



دانشکده مهندسی کامپیوتر

ارائه راهکاری برای یکپارچه سازی دستگاه‌های اینترنت
اشیاء با سکوی کوبرنیتزر

پروژه کارشناسی مهندسی کامپیوتر گرایش هوش مصنوعی

سینا شعبانی کومله

استاد راهنما

دکتر محسن شریفی

تابستان ۱۴۰۲

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: سینا شعبانی کومله

عنوان پروژه: ارائه راهکاری برای یکپارچه سازی دستگاه‌های اینترنت اشیاء با سکوی کوبرنیتز

تاریخ دفاع: تابستان ۱۴۰۲

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امض
۱	استاد راهنما	دکتر محسن شریفی	استاد تمام	دانشگاه علم و صنعت ایران	
۲	داور نهایی	دکتر TODO	دانشیار	دانشگاه علم و صنعت ایران	

ب

تأییدیهی صحت و اصالت نتایج

با اسمه تعالی

اینجانب سینا شعبانی کومله به شماره دانشجویی 97521351 رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیهی نتایج این پروژه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احراق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذیصلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: سینا شعبانی کومله

تاریخ و امضا:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنمای شرح زیر

تعیین می‌شود، بلامانع است:

- بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.
- بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنمای، بلامانع است.
- بهره‌برداری از این پایان‌نامه تا تاریخ ممنوع است.

استاد راهنمای: دکتر محسن شریفی

تاریخ:

امضا:

چکیده

در حال حاضر، مدیریت و نظارت بر دستگاه‌های اینترنت اشیاء^۱ به یک چالش عمدۀ تبدیل شده است. راه حل‌های موجود برای کنترل و نظارت بر این دستگاه‌ها اغلب ناسازگاری‌ها و محدودیت‌هایی دارند که موجب کاهش کارایی و پیچیدگی مدیریت در مقیاس بالا می‌شوند. به منظور حل این مسئله، این پروژه تلاش می‌کند تا با استفاده از کوبرنیت^۲ و پروژه کوبلت^۳ مجازی^۴ یک سازوکار جامع برای نظارت و مدیریت دستگاه‌های اینترنت اشیاء ارائه دهد. انگیزه اصلی پروژه متتمرکز کردن کنترل و نظارت بر دستگاه‌های اینترنت اشیاء به صورت یکپارچه و موثر است. راه حل‌های کنونی اغلب ناسازگاری‌هایی با استانداردها و فناوری‌های مختلف دستگاه‌های اینترنت اشیاء دارند و به تنها‌یی قادر به ارائه یک محیط یکپارچه برای مدیریت و نظارت نیستند. این پروژه شامل سه بخش اصلی، یعنی تامین‌کننده^۵، کنترلکننده^۶ و دستگاه‌ها^۷ این بخشها با یکدیگر ارتباط برقرار می‌کنند تا اطلاعات مفیدی درباره دستگاه‌های اینترنت اشیاء مورد کنترل ارائه دهند و این اطلاعات را در دسترس خوش کوبرنیت قرار دهند.

واژگان کلیدی: اینترنت اشیاء، کوبرنیت، کوبلت مجازی، نظارت یکپارچه، پایش مقیاس پذیر

(IoT) Things Of Internet^۱
Kubernetes^۲
Kubelet Virtual^۳
Provider^۴
Controller^۵
Device^۶

فهرست مطالب

ح

فهرست تصاویر

د

فهرست جداول

۱

فصل ۱: مقدمه

1	شرح مسأله	1-1
2	اهداف پژوهش	2-1
3	ساختار گزارش	3-1

۴

فصل ۲: مفاهیم پایه

4	مقدمه	1-2
4	شبکه‌های عصبی مصنوعی	2-2
5	شبکه‌های عصبی بازگشتی	3-2
6	حافظه طولانی کوتاه‌مدت	1-3-2
7	واحد بازگشتی دروازه‌دار	2-3-2
9	مدل‌های دنباله‌به‌دنباله	4-2
9	mekanizm توجه	5-2
10	Mekanizm توجه به‌اهدانا (Bahdanau)	1-5-2
11	Mekanizm توجه لوائگ (Luong)	2-5-2
11	Mekanizm توجه به خود	3-5-2
12	مدل‌های transformer و Mekanizm توجه آنها	6-2

ث

15	BERT	7-2
15	مدل‌های تک‌جریان	1-7-2
16	مدل‌های دو‌جریان	2-7-2
16	جمع‌بندی	8-2
17	کارهای مرتبط	فصل 3:
17	مقدمه	1-3
17	روش‌های متفاوت حل مسئله پرسش‌وپاسخ تصویری	2-3
17	روش‌های بر پایه Bilinear Pooling	1-2-3
18	روش‌های بر پایه توجه	2-2-3
19	ادغام روابط اشیاء	3-2-3
20	تولید پاسخ برای پرسش‌وپاسخ تصویری	3-3
21	مجموعه داده‌های منتشر شده در پرسش‌وپاسخ تصویری	4-3
22	مدل‌های از پیش آموزش داده شده	5-3
23	روش پیشنهادی	فصل 4:
23	مقدمه	1-4
23	معماری سیستم	2-4
24	ورودی اولیه و خروجی نهایی	3-4
25	Encoder مدل	4-4
25	[?] LXMERT مدل	1-4-4
26	[?] VisualBERT مدل	2-4-4
26	Decoder مدل	5-4
27	جمع‌بندی	6-4
28	ارزیابی روش پیشنهادی	فصل 5:
28	مقدمه	1-5

28	معیارهای برپایه شباهت نحوی	2-5
29	معیارهای برپایه بردارهای تعبیه	3-5
29	معیار [?]BERTScore	1-3-5
29	معیار Average Score	2-3-5
30	مجموعه‌داده‌ی مورد استفاده	4-5
30	نتایج ارزیابی سیستم پیشنهادی	5-5
30	ارزیابی و مقایسه معماری‌های ارجح	1-5-5
31	مقایسه شبکه‌های عصبی بازگشتی	2-5-5
31	ارزیابی و مقایسه شبکه‌های عصبی بازگشتی و مکانیزم توجه سراسری	3-5-5
32	ارزیابی و مقایسه مدل‌های برپایه تبدیل‌شونده‌ها	4-5-5
33	ارزیابی انسانی	6-5
36	جمع‌بندی	7-5
37	فصل 6: نتیجه‌گیری و کارهای آینده	
37	نتیجه‌گیری	1-6
38	دستاوردها	2-6
38	کارهای آینده	3-6
39	واژه‌نامه فارسی به انگلیسی	
41	واژه‌نامه انگلیسی به فارسی	

فهرست تصاویر

2	یک نمونه از جفت پرسش و تصویر همراه با پاسخ مورد انتظار	1-1
5	معماری کلی شبکه‌های عصبی بازگشتی	1-2
6	معماری حافظه کوتاه‌مدت طولانی‌مدت	2-2
8	معماری شبکه واحد بازگشتی دروازه‌دار	3-2
9	نمونه‌ای از معماری کدگذار-کدگشا برای حل مسئله ترجمه ماشینی	4-2
10	mekanizm توجه باهدانا در ترجمه ماشینی	5-2
12	نمونه‌ای از توجه به خود در یک جمله	6-2
13	مدل تبدیل‌شونده و معماری آن	7-2
14	توجه چندسر و توجه مقیاس‌شده بر پایه ضرب داخلی	8-2
15	نمونه‌ای از یک مدل تک‌جریان	9-2
16	نمونه‌ای از یک مدل دو‌جریان	10-2
18	روش MCB برای ادغام تصویر و زبان	1-3
19	توجه پشته‌ای برای حل مسئله پرسش و پاسخ تصویری	2-3
19	حل پرسش و پاسخ تصویری با ساختارهای گرافی	3-3
20	چهارچوب یکپارچه ارائه شده برای حل مسائل چند مازول به روش تولید متن	4-3
21	نمونه‌ای مجموعه داده VCR که علاوه بر ارائه پاسخ، نیازمند ارائه دلیلی برای پاسخ است	5-3
23	شماتیک سیستم پرسش و پاسخ تصویری	1-4
24	نمونه‌ای از کارکرد Tokenizer	2-4

فهرست تصاویر

خ		
25	معماری مدل LXMERT همراه با ورودی و خروجی	3-4
26	میزان توجه بخش متن به تصاویر در VisualBERT	4-4
27	ساختار مدل Autoregressive Decoder	5-4
35	نتایج ارزیابی انسانی	1-5

فهرست جداول

تاییج معماری‌های متفاوت بر مجموعه داده FSVQA. اعداد موجود در نام کدگشا نشانگر تعداد لایه‌های آن است.	1-5
31	
تاثیر معماری‌های متفاوت در کدگشاها بر پایه شبکه‌های عصبی بازگشتی	2-5
32	
بررسی شبکه‌های عصبی بازگشتی با مکانیزم توجه	3-5
33	
استفاده از مدل‌های تبدیل شونده به عنوان کدگشا	4-5
33	
چند نمونه از پاسخ‌ها و اشتباهات آن‌ها بر اساس دسته‌بندی	5-5
34	

فصل 1

مقدمه

1-1 شرح مسائله

یک عامل برای انجام اقدام مناسب در محیط، باید بتواند محیط اطراف خودش را درک کند. یکی از مهمترین روش‌ها، درک محیط اطراف از طریق دیدن است. در سال‌های اخیر، حوزه پردازش تصویر به خصوص با وجود ابزاری نظیر یادگیری عمیق پیشرفت چشمگیری داشته‌اند. با توجه به وجود داده‌های زیاد در این زمینه‌ها، محققان در حل مسائلی نظیر دسته‌بندی تصاویر [?] ، تشخیص اشیاء [?] و مسائل کاربردی مشابه به خوبی عمل کرده‌اند به‌گونه‌ای که می‌توان ادعا کرد سیستم توانایی حل مسئله دسته‌بندی تصاویر را با دقیقی بیشتر از انسان‌ها دارد [?] ! با این حال، این مسائل از جمله مسائل ساده هستند و نیاز به درک دقیق اجزاء تصویر و ارتباطات بین آنها ندارند. به عبارتی دیگر، به عنوان یک انسان، به سادگی می‌توانیم اشیاء موجود در تصویر را تشخیص دهیم و موقعیت آنها را تعیین کنیم.

یکی دیگر از توانایی‌های مهم و کلیدی عوامل مصنوعی، ارتباط با انسان‌ها است. یک عامل مصنوعی باید بتواند با انسان‌ها به ساده‌ترین روش، که همان زبان طبیعی انسان‌هاست، با آن‌ها ارتباط برقرار کند. پردازش زبان طبیعی را می‌توان زیرمجموعه‌ای از زبان‌شناسی بین انسان‌ها و ماشین‌ها در نظر گرفت.

مسائل متفاوتی نظیر عنوان‌نویسی تصاویر، پرسش‌وپاسخ تصویری و تولید تصاویر از عنایون وجود دارند که پردازش تصویر و پردازش زبان طبیعی را در کنار یکدیگر استفاده کرده‌اند. تا مدتی قبل، پرسش‌وپاسخ تصویری یک اغراق‌نمایی بود تا اینکه یک مجموعه داده در این رابطه منتشر شد. مجموعه داده پرسش‌وپاسخ تصویری [?] یک وظیفه جدید را برای حل با خود به همراه آورد که با توجه به پرسش‌ها، نیازمند زیر وظیفه‌های متفاوت نظیر تشخیص اشیاء (تصویر چه شیئی را

2-1. اهداف پژوهش

نشان میدهد؟)، تشخیص مشخصات (این شیء چه رنگی دارد؟) و موارد مشابه است. با کاوش کردن مجموعه داده پرسش‌پاسخ تصویری متوجه تک‌کلمه‌ای بودن پاسخ‌ها می‌شویم که در حال حاضر از الگوریتم‌های دسته‌بندی برای حل این مسئله استفاده می‌شود. پاسخ دادن به سوال از طریق دسته‌بندی بین کلمات از پیش تعیین شده طبیعی به نظر نمی‌رسد. از آنجایی که هدف هوش مصنوعی حل کردن مسائل به روش انسان‌ها است و انسان‌ها نیز به صورت جمله به سوالات پاسخ می‌دهند، ارائه دادن پاسخ‌های تک‌کلمه‌ای برای پرسش‌های تصویری اندکی غیر طبیعی به نظر می‌رسد. از این رو ارائه دادن ساز و کاری برای پاسخ دادن به سوالات به صورت جمله می‌تواند علاوه بر طبیعی کردن مسئله اطلاعات اضافی در اختیار کاربر قرار دهد که منجر به تشخیص بهتر خطاهای و رفع ابهامات بسیاری می‌شود.

