دانشکده مهندسی کامپیوتر

پرسش و پاسخ تصویری

گزارش سمینار کارشناسی ارشد در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی

مریم سادات هاشمی

استاد راهنما

دکتر سید صالح اعتمادی

دی ۱۳۹۹


بسم الله الرحمن الرحیم

چکیده

مسئله پرسش و پاسخ تصویری یک مسئله چالش برانگیز است که در سال‌های اخیر معرفی شده است و مورد توجه بسیاری از محققان دو حوزه پردازش زبان طبیعی و بینایی ماشین قرار گرفته است. هدف این مسئله پاسخ به پرسش مطرح شده در مورد تصویر ورودی است. یک سیستم پرسش و پاسخ تصویری سعی می‌کند با استفاده از عناصر بصری تصویر و استنتاج جمع‌آوری شده از سوال متنی، پاسخ صحیح را پیدا کند. در فصل اول این بررسی، به معرفی مسئله پرسش و پاسخ تصویری، کاربرد و اهمیت آن و چالش‌های این مسئله می‌پردازیم. پس از تعریف برخی مفاهیم مورد نیاز در فصل دوم، مجموعه‌دادگان، روش‌های حل مسئله پرسش و پاسخ و تصویری و معیارهای ارزیابی آن را بررسی می‌کنیم. با توجه به موفقیت یادگیری عمیق و مدل‌های از قبل آموزش دیده، رویکردهای حل مسئله پرسش و پاسخ تصویری را به دو دسته کلی رویکردهای یادگیری عمیق و رویکردهای مدل‌های از قبل آموزش‌دیده نقسیم‌بندی می‌کنیم. در فصل آخر، پس از نتیجه‌گیری در مورد ابعاد مختلف مسئله پرسش و پاسخ تصویری، در مورد مسیرهای تحقیق در آینده بحث می‌کنیم.

واژگان کلیدی: پرسش و پاسخ تصویری، پردازش زبان طبیعی، بینایی ماشین، یادگیری عمیق، مدل‌های از قبل آموزش‌دیده

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

در سال‌های اخیر پیشرفت‌های زیادی در مسائل هوش مصنوعی و یادگیری عمیق که در تقاطع دو حوزه پردازش زبان طبیعی و بینایی ماشین قرار می‌گیرند؛ رخ داده است. یکی از مسائلی که اخیراً مورد توجه قرارگرفته است؛ پرسش و پاسخ تصویری است. با توجه به یک تصویر و یک سؤال به زبان طبیعی، سیستم سعی می‌کند با استفاده از عناصر بصری تصویر و استنتاج جمع آوری شده از سوال متنی، پاسخ صحیح را پیدا کند [۳۹]. پرسش و پاسخ تصویری نسخه گسترش یافته مسئله پرسش و پاسخ متنی است است که اطلاعات بصری به مسئله اضافه شده است. شکل ۱ ـ ۱ گویای تفاوت این دو مسئله است.

در سیستم پرسش و پاسخ متنی، یک متن و یک سوال متنی به عنوان ورودی به سیستم داده می‌شود و انتظار می‌رود که سیستم با توجه به درک و تفسیری که از متن و سوال بدست می‌آورد؛ یک جواب متنی را خروجی دهد. اما در سیستم پرسش و پاسخ تصویری، یک تصویر و یک سوال متنی به ورودی سیستم داده می‌شود و انتظار می‌رود که سیستم بتواند با استفاده از عناصر بصری تصویر و تفسیری که از سوال بدست می‌آورد؛ یک پاسخ متنی را در خروجی نشان دهد.

مسئله پرسش و پاسخ تصویری پیچیدگی بیشتری نسبت به مسئله پرسش و پاسخ متنی دارد زیرا تصاویر بعد بالاتر و نویز بیشتری نسبت به متن دارند. علاوه بر این، تصاویر فاقد ساختار و قواعد دستوری زبان هستند. در نهایت هم، تصاویر غنای بیشتری از دنیای واقعی را ضبط می‌کنند، در حالی که زبان طبیعی در حال حاضر نشانگر سطح بالاتری از انتزاع دنیای واقعی است [۶۵].

شکل ۱ـ۱: مثالی از سیستم پرسش و پاسخ متنی و تصویری

۱ـ۱ کاربرد و اهمیت مسئله

در طی سال‌های متمادی، محققان به دنبال ساخت ماشین‌هایی بودند که به اندازه‌ی کافی باهوش باشند که از آن به طور موثر همانند انسان‌ها برای تعامل استفاده کنند. مسئله‌ی پرسش و پاسخ تصویری یکی از پله‌های رسیدن به این رویای هوش مصنوعی است و از این جهت حائز اهمیت است.

کاربردهای بسیاری برای پرسش و پاسخ تصویری وجود دارد. یکی از مهم‌ترین موارد دستیار هوشمند برای افراد کم‌بینا و نابینا است [۱۸]. علاوه بر این، در سال های اخیر دستیاران صوتی [۱] و عامل‌های گفتگو [۲] مانند Cortana ، Siri و Alexa در بازار عرضه شدند که می‌توانند با انسان‌ها با استفاده از زبان طبیعی ارتباط برقرار کنند. در حال حاضر این دستیاران با استفاده از صوت و متن این ارتباط را برقرار می‌کنند در نتیجه گفتگوی بین این دستیاران با انسان‌ها مشابه دنیای واقعی نمی‌باشد. این ارتباط را می‌توان با استفاده از داده‌های تصویری و ویدئویی به واقعیت نزدیک‌تر کرد. اینجاست که مسئله‌ی پرسش و پاسخ تصویری برای نزدیک کردن تعامل بین انسان و عامل‌های گفتگو به دنیای واقعی می‌تواند موثر باشد. همین موضوع را می‌توانیم به صورت گسترده‌تری در ربات‌ها مشاهده کنیم. برای این‌که ربات بتواند بهتر با انسان‌ها ارتباط برقرار کند و به سوالات و درخواست‌ها پاسخ دهد؛ نیاز دارد که درک و فهم درستی از اطراف داشته باشد که این مستلزم داشتن تصویری دقیق از پیرامون است. بنابراین این ربات می‌تواند برای پاسخ به پرسش‌ها از

[۱] Voice Assistants
[۲] Conversational Agents

دانشی که از طریق تصویر پیرامون خود بدست می‌آورد، جواب درستی را بدهد.

کاربرد دیگر این مسئله در پزشکی است. در بسیاری از موارد تحلیل تصاویر پزشکی مانند تصاویر CT اسکن و x-ray برای یک پزشک متخصص هم دشوار است. اما یک سیستم پرسش و پاسخ تصویری می‌تواند با تحلیل و تشخیص موارد غیرطبیعی موجود در تصویر، به عنوان نظر دوم به پزشک متخصص کمک کند. از طرفی ممکن است در بعضی اوقات بیمار دسترسی به پزشک را نداشته باشد تا شرح تصاویر را متوجه شود. وجود سیستم پرسش و پاسخ تصویری می‌تواند آگاهی بیمار را نسبت به بیماری افزایش دهد و از نگرانی او بکاهد [۵۸].

۱ـ۲ بررسی چالش‌های موجود در این مسئله

در مقایسه با مسائل دیگری که مشترک بین پردازش زبان طبیعی و بینایی ماشین است مانند توصیف تصویر [۳] و بازیابی متن به تصویر [۴] ، مسئله پرسش و پاسخ تصویری چالش‌برانگیزتر است زیرا (۱) سوالات از پیش تعیین نشده است. به این معنی که در مسئله‌ای مانند تشخیص اشیا، سوال این است که چه اشیایی در تصویر وجود دارد و این سوال از پیش تعیین شده است و در طول حل مسئله تغییر نمی‌کند و تنها تصویر تغییر می‌کند که منجر به پاسخ‌ها متفاوت می‌شود. اما در پرسش و پاسخ تصویری، برای هر تصویر سوالات متفاوت و مرتبط با همان تصویر پرسیده می‌شود که در زمان اجرا تعیین می‌شود. (۲) اطلاعات موجود در تصویر ابعاد بالایی دارد که پردازش آن‌ها به زمان و حافظه زیادی نیاز دارد. (۳) مسئله پرسش و پاسخ تصویری نیاز به حل مسائل پایه‌ای و فرعی دارد مانند تشخیص اشیا [۵] (آیا در تصویر سگ وجود دارد؟)، تشخیص فعالیت [۶] (آیا کودک گریه می‌کند؟)، طبقه‌بندی صفات [۷] (چتر چه رنگی است؟)، شمارش(چند نفر در تصویر وجود دارد؟)، طبقه‌بندی صحنه [۸] (هوا بارانی است؟) و روابط مکانی بین اشیا(چه چیزی بین گربه و مبل است؟).

[۳] Image Captioning
[۴] Text-to-image Retrieval
[۵] Object Detection
[۶] Activity Recognition
[۷] Attribute Classification
[۸] Scene Classification

فصل ۲

تعاریف و مفاهیم مبنایی

همان‌طور که قبلاً اشاره شد مسئله پرسش و پاسخ تصویری در تقاطع دو حوزه پردازش زبان طبیعی و بینایی ماشین قرار می‌گیرد. از این رو قبل از بررسی کارهای مرتبط با مسئله پرسش و پاسخ تصویری، نیاز است تا با مفاهیم مربوط به این دو حوزه آشنا شویم. در ادامه این فصل به شرح مفاهیم و تعاریف پایه می‌پردازیم.

۲ـ۱ پردازش زبان طبیعی

پردازش زبان طبیعی [۱] یکی از زیرشاخه‌های علوم کامپیوتر و هوش مصنوعی است که به تعامل بین کامپیوتر و زبان‌های (طبیعی) انسانی می‌پردازد. هدف اصلی در پردازش زبان طبیعی، تحلیل زبانهای طبیعی به منظور آسان‌تر ساختن فهم آن‌ها برای کامپیوتر می‌باشد. مسلماً درصورتی که کامپیوتر بتواند توسط زبان‌های طبیعی با انسان ارتباط برقرار کند، بسیاری از مشکلات تعامل انسان با کامپیوتر حل شده و زندگی برای انسان‌ها راحت‌تر خواهد شد. با پیشرفت تکنولوژی و بوجود آمدن نیازهای متفاوت برای انسان‌ها، کاربردهای جدیدی برای این حوزه تعریف می‌شود. ترجمه ماشینی، خلاصه‌سازی متون، تحلیل احساسات، طبقه‌بندی متون، سیستم‌های توصیه‌گر، غلط‌یابی متون و ... از جمله مهم‌ترین کاربردهای پردازش زبان طبیعی است.

[۱] Natural Language Processing

۲ـ۲ بینایی ماشین

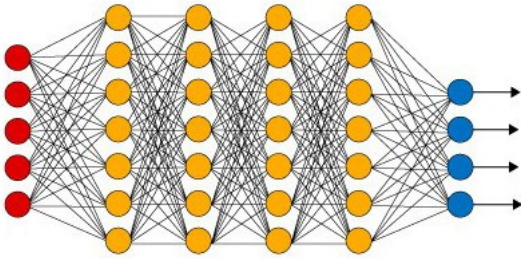
بینایی ماشین [۲] جز حوزه‌های در حال توسعه در علوم کامپیوتر و هوش مصنوعی محسوب می‌شود که سعی دارد از طریق پردازش تصاویر دوبعدی، جهان سه بعدی پیرامون را بازسازی و تفسیر کند. به بیان ساده، بینایی ماشین یعنی کامپیوترها بتوانند جهان را به کمک دوربین‌ها ببینند، بفهمند و حتی از بینایی انسان پیشی بگیرند. بینایی کامپیوتر دارای کاربردهای بسیار متنوعی مانند طبقه‌بندی اشیا، تشخیص اشیا، تقسیم‌بندی اشیا، تشخیص چهره و ... است.

۲ـ۳ یادگیری عمیق

یادگیری عمیق [۳] زیر شاخه‌ای از یادگیری ماشین [۴] است که تلاش می‌کند تا مفاهیم انتزاعی سطح بالا را با استفاده از نمونه‌های(دادگان) زیاد مدل نماید. بیشتر روش‌های یادگیری عمیق از معماری شبکه‌های عصبی مصنوعی [۵] استفاده می‌کنند. به همین دلیل است که اغلب از مدل‌های یادگیری عمیق به عنوان شبکه‌های عصبی عمیق یاد می‌شود. اصطلاح «عمیق» معمولاً به تعداد لایه‌های پنهان در شبکه عصبی اشاره دارد. شبکه‌های عصبی سنتی فقط شامل ۲ یا ۳ لایه پنهان هستند، در حالی که شبکه‌های عمیق می‌توانند تا ۱۵۰ لایه داشته باشند. مدل های یادگیری عمیق معمولا با استفاده از مجموعه‌های بزرگی از داده‌های دارای برچسب و معماری شبکه عصبی که ویژگی‌ها را مستقیماً از داده‌ها بدون نیاز به استخراج دستی ویژگی‌ها یاد می‌گیرند، آموزش می‌بینند.

در حالی که یادگیری عمیق برای اولین بار در دهه ۱۹۸۰ مطرح شد اما به دلیل تولید داده‌های زیاد، افزایش قدرت محاسباتی و پیشرفت الگوریتم‌های این حوزه شاهد پیشرفت چشمگیر یادگیری عمیق در سال‌های اخیر هستیم. در حال حاضر شبکه‌های عصبی عمیق در حوزه‌های زیادی از جمله پردازش زبان طبیعی، بینایی ماشین، پردازش گفتار و... کاربرد دارد. شبکه‌های عصبی کانولوشنی و شبکه‌های عصبی بازگشتی از مهم‌ترین و پرکاربردترین شبکه‌های یادگیری عمیق هستند.

[۲] Computer Vision
[۳] deep learning
[۴] machine learning
[۵] artificial neural networks



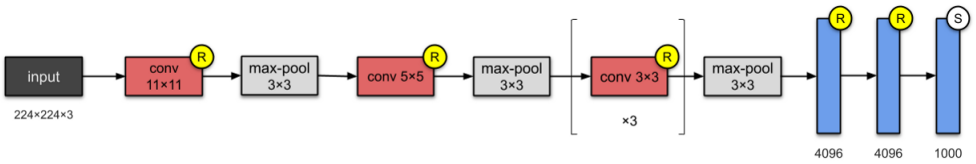
شکل ۲ـ۱: نمونه‌ای از شبکه عصبی عمیق. نورون‌های قرمز لایه‌ی ورودی ، نورون‌های نارنجی لایه‌ی مخفی و نورون‌های آبی لایه‌ی خروجی را نشان می‌دهند.

