# پرسش و پاسخ تصویری در فارسی

مریم سادات هاشمی
دانشکده مهندسی کامپیوتر
دانشگاه علم و صنعت ایران
m_hashemi94@cs.iust.ac.ir

علیرضا اصغری
دانشکده مهندسی کامپیوتر
دانشگاه علم و صنعت ایران
a_asghari@comp.iust.ac.ir

## چکیده

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 اخیرا مورد توجه قرارگرفته است؛ پرسش و پاسخ تصویری است. در این مسئله، 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. هدف ما در این پروژه، حل مسئله‌ی VQA در زبان فارسی است. بدین منظور مجموعه‌داده‌ای را فراهم کردیم و سه روش LSTM Q + norm I ، Stacked Attention Network و HieCoAttention را بر روی این مجموعه‌داده پیاده‌سازی و اجرا کردیم.

## ۱ مقدمه

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.

پرسش و پاسخ تصویری نسخه گسترش یافته مسئله پرسش و پاسخ متنی است که اطلاعات بصری به مسئله اضافه شده است. شکل 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


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

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 نشانگر سطح بالاتری از انتزاع دنیای واقعی است [۱۳].

کاربردهای بسیاری برای پرسش و پاسخ تصویری وجود دارد. یکی از مهم‌ترین موارد دستیار هوشمند برای افراد کم‌بینا و نابینا [۱] است [۳]. علاوه بر این، در سال‌های اخیر دستیاران صوتی [۲] و عامل‌های گفتگو [۳] مانند Siri ، Cortana و Alexa در بازار عرضه شدند که می‌توانند با انسان‌ها با استفاده از زبان طبیعی ارتباط برقرار کنند. در حال حاضر این دستیاران با استفاده از صوت و متن این ارتباط را برقرار می‌کنند در نتیجه گفتگوی بین این دستیاران با انسان‌ها مشابه دنیای واقعی نمی‌باشد. این ارتباط را می‌توان با استفاده از داده‌های تصویری و ویدئویی به واقعیت نزدیک‌تر کرد. اینجاست که مسئله‌ی پرسش و پاسخ تصویری برای نزدیک کردن تعامل بین انسان و عامل‌های گفتگو به دنیای واقعی می‌تواند موثر باشد. همین موضوع را می‌توانیم به صورت گسترده‌تری در ربات‌ها مشاهده کنیم. برای این‌که ربات بتواند بهتر با انسان‌ها ارتباط برقرار کند و به سوالات و درخواست‌ها پاسخ دهد؛ نیاز دارد که درک و فهم درستی از اطراف داشته باشد که این مستلزم داشتن تصویری دقیق از پیرامون است. بنابراین این ربات می‌تواند برای پاسخ به پرسش‌ها از دانشی که از طریق تصویر پیرامون خود بدست می‌آورد، جواب درستی را بدهد. کاربرد دیگر این مسئله در پزشکی است. در بسیاری 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 می تواند 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.

## ۲ کارهای مرتبط / پیش‌زمینه

در سال‌های اخیر، رویکردهای بیشماری برای VQA پیشنهاد شده است. همه رویکردهای موجود شامل موارد زیر است:

۱. رویکردهای مبتنی بر ترکیب ویژگی

۲. رویکردهای مبتنی بر attention

۳. رویکرد های مبتنی بر استدلال

در این پروژه ما از سه روش استفاده کرده‌ایم که روش LSTM Q + norm I مبتنی بر ترکیب ویژگی هاست و دو روش Stacked Attention Network و HieCoAttention مبتنی بر attention هستند. بر این اساس، کارهای انجام شده در این دو دسته را مرور خواهیم کرد.

