است.

در وظیفهٔ پرسش و پاسخ تصویری، علاوه بر لزوم ادراک تصویر و مفاهیم موجود در آن، نیاز به درک صحیح پرسش نیز هست. در صورتی که مدل نتواند همزمان درک خوبی از تصویر و پرسش داشته باشد، در پیش‌بینی

[1] Chance level

پاسخ دچار مشکل خواهد شد. مشکل مدل CLIP آزمایش شده این است که برای مسئلهٔ پرسش و پاسخ تصویری تنظیم دقیق نشده‌است، به همین دلیل سعی در اجرای همان کاری را دارد که برایش از پیش آموزش دیده است، یعنی تطابق تصویر و متن و انتخاب متنی که بیشترین شباهت را با تصویر دارد. از این رو در بعضی موارد بدون توجه به پرسش، پاسخی را انتخاب می‌کند که متن تشکیل شده از پرسش و پاسخ بیشترین شباهت را به تصویر داشته باشد. به طور مثال در شکل ۶ـ۲۶، تصویر دو ورزشکار در حال انجام بازی بسکتبال دیده می‌شود؛ اما پرسش مربوط به رنگ شی است. پاسخی که مدل به این پرسش داده «بسکتبال» است تا شباهت متن را به تصویر بیشتر کند. مثال‌های دیگر از این قبیل عبارتند از: شکل ۶ـ۳۳ که یک جشن و مراسم شادمانی را نشان می‌دهد اما پرسش از نوع بله/خیر بوده، شکل ۶ـ۳۴ که تصویری از یک مرکز خرید است، اما پرسش از نوع بله/خیر بوده، شکل ۶ـ۳۵ که در آن یک مرد مشاهده می‌شود اما پرسش از نوع بله/خیر بوده است. از دیگر نمونه‌ها می‌توان به شکل‌های ۶ـ۲۲ و ۶ـ۲۹ اشاره کرد که همهٔ پرامپت‌ها در هر دو نوع یادگیری پاسخ اشتباه و نامربوط تولید کرده‌اند. مدل به همین دلیل به طور خاص در پاسخ به سوالات بله/خیر نتوانسته‌است عملکرد خوبی از خود نشان دهد و همواره پاسخی غیر از بله / خیر را انتخاب کرده تا شباهت به تصویر را افزایش دهد.

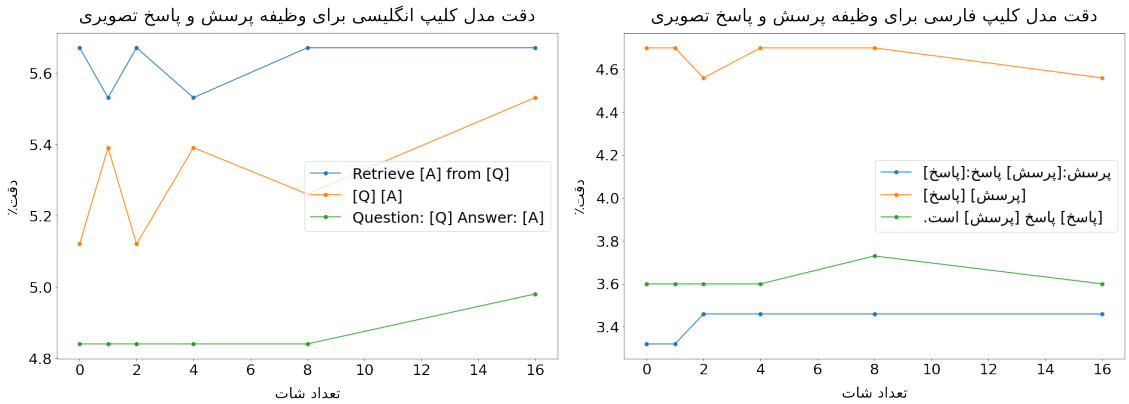
از سوی دیگر، مدل در بعضی نمونه‌ها با وجود درک پرسش، نتوانسته‌است شی مورد نظر را به خوبی تشخیص دهد که این می‌تواند به دلیل کافی نبودن آموزش آن باشد، به خصوص مدل فارسی که تنها با ۴۰۰ هزار داده آموزش دیده است. به طور مثال در شکل ۶ـ۲۵ سوال در رابطه با استدلال مکانی پرسیده شده است. با توجه به پاسخ‌های پرامپت اول و دوم، به نظر می‌رسد مدل سوال را به درستی فهمیده است، اما موفق نشده بر اساس تصویر به آن پاسخ صحیح دهد. مثال‌های دیگر از این نوع عبارتند از: شکل ۶ـ۳۰ که پرسش از نوع غذا و پاسخ پرامپت اول هم نشان‌دهنده یک غذاست، شکل ۶ـ۳۱ پرسش از رنگ ماشین ولی پاسخ پرامپت اول رنگ دیگری است، شکل ۶ـ۳۲ که پرسش از فعالیت ورزشی است اما پاسخ تمامی پرامپت‌ها به ورزشی دیگر اشاره دارد.