جدول ۲ـ۱: مقایسه مهم‌ترین شبکه‌های عصبی کانولوشنی آموزش‌دیده بر روی مجموعه‌داده ImageNet

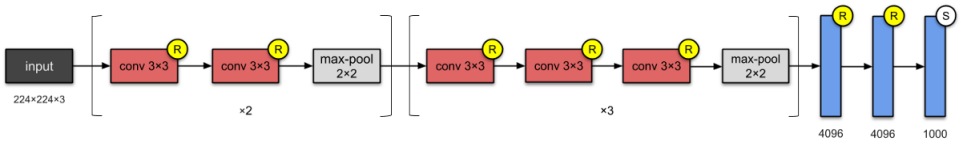
مدل CNN	سال	تعداد لایه‌ها	ابعاد ورودی	ابعاد خروجی	پارامترها
AlexNet [۲۰]	۲۰۱۲	۸	۲۲۷×۲۲۷	۴۰۹۶	۶۰ میلیون
VGGNet [۵۴]	۲۰۱۴	۱۹	۲۲۴×۲۲۴	۴۰۹۶	۱۳۸ میلیون
GoogleNet [۵۷]	۲۰۱۴	۲۲	۲۲۹×۲۲۹	۱۰۲۴	۵ میلیون
ResNet [۱۹]	۲۰۱۵	۱۵۲	۲۲۴×۲۲۴	۲۰۱۴۸	۲۶ میلیون

۲ـ۴ شبکه‌های عصبی کانولوشنی

شبکه‌های عصبی کانولوشنی [۶] دسته‌ای از شبکه‌های عصبی عمیق هستند که معمولاً برای تجزیه و تحلیل تصاویر استفاده می‌شوند. برخلاف شبکه‌های کاملاً متصل که هر نورون در یک لایه به همه نورون‌های لایه بعدی متصل است، در شبکه‌های عصبی کانولوشنی هر نورون تنها به بخشی از نورون‌های لایه‌ی بعدی متصل است. این خاصیت به دلیل انجام عملیات کانولوشن در شبکه‌های عصبی کانولوشنی است و باعث می‌شود که الگوهای محلی را از داده فرا بگیرند. در حالی که شبکه‌های کاملاً متصل الگوهای سراسری را یاد می‌گیرند. معمولا از شبکه‌های عصبی کانولوشنی برای استخراج ویژگی از تصویر استفاده می‌شود. در ادامه چند نمونه از برجسته‌ترین شبکه‌های عصبی کانولوشنی را معرفی می‌کنیم.



شکل ۲ـ۲: معماری شبکه AlexNet



شکل ۲ـ۳: معماری شبکه VGGNet

۲ـ۴ـ۱ AlexNet

در سال ۲۰۱۲، AlexNet به طور قابل توجهی بهتر از تمام رقبای قبلی عمل کرد. در این شبکه از ۵ لایه‌ی کانولوشنی و ۳ لایه‌ی کاملاً متصل استفاده شده است. معماری این شبکه در شکل ۲ـ۲ نمایش داده شده است.

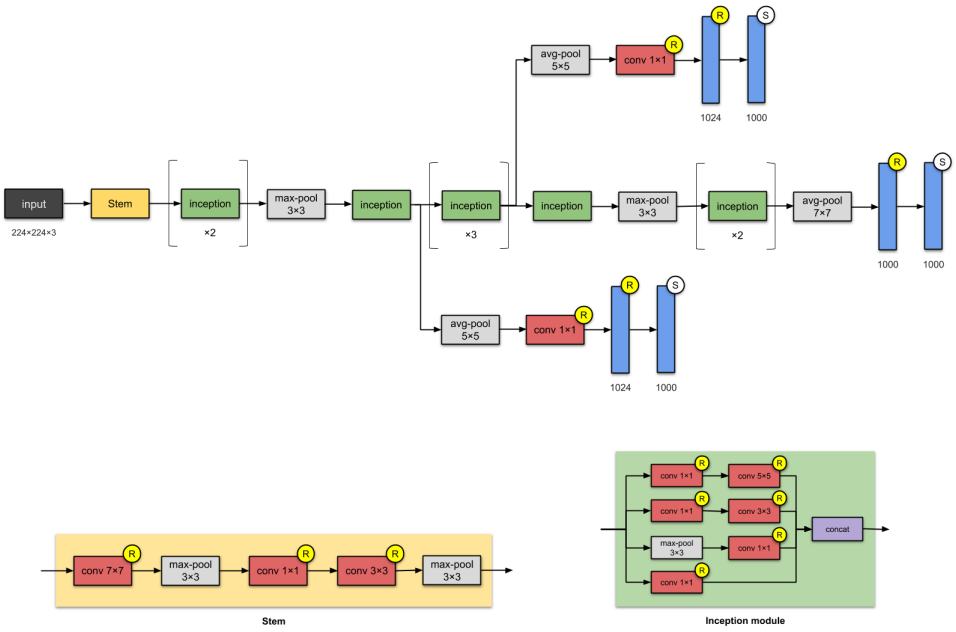
۲ـ۴ـ۲ VGGNet

شبکه VGGNet در سال ۲۰۱۴ معرفی شد. شبکه VGGNet از ۱۶ لایه کانولوشن تشکیل شده است و معماری بسیار یکنواختی دارد. شبکه VGGNet یکی از محبوب‌ترین شبکه‌ها برای استخراج ویژگی است. تعداد پارامترهای این شبکه برابر با ۱۳۸ میلیون است. معماری این شبکه در شکل ۲ـ۳ نشان داده شده است.

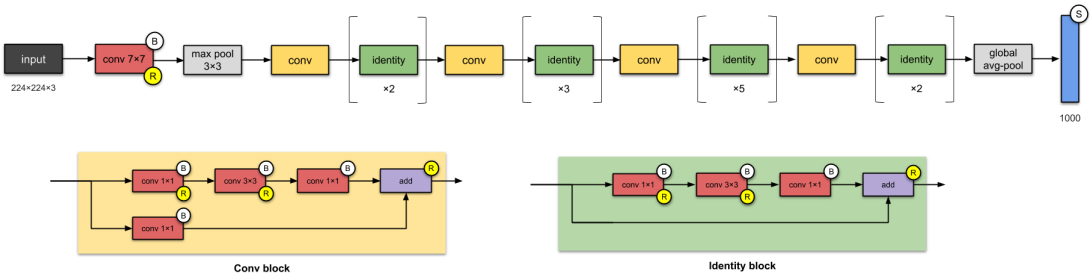
۲ـ۴ـ۳ GoogleNet

شبکه GoogleNet همزمان با شبکه VGGNet در سال ۲۰۱۴ معرفی شد و توانست بر شبکه VGGNet غلبه کند و دقت بالاتری را بر روی مجموعه داده ImageNet بدست آورد. مهم‌ترین علت موفقیت آن استفاده از ماژول inception بود که منجر به کاهش شدید تعداد پارامترها در این شبکه شد. شبکه GoogleNet از ۲۲ لایه با ۴ میلیون پارامتر تشکیل شده است. معماری این شبکه در شکل ۲ـ۴ نمایش داده شده است.

[۶] convolutional neural networks



شکل ۲ـ۴: معماری شبکه GoogleNet



شکل ۲ـ۵: معماری شبکه ResNet

۲ـ۴ـ۴ ResNet

شبکه ResNet با معرفی مفهوم جدید [skip connection] این امکان را برای شبکه‌های عصبی کانولوشنی ایجاد کرد که شبکه‌ها عمیق‌تر شوند و در عین حال آموزش در زمان کمتری انجام شود. با وجود اتصالات skip conection ورودی هر لایه بدون واسطه به لایه بعدی منتقل می‌شود بنابراین مشکل از بین رفتن گرادیان در شبکه‌های عمیق رفع می‌شود. ۱۵۲ لایه در شبکه ResNet به کار رفته است. معماری این شبکه در شکل ۲ـ۵ نشان داده شده است.

۲ ـ ۵ شبکه‌های عصبی بازگشتی

شبکه‌های عصبی بازگشتی [۷] دسته‌ی دیگری از شبکه‌های عصبی عمیق هستند که معمولا برای پردازش داده‌های دنباله‌دار مانند جملات، صوت و ویدئو استفاده می‌شوند. این شبکه‌ها دارای یک نوع حافظه هستند که اطلاعاتی تا کنون دیده‌اند را ضبط می‌کنند. در تئوری این‌طور به نظر می‌رسد که شبکه‌های عصبی بازگشتی می‌توانند اطلاعات موجود در یک دنباله طولانی را ضبط و از آن‌ها استفاده کنند اما در عمل این‌طور نیست و بسیار محدود هستند و فقط اطلاعات چند گام قبل را نگه می‌دارند. شبکه‌های عصبی بازگشتی پارامترهای مشابهی را بین همه گام‌های زمانی به اشتراک می‌گذارند . این بدین معنی است که در هر گام زمانی عملیات مشابهی را انجام می‌دهند و فقط ورودی‌ها متفاوت هستند. با این تکنیک تعداد کلی پارامترهایی که شبکه باید یاد بگیرد به شدت کاهش پیدا می‌کند. در ادامه این بخش به معرفی دو شبکه عصبی بازگشتی معروف می‌پردازیم.

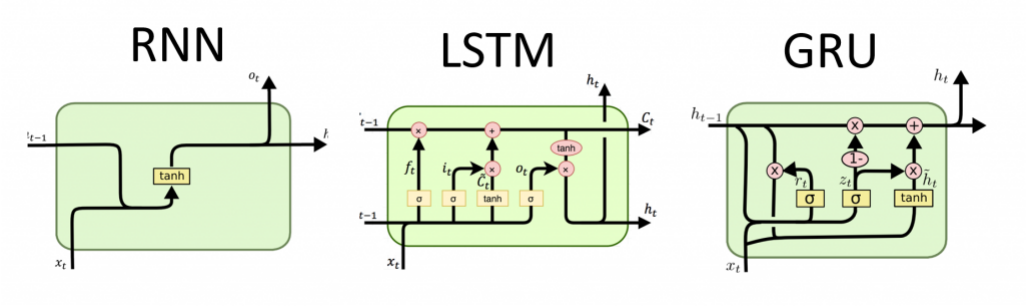
۲ ـ ۵ ـ ۱ LSTM

مدل LSTM [۸] در سال ۱۹۹۵ برای توسعه شبکه‌های عصبی بازگشتی ظهور پیدا کرد. شبکه LSTM برای حل مشکل ناپدید شدن گرادیان در شبکه‌های عصبی بازگشتی بوجود آمد. بزرگ‌ترین ویژگی LSTM امکان یادگیری وابستگی بلند مدت است که توسط شبکه‌های عصبی بازگشتی امکان‌پذیر نبود. برای پیش‌بینی گام زمانی بعدی نیاز است که مقادیر وزن‌ها در شبکه بروزرسانی شوند که این کار مستلزم حفظ اطلاعات گام‌های زمانی ابتدایی است. یک شبکه عصبی بازگشتی فقط می‌تواند تعداد محدودی از وابستگی‌های کوتاه مدت را یاد بگیرد ، اما سری‌های زمانی بلند مدت قایل یادگیری توسط شبکه‌های عصبی بازگشتی نیستند اما LSTM می‌تواند این وابستگی‌های بلند مدت را به درستی یاد بگیرند.

۲ ـ ۵ ـ ۲ GRU

یکی دیگر از شبکه‌های عصبی بازگشتی، GRU [۹] است که در سال ۲۰۱۴ معرفی شد. این شبکه نیز مانند LSTM مشکل ناپدیدشدن گرادیان در شبکه‌های عصبی بازگشتی را حل می‌کند. در واقع GRU نوع خاصی از LSTM است که با کم کردن تعداد دروازه‌ها، سرعت محاسبات را افزایش داده است.

[۷] recurrent neural networks
[۸] Long Short-Term Memory
[۹] Gated Recurrent Unit



شکل ۲ـ۶: مقایسه معماری شبکه‌های عصبی بازگشتی، LSTM و GRU [منبع]

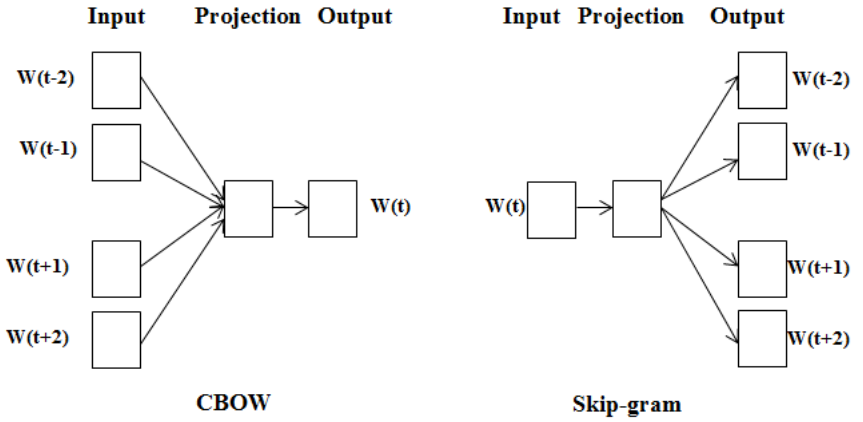
۲ـ۶ تعبیه کلمات

بیشتر الگوریتم‌های یادگیری ماشین و یادگیری عمیق قادر به پردازش متن به شکل خام وساده نیستند و برای بازنمایی متن‌ها نیاز به تعبیه کلمات [۱۰] دارند. تعبیه کلمات نگاشت کلمات یا عبارات از واژگان به بردارهای عددی است تا کامپیوترها بتوانند به راحتی آن‌ها را پردازش کنند. تعبیه کلمات عمدتاً برای مدل‌سازی زبان و یادگیری ویژگی در پردازش زبان طبیعی استفاده می‌شود. ایده اصلی در پشت تمام روش‌های تعبیه کلمات، گرفتن هرچه بیشتر اطلاعات معنایی و ریخت‌شناسی است. روش‌های تعبیه کلمات بسیاری در مسئله پرسش و پاسخ تصویری استفاده شده است. در ادامه به برجسته‌ترین و پرکاربردترین روش‌های تعبیه کلمات موجود و استفاده‌شده در مسئله پرسش و پاسخ تصویری می‌پردازیم و معایب و مزایای هر کدام را بررسی خواهیم کرد.

۲ـ۶ـ۱ کدگذاری one-hot

روش کدگذاری one-hot ساده‌ترین روش تعبیه کلمات است. در این روش یک لغت‌نامه از همه واژه‌های منحصربه‌فرد موجود در مجموعه‌داده ساخته‌می‌شود و اندیس یکتایی به هر واژه اختصاص می‌یابد. بنابراین برای هر واژه یک بردار به طول تعداد واژه‌ها ساخته می‌شود که تمامی مقادیر آن صفر است به جز اندیس مربوط به همان واژه که مقدار آن یک است. پیاده‌سازی این روش آسان است اما طول بردارها بزرگ است زیرا برابر با تعداد کل واژه‌های منحصر به فرد مجموعه داده است و هزینه زیادی برای ذخیره‌سازی دارد. بزرگترین عیب این روش این است که نمی‌توان از آن معنا و مفهوم استخراج کرد زیرا فاصله‌ی تمامی کلمات با هم یکسان است. در صورتی که ما انتظار داریم؛ کلماتی که مشابه هم هستند بردارهای نزدیک به هم یا مشابه هم داشته

[۱۰] word embedding



شکل ۲ـ۷: معماری شبکه CBOW و Skip-gram

باشند و کلملاتی که معنای متفاوتی با یکدیگر دارند تا حد امکان بردارهایشان از هم دور باشند.