### ۱.۲ رویکردهای مبتنی بر ترکیب ویژگی

این رویکردها هم ویژگی‌های تصویری و هم ویژگی‌های سوال را به یک فضای مشترک برای پیش‌بینی پاسخ منتقل می‌کنند. برای استخراج ویژگی‌های تصاویر، اکثر الگوریتم‌ها از CNN های از قبل آموزش دیده استفاده می‌کنند که بر روی مجموعه داده ImageNet آموزش داده شده‌اند. برخی از شبکه‌های رایج عبارتند از: GoogLeNet[۱۲] ، ResNet [۴] و VGGNet [۱۱]. برای استخراج ویژگی‌ها از سوالات، از روش‌هایی مانند کیسه کلمات (BOW) ، GRU [۲] و LSTM [۵] استفاده می‌شود. در این رویکرد عموماً مسئله VQA را یک مسئله طبقه‌بندی در نظر می‌گیرند و روش‌های متعددی برای ترکیب ویژگی‌های تصویر و سوال وجود دارد. بعضی از این روش‌ها ساده می‌باشند از جمله: concatenation ، elementwise addition ، elementwise multiplication و bilinear pooling . اما ممکن است از روش‌های پیچیده‌تری مانند Bayesian models نیز استفاده شود. دقتی که از روش‌های مبتنی بر این رویکرد بدست می‌آید متفاوت است و وابستگی زیادی به انتخاب هایپرپارامترها، پیکربندی سیستم و تنظیمات آزمایش‌ها دارد.

---

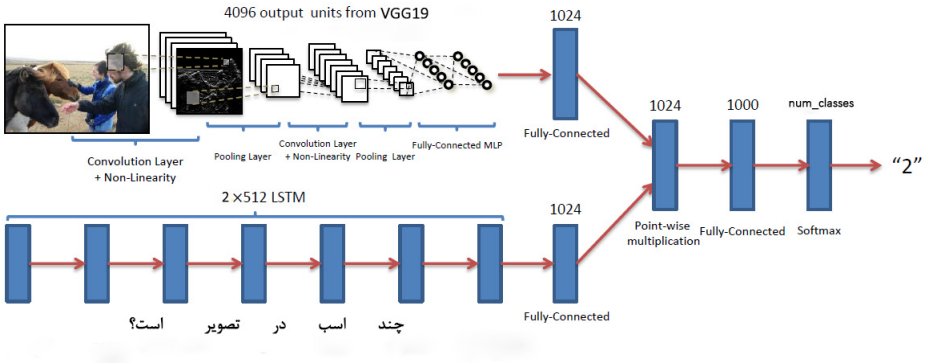
[۱] https://vizwiz.org/

[۲] Voice Assistant

[۳] Conversational Agents

| | تعداد تصاویر | تعدادسوالات | تعداد پاسخ‌ها |
|---|---|---|---|
| داده‌های آموزشی | ۸۲،۷۸۳ | ۲۴۸،۳۴۹ | ۲،۴۸۳،۴۹۰ |
| داده‌های ارزیابی | ۴۰،۵۰۴ | ۱۲۱،۵۱۲ | ۱،۲۱۵،۱۲۰ |
| داده‌های تست | ۸۱،۴۳۴ | ۲۴۴،۳۰۲ | |

جدول ۱: مشخصات مجموعه‌داده VQA




شکل ۲: ساختار کلی روش LSTM Q + norm I .

### ۲.۲ رویکردهای مبتنی بر attention

مدل‌های مبتنی بر attention به ناحیه‌هایی از تصاویر که مربوط به سوال است، توجه می‌کنند. مدل‌های موجود در این رویکرد یا به تصویر و یا به سوال و یا به هر دو توجه می‌کنند. به عنوان مثال، در [۱۰] مدلی را پیشنهاد داده است که با انتخاب یک منطقه تصویری که مربوط به متن سؤال باشد، پاسخ را پیش بینی می‌کند. در این روش به به تصویر توجه شده است. اما در مثالی دیگر [۸] از چندین لایه coattention استفاده می‌کند و هر کلمه از سوال با هر منطقه در تصویر در تعامل است و بالعکس. روش‌های پیشنهادی در این رویکرد بسیار است مانند linear Attention Network (BAN) [۶] و Question Type guided Attention (QTA) [۹] .

## ۳ مدل پیشنهاد شده

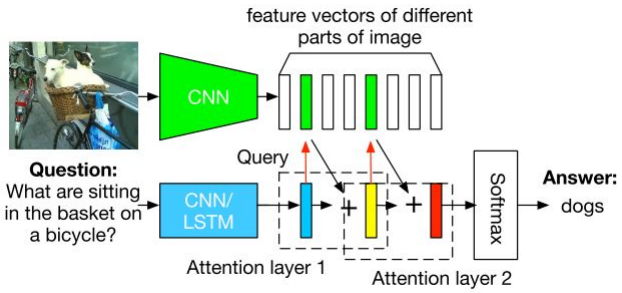
در این بخش ابتدا نحوه‌ی آماده‌سازی مجموعه‌داده را توضیح می‌دهیم و سپس به شرح روش‌های پیاده‌سازی شده در این پروژه می‌پردازیم.