در برخی نمونه‌ها، تنها بعضی از پرامپت‌ها پاسخ را به درستی پیش‌بینی کرده‌اند که این مثال‌ها می‌توانند نشان‌دهندهٔ تاثیر پرامپت در یادگیری و عملکرد مدل باشند. مثلا در شکل ۶ـ۲۷ مدلی که با پرامپت دوم آموزش دیده و در شکل ۶ـ۲۸ مدلی که با پرامپت اول آموزش دیده پاسخ صحیح را پیش‌بینی کرده‌اند.

نکاتی که در مورد نمونه‌های فارسی ذکر شد، در زبان انگلیسی نیز دیده می‌شود. شکل ۶ـ۳۶ تا ۶ـ۴۲ چند نمونه از پیش‌بینی مدل‌های انگلیسی هستند. در هنگام گردآوری مجموعه داده، سعی شد تا حد ممکن

برچسب‌های موجود در مجموعه داده معانی مشابه نداشته باشند.[۲]

اما در بعضی موارد اگر پاسخ‌ها با دقت بیشتری جمع‌آوری می‌شدند عملکرد مدل بهبود می‌یافت. به طور مثال در تصویر ۶ـ۳۹ برچسب worship (به معنی عبادت) و پیش‌بینی مدل با پرامپت دوم و سوم Congregational Prayer (به معنی نماز جماعت) است. در این نمونه پیش‌بینی مدل از برچسب اصلی موفق‌تر ظاهر شده است. در مثالی دیگر در شکل ۶ـ۴۱، برچسب carpet (به معنی فرش) و پیش‌بینی مدل‌های آموزش دیده با پرامپت اول و سوم rug (به معنی قالیچه) است؛ فرش و قالیچه تا حدودی شباهت دارند. وجود چنین مواردی می‌توانند از دلایل پایین آمدن دقت مدل باشند.



شکل ۶ـ۴۳: بررسی یادگیری بدون نمونه و با تعداد نمونه محدود مدل CLIP برای مسئلهٔ پرسش و پاسخ تصویری

[۲] به طور مثال تبدیل "سمت راست" به "راست" که در فصل پیش به آن اشاره شد.

فصل ۷

جمع‌بندی و پیشنهادها

۷ـ۱ مقدمه

در این پژوهش، به آزمایش عملکرد مدل‌هایی که تنها با تصاویر خارجی آموزش داده شده‌اند بر روی تصاویر مخصوص فرهنگ ایرانی برای دو مسئله پرسش و پاسخ تصویری و توصیف تصویر و در دو زبان فارسی و انگلیسی پرداخته شده است. برای اجرای آزمایش به کمک نیروی انسانی مجموعه دادگان کوچکی برای این دو مسئله ایجاد کرده‌ایم. نحوهٔ آزمایش، به کار گیری دو نوع یادگیری بدون نمونه و با نمونه‌های معدود است و برای داشتن نتیجهٔ قابل قبول از مدلی که در این دو نوع یادگیری عملکرد خوبی از خود نشان داده است (CLIP) استفاده کرده‌ایم. برای مدل پرسش و پاسخ تصویری نیاز به یک محرک داشتیم تا بتوانیم پرسش و پاسخ را با هم در یک متن جایگذاری کنیم بدین ترتیب سه محرک را برای هر زبان امتحان کردیم.