۲ـ۶ـ۲ CBOW و Skip-gram

برای رفع مشکلات کدگذاری one-hot ، دو روش CBOW [11] [۴۰] و Skip-gram [۴۰] پیشنهاد شد که از شبکه‌های عصبی به عنوان جز اصلی خود استفاده می‌کنند. این دو مدل بر عکس هم کار می‌کنند. در هر دو مدل، از یک شبکه عصبی سه لایه که شامل لایه ورودی، لایه پنهان ولایه خروجی است، استفاده شده است. درمدل CBOW کلمات اطراف و نزدیک به یک کلمه(n-1کلمه) به لایه ورودی داده می‌شود و مدل سعی می‌کند این کلمه (nامین کلمه) را حدس بزند. بعد از آموزش این شبکه، وزن بین لایه‌ی پنهان و لایه خروجی، کلمات مجموعه داده را بازنمایی می‌کند که هر ستون آن بردار مربوط به یک کلمه را نشان می‌دهد. در مدل skip-gram برعکس CBOW یک کلمه به شبکه ورودی داده می‌شود و شبکه باید کلمات اطراف و نزدیک به آن را حدس بزند. معماری CBOW و Skip-gram در شکل ۲ـ۷ آورده شده است.

۲ـ۶ـ۳ GloVe

یکی دیگر از تعبیه کلمات مشهور، مدل بردار سراسری یا به اختصار GloVe [12] است که توسط پنینگتون و همکاران [۴۴] در سال ۲۰۱۴ در تیم پردازش زبان‌های طبیعی دانشگاه استنفورد معرفی و توسعه داده شد.

[11] Continouse Bag Of Words
[12] Global Vector

در GloVe فاصله میان بردارها نشان‌دهنده شباهت معنایی میان آن بردارها است.

۲ـ۶ـ۴ CNN، LSTM و GRU

با پیشرفت یادگیری عمیق در دهه اخیر، محققان برای استخراج ویژگی و بازنمایی متن از CNN ، LSTM [۲۱] و GRU [۱۰] استفاده می‌کنند. برای استخراج ویژگی از متن با استفاده از CNN بردارهای کلمات در کنار هم قرار داده می‌شود سپس به لایه‌های کانولوشنی یک بعدی داده می‌شود و فیلترهای متفاوتی بر روی آن‌ها اعمال می‌شود و پس از عبور از لایه max-pooling ویژگی‌ها بدست می‌آید. همچنین برای استخراج ویژگی از متن با استفاده از LSTM و GRU کافی است، بردار کلمات یک جمله به عنوان ورودی به این لایه‌ها داده شود.سپس خروجی آخرین گام زمانی به عنوان ویژگی کل جمله خواهد بود.

فصل ۳

مروری بر کارهای مرتبط

۳ـ۱ بررسی مجموعه دادگان مطرح این حوزه

در این بخش به معرفی مجموعه‌داده‌های مشهور در حوزه‌ی پرسش و پاسخ تصویری می‌پردازیم و ویژگی‌های هر کدام را بررسی خواهیم کرد. در جدول ۳ـ۱ اطلاعات آماری این مجموعه‌داده‌ها به صورت خلاصه آمده‌است.

۳ـ۱ـ۱ مجموعه داده DAQUAR [۳۷]

DAQUAR مخفف Dataset for Question Answering on Real World Images است که توسط مالینوفسکی منتشرشده‌است. این اولین مجموعه‌داده‌ای است که برای مسئله VQA منتشرشده‌است. تصاویر از مجموعه‌داده NYU-Depth V2 [۵۳] گرفته‌شده‌است. اندازه این مجموعه‌داده کوچک است و در مجموع ۱۴۴۹ تصویر دارد. DAQUAR شامل ۱۲۴۶۸ زوج پرسش و پاسخ با ۲۴۸۳ سوال منحصربه‌فرد است. برای تولید پرسش و پاسخ‌ها از دو روش مصنوعی و انسانی استفاده‌شده‌است. در روش مصنوعی پرسش و پاسخ‌ها به صورت خودکار از الگوهای موجود در جدول ۳ـ۲ تولیدشده‌است. در روش دیگر از ۵ نفر انسان خواسته‌شده‌است تا پرسش و پاسخ تولید کنند. تعداد پرسش و پاسخ‌های آموزشی در این مجموعه‌داده ۶۷۹۴ و تعداد پرسش و پاسخ‌های تست ۵۶۴ است و به طور میانگین برای هر عکس تقریبا ۹ پرسش و پاسخ وجود دارد. این مجموعه‌داده با مشکل بایاس روبه‌رو است زیرا تصاویر این مجموعه تنها مربوط به داخل خانه است و بیش از

جدول ۳ـ۱: بررسی اجمالی مجموعه‌داده‌های معروف در حوزه پرسش و پاسخ تصویری.

مجموعه‌داده	تعداد تصاویر	تعدادسوالات	سال انتشار
DAQUAR [۳۷]	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1 [۴]	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
Visual Madlibs [۶۸]	۱۰۷۳۸	۳۶۰۰۰۱	۲۰۱۵
Visual7w [۷۱]	۴۷۳۰۰	۲۲۰۱۱۵۴	۲۰۱۶
VQA v2 [۱۷]	۲۰۴۷۲۱	۱۱۰۵۹۰۴	۲۰۱۷
CLEVR [۲۳]	۱۰۰۰۰۰	۸۵۳۵۵۴	۲۰۱۷
Tally-QA [۱]	۱۶۵۰۰۰	۳۰۶۹۰۷	۲۰۱۹
KVQA [۵۰]	۲۴۶۰۲	۱۸۳۰۰۷	۲۰۱۹

جدول ۳ـ۲: الگوهای استفاده شده برای تولید سوال در مجموعه‌داده DAQUAR. سوالات می‌تواند در مورد یک تصویر و یا مجموعه‌ای از تصاویر باشد [۳۷].

	توضیح	الگو	نمونه
منفرد	شمارشی	How many {object} are in {image id}?	How many cabinets are in image1?
منفرد	شمارشی و رنگ	How many {color} {object} are in {image id}?	How many gray cabinets are in image1?
منفرد	نوع اتاق	HWhich type of the room is depicted in {image id}?	Which type of the room is depicted in image1?
منفرد	صفات عالی	What is the largest {object} in {image id}?	What is the largest object in image1?
مجموعه‌ای	شمارشی و رنگ	How many {color} {object}?	How many black bags?
مجموعه‌ای	نفی نوع ۱	Which images do not have {object}?	Which images do not have sofa?
مجموعه‌ای	نفی نوع ۲	Which images are not {room type}?	Which images are not bedroom?
مجموعه‌ای	نفی نوع ۳	Which images have {object} but do not have a {object}?	Which images have desk but do not have a lamp?

۴۰۰ مورد وجود دارد که اشیایی مثل میز و صندلی در پاسخ‌ها تکرارشده‌است.

شکل ۳ـ۱: چند نمونه از مجموعه‌داده DAQUAR [۳۷]

۳ـ۱ـ۲ مجموعه داده VQA [۴] [۱۷]

مجموعه‌داده Visual Question Answering v1(VQA v1) [۱] یکی از پرکاربردترین مجموعه‌داده‌ها در زمینه پرسش و پاسخ تصویری است. این مجموعه‌داده شامل دو بخش است. یک بخش از تصاویر واقعی ساخته‌شده‌است که VQA-real نام‌دارد و بخش دیگر با تصاویر کارتونی ساخته‌شده‌است که با نام VQA-abstract از آن در مقالات یاد می‌شود.

VQA-real به ترتیب شامل ۱۲۳۲۸۷ تصویر آموزشی و ۸۱۴۳۴ تصویر آزمایشی است که این تصاویر از مجموعه‌داده MS-COCO [۳۱] تهیه شده است. برای جمع‌آوری پرسش و پاسخ از نیروی انسانی استفاده‌شده‌است. برای هر تصویر حداقل ۳ سوال منحصربه‌فرد وجود دارد و برای هر سوال ۱۰ پاسخ توسط کاربرهای منحصر به فرد جمع‌آوری‌شده‌است. این مجموعه‌داده شامل ۶۱۴۱۶۳ سوال به صورت open-ended و چندگزینه‌ای است. در [۴] بررسی دقیقی در مورد نوع سوالات، طول سوالات و پاسخ‌ها و غیره انجام‌شده‌است.

VQA-abstract به عنوان یک مجموعه‌داده جداگانه و مکمل در کنار VQA-real قرار دارد. هدف از این مجموعه‌داده از بین بردن نیاز به تجزیه و تحلیل تصاویر واقعی است تا مدل‌ها برای پاسخ به سوالات تمرکز خود را بر روی استدلال‌های سطح بالاتری بگذارند. تصاویر کارتونی در این مجموعه‌داده به صورت دستی توسط انسان‌ها و به وسیله‌ی رابط کاربری که از قبل آماده‌شده‌است؛ ساخته‌شده‌است. تصاویر می‌تواند دو حالت را نشان‌دهند: داخل خانه و خارج از خانه که هر کدام مجموعه متفاوتی از عناصر را شامل می‌شوند از جمله

[۱] https://visualqa.org/

حیوانات، اشیا و انسان‌ها با حالت‌های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجادشده‌است. مشابه VQA-real ، ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع‌آوری‌شده‌است.

مجموعه‌داده Visual Question Answering v2(VQA v2) در سال ۲۰۱۷ پس از مجموعه‌داده VQA v1 معرفی شد. VQA v2 نسبت به VQA v1 متوازن تر است و تعصبات زبانی در VQA v1 را کاهش داده است. اندازه‌ی مجموعه داده‌ی VQA v2 تقریبا دو برابر مجموعه‌داده‌ی VQA v1 است. در مجموعه‌داده‌ی VQA v2 تقریبا برای هر سوال دو تصویر مشابه وجود دارد که پاسخ‌های متفاوتی برای سوال دارند.

Q: What shape is the bench seat ?
A: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved

Q: What color is the stripe on the train ?
A: white, white, white, white, white, white, white, white, white

Q: Where are the magazines in this picture ?
A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

شکل ۳ـ۲: چند نمونه از مجموعه‌داده VQA v1 - real [۴]

Q: Who looks happier ?.
A: old person, man, man, man, old man, man, man, man, grandpa

Q: Where are the flowers ?
A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree

Q: How many pillows ?
A: 1, 2, 2, 2, 2, 2, 2, 2, 2, 2

شکل ۳ـ۳: چند نمونه از مجموعه‌داده VQA v1 - abstarct [۴]

۳ـ۱ـ۳ مجموعه داده Visual Madlibs [۶۸]

مجموعه‌داده Visual Madlibs شکل متفاوتی از پرسش و پاسخ را ارائه می‌دهد. برای هر تصویر جملاتی در نظرگرفته شده‌است و یک کلمه از آن که معمولا مربوط به آدم، اشیا و فعالیت‌های نمایش‌داده‌شده در تصویر است؛ از جمله حذف‌شده و به جای آن جای‌خالی قرارگرفته‌است. پاسخ‌ها کلماتی هستند که این جملات

شکل ۳ـ۴: چند نمونه از مجموعه‌داده VQA v2 [۱۷]

را تکمیل می‌کنند. برای مثال جمله "دو[جای‌خالی]در پارک [جای‌خالی] بازی‌می‌کنند."در وصف یک تصویر بیان‌شده‌است که با دو کلمه "مرد" و "فریزبی" می‌توان جاهای‌خالی را پرکرد. این مجموعه‌داده شامل ۱۰۷۳۸ تصویر از مجموعه‌داده MS-COCO [۳۱] و ۳۶۰۰۰۱ جمله با جای‌خالی است. جملات با جای‌خالی به طور خودکار و با استفاده از الگوهای از پیش‌تعیین‌شده تولیدشده‌اند. پاسخ‌ها در این مجموعه‌داده به هر دو شکل open-ended و چندگزینه‌ای است.

۳ـ۱ـ۴ مجموعه داده Visual7w [۷۱]

مجموعه‌داده Visual7W نیز بر اساس مجموعه‌داده MS-COCO [۳۱] ساخته‌شده‌است. این مجموعه‌داده شامل ۴۷۳۰۰ تصویر و ۳۲۷۹۳۹ جفت سوال و پاسخ است. این مجموعه‌داده همچنین از ۱۳۱۱۷۵۶ پرسش و پاسخ چندگزینه‌ای تشکیل‌شده‌است که هر سوال ۴ گزینه دارد و تنها یکی از گزینه‌ها پاسخ صحیح سوال است. برای جمع‌آوری سوالات چندگزینه‌ای توسط انسان‌ها از پلتفرم آنلاین Amazon Mechanical Turk استفاده‌شده‌است. نکته‌ی حائز اهمیت در این مجموعه‌داده این است که تمامی اشیایی که در متن پرسش یا پاسخ ذکرشده‌است، به نحوی به کادر محدودکننده‌ی آن شی در تصویر مرتبط شده‌است. مزیت این روش، رفع ابهام‌های موجود در متن است. همان‌طور که از نام این مجموعه‌داده پیداست؛ سوالات آن با ۷ کلمه‌ی پرسشی که حرف اول آن w است شروع می‌شود. این ۷ کلمه شامل what ، where ، when ، who ، why ، how و which است. پرسش‌های Visual7W نسبت به به مجموعه‌داده VQA v1 غنی‌تر و سخت‌تر است. همچنین پاسخ‌ها طولانی‌تر هستند.

1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will throw the frisbee.
6. Person A is wearing blue shorts.
7. Person A is in front of person B.
8. Person A is blocking person B.
9. Person B is a young man wearing an orange hat.
10. Person B is on a grassy field.
11. Person B is holding a frisbee.
12. The frisbee is white and round.
13. The frisbee is in the hand of the man with the orange cap.
14. People could throw the frisbee.
15. The people are playing with the frisbee.