### ۱.۳ تهیه‌ی مجموعه‌داده

مجموعه‌داده‌ای که برای حل این مسئله انتخاب کردیم؛ مجموعه داده VQA v1 است. مشخصات کامل مجموعه داده را می‌توانید در جدول ۱ مشاهده کنید. برای ترجمه مجموعه‌داده از دو ابزار Google و ترگمان استفاده کردیم. در این مجموعه‌داده برای هر تصویر سه سوال وجود دارد و برای هر سوال ۱۰ پاسخ موجود می‌باشد. در این مجموعه‌داده سه نوع سوال وجود دارد. نوع اول بله و خیر است. نوع دوم تعداد یک شی در تصویر است و نوع سوم مربوط به سوالات دیگر است. توزیع طول سوالات و پاسخ‌ها و ۳۰ پاسخ پرتکرار در این مجموعه‌داده را برای دو حالت ترجمه Google و ترگمان را در شکل های ؟؟ ، ؟؟ و ؟؟ می‌توانید مشاهده کنید.

### ۲.۳ LSTM Q + norm I [۱]

این روش ساده‌ترین روش یادگیری عمیق برای حل مسئله پرسش و پاسخ تصویری است. در اینجا مسئله VQA به عنوان یک مسئله طبقه‌بندی در نظر گرفته می‌شود که در آن ۱۰۰۰ پاسخ پرتکرار به عنوان کلاس‌ها انتخاب می‌شوند. ساختار کلی این شبکه در شکل ۲ نشان داده شده است. ابتدا با عبور دادن تصاویر از شبکه VGG19 برای هر تصویر یک بردار ویژگی ۴۰۹۶تایی در لایه‌ی ماقبل آخر در شبکه‌ی VGG19 تولید می‌شود. از طرفی دیگر با عبور سوال‌ها از لایه‌یEembedding برای هر کلمه موجود در سوال یک بردار ۳۰۰تایی تولید می‌شود. سپس از طریق ۲ لایه LSTM بردار ویژگی معنایی سوال استخراج می‌شود.




شکل ۳: ساختار کلی روش SAN .

هر یک از بردارهای ویژگی تصویر و سوال را به یک لایه‌ی Dense ۱۰۲۴ واحدی می‌دهیم تا ابعاد بردارها مشابه هم شوند. برای ترکیب بردار ویژگی سوال و تصویر از ضرب نقطه‌ای استفاده می‌کنیم. از این بردار ترکیب شده به عنوان ورودی برای لایه‌ی کاملا متصل استفاده می‌کنیم و در نهایت با عبور از یک لایه softmax کلاس(پاسخ) پیش‌بینی شده بدست می‌آید.

## ۳.۳ Stacked Attention Network [۱۴]

ایده‌ی اصلی روش SAN این است که ابتدا از سوال، یک بازنمایی معنایی و مفهومی استخراج می‌کند. سپس از آن به عنوان یک کوئری برای پیدا کردن مناطقی از تصویر که مرتبط با سوال است؛ استفاده می‌کند. غالباً در مسئله VQA نیاز است تا چندین مرحله استدلال صورت بگیرد. بنابراین در این شبکه از چندین لایه برای جستجو در تصویر استفاده می‌کنیم تا به تدریج به جواب مورد نظر برسیم. ساختار کلی شبکه SAN را در شکل ۳ می‌توانید مشاهده کنید. شبکه SAN از سه جز اصلی تشکیل شده است: ۱) مدل تصویر که با استفاده از CNN ویژگی‌های سطح بالایی را از تصویر استخراج می‌کند. ۲) مدل سوال که با استفاده از CNN یا LSTM ویژگی‌های معنایی سوال را استخراج می‌کند. ۳) مدل stacked attention که از طریق استدلال چند مرحله‌ای مناطقی از تصویر که مرتبط به سوال است را پیدا می‌کند تا پاسخ را پیش‌بینی کند.