۷ـ۲ نتیجه‌گیری

مدلی که برای مسئله پرسش و پاسخ تصویری استفاده شد، مدل CLIP به تنهایی و بدون هیچ شبکهٔ اضافی بود. این مدل در اصل برای تشخیص میزان شباهت یک تصویر و یک متن آموزش دیده است و برای مسئلهٔ پرسش و پاسخ تصویری تنظیم نشده است. ما در این پژوهش بدون تنظیم دقیق مدل برای این مسئله روی آن آزمایش اجرا کردیم تا عملکرد مدل اصلی CLIP را در مسائل خاص بررسی کنیم. با وجود این که دقت مدل از حالت انتخاب تصادفی یک پاسخ برای یک پرسش بالاتر است (حدود شش تا ده برابر بیشتر برای محرک‌های مختلف) اما دقتش بسیار پایین‌تر از آخرین مدل‌های ارائه شده برای این مسئله است. بدین ترتیب لازم است مدل برای مسئله خاص پرسش و پاسخ تصویری تنظیم دقیق شود تا بتوان نتایج آن را برای تصاویر فرهنگ ایرانی بررسی کرد.

اما مدل توصیف تصویر به نام CLIPCap، مدلی که از سه شبکهٔ مختلف (که یکی از آن‌ها CLIP است) ساخته شده است، برای این مسئله تنظیم دقیق شده (هر چند تنظیم دقیق آن می‌تواند با تعداد دورهای بیشتری انجام شود و جا برای پیشرفت دارد) و می‌تواند توصیف‌های مناسبی با رعایت دستور زبان برای یک تصویر ایجاد کند. از آزمایش این مدل با تصاویر مخصوص فرهنگ ایرانی می‌توان چند نتیجه گرفت. مدل تا حد خوبی می‌تواند جنسیت زنان باحجاب را تشخیص دهد، اما در شناخت مکان‌ها و مراسم مذهبی مشکل دارد و نمی‌تواند مذهبی بودن آن و یا دین مربوط به آن را به درستی تشخیص دهد. این مدل در مواجهه با تصاویر غذاهای ایرانی، ساده‌ترین ویژگی‌های آن غذا را بیان می‌کند و به تصاویر محلی مربوط به قوم‌های مختلف ایران

صفت «محلی» یا «سنتی» نسبت می‌دهد.

آخرین جنبهٔ قابل بررسی، مقایسهٔ یادگیری بدون نمونه و با نمونه‌های محدود است. بر خلاف انتظار، آموزش با چند نمونه حتی با وجود کم بودن تعداد آن، همیشه عملکرد مدل را بهبود نداده و در مواردی باعث بدتر شدن نتیجهٔ مدل می‌شد. هم‌چنین با آزمایش تعداد نمونه‌های (شات‌های) ۱، ۲، ۴، ۸ و ۱۶ برای هر کدام از مسائل می‌توان برداشت کرد که برای مدل توصیف تصویر تعداد نمونه‌های بیشتر از ۸ تاثیر زیادی در پیشرفت مدل نداشته است اما در مدل پرسش و پاسخ تصویری که نیاز به آموزش بسیار بیشتری برای داشتن عملکرد خوب دارد هر چه تعداد نمونه‌ها بیشتر شود تاثیر خود را در عملکرد مدل می‌گذارد.

در پایان می‌توان اشاره کرد که برای به کار بردن مدل CLIP در مسئلهٔ پرسش و پاسخ تصویری، بهترین محرک‌ها بین محرک‌های آزمایش شده "Retrieve [A] from [Q]" برای زبان انگلیسی ([A] همان پاسخ جاساز شده و [Q] پرسش جاساز شده است) و "[پرسش] [پاسخ]" برای زبان فارسی می‌باشد.

