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. 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 ...

۱ مقدمه

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.

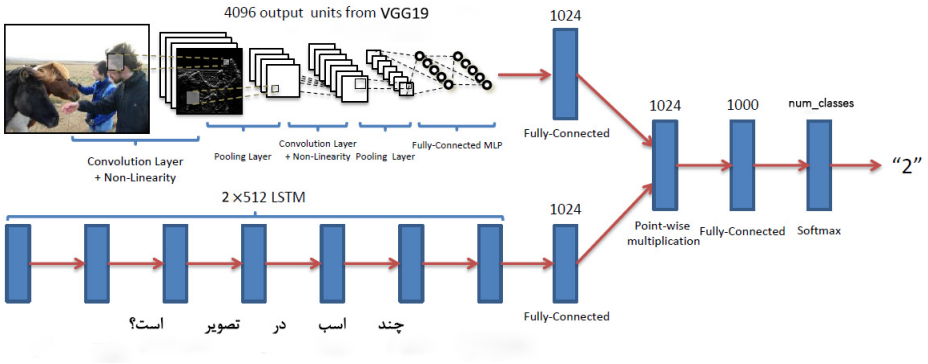
۲ کارهای مرتبط / پیش‌زمینه

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.

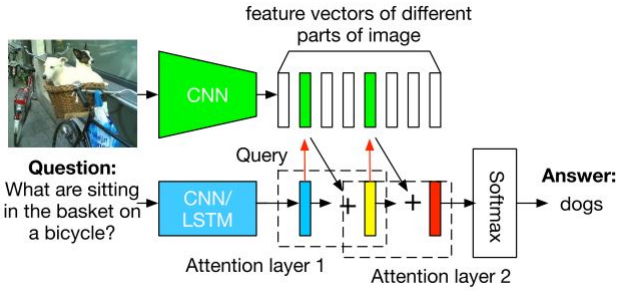
۳ مدل پیشنهاد شده

LSTM Q + norm I ۱.۳

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 ۱ 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. 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.



شکل ۱: ساختار کلی روش LSTM Q + norm I .



شکل ۲: ساختار کلی روش SAN .

۲.۳ Stacked Attention Network

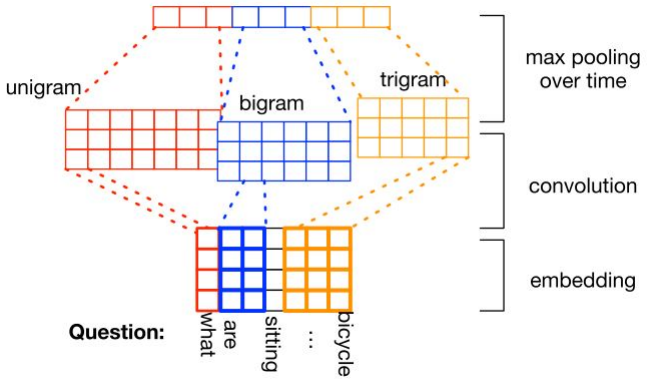
ایده‌ی اصلی روش SAN این است که ابتدا از سوال، یک بازنمایی معنایی و مفهومی استخراج می‌کند. سپس از آن به عنوان یک کوئری برای پیدا کردن مناطقی از تصویر که مرتبط با سوال است؛ استفاده می‌کند. غالباً در مسئله VQA نیاز است تا چندین مرحله استدلال صورت بگیرد. بنابراین در این شبکه از چندین لایه برای جستجو در تصویر استفاده می‌کنیم تا به تدریج به جواب مورد نظر برسیم. ساختار کلی شبکه SAN را در شکل ۲ 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.

۱.۲.۳ 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 448×448 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 $512\times14\times14$ حاصل می‌شود. در حقیقت، برای هر تصویر به تعداد 14×14 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.

۲.۲.۳ 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،



شکل ۳: مدل سوال براساس CNN .

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 ۳ مدل سوال بر اساس CNN را نشان می‌دهد.