### ۱.۳.۳ مدل تصویر

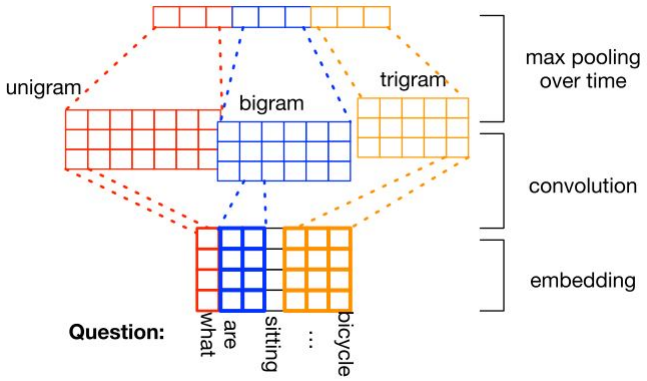
در این بخش برای استخراج ویژگی از شبکه‌ی VGG16 استفاده می‌کنیم و ویژگی‌ها را از آخرین لایه‌ی pooling شبکه بدست می‌آوریم. ابتدا تمام تصاویر را به $448\times448$ تغییر سایز می‌دهیم و بعد از این که تابع پیش‌پردازش موجود برای شبکه‌ی VGG16 را بر روی تصاویر اعمال کردیم، تصاویر را برای استخراج ویژگی به شبکه می‌دهیم. بنابراین برای هر تصویر یک ویژگی با ابعاد $14\times14\times512$ حاصل می‌شود. در حقیقت، برای هر تصویر به تعداد $14\times14$ منطقه استخراج می‌شود که هر منطقه به وسیله‌ی یک بردار ویژگی ۵۱۲تایی بازنمایی می‌شود. برای راحتی، از یک لایه‌ی Dense بعد از شبکه‌ی VGG16 استفاده می‌کنیم تا ابعاد بردار ویژگی مناطق مشابه با ابعاد بردار ویژگی سوال شود.

### ۲.۳.۳ مدل سوال

برای استخراج ویژگی‌های معنایی از سوال، از هر دو روش LSTM و CNN یک بعدی استفاده می‌کنیم. در هر دو روش ابتدا سوال را به یک دنباله‌ی عددی تبدیل می‌کنیم و سپس این دنباله‌ها را به یک لایه‌ی Embedding می‌دهیم. در روش LSTM خروجی لایه Embedding را به دو لایه‌ی LSTM می‌دهیم و خروجی آخرین لایه‌ی مخفی LSTM را به عنوان بردار ویژگی سوال در نظر می‌گیریم. در روش CNN خروجی Embedding را به سه لایه‌ی کانولوشنی یک بعدی با فیلترهایی با سایز ۱، ۲، ۳ می‌دهیم که به ترتیب ترکیب‌های یک کلمه‌ای ، دو کلمه‌ای و سه کلمه‌ای را برای ما استخراج می‌کند. در نهایت بر روی خروجی هر سه لایه تابع maxpooling را اعمال می‌کنیم و با قرار دادن این سه خروجی در کنار هم بردار ویژگی سوال بدست می‌آید. شکل ۴ مدل سوال بر اساس CNN را نشان می‌دهد.

### ۳.۳.۳ مدل stacked attention

در این بخش، مدل stacked attention با توجه به ماتریس ویژگی تصویر و بردار ویژگی سوال پاسخ را از طریق استدلال چند مرحله‌ای پیش‌بینی می‌کند. در بسیاری از موارد، یک پاسخ فقط مربوط به یک ناحیه کوچک از تصویر است. بنابراین، استفاده از یک ماتریس ویژگی کلی برای تصویر می‌تواند به دلیل وجود نویزهای مناطق بی‌ربط به پاسخ، منجر به نتایج نامطلوبی




شکل ۴: مدل سوال براساس CNN .

