



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

چکیده

چکیده

واژگان کلیدی: واژگان کلیدی

# فهرست مطالب

ج	فهرست تصاویر
چ	فهرست جداول
۱	فصل ۱: مقدمه
۲	۱-۱ کاربرد و اهمیت مسئله
۳	۱-۲ بررسی چالش‌های موجود در این مسئله
۴	فصل ۲: تعاریف و مفاهیم مبنایی
۵	فصل ۳: مروری بر کارهای مرتبط
۵	۳-۱ بررسی مجموعه داده‌گان مطرح این حوزه
۵	۳-۱-۱ مجموعه داده DAQUAR
۷	۳-۱-۲ مجموعه داده VQA
۸	۳-۱-۳ مجموعه داده Visual Madlibs
۹	۳-۱-۴ مجموعه داده Visual7w
۱۰	۳-۱-۵ مجموعه داده CLEVR
۱۰	۳-۱-۶ مجموعه داده Tally-QA
۱۲	۳-۱-۷ مجموعه داده KVQA
۱۳	۳-۲ تقویت مجموعه داده در مسئله پرسش و پاسخ تصویری
۱۵	۳-۳ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

۱۵	۳-۳-۱ فاز ۱ : استخراج ویژگی از تصویر و سوال
۱۹	۳-۳-۲ فاز ۲ : بازنمایی مشترک تصویر و سوال
۱۹	۳-۳-۲-۱ روش های پایه
۲۱	۳-۳-۲-۲ روش های مبتنی بر شبکه های عصبی
۲۱	۳-۳-۲-۳ روش های مبتنی بر توجه
۲۱	۳-۳-۳ فاز ۳ : تولید جواب
۲۱	۳-۴ مدل های از قبل آموزش دیده بر روی زبان طبیعی و تصویر
۲۲	۳-۴-۱ معماری تک جریان
۲۷	۳-۴-۲ معماری دو جریان
۳۱	۳-۵ معیارهای ارزیابی مسئله پرسش و پاسخ تصویری
۳۱	۳-۵-۱ معیار دقت
۳۲	۳-۵-۲ معیار شباهت Wu-Palmer
۳۲	۳-۵-۳ معیار اجماع
۳۲	۳-۵-۴ MPT
۳۳	۳-۵-۵ BLEU
۳۳	۳-۵-۶ METEOR

## فصل ۴: نتیجه گیری و کارهای آینده

۳۴	۴-۱ نتیجه گیری
۳۴	۴-۲ مسائل باز و کارهای قابل انجام

## مراجع

۴۲	واژه نامه فارسی به انگلیسی
۴۳	واژه نامه انگلیسی به فارسی

## فهرست تصاویر

۱-۱	مثالی از سیستم پرسش و پاسخ متنی و تصویری	۲
۱-۳	چند نمونه از مجموعه داده DAQUAR	۷
۲-۳	چند نمونه از مجموعه داده VQA v1 - real	۸
۳-۳	چند نمونه از مجموعه داده VQA v1 - abstract	۸
۴-۳	چند نمونه از مجموعه داده VQA v2	۹
۵-۳	یک نمونه از مجموعه داده Visual Madlibs	۱۰
۶-۳	چند نمونه از مجموعه داده Visual7W	۱۱
۷-۳	چند نمونه از مجموعه داده CLEVR	۱۲
۸-۳	چند نمونه از مجموعه داده Tally-QA	۱۲
۹-۳	چند نمونه از مجموعه داده KVQA	۱۳
۱۰-۳	۱۰ معماری شبکه CBOW و Skip-gram	۱۸
۱۱-۳	۱۱ معماری شبکه از قبل آموزش دیده VL-BERT	۲۳
۱۲-۳	۱۲ نحوه ورودی و خروجی شبکه VL-BERT برای آموزش در مسئله پرسش و پاسخ تصویری	۲۴
۱۳-۳	۱۳ معماری شبکه از قبل آموزش دیده UNITER	۲۴
۱۴-۳	۱۴ معماری شبکه از قبل آموزش دیده VLP	۲۵
۱۵-۳	۱۵ معماری شبکه از قبل آموزش دیده OSCAR	۲۷
۱۶-۳	۱۶ معماری شبکه از قبل آموزش دیده ViLBERT	۲۷
۱۷-۳	۱۷ ساختار لایه co-attentional transformer	۲۸
۱۸-۳	۱۸ معماری شبکه از قبل آموزش دیده LXMERT	۲۹

## فهرست جداول

۳-۱	بررسی اجمالی مجموعه داده های معروف در حوزه پرسش و پاسخ تصویری.	۶
۳-۲	الگوهای استفاده شده برای تولید سوال در مجموعه داده DAQUAR.	۶
۳-۳	بررسی اجمالی مهم ترین شبکه های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده اند.	۱۶
۳-۴	شبکه های عصبی کانولوشنی استفاده شده در مدل های پرسش و پاسخ تصویری.	۱۷
۳-۵	word embedding های استفاده شده در مدل های پرسش و پاسخ تصویری.	۲۰
۳-۶		۳۰
۳-۷	دقت شبکه های از قبل آموزش دیده بر روی مجموعه داده VQA v2.0 (test-std)	۳۰

# فصل ۱

## مقدمه

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 سؤال متنی، پاسخ صحیح را پیدا کند [۳۹]. 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 سؤال متنی به عنوان ورودی به سیستم داده می‌شود و انتظار می‌رود که سیستم با توجه به درک و تفسیری که از متن و سؤال بدست می‌آورد؛ یک جواب متنی را خروجی دهد. اما در سیستم پرسش و پاسخ تصویری، یک تصویر و یک سؤال متنی به ورودی سیستم داده می‌شود و انتظار می‌رود که سیستم بتواند با استفاده از عناصر بصری تصویر و تفسیری که از سؤال 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 نشانگر سطح بالاتری از انتزاع دنیای واقعی است [۶۵].





شکل ۱-۱: مثالی از سیستم پرسش و پاسخ متنی و تصویری

## ۱-۱ 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 [۱۸]. علاوه بر این، در سال‌های اخیر دستیاران صوتی<sup>۱</sup> و عامل‌های گفتگو<sup>۲</sup> 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

<sup>۱</sup> Voice Assistants  
<sup>۲</sup> Conversational Agents

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 [۵۸].

## ۱-۲. بررسی چالش‌های موجود در این مسئله

در مقایسه با مسائل دیگری که مشترک بین پردازش زبان طبیعی و بینایی ماشین است مانند توصیف تصویر<sup>۳</sup> و بازیابی متن به تصویر<sup>۴</sup>، مسئله پرسش و پاسخ تصویری چالش‌برانگیزتر است زیرا (۱) سوالات از پیش تعیین نشده است. به این معنی که در مسئله‌ای مانند تشخیص اشیا، سوال این است که چه اشیایی در تصویر وجود دارد و این سوال از پیش تعیین شده است و در طول حل مسئله تغییر نمی‌کند و تنها تصویر تغییر می‌کند که منجر به پاسخ‌ها متفاوت می‌شود. اما در پرسش و پاسخ تصویری، برای هر تصویر سوالات متفاوت و مرتبط با همان تصویر پرسیده می‌شود که در زمان اجرا تعیین می‌شود. (۲) اطلاعات موجود در تصویر ابعاد بالایی دارد که پردازش آن‌ها به زمان و حافظه زیادی نیاز دارد. (۳) مسئله پرسش و پاسخ تصویری نیاز به حل مسائل پایه‌ای و فرعی دارد مانند تشخیص اشیا<sup>۵</sup> (آیا در تصویر سگ وجود دارد؟)، تشخیص فعالیت<sup>۶</sup> (آیا کودک گریه می‌کند؟)، طبقه‌بندی صفات<sup>۷</sup> (چتر چه رنگی است؟)، شمارش (چند نفر در تصویر وجود دارد؟)، طبقه‌بندی صحنه<sup>۸</sup> (هوا بارانی است؟) و روابط مکانی بین اشیا (چه چیزی بین گربه و مبل است؟).

<sup>۳</sup> Image Captioning

<sup>۴</sup> Text-to-image Retrieval

<sup>۵</sup> Object Detection

<sup>۶</sup> Activity Recognition

<sup>۷</sup> Attribute Classification

<sup>۸</sup> Scene Classification

## فصل ۲

### تعاریف و مفاهیم مبنایی

## فصل ۳

### مروری بر کارهای مرتبط

#### ۳-۱ بررسی مجموعه دادگان 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 ۳-۱ اطلاعات آماری این مجموعه داده‌ها به صورت خلاصه آمده است.

#### ۳-۱-۱ مجموعه داده DAQUAR [۳۷]

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 [۵۳] 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 ۳-۲ 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

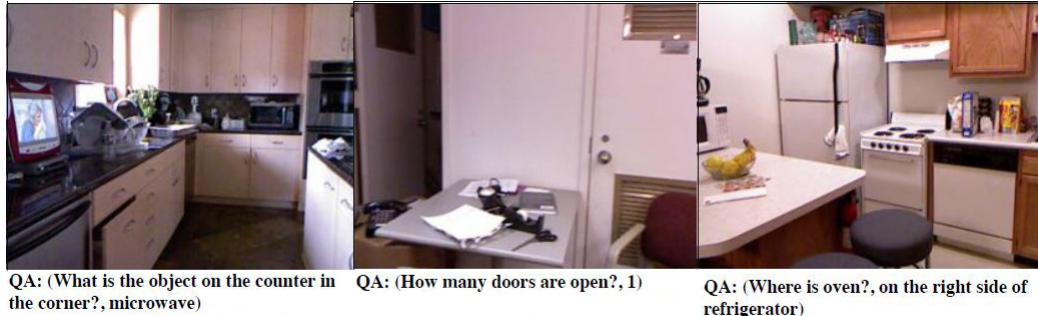
جدول ۳-۱: بررسی اجمالی مجموعه داده‌های معروف در حوزه پرسش و پاسخ تصویری.