۷ـ۳ نوآوری

در مسائل پرسش و پاسخ تصویری و توصیف تصویر، تا به حال به موضوع تصاویر مربوط به فرهنگ یک کشور پرداخته نشده‌است. در این پژوهش ما تلاش کردیم تا مدل‌هایی که تا به حال تنها به کمک تصاویر خارجی آموزش داده شده بودند را با تصاویر مخصوص کشور خود، ایران، آزمایش کنیم تا میزان شناخت آن‌ها را از مسائل مختلف فرهنگی و مذهبی که خود در ایران با آن روبه‌رو هستیم دریابیم.

هم‌چنین یادگیری بدون نمونه و با نمونه‌های معدود از موضوعات جدیدی هستند که قابل توسعه هستند. مدل CLIP که برای پیدا کردن تشابه تصویر و متن به کمک این دو نوع یادگیری نتیجهٔ شگفت‌انگیزی در مقایسه با دیگر مدل‌ها داشته است، به طور خاص در بسیاری از مسائل مربوط به تصویر و متن آزمایش نشده‌است. در این پژوهش ما این مدل را با دو نوع یادگیری نام برده شده برای دو مسئله پرسش و پاسخ تصویری و توصیف تصویر به کار بردیم.

در پایان می‌توان به کمبود مجموعه دادگان‌های با کیفیت فارسی برای این دو مسئله اشاره کرد. به همین دلیل ما مجموعهٔ کوچکی از نمونه‌ها را به کمک نیروی انسانی گردآوری و پردازش کرده‌ایم. از ویژگی‌های خاص این مجموعه داده، تصاویر گرفته شده از فرهنگ ایران و نقاط مختلف ایران است.

۷ـ۴ پیشنهادها

مسئله‌ای که در این پژوهش به آن پرداخته شد، مسئله جدیدی است، از این رو موضوعات و جنبه‌های بسیاری برای تحقیق و توسعه در آن وجود دارد. به طور خلاصه می‌توان به موارد زیر اشاره کرد:

- تنظیم دقیق مدل کلیپ به طور خاص برای مسئله پرسش و پاسخ تصویری با مجموعه دادگان مخصوص این مسئله و تکرار آزمایش‌ها برای بررسی میزان پیشرفت دقت مدل نسبت به مدل تنظیم نشده. ممکن است مدل تنظیم شده بتواند با پیشرفته‌ترین مدل‌های این حوزه رقابت کند.
- آزمایش محرک‌های بیشتر برای مسئله پرسش و پاسخ تصویری
- رفع نیازمندی مدل CLIP به استفاده از محرک‌های ساخته شدهٔ دستی برای حوزهٔ پرسش و پاسخ تصویری یا تولید محرک اتوماتیک با توجه به تصویر و دیگر شرایط مسئله
- جمع‌آوری توصیف‌های متنوع بیشتر برای پوشش حالات ممکن توصیف برای یک تصویر در مسئله توصیف تصویر
- دسته‌بندی همه جنبه‌های فرهنگ ایرانی و جدا کردن تصاویر مربوط به هر دسته به صورت جزئی‌تر و دقیق‌تر از دسته‌بندی ما در این پژوهش
- مقایسهٔ عملکرد یادگیری بدون نمونه یا با نمونه‌های معدود مدل‌های این پژوهش با دیگر مدل‌های ارائه شده برای مسائل پرسش و پاسخ تصویری و توصیف تصویر
- ساخت مدلی مشابه مدل توصیف تصویر برای مسئلهٔ پرسش و پاسخ تصویری به طوری که CLIP تنها بخشی از آن باشد که کار درک تصویر و پرسش را انجام داده و ویژگی‌های استخراج شده از CLIP به شبکه‌ای برای انتخاب پاسخ درست داده شود.
- تحقیق دربارهٔ دلیل نوسان در نمودارها و افزایش و کاهش امتیاز و دقت مدل‌ها هنگام افزایش تعداد نمونه‌ها در یادگیری با نمونه‌های معدود. ممکن است این نوسان با در نظر گرفتن seedهای مختلف برای انتخاب نمونه‌ها رفع شود.

مراجع

[1] The future of ai will be about less data, not more. https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more. Accessed: 2022-09-24.

[2] An introduction to few-shot learning. https://www.analyticsvidhya.com/blog/2021/05/an-introduction-to-few-shot-learning/. Accessed: 2022-10-02.