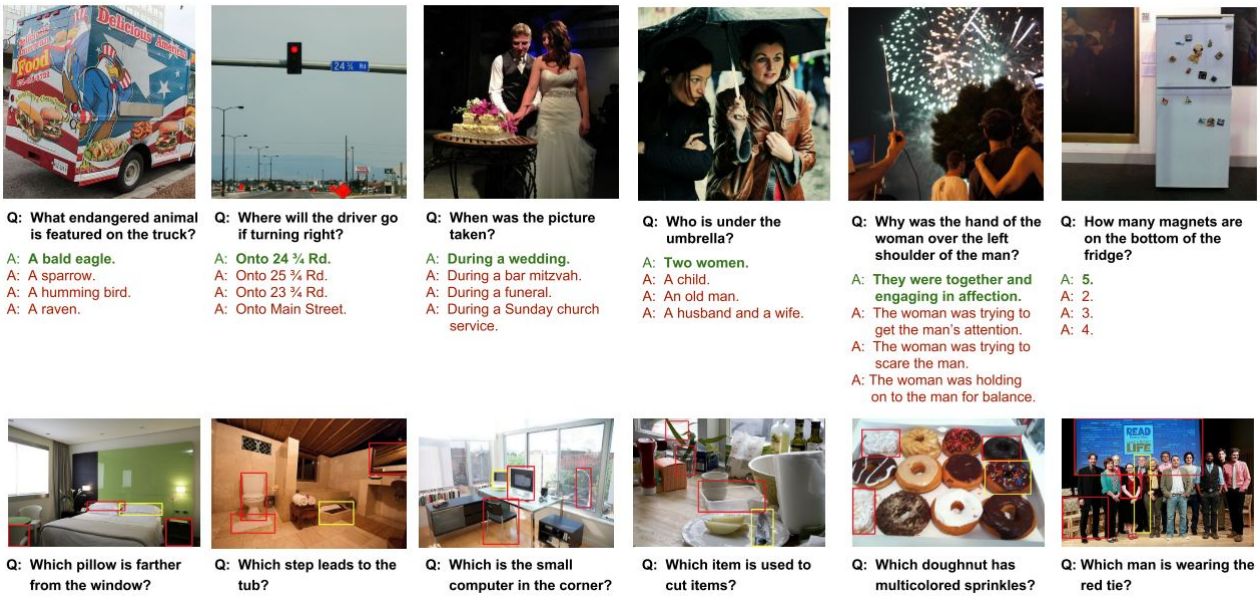
شکل ۳ـ۵: یک نمونه از مجموعه‌داده Visual Madlibs [۶۸]

۳ـ۱ـ۵ مجموعه داده CLEVR[۲۳]

CLEVR یک مجموعه‌داده برای ارزیابی درک بصری سیستم‌های VQA است. تصاویر این مجموعه‌داده با استفاده از سه شی استوانه، کره و مکعب تولیدشده‌است. برای هر کدام از این اشیا دو اندازه متفاوت، دو جنس متفاوت و هشت رنگ مختف در نظر گرفته شده است. سوالات هم به طور مصنوعی بر اساس مکانی که اشیا در تصویر قرار گرفته اند؛ ایجاد شده‌است. سوالات در CLEVR به گونه‌ای طراحی‌شده‌است که جنبه‌های مختلف استدلال بصری توسط سیستم‌های VQA را مورد ارزیابی قرار می‌دهد از جمله شناسایی ویژگی، شمارش اشیا، مقایسه، روابط مکانی اشیا و عملیات منطقی. در این مجموعه‌داده مکان تصاویر نیز با استفاده از یک مستطیل مشخص‌شده‌است.

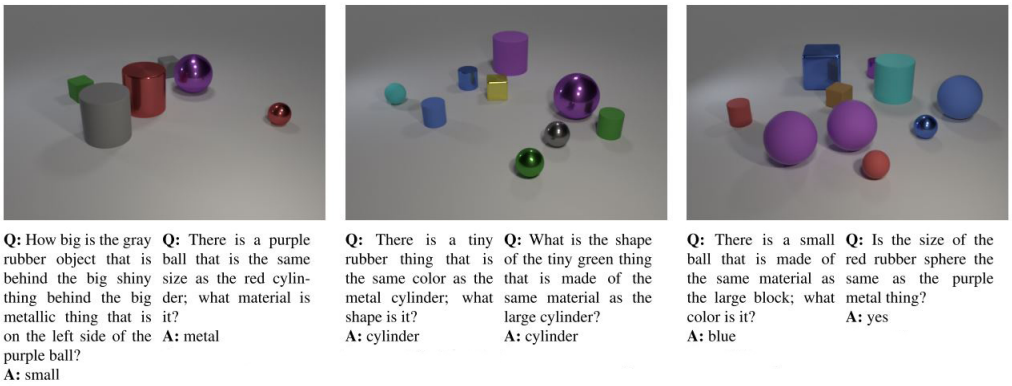
۳ـ۱ـ۶ مجموعه داده Tally-QA[۱]

در سال ۲۰۱۹، مجموعه‌داده Tally-QA منتشر شد که بزرگ‌ترین مجموعه‌داده پرسش و پاسخ تصویری برای شمارش اشیا است. اکثر مجموعه‌داده‌های شمارش اشیا در پرسش و پاسخ تصویری دارای سوالات ساده

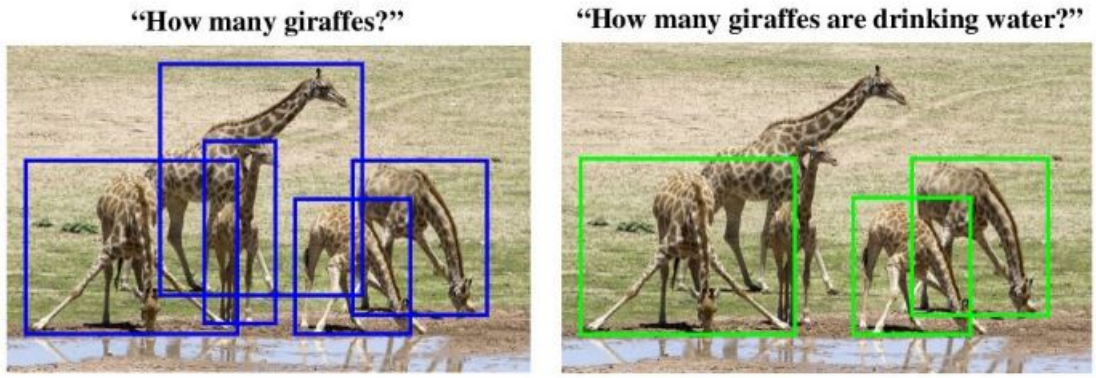


شکل ۳ـ۶: چند نمونه از مجموعه‌داده Visual7W [۷۱]. ردیف اول،پاسخ‌های سبز رنگ، پاسخ صحیح هستند و پاسخ‌های قرمز پاسخ‌های نادرست تولید شده توسط انسان است. ردیف دوم، کادر زرد جواب صحیح است و کادرهای قرمز پاسخ‌های اشتباه انسانی است.

هستند که برای پاسخ‌دادن به این سوال‌ها تنها کافی است که اشیا در تصویر تشخیص‌داده‌شوند. بنابراین، این موضوع باعث ایجاد مجموعه‌داده‌ی Tally-QA شد که علاوه بر سوالات ساده، سوالات پیچیده را نیز در بر می‌گیرد که برای پاسخ دادن به آن‌ها به استدلال بیشتری از تشخیص اشیا نیاز است. تعداد سوالات ساده در Tally-QA برابر با ۲۱۱۴۳۰ و تعداد سوالات پیچیده برابر با ۷۶۴۷۷ است. سوالات ساده این مجموعه‌داده از مجموعه‌داده‌های دیگری(VQA v2 [۱۷] و Visual Genome [۲۷]) برداشته‌شده‌است و سوالات پیچیده با استفاده از ۸۰۰ کاربر انسانی از طریق پلتفرم آنلاین Amazon Mechanical Turk جمع‌آوری شده‌است. مجموعه‌داده Tally-QA به سه بخش آموزش و تست_ساده و تست_پیچیده تقسیم می‌شود. بخش تست_ساده تنها شامل سوالات ساده و بخش تست_پیچیده تنها دارای سوالات پیچیده‌ای است که از Amazon Mechanical Turk جمع‌آوری شده‌است.



شکل ۳ـ۷: چند نمونه از مجموعه‌دادهCLEVR [۲۳].

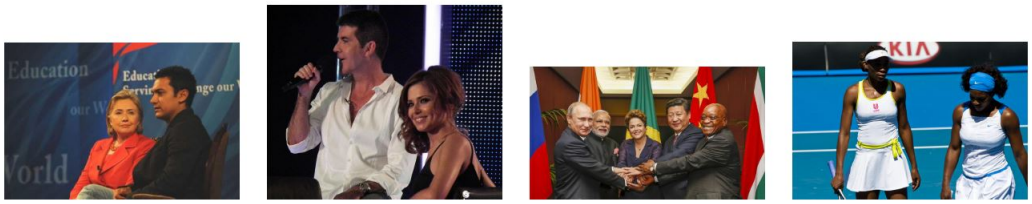


شکل ۳ـ۸: چند نمونه از مجموعه‌دادهTally-QA [۱].عکس سمت چپ یک نمونه از سوالات ساده و عکس سمت راست یک نمونه از سوالات پیچیده است.

۳ـ۱ـ۷ مجموعه داده KVQA[۵۰]

مجموعه داده KVQA که مخفف Knowledge-based Visual Question Answering است در سال ۲۰۱۹ طراحی شده است به طوری که بر خلاف مجموعه‌داده‌های قبلی، برای پیدا کردن پاسخ سوالات نیاز به دانش خارجی دارد. بدین منظور این مجموعه داده شامل ۱۸۳ هزار پرسش و پاسخ در مورد ۱۸ هزار شخص معروف شامل ورزشکاران، سیاستمداران و هنرمندان است. اطلاعات و تصاویر مرتبط با این اشخاص از Wikidata و Wikipedia استخراج شده است. KVQA شامل ۲۴ هزار تصویر است. این مجموعه‌داده به صورت تصادفی به سه بخش آموزش، ارزیابی و آزمون به ترتیب با نسبت های 0.7 ، 0.2 و 0.1 تقسیم شده است. تنوع پرسش

و پاسخ ها در KVQA به گونه‌ای در نظر گرفته شده است که مشکل همیشگی بایاس در مجموعه‌داده‌های پرسش و پاسخ تصویری، در این مجموعه داده وجود نداشته باشد.



(a) ***Wikipedia caption***: Khan with United States Secretary of State Hillary Clinton in 2009.

Q: Who is to the left of Hillary Clinton? (*spatial*)
A: **Aamir Khan**

Q: Do all the people in the image have a common occupation? (*multi-entity, intersection, 1-hop, Boolean*)
A: **No**

(b) ***Wikipedia caption***: Cheryl alongside Simon Cowell on The X Factor, London, June 2010.

Q: What is the age gap between the two people in the image? (*multi-entity, subtraction, 1-hop*)
A: **24 years**

Q: How many people in this image were born in United Kingdom? (*1-hop, multi-entity, counting*)
A: **2**

(c) ***Wikipedia caption***: BRICS leaders at the G-20 summit in Brisbane, Australia, 15 November 2014

Q: Were all the people in the image born in the same country? (*Boolean, multi-entity, intersection*)
A: **No**

Q: Who is the founder of the political party to which person second from left belongs to? (*spatial, multi-hop*)
A: **Syama Prasad Mookerjee**

(d) ***Wikipedia caption***: Serena Williams and Venus Williams, Australian Open 2009.

Q: Who among the people in the image is the eldest? (multi-entity, comparison)
A: **Person in the left**

Q: Who among the people in the image were born after the end of World War II? (multi-entity, multi-relation, comparison)
A: **Both**

شکل ۳ـ۹: چند نمونه از مجموعه‌دادهKVQA [۵۰]

.

۳ـ۲ تقویت مجموعه‌داده در مسئله پرسش و پاسخ تصویری

به لطف توسعه سریع شبکه‌های عصبی عمیق مسئله پرسش و پاسخ تصویری به موفقیت‌های بزرگی دست یافته است. مطالعات نشان می‌دهد که عملکرد شبکه‌های عصبی عمیق به میزان داده‌های آموزشی بستگی دارد و همیشه از داده‌های آموزشی بیشتر سود می‌برند. یکی از ترفندهای اصلی در شبکه‌های عصبی عمیق تقویت داده [۲] است که به طور گسترده در بسیاری از مسائل پردازش تصویر و بینایی ماشین مورد استفاده قرار می‌گیرد. اما مقالات کمی وجود دارد که مسئله تقویت داده را در پرسش و پاسخ تصویری بررسی کرده‌اند. یکی از چالش‌های تقویت داده در مسئله پرسش و پاسخ تصویری این است که هیچ یک از روش‌های تقویت داده مبتنی بر تصویر مانند چرخش [۳] و ورق زدن [۴] نمی‌توانند مستقیماً بر روی مسئله پرسش و پاسخ تصویری

[۲] data augmentation
[۳] rotation
[۴] flipping

اعمال شود زیرا ساختار معنایی آن حفظ نخواهد شد. به عنوان مثال با چرخش یک تصویر ممکن است پرسش و پاسخ مرتبط با آن(مانند «ماشین در سمت چپ یا راست سطل زباله است؟») دیگر درست نباشد.

در [۲۶] برای اولین بار دو روش برای تقویت داده در مسئله پرسش و پاسخ تصویری پیشنهاد شد. در روش اول برای تولید پرسش و پاسخ از الگو استفاده می‌شود. برای تولید الگو از حاشیه‌نویسی [۵] موجود در مجموعه‌داده استفاده می‌شود. با استفاده از این روش ۴ نوع سوال تولید می‌شود: (۱) سوالات بله و خیر (۲) سوالات شمارشی (۳) تشخیص شی، صحنه و یا فعالیت (۴) تشخیص ورزش. برای مثال برای تولید سوالات بله و خیر، با استفاده از حاشیه‌نویسی موجود در مجموعه داده لیستی از اشیا موجود در تصویر آماده می‌شود. سپس اگر محدوده مربوط به اشیا بزرگتر از ۲۰۰۰ پیکسل باشد، سوالی مانند « آیا [شی] در تصویر وجود دارد؟» تولید می‌شود که پاسخ آن هم «بله» است. به همین ترتیب با استفاده از دانشی که از مجموعه‌داده می‌توان بدست آورد؛ برای سایر انواع سوالات الگویی برای تولید سوال و پاسخ آن تولید می‌شود. یکی از مشکلات این روش برای تقویت داده این است که سوالات تولید شده انعطاف‌پذیر نیستند و ممکن است شباهت زیادی به نحوه‌ی طرح سوالات موجود در مجموعه‌داده نداشته باشند. به همین علت، روش دیگری در [۲۶] مبتنی بر LSTM برای تولید سوال برای هر تصویر پیشنهاد شده است. این شبکه از دو لایه LSTM تشکیل شده است که هر کدام دارای ۱۰۰۰ واحد مخفی است و پس از آن‌ها نیز دو لایه‌ی کاملا متصل که هر کدام ۷۰۰۰ نورون مخفی دارند(برابر با تعداد واژگان) ساخته شده است. برای تولید سوال، در ابتدا توکن شروع سوال به همراه ویژگی‌های تصویر به شبکه داده می‌شود. برای هر تصویر ۳۰ سوال تولید می‌شود که تنها سه تا از پرتکرارترین سوالات نگه داشته می‌شود. برای پیدا کردن جواب سوال‌های تولیده شده توسط شبکه LSTM از یک شبکه‌ی ساده MLP که در [۲۴] پیشنهاد شده است؛ استفاده شده است. در [۲۶] نشان دادند که استفاده از این دو روش برای تقویت داده‌ها منجر به بهبود عملکرد روش‌های موجود برای حل مسئله پرسش و پاسخ تصویری می‌شود.

