



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	۱-۱ شرح مسئله
۲-۱	۲-۱ کاربرد و اهمیت مسئله
۳-۱	۳-۱ بررسی چالشهای موجود در این مسئله
۴-۱	۴-۱ بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه
۴-۱	۴-۱-۱ مجموعه داده DAQUAR
۴-۱	۴-۱-۲ مجموعه داده VQA
۴-۱	۴-۱-۳ مجموعه داده Visual Madlibs
۴-۱	۴-۱-۴ مجموعه داده Visual7w
۴-۱	۴-۱-۵ مجموعه داده CLEVR
۴-۱	۴-۱-۶ مجموعه داده Tally-QA
۴-۱	۴-۱-۷ مجموعه داده KVQA
۵-۱	۵-۱ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری
۵-۱	۵-۱-۱ فاز ۱: استخراج ویژگی از تصویر و سوال
۵-۱	۵-۱-۲ فاز ۲: درک مشترک تصویر و سوال
۵-۱	۵-۱-۳ فاز ۳: تولید جواب
۶-۱	۶-۱ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری
۶-۱	۶-۱-۱ معیار دقت
۶-۱	۶-۱-۲ معیار شباهت Wu-Palmer

۱۴	۱-۶-۳ معیار اجماع
۱۴	۱-۶-۴ MPT
۱۴	۱-۶-۵ BLEU
۱۵	۱-۶-۶ METEOR
۱۵	۱-۷ 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

مجموعه داده	تعداد تصاویر	تعداد سوالات	سال انتشار
DAQUAR	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
VQA v2	.	.	.
Visual Madlibs	۱۰۷۳۸	۳۶۰۰۱	۲۰۱۵
Visual7w	۴۷۳۰۰	۲۲۰۱۱۵۴	۲۰۱۶
CLEVR	۱۰۰۰۰۰	۸۵۳۵۵۴	۲۰۱۷
Tally-QA	۱۶۵۰۰۰	۳۰۶۹۰۷	۲۰۱۹
KVQA	۲۴۶۰۲	۱۸۳۰۰۷	۲۰۱۹

جدول ۱-۱: بررسی اجمالی مجموعه داده‌های 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.

۱-۳ بررسی چالشهای موجود در این مسئله

باید تکمیل شود.

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

در این بخش به معرفی مجموعه داده‌های مشهور در حوزه پرسش و پاسخ تصویری می‌پردازیم و ویژگی‌های هر کدام را بررسی خواهیم کرد. در جدول ۱-۱ اطلاعات آماری این مجموعه داده‌ها به صورت خلاصه آمده است.

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 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 شده است.

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 شده است که VQA-real نام دارد و دیگری با تصاویر کارتونی ساخته شده است 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 شده است. برای جمع‌آوری 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 جمع‌آوری شده است. 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 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 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 شده است؛ ساخته شده است. تصاویر می‌تواند دو حالت را نشان دهند: 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 شده است. مشابه تصاویر واقعی ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع‌آوری 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، اشیاء 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 ۳۶۰۰۱ 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 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 چندگزینه‌ای تشکیل شده است 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 استفاده شده است. نکته‌ی حائز اهمیت در این مجموعه داده این است که تمامی اشیایی که در متن پرسش یا

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

۵-۴-۱ مجموعه داده CLEVR

CLEVR یک مجموعه داده برای ارزیابی درک بصری سیستم‌های 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 که اشیا در تصویر قرار گرفته اند؛ ایجاد شده است. سوالات در CLEVR به گونه‌ای طراحی شده است 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 مجموعه داده مکان تصاویر نیز با استفاده از یک مستطیل مشخص شده است.

۶-۴-۱ مجموعه داده Tally-QA

در سال ۲۰۱۹، مجموعه داده Tally-QA منتشر شد که بزرگ‌ترین مجموعه داده پرسش و پاسخ تصویری برای شمارش اشیا است. اکثر مجموعه داده‌های شمارش اشیا در پرسش و پاسخ تصویری دارای سوالات ساده هستند که برای پاسخ دادن به این سوال‌ها تنها کافی است که اشیا در تصویر تشخیص داده شوند. بنابراین، این موضوع باعث ایجاد مجموعه داده‌ی 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 مجموعه داده از مجموعه داده‌های 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 جمع‌آوری شده است. مجموعه داده 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.