[3] What is meta-learning in machine learning? https://machinelearningmastery.com/meta-learning-in-machine-learning/. Accessed: 2022-10-09.

[4] Al Faraby, H., Azad, M. M., Fedous, M. R., Morol, M. K., et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. in *2020 23rd international conference on computer and information technology (ICCIT)* (2020), IEEE, pp. 1–6.

[5] Ayoubi, S. Sajjadayoubi/clip-fa-vision. https://huggingface.co/SajjadAyoubi/clip-fa-vision. Accessed: 2022-08-20.

[6] Ayoubi, S. Cc3mfa dataset, 2021.

[7] Ayoubi, S. Cc3mfa dataset ver.2, 2021.

[8] Ayoubi, S. Coco-flickr farsi dataset, 2021.

[9] Ayoubi, S. Flickrfa dataset, 2021.

[10] Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., and Cucchiara, R. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2022), pp. 4661–4669.

[11] bolbolzaban. Gpt2-persian. https://huggingface.co/bolbolzaban/gpt2-persian. Accessed: 2022-08-28.

[12] Brownlee, J. A gentle introduction to calculating the bleu score for text in python. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/. Accessed: 2022-09-30.

[13] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[14] Chiusano, F. Two minutes nlp — learn the rouge metric by examples. https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499. Accessed: 2022-09-30.

[15] Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., and Bansal, M. Fine-grained image captioning with clip reward, 2022.

[16] Doshi, K. Foundations of nlp explained — bleu score and wer metrics. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b. Accessed: 2022-09-30.

[17] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems 28* (2015).

[18] Hitschler, J., Schamoni, S., and Riezler, S. Multimodal pivots for image caption translation. *arXiv preprint arXiv:1601.03916* (2016).

[19] Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., and Parikh, D. Pythia v0.1: the winning entry to the vqa challenge 2018. *ArXiv abs/1807.09956* (2018).

[20] Khan, H. R., Gupta, D., and Ekbal, A. Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation. *arXiv preprint arXiv:2109.04653* (2021).

[21] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. Ask me anything: Dynamic memory networks for natural language processing. in *Proceedings of The 33rd International Conference on Machine Learning* (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, eds. , volume 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 1378–1387.

[22] Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., and Xu, J. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia 21*, 9 (2019), 2347–2360.

[23] Luo, R. A better variant of self-critical sequence training, 2020.

[24] McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730* (2018).

[25] Mohtaj, S., Roshanfekr, B., Zafarian, A., and Asghari, H. Parsivar: A language processing toolkit for persian. in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[26] Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).

[27] OpenAI. Better language models and their implications. https://openai.com/blog/better-language-models/. Accessed: 2022-09-30.

[28] OpenAI. Clip: Connecting text and images. https://openai.com/blog/clip/. Accessed: 2022-09-20.

[29] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners.

[30] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21*, 140 (2020), 1–67.

[31] Rahman, M., Mohammed, N., Mansoor, N., and Momen, S. Chittron: An automatic bangla image captioning system. *Procedia Computer Science 154* (2019), 636–642.

[32] Rahman, S., Khan, S., and Porikli, F. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions on Image Processing 27*, 11 (2018), 5652–5667.

[33] Roshan-AI. hazm. https://github.com/sobhe/hazm, 2022.

[34] Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 2556–2565.

[35] Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can clip benefit vision-and-language tasks? *ArXiv abs/2107.06383* (2022).

[36] Song, H., Dong, L., Zhang, W., Liu, T., and Wei, F. Clip models are few-shot learners: Empirical studies on vqa and visual entailment.

[37] Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Proc. Neural Information Processing Systems* (2021).

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[39] Wolfe, C. Using clip to classify images without any labels. https://towardsdatascience.com/using-clip-to-classify-images-without-any-labels-b255bb7205de. Accessed: 2022-09-29.

[40] Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823* (2017).

[41] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78.

[42] Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. Deep modular co-attention networks for visual question answering. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 6274–6283.

