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


بسم الله الرحمن الرحیم

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

۱ـ۱ 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 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 [۳۰]. پرسش و پاسخ تصویری نسخه گسترش یافته مسئله پرسش و پاسخ متنی است است که اطلاعات بصری به مسئله اضافه شده است. شکل ۱ـ۱ 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

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

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

۱ ـ 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 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

[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].

۱ـ۳ 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

مجموعه‌داده	تعداد تصاویر	تعدادسوالات	سال انتشار
DAQUAR [۲۸]	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1 [۳]	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
Visual Madlibs [۵۳]	۱۰۷۳۸	۳۶۰۰۰۱	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

جدول ۱ـ۱: بررسی اجمالی مجموعه‌داده‌های معروف در حوزه پرسش و پاسخ تصویری.

۱ـ۴ 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 ۱ـ۱ اطلاعات آماری این مجموعه‌داده‌ها به صورت خلاصه آمده‌است.

۱ـ۴ـ۱ مجموعه داده DAQUAR [۲۸]

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 [۴۲] 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 ۱ـ۲ 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 ۴۰۰ مورد وجود دارد که اشیایی مثل میز و صندلی در پاسخ‌ها تکرارشده‌است.

	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?

جدول ۱ـ۲: الگوهای استفاده شده برای تولید سوال در مجموعه‌داده DAQUAR. سوالات می‌تواند در مورد یک تصویر و یا مجموعه‌ای از تصاویر باشد [۲۸].

شکل ۱ـ۲: چند نمونه از مجموعه‌داده DAQUAR [۲۸]

۱ـ۴ـ۲ مجموعه داده VQA [۳] [۱۳]

مجموعه‌داده Visual Question Answering v1(VQA v1) [۹] 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 [۲۳] 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 [۳] بررسی دقیقی در مورد نوع سوالات، طول سوالات و پاسخ‌ها و غیره انجام‌شده‌است.

[۹] https://visualqa.org/

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 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 ۱ـ۳: چند نمونه از مجموعه‌داده VQA v1 - real [۳]

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 ۱ـ۴: چند نمونه از مجموعه‌داده VQA v1 - abstarct [۳]

شکل ۱ـ۵: چند نمونه از مجموعه‌داده VQA v2 [۱۳]

۱ـ۴ـ۳ مجموعه داده Visual Madlibs [۵۳]

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 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 [۲۳] 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.

۱ـ۴ـ۴ مجموعه داده Visual7w [۵۵]

مجموعه‌داده Visual7W نیز بر اساس مجموعه‌داده MS-COCO [۲۳] 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

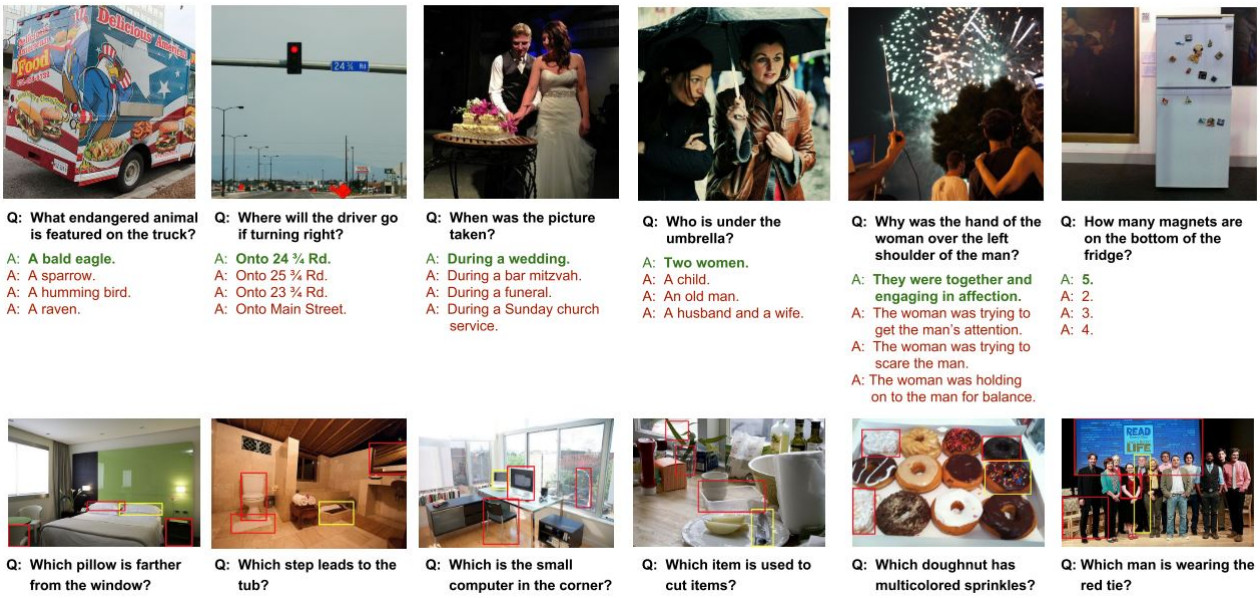
1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will 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 ۱ـ۶: یک نمونه از مجموعه‌داده Visual Madlibs [۵۳]

