



دانشکده مهندسی کامپیوتر

پرسش و پاسخ تصویری

گزارش سمینار برای دریافت درجه کارشناسی ارشد در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی

مریم سادات هاشمی

استاد راهنما

سید صالح اعتمادی

دی ۱۳۹۹

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

چکیده

چکیده

واژگان کلیدی: واژگان کلیدی

فهرست مطالب

ج	فهرست تصاویر
چ	فهرست جداول
۱	فصل ۱: مقدمه
۱-۱	۱-۱ شرح مسئله
۲-۱	۲-۱ کاربرد و اهمیت مسئله
۳-۱	۳-۱ بررسی چالش‌های موجود در این مسئله
۴-۱	۴-۱ بررسی مجموعه داده‌گان مطرح این حوزه
۴-۱	۴-۱-۱ مجموعه داده DAQUAR [۲۸]
۵-۱	۴-۱-۲ مجموعه داده VQA [۳] [۱۳]
۷-۱	۴-۱-۳ مجموعه داده Visual Madlibs [۵۳]
۷-۱	۴-۱-۴ مجموعه داده Visual7w [۵۵]
۸-۱	۴-۱-۵ مجموعه داده CLEVR [۱۹]
۹-۱	۴-۱-۶ مجموعه داده Tally-QA [۱]
۱۰-۱	۴-۱-۷ مجموعه داده KVQA [۳۹]
۱۱-۱	۵-۱ افزایش مجموعه داده در مسئله پرسش و پاسخ تصویری
۱۲-۱	۶-۱ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری
۱۲-۱	۶-۱-۱ فاز ۱: استخراج ویژگی از تصویر و سوال
۱۶-۱	۶-۱-۲ فاز ۲: بازنمایی مشترک تصویر و سوال

۱-۶-۳ : تولید جواب	۱۸
۱-۷ شبکه های از قبل آموزش دیده بر روی زبان طبیعی و تصویر	۱۹
۱-۸ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری	۲۰
۱-۸-۱ معیار دقت	۲۰
۱-۸-۲ معیار شباهت Wu-Palmer [۵۱]	۲۱
۱-۸-۳ معیار اجماع	۲۱
۱-۸-۴ MPT [۲۰]	۲۱
۱-۸-۵ BLEU [۳۳]	۲۲
۱-۸-۶ METEOR [۹]	۲۲
۱-۹ چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی	۲۲

۲۳

مراجع

۲۹

واژه نامه فارسی به انگلیسی

۳۰

واژه نامه انگلیسی به فارسی

فهرست تصاویر

- ۱-۱ مثالی از سیستم پرسش و پاسخ متنی و تصویری ۲
- ۱-۲ چند نمونه از مجموعه داده DAQUAR [۲۸] ۵
- ۱-۳ چند نمونه از مجموعه داده VQA v1 - real [۳] ۶
- ۱-۴ چند نمونه از مجموعه داده VQA v1 - abstaret [۳] ۶
- ۱-۵ چند نمونه از مجموعه داده VQA v2 [۱۳] ۷
- ۱-۶ یک نمونه از مجموعه داده Visual Madlibs [۵۳] ۸
- ۱-۷ چند نمونه از مجموعه داده Visual7W [۵۵]. 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. ۹
- ۱-۸ چند نمونه از مجموعه داده CLEVR [۱۹]. ۱۰
- ۱-۹ چند نمونه از مجموعه داده Tally-QA [۱]. عکس سمت چپ یک نمونه از سوالات ساده و عکس سمت راست یک نمونه از سوالات پیچیده است. ۱۰
- ۱-۱۰ چند نمونه از مجموعه داده KVQA [۳۹] ۱۱
- ۱-۱۱ معماری شبکه CBOW و Skip-gram ۱۵

فهرست جداول

- ۱-۱ بررسی اجمالی مجموعه داده های معروف در حوزه پرسش و پاسخ تصویری. ۴
- ۱-۲ الگوهای استفاده شده برای تولید سوال در مجموعه داده DAQUAR. سوالات می تواند در مورد یک تصویر و یا مجموعه ای از تصاویر باشد [۲۸] ۵
- ۱-۳ بررسی اجمالی مهم ترین شبکه های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده. ۱۳
- ۱-۴ شبکه های عصبی کانولوشنی استفاده شده در مدل های پرسش و پاسخ تصویری. ۱۴
- ۱-۵ word embedding های استفاده شده در مدل های پرسش و پاسخ تصویری. ۱۷

فصل ۱

مقدمه

۱-۱ شرح مسئله

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 قرار گرفته 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 جمع‌آوری شده از سؤال متنی، پاسخ صحیح را پیدا کند [۳۰]. پرسش و پاسخ تصویری نسخه گسترش یافته مسئله پرسش و پاسخ متنی است که اطلاعات بصری به مسئله اضافه شده است. شکل ۱-۱ گویای تفاوت این دو مسئله است.

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



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

نشانگر سطح بالاتری از انتزاع دنیای واقعی است [۵۰].

۱-۲ 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 [۱۴]. علاوه بر این، در سال‌های اخیر دستیاران صوتی^۱ و عامل‌های گفتگو^۲ مانند Siri، Cortana 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

^۱ Voice Assistants
^۲ Conversational Agents

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 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 [۴۵].

۱-۳. 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؟).

^۳Image Captioning

^۴Text-to-image Retrieval

^۵Object Detection

^۶Activity Recognition

^۷Attribute Classification

^۸Scene Classification

جدول ۱-۱: بررسی اجمالی مجموعه داده‌های معروف در حوزه پرسش و پاسخ تصویری.