شود. در عوض، استدلال از طریق چندین لایه توجه، قادر است به تدریج مناطق غیرمرتبط با جواب را فیلتر کند و از ماتریس ویژگی تصویر حذف کند. بدین منظور ماتریس ویژگی تصویر $v_I$ و بردار ویژگی سوال $v_Q$ ، را به یک لایه Dense می‌دهیم و خروجی این لایه را به یک تابع softmax می‌دهیم تا توزیع توجه را بر روی نواحی تصویر بدست آوریم. بنابراین داریم:

$$h_A = tanh(W_{I,A}v_I \oplus (W_{Q,A}v_Q + b_A)) \quad (۱)$$

$$p_I = softmax(W_P h_A + b_P) \quad (۲)$$

بر اساس توزیع توجه $p_i$ ، جمع وزن‌دار بردارهای تصویر را که هر کدام متناظر به یک منطقه هست را محاسبه می‌کنیم. سپس $\tilde{v}_I$ را با بردار ویژگی سوال ترکیب می‌کنیم و یک کوئری برای لایه‌ی بعدی توجه ایجاد می‌کنیم.

$$\tilde{v}_I = \sum_i p_i v_i, \quad (۳)$$

$$u = \tilde{v}_I + v_Q. \quad (۴)$$

این روش را به تعداد k بار تکرار می‌کنیم. در نهایت از u در لایه‌ی k برای پیش‌بینی پاسخ استفاده می‌کنیم:

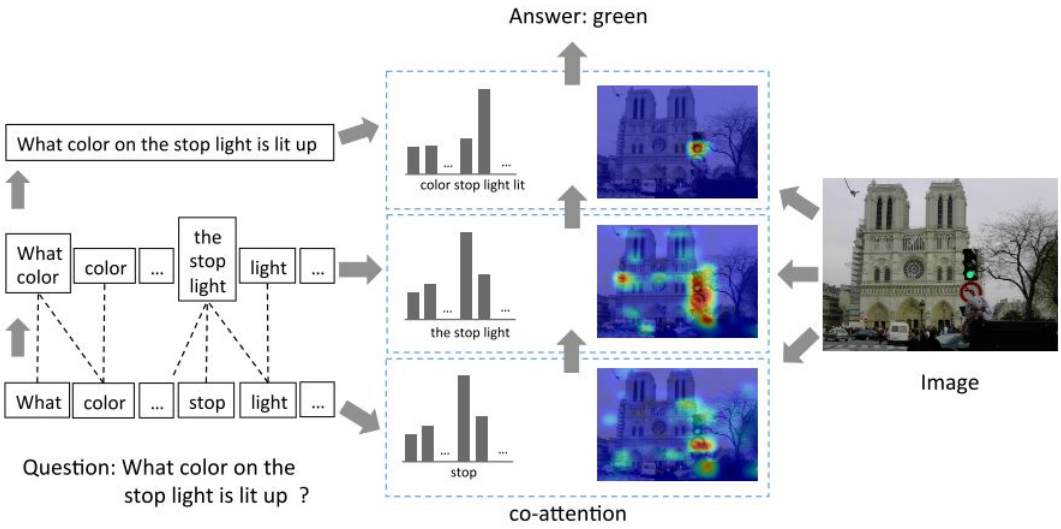
$$p_{ans} = softmax(W_u u^K + b_u) \quad (۵)$$

### ۴.۳ HieCoAttention [۷]

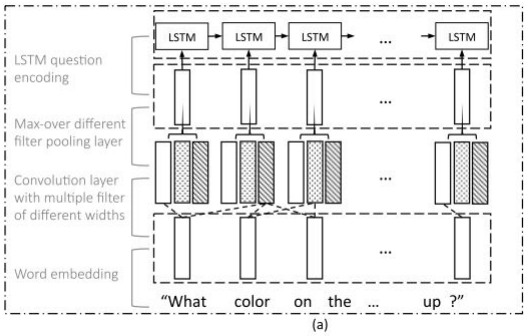
روش پیشنهاد شده در [] دارای دو ویژگی مهم است. ویژگی اول بازنمایی سلسله‌مراتبی سوال و ویژگی دوم مکانیزم coattention می‌باشد. در ادامه این دو خصوصیات را شرح می‌دهیم.