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.

۱ـ۴ـ۵ مجموعه داده CLEVR [۱۹]

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

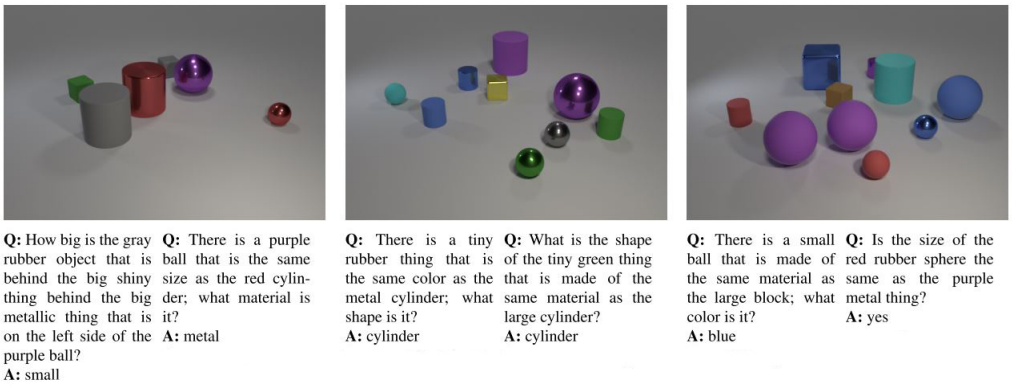


شکل ۱ـ۷: چند نمونه از مجموعه‌داده Visual7W [۵۵]. ردیف اول،پاسخ‌های سبز رنگ، پاسخ صحیح هستند و پاسخ‌های قرمز پاسخ‌های نادرست تولید شده توسط انسان است. ردیف دوم، کادر زرد جواب صحیح است و کادرهای قرمز پاسخ‌های اشتباه انسانی است.

استفاده از یک مستطیل مشخص‌شده‌است.

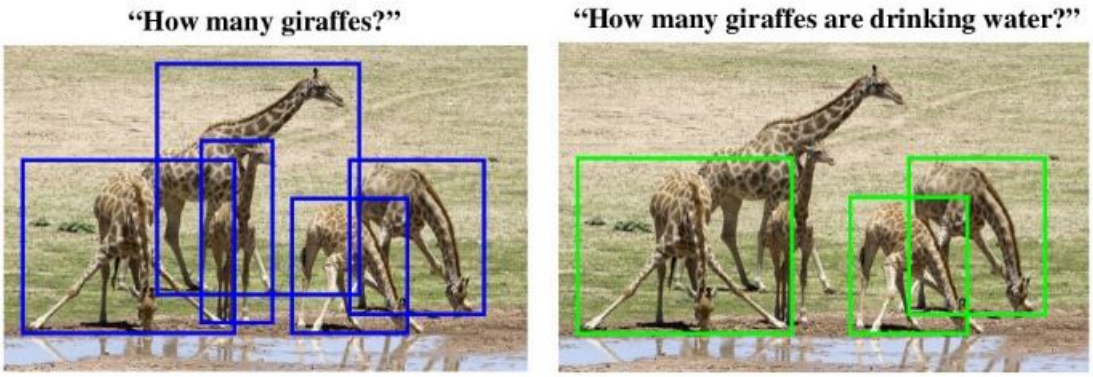
۱ـ۴ـ۶ مجموعه داده Tally-QA [۱]