اخیرا در [۶۰] برای تقویت داده روشی مبتنی بر تولید نمونه‌های خصمانه [۶] پیشنهاد شده است که بر خلاف کارهای قبلی، تقویت داده هم برای تصاویر و هم برای سوالات انجام می‌شود.

[۵] annotation

[۶] adversarial examples

۳ـ۳ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

بسیاری از محققان راه‌حل‌ها یا الگوریتم‌هایی را برای حل مسئله پرسش و پاسخ تصویری پیشنهادکرده‌اند که به طور کلی می‌توان آن را به یک فرآیند سه فازی تقسیم‌بندی کرد. فاز اول این فرآیند استخراج ویژگی از تصویر و سوالات است که راه‌حل‌های موفق در این فاز ریشه در روزهای باشکوه یادگیری عمیق دارد زیرا بیشتر راه‌حل‌های موفق در این حوزه از مدل‌های یادگیری عمیق استفاده می‌کنند مانند CNN ها برای استخراج ویژگی از تصویر و RNN ها و انواع آن(LSTM و GRU) برای استخراج ویژگی از سوالات. در فاز دوم که مهم‌ترین و اصلی‌ترین فاز می‌باشد، ویژگی‌های استخراج شده از تصویر و سوال باهم ترکیب می‌شوند. سپس از ترکیب ویژگی‌ها برای تولید پاسخ نهایی در فاز سوم استفاده می‌شود.

۳ـ۳ـ۱ فاز ۱ : استخراج ویژگی از تصویر و سوا ل

استخراج ویژگی از تصویر و سوال مرحله‌ی مقدماتی در پرسش و پاسخ تصویری است. ویژگی تصویر، تصویر را به عنوان یک بردار عددی توصیف می‌کند تا بتوان به راحتی عملیات‌های مختلف ریاضی را بر روی آن اعمال کرد. روش‌های زیادی وجود دارد که به صورت مستقیم از تصویر ویژگی استخراج می‌کنند مانند بردار ساده RGB ، SIFT ، تبدیل HAAR و HOG. اما با ظهور شبکه‌های یادگیری عمیق، نیاز به استخراج ویژگی به صورت مستقیم از بین رفت زیرا این شبکه‌ها قادر به یادگیری ویژگی هستند. آموزش مدل‌های یادگیری عمیق به منابع محاسباتی گران قمیت و مجموعه‌داده‌های بزرگ نیاز دارد. از این رو، استفاده از مدل‌های شبکه عصبی عمیق از قبل آموزش دیده، استخراج ویژگی از تصاویر را به راحتی امکان‌پذیر می‌کنند.

یکی از بهترین شبکه‌های عصبی برای استخراج ویژگی از تصویر، شبکه‌های عصبی کانولوشنی هستند.

در جدول ۲ـ۱ چند نمونه از برجسته‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه‌داده ImageNet [۱۱] آموزش داده شده‌اند؛ آورده شده است. بیشتر مدل‌های ارائه‌شده در پرسش و پاسخ تصویری از این شبکه‌های عصبی کانولوشنی استفاده می‌کنند تا محتوای تصویری خود را به بردارهایی عددی تبدیل کنند. جدول ۳ـ۳ لیستی از مدل‌های استفاده شده برای حل مسئله پرسش و پاسخ تصویری را نشان می‌دهد و مشخص می‌کند که هر کدام از این مدل‌ها برای استخراج ویژگی از تصویر از کدام یک از شبکه‌های عصبی کانولوشنی موجود در جدول ۲ـ۱ بهره می‌برد. همان‌طور که واضح است VGGNet و ResNet به طور گسترده‌ای در سیستم‌های پرسش و پاسخ تصویری مورد استفاده قرار گرفته‌اند. یکی از دلایلی که محققان

جدول ۳ـ۳: شبکه‌های عصبی کانولوشنی استفاده شده در مدل‌های پرسش و پاسخ تصویری.

مدل پرسش و پاسخ تصویری	AlexNet	VGGNet	GoogleNet	ResNet
Image_QA [۴۷]		✓		
Talk_to_Machine [۱۵]			✓	
VQA [۴]		✓		
Vis_Madlibs [۶۸]	✓			
VIS + LSTM [۴۶]		✓		
Ahab [۶۴]		✓		
ABC-CNN [۸]		✓		
Comp_QA [۳]		✓		
DPPNet [۴۱]		✓		
Answer_CNN [۳۵]		✓		
VQA-Caption [۳۲]		✓		
Re_Baseline [۲۲]				✓
MCB [۱۴]				✓
SMem-VQA [۶۷]			✓	
Region_VQA [۵۲]		✓		
Vis7W [۷۱]		✓		
Ask_Neuron [۳۸]	✓	✓	✓	✓
SCMC [۷]				✓
HAN [۳۶]				✓
StrSem [۶۹]		✓		
AVQAN [۴۹]				✓
CMF [۲۸]				✓
EnsAtt [۳۳]				✓
MetaVQA [۶۱]				✓
DA-NTN [۵]				✓
QGHC [۷]				✓
QTA [۵۱]				✓
WRAN [۴۳]				✓
QAR [۶۳]				✓

جدول ۳ـ۴: word embedding های استفاده شده در مدل‌های پرسش و پاسخ تصویری.

مدل پرسش و پاسخ تصویری	one-hot	CBOW	Skip-gram/Word2vec	GloVe	CNN	LSTM	GRU
Image_QA [۴۷]			✓				
Talk_to_Machine [۱۵]						✓	
VQA [۴]		✓					
Vis_Madlibs [۶۸]			✓				
VIS + LSTM [۴۶]						✓	
ABC-CNN [۸]						✓	
Comp_QA [۳]						✓	
DPPNet [۴۱]							✓
Answer_CNN [۳۵]					✓		
VQA-Caption [۳۲]						✓	
Re_Baseline [۲۲]			✓				
MCB [۱۴]						✓	
SMem-VQA [۶۷]		✓					
Region_VQA [۵۲]			✓				
Vis7W [۷۱]	✓						
Ask_Neuron [۳۸]		✓			✓	✓	✓
SCMC [۷]					✓		
HAN [۳۶]						✓	
StrSem [۶۹]						✓	
AVQAN [۴۹]	✓						
CMF [۲۸]				✓		✓	
EnsAtt [۳۳]				✓			
MetaVQA [۶۱]				✓			✓
DA-NTN [۵]							✓
QGHC [۷]							✓
WRAN [۴۳]							✓
QAR [۶۳]				✓			

VGGNet را ترجیح می‌دهند این است که ویژگی‌هایی را استخراج می‌کند که عمومیت بیشتری دارد و برای مجموعه‌داده‌هایی غیر از ImageNet که این مدل‌ها بر روی آن‌ها آموزش داده می‌شوند، موثرتر هستند. دلایل دیگر شامل همگرایی سریع در fine-tuning و پیاده‌سازی ساده در مقایسه با GoogLeNet و ResNet است. نکته‌ی قابل توجه دیگر در جدول ۳ـ۳ روند مهاجرت از VGGNet به ResNet در مقالات اخیر است. زیرا در سال‌های اخیر، منابع محاسباتی کافی با هزینه مناسب در دسترس محققان می‌باشد.

مدل‌های مختلف در مسئله پرسش و پاسخ تصویری از تعبیه کلمات ذکر شده در ؟؟ برای تولید بردار ویژگی برای سوال‌ها استفاده کرده‌اند. جدول ۳ـ۴ لیستی از مدل‌های پرسش و پاسخ تصویری به همراه word

embedding استفاده شده در آن‌ها را نمایش می‌دهد. با بررسی جدول ۳ـ۴ مشاهده می‌کنیم که محققان حوزه‌ی پرسش و پاسخ تصویری ترجیح می‌دهند؛ برای استخراج ویژگی از متن و بازنمایی آن از LSTM استفاده کنند. آن‌ها معتقد هستند که RNN ها عملکرد بهتری نسبت به روش‌های مستقل از دنباله‌ی کلمات مانند word2vec دارند. اما آموزش RNN ها نیاز به داده‌های برچسب خورده‌ی زیادی دارد.

۳ـ۳ـ۲ فاز ۲ : بازنمایی مشترک تصویر و سوا ل

در گام اول پرسش و پاسخ تصویری، تصویر و سوال به طور مستقل پردازش می‌شوند تا از آن‌ها ویژگی استخراج شود. روش‌های مختلف برای انجام این کار، در بخش ۳ـ۳ـ۱ به تفصیل بررسی شد. در گام بعدی، این ویژگی‌ها باید به یک فضای مشترک ترسیم شوند و یا به عبارتی ترکیب شوند تا آماده گام آخر(تولید پاسخ) شوند. در ادامه این بخش، به مرور روش‌های ترکیب ویژگی‌های استخراج شده از سوال و تصویر می‌پردازیم.

۳ـ۳ـ۲ـ۱ روش‌های پایه

ساده‌ترین و پایه‌ای‌ترین روش‌ها برای ترکیب ویژگی‌ها concatination ، جمع متناظر ویژگی‌ها [۷] و ضرب متناظر ویژگی‌ها [۸] است. مالینوفسکی در [۳۸] این سه روش را امتحان کرده است و دریافت کرد که ضرب متناظر ویژگی‌ها منجر به دقت بالاتری می‌شود. یافته مهم دیگر مالینوفسکی این است که نرمال‌سازی L2 ویژگی‌های تصویر، تأثیر قابل توجهی دارد به خصوص در روش‌های concatination و جمع متناظر ویژگی‌ها. با توجه به نتایج آن‌ها، جمع متناظر ویژگی‌ها پس از نرمال‌سازی از دقت بالاتری برخوردار است. در [۵۲] از ضرب نقطه‌ای(داخلی) بین ویژگی‌های استخراج شده از تصویر در سطح region و word embedding های حاصل از سوال استفاده شده است.

روش کلاسیک دیگر برای یافتن رابطه بین دو بردار که ریشه آن در علم آمار است، روش CCA [۹] است که برای ترکیب ویژگی‌های تصویر و سوال در VQA استفاده شده است. CCA بازنمایی مشترک بین بردار تصویر و بردار سوال را پیدا می‌کند. CCA یک نسخه نرمالیزه شده به نام nCCA [۱۰] نیز دارد که توسط [۱۶] پیشنهاد شده است. در [۶۸] و [۶۲] از هر دو مدل CCA و nCCA برای ترکیب بردارهای ویژگی سوال و تصویر

[۷] element-wise addition
[۸] element-wise multiplication
[۹] Analysis Correlation Canonical
[۱۰] Analysis Correlation Canonical normalized

استفاده کردند و دریافتند که روش nCCA به ویژه در مورد سوالات چندگزینه‌ای عملکرد بهتری دارد.

۳ـ۳ـ۲ـ۲ روش‌های مبتنی بر شبکه‌های عصبی

در اینجا، محققان شبکه‌های عصبی عمیق end-to-end را با لایه‌های خاص برای ترکیب ویژگی‌های تصویر و سوال آموزش می‌دهند. ساختار و عملکرد این لایه ممکن است برای مدل‌های مختلف پیشنهادشده متفاوت باشد.

ادامه اش باید تکمیل بشه ...

۳ـ۳ـ۲ـ۳ روش‌های مبتنی بر توجه

۳ـ۳ـ۳ فاز ۳ : تولید جواب

باید تکمیل شود.

۳ـ۴ مدل‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر [۱۱]

در سال‌های اخیر شاهد ظهور شبکه‌های از قبل آموزش‌دیده تنها بر روی داده‌های تصویری مثل ResNet[۱۹] و یا تنها بر روی داده‌های متنی مانند BERT[۱۳] ، GPT-2[۴۵] و GPT-3[۶] بوده‌ایم. استفاده از این شبکه‌ها منجر به بهبود مسائل موجود در بینایی ماشین و پردازش زبان‌های طبیعی شده است. با الهام از این موضوع، شبکه‌های از قبل آموزش‌دیده بر روی داده‌های تصویری و متنی نیز ایجاد شدند که هدف آن‌ها بازنمایی مشترک داده‌های تصویری و داده‌های زبانی است . بنابراین می‌توان از این شبکه‌ها برای بهبود عملکرد مسائل مشترک بین بینایی ماشین و پردازش زبان‌های طبیعی مانند پرسش و پاسخ تصویری نیز استفاده کرد. معماری شبکه‌های از قبل آموزش‌دیده برروی زبان طبیعی و تصویر به طور کلی به دو دسته تک جریان [۱۲] و دو جریان [۱۳] تقسیم می‌شود. در ادامه به بحث و بررسی هر یک از این دسته‌ها می‌پردازیم.