مجموعه داده	تعداد تصاویر	تعداد سوالات	سال انتشار
DAQUAR [۳۷]	۱۴۴۹	۱۲۴۶۸	۲۰۱۴
VQA v1 [۴]	۲۰۴۷۲۱	۶۱۴۱۶۳	۲۰۱۵
Visual Madlibs [۶۸]	۱۰۷۳۸	۳۶۰۰۰۱	۲۰۱۵
Visual7w [۷۱]	۴۷۳۰۰	۲۲۰۱۱۵۴	۲۰۱۶
VQA v2 [۱۷]	۲۰۴۷۲۱	۱۱۰۵۹۰۴	۲۰۱۷
CLEVR [۲۳]	۱۰۰۰۰۰	۸۵۳۵۵۴	۲۰۱۷
Tally-QA [۱]	۱۶۵۰۰۰	۳۰۶۹۰۷	۲۰۱۹
KVQA [۵۰]	۲۴۶۰۲	۱۸۳۰۰۷	۲۰۱۹

جدول ۳-۲: الگوهای استفاده شده برای تولید سوال در مجموعه داده DAQUAR. سوالات می‌تواند در مورد یک تصویر و یا مجموعه‌ای از تصاویر باشد [۳۷].

نمونه	الگو	توضیح	
How many cabinets are in image1?	How many {object} are in {image id}?	شماری	منفرد
How many gray cabinets are in image1?	How many {color} {object} are in {image id}?	شماری و رنگ	منفرد
Which type of the room is depicted in image1?	Which type of the room is depicted in {image id}?	نوع اتاق	منفرد
What is the largest object in image1?	What is the largest {object} in {image id}?	صفات عالی	منفرد
How many black bags?	How many {color} {object}?	شماری و رنگ	مجموعه‌ای
Which images do not have sofa?	Which images do not have {object}?	نفی نوع ۱	مجموعه‌ای
Which images are not bedroom?	Which images are not {room type}?	نفی نوع ۲	مجموعه‌ای
Which images have desk but do not have a lamp?	Which images have {object} but do not have a {object}?	نفی نوع ۳	مجموعه‌ای

۴۰۰ مورد وجود دارد که اشیایی مثل میز و صندلی در پاسخها تکرار شده است.



شکل ۳-۱: چند نمونه از مجموعه داده DAQUAR [۳۷]

### ۳-۱-۲ مجموعه داده VQA [۴] [۱۷]

مجموعه داده Visual Question Answering v1 (VQA v1)<sup>۱</sup> 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 بخش دیگر 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 [۳۱] تهیه شده است. برای جمع‌آوری پرسش و پاسخ 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 منحصر به فرد جمع‌آوری شده است. این مجموعه داده شامل ۶۱۴۱۶۳ سوال به صورت open-ended و چندگزینه‌ای است. در [۴] 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

<sup>۱</sup><https://visualqa.org/>

حیوانات، اشیاء و انسان‌ها با حالت‌های مختلف. در مجموع ۵۰۰۰۰ تصویر ایجاد شده است. مشابه VQA- real، ۳ سوال برای هر تصویر (یعنی در کل ۱۵۰۰۰۰ سوال) و برای هر سوال ۱۰ پاسخ جمع‌آوری شده است. 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 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.



Q: What shape is the bench seat ?

A: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved



Q: What color is the stripe on the train ?

A: white, white, white, white, white, white, white, white, white, white



Q: Where are the magazines in this picture ?

A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

شکل ۳-۲: چند نمونه از مجموعه داده real - VQA v1 [۴]



Q: Who looks happier ?

A: old person, man, man, man, old man, man, man, man, man, grandpa



Q: Where are the flowers ?

A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree



Q: How many pillows ?

A: 1, 2, 2, 2, 2, 2, 2, 2, 2, 2

شکل ۳-۳: چند نمونه از مجموعه داده abstarct - VQA v1 [۴]

### ۳-۱-۳ مجموعه داده Visual Madlibs [۶۸]

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



شکل ۳-۴: چند نمونه از مجموعه داده VQA v2 [۱۷]

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 [۳۱] 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.

### ۳-۱-۴ مجموعه داده Visual7w [۷۱]

مجموعه داده Visual7W نیز بر اساس مجموعه داده MS-COCO [۳۱] 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 محدودکننده‌ی 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 what ، where ، when ، who ، why ، how و which است. پرسش‌های Visual7W نسبت به مجموعه داده VQA v1 غنی‌تر و سخت‌تر است. همچنین پاسخ‌ها طولانی‌تر هستند.





1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will throw the frisbee.
6. Person A is wearing blue shorts.
7. Person A is in front of person B.
8. Person A is blocking person B.
9. Person B is a young man wearing an orange hat.
10. Person B is on a grassy field.
11. Person B is holding a frisbee.
12. The frisbee is white and round.
13. The frisbee is in the hand of the man with the orange cap.
14. People could throw the frisbee.
15. The people are playing with the frisbee.

شکل ۳-۵: یک نمونه از مجموعه داده Visual Madlibs [۶۸]

### ۳-۱-۵ مجموعه داده CLEVR [۲۳]

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 اشیاء دو اندازه متفاوت، دو جنس متفاوت و هشت رنگ مختلف در نظر گرفته شده است. سوالات هم به طور مصنوعی بر اساس مکانی که اشیاء 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 اشیاء، مقایسه، روابط مکانی اشیاء و عملیات منطقی. در این مجموعه داده مکان تصاویر نیز با استفاده از یک مستطیل مشخص شده است.

### ۳-۱-۶ مجموعه داده Tally-QA [۱]

در سال ۲۰۱۹، مجموعه داده Tally-QA منتشر شد که بزرگ‌ترین مجموعه داده پرسش و پاسخ تصویری برای شمارش اشیاء است. اکثر مجموعه داده‌های شمارش اشیاء در پرسش و پاسخ تصویری دارای سوالات ساده



Q: What endangered animal is featured on the truck?

A: A bald eagle.  
A: A sparrow.  
A: A hummingbird.  
A: A raven.



Q: Where will the driver go if turning right?

A: Onto 24 ¼ Rd.  
A: Onto 25 ¼ Rd.  
A: Onto 23 ¼ Rd.  
A: Onto Main Street.



Q: When was the picture taken?

A: During a wedding.  
A: During a bar mitzvah.  
A: During a funeral.  
A: During a Sunday church service.



Q: Who is under the umbrella?

A: Two women.  
A: A child.  
A: An old man.  
A: A husband and a wife.



Q: Why was the hand of the woman over the left shoulder of the man?

A: They were together and engaging in affection.  
A: The woman was trying to get the man's attention.  
A: The woman was trying to scare the man.  
A: The woman was holding on to the man for balance.



Q: How many magnets are on the bottom of the fridge?

A: 5.  
A: 2.  
A: 3.  
A: 4.



Q: Which pillow is farther from the window?



Q: Which step leads to the tub?



Q: Which is the small computer in the corner?



Q: Which item is used to cut items?



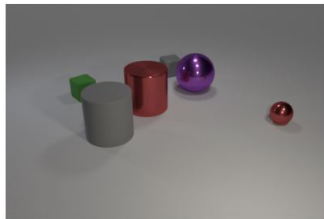
Q: Which doughnut has multicolored sprinkles?



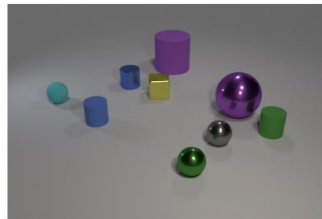
Q: Which man is wearing the red tie?

شکل ۳-۶: چند نمونه از مجموعه داده Visual7W [۷۱]. ردیف اول، پاسخ‌های سبز رنگ، پاسخ صحیح هستند و پاسخ‌های قرمز پاسخ‌های نادرست تولید شده توسط انسان است. ردیف دوم، کادر زرد جواب صحیح است و کادرهای قرمز پاسخ‌های اشتباه انسانی است.

هستند که برای پاسخ دادن به این سوال‌ها تنها کافی است که اشیاء 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 اشیاء 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 [۱۷] و Visual Genome [۲۷]) 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 شده است.



Q: How big is the gray rubber object that is behind the big shiny thing behind the big der; what material is on the left side of the purple ball?  
A: small



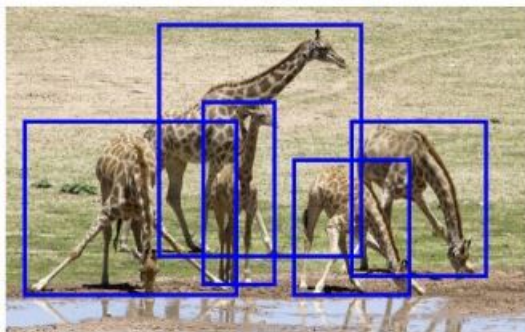
Q: There is a tiny rubber thing that is of the tiny green thing the same color as the metal cylinder; what shape is it?  
A: cylinder



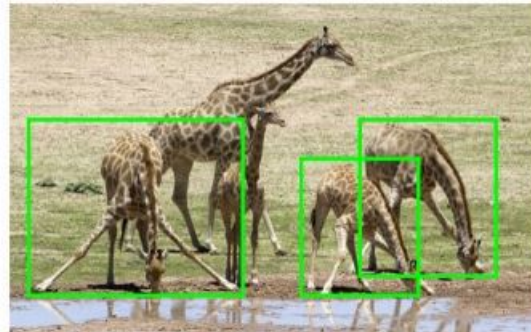
Q: There is a small ball that is made of red rubber sphere the same material as the purple the large block; what color is it?  
A: blue

شکل ۳-۷: چند نمونه از مجموعه داده CLEVR [۲۳].

“How many giraffes?”



“How many giraffes are drinking water?”



شکل ۳-۸: چند نمونه از مجموعه داده Tally-QA [۱]. عکس سمت چپ یک نمونه از سوالات ساده و عکس سمت راست یک نمونه از سوالات پیچیده است.

### ۳-۱-۷ مجموعه داده KVQA [۵۰]

مجموعه داده KVQA که مخفف Knowledge-based Visual Question Answering است در سال ۲۰۱۹ طراحی شده است به طوری که بر خلاف مجموعه داده های 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 ۰.۷، ۰.۲ و ۰.۱ تقسیم شده است. تنوع پرسش

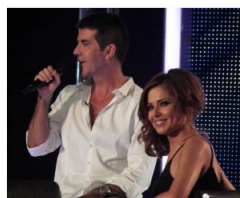
و پاسخ ها در KVQA به گونه ای در نظر گرفته شده است که مشکل همیشگی بایاس در مجموعه داده های پرسش و پاسخ تصویری، در این مجموعه داده وجود نداشته باشد.



(a) **Wikipedia caption:** Khan with United States Secretary of State Hillary Clinton in 2009.

