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

آبان ۱۳۹۹


بسم الله الرحمن الرحيم

فهرست مطالب

فصل ۱

۱ـ۱ بررسی کلی و تعریف مبحث پرسش و پاسخ تصویری

باید تکمیل شود.

۱ـ۲ کاربرد و اهمیت این مسئله

باید تکمیل شود.

۱ـ۳ بررسی چالشهای موجود در این مسئله

باید تکمیل شود.

۱ـ۴ بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه

باید تکمیل شود.

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

باید تکمیل شود.

۱ـ۴ـ۶ مجموعه داده Tally-QA

باید تکمیل شود.

۱ـ۴ـ۷ مجموعه داده KVQA

باید تکمیل شود.

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

باید تکمیل شود.

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

باید تکمیل شود.