پاسخ‌دهی خودکار به پرسش در سیستم پرسش و پاسخ بصری با استفاده از شبکه عصبی کانولوشنی ResNeXt و شبکه عصبی بازگشتی GRU دوسویه

امیر شکری [۱۱]، علیرضا غلام‌نیا [۲]

۱- دانش‌آموخته کارشناسی‌ارشد هوش‌مصنوعی، دانشگاه سمنان، amirsh.nll@gmail.com

۲- دانشجو کارشناسی‌ارشد هوش مصنوعی، دانشگاه سمنان، gholamniareza@gmail.com

چکیده

سیستم پرسش و پاسخ بصری از دو بخش کلی پردازش تصویر و پردازش متن تشکیل شده است. در این سیستم‌ها یک پرسش در مورد تصویر توسط کاربر مطرح شده و سیستم وظیفه دارد با توجه به تصویر پاسخ را پیش‌بینی کرده و به کاربر نمایش دهد. عمده تلاش محققان افزایش دقت صحت پاسخ پیش‌بینی شده در این سیستم‌ها است. از این رو در این مقاله نیز ما یک سیستم پرسش و پاسخ بصری معرفی و طراحی کردیم که برای پردازش تصویر ورودی از شبکه عصبی کانولوشنی با معماری ResNeXt با ۱۰۱ لایه و برای پردازش متن از شبکه عصبی بازگشتی از نوع GRU دوسویه استفاده می‌کند. همچنین از مدل زبانی Glove جهت تعبیه‌سازی کلمات متن ورودی استفاده خواهد شد. جهت نتیجه‌گیری بهتر در مدل پیشنهادی از سازوکار توجه چندسر نیز استفاده شده است. برای ارزیابی مدل پیشنهادی از هر دو نسخه مجموعه‌داده VQA، یعنی VQA1.0 و VQA2.0 استفاده می‌شود.

کلمات کلیدی: پرسش و پاسخ بصری، شبکه عصبی کانولوشنی، شبکه عصبی بازگشتی GRU دوسویه، مجموعه‌داده VQA، سازوکار توجه چندسر.

Email: amirsh.nll@gmail.com

۱. مقدمه

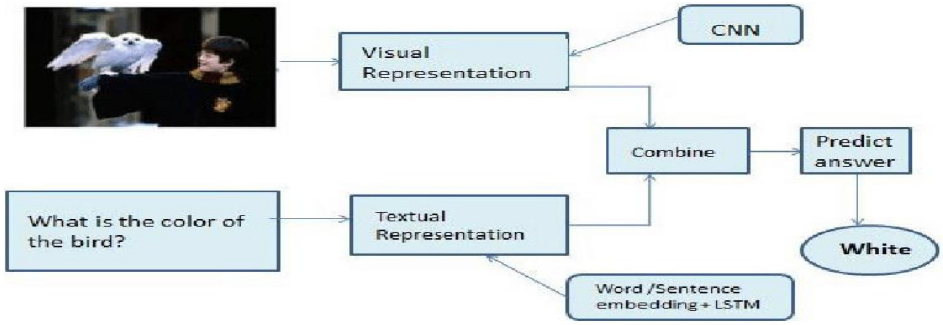
امروزه انسان‌ها بدنبال راهکارها و تکنولوژی‌های مختلفی جهت ارتباط و تعامل آسان‌تر با کامپیوتر و ماشین هستند. سیستم‌های پرسش و پاسخ بصری می‌تواند موجب تسهیل این ارتباط شود. این سیستم‌ها در حوزه‌های مختلفی کاربرد دارد. از این سیستم‌ها در صنایع مختلف و بخصوص در پزشکی می‌توان استفاده کرد. در این سیستم‌ها یک تصویر و یک پرسش مرتبط با تصویر به عنوان ورودی دریافت شده و سیستم وظیفه دارد تا پاسخ به پرسش مطرح شده را پیش‌بینی کند. پرسش و پاسخ بصری از دو بخش پردازش زبان طبیعی و بینایی کامپیوتر تشکیل شده است و یک مبحث میان‌رشته‌ای محبوب به‌حساب می‌آید[۱].