مجموعه داده	تعداد تصاویر	تعداد سوالات	سال انتشار
DAQUAR [۲۸]	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1 [۳]	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
Visual Madlibs [۵۳]	۱۰۷۳۸	۳۶۰۰۱	۲۰۱۵
Visual7w [۵۵]	۴۷۳۰۰	۲۲۰۱۱۵۴	۲۰۱۶
VQA v2 [۱۳]	۲۰۴۷۲۱	۱۱۰۵۹۰۴	۲۰۱۷
CLEVR [۱۹]	۱۰۰۰۰۰	۸۵۳۵۵۴	۲۰۱۷
Tally-QA [۱]	۱۶۵۰۰۰	۳۰۶۹۰۷	۲۰۱۹
KVQA [۳۹]	۲۴۶۰۲	۱۸۳۰۰۷	۲۰۱۹

۴-۱. بررسی مجموعه دادگان مطرح این حوزه

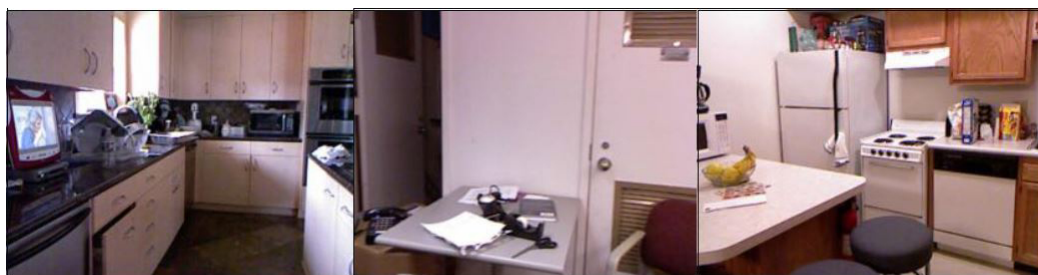
در این بخش به معرفی مجموعه داده‌های مشهور در حوزه پرسش و پاسخ تصویری می‌پردازیم و ویژگی‌های هر کدام را بررسی خواهیم کرد. در جدول ۱-۱ اطلاعات آماری این مجموعه داده‌ها به صورت خلاصه آمده است.

۱-۴-۱. مجموعه داده DAQUAR [۲۸]

DAQUAR مخفف Dataset for Question Answering on Real World Images است که توسط مالدینوفسکی منتشر شده است. این اولین مجموعه داده‌ای است که برای مسئله VQA منتشر شده است. تصاویر از مجموعه داده NYU-Depth V2 [۴۲] گرفته شده است. اندازه این مجموعه داده کوچک است و در مجموع ۱۴۴۹ تصویر دارد. DAQUAR شامل ۱۲۴۶۸ زوج پرسش و پاسخ با ۲۴۸۳ سوال منحصر به فرد است. برای تولید پرسش و پاسخ‌ها از دو روش مصنوعی و انسانی استفاده شده است. در روش مصنوعی پرسش و پاسخ‌ها به صورت خودکار از الگوهای موجود در جدول ۱-۲ تولید شده است. در روش دیگر از ۵ نفر انسان خواسته شده است تا پرسش و پاسخ تولید کنند. تعداد پرسش و پاسخ‌های آموزشی در این مجموعه داده ۶۷۹۴ و تعداد پرسش و پاسخ‌های تست ۵۶۴ است و به طور میانگین برای هر عکس تقریباً ۹ پرسش و پاسخ وجود دارد. این مجموعه داده با مشکل بایاس روبه‌رو است زیرا تصاویر این مجموعه تنها مربوط به داخل خانه است و بیش از ۴۰۰ مورد وجود دارد که اشیایی مثل میز و صندلی در پاسخ‌ها تکرار شده است.

جدول ۱-۲: الگوهای استفاده شده برای تولید سوال در مجموعه داده DAQUAR. سوالات می تواند در مورد یک تصویر و یا مجموعه ای از تصاویر باشد [۲۸].

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



QA: (What is the object on the counter in the corner?, microwave)

QA: (How many doors are open?, 1)

QA: (Where is oven?, on the right side of refrigerator)

شکل ۱-۲: چند نمونه از مجموعه داده DAQUAR [۲۸]

۱-۴-۲ مجموعه داده VQA [۳] [۱۳]

مجموعه داده Visual Question Answering v1 (VQA v1) ^۹ یکی از پرکاربردترین مجموعه داده ها در زمینه پرسش و پاسخ تصویری است. این مجموعه داده شامل دو بخش است. یک بخش از تصاویر واقعی ساخته شده است که VQA-real نام دارد و بخش دیگر با تصاویر کارتونی ساخته شده است که با نام VQA-abstract از آن در مقالات یاد می شود.

VQA-real به ترتیب شامل ۱۲۳۲۸۷ تصویر آموزشی و ۸۱۴۳۴ تصویر آزمایشی است که این تصاویر از مجموعه داده MS-COCO [۲۳] تهیه شده است. برای جمع آوری پرسش و پاسخ از نیروی انسانی استفاده شده است. برای هر تصویر حداقل ۳ سوال منحصر به فرد وجود دارد و برای هر سوال ۱۰ پاسخ توسط کاربرهای منحصر به فرد جمع آوری شده است. این مجموعه داده شامل ۶۱۴۱۶۳ سوال به صورت open-ended و چندگزینه ای است. در [۲] بررسی دقیقی در مورد نوع سوالات، طول سوالات و پاسخ ها و غیره انجام شده است.