2-1 اهداف پژوهش

هدف اصلی این پژوهش امکان‌سنجی، طراحی، پیاده‌سازی و ارزیابی سیستمی خودکار مبتنی بر روش یادگیری عمیق برای پاسخ‌گویی به سوالات مربوط به یک تصویر به صورت جمله است. در این مسئله ورودی یک تصویر به همراه یک دنباله‌ای از کلمات به عنوان سوال است. وظیفه سیستم تولید دنباله‌ای از کلمات است که بتواند پاسخ سوال مربوط به تصویر را پاسخ دهد. شکل 1-1 یک نمونه از پرسش‌پاسخ تصویر به همراه پاسخ مورد انتظار را نشان می‌دهد.



Q: Is this a modern train?
A: No, this is not a modern train.

شکل 1-1: یک نمونه از جفت پرسش و تصویر همراه با پاسخ مورد انتظار

3-1 ساختار گزارش

در این پژوهه هدف ارائه روشی نو برای حل مسئله پرسش و پاسخ تصویری است. در ابتدا به معرفی مفاهیم پایه استفاده شده در این پژوهش و سپس به معرفی روش‌ها و کارهای مرتبط پرداخته خواهد شد. پس از آن به معرفی روش و ارزیابی آن پرداخته شده است. در انتها نتیجه‌گیری و کارهای آینده معرفی می‌شوند.

فصل 2

مفاهیم پایه

1-2 مقدمه

در این بخش به معرفی مفاهیم پایه از یادگیری عمیق و برخی معماری‌هایی که در این پژوهش استفاده شده‌اند به مانند شبکه‌های عصبی بازگشتی، مکانیزم توجه و معماری ترانسفورمرز و مدل‌های کدکار و کدگشا پرداخته شده است.

2-2 شبکه‌های عصبی مصنوعی

شبکه‌های عصبی مصنوعی سیستم‌ها و روش‌های محاسباتی نوینی هستند که از گروهی راس به نام نورون‌ها تشکیل شده‌اند که توسط وزن‌هایی به یکدیگر متصلند. یک شبکه عصبی از یک لایه ورودی، چندین لایه مخفی و یک لایه خروجی تشکیل شده است. اگر از دید ریاضیات این شبکه‌ها را بررسی کنیم، می‌توانیم دو لایه را از طریق ماتریس وزن‌ها به یکدیگر متصل کنیم. هر لایه از شبکه‌های عصبی تبدیل به خصوصی را به ورودی‌های خود اعمال می‌کنند. در نهایت خروجی‌های هر لایه مطابق با معادله 1-2 محاسبه می‌شوند و به لایه‌های بعدی منتقل می‌شوند.

$$LayerOutput = A(XW + B) \quad (1-2)$$

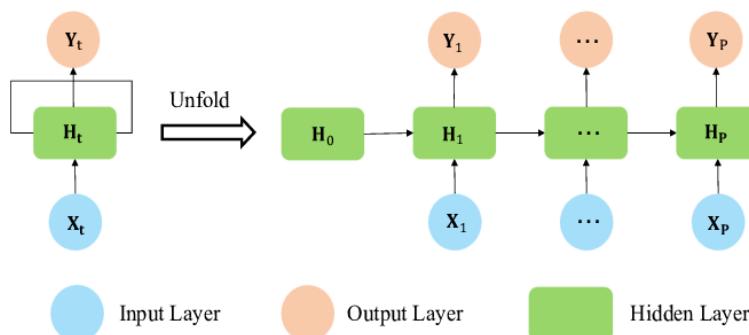
3-2. شبکه‌های عصبی بازگشتی

به هر شبکه مصنوعی یکتابع هزینه نسبت داده شده است که برای کاهش آن از مکانیزم‌های بهینه‌سازی استفاده می‌شود. این مکانیزم‌ها برای توابع مشتق‌پذیر برپایه گرادیان کاهشی و برای توابع مشتق‌نایپذیر برپایه الگوریتم‌های ژنتیک بنا شده‌اند. بر اساس قضیه تقریب، هر تابع پیوسته را در فضای \mathbb{R}^n را می‌توان با یک لایه مخفی و یک تابع فعال‌ساز تقریب زد. این قضیه را می‌توان نقطه قوت شبکه‌های عصبی مصنوعی دانست. دلایل اصلی پیشرفت‌های اخیر شبکه‌های عصبی می‌توان وجود داده در دسترس، توانایی آموزش مدل‌ها از طریق پردازنده‌های گرافیکی و وجود کتابخانه‌های مشتق‌گیر نظیر [?] و Tensorflow [?].

3-2 شبکه‌های عصبی بازگشتی

باید توجه داشت که استفاده از شبکه‌های عصبی مصنوعی برای داده‌های دنباله‌ای عملی نیست. به عبارتی دیگر در شبکه‌های عصبی مصنوعی فرض می‌کنیم که داده‌های ورودی به صورت کامل مستقل از یکدیگرند. با توجه به این فرضیه شبکه‌های عصبی مصنوعی عادی برای استخراج وابستگی بین دنباله‌ها (برای مثال کلمات درون یک جمله) کارآمد نیستند. برای در نظر گرفتن این وابستگی، از شبکه‌های عصبی بازگشتی استفاده می‌شود.

شبکه‌های عصبی بازگشتی را می‌توان همان شبکه‌های عصبی عادی با یک حافظه در نظر گرفت. در واقع مهم‌ترین تفاوت شبکه‌های عصبی بازگشتی وجود حالت‌های مخفی است که می‌توانند اطلاعات گذشته را در حافظه نگه دارند و سپس از طریق پس انتشار خطا در راستای زمان به کاهش تابع هزینه کمک کنند. قابلیت به یاد سپردن گذشته شبکه‌های عصبی بازگشتی را بسیار کارآمد کرده است. در واقع شبکه‌های عصبی بازگشتی با تعداد پارامترهای محدود تورینگ کامل هستند و توانایی پیاده‌سازی هر الگوریتمی را دارند. [?]



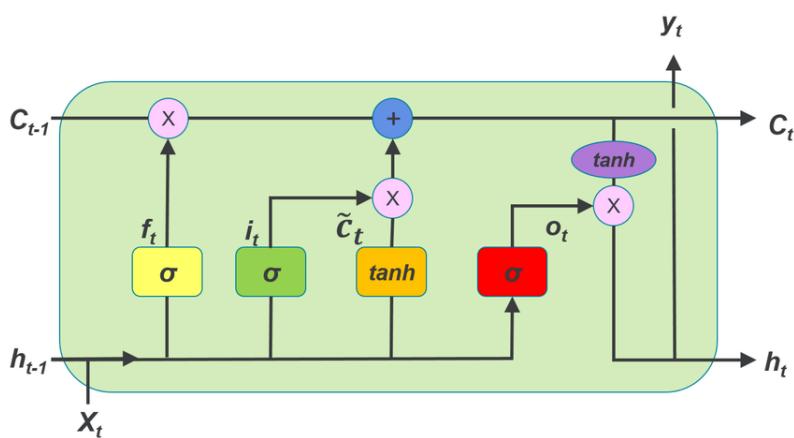
شکل 2-1: معماری کلی شبکه‌های عصبی بازگشتی

3-2. شبکه‌های عصبی بازگشتی

یکی از مشکلات متداول شبکه‌های عصبی بازگشتی، محو شدگی و انفجار گرادیان به صورت نمایی در گذر زمان است. این مسئله بدان معنی است که این معماری‌ها توانایی پردازش دنباله‌هایی با وابستگی‌های طولانی را ندارند. برای حل این مشکل معماری‌های متفاوتی ارائه شد که به بررسی مهم‌ترین آن‌ها می‌پردازیم.

1-3-2 حافظه طولانی کوتاه‌مدت

در سال 1977 یک معماری از شبکه‌های عصبی بازگشتی به نام حافظه طولانی کوتاه‌مدت ارائه شد. هر عنصر از شبکه شامل یک درگاه ورودی، یک سلول، یک درگاه فراموشی و یک درگاه خروجی است. سلول طولانی کوتاه‌مدت وظیفه یادآوری مقادیر دیده شده در گذشته را برعهده دارد، در حالی که درگاه‌های ورودی، فراموشی و خروجی وظیفه کنترل جریان داده‌ها را دارند.



شکل 2-2: معماری حافظه کوتاه‌مدت طولانی‌مدت

از آنجایی که هدف اصلی حل محو شدگی نمایی گرادیان در طول زمان بود، حافظه‌های طولانی کوتاه‌مدت قابلیت پردازش دنباله‌های طولانی‌تری را دارند. معادلات مرتبط با این شبکه‌ها در زیر آورده شده است.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2-2)$$

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \quad (3-2)$$

$$i_t = \sigma(W_i^T x_t + W_i^T h_{t-1} + b_i) \quad (4-2)$$

$$f_t = \sigma(W_f^T x_t + W_f^T h_{t-1} + b_f) \quad (5-2)$$

$$o_t = \sigma(W_o^T x_t + W_o^T h_{t-1} + b_o) \quad (6-2)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (7-2)$$

$$h_t = o_t \circ \tanh(c_t) \quad (8-2)$$

سلول (معادله 8-2) وابستگی‌های بین مقادیر ورودی در دنباله را نگه می‌دارد. درگاه ورودی (معادله 4-2) وظیفه انتخاب ویژگی‌های ورودی برای عبور به سلول را عهده‌دار است. درگاه فراموشی (معادله 5-2) میزان این که یک مقدار چقدر در سلول باقی بماند را برعهده دارد و در نهایت درگاه خروجی، میزان مشارکت مقادیر موجود در سلول را برای محاسبه خروجی را کنترل می‌کند.

2-3-2 واحد بازگشتی دروازه‌دار

در سال 2014 شبکه‌های واحد بازگشتی دروازه‌دار به عنوان معماری متفاوتی برای شبکه‌های عصبی بازگشتی ارائه شد تا مشکل محوش‌گی گرادیان در حال آموزش را حل کند. این واحد شباهت زیادی به حافظه طولانی کوتاه‌مدت دارد ولی در عین حال تعداد پارامترهای کمتری دارد و در نتیجه سریع‌تر آموزش می‌بیند.

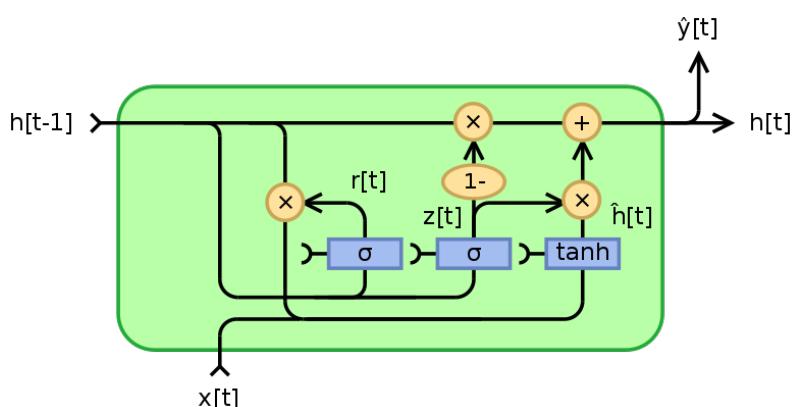
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (9-2)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (10-2)$$

$$\hat{h}_t = \tanh(W_h x_t + U_r(r_t \circ h_{t-1}) + b_h) \quad (11-2)$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \hat{h}_t \quad (12-2)$$

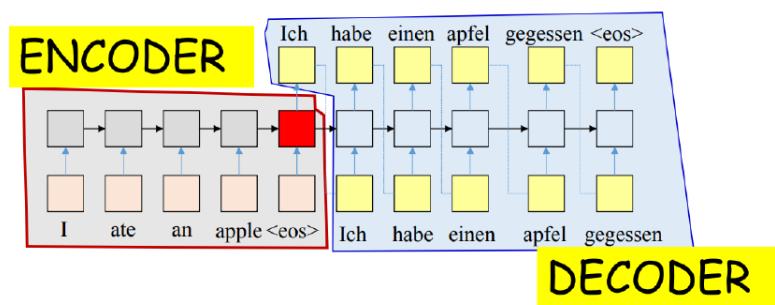
مطابق با معادله 9-2 درگاه z_t مسئولیت تعیین عبور مقادیر برای استفاده در آینده را دارد. به عبارتی اگر مقدار آن برابر 1 باشد، بدان معنی است که شبکه مقدار بردار h_t را به صورت کامل به روز و اطلاعات گذشته را فراموش می‌کند. متقابلاً اگر مقدار این درگاه به 0 نزدیک باشد، شبکه مقادیر h_t را از گذشته انتخاب می‌کند. درگاه فراموشی، مطابق با معادله 10-2 مسئولیت تعیین تاثیرگذاری مقادیر حالت قبلی در محاسبه حالت فعلی را بر عهده دارد.