Q: Who is to the left of Hillary Clinton? (*spatial*)  
A: **Aamir Khan**

Q: Do all the people in the image have a common occupation? (*multi-entity, intersection, 1-hop, Boolean*)  
A: **No**



(b) **Wikipedia caption:** Cheryl alongside Simon Cowell on The X Factor, London, June 2010.

Q: What is the age gap between the two people in the image? (*multi-entity, subtraction, 1-hop*)  
A: **24 years**

Q: How many people in this image were born in United Kingdom? (*1-hop, multi-entity, counting*)  
A: **2**



(c) **Wikipedia caption:** BRICS leaders at the G-20 summit in Brisbane, Australia, 15 November 2014

Q: Were all the people in the image born in the same country? (*Boolean, multi-entity, intersection*)  
A: **No**

Q: Who is the founder of the political party to which person second from left belongs to? (*spatial, multi-hop*)  
A: **Syama Prasad Mookerjee**



(d) **Wikipedia caption:** Serena Williams and Venus Williams, Australian Open 2009.

Q: Who among the people in the image is the eldest? (*multi-entity, comparison*)  
A: **Person in the left**

Q: Who among the people in the image were born after the end of World War II? (*multi-entity, multi-relation, comparison*)  
A: **Both**

شکل ۳-۹: چند نمونه از مجموعه داده KVQA [۵۰]

## ۲-۳ تقویت مجموعه داده در مسئله پرسش و پاسخ تصویری

به لطف توسعه سریع شبکه های عصبی عمیق مسئله پرسش و پاسخ تصویری به موفقیت های بزرگی دست یافته است. مطالعات نشان می دهد که عملکرد شبکه های عصبی عمیق به میزان داده های آموزشی بستگی دارد و همیشه از داده های آموزشی بیشتر سود می برند. یکی از ترندهای اصلی در شبکه های عصبی عمیق تقویت داده <sup>۲</sup> است که به طور گسترده در بسیاری از مسائل پردازش تصویر و بینایی ماشین مورد استفاده قرار می گیرد. اما مقالات کمی وجود دارد که مسئله تقویت داده را در پرسش و پاسخ تصویری بررسی کرده اند. یکی از چالش های تقویت داده در مسئله پرسش و پاسخ تصویری این است که هیچ یک از روش های تقویت داده مبتنی بر تصویر مانند چرخش <sup>۳</sup> و ورق زدن <sup>۴</sup> نمی توانند مستقیماً بر روی مسئله پرسش و پاسخ تصویری

<sup>۲</sup> data augmentation  
<sup>۳</sup> rotation  
<sup>۴</sup> flipping

اعمال شود زیرا ساختار معنایی آن حفظ نخواهد شد. به عنوان مثال با چرخش یک تصویر ممکن است پرسش و پاسخ مرتبط با آن (مانند «ماشین در سمت چپ یا راست سطل زباله است؟») دیگر درست نباشد.

در [۲۶] برای اولین بار دو روش برای تقویت داده در مسئله پرسش و پاسخ تصویری پیشنهاد شد. در روش اول برای تولید پرسش و پاسخ از الگو استفاده می شود. برای تولید الگو از حاشیه نویسی<sup>۵</sup> موجود در مجموعه داده استفاده می شود. با استفاده از این روش ۴ نوع سوال تولید می شود: (۱) سوالات بله و خیر (۲) سوالات شمارشی (۳) تشخیص شی، صحنه و یا فعالیت (۴) تشخیص ورزش. برای مثال برای تولید سوالات بله و خیر، با استفاده از حاشیه نویسی موجود در مجموعه داده لیستی از اشیا موجود در تصویر آماده می شود. سپس اگر محدوده مربوط به اشیا بزرگتر از ۲۰۰۰ پیکسل باشد، سوالی مانند «آیا [شی] در تصویر وجود دارد؟» تولید می شود که پاسخ آن هم «بله» است. به همین ترتیب با استفاده از دانشی که از مجموعه داده می توان بدست آورد؛ برای سایر انواع سوالات الگویی برای تولید سوال و پاسخ آن تولید می شود. یکی از مشکلات این روش برای تقویت داده این است که سوالات تولید شده انعطاف پذیر نیستند و ممکن است شباهت زیادی به نحوه ی طرح سوالات موجود در مجموعه داده نداشته باشند. به همین علت، روش دیگری در [۲۶] مبتنی بر LSTM برای تولید سوال برای هر تصویر پیشنهاد شده است. این شبکه از دو لایه LSTM تشکیل شده است که هر کدام دارای ۱۰۰۰ واحد مخفی است و پس از آن ها نیز دو لایه ی کاملاً متصل که هر کدام ۷۰۰۰ نورون مخفی دارند (برابر با تعداد واژگان) ساخته شده است. برای تولید سوال، در ابتدا توکن شروع سوال به همراه ویژگی های تصویر به شبکه داده می شود. برای هر تصویر ۳۰ سوال تولید می شود که تنها سه تا از پرتکرارترین سوالات نگه داشته می شود. برای پیدا کردن جواب سوال های تولید شده توسط شبکه LSTM از یک شبکه ی ساده MLP که در [۲۴] پیشنهاد شده است؛ استفاده شده است. در [۲۶] نشان دادند که استفاده از این دو روش برای تقویت داده ها منجر به بهبود عملکرد روش های موجود برای حل مسئله پرسش و پاسخ تصویری می شود.

اخیرا در [۶۰] برای تقویت داده روشی مبتنی بر تولید نمونه های خصمانه<sup>۶</sup> پیشنهاد شده است که بر خلاف کارهای قبلی، تقویت داده هم برای تصاویر و هم برای سوالات انجام می شود.

<sup>۵</sup> annotation  
<sup>۶</sup> adversarial examples



### ۳-۳ بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری

بسیاری از محققان راه حل ها یا الگوریتم هایی را برای حل مسئله پرسش و پاسخ تصویری پیشنهاد کرده اند که به طور کلی می توان آن را به یک فرآیند سه فازی تقسیم بندی کرد. فاز اول این فرآیند استخراج ویژگی از تصویر و سوالات است که راه حل های موفق در این فاز ریشه در روزهای باشکوه یادگیری عمیق دارد زیرا بیشتر راه حل های موفق در این حوزه از مدل های یادگیری عمیق استفاده می کنند 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 مهم ترین و اصلی ترین فاز می باشد، ویژگی های استخراج شده از تصویر و سوال باهم ترکیب می شوند. سپس از ترکیب ویژگی ها برای تولید پاسخ نهایی در فاز سوم استفاده می شود.

#### ۳-۳-۱ فاز ۱: استخراج ویژگی از تصویر و سوال

استخراج ویژگی از تصویر و سوال مرحله ی مقدماتی در پرسش و پاسخ تصویری است. ویژگی تصویر، تصویر را به عنوان یک بردار عددی توصیف می کند تا بتوان به راحتی عملیات های مختلف ریاضی را بر روی آن اعمال کرد. روش های زیادی وجود دارد که به صورت مستقیم از تصویر ویژگی استخراج می کنند مانند بردار ساده RGB، SIFT، تبدیل HAAR و HOG. اما با ظهور شبکه های یادگیری عمیق، نیاز به استخراج ویژگی به صورت مستقیم از بین رفت زیرا این شبکه ها قادر به یادگیری ویژگی هستند. آموزش مدل های یادگیری عمیق به منابع محاسباتی گران قیمت و مجموعه داده های بزرگ نیاز دارد. از این رو، استفاده از مدل های شبکه عصبی عمیق از قبل آموزش دیده، استخراج ویژگی از تصاویر را به راحتی امکان پذیر می کنند.

یکی از بهترین شبکه های عصبی برای استخراج ویژگی از تصویر، شبکه های عصبی کانولوشنی هستند. در جدول ۳-۳ چند نمونه از برجسته ترین شبکه های عصبی کانولوشنی که بر روی مجموعه داده ImageNet [۱۱] آموزش داده شده اند؛ آورده شده است. بیشتر مدل های ارائه شده در پرسش و پاسخ تصویری از این شبکه های عصبی کانولوشنی استفاده می کنند تا محتوای تصویری خود را به بردارهایی عددی تبدیل کنند. جدول ۳-۴ لیستی از مدل های استفاده شده برای حل مسئله پرسش و پاسخ تصویری را نشان می دهد و مشخص می کند که هر کدام از این مدل ها برای استخراج ویژگی از تصویر از کدام یک از شبکه های عصبی کانولوشنی موجود در جدول ۳-۳ بهره می برد. همان طور که واضح است VGGNet و ResNet به طور گسترده ای در سیستم های پرسش و پاسخ تصویری مورد استفاده قرار گرفته اند. یکی از دلایلی که محققان

جدول ۳-۳: بررسی اجمالی مهم‌ترین شبکه‌های عصبی کانولوشنی که بر روی مجموعه داده ImageNet آموزش داده شده‌اند.

مدل CNN	سال	تعداد لایه‌ها	ابعاد ورودی	ابعاد خروجی (تعداد ویژگی‌ها)
AlexNet [۲۰]	۲۰۱۲	۸	۲۲۷×۲۲۷	۴۰۹۶
VGGNet [۵۴]	۲۰۱۴	۱۹	۲۲۴×۲۲۴	۴۰۹۶
GoogLeNet [۵۷]	۲۰۱۴	۲۲	۲۲۹×۲۲۹	۱۰۲۴
ResNet [۱۹]	۲۰۱۵	۱۵۲	۲۲۴×۲۲۴	۲۰۱۴۸

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 ۳-۴ 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 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 پردازش 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.

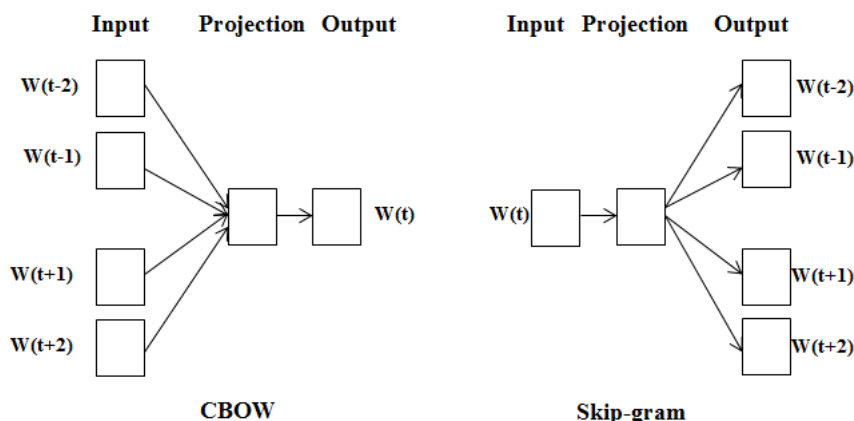
روش‌های word embedding بسیاری در مسئله پرسش و پاسخ تصویری استفاده شده است. در ادامه به برجسته‌ترین و پرکاربردترین روش‌های word embedding موجود و استفاده‌شده در مسئله پرسش و پاسخ تصویری می‌پردازیم و معایب و مزایای هر کدام را بررسی خواهیم کرد.