۱-۴-۷ 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 مجموعه داده 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.

۱-۵-۵ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

بسیاری از محققان راه‌حل‌ها یا الگوریتم‌هایی را برای حل مسئله پرسش و پاسخ تصویری پیشنهاد کرده‌اند 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.

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

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

مدل CNN	سال	تعداد لایه‌ها	ابعاد ورودی	ابعاد خروجی (تعداد ویژگی‌ها)
AlexNet	۲۰۱۲	۸	۲۲۷×۲۲۷	۴۰۹۶
VGGNet	۲۰۱۴	۱۹	۲۲۴×۲۲۴	۴۰۹۶
GoogleNet	۲۰۱۴	۲۲	۲۲۹×۲۲۹	۱۰۲۴
ResNet	۲۰۱۵	۱۵۲	۲۲۴×۲۲۴	۲۰۱۴۸

جدول ۱-۲: بررسی اجمالی مهم‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده.

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 قیمت و مجموعه داده‌های 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 ۱-۲ چند نمونه از برجسته‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده‌اند؛ آورده شده است. بیشتر مدل‌های ارائه شده در پرسش و پاسخ تصویری از این شبکه‌های عصبی کانولوشنی استفاده می‌کنند تا محتوای تصویری خود را به برداری‌هایی عددی تبدیل کنند. جدول ۱-۳ 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 ۱-۲ 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 مجموعه داده‌هایی 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 ۱-۳ 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 و ساده نیستند

ResNet	GoogleNet	VGGNet	AlexNet	مدل پرسش و پاسخ تصویری
		✓		Image_QA
	✓			Talk_to_Machine
		✓		VQA
			✓	Vis_Madlibs
		✓		VIS + LSTM
		✓		Ahab
		✓		ABCCNN
		✓		Comp_QA
		✓		DPPNet
		✓		Answer_CNN
		✓		VQACaption
✓				Re_Baseline
✓				MCB
	✓			SMem-VQA
		✓		Region_VQA
		✓		Vis7W
✓	✓	✓	✓	Ask_Neuron
✓				SCMC
✓				HAN
		✓		StrSem
✓				AVQAN
✓				CMF
✓				EnsAtt
✓				MetaVQA
✓				DA-NTN
✓				QGHC
✓				QTA
✓				WRAN
✓				QAR

جدول ۱-۳: شبکه‌های عصبی کانولوشنی استفاده شده در مدل‌های پرسش و پاسخ تصویری.

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 اداره 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 واژه‌های منحصر به فرد موجود در مجموعه داده ساخته می‌شود 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^۱ و skip-gram 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 و لایه خروجی است، استفاده شده است. در مدل 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 لایه 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 ۱-۵ آورده شده است.

^۱ Continouse Bag Of Words

یکی دیگر از word embedding های مشهور، مدل بردار سراسری یا به اختصار Glove^۲ است که توسط پنینگتون و همکاران در سال ۲۰۱۴ در تیم پردازش زبانهای طبیعی دانشگاه استنفورد معرفی و توسعه داده شد. ادامه اش باید تکمیل بشه ...

با پیشرفت یادگیری عمیق در دهه اخیر، محققان برای استخراج ویژگی و بازنمایی متن از LSTM، CNN و GRU 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 می شود سپس به لایه های کانولوشنی یک بعدی داده می شود و فیلترهای متفاوتی بر روی آنها اعمال می شود و پس از عبور از لایه max-pooling ویژگی ها بدست می آید.

توضیح LSTM ...

توضیح GRU ...

مدل های مختلف در مسئله پرسش و پاسخ تصویری از word embedding های ذکر شده در بالا برای تولید بردار ویژگی برای سوال ها استفاده کرده اند. جدول ۱-۴ لیستی از مدل های پرسش و پاسخ تصویری به همراه word embedding شده در آنها را نمایش می دهد. با بررسی جدول ۱-۴ مشاهده می کنیم که محققان حوزه پرسش و پاسخ تصویری ترجیح می دهند؛ برای استخراج ویژگی از متن و بازنمایی آن از LSTM استفاده کنند. آنها معتقد هستند که RNN ها عملکرد بهتری نسبت به روش های مستقل از دنباله ای کلمات مانند word2vec دارند. اما آموزش RNN ها نیاز به داده های برجسب خورده ی زیادی دارد.

