



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



فهرست مطالب

| | |
|-------|--|
| ۱ | فصل ۱: |
| ۱-۱ | بررسی کلی و تعریف مبحث پرسش و پاسخ تصویری |
| ۱-۲ | کاربرد و اهمیت این مسئله |
| ۱-۳ | بررسی چالشهای موجود در این مسئله |
| ۱-۴ | بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه |
| ۱-۴-۱ | مجموعه داده DAQUAR |
| ۱-۴-۲ | مجموعه داده VQA |
| ۱-۴-۳ | مجموعه داده Visual Madlibs |
| ۱-۴-۴ | مجموعه داده Visual 7w |
| ۱-۴-۵ | مجموعه داده CLEVR |
| ۱-۴-۶ | مجموعه داده Tally-QA |
| ۱-۴-۷ | مجموعه داده KVQA |
| ۱-۵ | بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری |
| ۱-۵-۱ | فاز ۱: استخراج ویژگی از تصویر و سوال |
| ۱-۵-۲ | فاز ۲: درک مشترک تصویر و سوال |
| ۱-۵-۳ | فاز ۳: تولید جواب |
| ۱-۶ | معیارهای ارزیابی مسئله پرسش و پاسخ تصویری |
| ۱-۷ | چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی |

فصل ۱

۱-۱ بررسی کلی و تعریف مبحث پرسش و پاسخ تصویری

باید تکمیل شود.

۲-۱ کاربرد و اهمیت این مسئله

باید تکمیل شود.

۳-۱ بررسی چالشهای موجود در این مسئله

باید تکمیل شود.

۴-۱ 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 مالدینوفسکی منتشر شده است. این اولین مجموعه داده ای 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 پاسخ ها از دو روش مصنوعی و انسانی استفاده شده است. در روش مصنوعی پرسش و پاسخ ها 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 پاسخ های آموزشی در این مجموعه داده ۶۷۹۴ و تعداد پرسش و پاسخ های تست ۵۶۴ است و به طور میانگین برای هر عکس تقریباً ۹ پرسش و پاسخ وجود دارد. این مجموعه داده با مشکل بایاس روبه رو 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 پاسخ ها تکرار شده است.

۱-۴-۲ مجموعه داده VQA

مجموعه داده Visual Question Answering v1 (VQA v1) 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 دیگری با تصاویر کارتوننی 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 تهیه شده است. برای جمع آوری پرسش و پاسخ هم 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 غیر تکراری جمع آوری شده است. این مجموعه داده شامل ۶۱۴۱۶۳ سوال به صورت open-ended و چندگزینه ای است. در (اشاره به مقاله) بررسی دقیقی در مورد نوع سوالات، طول سوالات و پاسخ ها و غیره انجام شده است. 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 مدل ها برای پاسخ به سوالات تمرکز خود را بر روی استدلال های سطح بالاتری بگذارند. تصاویر کارتوننی در این مجموعه داده به صورت دستی توسط انسان ها و به وسیله ی رابط کاربری که از قبل آماده شده است؛ ساخته شده است. تصاویر می تواند دو حالت را نشان دهند: داخل خانه و خارج از خانه که هر کدام مجموعه متفاوتی از عناصر را شامل می شوند از جمله حیوانات، اشیاء و انسان ها با حالت های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجاد شده است. مشابه تصاویر واقعی ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع آوری شده است.

مجموعه داده VQA v2 (Visual Question Answering 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.

۱-۴-۳ مجموعه داده Visual Madlibs

مجموعه داده Visual Madlibs شکل متفاوتی از پرسش و پاسخ را ارائه می‌دهد. برای هر تصویر جملاتی در نظر گرفته شده است و یک کلمه از آن که معمولاً مربوط به آدم، اشیا و فعالیت‌های نمایش داده شده در تصویر است؛ از جمله حذف شده و به جای آن جای خالی قرار گرفته است. پاسخ‌ها کلماتی هستند که این جملات را تکمیل می‌کنند. برای مثال جمله ”دو [جای خالی] در پارک [جای خالی] بازی می‌کنند.“ در وصف یک تصویر بیان شده است که با دو کلمه ”مرد“ و ”فریزبی“ می‌توان جاهای خالی را پرکرد. 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 شده‌اند. پاسخ‌ها در این مجموعه داده به هر دو شکل open-ended و چندگزینه‌ای است.

۱-۴-۴ مجموعه داده Visual 7w

باید تکمیل شود.

۱-۴-۵ مجموعه داده CLEVR

باید تکمیل شود.

۱-۴-۶ مجموعه داده Tally-QA

باید تکمیل شود.

۱-۴-۷ مجموعه داده KVQA

باید تکمیل شود.

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.