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.

جدول ۳-۴: شبکه‌های عصبی کانولوشنی استفاده شده در مدل‌های پرسش و پاسخ تصویری.

ResNet	GoogleNet	VGGNet	AlexNet	مدل پرسش و پاسخ تصویری
		✓		[۴۷]Image_QA
	✓			[۱۵]Talk_to_Machine
		✓		[۴]VQA
			✓	[۶۸]Vis_Madlibs
		✓		[۴۶]VIS + LSTM
		✓		[۶۴]Ahab
		✓		[۸]ABC-CNN
		✓		[۳]Comp_QA
		✓		[۴۱]DPPNet
		✓		[۳۵]Answer_CNN
		✓		[۳۲]VQA-Caption
✓				[۲۲]Re_Baseline
✓				[۱۴]MCB
	✓			[۶۷]SMem-VQA
		✓		[۵۲]Region_VQA
		✓		[۷۱]Vis7W
✓	✓	✓	✓	[۳۸]Ask_Neuron
✓				[۷]SCMC
✓				[۳۶]HAN
		✓		[۶۹]StrSem
✓				[۴۹]AVQAN
✓				[۲۸]CMF
✓				[۳۳]EnsAtt
✓				[۶۱]MetaVQA
✓				[۵]DA-NTN
✓				[۷]QGHC
✓				[۵۱]QTA
✓				[۴۳]WRAN
✓				[۶۳]QAR





شکل ۳-۱۰: معماری شبکه Skip-gram و CBOW

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<sup>۷</sup> [۴۰] و Skip-gram [۴۰] 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 لایه‌ی 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 ۳-۱۰ آورده شده است.

یکی دیگر از word embedding های مشهور، مدل بردار سراسری یا به اختصار GloVe<sup>۸</sup> است که توسط پنینگتون و همکاران [۴۴] در سال ۲۰۱۴ در تیم پردازش زبان‌های طبیعی دانشگاه استنفورد معرفی و توسعه داده شد. آیا نیاز به توضیح کامل این روش است؟

با پیشرفت یادگیری عمیق در دهه اخیر، محققان برای استخراج ویژگی و بازنمایی متن از LSTM، CNN،

<sup>۷</sup> Continouse Bag Of Words  
<sup>۸</sup> Global Vector

[۲۱] و [۱۰] 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 آن ها اعمال می شود و پس از عبور از لایه max-pooling ویژگی ها بدست می آید.

توضیح LSTM لازمه؟

توضیح GRU لازمه؟

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 ۳-۵ لیستی از مدل های پرسش و پاسخ تصویری به همراه word embedding استفاده شده در آن ها را نمایش می دهد. با بررسی جدول ۳-۵ 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. آن ها 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 برچسب خورده ی زیادی دارد.

### ۳-۳-۲ فاز ۲: بازنمایی مشترک تصویر و سوال

در گام اول پرسش و پاسخ تصویری، تصویر و سوال به طور مستقل پردازش می شوند تا از آن ها ویژگی استخراج شود. روش های مختلف برای انجام این کار، در بخش ۳-۳-۱ به تفصیل بررسی شد. در گام بعدی، این ویژگی ها باید به یک فضای مشترک ترسیم شوند و یا به عبارتی ترکیب شوند تا آماده گام آخر (تولید پاسخ) شوند. در ادامه این بخش، به مرور روش های ترکیب ویژگی های استخراج شده از سوال و تصویر می پردازیم.

#### ۳-۳-۲ روش های پایه

ساده ترین و پایه ای ترین روش ها برای ترکیب ویژگی ها concatenation، جمع متناظر ویژگی ها<sup>۹</sup> و ضرب متناظر ویژگی ها<sup>۱۰</sup> است. مالینوفسکی در [۳۸] این سه روش را امتحان کرده است و دریافت کرد که ضرب متناظر ویژگی ها منجر به دقت بالاتری می شود. یافته مهم دیگر مالینوفسکی این است که نرمال سازی L2 ویژگی های تصویر، تأثیر قابل توجهی دارد به خصوص در روش های concatenation و جمع متناظر ویژگی ها. با توجه

<sup>۹</sup> element-wise addition

<sup>۱۰</sup> element-wise multiplication

جدول ۳-۵: word embedding های استفاده شده در مدل های پرسش و پاسخ تصویری.

GRU	LSTM	CNN	GloVe	Skip-gram/Word2vec	CBOW	one-hot	مدل پرسش و پاسخ تصویری
				✓			[۴۷] Image_QA
	✓						[۱۵] Talk_to_Machine
					✓		[۴] VQA
				✓			[۶۸] Vis_Madlibs
	✓						[۴۶] VIS + LSTM
	✓						[۸] ABC-CNN
	✓						[۳] Comp_QA
✓							[۴۱] DPPNet
		✓					[۳۵] Answer_CNN
	✓						[۳۲] VQA-Caption
				✓			[۲۲] Re_Baseline
	✓						[۱۴] MCB
					✓		[۶۷] SMem-VQA
				✓			[۵۲] Region_VQA
						✓	[۷۱] Vis7W
✓	✓	✓			✓		[۳۸] Ask_Neuron
		✓					[۷] SCMC
	✓						[۳۶] HAN
	✓						[۶۹] StrSem
						✓	[۴۹] AVQAN
	✓		✓				[۲۸] CMF
			✓				[۳۳] EnsAtt
✓			✓				[۶۱] MetaVQA
✓							[۵] DA-NTN
✓							[۷] QGHC
✓							[۴۳] WRAN
			✓				[۶۳] QAR

فصل ۳. مروری بر کارهای مرتبط ۳-۴. مدل‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر

به نتایج آن‌ها، جمع متناظر ویژگی‌ها پس از نرمال‌سازی از دقت بالاتری برخوردار است. در [۵۲] از ضرب نقطه‌ای (داخلی) بین ویژگی‌های استخراج شده از تصویر در سطح region و word embedding های حاصل از سوال استفاده شده است.

روش کلاسیک دیگر برای یافتن رابطه بین دو بردار که ریشه آن در علم آمار است، روش CCA<sup>۱۱</sup> است که برای ترکیب ویژگی‌های تصویر و سوال در VQA استفاده شده است. CCA بازنمایی مشترک بین بردار تصویر و بردار سوال را پیدا می‌کند. CCA یک نسخه نرمالیزه شده به نام nCCA<sup>۱۲</sup> نیز دارد که توسط [۱۶] پیشنهاد شده است. در [۶۸] و [۶۲] از هر دو مدل CCA و nCCA برای ترکیب بردارهای ویژگی سوال و تصویر استفاده کردند و دریافتند که روش nCCA به ویژه در مورد سوالات چندگزینه‌ای عملکرد بهتری دارد.

۳-۲-۲ روش‌های مبتنی بر شبکه‌های عصبی

در اینجا، محققان شبکه‌های عصبی عمیق end-to-end را با لایه‌های خاص برای ترکیب ویژگی‌های تصویر و سوال آموزش می‌دهند. ساختار و عملکرد این لایه ممکن است برای مدل‌های مختلف پیشنهاد شده متفاوت باشد.

ادامه اش باید تکمیل بشه ...

۳-۲-۳ روش‌های مبتنی بر توجه

۳-۳-۳ فاز ۳: تولید جواب

باید تکمیل شود.

۳-۴ مدل‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر<sup>۱۳</sup>

در سال‌های اخیر شاهد ظهور شبکه‌های از قبل آموزش دیده تنها بر روی داده‌های تصویری مثل ResNet [۱۹] و یا تنها بر روی داده‌های متنی مانند BERT [۱۳]، GPT-2 [۴۵] و GPT-3 [۶] بوده‌ایم. استفاده از این شبکه‌ها منجر به بهبود مسائل موجود در بینایی ماشین و پردازش زبان‌های طبیعی شده است. با الهام از این موضوع،

<sup>۱۱</sup> Analysis Correlation Canonical

<sup>۱۲</sup> Analysis Correlation Canonical normalized

<sup>۱۳</sup> vision-and-language pretraining models

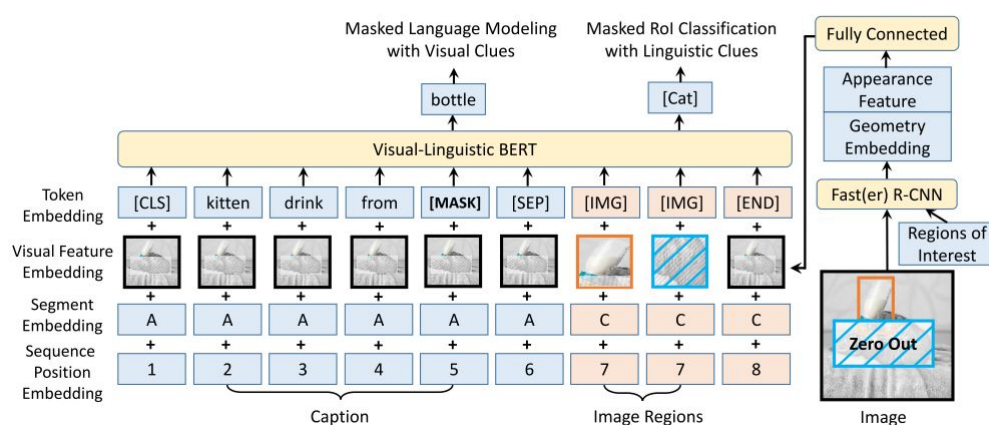
شبکه‌های از قبل آموزش دیده بر روی داده‌های تصویری و متنی نیز ایجاد شدند که هدف آن‌ها بازنمایی مشترک داده‌های تصویری و داده‌های زبانی است. بنابراین می‌توان از این شبکه‌ها برای بهبود عملکرد مسائل مشترک بین بینایی ماشین و پردازش زبان‌های طبیعی مانند پرسش و پاسخ تصویری نیز استفاده کرد. معماری شبکه‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر به طور کلی به دو دسته تک جریان<sup>۱۴</sup> و دو جریان<sup>۱۵</sup> تقسیم می‌شود. در ادامه به بحث و بررسی هر یک از این دسته‌ها می‌پردازیم.