شکل 2-3: معماری شبکه واحد بازگشتی دروازه‌دار

4-2 مدل‌های دنباله‌به‌دنباله

مدل‌های دنباله‌به‌دنباله وظیفه تبدیل دنباله‌ای دیگر را برعهده دارند. به طور کلی این مدل‌ها دنباله‌ای با طول متغیر را ورودی می‌گیرند و دنباله‌ای دیگر با طول متغیر بدل می‌کنند که لزماً طول آن با ورودی یکسان نیست. مدل‌های دنباله‌به‌دنباله دارای دو شبکه مجزا هستند. یک شبکه به عنوان کدگذار که وظیفه تبدیل ورودی‌ها به یک یا چند بردار ویژگی را برعهده دارد. این بردار ویژگی یک بازنمایی از دنباله ورودی است که حداکثر معانی قابل استخراج از دنباله ورودی را در بر دارد. شبکه دیگری با عنوان کدگشا برای تولید دنباله هدف آموزش می‌بیند. ورودی این بخش همان بردار ویژگی بدست آمده از کدگذار است. معمولاً شبکه‌های کدگذار و کدگشا با یکدیگر آموزش داده می‌شوند که به معنی پس انتشار خطا از کدگشا به کدگذار است. در شکل 4-2 نمونه‌ای از مدل‌های دنباله‌به‌دنباله را برای حل ترجمه ماشینی مشاهده می‌شود.



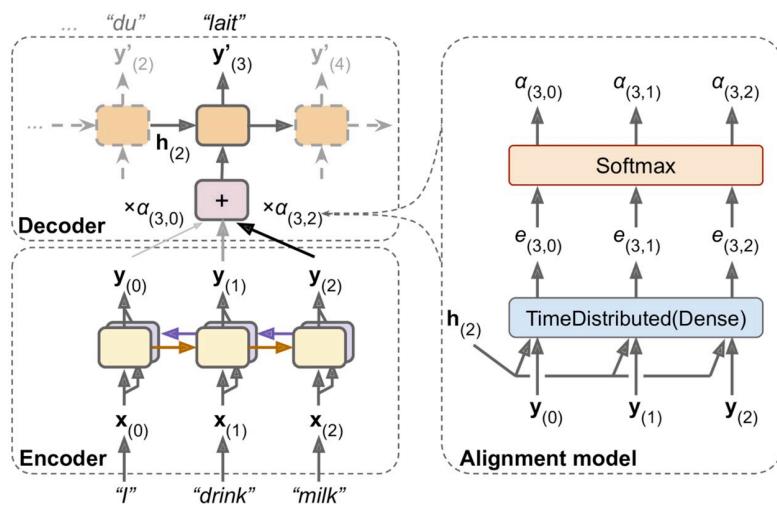
شکل 2-4: نمونه‌ای از معماری کدگذار-کدگشا برای حل مسئله ترجمه ماشینی

5-2 مکانیزم توجه

شبکه‌های عصبی بازگشتی و معماری‌های مشابه آن می‌توانند اطلاعات درون یک دنباله را پردازش کنند و بر اساس آن‌ها نتیجه گیری کنند. حال فرض کنیم که یک پاراگراف طولانی داریم و پس از خواندن آن سوالی پرسیده شود. پس از روبرو شدن با سوال، ممکن است متن را به صورت کامل به یاد نیاوریم و لازم باشد که دوباره با توجه بیشتری به برخی جملات متن را بخوانیم. بنابراین، توجه را می‌توانیم رفتار تمرکز بر روی بخش گستته‌ای از اطلاعات و در نظر نگرفتن دیگر اجزاء تعریف کنیم. در ادامه به ایده‌ها و انواع مکانیزم‌های توجه ارائه شده در طی سال‌های گذشته می‌پردازیم.

1-5-2 مکانیزم توجه باهدانا (Bahdanau)

در سال 2014 مکانیزمی [?] برای توجه به حالت‌های مخفی در محاسبه خروجی ارائه داد. در این معماری علاوه بر حالت سلول کدگذار آخر، تمام خروجی‌های کدگذار را به کدگشا می‌فرستیم تا در هر گام، کدگشا یک جمع وزن‌دار بین خروجی‌ها حساب کند. وزن‌های این جمع، مفهوم توجه را پیاده‌سازی می‌کنند. هر حالتی که وزن بیشتری داشته باشد توجه بیشتری را جلب می‌کند!



شکل 2-5: مکانیزم توجه باهدانا در ترجمه ماشینی

وزن‌های مذکور همان α موجود در شکل 5-2 که همراه با شبکه آموزش می‌بینند. با توجه به شکل متوجه می‌شویم به هنگام ترجمه کلمه lait، توجه بیشتری به کلمه متناظر آن که milk است، می‌شود. در ادامه به بررسی فرمول‌های مکانیزم توجه باهدانا می‌پردازیم.

$$e_{t,i} = a(s_{t-1}, h_i) \quad (13-2)$$

$$a(s_{t-1}, h_i) = v^T \tanh(W[h_i; s_{t-1}]) \quad (14-2)$$

$$a(s_{t-1}, h_i) = v^T \tanh(W[W_1 h_i + W_2 s_{t-1}]) \quad (15-2)$$

$$\alpha_{t,i} = \text{softmax}(e_{t,i}) \quad (16-2)$$

با توجه به معادله 16-2، برای محاسبه مقادیر α لازم است ازتابع تناظر a استفاده کنیم که برای محاسبه آن نیز دو روش الحق (معادله 14-2) و جمع (معادله 15-2) وجود دارد. در نهایت با استفاده از یک لایه softmax به محاسبه وزن‌ها می‌پردازیم.

2-5-2 مکانیزم توجه لوآنگ (Luong)

در سال 2015، مقاله‌ای دیگر [?] مدل مکانیزم دیگری ارائه داد. از آنجایی که هدف مکانیزم توجه، محاسبه شباهت‌های بین خروجی‌های کدگذار و حالت‌های پنهان مراحل قبل کدگشا است، در این مقاله نیز از ضرب کسینوسی استفاده شده است. همانطوری که از معادلات زیر مشخص است، توجه لوآنگ، برخلاف توجه باهدانا، از حالت پنهان s_t استفاده می‌کند.

$$a(s_t, h_i) = v_a^T \tanh(W[h_i; s_t]) \quad (17-2)$$

$$a(s_t, h_i) = s_t^T h_i \quad (18-2)$$

$$a(s_t, h_i) = s_t^T W_a h_i \quad (19-2)$$

3-5-2 مکانیزم توجه به خود

این مفهوم برای اولین بار در سال 2016 مطرح شد. [?] هدف این توجه این است که ارتباط اجزای موجود در یک دنباله را با یکدیگر بسنجد تا بتواند برداشت درست‌تری از کل دنباله داشته باشد. تنها تفاوت این مکانیزم با توجه عادی این

است میزان توجه اجزا دنباله را با خود همان دنباله می‌سنجیم. شکل ?? ارتباط کلمه قرمز رنگ با کلمه‌های قبلی در جمله بررسی شده و شدت آبی بودن هر کلمه، میزان ارتباط را نشان می‌دهد.

The FBI is chasing a criminal on the run .
 The **FBI** is chasing a criminal on the run .
 The **FBI** **is** chasing a criminal on the run .
 The **FBI** **is** **chasing** a criminal on the run .
 The **FBI** **is** **chasing** a **criminal** on the run .
 The **FBI** **is** **chasing** a **criminal** **on** the run .
 The **FBI** **is** **chasing** a **criminal** **on** **the** run .
 The **FBI** **is** **chasing** a **criminal** **on** **the** **run** .
 The **FBI** **is** **chasing** a **criminal** **on** **the** **run** .

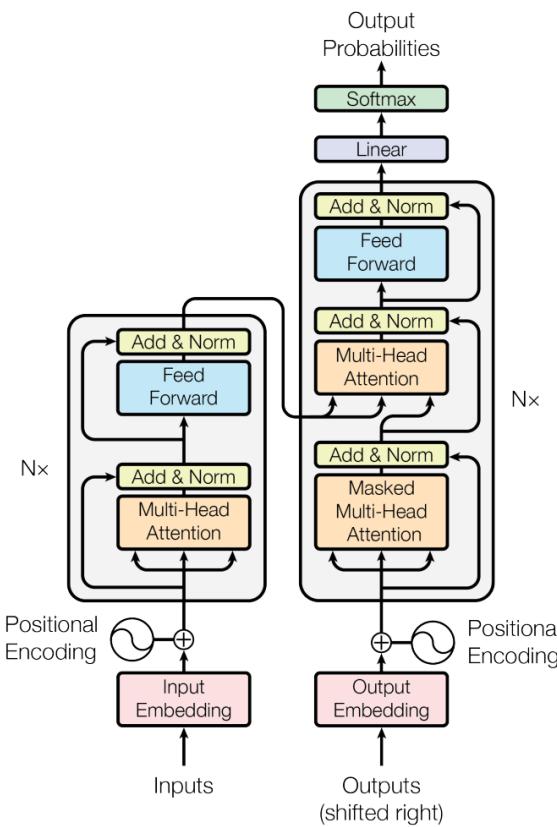
شکل 6-2: نمونه‌ای از توجه به خود در یک جمله

6-2 مدل‌های transformer و مکانیزم توجه آن‌ها

[مدل‌های تبدیل‌کننده روش جدیدی را برای استفاده از مفهوم توجه ارائه دادند. در مقاله‌ی «توجه و دیگر هیچ! »]
 که در سال 2017 ارائه شد، این مدل‌ها مطلقاً بر پایه مکانیزم توجه به خود تکیه کرده‌اند. اکثر این مدل‌ها به صورت مدل‌های دنباله‌به‌دنباله پیاده‌سازی شده‌اند به صورتی که دو بخش کدگذار و کدگشا دارند. مطابق با شکل 7-2 مشاهده می‌کنیم که بخشی به نام Positional Embedding شرایط پردازش کلمات با توجه به جایگاه آن‌ها در جمله را فراهم می‌کند. برای این کار از معادلات 20-2 و 21-2 بردارهای جایگاه محاسبه می‌شوند و به بردارهای embedding اضافه می‌شوند.

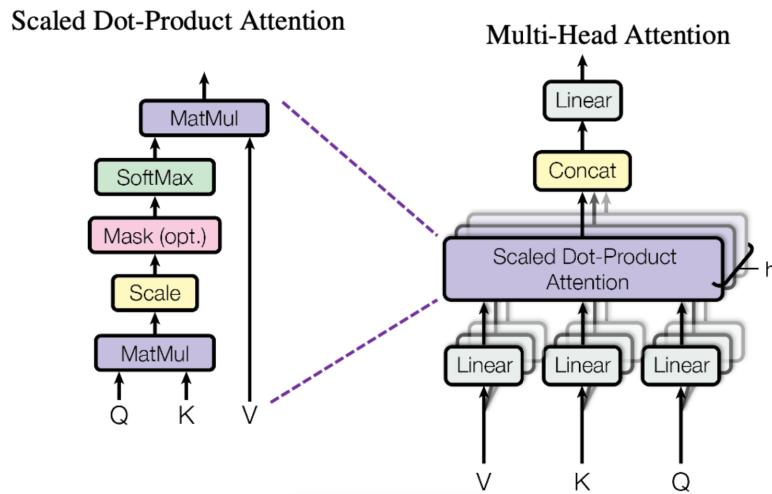
$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (20-2)$$

$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (21-2)$$



شکل 2-7: مدل تبدیل‌شونده و معماری آن

همان‌طوری که در شگل 7-2 مشخص است، شبکه‌های کدگذار و کدگشا از توجه چندسر استفاده می‌کنند. این نوع از مکانیزم توجه را می‌توان تعمیم‌یافته‌ی نسخه قبلی آن دانست. در توجه چندسر ما با سه موجودیت مختلف به نام کلید (Key)، مقدار (Value) و پرسش (Query) سروکار داریم. در مبحث توجه گفتیم که هدف این کار، تعیین میزان ارتباط بین هر جزء جمله خروجی با تمامی اعضای جمله ورودی است. اگر به شکل 7-2 توجه کنید، می‌بینید که هر سه ورودی این بلوک توجه از یک منبع که همان جمله‌ی ورودی است می‌آیند، پس از نوع توجه به خود است و هدف آن تعیین ارتباط و درک بهتر اجزای جمله‌ی زبان مبدا به یکدیگر است. پیش از بررسی ادامه‌ی روند مدل، لازم است تا الگوریتمی که در این نوع از مکانیزم توجه دنبال می‌شود را بررسی کنیم.