^۹<https://visualqa.org/>

VQA-abstract به عنوان یک مجموعه داده جداگانه و مکمل در کنار VQA-real قرار دارد. هدف از این مجموعه داده 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 مجموعه داده به صورت دستی توسط انسان‌ها و به وسیله‌ی رابط کاربری که از قبل آماده شده است؛ ساخته شده است. تصاویر می‌تواند دو حالت را نشان دهند: داخل خانه و خارج از خانه که هر کدام مجموعه متفاوتی از عناصر را شامل می‌شوند از جمله حیوانات، اشیاء و انسان‌ها با حالت‌های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجاد شده است. مشابه VQA-real، ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع‌آوری شده است. مجموعه داده Visual Question Answering v2 (VQA v2) در سال ۲۰۱۷ پس از مجموعه داده 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 تقریباً دو برابر مجموعه داده‌ی VQA v1 است. در مجموعه داده‌ی VQA v2 تقریباً 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, 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 ۱-۳: چند نمونه از مجموعه داده real - VQA v1 [۲]



Q: Who looks happier ?

A: old person, man, man, man, old man, 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 ۱-۴: چند نمونه از مجموعه داده abstract - VQA v1 [۳]



شکل ۱-۵: چند نمونه از مجموعه داده VQA v2 [۱۳]

۱-۴-۳ مجموعه داده Visual Madlibs [۵۳]

مجموعه داده Visual Madlibs شکل متفاوتی از پرسش و پاسخ را ارائه می دهد. برای هر تصویر جملاتی در نظر گرفته شده است و یک کلمه از آن که معمولاً مربوط به آدم، اشیا و فعالیت های نمایش داده شده در تصویر است؛ از جمله حذف شده و به جای آن جای خالی قرار گرفته است. پاسخ ها کلماتی هستند که این جملات را تکمیل می کنند. برای مثال جمله ”دو [جای خالی] در پارک [جای خالی] بازی می کنند.“ در وصف یک تصویر بیان شده است که با دو کلمه ”مرد“ و ”فریزبی“ می توان جاهای خالی را پر کرد. این مجموعه داده شامل ۱۰۷۳۸ تصویر از مجموعه داده MS-COCO [۲۳] و ۳۶۰۰۰۱ جمله با جای خالی است. جملات با جای خالی به طور خودکار و با استفاده از الگوهای از پیش تعیین شده تولید شده اند. پاسخ ها در این مجموعه داده به هر دو شکل open-ended و چندگزینه ای است.

۱-۴-۴ مجموعه داده Visual7w [۵۵]

مجموعه داده Visual7W نیز بر اساس مجموعه داده MS-COCO [۲۳] ساخته شده است. این مجموعه داده شامل ۴۷۳۰۰ تصویر و ۳۲۷۹۳۹ جفت سوال و پاسخ است. این مجموعه داده همچنین از ۱۳۱۱۷۵۶ پرسش و پاسخ چندگزینه ای تشکیل شده است که هر سوال ۴ گزینه دارد و تنها یکی از گزینه ها پاسخ صحیح سوال است. برای جمع آوری سوالات چندگزینه ای توسط انسان ها از پلتفرم آنلاین Amazon Mechanical Turk استفاده شده است. نکته ی حائز اهمیت در این مجموعه داده این است که تمامی اشیایی که در متن پرسش یا



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 ۱-۶: یک نمونه از مجموعه داده Visual Madlibs [۵۳]

پاسخ ذکر شده است، به نحوی به کادر محدودکننده آن شی در تصویر مرتبط شده است. مزیت این روش، رفع ابهام‌های موجود در متن است. همان‌طور که از نام این مجموعه داده پیداست؛ سوالات آن با ۷ کلمه‌ی پرسشی که حرف اول آن w است شروع می‌شود. این ۷ کلمه شامل why ، who ، when ، where ، what ، how و which است. پرسش‌های Visual7W نسبت به به مجموعه داده VQA v1 غنی‌تر و سخت‌تر است. همچنین پاسخ‌ها طولانی‌تر هستند.

۵-۴-۱ مجموعه داده CLEVR [۱۹]

CLEVR یک مجموعه داده برای ارزیابی درک بصری سیستم‌های VQA است. تصاویر این مجموعه داده با استفاده از سه شی استوانه، کره و مکعب تولید شده است. برای هر کدام از این اشیا دو اندازه متفاوت، دو جنس متفاوت و هشت رنگ مختلف در نظر گرفته شده است. سوالات هم به طور مصنوعی بر اساس مکانی که اشیا در تصویر قرار گرفته اند؛ ایجاد شده است. سوالات در CLEVR به گونه‌ای طراحی شده است 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 مجموعه داده مکان تصاویر نیز با



Q: What endangered animal is featured on the truck?

A: A bald eagle.
A: A sparrow.
A: A hummingbird.
A: A raven.



Q: Where will the driver go if turning right?

A: Onto 24 % Rd.
A: Onto 25 % Rd.
A: Onto 23 % Rd.
A: Onto Main Street.



Q: When was the picture taken?

A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.



Q: Who is under the umbrella?

A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.



Q: Why was the hand of the woman over the left shoulder of the man?

A: They were together and engaging in affection.
A: The woman was trying to get the man's attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.



Q: How many magnets are on the bottom of the fridge?

A: 5.
A: 2.
A: 3.
A: 4.



Q: Which pillow is farther from the window?



Q: Which step leads to the tub?



Q: Which is the small computer in the corner?



Q: Which item is used to cut items?



Q: Which doughnut has multicolored sprinkles?