واژه‌نامه فارسی به انگلیسی

امتیاز هم‌ترازی ..... Alignment Score
برچسب‌زنی بخش گفتار ..... Part of Speech Tagging
بینایی کامپیوتر ..... Computer Vision
پرسش و پاسخ تصویری ..... VQA (Visual Question Answering)
پیش‌پردازش ..... Pre-processing
تابع هزینه تقابلی ..... Contrastive loss
تجزیه وابستگی ..... Dependency Parsing
تنظیم دقیق ..... Fine-tuning
توصیف تصویر ..... Image Captioning
توکن ..... Token
جستجوی پرتو ..... Beam Search
دادهٔ برچسب‌دار ..... Labeled data
دسته بند ..... Classifier
زبان طبیعی ..... Natural Language
سوگیری ..... Bias
شناسایی اشیا ..... Object Recognition
شناسایی عواطف ..... Sentiment Analysis
فرایادگیری ..... Meta-Learning
قالب ..... Template
کدگذار ..... Encoder
کدگذاری ..... Encoding
کدگشا ..... Decoder
مجموعه دادگان ..... Dataset
مجموعه لغات ..... Vocabulary
محرک، پرامپت ..... Prompt
مدل از پیش آموزش دیده ..... Pre-trained Model

مدل زبانی ..... Language Model
مدل مبتنی بر ترنسفورمر ..... Transformer-based Model
مدل‌های تولیدکننده دنباله ..... Sequence Generation Models
وظایف پایین‌دستی ..... Downstream Tasks
هایپرپارامتر ..... Hyperparameter
یادگیری با ناظر ..... Supervised Learning
یادگیری با نمونه‌های معدود (آموزشی) ..... Few-shot Learning
یادگیری بدون ناظر ..... Unsupervised Learning
یادگیری بدون نمونه (آموزشی) ..... Zero-shot Learning
یادگیری بر پایهٔ محرک‌ها ..... Prompt-based Learning

واژه‌نامه انگلیسی به فارسی

بین حالتی ..... Cross-modal
جانمایی..... Embedding
ثابت کردن، غیر قابل آموزش کردن ..... Freeze
توصیف تصویر ..... Image Captioning
شبکه نگاشت..... Mapping network
پردازش زبان‌های طبیعی ..... Natural Language Processing
جانمایی‌های پیشوندها..... Prefix Embeddings
پیشرفته‌ترین مدل‌ها با بهترین عملکرد ..... State of The Art Models
مسئله، وظیفه، حوزه ..... Task
پرسش و پاسخ تصویری..... Visual Question Answering

Abstract:

One of the trends in the field of Artificial Intelligence nowadays is the ability to perceive images with the help of efficient models. The advent of machine learning solutions such as Image Captioning is a boon for blind and visually impaired people who are unable to understand visual images. Therefore, image descriptions can be provided to people by utilising an Image Captioning model so they can have a better understanding of their surroundings. In addition, with the help of Visual Question Answering models, user questions about an image can be answered appropriately. It is also possible to produce automatic question answering systems that have the ability to perceive images. These systems can communicate with humans and have many applications in daily life.

In recent years, many efforts have been made for both the Image Captioning and Visual Question Answering tasks, but most of them cannot perform well on images related to a specific culture. In this research, we intend to collect a dataset that represents the culture of Iran and evaluate the performance of one of the existing models (CLIP model) which has a better performance on understanding the relationship between an unseen image and text (better zero-shot performance) than most of similar models, on these two tasks with images related to Iranian culture in Persian and English. The types of learning tested are zero-shot and few-shot learning. In these experiments, we try to test the performance of the model on certain categories of images separately and improve its performance (in VQA) by trying different prompts.

Keywords: Artificial Intelligence, Natural Language Processing, Computer Vision, Visual Question Answering, Image Captioning, Zero-shot Learning, Few-shot Learning, Vision-&-Language Models, Persian Datasets, Analysis on the performance of models tested with images relating to Persian culture

Iran University of Science and Technology
Computer Engineering Department

Evaluation of Few-shot and Zero-shot Learning on a transformer-based model for Visual Question Answering and Image Captioning tasks

Bachelor of Science Thesis in Computer Engineering

By:

Shaghayegh Mobasher and Ghazal Zamaninejad

Supervisor:

Dr. Sayyed Sauleh Eetemadi

October 2022