شکل 2-8: توجه چندسر و توجه مقیاس شده بر پایه ضرب داخلی

در مرحله‌ی اول، بردار کلمات با گذر از سه لایه‌ی پیش خور (feed-forward layer) با وزن‌های متفاوت، وکتورهای K ، V و Q را می‌سازند. سپس این وکتورها، وارد بلوك توجه مقیاس شده‌ی بر پایه‌ی ضرب داخلی (Scaled Dot-Product Attention) می‌شوند. در این بخش، ابتدا ضرب داخلی وکتورهای Q و K محاسبه می‌شود تا مشخص شود این دو چقدر به هم شبیه‌ند. سپس، با تقسیم امتیاز حاصل بر جذر طول رشته‌ی ورودی، امتیاز نرمال می‌شود تا از بروز مشکل انفجار گرادیان جلوگیری شود. سپس، با اعمال تابع softmax بر روی این مقادیر، وزن هر کدام از کلیدها برای هر پرسش (query) تعیین می‌شود. در نهایت، وزن‌های محاسبه شده در مقادرهای (value) کلمات متفاوت ضرب می‌شوند تا خروجی مورد نظر تولید شود. این مراحل را می‌توانید در تصویر 2-8 مشاهده کنید. با توجه به این که کدگشا رشته‌ی خروجی را کلمه به کلمه تولید می‌کند، باید در نظر داشته باشیم که وزن‌های توجه به گونه‌ای نباشند که کلمات به واژه‌های بعد از خود توجه داشته باشند.

برای این کار، در روند توجه مقیاس شده‌ی بر پایه‌ی ضرب داخلی (Scaled Dot-Product Attention)، بعد از نرمال شدن امتیازها و قبل از اعمال تابع softmax، آن را با یک ماتریس اکیدا بالا مثلثی که مقادیر بالای قطر اصلی آن منفی بی‌نهایت هستند جمع می‌کنیم. این کار باعث می‌شود که بعد از اعمال تابع softmax، مقدار وزن توجه هر کلمه به کلمه‌های بعد از خودش صفر شود. نمونه‌ای از این عمل سرپوش گذاری را می‌توانید در معادله 22-2 مشاهده کنید.

$$(22-2)$$

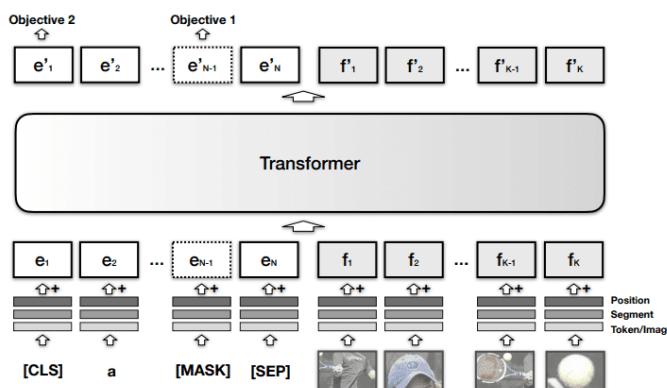
$$\begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.6 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.6 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix} + \begin{bmatrix} 0 & -inf & -inf & -inf \\ 0 & 0 & -inf & -inf \\ 0 & 0 & 0 & -inf \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.7 & -inf & -inf & -inf \\ 0.1 & 0.6 & -inf & -inf \\ 0.1 & 0.2 & 0.6 & -inf \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix}$$

7-2 مدل‌های زبان و تصویر بر پایه BERT

مدل‌های تصویر و زبان خیرا شهرت بسیاری پیدا کرده‌اند. بر اساس وظیفه اعمال شده، معماری‌های متفاوتی برای حل مسائل تصویر و زبان در کنار یکدیگر ارائه شده است. راجایی که تبدیل کنندگان در بسیاری از موارد به خوبی عمل کردند، استفاده از آن‌ها برای حل این مسائل اجتناب‌ناپذیر است. بسیاری از این مدل‌های رائه‌شده بر پایه BERT یا معماری مشابه با آن دارند. به طور کلی می‌توان این مدل‌ها را در دو دسته جای داد.

1-7-2 مدل‌های تک‌جریان

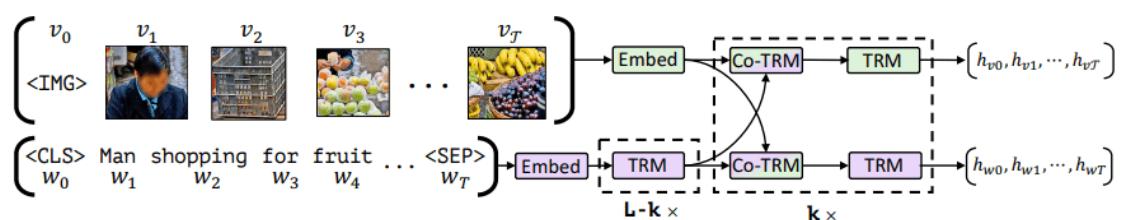
معماری‌هایی هستند که هر دو بخش تصویر و زبان را در یک مازول کدگذاری می‌شوند. همانطوری که در تصویر 9-2 مشخص است، بردارهای تصویر و زبان هردو از یک مازول تبدیل کننده عبور می‌کنند. مدل‌های تک‌جریان از لحاظ تعداد پارامترها بهینه هستند و از ابتدا به صورت تلفیقی بین بردارهای تعییه تصاویر و زبانی تناظر ایجاد می‌کنند.



شکل 2-9: نمونه‌ای از یک مدل تک‌جریان

2-7-2 مدل‌های دوچریان

معماری‌هایی هستند که بخش تصویر و زبان در مازول‌های جداگانه پردازش می‌شوند و سپس توسط مازول دیگری ارتباط بین بردارهای تعبیه شده زبانی با بردارهای تصویر را محاسبه می‌کند. با توجه به تصویر 2-10 هر دو بخش تصویر و زبان جداگانه جریان پیدا می‌کنند و سپس یک تناظر بین این دو جریان بدست می‌آوریم.



شکل 2-10: نمونه‌ای از یک مدل دوچریان

8-2 جمع‌بندی

در این فصل به شرح مفاهیم پایه‌ای پرداخته شد که در مراحل مختلف انجام این پژوهش مورد استفاده قرار گرفته و برای درک کامل خواننده‌ی این نوشتار مورد نیاز است. در این فصل معماری‌ها و مدل‌های استفاده شده در این پژوهش به صورت جزیی مورد بررسی قرار گرفت. سپس به معرفی انواع مدل‌های موجود برای حل مسائل تصویر و زبانی پرداخته شد.

فصل 3

کارهای مرتبط

1-3 مقدمه

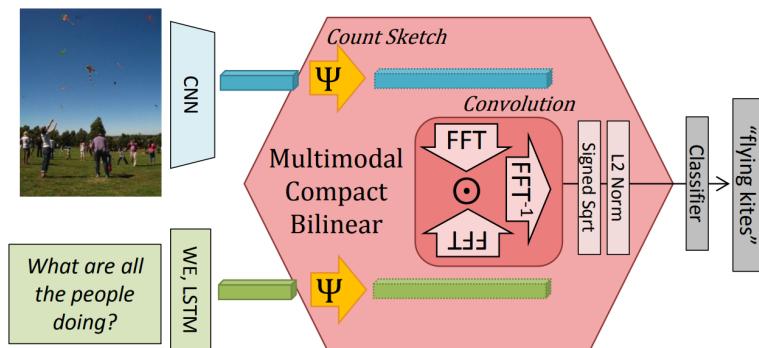
در این فصل به معرفی کارهای مشابه و شرح رویکردهای مورد استفاده برای حل مساله‌ی پرسش و پاسخ تصویری می‌پردازیم. در ابتدا به بررسی روش‌های متفاوتی که برای حل مسئله پرسش و پاسخ تصویری ارائه شد و سپس به کارهای مشابه با تولید پاسخ در پرسش و پاسخ تصویری و مجموعه داده‌های موجود در این زمینه پرداخته شده است. در نهایت به مدل‌های از پیش آموزش داده شده در این زمینه می‌پردازیم.

2-3 روش‌های متفاوت حل مسئله پرسش و پاسخ تصویری

تا قبل از سال 2019 روش‌ها و رویکردهای متفاوتی ارائه شده که بتوان ویژگی‌های تصویری و زبانی را در کنار یکدیگر قرار دهیم. به طور کلی این روش‌ها را می‌توان در 5 دسته طبقه‌بندی کرد.

1-2-3 روش‌های بر پایه Bilinear Pooling

روشی برای ایجاد توجه بین بردارهای زبانی و بردارهای تصویری است. روش‌های r بازنمایی خوبی را فراهم می‌کردند و در بسیاری از کارهای تصویری استفاده می‌شد. در سال 2016 روشی با نام MCB [?] برای ادغام تصویر و زبان ارائه شد. رویکرد آن‌ها در شکل 1-3 نشان داده شده است.

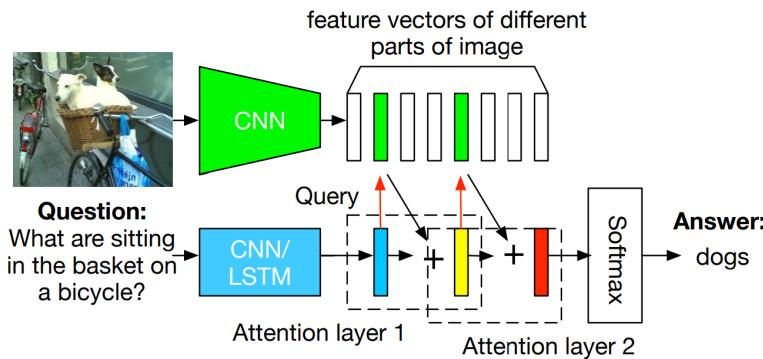


شکل 3-1: روش MCB برای ادغام تصویر و زبان

در سال 2017 نیز روشی از دیگر با نام [?] MLB Bilinear Pooling برپایه مسائل زبان و تصویر منتشر کرد که در زمان خود از مدل‌های پایه‌ای عملکرد بهتری نشان داد.

2-2-3 روش‌های بر پایه توجه

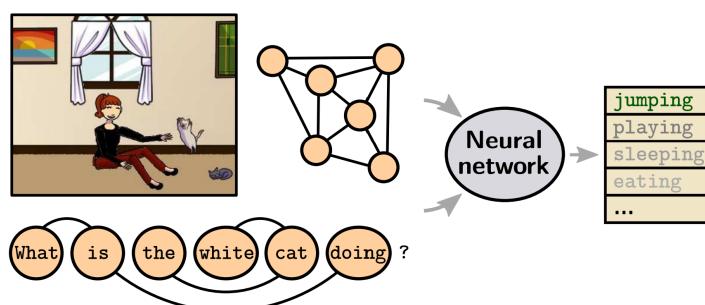
تا قبل از اینکه معماری‌های تبدیل کننده معرفی شوند، محققان از مکانیزم توجه بر شبکه‌های عصبی بازگشتی و شبکه‌های کانولوشن استفاده می‌کردند، برای مثال شبکه توجه پشتهدای [?]، توجه سلسه‌مراتبی [?] و توجه متقاضن [?] پس از ارائه مکانیزم‌های توجه منتشر شدند. همان‌طوری که از تصویر 2-3 مشخص است، توجه سلسه‌مراتبی برای پردازش تصویر از یک شبکه کانولوشن و برای پردازش سوال یا همان بخش زبانی، از یک LSTM استفاده شده است، سپس با استفاده از مکانیزم پیشنهادی به تناظر میان این دو بخش پرداخته شده است. در هر دو پژوهش مجموعه داده‌های مشهور زبان و تصویر ارزیابی شده و نسبت به مدل‌های SOTA به خوبی عمل کرده بود. با این حال، این مدل‌ها دارای اشتباهات بسیاری بودند و نیازمند بهبود بودند تا اینکه مدل‌های تبدیل شونده منتشر شدند و توانستند به دقت به نسبت بالاتری از این معماری‌ها برسند.