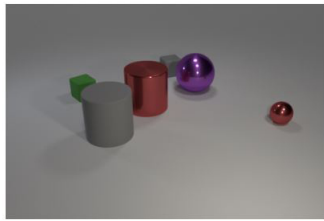
Q: Which man is wearing the red tie?

شکل ۱-۷: چند نمونه از مجموعه داده Visual7W [۵۵]. 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.

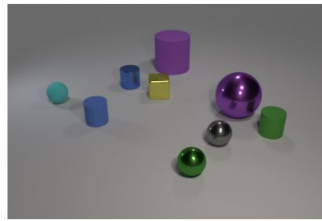
استفاده از یک مستطیل مشخص شده است.

۴-۶ مجموعه داده Tally-QA [۱]

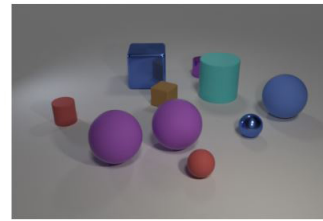
در سال ۲۰۱۹، مجموعه داده Tally-QA منتشر شد که بزرگ‌ترین مجموعه داده پرسش و پاسخ تصویری برای شمارش اشیاء است. اکثر مجموعه داده‌های شمارش اشیاء در پرسش و پاسخ تصویری دارای سوالات ساده هستند که برای پاسخ دادن به این سوال‌ها تنها کافی است که اشیاء در تصویر تشخیص داده شوند. بنابراین، این موضوع باعث ایجاد مجموعه داده‌ی 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 اشیاء نیاز است. تعداد سوالات ساده در Tally-QA برابر با ۲۱۱۴۳۰ و تعداد سوالات پیچیده برابر با ۷۶۴۷۷ است. سوالات ساده این مجموعه داده از مجموعه داده‌های دیگری (VQA v2 [۱۳] و Visual Genome [۲۱]) برداشته شده است و سوالات پیچیده با استفاده از ۸۰۰ کاربر انسانی از طریق پلتفرم آنلاین Amazon Mechanical Turk جمع‌آوری شده است. مجموعه داده Tally-QA به سه بخش آموزش و تست - ساده و تست - پیچیده تقسیم می‌شود. بخش تست -



Q: How big is the gray rubber object that is behind the big shiny thing behind the big der; what material is on the left side of the purple ball?
A: small



Q: There is a tiny rubber thing that is the same color as the metal cylinder; what shape is it?
A: cylinder

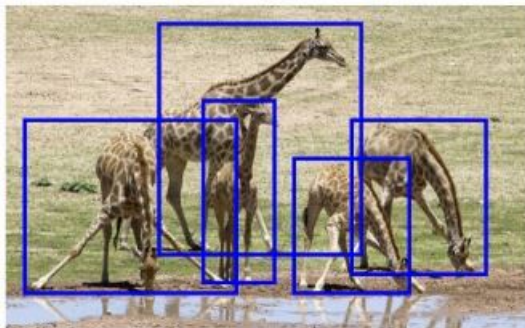


Q: There is a small ball that is made of red rubber sphere the same material as the purple the large block; what color is it?
A: blue

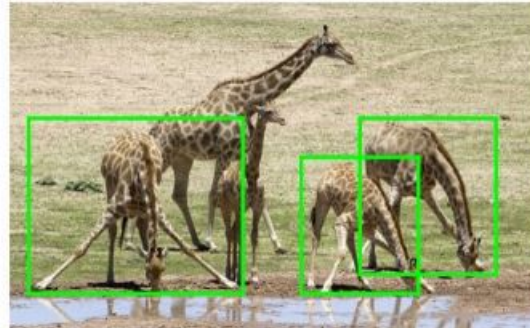
شکل ۱-۸: چند نمونه از مجموعه داده CLEVR [۱۹].

ساده تنها شامل سوالات ساده و بخش تست-پیچیده تنها دارای سوالات پیچیده‌ای است که از Amazon Mechanical Turk جمع‌آوری شده است.

“How many giraffes?”



“How many giraffes are drinking water?”



شکل ۱-۹: چند نمونه از مجموعه داده Tally-QA [۱]. عکس سمت چپ یک نمونه از سوالات ساده و عکس سمت راست یک نمونه از سوالات پیچیده است.

۱-۴-۷ مجموعه داده KVQA [۳۹]

مجموعه داده KVQA که مخفف Knowledge-based Visual Question Answering است در سال ۲۰۱۹ طراحی شده است به طوری که بر خلاف مجموعه داده‌های 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 مجموعه داده 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 گونه ای در نظر گرفته شده است که مشکل همیشگی بایاس در مجموعه داده های 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 ۱-۱۰: چند نمونه از مجموعه داده KVQA [۳۹]

۵-۱ افزایش مجموعه داده در مسئله پرسش و پاسخ تصویری

به لطف توسعه سریع شبکه های عصبی عمیق مسئله پرسش و پاسخ تصویری به موفقیت های بزرگی دست یافته است. مطالعات نشان می دهد که عملکرد شبکه های عصبی عمیق به میزان داده های آموزشی آن بستگی دارد و آن ها همیشه از داده های آموزشی پیشتر سود می برند. بنابراین یکی از ترفندهای اصلی در شبکه های عصبی عمیق افزایش داده است که به طور گسترده در بسیاری از مسائل پردازش تصویر و بینایی کامپیوتر مورد استفاده قرار می گیرد. اما مقالات کمی وجود دارد که مسئله افزایش داده را در پرسش و پاسخ تصویری بررسی کرده باشند. یکی از چالشهای افزایش داده در مسئله پرسش و پاسخ تصویری این است که هیچ یک از

روش های افزایش داده مبتنی بر تصویر مانند چرخش و ورق زدن نمی توانند مستقیماً بر روی مسئله پرسش و پاسخ تصویری اعمال شود زیرا ساختار معنایی آن به درستی حفظ نخواهد شد. به عنوان مثال با چرخش یک تصویر ممکن است پرسش و پاسخ مرتبط با آن دیگر درست نباشد.