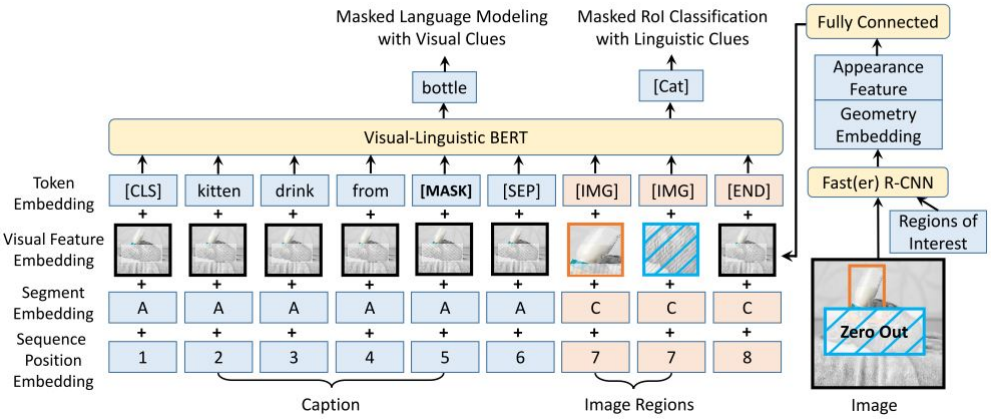
[۱۱] vision-and-language pretraining models
[۱۲] single-stream
[۱۳] two-stream

۳ـ۴ـ۱ معماری تک جریان

پایه و اساس این معماری شبیه معماری مدل BERT[۱۳] است که رمزگذاری متن [۱۴] و رمزگذاری تصویر [۱۵] را به طور همزمان انجام می‌دهد. در واقع برای یادگیری بازنمایی متن و تصویر از یک رمزگذار [۱۶] استفاده می‌کند. بنابراین ورودی مدل‌های پیشنهادشده در این معماری داده‌های چندحالته [۱۷] هستند که به صورت همزمان و یکجا به مدل داده می‌شوند برای مثال تصویر به همراه یک جمله توصیف کننده آن و یا یک فیلم به همراه زیرنویسش به این شبکه‌ها برای آموزش داده می‌شوند. به علاوه این مدل‌ها با ترکیبی از اهداف مختلف مانند visual-based Masked Language Model ، text-based Masked Language Model ، masked visual-feature modeling و visual-linguistic matching بهینه می‌شود. سپس از بازنمایی‌های آموخته‌شده توسط این مدل‌ها در مسائل پایین‌دستی understanding و یا generation استفاده می‌شود. به عنوان مثال، مدل VideoBERT[۵۶] برای مسائل generation مانند تولید توصیف فیلم طراحی شده است. در حالی که چندین مدل دیگر مانند B2T2[۲] ، Unicoder-VL[۲۹] ، VL-BERT[۱۸][۵۵] و UNITER[۹] وجود دارد که همگی برای مسائل understanding طراحی شده‌اند. مدل‌های دیگری مانند VLP[۷۰] و OSCAR[۳۰] مدل‌های یکپارچه‌ای هستند که هم در مسائل پایین‌دستی understanding و هم در مسائل generative کاربرد دارد. از بین این مدل‌ها، تنها از مدل‌های VL-BERT ، UNITER ، VLP و OSCAR می‌توان برای مسئله پرسش و پاسخ تصویری استفاده کرد. بنابراین در ادامه این بخش جزئیات هر کدام از این مدل‌ها را توضیح خواهیم داد.

شکل ۳ـ۱۰ معماری VL-BERT را نشان می‌دهد. مشابه BERT ، از کدگذارهای multi-layer bidirectional transformer استفاده شده است. اما برخلاف BERT که ورودی آن تنها کلمات جمله هستند، این شبکه به همراه کلمات یک جمله، مناطق مورد علاقه [۱۹] استخراج شده از تصویر و یا به اختصار ROI را نیز به عنوان ورودی می‌گیرد. برای استخراج ROI از تصویر از شبکه Faster RCNN[۴۸] استفاده شده است. هر ورودی این شبکه با توکن [CLS] آغاز می‌شود. سپس با کلمات جمله و ROI های تصویر ادامه می‌یابد و با توکن [END] خاتمه می‌یابد. از توکن [SEP] نیز برای جدا کردن جملات و یا جملات و تصویر از هم استفاده می‌شود. برای هر ورودی، تعبیه ویژگی [۲۰] آن جمع چهار نوع تعبیه است که در شکل ۳ـ۱۰ مشخص شده

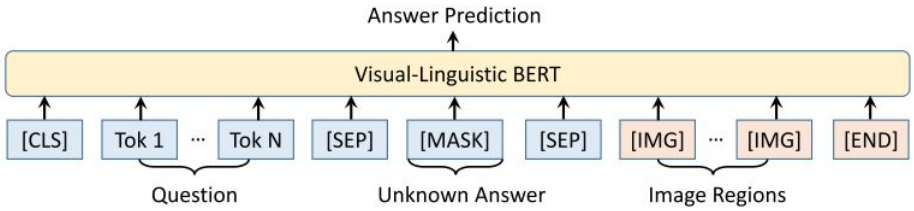
[۱۴] text encoding
[۱۵] image encoding
[۱۶] encoder
[۱۷] multimodal
[۱۸] Visual-Linguistic BERT
[۱۹] regions-of-interest
[۲۰] feature embedding



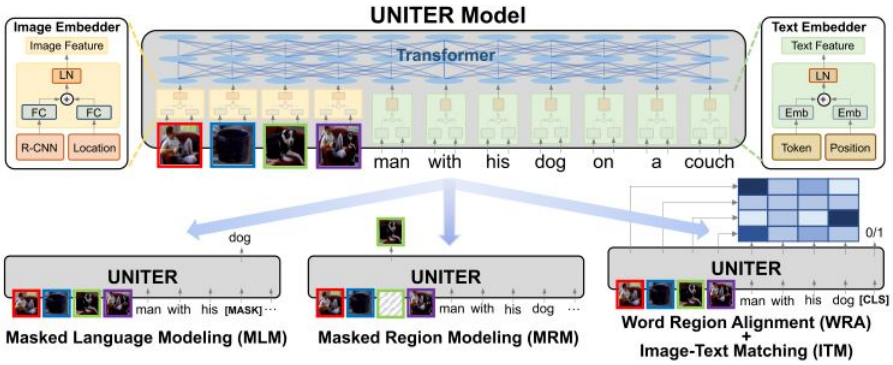
شکل ۳ـ۱۰: معماری شبکه از قبل آموزش‌دیده VL-BERT[۵۵]

است. در میان آن‌ها، تعبیه مربوط به ویژگی‌های تصویری [۲۱] به تازگی به شبکه اضافه شده است در حالی که سه تعبیه دیگر از قبل در مدل BERT وجود داشته است. برای آموزش VL-BERT از مجموعه داده Concep-tual Captions به عنوان مجموعه‌داده زبانی‌ـ تصویری استفاده شده است. علاوه بر این از دو مجموعه‌داده فقط زبانی به نام‌های BooksCorpus و English Wikipedia به منظور بهبود تعمیم‌دهی شبکه استفاده شده است. برای بهینه‌سازی شبکه VL-BERT از دو تابع هدف استفاده شده است: text-based Masked Lan-guage Model و visual-based Masked Language Model. در text-based Masked Language Model با احتمال ۱۵ درصد یکی از کلمات ورودی با توکن [MASK] جایگزین می‌شود. بنابراین شبکه باید سعی کند که این کلمه ماسک شده را با توجه به کلمات دیگر و ویژگی‌های تصویری در خروجی پیش‌بینی نماید. در visual-based Masked Language Model با احتمال ۱۵ درصد یکی از ROI ها ماسک می‌شود و شبکه باید سعی کند در خروجی برچسب گروه مربوط به آن ROI را باتوجه به کلمات و سایر ROI ها پیش‌بینی کند. دقت شود که همانطور که در قسمت سمت راست تصویر ۳ـ۱۰ مشخص است، ملاک برچسب گروه‌بندی درست برای ROI ها، خروجی شبکه Faster RCNN است. برای استفاده از شبکه از قبل آموزش‌دیده VL-BERT برای مسئله پرسش و پاسخ تصویری، مطابق شکل ۳ـ۱۱ سه تایی کلمات سوال، پاسخ و ROI های استخراج‌شده از تصویر توسط Faster RCNN در ورودی داده می‌شود که به جای پاسخ، [MASK] قرار گرفته که شبکه تلاش می‌کند؛ پاسخ را در خروجی پیش‌بینی کند.

[۲۱] visual feature embedding



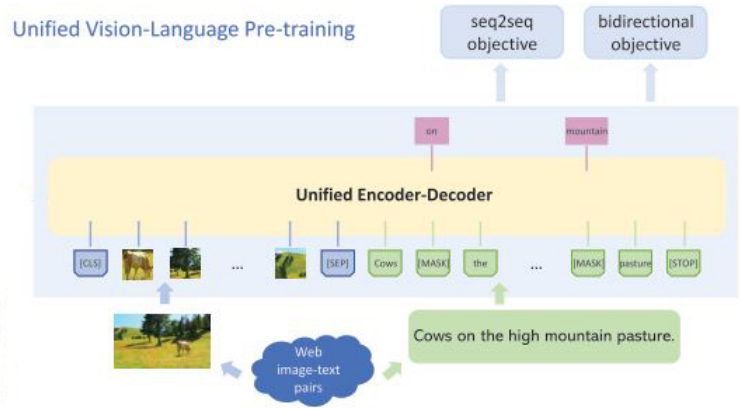
شکل ۳ـ۱۱: نحوه ورودی و خروجی شبکهVL-BERT برای آموزش در مسئله پرسش و پاسخ تصویری[۵۵]



شکل ۳ـ۱۲: معماری شبکه از قبل آموزش‌دیده UNITER[۹]

معماری مدل UNITER در شکل ۳ـ۱۲ نشان داده شده است. ورودی این مدل مانند VL-BERT ،کلمات یک جمله به همراه ROI های تصویر است. یکی از تفاوت‌های مدل UNITER با مدل VL-BERT این است که از ۴ مجموعه‌داده زبانی‌ـ تصویری برای آموزش استفاده کرده است:(۱) COCO ،(۲) Visual Genome (VG)، (۳) Conceptual Captions و(۴) SBU Captions. تفاوت دیگر این مدل با مدل VL-BERT در توابع هدف است که علاوه بر text-based Masked Language Model و visual-based Masked Language Model از دو تابع هدف دیگر به نام‌های Image-Text Matching و Word-Region Alignment نیز استفاده می‌کند. در Image-Text Matching هدف این است که مدل بتواند پیش‌بینی کند که آیا جمله و تصویر داده شده در ورودی با هم مطابقت دارند یا خیر. بدین منظور، یک جمله و ROI های تصویر به UNITER داده می‌شود و در خروجی بازنمایی مربوط به توکن [CLS] از یک تابع سیگموئید عبور داده می‌شود که یک مقدار بین صفر و یک را برمی‌گرداند که مقدار یک نشان می‌دهد که جمله و تصویر ورودی کاملا با هم مطابقت دارد و مقدار صفر به این معناست که جمله و تصویر ورودی با هم مطابقت ندارد. در UNITER علاوه بر در نظر گرفتن تطابق جمله و تصویر، از تطابق بین کلمات موجود در جمله و ROI های تصویر نیز برای آموزش استفاده می‌شود که این موضوع در قالب تابع هدف Word-Region Alignment در مدل مطرح شده است. زمان آموزش مدل UNITER به ازای هر دسته از داده‌های ورودی، یکی از ۴ تابع هدف نامبرده شده به صورت تصادفی انتخاب می‌شود و براساس آن تابع هدف، عملیات کاهش گرادیان برای شبکه انجام می‌شود. برای استفاده از شبکه از قبل آموزش‌دیده UNITER برای مسئله پرسش و پاسخ تصویری، بازنمایی حاصل از توکن [CLS] به یک شبکه MLP داده می‌شود و پاسخ را برای سوال و تصویر ورودی پیش‌بینی می‌کند. در واقع در این حالت، مسئله پرسش و پاسخ تصویری به عنوان یک مسئله طبقه‌بندی در نظر گرفته می‌شود.

شبکه از قبل آموزش‌دیده VLP نیز مانند دو شبکه‌ی قبلی از کلمات یک جمله و ROI های استخراج شده از تصویر به عنوان ورودی استفاده می‌کند. تفاوت اصلی این شبکه با دو شبکه VL-BERT و UNITER در این است که یک شبکه‌ی یکپارچه رمزگذار‌ـ رمزگشا است که نه تنها در مسائل understanding بلکه در مسائل generative به دلیل وجود رمزگشا قابل استفاده است. مدل VLP بر روی مجموعه‌داده Conceptual Captions آموزش داده شده است. دو تابع هدف در شبکه VLP استفاده شده است: (۱) bidirectional و seq2seq. در تابع هدف bidirectional یکی از کلمات موجود در جمله با توکن [MASK] جایگزین می‌شود و برای پیش‌بینی این کلمه ماسک شده در خروجی از تمامی کلمات و ROI های اطراف آن استفاده می‌شود. اما در تابع هدف seq2seq برای پیش‌بینی کلمه ماسک شده در خروجی، تنها از کلمات سمت چپ کلمه ماسک شده و ROI

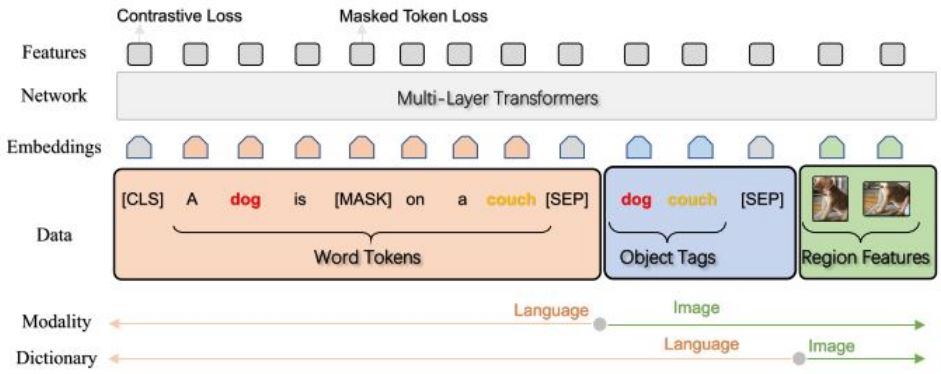


شکل ۳ـ۱۳: معماری شبکه از قبل آموزش‌دیدهVLP[۷۰]

های اطراف آن استفاده می‌کند. به عبارتی دیگر، برای پیش‌بینی کلمه ماسک شده نمی‌توان از کلماتی که بعد از آن و در آینده در جمله آمده است؛ استفاده کرد. معماری شبکه VLP در شکل ۳ـ۱۳ نشان داده شده است.