شکل 3-2: توجه پشت‌های برای حل مسئله پرسش‌پاسخ تصویری

3-2-3 ادغام روابط اشیاء

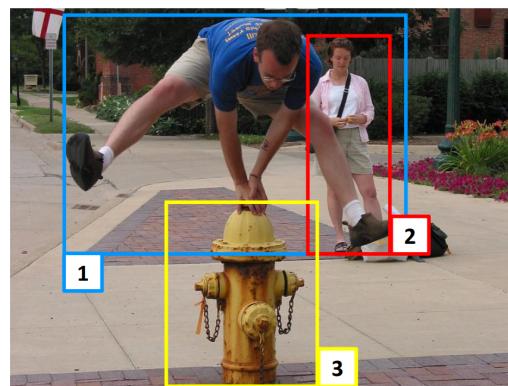
یکی از ایده‌های خلاقانه حل مسئله پرسش‌پاسخ تصویری استفاده از شبکه‌های رابطه‌ای است که در ابتدا روابط بین اشیاء در تصویر و روابط بین کلمات در سوال مشخص می‌شوند و سپس به شبکه‌ای برای پیش‌بینی پاسخ مورد نظر ورودی داده می‌شوند. از نمونه کارآمد این روش می‌توان به مقاله حل پرسش‌پاسخ تصویری از طریق ساختار گرافی [?] اشاره کرد که در ابتدا اشیاء موجود در تصویر و روابط بین آنها، و همچنین کلمات موجود در جمله سوال و روابط بین آنها را به صورت گراف توصیف کرده و سپس آنها به شبکه عصبی ورودی می‌دهیم تا بتوانند از طریق آنها به پاسخ مورد نظر برسد. برای بدست آوردن رابطه بین اشیاء در تصویر می‌توان از شبکه‌های رابطه‌ای و برای بدست آوردن گراف جمله می‌توان از گراف وابستگی گرامری استفاده کرد.



شکل 3-3: حل پرسش‌پاسخ تصویری با ساختارهای گرافی

3-3 تولید پاسخ برای پرسش و پاسخ تصویری

همان طوری که در بخش 2-3 توضیح داده شد، تلاش های بسیاری برای حل پرسش و پاسخ تصویری از طریق دسته بندی شده است، در حالی که حل مسئله از طریق تولید متن توجه چندانی از محققان را جلب نکرده است. راههایی برای حل پرسش و پاسخ تصویری با نامهای [mQA] و [AMA] ارائه شد که از شبکه های LSTM برای تولید پاسخ استفاده شده بود. پس از آن نیز سیستم هایی نظیر SimVLM [?] منتشر شد که هدف آن کاهش هزینه و پیچیدگی آموزش بود. در این سیستم ها حل کردن مسائل مرتبط با زبان و تصویر به هر دو روش تولید پاسخ و دسته بندی بررسی شده است. در نهایت، VL-Bart و VL-T5 [?] یک چهار چوب یکپارچه ای را در بخش های مختلف نشان می دهد. این مدل ها نشان دادند که یکدیگر را داشت. شکل 4-3 کارایی این چهار چوب را در بخش های مختلف نشان می دهد. با این حال، از آنجایی که حل کردن مسائل چند ماژول از طریق تولید متن عملکرد و کارایی بهینه تری از دسته بندی دارد. با این حال، از آنجایی که این مدل ها روی مجموعه داده پرسش و پاسخ تصویری [?] آموزش داده شده اند و پاسخ ها در این مجموعه داده به صورت دسته بندی داده شده است، همچنان پاسخ دادن به صورت جملات را در بر ندارند.



	Text Input	Text Output
Multimodal LM	"span prediction: A <text_1> is <text_2> over fire hydrant"	"<text_1> man <text_2> jumping"
Visual QA	"vqa: what is the man jumping over?"	"fire hydrant"
Visual Grounding	"visual grounding: yellow fire hydrant"	"<vis_3>"
Image-Text Matching	"image text match: A cat is lying on a bed"	"false"

شکل 3-4: چهار چوب یکپارچه ارائه شده برای حل مسائل چند ماژول به روش تولید متن

4-3 مجموعه داده‌های منتشر شده در پرسش و پاسخ تصویری

مجموعه داده‌های زیادی در حوزه پرسش و پاسخ تصویری ارائه شده است. با این حال، بسیاری از مجموعه داده‌ها دارای پاسخ‌های تک‌کلمه‌ای هستند و تعداد کمی از آن‌ها پاسخ‌ها را به صورت جمله نوشته‌اند. در سال 2019 مقاله VCR [?] ارائه شد که یک سوال چالشی از تصویری پرسیده شود و در جواب آن باید پاسخی همراه با دلیل انتخاب پاسخ خروجی داده شود. این مجموعه داده شامل 290 هزار پرسش و پاسخ چند‌گزینه‌ای است که از 110 هزار صحنه فیلم‌ها گرفته شده است. این مجموعه داده نشان داد که بر خلاف ساده بودن VCR برای انسان‌ها (دقیقی بالاتر از 90) ماشین‌ها توانایی چندانی در حل آن ندارند. نویسنده‌گان، مدلی نیز ارائه دادند که این مسئله را حل کند، ولی طبق ادعای آن‌ها این مسئله همچنان فاصله زیادی تا حل دارد.



شکل 3-5: نمونه‌ای مجموعه داده VCR که علاوه بر ارائه پاسخ، نیازمند ارائه دلیلی برای پاسخ است

مجموعه داده دیگری با نام VQA-E [?] منتشر شد که برگرفته در مجموعه داده رسمی پرسش و پاسخ تصویری بود. برای حل این مجموعه داده، علاوه بر ارائه پاسخ (تک‌کلمه‌ای) باید توضیحی برای پاسخ انتخاب شده ارائه شود. نحوه تولید این مجموعه داده به این صورت است که به ازای هر سه‌تایی پرسش، تصویر و پاسخ یک توضیحی با توجه به توضیحات تصویر تولید می‌شود. در این مجموعه داده، تلاش بر این بوده که توزیعی مشابه با مجموعه داده رسمی پرسش و پاسخ تصویری

داشته باشد. مجموعه داده‌های دیگری نیز به صورت جمعیت‌محور ارائه شده‌اند. این مدل مجموعه داده‌ها می‌توان به [?] A-OKVQA اشاره کرد که شامل 25 هزار پرسش است.

5-3 مدل‌های از پیش آموزش داده شده

پیش‌آموزش مدل‌های زبان-تصویر اخیراً مورد توجه بسیاری واقع شده است. مدل‌های موجود بیشتر بر پایه توجه به خود و معما ری تبدیل شونده که در بخش‌های 6-2 و 3-5 به صورت جزئی مورد بررسی قرار گرفته‌اند. این مدل‌ها در هر دو بخش تصویر و زبان به خوبی عمل کرده‌اند. همانطوری که در بخش 7-2 توضیح داده شد، این مدل‌ها را می‌توان به دو بخش تک‌جريان و دو جريان دسته‌بندی کرد. با توجه به اينکه اين مدل‌ها در تولید متن استفاده چندانی نداشته‌اند، در اين پژوهش تلاش بر اين بوده که مدل‌های از پیش آموزش داده شده را برای اين هدف استفاده کنیم.

فصل 4

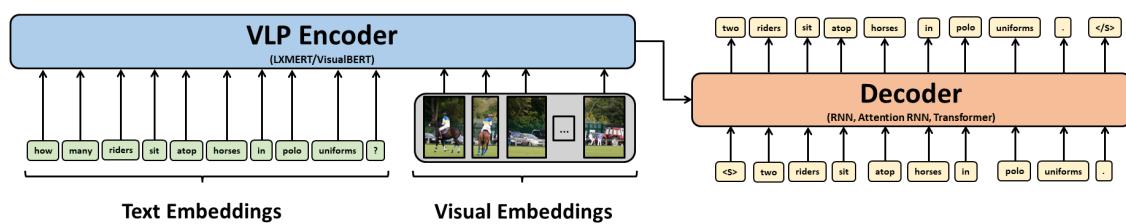
روش پیشنهادی

1-4 مقدمه

هدف از این پژوهش انتشار روشی برای تولید پاسخ‌ها به صورت جمله است. برای این مورد، از شبکه‌های از پیش‌آموزش داده شده استفاده شده است. از LXMERT [?] [?] به عنوان یک معماری دوچریان و از VisualBERT [?] به عنوان یک معماری تک‌چریان بهره‌برداری شده است.

2-4 معماری سیستم

برای حل این مسئله از معماری کدگذار-کدگشا استفاده شده است به گونه‌ای که از یک شبکه از پیش‌آموزش داده شده به عنوان کدگذار و از معماری‌های متفاوتی به عنوان کدگشا استفاده شده است.

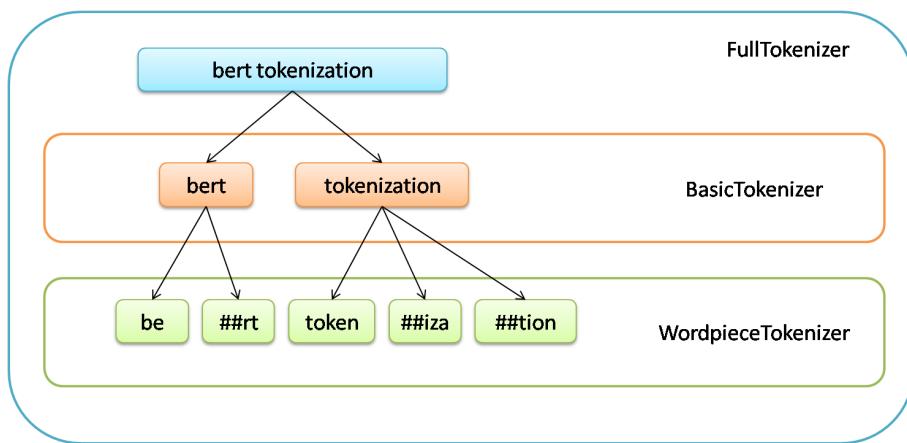


شکل 4-1: شماتیکی سیستم پرسش‌پاسخ تصویری

3-4 ورودی اولیه و خروجی نهایی

در ورودی سیستم باید بتوانیم تصویر و متن را به سیستم ورودی بدهیم. برای انجام این عمل باید هر دو بخش متن و تصاویر را به بردارهای ویژگی تبدیل کنیم.

1. **بردارهای متن:** برای پردازش دقیق پرسش‌ها، لازم است که پرسش‌ها را به بخش‌های کوچک‌تری تقسیم کنیم که شبکه‌های عصبی عمیق قادر به اعمال محاسبات لازم باشند. به قسمت‌های کوچک‌تر اصطلاحاً توکن گفته می‌شود. به عمل جداسازی قسمت‌های یک جمله Tokenization گفته می‌شود. عکس این عمل که همان اتصال توکن‌ها و تشکیل جمله است را De-Tokenization نامیده می‌شود. در شکل 4-2 کارکرد این جداسازی مشاهده می‌شود. با توجه به اینکه هر دو شبکه VisualBERT و LXMERT بر پایه شبکه BERT هستند، برای بدست آوردن بردارهای ویژگی متن ورودی از جداساز شبکه BERT استفاده شده است.



شکل 4-2: نمونه‌ای از کارکرد Tokenizer

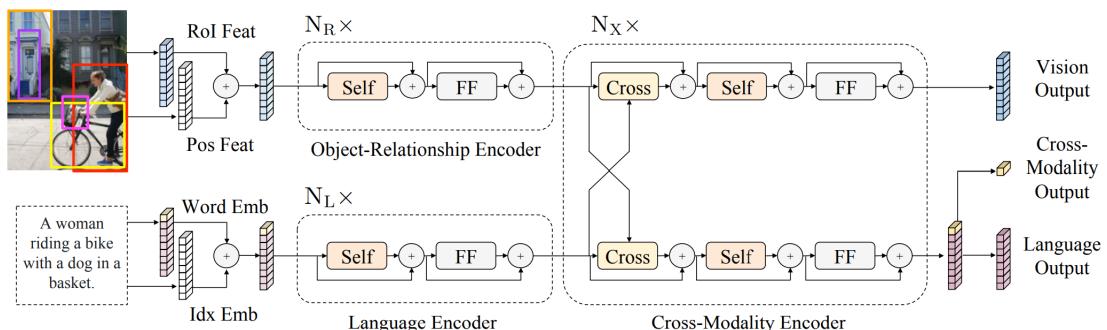
2. **بردارهای تصویر:** برای تبدیل تصاویر به بردارهای ویژگی مطابق با مقاله‌ای برای حل مسائل زبان-تصویر [?] می‌توانیم هر تصویر را مجموعاً به صورت 36 شیء در نظر بگیریم که از شبکه عصبی Faster-RCNN [?] تشخیص داده شده و هر شیء برداری با اندازه 2048 دارد.