۱-۶ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

بسیاری از محققان راه حل ها یا الگوریتم هایی را برای حل مسئله پرسش و پاسخ تصویری پیشنهاد کرده اند که به طور کلی می توان آن را به یک فرآیند سه فازی تقسیم بندی کرد. فاز اول این فرآیند استخراج ویژگی از تصویر و سوالات است که راه حل های موفق در این فاز ریشه در روش های باشکوه یادگیری عمیق دارد زیرا بیشتر راه حل های موفق در این حوزه از مدل های یادگیری عمیق استفاده می کنند 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 مهم ترین و اصلی ترین فاز می باشد، ویژگی های استخراج شده از تصویر و سوال باهم ترکیب می شوند. سپس از ترکیب ویژگی ها برای تولید پاسخ نهایی در فاز سوم استفاده می شود.

۱-۶-۱ فاز ۱: استخراج ویژگی از تصویر و سوال

استخراج ویژگی از تصویر و سوال مرحله ی مقدماتی در پرسش و پاسخ تصویری است. ویژگی تصویر، تصویر را به عنوان یک بردار عددی توصیف می کند تا بتوان به راحتی عملیات های مختلف ریاضی را بر روی آن اعمال کرد. روش های زیادی وجود دارد که به صورت مستقیم از تصویر ویژگی استخراج می کنند مانند بردار ساده RGB، SIFT، تبدیل HAAR و HOG. اما با ظهور شبکه های یادگیری عمیق، نیاز به استخراج ویژگی به صورت مستقیم از بین رفت زیرا این شبکه ها قادر به یادگیری ویژگی هستند. آموزش مدل های یادگیری عمیق به منابع محاسباتی گران قیمت و مجموعه داده های بزرگ نیاز دارد. از این رو، استفاده از مدل های شبکه عصبی عمیق از قبل آموزش دیده، استخراج ویژگی از تصاویر را به راحتی امکان پذیر می کنند.

یکی از بهترین شبکه های عصبی برای استخراج ویژگی از تصویر، شبکه های عصبی کانولوشنی هستند. در جدول ۱-۳ چند نمونه از برجسته ترین شبکه های عصبی کانولوشنی که بر روی مجموعه داده ImageNet [۸] آموزش داده شده اند؛ آورده شده است. بیشتر مدل های ارائه شده در پرسش و پاسخ تصویری از این شبکه های عصبی کانولوشنی استفاده می کنند تا محتوای تصویری خود را به بردارهایی عددی تبدیل کنند.

جدول ۱-۳: بررسی اجمالی مهم‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده.

مدل CNN	سال	تعداد لایه‌ها	ابعاد ورودی	ابعاد خروجی (تعداد ویژگی‌ها)
AlexNet [۱۶]	۲۰۱۲	۸	۲۲۷×۲۲۷	۴۰۹۶
VGGNet [۴۳]	۲۰۱۴	۱۹	۲۲۴×۲۲۴	۴۰۹۶
GoogleNet [۴۴]	۲۰۱۴	۲۲	۲۲۹×۲۲۹	۱۰۲۴
ResNet [۱۵]	۲۰۱۵	۱۵۲	۲۲۴×۲۲۴	۲۰۱۴۸

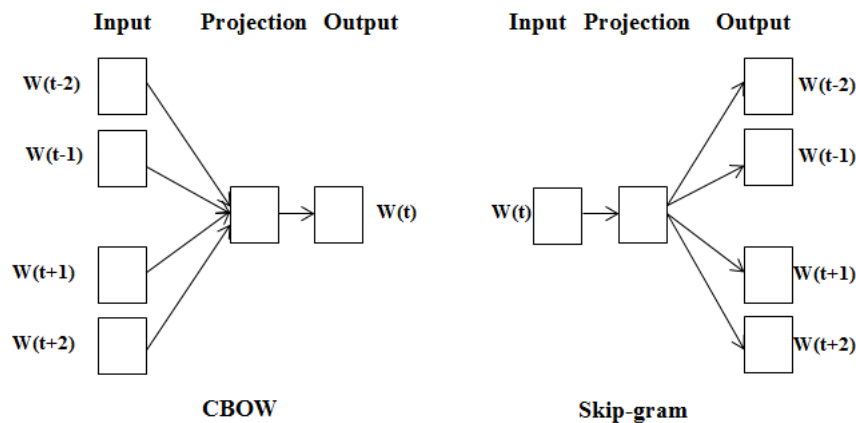
جدول ۱-۴ لیستی از مدل‌های استفاده شده برای حل مسئله پرسش و پاسخ تصویری را نشان می‌دهد و مشخص می‌کند که هر کدام از این مدل‌ها برای استخراج ویژگی از تصویر از کدام یک از شبکه‌های عصبی کانولوشنی موجود در جدول ۱-۳ بهره می‌برد. همان‌طور که واضح است VGGNet و ResNet به طور گسترده‌ای در سیستم‌های پرسش و پاسخ تصویری مورد استفاده قرار گرفته‌اند. یکی از دلایلی که محققان VGGNet را ترجیح می‌دهند این است که ویژگی‌هایی را استخراج می‌کند که عمومیت بیشتری دارد و برای مجموعه داده‌هایی 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 ۱-۴ روند مهاجرت از VGGNet به ResNet در مقالات اخیر است. زیرا در سال‌های اخیر، منابع محاسباتی کافی با هزینه مناسب در دسترس محققان می‌باشد.