### ۳-۴-۱ معماری تک جریان

پایه و اساس این معماری شبیه معماری مدل BERT [۱۳] است که رمزگذاری متن<sup>۱۶</sup> و رمزگذاری تصویر<sup>۱۷</sup> را به طور همزمان انجام می‌دهد. در واقع برای یادگیری بازنمایی متن و تصویر از یک رمزگذار<sup>۱۸</sup> استفاده می‌کند. بنابراین ورودی مدل‌های پیشنهادشده در این معماری داده‌های چندحالتی<sup>۱۹</sup> هستند که به صورت همزمان و یکجا به مدل داده می‌شوند برای مثال تصویر به همراه یک جمله توصیف کننده آن و یا یک فیلم به همراه زیرنویسش به این شبکه‌ها برای آموزش داده می‌شوند. به علاوه این مدل‌ها با ترکیبی از اهداف مختلف مانند masked visual-، text-based Masked Language Model، visual-based Masked Language Model، feature modeling و visual-linguistic matching بهینه می‌شود. سپس از بازنمایی‌های آموخته‌شده توسط این مدل‌ها در مسائل پایین‌دستی understanding و یا generation استفاده می‌شود. به عنوان مثال، مدل VideoBERT [۵۶] برای مسائل generation مانند تولید توصیف فیلم طراحی شده است. در حالی که چندین مدل دیگر مانند B2T2 [۲]، Unicoder-VL [۲۹]، VL-BERT<sup>۲۰</sup> [۵۵] و UNITER [۹] وجود دارد که همگی برای مسائل understanding طراحی شده‌اند. مدل‌های دیگری مانند VLP [۷۰] و OSCAR [۳۰] مدل‌های یکپارچه‌ای هستند که هم در مسائل پایین‌دستی understanding و هم در مسائل generative کاربرد دارد. از بین این مدل‌ها، تنها از مدل‌های VL-BERT، UNITER، VLP و OSCAR می‌توان برای مسئله پرسش و پاسخ تصویری استفاده کرد. بنابراین در ادامه این بخش جزئیات هر کدام از این مدل‌ها را توضیح خواهیم داد.

شکل ۳-۱۱ معماری VL-BERT را نشان می‌دهد. مشابه BERT، از کدگذارهای multi-layer bidi-

single-stream<sup>۱۴</sup>  
two-stream<sup>۱۵</sup>  
text encoding<sup>۱۶</sup>  
image encoding<sup>۱۷</sup>  
encoder<sup>۱۸</sup>  
multimodal<sup>۱۹</sup>  
Visual-Linguistic BERT<sup>۲۰</sup>



شکل ۳-۱۱: معماری شبکه از قبل آموزش دیده VL-BERT [۵۵]

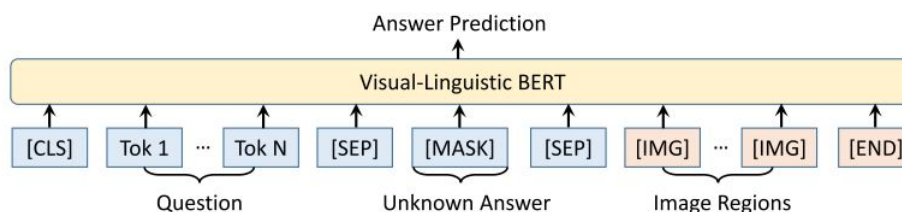
این rectional transformer استفاده شده است. اما برخلاف BERT که ورودی آن تنها کلمات جمله هستند، این شبکه به همراه کلمات یک جمله، مناطق مورد علاقه<sup>۲۱</sup> استخراج شده از تصویر و یا به اختصار ROI را نیز به عنوان ورودی می‌گیرد. برای استخراج ROI از تصویر از شبکه Faster RCNN [۴۸] استفاده شده است. هر ورودی این شبکه با توکن [CLS] آغاز می‌شود. سپس با کلمات جمله و ROI های تصویر ادامه می‌یابد و با توکن [END] خاتمه می‌یابد. از توکن [SEP] نیز برای جدا کردن جملات و یا جملات و تصویر از هم استفاده می‌شود. برای هر ورودی، تعبیه ویژگی<sup>۲۲</sup> آن جمع چهار نوع تعبیه است که در شکل ۳-۱۱ مشخص شده است. در میان آن‌ها، تعبیه مربوط به ویژگی‌های تصویری<sup>۲۳</sup> به تازگی به شبکه اضافه شده است در حالی که سه تعبیه دیگر از قبل در مدل BERT وجود داشته است. برای آموزش VL-BERT از مجموعه داده Conceptual Captions به عنوان مجموعه داده زبانی- تصویری استفاده شده است. علاوه بر این از دو مجموعه داده فقط زبانی به نام‌های BooksCorpus و English Wikipedia به منظور بهبود تعمیم‌دهی شبکه استفاده شده است. برای بهینه‌سازی شبکه VL-BERT از دو تابع هدف استفاده شده است: text-based Masked Language Model و visual-based Masked Language Model. در text-based Masked Language Model با احتمال ۱۵ درصد یکی از کلمات ورودی با توکن [MASK] جایگزین می‌شود. بنابراین شبکه باید سعی کند که این کلمه ماسک شده را با توجه به کلمات دیگر و ویژگی‌های تصویری در خروجی پیش‌بینی نماید. در

<sup>۲۱</sup> regions-of-interest

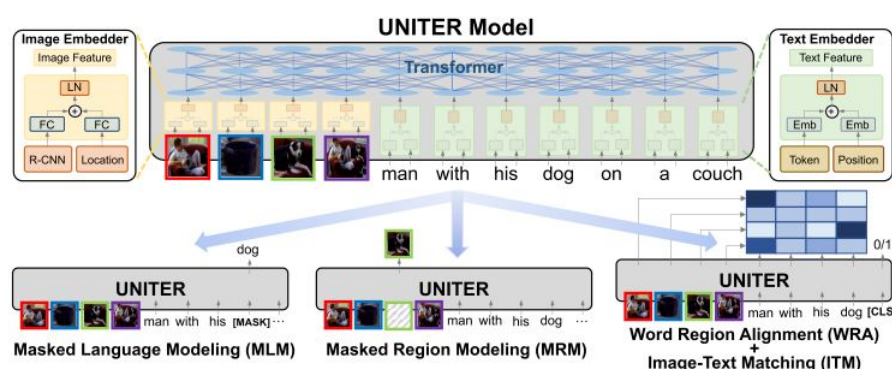
<sup>۲۲</sup> feature embedding

<sup>۲۳</sup> visual feature embedding

فصل ۳. مروری بر کارهای مرتبط ۳-۴. مدل‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر



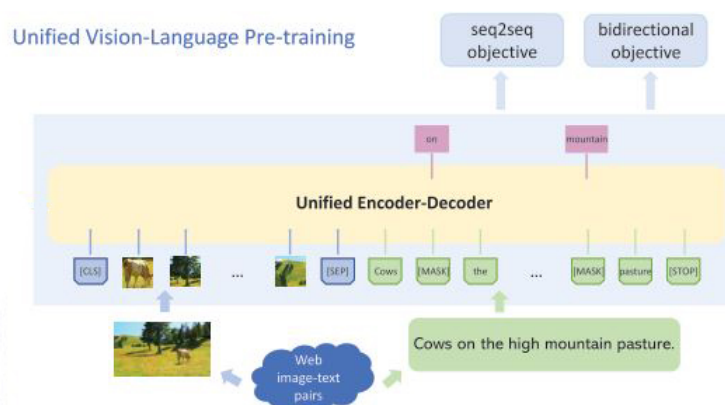
شکل ۳-۱۲: نحوه ورودی و خروجی شبکه VL-BERT برای آموزش در مسئله پرسش و پاسخ تصویری [۵۵]



شکل ۳-۱۳: معماری شبکه از قبل آموزش دیده UNITER [۹]

visual-based Masked Language Model با احتمال ۱۵ درصد یکی از ROI ها ماسک می‌شود و شبکه باید سعی کند در خروجی برچسب گروه مربوط به آن ROI را با توجه به کلمات و سایر ROI ها پیش‌بینی کند. دقت شود که همانطور که در قسمت سمت راست تصویر ۳-۱۱ مشخص است، ملاک برچسب گروه‌بندی درست برای ROI ها، خروجی شبکه Faster RCNN است. برای استفاده از شبکه از قبل آموزش دیده VL-BERT برای مسئله پرسش و پاسخ تصویری، مطابق شکل ۳-۱۲ سه تایی کلمات سوال، پاسخ و ROI های استخراج شده از تصویر توسط Faster RCNN در ورودی داده می‌شود که به جای پاسخ، [MASK] قرار گرفته که شبکه تلاش می‌کند؛ پاسخ را در خروجی پیش‌بینی کند.

معماری مدل UNITER در شکل ۳-۱۳ نشان داده شده است. ورودی این مدل مانند VL-BERT، کلمات یک جمله به همراه ROI های تصویر است. یکی از تفاوت‌های مدل UNITER با مدل VL-BERT این است که از ۴ مجموعه داده زبانی-تصویری برای آموزش استفاده کرده است: (۱) COCO، (۲) Visual Genome (VG)، (۳) Conceptual Captions و (۴) SBU Captions. تفاوت دیگر این مدل با مدل VL-BERT در توابع هدف است که علاوه بر text-based Masked Language Model و visual-based Masked Language Model



شکل ۳-۱۴: معماری شبکه از قبل آموزش دیده VLP [۷۰]

از دو تابع هدف دیگر به نام‌های Image-Text Matching و Word-Region Alignment نیز استفاده می‌کند. در Image-Text Matching هدف این است که مدل بتواند پیش‌بینی کند که آیا جمله و تصویر داده شده در ورودی با هم مطابقت دارند یا خیر. بدین منظور، یک جمله و ROI های تصویر به UNITER داده می‌شود و در خروجی بازنمایی مربوط به توکن [CLS] از یک تابع سیگموئید عبور داده می‌شود که یک مقدار بین صفر و یک را برمی‌گرداند که مقدار یک نشان می‌دهد که جمله و تصویر ورودی کاملاً با هم مطابقت دارد و مقدار صفر به این معناست که جمله و تصویر ورودی با هم مطابقت ندارد. در UNITER علاوه بر در نظر گرفتن تطابق جمله و تصویر، از تطابق بین کلمات موجود در جمله و ROI های تصویر نیز برای آموزش استفاده می‌شود که این موضوع در قالب تابع هدف Word-Region Alignment در مدل مطرح شده است. زمان آموزش مدل UNITER به ازای هر دسته از داده‌های ورودی، یکی از ۴ تابع هدف نامبرده شده به صورت تصادفی انتخاب می‌شود و براساس آن تابع هدف، عملیات کاهش گرادیان برای شبکه انجام می‌شود. برای استفاده از شبکه از قبل آموزش دیده UNITER برای مسئله پرسش و پاسخ تصویری، بازنمایی حاصل از توکن [CLS] به یک شبکه MLP داده می‌شود و پاسخ را برای سوال و تصویر ورودی پیش‌بینی می‌کند. در واقع در این حالت، مسئله پرسش و پاسخ تصویری به عنوان یک مسئله طبقه‌بندی در نظر گرفته می‌شود.