۳.۲.۳ 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 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:

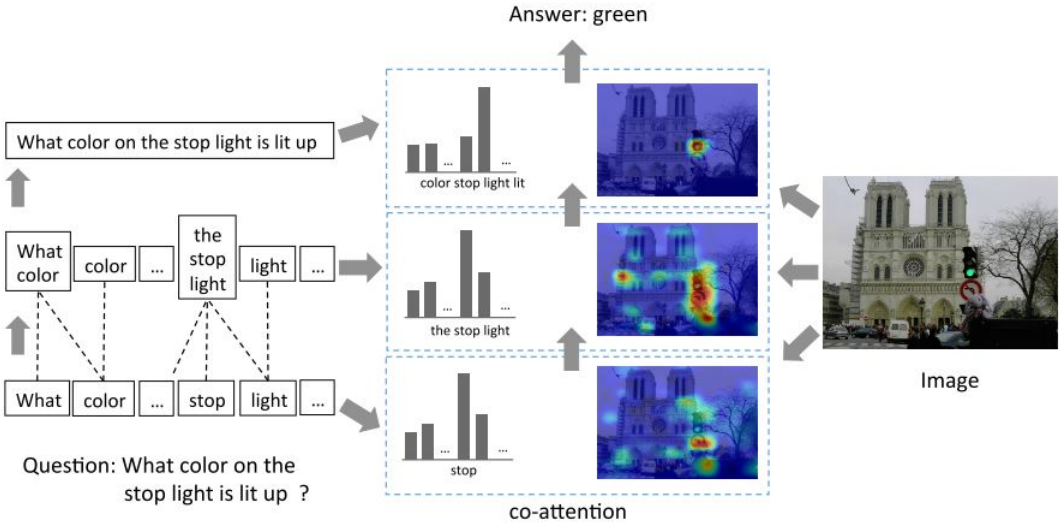
$$p_{ans} = softmax(W_u u^K + b_u) \quad (۵)$$

۳.۳ HieCoAttention

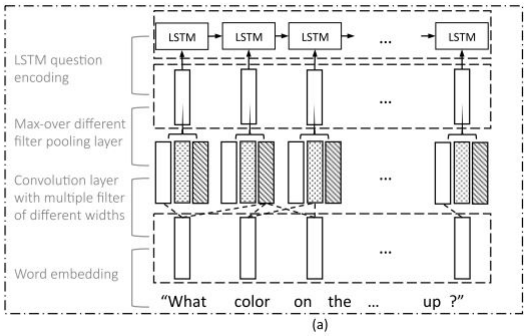
روش پیشنهاد شده در دارای دو ویژگی مهم است. ویژگی اول بازنمایی سلسله‌مراتبی سوال و ویژگی دوم مکانیزم coattention می‌باشد. در ادامه این دو خصوصیات را شرح می‌دهیم.

۱.۳.۳ 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



شکل ۴: ساختار کلی روش HieCoAttention



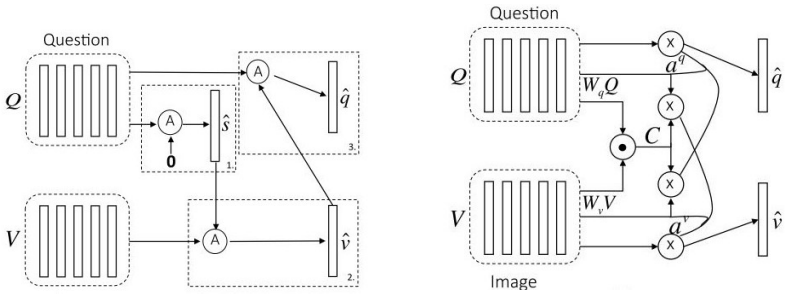
شکل ۵: بازنمایی سلسله‌مراتبی سوال.

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 ۵ به تصویر کشیده شده است.

۲.۳.۳ مکانیزم coattention

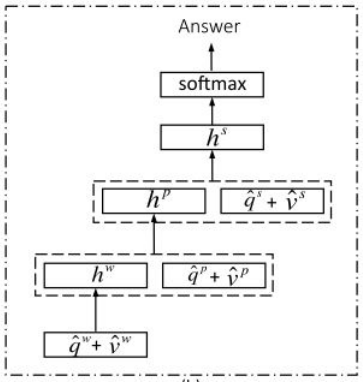
در دو مکانیزم برای coattention پیشنهاد شده است که از نظر ترتیب تولید attention map برای سوال و تصویر با هم تفاوت دارند. اولین مکانیزم که parallel coattention نامیده می‌شود، باعث تولید attention به طور همزمان برای سوال و تصویر می‌شود. به مکانیزم دوم alternating coattention می‌گویند که برای تولید attention برای سوال و تصویر به صورت تناوبی عمل می‌کند (شکل ۶). 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:

$$C = tanh(Q^T W_b V) \qquad (۶)$$



(ب) (آ)

شکل ۶: (آ) parallel coattention (ب) alternating coattention



شکل ۷: پیش‌بینی پاسخ

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:

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

۳.۳.۳ 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 ۷ 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

نتایجلورم 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.

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.

منابع