#### ۱.۴.۳ بازنمایی سلسله‌مراتبی سوال

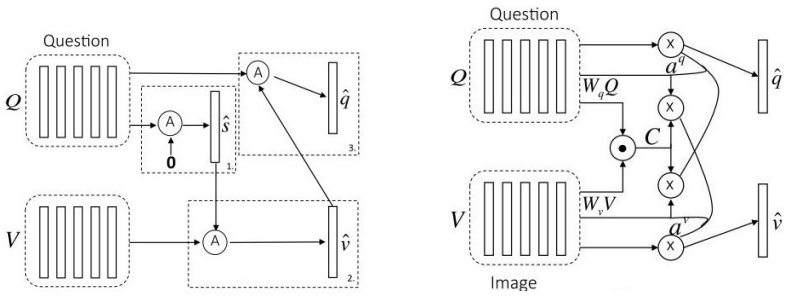
در این بخش برای هر سوال سه سطح Embedding را محاسبه می‌کنیم. اولین Embedding مربوط به کلمات است که بعد از این‌که سوال را به دنباله‌های عددی تبدیل کردیم؛ با عبور دادن این دنباله‌ها از لایه‌ی Embedding ، بردارهای Embedding کلمات بدست می‌آید. برای محاسبه سطح بعدی Embedding که مربوط به عبارات است از کانولوشن‌های یک بعدی با فیلترهایی با سایز ۱، ۲ و ۳ استفاده می‌کنیم و سپس با اعمال تابع Maxpooling بردار Embedding هر عبارت بوجود می‌آید. در نهایت از Embedding عبارات برای محاسبه‌ی Embedding کل سوال استفاده می‌کنیم. این کار توسط یک لایه LSTM انجام می‌شود. بنابراین برای هر سوال به صورت سلسله‌مراتبی سه سطح Embedding کلمه، عبارت و سوال تولید می‌شود. بازنمایی سلسله‌مراتبی سوال در شکل ۶ به تصویر کشیده شده است.



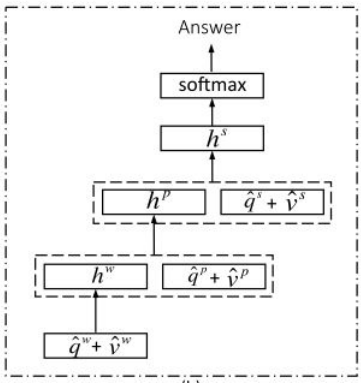

شکل ۵: ساختار کلی روش HieCoAttention




شکل ۶: بازنمایی سلسله‌مراتبی سوال.




شکل ۷: (آ) parallel coattention (ب) alternating coattention




شکل ۸: پیش‌بینی پاسخ

### ۲.۴.۳ مکانیزم coattention

در [] دو مکانیزم برای coattention پیشنهاد شده است که از نظر ترتیب تولید attention map برای سوال و تصویر با هم تفاوت دارند. اولین مکانیزم که parallel coattention نامیده می‌شود، باعث تولید attention به طور همزمان برای سوال و تصویر می‌شود. به مکانیزم دوم alternating coattention می‌گویند که برای تولید attention برای سوال و تصویر به صورت تناوبی عمل می‌کند (شکل ۷ ). این مکانیزم coattention در هر سه سطح سلسله‌مراتبی سؤال اجرا می‌شوند. در این پروژه ما از مکانیزم parallel coattention استفاده می‌کنیم. در این مکانیزم با محاسبه شباهت بین ویژگی‌های تصویر و سوال، تصویر و سؤال را به هم متصل می‌کنیم. اگر بردار ویژگی تصویر را با V و بازنمایی سوال را با Q نشان دهیم؛ ماتریس شباهت C به صورت زیر محاسبه می‌شود:

$$C = tanh(Q^T W_b V) \tag{۶}$$

پس از محاسبه ماتریس شباهت، برای محاسبه بردار وزن‌های attention برای تصویر و سوال از روابط زیر استفاده می‌کنیم:

$$\begin{aligned} H^v &= tanh(W_v V + (W_q Q)C), \quad & H^q &= tanh(W_q Q + (W_v V)CT) \\ a^v &= softmax(w_{hv}^T H^v), \quad & a^q &= softmax(w_{hq}^T H^q) \end{aligned} \tag{۷}$$