چالش اصلی در سیستم‌های پرسش و پاسخ بصری، دقت درستی پاسخ پیش‌بینی شده است. تلاش تمام محققان و پژوهشگران در این حوزه طراحی سیستمی است که بیشترین درصد دقت درستی را در پاسخ پیش‌بینی شده داشته باشد. برای رسیدن به این هدف باید شبکه‌های عصبی مناسبی جهت پردازش ورودی‌های سیستم انتخاب شوند و به روشی مناسبی خروجی شبکه‌های عصبی انتخاب شده، با یکدیگر ادغام شوند. در بیش‌تر پژوهش‌های انجام شده برای پردازش تصویر ورودی از معمارهای مختلف شبکه عصبی کانولوشنی[۲] استفاده می‌شود. ساختار شبکه عصبی کانولوشنی به گونه‌ای است که برای پردازش تصویر بسیار مناسب می‌باشد. از این رو محققان برای پردازش هر چه بهتر تصویر از معماری‌های مختلف این شبکه استفاده می‌کنند. همچنین برای پردازش متن ورودی از انواع مختلف شبکه‌های عصبی بازگشتی[۳] استفاده می‌شود[۲].

استفاده از مکانیزم‌های مختلف مانند انواع مکانیزم توجه[۴] نیز باعث می‌شود تا دقت درستی در سیستم‌های پرسش و پاسخ بصری افزایش پیدا کند. از انواع مختلف مکانیزم توجه می‌توان در قسمت پردازش تصویر و در قسمت پردازش متن استفاده کرد. از آنجایی که برای پاسخ‌دهی به پرسش مطرح شده معمولا نیاز به اطلاعات قسمتی از تصویر وجود دارد، از این رو می‌توان توجه بیشتری به قسمتی از تصویر که مرتبط با پرسش است، داشت. این امر باعث می‌شود تا در پردازش تصویر قسمت پراهمیت وزن بیشتری داشته باشد و در نتیجه حاصل شده، اثرگذاری بیشتری نسبت به بخش‌های دیگر تصویر داشته باشد. مکانیزم توجه دقیقا همین‌کار را انجام می‌دهد و با مشخص کردن قسمت‌هایی از تصویر که مرتبط با سوال است، به آن‌ها وزن بیشتری داده و باعث بهبود دقت درستی پاسخ می‌شود. از همین رو از انواع مختلف مکانیزم توجه در پژوهش‌های مرتبط در این حوزه استفاده شده است.

استفاده از مجموعه‌داده[۵] مناسب نیز تاثیر چشم‌گیری در دقت درستی پاسخ پیش‌بینی شده دارد. مجموعه‌داده انتخاب شده جهت آموزش مدل، بایستی بزرگ و مناسب باشد. استفاده از مجموعه‌داده نامناسب و کوچک در آموزش مدل پرسش و پاسخ بصری موجب پایین آمدن دقت درستی پاسخ پیش‌بینی شده خواهد شد. از این رو در انتخاب مجموعه‌داده مناسب باید بسیار دقت کرد. در شکل ۱، شماتیک کلی یک سیستم پرسش و پاسخ بصری نشان داده شده است.

[2] Convolutional Neural Network
[3] Recurrent Neural Network
[4] Attention Mechanism
[5] Dataset



شکل ۱: شماتیک کلی سیستم پرسش و پاسخ بصری

۲. کارهای مرتبط

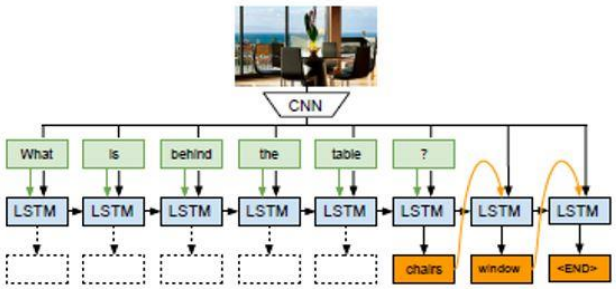
در حوزه‌ی پرسش و پاسخ بصری پژوهش‌های مختلفی انجام شده است که در ادامه به بررسی برخی از آن‌ها خواهیم پرداخت.