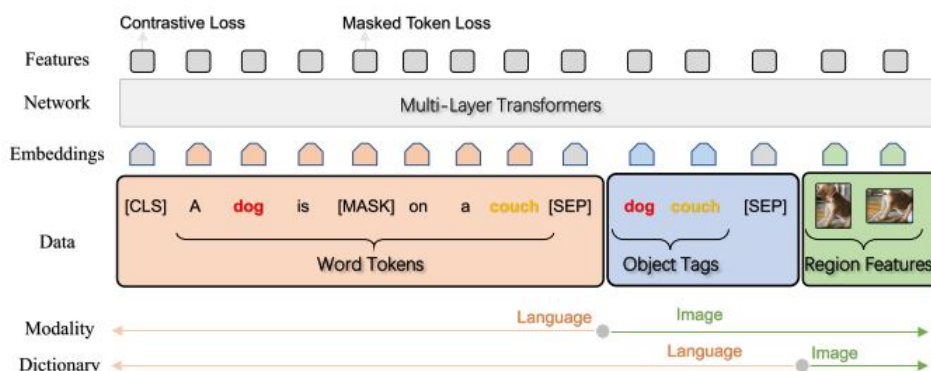
شبکه از قبل آموزش دیده VLP نیز مانند دو شبکه‌ی قبلی از کلمات یک جمله و ROI های استخراج شده از تصویر به عنوان ورودی استفاده می‌کند. تفاوت اصلی این شبکه با دو شبکه VL-BERT و UNITER در این است که یک شبکه‌ی یکپارچه رمزگذار-رمزگشا است که نه تنها در مسائل understanding بلکه در مسائل generative به دلیل وجود رمزگشا قابل استفاده است. مدل VLP بر روی مجموعه داده Conceptual Captions



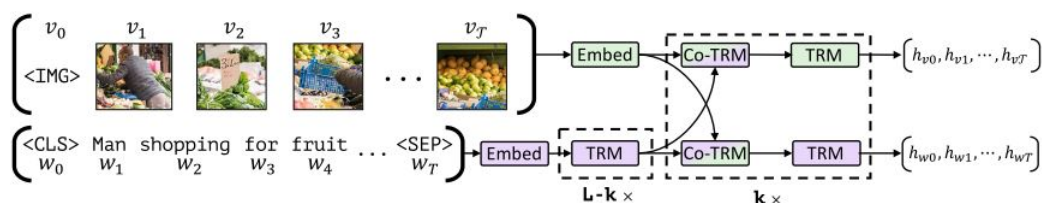
آموزش داده شده است. دو تابع هدف در شبکه VLP استفاده شده است: (۱) bidirectional و seq2seq. در تابع هدف bidirectional یکی از کلمات موجود در جمله با توکن [MASK] جایگزین می‌شود و برای پیش‌بینی این کلمه ماسک شده در خروجی از تمامی کلمات و ROI های اطراف آن استفاده می‌شود. اما در تابع هدف seq2seq برای پیش‌بینی کلمه ماسک شده در خروجی، تنها از کلمات سمت چپ کلمه ماسک شده و ROI های اطراف آن استفاده می‌کند. به عبارتی دیگر، برای پیش‌بینی کلمه ماسک شده نمی‌توان از کلماتی که بعد از آن و در آینده در جمله آمده است؛ استفاده کرد. معماری شبکه VLP در شکل ۳-۱۴ نشان داده شده است. ورودی سه مدل قبلی یعنی VL-BERT ، UNITER و VLP یک جمله به همراه ROI های استخراج شده از تصویر بود. در مدل OSCAR علاوه بر این دو ورودی از ورودی دیگری به نام برچسب اشیا<sup>۲۴</sup> استفاده می‌شود که اشیا‌یی که هم در تصویر وجود دارد و هم در جمله به آن اشاره شده است را نشان می‌دهد. در [۳۰] ادعا شده است که استفاده برچسب اشیا منجر به تولید بازنمایی بهتری از متن و تصویر می‌شود و در واقع از این برچسب‌ها به عنوان لنگر برای تطابق دادن فضای تصویر و متن استفاده می‌شود. در مدل OSCAR برای بدست آوردن ROI های تصویر و برچسب اشیا از شبکه Faster RCNN استفاده شده است. در مدل OSCAR به دو طریق می‌توان به ورودی‌ها نگاه کرد که در نتیجه دو تابع هدف برای آموزش این شبکه تعریف می‌شود. در روش اول، کلمات جمله و برچسب اشیا با هم در نظر گرفته می‌شود (دید Dictionary) و به احتمال ۱۵ در صد یکی از کلمات جمله و یا یکی از برچسب‌های اشیا با توکن [MASK] جایگزین می‌شود و مدل باید سعی کند این کلمه ماسک شده را در خروجی پیش‌بینی کند (Masked Token Loss). در روش دوم، ROI های تصویر و برچسب اشیا با هم در نظر گرفته می‌شود (دید Modality) و با احتمال ۵۰ درصد برچسب‌های اشیا با برچسب‌های دیگری تغییر می‌کند و مدل باید پیش‌بینی کند که آیا کلمات موجود در جمله با قسمت برچسب اشیا و ROI های تصویر مطابقت دارد یا نه. که بدین منظور خروجی شبکه برای توکن [CLS] به یک شبکه کاملاً متصل داده می‌شود و یک طبقه‌بندی باینری انجام می‌شود که یک به معنای تطابق کلمات جمله با ROI های تصویر و برچسب اشیاست و صفر نشان‌دهنده عدم تطابق است (Contrastive Loss). برای آموزش مدل OSCAR از مجموعه داده‌های COCO ، Conceptual Captions ، SBU captions ، flicker30 و GQA استفاده شده است. برای استفاده از شبکه از قبل آموزش دیده OSCAR برای مسئله پرسش و پاسخ تصویری، سوال به همراه برچسب اشیا و ROI های تصویر به ورودی شبکه داده می‌شود و خروجی توکن [CLS] به یک طبقه‌بند داده می‌شود تا پاسخ سوال و تصویر داده شده در تصویر بدست آید. در واقع در این روش، مسئله

<sup>۲۴</sup>object tag

فصل ۳. مروری بر کارهای مرتبط ۴-۳. مدل‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر



شکل ۳-۱۵: معماری شبکه از قبل آموزش دیده OSCAR [۳۰]



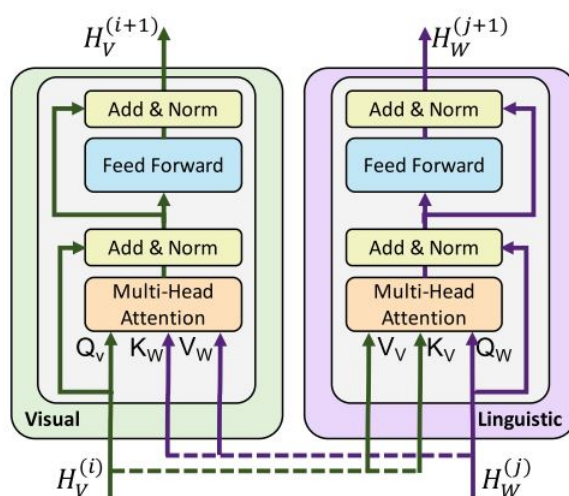
شکل ۳-۱۶: معماری شبکه از قبل آموزش دیده ViLBERT [۳۴]

پرسش و پاسخ تصویری به صورت یک مسئله طبقه‌بندی در نظر گرفته می‌شود. معماری شبکه OSCAR در شکل ۳-۱۵ نمایش داده شده است.

### ۳-۴-۲ معماری دو جریان

در مقابل معماری تک جریان، معماری دو جریان برای یادگیری هر کدام از بازنمایی‌های تصویر و متن از یک رمزگذار مستقل استفاده می‌کند. سپس از یک رمزگذار دیگر برای بدست آوردن بازنمایی مشترک متن و تصویر استفاده می‌کند. مشابه معماری تک جریان، معماری‌های دو جریان نیز مدل‌های خود را با visual-based Masked Language Model، text-based Masked Language Model و visual-linguistic matching بهینه می‌کنند. ViLBERT [۳۴] و LXMERT [۵۹] نمونه‌هایی از معماری دو جریان هستند که از این دو مدل می‌توان برای مسئله پرسش و پاسخ تصویری استفاده کرد. پس در ادامه این بخش، جزئیات این دو شبکه را بررسی خواهیم کرد.