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

روش کدگذاری one-hot ساده‌ترین روش word embedding است. در این روش یک لغت‌نامه از همه

جدول ۱-۴: شبکه‌های عصبی کانولوشنی استفاده شده در مدل‌های پرسش و پاسخ تصویری.

ResNet	GoogleNet	VGGNet	AlexNet	مدل پرسش و پاسخ تصویری
		✓		[۳۷]Image_QA
	✓			[۱۱]Talk_to_Machine
		✓		[۳]VQA
			✓	[۵۳]Vis_Madlibs
		✓		[۳۶]VIS + LSTM
		✓		[۴۹]Ahab
		✓		[۶]ABC-CNN
		✓		[۲]Comp_QA
		✓		[۳۲]DPPNet
		✓		[۲۶]Answer_CNN
		✓		[۲۴]VQA-Caption
✓				[۱۸]Re_Baseline
✓				[۱۰]MCB
	✓			[۵۲]SMem-VQA
		✓		[۴۱]Region_VQA
		✓		[۵۵]Vis7W
✓	✓	✓	✓	[۲۹]Ask_Neuron
✓				[۵]SCMC
✓				[۲۷]HAN
		✓		[۵۴]StrSem
✓				[۳۸]AVQAN
✓				[۲۲]CMF
✓				[۲۵]EnsAtt
✓				[۴۶]MetaVQA
✓				[۴]DA-NTN
✓				[۵]QGHC
✓				[۴۰]QTA
✓				[۳۴]WRAN
✓				[۴۸]QAR



شکل ۱-۱۱: معماری شبکه Skip-gram و CBOW

واژه‌های منحصر به فرد موجود در مجموعه داده ساخته می‌شود 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 باشند و کلماتی که معنای متفاوتی با یکدیگر دارند تا حد امکان بردارهایشان از هم دور باشند.

برای رفع مشکلات کدگذاری one-hot، دو روش CBOW^{۱۰} [۳۱] و Skip-gram [۳۱] 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 و لایه خروجی است، استفاده شده است. در مدل CBOW کلمات اطراف و نزدیک به یک کلمه (n-1 کلمه) به لایه ورودی داده می‌شود و مدل سعی می‌کند این کلمه (n امین 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 ۱-۱۱ آورده شده است.

^{۱۰} Continuous Bag Of Words

یکی دیگر از word embedding های مشهور، مدل بردار سراسری یا به اختصار GloVe^{۱۱} است که توسط پنینگتون و همکاران [۳۵] در سال ۲۰۱۴ در تیم پردازش زبان های طبیعی دانشگاه استنفورد معرفی و توسعه داده شد. آیا نیاز به توضیح کامل این روش است؟

با پیشرفت یادگیری عمیق در دهه اخیر، محققان برای استخراج ویژگی و بازنمایی متن از LSTM، CNN، [۱۷] و [۷] GRU استفاده کردند. در مسئله پرسش و پاسخ تصویری برای استخراج ویژگی از سوال با استفاده از CNN بردارهای کلمات سوال در کنار هم قرار داده می شود سپس به لایه های کانولوشنی یک بعدی داده می شود و فیلترهای متفاوتی بر روی آن ها اعمال می شود و پس از عبور از لایه max-pooling ویژگی ها بدست می آید.

توضیح LSTM لازمه؟

توضیح GRU لازمه؟

مدل های مختلف در مسئله پرسش و پاسخ تصویری از word embedding های ذکر شده در بالا برای تولید بردار ویژگی برای سوال ها استفاده کرده اند. جدول ۱-۵ لیستی از مدل های پرسش و پاسخ تصویری به همراه word embedding استفاده شده در آن ها را نمایش می دهد. با بررسی جدول ۱-۵ مشاهده می کنیم که محققان حوزه پرسش و پاسخ تصویری ترجیح می دهند؛ برای استخراج ویژگی از متن و بازنمایی آن از LSTM استفاده کنند. آن ها معتقد هستند که RNN ها عملکرد بهتری نسبت به روش های مستقل از دنباله ای کلمات مانند word2vec دارند. اما آموزش RNN ها نیاز به داده های برجسب خورده ی زیادی دارد.

۱-۶-۲ فاز ۲: بازنمایی مشترک تصویر و سوال

در گام اول پرسش و پاسخ تصویری، تصویر و سوال به طور مستقل پردازش می شوند تا از آن ها ویژگی استخراج شود. روش های مختلف برای انجام این کار، در بخش ۱-۶-۱ به تفصیل بررسی شد. در گام بعدی، این ویژگی ها 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 روش های ترکیب ویژگی های استخراج شده از سوال و تصویر می پردازیم.

^{۱۱}Global Vector

جدول ۱-۵: word embedding های استفاده شده در مدل های پرسش و پاسخ تصویری.