۱-۵-۲ فاز ۲: درک مشترک تصویر و سوال

باید تکمیل شود.

۱-۵-۳ فاز ۳: تولید جواب

باید تکمیل شود.

^۲Global Vector

GRU	LSTM	CNN	GloVe	Skip-gram/word2vec	CBOW	one-hot	مدل پرسش و پاسخ تصویری
				✓			Image_QA
	✓						Talk_to_Machine
					✓		VQA
				✓			Vis_Madlibs
	✓						VIS + LSTM
	✓						ABCCNN
	✓						Comp_QA
✓							DPPNet
		✓					Answer_CNN
	✓						VQA-Caption
				✓			Re_Baseline
	✓						MCB
					✓		SMem-VQA
				✓			Region_VQA
						✓	Vis7W
✓	✓	✓			✓		Ask_Neuron
		✓					SCMC
	✓						HAN
	✓						StrSem
						✓	AVQAN
	✓		✓				CMF
			✓				EnsAtt
✓			✓				MetaVQA
✓							DA-NTN
✓							QGHC
✓							WRAN
			✓				QAR

جدول ۱-۴: word embedding های استفاده شده در مدل های پرسش و پاسخ تصویری.

۱-۶ 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 قبلاً ذکر شد؛ معمولاً دو نوع سوال در مجموعه داده‌های پرسش و پاسخ تصویری در نظر گرفته می‌شود: سوالات open-ended و سوالات چندگزینه‌ای. در سوالات چندگزینه‌ای، برای هر سوال دقیقاً 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 مجموعه داده‌های پرسش و پاسخ تصویری پاسخ‌ها را محدود به چند کلمه (۱ تا ۳ 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 مجموعه داده و نحوه ساخت آن، میزان بایاس موجود در مجموعه داده و ... توجه نمود.

۱-۶-۱ 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 کاملاً 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 شده است.

۱-۶-۲ معیار شباهت Wu-Palmer

این معیار ارزیابی توسط مالینوفسکی 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 اندازه‌گیری کند. 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.

۱-۶-۳ 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 معمولاً از حالت کمترین اجماع استفاده می‌شود و آستانه را هم برابر ۳ قرار می‌دهند به این معنی که اگر پاسخ پیش‌بینی شده با ۳ یا بیشتر از ۳ پاسخ برابر باشد امتیاز کامل می‌گردد و در غیر این صورت هیچ امتیازی کسب نخواهد کرد. از معایب این روش می‌توان به هزینه زیاد جمع‌آوری پاسخ برای سوالات اشاره کرد.

۱-۶-۴ MPT

یکی از مشکلات مجموعه داده‌های پرسش و پاسخ تصویری توزیع غیریکنواخت انواع سوال‌هاست. در این مواقع، نمی‌توان از معیار دقت استفاده کرد. بنابراین در ... معیار جدیدی به نام MPT^۳ ارائه شده است که توزیع نامتوازن سوال‌ها را جبران می‌کند. معیار MPT میانگین دقت برای هر نوع سوال را محاسبه می‌کند. از نسخه‌ی نرمالایز شده‌ی این معیار نیز برای رفع مشکل بایاس در توزیع پاسخ‌ها استفاده می‌شود.

۱-۶-۵ BLEU

BLEU^۴ یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. در ... 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

^۳Mean Per Type
^۴BiLingual Evaluation Understudy

پاسخ پیش‌بینی شده و پاسخ صحیح را اندازه‌گیری می‌کند. معمولاً BLEU زمانی که جمله‌ها کوتاه باشند، با شکست مواجه می‌شود.

۱-۶-۶ METEOR

METEOR^۵ نیز همانند BLEU یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. به پیشنهاد فلان 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.

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

باید تکمیل شود.

^۵Metric for Evaluation of Translation with Explicit Ordering