Encoder مدل 4-4

از آنجایی که در بخش‌های قبل گفته شد، از مدل‌های از پیش آموزش داده شده استفاده می‌کنیم. به منظور مقایسه هر دو حالت تک‌جریان و دوچرخیان از دو مدل تبدیل شونده با معماری‌های متفاوت استفاده شده است تا بتوان مقایسه جامع و کاملی حول انواع مدل‌ها و عملکرد آن‌ها داشت. هدف از به کارگیری بخش کدگذار محاسبه یک بازنمایی از پرسش و تصویر همراه با یکدیگر و عبور دادن آن به بخش کدگشا است.

[?] LXMERT مدل 1-4-4

این مدل به عنوان یک مدل پایه برای حل مسئله پرسش و پاسخ تصویری ارائه شد. این مدل شامل سه بخش کدگذار تصویری، زبانی و میان‌ماژولی است. این مدل دو ورودی می‌گیرد، یک تصویر و یک متن که همان پرسش مربوط به تصویر است. با ترتیب لایه‌های توجه به خود و توجه میانی باعث می‌شود که مدل بتواند یک بازنمایی از تصویر، یک بازنمایی از متن و یک بازنمایی میان‌ماژولی از ورودی بدست بیاورد. محققان در این مدل از روش‌های متفاوتی نظری MOP¹، MLM² و تناظر میان‌ماژولی³ برای آموزش استفاده کرده‌اند که باعث شده است روابط درون‌ماژولی و میان‌ماژولی به خوبی تشخیص داده شود. همانطوری که از تصویر 3-4 مشخص است این مدل یک مدل دوچرخیان است.

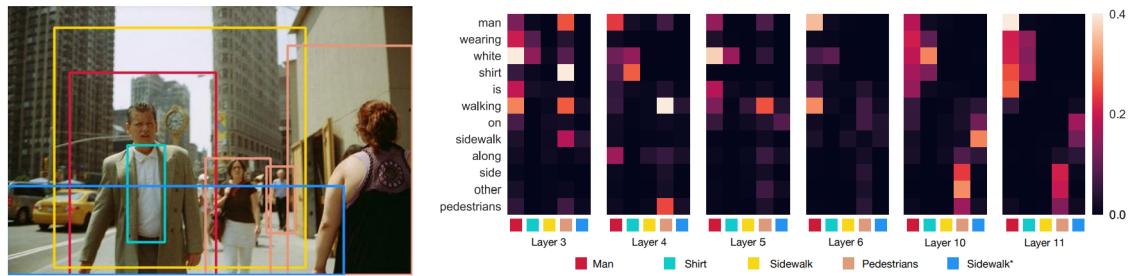


شکل 4-3: معماری مدل LXMERT همراه با ورودی و خروجی

Masked Language Modeling¹
Masked Object Prediction²
Cross-modality Matching³

2-4-4 مدل VisualBERT [?]

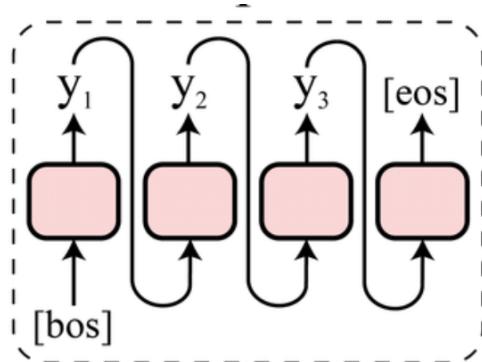
پس از انتشار LXMERT و پیشرفت چشمگیرش در حل مسئله پرسش و پاسخ تصویری پژوهش‌های بسیاری برای بهبود بیشتر انجام شد. از جمله این پژوهش‌ها بررسی مدل BERT و عملکرد آن در مسائل زبان-تصویر را میتوان نام برد که منجر به انتشار مدل VisualBERT [?] شد. این مدل به صورت تک‌جریان است و دنباله‌ای از بردارهای ویژگی متن همراه با دنباله‌ای از بردارهای ویژگی تصویر را ورودی می‌گیرد و خروجی را به صورت بردارهای بازنمایی می‌دهد. به علاوه، میزان توجه بردارهای ویژگی توکن‌ها و بردارهای ویژگی اشیاء موجود در تصویر را با یکدیگر اندازه‌گیری کردن و اشیاء با توکن‌های مربوطه به خوبی توجه خورده‌اند. برای مثال با توجه به شکل 4-4 در لایه 11 کلمه man به خوبی با محدوده مربوط به مرد موجود در تصویر مرتبط شده است!



شکل 4-4: میزان توجه بخش متن به تصاویر در VisualBERT

5-4 مدل Decoder

پس از آنکه بردارهای بازنمایی تصویر و پرسش توسط کدگذار محاسبه شد، لازم است که آن‌ها را برای تولید پاسخ به بخش کدگشا بدھیم. بخش کدگشای معماری‌های ارائه شده از دو معماری شبکه‌های عصبی بازگشتی و شبکه‌های تبدیل‌شونده استفاده شده است. برای تولید پاسخ از مکانیزم Autoregressive استفاده شده است. در این مکانیزم توکن‌ها بر اساس توکن‌های قبلی پیش‌بینی می‌شوند و سپس همان توکن جدید همرا با دنباله قبلی برای توکن جدید دیگری مطابق با شکل 5-4 به کدگشا ورودی داده می‌شوند.



شکل 4-5: ساختار مدل Autoregressive Decoder

6-4 جمع‌بندی

در این بخش از نوشتار به بررسی دقیق اجزای تشکیل‌دهندهٔ سیستم پیشنهادی برای حل مسالهٔ پرسش‌پاسخ تصویری پرداختیم. روشی نوبرای حل این مسئله که باعث رفع ابهامات بسیاری می‌شود. اشاره شد که از مدل‌های از پیش آموزش داده شده استفاده کردیم و همچنین برای مقایسه نتایج و بدست آوردن بهترین معماری، چندین حالت بررسی شد که در بخش 5 به مقایسه نتایج پرداخته خواهد شد.

فصل 5

ارزیابی روش پیشنهادی

1-5 مقدمه

در این فصل برای بررسی عملکرد روش پیشنهادی در فصل قبل، سیستم مورد نظر به طور کامل پیاده‌سازی شده و بر روی یک مجموعه‌داده‌ی شناخته‌شده اجرا شده است. برای ارزیابی روش پیشنهادی روش‌های متفاوتی ارائه شده است. برای انتخاب رویکرد مناسب ارزیابی از میان خیل عظیمی از روش‌های شناخته‌شده عوامل متعددی مورد بررسی قرار می‌گیرد و با توجه به آن‌ها یک یا چند روش انتخاب می‌شود. مهم‌ترین عامل ساختار و جنس داده‌های خروجی مدل و همچنین جنس داده‌های ارزیابی است. با توجه به مدل و مجموعه داده مورد استفاده، در این پژوهش نیاز به معیارهای ارزیابی برای مقایسه جملات داریم. از آنجایی که این روش، روشی نو است برای اثبات درستی آن از معیارهای ارزیابی متفاوتی استفاده کردیم که به طور کلی در دو بخش دسته‌بندی می‌شود.

2-5 معیارهای برپایه شباهت نحوی

معیارهای برپایه شباهت نحوی به یکسان بودن توکن‌ها در یک جمله تمرکز دارد. به عبارتی براساس تعداد توکن‌های انتخابی و برابری آن‌ها امتیازی بین دو جمله در نظر گرفته شده که نشانگر میزان شباهت دو جمله به یکدیگر است. در این پژوهش از معیارهای BLEU [?], METEOR [?], Rouge [?] به عنوان معیارهای شباهت برپایه توکن‌ها استفاده شده است. این معیارها معمولاً برای ترجمه‌های ماشینی استفاده می‌شوند و از آنجایی که هدف مقایسه دو جمله پاسخ و

جمله مرجع است استفاده از این معیارها برای سنجش سیستم راه حل مناسبی است.

3-5 معیارهای برپایه بردارهای تعبیه

در این معیارها با بهره‌برداری از بردارهای تعبیه¹ به محاسبه شباهت بین جملات پرداخته می‌شود. در این پژوهش از دو معیار $\text{BERTScore} [?]$ و $\text{Average Score} [?]$ به عنوان معیارهای بردار تعبیه استفاده شده است.

1-3-5 معیار $\text{BERTScore} [?]$

BERTScore یک معیار ارزیابی خودکار است که برای ارزیابی سیستم‌های تولید متن استفاده می‌شود. برخلاف روش‌های رایج موجود که شباهت نحوی توکن‌ها را محاسبه می‌کنند، BERTScore بر محاسبه شباهت معنایی بین توکن‌های مرجع و توکن‌های پیش‌بینی شده تمرکز دارد. نویسنده مقاله آن را بر روی ترجمه‌های ماشینی و وظایف توضیح تصاویر آزمایش کرد و دریافت که با قضاوت‌های انسانی ارتباط بهتری دارد.

از مزایای این معیار نسبت به مقایسه نحوی می‌توان به عدم محاسبه توکن‌های هم‌معنی در نظر گرفت. برای مثال معیارهای برمبانای شباهت نحوی، در صورت مغایرت توکن‌ها هیچ امتیازی در نظر نمی‌گیرند در حالی که BERTScore به نزدیکی معانی کلمات توجه می‌کند.

2-3-5 معیار Average Score

برای محاسبه این معیار، در ابتدا یک میانگین از بردارهای تعبیه جملات گرفته و سپس با استفاده از تابع شباهت کسینوسی میزان شباهت بین جملات مرجع و جملات پیش‌بینی شده محاسبه می‌شود. (معادله 2-5)

$$\text{Score}(s = [t_1, t_2, \dots, t_T]) = \frac{1}{T} \sum_{i=0}^T \text{Embedding}(t_i) \quad (1-5)$$

$$\text{sim}(s_1, s_2) = \frac{\vec{\text{Score}}(s_1) \cdot \vec{\text{Score}}(s_2)}{\|\vec{\text{Score}}(s_1)\| \times \|\vec{\text{Score}}(s_2)\|} \quad (2-5)$$

¹ Embedding vectors

4-5 مجموعه‌داده‌ی مورد استفاده

در بخش 4-3 به معرفی بخش اعظمی از مجموعه داده‌های موجود در پرسش و پاسخ تصویری پرداخته شد. با این حال اگر به این مجموعه داده‌ها با دید عمیق تری بنگریم متوجه این موضوع می‌شویم که در هیچ‌کدام از این مجموعه‌ها پاسخ‌ها به صورت جمله کامل نیستند، به عبارتی برای حل آن‌ها استفاده از دسته‌بندی کافی است. برای ارزیابی دقیق سیستم پیشنهادی نیازمندی مجموعه داده‌ای است که پاسخ‌های پرسش‌ها به صورت جمله باشد. مجموعه داده Full Sentence از [؟] این خاصیت را دارد و بهترین گزینه برای استفاده به عنوان مجموعه داده ارزابی است.

مجموعه داده FSVQA شامل یک تصویر و یک پرسش مربوط به آن است و جمله یه صورت پاسخ آن پرسش به عنوان هدف در نظر گرفته شده است. این پاسخ‌ها از دو مجموعه داده VQA [؟] و MSCOCO [؟] و بر اساس قوانین زبان انگلیسی به صورت خودکار تولید شده‌اند. توجه به این نکته حائز اهمیت است که تولید جملات از این روش امکان ایجاد اشکالاتی را در مجموعه داده بوجود می‌آورد و مدل‌ها نیز قادر به یادگیری الگوی پاسخ‌ها یا همان قوانین هستند.

5-5 نتایج ارزیابی سیستم پیشنهادی

براساس توضیحات داده‌شده سیستم پیشنهادی را با معماری‌های متفاوت پیاده‌سازی و برروی مجموعه داده معرفی شده در بخش 4-5 آموزش دادیم. سپس هر کدام از معماری‌ها برروی داده ارزیابی سنجیده شد. علاوه بر سنجش مدل‌ها و معماری‌های متفاوت میزان تأثیر تغییرات در هر معماری نیز مورد مطالعه قرار گرفته شده است.