که در عبارت ۷ $W_v$ ، $W_q$ ، $w_{hv}$ و $w_{hq}$ پارامترهای وزن هستند. $a_v$ و $a_q$ نیز به ترتیب وزن‌های attention برای تصویر و سوال هستند. با توجه به وزن‌های attention ، بردارهای توجه تصویر و سوال به وسیله‌ی جمع وزن‌دار ویژگی‌های تصویر و ویژگی‌های سوال با وزن‌های attention محاسبه می‌شوند:

$$\hat{v} = \sum_{n+1}^{N} a_n^v v_n, \quad \hat{q} = \sum_{t=1}^{T} a_t^q q_t \tag{۸}$$

### ۳.۴.۳ پیش‌بینی پاسخ

ما پاسخ را بر اساس coattention تصویر و سوال بدست آمده در هر سه سطح Embedding پیش‌بینی می‌کنیم. از یک پرسپترون چندلایه (MLP) استفاده می‌کنیم تا ویژگی‌های attention را همان طور که در شکل ۸ نشان داده شده است؛ ترکیب کنیم.

$$\begin{aligned} h^w &= tanh(W_w(\hat{q}^w + \hat{v}^w)) \\ h^p &= tanh(W_p[\hat{q}^p + \hat{v}^p), h^w]) \\ h^s &= tanh(W_s[(\hat{q}^s + \hat{v}^s), h^p]) \\ p &= softmax(W_h h^s) \end{aligned} \tag{۹}$$

$W_W$ ، $W_p$ ، $W_s$ و $W_h$ پارامترهای وزن هستند. p احتمال پاسخ نهایی است.

# ۴ نتایج

## ۱.۴ LSTM Q + norm I

| | Google Translation | | | | Targoman Translation | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** | **yes/no** | **Number** | **Other** | **All** |
| **lstm Q + VGG19** | 76.14 | 32.97 | 35.78 | 50.53 | 75.58 | 32.61 | 33.53 | 49.15 |

جدول ۲

| | English-paperToken | | | | English-kerasToken | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** | **yes/no** | **Number** | **Other** | **All** |
| **lstm Q + VGG19** | 78.43 | 33.7 | 37.99 | 52.58 | 78.53 | 31.91 | 38.78 | 52.79 |

جدول ۳

| | Google Translation | | | |
|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** |
| **BilstmQ+resNet152** | 76.46 | 31.63 | 38.6 | 51.89 |
| **lstmQ+resNet152** | 76.83 | 31.75 | 38.77 | 52.13 |
| **CNNQ+resNet152** | 78.34 | 31.91 | 38.98 | 52.82 |

جدول ۴

## ۲.۴ Stacked Attention Network

| | Google Translation | | | | Targoman Translation | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** | **yes/no** | **Number** | **Other** | **All** |
| **SAN_LSTM_2** | 77.83 | 33.19 | 39.08 | 52.84 | 75.95 | 31.61 | 36.82 | 50.81 |
| **SAN_CNN_2** | 77.49 | 33.17 | 39.18 | 52.76 | 76.48 | 32.29 | 37.37 | 51.37 |

جدول ۵

| | Google Translation | | | |
|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** |
| **SAN_LSTM_1** | 77.46 | 32.23 | 38.35 | 52.22 |
| **SAN_LSTM_2** | 77.83 | 33.19 | 39.08 | 52.84 |
| **SAN_LSTM_3** | 77.12 | 32.56 | 38.62 | 52.27 |

جدول ۶

## ۳.۴ HieCoAttention

| | Google Translation | | | | Targoman Translation | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **yes/no** | **Number** | **Other** | **All** | **yes/no** | **Number** | **Other** | **All** |
| **CoAttention** | 76.62 | 32.7 | 38.12 | 51.85 | 74.18 | 32.41 | 32.47 | 48.07 |

جدول ۷

## ۵ تحلیل

لورم ایپسوم متن ساختگی با تولید سادگی نامفهوم از صنعت چاپ و با استفاده از طراحان گرافیک است. چاپگرها و متون بلکه روزنامه و مجله در ستون و سطرآنچنان که لازم است و برای شرایط فعلی تکنولوژی مورد نیاز و کاربردهای متنوع با هدف بهبود ابزارهای کاربردی می باشد.

## مراجع

## References


[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[8] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.

[9] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–166, 2018.

[10] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[13] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[14] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.