شکل ۳-۱۶ معماری شبکه ViLBERT را نمایش می‌دهد. مدل ViLBERT شامل دو مدل موازی به

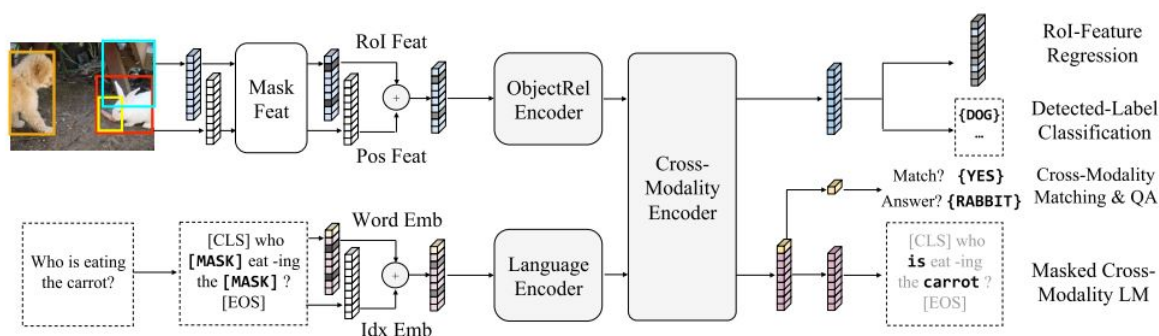


شکل ۳-۱۷: ساختار لایه co-attentional transformer [۳۴]

شبکه BERT است که به صورت جداگانه بر روی کلمات متن و ROI های تصویر اعمال می‌شود و از بلوک‌های ترنسفرمر در هر جریان استفاده شده است (در شکل ۳-۱۶ با TRM مشخص شده است). سپس برای بدست آوردن بازنمایی مشترک بین متن و تصویر از لایه‌های co-attentional transformer استفاده شده است (در شکل ۳-۱۶ با Co-TRM مشخص شده است). اساس لایه co-attentional transformer بر پایه‌ی ترنسفرمر است در واقع برای هر کدام از بخش‌های تصویری و متنی داده ورودی، یک ترنسفرمر در لایه co-attentional transformer در نظر گرفته شده است که پس از عبور متن و داده از جریان‌های مستقل خود و بدست آمدن query، key و value برای هر کدام، key و value متن به ترنسفرمر تصویر در co-attentional transformer داده می‌شود و به صورت متقابل key و value تصویر به ترنسفرمر متن داده می‌شود.

شکل ۳-۱۷ ساختار لایه co-attentional transformer را نشان می‌دهد. برای آموزش مدل ViLBERT از توابع هدف visual-based Masked Language Model، text-based Masked Language Model و linguistic matching استفاده شده است. شبکه ViLBERT بر روی مجموعه داده Conceptual Captions آموزش داده شده است. برای استفاده از شبکه از قبل آموزش دیده ViLBERT برای مسئله پرسش و پاسخ تصویری، ابتدا خروجی بازنمایی توکن [CLS] و بازنمایی تصویر ضرب متناظر می‌شوند. سپس با عبور از یک شبکه MLP دولایه پاسخ مربوط به سوال و تصویر حاصل می‌شود.

شکل ۳-۱۸ معماری مدل LXMERT را نشان می‌دهد. ورودی این شبکه کلمات جمله ورودی و ROI



شکل ۳-۱۸: معماری شبکه از قبل آموزش دیده LXMERT [۵۹]

های استخراج شده از تصویر است. همان‌طور که قبلاً اشاره شد؛ مدل LXMERT یک مدل دو جریان است به همین دلیل برای پردازش متن و تصویر از دو رمزگذار مجزا و مستقل استفاده شده است (در شکل ۳-۱۸ به ترتیب با عنوان‌های ObjectRel Encoder و Language Encoder برای تصویر و متن مشخص شده است. ( و سپس برای بدست آوردن بازنمایی مشترک از رمزگذار Cross-modality استفاده شده است. توابع هدف استفاده شده در مدل LXMERT مشابه شبکه ViLBERT است اما در LXMERT از تابع هدف دیگری به نام image question answering برای آموزش شبکه استفاده شده است. زیرا حدود ۱/۳ داده‌ای که برای آموزش این شبکه استفاده شده است؛ یک سوال در مورد تصویر ورودی است. بنابراین با تعریف تابع هدف image question answering مدل سعی می‌کند تا پاسخ این سوال را در خروجی پیش‌بینی کند. برای آموزش شبکه LXMERT از مجموعه داده‌های Visual Genome ، MS COCO ، VQA v2.0 ، GQA balanced version و VG-QA استفاده شده است.

در جدول ۳-۷ نتایج مدل‌های LXMERT ، ViL-BERT ، OSCAR ، VLP ، UNITER ، VL-BERT بر روی مجموعه داده VQA v2.0 نشان داده شده است. بهترین نتیجه بدست آمده برای مدل UNITER است. یکی از نکات قابل ملاحظه در این جدول این است که مدل‌های تک جریان نتایج بهتری نسبت به مدل‌های دو جریان بدست آوردند در حالی که تعداد پارامترهای مدل‌های تک جریان نسبت به مدل‌های دو جریان کمتر است.

فصل ۳. مروری بر کارهای مرتبط ۴-۳. مدل‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر

جدول ۳-۶:

روش	معماری	ورودی	مجموعه داده‌های استفاده شده برای آموزش	توابع هدف
<b>VL-BERT</b> [۵۵]	تک جریان	کلمات جمله + ROI های تصویر	Conceptual Captions + BooksCorpus + English Wikipedia	text-based MLM + visual-based MLM
<b>UNITER</b> [۹]	تک جریان	کلمات جمله + ROI های تصویر	COCO + Visual Genome + Conceptual Captions + SBU Captions	text-based MLM + visual-based MLM + Image-Text Matching + Word-Region Alignment
<b>OSCAR</b> [۳۰]	تک جریان	کلمات جمله + ROI های تصویر	Conceptual Captions	text-based MLM + visual-based MLM + Image-Text Matching + Word-Region Alignment
<b>VLP</b> [۷۰]	تک جریان	–	–	–
<b>ViL-BERT</b> [۳۴]	دو جریان	–	–	–
<b>LXMERT</b> [۵۹]	دو جریان	–	–	–

جدول ۳-۷: دقت شبکه‌های از قبل آموزش دیده بر روی مجموعه داده VQA v2.0 (test-std)

دقت کل	سایر سوالات	سوالات شمارشی	سوالات بله/خیر	روش
۷۰/۷	۶۰/۵	۵۲/۱	۸۷/۴	<b>VLP</b> [۷۰]
۷۰/۹۲	–	–	–	<b>ViL-BERT</b> [۳۴]
۷۲/۲۲	–	–	–	<b>VL-BERT</b> [۵۵]
۷۲/۵	۶۳/۱	۵۴/۲	۸۸/۲	<b>LXMERT</b> [۵۹]
۷۳/۸۲	–	–	–	<b>OSCAR</b> [۳۰]
۷۴/۰۲	–	–	–	<b>UNITER</b> [۹]

### ۳-۵ 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.

#### ۳-۵-۱ 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.

$$Accuracy = \frac{\text{Number of questions answered correctly}}{\text{Total questions}} \quad (۳-۱)$$

### ۳-۵-۲ معیار شباهت Wu-Palmer [۶۶]

این معیار ارزیابی توسط مالینوفسکی [۳۷] 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.

### ۳-۵-۳ 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 می‌گیرد و در غیر این صورت هیچ امتیازی کسب نخواهد کرد. از معایب این روش می‌توان به هزینه زیاد جمع‌آوری پاسخ برای سوالات اشاره کرد. آنتول و همکارانش از این معیار ارزیابی در [۴] استفاده کرده‌اند.

$$Accuracy_{VQA} = \min\left(\frac{n}{3}, 1\right) \quad (2-3)$$

### ۳-۵-۴ MPT [۲۵]

یکی از مشکلات مجموعه داده‌های پرسش و پاسخ تصویری توزیع غیریکنواخت انواع سوال‌هاست. در این مواقع، نمی‌توان از معیار دقت استفاده کرد. بنابراین در [۲۵] معیار جدیدی به نام MPT<sup>۲۵</sup> ارائه شده است که توزیع نامتوازن سوال‌ها را جبران می‌کند. معیار MPT میانگین دقت برای هر نوع سوال را محاسبه می‌کند.

Mean Per Type<sup>۲۵</sup>

از نسخه‌ی نرمالایز شده‌ی این معیار نیز برای رفع مشکل بایاس در توزیع پاسخ‌ها استفاده می‌شود.

### ۳-۵-۵ BLEU [۴۲]

BLEU<sup>۲۶</sup> یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. در [۱۸] 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 پاسخ پیش‌بینی شده و پاسخ صحیح را اندازه‌گیری می‌کند. معمولاً BLEU زمانی که جمله‌ها کوتاه باشند، با شکست مواجه می‌شود.

### ۳-۵-۶ METEOR [۱۲]

METEOR<sup>۲۷</sup> نیز همانند BLEU یکی از معیارهای ارزیابی خودکار ترجمه ماشینی است. به پیشنهاد [۱۸] 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.

---

<sup>۲۶</sup> BiLingual Evaluation Understudy  
<sup>۲۷</sup> Metric for Evaluation of Translation with Explicit ORdering



## فصل ۴

### نتیجه‌گیری و کارهای آینده

۴-۱ نتیجه‌گیری

۴-۲ مسائل باز و کارهای قابل انجام

- [1] ACHARYA, M., KAFLE, K., AND KANAN, C. Tallyqa: Answering complex counting questions. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), volume 33, pp. 8076–8084.
- [2] ALBERTI, C., LING, J., COLLINS, M., AND REITTER, D. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054* (2019).
- [3] ANDREAS, J., ROHRBACH, M., DARRELL, T., AND KLEIN, D. Deep compositional question answering with neural module networks. corr abs/1511.02799 (2015). *arXiv preprint arXiv:1511.02799* (2015).
- [4] ANTOL, S., AGRAWAL, A., LU, J., MITCHELL, M., BATRA, D., LAWRENCE ZITNICK, C., AND PARIKH, D. Vqa: Visual question answering. in *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2425–2433.
- [5] BAI, Y., FU, J., ZHAO, T., AND MEI, T. Deep attention neural tensor network for visual question answering. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 20–35.
- [6] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [7] CAO, L., GAO, L., SONG, J., XU, X., AND SHEN, H. T. Jointly learning attentions with semantic cross-modal correlation for visual question answering. in *Australasian Database Conference* (2017), Springer, pp. 248–260.
- [8] CHEN, K., WANG, J., CHEN, L.-C., GAO, H., XU, W., AND NEVATIA, R. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960* (2015).

