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

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. 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.

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

مجموعه‌داده	تعداد تصاویر	تعدادسوالات	سال انتشار
DAQUAR	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
VQA v2	۰	۰	۰
Visual Madlibs	۱۰۷۳۸	۳۶۰۰۰۱	۲۰۱۵
Visual7w	۴۷۳۰۰	۲۲۰۱۱۵۴	۲۰۱۶
CLEVR	۱۰۰۰۰۰	۸۵۳۵۵۴	۲۰۱۷
Tally-QA	۱۶۵۰۰۰	۳۰۶۹۰۷	۲۰۱۹
KVQA	۲۴۶۰۲	۱۸۳۰۰۷	۲۰۱۹

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

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.

۱ـ۳ بررسی چالشهای موجود در این مسئله

باید تکمیل شود.

۱ـ۴ بررسی مجموعه دادگان مطرح و مسابقات 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

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 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 فلان 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.

۱ـ۴ـ۲ مجموعه داده VQA

مجموعه‌داده Visual Question Answering v1(VQA v1) 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 دیگری 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 تهیه شده است. برای جمع‌آوری پرسش و پاسخ هم از نیروی انسانی استفاده‌شده‌است. برای هر تصویر حداقل ۳ سوال منحصربه‌فرد وجود دارد و برای هر سوال ۱۰ پاسخ توسط کاربرهای غیرتکراری جمع‌آوری‌شده‌است. این مجموعه‌داده شامل ۶۱۴۱۶۳ سوال به صورت open-ended و چندگزینه‌ای است. در (اشاره به مقاله) بررسی دقیقی در مورد نوع سوالات، طول سوالات و پاسخ‌ها و غیره انجام‌شده‌است.

VQA-abstract به عنوان یک مجموعه‌داده جداگانه و مکمل در کنار VQA-real قرار دارد. هدف از این مجموعه‌داده از بین بردن نیاز به تجزیه و تحلیل تصاویر واقعی است تا مدل‌ها برای پاسخ به سوالات تمرکز خود را بر روی استدلال‌های سطح بالاتری بگذارند. تصاویر کارتونی در این مجموعه‌داده به صورت دستی توسط

انسان‌ها و به وسیله‌ی رابط کاربری که از قبل آماده‌شده‌است؛ ساخته‌شده‌است. تصاویر می‌تواند دو حالت را نشان‌دهند: داخل خانه و خارج از خانه که هر کدام مجموعه متفاوتی از عناصر را شامل می‌شوند از جمله حیوانات، اشیا و انسان‌ها با حالت‌های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجادشده‌است. مشابه تصاویر واقعی 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.

۱ـ۴ـ۳ مجموعه داده Visual Madlibs

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 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

مجموعه‌داده Visual7W نیز بر اساس مجموعه‌داده MS-COCO ساخته‌شده‌است. این مجموعه‌داده شامل ۴۷۳۰۰ تصویر و ۳۲۷۹۳۹ جفت سوال و پاسخ است. این مجموعه‌داده همچنین از ۱۳۱۱۷۵۶ پرسش و پاسخ چندگزینه‌ای تشکیل‌شده‌است که هر سوال ۴ گزینه دارد و تنها یکی از گزینه‌ها پاسخ صحیح سوال است. برای جمع‌آوری سوالات چندگزینه‌ای توسط انسان‌ها از پلتفرم آنلاین Amazon Mechanical Turk استفاده‌شده‌است. نکته‌ی حائز اهمیت در این مجموعه‌داده این است که تمامی اشیایی که در متن پرسش یا

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

۱ـ۴ـ۵ مجموعه داده CLEVR

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 تصوبر قرار گرفته اند؛ ایجاد شده‌است. سوالات در CLEVR به گونه‌ای طراحی‌شده‌است که جنبه‌های مختلف استدلال بصری توسط سیستم‌های VQA را مورد ارزیابی قرار می‌دهد از جمله شناسایی ویژگی، شمارش اشیا، مقایسه، روابط مکانی اشیا و عملیات منطقی. در این مجموعه‌داده مکان تصاویر نیز با استفاده از یک مستطیل مشخص‌شده‌است.

۱ـ۴ـ۶ مجموعه داده Tally-QA

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 و Visual Genome) برداشته‌شده‌است و سوالات پیچیده با استفاده از ۸۰۰ کاربر انسانی از طریق پلتفرم آنلاین Amazon Mechanical Turk جمع‌آوری شده‌است. مجموعه‌داده Tally-QA به سه بخش آموزش و تست_ساده و تست_پیچیده تقسیم می‌شود. بخش تست_ساده تنها شامل

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

۱ـ۴ـ۷ مجموعه داده KVQA

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.

۱ـ۵ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

باید تکمیل شود.

۱ـ۵ـ۱ فاز ۱ : استخراج ویژگی از تصویر و سوا ل

باید تکمیل شود.

۱ـ۵ـ۲ فاز ۲ : درک مشترک تصویر و سوا ل

باید تکمیل شود.

۱ـ۵ـ۳ فاز ۳ : تولید جواب

باید تکمیل شود.

۱ـ۶ 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 بررسی کنیم. 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 جواب درست 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.

۱ـ۶ـ۱ 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.

۱ـ۶ـ۲ معیار شباهت Wu-Palmer

این معیار ارزیابی توسط مالینوفسکی برای پرسش و پاسخ تصویری استفاده 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.

۱ـ۶ـ۳ 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 می‌گبرد و در غیر این صورت هیچ امتیازی کسب نخواهد کرد. از معایب این روش می‌توان به هزینه زیاد جمع‌آوری پاسخ برای سوالات اشاره کرد.

۱ـ۷ چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی

باید تکمیل شود.