



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

فهرست مطالب

۱	فصل ۱: مقدمه
۱-۱	شرح مسئله
۲-۱	کاربرد و اهمیت مسئله
۳-۱	بررسی چالشهای موجود در این مسئله
۴-۱	بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه
۴-۱	مجموعه داده DAQUAR
۴-۲	مجموعه داده VQA
۴-۳	مجموعه داده Visual Madlibs
۴-۴	مجموعه داده Visual7w
۴-۵	مجموعه داده CLEVR
۴-۶	مجموعه داده Tally-QA
۴-۷	مجموعه داده KVQA
۵-۱	بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری
۵-۱	فاز ۱: استخراج ویژگی از تصویر و سوال
۵-۲	فاز ۲: درک مشترک تصویر و سوال
۵-۳	فاز ۳: تولید جواب
۶-۱	معیارهای ارزیابی مسئله پرسش و پاسخ تصویری
۷-۱	چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی

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 سوال متنی، پاسخ صحیح را پیدا کند. 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 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 ۱-۱: مثالی از سیستم پرسش و پاسخ متنی و تصویری

نشانگر سطح بالاتری از انتزاع دنیای واقعی است.

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. علاوه بر این، در سال‌های 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

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.

۱-۳ بررسی چالشهای موجود در این مسئله

باید تکمیل شود.

۱-۴ بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه

باید تکمیل شود.

۱-۴-۱ مجموعه داده DAQUAR

DAQUAR مخفف Dataset for Question Answering on Real World Images است که توسط مالینوفسکی منتشر شده است. این اولین مجموعه داده‌ای است که برای مسئله VQA منتشر شده است. تصاویر از مجموعه داده NYU-Depth V2 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 فلان 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 شده‌است.

۱-۴-۲ مجموعه داده VQA

مجموعه داده (VQA v1) Visual Question Answering یکی از پرکاربردترین مجموعه داده‌ها در زمینه پرسش و پاسخ تصویری است. این مجموعه داده شامل دو بخش است. یک بخش از تصاویر واقعی ساخته شده‌است که VQA-real نام دارد و دیگری با تصاویر کارتونی ساخته شده‌است 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 مجموعه داده MS-COCO تهیه شده‌است. برای جمع‌آوری پرسش و پاسخ هم از نیروی انسانی استفاده شده‌است. برای هر تصویر حداقل ۳ سوال منحصر به فرد وجود دارد و برای هر سوال ۱۰ پاسخ توسط کاربرهای غیر تکراری جمع‌آوری شده‌است. این مجموعه داده 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 انجام شده‌است.

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 شده‌است؛ ساخته شده‌است. 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، اشیاء و انسان‌ها با حالت‌های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجاد شده‌است. مشابه تصاویر واقعی ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع‌آوری شده‌است. 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 داده‌است. 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.

۱-۴-۳ مجموعه داده Visual Madlibs

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 ”فریزی“ می توان جاهای خالی را پر کرد. این مجموعه داده شامل ۱۰۷۳۸ تصویر از مجموعه داده MS-COCO 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

مجموعه داده Visual7W نیز بر اساس مجموعه داده MS-COCO 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 ای حائز اهمیت در این مجموعه داده این است که تمامی اشیا یا پرسش یا پاسخ ذکر شده است، به نحوی به کادر محدود کننده آن شی در تصویر مرتبط شده است. مزیت این روش، رفع ابهام های موجود در متن است. همان طور که از نام این مجموعه داده پیداست؛ سوالات آن با ۷ کلمه ی پرسشی که حرف اول آن w است شروع می شود. این ۷ کلمه شامل what ، where ، when ، who ، why ، how و which است. پرسش های Visual7W نسبت به به مجموعه داده VQA v1 غنی تر و سخت تر است. همچنین پاسخ ها طولانی تر هستند

۱-۴-۵ مجموعه داده CLEVR

CLEVR یک مجموعه داده برای ارزیابی درک بصری سیستم های 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 اشیا در تصویر قرار گرفته اند؛ ایجاد شده است. سوالات در CLEVR به گونه ای طراحی شده است که جنبه های مختلف استدلال بصری توسط سیستم های VQA را مورد ارزیابی قرار می دهد از جمله شناسایی ویژگی، شمارش اشیا، مقایسه، روابط مکانی اشیا و عملیات منطقی. در این مجموعه داده مکان تصاویر نیز با استفاده از یک مستطیل مشخص شده است.

۱-۴-۶ مجموعه داده Tally-QA

در سال ۲۰۱۹، مجموعه داده Tally-QA منتشر شد که بزرگ ترین مجموعه داده پرسش و پاسخ تصویری برای شمارش اشیا است. اکثر مجموعه داده های شمارش اشیا در پرسش و پاسخ تصویری دارای سوالات ساده هستند که برای پاسخ دادن به این سوال ها تنها کافی است که اشیا در تصویر تشخیص داده شوند. بنابراین، این موضوع باعث ایجاد مجموعه داده ای Tally-QA شد که علاوه بر سوالات ساده، سوالات پیچیده را نیز در بر می گیرد که برای پاسخ دادن به آن ها 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 مجموعه داده از مجموعه داده های دیگری (VQA v2 و Visual Genome) برداشته شده است و سوالات پیچیده با استفاده از ۸۰۰ کاربر انسانی از طریق پلتفرم آنلاین Amazon Mechanical Turk جمع آوری شده است. مجموعه داده Tally-QA به سه بخش آموزش و تست - ساده و تست - پیچیده تقسیم می شود. بخش تست - ساده تنها شامل سوالات ساده و بخش تست - پیچیده تنها دارای سوالات پیچیده ای است که از Amazon Mechanical Turk جمع آوری شده است.

۱-۴-۷ مجموعه داده KVQA

برای رسیدن به پاسخ سوالات موجود در این مجموعه داده، نیاز به استدلال های چند جانبه است. این مجموعه داده شامل ۲۴۰۰۰ تصویر و ۱۸۳۱۰۰ پرسش و پاسخ است.

۱-۵ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

باید تکمیل شود.

۱-۵-۱ فاز ۱: استخراج ویژگی از تصویر و سوال

باید تکمیل شود.

۱-۵-۲ فاز ۲: درک مشترک تصویر و سوال

باید تکمیل شود.

۱-۵-۳ فاز ۳: تولید جواب

باید تکمیل شود.

۱-۶ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری

باید تکمیل شود.

۱-۷ چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی

باید تکمیل شود.