- [9] CHEN, Y.-C., LI, L., YU, L., EL KHOLY, A., AHMED, F., GAN, Z., CHENG, Y., AND LIU, J. Uniter: Universal image-text representation learning. in *European Conference on Computer Vision* (2020), Springer, pp. 104–120.
- [10] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [11] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. in *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.
- [12] DENKOWSKI, M., AND LAVIE, A. Meteor universal: Language specific translation evaluation for any target language. in *Proceedings of the ninth workshop on statistical machine translation* (2014), pp. 376–380.
- [13] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [14] FUKUI, A., PARK, D. H., YANG, D., ROHRBACH, A., DARRELL, T., AND ROHRBACH, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).
- [15] GAO, H., MAO, J., ZHOU, J., HUANG, Z., WANG, L., AND XU, W. Are you talking to a machine? dataset and methods for multilingual image question. in *Advances in neural information processing systems* (2015), pp. 2296–2304.
- [16] GONG, Y., KE, Q., ISARD, M., AND LAZEBNIK, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106, 2 (2014), 210–233.
- [17] GOYAL, Y., KHOT, T., SUMMERS-STAY, D., BATRA, D., AND PARIKH, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 6904–6913.
- [18] GURARI, D., LI, Q., STANGL, A. J., GUO, A., LIN, C., GRAUMAN, K., LUO, J., AND BIGHAM, J. P. Vizwiz grand challenge: Answering visual questions from blind people. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3608–3617.

- [19] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.
- [20] HINTON, G. E., KRIZHEVSKY, A., AND SUTSKEVER, I. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1106–1114.
- [21] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [22] JABRI, A., JOULIN, A., AND VAN DER MAATEN, L. Revisiting visual question answering baselines. in *European conference on computer vision* (2016), Springer, pp. 727–739.
- [23] JOHNSON, J., HARIHARAN, B., VAN DER MAATEN, L., FEI-FEI, L., LAWRENCE ZITNICK, C., AND GIRSHICK, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2901–2910.
- [24] KAFLE, K., AND KANAN, C. Answer-type prediction for visual question answering. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4976–4984.
- [25] KAFLE, K., AND KANAN, C. An analysis of visual question answering algorithms. in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1965–1973.
- [26] KAFLE, K., YOUSEFHUSSEN, M., AND KANAN, C. Data augmentation for visual question answering. in *Proceedings of the 10th International Conference on Natural Language Generation* (2017), pp. 198–202.
- [27] KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANTIDIS, Y., LI, L.-J., SHAMMA, D. A., ET AL. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.
- [28] LAO, M., GUO, Y., WANG, H., AND ZHANG, X. Cross-modal multistep fusion network with co-attention for visual question answering. *IEEE Access* 6 (2018), 31516–31524.

- [29] LI, G., DUAN, N., FANG, Y., GONG, M., JIANG, D., AND ZHOU, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. in *AAAI* (2020), pp. 11336–11344.
- [30] LI, X., YIN, X., LI, C., ZHANG, P., HU, X., ZHANG, L., WANG, L., HU, H., DONG, L., WEI, F., ET AL. Oscar: Object-semantics aligned pre-training for vision-language tasks. in *European Conference on Computer Vision* (2020), Springer, pp. 121–137.
- [31] LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. Microsoft coco: Common objects in context. in *European conference on computer vision* (2014), Springer, pp. 740–755.
- [32] LIN, X., AND PARIKH, D. Leveraging visual question answering for image-caption ranking. in *European Conference on Computer Vision* (2016), Springer, pp. 261–277.
- [33] LIOUTAS, V., PASSALIS, N., AND TEFAS, A. Explicit ensemble attention learning for improving visual question answering. *Pattern Recognition Letters 111* (2018), 51–57.
- [34] LU, J., BATRA, D., PARIKH, D., AND LEE, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. in *Advances in Neural Information Processing Systems* (2019), pp. 13–23.
- [35] MA, L., LU, Z., AND LI, H. Learning to answer questions from image using convolutional neural network. in *AAAI* (2016).
- [36] MALINOWSKI, M., DOERSCH, C., SANTORO, A., AND BATTAGLIA, P. Learning visual question answering by bootstrapping hard attention. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–20.
- [37] MALINOWSKI, M., AND FRITZ, M. A multi-world approach to question answering about real-world scenes based on uncertain input. in *Advances in neural information processing systems* (2014), pp. 1682–1690.
- [38] MALINOWSKI, M., ROHRBACH, M., AND FRITZ, M. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision 125*, 1-3 (2017), 110–135.
- [39] MANMADHAN, S., AND KOVOOR, B. C. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review* (2020), 1–41.

- [40] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [41] NOH, H., HONGSUCK SEO, P., AND HAN, B. Image question answering using convolutional neural network with dynamic parameter prediction. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 30–38.
- [42] PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.
- [43] PENG, L., YANG, Y., BIN, Y., XIE, N., SHEN, F., JI, Y., AND XU, X. Word-to-region attention network for visual question answering. *Multimedia Tools and Applications* 78, 3 (2019), 3843–3858.
- [44] PENNINGTON, J., SOCHER, R., AND MANNING, C. D. Glove: Global vectors for word representation. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.
- [45] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., AND SUTSKEVER, I. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [46] REN, M., KIROS, R., AND ZEMEL, R. Exploring models and data for image question answering. in *Advances in neural information processing systems* (2015), pp. 2953–2961.
- [47] REN, M., KIROS, R., AND ZEMEL, R. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst* 1, 2 (2015), 5.
- [48] REN, S., HE, K., GIRSHICK, R., AND SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. in *Advances in neural information processing systems* (2015), pp. 91–99.
- [49] RUWA, N., MAO, Q., WANG, L., AND DONG, M. Affective visual question answering network. in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)* (2018), IEEE, pp. 170–173.
- [50] SHAH, S., MISHRA, A., YADATI, N., AND TALUKDAR, P. P. Kvqa: Knowledge-aware visual question answering. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), volume 33, pp. 8876–8884.

- [51] SHI, Y., FURLANELLO, T., ZHA, S., AND ANANDKUMAR, A. Question type guided attention in visual question answering. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 151–166.
- [52] SHIH, K. J., SINGH, S., AND HOIEM, D. Where to look: Focus regions for visual question answering. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4613–4621.
- [53] SILBERMAN, N., HOIEM, D., KOHLI, P., AND FERGUS, R. Indoor segmentation and support inference from rgb-d images. in *European conference on computer vision* (2012), Springer, pp. 746–760.
- [54] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [55] SU, W., ZHU, X., CAO, Y., LI, B., LU, L., WEI, F., AND DAI, J. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).
- [56] SUN, C., MYERS, A., VONDRICK, C., MURPHY, K., AND SCHMID, C. Videobert: A joint model for video and language representation learning. in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 7464–7473.
- [57] SZEGEDY, C., LIU, W., JIA, Y., Sermanet, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. Going deeper with convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.
- [58] TALAFHA, B., AND AL-AYYOUB, M. Just at vqa-med: A vgg-seq2seq model. in *CLEF (Working Notes)* (2018).
- [59] TAN, H., AND BANSAL, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
- [60] TANG, R., MA, C., ZHANG, W. E., WU, Q., AND YANG, X. Semantic equivalent adversarial data augmentation for visual question answering. in *European Conference on Computer Vision* (2020), Springer, pp. 437–453.
- [61] TENEY, D., AND VAN DEN HENGEL, A. Visual question answering as a meta learning task. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 219–235.

- [62] TOMMASI, T., MALLYA, A., PLUMMER, B., LAZEBNIK, S., BERG, A. C., AND BERG, T. L. Combining multiple cues for visual madlibs question answering. *International Journal of Computer Vision* 127, 1 (2019), 38–60.
- [63] TOOR, A. S., WECHSLER, H., AND NAPPI, M. Question action relevance and editing for visual question answering. *Multimedia Tools and Applications* 78, 3 (2019), 2921–2935.
- [64] WANG, P., WU, Q., SHEN, C., HENGEL, A. V. D., AND DICK, A. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).
- [65] WU, Q., TENEY, D., WANG, P., SHEN, C., DICK, A., AND VAN DEN HENGEL, A. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.
- [66] WU, Z., AND PALMER, M. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033* (1994).
- [67] XU, H., AND SAENKO, K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. in *European Conference on Computer Vision* (2016), Springer, pp. 451–466.
- [68] YU, L., PARK, E., BERG, A. C., AND BERG, T. L. Visual madlibs: Fill in the blank description generation and question answering. in *Proceedings of the ieee international conference on computer vision* (2015), pp. 2461–2469.
- [69] YU, Z., YU, J., XIANG, C., FAN, J., AND TAO, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* 29, 12 (2018), 5947–5959.
- [70] ZHOU, L., PALANGI, H., ZHANG, L., HU, H., CORSO, J. J., AND GAO, J. Unified vision-language pre-training for image captioning and vqa. in *AAAI* (2020), pp. 13041–13049.
- [71] ZHU, Y., GROTH, O., BERNSTEIN, M., AND FEI-FEI, L. Visual7w: Grounded question answering in images. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4995–5004.



# واژه‌نامه فارسی به انگلیسی

Voice assistants	دستیاران صوتی
Conversational agents	عامل‌های گفتگو
Image captioning	توصیف تصویر
Text-to-image retrieval	بازیابی متن به تصویر
Object detection	تشخیص اشیا
Activity detection	تشخیص فعالیت
Attribute classification	طبقه‌بندی صفات
Scene classification	طبقه‌بندی صحنه
Global vector	بردار سراسری
Element-wise addition	جمع متناظر
Element-wise multiplication	ضرب متناظر
Rotation	چرخش
Flipping	ورق زدن
Data augmentation	افزایش داده

# واژه‌نامه انگلیسی به فارسی

Voice assistants . . . . .	دستیاران صوتی . . . . .
Conversational agents . . . . .	عامل‌های گفتگو . . . . .
Image captioning . . . . .	توصیف تصویر . . . . .
Text-to-image retrieval . . . . .	بازیابی متن به تصویر . . . . .
Object detection . . . . .	تشخیص اشیا . . . . .
Activity detection . . . . .	تشخیص فعالیت . . . . .
Attribute classification . . . . .	طبقه‌بندی صفات . . . . .
Scene classification . . . . .	طبقه‌بندی صحنه . . . . .
Global vector . . . . .	بردار سراسری . . . . .
Element-wise addition . . . . .	جمع متناظر . . . . .
Element-wise multiplication . . . . .	ضرب متناظر . . . . .
rotation . . . . .	چرخش . . . . .
flipping . . . . .	ورق زدن . . . . .
Data augmentation . . . . .	افزایش داده . . . . .

**Abstract:**

abstract

**Keywords:** keywords



**Iran University of Science and Technology  
Computer Engineering Department**

# **Visual Question Answering**

**A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree  
of Master of Science in Computer Engineering**

**By:**

Maryam Sadat Hashemi

**Supervisor:**

**Sayyed Sauleh Eetemadi**

**December 2020**