GRU	LSTM	CNN	GloVe	Skip-gram/Word2vec	CBOW	one-hot	مدل پرسش و پاسخ تصویری
				✓			[۳۷]Image_QA
	✓						[۱۱]Talk_to_Machine
					✓		[۳]VQA
				✓			[۵۳]Vis_Madlibs
	✓						[۳۶]VIS + LSTM
	✓						[۶]ABC-CNN
	✓						[۲]Comp_QA
✓							[۳۲]DPPNet
		✓					[۲۶]Answer_CNN
	✓						[۲۴]VQA-Caption
				✓			[۱۸]Re_Baseline
	✓						[۱۰]MCB
					✓		[۵۲]SMem-VQA
				✓			[۴۱]Region_VQA
						✓	[۵۵]Vis7W
✓	✓	✓			✓		[۲۹]Ask_Neuron
		✓					[۵]SCMC
	✓						[۲۷]HAN
	✓						[۵۴]StrSem
						✓	[۳۸]AVQAN
	✓		✓				[۲۲]CMF
			✓				[۲۵]EnsAtt
✓			✓				[۴۶]MetaVQA
✓							[۴]DA-NTN
✓							[۵]QGHC
✓							[۳۴]WRAN
			✓				[۴۸]QAR

روش‌های پایه

ساده‌ترین و پایه‌ای‌ترین روش‌ها برای ترکیب ویژگی‌ها concatenation، جمع متناظر ویژگی‌ها^{۱۲} و ضرب متناظر ویژگی‌ها^{۱۳} است. مالینوفسکی در [۲۹] 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 concatenation و جمع متناظر ویژگی‌ها. با توجه به نتایج آن‌ها، جمع متناظر ویژگی‌ها پس از نرمال‌سازی از دقت بالاتری برخوردار است. در [۴۱] 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^{۱۴} 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^{۱۵} نیز دارد که توسط [۱۲] پیشنهاد شده است. در [۵۳] و [۴۷] از هر دو مدل CCA و nCCA برای ترکیب بردارهای ویژگی سوال و تصویر استفاده کردند و دریافتند که روش nCCA به ویژه در مورد سوالات چندگزینه‌ای عملکرد بهتری دارد.

روش‌های مبتنی بر شبکه‌های عصبی end-to-end

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 پیشنهاد شده متفاوت باشد.

ادامه اش باید تکمیل بشه ...

۱-۶-۳ فاز ۳: تولید جواب

باید تکمیل شود.

^{۱۲} element-wise addition

^{۱۳} element-wise multiplication

^{۱۴} Analysis Correlation Canonical

^{۱۵} Analysis Correlation Canonical normalized

۷-۱ شبکه های از قبل آموزش دیده بر روی زبان طبیعی و تصویر

در سال های اخیر شاهد ظهور شبکه های از قبل آموزش دیده تنها بر روی داده های تصویری و یا تنها بر روی داده های زبانی بوده ایم. استفاده از این شبکه ها منجر به بهبود مسائل موجود در بینایی ماشین و پردازش زبان های طبیعی شده است. با الهام از این موضوع شبکه های از قبل آموزش دیده بر روی داده های تصویری و داده های زبانی نیز ایجاد شدند که هدف آنها بازنمایی مشترک داده های تصویری و داده های زبانی است. بنابراین می توان از این شبکه ها برای بهبود عملکرد مسائل مشترک مابین بینایی ماشین و پردازش زبان های طبیعی مانند پرسش و پاسخ تصویری نیز استفاده کرد. در ادامه به بحث و بررسی چند نمونه از این شبکه ها می پردازیم. معماری این شبکه ها به طور کلی به دو دسته تک جریان و دو جریان تقسیم می شود.

معماری تک جریان پایه و اساس این معماری مدل برت است. به طور کلی مدل های پیشنهاد شده در این معماری از داده های چند حالت موازی برای آموزش استفاده می کنند برای مثال تصویر به همراه یک جمله توصیف کننده تصویر و یا یک فیلم به همراه زیرنویس. به علاوه این مدلها با ترکیبی از اهداف مختلف مانند فلان و فلان و فلان بهینه می شود. سپس از بازنمایی های آموخته شده در کارهای پایین دستی به طور کلی درک یا تولید استفاده می شود. به عنوان مثال، معماری VideoBERT برای یادگیری بازنمایی های زبان بینایی برای یک کار پایین دستی مانند تولید توصیف فیلم طراحی شده است. در حالی که چندین رویکرد دیگر مانند ۲T۲B، VL-BERT، Unocoder-VL و UNITER وجود دارد، همگی برای درک چند حالت طراحی شده اند و کارهای پایین دستی را تسهیل می کنند. آثاری مانند VLP و OSCAR مدل های واحدی ساخته اند که می توانند به طور مشترک داده های متقابل را درک و تولید کنند.

معماری دو جریان در مقابل تک جریان، معماری های دو جریان دو رمزگذار مستقل را برای یادگیری نمایش های تصویری و متنی به کار گرفتند. ViLBERT و LXMERT نمونه هایی از معماری دو جریان هستند که از اصول توجه به خود برای یادگیری مشترک بازنمایی از داده های تصویری و متنی استفاده می کنند. ViLBERT یک لایه ترانسفورماتور مشترک را ایجاد می کند، در حالی که LXMERT از یک رمزگذار متقابل استفاده می کند. مشابه معماری تک جریان، معماری های دو جریان نیز مدل های خود را با کارهای قبل از آموزش، مانند MLM و مطابقت بین متن، بهینه می کنند. گاهی اوقات آنها برای دستیابی به تعمیم بهتر جملات طولانی و پیچیده، از شرکای فقط متن دیگری استفاده می کنند.

و برای بهره برداری بهتر از عناصر بصری، تصاویر با استفاده از RoI یا تکنیک های بازیابی جعبه محدود

قبل از رمزگذاری توسط ترانسفورماتورهای آموزش دیده، به توالی مناطق تبدیل می شوند.