یانگ و همکاران [۳]، یک سیستم پرسش و پاسخ بصری پیشنهاد کرده‌اند. در سیستم پیشنهادی آن‌ها، برای پردازش متن از شبکه عصبی حافظه کوتاه مدت طولانی دوسویه استفاده می‌شود. همچنین برای پردازش تصویر ورودی از شبکه عصبی کانولوشنی استفاده شده است. در این مقاله همچنین از مکانیزم توجه Co-Attention که از یک مکانیزم توجه‌به‌خود[۶] و یک مکانیزم توجه تصویری تشکیل شده، استفاده شده است. در ابتدا در این مدل کلمات موجود در متن ورودی تعبیه‌سازی[۷] شده و جهت پردازش به شبکه عصبی بازگشتی حافظه کوتاه مدت طولانی داده می‌شود. تصویر ورودی نیز به شبکه عصبی Faster R-CNN داده شده و تصویر پردازش می‌شود. خروجی شبکه عصبی حافظه کوتاه مدت طولانی و شبکه عصبی Faster R-CNN به مکانیزم Co-Attention داده می‌شود، سپس خروجی شبکه حافظه کوتاه مدت طولانی دوسویه[۸] و مکانیزم Co-Attention با یکدیگر جمع می‌شوند. حاصل این دو جمع و نوع سوال که از قبل مشخص شده، در مرحله بعدی با یکدیگر ترکیب می‌شود و نتیجه آن جهت پیش‌بینی پاسخ نهایی به یک طبقه‌بندی کننده داده می‌شود. این مدل در مجموعه داده VQA1.0 به دقت ۶۶/۵۳ درصد و در مجموعه‌داده VQA2.0 به دقت ۶۵/۸۰ درصد برای درستی پاسخ پیش‌بینی شده رسیده است.

مالینوفسکی و همکاران [۴] یک سیستم پرسش و پاسخ بصری پیشنهاد کرده‌اند. در مدل پیشنهادی این مقاله، تصویر ورودی به یک شبکه عصبی کانولوشنی با معماری GoogleNet [۵] جهت پردازش داده شده و بازنمایی تصویر بدست می‌آید. در مرحله بعد، بازنمایی بدست آمده برای تصویر و متن ورودی به یک شبکه عصبی حافظه کوتاه مدت طولانی داده شده و خروجی این شبکه، یک بردار ویژگی با اندازه ثابت خواهد بود که به عنوان ورودی به شبکه عصبی حافظه کوتاه مدت طولانی بعدی داده می‌شود. طول پاسخ تولید شده نهائی توسط این شبکه متغیر خواهد بود و به ازای هر بار تکرار یک کلمه از پاسخ تولید می‌شود. در این مدل آخرین کلمه پیش‌بینی شده در هر مرحله، به عنوان ورودی به شبکه حافظه کوتاه مدت طولانی داده می‌شود تا در پیش‌بینی باقی پاسخ مود استفاده قرار گیرد. در نهایت با پیش‌بینی نماد خاص <END> فرآیند

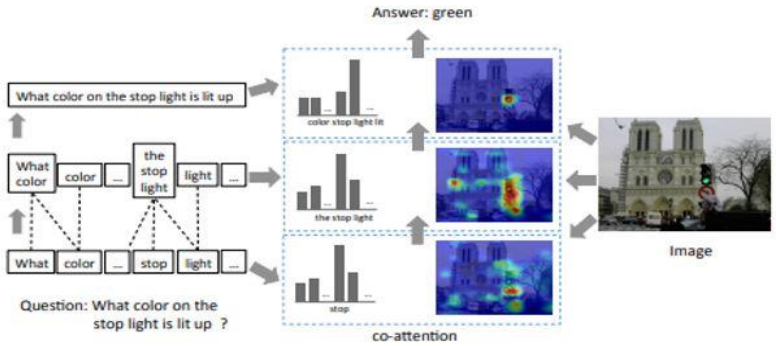
[6] Self-Attention
[7] Word Embedding
[8] Long Short Term Memory

پیش‌بینی پاسخ به اتمام می‌رسد. این نماد زمانی پیش‌بینی می‌شود که تمام پاسخ‌های مرتبط بدست آمده باشد. در شکل ۲، ساختار مدل پیشنهادی در مقاله [۴] مشخص شده است.