1-5-5 ارزیابی و مقایسه معماری‌های ارجح

پس از بدست آوردن نتایج، به مقایسه معماری‌های مورد استفاده پرداخته شده است. با توجه به جدول 1-5 همان‌طوری که انتظار می‌رفت، شبکه‌های عصبی Transformers همراه با شبکه LXMERT به عنوان کدگذار بهترین عملکرد را داشت. اگر اندکی نتایج را وارسی کنیم متوجه بالا بودن امتیازها می‌شویم که نسبت به مدل‌های مشابه در توضیح تصاویر مقادیر بسیار بالایی دارند. مطابق با آنچه در بخش 4-5 بحث شد، یکی از دلایل را می‌توان ساده بودن مجموعه داده در نظر گرفت. از آنجایی که مجموعه داده از قوانین از پیش تعیین شده و بصورت خودکار تولید شده‌اند، مدل‌ها و به خصوص کدگشا ممکن است این الگوها را آموزش دیده باشند. با این حال در همه حالات از دقتی که نویسنده‌گان مجموعه داده

گزارش کرده بودند عملکرد بهتری داشته‌ایم.

Embedding-based		Word-based			Method	
BERT Score	Average Score	ROUGE-L	METEOR	BLEU	Decoder	Encoder
-	-	-	23.3	23.9	LSTM	LSTM Q+I
79.19	86.50	56.38	56.65	32.19	1-LSTM	
82.07	89.63	61.50	62.99	39.37	1-GRU	
91.84	95.94	85.49	86.43	79.03	1-LSTM+Bahdanau attention	
91.90	<u>96.11</u>	86.25	86.96	79.54	1-LSTM+Luong(general) attention	LXMERT
89.32	95.42	82.62	83.26	71.40	1-GRU+Bahdanau attention	
90.37	95.73	83.84	84.71	73.92	1-GRU+Luong(dot) attention	
<u>95.01</u>	90.20	<u>90.60</u>	<u>91.18</u>	<u>86.73</u>	3-Transformer Decoder	
69.20	84.83	38.35	38.00	18.62	1-LSTM	
73.59	87.74	43.72	44.24	22.51	1-GRU	
93.50	97.11	87.28	88.07	84.27	1-LSTM+Bahdanau attention	
93.11	<u>97.17</u>	86.90	87.71	82.90	1-LSTM+Luong(concat) attention	VisualBERT
89.26	96.10	82.40	82.87	72.20	1-GRU+Bahdanau attention	
91.81	96.93	85.23	86.17	79.65	1-GRU+Luong(dot) attention	
<u>94.44</u>	91.94	<u>89.09</u>	<u>89.76</u>	<u>85.95</u>	3-Transformer Decoder	

جدول 5-1: نتایج معماری‌های متفاوت بر مجموعه داده FSVQA. اعداد موجود در نام کدگشا نشانگر تعداد لایه‌های آن است.

2-5-5 مقایسه شبکه‌های عصبی بازگشتی

در این بخش به بررسی تغییرات کدگشاها بر پایه شبکه‌های عصبی بازگشتی پرداخته شده است. برای بررسی بیشتر تعداد لایه‌های شبکه و نوع شبکه‌ها و همچنین جهت جریان اطلاعات در شبکه‌ها مورد بررسی قرار گرفت که نتایج با تمام جزئیات در جدول 2-5 آورده شده‌اند. با توجه به این جدول می‌توانیم نتیجه بگیریم که برترین معماری مربوط به استفاده از 3 لایه شبکه‌های LSTM به صورت دو طرفه است. اگر به نتایج کدگذارها دقت کنیم متوجه می‌شویم که استفاده از VisualBERT به عنوان کدگذار در این حالت به خوبی عمل نمی‌کند. در حالت کلی این معماری برا حل مسئله به روش و با توجه وجود معماری‌های نوبن، استفاده از این معماری پیشنهاد نمی‌شود.

3-5-5 ارزیابی و مقایسه شبکه‌های عصبی بازگشتی و مکانیزم توجه سراسری

بررسی نتایج بخش 2-5-2 نشان داد که استفاده از شبکه‌های عصبی بازگشتی به تنها یک پاسخگوی نیازهای مجموعه داده مورد نظر نیست. برای رفع این مشکل از مکانیزم‌های توجه سراسری نظیر باهدانا و لوآنگ که در بخش 2-5-1 و بخش 2-5-2 توضیح داده شده‌اند استفاده کرده و از هر سه متد الحق، جمع و ضرب آن‌ها را مورد بررسی قرار داده شده است. بررسی نتایج مشخص می‌کند که استفاده از این مکانیزم کمک بسیاری به شبکه کرده و امتیازهای نحوی و معنایی هر دو به حد قابل توجهی بالا رفته‌اند. از نتایج این بخش می‌توان نتیجه گرفت استفاده از شبکه کدگذار VisualBERT عملکرد

Embedding-based		Word-based			Method	
BERT Score	Average Score	ROUGE-L	METEOR	BLEU	Decoder	Encoder
79.19	86.50	56.38	56.65	32.19	1-LSTM	LXMERT
83.10	89.57	62.97	64.39	41.28	2-LSTM	
82.67	89.54	63.02	64.25	41.10	3-LSTM	
82.07	89.63	61.50	62.99	39.37	1-GRU	
79.68	88.56	58.24	59.57	35.23	2-GRU	
83.51	90.11	63.65	65.25	41.97	1-BiLSTM	
<u>84.09</u>	90.47	64.67	66.28	43.26	2-BiLSTM	
84.02	<u>90.58</u>	<u>64.74</u>	<u>66.39</u>	<u>43.54</u>	3-BiLSTM	
82.46	89.88	62.66	64.15	40.89	1-BiGRU	
79.27	88.74	57.29	58.54	34.32	2-BiGRU	
74.72	86.66	54.55	54.82	27.88	3-BiGRU	
69.20	84.83	38.35	38.00	18.62	1-LSTM	VisualBERT
70.40	85.30	38.42	38.62	18.92	2-LSTM	
71.35	85.87	38.80	39.33	19.30	3-LSTM	
73.59	87.74	43.72	44.24	22.51	1-GRU	
73.26	88.21	44.93	45.16	21.96	2-GRU	
67.15	84.23	33.99	32.34	11.36	3-GRU	
71.19	85.52	39.75	39.70	20.03	1-BiLSTM	
72.13	86.67	40.27	40.05	18.84	2-BiLSTM	
73.43	87.21	41.19	41.51	20.46	3-BiLSTM	
<u>74.32</u>	<u>88.42</u>	<u>45.26</u>	<u>45.66</u>	<u>22.72</u>	1-BiGRU	
73.33	88.04	41.84	42.21	19.70	2-BiGRU	
69.14	85.93	36.37	36.74	15.76	3-BiGRU	

جدول 5-2: تاثیر معماری‌های متفاوت در کدگشاھای بر پایه شبکه‌های عصبی بازگشتی

بهتری دارد و پیشنهاد می‌شود از این معماری استفاده کرد. به علاوه از بین مکانیزم‌های توجه، استفاده از مکانیزم باهدانا نتایج بهتری را از خود نشان داده اند. نکته‌ای که قابل ملاحظه و توجه است، استفاده از روش الحاقی در GRU و شبکه کدگذار VisualBERT مقدار قابل توجهی از معیارها را کاهش می‌دهد که می‌تواند نکته حائز اهمیت باشد و نیاز به مطالعه و پژوهش بیشتری دارد. اطلاعات دقیق این نتایج در جدول 5-3 گزارش شده است.

4-5-5 ارزیابی و مقایسه مدل‌های برپایه تبدیل‌شونده‌ها

مدل‌های تبدیل‌شونده اخیراً توانایی و قابلیت‌های زیادی از خود نشان داده اند به طوریکه در اکثر موارد جزء پیشروترین معماری مورد استفاده قرار گرفته‌اند که این خاصیت آن‌ها در کارهای زبان-تصویر نیز حائز اهمیت است. از این رو بخش

Transformers¹

Embedding-based		Word-based			Method	
Score BERT	Score Average	ROUGE-L	METEOR	BLEU	Decoder	Encoder
91.84	95.94	85.49	86.43	79.03	1-LSTM+Bahdanau attention	LXMERT
<u>91.94</u>	<u>96.20</u>	86.05	86.90	78.79	1-LSTM+Luong(dot) attention	
91.90	96.11	<u>86.25</u>	<u>86.96</u>	<u>79.54</u>	1-LSTM+Luong(general) attention	
89.32	95.42	82.62	83.26	71.40	1-GRU+Bahdanau attention	
90.37	95.73	83.84	84.71	73.92	1-GRU+Luong(dot) attention	
85.40	93.61	75.04	75.92	58.44	1-GRU+Luong(general) attention	
79.99	88.88	59.48	60.62	36.53	1-GRU+Luong(concat) attention	
<u>93.50</u>	97.11	<u>87.28</u>	<u>88.07</u>	<u>84.27</u>	1-LSTM+Bahdanau attention	
92.12	96.98	85.81	86.68	80.90	1-LSTM+Luong(dot) attention	
92.68	97.10	86.42	87.35	82.41	1-LSTM+Luong(general) attention	
93.11	<u>97.17</u>	86.90	87.71	82.90	1-LSTM+Luong(concat) attention	VisualBERT
89.26	96.10	82.40	82.87	72.20	1-GRU+Bahdanau attention	
91.81	96.93	85.23	86.17	79.65	1-GRU+Luong(dot) attention	
89.22	96.17	81.21	82.15	72.40	1-GRU+Luong(general) attention	
70.50	89.13	54.01	54.30	29.64	1-GRU+Luong(concat) attention	

جدول 5-3: بررسی شبکه‌های عصبی بازگشتی با مکانیزم توجه

اصلی این پژوهش استفاده از این معماری‌ها برای کدگشا است که نتایج این اجرایها در جدول 4 به صورت کامل آورده شده است. با توجه به جدول و همانطوری که انتظار می‌رفت، مدل‌های تبدیل‌شونده از دیگر مدل‌ها عملکرد بسیار بهتری داشتند و افزایش تعداد لایه‌ها تاثیر چندانی در بهبود آن نداشت.

Embedding-based		Word-based			Method	
BERT Score	Average Score	ROUGE-L	METEOR	BLEU	Decoder	Encoder
95.01	90.20	90.60	91.18	86.73	3-Transformer Decoder	LXMERT
94.79	90.13	90.33	90.91	85.98	4-Transformer Decoder	
94.44	91.94	89.09	89.76	85.95	3-Transformer Decoder	VisualBERT
<u>94.52</u>	<u>91.95</u>	<u>89.16</u>	<u>89.78</u>	<u>85.99</u>	4-Transformer Decoder	

جدول 5-4: استفاده از مدل‌های تبدیل‌شونده به عنوان کدگشا

6-5 ارزیابی انسانی

در این پژوهش علاوه بر معیارهای اندازه‌گیری یک ارزیابی انسانی نیز انجام شده است. برای انجام این مرحله، ۱۰۰ عنصر به صورت تصادفی از مجموعه داده انتخاب شد و بر روی سه مدل برتر برای هر کدام از معماری‌های توضیح داده شده در بخش 5 ارزیابی شده است. بررسی پاسخ‌های تولید شده توسط افراد، باعث شد پاسخ‌ها را بر اساس نوع اشتباہی که رخداده در 5 دسته دسته‌بندی کرد:

1. **تطابق دقیق (EM):** جملات پاسخ دقیقاً با جملات مرجع برابر هستند.

2. پاسخ اشتباه: (WA) پاسخ‌ها از لحاظ گرامری درست، اما از لحاظ منطقی اشتباه هستند.

3. اشتباه گرامری (GE): پاسخ‌ها از لحاظ گرامری اشتباه هستند، ولی پاسخ تک کلمه‌ای مرتبط با دیتا است را دارند.

4. توضیح اشتباه: (WD) پاسخ تک کلمه‌ای درست و از لحاظ گرامری نیز درست است، ولی توضیح پاسخ اشتباهاتی دارد.

5. پاسخ جایگزین: پاسخ‌ها از هر نظر درست هستن، ولی مطابقت دقیق با مرجع ندارند.

تعدادی از این مثال‌ها را می‌توان در جدول 5-5 مشاهده کرد.