۸-۱ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری

در این بخش می خواهیم به طور مختصر معیارهای ارزیابی شناخته شده در مسئله پرسش و پاسخ تصویری را بررسی کنیم. همان طور که قبلاً ذکر شد؛ معمولاً دو نوع سوال در مجموعه داده های پرسش و پاسخ تصویری در نظر گرفته می شود: سوالات open-ended و سوالات چندگزینه ای. در سوالات چندگزینه ای، برای هر سوال دقیقاً یک پاسخ صحیح وجود دارد. بنابراین ارزیابی آن ساده است زیرا می توان 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 مجموعه داده های پرسش و پاسخ تصویری پاسخ ها را محدود به چند کلمه (۱ تا ۳ کلمه) می کنند و یا پاسخ ها را از یک مجموعه بسته انتخاب می کنند.

در ادامه به بررسی مهم ترین معیارهای این حوزه می پردازیم. اما ارزیابی مسئله پرسش و پاسخ تصویری همچنان یک مسئله حل نشده است. هر کدام از روش ها و معیارهای ارزیابی موجود، مزیت ها و معایب خاص خود را دارند. بنابراین برای انتخاب معیار ارزیابی باید به مواردی همچون ساختار مجموعه داده و نحوه ساخت آن، میزان بایاس موجود در مجموعه داده و ... توجه نمود.

۸-۱-۱ معیار دقت

اگر چه در سوالات چندگزینه ای برای سنجش یک مدل معیار دقت کافی است اما در سوالات open-ended معیار دقت سخت گیرانه است زیرا فقط در حالتی که پاسخ مدل کاملاً مطابق با پاسخ در نظر گرفته شده باشد، پذیرفته می شود. برای مثال اگر صورت سوال «چه حیواناتی در تصویر است؟» باشد و پاسخ مدل به جای «سگ ها» پاسخ «سگ» باشد؛ غلط تلقی می شود. بنابراین به دلیل این محدودیت هایی که معیار دقت دارد؛ معیارهای دیگری برای ارزیابی این نوع سوالات پیشنهاد شده است.

$$Accuracy = \frac{\text{Number of questions answered correctly}}{\text{Total questions}} \quad (1-1)$$

۱-۸-۲ معیار شباهت Wu-Palmer [۵۱]

این معیار ارزیابی توسط مالینوفسکی [۲۸] 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 اندازه‌گیری کند. 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.

۱-۸-۳ 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 معمولاً از حالت کمترین اجماع استفاده می‌شود و آستانه را هم برابر ۳ قرار می‌دهند به این معنی که اگر پاسخ پیش‌بینی شده با ۳ 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 [۳] استفاده کرده‌اند.

$$Accuracy_{VQA} = \min\left(\frac{n}{3}, 1\right) \quad (2-1)$$

۱-۸-۴ MPT [۲۰]

یکی از مشکلات مجموعه داده‌های پرسش و پاسخ تصویری توزیع غیریکنواخت انواع سوال‌هاست. در این مواقع، نمی‌توان از معیار دقت استفاده کرد. بنابراین در [۲۰] معیار جدیدی به نام MPT^{۱۶} ارائه شده است که توزیع نامتوازن سوال‌ها را جبران می‌کند. معیار MPT میانگین دقت برای هر نوع سوال را محاسبه می‌کند.

^{۱۶} Mean Per Type

از نسخه‌ی نرمالایز شده‌ی این معیار نیز برای رفع مشکل بایاس در توزیع پاسخ‌ها استفاده می‌شود.

۱-۸-۵ BLEU [۳۳]

BLEU^{۱۷} یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. در [۱۴] 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. معمولاً BLEU زمانی که جمله‌ها کوتاه باشند، با شکست مواجه می‌شود.

۱-۸-۶ METEOR [۹]

METEOR^{۱۸} نیز همانند BLEU یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. به پیشنهاد [۱۴] 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.

۱-۹ چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی

باید تکمیل شود.

^{۱۷} BiLingual Evaluation Understudy
^{۱۸} Metric for Evaluation of Translation with Explicit Ordering

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] 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).
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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).
- [7] 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).

- [8] 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.
- [9] 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.
- [10] 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).
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] LIN, X., AND PARIKH, D. Leveraging visual question answering for image-caption ranking. in *European Conference on Computer Vision* (2016), Springer, pp. 261–277.
- [25] LIOUTAS, V., PASSALIS, N., AND TEFAS, A. Explicit ensemble attention learning for improving visual question answering. *Pattern Recognition Letters* 111 (2018), 51–57.
- [26] MA, L., LU, Z., AND LI, H. Learning to answer questions from image using convolutional neural network. in *AAAI* (2016).
- [27] 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.
- [28] 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.

- [29] 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.
- [30] MANMADHAN, S., AND KOVOOR, B. C. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review* (2020), 1–41.
- [31] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.

- [40] 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.
- [41] 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.
- [42] SILBERMAN, N., HOIEM, D., KOHLI, P., AND FERGUS, R. Indoor segmentation and support inference from rgb-d images. in *European conference on computer vision* (2012), Springer, pp. 746–760.
- [43] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [44] 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.
- [45] TALAFHA, B., AND AL-AYYOUB, M. Just at vqa-med: A vgg-seq2seq model. in *CLEF (Working Notes)* (2018).
- [46] 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.
- [47] TOMMASI, T., MALLIA, 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.
- [48] 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.
- [49] 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).
- [50] 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.

- [51] WU, Z., AND PALMER, M. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033* (1994).
- [52] 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.
- [53] 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.
- [54] 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.
- [55] 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

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

واژه‌نامه انگلیسی به فارسی

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

Abstract:

abstract

Keywords: keywords



**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:

Sayyed Sauleh Eetemadi

December 2020