ورودی سه مدل قبلی یعنی VL-BERT ، UNITER و VLP یک جمله به همراه ROI های استخراج شده از تصویر بود. در مدل OSCAR علاوه بر این دو ورودی از ورودی دیگری به نام برچسب اشیا [۲۲] استفاده می‌شود که اشیایی که هم در تصویر وجود دارد و هم در جمله به آن اشاره شده است را نشان می‌دهد. در [۳۰] ادعا شده است که استفاده برچسب اشیا منجر به تولید بازنمایی بهتری از متن و تصویر می‌شود و در واقع از این برچشب‌ها به عنوان لنگر برای تطابق دادن فضای تصویر و متن استفاده می‌شود. در مدل OSCAR برای بدست آوردن ROI های تصویر و برچسب اشیا از شبکه Faster RCNN استفاده شده است. در مدل OSCAR به دو طریق می‌توان به ورودی‌ها نگاه کرد که در نتیجه دو تابع هدف برای آموزش این شبکه تعریف می شود. در روش اول، کلمات جمله و برچسب اشیا با هم در نظر گرفته می‌شود (دید Dictionary) و به احتمال ۱۵ در صد یکی از کلمات جمله و یا یکی از برچسب‌های اشیا با توکن [MASK] جایگزین می‌شود و مدل باید سعی کند این کلمه ماسک شده را در خروجی پیش‌بینی کند (Masked Token Loss). در روش دوم، ROI های تصویر و برچسب اشیا با هم در نظر گرفته می‌شود(دید Modality) و با احتمال ۵۰ درصد برچسب‌های اشیا با برچسب‌های دیگری تغییر می‌کند و مدل باید پیش‌بینی کند که آیا کلمات موجود در جمله با قسمت برچسب اشیا و ROI های تصویر مطابقت دارد یا نه. که بدین منظور خروجی شبکه برای توکن [CLS] به یک شبکه کاملا متصل داده می‌شود و یک طبقه‌بندی باینری انجام می‌شود که یک به معنای تطابق کلمات جمله با

[۲۲] object tag



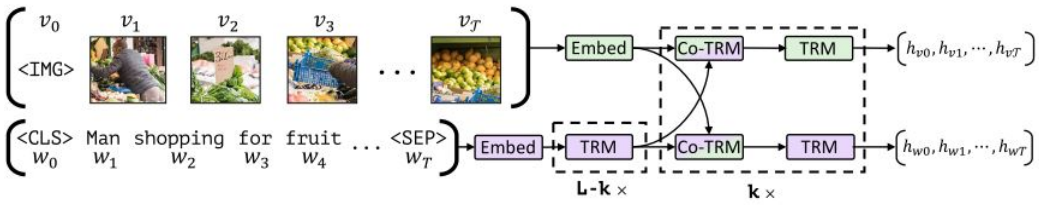
شکل ۳ـ۱۴: معماری شبکه از قبل آموزش‌دیدهOSCAR[۳۰]

ROI های تصویر و برچسب اشیاست و صفر نشان‌دهنده عدم تطابق است(Contrastive Loss). برای آموزش مدل OSCAR از مجموعه‌داده‌های COCO ، Conceptual Captions ، SBU captions ، flicker30 و GQA استفاده شده است. برای استفاده از شبکه از قبل آموزش‌دیده OSCAR برای مسئله پرسش و پاسخ تصویری، سوال به همراه برچسب اشیا و ROI های تصویر به ورودی شبکه داده می‌شود و خروجی توکن [CLS] به یک طبقه‌بند داده می‌شود تا پاسخ سوال و تصویر داده شده در تصویر بدست آید. در واقع در این روش، مسئله پرسش و پاسخ تصویری به صورت یک مسئله طبقه‌بندی در نظر گرفته می‌شود. معماری شبکه OSCAR در شکل ۳ـ۱۴ نمایش داده شده است.

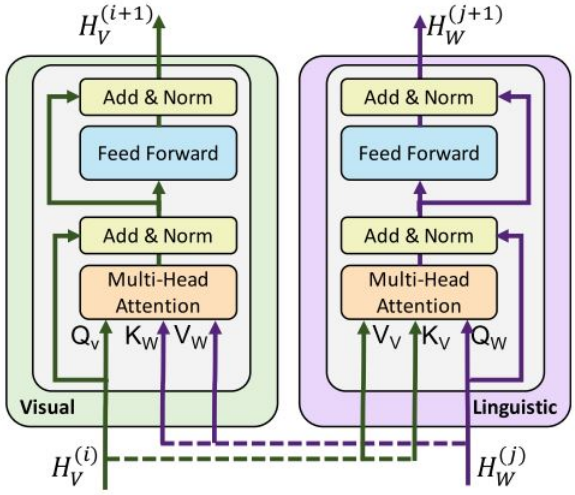
۳ـ۴ـ۲ معماری دو جریان

در مقابل معماری تک جریان، معماری دو جریان برای یادگیری هر کدام از بازنمایی‌های تصویر و متن از یک رمزگذار مستقل استفاده می‌کند. سپس از یک رمزگذار دیگر برای بدست آوردن بازنمایی مشترک متن و تصویر استفاده می‌کند.مشابه معماری تک جریان، معماری‌های دو جریان نیز مدل‌های خود را با visual-based Masked Language Model ، text-based Masked Language Model و visual-linguistic matching بهینه می‌کنند. ViLBERT[۳۴] و LXMERT[۵۹] نمونه‌هایی از معماری دو جریان هستند که از این دو مدل می‌توان برای مسئله پرسش و پاسخ تصویری استفاده کرد. پس در ادامه این بخش، جزئیات این دو شبکه را بررسی خواهیم کرد.

شکل ۳ـ۱۵ معماری شبکه ViLBERT را نمایش می‌دهد. مدل ViLBERT شامل دو مدل موازی به



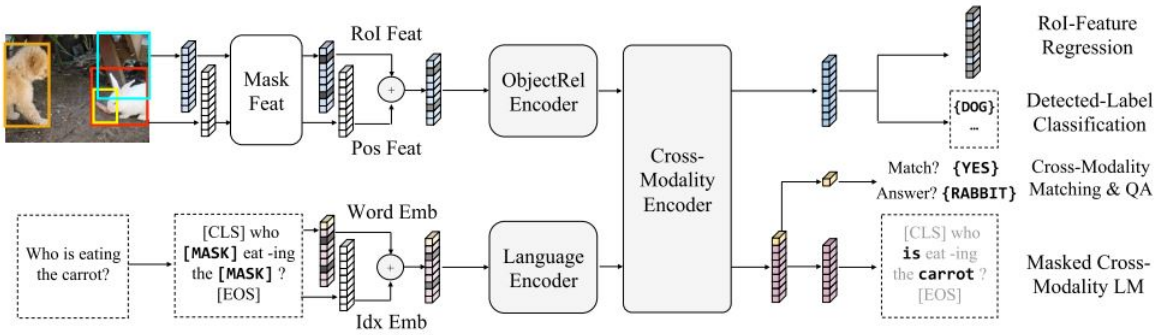
شکل ۳ـ۱۵: معماری شبکه از قبل آموزش‌دیده ViLBERT[۳۴]



شکل ۳ـ۱۶: ساختار لایه co-attentional transformer[۳۴]

سبک BERT است که به صورت جداگانه بر روی کلمات متن و ROI های تصویر اعمال می‌شود و از بلوک‌های ترنسفورمر در هر جریان استفاده شده است(در شکل ۳ـ۱۵ با TRM مشخص شده است.). سپس برای بدست آوردن بازنمایی مشترک بین متن و تصویر از لایه‌های co-attentional transformer استفاده شده است(در شکل ۳ـ۱۵ با Co-TRM مشخص شده است.). اساس لایه‌ی co-attentional transformer بر پایه‌ی ترنسفرمر است در واقع برای هر کدام از بخش‌های تصویری و متنی داده ورودی، یک ترنسفرمر در لایه co-attentional transformer در نظر گرفته شده است که پس از عبور متن و داده از جریان‌های مستقل خود و بدست آمدن query ، key و value برای هر کدام، key و value متن به ترنسفرمر تصویر در co-attentional transformer داده می‌شود و به صورت متقابل key و value تصویر به ترنسفرمر متن داده می‌شود.

شکل ۳ـ۱۶ ساختار لایه co-attentional transformer را نشان می‌دهد. برای آموزش مدل ViLBERT از توابع هدف visual-based Masked Language Model ، text-based Masked Language Model و visual-



شکل ۳ـ۱۷: معماری شبکه از قبل آموزش‌دیده LXMERT[۵۹]

linguistic matching استفاده شده است. شبکه ViLBERT بر روی مجموعه‌داده Conceptual Captions آموزش داده شده است. برای استفاده از شبکه از قبل آموزش‌دیده ViLBERT برای مسئله پرسش و پاسخ تصویری، ابتدا خروجی بازنمایی توکن [CLS] و بازنمایی تصویر ضرب متناظر می‌شوند. سپس با عبور از یک شبکه MLP دولایه پاسخ مربوط به سوال و تصویر حاصل می‌شود.

شکل ۳ـ۱۷ معماری مدل LXMERT را نشان می‌دهد. ورودی این شبکه کلمات جمله ورودی و ROI های استخراج شده از تصویر است. همان‌طور که قبلا اشاره شد؛ مدل LXMERT یک مدل دو جریان است به همین دلیل برای پردازش متن و تصویر از دو رمزگذار مجزا و مستقل استفاده شده است(در شکل ۳ـ۱۷ به ترتیب با عنوان‌های ObjectRel Encoder و Language Encoder برای تصویر و متن مشخص شده است.) و سپس برای بدست آوردن بازنمایی مشترک از رمزگذار Cross-modality استفاده شده است. توابع هدف استفاده شده در مدل LXMERT مشابه شبکه ViLBERT است اما در LXMERT از تابع هدف دیگری به نام image question answering برای آموزش شبکه استفاده شده است. زیرا حدود ۱/۳ داده‌ای که برای آموزش این شبکه استفاده شده است؛ یک سوال در مورد تصویر ورودی است. بنابراین با تعریف تابع هدف image question answering مدل سعی می‌کند تا پاسخ این سوال را در خروجی پیش‌بینی کند. برای آموزش شبکه LXMERT از مجموعه‌داده‌های MS COCO ، Visual Genome ، VQA v2.0 ، GQA balanced version و VG-QA استفاده شده است.

در جدول ۳ـ۵ مقایسه چند نمونه از مدل‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر که مسئله پرسش و پاسخ تصویری را پشتیبانی می‌کنند؛ آورده شده است. ورودی تمام این مدل‌ها، کلمات جمله و ROI

جدول ۳ـ۵: مقایسه بین شبکه‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر

روش	معماری	ورودی	مجموعه‌داده‌های استفاده شده برای آموزش	توابع هدف
VL-BERT[۵۵]	تک جریان	کلمات جمله + ROIهای تصویر	Conceptual Captions + BooksCorpus + English Wikipedia	text-based MLM + visual-based MLM
UNITER[۹]	تک جریان	کلمات جمله + ROIهای تصویر	COCO + Visual Genome + Conceptual Captions + SBU Captions	text-based MLM + visual-based MLM + Image-Text Matching + Word-Region Alignment
VLP[۷۰]	تک جریان	کلمات جمله + ROIهای تصویر	Conceptual Captions	bidirectional + seq2seq
OSCAR[۳۰]	تک جریان	کلمات جمله + ROIهای تصویر+ برچسب اشیا	COCO + Conceptual Captions + SBU captions + flicker30 + GQA	Masked Token Loss + Contrastive Loss
ViL-BERT[۳۴]	دو جریان	کلمات جمله + ROIهای تصویر	Conceptual Captions	text-based MLM + visual-based MLM + Image-Text Matching
LXMERT[۵۹]	دو جریان	کلمات جمله + ROIهای تصویر	MS COCO + Visual Genome + VQA v2.0 + GQA balanced version + VG-QA	text-based MLM + visual-based MLM + Image-Text Matching + Image Question Answering

جدول ۳ـ۶: دقت شبکه‌های از قبل آموزش‌دیده بر روی مجموعه‌داده VQA v2.0 (test-std)

روش	سوالات بله/خیر	سوالات شمارشی	سایر سوالات	دقت کل
VLP[۷۰]	۸۷٫۴	۵۲٫۱	۶۰٫۵	۷۰٫۷
ViL-BERT[۳۴]	–	–	–	۷۰٫۹۲
VL-BERT[۵۵]	–	–	–	۷۲٫۲۲
LXMERT[۵۹]	۸۸٫۲	۵۴٫۲	۶۳٫۱	۷۲٫۵
OSCAR[۳۰]	–	–	–	۷۳٫۸۲
UNITER[۹]	–	–	–	۷۴٫۰۲

های تصویر است به جز مدل OSCAR که علاوه بر این دو، برچسب اشیا را نیز به عنوان ورودی دریافت می‌کند. شباهت دیگر این مدل‌ها در استفاده از مجموعه‌داده Conceptual Captions برای آموزش است البته به جز مدل LXMERT که از این مجموعه‌داده استفاده نکرده است. نکته‌ی حائز اهمیت دیگر در این جدول استفاده تقریباً تمامی مدل‌ها از دو تابع هدف text-based Masked Language Model و visual-based Masked Language Model است.

در جدول ۳ـ۶ نتایج مدل‌های VL-BERT ، UNITER ، VLP ، OSCAR ، ViL-BERT و LXMERT بر روی مجموعه‌داده VQA v2.0 نشان داده شده است. بهترین نتیجه بدست آمده برای مدل UNITER است. یکی از نکات قابل ملاحظه در این جدول این است که مدل‌های تک جریان نتایج بهتری نسبت به مدل‌های دو جریان بدست آوردند در حالی که تعداد پارامترهای مدل‌های تک جریان نسبت به مدل‌های دو جریان کمتر است.

۳ـ۵ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری

در این بخش می‌خواهیم به طور مختصر معیارهای ارزیابی شناخته شده در مسئله پرسش و پاسخ تصویری را بررسی‌کنیم. همان‌طور که قبلا ذکر شد؛ معمولا دو نوع سوال در مجموعه‌داده‌های پرسش و پاسخ تصویری در نظر گرفته می‌شود: سوالات open-ended و سوالات چندگزینه‌ای. در سوالات چندگزینه‌ای، برای هر سوال دقیقا یک پاسخ صحیح وجود دارد. بنابراین ارزیابی آن ساده است زیرا می‌توان به راحتی از معیار دقت استفاده کرد. اما در سوالات open-ended این امکان وجود دارد که چندین پاسخ صحیح برای هر سوال وجود داشته باشد. بنابراین ارزیابی در این حالت ساده نخواهد بود. برای حل این موضوع، اکثر مجموعه‌داده‌های پرسش

و پاسخ تصویری پاسخ‌ها را محدود به چند کلمه(۱ تا۳ کلمه) می‌کنند و یا پاسخ‌ها را از یک مجموعه بسته انتخاب می‌کنند.

در ادامه به بررسی مهم‌ترین معیارهای این حوزه می‌پردازیم. اما ارزیابی مسئله پرسش و پاسخ تصویری همچنان یک مسئله حل نشده است. هر کدام از روش‌ها و معیارهای ارزیابی موجود، مزیت‌ها و معایب خاص خود را دارند. بنابراین برای انتخاب معیار ارزیابی باید به مواردی همچون ساختار مجموعه‌داده و نحوه‌ی ساخت آن، میزان بایاس موجود در مجموعه‌داده و ... توجه نمود.

۳ـ۵ـ۱ معیار دقت

اگر چه در سوالات چندگزینه‌ای برای سنجش یک مدل معیار دقت کافی است اما در سوالات open-ended معیار دقت سخت‌گیرانه است زیرا فقط در حالتی که پاسخ مدل کاملا مطابق با پاسخ در نظر گرفته شده باشد، پذیرفته می‌شود. برای مثال اگر صورت سوال «چه حیواناتی در تصویر است؟» باشد و پاسخ مدل به جای «سگ‌ها» پاسخ «سگ» باشد؛ غلط تلقی می‌شود. بنابراین به دلیل این محدودیت‌هایی که معیار دقت دارد؛ معیارهای دیگری برای ارزیابی این نوع سوالات پیشنهاد شده‌است.

$$Accuracy = \frac{Number\ of\ questions\ answered\ correctly}{Total\ questions} \qquad (۱ـ۳)$$

۳ـ۵ـ۲ معیار شباهت Wu-Palmer [۶۶]

این معیار ارزیابی توسط مالینوفسکی [۳۷] برای پرسش و پاسخ تصویری ارائه شد. این معیار از تئوری مجموعه‌های فازی الهام گرفته شده است و نسبت به معیار دقت سخت‌گیری کمتری دارد. معیار شباهت Wu-Palmer سعی می‌کند که تفاوت بین پاسخ پیش‌بینی شده با پاسخ صحیح را از لحاظ معنایی اندازه‌گیری‌کند. یکی از معایب این معیار این است که به پاسخ‌هایی که از لحاظ لغوی شبیه هم هستند ولی از لحاظ معنایی متفاوت هستند، امتیاز بالایی می‌دهد. زمانی که پاسخ‌های ما به صورت عبارت یا جمله باشد؛ این معیار عملکرد خوبی ندارد.