شکل ۲: ساختار مدل پیشنهادی پرسش و پاسخ بصری در مقاله [۵]

لی و همکاران [۶] یک سیستم پرسش و پاسخ بصری به نام Hi-co attention معرفی کرده‌اند. مدل پیشنهادی بر پایه مکانیزم توجه بوده و از سه مرحله تعبیه‌گذاری بر روی متن تشکیل شده است. ابتدا تعبیه‌گذاری در سطح کلمه، سپس در سطح عبارت و سپس در سطح جمله انجام می‌شود. تعبیه‌گذاری در سطح کل جمله با استفاده از تعبیه‌گذاری مرحله قبل یعنی تعبیه‌گذاری در سطح عبارت و شبکه عصبی حافظه کوتاه مدت طولانی بدست می‌آید. برای پردازش تصویر در مدل پیشنهادی از شبکه عصبی کانولوشنی استفاده شده است. مکانیزم توجه مورد استفاده در این مدل Co-attention parallel است که در هر سه سطح بازنمایی پرسش استفاده می‌شود و مقدار شباهت بین تصویر و پرسش را محاسبه کرده و موجب متصل شدن آن‌ها به یکدیگر می‌شود. با ترکیب ویژگی‌های خروجی مکانیزم توجه و تابع بیشینه هموار، جواب نهائی پیش‌بینی می‌شود. در شکل ۳، ساختار مدل پیشنهادی در مقاله [۶] نشان داده شده است.



شکل ۳: ساختار مدل پیشنهادی سیستم پرسش و پاسخ بصری در مقاله [۶]

۳. مجموعه‌داده

همانگونه که گفته شد یکی از مهم‌ترین مراحل در طراحی سیستم‌های پرسش و پاسخ بصری آموزش مدل با مجموعه‌داده مناسب و بزرگ است. مجموعه‌داده‌های مختلفی در این حوزه وجود دارد که با مطالعه پژوهش‌های پیشین ما از مجموعه‌داده VQA [۷] استفاده خواهیم کرد. این مجموعه‌داده در دو نسخه VQA1.0 و VQA2.0 منتشر شده است، که در این مقاله از هر دو نسخه جهت ارزیابی مدل پیشنهادی استفاده شده است.