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 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 [۱۳] و Visual Genome [۲۱]) 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_



شکل ۱ـ۸: چند نمونه از مجموعه‌دادهCLEVR [۱۹].

ساده تنها شامل سوالات ساده و بخش تست‌ـپیچیده تنها دارای سوالات پیچیده‌ای است که از Amazon Mechanical Turk جمع‌آوری شده‌است.

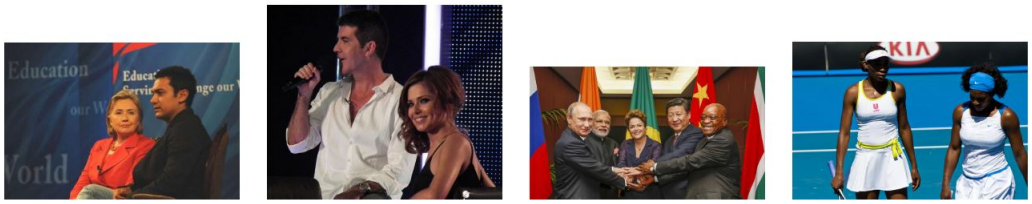


شکل ۱ـ۹: چند نمونه از مجموعه‌دادهTally-QA [۱].عکس سمت چپ یک نمونه از سوالات ساده و عکس سمت راست یک نمونه از سوالات پیچیده است.

۱ـ۴ـ۷ 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 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 ۱ـ۱۰: چند نمونه از مجموعه‌دادهKVQA [۳۹]

.

۱ـ۵ افزایش 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 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 ـ ۶ 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.

۱ ـ ۶ ـ ۱ 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 ۱ ـ ۳ چند نمونه از برجسته‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه‌داده ImageNet [۸] 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.

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 ۱ـ۳: بررسی اجمالی مهم‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه‌داده ImageNet آموزش داده شده.

جدول ۱ـ۴ 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 ۱ـ۳ 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 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 ۱ـ۴ روند مهاجرت از VGGNet به ResNet در مقالات اخیر است. زیرا در سال‌های اخیر، منابع محاسباتی کافی با هزینه مناسب در دسترس محققان می‌باشد.

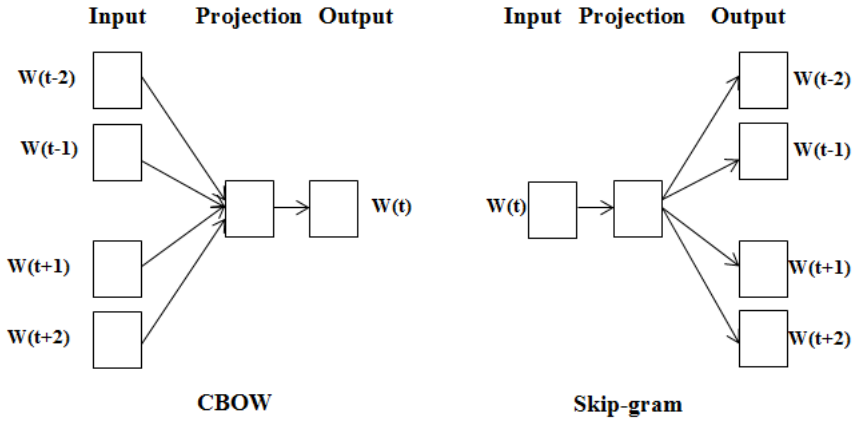
بیشتر الگوریتم‌های یادگیری ماشین و یادگیری عمیق قادر به پردازش متن به شکل خام وساده نیستند و برای بازنمایی متن‌ها نیاز به word embedding دارند. مسئله پرسش و پاسخ تصویری نیز از این قاعده مستثنا نیست و باید برای بازنمایی سوالات از word embedding استفاده کند. word embedding نگاشت کلمات یا عبارات از واژگان 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