۳ـ۵ـ۳ معیار اجماع

از این معیار زمانی استفاده می‌شود که هر سوال توسط کاربرهای انسانی متفاوتی پاسخ داده شود. در واقع برای هر سوال چندین پاسخ مستقل وجود داشته باشد. این معیار دو نوع دارد: میانگین اجماع و کمترین اجماع. در میانگین اجماع امتیاز نهایی برابر با میانگین وزندار پاسخ‌های وارد شده توسط کاربرهای متفاوت است و در کمترین اجماع پاسخ پیش‌بینی شده حداقل باید با یکی از پاسخ‌ها مطابقت داشته باشد. در مسئله‌ی پرسش و پاسخ تصویری معمولا از حالت کمترین اجماع استفاده می‌شود و آستانه را هم برابر ۳ قرار می‌دهند به این معنی که اگر پاسخ پیش‌بینی شده با ۳ و یا بیشتر از ۳ پاسخ برابر باشد امتیاز کامل می‌گیرد و در غیر این صورت هیچ امتیازی کسب نخواهد کرد. از معایب این روش می‌توان به هزینه زیاد جمع‌آوری پاسخ برای سوالات اشاره کرد. آنتول و همکارانش از این معیار ارزیابی در [۴] استفاده کرده‌اند.

$$Accuracy_{VQA} = min(\frac{n}{۳}, ۱) \qquad (۲ـ۳)$$

۳ـ۵ـ۴ MPT [۲۵]

یکی از مشکلات مجموعه‌داده‌های پرسش و پاسخ تصویری توزیع غیریکنواخت انواع سوال‌هاست. دراین مواقع، نمی‌توان از معیار دقت استفاده کرد. بنابراین در [۲۵] معیار جدیدی به نام MPT [۲۳] ارائه شده است که توزیع نامتوازن سوال‌ها را جبران می‌کند. معیار MPT میانگین دقت برای هر نوع سوال را محاسبه می‌کند. از نسخه‌ی نرمالایز شده‌ی این معیار نیز برای رفع مشکل بایاس در توزیع پاسخ‌ها استفاده می‌شود.

۳ـ۵ـ۵ BLEU [۴۲]

BLEU [۲۴] یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. در [۱۸] پیشنهاد داده شد که از این معیار نیز برای ارزیابی پرسش و پاسخ تصویری می‌توان استفاده کرد. معیار BLEU کنار هم قرار گرفتن n-gram های پاسخ پیش‌بینی شده و پاسخ صحیح را اندازه‌گیری می‌کند. معمولا BLEU زمانی که جمله‌ها کوتاه باشند، با شکست مواجه می‌شود.

[۲۳] Mean Per Type
[۲۴] BiLingual Evaluation Understudy

۳ـ۵ـ۶ METEOR [۱۲]

METEOR [۲۵] نیز همانند BLEU یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. به پیشنهاد [۱۸] از این معیار هم می‌توان برای پرسش و پاسخ تصویری نیز استفاده نمود. معیار METEOR سعی می‌کند که هم‌ترازی بین کلمات موجود در پاسخ پیش‌بینی شده و پاسخ صحیح را پیدا کند.

[۲۵] Metric for Evaluation of Translation with Explicit ORdering

فصل ۴

نتیجه‌گیری و کارهای آینده

۴ـ۱ نتیجه‌گیری

علی رغم این که از معرفی مسئله پرسش و پاسخ تصویری تنها چندین سال می‌گذرد، رشد آن در این چند سال قابل توجه بوده است. برای حل مسئله پرسش و پاسخ تصویری، رویکردهای یادگیری عمیق همچنان در مرکز توجه هستند. ما برجسته‌ترین مدل‌های یادگیری عمیق برای مسئله پرسش و پاسخ تصویری را بررسی کردیم. با معرفی شبکه‌های از قبل آموزش‌دیده بهبود چشمگیری در مسائل یادگیری عمیق رخ داد به طوری که بیشتر مسائل مختلف در یادگیری عمیق، بهترین نتیجه خود را با استفاده از شبکه‌های از قبل آموزش‌دیده بدست آورده‌اند. مسئله پرسش و پاسخ تصویری نیز از این قاعده مستثنی نیست و در حال حاضر شبکه‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر بهترین عملکرد را برای مجموعه‌دادگان پرسش و پاسخ تصویری رقم زده‌اند. چندین نمونه از این مدل‌ها را در بخش با جزئیات بحث کردیم. پیشرفت‌های زیادی که همچنان برای مجموعه‌دادگان مختلف در این حوزه اتفاق می‌افتد، به این معناست که هنوز فضای زیادی برای نوآوری در آینده در این کار وجود دارد.

۴ـ۲ مسائل باز و کارهای قابل انجام

مراجع

[1] Acharya, M., Kafle, K., and Kanan, C. Tallyqa: Answering complex counting questions. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), volume 33, pp. 8076–8084.

[2] Alberti, C., Ling, J., Collins, M., and Reitter, D. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054* (2019).

[3] Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Deep compositional question answering with neural module networks. corr abs/1511.02799 (2015). *arXiv preprint arXiv:1511.02799* (2015).

[4] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. Vqa: Visual question answering. in *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2425–2433.

[5] Bai, Y., Fu, J., Zhao, T., and Mei, T. Deep attention neural tensor network for visual question answering. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 20–35.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[7] Cao, L., Gao, L., Song, J., Xu, X., and Shen, H. T. Jointly learning attentions with semantic cross-modal correlation for visual question answering. in *Australasian Database Conference* (2017), Springer, pp. 248–260.

[8] Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., and Nevatia, R. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960* (2015).

[9] Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. in *European Conference on Computer Vision* (2020), Springer, pp. 104–120.

[10] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. in *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.

[12] Denkowski, M., and Lavie, A. Meteor universal: Language specific translation evaluation for any target language. in *Proceedings of the ninth workshop on statistical machine translation* (2014), pp. 376–380.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[15] 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. in *Advances in neural information processing systems* (2015), pp. 2296–2304.

[16] Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision 106*, 2 (2014), 210–233.

[17] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 6904–6913.

[18] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3608–3617.

[19] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[20] Hinton, G. E., Krizhevsky, A., and Sutskever, I. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 25* (2012), 1106–1114.

[21] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[22] Jabri, A., Joulin, A., and Van Der Maaten, L. Revisiting visual question answering baselines. in *European conference on computer vision* (2016), Springer, pp. 727–739.

[23] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2901–2910.

[24] Kafle, K., and Kanan, C. Answer-type prediction for visual question answering. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4976–4984.

[25] Kafle, K., and Kanan, C. An analysis of visual question answering algorithms. in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1965–1973.

[26] Kafle, K., Yousefhussien, M., and Kanan, C. Data augmentation for visual question answering. in *Proceedings of the 10th International Conference on Natural Language Generation* (2017), pp. 198–202.

[27] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision 123*, 1 (2017), 32–73.

[28] Lao, M., Guo, Y., Wang, H., and Zhang, X. Cross-modal multistep fusion network with co-attention for visual question answering. *IEEE Access 6* (2018), 31516–31524.

[29] Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., and Zhou, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. in *AAAI* (2020), pp. 11336–11344.

[30] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. in *European Conference on Computer Vision* (2020), Springer, pp. 121–137.

[31] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. in *European conference on computer vision* (2014), Springer, pp. 740–755.

[32] Lin, X., and Parikh, D. Leveraging visual question answering for image-caption ranking. in *European Conference on Computer Vision* (2016), Springer, pp. 261–277.

[33] Lioutas, V., Passalis, N., and Tefas, A. Explicit ensemble attention learning for improving visual question answering. *Pattern Recognition Letters 111* (2018), 51–57.

[34] Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. in *Advances in Neural Information Processing Systems* (2019), pp. 13–23.

[35] Ma, L., Lu, Z., and Li, H. Learning to answer questions from image using convolutional neural network. in *AAAI* (2016).

[36] Malinowski, M., Doersch, C., Santoro, A., and Battaglia, P. Learning visual question answering by bootstrapping hard attention. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–20.

[37] Malinowski, M., and Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. in *Advances in neural information processing systems* (2014), pp. 1682–1690.

[38] Malinowski, M., Rohrbach, M., and Fritz, M. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision 125*, 1-3 (2017), 110–135.

[39] Manmadhan, S., and Kovoor, B. C. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review* (2020), 1–41.

[40] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[41] Noh, H., Hongsuck Seo, P., and Han, B. Image question answering using convolutional neural network with dynamic parameter prediction. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 30–38.

[42] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.

[43] Peng, L., Yang, Y., Bin, Y., Xie, N., Shen, F., Ji, Y., and Xu, X. Word-to-region attention network for visual question answering. *Multimedia Tools and Applications 78*, 3 (2019), 3843–3858.

[44] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[45] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[46] Ren, M., Kiros, R., and Zemel, R. Exploring models and data for image question answering. in *Advances in neural information processing systems* (2015), pp. 2953–2961.

[47] Ren, M., Kiros, R., and Zemel, R. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst 1*, 2 (2015), 5.

[48] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. in *Advances in neural information processing systems* (2015), pp. 91–99.

[49] Ruwa, N., Mao, Q., Wang, L., and Dong, M. Affective visual question answering network. in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)* (2018), IEEE, pp. 170–173.

[50] Shah, S., Mishra, A., Yadati, N., and Talukdar, P. P. Kvqa: Knowledge-aware visual question answering. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), volume 33, pp. 8876–8884.

[51] Shi, Y., Furlanello, T., Zha, S., and Anandkumar, A. Question type guided attention in visual question answering. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 151–166.

[52] Shih, K. J., Singh, S., and Hoiem, D. Where to look: Focus regions for visual question answering. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4613–4621.

[53] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. in *European conference on computer vision* (2012), Springer, pp. 746–760.

[54] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[55] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[56] Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. Videobert: A joint model for video and language representation learning. in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 7464–7473.

[57] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[58] Talafha, B., and Al-Ayyoub, M. Just at vqa-med: A vgg-seq2seq model. in *CLEF (Working Notes)* (2018).

[59] Tan, H., and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[60] Tang, R., Ma, C., Zhang, W. E., Wu, Q., and Yang, X. Semantic equivalent adversarial data augmentation for visual question answering. in *European Conference on Computer Vision* (2020), Springer, pp. 437–453.

[61] Teney, D., and van den Hengel, A. Visual question answering as a meta learning task. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 219–235.

[62] Tommasi, T., Mallya, A., Plummer, B., Lazebnik, S., Berg, A. C., and Berg, T. L. Combining multiple cues for visual madlibs question answering. *International Journal of Computer Vision 127*, 1 (2019), 38–60.

[63] Toor, A. S., Wechsler, H., and Nappi, M. Question action relevance and editing for visual question answering. *Multimedia Tools and Applications 78*, 3 (2019), 2921–2935.

[64] Wang, P., Wu, Q., Shen, C., Hengel, A. v. d., and Dick, A. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).

[65] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., and van den Hengel, A. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding 163* (2017), 21–40.

[66] Wu, Z., and Palmer, M. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033* (1994).

[67] Xu, H., and Saenko, K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. in *European Conference on Computer Vision* (2016), Springer, pp. 451–466.

[68] Yu, L., Park, E., Berg, A. C., and Berg, T. L. Visual madlibs: Fill in the blank description generation and question answering. in *Proceedings of the ieee international conference on computer vision* (2015), pp. 2461–2469.

[69] Yu, Z., Yu, J., Xiang, C., Fan, J., and Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems 29*, 12 (2018), 5947–5959.

[70] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. Unified vision-language pre-training for image captioning and vqa. in *AAAI* (2020), pp. 13041–13049.

[71] Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. Visual7w: Grounded question answering in images. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4995–5004.

واژه‌نامه فارسی به انگلیسی

دستیاران صوتی ..... Voice assistants
عامل‌های گفتگو ..... Conversational agents
توصیف تصویر ..... Image captioning
بازیابی متن به تصویر ..... Text-to-image retrieval
تشخیص اشیا ..... Object detection
تشخیص فعالیت ..... Activity detection
طبقه‌بندی صفات ..... Attribute classification
طبقه‌بندی صحنه ..... Scene classification
بردار سراسری ..... Global vector
جمع متناظر ..... Element-wise addition
ضرب متناظر ..... Element-wise multiplication
چرخش ..... Rotation
ورق زدن ..... Flipping
افزایش داده ..... Data augmentation

واژه‌نامه انگلیسی به فارسی

دستیاران صوتی ..... Voice assistants
عامل‌های گفتگو ..... Conversational agents
توصیف تصویر ..... Image captioning
بازیابی متن به تصویر ..... Text-to-image retrieval
تشخیص اشیا ..... Object detection
تشخیص فعالیت ..... Activity detection
طبقه‌بندی صفات ..... Attribute classification
طبقه‌بندی صحنه ..... Scene classification
بردار سراسری ..... Global vector
جمع متناظر ..... Element-wise addition
ضرب متناظر ..... Element-wise multiplication
چرخش ..... rotation
ورق زدن ..... flipping
افزایش داده ..... Data augmentation

Abstract:

Visual Question Answering(VQA) is a challenging task that has been introduced in recent years and has received increasing attention from both the computer vision and the natural language processing communities. Visual Question Answering aims to answer the questions about given images. A VQA system tries to find the correct answer to questions using visual elements of the image and inference gathered from textual questions. In the first chapter of this review, we present the Visual Question Answering task, applications, and challenges. After defining some concepts in the second chapter, we discuss various datasets for VQA, methods, and evaluation metrics in chapter 3. Due to the success of deep learning and pre-trained models, we classify VQA methods into two general approaches: deep learning and pre-trained models. In the last chapter, after concluding on the different aspects of VQA, we provide some directions for future work.

Keywords: Visual Question Answering, Natural Language Processing, Computer Vision, Deep Learning, pretrained models

Iran University of Science and Technology
Computer Engineering Department

Visual Question Answering

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering

By:

Maryam Sadat Hashemi

Supervisor:

Dr. Sayyed Sauleh Eetemadi

December 2020