۳-۱. پیش‌پردازش مجموعه‌داده

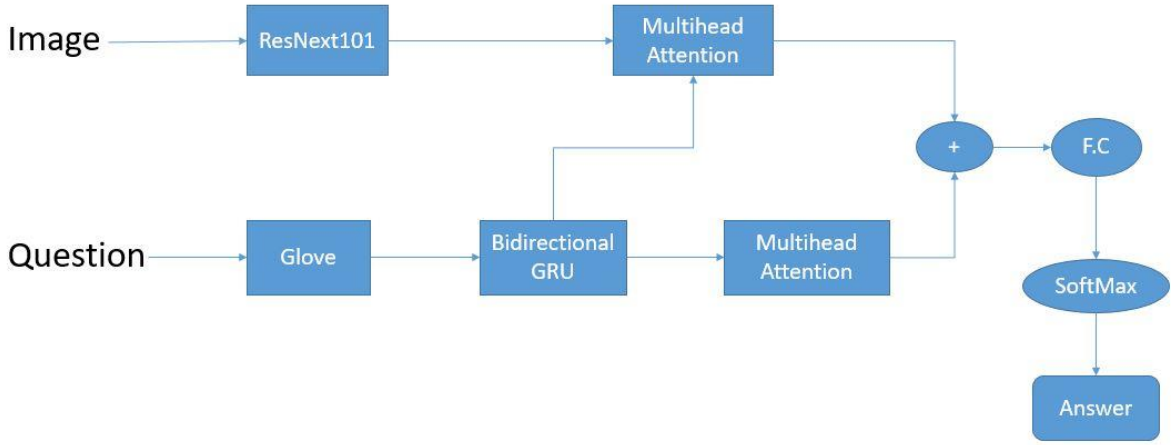
در ابتدا تمام کلمات و پرسش‌های موجود در هر دو نسخه مجموعه‌داده مورد استفاده Tokenize شده و طول سوالات ۱۴ در نظر گرفته می‌شود (حداکثر طول سوالات). جهت تعبیه‌سازی کلمات موجود در پرسش ورودی از مدل زبانی [۸] Glove استفاده خواهد شد. در مرحله بعد تصاویر را تغییر اندازه داده و به ۲۲۴*۲۲۴ تبدیل می‌کنیم و آن را به شبکه عصبی کانولوشنی مورد نظر می‌دهیم. در این شبکه تصاویر پردازش می‌شود و با توجه به کانولوشن چهارم خروجی بدست می‌آید.

۴. روش پیشنهادی

در ابتدا کلمات متن اصلی با استفاده از مدل زبانی[۸] Glove تعبیه‌سازی می‌شود. خروجی بدست آمده از این تعبیه‌سازی به یک شبکه عصبی بازگشتی از نوع GRU دوسویه[۹] داده می‌شود و بازنمایی پرسش ورودی بدست می‌آید. تصویر ورودی نیز به یک شبکه عصبی کانولوشنی با معماری[۱۰] ResNeXt با ۱۰۱ لایه داده می‌شود و ویژگی‌های تصویر استخراج شده و بازنمایی آن بدست می‌آید. خروجی در این شبکه ۲۰۴۸*۱۴*۱۴ است.

در مرحله بعد بازنمایی تصویر و بازنمایی بدست آمده پرسش به عنوان ورودی به یک مکانیزم توجه چندسر[۹] [۱۱] داده می‌شود، همچنین به طور موازی خروجی شبکه عصبی بازگشتی مورد استفاده نیز به یک مکانیزم توجه چندسر دیگر داده می‌شود. در نهایت خروجی‌های بدست آمده از دو مکانیزم توجه چند سر با هم جمع شده و با استفاده از لایه تماما متصل و تابع بیشینه هموار[۱۰] [۱۲] پاسخ نهائی پیش‌بینی می‌شود. در شکل ۴، ساختار مدل پیشنهادی پرسش و پاسخ بصری در این مقاله نشان داده شده است.

[9] Multihead Attention
[10] Softmax



شکل ۴: ساختار مدل پیشنهادی پرسش و پاسخ بصری

۵. پیاده‌سازی روش پیشنهادی

۵-۱. محیط پیاده‌سازی

زبان برنامه‌نویسی مورد استفاده برای پیاده‌سازی روش پیشنهادی، زبان برنامه‌نویسی پرکاربرد پایتون است. از کتابخانه‌های[۱۳] Tensorflow و Keras نیز در این پیاده‌سازی استفاده شده است. همچنین جهت پیاده‌سازی مدل پیشنهادی از نسخه گوگل کلب پرو[۱۱] استفاده شده است. حجم محاسبات در این پروژه سنگین است و نسخه معمولی گوگل کلب بدلیل محدودیت زمانی که دارد، نمی‌تواند مثمرثمر باشد. از همین رو از نسخه پرو آن استفاده کردیم. در گوگل کلب پرو، GPU مورد استفاده K 80 است. همچنین مقدار RAM مورد استفاده برای پیاده‌سازی، آموزش و تست روش پیشنهادی برابر با 25 گیگابایت است.