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

جدول ۱ـ۴: شبکه‌های عصبی کانولوشنی استفاده شده در مدل‌های پرسش و پاسخ تصویری.



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

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 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 ولایه خروجی است، استفاده شده است. درمدل 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 ـ ۱۱ آورده شده است.

[۱۰] Continouse Bag Of Words

یکی دیگر از word embedding های مشهور، مدل بردار سراسری یا به اختصار GloVe [11] است که توسط پنینگتون و همکاران [۳۵] در سال ۲۰۱۴ در تیم پردازش زبان‌های 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 CNN ، LSTM [۱۷] و GRU [۷] 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 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.

توضیح LSTM لازمه؟

توضیح GRU لازمه؟

مدل‌های مختلف در مسئله پرسش و پاسخ تصویری از word embedding های ذکر شده در بالا برای تولید بردار ویژگی برای سوال‌ها استفاده کرده‌اند. جدول ۱ـ۵ لیستی از مدل‌های پرسش و پاسخ تصویری به همراه word embedding استفاده شده در آن‌ها را نمایش می‌دهد. با بررسی جدول ۱ـ۵ 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.

۱ـ۶ـ۲ فاز ۲ : بازنمایی مشترک تصویر و سوال

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 ۱ـ۶ـ۱ 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.

[11] Global Vector

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

جدول ۱ـ۵: word embedding های استفاده شده در مدل‌های پرسش و پاسخ تصویری.

روش‌های پایه

ساده‌ترین و پایه‌ای‌ترین روش‌ها برای ترکیب ویژگی‌ها concatination ، 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 concatination 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 پیشنهادشده متفاوت باشد.

ادامه اش باید تکمیل بشه ...

۱ـ۶ـ۳ فاز ۳ : تولید جواب

باید تکمیل شود.

[۱۲] element-wise addition
[۱۳] element-wise multiplication
[۱۴] Analysis Correlation Canonical
[۱۵] Analysis Correlation Canonical normalized

۱ـ۷ شبکه های از قبل آموزش دیده بر روی زبان طبیعی و تصویر

در سال‌های اخیر شاهد ظهور شبکه‌های از قبل آموزش‌دیده 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 زبان‌های 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 حالته 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 ، Unicoder-VL ، VL-BERT و UNITER وجود دارد ، همگی برای درک چند حالته 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.

۱ـ۸ 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.

۱ـ۸ـ۱ 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.

$$Accuracy = \frac{Number\ of\ questions\ answered\ correctly}{Total\ questions} \qquad (۱ـ۱)$$

۱ـ۸ـ۲ 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 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.

۱ـ۸ـ۳ 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 [۳] استفاده کرده‌اند.

$$Accuracy_{VQA} = min(\frac{n}{۳}, ۱) \qquad (۱ـ۲)$$

۱ـ۸ـ۴ MPT [۲۰]

یکی از مشکلات مجموعه‌داده‌های پرسش و پاسخ تصویری توزیع غیریکنواخت انواع سوال‌هاست. دراین 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.

۱ـ۸ـ۵ 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. معمولا BLEU زمانی که جمله‌ها کوتاه باشند، با شکست مواجه می‌شود.

۱ـ۸ـ۶ 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.

۱ـ۹ 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 rgbd 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., 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.

[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.