تصویر	پرسش و پاسخ	پاسخ تولید شده	نوع خطأ
	Q: where is a banana? A: a banana sits between a full coffee cup and a granola bar.	LXMERT-3BiLSTM : a banana is on a table with a bowl of scissors. VisualBERT-BahdanauLSTM : a banana sits on a of a a. LXMERT-3Transformer : a banana sits on a table next to a table.	WD GE WD
	Q: what color is the umbrella? A : the umbrella is blue.	LXMERT-3BiLSTM : the umbrella is blue. VisualBERT-BahdanauLSTM : the umbrella is red. LXMERT-3Transformer : the umbrella is blue.	EM WA EM
	Q: what kind of appliance is this? A : this is refrigerator.	LXMERT-3BiLSTM : refrigerator is pictured. VisualBERT-BahdanauLSTM : this is blender. LXMERT-3Transformer : this is refrigerator.	AA E EM

جدول 5-5: چند نمونه از پاسخ‌ها و اشتباهات آن‌ها بر اساس دسته‌بندی

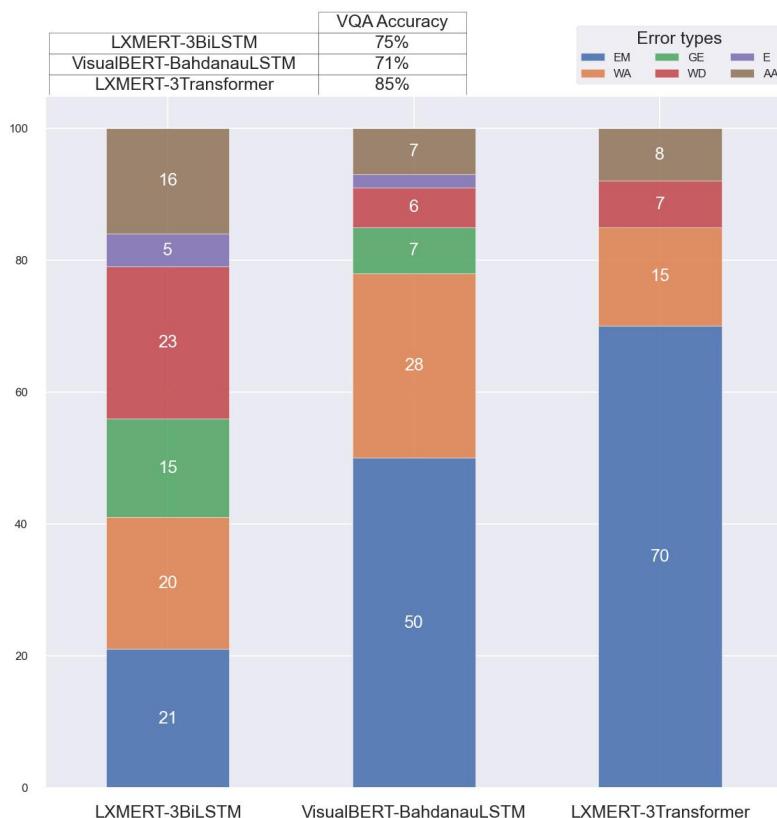
تحلیل این 100 عضو نشان داد که تولید پاسخ‌ها باعث کاهش ابهام می‌شود. به عبارت دیگر ارائه دادن توضیح اضافه بر کلمه پاسخ باعث می‌شود کاربر از اشتباهاتی که سیتم ممکن است منجر بشود را آشکار می‌کند. برای مثال فرض کنید پرسش و پاسخ‌های زیر تولید شده باشد:

Question : Is the dog baring its teeth?

Reference : No, the dog is not baring its teeth.

Generated : Yes, the dog is grillening its teeth.

از آنجایی که کلمه grillening کلمه‌ای متناسب با حیوانات نیست و حیوانات قادر به انجام آن نیستند، می‌توان از این توضیح اشتباه به درست بودن پاسخ نیز شک کرد. از این رو ارائه توضیح اضافی منجر به رفع ابهامات در پرسش‌وپاسخ تصویری می‌شود.



شکل 5-1: نتایج ارزیابی انسانی

شکل 5-1 نشان‌دهنده عملکرد خوب سیستم را نیز به خوبی نشان می‌دهد، به خصوص اینکه در حالت استفاده مدل‌های تبدیل‌شونده به عنوان کدگشا، هیچ اشتباه گرامی ندارد. یکی دیگر از نتایجی که می‌توان از این تحلیل‌ها گرفت، تاثیر کددگار است که با توجه به دقیق در پرسش‌وپاسخ تصویری، با حضور اشتباهات، همچنان مدلی که LXMERT را به عنوان کددگار استفاده کرده است، دارای دقیق‌تری از مدل‌های بر پایه VisualBERT است.

7-5 جمع‌بندی

با بررسی نتایج حاصل بر روی مجموعه داده FSVQA و سیستم‌های ارائه شده می‌توان نتیجه گرفت استفاده از مدل‌های تبدیل‌شونده به عنوان کدگذار برای تولید پاسخ در پرسش‌وپاسخ تصویری عملکرد بسیار خوبی از خود نشان داده است. علاوه بر عملکرد خوب سیتم بر مجموعه داده، با ارزیابی انسانی ثابت شد ارائه دادن توضیح اضافی هنگام پاسخ باعث کمک بیشتر برای فهم پاسخ و حتی در حالاتی منجر به رفع ابهامات می‌شود. با این حال همچنان مسیر بسیاری تا حل مسئله پرسش‌وپاسخ تصویری باقی‌مانده است. امتیازات بالایی که در این پژوهش بدست آمده‌اند در واقع به علت ساده بودن مجموعه داده است و لازم است پرسش‌هایی با پاسخ‌های طبیعی‌تر که از قوانین خاصی پیروی نمی‌کنند موجود باشند تا بتوان مقادیر واقعی این معیارها را در محیط‌های واقعی‌تر محاسبه کرد.

فصل ۶

نتیجه‌گیری و کارهای آینده

1-6 نتیجه‌گیری

در این پژوهه ابتدا روش‌های مختلف حل پرسش و پاسخ تصویری مورد تحلیل واقع شد. سپس روشی نو برای حل ان ارائه شد که به درک انسان از سوالات و پاسخ دادن به آن‌ها نزدیکتر است. تولید پاسخ‌ها بدون هیچ دانش خارجی و به صورت جمله در پاسخ به سوالات مربوط به یک تصویر نسبت به پاسخ به صورت تک‌کلمه روشی معممول‌تر و رایج‌تر است و حل این مسئله به این روش می‌تواند سیستم را به روش‌های انسانی نزدیک‌تر کند.

در هسته‌ی اصلی سیستم پیشنهادی مدلی با معماری کدگذار-کدگشا قرار دارد. قسمت کدگذار که عمدۀ وظیفه تحلیل داده ورودی را دارد از مدل‌های از پیش آموزش دیده در مسائل زبان-تصویر است که در این پژوهش از دو مدل VisualBERT و LXMERT استفاده شده است. مدل کدگشا نیز معماری‌های متفاوتی در نظر گرفته شد که مدل‌های تبدیل شونده بهترین عملکرد را داشتند و به دقت مناسبی بر مجموعه داده رسیدند.

برای ارزیابی هر معماری، چندین معیار در نظر گرفته شد که از ابعاد معنایی و نحوی مقادیر خروجی را ارزیابی می‌کنند. پس از ارزیابی برترین معماری‌ها، این نتیجه حاصل شد که استفاده از این معماری‌ها در مجموعه داده مدنظر مفید واقع شده و توانسته‌اند مسئله را با دقت خوبی حل کنند. سپس با روش‌های ارزیابی انسانی ثابت شد که این روش به رفع ابهامات در پرسش کمک بسیاری می‌کند.

در نظر گرفتن این نکته حائز اهمیت است که این دقت بالا در معیارها واقعی نیست زیرا مجموعه داده از داده‌های موجود در دنیای بیرون بسیار متفاوت‌تر است و به صورت خودکار تولید شده‌اند و حالت طبیعی خود را ندارند. لیکن این

فصل 6. نتیجه‌گیری و کارهای آینده

2-6. دستاوردها

به معنای حل کامل مسئله نیست و لازم به اعمال بسیاری برای ادعای حل کامل مسئله پرسش و پاسخ تصویری به صورت تولید متن است.

2-6 دستاوردها

در تمامی قسمت‌های این پژوهش اعم از قسمت‌های طراحی، پیاده‌سازی و ارزیابی نوآوری‌هایی مطرح شد که به شرح زیر است:

1. ارائه روشی نو برای حل مسئله پرسش و پاسخ تصویری که در صورت حل، روشی کامل‌تر از روش‌های فعلی است.
2. معرفی و استفاده از معماری‌های نو مناسب با روش پیشنهادی
3. اثبات درستی و فواید استفاده از این روش برای حل مسئله

3-6 کارهای آینده

همانطور که در قسمت ارزیابی مشاهده شد، با وجود اعداد و ارقام به نسبت بالا، همچنان نمی‌توان ادعا کرد که این مسئله حل شده است. همچنین معماری‌های اراسه شده به نسبت قدیمی هستند و روش‌های بسیار نوآورانه‌تری معرفی شده‌اند که می‌توانند این مسئله را با سادگی بیشتری حل کنند. چندی از این ایده‌ها که بهبود این مسئله کمک بسیاری می‌کنند در زیر آورده شده است.

1. ارائه مجموعه داده‌ای جامع و کامل، به طوری که پاسخ‌ها به صورت جمله و طبیعی باشند. اولین قدم برای بهبود این سیستم ارائه مجموعه داده‌ی کامل‌تری است که بتوان معماری‌های پیچیده‌تر را بر آن آموزش داد و همچنین به نتایج بدست آمده و نزدیک بودن آن به دنیای واقعی اطمینان بیشتری داشت.
2. استفاده از کدگشاها از پیش آموزش داده شده، یکی دیگر از راه‌های بهبود حل، ارائه معماری‌هایی غنی‌تر با پیچیدگی بالاتر است که قادر به حل مسئله به صورت جامع‌تر باشند. از این معماری‌ها می‌توان استفاده از مدل‌هایی نظیر BART و یا GPT را نام برد. استفاده از این قبیل معماری‌ها باعث عمومی‌سازی بیشتر سیستم می‌شود.

واژه‌نامه فارسی به انگلیسی

Visual Question Answering	پرسش و پاسخ تصویری
Deep Learning	یادگیری عمیق
Attention	توجه
Transfer Learning	یادگیری انتقالی
Error	خطا
Specification	مشخصات
Representation Learning	یادگیری بازنمایی
Recurrent Neural Network	شبکه عصبی بازگشتی
Encoder	کدگذار
Decoder	کدگشا
Sequence-to-Sequence	دنباله به دنباله
End-to-End	انتها به انتها
Transformer	تبديل‌شونده
Tokenization	نشانه‌گذاری
Embedding	تعبیه
Key	کلید
Query	پرسش
Value	مقدار
Multi-Head Attention	توجه چندسر
Masked Language Model	مدل زبانی ماسک‌دار
Forward Propagation	گسترش رو به جلو
Gradient Decent	گرادیان کاهشی
Self-attention	توجه به خود
Single-stream model	مدل تک‌جريان
Dual-stream model	مدل دو‌جريان
Word-similarity	شباخت نحوی

ارزیابی انسانی Human Evaluation

واژه‌نامه انگلیسی به فارسی

پرسش و پاسخ تصویری	Visual Question Answering
یادگیری عمیق	Deep Learning
توجه	Attention
یادگیری انتقالی	Transfer Learning
خطا	Error
مشخصات	Specification
یادگیری بازنمایی	Representation Learning
شبکه عصبی بازگشتی	Recurrent Neural Network
کدگذار	Encoder
کدگشا	Decoder
دبالت به دنباله	Sequence-to-Sequence
انتها به انتها	End-to-End
تبدیل‌شونده	Transformer
نشانه‌گذاری	Tokenization
تعابیه	Embedding
کلید	Key
پرسش	Query
مقدار	Value
توجه چندسر	Multi-Head Attention
مدل زبانی ماسکدار	Masked Language Model
گسترش رو به جلو	Forward Propagation
گرادیان کاهشی	Gradient Decent
توجه به خود	Self-attention
مدل تک‌جریان	Single-stream model
مدل دو‌جریان	Dual-stream model
شباخت نحوی	Word-similarity

..... ارزیابی انسانی Human Evaluation

Abstract

Visual Question Answering is a multi-modal task under the consideration of both the Vision and Language communities. Present VQA models are limited to classification answers and cannot provide answers for reasoning questions. In this work, we introduce an encoder-decoder model using vision-and-language pre-trained embedding, which delivers multi-word generated sentences as answers. We utilise LXMERT and VisualBERT embedding space with three different generative decoder heads, including RNNs, Attention RNNs and Transformers. Extensive experiments show competitive performance on the FSVQA dataset through qualitative and quantitative evaluation and a Human Error Analysis.

Keywords : Visual Question Answering, Natural Language Generation, NLG, VQA



Iran University of Science and Technology
Computer Engineering Department

Generate Answer to Visual Questions with Pre-trained Vision-and-Language Embeddings

Bachelor of Science Thesis in Computer Engineering - Artificial Intelligence

By:

Hadi Sheikhi

Supervisor:

Dr. Sauleh Eetemadi

Feb 2023