۵-۲. جزئیات پیاده‌سازی

در پیاده‌سازی مدل پیشنهادی از الگوریتم بهینه‌ساز ADAM [۱۴] با نرخ یادگیری[۱۲] ۰,۰۰۰۱ استفاده شده است. اندازه دسته[۱۳] ۶۴ در نظر گرفته شده و آموزش شبکه در ۸۰ دوره انجام می‌شود. همچنین از نرمالسازی دسته‌ای[۱۴] و Dropout با نرخ ۰,۳ استفاده شده است.

[11] Google Colab Pro
[12] Learning Rate
[13] Batch size
[14] Batch Normalization

۶. نتایج بدست آمده مدل پیشنهادی

پرسش‌هایی که در مدل پیشنهادی مورد نظر است به سه دسته زیر تقسیم می‌شوند:

۱- پرسش‌هایی که پاسخ آن‌ها بله یا خیر است.

۲- پرسش‌هایی که جواب آن‌ها اعداد هستند.

۳- پرسش‌هایی که پاسخ آن‌ها غیره (پاسخ‌های یک کلمه‌ای و دو کلمه‌ای جز دو مورد بالا) است.

در جدول ۱ و ۲، نتایج حاصل شده از مدل پیشنهادی در دو مجموعه‌داده مورد استفاده نشان داده شده است.

جدول ۱: نتایج مدل پیشنهادی با استفاده از مجموعه‌داده VQA1.0

نوع پاسخ	بله/خیر	اعداد	غیره	دقت کل
دقت پیش بینی پاسخ (درصد)	۸۶/۵	۴۴/۸	۵۷/۴	۶۷/۲

جدول ۲: نتایج مدل پیشنهادی با استفاده از مجموعه‌داده VQA2.0

نوع پاسخ	بله/خیر	اعداد	غیره	دقت کل
دقت پیش بینی پاسخ (درصد)	۸۱/۳	۴۱/۹	۵۲/۸	۶۳/۳

7. ارزیابی مدل پیشنهادی

در این مقاله یک معماری جدید برای پرسش و پاسخ تصویری با استفاده از دو مجموعه‌داده VQA1.0 و VQA2.0 معرفی شده است. در جدول ۳، به بررسی و مقایسه نتایج حاصل شده در مدل پیشنهادی و نتایج دیگر پژوهش‌ها در حوزه پرسش و پاسخ بصری با استفاده از مجموعه‌داده VQA1.0 پرداخته شده است. همچنین در جدول ۴، به بررسی و مقایسه نتایج حاصل شده در مدل پیشنهادی و نتایج دیگر پژوهش‌ها در حوزه پرسش و پاسخ بصری با استفاده از مجموعه‌داده VQA2.0 پرداخته شده است.

جدول ۳: نتایج مدل پیشنهادی در این مقاله و دیگر مدل‌های پرسش و پاسخ تصویری با استفاده از مجموعه‌داده VQA1.0

	نوع پاسخ			
روش	بله/ خیر	اعداد	غیره	دقت کل
DMN+ [۱۵]	۸۰/۵	۳۶/۸	۴۸/۳	۶۰/۴
NMN [۱۶]	۷۷/۷	۳۷/۲	۳۹/۳	۵۵/۱
MCB [۱۷]	۸۲/۲	۳۷/۷	۵۴/۸	۶۴/۲

SAN [۱۸]	۷۹/۳	۳۶/۶	۴۶/۱	۵۸/۷
AYN [۱۹]	۷۸/۴	۳۶/۴	۴۶/۳	۵۸/۴
HieCoAttention [۲۰]	۷۹/۷	۳۸/۷	۵۱/۷	۶۱/۸
MRN [۲۱]	۸۲/۳	۳۸/۲	۴۹/۴	۶۱/۸
مدل پیشنهادی	**۸۶/۵**	**۴۴/۸**	**۵۷/۴**	**۶۷/۲**

جدول ۴: نتایج مدل پیشنهادی در این مقاله و دیگر مدل‌های پرسش و پاسخ تصویری با استفاده از مجموعه‌داده VQA2.0

	نوع پاسخ			
روش	بله/ خیر	اعداد	غیره	دقت کل
MCB [۱۷]	۸۱	۳۵/۸	۴۵/۱	۵۹/۱۴
ConceptBert [۲۲]	۸۳/۹	**۵۵/۲**	۷۰/۵	۶۹/۹
HieCoAttention [۲۰]	۸۰/۲	۳۷/۱	۴۲/۸	۵۴/۵۷
VilBert [۲۳]	۸۲/۶	۵۴/۳	۶۷/۱	۶۷/۹
In Defense of Grid Features [۲۴]	**۸۶/۴**	۵۵/۱	**۷۴/۲**	**۷۲/۷**
مدل پیشنهادی	۸۱/۳	۴۱/۹	۵۲/۸	۶۳/۳

۸. نتیجه‌گیری

در این مقاله یک سیستم پرسش و پاسخ بصری جدید طراحی کردیم. در سیستم پیشنهادی این مقاله از یک شبکه عصبی کانولوشنی با معماری ResNeXt با ۱۰۱ لایه برای پردازش تصویر استفاده کردیم. همچنین از یک شبکه عصبی بازگشتی GRU دوسویه جهت پردازش متن ورودی بهره بردیم. برای تعبیه‌سازی کلمات متن ورودی از مدل زبانی Glove استفاده کردیم. همچنین برای بدست آمدن نتیجه بهتر و کاراتر بودن مدل پیشنهادی از دو مکانیزم توجه Multihead Attention در مدل پیشنهادی استفاده کردیم. برای ارزیابی مدل پیشنهادی از دو نسخه مجموعه‌داده VQA یعنی VQA1.0 و VQA2.0 استفاده شده است. نتایج بدست آمده نشان‌دهنده‌ی این موضوع است که مدل پیشنهادی این مقاله با استفاده از مجموعه‌داده VQA1.0 بیشترین میزان دقت درستی پاسخ پیش‌بینی شده را نسبت به پژوهش‌های پیشین پرسش و پاسخ بصری که از مجموعه‌داده VQA1.0 استفاده کرده‌اند، بدست آورده است. اما نتایج بدست آمده برای مدل پیشنهادی مقاله با استفاده از مجموعه‌داده VQA2.0 نشان‌دهنده‌ی این است که تعدادی از پژوهش‌های دیگر در حوزه‌ی پرسش و پاسخ بصری که از مجموعه‌داده VQA2.0 استفاده کرده‌اند، میزان دقت بیشتری را دارا می‌باشند.

۹. مراجع

۱) Andreas,J. Rohrbach,M. Darrell,T. (2016). “Learning to compose neural networks for question answering”. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 760-772.

۲) Kafle,K and Kanan,C. (2016). “Answer-Type Prediction for Visual Question Answering.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4976-4984 .

۳) Yang,C. Jiang,M. Jiang,B. Zhou,W. (2016). “Co-Attention Network With Question Type for Visual Question Answering ”, Conference on Computer Vision and Pattern Recognition (CVPR), pp.4832-4841.

۴) Malinowski,M and Fritz,M. (2014). “A multi-world approach to question answering about real-world scenes based on uncertain input.” In Proc. Advances in Neural Inf. Process. Syst, pages1682–1690.

۵) Szegedy,C. Liu,W. Jia,Y. Sermanet,P and Rabinovich, A. (2015). “Going deeper with convolutions”. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

۶) Lu,J. Yang,J. Batra,D and Parikh,D. (2016). “Hierarchical question-image co-attention for visual question answering”. arXiv preprint arXiv: 1606.00061.

۷) Ren,M. Kiros,R and Zemel,R. (2015). “Image Question Answering: A Visual Semantic Embedding Model and a New Dataset”. In Proc. Advances in Neural Inf. Process. Syst . pages1552-1564.

۸) Pennington,J. Socher,R and Manning,C. (2014). “Glove: Global vectors for word representation.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

۹) Bhuvaneswari, A. Thomas, J. Kesavan, P. (2019). “Embedded Bi-directional GRU and LSTM Learning Models to Predict Disasterson Twitter Data”. International conference on recent trends in advanced computing(ICRTAC).165, 511-516.

۱۰) xie,S. Gitshick,R. Dollar,P. Tu,Z and He,K. (2017). “Aggregated Residual Transformations for Deep Neural Networks”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

۱۱) Niu,Z. Zhong,G. Yiu,H. (2021). “ A review on the attention mechanism of deep learning”, Neurocomputing, Volume 452, Pages 48-62.

۱۲) Alabassy,B and Safar,M. (2016). “A High-Accuracy Implementation for Softmax Layer in Deep Neural Networks.” In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 335–340.

۱۳) Abadi,M and et al. (2016). “Tensorflow: A system for large-scale machine learning”. in 12th {USENIX} symposium on operating systems design and implementation ({OSDI} .

۱۴) Kingma,D and Ba,J. (2014). “Adam: A method for stochastic optimization,” International Conference on Learning Representations.

۱۵) Xiong,C. Merity,S. and Socher,R. (2016). “Dynamic memory networks for visual and textual question answering”. In ICML.

۱۶) Andreas,J. Rohrbach,M. Darrell,T and Klein,D. (2016). “Deep Compositional Question Answering with Neural Module Networks” . In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

۱۷) Fukui,D. Park,H. Darrell,T and Rohrbach,M. (2016). “Multimodal compact bilinear pooling for visual question answering and visual grounding.” In EMNLP.

۱۸) Yang,Z and et al. (2015). “Stacked attention networks for image question answering.” arXiv preprint arXiv: 1511.02274.

۱۹) Malinowski,M. Rohrbach,M and Fritz,M. (2016). “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering.” arXiv preprint arXiv: 1605.02697.

۲۰) Lu,J. Yang,J. Batra,D and Parikh,D. (2016). “Hierarchical Co-Attention for Visual Question Answering.” In Advances in Neural Information Processing Systems (NIPS).

۲۱) Kim,J. Lee,W. Kwak,D. Kim, J and Zhang,T. (2016). “Multimodal residual learning for visual qa.” In Advances in Neural Information Processing Systems, pages361–369.

۲۲) Garderes,F and et al. (2020). “ConceptBert: Concept-Aware Representation for Visual Question Answering.” Findings of the Association for Computational Linguistics: EMNLP.

۲۳) Lu,J. Batra,D. Parikh,D and Lee,S. (2019). “Vilbert: Pretraining task-agnosticvisiolinguistic representations for vision-and-language tasks. 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, pp.155-167.

۲۴) Jiang,H and et al. (2020). “In Defense of Grid Features for Visual Question Answering.” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

۲۵) Conlin,R. Erickson,K. Abbate,J. Koleman,K. (2020). “Keras2c: A library for converting Keras neural networks to real-time compatible C”. Engineering Applications of Artificial Intelligence100: 2140 .

۲۶) Gholamalinezhad, H and Khosravi,H. (2020). “Pooling Methods in Deep Neural Networks, a Review”. arXiv preprint arXiv: 2009.07485.

۲۷) Szandała,T. (2020). “Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks”. international Conference on Dependability and Complex Systems, pp.498–505.

Visual question answering based on ResNeXt and Bidirectional GRU neural network

Amir Shokri, Alireza Gholamnia
amirsh.nll@gmail.com – gholamniareza@gmail.com

Abstract

This paper introduces a visual question-and-answer (VQA) system that utilizes a convolutional neural network (CNN) with ResNeXt architecture, along with a two-way GRU-type recurrent neural network (RNN) for text processing. To enhance the accuracy of predicted answers, the Glove language model is employed to embed input text words, and a multi-headed attention mechanism is incorporated into the proposed model. Both versions of the VQA dataset, VQA1.0 and VQA2.0, are employed to evaluate the performance of the proposed system. By combining image and text processing, the VQA system can effectively predict answers to user queries about images. The proposed system shows promising results and provides a foundation for further improvements in VQA research.

Keywords: visual question answering, convolutional neural networks, bidirectional gated recurrent unit neural network, visual question answering dataset, multi-head attention mechanism